Interpretable Artificial Intelligence and Algorithmic Decision Making Processes to Analyze User Behaviors and Their Socio-Economic Impacts

By

Ehsan Gholami
Dissertation

Submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

Ashwin Aravindakshan, Chair

Chen-Nee Chuah

Yong Jae Lee

Committee in Charge

2022

Contents

List of Figures

List of Tables

Abstract

Interpretable Artificial Intelligence and Algorithmic Decision Making Processes to Analyze User Behaviors and Their Socio-Economic Impacts

Understanding and predicting users' behavior is crucial to decision-making processes in various applications. With the growing scale of the user population in numerous domains, autonomous approaches like Artificial Intelligence (AI) and algorithmic designs have become an essential part of many user behavior analyses. Intelligent systems outperform humans on specific tasks [1] and facilitate human decisions by improving their perception. For practical, social, and in many cases, legal reasons, the adoption of intelligent systems in many domains hinges on their ability to provide explanations to developers, users, and regulators. Example domains are financial and e-commerce services, medical diagnosis, social media, and government policies. Such explanations increase users' trust in *why* and *how* decisions are made [2, 3], and enable their designers to increase robustness and fairness of the systems and reduce bias and discrimination [4, 5].

Data-driven systems often use logs of captured human interactions with applications, services, and systems for their user behavior analyses. These records represent naturalistic human behavior uninfluenced by observers and typically include many kinds of users worldwide. However, they usually provide information about *what* people do and less about *why* they do so. Some data-driven AI techniques like neural networks remain primarily black boxes despite their performance achievement. Designing a model capable of simultaneously achieving high performance in prediction tasks and explaining the underlying reasoning of those predictions is desirable.

This dissertation aims to design predictive models that bridge the gap between performance in predicting user behavior and the interpretability of the model in multiple application domains. We investigate several data-driven approaches, from stochastic models to neural networks. We propose effective methods that offer explanations along with performance and discuss their limitations.

The domains studied in this dissertation are:

- First, we look at the case of recommender systems. They play critical roles by offering items of interest to the users, thereby narrowing down a vast search space that comprises hundreds of thousands of products. We propose an architecture that relies on common patterns as well as individual behaviors to tailor its recommendations for each person. Simulations under a controlled environment show that our proposed model learns *interpretable* personalized user behaviors. Our empirical results on Nielsen Consumer Panel dataset indicate that the proposed approach achieves up to 27.9% performance improvement compared to the state-of-the-art.

- Second, we investigate how human behavior in the presence of Non-Pharmaceutical Interventions (NPI) (e.g., limited public gatherings and masks) impacts the spread of a contagious disease. A deeper understanding of the policy effects on human behavior and, subsequently, disease containment allows a more accurate forecast of disease spread when NPIs are partially loosened and provides policymakers with better data for making informed decisions. We adapt the Susceptible–Exposed–Infected–Recovered (SEIR) model for disease propagation in a network of interconnecting humans to incorporate human movement and social distancing. Even though the proposed SEIR model can estimate the disease propagation accurately, it singly does not explain the effect of NPIs on the disease spread. We measure the impact of NPIs on human behavior and, therefore, on mitigating COVID-19 spread by exploiting the spatio-temporal variations in policy measures across the 16 states of Germany. Our model finds that German policies that mandated contact restrictions (e.g., limited movement in public space) and closure of educational institutions are associated with the sharpest drops in movement within and across states. While this quasi-experiment does not allow for causal identification, each policy's effect on reducing disease spread provides meaningful insights. By combining the SEIR model with a model that measures the policy contributions to mobility reduction, we forecast scenarios for relaxing various types of NPIs. In another study, we separately use a reduced form econometric model to relate population-wide changes in mask-wearing to the growth rate of airborne disease infections in the presence of other NPIs. We use the estimated growth rate to predict COVID-19 spread

across 24 countries using the Susceptible–Infected–Recovered (SIR) network model.

- Lastly, we study the propagation of malicious user activity in Online Social Networks (OSN). OSNs spread content widely and rapidly among users; thus, they can be used as vectors to disseminate malicious content, e.g., spam. We analyze malicious behaviors by conducting case studies on a content-driven OSN, Pinterest. Based on the insights gained from our analyses, we develop learning-based models to detect whether posted content is malicious. We investigate the role of various features in the prediction task and show that the observed properties when a content is posted can be used to protect users from potential risks.

Acknowledgments

I would like to express my deepest gratitude and appreciation to my advisor, Professor Ashwin Aravindakshan, for his consistent support, invaluable guidance, and inspiring suggestions. Ashwin's advice steered me throughout my research. I cannot express how thankful I am for patiently providing encouragement and dedicated guidance that facilitated my success.

I am incredibly grateful for Professor Chen-Nee Chuah, without whom I would not have been able to complete this research. Her continued support and advice have propelled me throughout my academic endeavors.

I would like to especially thank my great friend Mohammad Motamedi for his invaluable help. Thank you for your brilliant advice during our countless chats. I would also like to thank Jörn Boehnke, Ashutosh Nayak, and Kourosh Vali, whom I had the pleasure of working with and who had helped me tremendously in different parts of this work. I am grateful to Professor Yong Jae Lee and Professor Soheil Ghiasi for their insightful comments that improved the quality of this work.

Finally, I am especially grateful for Erin Nannery's tremendous help, love, and support, and my family's unconditional love and support.

Chapter 1

Introduction

Machine Learning (ML) and algorithmic designs have achieved remarkable accuracy in various prediction tasks [1, 6, 7]. However, many of those techniques remain primarily black boxes[8]. Designing a model capable of explaining the prediction process has become a key factor in deploying such methods. The interpretability of models increases the user's trust in *why* and *how* decisions are made.

Bridging the gap between high-performance models and their explainability would make them viable intelligent decision-making processes. In this dissertation, we propose interpretable models in multiple application domains. In the first section of this chapter, we review state-of-the-art recommender systems, discuss their limitations and propose a recommender system that achieves state-of-the-art performance and can learn personalized interpretable user behavior. In the second and third sections, we discuss the impact of Non-Pharmaceutical Interventions (NPI) on mitigating the spread of contagious diseases in the early phases of the pandemic. In section two, we give an introduction to a spatio-temporal model to understand the impact of each NPI on COVID-19 spread and summarize forecasts and insights for relaxing various types of NPIs. The subsequent section introduces a reduced-form econometric model to relate population-wide changes in mask-wearing to the growth rate of confirmed COVID-19 cases. Finally, in the last section, we give an introduction to malicious content propagation in online social networks and list our contributions to malicious content detection.

1.1 PARSRec: Explainable Personalized Attention-fused Recurrent Sequential Recommendation Using Session Partial Actions

The task of recommender systems is to delineate users' interests accurately. Recommender systems help providers offer users viable alternatives as they navigate a vast number of available choices. They achieve this by leveraging users' historical behavior to extract meaningful patterns that help predict users' future interests. These patterns often change over time and are heterogeneous across users. Therefore, deriving functional patterns becomes increasingly challenging with growing numbers of users, items, and user-item actions. Recommender systems focus on capturing these evolving, diverse, and high-dimensional behaviors.

Two types of recommender systems have gained popularity in recent research, i) sequential and ii) session-based recommenders. Sequential recommenders often consider all historical user actions as a single ordered sequence and try to successively infer each user action based on the user's prior actions in the sequence. Session-based recommenders only leverage the user's most recent actions called session (e.g., anonymous online shopper without an existing historical behavior). The state-of-the-art approaches benefit from deep neural nets to enhance the performance of recommendation tasks. Recurrent Neural Networks (RNN) [9, 10] and their improved variants such as Gated Recurrent Units (GRU) [11, 12] and Long Short-Term Memory (LSTM) [13] often capture all the previous actions of the user in the past via hidden states. This strategy allows them to understand complex user behaviors. However, RNNs suffer from long-range dependencies because long-range back-propagated gradients can vanish or explode. LSTM and GRU prove effective in some fields by resolving this issue but have their limitations in the field of sequential recommendation. They tend to summarize the session information into a single representation. For example, in machine translation, words within a sentence are related to each other (in various degrees). However, not all items within a shopping basket are necessarily related. For example, in the shopping basket *(milk, cereal, laundry detergent)*, the choice of laundry detergent could be completely independent of milk and cereal.

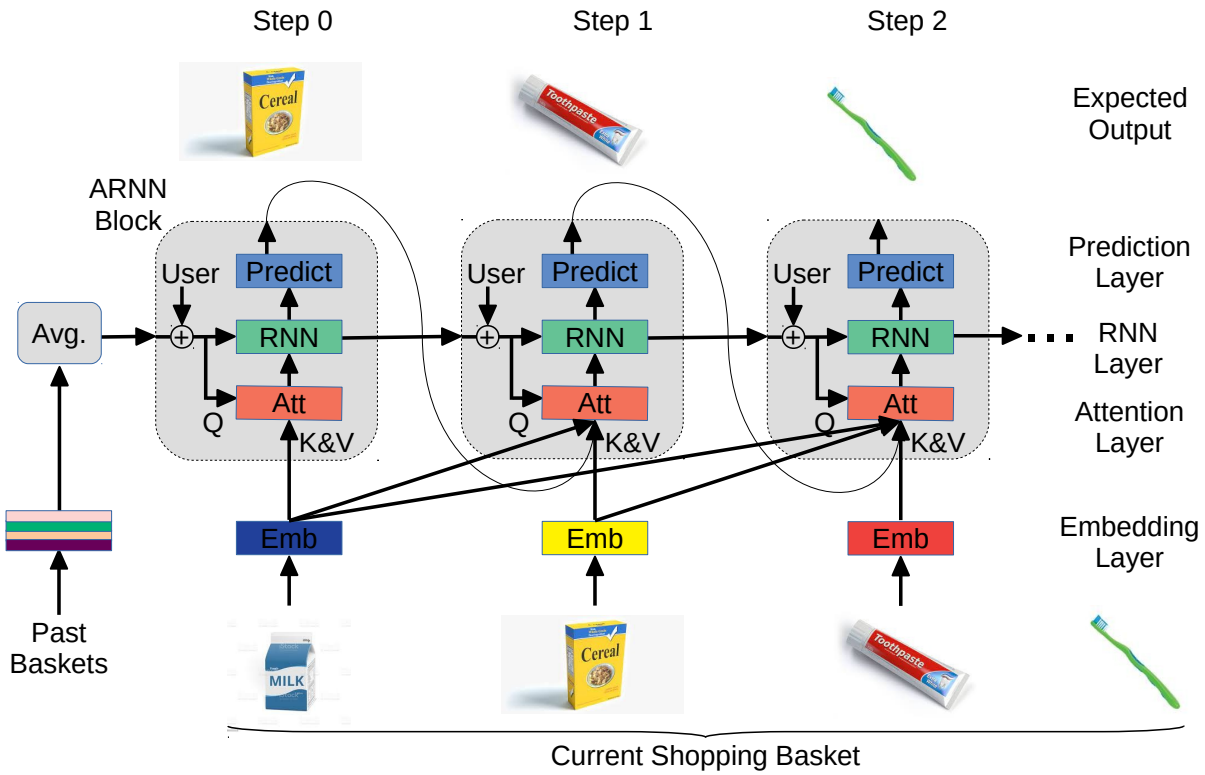


Figure 1.1: Coarse architecture of PARSRec: In every step, first, the Attention module investigates the items that currently exist in the shopping basket to identify those items that will considerably impact the selection of the next item. Then, it represents such items in the latent space as a hint to the recurrent network. Furthermore, the user-aware nature of the proposed architecture makes it possible to leverage users' idiosyncrasies for predicting and suggesting the next item.

Encoding the entire session into one (or a few) representation(s) will entangle irrelevant information together and would possibly make it harder for the decoder to detangle them. These methods assume a natural order to historical user actions, which does not always hold in real-world applications. Moreover, they tend to become slow when processing long sequences due to their sequential nature.

More recent recommender systems use attention mechanisms to overcome the issues mentioned above by identifying a smaller number of actions most relevant to the next item recommendation. They are effective in many applications and can provide helpful interpretable visuals for item-item relationships. The self-attention mechanism in Transformers learns long-range global item-item relationships and utilizes that along with the

items in the user's historical actions to learn heterogeneous user behaviors implicitly. They require less dense data and can run faster in parallel. However, the state-of-the-art transformer recommender systems usually can handle limited input length, limiting the number of user actions[14, 15, 16]. This is due to the extra positional embedding required to capture the relative or absolute order of the items in a sequence. On top of that, item relations generally differ from one user to another. For instance, user A might purchase milk for cookies, while user B might purchase milk for cereal and independently buy cookies simultaneously. Most attention-based models capture only the universal item-item relationships based only on item sequence and lack personalized interpretable user behavior representations.

Inspired by these methods, we propose PARSRec, a **P**ersonalized **A**ttention-based **R**ecurrent **S**equential **Rec**ommender that fuses the attention mechanism with RNNs, illustrated in Figure 1.1, to address the limitations mentioned above. Our framework partitions a user's actions into two groups: *i)* information from past sessions and *ii)* items interacted with so far during the current session. There are multiple reasons for this approach. In a wide range of applications, the partial knowledge of the current session provides much richer information than the previous sessions [17]. For example, one may purchase cereal and milk on a trip to a grocery store. Knowing the partial information of the cereal is more likely to help predict the next item, milk, accurately compared to the shopping baskets in the past. In another example, the list of the music tracks that a user listens to in a session is more likely to be related to other songs in the same session than tracks in other sessions [18]. Second, in many real-world cases, there is no order to the items in a session (e.g., items within a basket), whereas the sessions themselves may follow a chronological order. Lastly, our model eliminates the need for positional embeddings by such partitioning of the historical interactions. This, in turn, reduces the network's memory footprint.

The hidden states of the RNN network in our model carry the user information and the user's historical behavior. The attention layer, which is agnostic to orders, uses the hidden state to determine which items within the current session are more relevant in predicting

the next item. PARSRec outperforms the state-of-the-art methods on Nielsen's real-world Consumer Panel dataset, as detailed in Section 2.4. The model extracts interpretable personalized user behavior by using explicit user representations in the attention layer's queries. We show that PARSRec can accurately explain personalized user behavior in a controlled environment on a synthetic dataset. This powerful explanation allows the provider to fully understand the underlying user behavior beyond a simple next item recommendation to make informed decisions on many tasks. Examples of tasks that can benefit from this knowledge are assortment optimization, assortment allocation (what items go together on the same shelf or a webpage design), and personalized coupons, discounts, and displays. Our recurrent model capacity is independent of the sequence length. Its complexity depends only on the number of items within a session which is usually small, as detailed in Section 2.4. The network can utilize any length of user history without increasing the capacity or complexity of the model. The key contributions of our work are:

1. We propose a model that uses attention layers combined with RNNs for the task of sequential recommendation. We show that our model outperforms various state-of-the-art methods on synthetic and real-world data under different evaluation metrics.
2. We test our model in a controlled environment of a synthetic dataset. We show that our model learns personalized user behaviors and offers interpretable results through visualizing item relationships.
3. We conduct an ablation study on variations of the proposed model to evaluate the contribution of components of the model and report the most effective architecture.

1.2 Preparing for a Second COVID-19 Wave: Insights and limitations from a data-driven evaluation of non-pharmaceutical interventions in Germany

In response to the COVID-19 pandemic, governments around the world implemented varying degrees of non-pharmaceutical interventions (NPIs) to control the spread of the

disease [19, 20, 21, 22, 23]. These policies severely restricted movement, public gatherings, national and international travel, and shut down large parts of the economy including schools and non-essential businesses. Multiple studies have investigated the role that these lockdowns played in delaying the spread and reducing the severity of the pandemic [24, 25, 26, 27, 28, 29, 30, 31, 32, 33]. The lockdowns also created tremendous hardships for individuals and businesses [34, 35, 36]. As the spread of COVID-19 decelerated across countries, governments began relaxing the NPIs to help balance the need for economic security against the risk of growing infection numbers [37]. Nevertheless, beyond the knowledge of the possibility of a second wave [38], there is still limited understanding of the association of loosening different policies with changes in mobility that eventually relate to the spread of the disease.

In this study, we explore this association by estimating each NPI's connection to social mobility and the resulting disease mitigation. The proposed methodology permits the forecasting of disease spread under different policy scenarios of implementation and relaxation by associating a policy with a change in mobility. The presented model allows policymakers to forecast the impacts of removing different types of restrictions on mobility and disease mitigation.

Initial analysis of the impact of policy restrictions in China suggests that NPIs that significantly affected human mobility (e.g., household quarantine) reduced the spread of the disease [39, 40], even more than restrictions that limited national and international travel [41]. Additionally, simulations of NPIs in Wuhan [31] show that maintaining restrictions helped delay the epidemic peak. The results also suggest that an early end to such interventions leads to an earlier secondary peak, which can be flattened by relaxing the social mixing [31]. To the best of our knowledge, no study quantifies the effects of the types and timings of the implementation and relaxation of government policy interventions in reducing mobility and, in turn, decreasing the spread of COVID-19. Our estimates allow for projections of the impact of easing individual interventions on the spread of the disease. These projections act as aids for policymakers to determine how lifting certain policies will change social mobility and, in turn, affect the number of new

COVID-19 cases.

Using data from the 16 states of Germany, we explore the effectiveness of different NPIs in reducing social mobility and, in turn, affecting the spread of the disease. Because German states enforced and relaxed policies to varying degrees at different points in time, the variations in implementation allow us to capture the incremental effectiveness of these policies at reducing social mobility in the general population. It is important to note that we cannot identify each policy's causal effect on reducing disease spread. While each German state implemented its own COVID-19 interventions, the data do not provide sufficient variation in the sequencing of policies to uniquely identify the effect of each intervention. However, the existing variation allows for a necessary understanding of the relative magnitude of each of the interventions and provides meaningful insights into each policy's contribution to reducing disease spread.

To determine how policy enforcement impacted mobility and disease spread, we associate the type and timing of the policy intervention to actual social mobility as recorded in the data released by Google [42]. Next, using our predictions of social mobility based on the policy interventions, we predict the spread of COVID-19 by modifying the SEIR model presented in [39] to include social distancing and other forms of mobility data (e.g., travel by air, bus, rail, and road). Finally, we project the impact of relaxing a policy on the number of new cases across Germany and compare how differences in start times for policy relaxations alter the cumulative number of expected cases over a 90-day time span.

Our findings suggest that not implementing social distancing in Germany is associated with a 24.6-fold (IQR: 20 to 29-fold) increase in cumulative infected case counts as of May 7, 2020. In other words, social distancing appears to reduce case counts by about 96% (IQR: 95%-96.6%). We also found that policies were not equal in their effectiveness at reducing new cases. Compared to keeping the restrictions in place, the lifting of contact restrictions appears to be the most punitive, with an associated 150% (IQR: 144-156%) increase in daily case numbers over a 90-day period. This is followed by the lifting educational facilities closure and the opening of retail outlets (46.1% IQR: 44.0-48.1%, and 33.9% IQR: 33.0-34.8%), respectively.

1.3 The Impact of Mask-Wearing in Mitigating the Spread of COVID-19 During the Early Phases of the Pandemic

In response to the COVID-19 pandemic, multiple countries curbed the spread of the disease by enforcing strict policy measures such as lockdowns and shelter-in-place orders [43]. The non-pharmaceutical interventions (NPIs) included closures of schools, restaurants, bars, retail outlets, and other non-essential businesses, as well as shelter-in-place policies and the prohibition of large gatherings (e.g., limited to 10 people) [44]. These institutional measures aimed to reduce the exposure of susceptible individuals to symptomatic and asymptomatic infected individuals by decreasing outdoor mobility (e.g., going out to movies, concerts, and restaurants and assembling in large groups) and encouraging social distancing. (e.g., 1m-2m physical distancing) [45, 28].

Unlike the widespread and proactive implementation of lockdowns and physical distancing measures, the usage of masks varied widely across countries. Some countries quickly adopted guidelines for mask usage (e.g., Malaysia, Singapore, Taiwan, and Thailand) while others did not recommend using face masks unless sick [46, 47, 48]. Indeed, the World Health Organization updated its mask-wearing guidelines only on June 5, 2020 [49], to recommend that "The general public should wear non-medical masks where there is widespread transmission and when physical distancing is difficult, such as on public transport, in shops or other confined or crowded environments.". Due to these changing guidelines and uneven implementations, mask-wearing varied dramatically across countries and over time during the early phases of the pandemic [50].

Multiple studies have investigated the impact of various governmental NPIs [24, 30, 45, 51, 52, 53], that encourage physical distancing and other restrictions. In each case, the studies find that NPIs and physical distancing reduce the transmission of COVID-19. Studies on the effectiveness of face masks [54, 55, 56] also show that face masks could contribute to the mitigation of COVID-19. However, a recent study [57] uses a randomized control trial to investigate the effect of masks. The authors find that infection with SARS-CoV-2 occurred in 1.8% of the participants in the treated group (recommended masks for

3 hours per day) versus 2.1% of the participants in the control group. A difference of about 17% over 60 days appears to be statistically insignificant. Despite this conclusion, as noted by [58], the trial in [57] points to "a likely benefit of mask-wearing to the wearer—it did not examine the wider potential benefit of the reduced spread of infection to others—and this even in a population where mask-wearing isn't mandatory and prevalence of infection is low." In addition, the interventions, government policies, individual measures, and exposures to infection due to outdoor mobility seldom act in isolation. Treating these measures in isolation could lead to under- or over-estimation of their effectiveness at reducing the spread of the disease, biasing the assessments of the measure's impact. This study investigates the association of population-wide mask-wearing with the number of COVID-19 cases concurrent with other individual and institutional measures.

In sum, because mask-wearing varied dramatically in early 2020, we restrict this study to examine the mitigating role that mask-wearing played during the early phases of the pandemic. Specifically, we expand on the current stream of research by simultaneously considering the effects of NPIs and outdoor mobility in combination with a population's reported usage of face masks in public places in a reduced-form econometric model (see examples in [30] and [45]). Using data from 24 countries, we identify the effect of each measure by exploiting the country-wise differences in the (1) percentage of the population who report wearing a face mask in public places (YouGov Survey Data [50]), (2) outdoor mobility across multiple categories such as Parks and Transit Locations (using Google Mobility Reports [42]) and (3) the NPI implementations (using CoronaNet-Project [43]). The results re-affirm the importance of mask-wearing in combating the spread of COVID-19.

1.4 Uncovering the Footprints of Malicious Content Propagation in Pinterest

Online social networks (OSNs) have become one of the most popular platforms to disseminate information, content, or news in the last decade. Therefore, many researchers as well as companies have investigated how such data propagate in various

OSNs [59, 60, 61, 62, 63]. Information may get reshared multiple times, e.g., via retweets on Twitter or repins on Pinterest, which results in a *cascade* that ultimately reaches a large number of users.

Due to the nature of OSNs, which spreads information widely and quickly, some suspicious users use such networks as vectors to disseminate unwanted spam, phishing pages, viral marketing advertisements, or malware downloads. These malicious activities, in turn, have prompted the research community to study how such threats spread in popular friendship-based OSNs, including Facebook [64, 65, 66, 67], revealing valuable insight into malicious activities in OSNs. However, little attention has been paid to understanding the propagation patterns of malicious content and the behavior of users who disseminate malicious content in content-driven rather than friendship-based OSNs such as Pinterest. Insights derived from fine-grained analyses can be used to develop models for detecting malicious content and protecting end-users.

To shed light on these issues, we investigate the characteristics of malicious content propagation on Pinterest, a popular content-centric and interest-driven OSN [68, 60, 69, 70]. Content in Pinterest is shared via a 'pin', and other users can further share it as a 'repin' or recently known as 'save', which is similar to a retweet on Twitter. A user can upload a pin or fetch a pin from external websites (e.g., `IMDB.com`). In the latter case, we refer to the URL containing the image or content appearing in the pin as a source. Each pin belongs to one of the 33 Pinterest-defined categories like sports, education, technology, etc.

We analyze a large-scale dataset of Pinterest [60], which contains over 1.5 M pins/repins, with each pin associated with a URL shared by 900 K users. Later, we perform a similar analysis on a newly collected dataset of 1.3 M pins/repins shared by 1.1 M users in 2017 to check whether our results are valid in the data collected in 2017, as the malicious content tends to change rapidly. We use a popular commercial URL scanner (i.e., McAfee TrustedSource scanner [71]) and the Virustotal [72], which includes 64 different antivirus engines and website scanners, to detect whether the URL content associated

with the corresponding pin is malicious[1]. In order to reduce the false-positive detection rate, a website is identified as malicious if it has been detected malicious by at least two different antivirus engines or website scanners. Among the 1.5 M pins/repins, 10 K pins/repins are identified as malicious content, accounting for 0.68% of all the pins/repins in our dataset. We investigate how malicious content propagates on Pinterest compared to benign content based on the identified malicious content. We next identify *'malicious users'* who publish at least one malicious content during our measurement period (See Section 5.3.2 for details) and examine their characteristics. We compare identified malicious users with benign ones on Pinterest regarding their activities, interest diversity, and following/follower relationships. Based on the insights gained from our analyses, we develop a learning-based model to detect whether a posted pin is malicious or not without using commercial URL scanners like McAfee or Virustotal, as they are typically too costly (in terms of expense and time) to scan every URL.

We highlight the main contributions and key findings of our work as follows:

1. *Malicious content propagation*: We observe that there are distinct sets of categories in which the majority of the malicious pins are posted. For instance, spammers are likely to target the 'home & decor' category; malicious pins for phishing are mostly shared in the 'design' category. Our analysis also reveals that many benign pins come from popular websites, while malicious pins are likely to come from a small set of non-popular websites.

2. *Characteristics of malicious users*: We find that malicious users tend to be more active than benign users on Pinterest. They tend to post pins more often compared to benign users and are likely to have larger numbers of followings/followers than benign users. Malicious users tend to have pins in more diverse categories, whereas benign users are likely to have pins in a few categories. Considering malicious users tend to be more active than benign users on Pinterest, they are not significantly cost-effective in distributing malicious content.

[1]The latter dataset is scanned only by Virustotal with 64 antivirus engines and website scanners.

3. *Detecting malicious content*: Based on lessons learned from our analysis, we develop learning-based models to detect whether a posted pin is malicious or not. Our results show that the prediction model based on the characteristics of the pins alone is more accurate than a more complicated model that incorporates user attributes. This finding implies that the properties that can be observed when a pin is posted provide sufficient information to predict whether or not that pin is malicious before it goes viral. Our model can achieve high accuracy (AUC: 0.95), which is highly useful in protecting users from potential risks in Pinterest-like OSNs.

Chapter 2

PARSRec: Explainable Personalized Attention-fused Recurrent Sequential Recommendation Using Session Partial Actions

The emerging meta- and multi-verse landscape is yet another step towards the more prevalent use of already ubiquitous online markets. In such markets, recommender systems play critical roles. Recommender systems help providers offer the optimal set of products that interest users. They also help users navigate through a vast number of items available. Recommender systems are usually designed to learn common user behaviors and rely on them for inference. This approach, while effective, is oblivious to subtle idiosyncrasies that differentiate humans from each other. Focusing on this observation, we propose an architecture that utilizes both common patterns and individual behaviors to choose its recommendations for each user. We illustrate how the proposed architecture can explain the recommendations. We start by reviewing the related work to our study. Then we define the problem, explain our proposed architecture, and discuss its findings in this chapter.

2.1 Related Work

We review the related works on recommender systems close to our framework.

2.1.1 General Recommendation

Collaborative Filtering (CF) is one of the classic approaches in the field of recommender systems [73, 74, 75]. CF infers user preferences from their historical interactions. Matrix Factorization (MF) is one of the successful CF methods that use a shared space to represent both users and items [76, 77]. MF estimates the user's preference for an item by the inner product of the user and item vectors. Another approach uses an item-similarity matrix. It estimates the preference of a user to an item by measuring the similarity of that item to the user's historical purchases without using an explicit user representation [78, 79]. More recent approaches use deep learning to improve the effectiveness of models. One line of work, utilizes extracted features from item meta-data (e.g., images [80, 81, 82], texts [83, 84, 85], acoustics [86], or temporal [18]) in context-aware recommender systems. A different approach replaces MF using neural networks. E.g., Neural Collaborative Filtering [87] uses Multi-Layer Perception (MLP) to infer user preferences. DLRM [88] uses MF and MLP to extract feature interactions for the task of recommendation. Autoencoders are another deep learning design used in recommender systems (e.g., AutoRec [89], CDAE [90], and VAE [91].)

2.1.2 Sequential Recommendation

Studies of sequential recommendation aim to extract item transitions in a sequence of items a user interacts with. Markov Chain (MC) models capture such transitions. Factorizing Personalized Markov Chain (FPMC) [92] and its extension Hierarchical Representation Model (HRM) [93] combine MF and MC to extract a personalized item transition and essentially utilize sequences similarities. Higher-order MCs have been studied for longer range dependencies [94]. Convolutional Sequence Embedding (Caser) uses convolutional neural networks to model fixed-length long-range dependencies [95]. Recurrent models have also shown promising results in the field of sequential recommendation. RNNs and their variants (e.g., GRU and LSTM) have been used for modeling user interaction sequences. Most RNN models encode users' historical behavior into a representation vector and use that along with the current interaction as the input to the model and predict the next action. DREAM [96] adopted MCs in a recurrent setting to create a dynamic

representation of users. GRU4Rec [11] and GRU4Rec+ [12] use Gated Recurrent Units for session-based recommendations. Some works utilize memory networks to store users' actions in RNNs for recommendation [97, 98]. BINN [99] uses two components to capture long and short-term preferences in the RNN setting.

2.1.3 Attention-based Recommendation

Attention mechanism has gained popularity in many fields (e.g., machine translation) due to its promising performance and interpretability. Recent state-of-the-art recommender systems use the attention mechanism. NARM uses attention with encoder to model the user's sequential behavior and the user's session purpose [100]. ACA-GRU leverages attention mechanism to build a context-aware recommender system [101]. STAMP uses attention to capture users' general interests from the long-term memory of a session context and users' current interests from the short-term memory [102]. KGAT [103] uses the attention mechanism on knowledge graphs for the recommendation. SASRec [15] and BERT4Rec [14] use uni-directional and bi-directional self-attention mechanism (i.e., Transformers) to capture item-item relationships and have achieved state-of-the-art performances. TiSASRec [104] incorporated time-intervals in attention mechanism for a time-aware recommender system.

Existing methods that use attention mechanisms often learn an implicit representation of users. Users' representations are commonly learned based on the sequence of user actions and from the global item-item relationships. The methods mentioned above often limit the length of the input sequence (number of user's historical actions) and require cropping the input to a pre-set max-length. Some methods also assume a rigid order to the sequence of user actions. Some real-world applications do not provide a rigid order to user actions (e.g., items within a grocery shopping basket). We seek to design a model that addresses these limitations. By separating the user's current actions (session) from the user's past actions, we eliminate the limitation on input size. We use the attention mechanism to extract the most relevant current actions for predicting the next item. Our model learns explainable personalized item-item relationships that provide insights into user choices.

2.2 Problem Statement

In recommender systems, a set of users, $U = \{u_1, u_2, ..., u_{|U|}\}$, interact with a set of items, $V = \{v_1, v_2, ..., v_{|V|}\}$. Examples of user-item interactions are purchasing an item, listening to a track, watching a movie, or clicking on a link. A sequence of interactions by user $u \in U$, denoted by S_u, is partitioned into sessions $(S^u_{t_1}, S^u_{t_2}, ..., S^u_{t_{|S_u|}})$. Each session $S^u_{t_i} \subseteq V$ is the interaction of user u with a set of items $\{v^{(S)}_j | v^{(S)}_j \in V, 1 < j < n^u_i\}$ at time t_i. The total number of items in the session is denoted by $n^u_i = |S^u_{t_i}|$, and the sessions are in chronological order $(t_1 < t_2 < ... < t_{|S_u|})$. An example of $S^u_{t_i}$ would be a shopping basket or a session of listening to music tracks. In this chapter, we use the terms session and basket interchangeably. We assume there is no specific chronological order to items within a session. For instance, a shopping basket at a brick-and-mortar store does not provide a meaningful order. However, if there is a meaningful order to items within a session, an ideal solution would be able to capture that relationship as well. The task of sequential recommendation is to predict the next item in the session $v^{(S)}_{j+1}$ given the history of user behavior $(S^u_{t_1}, ..., S^u_{t_{i-1}})$, and the subset of items $[v^{(S)}_1, ..., v^{(S)}_j]$ that user has interacted with so far during the current session $S^u_{t_i}$. In other words, the sequential recommendation is the task of finding the item that has the highest probability of being interacted with next:

$$\hat{v}^{(S)}_{j+1} = \underset{v \in V}{\operatorname{argmax}}\, P(v^{(S)}_{j+1} = v | (S^u_{t_1}, ..., S^u_{t_{i-1}}), \{v^{(S)}_1, ..., v^{(S)}_j\}) \tag{2.1}$$

2.3 Model Architecture

In this section, we introduce a new sequential recommendation model called **PARSRec**. A **P**ersonalized **A**ttention-based **R**ecurrent **S**equential **Rec**ommender that combines the power of recurrent neural networks (RNN) and scaled dot-product attention into a modified **A**ttention-fused **RNN**: **ARNN** (Figure 2.1). The outer structure of the model is similar to vanilla RNN, i.e., each ARNN block receives a hidden state and an input and produces an output and the next hidden state. However, the inner architecture of the ARNN block is different from that of the vanilla RNN. All ARNN blocks in Figure 2.1 use the same parameters. Next, we detail the architecture.

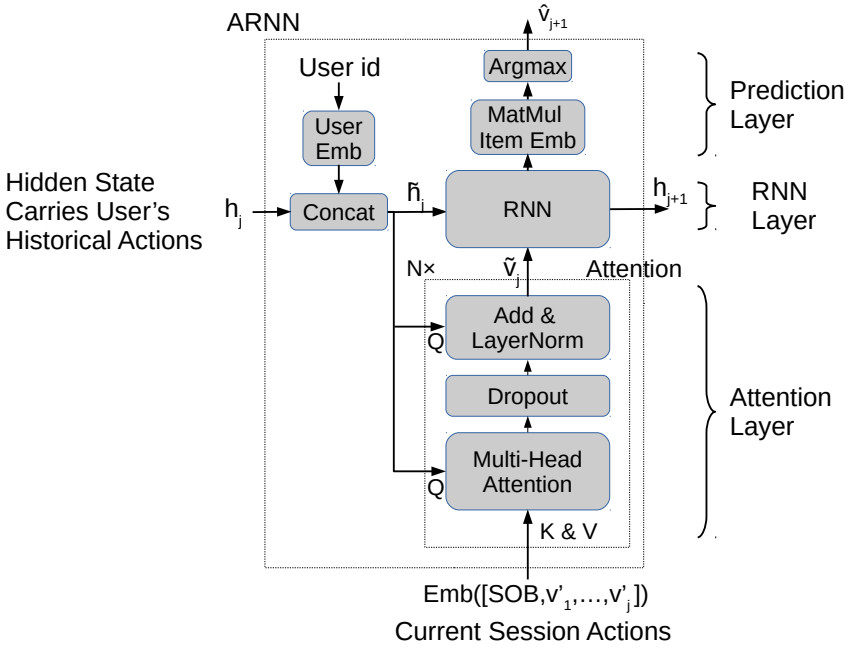


Figure 2.1: ARNN: Attention-fused RNN block

2.3.1 Embedding Layer

We build an *item embedding* matrix $\mathbf{E}^V \in \mathbb{R}^{|V| \times d_v}$ where d_v is the latent dimension of items. We use this embedding matrix wherever we need to convert item indices to embedding vectors in our model.

- We convert item indices within the current basket $S^u_{t_i} = [v^{(S)}_1, ..., v^{(S)}_{|n^u_i|}]$ to input embedding vectors $\mathbf{K}^u_{t_i} \in \mathbb{R}^{n^u_i \times d_v}$ where the j-th row of $\mathbf{K}^u_{t_i}$ is the $v^{(S)}_j$-th row of embedding matrix $\mathbf{E}^V$. We use subsets of $\mathbf{K}^u_{t_i}$ as keys and values to attention layer. The details are explained in section 2.3.2.

- We convert item indices of the previous sessions, $(S^u_{t_1}, ..., S^u_{t_{i-1}})$ to their embedding representations using $\mathbf{E}^V$ and reduce them to a single representation vector using a weighted sum operator, $\mathbf{H}^u_{t_{i-1}} \in \mathbb{R}^{1 \times d_v}$:

$$\mathbf{H}^u_{t_{i-1}} = \sum_{v \in \{w \in s | s \in \{S^u_{t_1}, ..., S^u_{t_{i-1}}\}\}} \mathbf{E}^V_v \tag{2.2}$$

  where $\mathbf{E}^V_v$ is v-th row of $\mathbf{E}^V$. Note that if an item v appears $m > 1$ times in previous sessions, its embedding vector $\mathbf{E}^V_v$ will also be added m times in $\mathbf{H}^u_{t_{i-1}}$. We observe

that this implicitly captures the frequency of user u interacting with item v. We use $\mathbf{H}^u_{t_{i-1}}$ as the initial hidden state h_0 to the **ARNN** block.

User Embedding

Similar to item embedding, the network learns a *user embedding* matrix $\mathbf{E}^U \in \mathbb{R}^{|U| \times d_u}$, where d_u is the latent dimension of user representations.

In cases where we have extra features for users, we can merge that information with the user embedding to create a combined user input to our model:

$$\hat{\mathbf{E}}^U_u = combine(\mathbf{E}^U_u, \mathbf{F}^U_u) \tag{2.3}$$

where $\mathbf{E}^U_u$, $\mathbf{F}^U_u$, and $\hat{\mathbf{E}}^U_u$ are embedding, features, and combined input of user u, respectively, and $combine(.)$ is any function that merges two vectors into one. Examples are (weighted) addition, concatenation, and element-wise multiplication. In our ablation study, we investigated the choices of concatenation. We concatenate $\hat{\mathbf{E}}^U_u$ with the hidden state output from **ARNN** at step $j-1$, h_j, to create the input hidden state to next **ARNN** block, $\tilde{h}_j$, as well as query, $\mathbf{Q}$ in the attention layer.

$$\mathbf{Q} = \tilde{h}_j = Concat(\hat{\mathbf{E}}^U_u, h_j) \tag{2.4}$$

User features can be of two types:

Continuous User Features: If we have continuous user features, x_u, we can pass them through a Multi-Layer Perceptron to extract the relevant features and also adjust it to the proper vector size $\mathbf{F}^U_u$:

$$\mathbf{F}^U_u = ReLU(x_u\mathbf{W} + b) \tag{2.5}$$

where $\mathbf{W}, b$ are weight matrix and bias learned by the model.

Categorical User Features: For every categorical feature, x_u, we can learn a new embedding matrix $\mathbf{F}^U \in \mathbb{R}^{|x_u| \times d_u}$, that maps the categorical feature to a representation vector $\mathbf{F}^U_u$.

2.3.2 Attention Layer

Attention layer have become a popular part of the state-of-the-art models in many research areas due to their promising performances [7]. The attention layer can learn dependencies between two representations regardless of their position in a sequence. We use the attention layer to identify the existing items in the current session that will considerably impact the selection of the next merchandise. The attention module in our model consists of three sub-layers:

Multi-Head Attention Block: We adopt the multi-head attention module described in [7]. It is shown that learning from multiple sub-spaces of representations is more flexible than a single representation. Multi-head attention inherently breaks down the representations into smaller sub-spaces, applies attention to each sub-space, and then concatenates the outputs back into a single representation. The multi-head attention with h heads can be summarized as:

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(head_1, ..., head_h)\mathbf{W}^O \tag{2.6}$$

$$\begin{aligned} head_i &= Attention(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \\ &= softmax\Big(\frac{(\mathbf{Q}\mathbf{W}_i^Q)(\mathbf{K}\mathbf{W}_i^K)^T}{\sqrt{d_k}}\Big)(\mathbf{V}\mathbf{W}_i^V) \end{aligned} \tag{2.7}$$

where the projection matrices $\mathbf{W}^O \in \mathbb{R}^{hd_v \times d_v}$, $\mathbf{W}_i^Q \in \mathbb{R}^{d_q \times d_q}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_k \times d_q}$, and $\mathbf{W}_i^V \in \mathbb{R}^{d_v \times d_q}$ are learned parameters. Query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ are matrices, d_k is the dimension of the key and value, and d_q is the dimension of the query. The attention block is a scaled dot-product of three vectors, key $\mathbf{K}$, value $\mathbf{V}$, and query $\mathbf{Q}$. The role of $\mathbf{Q}\mathbf{K}^T$ is to calculate the relationship between rows in $\mathbf{Q}$ and rows in $\mathbf{K}$. The scaling factor $\sqrt{d_k}$ restrains small gradients from happening by preventing the occurrence of large values in the dot products [7]. Intuitively, attention is a weighted sum of rows in $\mathbf{V}$ and the similarity of rows in matrices $\mathbf{Q}$ and $\mathbf{K}$ defines the weights. The projection matrices allow for various query, key, and value sizes to fit into the attention module facilitating a more flexible and asymmetric relationship learning. In a self-attention setting, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the same. However, we use the attention model differently. In our model, $\mathbf{K}$ and $\mathbf{V}$ are the same and their rows are embeddings of items $\{v_1^{(S)}, ..., v_{n_i^u}^{(S)}\}$. The matrix $\mathbf{Q}$,

however, is different and is the concatenation of user embedding and the previous hidden state.

$$\mathbf{Q} = Concat(\mathbf{E}_u^U, h_j) \tag{2.8}$$

$$m\text{-th row}(\mathbf{K}) = m\text{-th row}(\mathbf{V}) = v_m^{(S)}\text{-th row}(\mathbf{E}^V) \tag{2.9}$$

Similar to [7], to alleviate overfitting, help with stability, and speed up the training process, we add Layer Normalization (LN) and Dropout. We refer to the described architecture as the attention layer in the rest of this chapter.

Layer Normalization: We perform layer normalization on input to stabilize the learning [105]. This layer normalizes each mini-batch, x, across features to zero-mean and unit variance using learning parameters.

$$LayerNorm(x) = \gamma \odot \frac{x - \bar{x}}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{2.10}$$

where $\bar{x}$ and σ^2 are mean and variance of x, $\odot$ is element-wise multiplication, γ and β are learned parameters, and ϵ is a small constant value for numerical stability.

Dropout: It is established that dropout regularization is effective in resolving the overfitting issue. Dropout randomly turns off neurons with probability p during training time and uses all neurons during testing. Intuitively, one can imagine these randomly turned-off versions of the model during training as an ensemble of models that share part of their parameters. We apply dropout in every step of the model.

Stacking Attention Layers: Utilizing hierarchically more numbers ($N > 1$) of attention layers allows the model to learn richer and more complex relationships. We investigate this structure by stacking multiple attention layers using the same query $\mathbf{Q}$ and feeding the output of the lower layer as the next layer's key $\mathbf{K}$ and value $\mathbf{V}$. We find that a single attention layer with two heads is powerful enough to learn the item-item relationships for each individual. It may be beneficial to use multiple attention layers depending on the application and complexity of the relationships.

2.3.3 Recurrent Architecture

We combine the attention layer with a vanilla RNN by using the attention layer output as the input $\tilde{v}_j$ to the RNN block and attention query $\mathbf{Q}$ as the input hidden state $\tilde{h}_j$ to

the RNN block:

$$h_{j+1} = \tilde{v}_j \mathbf{W}^{(1)} + \tilde{h}_j \mathbf{W}^{(2)} + \mathbf{b}^{(1)} \tag{2.11}$$

$$\hat{v}_{j+1} = \text{argmax}\,(\tilde{v}_j \mathbf{W}^{(3)} + \tilde{h}_j \mathbf{W}^{(4)} + \mathbf{b}^{(2)}) \tag{2.12}$$

where $\mathbf{W}^{(1)} \in \mathbb{R}^{d_v \times d_v}$, $\mathbf{W}^{(2)} \in \mathbb{R}^{(d_u+d_v)\times d_v}$, $\mathbf{W}^{(3)} \in \mathbb{R}^{d_v \times |V|}$, $\mathbf{W}^{(4)} \in \mathbb{R}^{(d_u+d_v)\times |V|}$ are matrices, $\mathbf{b}^{(1)}$ is d_v, and $\mathbf{b}^{(2)}$ is $|V|$ dimensional vectors.

The hidden state $\tilde{h}_j$ carries the user information and the current state of the session. We concatenate user embedding to the hidden state at every step to ensure it does not vanish as we progress. The initial hidden state encodes the interaction history, which is the weighted average of embeddings of items in the users' historical actions. The attention layer identifies the items whose presence in the basket is expected to impact the next item selection highly. Subsequently, these items are offered to the RNN to enhance its prediction accuracy. The role of the attention layer in our model at each step j is, given a user and their history at step j (previous sessions plus items interacted within steps $1, ..., j$), find the most relevant items in the session so far $(v_1, ..., v_j)$ to another item in the rest of the session. For instance, when a user starts a session of searching videos on the web S_i^u, the next video v_{j+1} that the user will search likely depends on the user searches so far during that session $(v_1, ..., v_j)$. However, it is also important to know the user u and their search history $(S_1^u, ..., S_{i-1}^u)$ to know how much each of $(v_1, ..., v_j)$ contribute to the next search. The role of attention layer in ARNN is to extract $\tilde{v}_j$, a relevant weighted sum of $(v_1, ..., v_j)$ (i.e. keys and values), by looking at the user, user history and session information so far $\{u, (S_1^u, ..., S_{i-1}^u), (v_1, ..., v_j)\}$ (i.e. query).

Prediction Layer: We use the output of the RNN block at each step to predict the next item. The output of the RNN layer is multiplied by $\mathbf{E}^V$ to provide a similarity vector of size $|V|$. The indices with higher values represent items that are more likely to be interacted with next. We can rank this vector and make recommendations based on that. The objective loss function will convert the output vector to a probability vector using a softmax layer, discussed in the following subsection.

2.3.4 Loss Function

We adopt cross entropy loss function as our objective function:

$$-log\Big(\frac{exp(y[target])}{\sum_k exp(y[k])}\Big) \tag{2.13}$$

where y is the output of ARNN layer and $target$ is the ground truth item v'_{j+1} to be predicted at step j. This objective function combines LogSoftmax and negative log-likelihood loss. The loss is averaged across observations of each mini-batch. We also ignore any EOB padded items in training.

2.3.5 Training

We train the model to predict the items in the current session S_i^u sequentially given a user and their historical behavior. A basket of size n_i^u will have n_i^u prediction steps. Following a common practice in sequential recommender systems [15] and machine translation models [7], we benefit from teacher enforcing during the training. However, since we do not assume any natural order for items in the session, it is beneficial to know which items in the basket are the most related items to predict the next item j. Some recommenders use bidirectional learning that utilizes both left and right items for prediction [14]. The idea is to randomly mask an item (or items) and use remaining information on both left and right to predict masked item(s) [106, 107]. To alleviate the same issue, we take a different approach and modify our teacher enforcing at each step as follows:

$$v'_j = \begin{cases} SOB & & j = 0 \\ \begin{cases} \hat{v}_j & if\ \hat{v}_j \in \{v'_{j+1}, ..., v'_{n_i^u}\} \\ rand(\{v'_{j+1}, ..., v'_{n_i^u}\}) & if\ \hat{v}_j \notin \{v'_{j+1}, ..., v'_{n_i^u}\} \end{cases} & & 0 < j < n_i^u \\ EOB & & j \geq n_i^u \end{cases} \tag{2.14}$$

where SOB and EOB are **S**tart **O**f **B**asket and **E**nd **O**f **B**asket tokens, respectively. In other words, we start with SOB as input of step 0, and if the predicted output in each step is in the rest of the basket, we add that to inputs of the next step. Otherwise, we randomly pick an item from the remaining items in the basket to perform the teacher enforcing. This method of teacher enforcing has the benefit of bringing related items

within a session closer to each other in position and the attention layer can capture those relationships. For instance, a basket in the dataset might be *(cereal, toothbrush, milk, toothpaste)*. The aforementioned teacher enforcing method brings *(cereal, milk)* and *(toothbrush, toothpaste)* closer in position by accurately predicting the item-item relationships.

To perform training in mini-batches, we randomly put baskets of similar sizes in the same mini-batch to avoid unnecessary calculation steps. We pad all baskets on the left with SOB. If some baskets within the mini-batch are of different sizes, we pad them with EOB on the right to match the size of the largest basket within that minibatch. We use *Adam* and *Sparse Adam* optimizers to optimize network's non-sparse and embedding parameters, respectively. Both optimizers adaptively estimate the moments. *Sparse Adam* is a lazy version of *Adam* optimizer used for sparse tensors such as embedding tensors. *Sparse Adam* optimizer speeds up updating process by applying gradients to only those indices that show up in the current mini-batch. We also examined the Stochastic Gradient Descent (SGD) optimizer with various decaying learning rates over decaying steps and found similar results. (and convergence points) as *Adam-Sparse Adam*.

Causality: To avoid leaking information from the future to the prediction of the current step, we only feed the model with the inputs seen at previous steps. The only added item will be either the prediction of the last step or a randomly selected item from the remaining basket. We do not use any extra information from the remaining of the session (including the item we are predicting at the current step). Our modified teacher enforcement makes sure no future information leaks to the past.

2.4 Experimental Results

2.4.1 Datasets

We evaluate our model on two datasets: a controlled synthetic dataset and a real-world dataset. We discuss the details of each data preprocessing next.

Data Synthesis: Verifying a novel method for capturing item-item relationships in an empirical setting is challenging due to the lack of ground truth. Often these rela-

Dataset	#users	#items	#actions	avg. actions/user	Density
Synthetic	8,192	2,000	5.8M	710	0.10
Nielsen	12,800	1,302	15.1M	1,179	0.21

Table 2.1: Post preprocessing datasets statistics

tionships are not known and could vary drastically from user to user. Owing to these issues, most studies primarily evaluate the face validity of the model [15, 97]. The synthetic data allows us to test various effects and evaluate our model under a controlled environment. We start by examining our proposed model on synthetic data, comparing it to benchmarks, and evaluating the characteristics of the model (e.g., personalized and universal item-item relationships). Retail market basket data is one of the well-studied areas [108]. We follow common practice in synthesizing sessions to recreate heterogeneous user behaviors and various item-item relationships [109, 110]. We extract personalized user behaviors from our model using only the basket data, withholding any information regarding items' relations or user choice models. We compare the extracted user behaviors to known values in our simulation. The simulation results validate the model hypothesis and provide an additional basis for performance under empirical data, where results are similar and consistently outperform state-of-the-art models. To the best of our knowledge, no proposed sequential recommender model provides a validated study on a personalized item-item relationship. We next discuss the data generation schema.

The simulation of baskets is as follows: for a basket, S_t^u, with size n_t^u purchased by user u at time t, first the user chooses n_t^u categories from a set of available categories. Then from each chosen category, the user chooses a product j. A category c is a set of similar items (e.g., various types of cereal, milk) and categories are disjoint. For simplicity, each basket can contain at most one item from a category, and not all categories need to be purchased in a single basket. We use the multivariate normal distribution for simulating category choice model to capture various types of category-category relationships (e.g., complements vs. substitutes):

$$p_{ct}^u = \alpha_c + \epsilon_{ct}^u \tag{2.15}$$

where p_{ct}^{u} is probability of purchasing from category c by user u at time t, α_c is a constant category specific utility, and $\epsilon_{ct}^{u} \sim \mathcal{N}(0, \Sigma)$. The user chooses n_t^u categories with highest probability to purchase from at time t. Next, we use multinomial probit model [111] to simulate product choice within each category, $j_t^{u,c}$:

$$\begin{aligned} j_t^{u,c} &= \underset{j \in c}{\operatorname{argmax}}\, \eta_{jt}^{u,c} \\ \eta_{jt}^{u,c} &= \omega_j^{u,c} - \beta^c \nu_j^c + \gamma_{jt}^{u,c} \end{aligned} \tag{2.16}$$

where $\eta_{jt}^{u,c}$ is utility of item j in category c for user u at time t, $\omega_j^{u,c} \sim \mathcal{N}(0, \Omega^c)$ is the product base utility, $\beta^c \nu_j^c$ is a disutility for paying price ν_j^c, and $\gamma_{jt}^{u,c} \sim \mathcal{N}(0, \sigma^c)$ is a random term to capture any uncontrollable parameter affecting the product choice. The simulation first chooses categories based on Eq. (2.15), and the user chooses the product with the highest utility in each category based on Eq. (2.16). We use $C = 20$ categories, each with $|V^c| = 50$ products. Basket size n_t^u is sampled from Weibull(0.80, 1.47) to represent the basket size distribution in our real-world data. Single item baskets and large (> 10) baskets (tail of the distribution) are filtered. We set $\sigma^c = 1, \beta^c = 0.1$, and $\alpha_c = -0.5$. Category specific covariance matrix $\Omega^c = \tau_c^2 \Omega_0^c$ where $\tau_c = 2$ is the standard deviation and Ω_0^c is the correlation matrix. Ω_0^c is a positive-semidefinite matrix generated using vine method [112] under Beta(0.2, 1) distribution to simulate various degrees of product competition. Product prices $\nu_j^c \sim Uniform(\frac{\nu^c}{2}, 2\nu^c)$ where category base price $\nu^c \sim LogNormal(0.5, 0.1)$ is set to mimic the real-world data prices.

A positive value in the covariance matrix Σ indicates two categories are frequently purchased together (e.g., milk and cereal), while a negative value represents contrary (e.g., fresh vs. frozen meat). We manually chose a block diagonal form for the covariance matrix Σ. Each block represents categories that have relations, while categories from separate blocks are independent of each other. We chose various sizes for different blocks (2-3-4) to represent low-mid-high order relationships between categories. For off-diagonal values of each block, we manually chose various positive/zero/negative values to represent complementary/coincidence/substitute product relationships. Further, we divide users into multiple disjoint groups. Each group follows a different Σ to illustrate a deterministic user heterogeneity additional to the unknown user-specific choice error terms. Heatmaps

of Σ are illustrated in Fig. 2.2. We synthesize a less dense dataset compared to the real-world data with 8,192 users, 2,000 items, and 5.8M million user-item interactions. Data statistics are summarized in Table 2.1.

Empirical: We use the Nielsen Consumer Panel[1] dataset. The dataset comprises a representative panel of households. It includes all household purchases (from any outlet) intended for personal and at-home use [113]. This dataset is a particularly challenging one in terms of sequential recommendation. Households are geographically dispersed over all states and demographically balanced to accurately represent the market in each area. The sessions are from different retailers. Over 4.3 million products in the dataset cover a variety of items, including groceries, health and beauty aids, and alcohol. The majority of sequential recommenders evaluate their models on a limited set of users (e.g., within a state), a single retailer (e.g., a clothing retail or content provider), and a related set of items (e.g., all movies, or all clothing). We use data from 2014$\sim$2019 and randomly select 12,800 actively participating users with 15M+ user-item interactions from 50,000+ retailers. The products are grouped into hierarchical categories:

- Departments (10 departments)
  - Groups (118 groups)
    - Modules (1,305 modules)
      - Unique Product Code (4.3+ million UPCs)

The 10 Departments of products are dry grocery, frozen foods, dairy, deli, packaged meat, fresh produce, nonfood grocery, alcohol, general merchandise, and health and beauty aids. We use *Module* level as items in our experiment to reduce the sparsity of purchase patterns and will refer to it as products in the rest of the chapter. We keep one copy of the same item purchased in multiple quantities in each basket. We exclude items and users with less than ten records in the entire dataset. The statistics of data are summarized in Table 2.1.

[1] `https://www.chicagobooth.edu/research/kilts/datasets/nielseniq-nielsen`

2.4.2 Experimental Setup

The optimal architecture of PARSRec includes $N = 1$ attention layer with $h = 2$ heads. We observe that adding extra attention layers does not increase the performance significantly, presumably because the session lengths are short and a single transformer is able to induce all the necessary relationships within a session. Item and user embeddings are the size of $d_v = d_u = 128$. We initialize the item and user embedding matrices with uniform distributions of range $[-\frac{1}{\sqrt{|V|}}, \frac{1}{\sqrt{|V|}}]$ and $[-\frac{1}{\sqrt{|U|}}, \frac{1}{\sqrt{|U|}}]$, respectively. Other matrix parameters with size $n \times m$ are initialized with $\sim \mathcal{N}(0, \frac{2}{n+m})$. Dropout rate is set to 0.1 and mini-batch size is 256. We used Adam and Sparse Adam optimizers with learning rate 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The gradient is clipped at 30. We use PyTorch=v1.10.1 to implement the model.

2.4.3 Evaluation Metrics

We use users' most recent session $S^u_{t_{|S_u|}}$ for testing, their second to last session $S^u_{t_{|S_u|-1}}$ for validation, and the rest for training. Note that the test set will also have validation in its historical data. We predict items in the user's test basket sequentially (basket size > 1). For consistency, we adopt a similar strategy to [15, 14, 87, 95], where they use a set of randomly sampled negative items plus ground truth item for evaluation purposes. We use a set of 100 randomly sampled items (negative and positive) plus ground truth items and rank these items to get the highest recommendations for the next item. Negative sampling is suitable in applications where users are less likely to interact with an item more than once (e.g., watching movies). However, in applications where users frequently interact with an item (e.g., grocery shopping, listening to music), evaluating the model on a pool of negative items (items that the user never interacts with) would make the prediction task more trivial. We observe that in the real-world dataset, about 81% of test basket items also appear in the user's purchase history. We include positive items in the sampling to resolve this issue.

For evaluation metrics, we report common top-N metrics [15, 14]: Hit Ratio (HR@k) and Normalized Discounted Cumulative Gain (NDCG@k) with $k \in \{1, 5, 10\}$. NDCG is a rank-aware metric that penalizes lower-ranked recommendations. Note that NDCG@1

equals HR@1. Additionally, we report another metric for session-level evaluation: session precision (**Sess-Prec@k**), the average ratio of items predicted accurately within a session:

$$session\ precision = \frac{|\{\hat{v}_j | \hat{v}_j \in S_t^u\}|}{|S_t^u|} \tag{2.17}$$

In other words, session precision is the number of items that are recommended accurately normalized by the basket size.

2.4.4 Benchmarks

To validate the effectiveness of our proposed model, we benchmark it against the following baselines:

- **POPRec**: A basic benchmark that recommends items based on their frequency of interactions by users.
- **SASRec** [15]: It uses one-directional transformers for sequential recommendation. We modified negative sampling to sampling explained in Section 2.4.3 for fair comparison.
- **BERT4Rec** [14]: It applies bi-directional transformers with Cloze objective to sequential recommendation, and achieves state-of-the-art performance on many datasets.

We omit comparison with some other benchmarks like GRU4Rec [11], GRU4Rec$^+$ [12], BPR [114], NCF[87], and FPMC [92] that have been outperformed by above methods on various real-world datasets [15, 14]. We use the code provided by the corresponding authors for benchmark models. We consider latent dimension sizes $d_v \in \{16, 32, 64, 128, 256\}$ for methods that include embeddings. The dropout rate is chosen from $\{0, 0.1, ..., 0.5\}$. We set $N = 820$ (full sequence length) for synthetic dataset and $N = 1000$ ($\sim$median of sequence length) for Nielsen dataset in SASRec and BERT4Rec that require setting sequence length. We either tuned all hyper-parameters on the validation set or referred to benchmarks' authors suggestions for optimal values. The reported results are under models' optimal set of hyper-parameter values.

Dataset	Metric	POPRec	SASRec	BERT4Rec	PARSRec	Improvement
Synthetic	HR@1	0.0005	0.1816	0.1963	**0.2136**	8.8%
	HR@5	0.0028	0.3919	0.4150	**0.4791**	15.4%
	HR@10	0.0066	0.5133	0.5441	**0.5825**	7.0%
	NDCG@5	0.0013	0.2872	0.3061	**0.3489**	13.9%
	NDCG@10	0.0026	0.3253	0.3426	**0.3800**	10.9%
	Sess-Prec@1	0.0005	0.1826	0.1974	**0.2249**	13.9%
	Sess-Prec@5	0.0029	0.3931	0.4163	**0.4784**	14.9%
	Sess-Prec@10	0.0068	0.5151	0.5357	**0.5792**	8.1%
Nielsen	HR@1	0.0008	0.1663	0.1761	**0.2444**	38.8%
	HR@5	0.0921	0.4235	0.4833	**0.5934**	22.7%
	HR@10	0.1620	0.5771	0.6580	**0.7355**	11.7%
	NDCG@5	0.0454	0.2991	0.3290	**0.4208**	27.9%
	NDCG@10	0.0680	0.3497	0.3987	**0.4632**	16.1%
	Sess-Prec@1	0.0008	0.1669	0.1856	**0.2753**	48.3%
	Sess-Prec@5	0.0919	0.4197	0.4790	**0.6093**	27.2%
	Sess-Prec@10	0.1632	0.5729	0.6344	**0.7439**	17.2%

Table 2.2: Performance comparison of all benchmarks. Boldface and underlined values in each row represent the best and second best performances, respectively. Improvements of best to second best model are presented in the last column.

Performance Comparison: Table 2.2 summarizes the recommendation performance of all benchmarks as well as our proposed model on synthetic and Nielsen datasets. The last column shows the statistically significant improvements of the best model to the second-best model. We observe that POPRec has the lowest performance, presumably because it does not look into user-specific behaviors.

SASRec uses transformers and performs lower than BERT4Rec because it only looks at item relationships from left to right. BERT4Rec also uses transformers but in a bidirectional setting that allows learning from both right and left. PARSRec performs consistently better than all methods on both datasets. It is likely because PARSRec differentiates between interactions within a session and interactions in the past sessions by using RNN blocks. It also uses explicit user queries in its attention blocks to learn personalized item-item relationships, which is explained more in detail in Section 2.4.5. PARSRec gains an average of 19.1% HR@5, 20.9% NDCG@5, and 21.1% Sess-Prec@5 improvement over the second-best benchmark.

2.4.5 Discussion

The key deliverable of our study was to develop a personalized recommendation system that accounted for patterns of individual preferences over time and item relationships. The model developed here can be applied to multiple scenarios, from populating song lists to basket completion exercises. The model's success hinges on its ability to address two key questions: (1) how does PARSRec capture personalized item-item relationships? And (2) how does learning personalized item-item relationships affect the recommendation?

2.4.5.1 Extracting Personalized Item Relationships

Recommender systems follow one of the two common methods to capture heterogeneous user behaviors: *i)* learn an explicit user representation based on the user behavior [95, 92]; *ii)* implicitly represent users by aggregating the embeddings of user interacted items [15, 14, 11, 12, 94]. PARSRec takes the former approach by learning an explicit user embedding. Recently, the attention layers of transformers in the state-of-the-art recommender systems provide an explainable visual for item-item relationships. These models (e.g., BERT4Rec [14], and SASRec [15]) often use a self-attention layer where all key,

value, and query parameters consist of item embeddings (implicit user representation). Hence, they learn global item-item relationships that differentiate users via their purchase history. However, in PARSRec, the query consists of explicit user embeddings (plus user purchase history, if available). This approach allows the network to learn and visualize the personalized item-item relationships explicitly. We conduct a controlled simulation on the synthetic dataset to validate this capability. We split users into various groups, and each group has a different category covariance matrix, Σ, during data synthesis. Figures 2.2a, 2.2c, 2.2e and 2.2g illustrate the heatmaps (of a subset) of Σ for two different user groups A, and B (along with their average and difference). We then extract the highest attention weights of each user at every step of the training phase (averaged on attention heads). Note that we only extract weights of the valid non-padding items. We aggregate the item-item attention weights to category level for ease of comparison and readability. We perform row-wise sum followed by column-wise average within each category, then row-wise normalization over all items for each user individually. Figures 2.2b, 2.2d, 2.2f and 2.2h show the heatmaps of the category level attention weights for two different user groups A and B, and their average and difference. Further, Figure 2.2i represents the heatmap of dot-product of item embedding with itself ($\mathbf{E}^V \cdot \mathbf{E}^{V^T}$). We observe the following traits from the figures:

- (2.2a) vs. (2.2b): We expect the attention weights in (2.2b) to capture the average category relations for all users corresponding to Figure (2.2a). We observe that there is a correlation between attention weights and covariance matrices. Higher attention weights correspond to larger positive covariance values, and near-zero attention weights correspond to the negative covariance values. We also observe that attention weights capture category independence in the form of block-diagonal matrices. Note that the network has no prior knowledge of item categories, and user-item interactions are the only information introduced to the network.
- (2.2c, 2.2e, 2.2g) vs. (2.2d, 2.2f, 2.2h): We observe that PARSRec can learn user heterogeneity by accurately extracting different user groups' attention weights. We see a one-to-one match between user groups' covariance matrices in (2.2c, 2.2e) and their

corresponding attention weights in (2.2d, 2.2f). For easier comparison, we include their differences in (2.2h). The cells with higher absolute values in Figure (2.2h) match the differences in Figure (2.2g). Note that the network has no prior knowledge of user groups and infers their difference via user behavior.

- (2.2a) vs. (2.2i): The global item-item similarity (i.e. $\mathbf{E}^V \cdot \mathbf{E}^{V^T}$) matches the average relationships of items (i.e. the average covariance matrix). We observe the similarity of items within each category illustrated by blue colors in Figure (2.2i). This is because, in data synthesis, a user interacts with at most one item per category during a session, and items within a category are expected to have similar embedding. Figure (2.2i) also shows a matching pattern with global category correlations where red cells match high positive category covariance values, and blue cells match low positive or negative covariance values. Note that, unlike attention weights, the model learns the item embeddings relative to other item embeddings. E.g., categories 14 and 16 show similarities in Figure (2.2i) (blue color) even though they have a positive but *relatively* lower correlation *compared to* category pairs 14-15 and 15-16 (red color). However, attention weights of categories 14 and 16 are correctly captured positively in Figure (2.2b). Figure (2.2i) shows that the model learns the global item similarities to some extent via item embedding. However, item representations alone cannot present personalized item relationships for each user group like the attention layer does in Figures 2.2d and 2.2f.

Empirical Attention Between Items: The empirical dataset lacks the ground truth for user behavior. However, we can look into item relationships learned by the network. Figure 2.2j shows the average of a subset of attention weights between items for a group of users who have babies in the Nielsen dataset. We observe that the network learns to separate various product modules related to each other without prior knowledge of users or item categories (e.g., {nail polish, nail polish remover} vs. {adult cold remedy, cough syrups}). The network also identifies subgroups of related items within a category (e.g., {baby bath products, baby lotion} vs. {baby bottle, baby burp cloth, baby accessory}) even though they occur in the same sessions.

2.4.5.2 Effect of Personalized Recommendation

We conducted a controlled simulation on the synthetic dataset to investigate the importance of learning personalized item-item relationships in making recommendations. We drop a product category from the pool of products and observe its consequences on other categories' sales. For example, it is observed that when a retailer discontinues tobacco and cigarettes, they face a cross-category spillover effect on other products such as alcoholic beverages [115]. We mimic the real-world behavior in our simulation as follows: 1) we assume the assortment contains all products during the training and validation phase, 2) we remove baskets that contain the dropped category C_0 from the *test set* and also from the pool of products *during testing phase*, and 3) we estimate the unit sales of two types of categories for two user groups during testing phase (k=10). Category C_1 is independent of C_0 for both user groups. Category C_2 is highly correlated with C_0 for user group A (cov=0.6), but independent of C_0 for user group B (cov=0). Figure 2.3a shows the estimates of category sales for each user group, and Figure 2.3b shows Mean Absolute Percentage Error (MAPE) of sales. We observe that PARSRec accurately distinguishes the spillover effect of C_0 removal on categories C_1 and C_2. PARSRec estimates the cross-category drop in sales of category C_2 for user group A (MAPE=3.5%) and uninfluenced sales of C_2 for user group B (MAPE=1%). It also accurately estimates the sales of the independent category C_1 for both user groups (MAPE=3.8% and 3.7%). It is important to note that all training/validation sets include category C_0 and only the test set is subject to change. Without any prior training on category removal effects, the proposed model can predict user behavior by learning personalized item-item relationships. This method allows retailers to conduct simulated experiments on assortment modifications and estimate their effect on different customers without the costs of applied experiments.

2.4.6 Ablation Study

We conduct an ablation study to understand the impact of various components of PARSRec and some variations in the design of the network. Table 2.3 summarizes the performance of the optimal model and eight variants on both synthetic and empirical datasets. All other hyper-parameters are unchanged. The variants are as follows:

Architecture	Synthetic		Nielsen	
	Sess-Prec@10	NDCG@10	Sess-Prec@10	NDCG@10
PARSRec default	0.5792	0.3800	0.7439	0.4632
Remove LN	0.5756	0.3746	0.6775	0.4183
Remove Dropout	0.5780	0.3775	0.7356	0.4553
Remove Q in LN	0.5643	0.3745	0.7400	0.4601
#Att Layers=2	0.5506	0.3622	0.7320	0.4543
#heads=1	0.5769	0.3781	0.7399	0.4602
#heads=4	0.5790	0.3799	0.7431	0.4628
FFN pre-RNN	0.5723	0.3775	0.7426	**0.4637**
FFN post-RNN	0.5009	0.3362	0.6900	0.4189
Extra Features	-	-	**0.7591**	**0.4759**

Table 2.3: Ablation analysis (NDCG@10 and Sess-Prec@10) of synthetic and empirical dataset. An increase in performance from the default PARSRec setting is emphasized in bold.

1. *Remove LN, Dropout*: Including both components help improve the performance, with LN being more prominent on the empirical dataset.

2. *Remove* $\boldsymbol{Q}$ *from LN*: Adding $\mathbf{Q}$ at LN increases the accuracy. It carries the user history and is beneficial to add it to RNN input.

3. *Number of Attention Layers*: Stacking more attention layers achieves similar performance on the empirical dataset but performs lower on the synthetic dataset. Presumably, because the sessions are short and, single layer attention is sufficient to learn the relationship of items within a session. Longer sequences might pose more complex patterns and require more layers.

4. *Number of heads in Multi-Head Attention*: The results show that increasing the number of heads beyond two does not significantly boost performance, presumably because of small session lengths. Longer sessions might benefit from more number of heads.

5. *FFN pre-RNN input*: We explore adding layers of Feed-Forward Network right before the input of RNN. The results show that the added FFN has a similar performance. It is likely that the projection layer of the attention mechanism extracts the required features, and adding extra layers is redundant.

6. *FFN post-RNN output*: Adding FFN after RNN output significantly decreases the performance, possibly due to overfitting. The optimal dimension of RNN is satisfactory for output prediction.

7. *Extra User Features*: We explore adding extra user and session features from the empirical dataset. We pass continuous features through a Multi-Layer Perceptron and further concatenate them with the user and categorical features' embeddings to use as the initial hidden state. Features include income, age, gender, location, retailer id, day of the week of session, household size and composition, user education, marital status, and race. These features contribute slightly to improving performance ($\sim$2% increase). This confirms that the personalized model learns user features that contribute to the recommendation task via user embedding.

2.5 Future Work

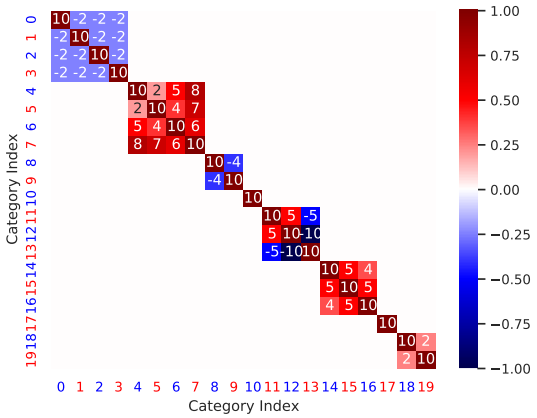
We proposed PARSRec, a sequential recommender model that combines attention mechanism and RNN. PARSRec learns personalized item relationships by using explicit user embeddings in the query of attention mechanism. We conduct a controlled simulation on a synthetic dataset to validate user behavior learning. Empirical results on Nielsen's Consumer Panel dataset show that PARSRec outperforms state-of-the-art self-attention models. One future direction is incorporating the attention mechanism with GRU/LSTM networks for longer sessions and exploring the impact of a different loss function such as Bayesian personalized ranking. Another direction would be to leverage item features (e.g., textual information, price, flavor) in the attention mechanism.

2.6 Data and Code Availability

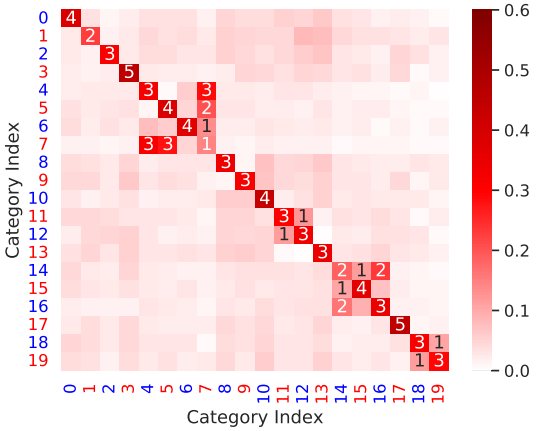
All codes are in python 3.7 programming language and use PyTorch=v1.10.1. Simulation data and codes are available at the public GitHub repository at: `https://github.com/ehgh/PARSRec`.

Empirical data is available by request from Nielsen: `https://www.chicagobooth.edu/research/kilts/datasets/nielseniq-nielsen`.

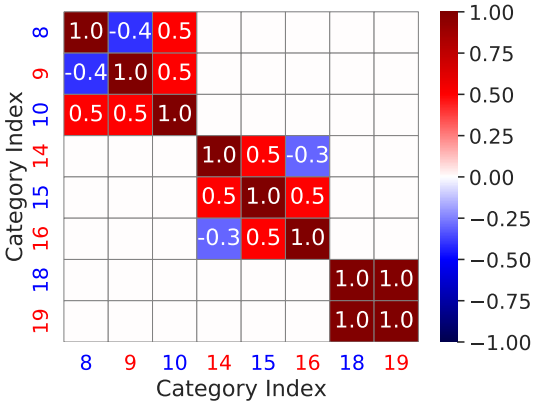
"Researcher(s)' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business." "The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein."



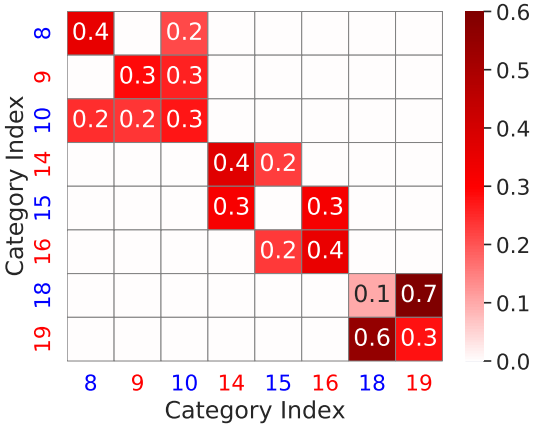
(a) Average category purchase covariance matrix Σ of all users in data synthesis



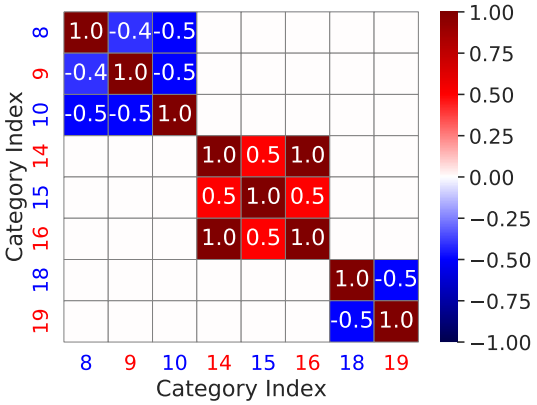
(b) Average category attention weights learned by PARSRec



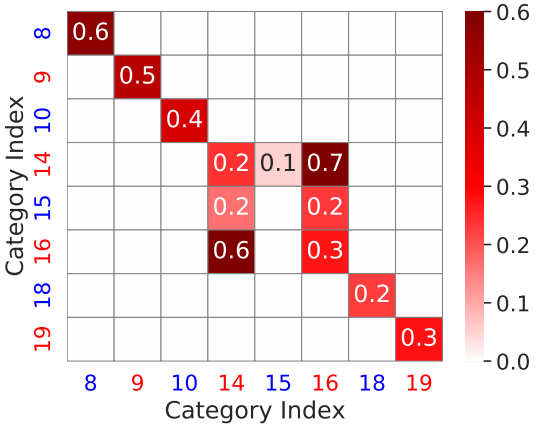
(c) A submatrix of Σ_A of user group A in data synthesis



(d) Attention weights of corresponding categories in (2.2c) of user group A

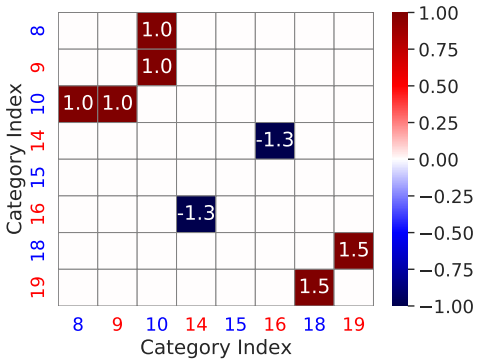


(e) A submatrix of Σ_B of user group B in data synthesis

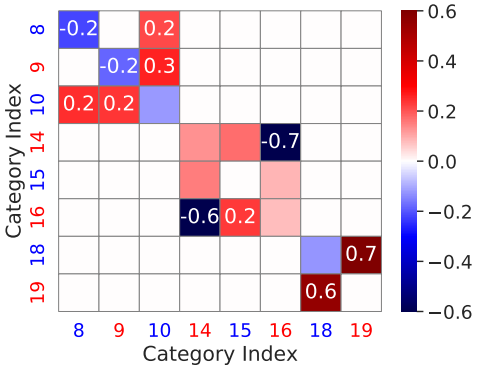


(f) Attention weights of corresponding categories in (2.2e) of user group B

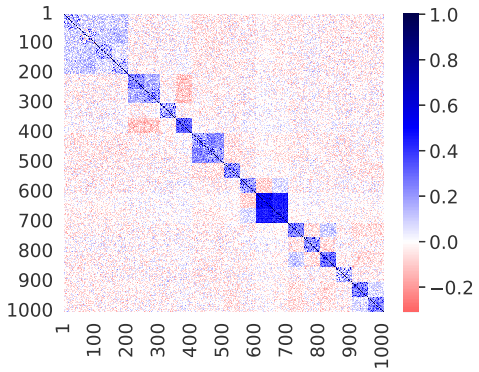
Figure 2.2: Simulation and empirical heatmaps (Continued on the next page.)



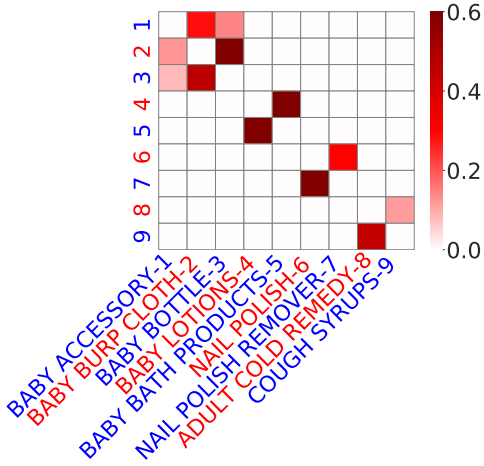
(g) $\Sigma_A - \Sigma_B$, difference of user groups A and B in data synthesis



(h) Difference of category attention weights of user groups A and B



(i) Item embeddings self-dot product heatmap ($\mathbf{E}^V \cdot \mathbf{E}^{V^T}$)



(j) Empirical data: Average attention weights of items for users with babies

Figure 2.2: Simulation and empirical heatmaps (cont.): (a) Average covariance matrix Σ of category purchases $p^u_{ct} \sim \mathcal{N}(\alpha_c, \Sigma)$ in data synthesis. A positive value in Σ indicates two categories are purchased together frequently (e.g. milk and cereal), while a negative value represents contrary of that. (b) Average category attention weights learned by the model. Higher values correspond to positive values in (a) while lower values correspond to zero/negative values in (a). (c, e) A subset of covariance matrix for user groups A and B, (g) difference of (c) and (e). (d, f) Category attention weights for selected categories in (c, e), the matching of (c↔d) and (e↔f) illustrates that model can learn personalized attention. Values < 0.05 are filtered for clarity, (h) difference of (d) and (f). The high values matches Subfigure (g). (i) Item embedding self-dot-product shows embedding has learned global item categories and relationships. (j) Attention weights of sample products in empirical data for a group of users with babies. The network learns correct relations without prior knowledge of users or products (Color represents sign and intensity represents value. Annotations in (a) and (b) are scaled by 10 for readability. All other annotations illustrate the true values of the cell. Figure best viewed in color. Axis label colors are for readability.)

(a) Sales estimation of different categories for user groups A and B

(b) Sales estimation MAPE

Figure 2.3: Estimating the influence of removing category C_0 (present at training phase) from the assortment at testing phase on sales of categories C_1 and C_2 for two user groups. Cov(C_0,C_2)=0.6 for user group A. Other covariances are zero. PARSRec captures category spillover effects.

Chapter 3

Preparing for a Second COVID-19 Wave: Insights and limitations from a data-driven evaluation of non-pharmaceutical interventions in Germany

In this chapter, we build an epidemiology model (Susceptible-Exposed-Infectious-Recovered, SEIR) to understand the impact of Non-Pharmaceutical Interventions (NPIs) on the spread of contagious diseases. We conducted a case study on COVID-19 across Germany. We postulate that the introduction of NPIs led to an increase in social distancing that helped contain the spread of the virus. We use Community Movement as an indicator of social distancing. First, we build an SEIR model that incorporates social distancing as a parameter. Then we use a regression model to estimate the impact of different NPIs on social distancing. We include other covariates that could simultaneously impact social distancing (e.g., awareness, weather, state, week, day of week fixed effects). We use the coefficients from the regression model to simulate different scenarios and quantify the impact of NPIs using the proposed SEIR model. We first discuss the data used in this study, followed by our analysis's methods, results, and limitations.

3.1 Method

3.1.1 COVID-19 Case Data

Interconnected air, land, and sea transportation networks led to the spreading of COVID-19 from Wuhan, China, to the rest of China, and eventually to most countries around the world [116, 117]. To accurately model the spatial spread of the disease into Germany, we collected three types of daily mobility data: (i) daily air transportation data to capture the movement within and between Germany and 142 other countries; (ii) daily ground transportation data between the nine countries that share borders with Germany; and (iii) daily inter-state ground transportation. The daily COVID-19 case data were obtained from the Johns Hopkins Coronavirus Resource Center [118] and Robert Koch Institute [119] for all countries in our dataset as well as the 16 German states. Figures 3.1 and 3.2 introduces the states and the cumulative case numbers for all German states and 143 countries.

3.1.2 Government Policies

To encourage and enforce physical distancing, governments across all 16 German states introduced a variety of NPIs at different points in time. Figure 3.3 provides an overview of the timeline. Data for these policies were collected from [120, 121], and Table 3.3 in Section 3.6 describes those policies. To understand the impact of each policy in containing the spread of the disease, we analyze (i) what would have happened if individuals did not reduce their mobility (i.e., if social distancing norms were never introduced), and (ii) what would happen if a policy (p) is relaxed in the future. While the former can help governments implement critical policies in case of future epidemics, the latter can help governments decide which policies to relax first during current and future epidemics.

3.1.3 Community Mobility

Google's COVID-19 Community Mobility Reports [42] detail how movement trends change over time as public awareness increases and NPIs are introduced (Figure 3.4). The report tracks movement trends over time by geography across different categories such as retail and recreation, groceries and pharmacies, parks, transit stations, work-

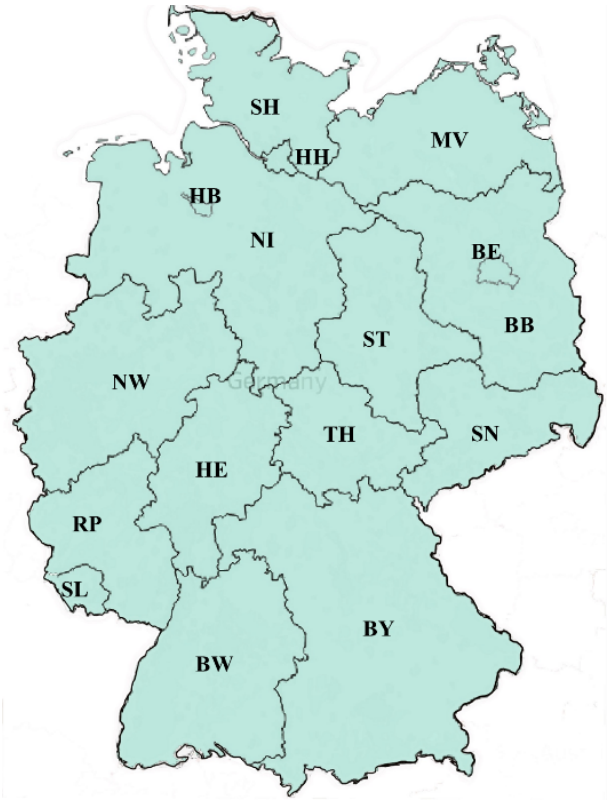


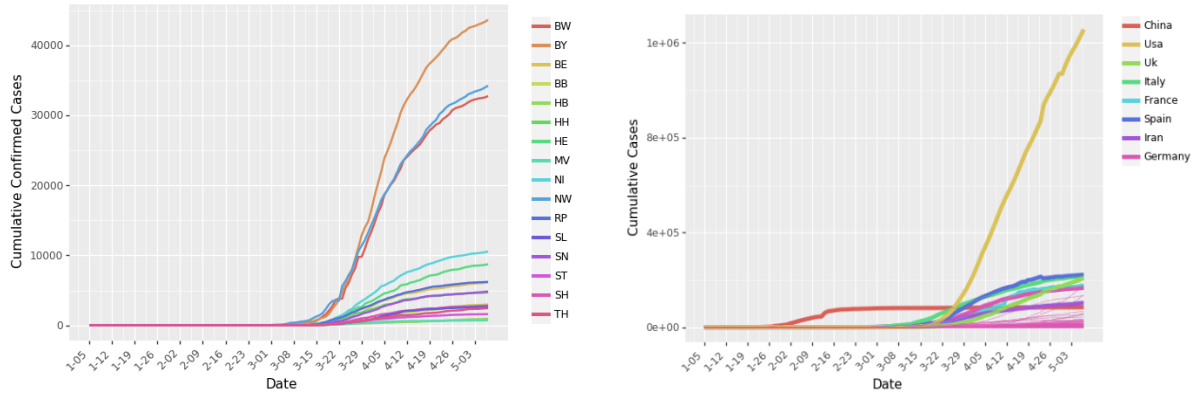
Figure 3.1: Map of Germany.

State	Abbreviation
Brandenburg	BB
Berlin	BE
Baden-Württemberg	BW
Bavaria	BY
Bremen	HB
Hessen	HE
Hamburg	HH
Mecklenburg-Vorpommern	MV
Lower Saxony	NI
North Rhine-Westphalia	NW
Rhineland-Palatinate	RP
Schleswig-Holstein	SH
Saarland	SL
Saxony	SN
Saxony-Anhalt	ST
Thuringia	TH

Table 3.1: Abbreviations of the 16 German states.

places, and residential. We also consider Apple's community movement reports [122]. As movement trends across locations are highly correlated (e.g., 0.92 between Apple's "Driving" and Google's "Retail and Recreation"), we only consider community mobility trends "Retail and Recreation" as a measure of social distancing (details in Section 3.6.3.1). We define social distancing as $sd_i = -C_i/100$ where C_i is the community mobility trend in state i.

While the community mobility data provides information on changes in the local movement, it does not provide information on the inter-state movement. Ground transportation accounts for the vast majority of the movement, with cars accounting for 85% of total ground transportation in Germany [123]. We collected detailed traffic data from January 1, 2013, to December 31, 2018, from the German Bundesanstalt für Straßenwesen (Federal Institute for Roadways). The dataset contains the hourly count of the number of vehicles crossing different checkpoints along highways across Germany. The institute



(a) Cumulative Cases in all 16 German States (b) Cumulative Cases in 143 Countries.

Figure 3.2: Confirmed positive cases for COVID-19 in Germany and the world. Panel (a) shows the cumulative curves for the 16 states of Germany. Panel (b) shows the cumulative curves from Germany and other 142 countries. We highlight the countries with very high number of cases – China, Italy, USA, UK, France, Spain, and Iran.

used sensors to identify the type of vehicle, which we include in our analysis to estimate the number of individuals. We construct correction factors to extrapolate hourly traffic for January 1, 2020, to April 30, 2020 (details in Section 3.6), and the model includes public holidays, day of the week, and state population as control variables. According to a survey, 61% of people in Germany use cars to commute to their workplace [124]. So, to control for changes in car movement during the period of the study, we adjust the predicted daily car movement between states using Google's community mobility data for workplaces (Figure 3.5).

We used Deutsche Bahn's timetables (www.bahn.com) for all major train stations in Germany to estimate the number of daily rail travelers moving across states and arriving from neighboring countries. To account for the changes in movement due to COVID-19 and cancelations of several trains, we adjust the number of passengers moving across states by using the community mobility data for transit stations (Figure 3.5).

We obtained the search history of a large European bus and train comparison platform to estimate the number of passengers moving across cities and states in Germany and passengers traveling to Germany from neighboring countries. We set bus transport to zero after March 16, 2020, as all bus movement in Germany stopped on that day.

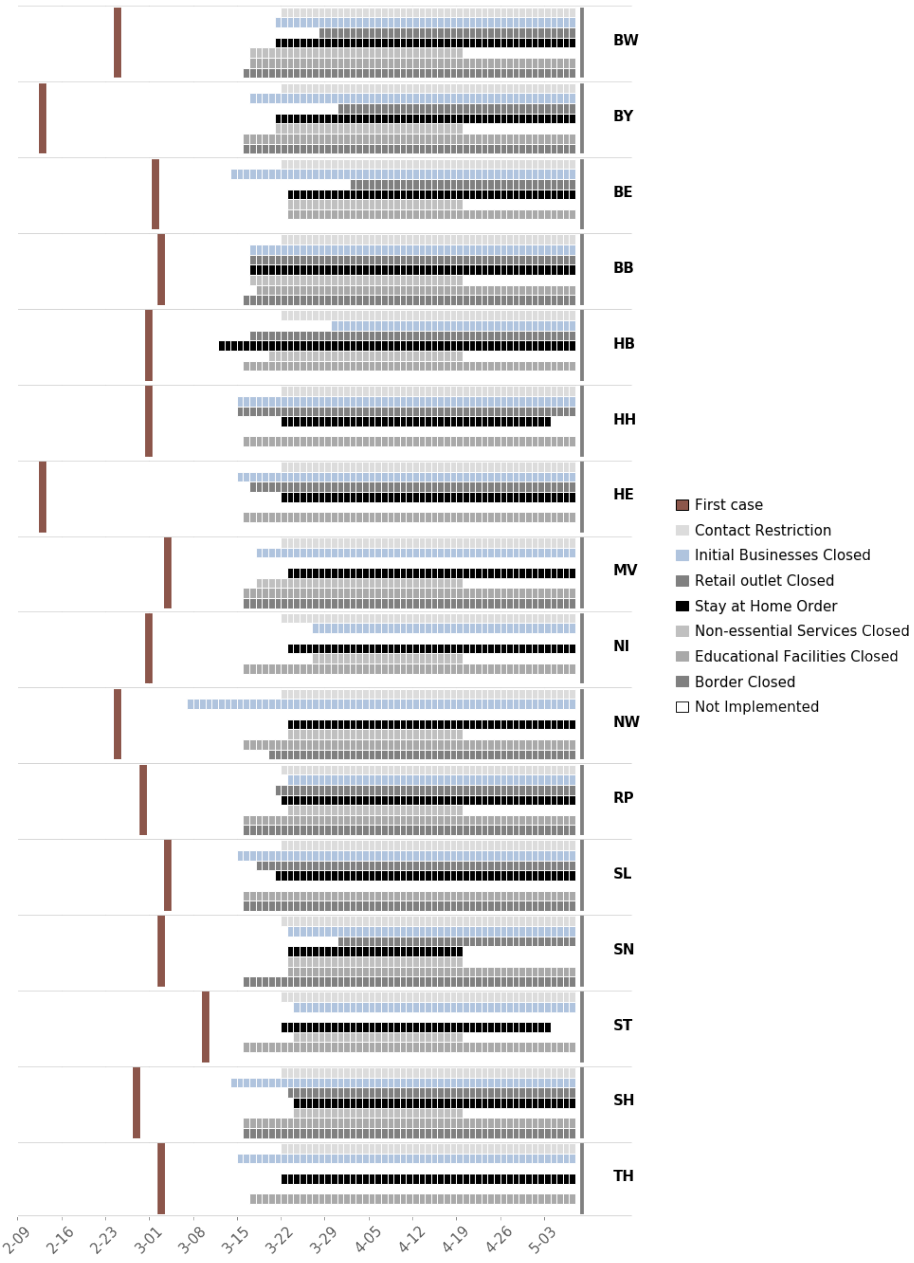
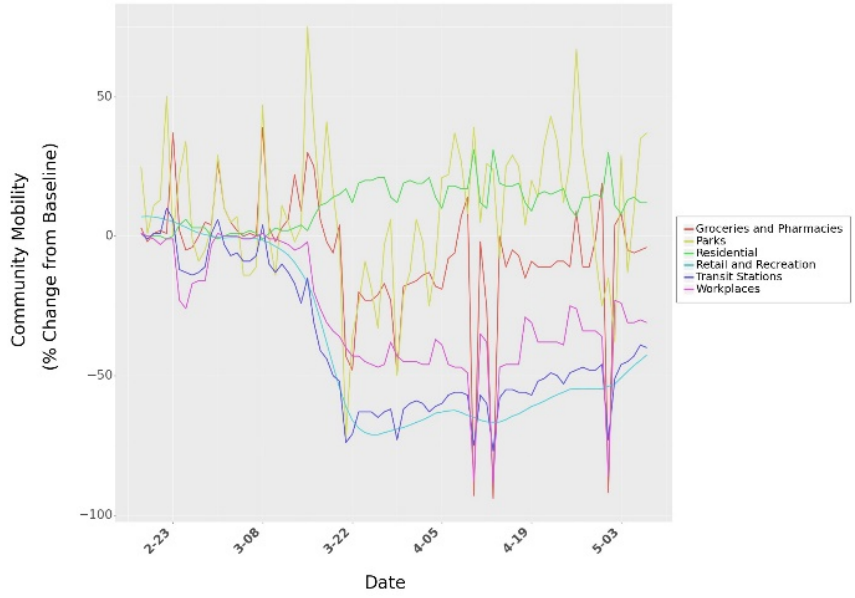
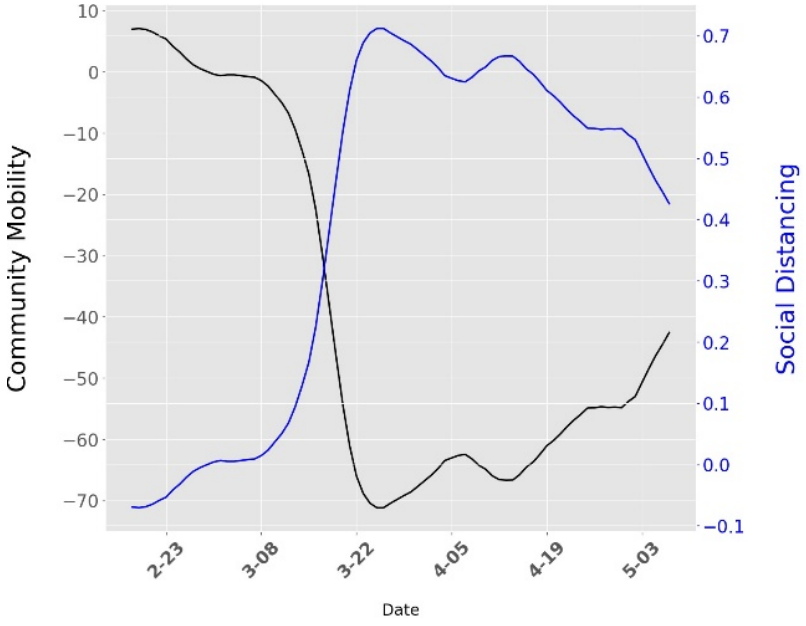


Figure 3.3: Timeline for implementation of NPIs across states. The policies are visualized in the same order from top to bottom for each state. These policies include, in this order: contact restrictions (movement in public space is limited to two persons or people co-living), initial business closures (e.g., restaurants), retail outlet closures, stay at home orders, non-essential business closures (e.g., trade shows), closure of educational institutes (e.g., schools and universities), and border closures (closing international borders). Border closures apply to 10 states sharing international borders. In our study, we use data from February 18, 2020, to May 7, 2020. Each state did not implement every policy as of April 20, 2020, and none of the implemented policies were relaxed until April 20. State governments started relaxing these policies after April 20, 2020.

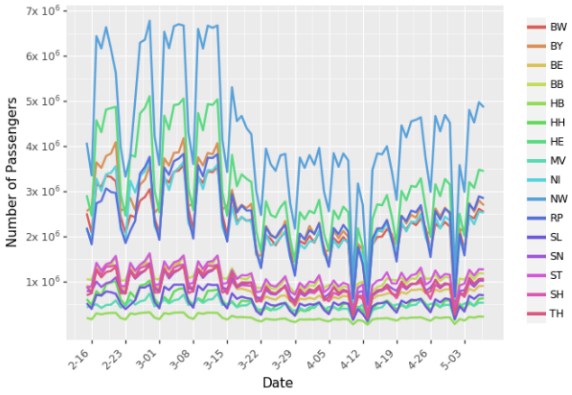


(a) Community Mobility in Bavaria

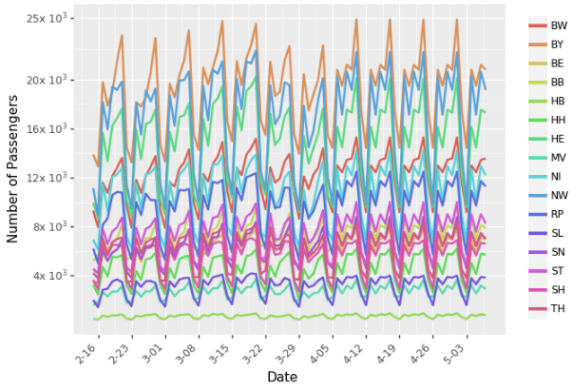


(b) Mobility and Social Distancing in Bavaria

Figure 3.4: Community mobility and social distancing. Community mobility data charts the difference in foot traffic in different locations as compared to baseline from historical data. Panel (a) shows the movement trend across different categories of places for Bavaria. Similar movement trends for all the other states are shown in the Section 3.6 Material. Panel (b) shows the community mobility and social distancing in Bavaria.



(a) Car Movement

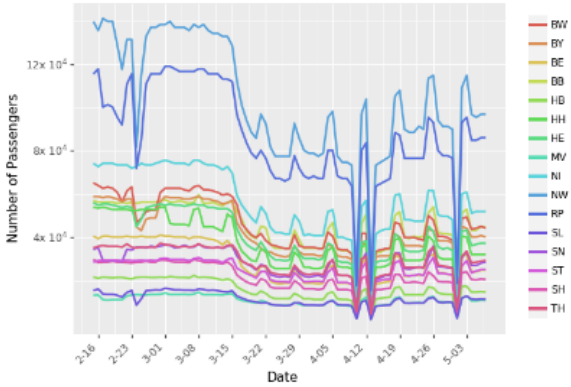


(b) Truck Movement

Abfahrt *Departure* **Berlin Hbf**

Zeit *Time*	Zug *Train*	in Richtung *Destination*	Gleis *Track*
		0:00	
0:16	**ICE** 948	Brandenburg 1:06 – Magdeburg 1:47 – Braunschweig 2:34 – Hannover 3:09 – Minden 4:10 – Bielefeld 4:34 ⊙ Hamm(Westf)Hbf 5:10 – Dortmund 5:35 – Essen 6:02 – Duisburg 6:16 – Düsseldorf 6:30 – **Köln 6:57** ➔ 14. bis 18. Jan kein Halt in Brandenburg, Magdeburg	7
0:26 Mo*	**IC** 1950 (R)	Lu. Wittenberg 1:11 – Bitterfeld 1:28 – Halle(Saale) 1:48 – Leipzig 2:13 – Weimar 3:35 – Erfurt 3:51 – Gotha 4:14 – Eisenach 4:30 ⊙ Fulda 5:38 – Hanau 6:22 – **Frankfurt (M) Hbf 6:40** *auch 2. Jan, 14. Apr, 2. Jun; nicht 13. Apr, 1. Jun	1
0:26	**RE 7** **RE** 3743 (RB)	B Zoologischer Garten 0:30 – B-Charlottenburg 0:34 – B-Wannsee 0:44 – P Medienstadt Babelsberg 0:49 – P-Rehbrücke 0:52 ⊙ Michendorf 0:59 – Seddin 1:03 – Beelitz-Heilstätten 1:07 – **Bad Belzig 1:27**	14
0:41	**RE 1** **RE** 3154 (ⓐ RB)	B Zoologischer Garten 0:46 – B-Charlottenburg 0:50 – B-Wannsee 0:59 – Potsdam Hbf 1:06 – P Charlottenhof 1:10 – P Park Sanssouci 1:13 – Werder 1:19 ⊙ Brandenburg 1:37 – Wusterwitz 1:49 – Genthin 1:57 – **Magdeburg 2:28** ➔ Mo – Fr, nicht 26. Dez, 1. Jan, 10., 13. Apr, 1., 21. Mai, 1. Jun nur bis **Brandenburg 1:37** ➔ Brandenburg – Magdeburg als **RE** 3156	14
0:42	**RE 3** **RE** 3363 (ⓐ RB)	B Potsdamer Platz 0:44 – B Südkreuz 0:49 – B-Lichterfelde Ost 0:56 – Teltow 1:01 – Großbeeren 1:05 ⊙ Ludwigsfelde 1:12 – Thyrow 1:17 – Trebbin 1:21 – Luckenwalde 1:33 – Jüterbog 1:41 – Lu. Wittenberg 1:58 – Dessau 2:24 – Bitterfeld 2:39 – **Halle(Saale) 2:57** ➔ Mo – Fr, nicht 25., 26. Dez, 1. Jan, 10., 13. Apr, 1., 21. Mai, 1. Jun nur bis **Jüterbog 1:41**	3
1:58 Sa, So*	**RE 3** **RE** 3366 (ⓐ RB)	B Gesundbrunnen 2:02 – Bernau 2:18 – Rüdnitz 2:23 ⊙ Melchow 2:31 – **Eberswalde 2:39** *auch 25., 26. Dez, 1. Jan, 10., 13. Apr, 1., 21. Mai, 1. Jun	6
		2:00	
2:15 Sa, So*	**RE 7** **RE** 3745 (RB)	B Zoologischer Garten 2:20 – B-Charlottenburg 2:23 – B-Wannsee 2:33 – P Medienstadt Babelsberg 2:38 – P-Rehbrücke 2:42 ⊙ Michendorf 2:49 – Seddin 2:53 – Beelitz-Heilstätten 2:57 – **Bad Belzig 3:16** *auch 25., 26. Dez, 1. Jan, 10., 13. Apr, 1., 21. Mai, 1. Jun	14
2:32 Sa, So*	**RE 1** **RE** 3143 (ⓐ RB)	B Friedrichstraße 2:35 – B Alexanderplatz 2:38 – B Ostbahnhof 2:43 – Berlin Ostkreuz 2:48 – Erkner 3:04 – Fangschleuse 3:09 – Hangelsberg 3:13 – Fürstenwalde 3:19 ⊙ **Frankfurt (O) 3:48** *auch 25., 26. Dez, 1. Jan, 10., 13. Apr, 1., 21. Mai, 1. Jun	11
2:32 Sa, So*	**RE 3** **RE** 3365 (ⓐ RB)	B Potsdamer Platz 2:34 – B Südkreuz 2:39 – B-Lichterfelde Ost 2:44 – Teltow 2:49 – Großbeeren 2:53 ⊙ Ludwigsfelde 2:59 – Thyrow 3:04 – Trebbin 3:08 – Luckenwalde 3:19 – **Jüterbog 3:29** *auch 25., 26. Dez, 1. Jan, 10., 13. Apr, 1., 21. Mai, 1. Jun	5
2:41 Mo*	**ICE** 781	Brandenburg 3:21 – Magdeburg 4:04 – Braunschweig 4:52 – Hildesheim 5:16 – Göttingen 5:48 – Kassel-Wilhelmshöhe 6:21 – Fulda 6:54 – Würzburg 7:28 – Nürnberg 8:24 – Ingolstadt 8:58 – **München 9:38** ⊙ *auch 1., 2. Jan, 14. Apr, 2. Jun; nicht 13. Apr, 1. Jun	13
2:42	**RB 14** **RB** 18647 **RB** 18643	**AIRPORTEXPRESS** B Friedrichstraße 2:44 – B Alexanderplatz 2:47 –	12

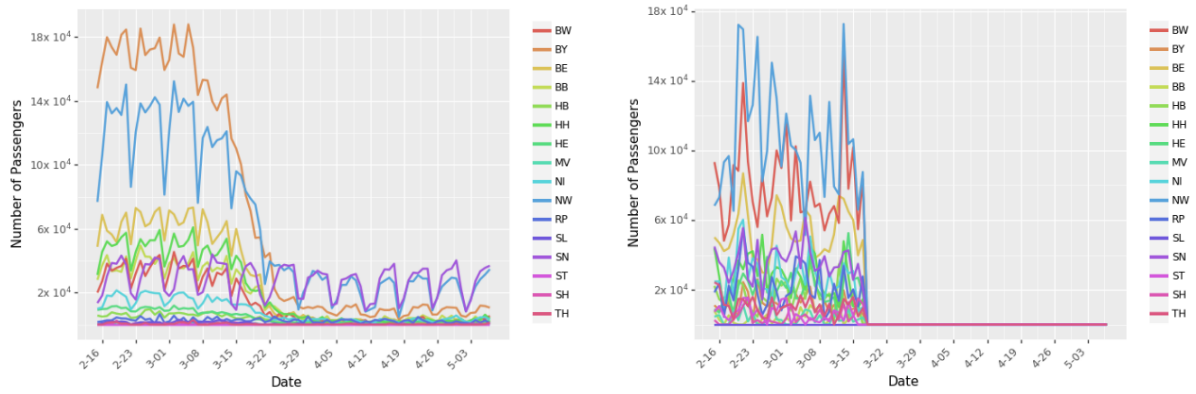
(c) Departing from Berlin Central Station



(d) Train Movement

Figure 3.5: Ground transportation using cars, trucks, and trains in Germany. Panel (a) shows the number of passengers arriving in different states by car. Panel (b) shows the number of passengers arriving in different states by truck. During the period of this study, there was no restriction on truck movement. Panel (c) shows part of the train timetable available in all German train stations. More specifically, the sample snippet displayed in Panel (c) lists the departure times of trains leaving Berlin Hbf (Berlin Central Station). We parsed 538 timetable files to obtain the train schedule for all of Germany. We combine information from the arrival and departure timetables to construct the complete route of a train. Panel (d) shows the number of passengers arriving to different states by train.

We also obtained flight transportation information from the Opensky Network [125], whose database utilizes Automatic Dependent Surveillance Broadcast (ADS-B) flight trajectories to identify the departure and arrival airport of flights (Figure 3.6).

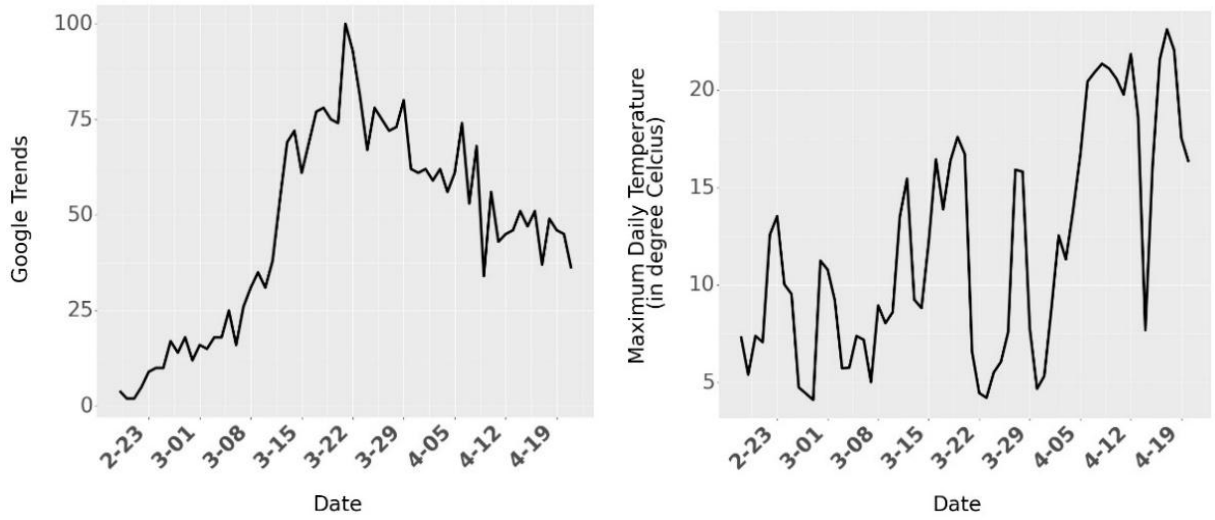


(a) Flight Movement

(b) Bus Movement

Figure 3.6: Transportation data in Germany. Panel (a) shows the number of air passengers arriving in different states in Germany. We assume 200 passengers for domestic and 500 passengers for international flights. Panel (b) shows the number of bus passengers arriving to different states from other states in Germany and neighboring countries. We assume 20 passengers per trip.

We also use additional controls to isolate the effect of individual policies. We use Google Trends data for the search term "COVID-19" to control for awareness over time (Figure 3.7a). Severe movement restrictions due to increased enforcement of these policy restrictions led to a greater sense of unease and dissatisfaction amongst some sections of the population (e.g., [126]. While such protests are small, prolonged enforcement of restrictions could increase dissatisfaction). We use weather data (maximum daily temperature) from wetterkontor.de as a control to account for the propensity of the population to leave their homes as the weather improves (Figure 3.7b). Finally, to account for unobserved heterogeneity across states and over time, we include state, week, and day of week fixed effects. Including these effects allows us to control for variations that the model does not include explicitly.



(a) Google Trends in Bavaria

(b) Daily Temperature in Bavaria

Figure 3.7: Covariates in Linear Regression Model. Panel (a) shows the Google Trends for the search term "COVID-19" for Bavaria. Google Trends numbers indicate search interest of a topic over time as a proportion of all other searches at the same time. We use Google Trends to account for increased awareness over time. Panel (b) shows the maximum temperature recorded in a day for Bavaria. We show the Google Trends and Maximum Daily Temperature for other states in the Section 3.6 (Figures 3.18 and 3.19 respectively).

3.2 Models

3.2.1 Susceptible-Exposed-Infected-Recovered/Removed (SEIR) Model

The SEIR compartmental model is a transmission model of an infectious disease. Inspired by [39], our proposed SEIR transmission model integrates the effect of human mobility

on disease spread through a set of evolutionary equations. The model is as follows:

$$\frac{dS_i}{dt} = -\frac{(1-\gamma\times sd_i)\beta S_i I_i^d}{N_i} - \frac{(1-\gamma\times sd_i)\mu\beta S_i I_i^u}{N_i} + \sum_\nu \left(\frac{\theta_\nu \sum_j M_{ij}^\nu \times S_j}{N_j - I_j^d} - \frac{\theta_\nu \sum_j M_{ji}^\nu \times S_i}{N_i - I_i^d}\right) \tag{3.1}$$

$$\frac{dE_i}{dt} = \frac{(1-\gamma\times sd_i)\beta S_i I_i^d}{N_i} + \frac{(1-\gamma\times sd_i)\mu\beta S_i I_i^u}{N_i} - \frac{E_i}{Z} + \sum_\nu \left(\frac{\theta_\nu \sum_j M_{ij}^\nu \times E_j}{N_j - I_j^d} - \frac{\theta_\nu \sum_j M_{ji}^\nu \times E_i}{N_i - I_i^d}\right) \tag{3.2}$$

$$\frac{dI_i^d}{dt} = \frac{\alpha_i E_i}{Z} - \frac{I_i^d}{D} \tag{3.3}$$

$$\frac{dI_i^u}{dt} = \frac{(1-\alpha_i)E_i}{Z} - \frac{I_i^u}{D} + \sum_\nu \left(\frac{\theta_\nu \sum_j M_{ij}^\nu \times I_j^u}{N_j - I_j^d} - \frac{\theta_\nu \sum_j M_{ji}^\nu \times I_i^u}{N_i - I_i^d}\right) \tag{3.4}$$

$$N_i = N_i + \sum_\nu \left(\theta_\nu \sum_j M_{ij}^\nu - \theta_\nu \sum_j M_{ji}^\nu\right) \tag{3.5}$$

where S_i, E_i, and N_i represent the susceptible, exposed, and the total population of state i, respectively. I_i^d represents the documented infected individuals, the subset of the infected population with severe enough symptoms to be diagnosed with the illness. I_i^u is the rest of the infected population known as the undocumented infected individuals. We consider values for SEIR metapopulation state variables on the day t. For the rest of this chapter, we omit the time index t that represents the time dependency of variables. β is the transmission rate of the disease from a documented infected individual to a susceptible individual under normal population mobility. The transmission rate due to undocumented infected is captured by $\mu\beta$ with μ being the reduction coefficient ($\mu < 1$). Additionally, the variable sd_i ($sd_i < 1$) defines the daily change in social mobility (or degree of social distancing) in the state i, and the coefficient $(1 - \gamma \times sd_i)$ is the decrease/increase in transmission rate due to changes in socializing and population mobility ($\gamma \leq 1$). The ratio of documented to total infected individuals in the state i is α_i, which varies among states with different population demographics based on age and gender [127]. Z is the average incubation time, and D is the infection period. More precisely, D captures the effective period in which the infected individual moves out of the chain of disease transmission by perishing, recovering, or entering quarantine. We also control for the potential of

spillovers between states by accounting for interstate travel. The number of interstate travelers from state j to state i via transportation network ν, is M_{ij}^{ν} on a given day, with θ_ν to fix the underreporting of transportation ($\theta_\nu \geq 1$). Two transportation networks represented by G and A account for ground and air mobility. The ground transportation M_{ij}^{G} includes movement of individuals by four different sub-networks: $M_{ij}^{G} = \sum_g M_{ij}^{g}$. These sub-networks are cars, trains, trucks, and buses. Similarly, M_{ij}^{A} is the number of people traveling via flights. In this model, we assume documented infected patients do not travel between states, while the asymptomatic undocumented infected individuals can move from one state to another. Finally, we note that the SEIR model works under the assumption of exponentially distributed incubation and infectious times, and the relative change in the total population is negligible for the period of the disease spread.

We estimate the parameters of this model via Ensemble Adjustment Kalman Filter (EAKF), which is suitable for models with a high number of parameters. We also adjust for various delays (e.g., the latency period from the onset of symptoms to diagnosis, the time between the exposure and becoming contagious, and the incubation period). Further details are provided in Section 3.6.1 and Table 3.2. We use the parameters estimated from the SEIR model to predict the number of new case counts under the different scenarios we consider for policy (NPIs) easing.

3.2.2 Linear Regression: NPIs and Social Mobility

We use a linear regression model to estimate the association of NPIs with community mobility ($C_{j,t}$) as shown in Equation 3.6 where K is a constant. We use a binary variable $x_{j,p,t} = 1$ if policy p is active in state j on day t, $trend_{j,t}$ is the exponentially smoothed Google Trends number for search term "COVID-19," and $temp_{j,t}$ is the daily maximum temperature in degree Celsius in state j on day t. To account for state-level unobserved heterogeneity, we use state fixed effects ($state_j$). In addition to state-level differences, we also control for week-based differences and day of week-based differences by incorporating week fixed effects ($week_t$) and day of week fixed effects (day_t). For stable parameter estimation, we consider state Thuringia as our base state (0 state fixed effects), week 9 as our base week (0 week fixed effect) and Monday as our base day (0 day of week

fixed effect). Fixed effects in Equation 3.6 are modeled as indicator functions. (details in Section 3.6.3).

$$C_{j,t} = K + \sum_{p=1}^{7} \beta_p^{policy} x_{j,p,t} + \beta_t^{trend} trend_{j,t} + \beta_m^{temp} temp_{j,t} + \sum_{j=1}^{16-1} \beta_j^{state} state_j + \sum_{w=1}^{9-1} \beta_w^{week} week_t + \sum_{d=1}^{7-1} \beta_d^{day} day_t + \epsilon_{j,t} \tag{3.6}$$

This specification models states to be independent, which may lead to the failure of the SEIR model in capturing spillovers across states. To address spillovers from different states, we use movement across states in the SEIR model (car, bus, train, truck, and flight) and state fixed effects. We assume that Google Trends might be able to capture spillover effects (if cases increase in one state, the population in neighboring states will be more self-aware). Further details are presented in Section 3.6.3.3.

3.2.3 Limitations: Association as opposed to Identification

Our approach rests on spatial-temporal variations in NPIs contributing to variations in social distancing, which, in turn, is linked to variations in COVID-19 case numbers. We do not claim causality when analyzing this chain of events, as there are ways in which causal identification may fail. For example, if the data do not provide adequate variation in the sequencing of policy interventions, then our estimates cannot uniquely identify the effect of each intervention. Figure 3.4 shows, for example, that Contact Restriction never occurred first in any state, and thus our associative effect estimate likely contains bias. Combinatorically, for the seven NPIs observed, there exist 13,700 possible sequences (i.e., there are $\sum_{k=0}^{7} P(7,k) = 13,700$ possible orders / permutations in which the seven NPIs are implemented in each state). However, in real-world settings, most countries adopted similar sets of policies within a short period despite these vast sets of possibilities. For example, Sebhatu et al. [23] note that "almost 80% of OECD countries adopted the same COVID-19 NPIs within a span of 2 w[ee]k[s]." They find that policymakers set policies based on "...the number of earlier adopters in the same region rather than accounting for country-specific characteristics." All policy level analyses suffer from estimation biases

due to the lack of variation in the adoption and implementation of policies despite the heterogeneity of country and state characteristics. Even though we observe temporal and sequencing differences across states, the variation is not enough for us to attribute causality to these interventions. Ultimately, we would like to examine interventions at a more granular, perhaps even individual level (e.g., assess the probability of a person contracting COVID-19 when certain NPIs are enacted) to obtain a greater variation in the sequencing of interventions. Unfortunately, this is not how the policy decisions work: they cover wide jurisdictions, and their historical sequence is limited.

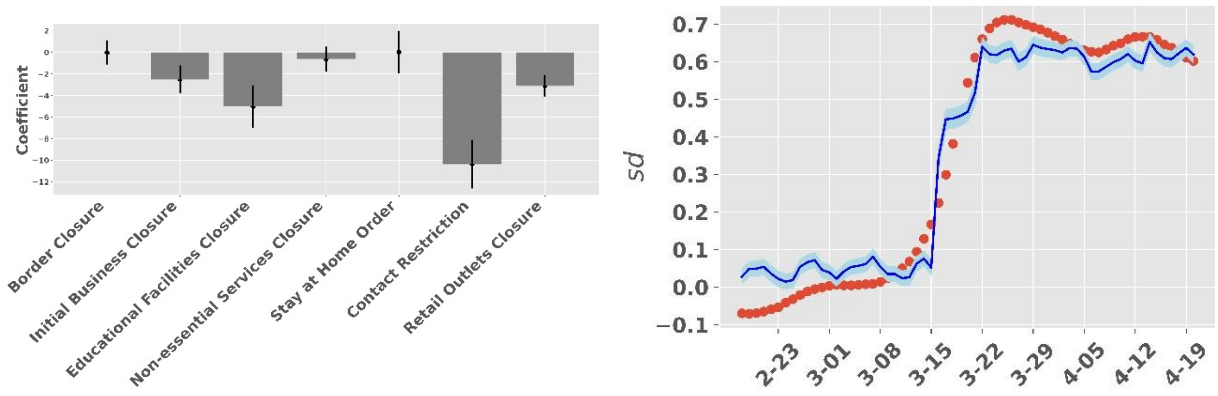
A second thread to identification is spillover effects across states that may result in biased estimates. We try our best to account for spillovers between states in the following ways: (i) implementing most interstate movement within the SEIR model (car, flight, bus, train, and truck), (ii) including Google Trends in the linear regression model (people in a state would be more cautious and search for "COVID-19" more frequently if case numbers rise in neighboring states), and (iii) including state fixed effects to account for state-level heterogeneity (may help alleviate potential spillover of a particular state from its neighbors). Moreover, Section 3.6.3.3 addresses the issue of potential spillover across states in more detail by evaluating two alternative specifications. However, we still cannot completely rule out the possibility of spillovers.

Due to the data-imposed inability to identify the causal impact of NPIs on COVID-19 cases, this study only proposes associative effects for each NPI.

3.3 Results

3.3.1 Quantifying Policy Contributions

To determine the impact of different state policies, we use a linear regression model to predict changes in mobility ($C_{j,t}$) due to policy p. Coefficients from the linear regression model are shown in Figure 3.8a. We calculate social distancing ($sd_{j,t}$) from mobility in Figure 3.8b. As explained above, we cannot identify each policy's causal effect on reducing disease spread. However, we can still provide meaningful insights into each policy's contribution in reducing the disease spread.



(a) Linear Regression Coefficients

(b) Predicting Social Distancing for Bavaria

Figure 3.8: Linear regression model for estimating the marginal contribution of different NPIs on social distancing. Panel (a) shows the linear regression coefficients for different NPIs. The vertical bars show the 95% confidence interval bounds for coefficients. Our regression model has an R-square value of 0.978. Panel (b) shows the smoothed social distancing (red dots) and predicted social distancing (blue line) for Bavaria. Blue shade shows the 95% confidence bounds around prediction for social distancing. The plots for all other states are included in the Section 3.6 (Figure 3.20).

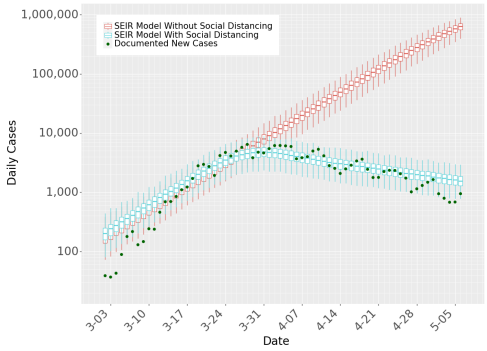
Figure 3.8a shows the parameter estimates of various NPIs on social mobility. Complete results for the parameter estimates are shown in Section 3.6 (Table 3.4). The estimates capture the average incremental effect of these policies across Germany's 16 states in reducing mobility. The graph shows that contact restrictions have a substantial negative effect on mobility (10.3 percentage point drop in mobility, 95% Confidence Interval (CI): 8.2 to 12.5). COVID-19 is mainly spread through airborne transmission among individuals in close proximity. The policy of contact restrictions limits the number of people that may congregate and requires all people to stay a minimum of 6 feet apart from one another to ensure social distancing. Educational Facilities Closure also precipitates a significant drop in mobility (5 percentage point drop; CI: 3.1 to 6.9). A significant drop is possible as it indirectly leads to decreased mobility (for example, children staying home without caregivers forces parents to work from home). The other policies listed in order of impact on mobility (estimate of the percentage point drop in mobility from the regression and 95% Confidence Interval in parenthesis) are Retail Outlets Closures (3.1, CI: 2.1 to 4.1), Initial Business Closures (3.4; CI: 1.2 to 3.7), Border Closures (0.03, CI: -1 to 1), Non-essential Services Closures (0.62; CI: -0.5 to 1.7), and Stay at Home Orders (-0.01;

CI: -1.9 to 1.8). We observe that parameter estimates for Border Closures, Non-essential Services Closure, and Stay at Home Orders are statistically insignificant (confidence intervals include 0). However, it does not conclude that these NPIs had no effect. For example, international travel might have played a significant role in seeding the disease. However, once the Borders were closed, the Border Closure NPI played a minimal role in mitigating the spread of the disease. As discussed in the previous section ("Limitations: Association as opposed to Identification"), we estimate associations as opposed to causal effects.

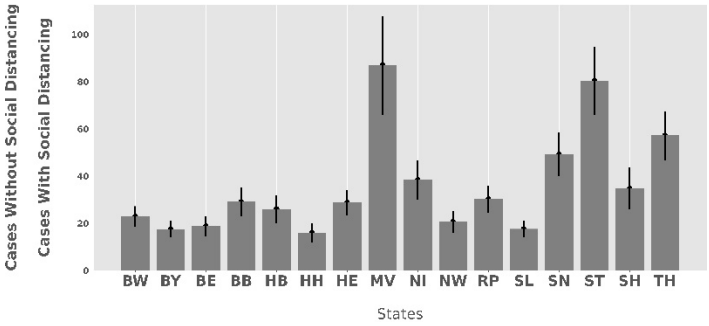
3.3.2 Predicting Disease Spread

We use the predicted mobility from the linear regression to predict new case counts. We investigate the contribution of each policy to the mitigation of disease spread by determining the role of social distancing in the estimation of the number of susceptible and exposed individuals in a given population. We modify the SEIR model used in [39] to include different transportation networks and predicted mobility for each state (Equations (3.1) to (3.5)). Using the estimation procedure in [39], we find the model parameters to predict disease spread for all 16 states in Germany — the model accounts for documented as well as undocumented infected cases. As shown in Figure 3.29, the proportion of documented infected (I^d) as a function of total cases increases over time. This finding aligns with expectations because of the rapid increase in testing across Germany [128].

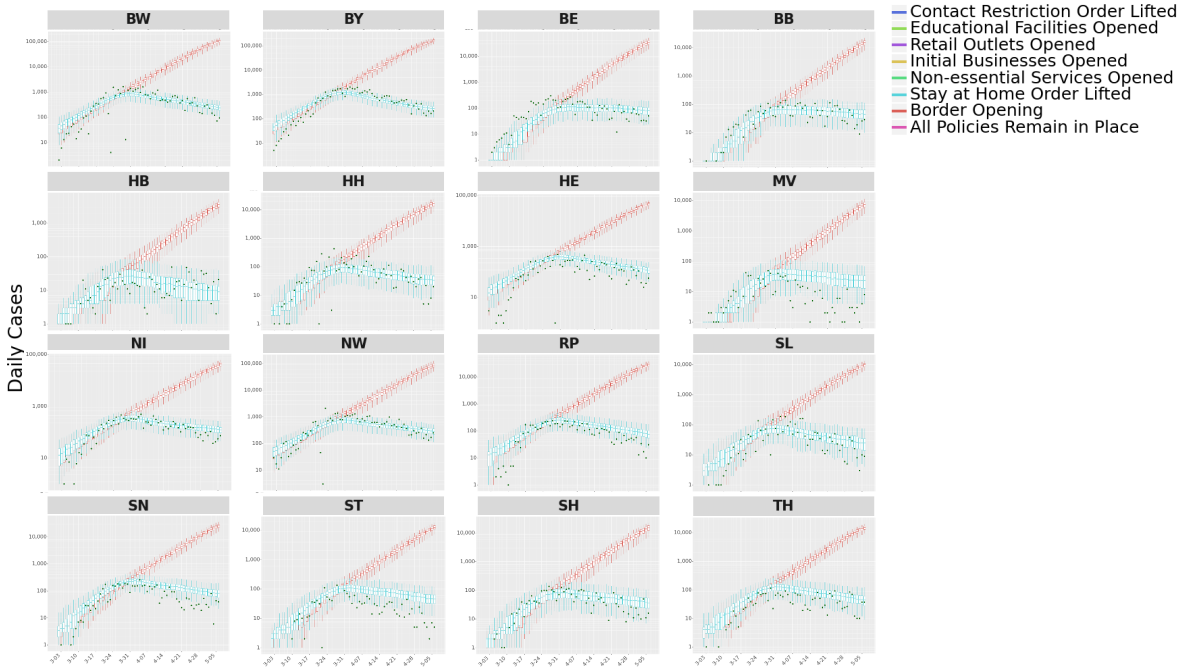
Figure 3.9a shows the actual disease progression in Germany; the disease spread as predicted by our model in the presence of predicted social distancing, as well as disease spread as predicted by our model when mobility remained unchanged with no social distancing measures. Similar predictions for the states are provided in Figure 3.9c. We use the period of February 18, 2020 – April 20, 2020, to infer model parameters. This period includes the early stages of the COVID-19 epidemic in Germany and the time that state policies are enacted. We use these parameters to estimate the number of daily documented cases during the time interval of February 18, 2020 - May 7, 2020, which corresponds to 17 days out of sample forecasts. The model finds 172,922 (IQR: 140,301 – 204,952) cumulative documented cases in Germany as of May 7, 2020 (actual reported



(a) Disease Progression in Germany



(b) Change in Disease Progression without Social Distancing



(c) Disease Progression across the 16 states of Germany

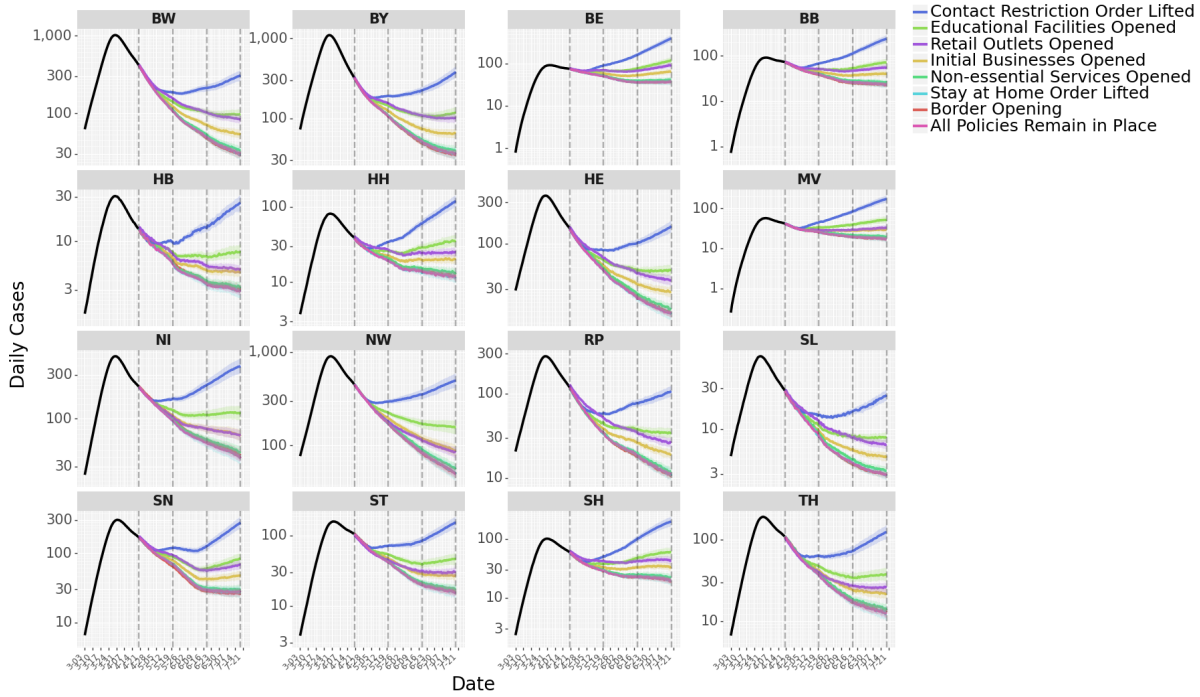
Figure 3.9: SEIR model predictions. Panel (a) shows the predictions of new case counts in Germany with and without any distancing measures, against actual cases for the same time period. Panel (b) presents the expected increase of cumulative cases (in multiples) when there is no social distancing as compared to the actual cases numbers. Panel (c) plots the predictions of new case counts with and without any distancing measures, against actual case numbers for all states. The box and whiskers provide the median, interquartile range, and up to a maximum of 1.5*IQR from the box hinges derived from 4,000 simulations using the inferred model parameters. Our predictive accuracy is higher for states with high case numbers such as Baden-Württemberg (BW) and Bavaria (BY) than for the small city-states Berlin (BE), Bremen (HB), and Hamburg (HB). We note that the last 17 days of Panels (a) and (c) correspond to out-of-sample forecasts from the SEIR model.

cases: 166,069) with the estimated average error rate of 4%. Figure 3.9b shows the amount of expected increase in the number of cases across the states of Germany and the nation if no social distancing was practiced. Across Germany, one would expect a 24.6-fold (IQR: 20 to 29) increase in the number of cases without any social distancing (i.e., $sd_i = 0$), the effect varying significantly by states from a low of 16-fold (IQR:12.2 to 19.7) in Hamburg to a high of 86-fold (IQR: 66.3 to 107.3) in Mecklenburg-Vorpommern.

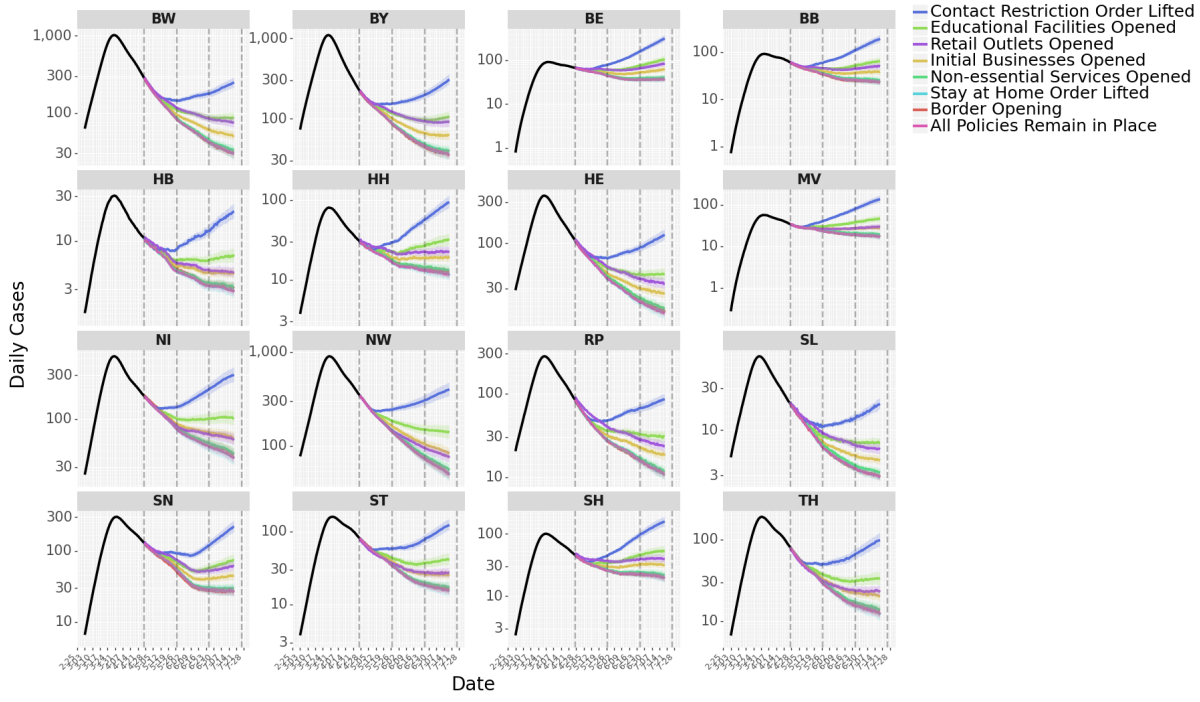
3.3.3 The Lifting of Restrictions

We simulate what-if scenarios to determine the impact of lifting restrictions on new cases in each state. Under the scenario that a restriction has been relaxed while others remained operational, we forecast mobility using Equation 3.6, and we subsequently project the new case count using the predicted mobility. We repeat this step across all restrictions, each relaxed individually.

In Scenario 1, we project all changes from April 21, 2020, to July 19, 2020, assuming that the restriction was relaxed on April 21, 2020. Figures 10(a) and (c) show the projections of case counts over a 90-day period if a restriction was relaxed exclusively. Figures 10(b) and (d) show Scenario 2, which projects all changes from April 21, 2020, to July 19, 2020, assuming that the restriction was relaxed on April 28, 2020. Because the two scenarios are exactly one week apart, it allows us to determine the impact of delaying lifting a restriction by one week. From the analysis, the lifting of contact restrictions, i.e., the rule limiting movement in public spaces, had the most significant impact on new case counts. Compared to keeping the restrictions in place, lifting contact restrictions is associated with a 150% (IQR: 144-156%) increase in daily case numbers in Scenario 1 and a 108% (IQR: 103.7-112.5%) increase in Scenario 2. However, lifting educational facilities closure is associated with an increase in daily case numbers by 46.1% (IQR: 44.0-48.1%) in Scenario 1 and 34.4% (IQR: 32.7-36.2%) in Scenario 2. Opening the retail outlets is linked to a 33.9% (IQR: 33.0-34.8%) increase in daily case numbers in Scenario 1 and a 24.5% (IQR: 23.4-25.6%) increase in Scenario 2. Lifting restrictions on initial business closures is associated with an increase in daily case numbers by 18.6% (IQR: 17.8-19.5%) in Scenario 1 and 14.4% (IQR: 13.7-15.0%) in Scenario 2, and easing non-essential service

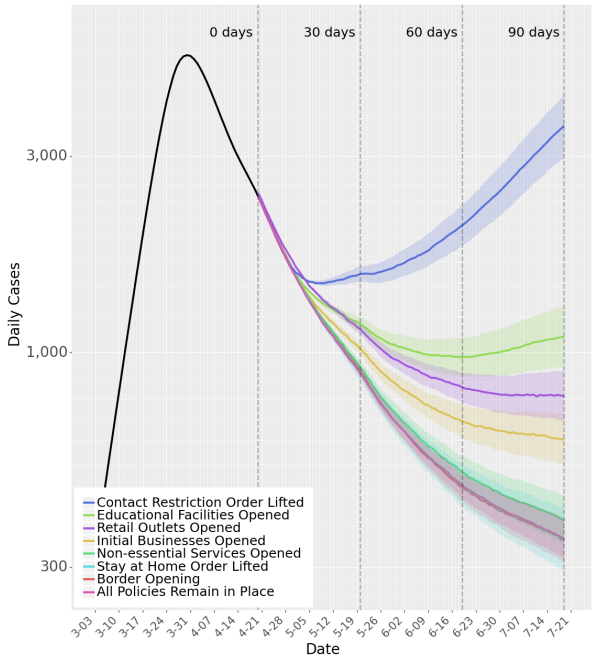


(a) New Cases Across All States due to Policy Easing from April 21, 2020

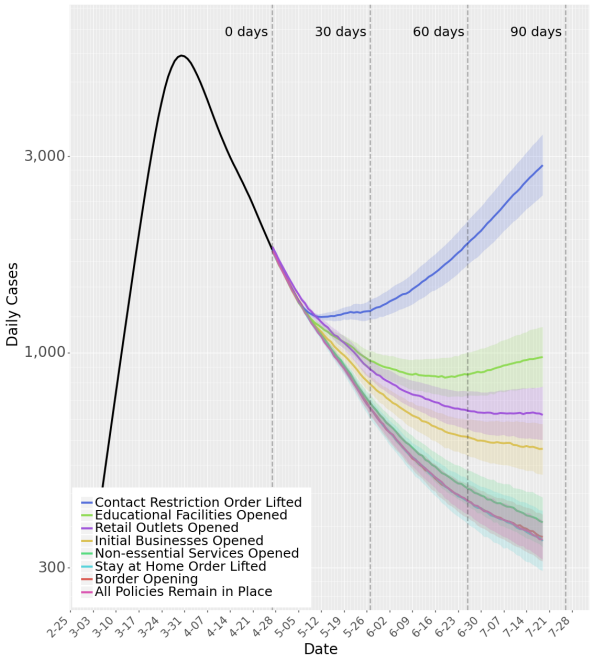


(b) New Cases Across All States due to Policy Easing from April 28, 2020

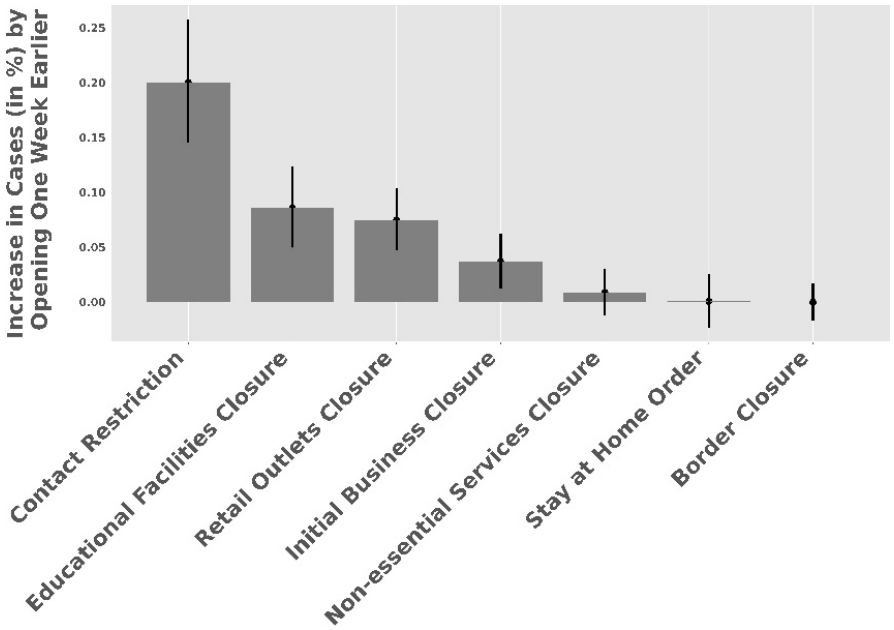
Figure 3.10: Effect of lifting policy restrictions (continued on the next page.)



(c) New Cases in Germany due to Policy Easing from April 21, 2020



(d) New Cases in Germany due to Policy Easing from April 28, 2020



(e) Predicted Increase in Cases due to Policy Easing a Week Earlier

Figure 3.10: Lifting of policy restrictions (cont.): Panels (a) and (b) show the predictions of new case counts across all states of Germany due to lifting of policy restrictions from April 21, 2020 to July 19, 2020. In Panel (a) policy restrictions are relaxed, one-at-a-time on April 21, 2020, in Panel (b) policy restrictions are relaxed, one-at-a-time on April 28, 2020. Panels (c) and (d) illustrate the same for all of Germany. Panel (e) charts the predicted increase in cases due to policy easing a week earlier.

closures is associated with an increase in daily case numbers by 3.6% (IQR: 3.1-4.1%) in Scenario 1 and 2.6% (IQR: 2.2-3.0%) in Scenario 2. These results show that NPIs have differential impacts on lowering disease spread and suggest a measured approach to lifting restrictions. For example, the opening of retail outlets could be balanced by maintaining the restrictions around limiting the number of individuals in a given place or store (e.g., controlling entry) – thereby allowing for the resumption of economic activity while limiting the risk of contagion.

Figure 3.10e shows the increases in the expected number of cases if a restriction was lifted on April 21 versus on April 28. Delaying the lifting of certain restrictions by one week could also significantly impact the total case counts. This impact occurs due to the delay and because the number of infected individuals that a person could come in contact with decreases over the week. For example, delaying the lifting of contact restrictions by one week is associated with a reduction in the number of new cases over the 90-day forecast by an average of 20% (IQR: 14.6% to 25.7%). We also observe that lifting restrictions on educational facilities closure and retail outlets closure is linked to an average of 8.9% (IQR: 5.1% to 12.2%) and 7.5% (IQR: 4.8% to 10.2%) increase in total case numbers over the 90-day forecast period.

3.4 Discussion

This study explores the role of NPIs in reducing the spread of COVID-19. We extend the spatio-temporal SEIR model in [39] by incorporating daily social distancing numbers from transportation data and mobility patterns. While our analysis does not allow for causal identification, each policy's association with the mitigation in disease spread provides meaningful insights. We first relate the NPIs to social mobility changes across the 16 states. Next, we link these changes in social mobility to the spread of COVID-19 by reconstructing patterns of disease spread across Germany's 16 states. Using this link with social mobility changes, we investigate the marginal association of each NPIs implemented by state governments in Germany with disease mitigation. The model suggests that without NPIs, COVID-19 cases may have shown a 24.6-fold increase across Germany as

of May 7, 2020. Finally, we forecast the number of new cases when policies are relaxed one at a time, suggesting that specific policies have a more significant impact on disease spread than others. Our model forecasts suggest that early relaxation of some NPIs could increase the number of cases, potentially leading to a second wave. An estimated increase in the effective reproduction number confirms this observation (Section 3.6). We also compare case counts for policies relaxed with a one-week delay; keeping some NPIs in place for an extra week is associated with reducing COVID-19 cases by up to 20% (as of July 19, 2020). The results confirm that policy restrictions are not all equal in affecting disease spread. The policy restricting mass gatherings (Contact Restriction) is estimated to be the most effective NPI to contain COVID-19, followed by closures of various businesses and stay-at-home orders. Due to this variation in effect, it is advisable to lift restrictions with minimal effects first, gradually easing restrictions that potentially lead to higher case numbers. This study presents a comprehensive quantitative analysis that includes individual effects of NPIs on the transmission of COVID-19. To the best of our knowledge, this is the first study that uses variations in policy interventions by governments to discover their differential impacts at reducing mobility, which, in turn, reduces disease spread. Prolonged lockdowns and restrictive policies can have devastating social and economic consequences; however, opening too soon could result in rapid disease spread. Therefore, governments need to develop cautious approaches to lifting restrictions to return to normalcy [129]. The approach presented in this chapter allows for a deeper understanding of the policy effects on mitigating the spread of COVID-19. The forecasts of disease spread when NPIs are partially loosened guide policymakers towards the appropriate strategy when reversing the restrictions.

3.5 Data availability

All code and data are available in the Section 3.6 and in a public GitHub repository `https://github.com/ehgh/COVID-19-case-estimation-and-policy-efects`.

3.6 supplement

3.6.1 Model Configuration and Initialization

In this section, we detail the proposed SEIR model configurations (introduced in Section 3.2). We estimate the parameters of this model using the procedure described in [39]. The model parameters of the SEIR model are inferred via Iterative Filtering of stochastic Ensemble Kalman Filter (EnKF-IF). The EnKF is a Monte Carlo (MC) approximation of the Kalman filter. We specifically used the Ensemble Adjustment Kalman Filter (EAKF), suitable for models with a high number of parameters. In this technique, we adjust a presumably Gaussian distributed ensemble of state vectors to posterior distribution via the Bayes rule in each iteration. The state vector includes model parameters and meta-population values. We use the maximum likelihood approach to determine the final values of the state vector in the algorithm. The daily documented cases act as observations in the model. We find that a few hundreds of ensembles are sufficient to infer the model parameters accurately.

We introduce a randomly drawn number for infection detection delay for each documented case. This additional delay captures the latency period from the onset of symptoms to diagnosis and the time it takes an individual to become contagious from an initial exposure [127]. The model randomly adds delay with Gamma-distributed values of shape $a = 1.78$ and an average of $T = 6$ days. This distribution fits with the information from confirmed cases in China and South Korea [39, 130, 19]. We examined Gamma distributions with various averages and a constant shape to capture the most accurate average delay time for estimation purposes.

Furthermore, we introduce another delay for the effect of social distancing on daily confirmed cases. There is an average of 6 days delay between the change in human mobility (sd_i) and the corresponding change in the number of daily documented infected cases across all states. This constant shift accounts for the time that the change in mobility starts showing an effect on the disease transmission rate.

Table 3.2 represents the configuration of the model and initial ranges for parameters and values. Parameter's prior range does not limit the search space of the EAKF al-

Parameters	Initial Range	Inferred Value [95% CI]
β	0.8-1.2	0.95 [0.92 – 1.01]
μ	0.2-1.0	0.22 [0.22 – 0.27]
θ_G	1.0-1.8	1.08 [1.08 – 1.16]
θ_A	1.0-1.8	1.63 [1.61 – 1.63]
Z	2.0-5.0	2.41 [2.41 – 2.67]
γ	0.6-1.0	0.89 [0.89 – 0.91]
D	2.0-5.0	2.52 [2.52 – 2.91]
α_i	0.5-1.0	(Per State Estimate Follows)
α_{BW}		0.71 [0.69 – 0.72]
α_{BY}		0.76 [0.75 – 0.77]
α_{BE}		0.85 [0.84 – 0.87]
α_{BB}		0.64 [0.61 – 0.64]
α_{HB}		0.67 [0.65 – 0.67]
α_{HH}		0.85 [0.84 – 0.86]
α_{HE}		0.65 [0.63 – 0.65]
α_{MV}		0.65 [0.62 – 0.65]
α_{NI}		0.66 [0.64 – 0.67]
α_{NW}		0.64 [0.61 – 0.64]
α_{RP}		0.66 [0.64 – 0.66]
α_{SL}		0.75 [0.73 – 0.75]
α_{SN}		0.67 [0.66 – 0.69]
α_{ST}		0.63 [0.6 – 0.63]
α_{SH}		0.65 [0.63 – 0.66]
α_{TH}		0.64 [0.61 – 0.64]

Table 3.2: Model parameters inference and prior ranges. The initial range is the prior range for parameters in EAKF algorithm. The selected set of inferred values along with 95% confidence interval is reported.

gorithm, and it can move outside of the prior range to find the optimum solution. We choose a suitable prior range for state vector initial values to facilitate the convergence while covering most of the possible values for parameters.

The prior range of μ covers a wide range of possible values $[0, 1]$. We set the prior range for α to include most of its possible values $(0, 1]$ with a lower bound set to 0.5 to account for the high volume of Covid-19 testing in Germany. The prior range for β is set to cover a wide range of values for R_0, i.e. $[0, 12]$. Prior range of Z and D are chosen to cover the known average incubation and infection period for Covid-19 $[3, 5]$. We set the prior range of θ to capture most of the possible range of underreporting of transportation. The range limits the number of each state's travelers to its population (Figures 3.5 and 3.6). The prior range for γ covers most of the possible range for the effect of human mobility and social distancing on transmission rate, i.e., $(0, 1]$. The lower bound allows for at least a 50% drop in transmission rate if social mobility drops 100%.

We use February 18, 2020 – April 20, 2020, as the period for model inference. February 18, 2020, is the early stages of the COVID-19 epidemic in Germany, with only two states with reported cases. The initial susceptible population of each state, S_i, is set to its initial total population. The initial documented cases, I_i^d, are set to the reported cases on February 18, 2020. We set the exposed and undocumented infected initial values in the following way. Based on reported daily cases, three states are early hubs of Covid-19 in Germany. Nordrhein-Westfalen (NW) is the pioneer state linked to large carnival events, followed by states Baden-Wurttemberg (BW) and Bayern (BY) via outbreak in Italy [127]. The first significant cluster of reported cases in Nordrhein-Westfalen was reported on February 28, 2020, with 25 cases. Accounting for the average doubling time of 6.4 days [95% CI: 5.8-7.1 days] for cases [131] and an average of 86% [95%CI: 83%, 90%] undocumented cases in early stages [39], we estimate to have 562 [95% CI: 390-825] cases on the initial day of the model in Nordrhein-Westfalen. We set the initial value range for E_{NW}, I_{NW}^u to be $[0 - 400]$ and uniformly select random values from this range. Similarly, we set the initial range of E_{BW}, I_{BW}^u, and E_{BY}, I_{BY}^u for Baden-Wurttemberg and Bayern to $[0, 200]$. Excluding the states mentioned above, the rest of the German

states' initial values for E and I^u were drawn uniformly from $[0, C_{max}]$ with C_{max} being the aggregated number of undocumented infected travelers from the three hub states Nordrhein-Westfalen, Baden-Wurttemberg, and Bayern on February 18:

$$C_{i_{max}}^{t=0} = \sum_{\nu} \sum_{i \in \{BW,BY,NW\}} M_{ji}^{\nu} \frac{E_i}{N_i} \tag{3.7}$$

All ensembles' initial state vector values are drawn randomly via Latin Hypercube Sampling with uniform distribution from the initial ranges.

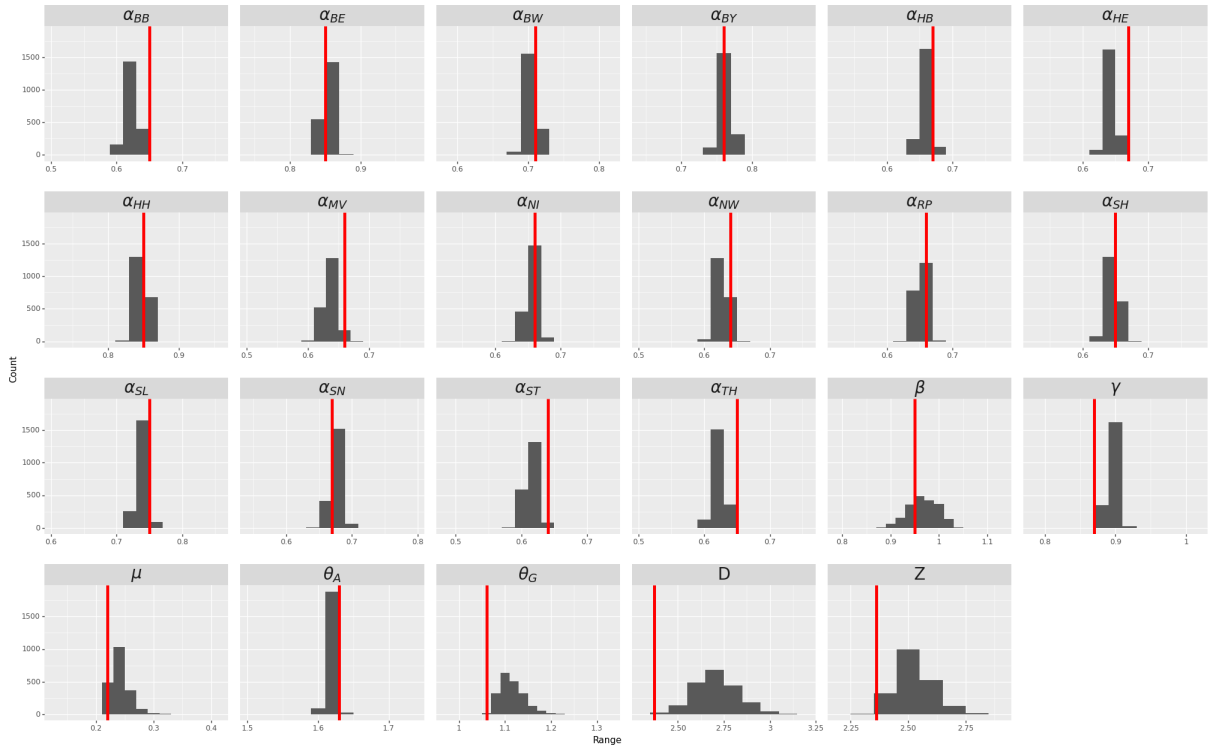
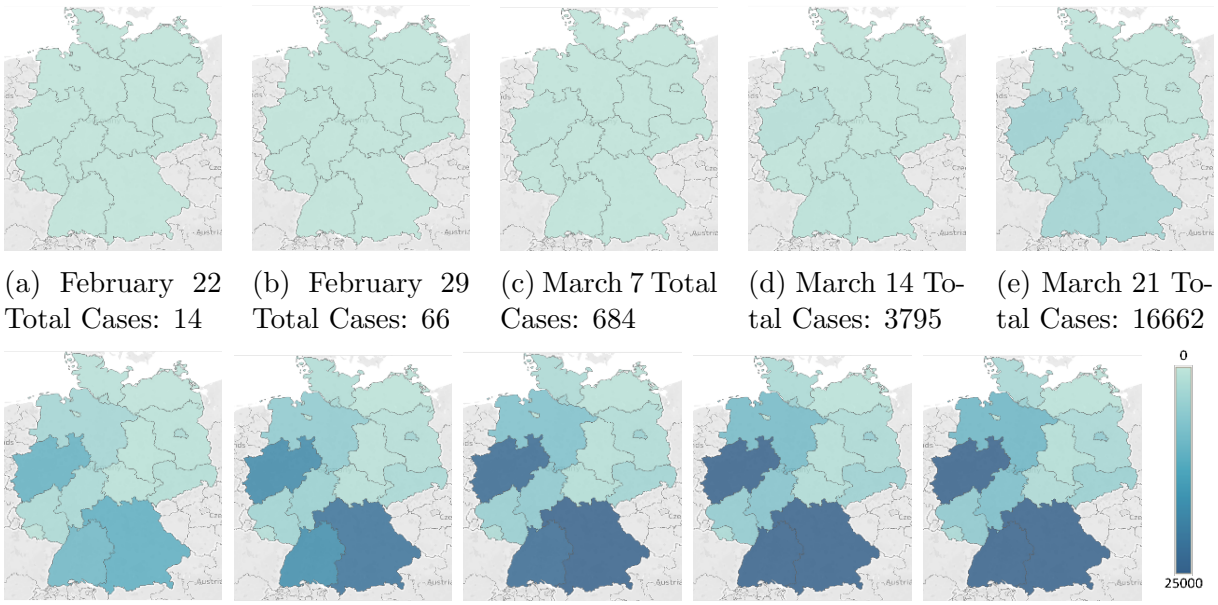


Figure 3.11: Model parameter inference. Initial day for the model is set to Feb 18, 2020. Histograms show distribution of model parameters inference over 2000 runs, each with 500 ensembles. Each run provides a set of parameters as a group that is optimized together and should be used as one. Red lines show the set of inferred parameters of a randomly selected run used for estimation.

Figure 3.11 shows the histograms of inferred model parameters. Red lines in Figure 3.11 depict the model parameter values selected from the set of inferred parameters. Please note that parameters sets are complex, and while one parameter might have a high inferred frequency in a set, another parameter in the same set might not have the highest inferred count.



(a) February 22 Total Cases: 14 (b) February 29 Total Cases: 66 (c) March 7 Total Cases: 684 (d) March 14 Total Cases: 3795 (e) March 21 Total Cases: 16662

(f) March 28 Total Cases: 48582 (g) April 4 Total Cases: 85778 (h) April 11 Total Cases: 117658 (i) April 18 Total Cases: 137439 (j) April 25 Total Cases: 152438

Figure 3.12: Spread of COVID-19 across Germany over time (Figures generated using Heat Map feature in Tableau Software, Version 2020.2 [132]).

3.6.2 Data Collection

3.6.2.1 Observations of confirmed COVID-19 Cases

The daily number of confirmed COVID-19 cases in Germany has been collated from the Robert Koch Institute (RKI) data in Berlin, Germany. Figure 3.12 shows how COVID-19 spread across Germany over time. When individuals contract COVID-19, they may remain asymptomatic. During this period, they can be active carriers of the virus. Symptomatic and asymptomatic individuals are identified through clinical testing. In Germany, COVID-19 testing required a primary care physician's referral at the time. More than 100 laboratories were contracted to conduct testing for the COVID-19 virus. The daily cases in RKI include the number of confirmed positive cases.

The RKI published the daily case reports at 10:00 before March 1, 2020. Between March 1, 2020, and March 9, 2020, the RKI reported cases at 10:00 and 15:00. Starting March 10, 2020, due to continuously rising case numbers, RKI switched to adopting the numbers transmitted electronically from testing centers across Germany. The new case

numbers were published once at 15:00. Starting March 17, 2020, the daily cases are published at midnight for the previous day. To account for the discrepancy in reporting time, we used a constrained cubic spline interpolation method to obtain the number of cases prior to March 17, adjusted for midnight reporting. Figure 3.13 shows the daily cases using cubic spline interpolation.

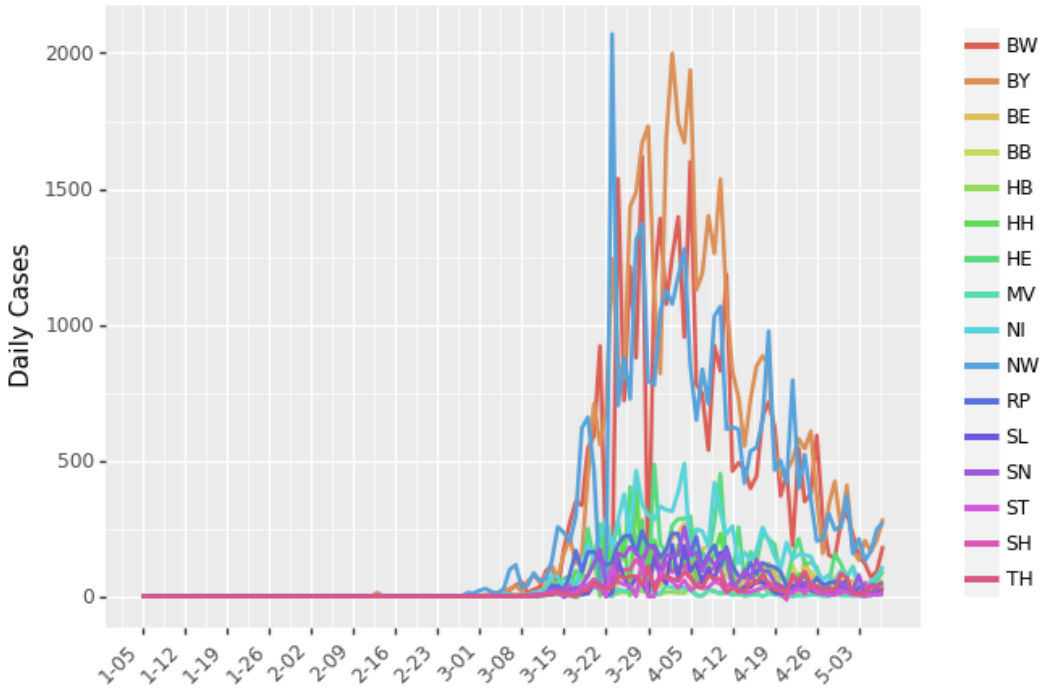


Figure 3.13: Daily new cases from RKI (adjusted for midnight reporting).

3.6.2.2 Community Mobility Data Trends

Google aggregates data from users' anonymized location history (for users who switched on the location history settings in their android mobile phones) to estimate foot traffic across six different location categories – retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential (Figure 3.14). With the outbreak of COVID-19, Google released data for changes in foot traffic (in percentage points) for the six location categories from February 15, 2020. These percentage changes in foot traffic are reported as community mobility trends. We use trends in retail and recreation to measure social distancing.

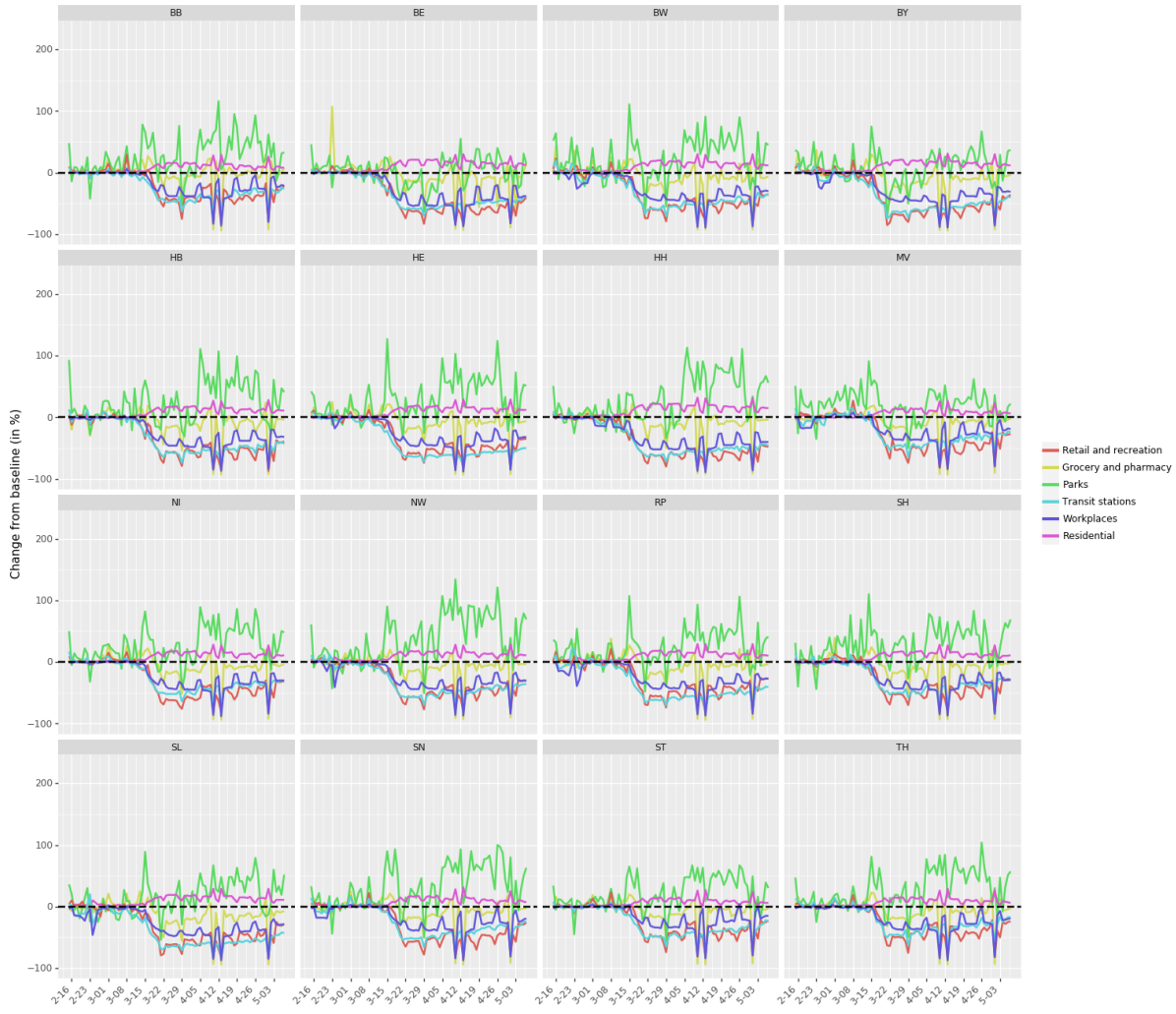


Figure 3.14: Google Community Mobility. Six location categories are retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential

3.6.2.3 Mobility Data

COVID-19 started in Wuhan, China, and spread to Germany and other parts of the world through cross-border human movement. We use air and different types of ground transportation to accurately collect the movement data across different states in Germany. For international travel into Germany from other countries, we consider traffic from 142 countries (including nine countries that share international borders with Germany) through ground and air transportation. Next, we discuss the movement data for different models.

3.6.2.4 Car Mobility Data and Truck Mobility Data

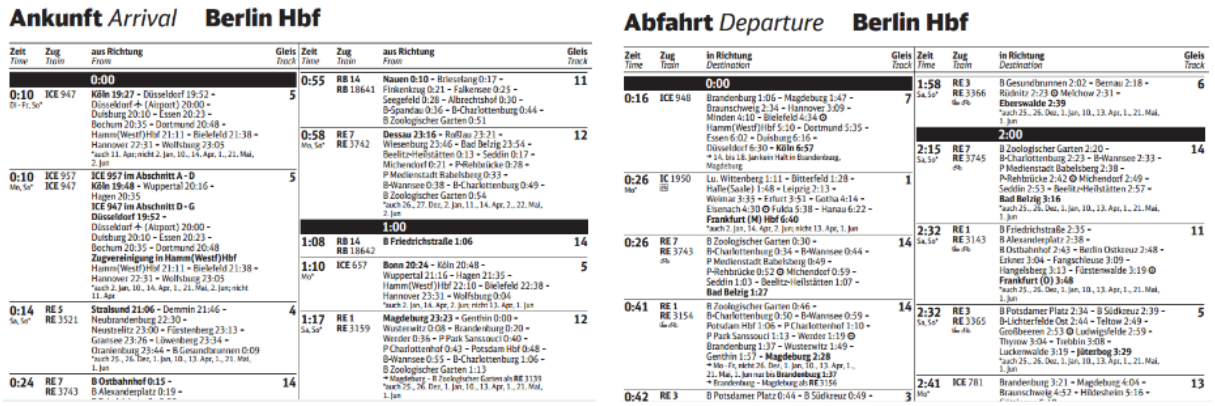
We collected detailed five-year highway traffic data from January 1, 2013, to December 31, 2018, provided by the German Bundesanstalt für Straßenwesen (Federal Institute for Roadways). The dataset contains the hourly count of vehicles passing through one of about 2,800 automatic counting stations along highways and state streets. Each station records traffic flow in the two directions (z_1 and z_2). The dataset contains geographical coordinates of the locations for these checkpoints. Sensors were used to identify the vehicles as cars, buses, and trucks. We use correction factors for public holidays, the day of the week, and state population to extrapolate hourly traffic for January 1, 2020, to May 7, 2020. We add the hourly data to get the daily movement data. We calculate the mean of daily movement data (W_h, z_1, z_2) for highway h (both directions). We also calculate the mean of the movement data for each highway during public holidays (W_h^{ph}, z_1, z_2) and day of the week (W_h^{dw}, z_1, z_2). We use the ratio, $\frac{(W_h^{ph},z_1,z_2)}{(W_h,z_1,z_2)}$ as correction factor for public holidays. We also use a similar correction factor for the day of the week. Finally, we correct for the increase in population by using the population ratio of 2020 to 2018. We build a similar model for trucks to predict the number of trucks moving from January 1, 2020, to May 7, 2020.

To adjust for changes in car movement due to COVID-19, we use the daily google mobility trends for workplaces. We multiply the projected car movement for 2020 with google mobility to obtain adjusted car movement for 2020. As part of the essential services to keep the supply chain from breaking, there were no restrictions on the truck movement. Therefore, we did not adjust for changes in the truck movement for the period of our study.

3.6.2.5 Train mobility data

We use Deutsche Bahn's public timetable to determine all major train routes in Germany. We first identify the 110 biggest cities in Germany and their respective states. We use these major cities to identify movement across states and neighboring countries. Each train station has two timetables for that station (See Figure 3.15) – one table shows the arrival time of all the trains to that station (including the departure time from its previous stations), and the other table shows the departure time of all the trains from that station

(including the arrival time of its previous stations). Using both schedules for each train, we find a complete route (based on the 100 biggest cities in Germany). The number of passengers boarding a train for the next station is kept proportional to the sequence of the cities in the train route. We identified 14,712 trains. There are 33 types of trains (based on speed, distance travel, and capacity). We assume the total number of passengers in a train to be 400. For long-distance trains, e.g., ICE, THA, and TGV, we assume the number of passengers in a train to be 600.



Ankunft *Arrival* **Berlin Hbf**

Zeit *Time*	Zug *Train*	aus Richtung *From*	Gleis *Track*
		0:00	
0:10 Di - Fr, So*	ICE 947	**Köln 19:27** - Düsseldorf 19:52 - Düsseldorf ✈ (Airport) 20:00 - Duisburg 20:10 - Essen 20:23 - Bochum 20:35 - Dortmund 20:48 - Hamm(Westf)Hbf 21:11 - Bielefeld 21:38 - Hannover 22:31 - Wolfsburg 23:05 *auch 11. Apr; nicht 2. Jan, 10., 14. Apr, 1., 21. Mai, 2. Jun	5
0:10 Mo, So*	ICE 957 ICE 947	**ICE 957 im Abschnitt A - D** **Köln 19:48** - Wuppertal 20:16 - Hagen 20:35 **ICE 947 im Abschnitt D - G** **Düsseldorf 19:52** - Düsseldorf ✈ (Airport) 20:00 - Duisburg 20:10 - Essen 20:23 - Bochum 20:35 - Dortmund 20:48 **Zugvereinigung in Hamm(Westf)Hbf** Hamm(Westf)Hbf 21:11 - Bielefeld 21:38 - Hannover 22:31 - Wolfsburg 23:05 *auch 2. Jan, 10., 14. Apr, 1., 21. Mai, 2. Jun; nicht 11. Apr	5
0:14 Sa, So*	RE 5 RE 3521	**Stralsund 21:06** - Demmin 21:46 - Neubrandenburg 22:30 - Neustrelitz 23:00 - Fürstenberg 23:13 - Gransee 23:26 - Löwenberg 23:34 - Oranienburg 23:44 - B Gesundbrunnen 0:09 *auch 25., 26. Dez, 1. Jan, 10., 13. Apr, 1., 21. Mai, 1. Jun	4
0:24	RE 7 RE 3743	**B Ostbahnhof 0:15** - B Alexanderplatz 0:19 -	14
0:55	RB 14 RB 18641	**Nauen 0:10** - Brieselang 0:17 - Finkenkrug 0:21 - Falkensee 0:25 - Seegefeld 0:28 - Albrechtshof 0:30 - B-Spandau 0:36 - B-Charlottenburg 0:44 - B Zoologischer Garten 0:51	11
0:58 Mo, So*	RE 7 RE 3742	**Dessau 23:16** - Roßlau 23:21 - Wiesenburg 23:46 - Bad Belzig 23:54 - Beelitz-Heilstätten 0:13 - Seddin 0:17 - Michendorf 0:21 - P-Rehbrücke 0:28 - P Medienstadt Babelsberg 0:33 - B-Wannsee 0:38 - B-Charlottenburg 0:49 - B Zoologischer Garten 0:54 *auch 26., 27. Dez, 2. Jan, 11., 14. Apr, 2., 22. Mai, 2. Jun	12
		1:00	
1:08	RB 14 RB 18642	**B Friedrichstraße 1:06**	14
1:10 Mo*	ICE 657	**Bonn 20:24** - Köln 20:48 - Wuppertal 21:16 - Hagen 21:35 - Hamm(Westf)Hbf 22:10 - Bielefeld 22:38 - Hannover 23:31 - Wolfsburg 0:04 *auch 2. Jan, 14. Apr, 2. Jun; nicht 13. Apr, 1. Jun	5
1:17 Sa, So*	RE 1 RE 3159	**Magdeburg 23:23** - Genthin 0:00 - Wusterwitz 0:08 - Brandenburg 0:20 - Werder 0:36 - P Park Sanssouci 0:40 - P Charlottenhof 0:43 - Potsdam Hbf 0:48 - B-Wannsee 0:55 - B-Charlottenburg 1:06 - B Zoologischer Garten 1:13 → Magdeburg - B Zoologischer Garten als RE 3139 *auch 25., 26. Dez, 1. Jan, 10., 13. Apr, 1., 21. Mai, 1. Jun	12

Abfahrt *Departure* **Berlin Hbf**

Zeit *Time*	Zug *Train*	in Richtung *Destination*	Gleis *Track*
		0:00	
0:16	ICE 948	Brandenburg 1:06 - Magdeburg 1:47 - Braunschweig 2:34 - Hannover 3:09 - Minden 4:10 - Bielefeld 4:34 ⚙ Hamm(Westf)Hbf 5:10 - Dortmund 5:35 - Essen 6:02 - Duisburg 6:16 - Düsseldorf 6:30 - **Köln 6:57** → 14. bis 18. Januar Halt in Brandenburg, Magdeburg	7
0:26 Mo*	IC 1950	Lu. Wittenberg 1:11 - Bitterfeld 1:28 - Halle(Saale) 1:48 - Leipzig 2:13 - Weimar 3:35 - Erfurt 3:51 - Gotha 4:14 - Eisenach 4:39 ⚙ Fulda 5:38 - Hanau 6:22 - **Frankfurt (M) Hbf 6:40** *auch 2. Jan, 14. Apr, 2. Jun; nicht 13. Apr, 1. Jun	1
0:26	RE 7 RE 3743	B Zoologischer Garten 0:30 - B-Charlottenburg 0:34 - B-Wannsee 0:44 - P Medienstadt Babelsberg 0:49 - P-Rehbrücke 0:52 ⚙ Michendorf 0:59 - Seddin 1:03 - Beelitz-Heilstätten 1:07 - **Bad Belzig 1:27**	14
0:41	RE 1 RE 3154	B Zoologischer Garten 0:46 - B-Charlottenburg 0:50 - B-Wannsee 0:59 - Potsdam Hbf 1:06 - P Charlottenhof 1:10 - P Park Sanssouci 1:13 - Werder 1:19 ⚙ Brandenburg 1:37 - Wusterwitz 1:49 - Genthin 1:57 - **Magdeburg 2:28** → Mo - Fr, nicht 26. Dez, 1. Jan, 10., 13. Apr, 1., 21. Mai, 1. Jun nur bis **Brandenburg 1:37** → Brandenburg - Magdeburg als RE 3156	14
0:42	RE 3	B Potsdamer Platz 0:44 - B Südkreuz 0:49 -	3
1:58 Sa, So*	RE 3 RE 3366	B Gesundbrunnen 2:02 - Bernau 2:18 - Ridnitz 2:23 ⚙ Melchow 2:31 - **Eberswalde 2:39** *auch 25., 26. Dez, 1. Jan, 10., 13. Apr, 1., 21. Mai, 1. Jun	6
		2:00	
2:15 Sa, So*	RE 7 RE 3745	B Zoologischer Garten 2:20 - B-Charlottenburg 2:23 - B-Wannsee 2:33 - P Medienstadt Babelsberg 2:38 - P-Rehbrücke 2:42 ⚙ Michendorf 2:49 - Seddin 2:53 - Beelitz-Heilstätten 2:57 - **Bad Belzig 3:16** *auch 25., 26. Dez, 1. Jan, 10., 13. Apr, 1., 21. Mai, 1. Jun	14
2:32 Sa, So*	RE 1 RE 3143	B Friedrichstraße 2:35 - B Alexanderplatz 2:38 - B Ostbahnhof 2:43 - Berlin Ostkreuz 2:48 - Erkner 3:04 - Fangschleuse 3:09 - Hangelsberg 3:13 - Fürstenwalde 3:19 ⚙ **Frankfurt (O) 3:48** *auch 25., 26. Dez, 1. Jan, 10., 13. Apr, 1., 21. Mai, 1. Jun	11
2:32 Sa, So*	RE 3 RE 3365	B Potsdamer Platz 2:34 - B Südkreuz 2:39 - B-Lichterfelde Ost 2:44 - Teltow 2:49 - Großbeeren 2:53 ⚙ Ludwigsfelde 2:59 - Thyrow 3:04 - Trebbin 3:08 - Luckenwalde 3:19 - **Jüterbog 3:29** *auch 25., 26. Dez, 1. Jan, 10., 13. Apr, 1., 21. Mai, 1. Jun	5
2:41 Mo*	ICE 781	Brandenburg 3:21 - Magdeburg 4:04 - Braunschweig 4:57 - Hildesheim 5:16 -	13

Figure 3.15: Ground transportation using trains in Germany. Panel (a) shows part of the train timetable available in all German train stations. More specifically, the sample snippet displayed in Panel (a) lists the departure times of trains leaving Berlin Hbf. We parse 538 timetable files to obtain the train schedule for all of Germany. We combine information from the arrival and departure timetables to construct the complete route of a train. Panel (b) lists the departure time for all the trains from Berlin with arrival time of all the succeeding stations for that train. We combine information from Panel (a) and Panel (b) to construct the complete route of a train.

Several trains were canceled due to COVID-19 and state policies, and train movement declined. We adjust the number of passengers using Google mobility data to account for these changes. We estimate the number of passengers traveling across states from Deutsche's Ban Schedule. We adjust the number of passengers traveling by train after the outbreak of COVID-19 using Google mobility trends at transit stations. Community mobility in transit stations is the change in the number of users in and around transit stations compared to the baseline defined by Google.

3.6.2.6 Bus mobility data

We use the travel search history provided by a large third-party European bus and train price comparison and booking company to estimate the number of passengers moving across cities (states) in Germany and passengers traveling to Germany from neighboring countries. The bus data contains the number of searches for a route (departure city to arrival city) aggregated by the day. For example, 33 people searched for buses from Frankfurt to Heidelberg on January 25, 2020. The data does not show the actual number of travelers on the bus, but we use this data as an indicator for bus movement across Germany and its neighboring countries. We only include connections that arrive at or depart from Germany. We define a bus route as the tuple [departure city, arrival city, date]. The dataset contains the history for 857,159 unique connections (aggregated by day) from December 1, 2019, to May 7, 2020. Of these, 191,356 routes had either their origin or destination in Germany. A subset of 116,706 routes originated and ended in Germany. We assume a capacity of 20 passengers in each bus to estimate the number of travelers. On March 16, 2020, all bus trips were halted in Germany.

3.6.2.7 Flight mobility data

We use flight transportation information from `https://opensky-network.org`. It is an open-source platform containing historical information on all airborne flights. The database uses Automatic Dependent Surveillance-Broadcast (ADS-B) trajectories and maps them with airport International Civil Aviation Organization (ICAO) codes to identify the departure and arrival airport. The database also maintains the UNIX timestamp for each contact signal (trajectory recorded). We use the last UNIX timestamp of a flight to identify its date of arrival and departure. We only consider flights with arrival or departure airports in Germany. We use ICAO airport codes to identify the state of each airport. We ignore flights for which neither the departure nor the arrival airport can be established. The dataset has 187,526 flights (with arrival or departure in Germany) from December 1, 2019, to May 7, 2020. Figure 3.16 shows the variation in the number of flights over time. We assume a capacity of 200 passengers for a domestic flight and 500 passengers for an international flight to determine the number of individuals on a flight.

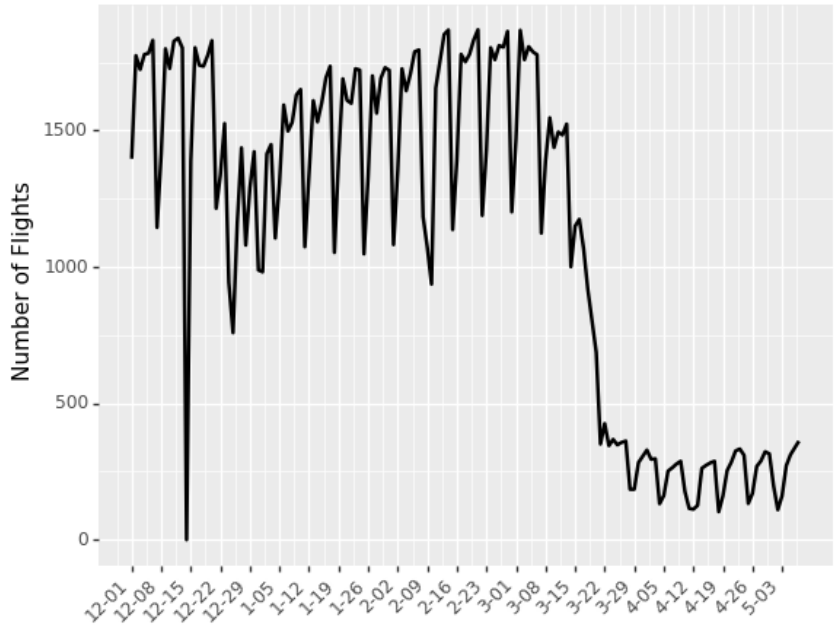


Figure 3.16: Daily flights arriving to different airports in Germany.

3.6.3 Effect of NPIs on Social Distancing

To understand the effect of policy interventions on social distancing, we build a linear regression model. We explain the details of the model here. We use data from February 18, 2020, to April 20, 2020, to estimate the coefficients for different policies. Note that not each state implemented every policy as of April 20, 2020, and none of the implemented policies were relaxed until April 20. State governments started relaxing these policies after April 20, 2020. So we use data up to April 20 to estimate the coefficients of different policies. We use Google community mobility ($C_{j,t}$) for retail and recreation to create a measure for deriving social distancing ($sd_{j,t}$) in state j on day t. We use smoothed (7-day moving average) social distancing in our linear regression model in Equation 3.6. Figure 3.17 shows raw and smoothed social distancing numbers for 16 states.

We also use Google Trends, weather data, and general dissatisfaction as additional control variables to account for latent awareness levels in the population, the tendency to leave one's home due to the higher temperatures during Spring in Germany, and the distress felt due to confining oneself. We use Google Trends data for the search term "COVID-19". Google Trends is an indicator of the search interest of a topic over time.

Policy	Description	Coefficients
Border Closure	Closure of International Borders. The border between Germany and Netherlands was never closed.	β_1^{policy}
Initial Business Closure	Closure of dance events, trade fairs, exhibitions, special markets, arcades, casinos, betting shops, and similar companies; amusement places; prostitution institutions, concert halls, fairs, leisure, and animal parks, providers of leisure activities, and similar facilities	β_2^{policy}
Educational Facilities Closure	Closure of All Educational Facilities, K-12 and University	β_3^{policy}
Non-Essential Services Closure	Closure of bars, clubs, cinemas, theatres, museums, florists, garages, fashion stores, and churches.	β_4^{policy}
Stay at Home Order	Residents asked to shelter-in-place	β_5^{policy}
Contact Restriction	Gatherings limited to no more than 5 people, unless family	β_6^{policy}
Retail Outlets Closure	Does not apply to retail for food, weekly markets, pick-up and delivery services, beverage markets, pharmacies, medical supply stores, drug stores, petrol stations, banks and savings banks, post offices, hairdressers, dry cleaners, Laundromats, newspaper sales, DIY and garden centers, pet supplies and wholesale, craftsmen and craft like trades.	β_7^{policy}

Table 3.3: Description of different NPIs

Figure 3.17: Raw and smoothed social distancing in different states.

It calculates the proportion of all other searches simultaneously and normalizes it to the range of 0 - 100. We normalize the data to $0 - 1$ in our model. High Google Trend numbers indicate high interest in the topic during that time. As more cases were observed worldwide and in Germany, public interest in COVID-19 increased. Increased search is also an indicator of public awareness towards increasing social distancing. Figure 3.18 shows the Google Trends data for all the sixteen states. We use seven days exponentially smoothed Google Trends in our analysis to account for increased awareness over time. We

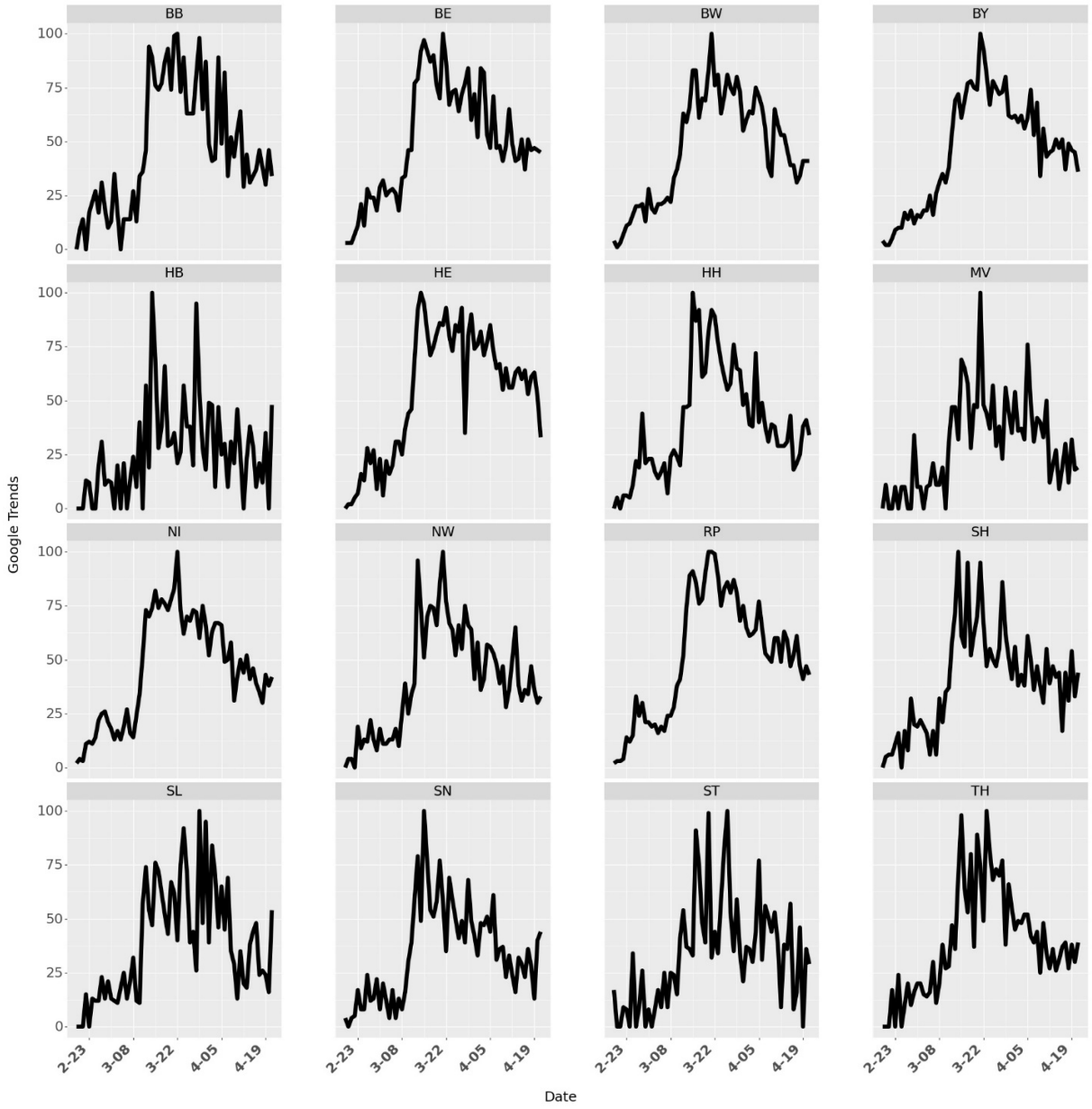


Figure 3.18: Google Trends data for different states in Germany.

use 7-day exponential smoothing to indicate higher awareness through active searching instead of search history. The smoothing function is:

$$trend_{j,t} = \frac{\sum_{d=1}^{7} U^{7-d} trend_{j,t}}{\sum_{d=1}^{7} U^{7-d}} \tag{3.8}$$

where $trend_{j,t}$ are the raw numbers of Google Trends for search term in state j and U is a constant. For simplicity, we use the same notation $trend_{j,t}$ in our model in Equation.

3.6 for the 7-day average of Google Trends term. The model is robust to changing the smoothing function. Also, we use maximum temperature recorded in a day to account for increased public interest in going out as summer approaches. Maximum Daily temperature for all the sixteen states is shown in Figure 3.19.

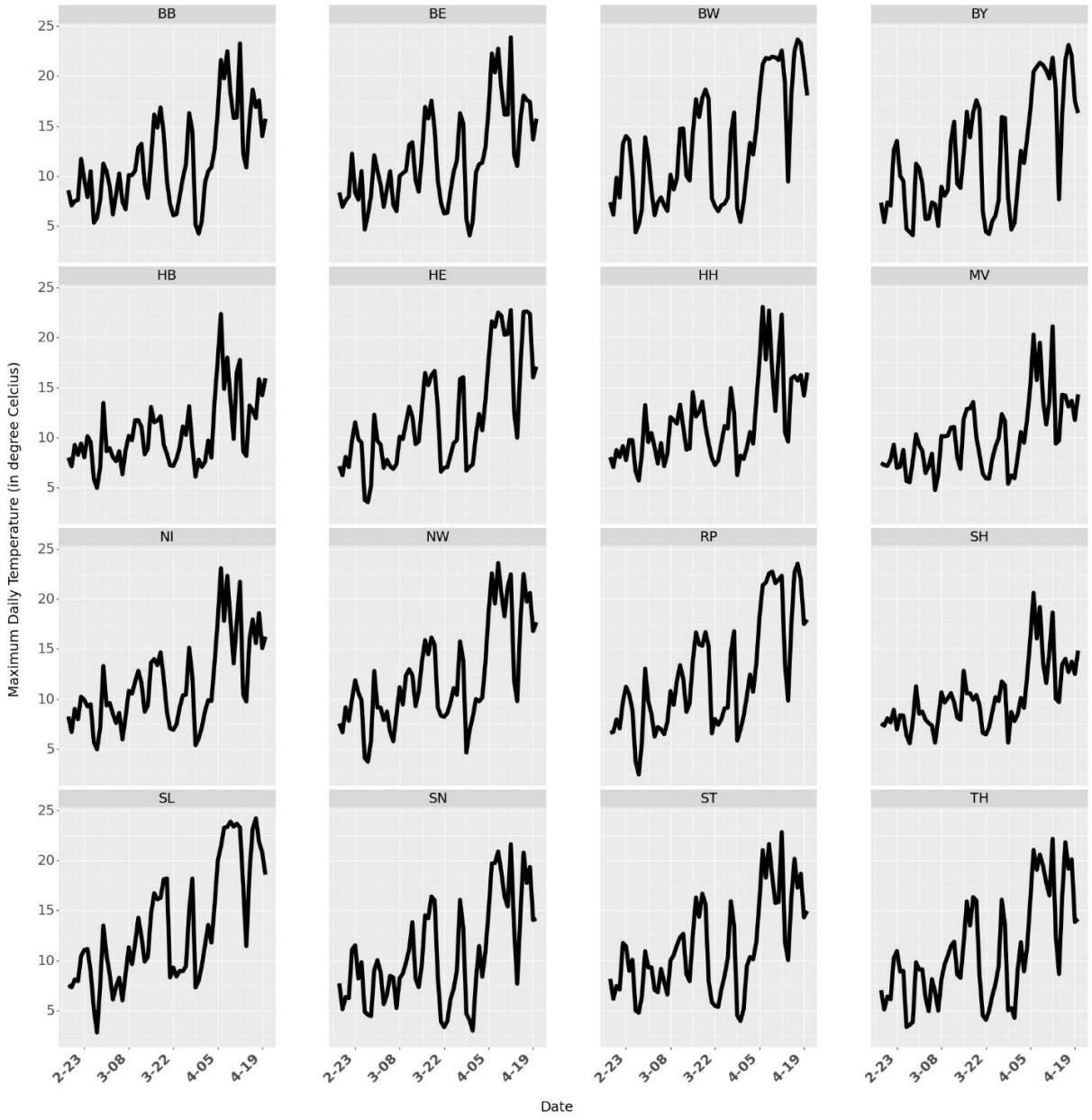


Figure 3.19: Maximum daily temperature data for different states in Germany.

The linear regression model used to measure the effect of NPIs on mobility $C_{j,t}$ is shown in Equation 3.6. The model is robust to changes in the value of the constant K.

We use a binary variable $x_{j,p,t} = 1$ if policy p is active in state j on day t. $temp_{j,t}$ is the daily maximum temperature in degrees Celsius. We use t as the index for days. To account for state-level heterogeneity we use state fixed effects ($state_j$ in Equation 3.6). $state_j$ assumes a value of one if the data considered are specific to that state and zero otherwise. This approach allows for control of state-level characteristics that are not in the model and helps reduce the errors due to omitted variables in our analysis. In addition to state-level differences, we also control for week-based and day-of-week differences by incorporating week-fixed and day-of-week fixed effects. The variable $week_t$ takes a value of 1 if the data considered is in the w^{th} week from February 18, 2020 (For example, all days from February 18, 2020, to February 24, 2020, have $week_0 = 1$). Variable day_t takes a value of 1 if the data considered are specific to that day of the week. Since we consider data for 63 days in our linear regression model, we consider eight weeks (week nine is considered as the base week with 0 fixed effect). We consider state Thuringia as our base state (0 state fixed effects) and Monday as our base day (0 day of week fixed effect) for stable parameter estimation.

Results for the parameter estimates and summary of the linear regression model is provided in Table 3.4. The coefficients (and 95% confidence intervals) for different NPIs is given in Figure 3.8a. Predicted and social mobility numbers ($sd_i = \frac{-C_i}{100}$) for different states in Germany is shown in Figure 3.20.

3.6.3.1 Selecting Measure for Social Distancing

In our analysis, we use community mobility to derive our measure for social distancing $sd_i = \frac{-C_i}{100}$. We select Google community mobility for retail and recreation to create a measure for social distancing. Google community mobility reports collect data from android users. We also collect data for social mobility from Apple for iOS users (Apple's Community Mobility Report [122]. Apple's community mobility data includes the change in trend in movement as compared to January 13, 2020. They collect data on driving, transit, and walking. Apple's community mobility provides state-wise data for Germany only for driving (shown in Figure 3.21). Table 3.5 shows the correlation between the six measures of community mobility by Google and Apple. Community mobility for Retail

Model Summary			
Number of Observations		1008	
Degree of Freedom (Residuals)		969	
Degree of Freedom (Model)		38	
R-square Value		0.978	
Coefficient Estimates			
Variables	Lower Confidence Bound	Estimate	Upper Confidence Bound
Constant (C)	-29.559	-25.6761 ***	-21.793
Border Closure (β_1^{policy})	-1.099	-0.0343	1.03
Initial Business Closure (β_2^{policy})	-3.736	-2.5041 ***	-1.272
Educational Facilities Closure (β_3^{policy})	-6.91	-5.0164 ***	-3.123
Non-essential Service Closure (β_4^{policy})	-1.753	-0.6261	0.501
Stay at Home Order (β_5^{policy})	-1.873	0.0142	1.901
Contact Restriction (β_6^{policy})	-12.519	-10.3598 ***	-8.2
Retail Outlets Closure (β_7^{policy})	-4.076	-3.1127 ***	-2.149
Google Trends (β^{trend})	-4.356	-3.797 ***	-3.238
T_{max} (β^{temp})	0.366	0.4473 ***	0.529
Baden-wurttemberg ($state_1$)	-7.389	-5.7414 ***	-4.094
Bayern ($state_2$)	-9.972	-8.358 ***	-6.744
Berlin ($state_3$)	-11.089	-9.5171 ***	-7.945
Brandenburg ($state_4$)	1.258	2.9328 ***	4.608
Bremen ($state_5$)	-7.969	-6.2889 ***	-4.608
Hamburg ($state_6$)	-13.172	-11.6518 ***	-10.131
Hessen ($state_7$)	-5.092	-3.5518 ***	-2.012
Mecklenburg-Vorpommern ($state_8$)	-3.879	-2.284 ***	-0.689
Niedersachsen ($state_9$)	-6.538	-5.0077 ***	-3.477
Nordrhein-Westfalen ($state_{10}$)	-6.396	-4.8454 ***	-3.295
Rheinland-Pfalz ($state_{11}$)	-4.333	-2.6913 ***	-1.05
Saarland ($state_{12}$)	-8.94	-7.3416 ***	-5.743
Sachsen ($state_{13}$)	-3.631	-1.9973 **	-0.363
Sachsen-anhalt ($state_{14}$)	-0.77	0.7554	2.281
Schleswig-holstein ($state_{15}$)	-2.55	-0.9342	0.681
Week 0 ($week_0$)	24.024	27.3614 ***	30.699
Week 1 ($week_1$)	27.835	31.0584 ***	34.282
Week 2 ($week_2$)	29.941	33.1193 ***	36.298
Week 3 ($week_3$)	22.275	25.1858 ***	28.097
Week 4 ($week_4$)	7.409	9.7545 ***	12.1
Week 5 ($week_5$)	3.094	4.7599 ***	6.426
Week 6 ($week_6$)	5.694	7.0033 ***	8.312
Week 7 ($week_7$)	0.018	1.1626 **	2.307
Tuesday (day_1)	1.98	2.9742 ***	3.969
Wednesday (day_2)	0.78	1.7576 ***	2.735
Thursday (day_3)	-0.152	0.8219 *	1.795
Friday (day_4)	-0.541	0.4362	1.414
Saturday (day_5)	-1.477	-0.5065	0.464
Sunday (day_6)	-0.909	0.0431	0.995
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$			

Table 3.4: Linear regression summary

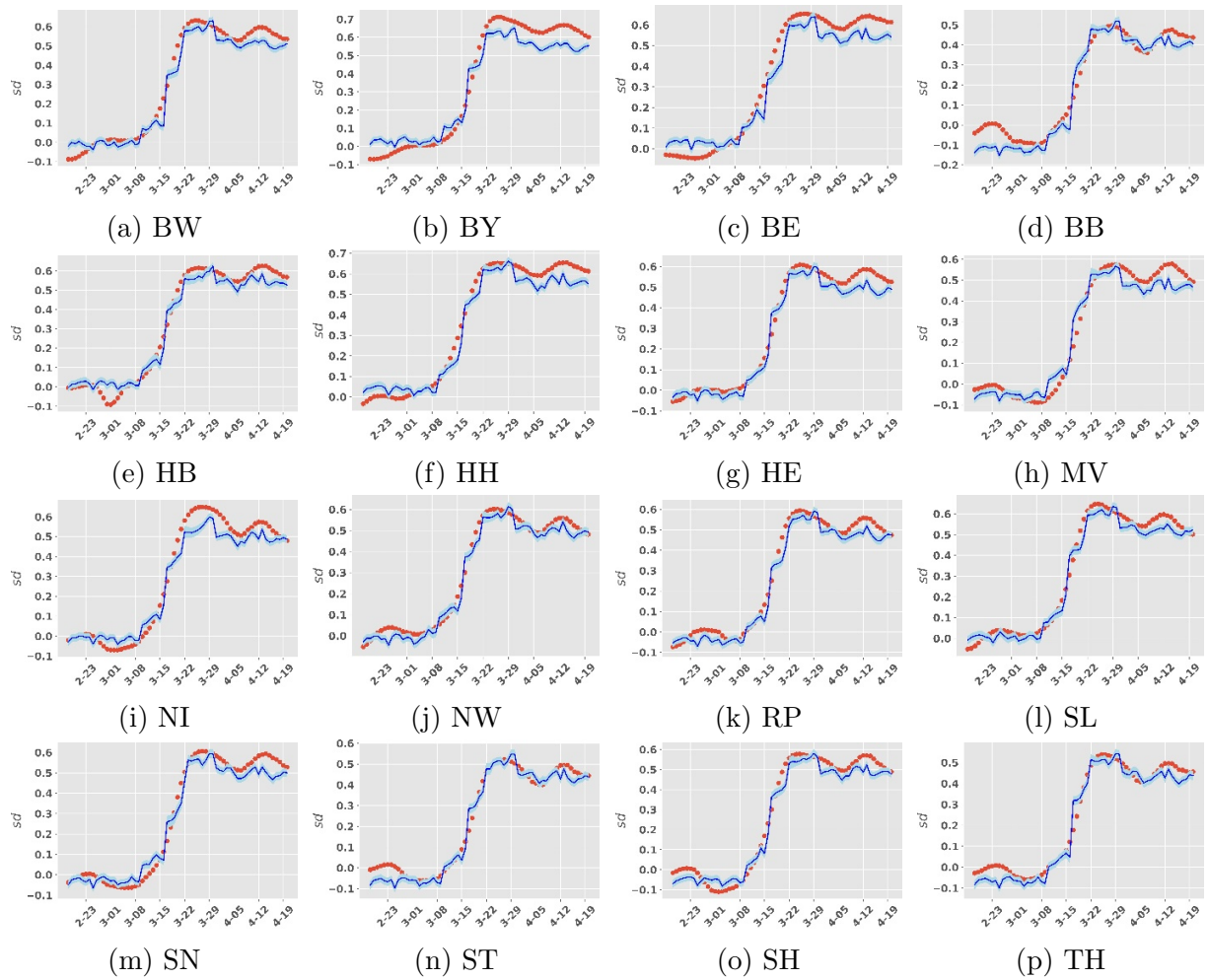


Figure 3.20: Predicting social distancing for different states in Germany. Red dots shows smoothed social distancing. Solid blue line shows the estimate for predicted *sd*. Light blue shade is the 95%confidence intervals around the predicted *sd*. The model achieves an adjusted R-square value of 0.978.

and Recreation is highly correlated with all the other indicators of community mobility (with Pearson correlation > 0.5 for grocery and pharmacy, transit stations, workplaces, and residential in Google community movement data and Driving in Apple community movement data). It is highly correlated with all other indicators of community mobility except for Google community mobility for Parks. However, few people visited parks during the pandemic. Hence, we consider Community mobility for Retail and Recreation as a measure for social distancing. Note that the community mobility provided by Google and Apple community movement reports do not include actual movement data (or actual

community mobility). It is an indicator of changes in trends in the movement as compared to their respective baselines.

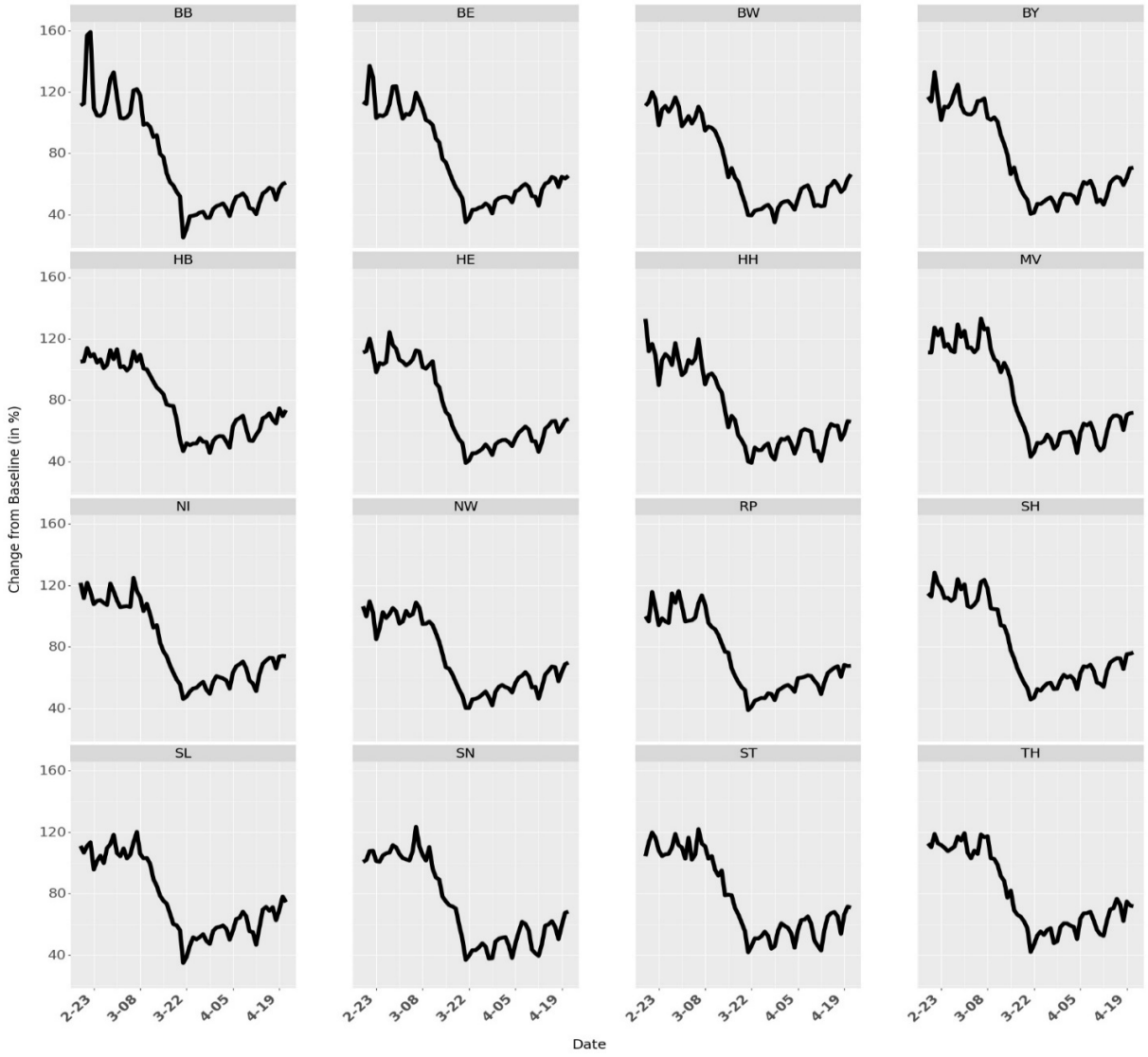


Figure 3.21: Apple Community Mobility (Driving).

3.6.3.2 Robustness Check: Lasso Regression Model

Lasso regression places a penalty on the sum over the absolute value of coefficients. As some of the NPIs were introduced simultaneously, this could lead to multicollinearity in NPIs. We use a penalized regression model (Equation 3.9) to check the robustness of the estimate from linear regression model in Equation 3.6 ($|l|$= magnitude of scalar l). We use 5-fold cross-validation to estimate the parameter coefficients using Lasso regression.

	Google: Retail and Recreation	Google: Grocery and Pharmacy	Google: Parks	Google: Transit Stations	Google: Workplace	Google: Residential	Apple: Driving
Google: Retail and Recreation	1	0.503	-0.178	0.958	0.855	-0.863	0.922
Google: Grocery and Pharmacy	0.503	1	-0.018	0.6	0.66	-0.622	0.457
Google: Parks	-0.178	-0.018	1	-0.111	-0.166	0.106	-0.165
Google: Transit Stations	0.958	0.6	-0.111	1	0.899	-0.9	0.913
Google: Workplace	0.855	0.66	-0.166	0.899	1	-0.97	0.77
Google: Residential	-0.863	-0.622	0.106	-0.9	-0.97	1	-0.788
Apple: Driving	0.922	0.457	-0.165	0.913	0.77	-0.788	1

Table 3.5: Correlation matrix for community mobility

Table 3.6 shows the estimates. Robustness check using Lasso shows that the parameter estimation from the Linear regression model in Equation 3.6 is robust.

$$
\begin{aligned}
C_{j,t} = K + \sum_{p=1}^{7} \beta_p^{policy} x_{j,p,t} + \beta_t^{trend} trend_{j,t} + \beta_m^{temp} temp_{j,t} + \\
\sum_{j=1}^{16-1} \beta_j^{state} state_j + \sum_{w=1}^{9-1} \beta_w^{week} week_t + \sum_{d=1}^{7-1} \beta_d^{day} day_t + \epsilon_{j,t} + \\
\lambda\Big(\sum_{p=1}^{7} |\beta_p^{policy}| + |\beta^{trend}| + |\beta^{temp}| + \\
\sum_{j=1}^{16-1} |\beta_j^{state}| + \sum_{w=1}^{9-1} |\beta_w^{week}| + \sum_{d=1}^{7-1} |\beta_d^{day}| \Big)
\end{aligned}
\tag{3.9}
$$

Lasso regression shrinks the coefficients to 0 such that only significant predictor variables affect the prediction. Therefore, Lasso regression does not provide a confidence interval on the estimates of the coefficients. Also, estimates of the confidence interval for weighted linear regression models are biased, so we do not provide confidence intervals for Lasso regression. However, we adopt a bootstrapping method to obtain the Interquartile range (IQR) for the parameter estimates. In bootstrapping, we run our Lasso regression

Coefficient Estimates			
Variables	25^{th} Percentile	50^{th} Percentile	75^{th} Percentile
Constant (C)	0	0	0
Border Closure (β_1^{policy})	0	0	0
Initial Business Closure (β_2^{policy})	-4.644155	-4.43152	-4.22604
Educational Facilities Closure (β_3^{policy})	-11.54166	-11.1959	-10.7803
Non-essential Service Closure (β_4^{policy})	-1.842748	-1.71596	-1.58752
Stay at Home Order (β_5^{policy})	-2.548597	-2.22569	-1.88736
Contact Restriction (β_6^{policy})	-16.99543	-16.6377	-16.2624
Retail Outlets Closure (β_7^{policy})	-4.559506	-4.43518	-4.31558
Google Trends (β^{trend})	-4.878719	-4.78912	-4.71101
T_{max} (β^{temp})	0.070563	0.081985	0.092904
Baden-wurttemberg ($state_1$)	-0.395887	-0.28022	-0.14904
Bayern ($state_2$)	-3.119428	-2.99275	-2.82677
Berlin ($state_3$)	-5.464731	-5.27469	-5.11732
Brandenburg ($state_4$)	5.239694	5.425942	5.595277
Bremen ($state_5$)	-1.799361	-1.6243	-1.42398
Hamburg ($state_6$)	-7.498475	-7.344	-7.18164
Hessen ($state_7$)	0	0	0
Mecklenburg-Vorpommern ($state_8$)	0	0	0
Niedersachsen ($state_9$)	-1.002236	-0.86896	-0.72117
Nordrhein-Westfalen ($state_{10}$)	0	0	0
Rheinland-Pfalz ($state_{11}$)	0.1722964	0.320479	0.456829
Saarland ($state_{12}$)	-2.159837	-2.03661	-1.90153
Sachsen ($state_{13}$)	0	0	0
Sachsen-anhalt ($state_{14}$)	2.5614361	2.731737	2.880114
Schleswig-holstein ($state_{15}$)	1.3204712	1.470629	1.625935
Week 0 ($week_0$)	0	0	0
Week 1 ($week_1$)	4.0150372	4.184664	4.360424
Week 2 ($week_2$)	6.5606947	6.709731	6.863293
Week 3 ($week_3$)	2.0989409	2.313502	2.560254
Week 4 ($week_4$)	-1.081132	-0.76835	-0.42437
Week 5 ($week_5$)	0.4573945	0.644298	0.871296
Week 6 ($week_6$)	3.4641791	3.616409	3.774666
Week 7 ($week_7$)	0	0	0
Tuesday (day_1)	0.3618764	0.477806	0.581058
Wednesday (day_2)	0	0	0
Thursday (day_3)	-0.007188	0	0
Friday (day_4)	-0.609046	-0.5159	-0.41635
Saturday (day_5)	-1.055083	-0.9451	-0.82874
Sunday (day_6)	0	0	0

Table 3.6: Parameter estimates from Lasso regression model

model 10,000 times with 90% of the data and obtain the estimates for the coefficients. We calculate IQR from the parameter estimates of these 10,000 runs. Table 3.6 provides the results for the parameter estimates from the Lasso regression model.

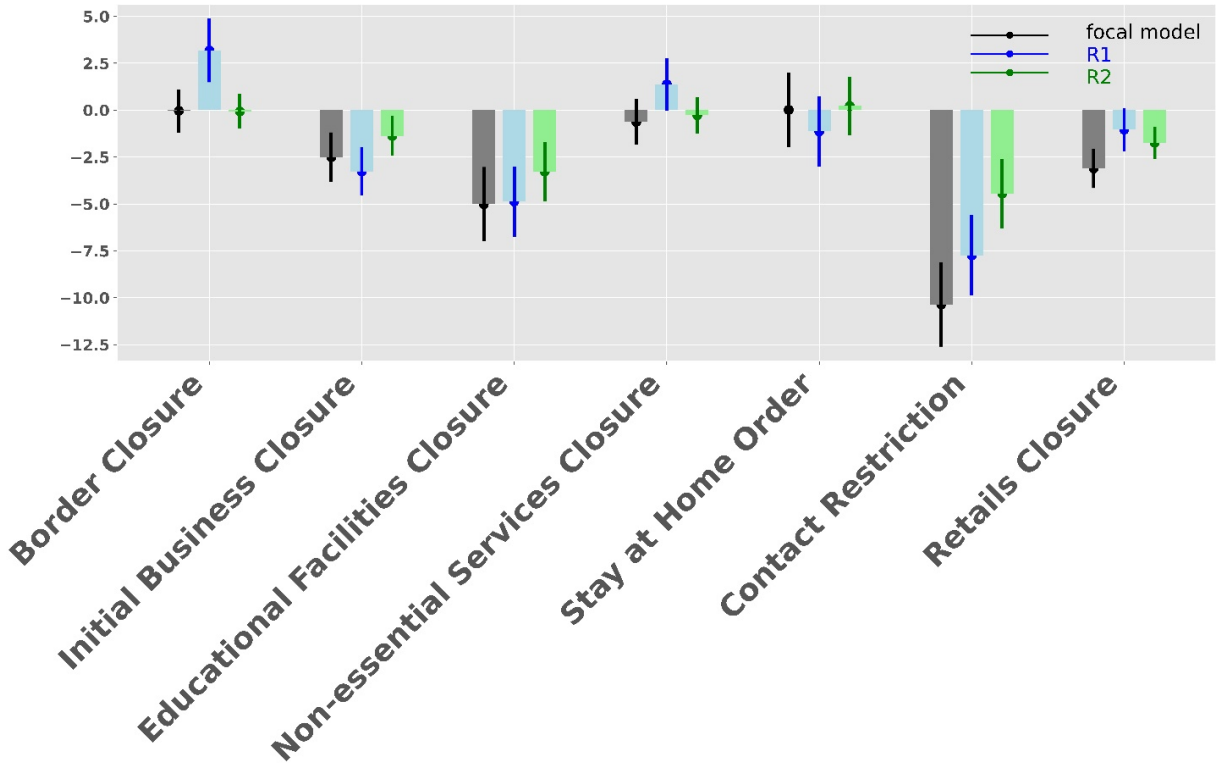


Figure 3.22: Parameter estimates from two robustness check models R1 and R2.

3.6.3.3 Robustness Check: Potential Spillover from Neighboring States

We already account for spillovers between states within the SEIR model by implementing most interstate movements (car, flight, bus, train, and truck). Furthermore, spillovers will also be accounted for by Google Trends, as people in a state would be more cautious and search for "COVID-19" on the internet more frequently if cases are rising in neighboring states. Moreover, we now include state-fixed effects to account for state-level heterogeneity. This new specification may also help alleviate potential spillover of a particular state from its neighbors.

Social distancing may be influenced by control variables outside of those considered in our focal model (Equation 3.6). We test two alternative model specifications to check the robustness of the parameter estimation. In the first alternative specification (R1), we

use social distancing numbers from neighboring states to account for spillovers from other states (Equation 3.10). We only consider states in Germany for neighboring states. NE_j is the number of neighbors for state j and $j(ne)$ represents the ne^{th} neighbor of state j.

$$C_{j,t} = K + \sum_{p=1}^{7} \beta_p^{policy} x_{j,p,t} + \beta_t^{trend} trend_{j,t} + \beta_m^{temp} temp_{j,t} + \sum_{j=1}^{16-1} \beta_j^{state} state_j + \sum_{w=1}^{9-1} \beta_w^{week} week_t + \sum_{d=1}^{7-1} \beta_d^{day} day_t + \sum_{ne=1}^{NE_j} \beta_{ne}^{neighbor} sd_{j(ne),t} + \epsilon_{j,t} \tag{3.10}$$

In the second model (R2), we use social distancing from a day before as a covariate in our model as shown in Equation 3.11. It controls for awareness in the previous day (i.e., if an individual did not go out due to COVID-19 yesterday, they will also not go out today due to COVID-19).

$$C_{j,t} = K + \sum_{p=1}^{7} \beta_p^{policy} x_{j,p,t} + \beta_t^{trend} trend_{j,t} + \beta_m^{temp} temp_{j,t} + \sum_{j=1}^{16-1} \beta_j^{state} state_j + \sum_{w=1}^{9-1} \beta_w^{week} week_t + \sum_{d=1}^{7-1} \beta_d^{day} day_t + \beta^{yesterday} sd_{j,t-1} + \epsilon_{j,t} \tag{3.11}$$

The coefficients of the focal model and both alternative specifications R1 and R2 are shown in Figure 3.22. We find that the coefficient estimates of all three models are close to one another. Given that there may be more covariates that could affect social mobility and are not included in our model, we cannot claim causality. However, these coefficients still provide significant insights into the potential of an NPI to induce social distancing in the community.

3.6.3.4 Leaving Out One Policy at a Time

NPIs were implemented across states in different sequence (Figure 3.1). However, the sequence they were implemented was not random; hence we cannot claim causality on the impact of an individual NPI (using coefficients from linear regression models) on social distancing. To check the robustness of the estimation of coefficients from the linear regression model, we simulate different scenarios when a particular policy (randomly selected) was never introduced in one of the states (randomly selected). Results for the parameter estimates from 10,000 simulations is shown in Figure 3.23. We design this

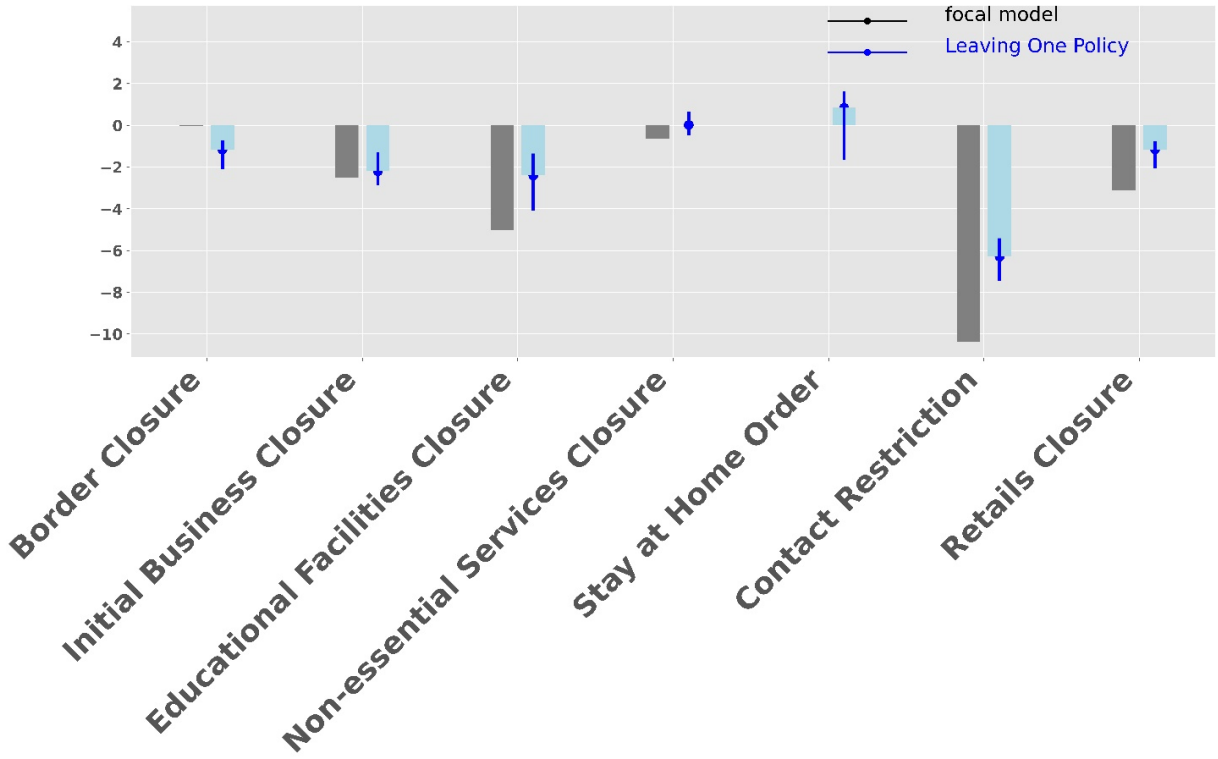


Figure 3.23: Parameter estimates from Linear Regression when one NPI is never implemented in one of the states (one state and one NPI selected randomly in one simulation). The grey bar shows the coefficients from linear regression. Green bars show the 50^{th} percentile of coefficients from the simulation. Vertical blue lines show the 25^{th} and 75^{th} percentile of the coefficients from the simulation.

simulation to bias the estimate strongly. It assumes that one randomly chosen NPI did not occur, while we kept the population's mobility changes in that state as observed in the data. We hope that the parameter estimates remain similar to the original ones in a robust model specification even though these strong biases are introduced. Results in Figure 3.23 show that the coefficients from the linear regression simulation by leaving our one policy in a state (where it was implemented) are close to the coefficients estimated from our focal model in Equation 3.6. We find that the estimated (biased) coefficients are comparable to the ones of the focal model and conclude that our model is robust to minor random changes in the sequencing on NPIs.

3.6.4 Generating Different What-If Scenarios

To understand the contribution of different policies in containing the spread of COVID-19, we consider different sets of scenarios for which policies were implemented and relaxed. State governments started introducing different policies around Mid-March. They started relaxing some of these policies on April 20, 2020 (Figure 3.4). We use data from February 18, 2020, to April 20, 2020. We create test scenarios with all the NPIs – border closures, educational activities closure, contact restriction order lifted, initial businesses opened, stay at home orders lifted, non-essential services opened, retail outlets opened, and all policies in place.

Figure 3.4 shows the timeline for policy introduction across different states in Germany. Some states did not introduce all the seven policies and, in some cases, introduced certain policies on varying dates, allowing for a quasi-experimental set-up to test the effect of policies on social distancing and the spread of disease subsequently. Next, we create eight counterfactual scenarios (one for each policy lifted and one for all policies in place) to study what will happen if the states lift the focal policy.

Different states across Germany started relaxing some policies on April 20, 2020. To determine the effect of lifting a policy, we examine policy relaxation in two scenarios: What would have happened if the policies were relaxed on April 21 or April 28 by easing one policy at a time to estimate their marginal effect. The week-long delay helps determine the increase in cases by relaxing a policy one week earlier.

3.6.4.1 Delay in NPI Relaxation

The differential effects allow for rank-ordering the policies by order of their impacts on disease spread. If a policy were relaxed on April 21 (without changing other policies), the social distancing would decrease after April 21. Similarly, if a policy was relaxed on April 28 (without changing other policies), social distancing will decrease after April 28 (without changing policies from April 21 to April 27 and keeping it as is). In this analysis, we simulate scenarios when the policies are relaxed (one at a time) on either April 21, 2020, or April 28, 2020, to understand the impact of delay in reopening. We predict mobility (hence social distancing) for 90 days from April 20, 2020. The social distancing

under different scenarios for different states are shown in Figure 3.24 and Figure 3.25.

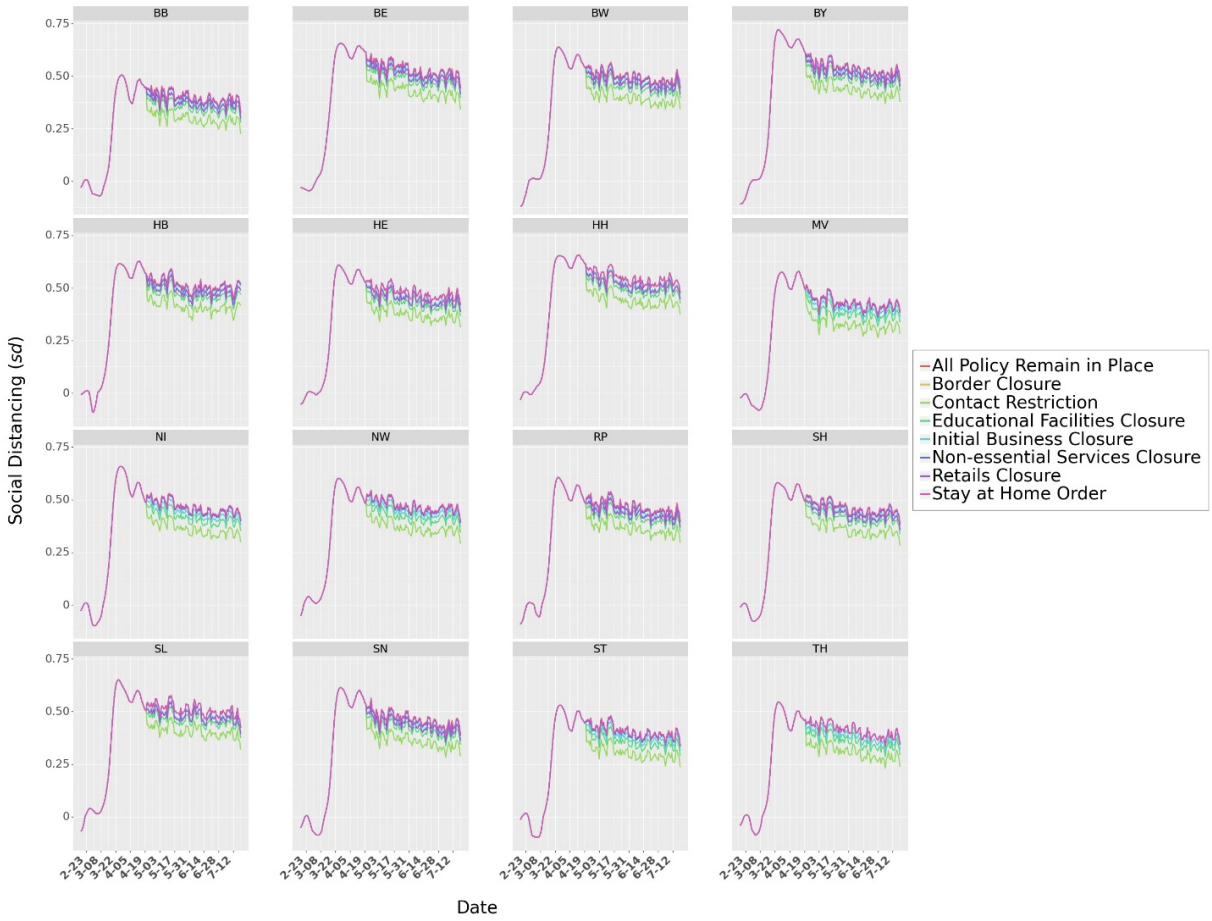


Figure 3.24: Scenario 1: Social distancing sd_i when policies are relaxed on April 21, 2020.

We use the Krinsky-Robb method [133] to estimate confidence intervals around the predictions. The Krinsky-Robb method assumes that coefficients of linear regression follow a multivariate normal distribution. Krinsky-Robb method then uses Cholesky decomposition of the covariance matrix of the coefficients to get random draws from a multivariate normal distribution. We use Monte Carlo simulation draws to get the confidence interval across the predictions (in different simulation draws). In our analysis, we use 10,000 random draws to estimate the confidence bounds for the predictions. Note that we use week fixed effects in our linear regression model, which is not available when projecting in the future. When generating scenarios, we predict based on 0 week fixed effects when predicting from April 21 onwards.

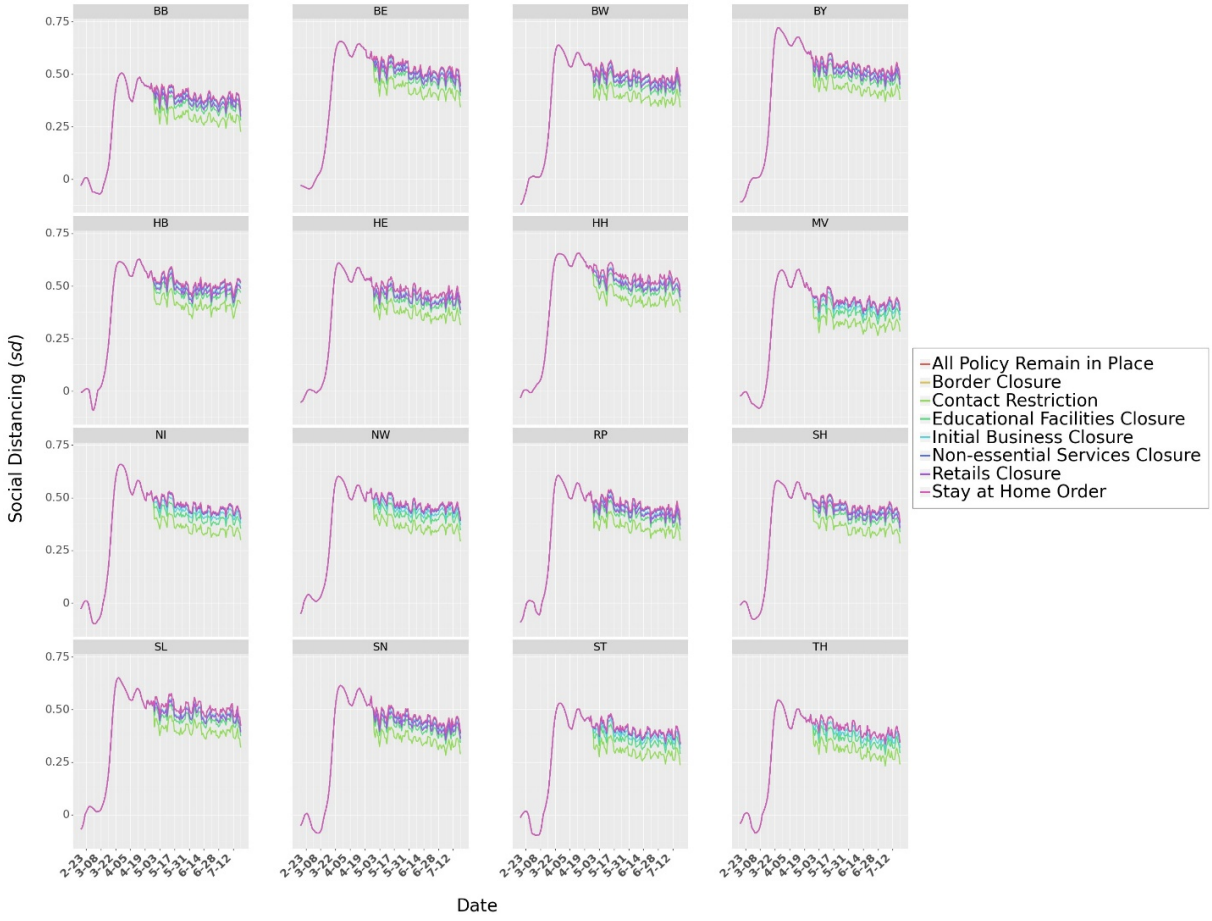
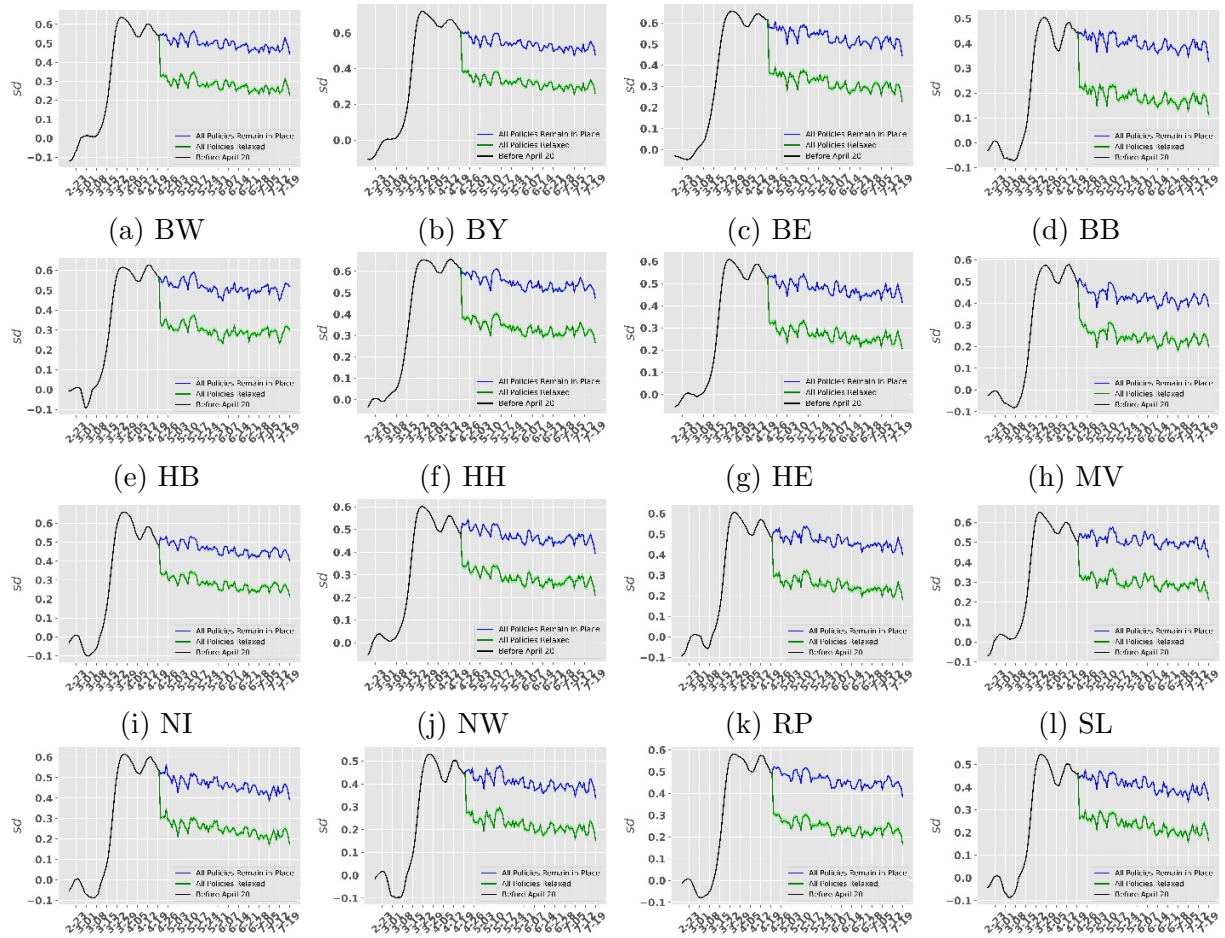


Figure 3.25: Scenario 2: Social distancing sd_i when policies are relaxed on April 28, 2020.

3.6.4.2 Mobility Bounce Back after NPI Relaxation

In this work, we postulate that the introduction of NPIs led to an increase in social distancing that helped contain the spread of the virus. However, NPIs cannot be implemented indefinitely and must be relaxed at some time. Due to increased awareness over time, we expect that mobility will not go back to normal ($C_{j,t} = 0$) at least in the near future (or right after a pandemic peak). Thus, we expect the social distancing to be greater than 0 even when all the NPIs are relaxed. To verify potential bounce back in community mobility, we predict community mobility if all the policies were relaxed on April 20, 2020. Social distancing for all the 16 states under all policies relaxed on April 20, 2020, is shown in Figure 3.26. It shows that even though social distancing decreases sharply when all the NPIs are relaxed, social distancing does not go down to 0 immediately, i.e.,

community mobility does not bounce back to baseline levels. This phenomenon is due to other covariates included in the model in Equation 3.6 e.g., Google Trends, maximum daily temperature, and various fixed effects. This prediction verifies that mobility does not go back to normal immediately, even when all the policies are relaxed.



(a) BW (b) BY (c) BE (d) BB

(e) HB (f) HH (g) HE (h) MV

(i) NI (j) NW (k) RP (l) SL

(m) SN (n) ST (o) SH (p) TH

Figure 3.26: Predicting social distancing under different scenarios. Green line shows the predicted social distancing when all the policies are relaxed on April 20, 2020. Blue line shows the predicted social distancing when all the policies remain in place till July 19, 2020. Black line shows actual social distancing numbers till April 20, 2020, derived from community mobility reports.

3.6.5 Effective Reproduction Number

The effective reproduction number R_e is the average number of newly infected cases caused by a single infected case on a given day. The effective reproduction number R_e for the

SEIR compartmental model can be calculated with the method introduced in [134] and later expanded in [135]. The value of R_e changes by increasing or decreasing the social mobility of both the infected and susceptible populations. Other factors impacting the value of R_e include environmental conditions and the drop in susceptible population over time.

In the SEIR model, $X = [E, I^d, I^u]$ are considered the infected compartments of the model. R_e can be calculated as the leading eigenvalue of the next generation matrix (NGM), $K = FV^{-1}$. $F = \frac{\partial F_i(x)}{\partial x_i}$ is the Jacobian matrix of the rate of new infections in infected compartments and $V = \frac{\partial V_i(x)}{\partial x_i}$ is the Jacobian matrix of the rate of transitions between infected compartments. Using Equations (3.1) to (3.4), we have:

$$F = \begin{bmatrix} \frac{(1-\gamma\times sd)\beta S I^d}{N} + \frac{(1-\gamma\times sd)\mu\beta S I^u}{N} \\ 0 \\ 0 \end{bmatrix}, \qquad V = \begin{bmatrix} \frac{E}{Z} \\ \frac{I^d}{D} - \frac{\alpha E}{Z} \\ \frac{I^u}{D} - \frac{(1-\alpha)E}{Z} \end{bmatrix} \tag{3.12}$$

$$F = \begin{bmatrix} 0 & \frac{(1-\gamma\times sd)\beta S}{N} & \frac{(1-\gamma\times sd)\mu\beta S}{N} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad V = \begin{bmatrix} \frac{1}{Z} & 0 & 0 \\ \frac{-\alpha}{Z} & \frac{1}{D} & 0 \\ \frac{-1-\alpha}{Z} & 0 & \frac{1}{D} \end{bmatrix} \tag{3.13}$$

And calculating R_e:

$$R_e = ((1-\gamma\times sd)\alpha\beta D + (1-\gamma\times sd)(1-\alpha)\mu\beta D)\frac{S}{N} \tag{3.14}$$

Figure 3.27 shows the effective reproduction number over time in different states of Germany using inferred parameters values. These values are in-line with the reported values for Germany 0.79 [95%CI: 0.66 – 0.90] using nowcasting approach as a moving 4-day average by Robert Koch Institute [127]. As depicted in this figure, the reproduction number at the beginning of the epidemic in Germany is above 1 in all states. Once the social distancing policies are in effect, this rate drops to below 1 in all states. When $R_e > 1$, the disease starts to spread throughout the population. Equation 3.14 shows that R_e also decreases with a drop in the susceptible fraction of population. A large proportion of population must transition from susceptible to infected, immune, or dead for this factor to be considerable.

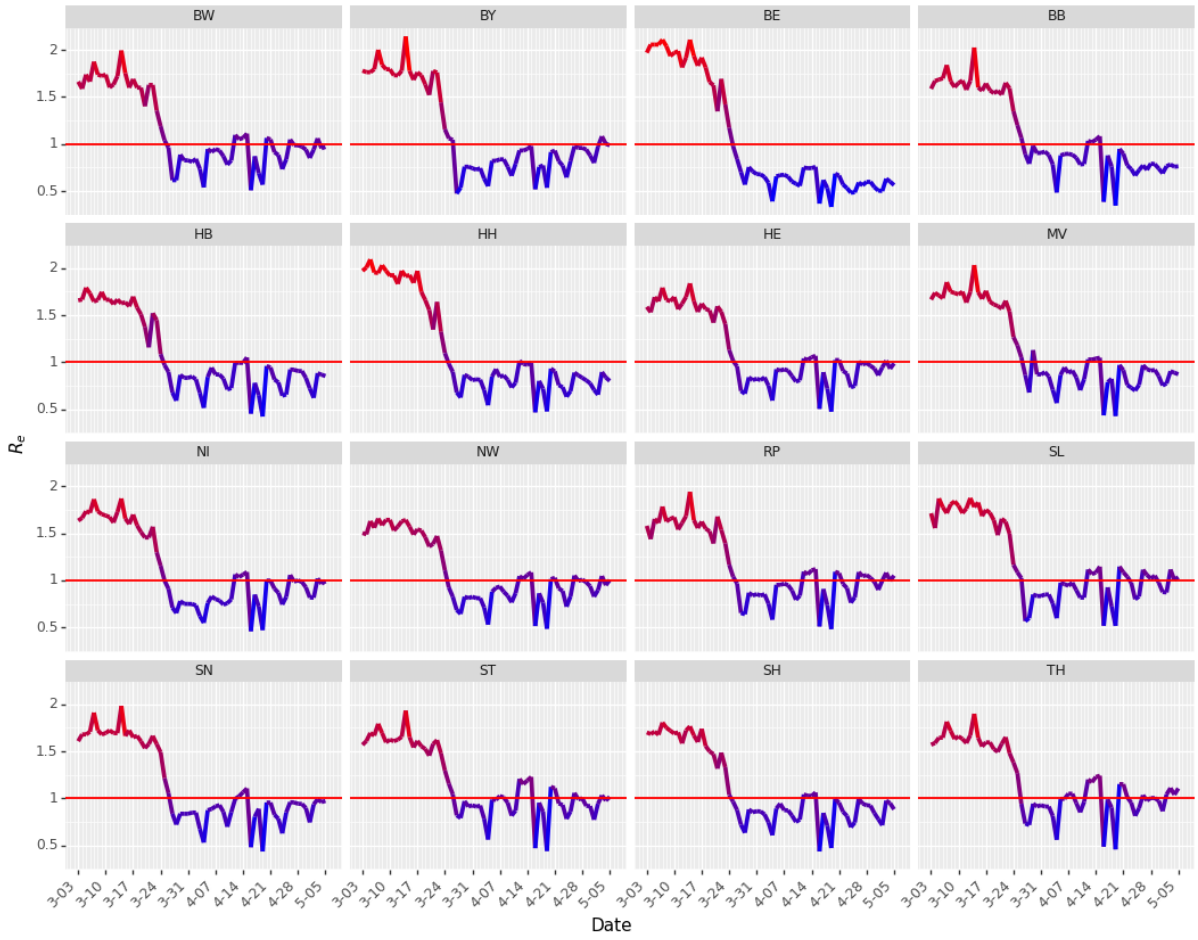


Figure 3.27: Effective reproduction number. The values of R_e drops from above the red line $R_e = 1$ to below $R_e = 1$ after policies are in effect.

To understand the effects of social distancing on R_e, we simulate the effect of lifting different policies on the value of R_e. We simulate different social distancing policies to be removed on April 21, 2020, and April 28, 2020. Removing some policies would increase R_e to above 1 in some states. For example, in large states such as Baden-Wurttemberg (BW) and Bayern (BY), R_e will rise above one if the contact restriction order is lifted or initial businesses are opened. Specific policies must be kept in effect until the susceptible proportion of the population drops significantly for R_e to remain below 1.

The dependency of R_e to different combinations of parameters α, γ, and sd is depicted in Figure 3.28. These figures represent the basic reproduction number where all population is assumed to be susceptible. Figure 3.28a represents the changes in R_e with respect to

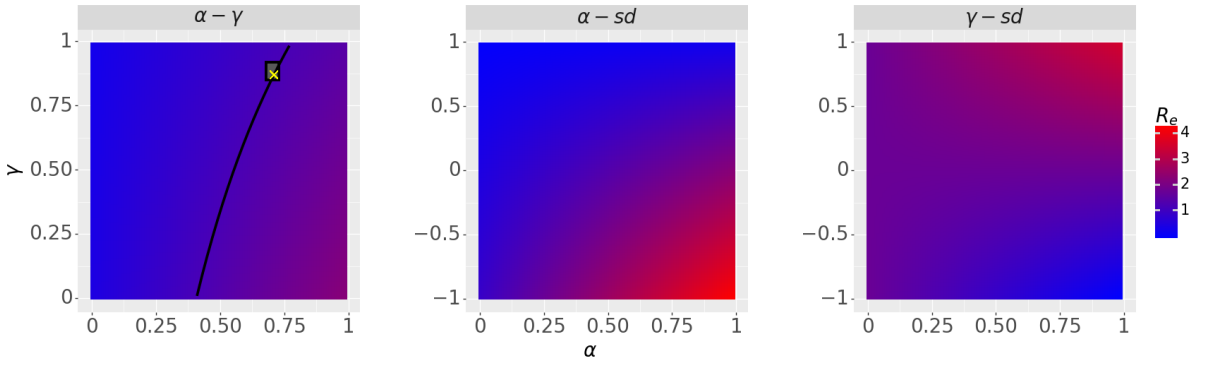


Figure 3.28: The dependency of R_e on model parameters α and γ and social mobility sd. All other variables are held constant with values represented in Table 3.2 for state BW. $\alpha-\gamma$: The impact of α and γ on R_e. The value of sd is set to $sd = 0.35$. Black curve represents the constant $R_e = 1.21$. The box represents the interval of inferred parameters that have the maximum likelihood and the yellow 'X' mark shows the parameter combination used for estimation ($\alpha = 0.71$, $\gamma = 0.89$). $\alpha - sd$: The impact of α and social mobility sd on R_e. Note that sd is not a model parameter. $\gamma - sd$: The impact of γ and social mobility sd on R_e.

possible values of parameters α and γ. High α and low γ combinations result in the highest R_e value. Figure 3.28b shows that R_e monotonically increases with decrease in sd and increase in α. Figure 3.28c shows that R_e will increase in combinations of high γ and low sd, while its value drops with high γ and high sd combinations. As γ decreases, the dependency of R_e to sd also declines.

Figure 3.29 shows the ratio of undocumented cases to daily infected cases across Germany from our model in Equations (3.1) to (3.4). The estimation shows that as the disease spreads, the documented infected portion of daily infected cases converges to a high value. This is likely due to the increasing awareness and higher availability of testing. One can also use the undocumented cases from the model to estimate the effective reproductivity number, R_e.

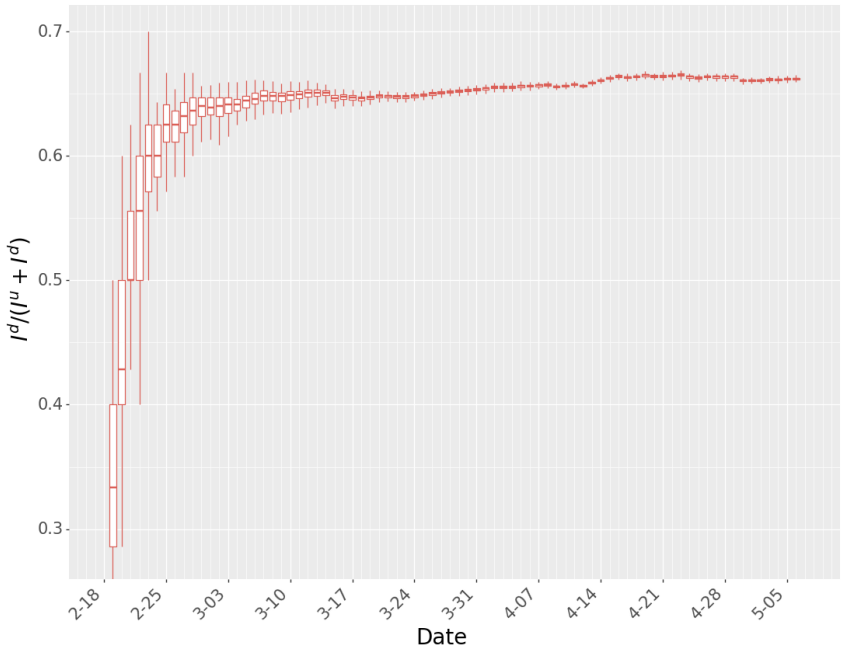


Figure 3.29: The ratio of documented infected cases to total daily (documented and undocumented) infected cases for Germany using SEIR model.

Chapter 4

The Impact of Mask-Wearing in Mitigating the Spread of COVID-19 During the Early Phases of the Pandemic

Similar to Chapter 3, this chapter aims to understand the impact of Non-Pharmaceutical Interventions (NPIs) in the spread of disease. However, the study in this chapter has some key differences from Chapter 3. In this chapter, we build a different epidemiology model (Susceptible-Exposed-Recovered, SIR) to estimate the spread of disease. Then we conduct a cross-sectional analysis of the effects of personal and governmental measures on mitigating disease spread. The dataset we use in this chapter is across 24 countries. During the dataset period, mask-wearing becomes a personal (and in some cases governmental) intervention in some countries to mitigate the spread of COVID-19. Therefore, we aim to understand the impact of mask-wearing in containing airborne diseases like COVID-19 while other covariates like NPIs and social distancing measures are simultaneously present. Another key difference is that the model proposed in this chapter estimates the growth rate of active daily cases. We discuss the advantages and limitations of the model in this chapter. The relevant findings in this chapter are aligned with corresponding results from the previous chapter.

4.1 Methods

This study is a cross-sectional analysis of the effects of personal and governmental measures across 24 countries on mitigating COVID-19 disease spread. The data used in this study were collected from February 21, 2020, to July 8, 2020, representing 139 days of data for each country. All analysis presented in this chapter uses publicly available data. Subsequently, we present the data on the three measures, namely, mask-wearing, outdoor mobility, and NPIs, and then discuss the model-based analysis.

4.1.1 Key Variables of Interest

Mask-Wearing: We study the impact of mask-wearing behavior using survey data across multiple countries released by the Institute of Global Health Innovation (IGHI) at Imperial College London, and YouGov [50]. The survey covers 26 countries (as of July 8, 2020), with around 21,000 people interviewed each week. Further details about the survey design can be found in Section 4.6.2.1. We restrict our analysis to 24 countries because two countries – China and Hong Kong – do not have publicly available data on outdoor mobility, which we control for in this study. The data present global insights on people's reported behavior in response to COVID-19. The dataset provides the percentage of the population in each country who report wearing a mask in public places. Because these surveys were conducted at an interval of several days, we interpolate (linearly) to estimate the percentage of the population that would wear masks in public spaces for days when the data are unavailable (Figure 4.1). We use the significant variation of mask-wearing across countries to measure the association of people reporting mask-wearing to the spread of COVID-19.

Outdoor mobility: Google Community Mobility Reports provide data on relative mobility changes with respect to an internal baseline across multiple categories, namely, retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential (Figure 4.2). A summary of the community mobility is shown in Table 4.2 in Section 4.6. Apart from the Google Mobility reports, we also utilize mobility data from Apple to test the robustness of the model to different measures of mobility. We note that neither Google nor Apple provides absolute mobility measures but rather presents relative

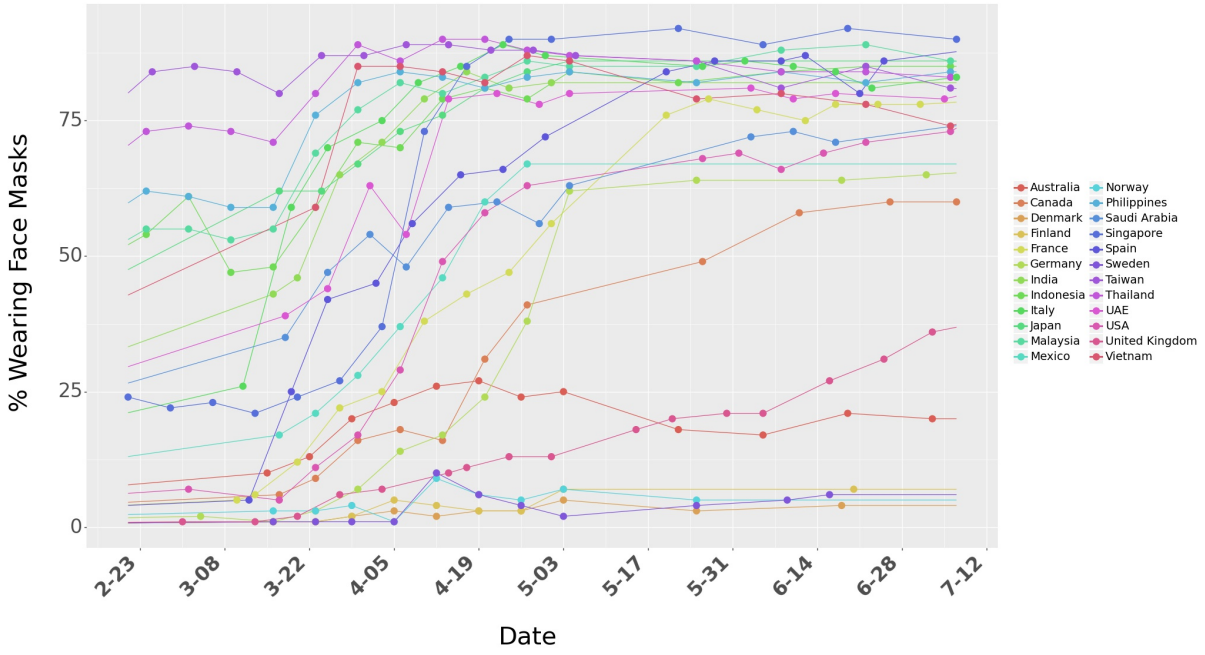
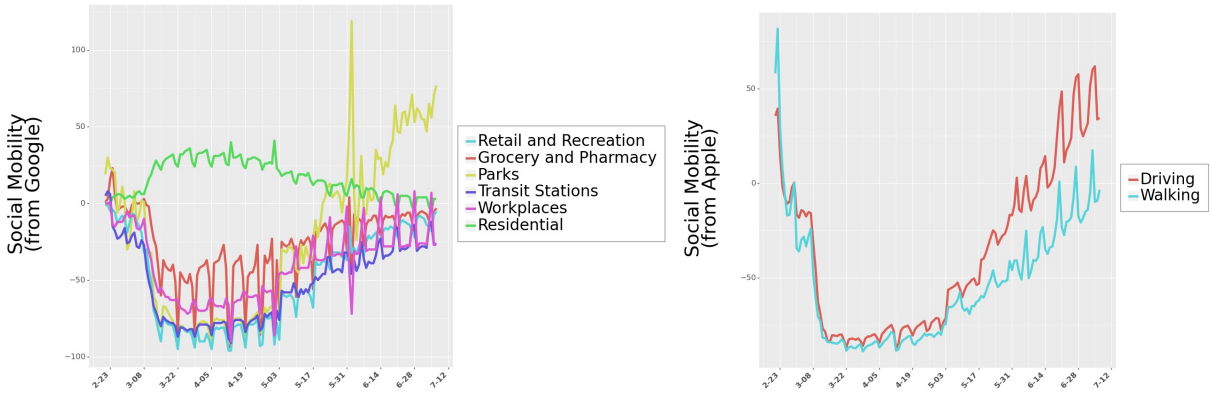


Figure 4.1: Percentage of people who say they are wearing a face mask in public spaces.

changes to benchmarks they use internally. Finally, drops in mobility could be driven by both individual actions (e.g., cautious behavior) and institutional actions due to NPIs enacted by governments. To control for mobility declines due to institutional actions, we also include country-specific interventions enacted both nationally and provincially.



(a) Outdoor mobility for Italy from Google Community Mobility reports

(b) Outdoor mobility for Italy from Apple Community Mobility reports

Figure 4.2: Outdoor mobility from Google Mobility Reports and Apple Mobility Reports for Italy. Outdoor mobility for all 24 countries is shown in Figure 4.14 in Section 4.6.

Non-Pharmaceutical Interventions: Governments across the 24 countries enforced

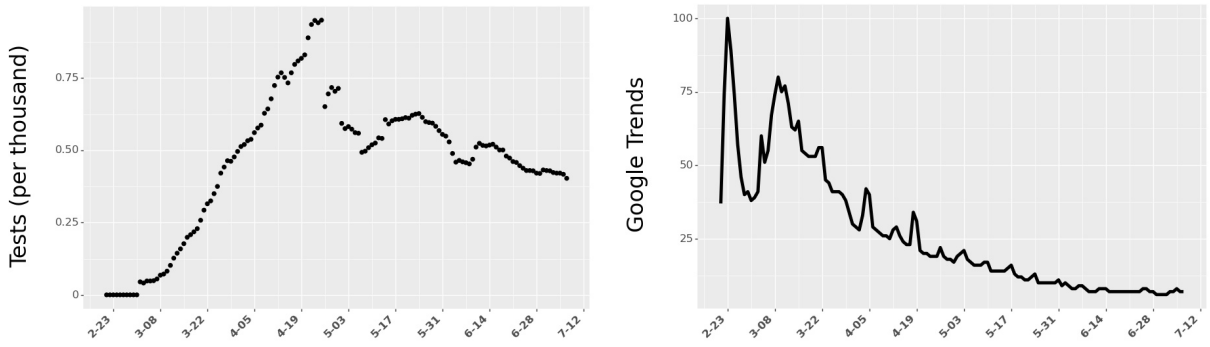
different policies to control the spread of COVID-19. Prior research has shown that these policies played a significant role in reducing the human to human physical contact and led to a slowdown in the spread of the disease [136]. However, these policies were implemented at different levels, some nationally, some provincially. We use data from COVID-19 Government Response Event Dataset [43] to control for government policies in estimating the effect of masks. Figure 4.16 lists the types and counts of national and provincial government policies implemented across the 24 countries we consider in this study. The dataset contains 5,816 entries on policies at the National and Provincial levels. Finally, the inclusion of these interventions helps control for some of the observed drops in mobility that are not necessarily associated with individual actions but with the presence of institutional policies. Section 4.6.2.5 provides detailed information about the interventions.

4.1.2 Covariates

Because the data span multiple countries and weeks, we include time and country fixed effects in the model. The model controls for country-level heterogeneity using fixed-effects, where the variable for a country assumes a value of one if the data considered are specific to that country and zero otherwise. This variable allows for control of country-level characteristics that are not in the model and helps reduce the errors due to omitted variables in our analysis. In addition to country-level differences, we also control for time-based differences (e.g., people are more aware and cautious over time) by incorporating time-fixed effects, where the variable $week_t$ takes a value of 1 if the data are from week t (where $t = 1$ represents the first week for a given country in the data). In addition, we control for each country's testing capability (Figure 4.3a) by accounting for the country's total number of daily tests. Finally, we also control for people's actions to educate themselves by including the Google Trends (Figure 4.3b) data for the search term 'coronavirus'.

4.1.3 Outcome Variable

Data for the number of active daily cases in each country were obtained from the Johns Hopkins University School of Public Health [137]. We use a seven-day moving average of

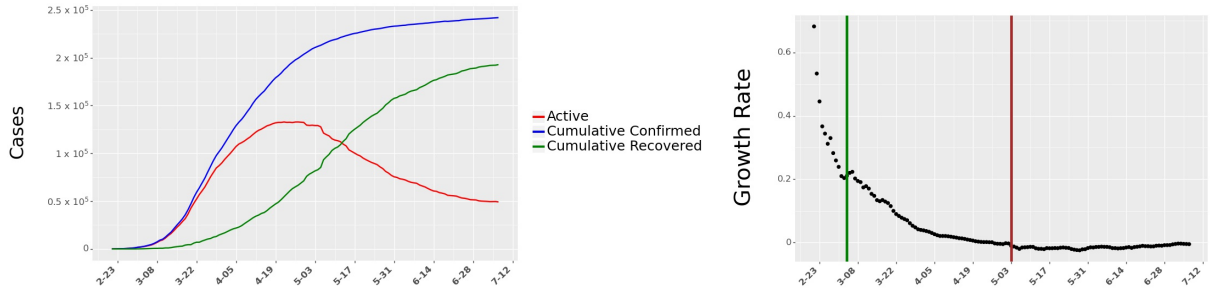


(a) Number of COVID-19 tests per thousand people in Italy

(b) Google Trends for search term coronavirus in Italy

Figure 4.3: Number of COVID-19 Tests per thousand people and Google Trends for Italy. Data for all the 24 countries is shown in Figures 4.19 and 4.20 of Section 4.6, respectively.

cumulative confirmed cases and cumulative recovered cases to compute daily active cases and daily growth rates. The daily growth rate is the ratio of active infections today to active infections the day before. The dataset aggregates this information across multiple national, state, and local health departments within each country. The daily growth rate is then related to the independent variables described earlier through a reduced-form econometric model. We describe the derivation in detail in Section 4.6.1.1. We illustrate the daily cases and growth rate for one country, Italy, in Figure 4.4.



(a) CoVID-19 Cases in Italy

(b) Growth rate in Italy

Figure 4.4: Daily active cases and growth rate of active cases for Italy. The vertical green line in (b) shows the start of data collection for Italy. The vertical red line in (b) shows the end of 60 days of data collection for Italy. Data for all 24 countries are shown in Figures 4.12 and 4.13 of the Section 4.6, respectively.

4.1.4 Analysis

A reduced form econometrics model was used to relate the growth rate of daily active infections to the independent variables described earlier. Similar models have been used by [45] to determine the effect of anti-contagion policies on the spread of COVID-19. In brief, the model assumes that the daily growth rate (ratio of active infections today to active infections the day before) is affected by institutional measures such as NPIs and individual measures such as outdoor mobility and mask-wearing. The covariates listed above help control for other factors that could affect growth over time. Because the epidemiological parameters for new diseases such as COVID-19 might not be well understood, reduced form techniques allow for the estimation of the impact of governmental and personal measures to help contain the spread of the virus. To filter out the high variation in growth rates when the number of cases is very low at the beginning of the pandemic, our model initializes when a country reaches 20% of peak new cases as observed by July 8, 2020. For robustness, we also test other starting times in Section 4.6 and find results in line with the ones presented here. Section 4.6 also provides further details about this chapter's methodological approach and model formulation. We provide some brief notes on the operationalization of the independent variables and the model initialization below:

1. Responses to the survey about mask-wearing are subject to biases. For example, individuals might overestimate the efficacy of their mask or their wearing pattern. To alleviate some of these concerns, we compute the natural log of the mask-wearing variable to discount its impact on the growth rate of daily active cases. This transformation yields a curve that grows at a slower rate as the values of mask-wearing increase, thereby diminishing the impact of higher levels of mask-wearing. We also test for other functional forms (square-root and linear) and present those results in the Section 4.6 (Table 4.7).

2. Due to the high correlation across the different mobility data categories obtained from Google, we only include mobility categories 'Parks' and 'Transit Stations' in the model. Because we are interested in determining the impact of mobility in general, the two mobility variables suffice in capturing the individual's movement

patterns during this time. In Section 4.6.3.4, we present results including other mobility types and also run the model with Apple Mobility data in place of Google Mobility Reports.

3. The CoronaNet dataset from [43] collected information on all the government policies introduced by different countries across the world. They categorized the policies into 19 different policy types. We use their categorization in the model. From February 21, 2020, to July 8, 2020, we check if a policy p was implemented in a country j on the day t. If the policy was implemented, we assign a value of 1 to $s_{j,t,p}$, where s represents the level of policy coverage. If the policy was introduced at a provincial level, we normalize $s_{j,t,p}$ by the population of the state. Because several policies were introduced simultaneously or close together, they suffered from collinearity issues. To minimize multicollinearity issues, we choose only a specific set of policies to include in the analysis. Section 4.6.2.5 discusses this selection mechanism.

4. Due to the uncertainty of the lag in COVID-19 incidences and the difficulties in detection during the early days of the disease [138], similar to prior research, we tested the focal model across multiple lag periods ($shift$) from zero to 14 days and for different initialization thresholds (th) for each country (zero percent to 20% of a country's peak daily cases by July 08, 2020). We chose the best $shift$ and th values using a k-fold cross-validation process (k=5). The chosen model had the highest maximum likelihood estimate of the data and the lowest prediction error. We discuss this procedure in Section 4.6.4.1. The results presented in the next section correspond to a model with a $shift$ of nine days and a th of 20% of peak new cases by July 12, 2020. Finally, we train the model on 1,422 observations across 24 countries. We restrict our analysis to the first 60 days after model initialization based on th. However, we test the robustness of the findings for other lengths of data. This allows for greater variation in mask usage within the data.

In the next section, we describe our results and their policy implications.

4.2 Results

The results indicate that individual measures such as mask-wearing and outdoor mobility combined with institutional measures (NPIs) play a role in mitigating the spread of COVID-19. Figure 4.5 shows the estimates from the focal reduced-form model for these measures and their corresponding confidence intervals. Section 4.6.3 provides the full table of results, along with results for all robustness checks. We first list the results of the key measures we consider and then discuss their implications.

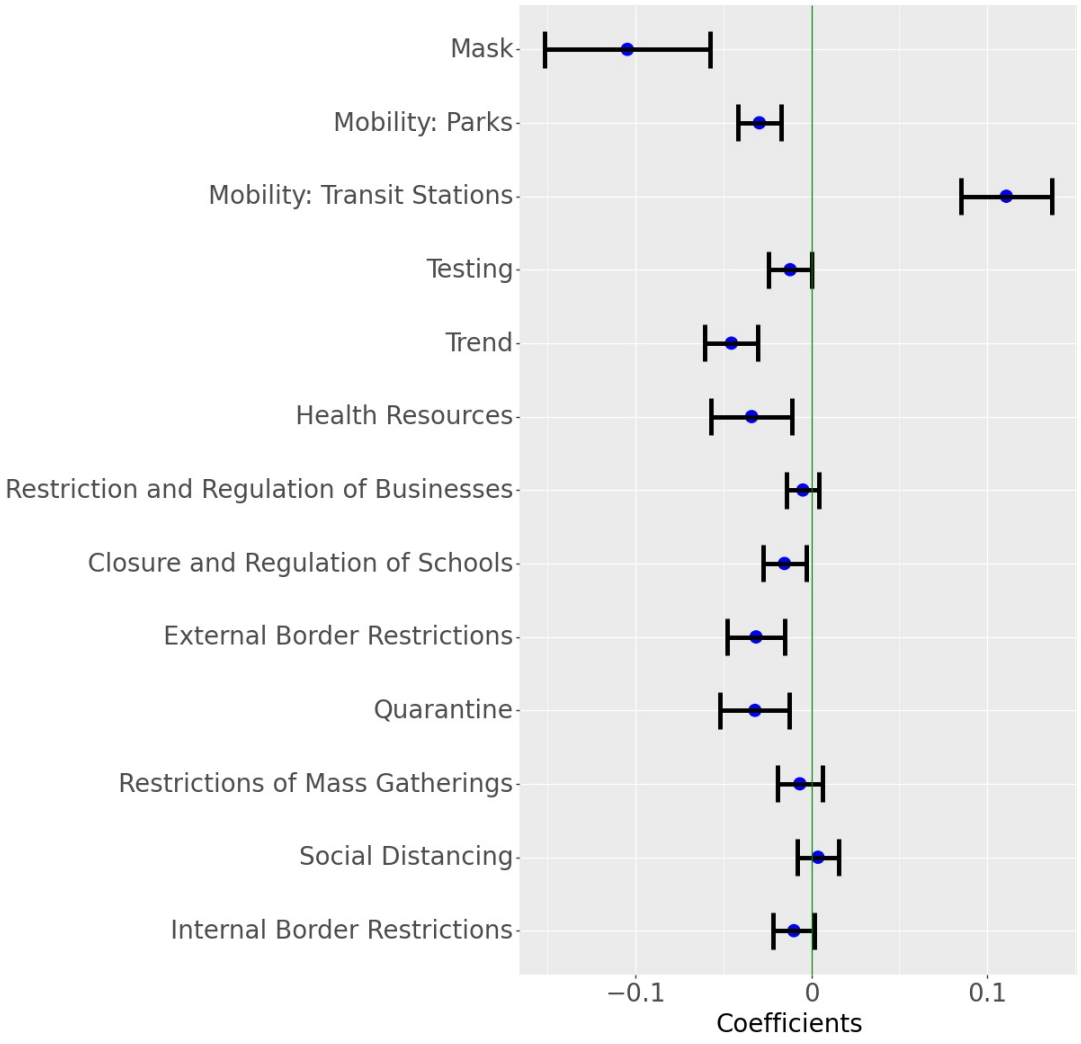


Figure 4.5: Parameter estimates for the Growth rate model. The blue dot indicates the point estimates and the horizontal lines indicate 95% confidence interval around the estimate (Masks transformed as ln(1+mask)). Results for different transformations of Mask and other covariates are shown in Figure 4.23.

Mask-Wearing: The model finds that a reported mask-wearing of 100% associates with an average 7% (95% CI: 3.94% — 9.99%) drop in the daily growth rate of COVID-19 cases. While this daily effect appears small, 100% reported mask-wearing leads to approximately 88.5% (95% CI: 68.7% — 89.2%) decline in active cases over 30 days compared to the situation where 0% of the people report wearing masks (all else remaining the same across the two scenarios). Modifying the functional form of the mask variable did not appreciably change the association. For example, in the linear model, masks are associated with an average 8.69% (95% CI: 5.63% — 11.66%) drop in daily growth rate, and for the square root model, the expected daily drop in growth rate was 7.89% (95% CI: 4.81% — 10.87%). The stability of the results indicates that mask-wearing plays a significant role in mitigating the spread of the disease. Figure 4.5 also illustrates that widespread mask-wearing, as an intervention by itself, has the most significant association (by magnitude) with the growth rate of active COVID-19 cases. Figure 4.6 plots the ratio of active cases under different proportions of respondents who claim to wear masks against no mask-wearing and for various periods.

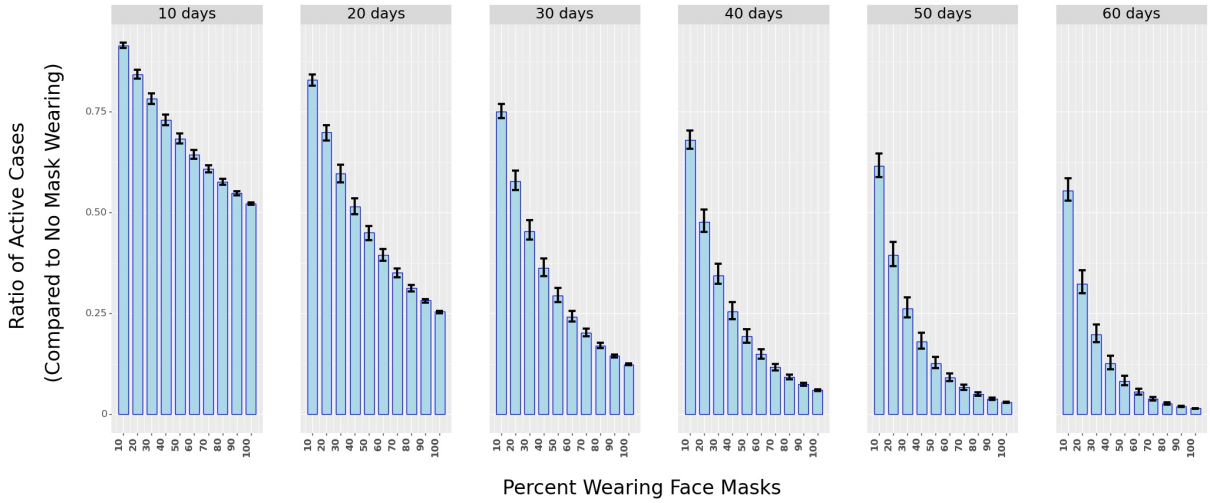


Figure 4.6: The ratio of active cases under different percentages of mask-wearing in public spaces as compared to 0% mask-wearing over different periods (in days). The shaded bars represent in ratio while the black vertical lines represent the 25^{th} and 75^{th} percentile of the ratio (from simulations using Krinsky-Robb method).

Mobility and NPIs: As expected, the model finds that a rise in mobility links with

a rise in the number of cases. Specifically, the selected mobility variables associate with a combined 8.1% (95% CI: 5.6% - 10.6%) drop in daily case numbers. Similarly, we find that the implementation of NPIs is also associated with a drop in daily growth rates across countries. After accounting for mobility declines, the NPI measures 'Quarantine', 'External Border Restrictions', and 'Closure and Regulation of Schools' link with the highest declines in the growth rate of daily active cases. Overall, all NPIs included in the model led to a decrease in the growth rate of COVID-19. This finding confirms multiple studies that investigated the effects of NPIs at limiting the spread of COVID-19 [30, 52, 27]. Overall, we find that if the NPIs were enacted uniformly across the whole country, then the combined association of the NPIs with the decline of growth in daily cases of COVID-19 would average 13% (95% CI: 9.2% - 16.2%). We determine the combined effect using the Krinsky-Robb method, a Monte Carlo simulation used to draw samples from a multivariate normal distribution. Section 4.6.3.1 provides more details on this method.

4.2.1 Controlling for Endogeneity using Control Functions

Due to nearly concurrent enactments and blanket coverage of policies and precautionary behaviors within countries, the individual (e.g., masks, limiting mobility) and institutional (NPIs) measures correlate in time. This correlation precludes the causal identification of each measure's effect on disease mitigation. In other words, because mask-wearing, mobility reductions, and NPIs occur at similar times, their effects are intertwined and difficult to determine separately. For some variables such as mobility and NPIs, we lack the necessary data to fully control for these issues. In the case of mask-wearing, even though we cannot eliminate all the possible endogeneity issues, we attempt to alleviate some of the concerns of confounding variables by employing control functions [139]. As noted in [140], control functions make the intervention exogenous in a regression equation. To create a control function, we use mortality data for prior outbreaks of SARS, H1N1, and MERS in each country as instrumental variables to predict the proportion of mask-wearing in each country (see Section 4.6.4.5 for more details). We posit that exposure to prior pandemics would have resulted in a more aware populace that could be amenable to

precautionary behaviors such as mask-wearing. Next, we compute the control function by determining the predicted mask-wearing residuals (computed via determining "Predicted Mask-Wearing minus Reported Mask-Wearing"), allowing for better identification of the effect of reported mask-wearing on COVID-19 case numbers.

Using this procedure, we find that if 100% of the population claimed to wear masks, then mask-wearing relates with an average of 4.95% (95% CI: 2.26% — 7.53%) drop in the daily growth rate of COVID-19, when compared to zero percent reported mask-wearing. Over 30 days, this translates to a 70.4% (95% CI: 62.3% — 72.7%) drop in new COVID-19 cases. While we are careful to note that this estimate could still be affected by confounding variables, this result lends stronger support to the magnitude of the disease mitigation that mask-wearing in the general population provides. In summary, widespread mask-wearing leads to a significant decline in the spread of COVID-19.

4.2.2 Robustness Checks

To help determine the accuracy and stability of the results, we run several robustness checks (see Section 4.6.4 for details):

1. We vary the lag period ($shift$) from 0 to 14 days. The results show that the estimates of the individual and institutional measures are relatively stable.

2. We also vary the length of time we consider in the analysis. The model considered 60 days of data for each country. We vary this to estimate the model on 35, 45, 55, 65, 75, and 85 days of data and find that the results remain stable to these variations.

3. We replace Google mobility data with Apple mobility data. The model estimates remain robust to this change.

4. We vary the functional form of how mask-wearing relates to the spread of COVID-19. The results are not statistically different in these cases.

5. We also test the robustness of the analysis by modifying the data using exponential smoothing. Specifically, for any day t, the focal model in Equation 4.8 ignores

the value of the independent variables from days $t - shift + 1$ to t (discussed in Figure 4.7). In the model we use for the robustness check, we do not ignore values between $t - shift$ and t and use exponential smoothing to average the intervening data. Finally, we also modify the interpolation method of mask-wearing data from linear (current) to quadratic. We find that the results are stable with all these modifications.

Section 4.6 details all the robustness checks and simulations as well as their results.

4.3 Discussion

Over the past few months, several studies have investigated the efficacy of masks at minimizing droplet dispersion [141, 142] and the potential consequences of their use [143, 54] in the general population. Although a randomized control trial on the efficacy of face mask usage appears to indicate inconclusive results in the general population [57], [56, 55] provide evidence for the benefits of face mask usage through a systematic review of the multiple observational studies and the evidence thus far. While the type of face mask, as well as the timing and length of use, can affect its efficacy, its use as a precautionary principle has been strongly advised [144]. Despite the abundant scholarly and some anecdotal evidence [46], face mask use in some countries like Sweden and the United States remains controversial [145, 146, 147]. Additionally, as observed in the data, even in countries where masks do not face similar headwinds and as support for mask usage gathers further evidence, face mask use is not as commonplace (e.g., Denmark, Norway, Sweden, Finland), even as a precautionary principle.

This study links the growth rate in active cases of COVID-19 in a country to a population's reported wearing of face masks in public places over time. The model also includes other measures that could simultaneously impact the spread of the disease as face mask usage changes over time. After accounting for these measures and controlling for other covariates, the results indicate that reported face mask use is associated with a decline in the growth of COVID-19. More precisely, if 100% of the population claimed to wear masks, then mask-wearing is associated with an average 7% decline in the growth of daily

active cases of COVID-19. This association persists across multiple robustness checks and model formulations. A decline of 7% corresponds to an 88.5% drop in the number of active cases 30 days later. Together with the other measures (mobility changes, NPIs), the combined association of individual and institutional measures on the decline in the growth rate of daily active cases of COVID-19 is 28.1% (95% CI: 24.2%-32%).

4.3.1 Limitations

Countries enacted multiple NPIs simultaneously. This precludes us from identifying the effectiveness of NPIs separately. Second, the mobility data provided by Google and Apple are only indicative of the relative changes from a benchmark, so their association with disease spread should be interpreted with precaution. Third, we rely on the accuracy of data collected by third parties. Inconsistencies in testing, reporting, and recording the data could lead to errors in the results obtained. Additionally, mask-types and mask-wearing patterns could vary across countries, individuals, and time. This limitation affects all observational COVID-19 population-based studies.

4.4 Conclusions

The population-wide usage of face masks as a preventative measure against the transmission of COVID-19 varies widely across countries. Using data from 24 countries, this study finds that face mask usage is associated with a decline in the growth rate of daily active cases of COVID-19. Over 30 days, mask-wearing associates with an 88.5% decline in the number of daily active cases. This result re-affirms the paramount importance of masks in combating the spread of COVID-19.

4.5 Data and Code Availability

All codes are written in the Python 3.7 programming language. All data and codes are available at the public GitHub repository at: `https://github.com/ashutoshnayakIE/COVID-masks`. Data is in python's numpy format. However, the raw data can be procured from the sources mentioned in the references above (link for raw data set is also provided in the Github repository).

4.6 supplement

4.6.1 Method

This section provides a detailed description of the reduced form econometrics model considered in our analysis. The model is derived from Susceptible-Infectious-Recovered (SIR) epidemiology models.

4.6.1.1 SIR Growth Rate Model

Similar to the work by Hsiang et al. [45], we employ a reduced form econometrics technique that relates the growth rate of active COVID-19 cases to the individual and institutional measures such as masks, social mobility, and governmental Non-Pharmaceutical Interventions (NPIs). The growth rate in econometrics is the first difference in the log of economic outputs in different periods. The growth rate model is a well-established method in econometrics where growth rates of economic output can be affected by different factors, e.g., policy. Similar to economic output, we model the growth rate of daily active cases and estimate how it is affected by wearing masks in public spaces, social mobility, and NPIs. The method also has roots in epidemiology models – SIR (Susceptible, Infection, Recover). We do not consider deaths and reinfection in this analysis.

Equations (4.1) to (4.4) describe the SIR model where $S_{j,t}$, $I_{j,t}$, and $R_{j,t}$ show the active susceptible, infectious and recovered population at time t in country j. β_j is the rate of transmission and γ_j is the rate of recovery in country j. Since we do not consider reinfection and deaths, γ_j can be considered as the rate of removal from infectious population. N_j is the total population of the country j. Equation 4.1 shows how infections spread from the infectious individuals to susceptible individuals. Equation 4.2 shows how infectious population changes over time as some susceptible individuals contract the disease while some already infectious individuals recover from the disease and test negative. Equation 4.3 shows how the number of recovered individuals increase over time as individuals recover after testing negative for the virus. Equation 4.4 is a feasibility constraint which ensures that the total population is accounted for in the model. Addition

of Equations (4.1) to (4.3) yields Equation 4.4.

$$\frac{dS_{j,t}}{dt} = -\frac{\beta_j I_{j,t} S_{j,t}}{N_j} \tag{4.1}$$

$$\frac{dI_{j,t}}{dt} = \frac{\beta_j I_{j,t} S_{j,t}}{N_j} - \gamma_j I_{j,t} \tag{4.2}$$

$$\frac{dR_{j,t}}{dt} = \gamma_j I_{j,t} \tag{4.3}$$

$$\frac{dS_{j,t}}{dt} + \frac{dI_{j,t}}{dt} + \frac{dR_{j,t}}{dt} = 0 \tag{4.4}$$

Since we model only the growth rate in the total confirmed cases, we consider Equation 4.2 in our analysis. Assuming $S_{j,t} \approx N_j$, we can rewrite Equation 4.2 as shown in Equation 4.5. It can be solved by integration as shown in Equation 4.6. If we consider daily growth rate ($t_2 - t_1 = 1$), Equation 4.6 can be simplified as shown in Equation 4.7, where g_j is the growth rate and it is given by $\beta_j - \gamma_j$.

$$\frac{dI_{j,t}}{dt} = (\beta_j - \gamma_j) I_{j,t} \tag{4.5}$$

$$\int_{t_1}^{t_2} \frac{dI_{j,t}}{I_{j,t}} = log(I_{j,t_2}) - log(I_{j,t_1}) = (\beta_j - \gamma_j)(t_2 - t_1) \tag{4.6}$$

$$log(I_{j,t_2}) - log(I_{j,t_1}) = g_j(t_2 - t_1) \tag{4.7}$$

Wearing face masks, reducing social mobility, and implementing NPIs can alter the growth rate by changing g_j. We include country fixed effects ($country_j$) to account for country-specific heterogeneity in g_j. Equation 4.8 represents the growth rate model (P is the set of policies; M is the set of indicators of social mobility, W is the set of weeks during the period of our analysis, and J is the countries in our analysis). $mobilityy_{j,t,m}$ is the m^{th} indicator for social mobility, $week_{j,t,w} = 1$ if day t in country j is in week w after the initialization point for country j. To account for other factors, we consider several control variables (e.g., testing, Google Trends, fixed effect for week) as discussed in the next section.

$$\begin{aligned} g_{j,t+shift} = \theta_0 + \theta_c mask_{j,t} + \sum_{p \in P} \theta_p policy_{j,t,p} + \sum_{m \in M} \theta_m mobility_{j,t,m} + \theta_e testing_{j,t} + \\ \sum_{w \in W} \theta_w week_{j,t,w} + \theta_r trend_{j,t} + \sum_{j \in J} \theta_j country_j + \epsilon_t \end{aligned} \tag{4.8}$$

We use government announcements on health resources, health monitoring, and tests conducted (per thousand individuals) to account for increased awareness and testing over time. We also use Google Trends on the keyword '*coronavirus*' to account for public self-awareness. We will discuss NPIs, Testing, and Google Trends later.

The econometrics approach of using the growth rate to estimate the effects of masks, social mobility, and NPIs has several advantages. The model can estimate the effect of the exogenous independent variables on the dependent outcome variable (growth rate). Since the right-hand side of Equation 4.5 can be empirically calculated, it does not explicitly require the knowledge of the relationship between exogenous variables and $I_{j,t}$. Thus, the model does not need to know the link between masks, NPIs, and social mobility on daily active cases (or cumulative conformed cases) but can still estimate their effect on the growth rate of infectious cases. Using the growth rate, $I_{j,t}$ can be estimated by integrating it from time 0 (or using previous integration up to the day $t-1$). Thus, this model is forward-looking.

The model is also able to handle underreporting in COVID-19. In the COVID-19 pandemic, data for an individual is recorded when only they are tested. Total confirmed cases (deaths and recovered cases) in publicly available datasets only provide information on the individuals who got themselves tested. Due to various reasons, e.g., lack of testing, lack of motivation to get tested, or lack of visible symptoms in symptomatic cases, it is being estimated that there is a massive underreporting in total confirmed positive cases. However, the growth rate model is agnostic to underreporting as it models the first difference in the log of confirmed cases. If the underreporting remains constant, we can multiply $I_{j,t}$ and $I_{j,t-1}$ with a constant over the period considered, and it would not affect our estimation of growth rate (Equation 4.5).

Multiple studies have reported delays between the association of policies with COVID-19 spread [45, 53]. This delay could be due to several reasons. One of the most commonly noted reasons is the incubation period (time between getting infected and onset of symptoms/knowing that individual is confirmed for COVID-19). During the incubation period, an individual may be asymptomatic. Incubation period is estimated to be 4 days to 14

days [130]. Another reason could be the testing time – the time it takes to get the confirmation of results. To model this delay, we use a lag variable. We use the cross-validation method to find the lag with the best fit for the data. We test and observe that the model performs best with a lag of 9 days.

4.6.1.2 Robustness and Control Function Models

Robustness Model: As a robustness check to the growth rate model discussed above, we also use an exponential smoothing model to estimate the effect of masks, social mobility, and NPIs to validate the results from the base model in Equation 4.8. In this model, we use exponentially smoothed data for the right-hand side of Equation 4.8 for the past w days, without using the lag variable. Equation 4.9 shows this model where $< x >$ is exponentially smoothed over the last w days. Smoothing function is shown in Equation 4.10. This method considers data in the recent future of day instead of considering all the data leading up to the day or data observed lag days before as in the growth rate model in Equation 4.8. This model is analogous to counting the number of days a policy was active in the past w days. We use exponential smoothing to include the lag effect of masks, NPIs, and mobility on the growth rate.

$$\begin{aligned}\overline{g_{j,t+shift}} = \theta_0 + \theta_c < mask_{j,t} >_w + \sum_{p \in P} \theta_p < policy_{j,t,p} >_w + \sum_{m \in M} \theta_m < mobility_{j,t,m} >_w + \\ \theta_e < testing_{j,t} >_w + \sum_{w \in W} \theta_w week_{j,t,w} + \\ \theta_r < trend_{j,t} >_w + \sum_{j \in J} \theta_j country_j + \epsilon_t\end{aligned} \tag{4.9}$$

$$< x >_w = \frac{\sum_{l=1}^{w} 0.8^{w-l} x_{t-w}}{\sum_{l=1}^{w} 0.8^{w-l}} \tag{4.10}$$

Figure 4.7 shows the idea behind the exponentially smoothed model. In the growth rate model, we consider events that happened at a lag of $shift$ days. Thus, on day t we assign 0 weights to data from $t + shift - 1$ to t. In the exponentially smoothed model, we assign exponentially reducing but non-zero weights to data for days $t + shift - 1$ to t.

Control Function Model: We also consider a control function approach to check the robustness of the mask parameter from Equation 4.8. Countries have had different expe-

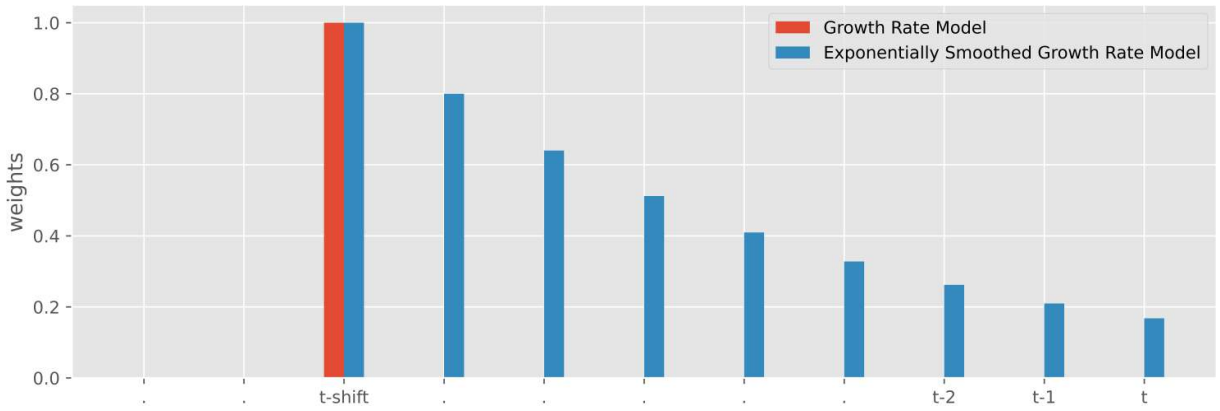


Figure 4.7: Weights given to different data points for two models – growth rate model in Equation 4.9 (red) and exponentially smoothed growth rate model (blue) in Equation 4.10. Exponentially smoothed growth rate model consider all the data points in recent history to day t instead of considering the events only on day $t - shift$. Weights decrease as we move closer to day t to incorporate the delay in observing the effect of events in the recent future.

riences with airborne diseases due to multiple outbreaks in the past, e.g., Severe Acute Respiratory Syndrome (SARS), Middle East Respiratory Syndrome Coronavirus (MERS-CoV), and H1N1 Influenza (Swine Flu). Countries with a severe outbreak of these airborne viruses were quick to adopt wearing face masks in public. This effect could potentially confound with the effect of the mask discussed in this chapter. Therefore, we use a control function approach to isolate the effect of masks. Since the growth rate in COVID-19 is independent of the number of deaths per thousand people from SARS, MERS, and H1N1, it may affect the percentage of the population wearing mask $mask_{j,t}$ but does not affect the growth rate. Thus, we use deaths from previous diseases as a control function. Figure 4.8 shows the total deaths per thousand people for different countries.

In the control function model, we first predict average mask wearing in country j, $\widehat{mask_j}$ using the number of deaths from SARS, H1N1 and MERS in country j as covariates in ordinary least square linear regression estimation. We use $d_{j,dis}$ as predictor variable where $dis \in$[SARS,H1N1,MERS] and $d_{j,dis}$ is the number of deaths per thousand people in country j from disease dis. Specifically we use $\widehat{d}_{j,dis} = 1$ if $d_{j,dis} > median(d_{dis})$, and 0 otherwise, where $median(d_{dis})$ is median among countries for a specific disease dis. The model to predict $mask_{j,t}$ is shown in Equation 4.11. After estimating $\widehat{mask_{j,t}}$, we use the

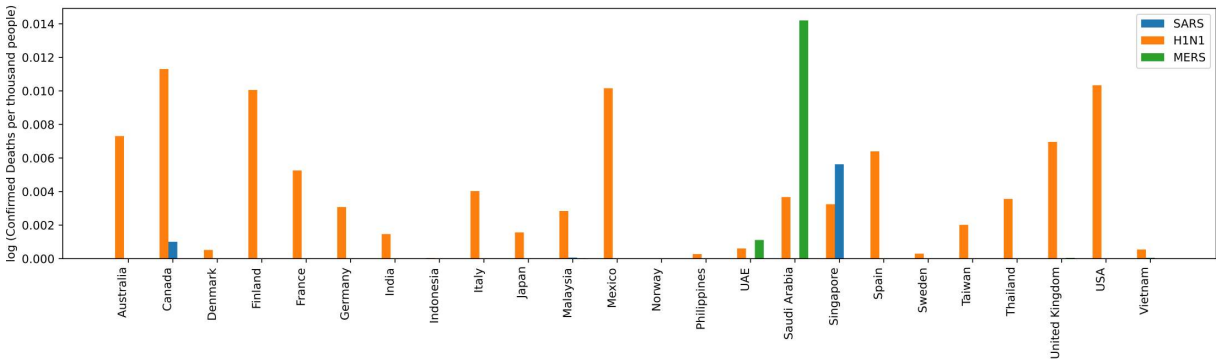


Figure 4.8: Logarithm of number of confirmed deaths with SARS, H1N1, and MERS-CoV per thousand people across 24 countries considered in this study.

error $mask_{j,t} - \widehat{mask_j}$ in Equation 4.9 as a covariate. Control function model is shown in Equation 4.12.

$$\widehat{mask_j} = \widehat{d}_{j,SARS} + \widehat{d}_{j,H1N1} + \widehat{d}_{j,MERS} + \epsilon_m \tag{4.11}$$

$$g_{j,t+shift} = \theta_0 + \theta_c mask_{j,t} + \sum_{p \in P} \theta_p policy_{j,t,p} + \sum_{m \in M} \theta_m mobility_{j,t,m} + \theta_e testing_{j,t} + \sum_{w \in W} \theta_w week_{j,t,w} + \theta_r trend_{j,t} + \sum_{j \in J} \theta_j country_j + \theta_{md}(mask_{j,t} - \widehat{mask_j}) + \epsilon_t \tag{4.12}$$

Wearing face masks in public is common in many Asian countries, as compared to countries in Europe or America [148, 149]. One of the reasons is their recent experience with airborne diseases. Another reason could be air pollution or the culture of wearing face masks. We do not account for the different trends in wearing face masks among countries due to pollution or culture. However, we believe the country-fixed effects could capture the country-wise trends in wearing face masks.

4.6.2 Data Collection and Processing

We model the effect of wearing face masks, change in social mobility, and government-enforced Non-Pharmaceutical Interventions (NPIs) in infection growth rate. We select the countries with a publicly available dataset for wearing face masks and community mobility. We collect data from February 21, 2020, to July 8, 2020. This section discusses the different datasets used in our analysis to isolate the effect of masks, social mobility, and NPIs in containing the spread of the contagious SARS-CoV-2 (COVID-19) virus.

4.6.2.1 Masks

We collect mask data from surveys conducted by YouGov [150]. YouGov is an international internet-based market research company specializing in opinion polls through online methods. YouGov used online surveys as their COVID-19 behavior change tracker. They conducted surveys periodically in some countries of the world to estimate the propensity of the percentage of people that wear face masks when they go out in public spaces. These surveys were conducted every week. Figure 4.9 shows the raw survey numbers from YouGov. As the surveys were conducted periodically and not every day, we used linear interpolation to estimate the percentage of the population that wear face masks in public spaces (Figure 4.10). Note that the online survey does not include data for type (quality) of masks or how people wear masks (insufficient quality or incorrect method of covering face masks, e.g., touching the surface, not covering nose or mouth – might not be effective in controlling the spread of virus). Thus, our estimation of mask effects in this work would be an estimation of the behavior of wearing masks. In our analysis, we normalize the number for mask-wearing such that $0 \leq mask_{j,t} \leq 1, \forall j, \forall t$.

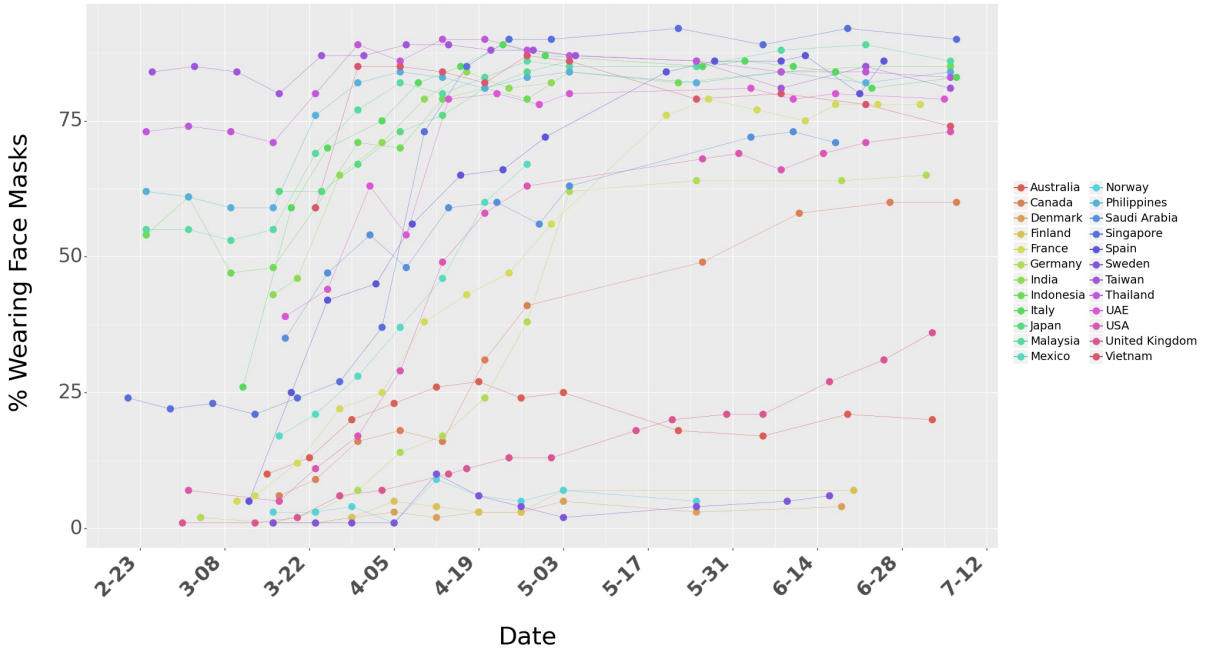


Figure 4.9: Raw data for surveys on the percentage of people who say they wear a face mask when in public spaces. The dots represent the raw numbers from the survey data from YouGov. The lines are shown for better visualization.

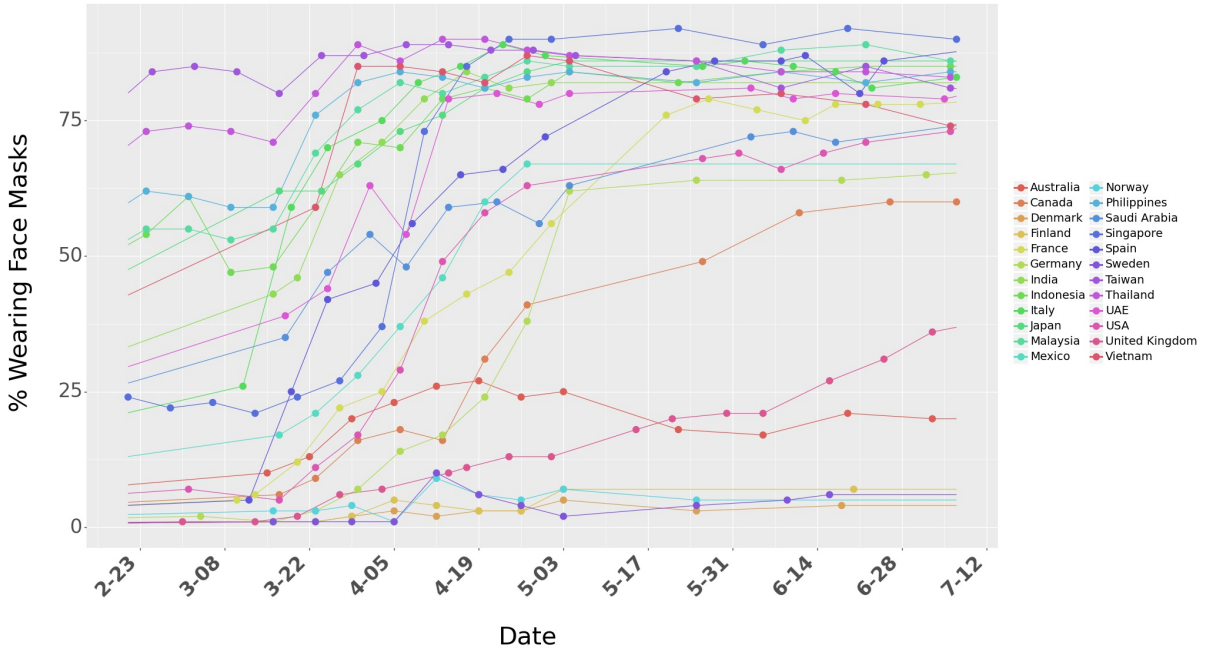


Figure 4.10: Survey data on percentage of people who say they wear a face mask when in public spaces. We use linear interpolation to consider mask numbers for days between surveys days. The dots represent the raw numbers from surveys.

4.6.2.2 Active Cases

We use the timeline for total confirmed cases, and total recovered cases from Johns Hopkins Coronavirus Research Center [151] to find the daily active cases across different countries. We use the daily active cases to calculate the outcome variable of our model – growth rate. Figure 4.11 shows the cumulative confirmed cases, Cumulative recovered cases, and daily active cases for different countries. We use a 7-day moving average for daily cumulative confirmed cases and cumulative recovered cases.

4.6.2.3 Growth Rate

We use daily active cases to estimate the daily growth rate for these countries. Growth Rate (Equation 4.7) can be very volatile at the start of the pandemic due to the low number of cases in the early stages. For example, a unit increase in $I_{j,t}$ will record a growth rate of 0.4 when $I_{j,t} = 2$ as compared to a growth rate of 0.0004 when $I_{j,t} = 1000$ (growth rate is calculated as the first difference in log of active cases in consecutive days). Similarly, during the later stages of the pandemic (at least when the first wave is slowed

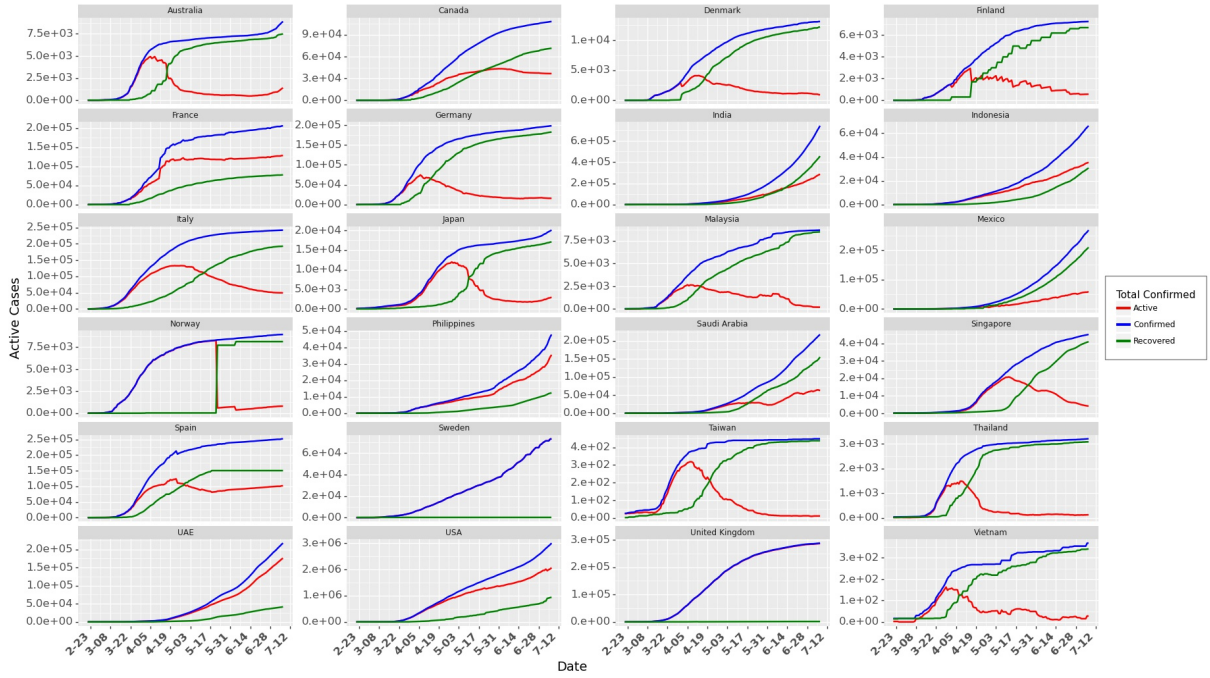


Figure 4.11: Cumulative confirmed cases, cumulative recovered cases, and daily active cases across 24 countries. We observe data reporting issues in Norway, Sweden, and United Kingdom. We use cumulative confirmed cases for calculating the growth rates for these countries.

down for some countries), the growth rate could be affected by multiple other factors such as awareness or changed individual behavior. To avoid these issues, we use the data for the first 60 days for a country (after we start collecting data for a country following the 'th' - will be discussed shortly).

Unlike Hsiang et al. [45], we use data for 60 days and do not restrict to the initial phase when the cases rise exponentially. Section 4.6.4 discusses the model's performance (and changes in model parameter estimates) as we add more/fewer data in the model from 24 countries. To filter out the volatile growth rate during the start of the pandemic, we consider data for each country when the daily new cases cross a threshold th. We define this threshold as the day when the seven-day average of daily new cases in a country crosses $th = 20\%$ of the peak case observed in that country (till July 8, 2020). We select th based on the maximum likelihood estimate of the growth rate model (mechanism for selecting th is discussed later in Section 4.6.4. Figure 4.12 shows the daily new cases, and

Figure 4.13 shows the respective growth rate. We use data for a maximum of 60 days for a country, from the day its daily cases cross the threshold. Thus, our dataset contains unbalanced panel data from 24 countries.

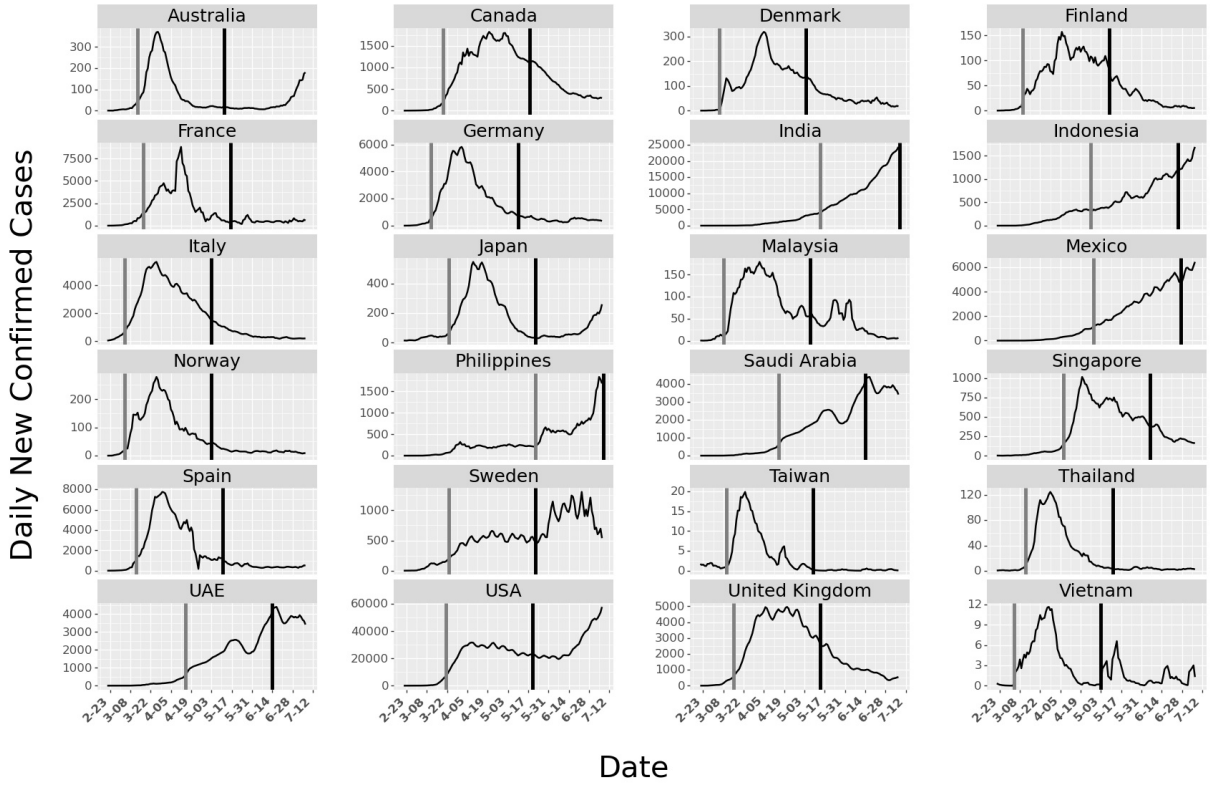


Figure 4.12: Daily new cases across countries. The light gray vertical line marks the day when daily new cases in that country crossed the threshold. The black vertical line shows the end of 60 days of data collected for each country. For countries where cases are still increasing vis-à-vis India and the Philippines, we collected fewer data points than 60 days.

4.6.2.4 Community Mobility

Google's COVID-19 Community Mobility Reports [42] provides information on how movement trends change over time across different types of locations in different countries. The mobility numbers are calculated based on the change in trend from the baseline (details in the report on how Google calculates the baseline). The report tracks movement trends over time by geography across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. Figure 4.14 shows the community mobility across different countries.

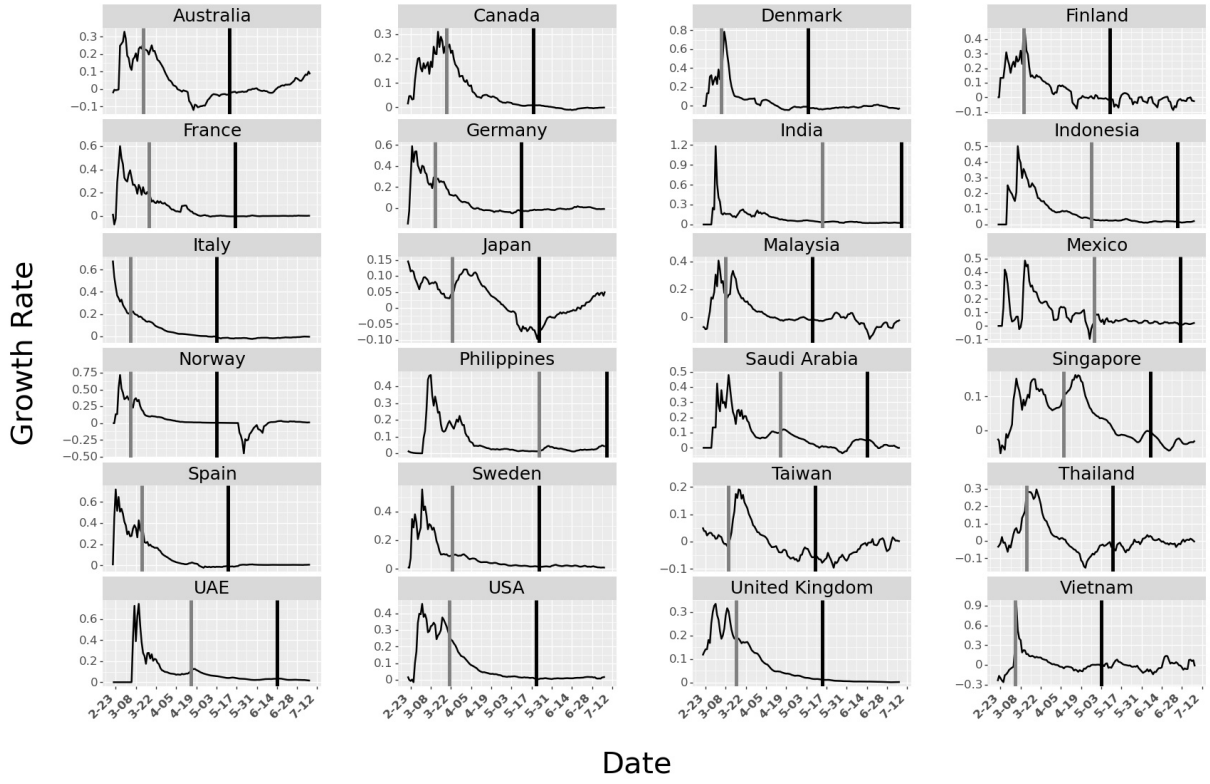


Figure 4.13: Growth rate across countries. The light gray vertical line marks the day when daily new cases in that country crossed the threshold. The black vertical line shows the end of 60 days of data collected for each country. We use cumulative confirmed cases to calculate the growth rate for Norway, Sweden, and United Kingdom. Collecting data after the light gray line allows us to filter the initial noisy growth rate from that country.

We observed a high correlation between the social mobility numbers from Google across different types of locations. This may lead to unstable parameter estimates due to multicollinearity in parameter estimation using ordinary least squares. Based on the correlations, we consider the mobility in Parks and Transit stations as our measure for mobility. We also confirm these two categories using a Lasso regression (more details in Section 4.6.4 in Section 4.5.3). The Lasso regression model pushes the coefficients of correlated variables (variables that do not add much information to the model) to 0 and gives non-zero weights to only two mobilities: Parks and Transit stations.

Table 4.1 shows the correlation matrix between the mobility across different locations. The correlation matrix shows that transit stations is highly correlated with mobility in Retail and Recreation, Grocery and Pharmacy, and Residential. Mobility in transit sta-

	Retail and Recreation	Grocery and Pharmacy	Parks	Transit Stations	Workplace	Residential	Driving	Walking
Retail and Recreation	1.00	0.85	0.53	0.90	0.75	-0.85	0.81	0.82
Grocery and Pharmacy	0.85	1.00	0.45	0.78	0.67	-0.74	0.69	0.67
Parks	0.53	0.45	1.00	0.42	0.16	-0.51	0.73	0.67
Transit Stations	0.90	0.78	0.42	1.00	0.84	-0.89	0.73	0.78
Workplace	0.75	0.67	0.16	0.84	1.00	-0.87	0.53	0.58
Residential	-0.85	-0.74	-0.51	-0.89	-0.87	1.00	-0.74	-0.75
Driving	0.81	0.69	0.73	0.73	0.53	-0.74	1.00	0.91
Walking	0.82	0.67	0.67	0.78	0.58	-0.75	0.91	1.00

Table 4.1: Correlation matrix for community mobility in different locations from Google and Apple.

Mobility	Min	Mean	Max	Std Dev	25^{th} percentile	75^{th} percentile
Retail and Recreation	-96 %	-32.3 %	23 %	27.1 %	-53 %	-10 %
Grocery and Pharmacy	-94 %	-10.1 %	51 %	18.7 %	-19 %	2 %
Parks	-91 %	4.7 %	517 %	65.5 %	-38 %	24 %
Transit Stations	-92 %	-38.8 %	14 %	23.9 %	-57 %	-20 %
Workplaces	-90 %	-27.7 %	57 %	23.8 %	-45 %	-7 %
Residential	-13 %	12.8 %	55 %	10.4 %	4 %	19 %

Table 4.2: Summary statistics on mobility trends from Google.

tions is negatively correlated with mobility in Residential - as fewer people travel, more people are staying home. Thus, mobility in transit stations can capture the information from mobility across all other locations except Parks. Henceforth, we include mobility in Parks and Transit stations as a measure of mobility (as also selected by the Lasso Regression model). In our analysis, we normalize the number for social mobility such that $0 \leq mobility_{j,t,m} \leq 1, \forall j, \forall t, \forall m$.

Table 4.2 shows the summary statistics for the community mobility. Google community mobility reports use data from users with the android operating system. Also, they collect data from users who allow location sharing (Google does not disclose any personal information in Community Reports). To check the robustness of the model and our estimates on the effect of mobility, we also use mobility data from Apple Mobility

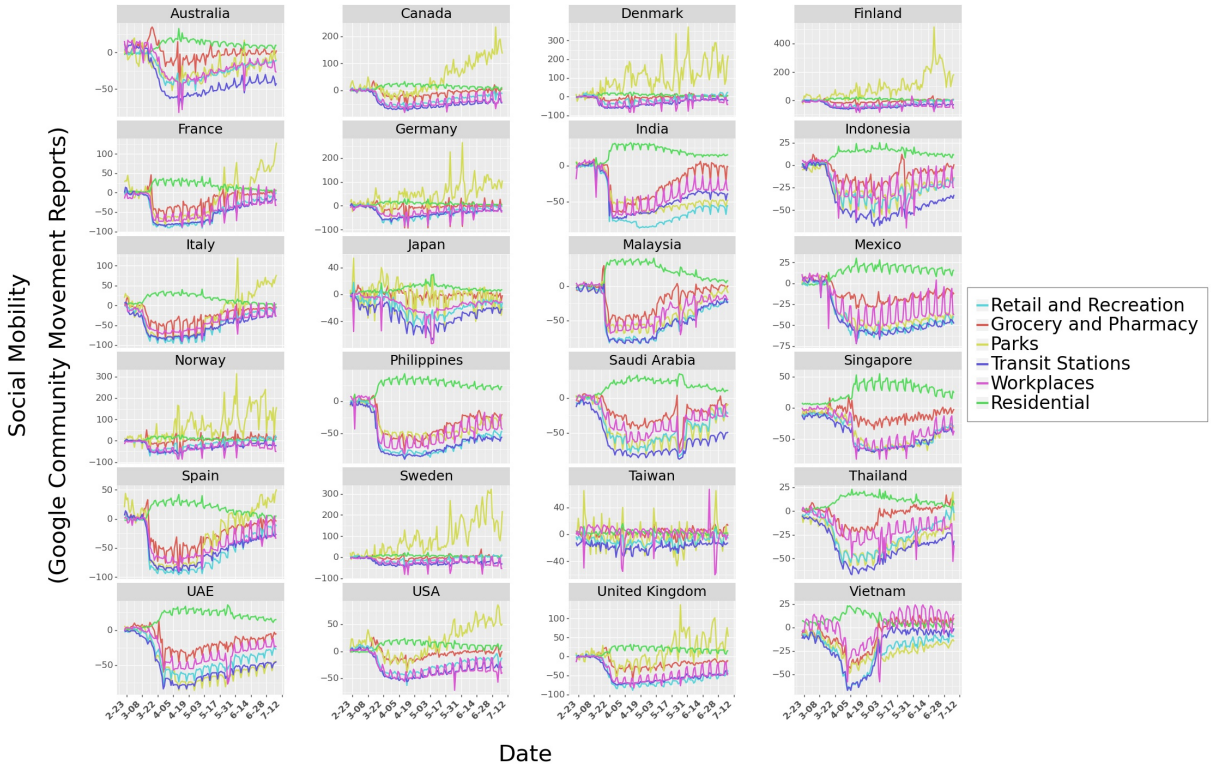


Figure 4.14: Community mobility trend from Google using Android operating system in different countries.

Mobility	Min	Mean	Max	Std Dev	25^{th} percentile	75^{th} percentile
Driving	-100 %	-28.2 %	184 .9%	48.5 %	-64.9 %	-6.2%
Walking	-100 %	-35.7 %	94.4 %	44.1 %	-72 %	0.58 %

Table 4.3: Summary statistics on mobility trends from Apple

Trend reports [122] to validate the results from Google Community Mobility reports. The Apple Mobility reports show a relative volume of direction requests per country/region, sub-region, or city compared to a baseline volume on January 13, 2020. Apple compares the relative volume for Driving, Transit, and Walking in their dataset. However, we could not find the data on Transit for all the 24 countries considered in this chapter. So we use the timeline for Driving and Walking as a proxy for measuring social mobility. Table 4.3 provides the summary statistics on Apple Mobility Trends. The mobility trends for Driving and Walking across 24 countries are shown in Figure 4.15.

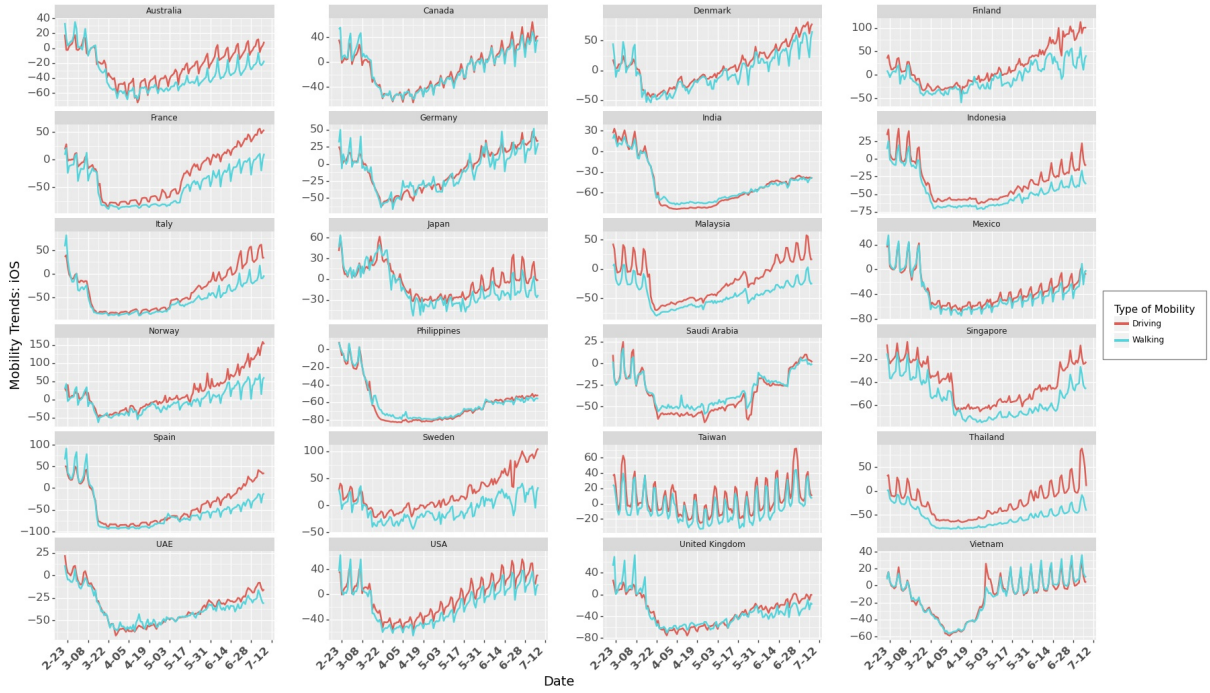


Figure 4.15: Community Mobility Trend from Apple using iOS.

4.6.2.5 Non-Pharmaceutical Interventions (NPIs)

Governments (and their policies) play a critical role in fighting a pandemic. Vaccines may take a long time to be available, particularly for a new disease, e.g., COVID-19. In an ongoing pandemic, we cannot depend only on vaccines but need vital government interventions (institutional measures) to control the spread of the disease. During such times, governments must take various measures, e.g., increasing testing infrastructure to control the spread of infections. These NPIs help decrease mobility (for example, travel bans imposed restrictions on travel across states/regions/countries while it also helped restrict mass gatherings in places such as transit stations. Since COVID-19 spreads through person-to-person physical interaction (or prolonged proximity), governments introduced various policies (Non-Pharmaceutical Interventions or NPIs) for social distancing to minimize person-to-person interaction. They also introduced closures of places where people gather together simultaneously, e.g., schools or businesses. However, these government policies seriously affected businesses [152], leading to economic shutdowns which adversely affected the poor community [153]. The effect of these shutdowns may also lead to pro-

longed economic hardships, e.g., closure of some businesses and employment [154].

Since these policies directly affect the livelihood of a majority of the population across the world, it is crucial to investigate their impact on controlling the spread of the disease. As these policies were implemented at different times across different countries, it allows us to explore the combined effect of these policies. Estimating the combined effect of these policies could help the governments in future (or current) pandemics to introduce effective policies and may not necessarily lead to complete lockdown unless extremely necessary. Please note that some of these policies were introduced simultaneously, or some were implemented first, and some were implemented consistently after some other policies. We do not claim any causal effect of the policy on the growth rate. It is difficult to isolate the effect of individual policies as the implementation of policies was not randomly sequenced across countries.

We use Coronanet dataset from Cheng at al [43] for NPIs. They collected information on all the government policies introduced by different countries worldwide. They categorized the policies into 19 different policy types. We use their categorization to build our model. The policies were implemented at different levels – National, Provincial and Municipal. This work considers the policies implemented at the National and Provincial levels. From February 21, 2020, to July 8, 2020, we check if a policy p was implemented in a country j or not on the day t. If the policy was implemented, we assign a value of 1 to $S_{j,t,p}$. If the policy was introduced at a provincial level (could be introduced by the central government or a respective state government), we increase $S_{j,t,p}$ using the population of the state. Equation 4.13 explains $S_{j,t,p}$ if a policy p is employed at a provincial level where $N_{j,s}$ is the population of state s in country j. After identifying $S_{j,t,p}$ for all countries over the period of our analysis (considering all the entries in the dataset), we use normalization using maximum value in a country such that $0 \leq S_{j,t,p} \leq 1, \forall j, \forall t, \forall p$, as shown in Equation 4.14.

$$S_{j,t,p} = S_{j,t,p} + \frac{N_{j,s}}{N_{j,s} + N_j} \tag{4.13}$$

$$S_{j,t,p} = \frac{S_{j,t,p}}{\max_{t,p} S_{j,t,p}} \tag{4.14}$$

The dataset contains 5,816 entries on policies (some of the policies were announcements/recommendations/new entry or an update to existing policy) at National and Provincial level. The dataset provides detailed information on the type of the data entry (e.g. policy type, description of the policy). The statistics on the types of policies is shown in Figure 4.16. Figure 4.16 also provides a count of entries of each policy type. The data set contains 20 policy types. Figure 4.16 shows how many countries implemented (light grey bars) a particular policy type. It also shows how many countries implemented a particular policy type at national or provincial level (light blue bars and dark blue bars respectively). The dark grey bars show the total number of entries (divided by 100 for visualization) for all policy types.

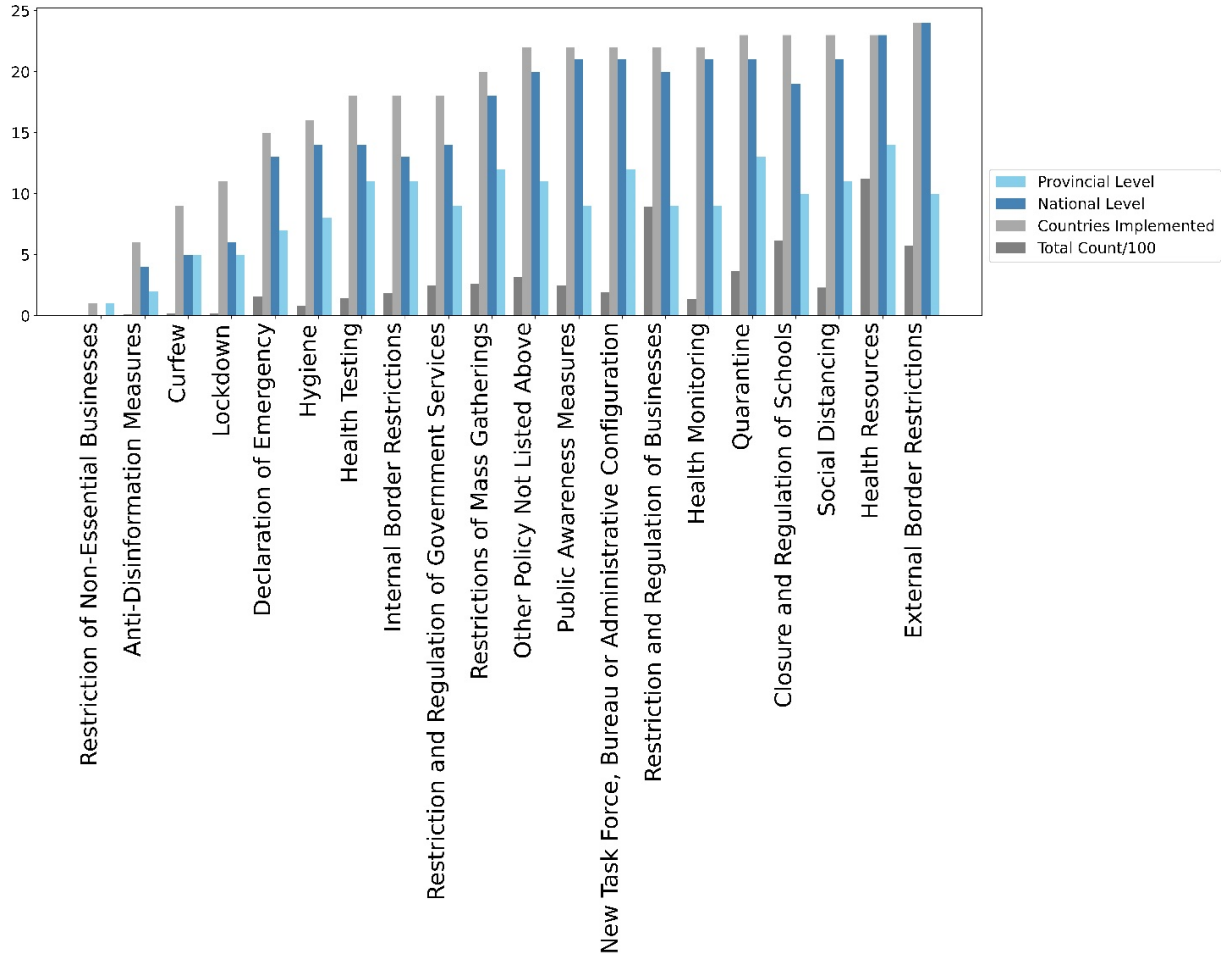


Figure 4.16: Policy implementation at national and provincial levels across different countries.

We use the text description of the policy to identify if an entry was an update, recommendation, or actual implementation. If the entry was an announcement or an update for a policy with a start date and an end date, we give a weight of 0 to that entry in the dataset because if there is a policy update, it could recount that policy. The policies could have been implemented differently across different countries, even if they were categorized in the same policy type. For example, a country may impose Social distancing rules from 4 pm – 8 pm while another may impose them from 6 am – 6 pm. It may differ across different states in the country. However, we do not consider the variations in the implementations of policies in this research.

As we have survey numbers for wearing face masks at a national level, we consider policy types implemented across most countries. Therefore, we do not consider Anti-disinformation measures, Curfew and Lockdowns. Curfew and Lockdowns are similar to Quarantine and Restrictions of Mass Gatherings (which lead to the closure of places of mass gatherings), so we can ignore them for the purpose of this research. Moreover, Curfew and Lockdowns affect community mobility, which can be accounted for by social mobility trends from Google Community Mobility Reports (discussed in the previous section).

We also do not consider Hygiene Announcements and New Task Force policy as these were administrative announcements and did not have much effect on the growth rate of the infection. Some policies did not have a start date and end date. We calculate the cumulative number of times announcements were made for such categories. Health Testing, Health Monitoring, and Health Resources are administrative announcements, so we combine them into one Health Resources policy. Using linear models, a linear combination (addition of three policies) does not affect our analysis. It further reduces the number of parameters to estimate. Similarly, we combined Restrictions and Regulations of Businesses and Restriction and Regulation of Government Services. Health resources can also be used as a proxy for increased awareness among governments and citizens. So, we do not consider "Public Awareness Measures" announcements to avoid multicollinearity in the set of predictor variables.

	Health Resources	Restriction and Regulation of Businesses	Closure and Regulation of Schools	External Border Restrictions	Quarantine	Restrictions of Mass Gatherings	Social Distancing	Internal Border Restrictions	Mobility Parks	Mobility Transit Stations
Health Resources	1.00	0.32	0.04	0.56	0.34	0.26	0.35	0.49	-0.26	-0.26
Restriction and Regulation of Businesses	0.32	1.00	0.55	0.46	0.35	0.58	0.44	0.43	-0.01	-0.44
Closure and Regulation of Schools	0.04	0.55	1.00	0.16	0.25	0.52	0.15	0.40	0.02	-0.40
External Border Restrictions	0.56	0.46	0.16	1.00	0.28	0.38	0.47	0.26	-0.19	-0.54
Quarantine	0.34	0.35	0.25	0.28	1.00	0.34	0.32	0.26	0.14	0.02
Restrictions of Mass Gatherings	0.26	0.58	0.52	0.38	0.34	1.00	0.35	0.37	-0.01	-0.38
Social Distancing	0.35	0.44	0.15	0.47	0.32	0.35	1.00	0.24	-0.14	-0.41
Internal Border Restrictions	0.49	0.43	0.40	0.26	0.26	0.37	0.24	1.00	-0.21	-0.26
Mobility Parks	-0.26	-0.01	0.02	-0.19	0.14	-0.01	-0.14	-0.21	1.00	0.51
Mobility Transit Stations	-0.26	-0.44	-0.40	-0.54	0.02	-0.38	-0.41	-0.26	0.51	1.00

Table 4.4: Correlation between NPIs

Table 4.4 shows the correlation between different government policies implemented across countries. The correlation value between pairs of any two NPIs is not high (> 0.7 as observed with community mobility across different locations in Google Community Mobility Reports), so we include all the following eight government policies in our model in Equation 4.8. We also include the Social Mobility in Parks and Transit Stations to check its correlation with NPIs. Changes (reduction during the early stages of the pandemic) in social mobility were induced by the introduction of NPIs. However, social mobility is a combination of institutional measures, e.g., NPIs, and individual measures, e.g., social mobility. The correlation between social mobility and any NPIs as shown in Table 4.4 is not high (> 0.7), so we do not reject any further NPIs.

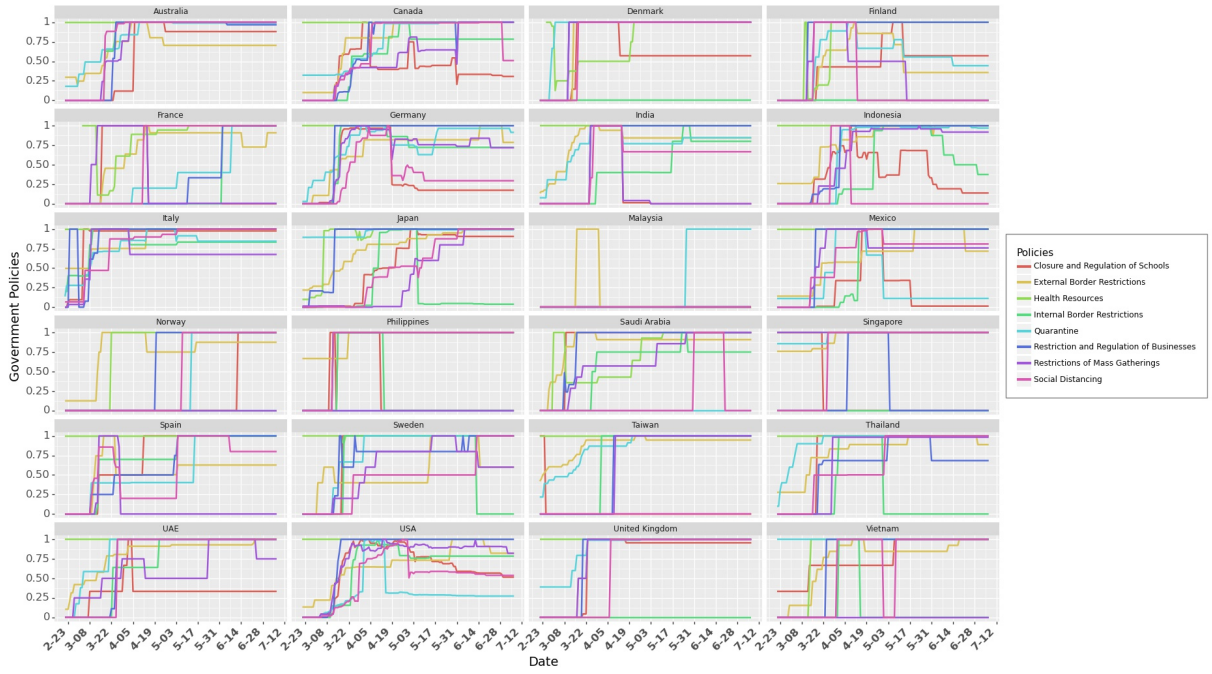


Figure 4.17: Policy implementation across different countries. This graph shows how some countries e.g., Taiwan and Malaysia employed only a few of the policies. It also shows that countries e.g., USA and Mexico employed region specific policies (at provincial level).

Figure 4.17 shows different policy implementation of each country for policies considered in this work (as described above). Figure 4.18 illustrates different countries that implemented each policy. Note that we normalize $S_{j,t,p}$ such that $0 \leq S_{j,t,p} \leq 1, \forall j, \forall t, \forall p$.

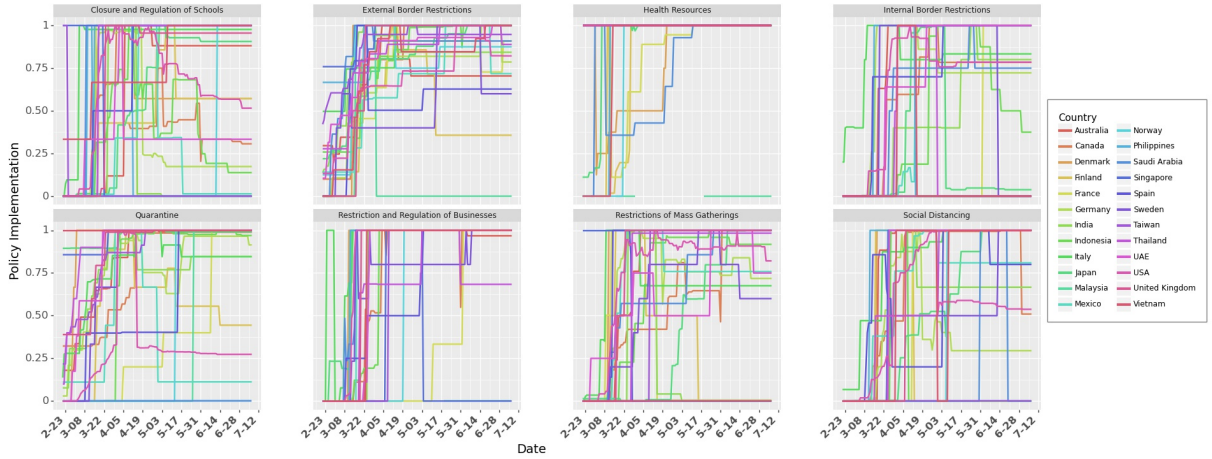


Figure 4.18: Country wise implementation of different policies. Figure shows how majority of the government policies were introduced at similar times across different countries. This makes it difficult to estimate the causal impact of each NPI on the spread of COVID-19.

4.6.2.6 Lag in Observation of the Effects of Control Variables

Studies have reported a delay in observing the effects of policies on the events of a given day. This delay could be due to several reasons. One of the most prominent reasons is the incubation period (time between getting infected and onset of symptoms/knowing that individual is confirmed for COVID-19). During the incubation period, an individual may be asymptomatic. Incubation period is estimated to be 4 days to 14 days [130]. Another reason could be the testing time – the time it takes to get the confirmation of results. Due to limited healthcare professionals, there could be a long queue to get tested to get the results from testing centers.

To model this delay, we use a lag variable $shift$. We use the cross-validation method to find the lag with the best fit for the data. We test $shift \in [0, 14]$ and observe that the model performs best at a shift of 9 days. We discuss this further in Section 4.6.4.

4.6.2.7 Testing

Testing is critical in identifying infectious individuals. Once identified, these individuals can be quarantined or isolated from the public so that they do not spread to susceptible individuals. While people can get tested when they start showing symptoms, evidence reports that even asymptomatic individuals can spread the virus (50% of cases can be attributed to asymptomatic cases [155]). Since they do not show any symptoms, people

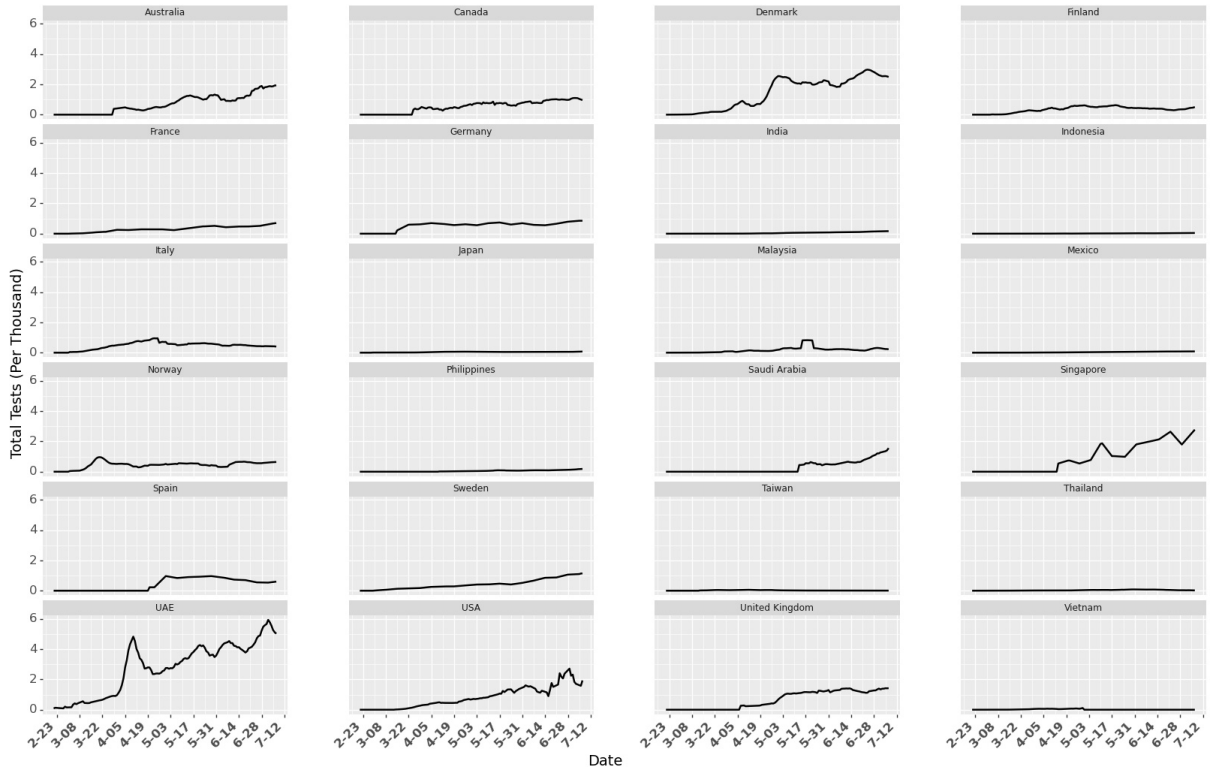


Figure 4.19: Tests per thousand people in different countries over time. The numbers indicate the total number of tests conducted per thousand people in a country, but the data set does not provide information on how many individuals were tested for COVID-19. Some individuals might get tested multiple times.

around them (e.g., an asymptomatic young adult living with family) are less cautious and may get infected through them. It is critical to identify asymptomatic individuals as they can spread the virus unknowingly. This can be done by increased testing and contact tracing the individuals who have come in contact with positively tested for the COVID-19. Testing can be crucial in identifying COVID-19 positive individuals so that they can be quarantined (hospital or home isolation) or treated early when symptoms start showing.

As testing increases, the probability that more confirmed positive cases would be identified increases too. This will increase the empirical growth rate over time as more confirmed cases will be reported. This effect shows that a change in testing patterns over time could lead to bias (or underestimating the effect of NPIs and masks). We use testing data to account for increased testing over time to counter this time-sensitive

bias. Figure 4.19 shows the data on total tests (per thousand people) in a country from ourworldindata.org [156]. In our analysis, we normalize the number for testing such that $0 \leq testing_{j,t} \leq 1, \forall j, \forall t$.

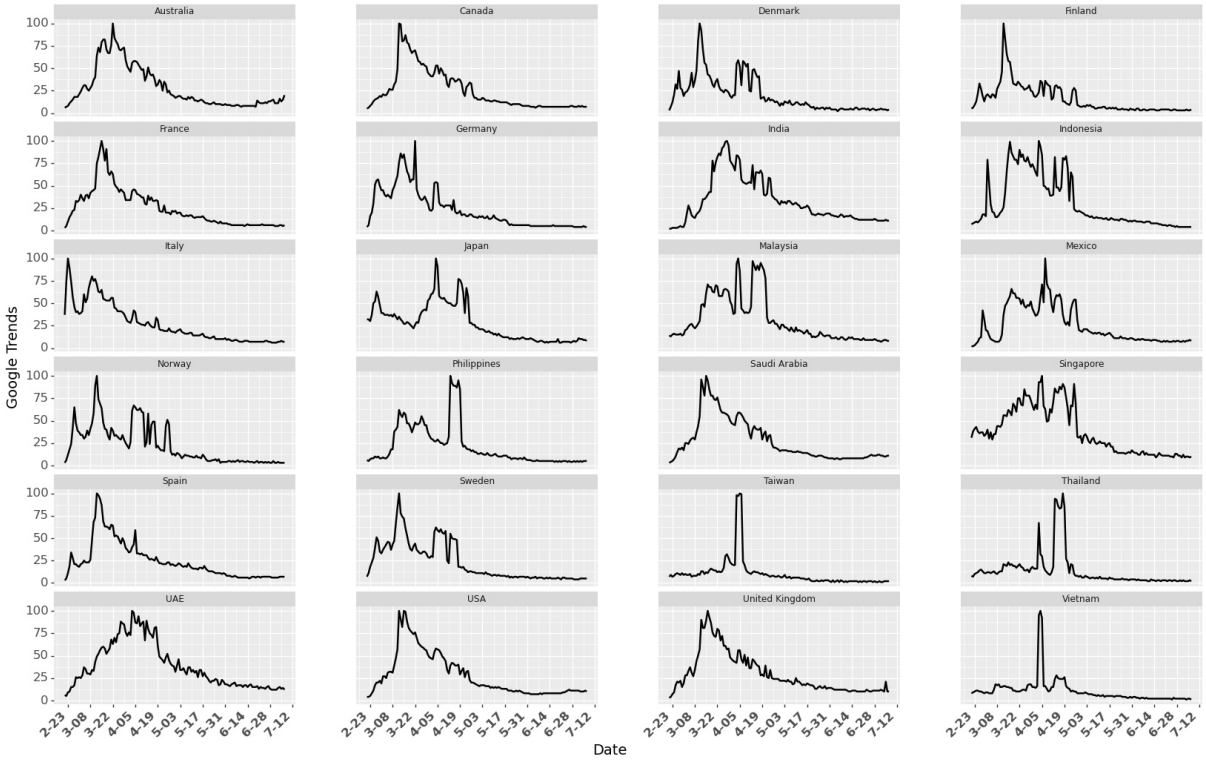


Figure 4.20: Google Trends for the search term 'coronavirus' in different countries.

4.6.2.8 Google Trends

As the number of cases increases, awareness increases in public (e.g., washing hands more often). We use Google Trends [157] to account for the increase in active awareness over time (Figure 4.20). Google Trends numbers indicate the search interest of a topic over time as a proportion of all other searches at the same time. In our analysis, we normalize the number for Google Trends such that $0 \leq trend_{j,t} \leq 1, \forall j, \forall t$.

4.6.2.9 Week Fixed Effects

COVID-19 containment changes over time. It includes increasing public awareness or better understanding of the virus as more studies and research come to public attention. Not only do citizens understand how to be more careful (or more informed), healthcare

providers also learn more about the disease for more efficient treatment of COVID-19 patients (e.g., creating new wards for COVID-19 patients, treating them by wearing Personal Protection Kits, PPE). It also involves improved infrastructure, e.g., testing or converting existing medical facilities to dedicated COVID-19 centers. To account for all the time-sensitive fixed effects (other than the controls we discussed before), we use fixed effects for weeks (from the day that country reaches th in our analysis).

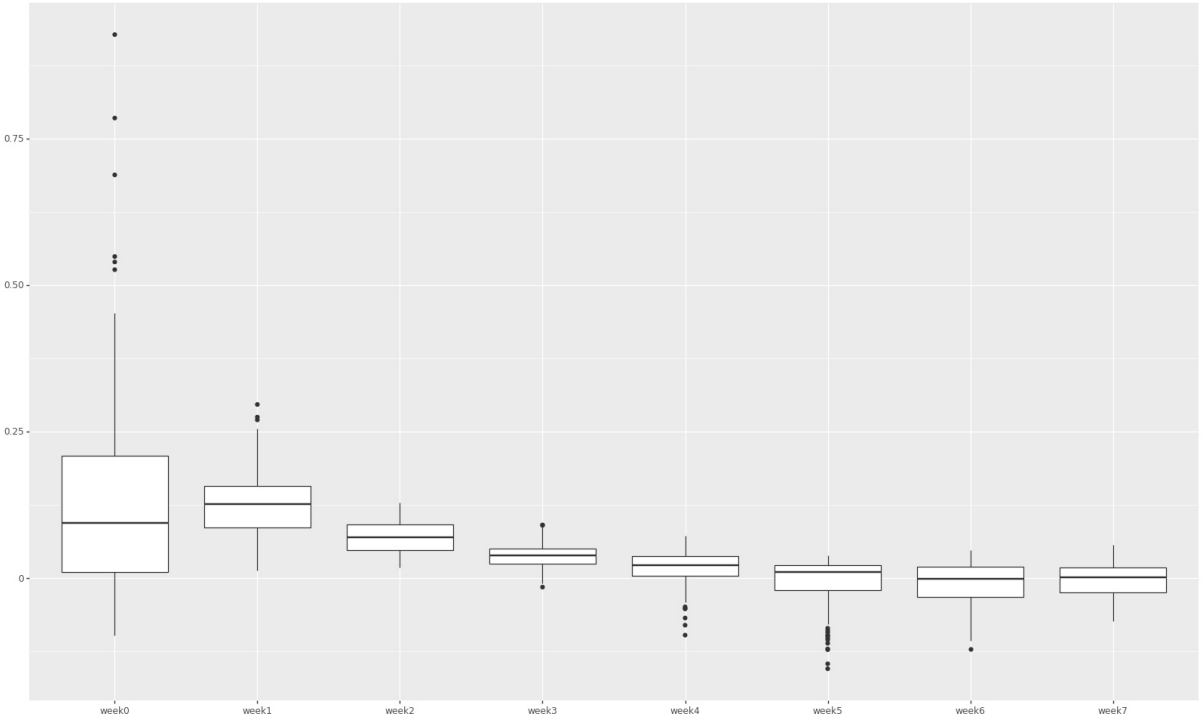


Figure 4.21: Box plot for growth rate of different countries across different weeks.

Figure 4.21 shows how the growth rate changes across different weeks (Figure 4.22 provides the same information across different countries). Even after removing the initial noisy data, we observe the highest variance in the growth rates during week 1 in most countries (Figure 4.21 and Figure 4.22). We use one-hot vector to denote week ($week_{j,t,w}$ = 1 if day t lies in week w for country j). Note that due to different starting times for each country (Figure 4.12), a day may come under a different week for a different country. For example, days in week 1 for Vietnam are earlier in the calendar than the days in week 1 in India. Note that the growth rate is higher during the initial weeks and slows down with time. To capture this effect, we use fixed effects for weeks. Results in Section 4.6.3

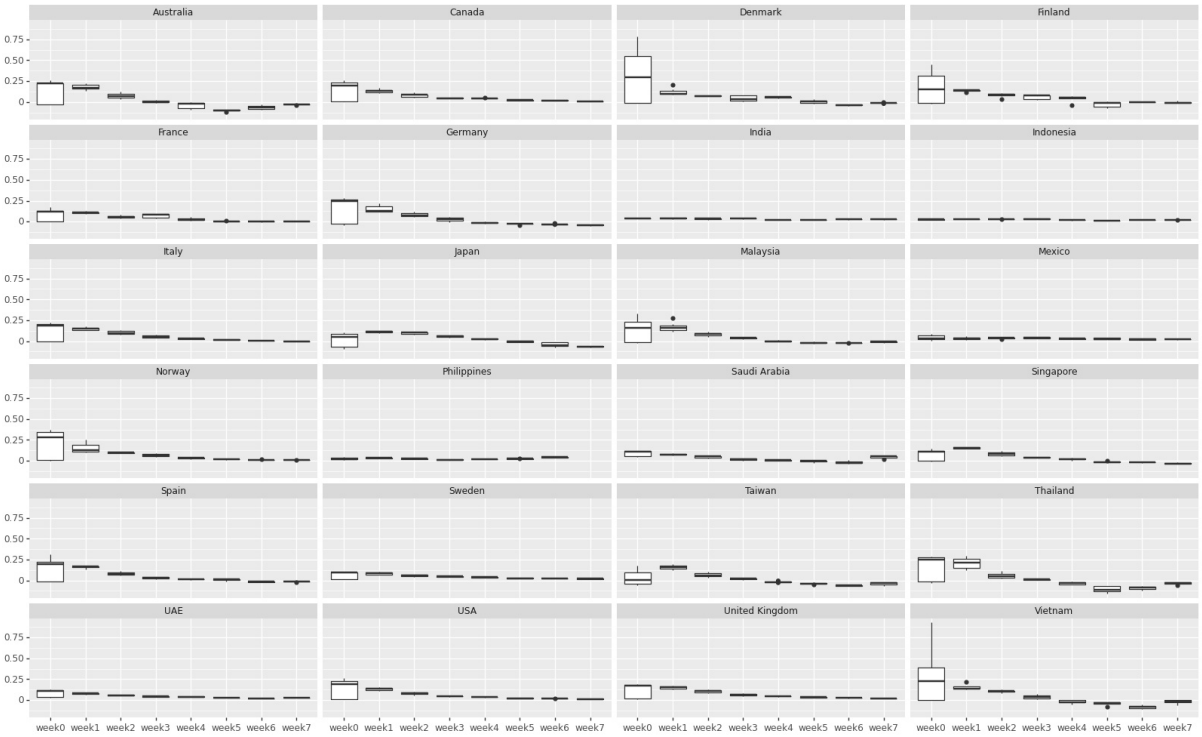


Figure 4.22: Box plot for growth rate in different weeks across different countries.

show that the magnitude of coefficient for week 0 is higher than the magnitude of the coefficient for week one and so on.

4.6.2.10 Country Fixed Effects

We observe heterogeneity across countries with many aspects in handling COVID-19. Multiple factors affect the spread and handling of a disease in a country. Heterogeneity may be observed at different levels. For example, heterogeneity at the government level includes the difference in reporting cases, testing infrastructure, strictness is reducing social mobility, and implementation of NPIs. Population-wise heterogeneity includes population density in a country or percentage of the population living in high-density urban regions or poor neighborhoods with shared sanitation facilities. It may also include cities with international airports or travelers (particularly from countries hard hit with COVID-19 in early 2020, e.g., China and Iran). It also includes heterogeneity at the level of education (awareness about COVID-19, responsibility in understanding the severity of precautions), poverty (health insurance, ability to purchase sanitizers or high-quality masks), primary

R-squared	0.738
Adj. R-squared	0.729
F-statistic	88.03
Prob (F-statistic)	0
Log-Likelihood	2400.8
AIC	-4712
BIC	-4475
No. Observations	1422
Degree of Freedom, Residuals	1377
Degree of Freedom, Model	44

Table 4.5: Model statistics

health care facilities (drinking water, sanitation, shared places) or family structure (number of young adults in a family or size of the family residing in a residential complex). To control for all this heterogeneity among countries which may lead to country-level effects in the growth rate of COVD-19, we use country fixed effects.

Note that since we use a constant in our model, we consider fixed effects for 23 countries, and we consider the last country (Vietnam) as our base country (with 0 fixed country effect). This ensures that parameter estimates are stable.

4.6.3 Results

We use the growth rate model to study the effect of Masks, Social Mobility, and Non-Pharmaceutical Interventions (NPIs). Since we do not have real numbers on how many people wear masks (or wear masks that could be effective), we use different transformations of the mask numbers from the surveys. We use growth rate model with masks transformed as $ln(1 + mask_{j,t})$ as our focal model. We use threshold $th = 0.2$ and $shift = 9$ days. Details on selection of th and $shift$ is provided in Section 4.6.4. Model statistics are given in Table 4.5. Table 4.6 shows the parameter estimates for the model with different transformations.

Figure 4.23 shows the complete results for the parameter estimation under different transformations for masks. Results for parameter estimation in Table 4.6 show consistency in the estimates for social mobility and NPIs. They also show consistent parameter estimates for fixed effects (week and countries). Fixed effect of week can capture the trend

variable	coefficient	std error	t-value	p-value	lower limit	upper limit
constant	0.1999	0.017	11.745	0	0.167	0.233
log(Mask)	-0.1047	0.024	-4.385	0	-0.152	-0.058
Mobility Parks	-0.0296	0.006	-4.677	0	-0.042	-0.017
Mobility Transit Stations	0.1109	0.013	8.398	0	0.085	0.137
week0	0.0981	0.009	11.025	0	0.081	0.116
week1	0.0589	0.009	6.872	0	0.042	0.076
week2	0.0411	0.008	5.146	0	0.025	0.057
week3	0.0324	0.007	4.335	0	0.018	0.047
week4	0.018	0.007	2.575	0.01	0.004	0.032
week5	0.0039	0.007	0.592	0.554	-0.009	0.017
week6	-0.0013	0.006	-0.209	0.834	-0.014	0.011
week7	0.0021	0.006	0.343	0.732	-0.01	0.014
Testing	-0.0121	0.006	-1.938	0.053	-0.024	0
Trend	-0.0455	0.008	-5.933	0	-0.061	-0.03
Health Resources	-0.034	0.012	-2.881	0.004	-0.057	-0.011
Restriction of Businesses	-0.0049	0.005	-1.03	0.303	-0.014	0.004
Closure of Schools	-0.0153	0.006	-2.436	0.015	-0.028	-0.003
External Border Restrictions	-0.0315	0.008	-3.807	0	-0.048	-0.015
Quarantine	-0.0321	0.01	-3.194	0.001	-0.052	-0.012
Restrictions of Mass Gatherings	-0.0066	0.007	-1.007	0.314	-0.019	0.006
Social Distancing	0.0038	0.006	0.618	0.536	-0.008	0.016
Internal Border Restrictions	-0.01	0.006	-1.639	0.101	-0.022	0.002
Australia	-0.0338	0.013	-2.531	0.011	-0.06	-0.008
Canada	0.0205	0.012	1.645	0.1	-0.004	0.045
Denmark	0.0167	0.014	1.23	0.219	-0.01	0.043
Finland	-0.02	0.016	-1.27	0.204	-0.051	0.011
France	-0.032	0.014	-2.209	0.027	-0.06	-0.004
Germany	-0.0101	0.014	-0.739	0.46	-0.037	0.017
India	-0.0038	0.013	-0.295	0.768	-0.029	0.022
Indonesia	0.0353	0.013	2.71	0.007	0.01	0.061
Italy	0.0481	0.011	4.308	0	0.026	0.07
Japan	-0.0034	0.01	-0.348	0.728	-0.022	0.016
Malaysia	-0.0397	0.013	-2.944	0.003	-0.066	-0.013
Mexico	0.0002	0.017	0.014	0.989	-0.032	0.033
Norway	-0.0373	0.016	-2.284	0.023	-0.069	-0.005
Philippines	-0.0315	0.017	-1.88	0.06	-0.064	0.001
UAE	0.1119	0.025	4.418	0	0.062	0.162
Saudi Arabia	0.015	0.016	0.949	0.343	-0.016	0.046
Singapore	0.0461	0.015	3.124	0.002	0.017	0.075
Spain	0.0073	0.014	0.543	0.588	-0.019	0.034
Sweden	-0.0248	0.014	-1.776	0.076	-0.052	0.003
Taiwan	-0.0294	0.013	-2.322	0.02	-0.054	-0.005
Thailand	0.0209	0.012	1.81	0.071	-0.002	0.044
United Kingdom	0.0215	0.014	1.547	0.122	-0.006	0.049
USA	0.0252	0.013	1.872	0.061	-0.001	0.052

Table 4.6: Parameter estimates for growth rate model

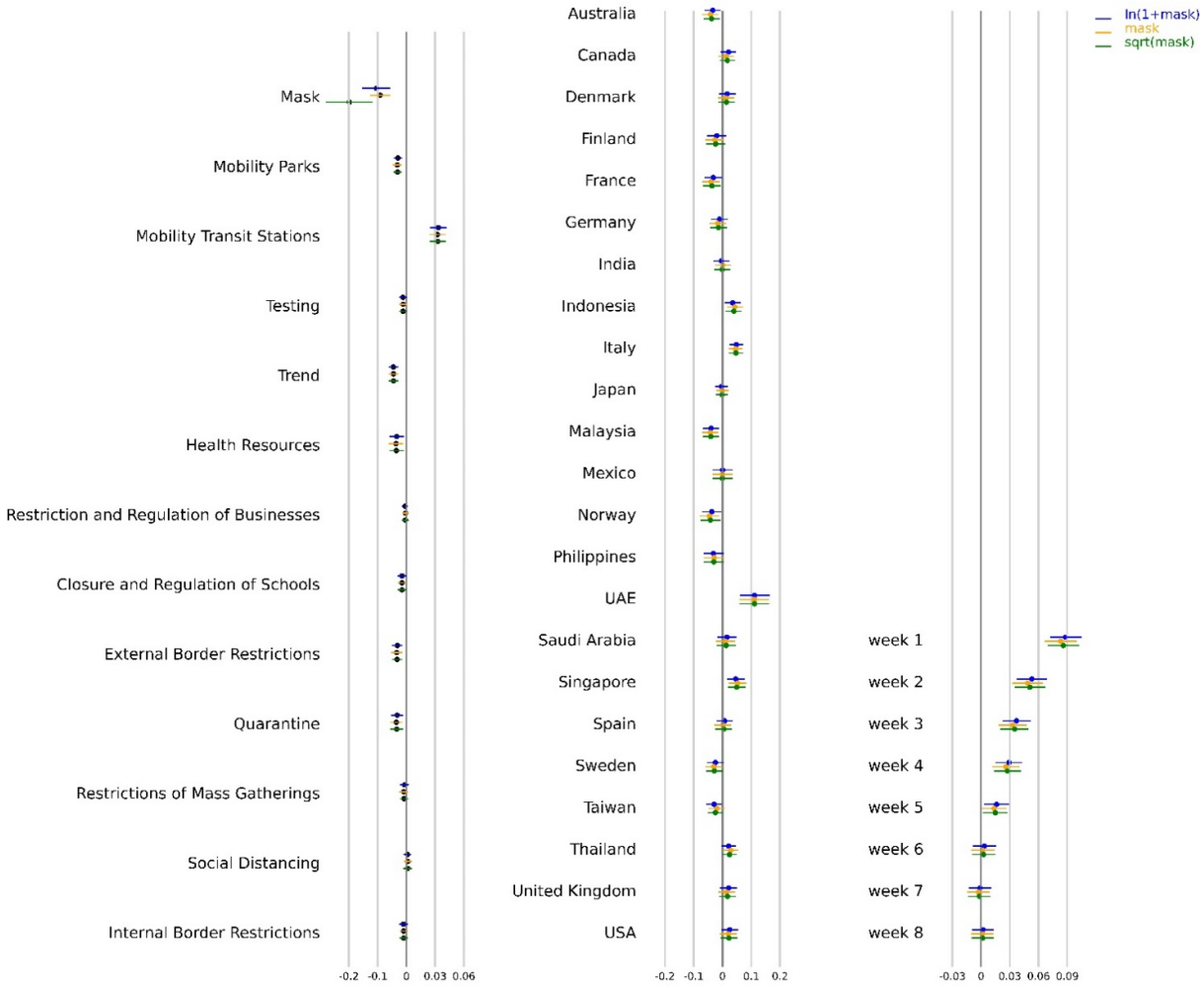


Figure 4.23: Parameter estimates for models with different transformations for masks. We use $th = 0.2$ and $shift = 9$ days.

in awareness or infrastructure change over time. At the beginning of the pandemic in a country, the growth rates were higher. This higher growth rate corroborates positive and statistically significant values for the fixed effects of weeks (decreases as weeks increase). The coefficients for masks seem different across the different transformations, but due to their transformation, it has to be interpreted differently (as we discuss next in the interpretation of results). However, we cannot claim causality from these results as the NPIs were not randomly introduced in different countries. Nonetheless, we can estimate the combined effect of different mobilities and NPIs.

4.6.3.1 Krinsky-Robb method

In the rest of our analysis, we use the Krinsky-Robb method to estimate confidence intervals for the combined effect of masks, social mobility, and NPIs. We also use this method to obtain confidence bounds across the growth rate predictions and active cases in a country. Krinsky-Robb method is a Monte Carlo simulation method used to draw samples from a multivariate normal distribution. We use the ordinary least square method to estimate the coefficients θ in Equation 4.8. The ordinary least square method for multiple linear regression assumes a multivariate normal distribution of θ. Krinsky-Robb method takes advantage of this assumption to sample random draws for θ using Cholesky decomposition and standard normal variates. Steps in the Krinsky-Robb method are:

1. Find Cholesky decomposition matrix C for the covariance matrix of Σ_θ.

2. Draw $|\theta| \times n$ random samples, $x_{samples}$, from standard normal distribution ($|\theta|$ is the cardinality of θ).

3. $\theta_{samples} = \hat{\theta} + C \times x_{samples}$ ($\hat{\theta}$ is the estimated coefficient or variable).

4. Calculate confidence interval based on $\theta_{samples}$.

We use this method to get confidence interval bounds for the sum of the coefficients of mobility and NPIs to get the combined effect. We also use this method to predict confidence intervals of growth rate and daily active cases under different scenarios, as we discuss next. First, we discuss the model performance and then discuss the interpretation of the coefficients in Table 4.6.

4.6.3.2 Model Performance

We can use the coefficients from our model to predict the growth rate for different countries. Figure 4.24 shows the actual growth rate (green dots) with the predicted growth rate (blue line) and its confidence interval (blue shade). We use the Krinsky-Robb method to estimate the confidence interval bounds around the prediction. Results show that the model can accurately predict the growth rate of daily infections across different countries.

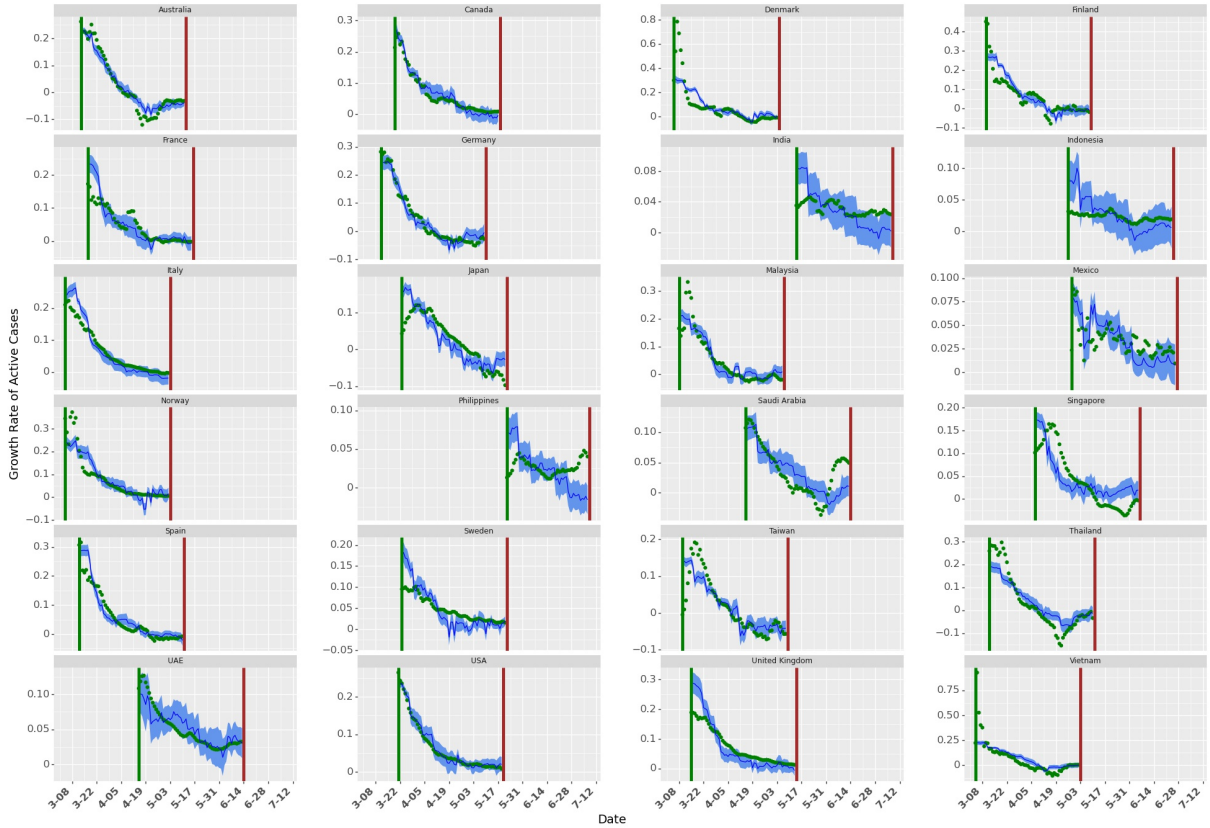


Figure 4.24: Growth rate predictions for different countries using current data. The green dots show the actual growth rates across countries. The dark blue line shows the mean of growth rate prediction (for 10000 samples in the Krinksky-Robb method). The blue shaded area shows the confidence bounds around the mean prediction.

Green and Brown vertical lines indicate the 60 days period for which data was collected for that country.

Since growth rate is a forward-looking model, we can also use growth rates to estimate active infectious population by $I_{j,t} = I_{j,t-1} \times g_{j,t}$. Figure 4.25 shows the results for daily active cases. Note that we estimate active cases using an exponential model. Thus, as the number of days in the prediction model increases, confidence interval bounds around predictions increase. However, the mean prediction for active cases closely approximates the actual active cases for different countries.

4.6.3.3 Effect of Masks, Social Mobility, and NPIs

We model growth rate as the first difference of log of daily active confirmed infectious cases as shown in Equation 4.7. Thus, the exponential of the coefficients in Table 4.6

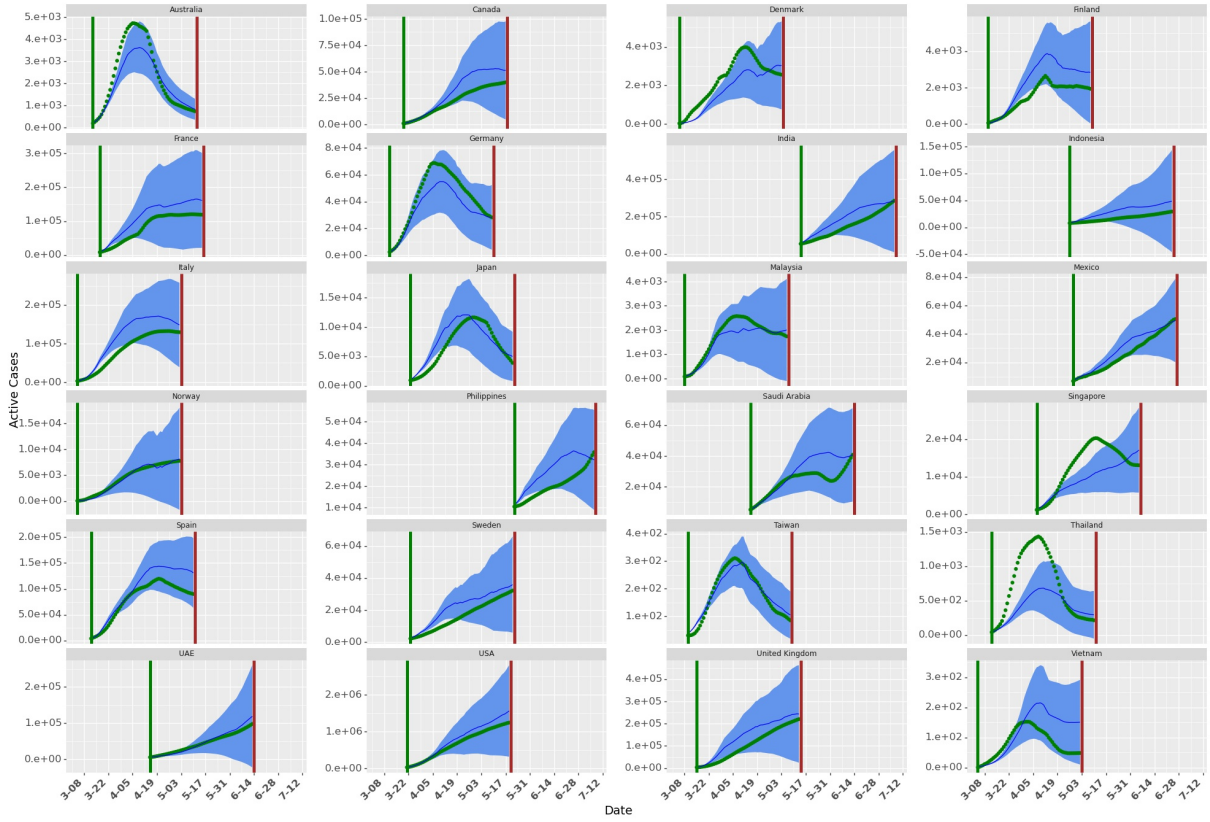


Figure 4.25: Simulating daily active cases using forward-looking growth rate model. The green dots show the actual daily active cases across countries. The dark blue line shows the mean of daily active cases prediction (for 10000 samples in the Krinksky-Robb method). The blue shaded area shows the confidence bounds around the mean prediction. As the days increase, the confidence bounds increase due to the multiplicative nature of forward-looking growth rate models.

(other than the mask as we discuss next) estimates % drop in active cases on the day t (as compared to active cases on the day $t-1$). Using the coefficients in Table 4.6, we can estimate the combined effect of masks, social mobility, and NPIs by using the Krinsky-Robb method.

Mask: Negative and statistically significant coefficient for masks show that increased mask-wearing behavior may lead to a decrease in the growth rate of COVID-19. As we use different transformations, we should interpret the coefficient for masks differently, too. When masks are transformed as $ln(1+mask)$, a coefficient of θ_m shows that if 100% of the population wears masks, it would lead to a daily drop of $1-e^{\theta_m(ln(1+1)-ln(1-0))}\% = (1-e^{\theta_m ln(2)})\%$ in the growth rate as compared with the scenario when no one wears

Masks	Lower Limit of Daily Drop	Daily Drop	Upper Limit of Daily Drop
$ln(1+mask)$	3.8%	6.9%	10.1%
$mask$	5.7%	9.1%	12.5%
$\sqrt{1+mask}$	4.9%	8.2%	11.4%

Table 4.7: Daily drop in growth rate when additional 100% people wear masks

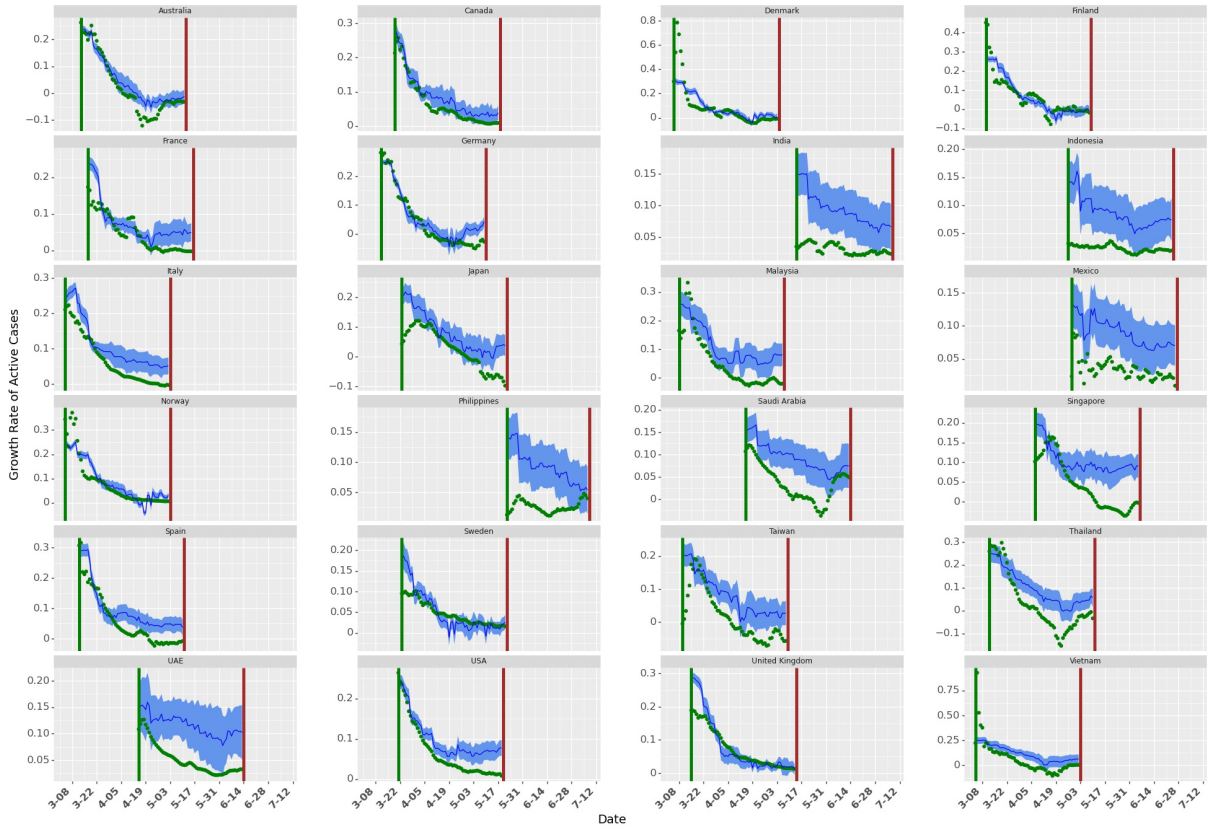


Figure 4.26: Growth rate with no mask-wearing. The green dots show the actual growth rates across countries. The dark blue line shows the mean of growth rate prediction (for 10000 samples in the Krinksky-Robb method). The blue shaded area shows the confidence bounds around the mean prediction. We transform the masks to $ln(1+mask)$ to predict the growth rate.

face masks. For raw mask numbers, the effect of masks can be interpreted directly as $(1-e^{\theta_m})\%$ drop in daily total infectious cases when everyone wears masks as compared to no one wearing masks. When masks are transformed as $\sqrt{1+mask}$, a coefficient of θ_m shows that if 100% of the population wears masks, it would lead to a daily drop of

$(1 - e^{\theta_m(\sqrt{1+1}-\sqrt{1+0})})\% = (1 - e^{\theta_m\sqrt{2}})\%$ in the growth rate as compared with the scenario when no one wears face mask. Similarly, we can estimate the bounds for the effect of coefficients of masks. Table 4.7 provides the estimate for the decrease in daily growth rate when 100% percent additional population wears face masks in public spaces (under different transformations).

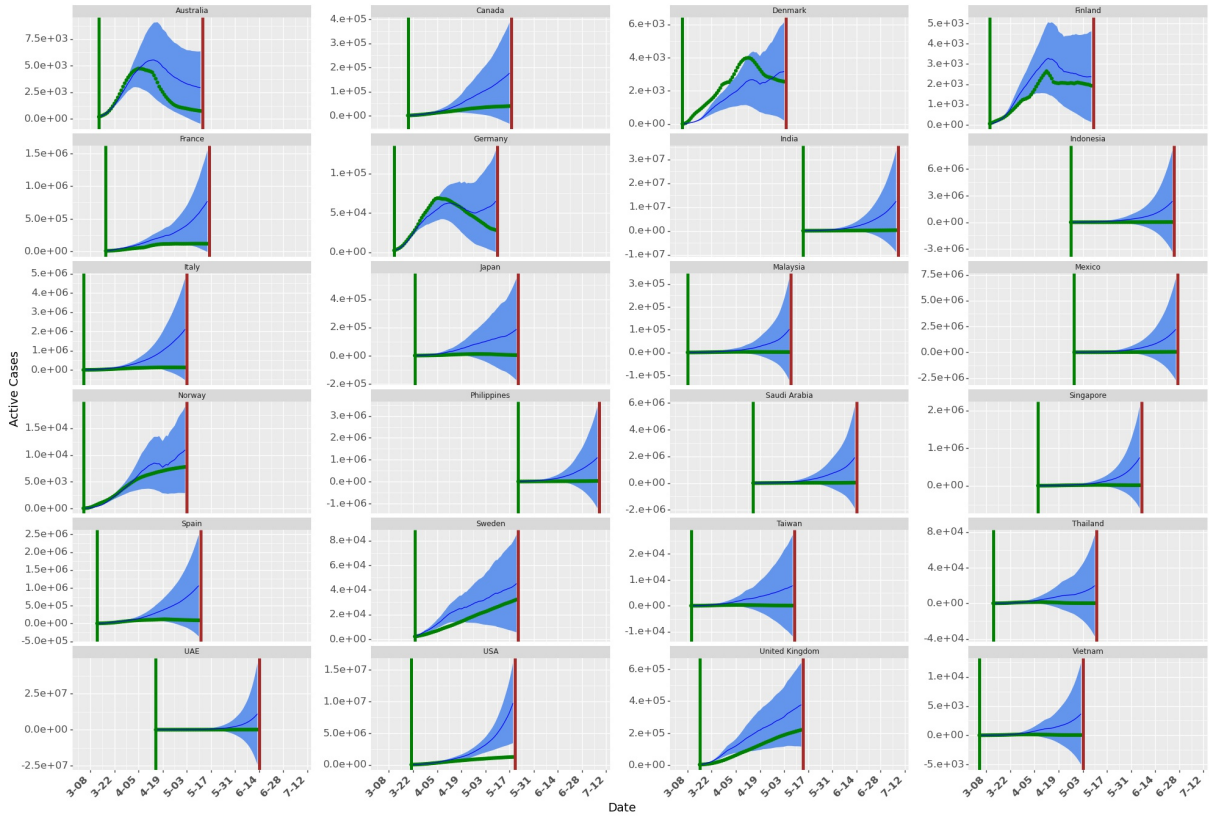


Figure 4.27: Simulation with no mask-wearing. The green dots show the actual active daily cases across countries. The dark blue line shows the mean of active cases prediction (for 10000 samples in the Krinksky-Robb method). The blue shaded area shows the confidence bounds around the mean prediction.

Figure 4.26 illustrates the effects of not wearing masks in each country. The results are not significantly different for Denmark, Finland, Norway, and Sweden as these countries already had very low numbers for mask-wearing in public spaces. Similarly, the effect is much more substantial for countries, e.g., Japan, Thailand, and Vietnam, with higher percentages of people wearing face masks.

Similar to Figure 4.25, we can predict daily active cases with zero percent mask-wearing

as shown in Figure 4.27. The results show that masks lead to a significant reduction in total cases as without these measures, the number of cases could exponentially increase over time (more discussion later on country-wise effect).

We build five simulation models to further understand the impact of masks. In the first simulation model, we consider a hypothetical country with a constant value for all the covariates in Equation 4.8. In the second simulation model, we check the change in active cases at the end of 60 days when mask-wearing in a country is $m\%$ where $m \in [0, 10, 20, ..., 100]$. In the third model, we check the change in active cases at the end of 60 days if the current levels of mask-wearing are multiplied by a factor of $x \in [0, 0.2, 0.4, ..., 2]$. In the fourth model, we present results for active cases at the end of 60 days when mask-wearing percentage increases by $a\%$ compared to the current levels in that country where $a \in [1, 2, ..., 10]$. In the fifth model, we exchange the mask-wearing numbers between countries with minimum and maximum average mask-wearing through the period of our analysis.

Simulation Model 1 helps isolate the effect of masks in this analysis; we do not consider any other covariates (as if no individual or institutional measures were taken apart from wearing masks in public areas). We construct data for a hypothetical country with these numbers to quantify the effect of masks in our analysis. In Simulation Model 2-4, we study the country-wise association of masks with growth rate. In these models, we do not change the numbers of any other covariates other than masks. These results show the potential change in active cases in that country for different percentages of people wearing masks in public. Simulation model 5 is used to build an approximate counterfactual model for masks by exchanging mask-wearing in countries with minimum and maximum average mask-wearing during our analysis. We discuss the results of these simulation models next.

Simulation Model 1: Average country

We simulate a hypothetical country with no mask-wearing ($mask_{j,t} = 0$), no active awareness ($trend_{j,t} = 0$), no testing ($testing_{j,t} = 0$), and no government implemented NPIs ($s_{j,t,p} = 0$). We predict the active cases at the end of 60 days at different levels of

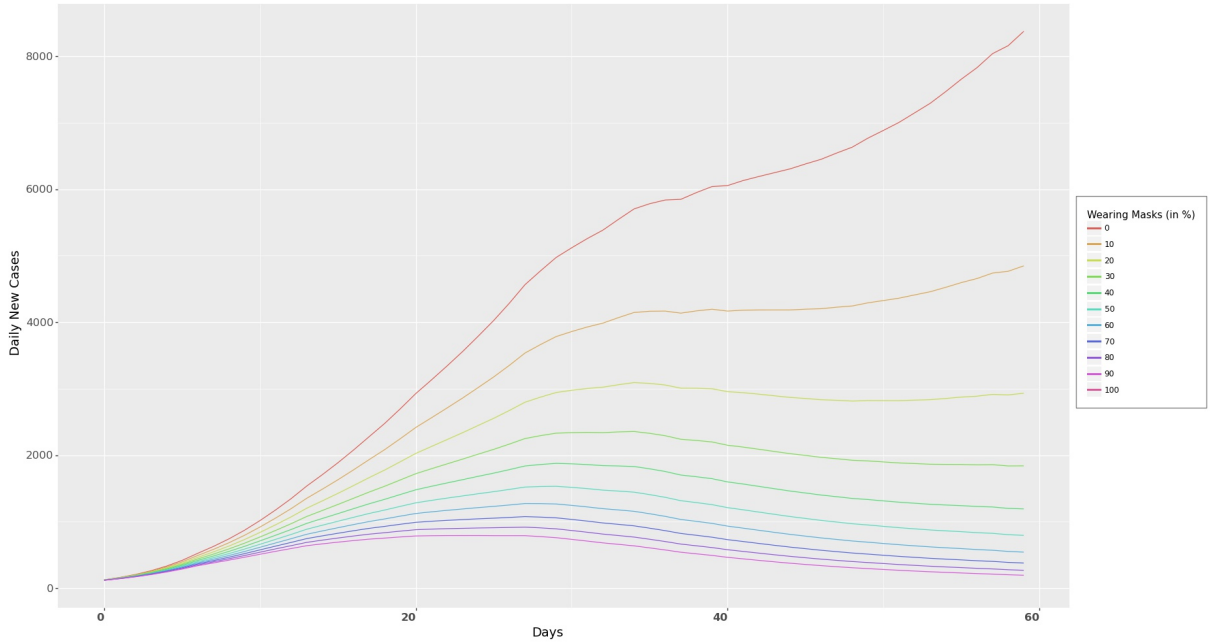


Figure 4.28: Daily active cases for the average country. As more people wear masks, the faster the curve can be flattened. Note that the graph shows the mean prediction under different levels of mask-wearing. Also, it shows the daily active cases. Flattening of a curve has been synonymous with daily new cases, but if we use $I_t - I_{t-1}$, we can approximate the growth rate.

mask-wearing. We consider the average country with 0 country fixed effects for prediction. Figure 4.28 shows the daily active cases for this average country at different levels of mask-wearing. We assume that the average country has 100 cases on day 0 in the simulation. Results show that increasing the mask number can help flatten the curve (even when social mobility and NPIs remain unchanged). As the percentage of people wearing face masks increases, the daily active cases go down compared to no mask-wearing. Results also show that social mobility and NPI can play a significant role in flattening the curve (the daily active curve flatten even with no mask-wearing, albeit slower). Cases start rising again after initial flattening for most cases as social mobility increases and NPIs are relaxed (Figure 4.14). The results imply that if masks are mandated and become widespread, complete lockdowns may be eased to help alleviate the associated economic hardships.

Simulation Model 2: Changing mask levels

In this simulation, we predict the number of active cases in each country by changing the levels of mask-wearing. Figure 4.29 shows the ratio of active cases at the end of 60 days under different levels of mask-wearing as compared to active cases at the current levels of masks. As the mask levels increase, the ratio of active cases to the actual active cases at the end of 60 days decreases.

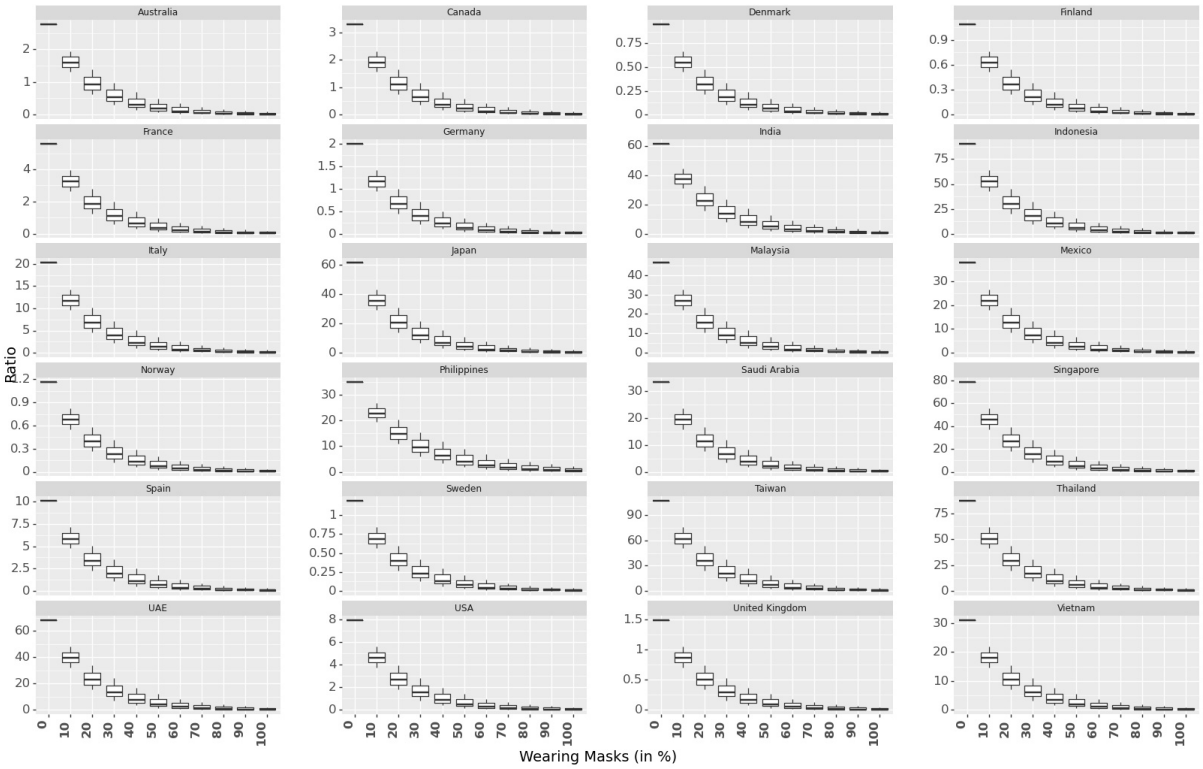


Figure 4.29: Box plot on ratio of predicted active cases to actual cases at the end of 60 days under different levels of mask wearing.

Simulation Model 3: Multiplying a constant to current mask levels

In this simulation, we multiply the current mask-wearing levels with a constant multiplication factor $\in [0.2, 0.4, ..., 2]$ to predict active cases at the end of 60 days as compared to actual cases across 24 countries (Figure 4.30). Similar to Figure 4.29, we observe that as we increase mask levels, the ratio decreases significantly. However, the effect is different across different countries.

Similar to results in Figure 4.29, when the mask levels are much lower than the current levels (e.g., countries like Thailand, Vietnam, Singapore), the ratio of active cases at the

end of 60 days to their actual cases is much higher compared to countries with lower current mask rates (e.g., Sweden, Norway).

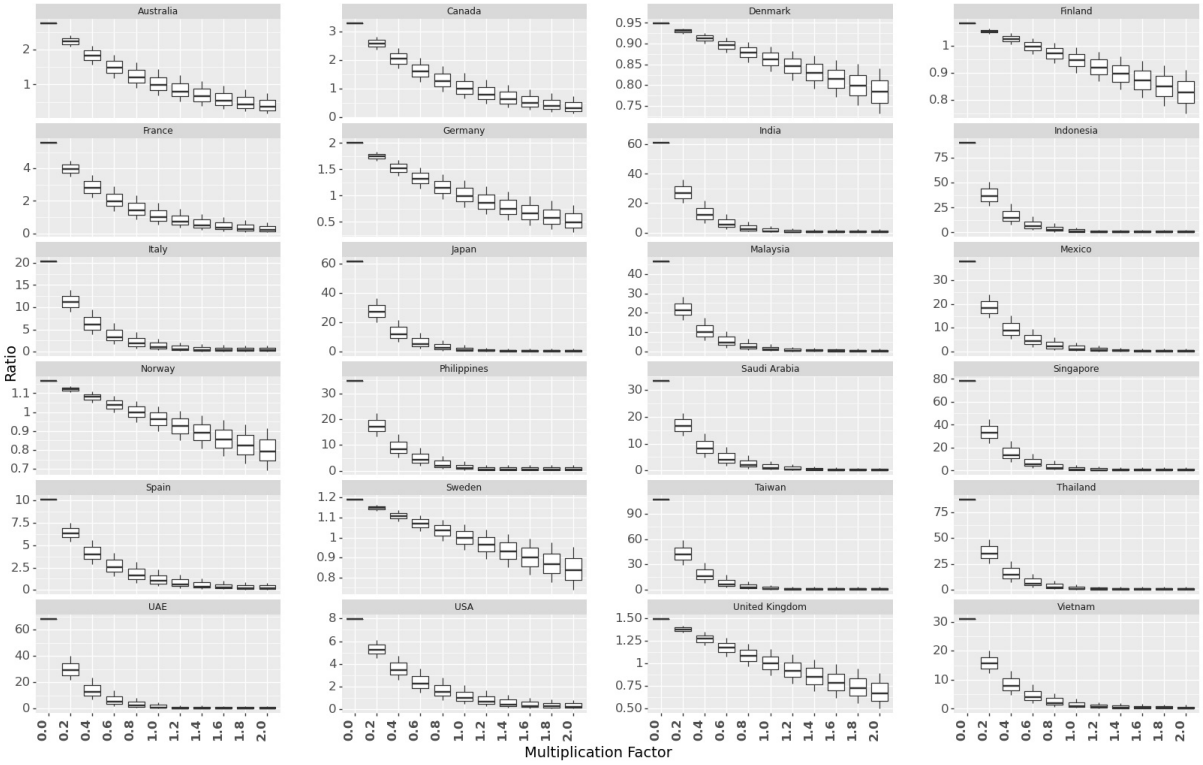


Figure 4.30: Box plot on ratio of predicted active cases to actual cases at the end of 60 days under different levels of mask wearing obtained by multiplying the current levels of mask wearing with a constant.

Simulation Model 4: Adding a constant to current mask levels

In this simulation, we predict the ratio of active cases at the end of 60 days when mask-wearing in a country is increased by different percentage points (0%, 1%, 2%, . . . , nine %). Figure 4.31 plots the ratio of active cases at the end of 60 days with simulation for increased mask-wearing to the actual active cases. This could help the government from policies that if a% of more people follow the mask-wearing guidelines, which NPIs they could relax while still controlling the spread of the virus. Enforcing a mask-wearing policy could be particularly useful in countries with low mask-wearing.

Simulation Model 5: Exchanging the current mask levels

Countries have observable heterogeneity in their culture of mask-wearing (Figure 4.10). Mask wearing has been more common in Asian countries than in Scandinavian countries.

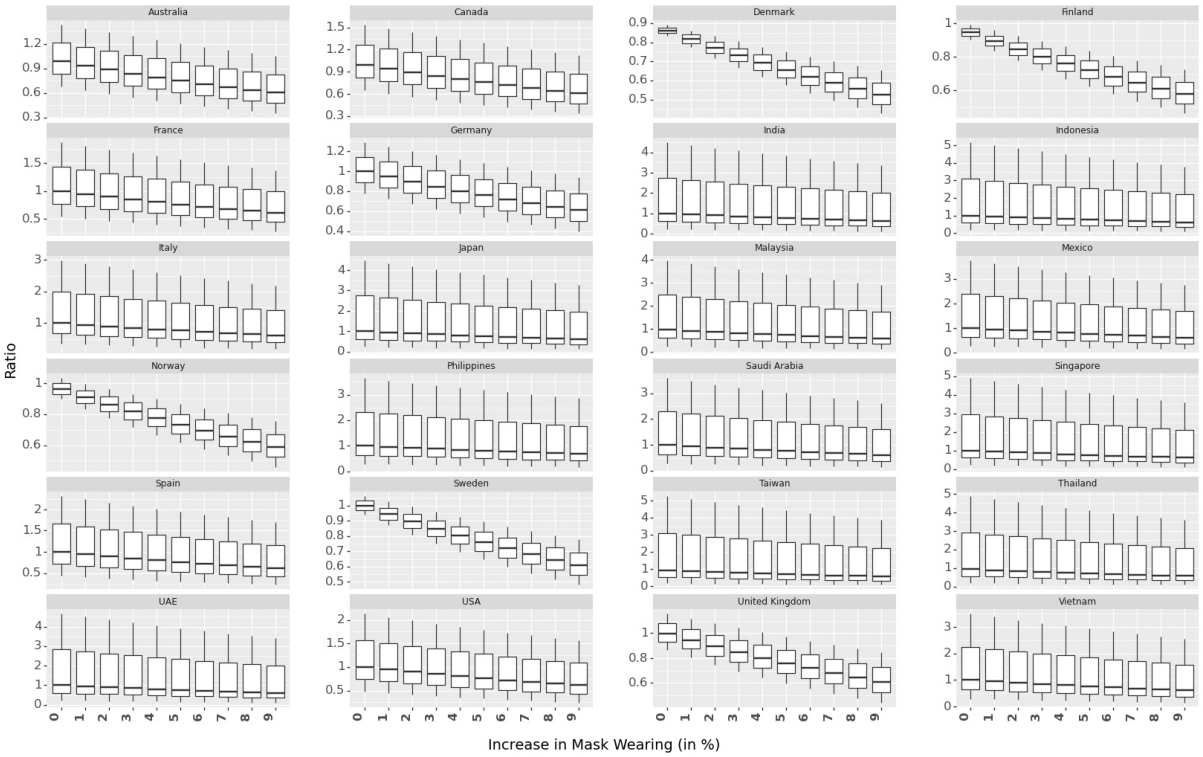


Figure 4.31: Box plot on ratio of predicted active cases to actual cases at the end of 60 days under different levels of mask wearing obtained by increasing the current levels of mask wearing by different percentage points.

In this simulation, we exchange the mask-wearing numbers between 8 countries (4 Asian with highest average mask-wearing among 24 countries and 4 Scandinavian countries with lowest average mask-wearing among 24 countries). The ratio of active cases under new mask-wearing as compared to actual active cases at the end of 60 days is shown in Table 4.8. Results show that Scandinavian countries could have reduced their confirmed cases significantly if they had enforced people to wear face masks in public. We find that the Scandinavian countries could have reduced the active cases by up to 50 times in 60 days if the citizens were wearing masks at Asian countries' levels.

Next, we present the combined effect of social mobility and NPIs. We provide a combined effect for social mobility and NPIs as it is difficult to estimate the causal analysis for individual variables. We use the Krinsky-Robb method to estimate the combined effect of social mobility and NPIs. After drawing samples of coefficients of social mobility and NPIs, we add the random samples draw and present the mean and confidence interval

	country	exchanged with	Lower Limit for 95% CI	Ratio	Upper Limit for 95% CI
max	Malaysia	Denmark	23.86516	25.13209	26.44875
max	Philippines	Finland	12.00857	13.95786	16.19232
max	Taiwan	Norway	35.53529	39.52427	43.90114
max	Thailand	Sweden	33.90308	36.76336	39.82361
min	Denmark	Malaysia	0.009502	0.041856	0.180894
min	Finland	Philippines	0.00727	0.035283	0.167807
min	Norway	Taiwan	0.00432	0.025148	0.143133
min	Sweden	Thailand	0.004408	0.025332	0.142346

Table 4.8: Ratio of active cases at the end of 60 days after exchanging mask wearing numbers

Social Mobility	Country	Exchanged with	Lower Limit for 95% confidence Interval	Ratio	Upper Limit for 95% confidence Interval
High	Philippines	Taiwan	1.81473	6.940316	26.09045
High	Denmark	United Kingdom	2.495734	2.620681	2.750162
High	Sweden	India	1.163527	1.270406	1.385541
High	Norway	Spain	0.998357	1.100658	1.211927
Low	Spain	Norway	0.363724	0.947547	2.438379
Low	India	Sweden	0.026063	0.123776	0.576219
Low	United Kingdom	Denmark	0.314838	0.384042	0.467267
Low	Taiwan	Philippines	0.016745	0.098078	0.561585

Table 4.9: Ratio of active cases at the end of 60 days after exchanging social mobility numbers

bounds of these samples as the combined effect and confidence interval bounds of that combined effect.

Social Mobility: Parameter coefficients in Table 4.6 show that growth rate increases as mobility increases. If people travel more or move to places with the potential of public gatherings, infected individuals can spread the virus to the susceptible population. Governments, therefore, imposed strict restrictions to reduce mobility. We report the effect of mobility as negative of the coefficients in Table 4.6. Thus, we report the effect of mobility if the mobility numbers were 0 (no mobility change). Results in Figure 4.32

indicates that 0 change in mobility trends (no change in individual mobility trend indicates if people move around as they were before COVID-19) is associated with a daily increase in growth rate by 8.1% (5.6% - 10.6%) as compared to actual cases. Note that a decrease in mobility can also attribute to NPIs. No causality can be claimed on the effect of increased mobility on growth rate.

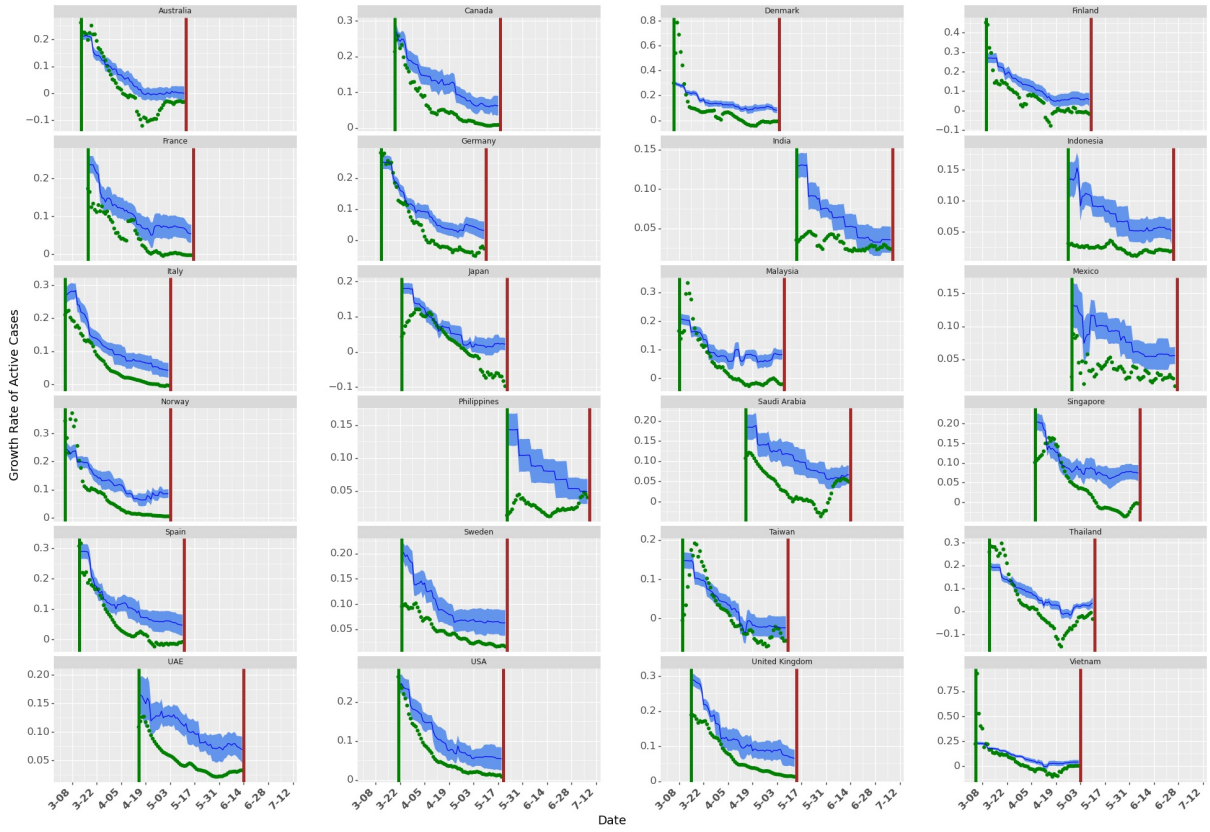


Figure 4.32: Growth rate with full mobility. The green dots show the actual growth rates across countries. The dark blue line shows the mean of growth rate prediction (for 10000 samples in the Krinksky-Robb method). The blue shaded area shows the confidence bounds around the mean prediction.

Figure 4.32 depicts the effect of full mobility across different countries. It shows that even with mask numbers remaining unchanged and NPIs being implemented as they were implemented in that country, increasing mobility can lead to a significant increase in growth rate. Similar to Figure 4.27, we can predict daily active cases with full mobility as shown in Figure 4.33.

Similar to exchanging the mask-wearing numbers among the countries with the lowest

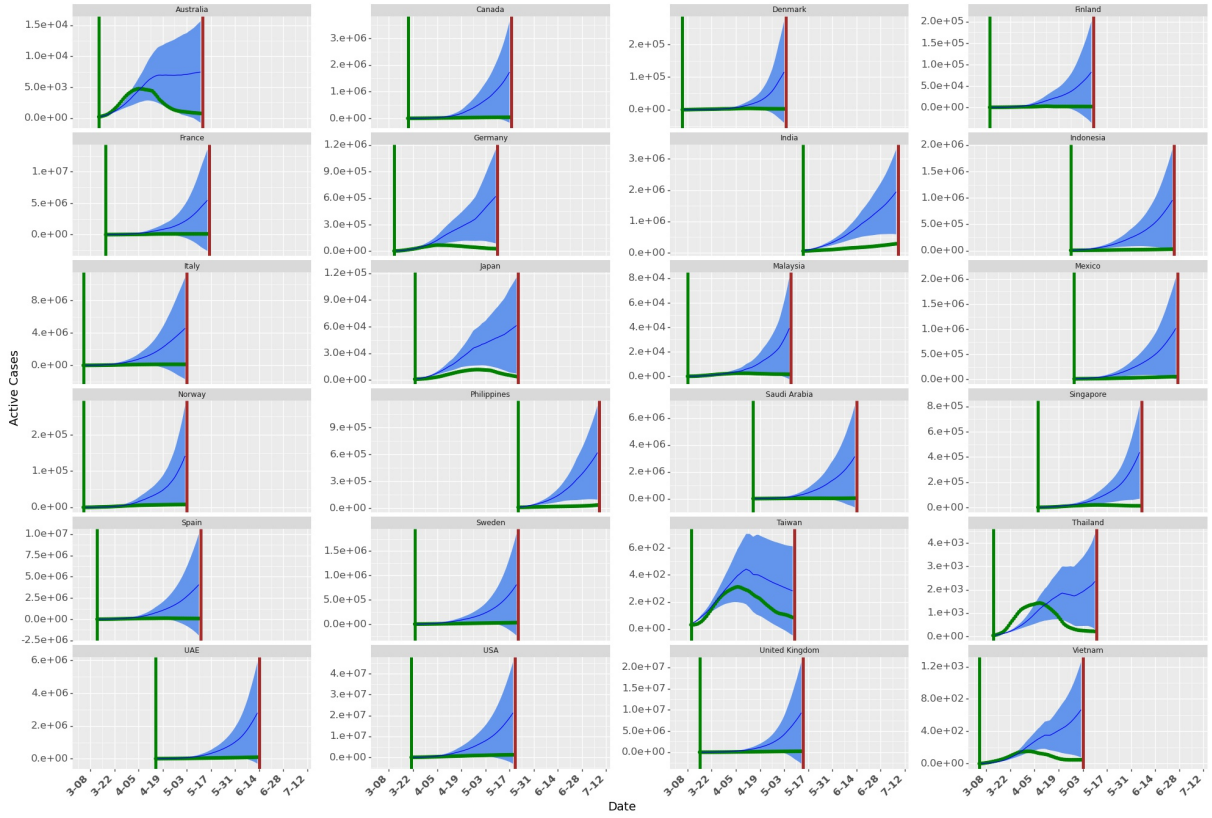


Figure 4.33: Simulation with no change in mobility (as compared to pre-COVID-19 mobility). The green dots show the actual daily active cases across countries. The dark blue line shows the mean of active daily cases prediction (for 10000 samples in the Krinksky-Robb method). The blue shaded area shows the confidence bounds around the mean prediction.

and highest mask-wearing percentages, we simulate the active cases at the end of 60 days by exchanging social mobility among countries with the highest and lowest social mobility. Table 4.9 summarizes the results.

Non-Pharmaceutical Interventions (NPIs): Negative and statistically significant estimates for the combined effect of NPIs show that NPIs helped in controlling the spread of the virus. Results in Table 4.6 indicate that if mask-wearing and mobility remain unchanged, implementing NPIs is associated with a daily drop of infectious cases by 13% (9.2% - 16.2%). Predicted growth rate and daily active cases with no NPIs are shown in Figures 4.34 and 4.35, respectively.

Similar to exchanging the mask-wearing numbers among the countries with the lowest

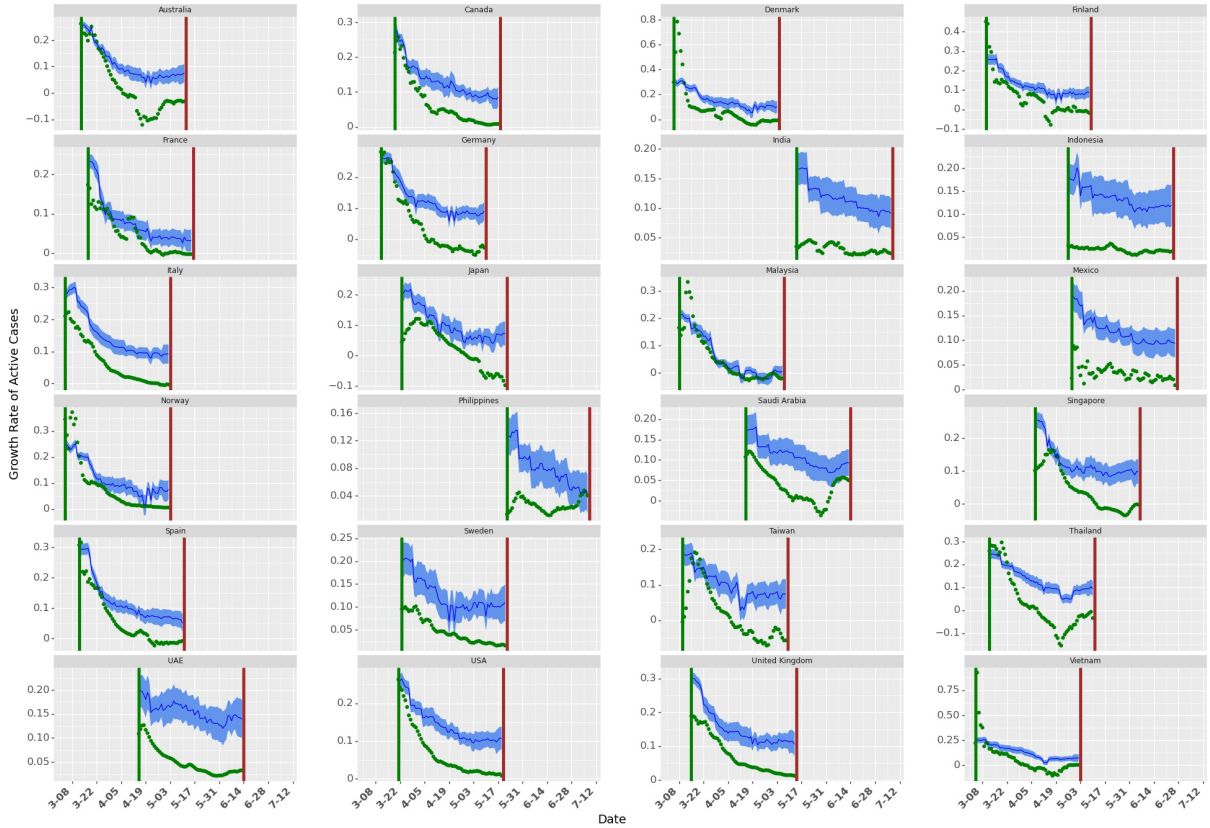


Figure 4.34: Growth rate with no NPIs. The green dots show the actual growth rates across countries. The dark blue line shows the mean of growth rate prediction (for 10000 samples in Krinksky-Robb method). The blue shaded area shows the confidence bounds around the mean prediction.

and highest mask-wearing percentages, we simulate the active cases at the end of 60 days by exchanging NPI numbers among countries with the highest and lowest number of NPIs introduced across the country. The results are shown in Table 4.10.

Combined Effect of Mask, Social Mobility and NPIs: Similar to presenting the combined effect of mobility and NPIs, we use the Krinsky-Robb method to present the effects for the combined effect of masks, social mobility, and NPIs in the daily drop of growth rate in Figure 4.36. Figure 4.36 also shows the robustness of the model across different values of $shift$. It also shows the Mean Absolute Percentage Error for 10–fold cross validation used to get the $shift$ that best fit the data. We observe best data fit for a lag of 9 days. Grey vertical lines indicate a shift of 7 days and 11 days. The combined effect of masks, social mobility and NPIs is estimated to be a 28.1% (24.2%-32%) drop in

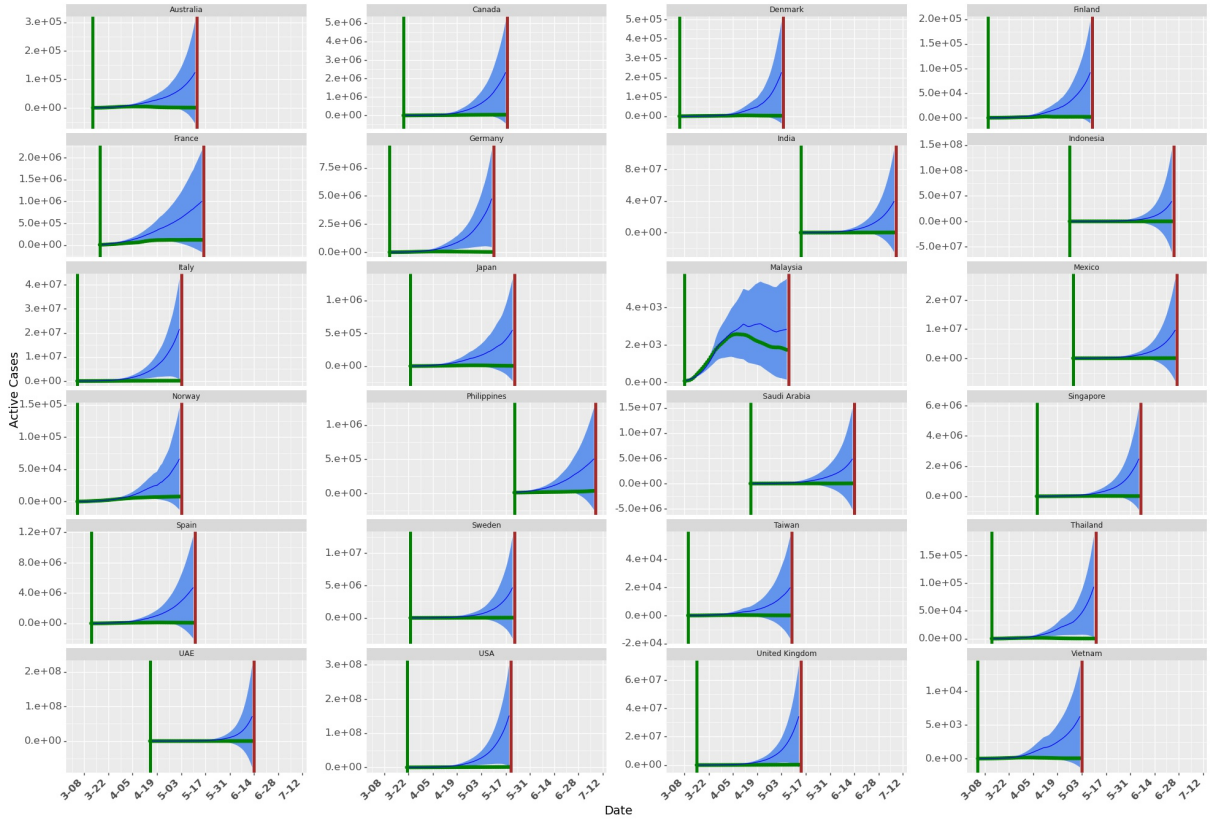


Figure 4.35: Simulation with no NPIs implemented. The green dots show the actual daily cases across countries. The dark blue line shows the mean of active cases prediction (for 10000 samples in Krinksky-Robb method). The blue shaded area shows the confidence bounds around the mean prediction.

daily growth rate.

Results show that the effect of masks remains consistent across different transformations and different $shift$. We also observe consistency across different transformations of mask numbers. Furthermore, the total combined effect of masks, social mobility, and NPIs remain consistent as we change $shift$.

The results show that masks, social mobility, and NPIs lead to a significant reduction in total cases, and without these measures, the number of cases could exponentially increase over time.

4.6.3.4 Testing and Google Trends

The negative and statistically significant coefficient for testing and Google Trends indicates a daily growth rate drop as these numbers increase. As testing increases, it

NPI	Country	Exchanged with	Lower Limit for 95% confidence Interval	Ratio	Upper Limit for 95% confidence Interval
High	Italy	Finland	2.551666	8.34897	26.90575
High	Australia	Norway	10.81348	17.51927	28.20859
High	Thailand	Malaysia	40.38261	209.2452	1061.602
High	Singapore	France	10.7138	59.20919	320.1261
Low	Finland	Italy	0.12681	0.13582	0.145342
Low	Norway	Australia	0.084334	0.092976	0.102375
Low	Malaysia	Thailand	0.001354	0.006118	0.027119
Low	France	Singapore	0.00753	0.015919	0.033334

Table 4.10: Ratio of active cases at the end of 60 days after exchanging NPI numbers

may show increased daily confirmed cases (as more people get tested and it can discover asymptomatic cases). However, our model uses a lag of 9 days. Thus the cases may not be affected by testing immediately. Thus, a negative coefficient with a lag shows that testing helps achieve a daily drop of growth rate in active infectious cases. Similar to testing, as our model uses a lag of 9 days, increased google trends indicate increased awareness among the citizens regarding COVID-19, leading to more caution against COVID-19.

4.6.4 Robustness Checks

4.6.4.1 Selection of th and $shift$

To filter the initial volatile growth rate, we use a threshold in the model (one for each country). We start collecting data for each country from the day it reaches its threshold. We define threshold as the day after which the 7-day average of daily new cases were th % of the peak daily new cases observed in that country. Decreasing th will add noise to the model due to high volatility in early growth rate values. However, if the threshold is high, we miss out on important data, particularly during the initial phase when the growth in infections is exponential.

Along with th, we also use $shift$ in the growth rate model to capture the delay in the effect of the mask, Non-Pharmaceutical Interventions (NPIs), and mobility. First, we select the value of th and then use that th to find optimal $shift$ for our analysis. We

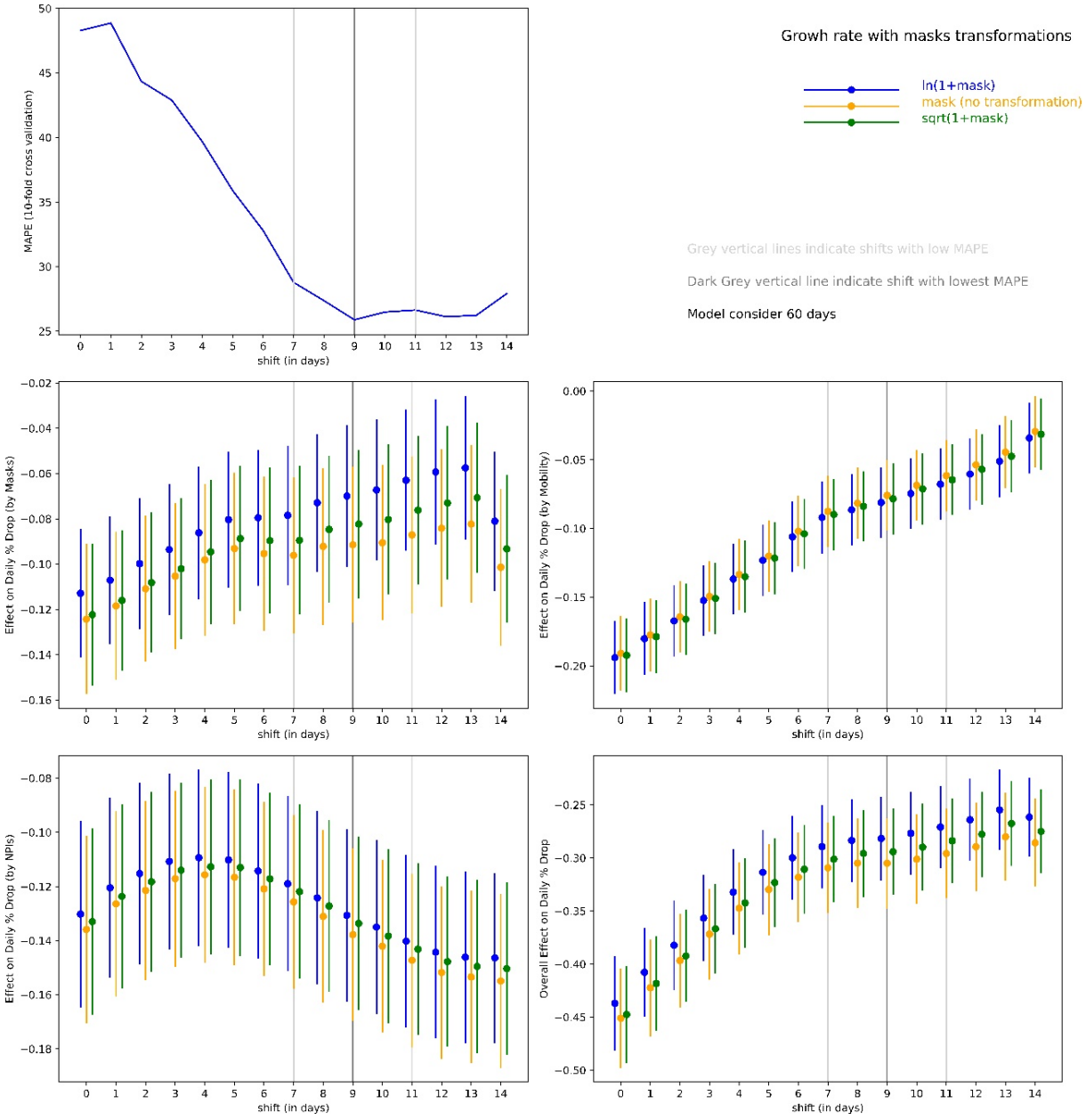


Figure 4.36: Top left: Mean Absolute Percentage Error under various $shift$ values. Bottom four: Combined effect of mask, mobility, and NPIs in daily % drop of infectious cases under various $shift$ values.

calculate log-likelihood for growth rate model under different $th \in [0.01, 0.02, ..., 0.3]$. For each th, we run the model for different $shift$. Using multiple lags ensures that the model performance is consistent with different values of $shift$. It also ensures that we do not select th that performs well by chance. We select th based on maximum average log-

likelihood for different $shift$. We start with a threshold value of 0.01 and start increasing it. Figure 4.37 shows that the average log-likelihood for growth rate does not change much after $th = 0.18$. We use a $th = 0.2$ in the rest of the chapter. We use a sensitivity test to check the consistency of the model.

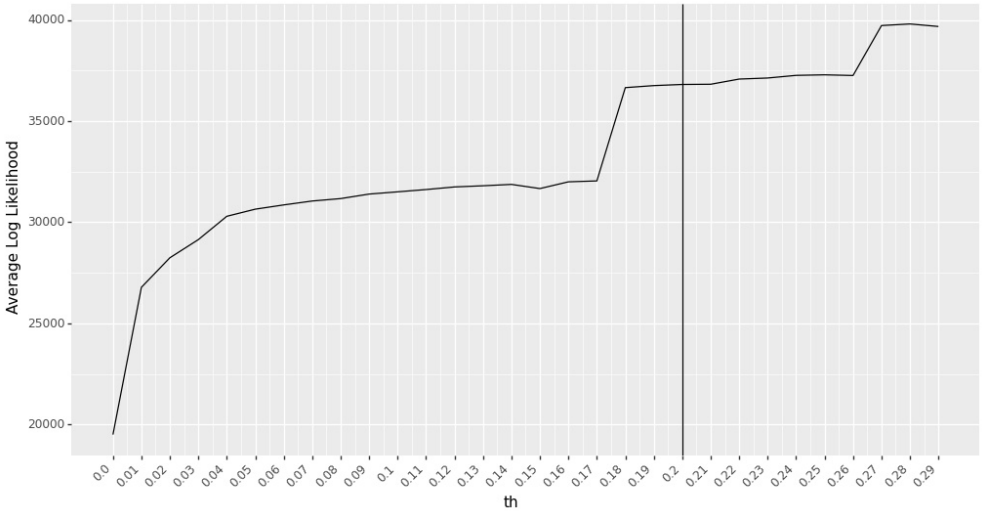


Figure 4.37: Average log-likelihood values for different th. The model performs the best (based on maximum likelihood function) when $th = 0.28$. However, the performance change is not significant after $th = 0.18$. So we use $th = 0.2$ in our analysis in this research.

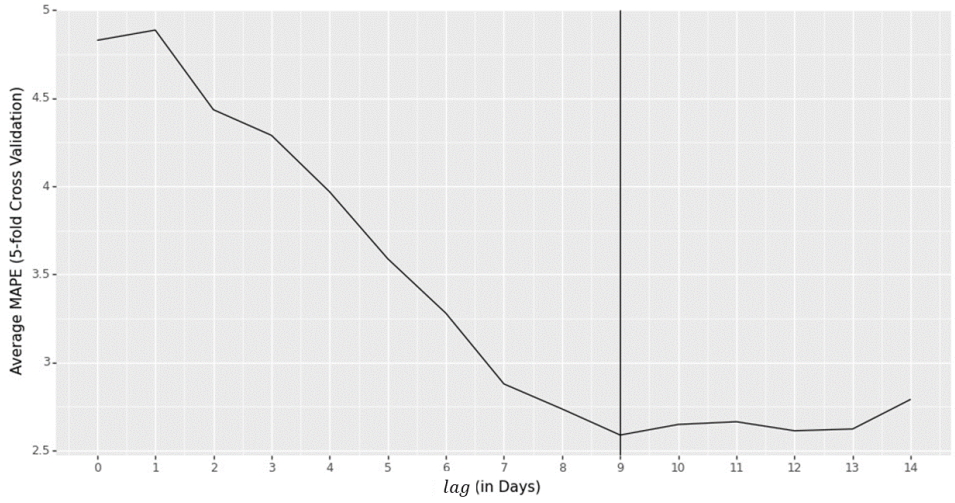


Figure 4.38: Average MAPE from 10-fold cross validation of different lags. Minimum value for MAPE is obtained at a lag=9days.

In the growth rate model, we use $shift$ to estimate the parameter coefficients as shown in Equation 4.8. After selecting th, we use cross validation to find the best lag that fits the data. We use 10-fold cross-validation with Mean Average Percentage Error (MAPE) as a metric for out-of-sample data points to select $shift$ with the best fit. Figure 4.38 shows the average MAPE (for 10-fold cross validation) for different lags. $shift = 9$ days shows the best fit with MAPE's minimum value.

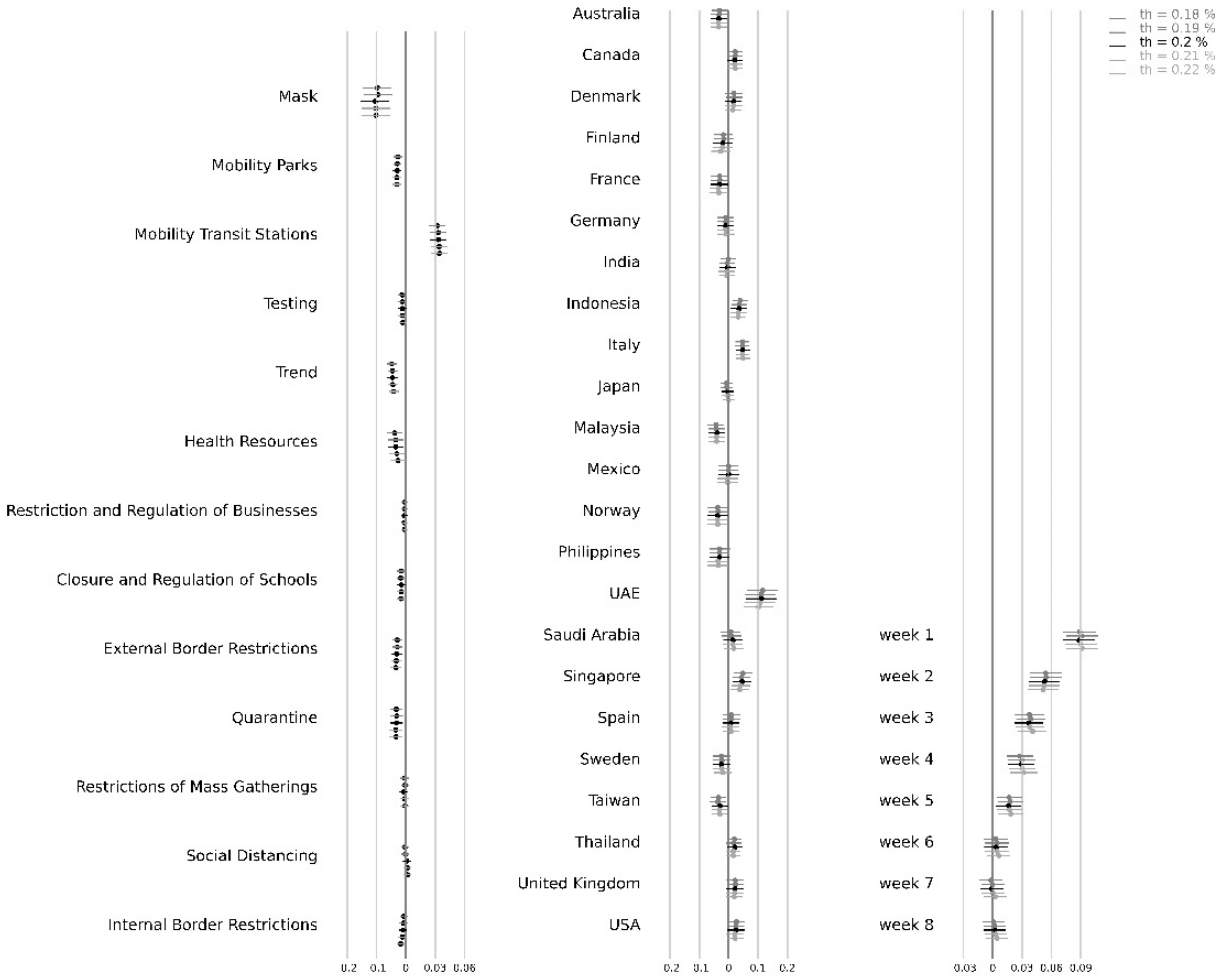


Figure 4.39: Parameter estimates for growth rate model for different th. Horizontal lines represent the upper and lower confidence interval bounds for the parameter estimates. We show the results for a lag of 9 days with $th \in [0.18, 0.19, 0.2, 0.21, 0.22]$. The results indicate that the model is robust to different th as the parameter estimates show consistency. Vietnam is kept as the base country (the fixed country effect is 0).

4.6.4.2 Model Estimation Sensitivity to th and $shift$

We use sensitivity test to check the consistency of the parameter estimates for different values of th and $shift$. The performance of the model remains consistent on changing the values of th from 0.18 to 0.22 as shown from parameter estimates in Figure 4.39 (we use a $shift$ of 9 days). The performance of the model remains consistent on changing the values of $shift$ from 7 days to 11 days as shown from parameter estimates in Figure 4.40 (we use $th = 0.2$). We transform masks as $ln(1 + mask_{j,t})$.

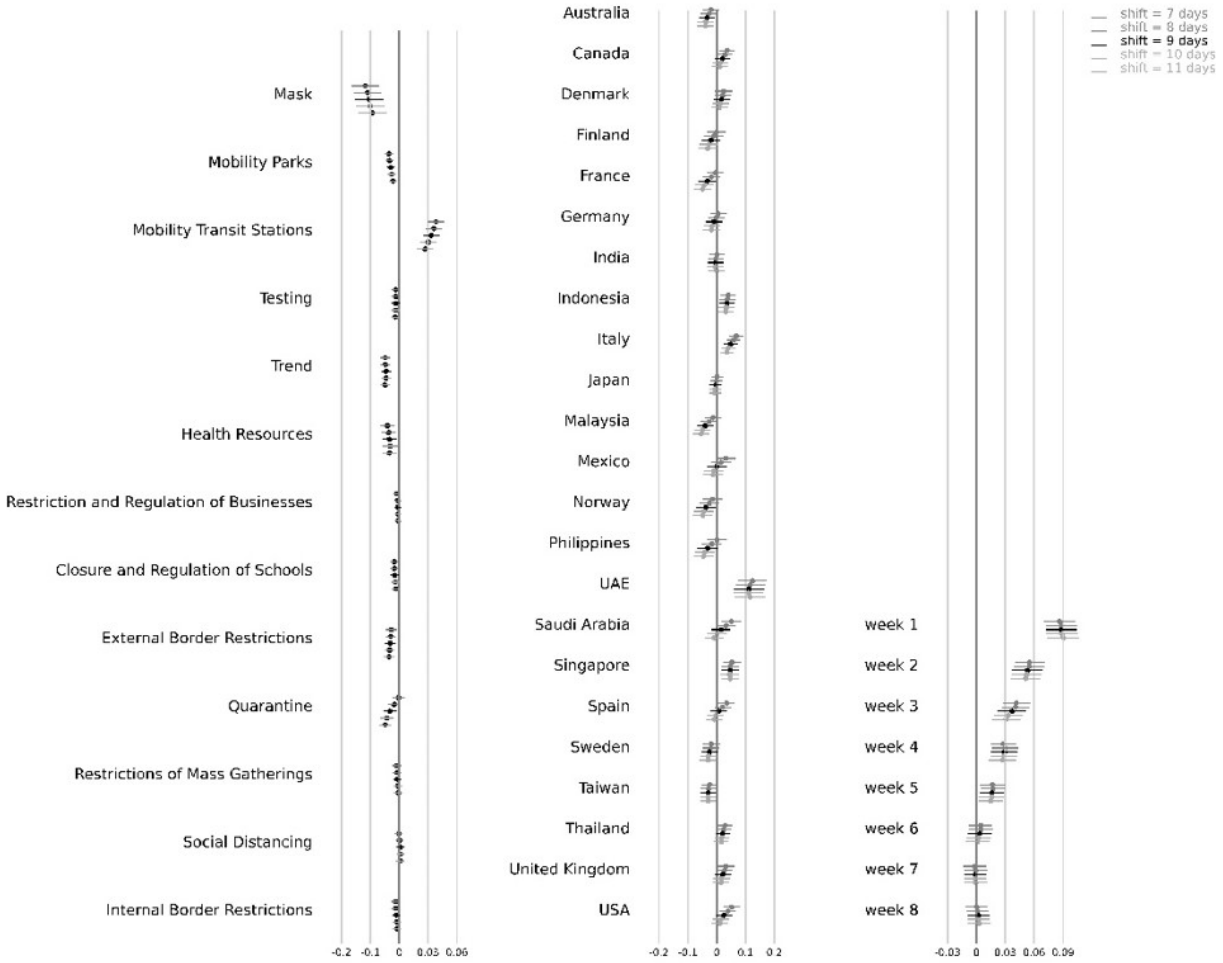


Figure 4.40: Parameter estimates for growth rate model for different $shift$. Horizontal lines represent the upper and lower confidence interval bounds for the parameter estimates. We show the results for a lag of 7 days to 11 days with $th = 0.2$. The results indicate that the model is robust to different $shift$ as the parameter estimates show consistency. Vietnam is kept as the base country (the fixed country effect is 0).

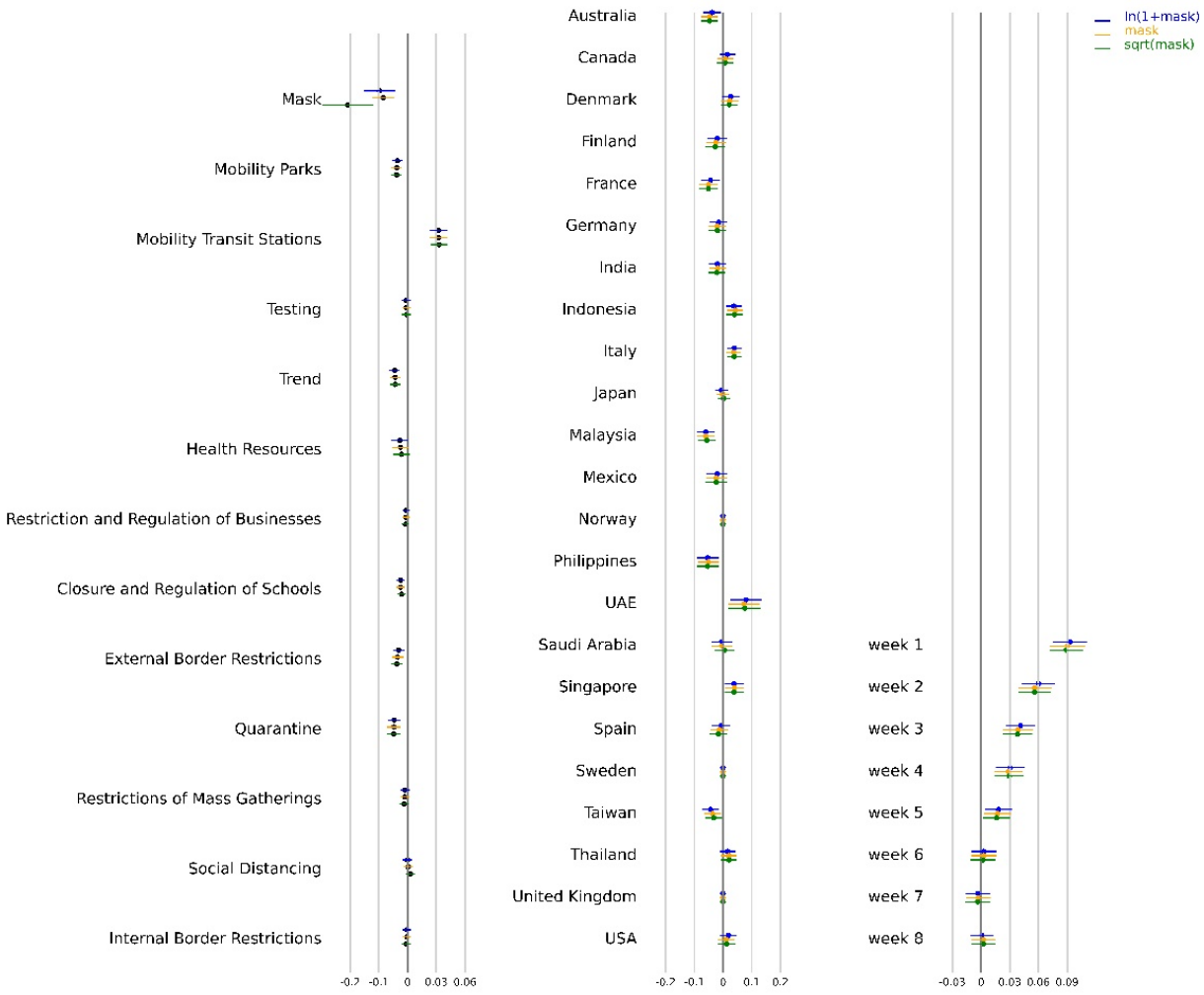


Figure 4.41: Parameter estimates for growth rate model after excluding countries Norway, Sweden, and United Kingdom from the analysis. Horizontal lines represent the upper and lower confidence interval bounds for the parameter estimates. We show the results for different transformations of $mask_{j,t}$. The results indicate that the model is robust to different transformations as the parameter estimates show consistency. Vietnam is kept as the base country (the fixed country effect is 0).

4.6.4.3 Handling Data Error in Active Cases

In Figure 4.11, we observe data reporting issues in Norway, Sweden, and United Kingdom. The recovered cases are reported late in Norway, not reported in Sweden, and recorded very low in the United Kingdom. We use data from Johns Hopkins Resource Center. Our analysis uses total confirmed cases to find active cases for these three countries. However,

this may lead to bias in the results. To check the bias, we run the model without these three countries. Figure 4.41 shows the parameter estimates after excluding these countries. In Figure 4.41, we show the parameter coefficients when $shift = 9$ days and $th = 0.2$.

The results show that the model parameter estimates are robust to excluding Norway, Sweden, and the United Kingdom from the model. However, the model slightly overestimates the coefficient of masks after excluding Norway, Sweden, and United Kingdom (compared to the growth model with data from all the 24 countries). All three countries are in Europe, where wearing face masks is not as common as in Asian countries. Moreover, Norway and Sweden (along with Denmark and Finland) have the lowest percentage of people who wear face masks in public in our data set.

4.6.4.4 Robustness Check for Mobility

In our analysis, we used Google's Community mobility reports numbers as a measure of social mobility as android operating devices are more common than iOS, particularly in Asian countries [156]. Also, Apple's Community Mobility Reports record data only when an individual opens Apple Maps. To check the robustness of the combined effect of masks, social mobility, and NPIs, we also consider Apple's community mobility numbers as a measure of social mobility. Apple released data for change in trends for driving and walking for all the 24 countries considered in this work. Figure 4.42 shows the combined effect after substituting with apple's mobility report. The consistency of the results confirms that the estimates for the model are robust.

4.6.4.5 Alternative Specifications

We build two robustness models to check the consistency and reliability of the parameter estimates of the growth rate model. In the first robustness check, we use exponential smoothing as shown in Equations (4.9) and (4.10). In the second model, we use a control function approach to identify the impact of masks on the spread of COVID-19. We discuss it in detail in this subsection.

Model 1: Exponentially Smoothed Variates for Growth Rate

In the first specification, we use exponential smoothing to estimate masks, NPIs, and social mobility parameters. In our base model in Equation 4.8, the growth rate is defined

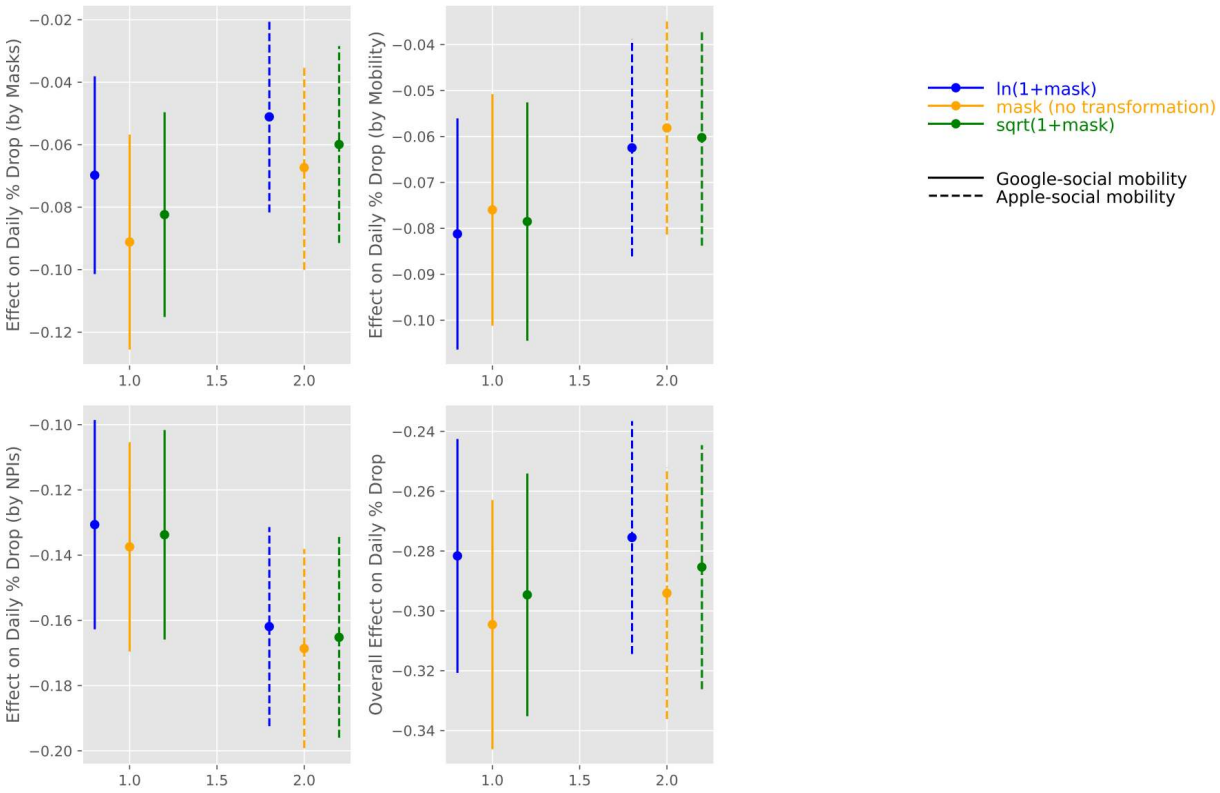


Figure 4.42: Combined effect of masks, social mobility, and NPIs with mobility numbers from Google and Apple.

as a function of masks, NPIs, social mobility, trend, and testing at a lag of $shift$ days. On any day t, this model ignores the value of the variates from days $t - shift + 1$ to t (discussed in Figure 4.7). In this model, we do not ignore variates between $t - shift$ and t and use exponential smoothing average to check the robustness of our model. To check the consistency of the parameter estimates for different transformations of masks, mobility, NPIs, and the combined effect of masks, social mobility, and NPIs is shown in Figure 4.43.

We also show the parameter estimates from the growth model for comparison. The results show that the parameter estimates for both models are close and consistent, thus showing the robustness of the results in Table 4.6.

In the next model, we use a control function approach to check the robustness of our estimates in Table 4.6. The control function approach considers an error variable based on an exogenous variable that is not correlated with the response variable but with

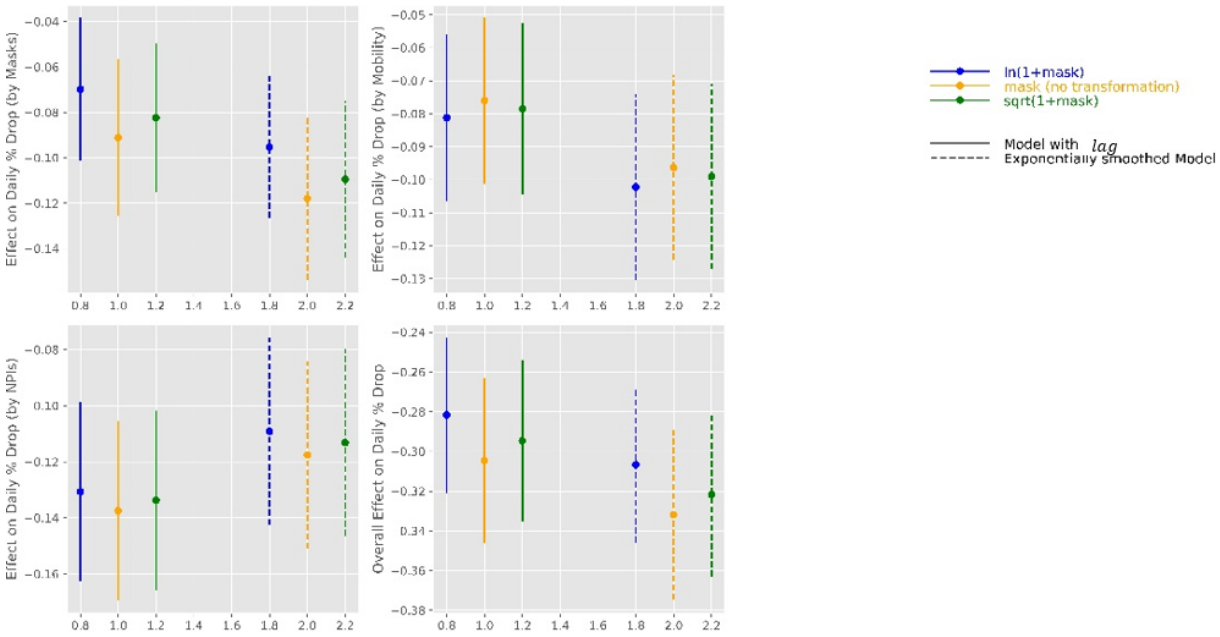


Figure 4.43: Effect on daily % drop in growth rate by masks (top left), mobility (top right), NPIs (bottom left), and the Combined effect of masks, social mobility and NPIs (bottom right) with $shift = 9$ days for exponentially smoothed model.

an instrumental variable. We use the number of deaths per thousand people for SARS, H1N1, and MERS CoV as our exogenous variables.

Model 2: Control function Approach to Growth Rate

In the control function approach, we first predict the average value of $mask_{j,t}$ by using the number of deaths per thousand people in each country by SARS, MERS-CoV, and H1N1. Results from predicting masks using disease per thousand people are shown in Table 4.11 and the parameter estimates are shown in Table 4.12.

We consider all the available data sets (from February 21, 2020, to July 8, 2020) to estimate SARS, H1N1, and MERS coefficients. We convert the numbers for deaths due to SARS, H1N1, and MERS into a binary variable (1 if the number for a country is greater than the median).

We use $mask_{j,t}$ along with the residuals from prediction model, $e = mask_{j,t} - \widehat{mask_{j,t}}$, as control function in Equation 4.11. Figure 4.44 illustrates the results for the combined effect of masks, social mobility, and NPIs. We use a $shift$ of 9 days. We also show the combined effect of masks, social mobility, and NPIs without control function (the focal

R-squared	0.056
Adj. R-squared	0.055
F-statistic	65.24
Probability (F-statistic)	0
Log-Likelihood	-8108
AIC	16200
BIC	16250
No. Observations	3312
Degree of Freedom: Residuals	3308
Degree of Freedom: Model	3

Table 4.11: Results statistics for predicting $log(1+mask)$ using SARS, H1N1, and MERS

coef	std	err	t	P> $\|t\|$	[0.025	0.975]
Sars	0.5013	0.179	2.797	0.005	0.15	0.853
H1N1	-1.3904	0.101	-13.78	0	-1.588	-1.193
Mers	0.1556	0.114	1.363	0.173	-0.068	0.38
const	6.9685	0.071	98.706	0	6.83	7.107

Table 4.12: Results for predicting $log(1+mask)$ using SARS, H1N1 and MERS

model in this chapter) to show that the combined effect of masks, social mobility, and NPIs estimated in Figure 4.36 are not appreciably different.

Lasso Regression

Governments worldwide introduced NPIs to enforce social distancing through policies like quarantine, restriction on mass gatherings or closure of schools and businesses. NPIs led to decreased social mobility. For example, there were no major gatherings in railway or bus stations as rails and buses were closed down. In our analysis, NPIs and social mobility across different location types are correlated. This correlation may lead to multicollinearity that may lead to unstable coefficients. We use penalized linear regression

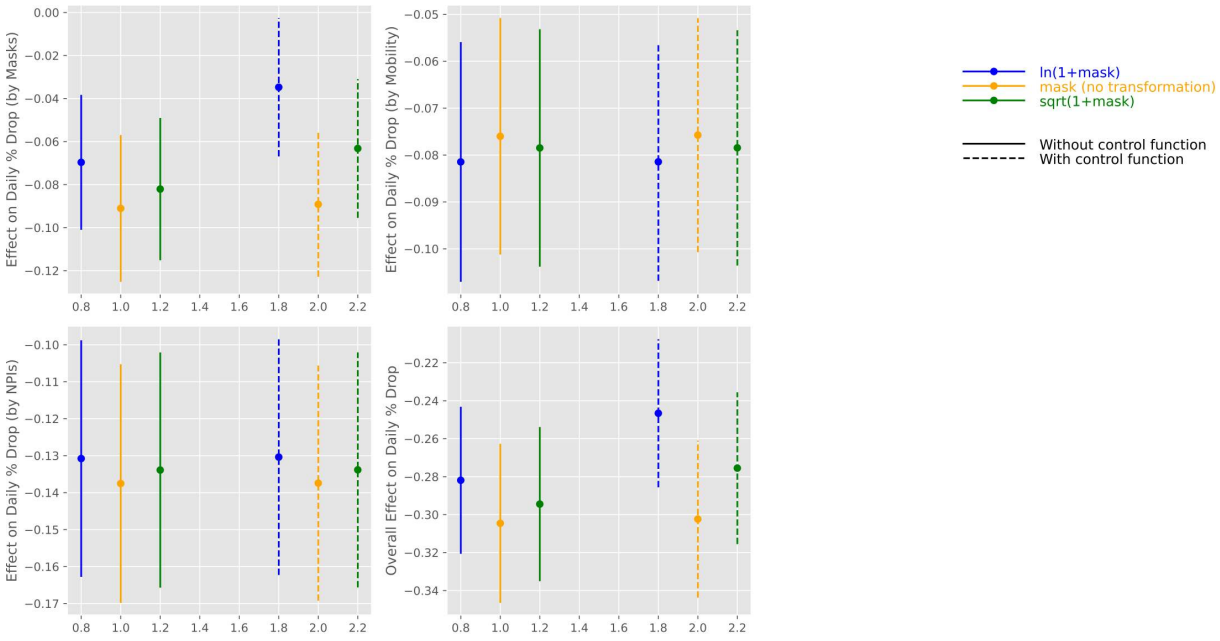


Figure 4.44: Combined effect of mask, social mobility, and NPIs with and without control functions. We use growth rate model with a $shift$ of 9 days.

(Lasso regression) as a robustness check for shrinking the coefficients of highly correlated variates.

Lasso regression can also handle multicollinearity in the data as it shrinks the coefficients to 0 using L1-norm. Lasso regression pushes the coefficients of insignificant variables to 0, thereby introducing sparsity in the model. Our analyses show that the Lasso regression pushes the coefficients for all social mobility indicators to zero except mobility in parks and transit stations. As we observe the correlation between different indicators of social mobility in Table 4.2, Lasso regression provides validation for the selection of two (out of 6) indicators of mobility. Figure 4.45 shows the coefficients for the Lasso growth model. Equation 4.15 represents a Lasso regression model where n is the sample size, β is a vector of coefficients, X is input variables, and Y is the outcome variable. We use 5-fold cross-validation to find λ that best fits the out-of-sample test data.

$$\beta = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{(Y - X\beta)^2}{n} + \lambda \|\beta\|_1 \tag{4.15}$$

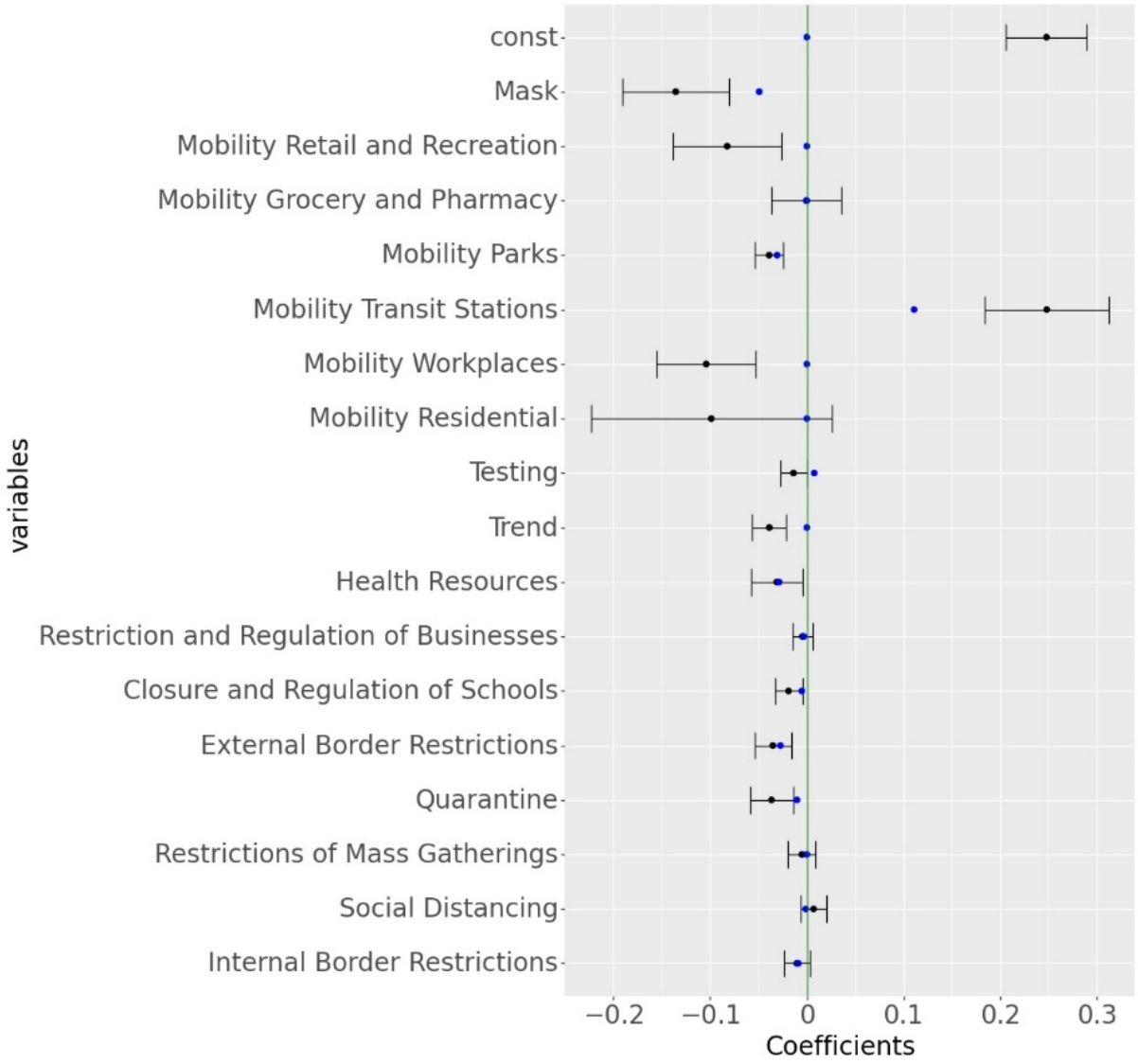


Figure 4.45: Parameter estimation from Ordinary Least Square and Lasso Regression model for growth rate. The blue dot represents the coefficients estimated from the Linear Regression model. The error bars represent the upper and lower confidence interval for the coefficients obtained from Ordinary Least Squares. The blue dot represents the coefficients estimated from the Lasso Regression model.

4.6.4.6 Selecting Period of Analysis

To filter out initial volatile growth rates during the start of the pandemic, we use a threshold th as discussed before. We collect data for up to 60 days for a country, from the day it reaches th percent of peak daily cases in that country. However, the model estimates could be biased and fit the given set of data points. To estimate the robustness

of the model, we estimate the model parameters by collecting data for up to D days from the day that country reaches threshold th. The results for the combined effect of mask, social mobility and NPIs for different $D \in [35, 45, 55, 65, 75, 85]$ is shown in Figure 4.46.

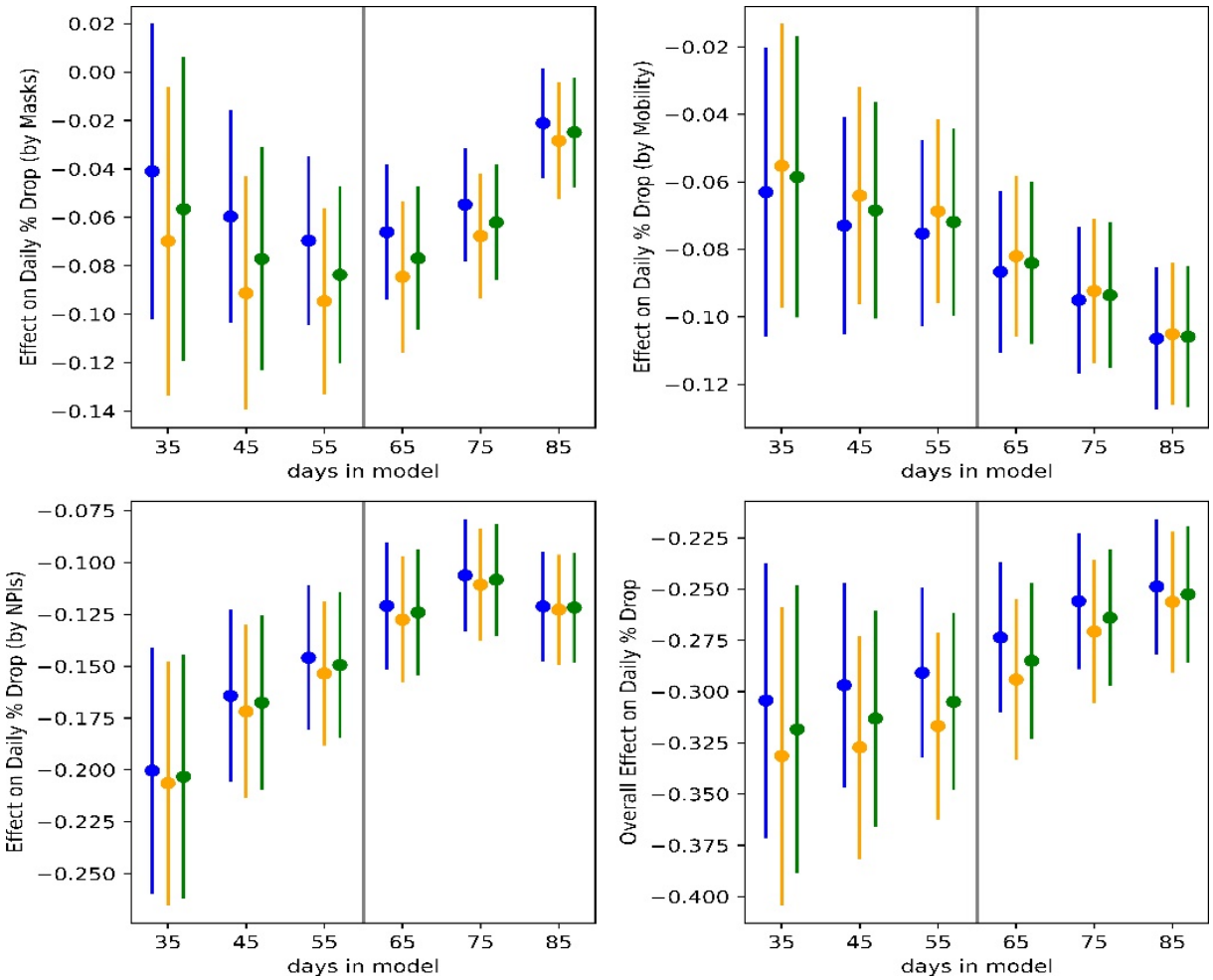


Figure 4.46: Combined effect of mask, social mobility, and NPIs when data is collected for different numbers of days. We consider a shift of 9 days for these results (Note that a shift of 9 days was the best fit for a model that used data for 60 days. The results show consistency within the bounds of the combined effects.

4.6.4.7 Interpolating Mask Survey Numbers Between Survey Days

We use survey data released by the Institute of Global Health Innovation (IGHI) at Imperial College London and YouGov4 for reported mask-wearing across multiple countries. The data present global insights on people's reported behavior in response to COVID-19. The dataset provides the percentage of the population in each country who report

wearing a mask in public places. Because these surveys were conducted at an interval of several days, we used linear interpolation to estimate the percentage of the population that would wear masks in public spaces for days when the data were unavailable (Figure 4.10). To check the robustness of estimates from the model, we use a quadratic interpolation method to estimate the percentage of the population that would wear masks in public spaces for days between surveys. The estimate for stated mask-wearing using quadratic interpolation is shown in Figure 4.47.

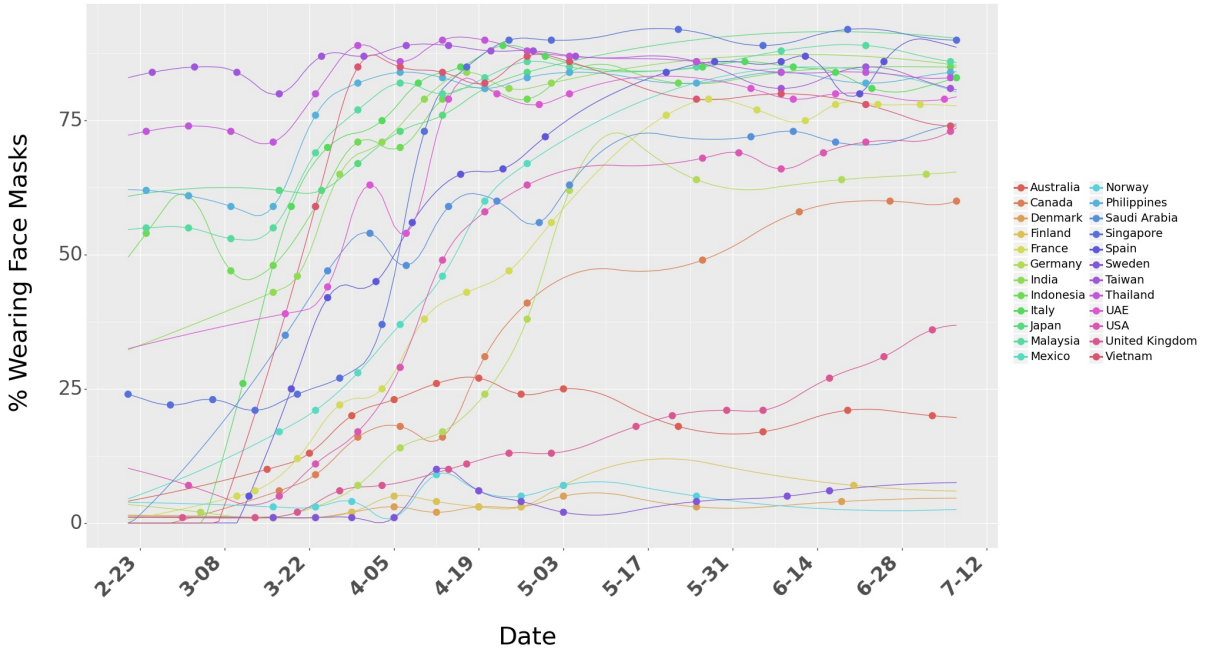


Figure 4.47: Survey data on percentage of people who say they wear a face mask when in public spaces. We use quadratic interpolation to consider mask numbers for days between surveys days. The dots represent the raw numbers from surveys.

Figure 4.48 shows the results for the association of mask, social mobility, NPIs, and the combined effect of masks, social mobility, and NPIs on the growth rate for quadratic interpolation. We use a $shift$ of 9 days and transformed masks as our focal model ($ln(1 + mask)$). We also show results for linear interpolation (the focal model in this chapter) to show that the parameter estimates are not appreciably different.

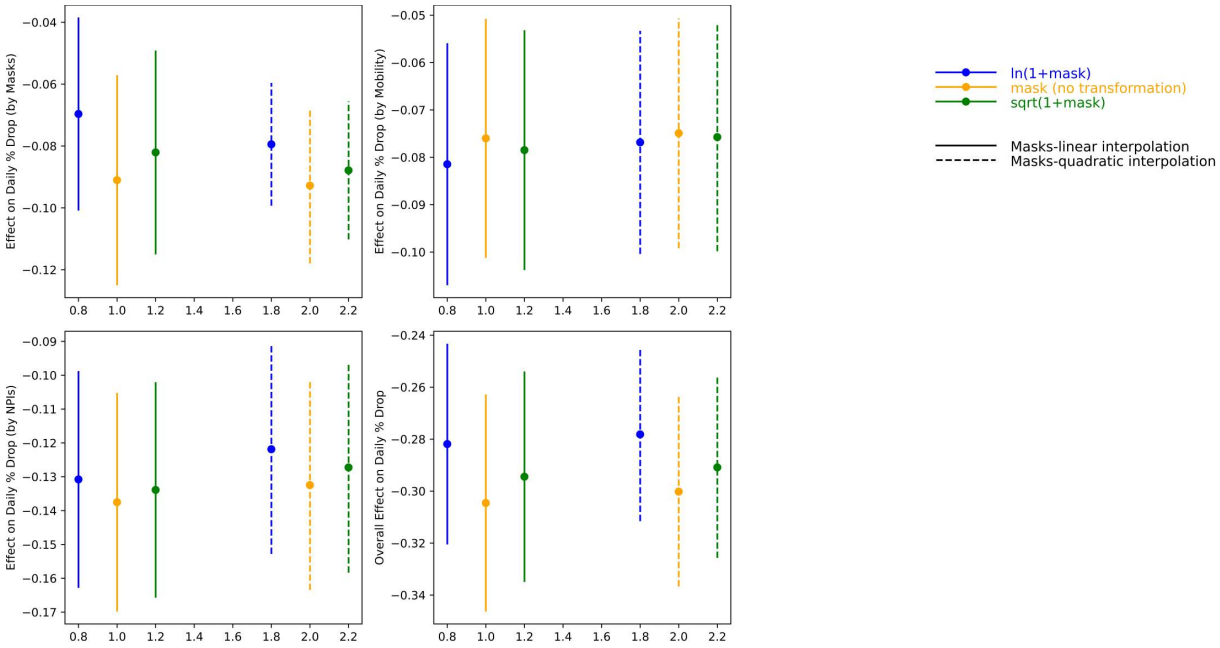


Figure 4.48: Combined effect of mask, social mobility, and NPIs under different interpolation for mask survey numbers. We use growth rate model with a $shift$ of 9 days and consider data for 60 days.

Chapter 5

Interest Meets Pinterest: Identifying Important User Roles in Malicious Content Propagation

Online social networks (OSNs) are platforms that disseminate content widely and rapidly, which makes them a target environment to spread malicious content, e.g., phishing attacks. This chapter studies the detailed characteristics of malicious content propagation. We perform a case study on Pinterest using two large-scale datasets that contain over 2.8 M pins/repins (content in Pinterest) and 2 M users combined. We propose learning-based models to detect whether a posted pin is malicious or not. Our proposed model can be used to protect users from potentially malicious content on Pinterest by preventing the spread of those contents at their creation/posting time. In this chapter, after reviewing related work in Section 5.1, we present our methodology in Section 5.2. In Section 5.3.1, we analyze how malicious pins propagate in Pinterest. We then analyze the characteristics of malicious users in Pinterest in Section 5.3.2. Section 5.3.3 presents our machine-learning-based models for detecting whether a pin is malicious or not.

5.1 Related Work

5.1.1 Information Cascades in OSNs

As OSNs have become one of the most popular spaces to spread content, product information, or news, there have been great efforts in analyzing and modeling information cascades in OSNs [158, 62, 159, 60]. Dow *et al.* discussed the diffusion characteristics of viral content such as memes in Facebook [158]. Cheng *et al.* applied machine-learning techniques for detecting the growth of image cascades in Facebook [62]. Choi *et al.* studied conversation patterns in Reddit and characterized conversation cascades (generated by commenting) in terms of volume, virality, and responsiveness [160]. Rahman *et al.* investigated the adoption cascades of Facebook applications and revealed that the evolutionary properties of cascades (e.g., initial growth rate) are good predictors of the ultimate cascade size [159]. Bakshy *et al.* studied user roles in information cascades in Facebook and showed that weak ties play a substantial role in information dissemination [161]. Zhong *et al.* applied machine-learning models for predicting whether a user will be interested in a given pin [162]. While these studies provide important insight for understanding 'general' information cascades in OSNs, this chapter focuses on 'malicious' content cascades and behaviors of users who disseminate such malicious content in OSNs by comparing malicious to benign (or general) content cascades.

5.1.2 Malicious Activities in OSNs

With the increasing threat of malicious users who seek to disseminate unwanted spam, phishing pages, or malware downloads, there have been significant efforts to understand how such threats spread in OSNs [163, 164, 165, 166, 67, 167]. Wu *et al.* [163] investigated the spamming activities in Twitter and classified those activities using machine learning techniques. Similarly, phishing patterns were also analyzed in Twitter [164]. Gong *et al.* [166] proposed DeepScan, a malicious account detection system in location-based OSNs. Stringhini *et al.* [65] developed a spammer detector based on the anomalous behavior of users in Twitter. Lee *et al.* [165] developed an algorithmic spammer account detector based on the account names. Cao *et al.* [167] utilized forwarding-based features

along with graph-based features for detecting Malicious URLs on Sina Weibo. Lee *et al.* developed a system for detecting suspicious URLs in Twitter called WARNINGBIRD [67]. While these works revealed critical insights into malicious activities in OSNs, little attention has been paid to understanding how different types of malicious content (e.g., spam, phishing, malware) propagate and how users who spread such malicious content show distinctive characteristics in Pinterest.

5.1.3 Pinterest - a Content-centric and Interest-driven OSN

There have been many works focused on understanding user behavior on Pinterest. One of the interesting properties of Pinterest is its *content-centric* and *interest-driven* nature [68, 70, 69, 63, 168, 169]. Han *et al.* established that pin propagation is mostly driven by pin's properties (e.g., its topic), not by user's characteristics (e.g., his/her number of followers) [60]. Forte *et al.* showed a survey result that Pinterest is thought of as an OSN about 'what' users enjoy, not about 'who' they are [68]. Gelley and John revealed that 'following' is not a factor in sharing pins in Pinterest [70], suggesting the interest-driven nature of Pinterest. Han *et al.* focused on 'topic' (e.g., humor, education) in Pinterest and investigated the differences in pinning and repinning behaviors on different topics [69]. Lo *et al.* studied factors that affect the growth in content collection over a long period in Pinterest [168]. Han *et al.* [170] proposed machine-learning models for predicting popular and viral images in Pinterest [63]. Hu *et al.* [170] developed a model for predicting image diffusion paths in Pinterest. Mittal *et al.* [169] performed a characterization of users and contents and used the derived features for predicting the gender of American users. Our work focuses on how 'malicious' content is shared and how it can be detected in an interest-driven OSN.

5.2 Research Methodology

This section describes the dataset used in this chapter and our pin cascade model. We then describe how we identify malicious pins for our analysis.

5.2.1 Dataset

We analyze a detailed dataset [60] that had been collected from Pinterest for the duration of 44 days from June 5 to July 18, 2013. The dataset consists of content information (e.g., their categories, sources, pinners) of 337,345 (original) pins and 1,179,092 repins. The number of users associated with those pins is 915,411, collectively having 161,309,627 followers and 511,737,156 followings. At the time of data collection, each pin belongs to one of the 33 Pinterest-defined categories such as 'sports', 'education', or 'women's fashion'.

A second dataset to validate our analysis is collected from Pinterest over the 12 days from September 21 to October 2 in 2017. Similarly, the dataset consists of content and user information. However, some of the previously accessible metadata were not available anymore at the moment of data collection. This missing information includes the chain of who repins from which users. As a result, there is no distinction between original pins and their repins in the second dataset. Also, Pinterest deactivated the like button in 2017. The rest of the metadata for the pins and the users who posted the pins were still available at the time of data collection. We show that the currently missing metadata has no impact on malicious content detection (See Section 5.3.3 for details). The second dataset consists of 1,287,158 pins with 1,110,703 associated users. The users collectively have 286,202,687 followers and 630,772,029 followings at the time of data collection. As of October 2017, Pinterest provides 32 categories with similar names of categories shown in the year of 2013, with an exception of the category 'products'. Note that we mostly report our analysis on the first dataset (2013) as the results are similar for the second one.

5.2.2 Pin Cascade Model

We define a pin cascade model as a directed graph, $G = (V, E)$, where V is the set of users (nodes) and E is the set of repins (edges). A pinner who posts original content (e.g., image, video, or URL) into his/her board is the root of a pin cascade. If user j repins the pin from user i, an edge $E(V_i, V_j)$ exists from user V_i to user V_j. In this way, a pin can be propagated multiple hops, e.g., from user i to user j to user k. Therefore, each pin

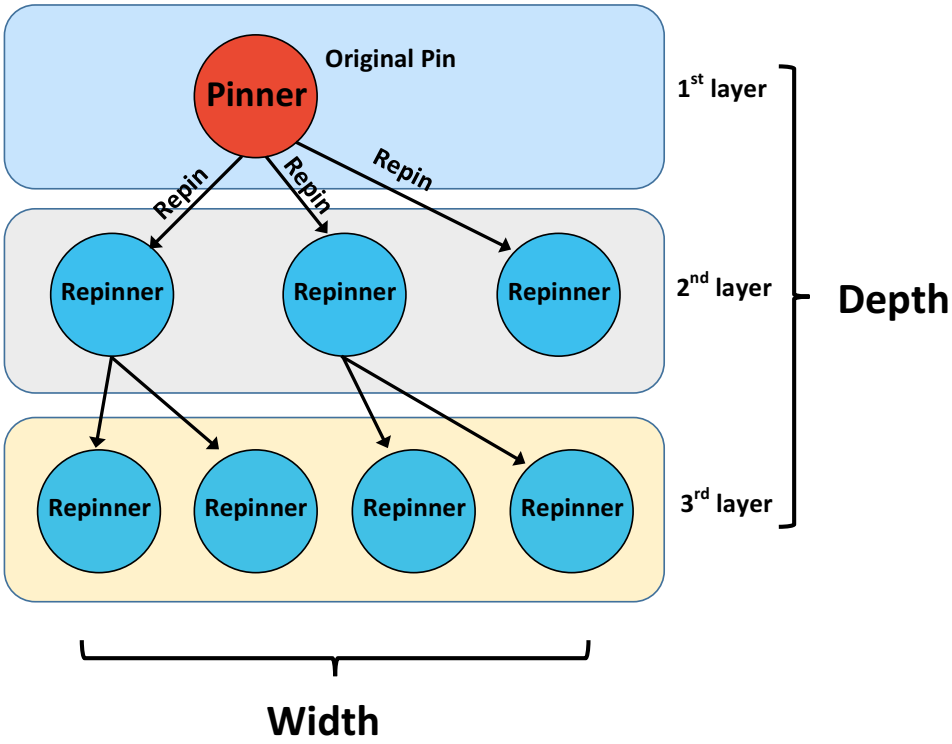


Figure 5.1: An example of pin cascade model, which consists of an original pinner, followed by multiple repinners. The size, depth, and width of the pin cascade in this example are 8, 2, and 4, respectively.

cascade includes a pinner, followed by some repinners, forming a 'pin cascade' together. Note that every pin cascade belongs to a certain category.

Figure 5.1 shows an example of a pin cascade model. In this model, we define four properties that can describe the structural pattern of the cascade: (i) size, (ii) depth, (iii) width, and (iv) inter-repin time. The size of a cascade is the total number of users in the cascade. The cascade depth is the longest distance of the nodes from the root. The width is the maximum number of nodes that have the same distance from the root (i.e., in the same layer in Figure 5.1). The inter-repin time is defined as the time difference between the root's pinning and his/her repinners' repins or two consecutive repins, which quantifies how quickly the corresponding content propagates, or in other words, how quickly repinners respond to the content. Since a pin cascade can consist of multiple inter-repin times, we measure (i) the average inter-repin time of a cascade and (ii) the first inter-repin time of a cascade, which is the time difference between pinning and the

Malicious Types	# of pins	# of repins	# of users
Malicious Downloads	897	4,778	5,440
Malicious Sites	567	2,417	2,917
Spams	682	1,584	1,832
Phishing	43	48	75

Table 5.1: Four attack types of malicious content in Pinterest: malicious downloads, malicious sites, spams, and phishing.

first repinning.

5.2.3 Identifying Malicious Pins

To identify whether pins are malicious or not, we use a group of URL scanner engines: (i) The commercial McAfee URL scan engine [171] and (ii) Virustotal [72] which consists of 64 different antivirus engines and website scanners in 2017[1]. All the source URLs of 337,345 original pins (and, similarly, source URLs of all the pins in the dataset collected in 2017) are scanned by the engines, which generates detailed information about each URL's risk level and its type. Note that each pin has a source URL that the content belongs to. For example, if a pinned content is an image, Pinterest provides a source URL to a website that originally has the image. In order to reduce the false-positive rate, a URL is considered to be malicious if it has been detected as malicious content by at least two different engines. Out of 337,345 pins, 2005 pins were identified as malicious, which accounts for 0.59% of the total pins. These malicious pins, posted by 1,489 users, generate 8,111 repins by 7,775 users. Similarly, in the 2017 dataset, 12,138 pins/repins generated by 10,968 users were detected malicious, which accounts for 0.96% of total pins/repins in the dataset.

The identified malicious pins are further categorized into four different attack types by the McAfee engine [171][2]:

- **Malicious Downloads** (44% of all malicious content in Pinterest): This attack

[1]The dataset collected in 2017 has been only scanned by 64 engines on Virustotal.

[2]Virustotal provides a similar categorization into different types like 'Malware', 'Phishing', etc. We use McAfee categorization as the reference since it provides more detail.

type deceives users to download codes. e.g., toolbars or screensavers, resulting in the installation of adware or spyware that can steal passwords or damage files.

- **Malicious Sites** (28%): The malicious sites deploy codes to hijack a computer's settings or activity, e.g., Trojan horses.

- **Spam** (34%): These refer to a wide range of spam URLs leading to offensive and fraudulent websites and unwanted product marketing.

- **Phishing** (2.1%): This attack type obtains sensitive information of the user by acting as legal websites.

Table 5.1 summarizes the basic information of the four attack types. In practice, it is typically not feasible to scan a substantial number of URLs through a commercial engine due to the expense and time required. In addition, most detection engines rely on signature-based detectors, which will not be able to detect previously unseen threats. Therefore, we will introduce a detection model for identifying malicious content without such engines, using only statistical features of content and users in Pinterest. The identified data by the McAfee engine and Virustotal will be used as ground truth to evaluate the proposed model.

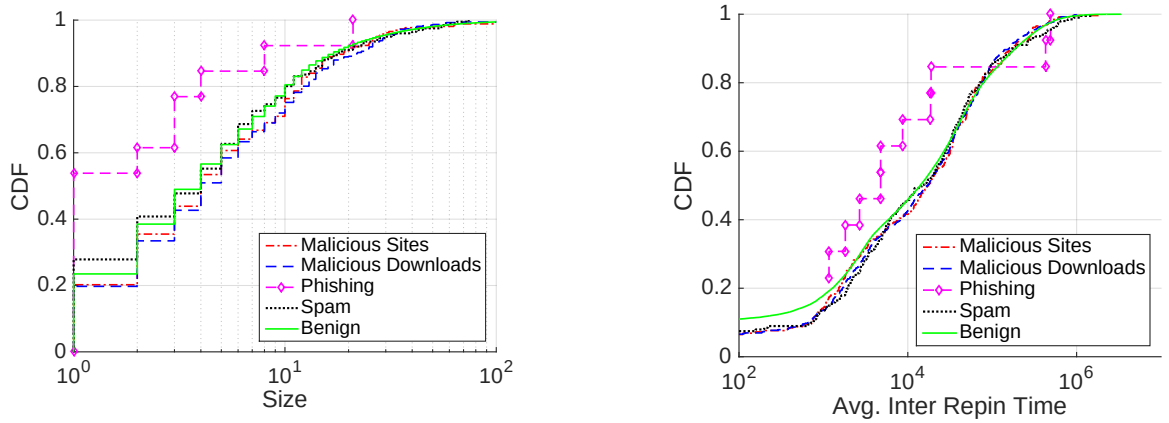
5.3 Results

5.3.1 Malicious Content Propagation

In this section, we investigate malicious content propagation from the following three perspectives: (i) how malicious pins propagate, (ii) what categories malicious pins belong to, and (iii) what websites malicious pins come from. To further explore malicious content, we analyze four attack types: malicious downloads, malicious sites, spam, and phishing.

5.3.1.1 Structural Analysis

We first investigate the structural characteristics of pin cascades generated by malicious and benign content in Figure 5.2. Since the distribution of depth and width are almost similar to that of size, we only report the size distribution in Figure 5.2a. The distribution



(a) Size

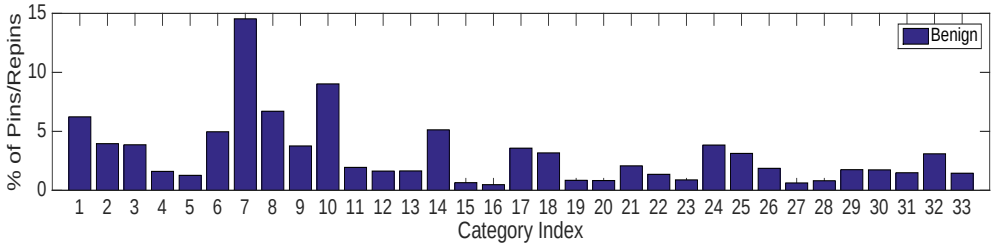
(b) Avg. inter-repin time

Figure 5.2: Structural characteristics of pin cascades: malicious vs. benign.

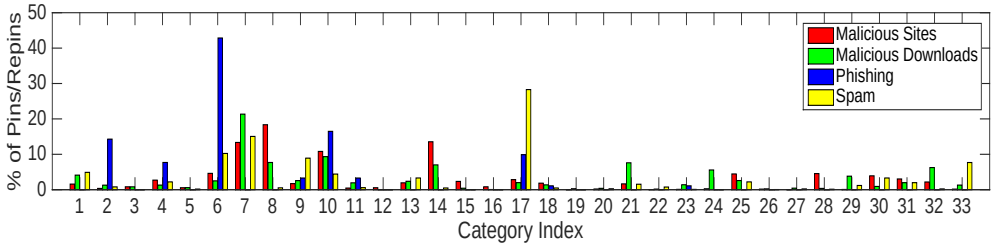
of the average inter-repin time is plotted in Figure 5.2b as the distribution of the first-repin time shows a similar pattern. Interestingly, we find that cascade sizes of malicious content are almost similar to those of benign content except for the phishing type. This indicates that malicious and benign content propagate similarly among Pinterest users. Phishing attacks tend to propagate less, which may be due to the increased awareness of phishing attacks. When we look at Figure 5.2b, we find that average inter-repin and first-repin times of malicious content are almost similar to those of benign ones except for the phishing type. Pins containing phishing attacks tend to spread more quickly than others but among a smaller set of users. Note that this structural information is only available in the dataset collected in 2013. The dataset collected in 2017 only includes the size information of cascades. Fortunately, the absence of such information has no significant impact on detecting malicious content (see Section 5.3.3 for details).

5.3.1.2 Category Analysis

We next investigate how users post/share malicious and benign pins across different categories on Pinterest. Recall that Pinterest provides 33 categories (as shown in Table 5.2), and each pin belongs to one of the 33 categories. Figure 5.3a shows the distribution of benign pins across the different categories for the 2013 dataset. Different categories exhibit different popularity. For example, 'DIY & crafts (Category Index (CI) 7)' accounts for almost 15% of the entire pins while 'holidays & events (CI 16)' is only responsible for



(a) Benign



(b) Malicious

Figure 5.3: Percentages of pins in each category for benign content and four malicious attack types (i.e., malicious downloads, malicious sites, spam, and phishing).

1% of all the pins in Pinterest. The top 3 popular categories are 'DIY & crafts (CI 7)', 'food & drink (CI 10)' and 'education (CI 8)'. Interestingly, in the year 2017 dataset, there is a similar pattern in pinning behavior. The top 3 categories for that dataset are 'DIY', 'Food', and 'Humor'.

1	animals	8	education	15	history	22	outdoors	29	technology
2	architecture	9	film/music/books	16	holidays & events	23	photography	30	travel
3	art	10	food & drink	17	home & decor	24	products	31	weddings
4	cars & motorcycles	11	gardening	18	humor	25	quotes	32	women's fashion
5	celebrities	12	geek	19	illustrations & posters	26	science & nature	33	shop
6	design	13	hair & beauty	20	kids	27	sports		
7	DIY & crafts	14	health & fitness	21	men's fashion	28	tattoos		

Table 5.2: 33 Pinterest categories with their indexes.

Figure 5.3b shows the percentages of pins in each category for four attack types (i.e., malicious downloads, malicious sites, spam, and phishing), respectively. Interestingly, the attack type of malicious downloads shows a similar pattern to the benign case as shown in Figure 5.3b. The top 3 categories that contain pins of attack type malicious downloads are 'DIY & crafts (CI 7)', 'food & drink (CI 10)' and 'education (CI 8)', which

is identical to the top 3 categories for benign pins. On the other hand, the other attack types show substantially different patterns from benign cases. The 'education (CI 8)' is the most popular category (accounting for 20% of the associated pins) for the attack type associated with malicious sites. Spamming and phishing attacks show more distinctive patterns. The malicious pins for spamming are usually shared in 'home & decor (CI 17)', where it is the main target by spammers who seek to spread their products on Pinterest. The malicious pins for phishing are mostly shared in 'design (CI 6)', which is not so popular in benign cases. One possible explanation is that these phishing attacks may be targeting Pinterest users who are interested in luxury designs and products. Our findings imply that users who seek to distribute malicious content on Pinterest are likely to utilize different categories for crafting different types of attacks.

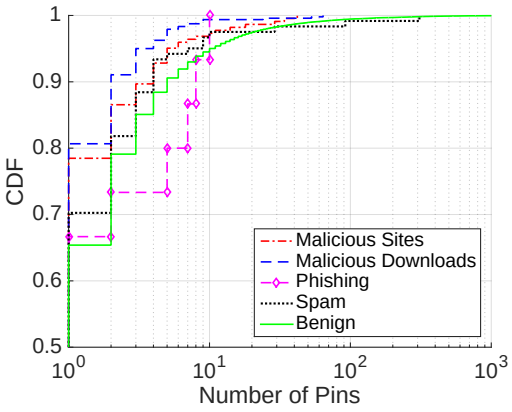
Similarly, in the dataset collected in 2017, malicious pins are distributed unbalanced amongst different categories compared to benign pins. The top 3 categories for the malicious pins are 'Art', 'DIY', and 'Travel'.

5.3.1.3 Domain Analysis

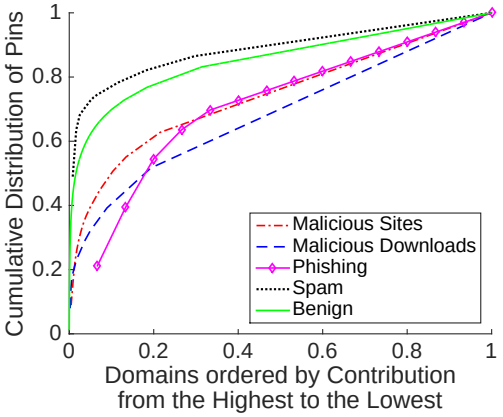
We also investigate where (or from what websites) Pinterest content originates from by analyzing 'sources' at a domain level. We aggregate all the source URLs of pins into domain-level URLs; e.g., `IMDB.COM/aa`, `IMDB.COM/bb`, and `IMDB.COM/cc` are aggregated into `IMDB.COM`. Here, a domain can be interpreted as a content provider; e.g., for the pins from `IMDB.COM`, `IMDB.COM` can be regarded as a content provider for those pins.

We first investigate the number of posted pins for each domain in Figure 5.4a. Overall, only one pin is fetched from over 60% of all the domains. We observe that a benign domain tends to have a higher number of pins than a malicious domain except the phishing, implying that malicious domains often provide a small number of pins to Pinterest compared to benign domains. Interestingly, the domains for phishing show the highest number of pins, which implies that the websites for phishing attacks tend to provide more content to Pinterest.

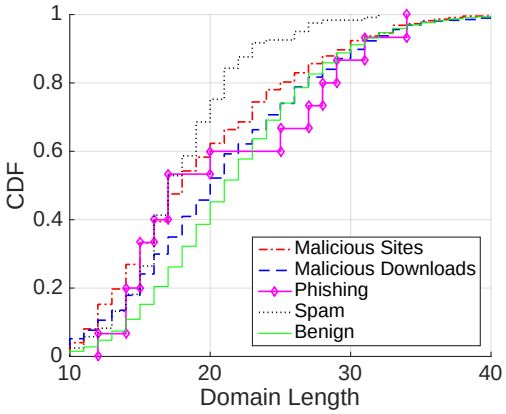
To investigate how different domains contribute to spreading pins in Pinterest, we plot the cumulative contributions of domains in terms of the number of pins, sorted in



(a) Number of Pins



(b) Domain Contribution



(c) Domain Length

Figure 5.4: Domain characteristics of pin cascades: malicious vs. benign.

descending order, starting from domains with the highest contribution in Figure 5.4b. In other words, the horizontal axis in Figure 5.4b represents the domains from the highest contribution to the lowest, and the vertical axis represents the (cumulative) fraction of pins that belong to those domains.

As shown in Figure 5.4b, domains for benign and malicious content show significantly different patterns. Overall, the domains for malicious content (except the spam) exhibit more even contributions than the domains for benign content. This indicates that a small set of domains are trendy for benign content, contributing significantly to providing content in Pinterest. Note that the top 20% domains (for benign content) contribute to almost 80% of benign pins in Pinterest.

On the other hand, the domains for the three attack types (i.e., malicious downloads, malicious sites, and phishing) contribute evenly to pins in Pinterest. Meanwhile, the domains associated with spamming attacks exhibit a very different pattern than other attack types; a small set of domains for spamming contribute significantly to malicious content in Pinterest. Note that one domain (i.e., `RAWINSPRE.COM`) is responsible for almost half of the spamming content in Pinterest.

We next explore the structural pattern of each domain in terms of length and domain digit ratio. The domain length is the number of letters and characters used in the domain name. For instance, the domain length of `WWW.GOOGLE.COM` is 14. The domain digit ratio for a domain is the number of numerical characters (e.g., 1, 2) used in its domain name divided by its domain length. For example, since the number of numerical characters and domain length for the domain `D97COOLTOOLS.BLOGSPOT.CA` are 2 (i.e., 9 and 7) and 24, respectively, the domain digit ratio would be $\frac{2}{24} \approx 0.083$.

Figure 5.4c shows the distribution of domain length for benign and malicious domains. Overall, the malicious domains tend to have a shorter length than the benign domains except the phishing. While benign domains are usually human-understandable, e.g., `DESIGNSPIRATION.NET`, many malicious domains often use short and meaningless names (e.g., `rtm.cc` or `atmr.ch`). When we look at domain digit ratio, malicious domains tend to use a slightly higher domain digit ratio than benign ones, but their difference is marginal; hence we do not show the figure for domain digit ratio.

5.3.2 Activities of Malicious Users

This section first describes how we identify 'malicious users' on Pinterest. We then analyze the activities of such users on Pinterest.

5.3.2.1 Identifying Malicious Users

We define a malicious user as any user who posts/pins or repins at least one malicious pin on Pinterest. Note that posting a pin usually requires a substantial effort in Pinterest [69]. Thus, we assume that users who post the original malicious content do so intentionally. We acknowledge that users who repin malicious content may be oblivious victims themselves. Nevertheless, their actions still pose potential risks to other users.

In the dataset collected in 2013, out of 915,411 users, 1.01% of users are tagged as malicious. In the dataset collected in 2017, 0.98% of users out of 1,110,703 users are malicious. We also find that 83% of malicious users post only one malicious pin during the data collection period.

5.3.2.2 Characteristics of Malicious Users

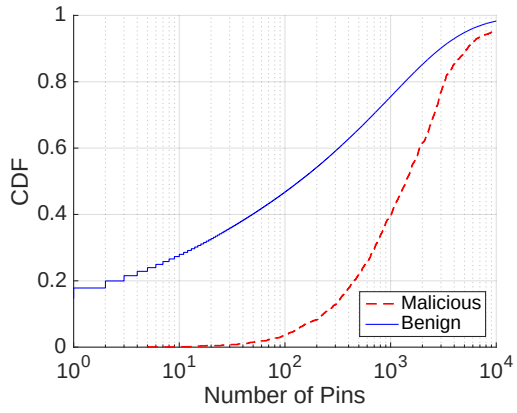
We now analyze the characteristics of malicious users in terms of the following: (i) number of pins that they post, (ii) number of likes that they give to others' pins (note: this data is not available in the 2017 dataset), (iii) number of individuals that they follow, (iv) number of followers who follow them, (v) category entropies that quantify how their pins distribute across different categories, and (vi) domain entropies that show how evenly they fetch pins from multiple domains. We compare such characteristics between malicious users and benign (non-malicious) users in Figure 5.5.

As shown in Figure 5.5a, the number of pins by malicious users is much higher than those by benign users, which implies that malicious users tend to be more active than benign users in Pinterest. They tend to post pins more often compared to benign users. Note that 60% of malicious users have more than 1,000 pins while 60% of benign users have more than 60 pins, which shows a two-order-of-magnitude difference. Similarly, as shown in Figure 5.5b, malicious users tend to give a higher number of likes to others' pins than benign users. When we look at following/follower relationships in Figures 5.5c and 5.5d, we find that malicious users tend to follow more individuals and also have more followers than benign users. This implies that malicious users tend to have active connectivity with Pinterest users. Interestingly, malicious users are likely to have a much larger number of followings than benign users, implying that malicious users seek to connect other users for their purposes.

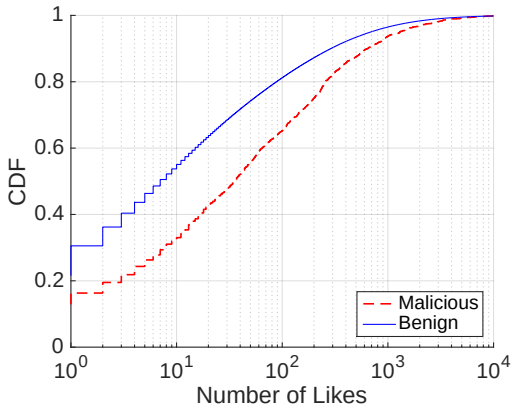
We next investigate the category entropies of malicious and benign users in Figure 5.5e. Here, the normalized category entropy of user U is defined as:

$$H_{category}(U) = -\sum_{i=1}^{C_U} \frac{p_i^U ln(p_i^U)}{ln(C_U)} \tag{5.1}$$

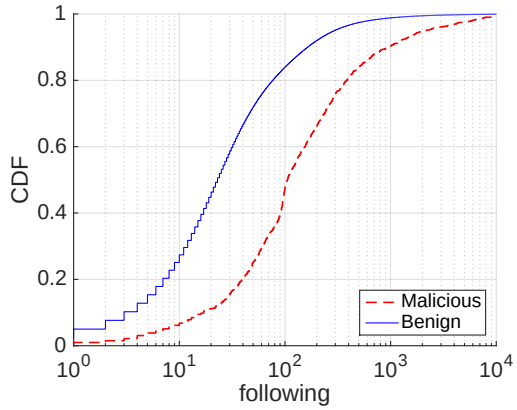
where C_U is the total number of categories that user U has pinned/repinned in, and p_i^U is



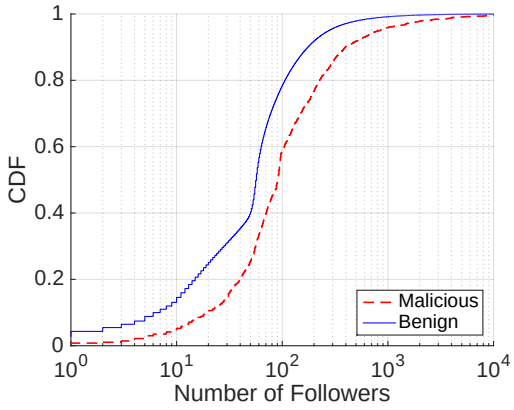
(a) Number of pins



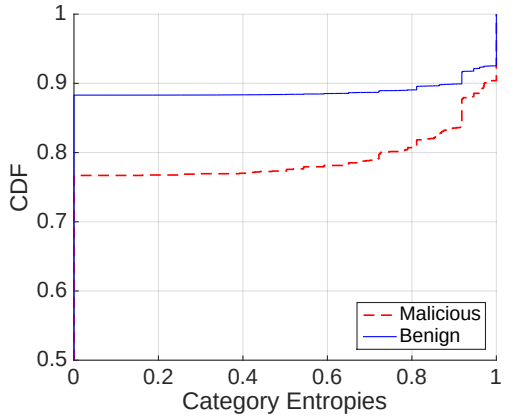
(b) Number of likes



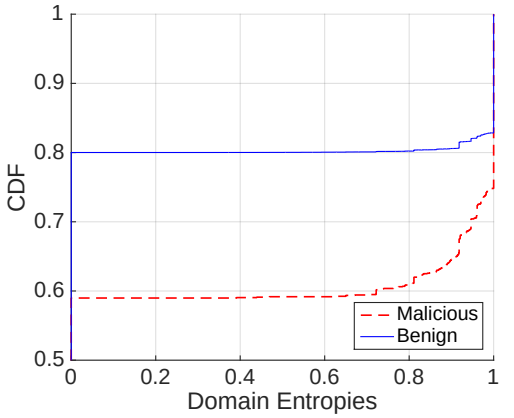
(c) Number of followings



(d) Number of followers



(e) Category entropies



(f) Domain entropies

Figure 5.5: Characteristics of malicious users in terms of their number of pins, likes, followings, followers, category, and domain entropies.

the portion of pins/repins of user U in category i. If a user has pins only in a particular category, the category entropy of him/her is zero. On the other hand, for a user with evenly distributed pins across multiple categories, the category entropy equals one. As shown in Figure 5.5e, category entropies of malicious users are higher than those of benign users, which implies that malicious users tend to have pins in diverse categories (hence possibly spreading their malicious pins widely), whereas benign users are likely to have pins in few numbers of their favorite categories.

To investigate how evenly a user fetches pins from multiple domains, we calculate the normalized domain entropy, $H_{domain}(U)$ for user U in a similar way to Equation 5.1. As shown in Figure 5.5f, domain entropies of malicious users are higher than those of benign users, and the gap between malicious and benign users in terms of domain entropies is even wider than the one of category entropies. Over 40% of malicious users tend to almost equally fetch content from different domains, whereas less than 20% of benign users tend to fetch content from different domains evenly. This implies that malicious users may want to deceive other users using various domains, while benign users often fetch content from their particular preferable domains. We observe similar patterns in the dataset collected in 2017.

5.3.3 Prediction on Malicious Content

Our observations on the content properties of malicious pins and the characteristics of malicious users who post such pins suggest that there exists a set of distinctive features that can be combined to detect whether a given pin is malicious or not. This section seeks to answer the following question: How can we detect previously unknown malicious pins without expensive scanning techniques? Leveraging the insight gained by our analysis, we propose a machine-learning-based model to identify malicious pins by observing the relevant features of the pin as well as their pinners. Such a model has important implications both for users and OSN service providers.

5.3.3.1 Problem Definition

Our goal is to identify whether a posted pin is malicious or not based on the observed features (as described below). To this end, we define two random variables: (i)

$X = (X_1, X_2, ..., X_n)$ is a set of features where X_i indicates the $i - th$ feature of the corresponding pin, and (ii) Y is a target variable where Y is 1 if the pin is malicious, and 0 otherwise. We use the Virustotal and Mcafee scan data described earlier as our ground truth.

To detect whether the posted pin is malicious or not, we extract a set of distinct features associated with content properties (discussed in Section 5.3.1) and pinner's characteristics (described in Section 5.3.2). Before building the classification model, we need to identify the relevant feature set that will contribute to the highest accuracy in malicious content prediction. Our set of selected features include:

- **Structural:** features are associated with properties of the cascade: cascade size, width, depth, average inter-repin time, and first inter-repin time.
- **Content:** features are associated with properties of the pin itself: number of likes, pin's category, number of pins associated with the corresponding domain, domain length, source length, and domain digit ratio.
- **User:** features are associated with properties of pinner of the content: number of pins, number of likes, number of following/followers, category entropy, and domain entropy.
- **All:** considers all **Structural**, **Content**, and **User** features.

5.3.3.2 Feature Selection

In order to find the proper subset of features for a classifier, redundant features need to be filtered. One of the popular filtering techniques is excluding highly cross-correlated features. We calculate the pairwise Pearson correlation between every pair of features. The higher correlation shows the redundancy of such a pair of features together. Table 5.3 includes highly correlated pairs of features.

We found that structural properties were shown to be less important features. Interestingly, domain and category entropies are also highly correlated (correlation = 0.534). Intuitively, every domain dedicates most of its content to a few specific categories, and

Feature Pair	Correlation
size – depth	0.600
size – width	0.941
depth – width	0.505
average repin time – first repin time	0.531
domain entropy – category entropy	0.534

Table 5.3: Highly correlated pairs of features (correlation > 0.5).

thus a high correlation between category and domain entropies can be expected. For every pair of features in Table 5.3, we choose the feature that has the highest correlation to the output class (malicious/benign).

To select the best subset of features out of remaining attributes, we perform the multi-variate (i) CFS (Correlation Feature Subset) [172] and (ii) RFE (Recursive Feature Elimination) [173] evaluations. In the CFS method, instead of a single pairwise correlation, we try to find the subset of attributes that has a high class-feature correlation, while minimizing the feature-feature correlation within the subset. The metric used in this method is

$$CFS = \max_{S_k} \frac{k\overline{r_{cf}}}{\sqrt{(k + k(k-1)\overline{r_{ff}})}} \tag{5.2}$$

where $\overline{r_{ff}}$ and $\overline{r_{cf}}$ are average feature-feature and class-feature correlation of all the attributes in the subset S_k, respectively. k is the number of features selected in the subset S_k and max is taken over all possible subsets of features. The optimal subset of features found in this method includes the 7 following features:

$S_* = \{$likes, category, domain length, domain count, source length, category entropy, domain digit ratio$\}$

where six out of seven selected features are content ones, and the only user feature selected in this method is 'category entropy'. Note that no structural attribute is selected in this method, as is expected from Section 5.3.1. The RFE method is an embedded technique that builds a classification model repeatedly and recursively eliminates features that have

less impact on the prediction accuracy. Even though this method is computationally complex compared to other filtering methods, it can accurately capture the importance of subsets of features.

Figure 5.6 shows the accuracy of the model using the best subset of features selected via RFE. In this analysis, we use the Random Forest model [174] as the classifier. We perform various models over the same dataset to achieve the best accuracy for malicious content detection. The Random Forest classifier is proven to be the highest accurate model by the AUC (Area Under the ROC (Receiver Operating Characteristic) Curve) measure [175]. As shown in Figure 5.6, the optimal subset size for our model is nine; however, the selected subset of the size seven shows a comparable accuracy. Thus, we choose to use seven features in our model. The selected subset with seven features is:

S_* = {likes, category, domain length, domain count, source length, cascade size, domain digit ratio}.

Note that the feature set selected by RFE has one different feature from the feature set selected by the CFS evaluation; 'cascade size' substitutes 'category entropy'. As two feature sets have the same size, we can compare the prediction performance of the two approaches mentioned above.

5.3.3.3 Addressing the Class Imbalance Problem and Performance Metrics

We build a classification model using the Random Forest ensemble learning algorithm [174]. To address the class-imbalance issue, we apply the SMOTE (Synthetic Minority Over-sampling TEchnique) [176], which allows us to learn with over-sampled instances from the minority class (i.e., malicious pins in our case). We randomly under-sampled instances from the majority class (i.e., benign pins in our case). Note that the numbers of benign and malicious pins are 335,340 and 2,005, respectively.

We report various performance metrics including precision, true positive rate (TPR), false positive rate (FPR), F-measure, and AUC [175]. Note that AUC implies the effectiveness of a prediction model; a perfect model has $AUC = 1$. To ensure the consistency of our model, we perform 10-fold cross-validation. A sensitivity analysis shows that a model with the Random Forest classifier consisting of 100 trees and their depths as 5 is

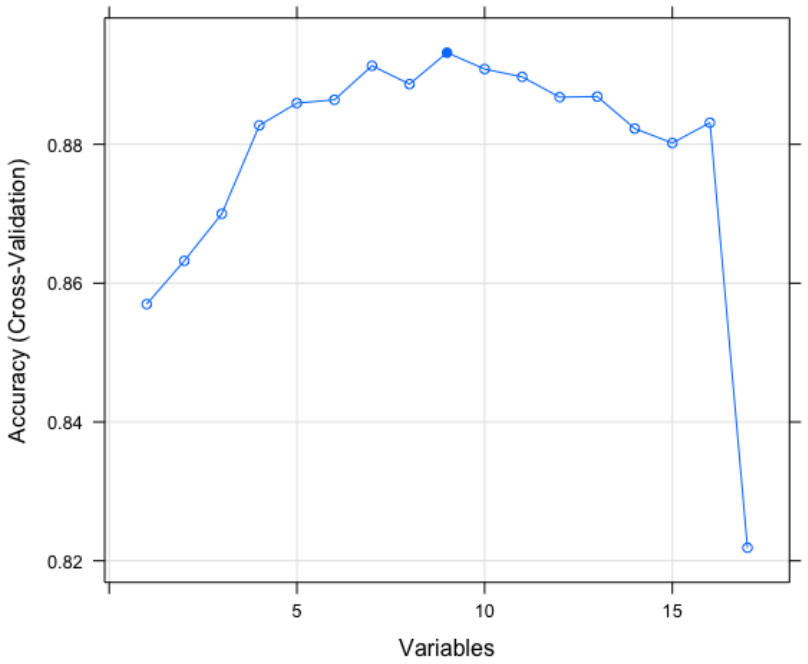


Figure 5.6: Accuracy vs. feature subset size.

Method	Class	Precision	TPR	FPR	F-Measure	AUC
RFE	Benign	0.881	0.855	0.115	0.868	0.949
	Malicious	0.859	0.885	0.145	0.872	0.949
	Weighted Avg.	0.871	0.870	0.130	0.870	0.949
CFS	Benign	0.799	0.811	0.204	0.805	0.934
	Malicious	0.808	0.796	0.189	0.802	0.934
	Weighted Avg.	0.804	0.804	0.197	0.804	0.934

Table 5.4: Prediction results using the top ranked features selected by RFE and CFS methods, respectively.

sufficient to achieve high performance.

5.3.3.4 Performance Results

Table 5.4 first represents the performance results using the top 7 features selected by RFE and CFS methods, respectively, based on the dataset collected in 2013. We find that each method also shows a similar level of performance with the dataset collected in 2017. We find that the selected subset by the RFE method performs marginally better than the

Features	Class	Precision	TPR	FPR	F-Measure	AUC
S_*	Benign	0.881	0.855	0.115	0.868	0.949
	Malicious	0.859	0.885	0.145	0.872	0.949
Content	benign	0.850	0.854	0.151	0.852	0.942
	malicious	0.853	0.849	0.146	0.851	0.942
User	Benign	0.724	0.827	0.316	0.772	0.828
	Malicious	0.798	0.684	0.173	0.737	0.828
All	Benign	0.845	0.853	0.156	0.849	0.932
	Malicious	0.852	0.844	0.147	0.848	0.932

Table 5.5: Performance results of the models based on S_*, content, user, and all features with sampling ratio as 1:1.

Features	Class	Precision	TPR	FPR	F-Measure	AUC
Content	benign	0.816	0.951	0.214	0.878	0.949
	malicious	0.941	0.786	0.049	0.857	0.949

Table 5.6: Performance results of the models based on content features with sampling ratio 1:1 with the dataset collected in 2017.

ones by the CFS method.

Table 5.5 next summarizes the results of the models based on optimal feature subset, S_*, as well as content, user, and all features for a comparison purpose. We exclude the results using structural features as they were not accurate enough. Remind that the models learn the same number of instances from the minority and majority classes, i.e., the sampling ratio is 1:1. We report the results both for benign and malicious classes. We find that the model based on content features performs better than the model based on user features (AUC: 0.949 vs. 0.828). This implies that content features are more significant predictors than user features in detecting malicious content. Interestingly,

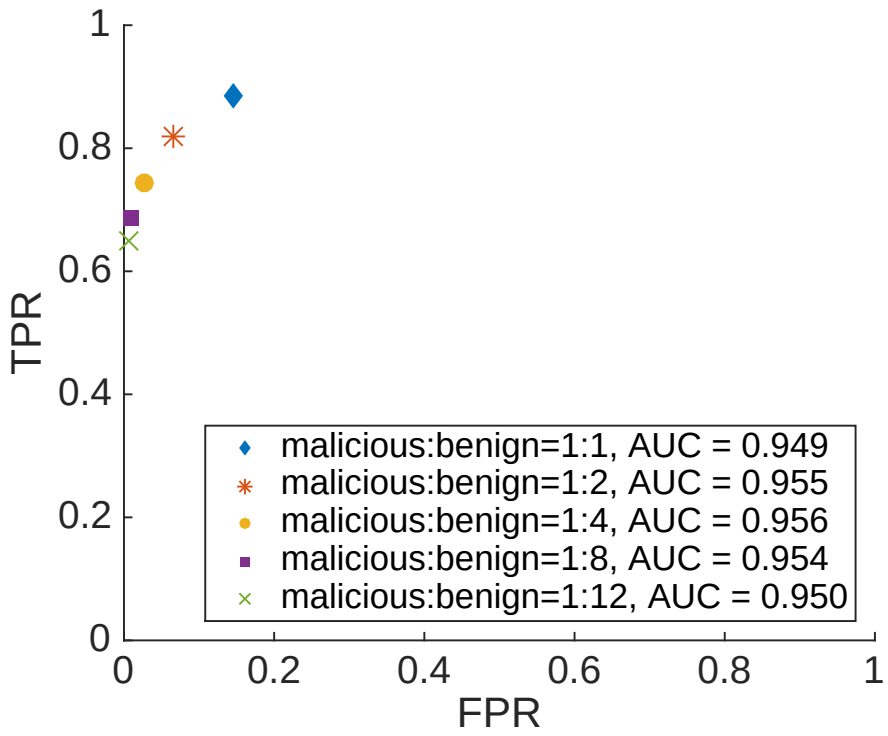


Figure 5.7: ROC for different sampling ratios from 1:1 to 1:12 in detecting malicious content.

the model based on both content and user features (i.e., 'All') performs worse than the model based only on content features, which indicates that users and structural features do not contribute much to improving the performance of detecting malicious content in Pinterest. Surprisingly, the model using only content features performs comparably to the model based on the optimal subset of features (AUC: 0.949 vs. 0.942). This implies that, by observing only pin information that can be captured at the time of pinning, our model can effectively detect whether the posted pin is malicious or not (before it goes viral), which has a significant implication on protecting users from potential risks in Pinterest. Note that the prediction model on the dataset collected in 2017 shows a similar performance trend. Here, we only report the results based on (available) content features for the dataset collected in 2017 in Table 5.6.

To investigate how the sampling ratio (for addressing the class-imbalance problem) affects the performance of our proposed model (based on content features) in detecting

malicious content, we vary the sampling ratio from 1:1 to 1:12 and report the ROC for different sampling ratios in Figure 5.7. Note that the ideal ROC would lie close to the upper-left corner, i.e., TPR closest to 100% and FPR closest to 0%. As shown in Figure 5.7, there is a trade-off between TPR and FRP for different sampling ratios. That is, if we learn more instances from the majority class (i.e., benign pins) such as 1:8 or 1:12, we can achieve lower TPRs but better (i.e., lower) FPRs. On the other hand, if we learn a similar number of instances from the minority and majority classes (e.g., 1:1 or 1:2), their TPRs are better, but FPRs are worse than others. Overall, regardless of the sampling ratio, we can achieve significantly high performance (i.e., AUC is 0.95). To the best of our knowledge, there has not been a study on malicious content detection over content-driven OSNs. However, malicious content detection in friendship-based OSNs has a similar performance, using both user and content features.

5.4 Summary

This chapter studied the detailed analysis of malicious content propagation in Pinterest. By analyzing how malicious pins propagate in Pinterest, compared to benign content, we revealed that malicious content propagation is substantially different from benign ones. We found that there are some particular categories where malicious pins are mainly published. Our analysis further revealed that a vast portion of benign pins come from popular websites, while malicious pins are likely to come from a small set of non-popular websites. When we looked at malicious users, we found they tend to be more active than benign users. They are likely to post pins more often and are likely to have a larger number of followings/followers than benign users. However, we also revealed that their effort in distributing malicious content is not so effective on Pinterest. Based on lessons learned from our analysis, we developed machine-learning-based models to detect whether a posted pin is malicious or not, which can achieve high performance (i.e., AUC is 0.95). We showed that content features are essential predictors in detecting malicious content on Pinterest. We believe our work has a great utility in understanding and detecting malicious content for protecting users from potential risks in Pinterest-like OSNs.

5.5 Data Availability

The dataset is publicly available at `https://mmlab.snu.ac.kr/traces/pinterest/` [60]. We only use anonymized user IDs and activity data for this research, and no personally identifiable information is used.

Chapter 6

Summary and Conclusion

6.1 Summary

This dissertation aims to bridge the gap in the explainability of data-driven user behavior models by proposing interpretable approaches in various domains. First, in the realm of deep learning models for sequential recommendations, we introduce a model that can learn *interpretable* personalized user behaviors (Chapter 2). Our model uses an attention mechanism at the individual level to learn and explain user behaviors. The proposed model outperforms the state-of-the-art sequential recommenders on simulated and empirical datasets.

Next, we measure the impact of individuals' behavior, namely social distancing, on mitigating the spread of a contagious disease. Chapter 3 proposes a model to understand the impact of Non-Pharmaceutical Interventions (NPI) on containing a disease. We use users' social distancing as a proxy to estimate the effect of NPIs. We then use these estimates to predict the spread of a disease using a proposed SEIR model that incorporates social distancing and human movements. We discuss the limitations of such a model when using log data for analysis. Based on our model, we provide an analysis of what would happen if different NPIs were lifted. Such interpretation of the impact of NPIs provides policymakers insights into what policies to loosen to reduce the socio-economic costs of NPIs while containing the spread of a contagious disease in society.

In Chapter 4, we expanded the analysis by including another personal intervention,

namely mask-wearing, by individuals to mitigate the spread of airborne disease. Inspired by SIR epidemiological model, we introduce a model that estimates the growth rate of active daily cases. We discuss the advantages of such a model when noises are present in data, especially at the early stages of a pandemic. Using the growth rate model, we are able to understand the effect of mask-wearing while other covariates like NPIs and social distancing exist. We performed a case study on COVID-19 using data collected across 24 countries. Such analysis benefits policymakers in countries where mask-wearing is not an accepted trend in society. We also discuss the limitations that the nature of log data introduces to the model.

Finally, we aim to understand the malicious behavior in Online Social Networks (OSN). Many studies are conducted on friendship-based OSNs like Facebook (now Meta) and Twitter. Little attention has been paid to malicious content dissemination in content-driven OSNs like Pinterest. We conduct case studies on two datasets from Pinterest in the early stages of its growth to understand the behavior of malicious users and the mechanism of malicious content dissemination. Chapter 5 provides insights on malicious content propagation and characteristic behaviors between malicious and benign users. We propose a machine learning model to predict whether a content is malicious. We show that in content-driven OSNs like Pinterest, a model that uses only the content features available when the content is posted on OSN is accurate enough to detect malicious content. Such a model offers a critical tool to the provider to protect its users from malicious content before the content spreads in the network.

References

[1] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[2] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai," *International Journal of Human-Computer Studies*, vol. 146, p. 102551, 2021.

[3] S. A. Neslin and H. J. Van Heerde, *Promotion dynamics*. Now Publishers Inc, 2009.

[4] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[5] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, "Social data: Biases, methodological pitfalls, and ethical boundaries," *Frontiers in Big Data*, vol. 2, p. 13, 2019.

[6] M. Motamedi, N. Sakharnykh, and T. Kaldewey, "A data-centric approach for training deep neural networks with less data," *arXiv preprint arXiv:2110.03613*, 2021.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[8] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A historical perspective of explainable artificial intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 1, p. e1391, 2021.

[9] T. Donkers, B. Loepp, and J. Ziegler, "Sequential user-based recurrent neural network recommendations," in *Proceedings of the eleventh ACM conference on recommender systems*, 2017, pp. 152–160.

[10] M. Quadrana, A. Karatzoglou, B. Hidasi, and P. Cremonesi, "Personalizing session-based recommendations with hierarchical recurrent neural networks," in *proceedings of the Eleventh ACM Conference on Recommender Systems*, 2017, pp. 130–137.

[11] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," *arXiv preprint arXiv:1511.06939*, 2015.

[12] B. Hidasi and A. Karatzoglou, "Recurrent neural networks with top-k gains for session-based recommendations," in *Proceedings of the 27th ACM international conference on information and knowledge management*, 2018, pp. 843–852.

[13] Y. Feng, F. Lv, W. Shen, M. Wang, F. Sun, Y. Zhu, and K. Yang, "Deep session interest network for click-through rate prediction," *arXiv preprint arXiv:1905.06482*, 2019.

[14] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1441–1450.

[15] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 197–206.

[16] Q. Chen, H. Zhao, W. Li, P. Huang, and W. Ou, "Behavior sequence transformer for e-commerce recommendation in alibaba," in *Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data*, 2019, pp. 1–4.

[17] R. He, W.-C. Kang, and J. McAuley, "Translation-based recommendation," in *Proceedings of the eleventh ACM conference on recommender systems*, 2017, pp. 161–169.

[18] C. Hansen, C. Hansen, L. Maystre, R. Mehrotra, B. Brost, F. Tomasi, and M. Lalmas, "Contextual and sequential user embeddings for large-scale music recommendation," in *Fourteenth ACM Conference on Recommender Systems*, 2020, pp. 53–62.

[19] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong *et al.*, "Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia," *New England journal of medicine*, 2020.

[20] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu *et al.*, "A novel coronavirus from patients with pneumonia in china, 2019," *New England journal of medicine*, 2020.

[21] "Who covid 19 timeline," https://www.who.int/news-room/detail/27-04-2020-who-timeline-covid-19, accessed 6 May 2020.

[22] "Bbc news," https://www.bbc.com/news/world-52103747, accessed 6 May 2020.

[23] A. Sebhatu, K. Wennberg, S. Arora-Jonsson, and S. I. Lindberg, "Explaining the homogeneous diffusion of covid-19 nonpharmaceutical interventions across heterogeneous countries," *Proceedings of the National Academy of Sciences*, vol. 117, no. 35, pp. 21 201–21 208, 2020.

[24] C. Courtemanche, J. Garuccio, A. Le, J. Pinkston, and A. Yelowitz, "Strong social distancing measures in the united states reduced the covid-19 growth rate: Study evaluates the impact of social distancing measures on the growth rate of confirmed covid-19 cases across the united states." *Health affairs*, vol. 39, no. 7, pp. 1237–1246, 2020.

[25] S. T. Ali, L. Wang, E. H. Lau, X.-K. Xu, Z. Du, Y. Wu, G. M. Leung, and B. J. Cowling, "Serial interval of sars-cov-2 was shortened over time by nonpharmaceutical interventions," *Science*, vol. 369, no. 6507, pp. 1106–1109, 2020.

[26] B. J. Cowling, S. T. Ali, T. W. Ng, T. K. Tsang, J. C. Li, M. W. Fong, Q. Liao, M. Y. Kwan, S. L. Lee, S. S. Chiu *et al.*, "Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in hong kong: an observational study," *The Lancet Public Health*, vol. 5, no. 5, pp. e279–e288, 2020.

[27] N. G. Davies, A. J. Kucharski, R. M. Eggo, A. Gimma, W. J. Edmunds, T. Jombart, K. O'Reilly, A. Endo, J. Hellewell, E. S. Nightingale *et al.*, "Effects of non-pharmaceutical interventions on covid-19 cases, deaths, and demand for hospital services in the uk: a modelling study," *The Lancet Public Health*, vol. 5, no. 7, pp. e375–e385, 2020.

[28] S. G. Benzell, A. Collis, and C. Nicolaides, "Rationing social contact during the covid-19 pandemic: Transmission risk and social benefits of us locations," *Proceedings of the National Academy of Sciences*, vol. 117, no. 26, pp. 14 642–14 644, 2020.

[29] R. M. Anderson, H. Heesterbeek, D. Klinkenberg, and T. D. Hollingsworth, "How will country-based mitigation measures influence the course of the covid-19 epidemic?" *The lancet*, vol. 395, no. 10228, pp. 931–934, 2020.

[30] S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton *et al.*, "Estimating the effects of non-pharmaceutical interventions on covid-19 in europe," *Nature*, vol. 584, no. 7820, pp. 257–261, 2020.

[31] K. Prem, Y. Liu, T. W. Russell, A. J. Kucharski, R. M. Eggo, N. Davies, S. Flasche, S. Clifford, C. A. Pearson, J. D. Munday *et al.*, "The effect of control strategies to reduce social mixing on outcomes of the covid-19 epidemic in wuhan, china: a modelling study," *The Lancet Public Health*, vol. 5, no. 5, pp. e261–e270, 2020.

[32] S. Lai, N. W. Ruktanonchai, L. Zhou, O. Prosper, W. Luo, J. R. Floyd, A. Wesolowski, M. Santillana, C. Zhang, X. Du *et al.*, "Effect of non-pharmaceutical interventions to contain covid-19 in china," *nature*, vol. 585, no. 7825, pp. 410–413, 2020.

[33] A. Aravindakshan, J. Boehnke, E. Gholami, and A. Nayak, "Mask-wearing during the covid-19 pandemic," *Available at SSRN 3700900*, 2020.

[34] "World economic forum," https://www.weforum.org/agenda/2020/04/coronavirus-lockdowns-global/, accessed 6 May 2020.

[35] A. Atkeson, "What will be the economic impact of covid-19 in the us? rough estimates of disease scenarios," National Bureau of Economic Research, Tech. Rep., 2020.

[36] R. Chetty, J. N. Friedman, N. Hendren, M. Stepner *et al.*, "The economic impacts of covid-19: Evidence from a new public database built using private sector data," national Bureau of economic research, Tech. Rep., 2020.

[37] "The health and economic impacts of covid-19 interventions," https://www.rand.org/pubs/tools/TLA173-1/tool.html, accessed 6 September 2020.

[38] K. Leung, J. T. Wu, D. Liu, and G. M. Leung, "First-wave covid-19 transmissibility and severity in china outside hubei after control measures, and second-wave scenario planning: a modelling impact assessment," *The Lancet*, vol. 395, no. 10233, pp. 1382–1393, 2020.

[39] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman, "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2)," *Science*, vol. 368, no. 6490, pp. 489–493, 2020.

[40] M. U. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, O. C.-. D. W. Group†, L. du Plessis, N. R. Faria, R. Li *et al.*, "The effect of human mobility and control measures on the covid-19 epidemic in china," *Science*, vol. 368, no. 6490, pp. 493–497, 2020.

[41] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. Pastore y Piontti, K. Mu, L. Rossi, K. Sun *et al.*, "The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak," *Science*, vol. 368, no. 6489, pp. 395–400, 2020.

[42] "Google llc "google covid-19 community mobility reports."," https://www.google.com/covid19/mobility/, accessed 7 May 2020.

[43] C. Cheng, J. Barceló, A. S. Hartnett, R. Kubinec, and L. Messerschmidt, "Covid-19 government response event dataset (coronanet v. 1.0)," *Nature human behaviour*, vol. 4, no. 7, pp. 756–768, 2020.

[44] A. Pan, L. Liu, C. Wang, H. Guo, X. Hao, Q. Wang, J. Huang, N. He, H. Yu, X. Lin *et al.*, "Association of public health interventions with the epidemiology of the covid-19 outbreak in wuhan, china," *Jama*, vol. 323, no. 19, pp. 1915–1923, 2020.

[45] S. Hsiang, D. Allen, S. Annan-Phan, K. Bell, I. Bolliger, T. Chong, H. Druckenmiller, L. Y. Huang, A. Hultgren, E. Krasovich *et al.*, "The effect of large-scale anti-contagion policies on the covid-19 pandemic," *Nature*, vol. 584, no. 7820, pp. 262–267, 2020.

[46] "Cdc updata on novel coronavirus," https://www.cdc.gov/media/releases/2020/t0212-cdc-telebriefing-transcript.html, accessed 9 September 2020.

[47] "Surgeon general," https://twitter.com/Surgeon_General/status/1233725785283932160?s=20, accessed 9 September 2020.

[48] S. Feng, C. Shen, N. Xia, W. Song, M. Fan, and B. J. Cowling, "Rational use of face masks in the covid-19 pandemic," *The Lancet Respiratory Medicine*, vol. 8, no. 5, pp. 434–436, 2020.

[49] "Advice on the use of masks in the context of covid-19," https://apps.who.int/iris/handle/10665/332293, accessed 9 September 2020.

[50] "Personal measures taken to avoid covid-19," https://yougov.co.uk/topics/international/articles-reports/2020/03/17/personal-measures-taken-avoid-covid-19, accessed 9 September 2020.

[51] K. A. Auger, S. S. Shah, T. Richardson, D. Hartley, M. Hall, A. Warniment, K. Timmons, D. Bosse, S. A. Ferris, P. W. Brady *et al.*, "Association between statewide school closure and covid-19 incidence and mortality in the us," *Jama*, vol. 324, no. 9, pp. 859–870, 2020.

[52] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec, "Mobility network models of covid-19 explain inequities and inform reopening," *Nature*, vol. 589, no. 7840, pp. 82–87, 2021.

[53] A. Aravindakshan, J. Boehnke, E. Gholami, and A. Nayak, "Preparing for a future covid-19 wave: insights and limitations from a data-driven evaluation of non-pharmaceutical interventions in germany," *Scientific reports*, vol. 10, no. 1, pp. 1–14, 2020.

[54] W. Lyu and G. L. Wehby, "Community use of face masks and covid-19: Evidence from a natural experiment of state mandates in the us: Study examines impact on covid-19 growth rates associated with state government mandates requiring face mask use in public." *Health affairs*, vol. 39, no. 8, pp. 1419–1425, 2020.

[55] D. K. Chu, E. A. Akl, S. Duda, K. Solo, S. Yaacoub, H. J. Schünemann, A. El-harakeh, A. Bognanni, T. Lotfi, M. Loeb *et al.*, "Physical distancing, face masks, and eye protection to prevent person-to-person transmission of sars-cov-2 and covid-19: a systematic review and meta-analysis," *The lancet*, vol. 395, no. 10242, pp. 1973–1987, 2020.

[56] J. Howard, A. Huang, Z. Li, Z. Tufekci, V. Zdimal, H.-M. van der Westhuizen, A. von Delft, A. Price, L. Fridman, L.-H. Tang *et al.*, "An evidence review of face masks against covid-19," *Proceedings of the National Academy of Sciences*, vol. 118, no. 4, 2021.

[57] H. Bundgaard, J. S. Bundgaard, D. E. T. Raaschou-Pedersen, C. von Buchwald, T. Todsen, J. B. Norsk, M. M. Pries-Heje, C. R. Vissing, P. B. Nielsen, U. C. Winsløw *et al.*, "Effectiveness of adding a mask recommendation to other public health measures to prevent sars-cov-2 infection in danish mask wearers: a randomized controlled trial," *Annals of internal medicine*, vol. 174, no. 3, pp. 335–343, 2021.

[58] K. Abbasi, "The curious case of the danish mask study," 2020.

[59] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 721–730.

[60] J. Han, D. Choi, B.-G. Chun, T. Kwon, H.-c. Kim, and Y. Choi, "Collecting, organizing, and sharing pins in pinterest: interest-driven or social-driven?" *ACM SIGMETRICS Performance Evaluation Review*, vol. 42, no. 1, pp. 15–27, 2014.

[61] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev, "Prediction of retweet cascade size over time," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 2335–2338.

[62] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 925–936.

[63] J. Han, D. Choi, J. Joo, and C.-N. Chuah, "Predicting popular and viral image cascades in pinterest." in *ICWSM*, 2017, pp. 82–91.

[64] H. Gao, J. Hu, T. Huang, J. Wang, and Y. Chen, "Security Issues in Online Social Networks," *Internet Computing, IEEE*, vol. 15, no. 4, pp. 56–63, 2011.

[65] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th annual computer security applications conference*. ACM, 2010, pp. 1–9.

[66] G. Yan, G. Chen, S. Eidenbenz, and N. Li, "Malware propagation in online social networks: nature, dynamics, and defense implications," in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*. ACM, 2011, pp. 196–206.

[67] S. Lee and J. Kim, "Warningbird: Detecting suspicious urls in twitter stream." in *NDSS*, vol. 12, 2012, pp. 1–13.

[68] M. Z. C. H. A. Forte, C. Hall, and M. Zarro, "Wedding dresses and wanted criminals: Pinterest. com as an infrastructure for repository building," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013, pp. 650–658.

[69] J. Han, D. Choi, A. Choi, J. Choi, T. Chung, T. T. Kwon, J.-Y. Rha, C.-N. Chuah *et al.*, "Sharing topics in pinterest: understanding content creation and diffusion behaviors," in *Proceedings of the 2015 ACM on Conference on Online Social Networks*. ACM, 2015, pp. 245–255.

[70] B. Gelley and A. John, "Do i need to follow you?: Examining the utility of the pinterest follow mechanism," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 2015, pp. 1751–1762.

[71] "McAfee Trustedsource Scanner," https://www.trustedsource.org.

[72] "VirusTotal Scanner," https://www.virustotal.com/en/about/.

[73] Y. Koren and R. Bell, "Advances in collaborative filtering," *Recommender systems handbook*, pp. 77–118, 2015.

[74] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in *Recommender systems handbook*. Springer, 2011, pp. 1–35.

[75] B. Hidasi, "Cutting-edge collaborative recommendation algorithms: Deep learning," in *COLLABORATIVE RECOMMENDATIONS: Algorithms, Practical Challenges and Applications*. World Scientific, 2019, pp. 79–126.

[76] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[77] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," *Advances in neural information processing systems*, vol. 20, 2007.

[78] S. Kabbur, X. Ning, and G. Karypis, "Fism: factored item similarity models for top-n recommender systems," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 659–667.

[79] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," *IEEE Internet computing*, vol. 7, no. 1, pp. 76–80, 2003.

[80] S. Wang, Y. Wang, J. Tang, K. Shu, S. Ranganath, and H. Liu, "What your images reveal: Exploiting visual contents for point-of-interest recommendation," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 391–400.

[81] W.-C. Kang, C. Fang, Z. Wang, and J. McAuley, "Visually-aware fashion recommendation and design with generative image models," in *2017 IEEE international conference on data mining (ICDM)*. IEEE, 2017, pp. 207–216.

[82] W. Yu, X. He, J. Pei, X. Chen, L. Xiong, J. Liu, and Z. Qin, "Visually aware recommendation with aesthetic features," *The VLDB Journal*, vol. 30, no. 4, pp. 495–513, 2021.

[83] D. Kim, C. Park, J. Oh, S. Lee, and H. Yu, "Convolutional matrix factorization for document context-aware recommendation," in *Proceedings of the 10th ACM conference on recommender systems*, 2016, pp. 233–240.

[84] N. Hariri, B. Mobasher, R. Burke, and Y. Zheng, "Context-aware recommendation based on review mining," in *ITWP@ IJCAI*, 2011.

[85] L. Guo, Y. Han, H. Jiang, X. Yang, X. Wang, and X. Liu, "Learning to make document context-aware recommendation with joint convolutional matrix factorization," *Complexity*, vol. 2020, 2020.

[86] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," *Advances in neural information processing systems*, vol. 26, 2013.

[87] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 173–182.

[88] M. Naumov, D. Mudigere, H.-J. M. Shi, J. Huang, N. Sundaraman, J. Park, X. Wang, U. Gupta, C.-J. Wu, A. G. Azzolini *et al.*, "Deep learning recommendation model for personalization and recommendation systems," *arXiv preprint arXiv:1906.00091*, 2019.

[89] S. Sedhain, A. K. Menon, S. Sanner, and L. Xie, "Autorec: Autoencoders meet collaborative filtering," in *Proceedings of the 24th international conference on World Wide Web*, 2015, pp. 111–112.

[90] Y. Wu, C. DuBois, A. X. Zheng, and M. Ester, "Collaborative denoising auto-encoders for top-n recommender systems," in *Proceedings of the ninth ACM international conference on web search and data mining*, 2016, pp. 153–162.

[91] Y. Chen and M. de Rijke, "A collective variational autoencoder for top-n recommendation with side information," in *Proceedings of the 3rd workshop on deep learning for recommender systems*, 2018, pp. 3–9.

[92] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, "Factorizing personalized markov chains for next-basket recommendation," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 811–820.

[93] P. Wang, J. Guo, Y. Lan, J. Xu, S. Wan, and X. Cheng, "Learning hierarchical representation model for nextbasket recommendation," in *Proceedings of the 38th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2015, pp. 403–412.

[94] R. He and J. McAuley, "Fusing similarity models with markov chains for sparse sequential recommendation," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 2016, pp. 191–200.

[95] J. Tang and K. Wang, "Personalized top-n sequential recommendation via convolutional sequence embedding," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 565–573.

[96] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, "A dynamic recurrent model for next basket recommendation," in *Proceedings of the 39th International ACM SIGIR*

conference on Research and Development in Information Retrieval, 2016, pp. 729–732.

[97] X. Chen, H. Xu, Y. Zhang, J. Tang, Y. Cao, Z. Qin, and H. Zha, "Sequential recommendation with user memory networks," in *Proceedings of the eleventh ACM international conference on web search and data mining*, 2018, pp. 108–116.

[98] J. Huang, W. X. Zhao, H. Dou, J.-R. Wen, and E. Y. Chang, "Improving sequential recommendation with knowledge-enhanced memory networks," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 505–514.

[99] Z. Li, H. Zhao, Q. Liu, Z. Huang, T. Mei, and E. Chen, "Learning from history and present: Next-item recommendation via discriminatively exploiting user behaviors," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1734–1743.

[100] J. Li, P. Ren, Z. Chen, Z. Ren, T. Lian, and J. Ma, "Neural attentive session-based recommendation," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1419–1428.

[101] W. Yuan, H. Wang, X. Yu, N. Liu, and Z. Li, "Attention-based context-aware sequential recommendation model," *Information Sciences*, vol. 510, pp. 122–134, 2020.

[102] Q. Liu, Y. Zeng, R. Mokhosi, and H. Zhang, "Stamp: short-term attention/memory priority model for session-based recommendation," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1831–1839.

[103] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua, "Kgat: Knowledge graph attention network for recommendation," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 950–958.

[104] J. Li, Y. Wang, and J. McAuley, "Time interval aware self-attention for sequential recommendation," in *Proceedings of the 13th international conference on web search and data mining*, 2020, pp. 322–330.

[105] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[106] A. Baevski, S. Edunov, Y. Liu, L. Zettlemoyer, and M. Auli, "Cloze-driven pretraining of self-attention networks," *arXiv preprint arXiv:1903.07785*, 2019.

[107] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[108] P. Bordalo, N. Gennaioli, and A. Shleifer, "Salience and consumer choice," *Journal of Political Economy*, vol. 121, no. 5, pp. 803–843, 2013.

[109] S. Gabel, D. Guhl, and D. Klapper, "P2v-map: Mapping market structures for large retail assortments," *Journal of Marketing Research*, vol. 56, no. 4, pp. 557–580, 2019.

[110] P. Manchanda, A. Ansari, and S. Gupta, "The "shopping basket": A model for multicategory purchase incidence decisions," *Marketing science*, vol. 18, no. 2, pp. 95–114, 1999.

[111] W. H. Greene, *Econometric analysis*. Pearson Education India, 2003.

[112] D. Lewandowski, D. Kurowicka, and H. Joe, "Generating random correlation matrices based on vines and extended onion method," *Journal of multivariate analysis*, vol. 100, no. 9, pp. 1989–2001, 2009.

[113] https://www.chicagobooth.edu/research/kilts/datasets/nielseniq nielsen, "Nielsen consumer panel dataset," Mar. 2020. [Online]. Available: https://www.chicagobooth.edu/research/kilts/datasets/nielseniq-nielsen

[114] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," *arXiv preprint arXiv:1205.2618*, 2012.

[115] A. Goli and P. K. Chintagunta, "What happens when a retailer drops a product category? investigating the consequences of ending tobacco sales," *Marketing Science*, vol. 40, no. 6, pp. 1169–1198, 2021.

[116] "How the virus got out," https://www.nytimes.com/interactive/2020/03/22/world/coronavirus-spread.html, accessed 20 May 2020.

[117] "Coronavirus: the first three months as it happened," https://www.nature.com/articles/d41586-020-00154-w, accessed 20 May 2020.

[118] "Johns hopkings coronavirus resource center," https://coronavirus.jhu.edu/, accessed 20 April 2020.

[119] "Robert koch institute "current situation report of the rki to covid-19"," https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Gesamt.html, accessed 9 May 2020.

[120] "Covid-19 projections," https://covid19.healthdata.org/germany/, accessed 12 May 2020.

[121] "Covid 19 government measures dataset," https://data.humdata.org/dataset/acaps-covid-19-government-measures-dataset, accessed 25 April 2020.

[122] "Apple community mobility reports," https://covid19.apple.com/mobility, accessed 21 July 2020.

[123] "Transport — passenger transport demand and modal split," https://www.eea.europa.eu/soer/2015/countries-comparison/transport, accessed 21 May 2020.

[124] "And which modes of transportation do you use for your daily commute to work/school/university?" https://www.statista.com/forecasts/998727/modes-of-transportation-for-commuting-in-germany, accessed 3 September 2020.

[125] M. Schäfer, M. Strohmeier, V. Lenders, I. Martinovic, and M. Wilhelm, "Bringing up opensky: A large-scale ads-b sensor network for research," in *IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*. IEEE, 2014, pp. 83–94.

[126] "Coronavirus: Berlin march against lockdown measures," https://www.bbc.com/news/av/world-europe-52426378/coronavirus-berlin-march-against-lockdown-measures, accessed 9 May 2020.

[127] "Robert koch institute "current situation report of the rki to covid-19"," https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Gesamt.html, accessed: May 9, 2020.

[128] "The number of tests performed, germany," https://ourworldindata.org/coronavirus\-testing\#germany, accessed 17 May 2020.

[129] "Dr. anthony fauci says staying closed for too long could cause "irreparable damage"," https://www.cnbc.com/2020/05/22/dr-anthony-fauci-says-staying-closed-for-too-long-could-cause-irreparable-damage.html, accessed 24 May 2020.

[130] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, and J. Lessler, "The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application," *Annals of internal medicine*, vol. 172, no. 9, pp. 577–582, 2020.

[131] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study," *The Lancet*, vol. 395, no. 10225, pp. 689–697, 2020.

[132] A. Deardorff, "Tableau (version. 9.1)," *Journal of the Medical Library Association*, vol. 104, no. 2, pp. 182–183, 2016.

[133] B. E. Dowd, W. H. Greene, and E. C. Norton, "Computation of standard errors," *Health services research*, vol. 49, no. 2, pp. 731–750, 2014.

[134] O. Diekmann, J. A. P. Heesterbeek, and J. A. Metz, "On the definition and the computation of the basic reproduction ratio r 0 in models for infectious diseases in heterogeneous populations," *Journal of mathematical biology*, vol. 28, no. 4, pp. 365–382, 1990.

[135] P. Van den Driessche and J. Watmough, "Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission," *Mathematical biosciences*, vol. 180, no. 1-2, pp. 29–48, 2002.

[136] A. Aravindakshan, J. Boehnke, E. Gholami, and A. Nayak, "Restarting after covid-19: A data-driven evaluation of opening scenarios," *medRxiv*, 2020.

[137] S. Woloshin, N. Patel, and A. S. Kesselheim, "False negative tests for sars-cov-2 infection—challenges and implications," *New England Journal of Medicine*, vol. 383, no. 6, p. e38, 2020.

[138] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," *The Lancet infectious diseases*, vol. 20, no. 5, pp. 533–534, 2020.

[139] J. M. Wooldridge, "Control function methods in applied econometrics," *Journal of Human Resources*, vol. 50, no. 2, pp. 420–445, 2015.

[140] J. J. Heckman and R. Robb Jr, "Alternative methods for evaluating the impact of interventions: An overview," *Journal of econometrics*, vol. 30, no. 1-2, pp. 239–267, 1985.

[141] P. Bahl, S. Bhattacharjee, C. de Silva, A. A. Chughtai, C. Doolan, and C. R. MacIntyre, "Face coverings and mask to minimise droplet dispersion and aerosolisation: a video case study," *Thorax*, vol. 75, no. 11, pp. 1024–1025, 2020.

[142] M. Riediker and D.-H. Tsai, "Estimation of viral aerosol emissions from simulated individuals with asymptomatic to moderate coronavirus disease 2019," *JAMA network open*, vol. 3, no. 7, pp. e2 013 807–e2 013 807, 2020.

[143] C. J. Worby and H.-H. Chang, "Face mask use in the general population and optimal resource allocation during the covid-19 pandemic," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[144] T. Greenhalgh, M. B. Schmid, T. Czypionka, D. Bassler, and L. Gruer, "Face masks for the public during the covid-19 crisis," *Bmj*, vol. 369, 2020.

[145] "Why scientists say wearing masks shouldn't be controversial," https://www.sciencenews.org/article/covid-19-coronavirus-why-wearing-masks-controversial, accessed 9 September 2020.

[146] "Masks could help stop coronavirus. so why are they still controversial?" https://www.wsj.com/articles/masks-could-help-stop-coronavirus-so-why-are-they-still-controversial-11593336601, accessed 9 September 2020.

[147] "Sweden's gamble," https://www.sciencemag.org/news/2020/10/it-s-been-so-so-surreal-critics-sweden-s-lax-pandemic-policies-face-fierce-backlash, accessed 1 December 2020.

[148] "Why was mask wearing popular in asia even before covid-19?" https://www.psychologytoday.com/us/blog/culture-shocked/202005/why-was-mask-wearing-popular-in-asia-even-covid-19, accessed 28 August 2020.

[149] "A quick history of why asians wear surgical masks in public," https://qz.com/299003/a-quick-history-of-why-asians-wear-surgical-masks-in-public/, accessed 28 August 2020.

[150] "Personal measures taken to avoid covid-19," https://today.yougov.com/topics/international/articles-reports/2020/03/17/personal-measures-taken-avoid-covid-19, accessed 28 July 2020.

[151] "Jhu csse covid-19 data," https://github.com/CSSEGISandData/COVID-19, accessed 15 July 2020.

[152] "How many u.s. workers have lost jobs during coronavirus pandemic? there are several ways to count," https://www.wsj.com/articles/how-many-u-s-workers-have-lost-jobs-during-coronavirus-pandemic-there-are-several-ways-to-count-11591176601, accessed 14 August 2020.

[153] "Hispanic women, immigrants, young adults, those with less education hit hardest by covid-19 job losses," https://www.pewresearch.org/fact-tank/2020/06/09/hispanic-women-immigrants-young-adults-those-with-less-education-hit-hardest-by-covid-19-job-losses/, accessed 10 July 2020.

[154] "Coronavirus layoffs - job losses and furloughs are even impacting 'safe' jobs," https://www.forbes.com/sites/ryanguina/2020/05/05/coronavirus-layoffs-are-impacting-safe-jobs/#19f99ab53e17, accessed 10 August 2020.

[155] "How many coronavirus cases are asymptomatic? cdc and other data range as high as 50%," https://www.miamiherald.com/news/coronavirus/article241703806.html, accessed 28 August 2020.

[156] "Coronavirus (covid-19) testing," https://ourworldindata.org/coronavirus-testing, accessed 18 July 2020.

[157] "Google trends," https://trends.google.com/trends/?geo=US, accessed 14 July 2020.

[158] P. A. Dow, L. A. Adamic, and A. Friggeri, "The anatomy of large facebook cascades." *ICWSM*, vol. 1, no. 2, p. 12, 2013.

[159] M. R. Rahman, J. Han, and C.-N. Chuah, "Unveiling the adoption and cascading process of osn-based gifting applications," in *Computer Communications (INFOCOM), 2015 IEEE Conference on*. IEEE, 2015, pp. 2596–2604.

[160] D. Choi, J. Han, T. Chung, Y.-Y. Ahn, B.-G. Chun, and T. T. Kwon, "Characterizing Conversation Patterns in Reddit: From the Perspectives of Content Properties and User Participation Behaviors," in *ACM Conference on Online Social Networks (COSN)*, 2015.

[161] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 519–528.

[162] C. Zhong, D. Karamshuk, and N. Sastry, "Predicting Pinterest: Automating a Distributed Human Computation," in *International conference on World Wide Web (WWW)*, 2015.

[163] T. Wu, S. Wen, S. Liu, J. Zhang, Y. Xiang, M. Alrubaian, and M. M. Hassan, "Detecting spamming activities in twitter based on deep-learning technique," *Concurrency and Computation: Practice and Experience*, vol. 29, no. 19, p. e4209, 2017.

[164] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, "Phishari: Automatic realtime phishing detection on twitter," in *eCrime Researchers Summit (eCrime), 2012*, Oct 2012, pp. 1–12.

[165] S. Lee and J. Kim, "Early filtering of ephemeral malicious accounts on twitter," *Computer Communications*, vol. 54, pp. 48–57, 2014.

[166] Q. Gong, Y. Chen, X. He, Z. Zhuang, T. Wang, H. Huang, X. Wang, and X. Fu, "Deepscan: Exploiting deep learning for malicious account detection in location-based social networks," *IEEE Communications Magazine, Feature Topic on Mobile Big Data for Urban Analytics*, vol. 56, no. 1, 2018.

[167] J. Cao, Q. Li, Y. Ji, Y. He, and D. Guo, "Detection of forwarding-based malicious urls in online social networks," *International Journal of Parallel Programming*, vol. 44, no. 1, pp. 163–180, 2016.

[168] C. Lo, J. Cheng, and J. Leskovec, "Understanding online collection growth over time: A case study of pinterest," in *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2017, pp. 545–554.

[169] S. Mittal, N. Gupta, P. Dewan, and P. Kumaraguru, "Pinned it! a large scale study of the pinterest network," in *Proceedings of the 1st IKDD Conference on Data Sciences*. ACM, 2014, pp. 1–10.

[170] W. Hu, K. K. Singh, F. Xiao, J. Han, C.-N. Chuah, and Y. J. Lee, "Who will share my image?: Predicting the content diffusion path in online social networks," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 252–260.

[171] "McAfee, Reference Guide McAfee TrustedSource Web Database, Category Set 4," 2010, https://www.trustedsource.org/download/ts_wd_reference_guide.pdf.

[172] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[173] I. Guyon and A. Elisseeff, "An introduction to feature extraction," *Feature extraction*, pp. 1–25, 2006.

[174] L. Breiman, "Random Forests," *Springer Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[175] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[176] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, pp. 321–357, 2002.