Changes in Reproductive Ratio of SARS-CoV-2 Due to Implementation and Rollback of Non-pharmaceutical Interventions in 1,904 United States Counties

Jie Ying Wu (jieying@jhu.edu)
Johns Hopkins University https://orcid.org/0000-0002-7306-8140
Benjamin Killeen
Johns Hopkins University https://orcid.org/0000-0003-2511-7929
Philipp Nikutta
Johns Hopkins University
Shreya Chakraborty
Johns Hopkins University
Anna Zapaishchykova
Johns Hopkins University
Mareike Thies
Johns Hopkins University https://orcid.org/0000-0002-1364-4337
Mathias Unberath
Johns Hopkins University

Article

Keywords: SARS-CoV-2, non-pharmaceutical interventions, reproductive ratio

Posted Date: September 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-71225/v1

Changes in Reproductive Ratio of SARS-CoV-2 Due to Implementation and Rollback of Non-pharmaceutical Interventions in 1,904 United States Counties

Jie Ying Wu*†[1,2], Benjamin D. Killeen†[1,2], Philipp Nikutta[2], Shreya Chakraborty[2], Anna Zapaishchykova[2], Mareike Thies[2], and Mathias Unberath[1,2,3]

[1]Department of Computer Science, The Johns Hopkins University
[2]Laboratory for Computational Sensing and Robotics, The Johns Hopkins University
[3]The Malone Center for Engineering in Healthcare, The Johns Hopkins University

Abstract

In response to the rapid spread of the novel coronavirus, SARS-CoV-2, the U.S. has largely delegated implementation and rollback of non-pharmaceutical interventions (NPIs) to local governments on the state and county level. This asynchronous response combined with the heterogeneity of the U.S. complicates quantification of the effect of NPIs on the reproductive ratio of SARS-CoV-2 on a national level.

We describe a data-driven approach to quantify the effect of NPIs that relies on county-level similarities to specialize a Bayesian mechanistic model based on observed fatalities. Using this approach, we estimate the effect of NPIs on the reproductive ratio R_t in 1,904 U.S. counties incorporating implementation, subsequent rollback, and mask mandate efficacy.

We estimate that at some point before August 2, 2020, 1,808 out of the considered 1,904 U.S. counties had reduced the reproductive ratio of SARS-CoV-2 to below 1.0. However, on August 2, the reproductive ration remained below that threshold for only 702 counties.

*Corresponding author.
†Equal contribution.

The estimated effect of any individual NPI is different across counties. Public school closings were estimated to be effective in metropolitan, urban, and suburban counties, while advisory NPIs were estimated to be effective in more rural counties. The cumulative prevalence predicted by the model ranges from 0 to 58.6% across the counties examined. The median is 2.6% while the 25th and 75th percentile are 1.3% and 44.6% respectively, indicating that most counties are far from herd immunity.

Our results suggest that local conditions, including socioeconomic, demographic and infrastructural factors, in addition to the cumulative prevalence are pertinent to containment and re-opening decisions.

1 Introduction

As of September 3, 2020, the United States has reported more than 6,100,000 cases of novel coronavirus 2019 (COVID-19).[1] The disease, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, has led to more than 183,000 deaths in the U.S..[1] As pharmaceutical interventions remain unavailable, non-pharmaceutical interventions (NPIs) have been a critical component of the public health effort to slow the spread of COVID-19. NPIs include guidelines for hand hygiene, mask mandates, cancellations of mass events, school closures, closure of non-essential businesses, and stay-at-home orders. The most drastic measures are designed to reduce transmission rapidly, buying time to expand healthcare capacity, develop effective testing and tracing mechanisms, and research pharmaceutical options, such as a vaccine.

Indeed, the implementation of NPIs coincides with a measurable decline in new cases and deaths.[1] Conversely, the rollback of NPIs coincides with a surge of new cases in certain areas,[1] but not all. This work aims to quantify how each type of NPI affects disease transmission on the U.S. county level. In the short term, these results may inform maintenance or re-implementation of NPIs as necessary; in the long term, they may guide the safe rollback of NPIs while minimizing adverse affects. Prior works have quantified the effect of implementing NPIs on the reproductive ratio R_t of SARS-CoV-2 in China,[2] UK,[3] Brazil,[4] and 14 European countries, including Italy, Spain, and Germany.[5,6] In the U.S., these effects have been quantified on the state level,[7] and in

some large urban areas[8], but the effect of county-level implementations and rollbacks has yet to be addressed. This is vital, since the U.S. delegated NPI implementation to local governments rather than establishing a unified, federal approach, and most states have done likewise.[9] Moreover, to our knowledge, no peer-reviewed work has estimated the effects of NPI rollbacks in the U.S..

As a result, we confront the United States' county-level implementations with a county-level approach. At the same time, we contend with the limited number of documented fatalities in many counties, which would otherwise limit this analysis to urban areas with sufficient data for epidemiological modeling.

We estimate the change in R_t due to implementing and rolling back NPIs in 1,904 U.S. counties that exhibit substantially different characteristics regarding population density, economy, demographics, and infrastructure. To contend with this heterogeneity, we develop an approach that establishes county similarity based on factors known to affect community transmission of infectious diseases.[10] Within groups of similar counties, we jointly optimize the parameters of a Bayesian mechanistic model to the observed deaths in every county under the assumption that the same NPI attains comparable effects across counties that are similar with respect to aforementioned factors. These counties may still differ in their initial reproductive ratio and the response of their citizens such as through mask wearing, which allows for county-level differentiation within the model. Furthermore, in order to incorporate rollbacks of NPIs in our analysis, we assume that the effect of the rollback is identical to the corresponding implementation. This follows from the fundamental assumption that rollbacks lead to a resumption of the original behavior, like returning to schools and workplaces, as they are intended to. However, differences arise due to additional behaviors that also affect transmission, which individuals may have adopted in the meantime. To this end, we consider the efficacy of face covering mandates as an independent NPI.

Face coverings have been shown to be an effective piece of personal protective equipment to reduce the likelihood of disease transmission,[11] and they are especially important for slowing the

spread of COVID-19 due to the prevalence of aerosol transmission, with or without presenting symptoms.[12,13] They may consist of folded handkerchiefs, knitted cotton masks, surgical masks, or other respirators, each of which reduces the distance that exhaled particles travel from an individual, albeit to varying degrees.[14] This work estimates the effect of mask mandates as an NPI, which the public may or may not adhere to, rather than of masks as individual protection. Mask mandates require individuals to wear face coverings in certain public settings, either at all times, when indoors, or when social distancing is not possible. The details of these requirements vary from state to state, an issue which is further complicated by differing public attitude toward, awareness of, or adherence to the mandate.[15] In a given state, a single mask mandate may have widely varying effectiveness from county to county. This work estimates the effectiveness of mask mandates as specific to each county, as a step toward understanding what factors may allow a county to safely reopen.

2 Conclusions

Based on the model, 1,808 (95%) of the considered 1,904 U.S. counties are estimated to have reduced the reproductive ratio of novel coronavirus to below 1.0 between Jan 22nd and Aug 2nd via the implementation of NPIs, as of August 2, 2020. Some of these counties have been able to reopen while keeping their reproductive ratio low, while others have seen a second rise in R_t upon reopening. On August 2 itself, our model estimates that 702 counties (36.9%) had a reproductive ratio below 1.0, meaning 1,202 (63.1%) did not. This is partially because of the different behavior changes adopted by the inhabitants of each county. In our model, this is reflected by rollbacks only taking place in some counties, or an estimated efficacy of mask mandate which mitigates the effects of rollbacks.

We observe that for metropolitan and urban counties, the most substantial reduction is attributed to the closing of public schools while less restrictive NPIs were estimated to be effective in more rural

counties. Further, the expected level of infection predicted by the model is far from herd immunity even in counties with advanced spread. Of the counties examined, the median percentage of the population infected is estimated to be 2.6%, with the 25th and 75th percentile at 1.3 and 44.6%, respectively. This indicates that few counties are driving the alarming increase in national case count, while most counties remain far below the prevalence necessary for herd immunity. While the model explains the observed trends in fatalities well, the rapid succession of different NPIs being implemented, and to a lesser extent being rolled back, in most counties complicates the disentanglement of the effects of any individual NPI.

Despite these limitations, our results suggest that strategies for shutdown as well as re-opening require careful consideration of county conditions in addition to state and national trends, with the responsiveness to mask mandates strongly affecting a county's ability to reopen safely.

3 Results

3.1 Characterizing Groups of Similar Counties as Clusters

Since our hypothesis is that local conditions affect the spread of COVID-19, we differentiate among groups of counties using a data-driven approach known as clustering. Each group — or cluster — of counties is characterized by having similar demographic and socioeconomic qualities. For instance, cluster 1 consists of low-population, mostly rural counties with little public transit capacity and the lowest median household income. Cluster 2 and 3 are similar in size, having a mean population size of 45,000 and 52,000 respectively, but cluster 2 includes higher-income, suburban areas where-as cluster 3 has lower income areas with a large land area. Cluster 4 consists of the densest metropolitan areas with a high proportion of 18- to 65-year-olds, high household income, a high public transit score, and small land area. Finally, cluster 5 has the highest mean population size besides cluster 4, but is less densely populated, has poor public transit, and a lower household income on average. It is important to note that although we refer to these clusters with numbers

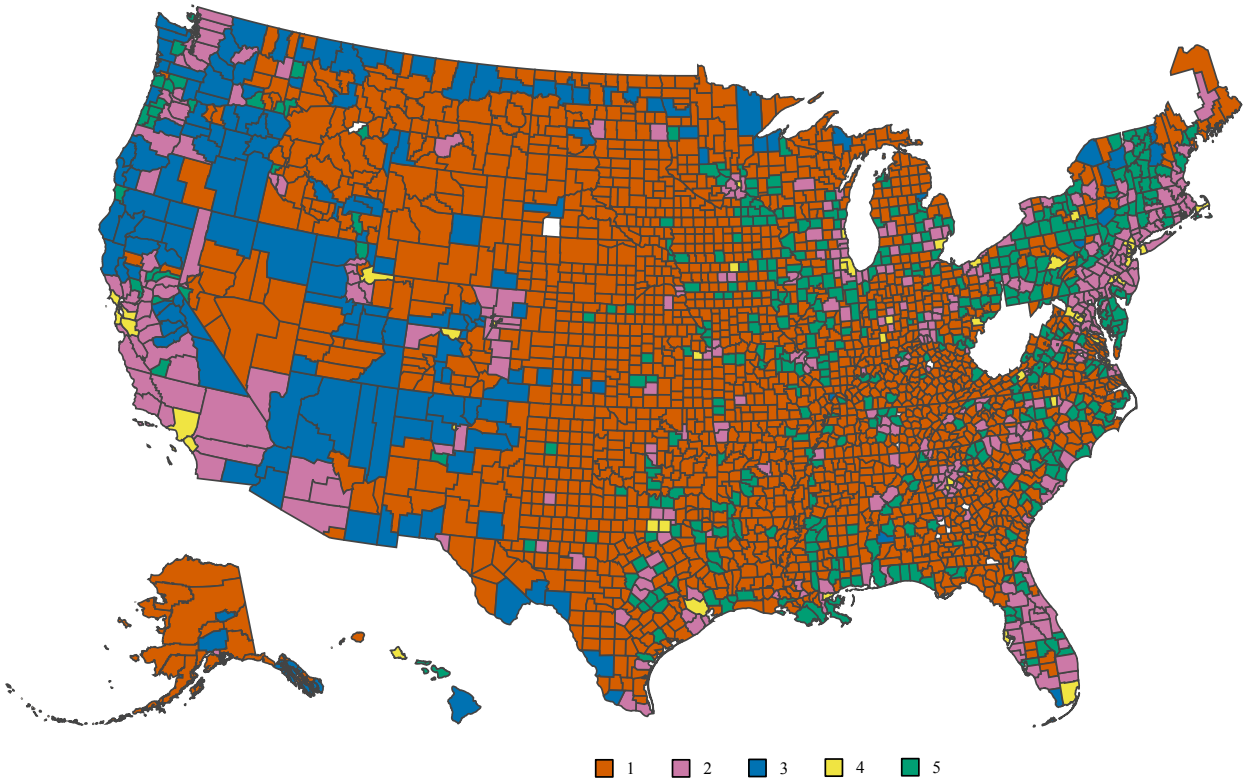

Figure 1: Cluster labels based on demographic and socioeconomic conditions are used to aggregate data and specialize epidemiological models. Here, one can see how cluster 1 and 3 primarily cover rural areas, while clusters 5, 2, and 4 consist of increasingly urban counties.

1-5, these are merely labels returned by the clustering algorithm without any meaning inherent to the ordering. Figure 1 shows our clustering for all U.S. counties with this data available, including those without any incidence of COVID-19.

Once we have identified the clusters, we infer the reproductive ratio of SARS-CoV-2 over time by jointly optimizing the parameters of a Bayesian mechanistic model on each cluster. This process is described in greater detail in Section 5. Figure 2 shows the reproductive ratio at select dates for all modeled counties associated with their public transportation use.

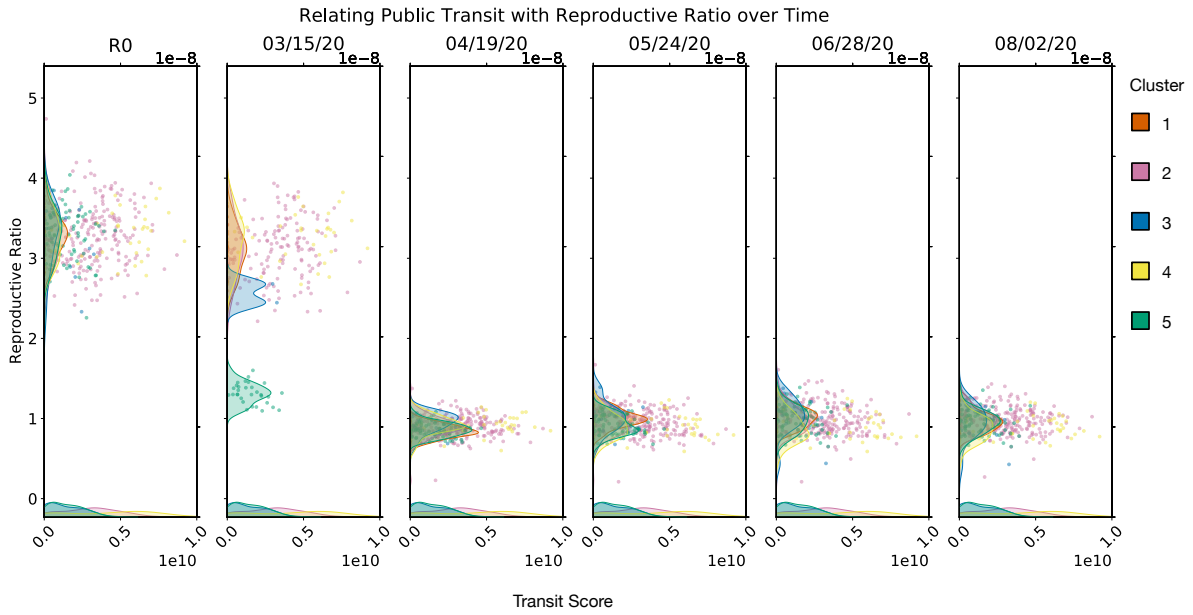

Figure 2: Relationship between public transit capacity and the time-dependent reproductive ratio of SARS-CoV-2 for U.S. counties, through August 2, 2020. In Clusters 2 and 5, which have lower transit scores than cluster 4, we observe a noticeable drop in reproductive ratio by March 15, indicating the stronger effect of advisory NPIs in effect at that time.

We observe that although the basic reproductive ratio of SARS-CoV-2 starts at a similar level for all clusters, the speed at which counties in each cluster respond to the disease and reduce its R_t differs. The reproductive ratio in metropolitan counties (cluster 4), which tends to have higher reliance on public transportation, decreases over the entire period. This is especially apparent going from March 15 to March 25. On the other hand, clusters with little transportation use, such as cluster 5, were estimated to have quickly reduced transmission rates. This suggests that if people have a higher reliance on public infrastructure, more stringent interventions are necessary to reduce transmission. The staggered drops in reproductive ratios are attributed to both varied efficacy of NPIs (see Section 3.3) and counties implementing NPIs at different times. The resurgence

of the virus in some areas can be seen as the reproductive ratio increases above 1 by August 2. Comparisons with more features are shown in Section 4.2.

3.2 Estimates of Initial and Current Reproductive ratios and Number of Infected

From Table 2, we observe that most counties exhibited an initial reproductive ratio R_0 above 3. As of August 2, however, most have successfully reduced the reproductive ratio to $R_t \approx 1$ after implementing NPIs. It is estimated that 17.46 to 31.83% of the population (95% confidence interval) has been infected in New York, NY, which has the highest number of cases by August 2. Based on the initial reproductive ratios between 2 and 4, herd immunity is reached only after 50 - 70% of the population has recovered,[16,17] suggesting that all U.S. counties are far from achieving herd immunity. Consequently, easing restrictions is likely to result in, and in some counties have resulted in, subsequent waves of the epidemic. We state these findings for 15 representative counties in Table 2 and provide the same metrics for all 1,904 counties online at github.com/JieYingWu/npi-model, where we also provide code and data required for reproduction.

3.3 Learned Effects of NPIs

We quantify the effectiveness of NPIs for counties in each cluster, as shown in Table 3, and the county-specific effectiveness of mask mandates where implemented, shown in Figure 3.

Cluster	County	R_0 (95% CI)	$R_{8/2}$ (95% CI)	# (%) estimated cases	# (%) measured cases	Population (2018)	Fatality rate (measured death/cases)
1	Tallapoosa County, AL	1.093 (0.989, 1.196)	0.752 (0.217, 0.910)	8,277 (20.4)	855 (2.1)	40,497	9.24%
1	Washington County, FL	3.647 (2.553, 4.073)	1.276 (1.218, 1.330)	1,931 (7.8)	867 (3.5)	24,880	1.61%
1	Iredell County, NC	3.679 (3.441, 3.860)	1.032 (0.868, 1.094)	1,897 (1.1)	1,820 (1.0)	178,435	0.99%
2	Putnam County, NY	0.894 (0.699, 1.141)	0.270 (0.026, 0.539)	8,074 (8.2)	1,449 (1.5)	98,892	4.35%
2	Brevard County, FL	1.216 (1.088, 1.323)	1.428 (1.296, 1.529)	20,338 (3.4)	6,064 (1.0)	596,849	2.49%
2	Maricopa County, AZ	3.941 (3.789, 4.106)	1.099 (1.067, 1.141)	361,572 (8.2)	126,053 (2.9)	4,410,824	1.87%
3	Yakima County, WA	3.050 (2.683, 3.423)	0.427 (0.023, 1.026)	31,322 (12.5)	10,325 (4.1)	251,446	2.04%
3	Yavapai County, AZ	1.132 (0.874, 1.465)	1.240 (0.992, 1.462)	15,425 (6.6)	2,004 (0.9)	231,993	3.34%
3	Webb County, TX	2.175 (1.360, 3.582)	1.160 (0.971, 1.246)	48,781 (17.7)	7,888 (2.9)	275,910	2.14%
4	Arlington County, VA	0.921 (0.831, 1.035)	0.500 (0.178, 0.742)	26,549 (11.2)	3,077 (1.3)	237,521	4.39%
4	Pinellas County, FL	3.784 (3.626, 4.005)	1.201 (1.141, 1.258)	59,300 (6.1)	17,879 (1.8)	975,280	2.80%
4	Los Angeles County, CA	3.380 (3.216, 3.543)	1.030 (0.840, 1.099)	710,834 (7.0)	208,563 (2.1)	10,105,518	2.39%
5	Saginaw County, MI	1.104 (0.935, 1.510)	0.649 (0.105, 0.853)	15,188 (8.0)	2,010 (1.1)	190,800	6.52%
5	Platte County, MO	1.094 (0.981, 1.281)	1.352 (1.255, 1.448)	1,720 (1.7)	363 (0.4)	102,985	2.75%
5	Leon County, FL	3.837 (3.667, 4.022)	1.178 (1.117, 1.243)	1,950 (0.7)	4,965 (1.7)	292,502	0.48%

Table 2: Estimated initial and current reproductive ratio as of August 2, 2020, and the number of cases for selected counties. We show the counties with the highest and lowest reproductive ratio as of August 2, as well as the largest county by population in each cluster. This is compared to the measured number of cases and fatality rates. Estimates for all 1,904 counties are available at github.com/JieYingWu/npi-model.

Intervention	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
I_1: Stay at home	0.101 (0.044, 0.168)	-0.004 (-0.008, 0.022)	0.019 (-0.008, 0.152)	-0.005 (-0.008, 0.012)	-0.005 (-0.008, 0.015)
I_2: > 50 gathering	0.008 (-0.008, 0.071)	0.001 (-0.008, 0.030)	0.085 (-0.008, 0.273)	0.001 (-0.008, 0.049)	0.141 (0.067, 0.212)
I_3: > 500 gathering	0.008 (-0.008, 0.076)	0.048 (0.003, 0.108)	0.075 (-0.008, 0.264)	0.001 (-0.008, 0.058)	0.006 (-0.008, 0.083)
I_4: Public schools	0.157 (-0.008, 0.815)	0.791 (0.312, 1.017)	0.284 (-0.008, 0.841)	0.688 (-0.008, 1.065)	0.034 (-0.008, 0.264)
I_5: Restaurant dine-in	0.004 (-0.008, 0.077)	-0.006 (-0.008, 0.015)	0.137 (-0.008, 0.379)	-0.001 (-0.008, 0.039)	0.103 (-0.008, 0.197)
I_6: Entertain-ment/gym	0.163 (0.071, 0.224)	0.204 (0.176, 0.235)	0.132 (-0.008, 0.369)	0.113 (0.062, 0.163)	0.049 (-0.008, 0.193)
I_7: Federal guidelines	0.823 (0.018, 1.095)	0.1174 (-0.008, 0.668)	0.281 (-0.008, 0.847)	0.391 (-0.008, 1.133)	0.046 (-0.008, 0.376)
I_8: Foreign travel ban	0.070 (-0.008, 0.407)	0.030 (-0.008, 0.204)	0.170 (-0.008, 0.778)	0.005 (-0.008, 0.084)	0.934 (0.623, 1.074)

Table 3: The mean and 95% confidence interval of learned α-values (see Equation 1) of the interventions and rollbacks for each cluster-specialized model. Higher α-values correspond to a greater reduction of the reproductive number R_t following the implementation of an NPI or, equivalently, greater increase following its rollback.

We note that our model estimates different behavior across counties. While all counties have

implemented a similar set of interventions, their estimated effects are substantially different in each respective cluster. For example, metropolitan counties (cluster 4) and suburban counties (cluster 2) were estimated to have a strong response to public school closings, while rural areas (clusters 1, 3, and 5) responded more to national-level interventions according to our model. One surprising observation is that stay-at-home orders are not given the same importance as seen in other works. This may be because of the concurrent estimation of its rollback effect and the mitigating effects of masks in the meantime. With non-essential businesses closed, thereby limiting indoor places to congregate, and masks being worn at all times, the risk of being in public spaces may remain low. Additionally, the effects of public school closing may confound the estimated effectiveness of stay-at-home orders as parents adopt stay-at-home-like practices even without mandated stay-at-home orders out of necessity.

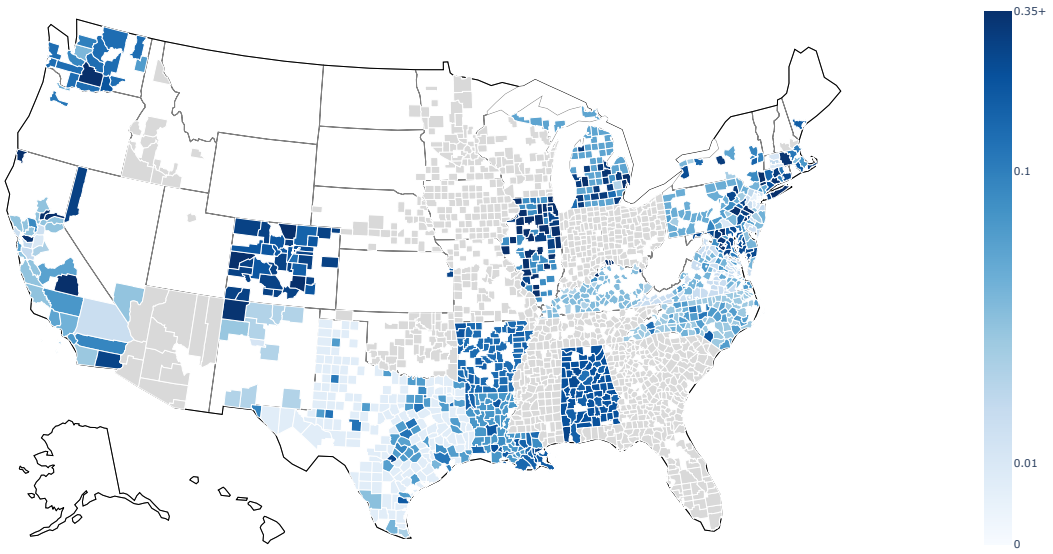

Figure 3: The quantified effectiveness $\alpha_{\texttt{mask},m}$ of mask mandates for each county included in our model. Counties which never implemented mask mandates, as of August 2, but are still included in our model are shown in gray. Darker colors correspond to higher values, where masks mandates were estimated to have greater effect toward reducing the reproductive ratio. Since mask mandates are implemented universally across a given state, these values are best interpreted in comparison with counties from the same state.

Figure 3 shows the county-specific estimates for the effectiveness of mask mandates. Because these mandates have been issued at the state level, these values tend to be similar across a given state. Our cluster-specialized model in particular gives rise to county-level variation for the effectiveness of a single mask mandate, which may not otherwise be apparent when modeling at the state level. In some states, such as Texas, we observe greater effect from these mandates in urban and suburban areas (clusters 2 and 4 resp.) compared to the surrounding counties. This could indicate greater

adherence to mask mandates, but it also likely describes the greater necessity of masks in a densely populated region, where social distancing is more difficult.

One caveat is that the effects of interventions that were implemented in short temporal succession are difficult to disentangle. Many local governments implemented formal NPIs in immediate response to the federal guidelines, leading a quick succession of NPIs coming into effect. While simultaneously estimating the rollbacks may help in disentanglement, rollbacks have often been implemented at the state level. Additionally, changes in people's behaviors such as mask wearing complicates disentanglement. Disentanglement of the effects of NPIs is explored more in Section B.

Another limitation of our model is that the federal level, advisory interventions may have come before some counties have seen any cases. This introduces ambiguity between R_0 estimates and the weight attributed to federal NPIs, i.e. federal guidelines and travel ban. Similar ambiguity is less prominent for other interventions, since they are not generally implemented before a county has seen cases. Additionally, since the epidemiological model used here is mechanistic, it only allows for changes of the transmission rate at the time of implementation of interventions. Other events, such as high-profile cases and cancellations of prominent festivals, likely contributed to increasing awareness of the disease and may also have effects on individual behavior, and therefore the R_t. These effects cannot be attributed to a specific date and may thus affect the weights of interventions that come into effect at around the same time.

4 Methods

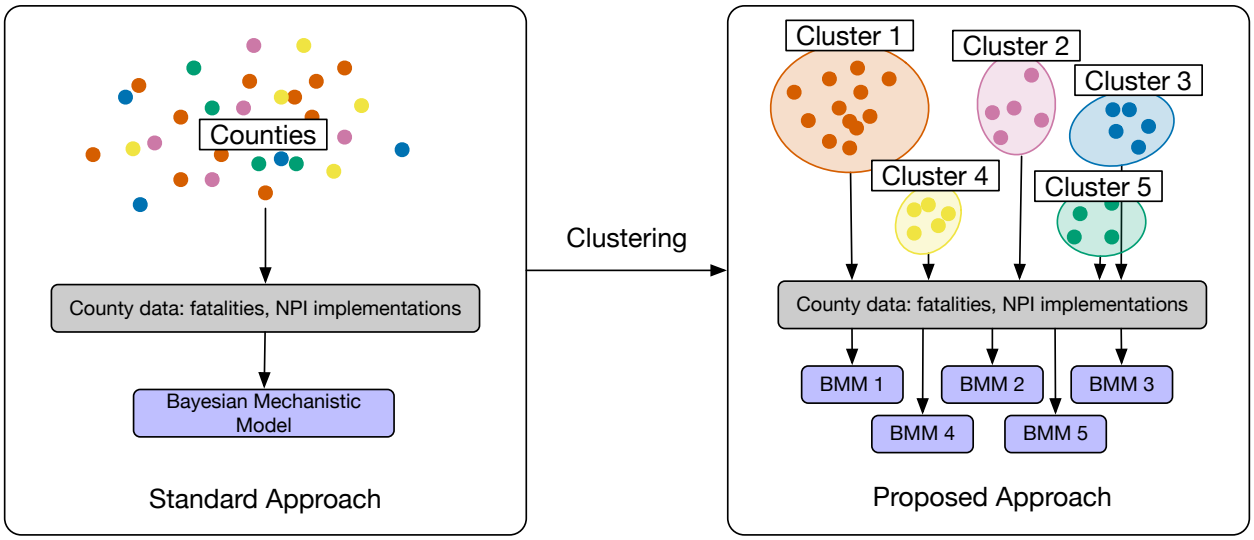

Figure 4: The fitting process for our model, compared to the standard approach. We fit Bayesian mechanistic models (BMM 1-5) to each one of five clusters of counties, based on similarity in socioeconomic, demographic, and other factors.

Figure 4 shows an overview of the proposed method, which adapts a well-established approach to account for the heterogeneity of the United States when estimating the effects of NPIs on disease spread. Naively, one may account for this heterogeneity by estimating NPI effects for each individual county, but this would limit analysis to counties with a large number of cases, without the potential to generalize to counties with fewer cases.

We aim to understand the effect of each type of NPI in general, not as it depends on each individual county's implementation, enforcement, or public awareness, so that this work may inform future implementations in response to the outbreak of COVID-19 or another disease. Therefore, we cluster counties based on variables known to affect disease spread and rely on the fundamental assumption that NPI effects can be estimated for each group of counties.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
Super-counties	24	16	4	1	22	68
Counties in super-counties	1,137	87	28	22	296	1,570
Counties not in super-counties	26	204	11	42	51	334
Total Counties Considered	1,163	291	39	64	347	1,904

Table 4: Number of counties or super-counties included in our study. Each county or super-county included has 50 or more cumulative fatalities by August 2, 2020.

Note that we exempt mask mandates from this assumption, since these appear to vary based on factors unrelated to disease transmission, such as political affiliation.[15] We train a Bayesian mechanistic model (see Sec. 5) for each cluster, allowing the effects of NPIs to differ based on epidemiological factors in a county rather than state or national boundaries.

4.1 Super-counties

Quantifying changes in COVID-19's reproductive ratio is complicated when considering differences at the county — rather than the national — level, which is necessary due to the distributed nature in which the United States has allowed individual counties to implement NPIs. Our Bayesian mechanistic model is optimized to describe fatality counts, the volume of which decreases significantly outside of highly populated regions. To make up for this scarcity when fitting an epidemiological model, we leverage a balanced clustering of U.S. counties to aggregate data from similar counties in the same state, treating them as a single entity or "super-county." This has the advantage of considering counties that would otherwise be excluded without assuming that the spread of the disease in those counties follows the same trend of more advanced regions in the same state or country. Counties with 50 or more cumulative fatalities as of August 2 are included independently. Counties with fewer than 50 fatalities are aggregated into a super-county if they are in the same state, belong to the same cluster, and implemented NPIs at the same time. Table 4 summarizes the number of counties in each of these categories for each cluster, and Figure 5 visualizes this process for counties in cluster 1 in Texas.

In addition to this data aggregation strategy, we fit a cluster-specialized model to each group of counties, quantifying the possibly disparate effects of the NPIs in each type of county, as detailed below.

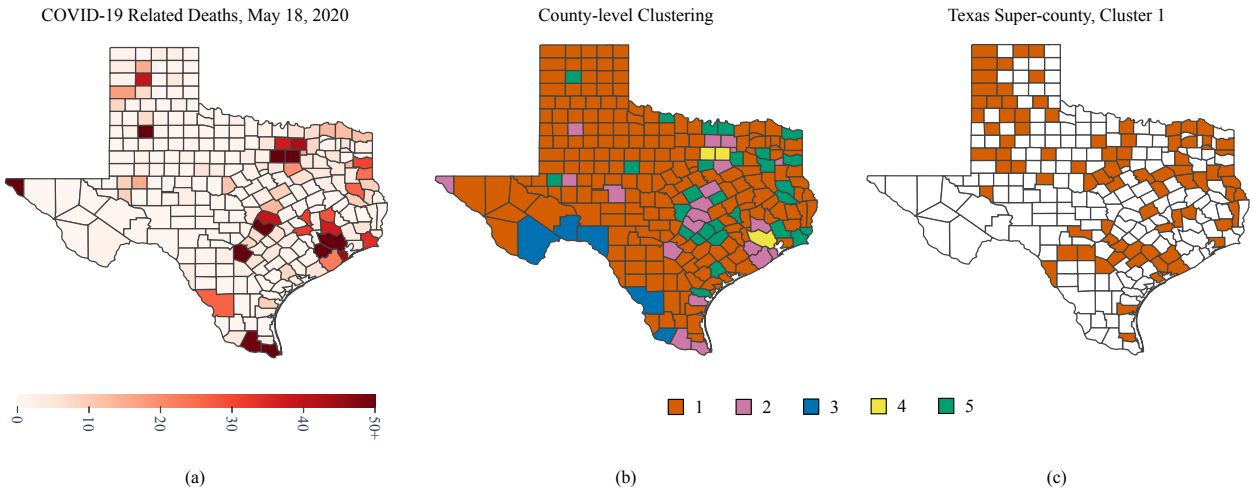

Figure 5: **(a)** The total confirmed deaths caused by COVID-19 for counties in Texas, as of August 2. **(b)** Cluster labels for each Texas county, based on demographics, education, density, and other factors. **(c)** Texas counties in cluster 1 having 1-49 cumulative deaths as of August 2, 2020 and the same NPI implementation dates. To enable robust epidemiological models, these counties are treated as a single "super-county." A county or super-county must have 50 or more cumulative deaths as of August 2 to be considered.

4.2 Clustering Counties

To generate the clustering, we partition 3,059 U.S. counties into five groups based on variables which directly affect disease spread.[10] Although these are not the only factors which may affect spread, they provide a meaningful basis for aggregation and separation of epidemiological data and parameters. Table 5 summarizes the variables used for clustering, which include demographic, economic, and public transit capacities that we have gathered, processed for machine readability,

Variable	Cluster 1 Mean (std)	Cluster 2 Mean (std)	Cluster 3 Mean (std)	Cluster 4 Mean (std)	Cluster 5 Mean (std)
Population	23,522 (23,217)	450,789 (464,001)	52,425 (57,060)	759,468 (1,418,561)	84,595 (59,674)
Fraction of population male, age 0-17	0.114 (0.018)	0.116 (0.015)	0.113 (0.026)	0.105 (0.020)	0.110 (0.014)
Fraction of population female, age 0-17	0.108 (0.018)	0.111 (0.014)	0.108 (0.025)	0.101 (0.019)	0.105 (0.014)
Fraction of population male, age 18-64	0.297 (0.035)	0.304 (0.014)	0.298 (0.032)	0.320 (0.033)	0.300 (0.024)
Fraction of population female, age 18-64	0.2803 (0.022)	0.309 (0.014)	0.282 (0.019)	0.325 (0.027)	0.296 (0.020)
Fraction of population male, age 65+	0.093 (0.023)	0.071 (0.017)	0.097 (0.032)	0.064 (0.016)	0.085 (0.021)
Fraction of population female, age 65+	0.1075 (0.022)	0.088 (0.019)	0.102 (0.032)	0.085 (0.022)	0.103 (0.023)
Fraction of population number with some college or associate's degree	0.213 (0.040)	0.198 (0.028)	0.238 (0.044)	0.167 (0.041)	0.213 (0.030)
Fraction of population in poverty	0.153 (0.057)	0.109 (0.042)	0.146 (0.056)	0.140 (0.059)	0.135 (0.047)
Fraction of population unemployed	0.018 (0.005)	0.018 (0.004)	0.023 (0.010)	0.018 (0.004)	0.019 (0.004)
Median household income	49,085 (10,377)	69,118 (16,914)	53,606 (11,240)	68,624 (27,320)	54,430 (10,073)
Population density (persons per sq. mile)	42.3 (44.5)	626.6 (469.6)	18.9 (17.4)	3789.1 (7,947.8)	132.6 (88.1)
Number of housing units (per capita)	0.498 (0.114)	0.389 (0.050)	0.539 (0.247)	0.415 (0.084)	0.454 (0.090)
Land area (sq. miles)	1,120.42 (4,296.43)	993.03 (1,591.41)	3,409.26 (3,301.91)	355.80 (583.55)	650.96 (239.58)
Population-weighted transit score	0.0 (0.0)	2.70e9 (1.92e9)	1.18e+9 (1.25e9)	2.86e+9 (2.98e9)	1.13e+9 (1.19e9)

Table 5: Average values and standard deviations for each of the 16 variables considered in our clustering, capturing demographic and socioeconomic information as well as transit capacity. As can be seen, certain variables are highly distinct among clusters, whereas others exhibit more overlap.

and released in a publicly available dataset.[9] Sources include the United States Census Bureau, the United States Department of Agriculture Economic Research Service and the Center for Neighborhood Technology. A full list of sources can be found on the corresponding website.[9] To incorporate potential exposure, we consider county population, density, housing density, and land area. Additionally, we consider portions of the population for age- and gender-based demographic categories, due to COVID-19's disparate effects on these groups.[18–22] Our clustering is also based on socioeconomic variables, which may indicate behavioral traits relevant to the spread of COVID-19. For instance, workers with tertiary education are more likely to hold office-type jobs which can be done from home.[23] At the same time, many secondary-education jobs have been deemed essential, resulting in a high contact rate, which in turn increases the likelihood of infection. Thus, our clustering considers college education, poverty, unemployment, and median household income for each county as a proxy of the overall job composition in the local area. We include a population-weighted transit score from the Center for Neighborhood Technology,[9] which conducts research on urban infrastructure, due to the likelihood of transmission in the enclosed, possibly crowded space that public transport entails. Altogether, these variables are used to separate counties into five clusters using a Gaussian mixture model.[24]

Figure 6 and Figure 7 show how R_t relates to median household income and population density for counties and super-counties in the different clusters. Super-counties are visualized as a single point using their population-weighted average for that feature. The plots show the distribution of the cluster over the features and its correlation with how R_t changes.

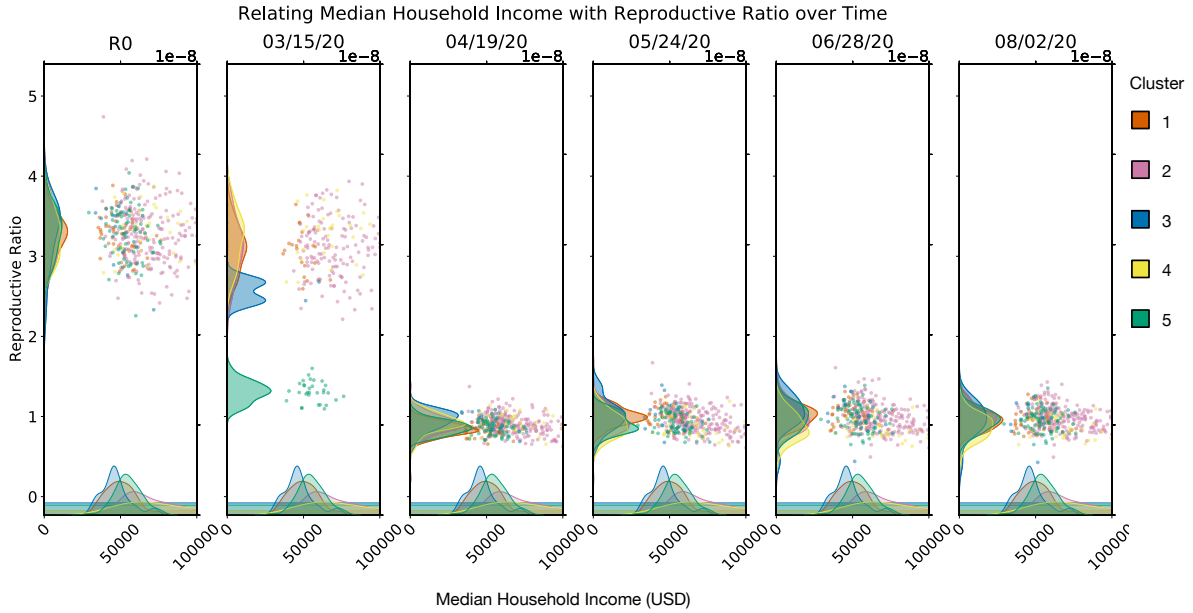

Figure 6: Scatter plot and density distribution plot for counties and super-counties comparing R_t over time to median household income. Colors indicate which cluster the county or super-county belongs in.

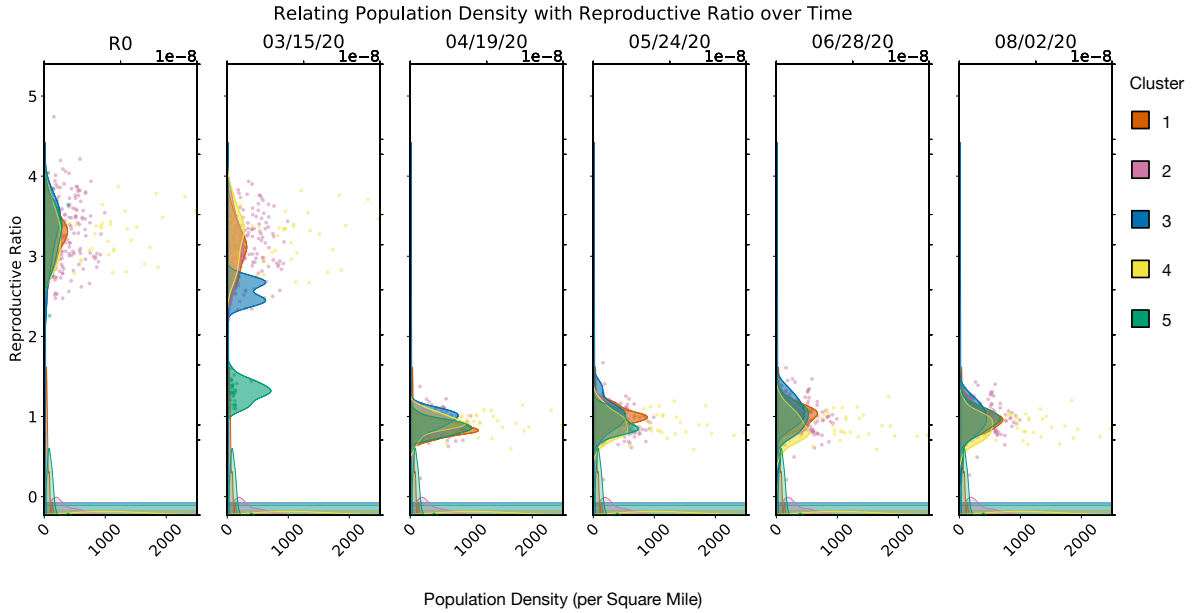

Figure 7: Scatter plot and density distribution plot for counties and super-counties comparing R_t over time to population density. Colors indicate which cluster the county or super-county belongs in.

Notably, we exclude ethnic demographics from the variables considered during clustering because we assume that no direct relationship exists between race and incidence of COVID-19. Rather, such a relationship would be mediated by socioeconomic factors that are already included for similarity assessment.

5 Modeling the Effects of NPIs with Bayesian mechanistic models

As counties have started to reopen, some have experienced alarming spikes in the number of cases and deaths from COVID-19.[1] We extend our model to include the effects of rollbacks in an effort to find patterns in which counties can safely reopen and which counties would experience spikes. We

separate this analysis from our primary analysis of the effect of NPIs as this is complicated by the asymmetric effect of interventions and their rollback. For example, counties generally mandated all restaurants to shut at the same time and that uniformly means they can only operate take-out and delivery. When reopening restaurants however, different phases have been implemented from 25% occupancy and outdoor-only seating to full occupancy with indoor service. Additionally, many restaurants continue to operate as take-out and delivery only even as the county has relaxed its restrictions as it may not make financial sense to operate at reduced capacity or out of fear of spreading the disease.[25] Given these caveats, we consider county-level modeling of even crudely lumped interventions a step in informing conversations about reopening.

5.0.1 Data Processing

We use cumulative fatality and infection counts from the JHU CSSE COVID-19 Dashboard, which has been tracking COVID-19 since January.[1] When fitting our model, we use measured fatality rates, which are generally considered more reliable than confirmed infections because of limited testing and the prevalence of asymptomatic cases. Thus, we use population-weighted fatality rates to estimate the true cases count. Obtaining a reasonable estimate for this ratio is crucial to realistically model the numbers of total infections. However, due to asymptomatic cases, undertesting and biased reporting, this parameter cannot be measured directly, but has to be inferred from observable data.[26–28] Previous studies all report fatality rates with substantial uncertainty but agree on the fact that fatality for COVID-19 depends strongly on the age of the infected person. Therefore, we adapt the fatality rates per age group presented in Verity et al.[26] for each county with respect to its demographic age distribution. Based on U.S. Census data, a per-county weighted fatality rate is computed using the share of each age group in the overall population.

To collect reopening dates, we start with the IHME database of rollback dates on the state level.[29] Since IHME's categories are rollback of stay at home, gatherings, essential business, and non-essential business, we match the former two rollbacks with our NPI intervention and associate

non-essential business with reopening of restaurants and gyms. We choose to match the rollbacks to the interventions so we can measure the asymmetry of an intervention's implementation and rollback. Then, we compare the IHME state-level results with reopening dates from the New York Times[30] and update any discrepancies, such as if the opening of the first non-essential business did not include restaurants or gyms. The latter data source also points out states with counties that have implemented rollbacks differently.[30] To collect county-level data, we look into every state the Times points out with county-level discrepancy, the states that have seen recent increases in cases such as Texas and Florida, and states that have driven the COVID trends in the US such as New York.[30]

5.0.2 Model

We estimate the effective reproductive ratio using a semi-mechanistic Bayesian mechanistic model proposed in Flaxman et al.[5], that infers the impact of a predefined set of interventions and estimates the number of infections over time. We use the priors given by the model and refer to Flaxman et al. for a complete description and exhaustive validation of these design choices.[5] Here, we outline the most relevant parameters for our work, most notably (1) the inclusion of NPI rollbacks and (2) their dependence on a county-specific effectiveness for mask mandates.

Eq 1 describes the model for the reproductive ratio $R_{t,m}$ in each county m at time t. The model estimates a county-specific initial reproductive ratio $R_{0,m}$, which captures county-level variations which are constant in time. In the mechanistic model, this initial reproductive ratio remains unchanged until an intervention takes effect at time t, represented by the binary indicator $I_{t,m}^{(i)}$ for the ith intervention. This indicator conveniently captures reversal of NPIs, reverting to 0 after a rollback. The model estimates the effects of NPIs as weights α_i for $i \in \{1, \ldots, n\}$. Additionally, the model estimates a county-specific weight $\alpha_{\texttt{mask},m}$ to account for differing mask usage in each

county.[15] Together, the effects of NPIs are modeled as multiplicative:

$$R_{t,m} = R_{0,m} \exp\left(-\left[I^{(1)}_{t,m}\alpha_1 + I^{(2)}_{t,m}\alpha_2 + \cdots + I^{(n)}_{t,m}\alpha_n\right] - I^{(\texttt{mask})}_{t,m}\alpha_{\texttt{mask},m}\right) \tag{1}$$

We describe the data collection process for NPI implementations and rollbacks in Section 5.0.1. Aside from mask mandates, our experiments consider $n = 8$ NPI types, as enumerated in Table 3.

The model assumes a normal distribution truncated at 0 as the prior for the R_0. For each initial reproductive ratio, we use the prior $R_{0,m} \sim N(3.28, \kappa)$ where $\kappa \sim N^+(0, 0.5)$, in accordance with the analysis presented in Liu et al.[31] We draw the intervention weights from an offset Gamma distribution following previous work.[5]

$$\alpha_i \sim \Gamma(0.1666, 1) - \frac{\log(1.05)}{6} \tag{2}$$

Starting from the time-varying R_t, a latent function of daily infections is modeled depending on two factors: a generation distribution $g \sim \Gamma(6.5, 0.62)$ that models the time between spread of infection from one individual to the next and the number of susceptible individuals left in the population.

Model fitting is driven by the timeseries of observed daily deaths. These are linked to the modeled number of infections by the county specific weighted fatality rate. We model the noise the weighted fatality rate in fitting the model as $\texttt{noise} \sim N(1, 0.1)$. The time from infection to death is given by the sum of the infection-to-onset distribution and the onset-to-death distribution, where the two distributions are independent of each other. The former is modeled as a Gamma distribution with mean 5.1 days and coefficient of variation 0.86, while the latter is modeled as a Gamma distribution with mean 18.8 days and a coefficient of variation 0.45. The infection-to-death distribution can

therefore be modeled as

$$\pi \sim \Gamma(5.1, 0.86) + \Gamma(18.8, 0.45) \tag{3}$$

We use this to model the probability of death by taking into account the number of people infected in all previous days, the number of days since infection. The number of cases at time t in region m is given by $c_{t,m}$.

$$c_{t,m} = R_{t,m} \sum_{\tau=0}^{t-1} c_{\tau,m} g_{t-\tau} \tag{4}$$

$$g_s = \int_{\tau=s-0.5}^{s+0.5} g(\tau) d\tau \tag{5}$$

$c_{t,m}$, along with the weighted probability of death, gives the number of deaths on a day for a given county as given by $d_{t,m}$.

$$d_{t,m} = \sum_{\tau=0}^{t-1} c_{\tau,m} \pi_{t-\tau,m}, \tag{6}$$

$$\pi_{s,m} = \int_{\tau=s-0.5}^{s+0.5} \pi_m(\tau) d\tau \tag{7}$$

We compare the model's expected number of deaths $d_{t,m}$ for region m on day t to the measured deaths $D_{t,m}$.

$$D_{t,m} \sim \texttt{NegativeBinomial}(d_{t,m}, d_{t,m} + \frac{d_{t,m}^2}{\phi}) \tag{8}$$

where $\phi \sim N^+(0, 5)$. To ensure that the deaths accounted for are from locally acquired infections, we include observed deaths in a county only after the cumulative count has exceeded 10. The seeding of new infections is assumed to be a month prior to that. For each of the first 6 days of infection that the model considers, the model optimizes the number of cases to a number drawn from the following prior: $\exp\left(\frac{1}{\tau}\right)$ where $\tau \sim \exp(0.3)$.

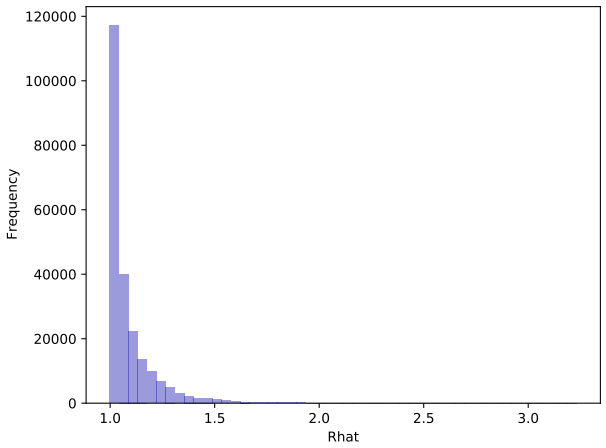

Figure 8: Histogram of Rhat for the Pystan model fit of cases and deaths for national-level counties and super-counties. Values close to 1 indicate convergence.

All parameters are estimated jointly using an adaptive Hamiltonian Monte Carlo (HMC) sampler in the probabilistic programming language Stan.[32] We fit the model over 2000 total iterations, with 1000 of those as warm-up iterations. We use 6 chains, set adapt delta to 0.99, and the maximum tree depth to 15 for all the following experiments. We show the histogram for the `rhat` convergence indicator below for our over all 5 clusters. The values are close to 1, indicating convergence of our model.

References

[1] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 0(0), February 2020. ISSN 1473-3099, 1474-4457. doi: 10.1016/S1473-3099(20)30120-1.

[2] Shengjie Lai, Nick W Ruktanonchai, Liangcai Zhou, Olivia Prosper, Wei Luo, Jessica R Floyd, Amy Wesolowski, Chi Zhang, Xiangjun Du, Hongjie Yu, et al. Effect of non-pharmaceutical interventions for containing the COVID-19 outbreak: An observational and modelling study. *medRxiv*, 2020.

[3] Nicholas G Davies, Adam J Kucharski, Rosalind M Eggo, Amy Gimma, W John Edmunds, Thibaut Jombart, Kathleen O'Reilly, Akira Endo, Joel Hellewell, Emily 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*, 2020.

[4] Thomas A Mellan, Henrique H Hoeltgebaum, Swapnil Mishra, Charlie Whittaker, Ricardo P Schnekenberg, Axel Gandy, H Juliette T Unwin, Michaela A C Vollmer, Helen Coupland, Iwona Hawryluk, Nuno Rodrigues Faria, Juan Vesga, Harrison Zhu, Michael Hutchinson, Oliver Ratmann, Melodie Monod, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Nicholas Brazeau, Giovanni Charles, Laura V Cooper, Zulma Cucunuba, Gina Cuomo-Dannenburg, Amy Dighe, Bimandra Djaafara, Jeff Eaton, Sabine L van Elsland, Richard FitzJohn, Keith Fraser, Katy Gaythorpe, Will Green, Sarah Hayes, Natsuko Imai, Ben Jeffrey, Edward Knock, Daniel Laydon, John Lees, Tara Mangal, Andria Mousa, Gemma Nedjati-Gilani, Pierre Nouvellet, Daniela Olivera, Kris V Parag, Michael Pickles, Hayley A Thompson, Robert Verity, Caroline Walters, Haowei Wang, Yuanrong Wang, Oliver J Watson, Lilith Whittles, Xiaoyue Xi, Lucy Okell, Ilaria Dorigatti, Patrick Walker, Azra Ghani, Steven M Riley, Neil M Ferguson, Christl A Donnelly, Seth Flaxman, and Samir Bhatt. Report 21: Estimating covid-19 cases and reproduction number in brazil. *medRxiv*, 2020. doi: 10.1101/2020.05.09.20096701. URL https://www.medrxiv.org/content/early/2020/05/18/2020.05.09.20096701.

[5] Seth Flaxman, Swapnil Mishra, Axel Gandy, H. Juliette T. Unwin, Helen Coupland, Thomas A. Mellan, Harrison Zhu, Tresnia Berah, Jeffrey W. Eaton, Pablo N. P. Guzman, Nora Schmit, Lucia Callizo, Imperial College COVID-19 Response Team, Charles Whittaker, Peter Winskill, Xiaoyue Xi, Azra Ghani, Christl A. Donnelly, Steven Riley, Lucy C. Okell, Michaela A. C. Vollmer, Neil M. Ferguson, and Samir Bhatt. Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in European

countries: Technical description update. *arXiv:2004.11342 [stat]*, April 2020.

[6] Michaela AC Vollmer, Swapnil Mishra, H Juliette T Unwin, Axel Gandy, Thomas A Mellan, Valerie Bradley, Harrison Zhu, Helen Coupland, Iwona Hawryluk, Michael Hutchinson, et al. A sub-national analysis of the rate of transmission of covid-19 in italy. *medRxiv*, 2020.

[7] H Juliette T Unwin, Swapnil Mishra, Valerie C Bradley, Axel Gandy, Thomas A Mellan, Helen Coupland, Jonathan Ish-Horowicz, Michaela Andrea Christine Vollmer, Charles Whittaker, Sarah L Filippi, et al. State-level tracking of covid-19 in the united states. *medRxiv*, 2020.

[8] Jesús Fernández-Villaverde and Charles I Jones. Estimating and simulating a sird model of covid-19 for many countries, states, and cities. Technical report, National Bureau of Economic Research, 2020.

[9] Benjamin D. Killeen, Jie Ying Wu, Kinjal Shah, Anna Zapaishchykova, Philipp Nikutta, Aniruddha Tamhane, Shreya Chakraborty, Jinchi Wei, Tiger Gao, Mareike Thies, and Mathias Unberath. A County-level Dataset for Informing the United States' Response to COVID-19. *arXiv:2004.00756 [physics, q-bio]*, April 2020.

[10] CDC. CDC COVID-19 Global Response. https://www.cdc.gov/coronavirus/2019-ncov/global-covid-19/community-mitigation-measures.html, June 2020.

[11] C Raina MacIntyre, Simon Cauchemez, Dominic E Dwyer, Holly Seale, Pamela Cheung, Gary Browne, Michael Fasher, James Wood, Zhanhai Gao, Robert Booy, et al. Face mask use and control of respiratory virus transmission in households. *Emerging infectious diseases*, 15(2):233, 2009.

[12] Monica Gandhi, Deborah S Yokoe, and Diane V Havlir. Asymptomatic transmission, the achilles' heel of current strategies to control covid-19, 2020.

[13] Xi He, Eric HY Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y Wong, Yujuan Guan, Xinghua Tan, et al. Temporal dynamics in viral shedding and transmissibility of covid-19. *Nature medicine*, 26(5):672–675, 2020.

[14] Siddhartha Verma, Manhar Dhanak, and John Frankenfield. Visualizing the effectiveness of face masks in obstructing respiratory jets. *Physics of Fluids*, 32(6):061708, 2020.

[15] Matthew H Goldberg, Abel Gustafson, Edward W Maibach, Matthew T Ballew, Parrish Bergquist, John E Kotcher, Jennifer R Marlon, Seth A Rosenthal, and Anthony Leiserowitz. Mask-wearing increased after a government recommendation: A natural experiment in the us during the covid-19 pandemic. *Frontiers in Communication*, 5:44, 2020.

[16] Haley E Randolph and Luis B Barreiro. Herd immunity: Understanding covid-19. *Immunity*, 52(5):737–741, 2020.

[17] Kin On Kwok, Florence Lai, Wan In Wei, Samuel Yeung Shan Wong, and Julian WT Tang. Herd immunity–estimating the level required to halt the covid-19 epidemics in affected countries. *Journal of Infection*, 80(6):e32–e33, 2020.

[18] CDC COVID and Response Team. Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, February 12–March 16, 2020. *MMWR. Morbidity and Mortality Weekly Report*, 69(12):343–346, 2020.

[19] Andrea Remuzzi and Giuseppe Remuzzi. COVID-19 and Italy: What next? *The Lancet*, 395 (10231):1225–1228, April 2020. ISSN 0140-6736. doi: 10.1016/S0140-6736(20)30627-9.

[20] Epidemiology Working Group for NCIP Epidemic Response, Chinese Center for Disease Control and Prevention. [The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China]. *Zhonghua Liu Xing Bing Xue Za Zhi = Zhonghua Liuxingbingxue Zazhi*, 41(2):145–151, February 2020. ISSN 0254-6450. doi: 10.3760/cma.j.issn.0254-6450.2020.02.003.

[21] Ping-Ing Lee, Ya-Li Hu, Po-Yen Chen, Yhu-Chering Huang, and Po-Ren Hsueh. Are children less susceptible to COVID-19? *Journal of Microbiology, Immunology, and Infection*, February 2020. ISSN 1684-1182. doi: 10.1016/j.jmii.2020.02.011.

[22] Qiurong Ruan, Kun Yang, Wenxia Wang, Lingyu Jiang, and Jianxin Song. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. *Intensive Care Medicine*, 46(5):846–848, May 2020. ISSN 1432-1238. doi: 10.1007/s00134-020-05991-x.

[23] Hans-Martin von Gaudecker, Radost Holler, Lena Janys, Bettina Siflinger, and Christian Zimpelmann. Labour Supply in the Early Stages of the COVID-19 Pandemic: Empirical Evidence on Hours, Home Office, and Expectations. SSRN Scholarly Paper ID 3579251, Social Science Research Network, Rochester, NY, April 2020.

[24] Douglas A Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741, 2009.

[25] Yvonne Wenger. Baltimore eases restrictions on restaurants, churches, gyms, as city moves to new phase of COVID-19 recovery. https://www.baltimoresun.com/maryland/baltimore-city/bs-md-ci-coronavirus-baltimore-recovery-20200619-kwd6wz3r4bbdhbman64stpwjr4-story.html.

[26] Robert Verity, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick GT Walker, Han Fu, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*, 2020.

[27] Timothy W Russell, Joel Hellewell, Christopher I Jarvis, Kevin Van Zandvoort, Sam Abbott, Ruwan Ratnayake, Stefan Flasche, Rosalind M Eggo, W John Edmunds, Adam J Kucharski, et al. Estimating the infection and case fatality ratio for coronavirus disease (covid-19) using

age-adjusted data from the outbreak on the diamond princess cruise ship, february 2020. *Eurosurveillance*, 25(12):2000256, 2020.

[28] Gianluca Rinaldi and Matteo Paradisi. An empirical estimate of the infection fatality rate of covid-19 from the first italian outbreak. *medRxiv*, 2020.

[29] Bo Xu, Bernardo Gutierrez, Sumiko Mekaru, Kara Sewalk, Lauren Goodwin, Alyssa Loskill, Emily L Cohn, Yulin Hswen, Sarah C Hill, Maria M Cobo, et al. Epidemiological data from the covid-19 outbreak, real-time case information. *Scientific data*, 7(1):1–6, 2020.

[30] Jasmine Lee, Sarah Mervosh, Yuriria Avila, Barbara Harvey, and Alex Leeds Matthews. See How All 50 States Are Reopening (and Closing Again). *The New York Times*, September 2020.

[31] Ying Liu, Albert A Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. The reproductive number of covid-19 is higher compared to sars coronavirus. *Journal of travel medicine*, 2020.

[32] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[33] Ahmet Aktay, Shailesh Bavadekar, Gwen Cossoul, John Davis, Damien Desfontaines, Alex Fabrikant, Evgeniy Gabrilovich, Krishna Gadepalli, Bryant Gipson, Miguel Guevara, Chaitanya Kamath, Mansi Kansal, Ali Lange, Chinmoy Mandayam, Andrew Oplinger, Christopher Pluntke, Thomas Roessler, Arran Schlosberg, Tomer Shekel, Swapnil Vispute, Mia Vu, Gregory Wellenius, Brian Williams, and Royce J Wilson. Google covid-19 community mobility reports: Anonymization process description (version 1.0), 2020.

A Validation

Although prior work has validated the model,[5,7] we perform two experiments to show the value and consistency of our results. First we compare the fatality estimates of our model for a given county to validation models which observe data only from the given model. Naturally, these validation models have less signal to accurately estimate NPI effects, but by fixing the effects of NPIs, we can observe how comparable fatality estimates are nevertheless obtained using the appropriate, cluster-specific is. Second, we correlate the estimated reproductive ratio with publicly available mobility data from Google.[33] This shows the consistency of our results, which are based solely on official NPI implementations, with actual observations of the behavior those NPIs address. This validates our fundamental assumption that NPIs can influence behaviors that affect disease transmission, such as mobility.

A.1 Validating NPI Effects

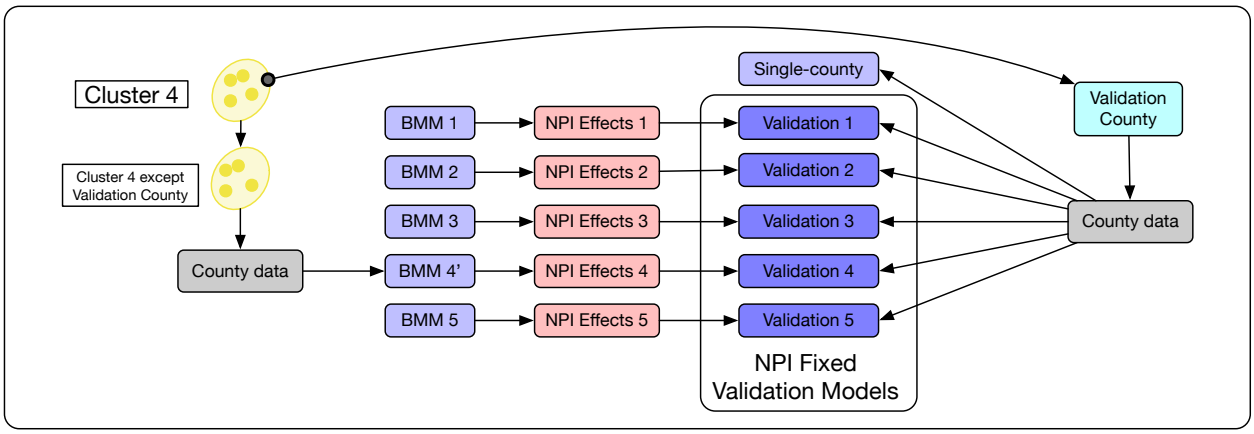

Figure 9: An overview of our process for validation the advantages of clustering when estimating the effects of NPIs. Validation models (dark blue) use NPI-effects from each cluster-specialized model. We expect the estimates from Validation 4 to closely match those of BMM 4, but others may vary significantly. BMM 1, 2, 3, and 5 match the models from Figure 4. BMM 4' is similar to BMM 4, with the difference being the exclusion of the validation county during fit.

To validate the advantages of clustering, we fit several models to a single county and compare the fatality estimates and $\alpha_{\texttt{mask}}$ values with the main model. We choose the largest county by population, Los Angeles County, both to ensure adequate signal for the model and also because Los Angeles has experienced a resurgence of the virus as of August 2, 2020.[1] For validation, we use the model from Equation 1 with the NPI effects $\alpha_1, \ldots, \alpha_8$ fixed to the mean values estimated by the main model for each cluster, as shown in Table 3. For Validation 4, which uses α_i values from cluster 4, we exclude L.A. county from the model fit. Additionally, we fit a "single-county" model to Los Angeles County on its own, without fixing NPI effects. Figure 9 gives an overview of this validation process. The effect of mask mandates is not fixed for any model, since we consider mask mandates to be county-specific. As a result, each validation model is able to converge by

manipulating the effects of masks, showing similar overall trends for the course of the outbreak in that county. However, they may estimate very different values for $\alpha_{\texttt{mask}}$ when we take the α_is from clusters 1, 2, 3, and 5, which L.A. county does not belong to. For instance, the Validation 1 and 3 models estimate $\alpha_{\texttt{mask}} = 0.193, 0.204$ respectively, to compensate for the difference in other NPI effects. When using fixed α_i's from cluster 4, on the other hand, the Validation 4 model estimates $\alpha_{\texttt{mask}}$ to be 0.018, much closer to the main model's estimate. This highlights how NPI can be highly specific to each cluster, since using the wrong values for α resulted in markedly different fatality estimates and other parameters. Clustering counties provides ample signal for the model to estimate NPI effects in each cluster while distinguishing between regions that exhibit markedly different values for these effects. This shows the advantages of clustering counties based on the variables we describe in section 4.2

Model	Fixed NPI Effects	Counties for NPI Effects (if Fixed)	Estimated R_0 (95% CI)	$\alpha_{\texttt{mask}}$ (95% CI)	Fatality Estimates Mean Error (std)
BMM 4	✗	—	3.380 (3.216, 3.543)	0.040 (0.0, 0.299)	11.5 (10.3)
Single-county	✗	—	4.634 (3.285, 6.787)	0.021 (0.0, 0.124)	11.2 (10.5)
Validation 1	✓	Cluster 1	3.739 (3.710, 3.922)	0.193 (0.058, 0.328)	12.2 (10.5)
Validation 2	✓	Cluster 2	3.514 (3.428, 3.607)	0.074 (0.0, 0.197)	11.4 (10.2)
Validation 3	✓	Cluster 3	3.321 (0.088, 0.329)	0.204 (0.088, 0.329)	11.8 (10.2)
Validation 4	✓	Cluster 4 (except L.A. County)	3.427 (3.358, 3.507)	0.018 (0.0, 0.153)	11.7 (10.2)
Validation 5	✓	Cluster 5	3.808 (3.710, 3.922)	0.046 (0.0, 0.175)	12.2 (10.1)

Table 6: Mean error and standard deviation for fatality estimates in Los Angeles County, using the Cluster 4 model as well as validation models. Validation models are fit to L.A. county on its own, either in the same manner as described in Section 5 or as described above, with fixed NPI effects. Those with fixed NPI effects still estimate the effect of mask mandates, as this is county-specific, and these are reported as well.

Figure 10 shows the fatality estimates for Los Angeles County, using the Cluster 4 specialized model, a single-county model, and two validation models with fixed NPI effects. As expected, Figure 10a and 10b show similar trends for the disease, although the Cluster 4 model has a larger confidence interval, since its parameters are fit to a greater number of counties. The model in

Figure 10c uses fixed NPI effects, with $\alpha_1, \ldots, \alpha_8$ taken from a model fit to the same counties as 10, except Los Angeles County itself. This looks similar to the single-county model trained on Los Angeles County, indicating the compatibility of estimated NPI effects with those learned across the whole cluster. This contrasts with Figure 10d, which uses NPI effects from cluster 1, resulting in a worse fit in terms of mean error.

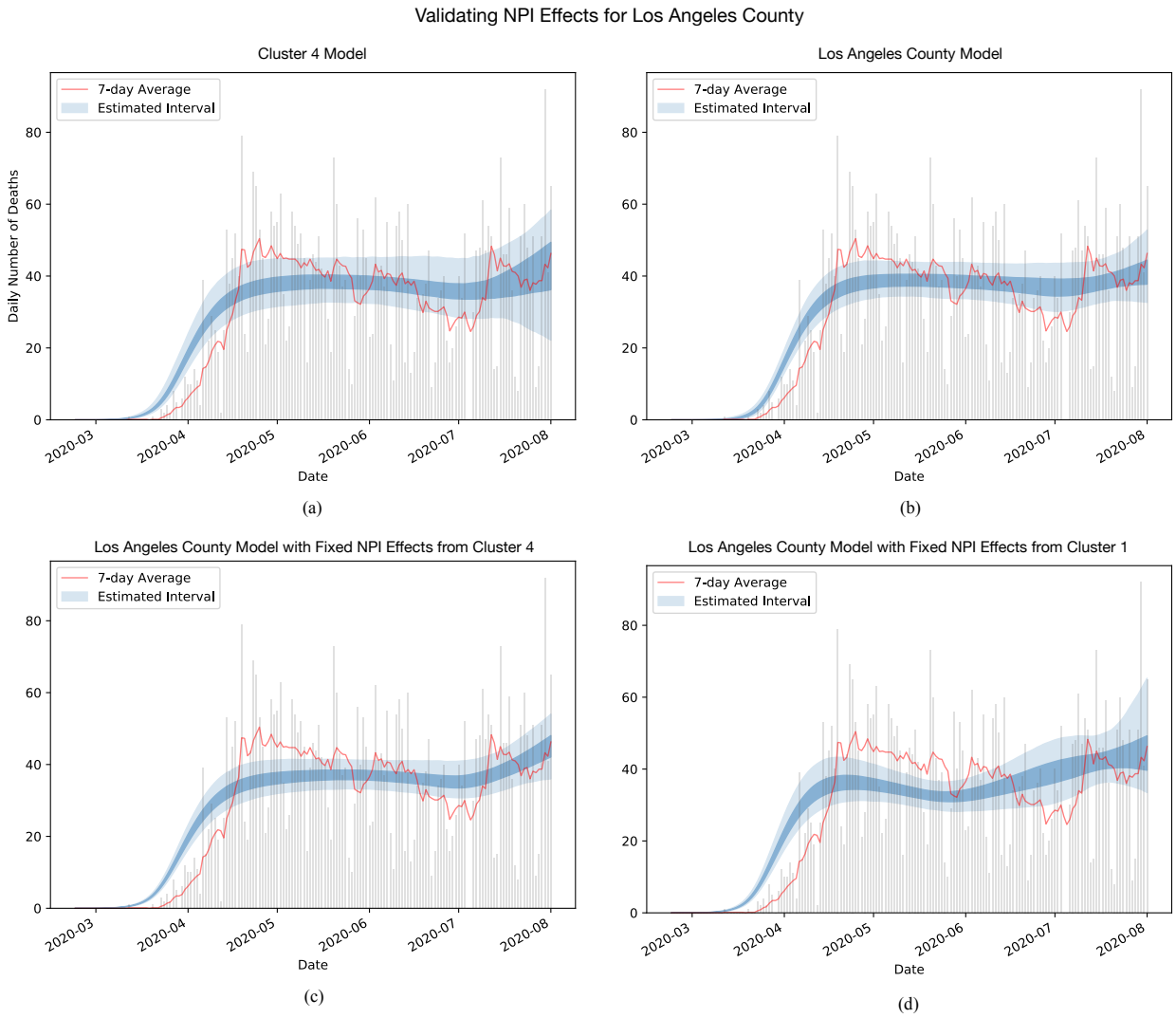

Figure 10: Estimated fatalities in Los Angeles County up to August 2, 2020. **(a)** estimates from the model described in Section 5, fit to counties from Cluster 4 (see Section 4.2. **(b)** estimates from a fixed-NPI model, using α_i values from Cluster 4, excepting Los Angeles County. **(c)** estimates from a fixed-NPI model, using α_i values from Cluster 1, which does not include Los Angeles County.

A.2 Correlating Estimated Reproductive Ratio with Mobility

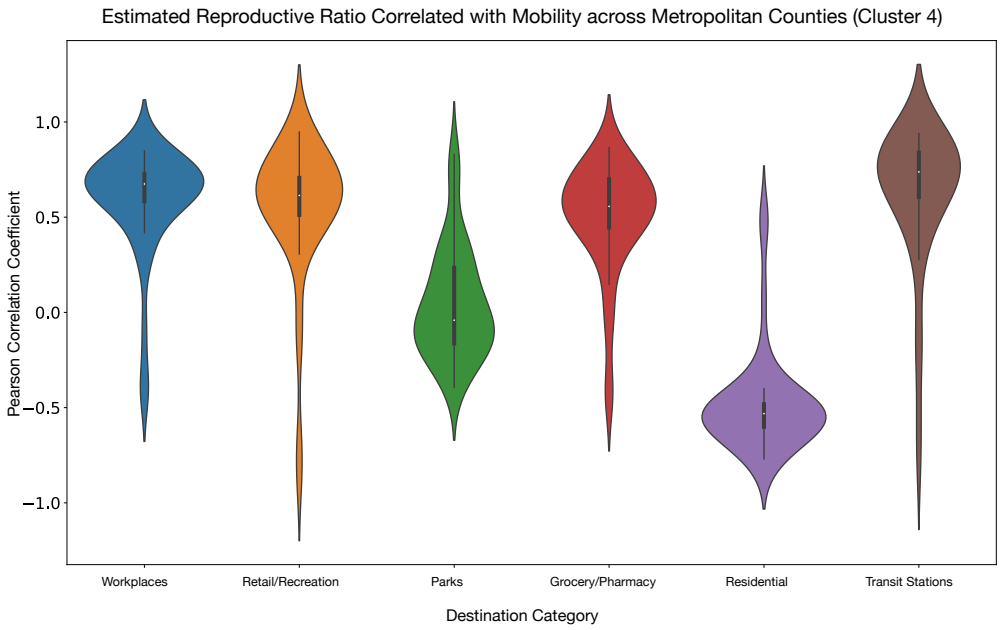

Figure 11: Distribution of correlations between the estimated reproductive ratio R_t and the number of visits to various destination types, collected from smart phone data.[33] Higher mobility in residential areas are negatively correlated with the R_t estimates, which is consistent with the expectation that staying home reduces disease spread. Visits to workplaces, retail, and transit stations, on the other hand, are positively correlated with R_t.

Destination Category	Mean Correlation (std)
Workplaces	0.195 (0.289)
Retail/Recreation	0.206 (0.290)
Parks	-0.002 (0.378)
Grocery/Pharmacy	0.246 (0.226)
Residential	-0.175 (0.347)
Transit	0.186 (0.343)

Table 7: Pearson correlation coefficient between mobility in each destination category and estimated reproductive ratio, based on data up to August 2, for metropolitan counties (Cluster 4).

Although we do not incorporate mobility data into our model, we nevertheless observe a strong correlation between the estimated reproductive ratio and activity outside the home, based on data up to August 2. For each county, we compute the Pearson correlation coefficient of R_t with daily mobility data for six destination categories: workplaces, retail/recreation, parks, grocery/pharmacy, and residential, and transit. Table 7 shows the mean and standard deviation of these correlation values across metropolitan counties (Cluster 4), and Figure 11 shows each distribution. As is evident, visits to the workplace have, on average, a strong positive correlation of 0.701 with the estimated reproductive ratio, followed by visits to retail/recreation and grocery/pharmacy. Visits to residential destinations suggests individuals are likely staying home, which has a negative correlation with our reproductive ratio, consistent with claims that stay-at-home orders and social distancing mitigates disease spread. Finally visits to parks, which are outdoors, have correlation values distributed about 0. This does not indicate visits to parks are free of disease spread, but it is consistent with public health guidance that open air activities carry less risk compared to indoor activities.

We make these observations not to establish the causal relationship between greater mobility and disease spread — that much has been broadly established — but rather as a validation of our

estimates, which are consistent with broadly established expectations despite mobility not being explicitly considered by our model.

B Disentanglement

I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
	6.76	8.79	10.4	9.49	7.99	12.48	17.48
6.76		1.85	4.07	4.13	5.37	5.89	10.83
8.79	1.85		4.86	4.57	6.01	5.33	8.86
10.4	4.07	4.86		3.74	4.84	2.38	7.37
9.49	4.13	4.57	3.74		1.9	3.82	8.68
7.88	5.27	6.01	4.84	1.9		4.89	9.79
12.48	5.89	5.33	2.38	3.82	4.89		5
17.48	10.83	8.86	7.37	8.68	9.79	5	

Table 8: Average number of days between intervention types for all counties in the US that have implemented the two interventions being compared. The interventions are in the order defined in Table 3

.

One drawback of a mechanistic model is that it cannot disentangle implementations that came into effect at the same time. For example, states often closed public schools at the same time as federal guidelines were issued so it is difficult to discern the individual effect on reducing R_t. Additionally, in Table 8, we observe that banning gatherings of 50 or more people often occurs at the same time as banning gatherings of 500 or more, and restaurants and entertainment venues are often closed together. This suggests that these pairs of interventions may not or be poorly disentangled. This holds true even as interventions are rolled back.

	Stay at Home	*>50 Gatherings*	*>500 Gatherings*	*Restaurant dine-in*	*Entertainment / Gym*
Stay at Home		15.07	15.38	13.93	14.54
>50 Gatherings	15.07		0.34	16.43	13.14
>500 Gatherings	15.38	0.34		14.90	12.99
Restaurant dine-in	13.93	16.43	14.90		1.28
Entertainment / Gym	14.54	13.14	12.99	1.28	

Table 9: Average number of days between rollbacks of interventions for all counties in the US that have implemented the two rollbacks being compared.

To further investigate the model's ability to disentangle intervention weights, we create simulated trajectories of counties' deaths and cases counts based on their R_0 and the dates on which the interventions came into effect. Using all counties that have more than 500 cumulative deaths on August 2without super-counties, we seed each county with 200 cases in each of the first 6 days. The higher threshold aims to limit the disentanglement analysis to counties that have more cases and are therefore likely to implement their own set of NPIs rather than follow the state timeline. To simulate county-specific trajectories, we construct a set of generated time series. We assign intervention weights α_i to be randomly generated from a Gamma distribution, the same distribution as our prior on the Bayesian mechanistic model adjusted to be in the range of our learned weights. To reduce the complexity in disentanglement, we do not use the county-specific mask term when generating the trajectories and do not fit to that term in our disentanglement model. We then calculate what the R_t on each day must have been based on the R_0 and the interventions in place. Once we have the seeded infection and the R_t trajectory for each county, we can calculate daily infections and thus expected fatalities. Since fewer regions are considered for disentanglement runs, we fit for 1000 warmup iterations, 1800 iterations in total, and use 4 chains. Using the simulated trajectories, we fit the model. Table 10 compares the weights used for generation with

the weights that the model learned.

	Intervention weights	Learned weights (95% confidence interval)
I_1	0.803	1.413 (1.211,2.541)
I_2	0.039	0.311 (0.017,0.531)
I_3	0.066	0.021 (-0.008,0.215)
I_4	0.329	0.468 (0.087,0.805)
I_5	0.115	0.011 (-0.008,0.096)
I_6	0.013	0.010 (-0.008, 0.113)
I_7	0.472	0.045 (-0.008, 0.428)
I_8	0.055	0.008 (-0.008, 0.114)

Table 10: By setting the intervention weights, we can generate simulated timeseries of cases and deaths counts and have the model learn the weights. The learned values differ from the ground truth intervention weights, showing that the model does not disentangle the contribution of each intervention well in cases where interventions were implemented close together.

We observe that the effects of individual NPIs are not well disentangled in general. The model tends to attribute more weight to few NPIs rather than spread out the weight evenly. Specifically, the model tends to put more weight on stay-at-home orders. This may be because interventions I_2 to I_8 are often implemented close together (see Table 8) and it is difficult to attribute effect to any single one of them on a national scale. Although rollbacks are implemented with more variability in time, they show similar groupings as the NPI implementation between ¿50 and ¿500 gatherings, and between restaurant and entertainment/gym re-openings. The region-specific mask factor further entangles the rollback effects. While we can conclude that the trajectories the model predicts are reliable, due to their match to measured death, and therefore the overall change in R_t is reliable, attributing variation to any individual NPI is challenging whenever the difference

in implementation date is small. This observation seems to be in line with the similarly large confidence intervals reported in previous work on varied models and regions.[2–6,8]

Figures

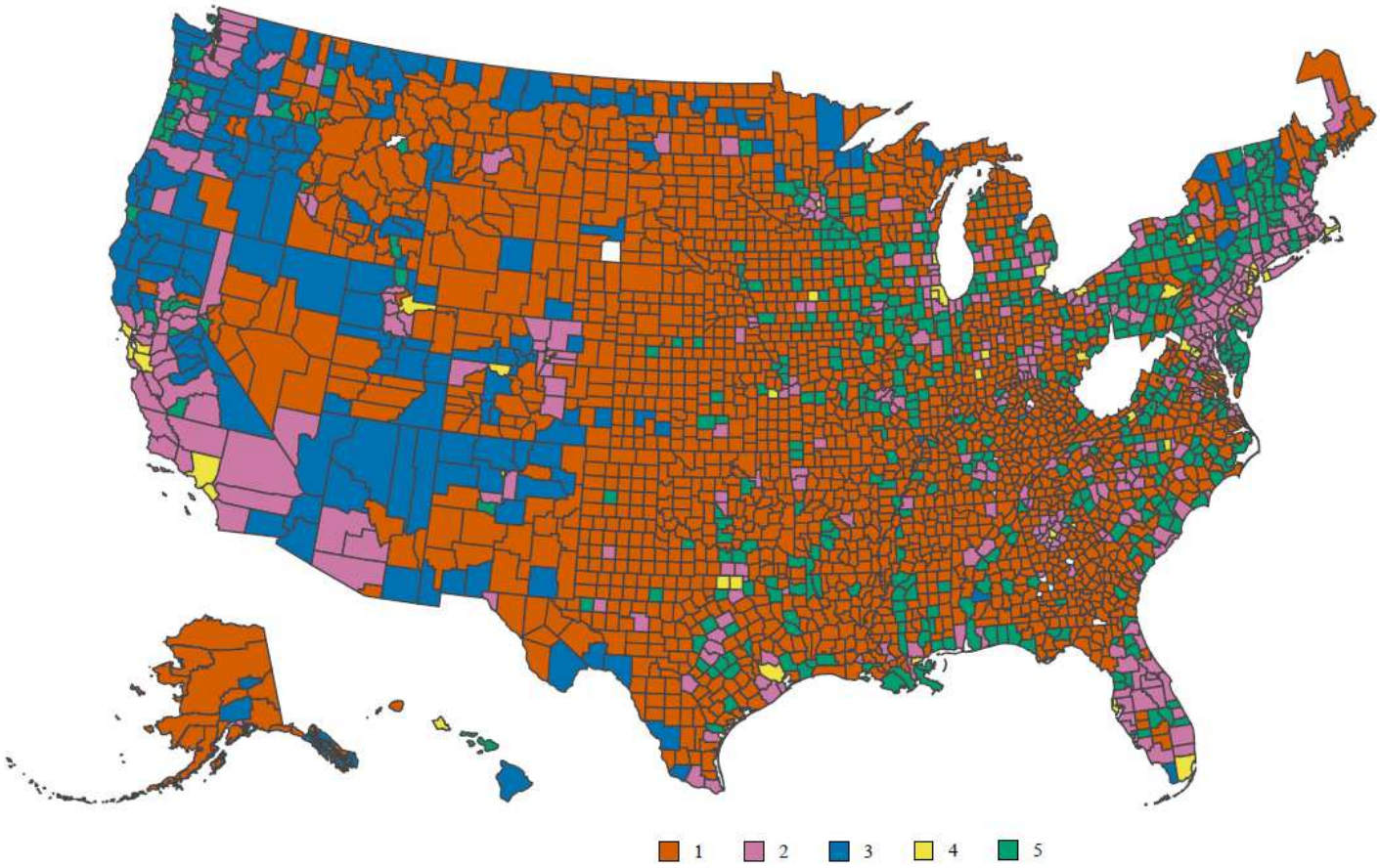

Figure 1

Cluster labels based on demographic and socioeconomic conditions are used to aggregate data and specialize epidemiological models. Here, one can see how cluster 1 and 3 primarily cover rural areas, while clusters 5, 2, and 4 consist of increasingly urban counties.

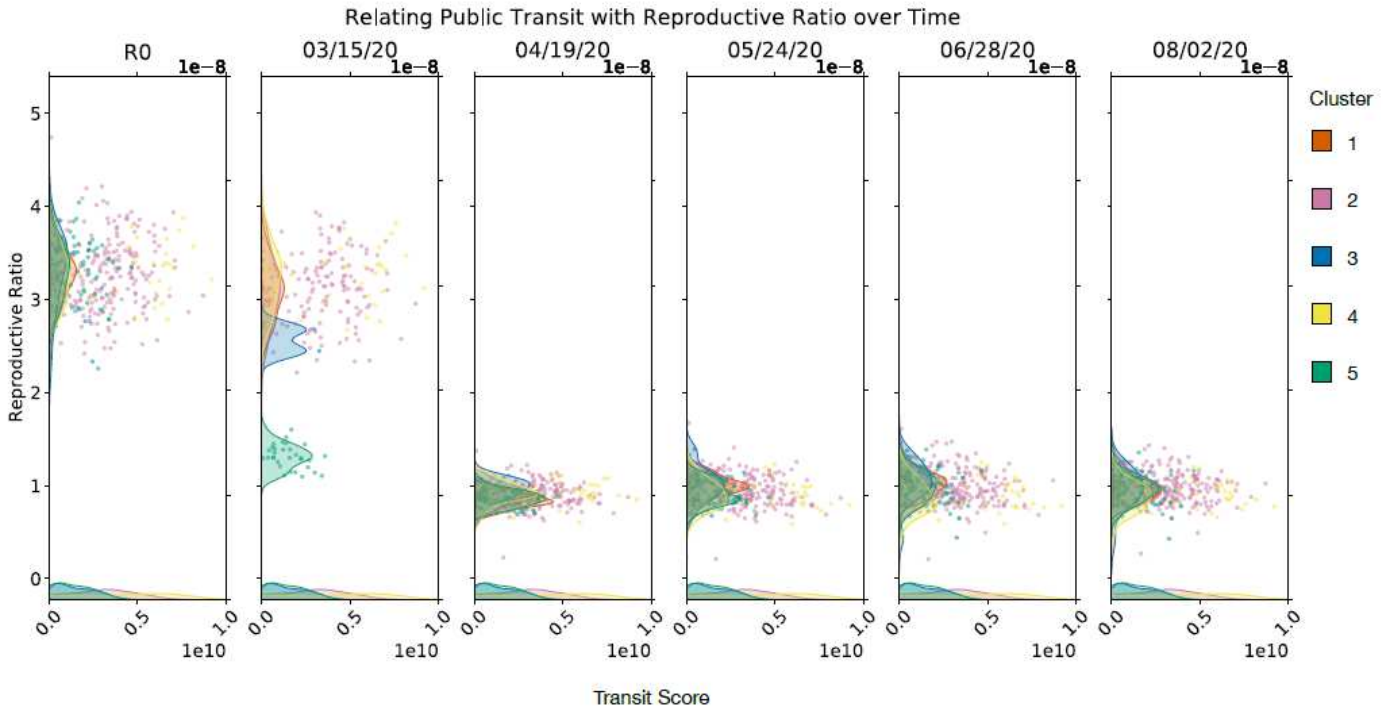

Figure 2

Relationship between public transit capacity and the time-dependent reproductive ratio of SARS-CoV-2 for U.S. counties, through August 2, 2020. In Clusters 2 and 5, which have lower transit scores than cluster 4, we observe a noticeable drop in reproductive ratio by March 15, indicating the stronger effect of advisory NPIs in effect at that time.

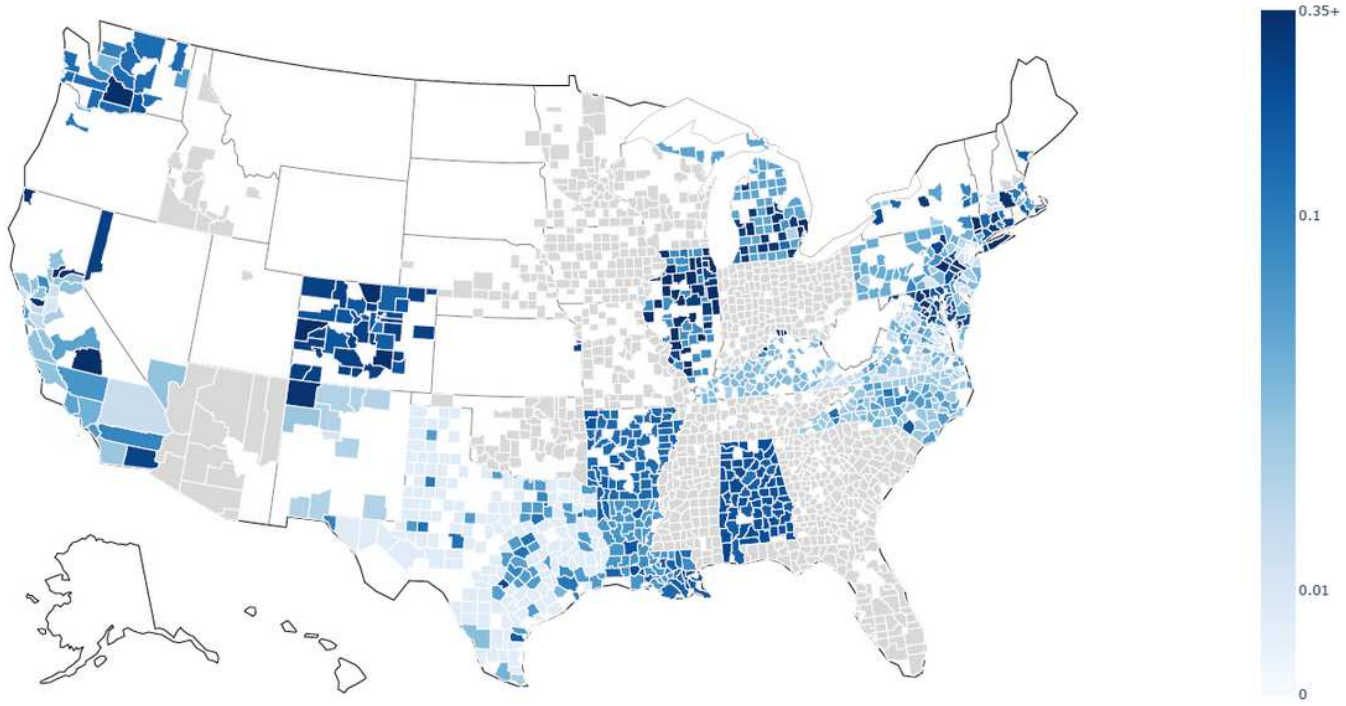

Figure 3

The quantified effectiveness αmask;m of mask mandates for each county included in our model. Counties which never implemented mask mandates, as of August 2, but are still included in our model are shown in gray. Darker colors correspond to higher values, where masks mandates were estimated to have greater effect toward reducing the reproductive ratio. Since mask mandates are implemented universally across a given state, these values are best interpreted in comparison with counties from the same state.

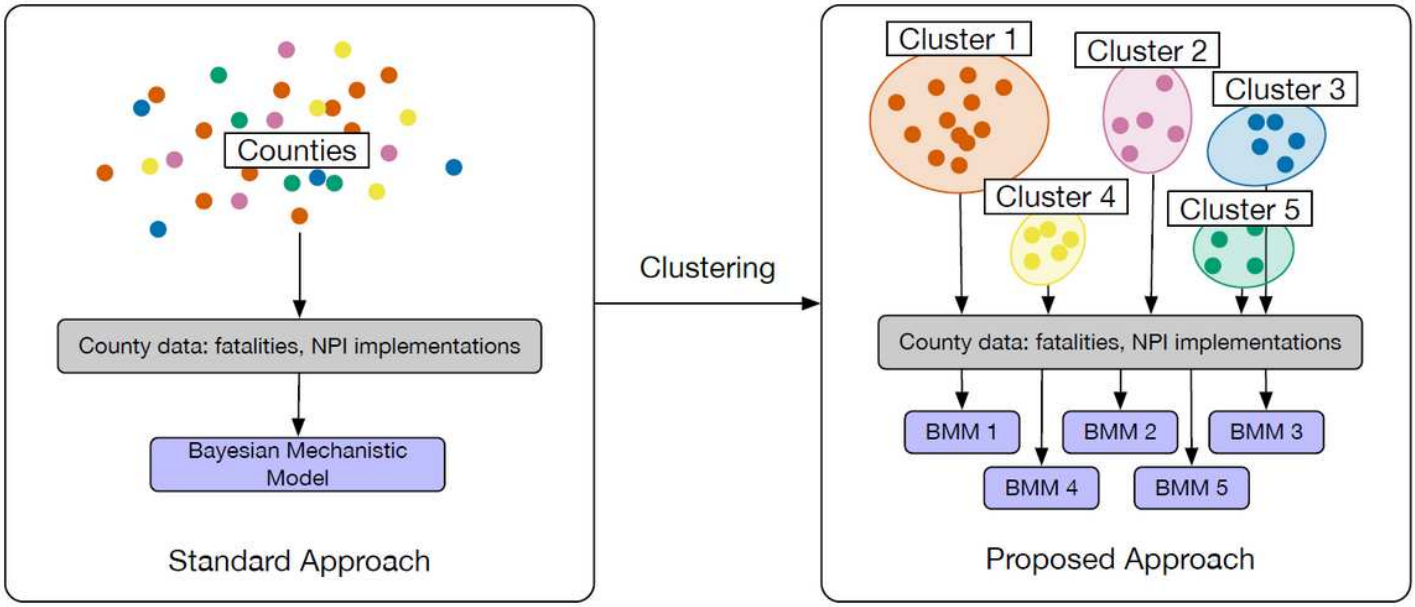

Figure 4

The fitting process for our model, compared to the standard approach. We fit Bayesian mechanistic models (BMM 1-5) to each one of five clusters of counties, based on similarity in socioeconomic, demographic, and other factors.

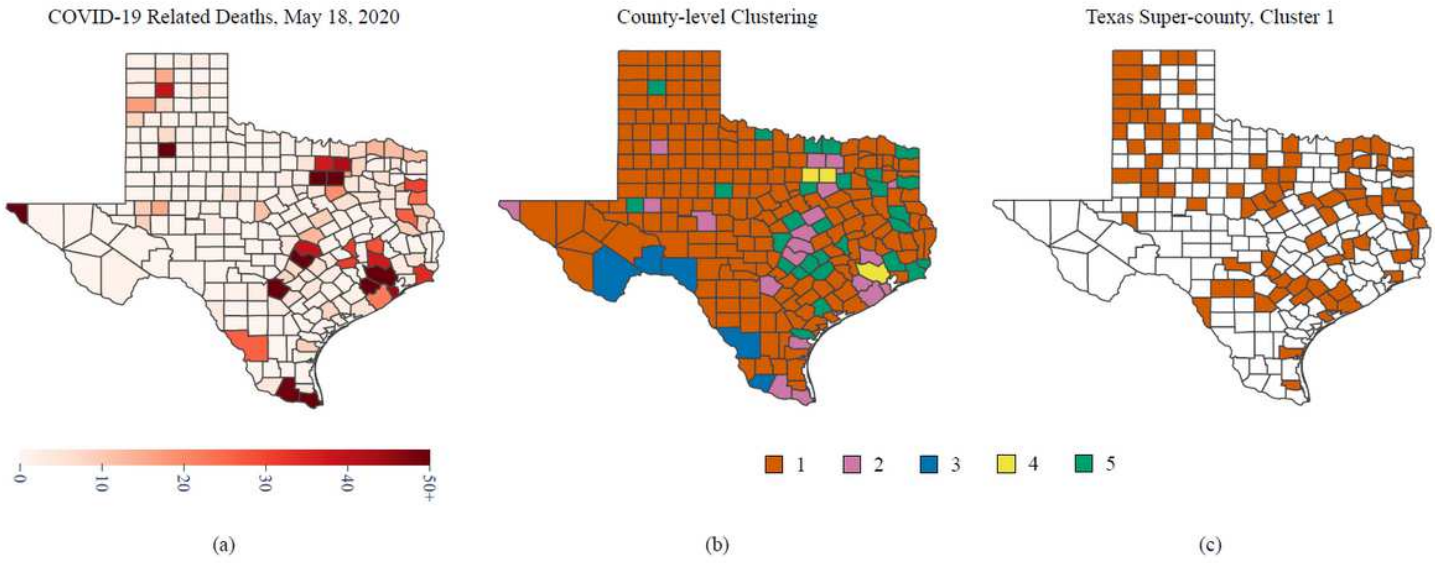

Figure 5

(a) The total confirmed deaths caused by COVID-19 for counties in Texas, as of August 2. (b) Cluster labels for each Texas county, based on demographics, education, density, and other factors. (c) Texas counties in cluster 1 having 1-49 cumulative deaths as of August 2, 2020 and the same NPI implementation dates. To enable robust epidemiological models, these counties are treated as a single “super-county.” A county or super-county must have 50 or more cumulative deaths as of August 2 to be considered.

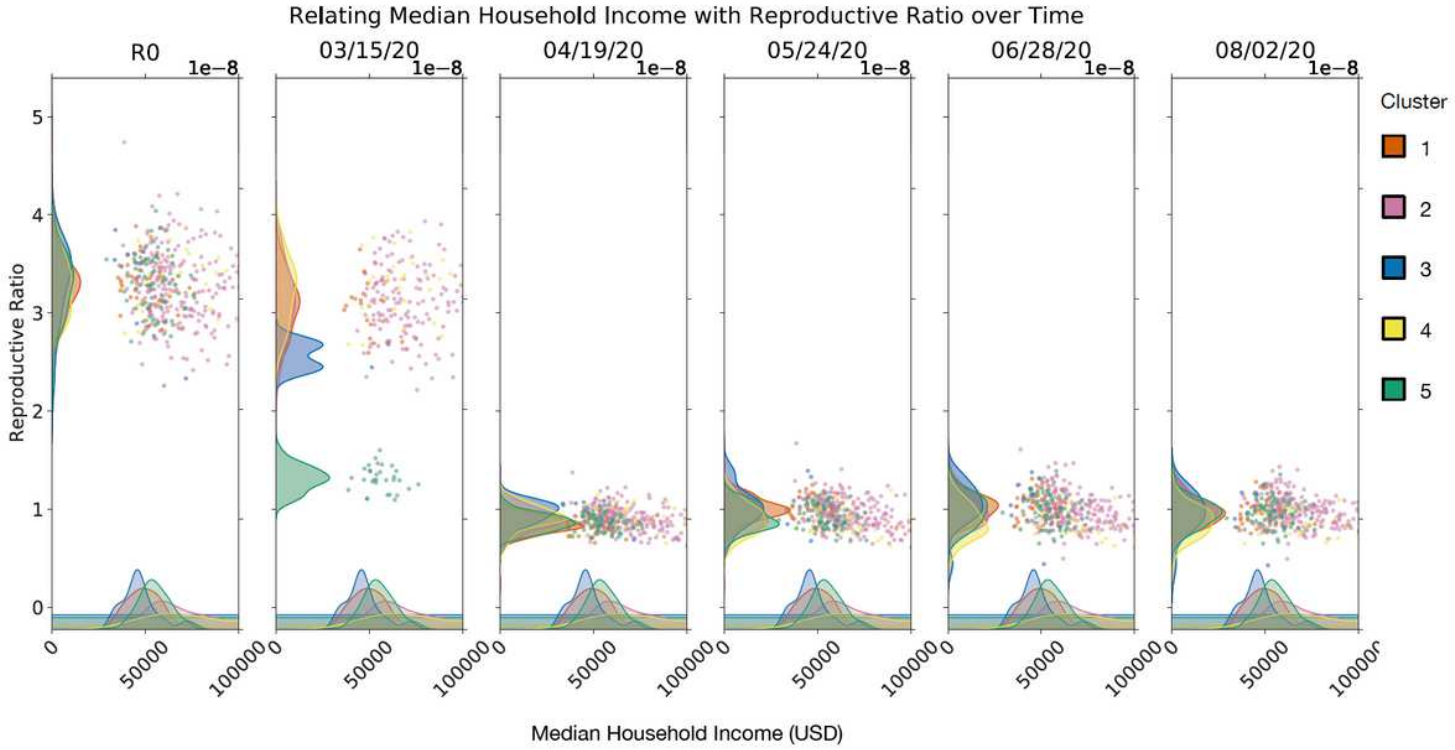

Figure 6

Scatter plot and density distribution plot for counties and super-counties comparing Rt over time to median household income. Colors indicate which cluster the county or super-county belongs in.

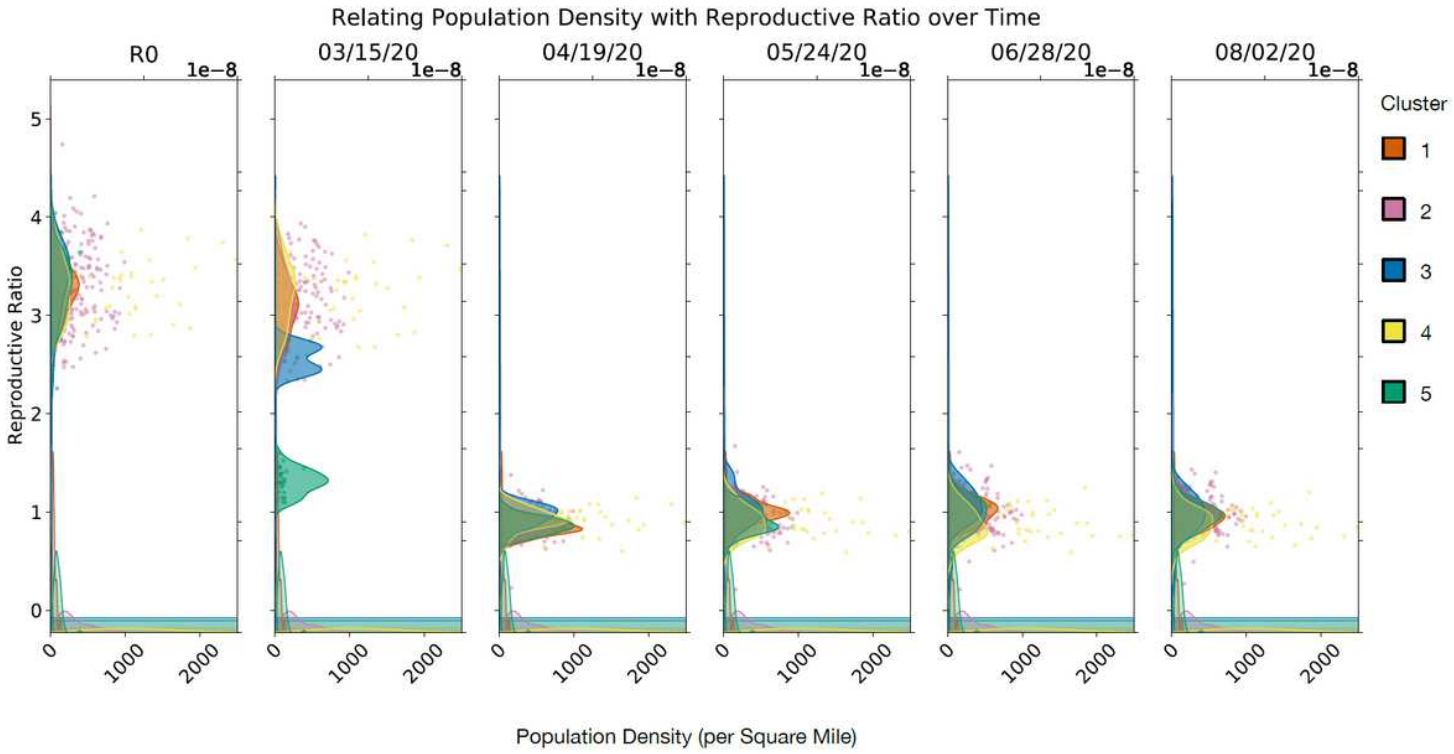

Figure 7

Scatter plot and density distribution plot for counties and super-counties comparing Rt over time to population density. Colors indicate which cluster the county or super-county belongs in.

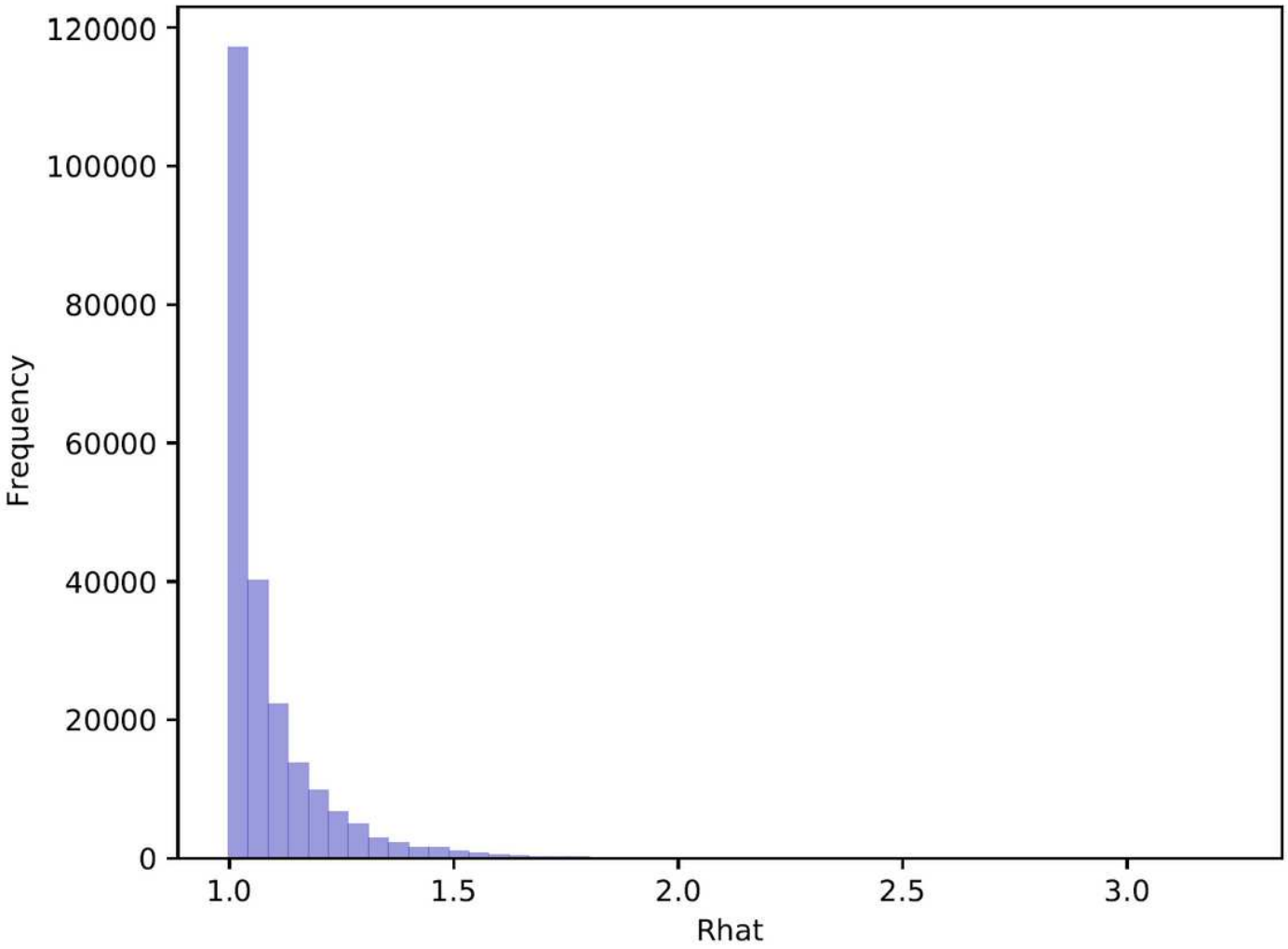

Figure 8

Histogram of Rhat for the Pystan model fit of cases and deaths for national-level counties and super-counties. Values close to 1 indicate convergence.

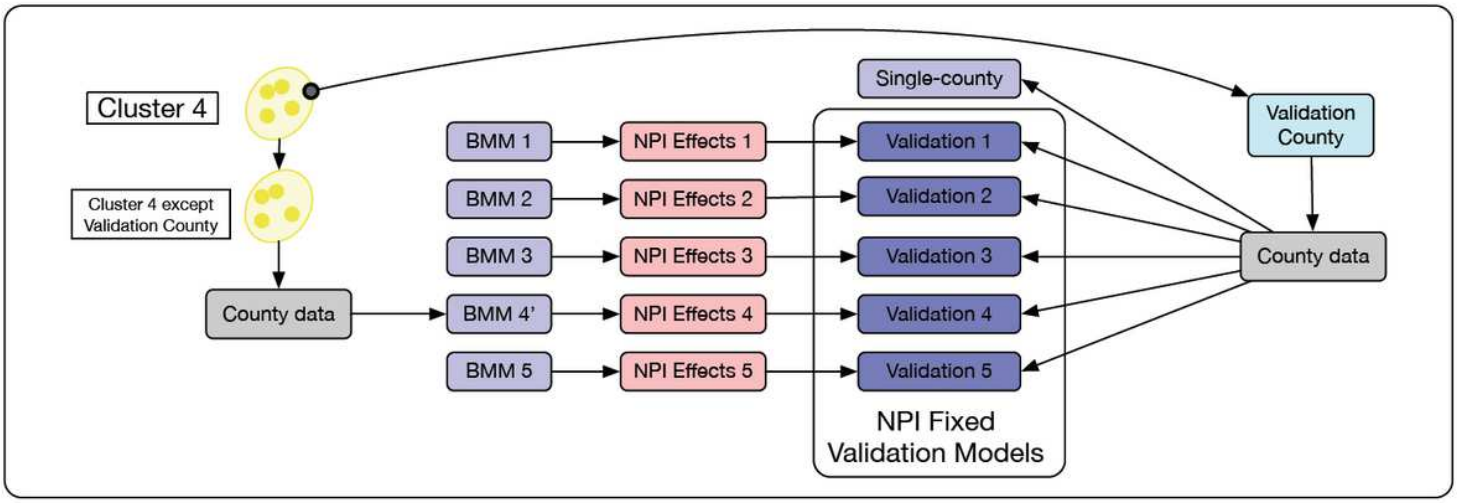

Figure 9

An overview of our process for validation the advantages of clustering when estimating the effects of NPIs. Validation models (dark blue) use NPI-effects from each cluster-specialized model. We expect the estimates from Validation 4 to closely match those of BMM 4, but others may vary significantly. BMM 1, 2, 3, and 5 match the models from Figure 4. BMM 4' is similar to BMM 4, with the difference being the exclusion of the validation county during fit.

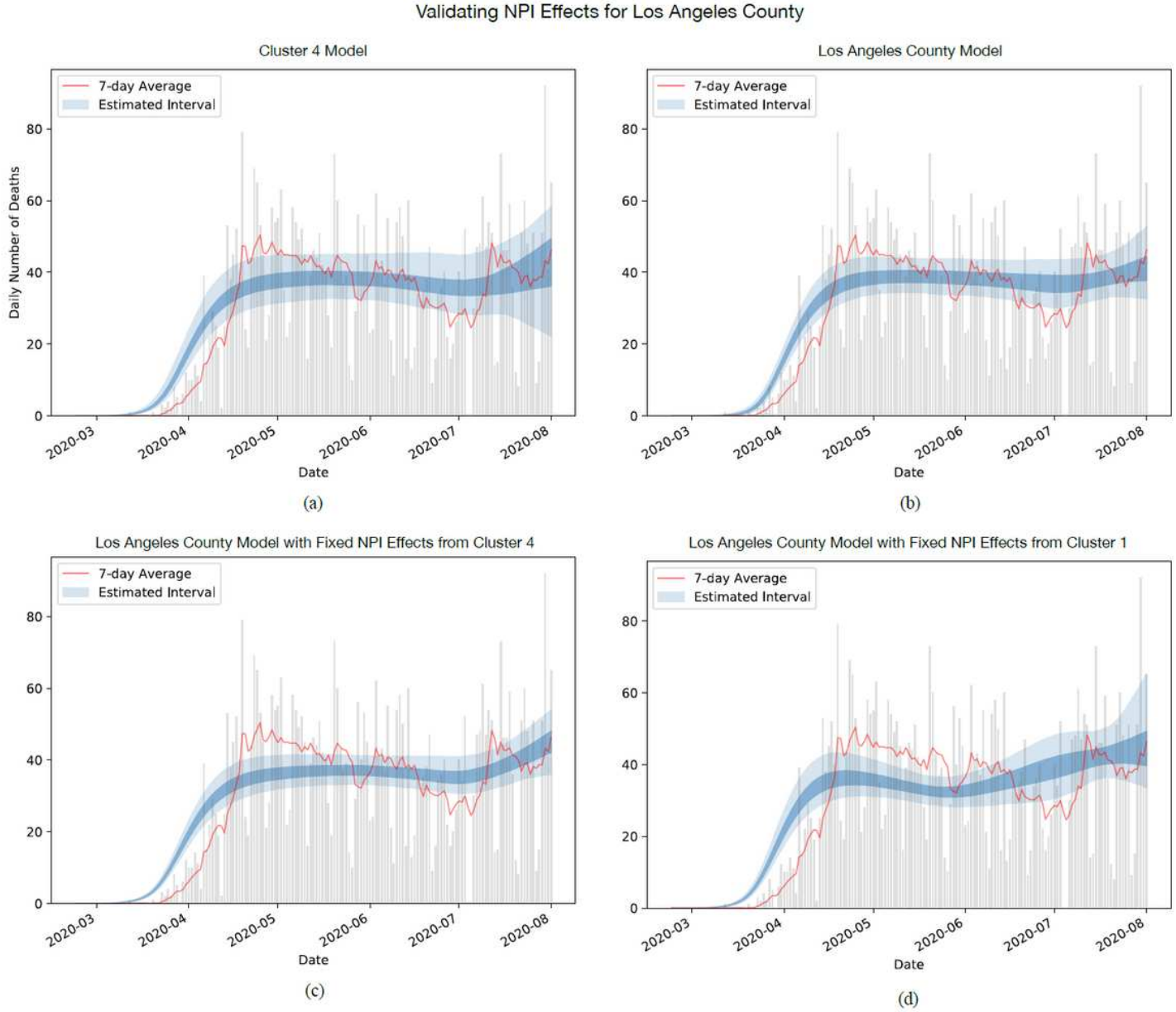

Figure 10

Estimated fatalities in Los Angeles County up to August 2, 2020. (a) estimates from the model described in Section 5, fit to counties from Cluster 4 (see Section 4.2. (b) estimates from a fixed-NPI model, using αi values from Cluster 4, excepting Los Angeles County. (c) estimates from a fixed-NPI model, using αi values from Cluster 1, which does not include Los Angeles County.

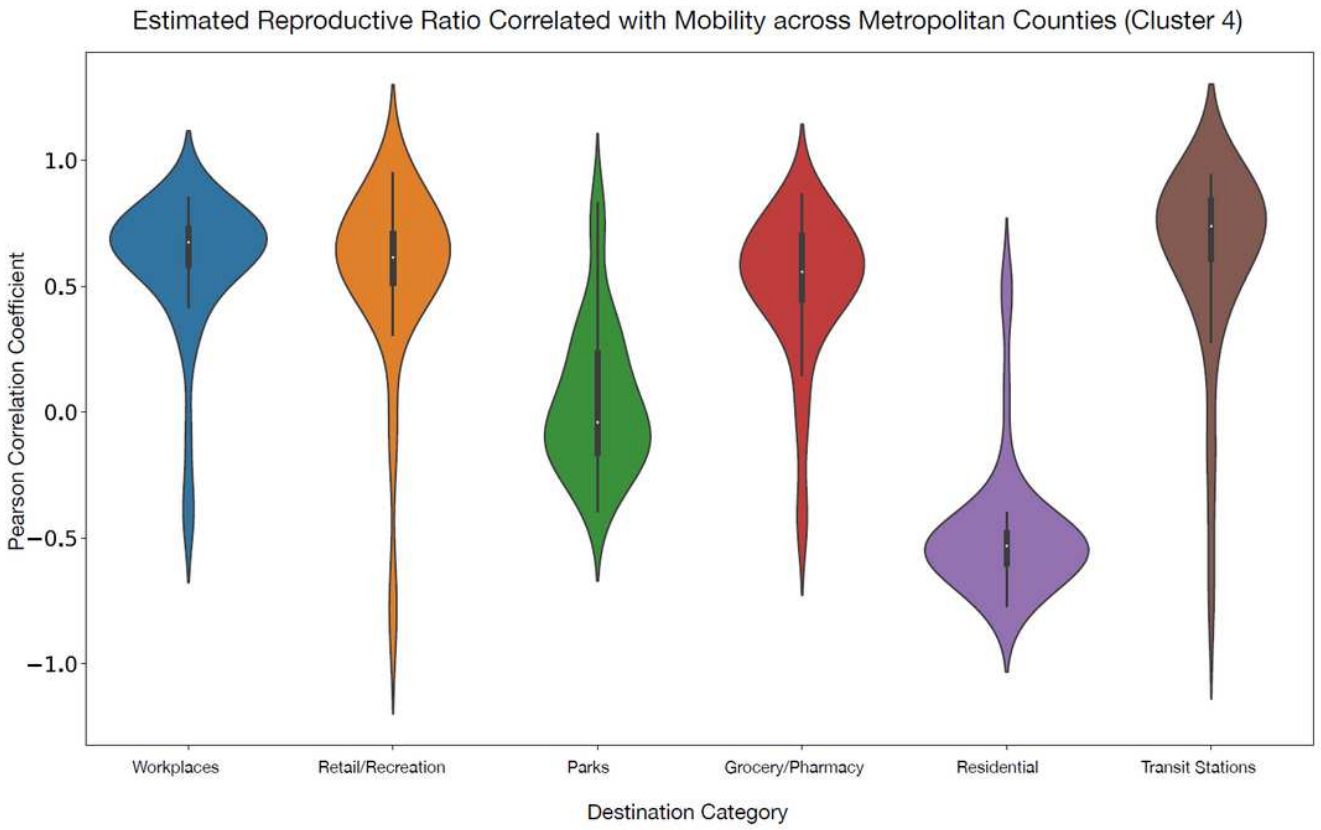

Figure 11

Distribution of correlations between the estimated reproductive ratio Rt and the number of visits to various destination types, collected from smart phone data.33 Higher mobility in residential areas are negatively correlated with the Rt estimates, which is consistent with the expectation that staying home reduces disease spread. Visits to workplaces, retail, and transit stations, on the other hand, are positively correlated with Rt.