The University of Manchester
Global Development Institute

Global Development Institute

Working Paper Series

2021-053

April 2021

ISBN: 978-1-912607-11-2

Birds of a feather flock together? Diversity and the spread of Covid-19 cases in India

Udayan Rathore[1]

[1] Oxford Policy Management, UK

Email: rathore.udayan@gmail.com

Upasak Das[2]

[1] Global Development Institute, The University of Manchester, UK

Email: upasak.das@manchester.ac.uk

Prasenjit Sarkhel[3]

[1] Department of Economics, University of Kalyani, India.

Email: prasenjitsarkhel@gmail.com

Cite this paper as:

Rathore, U., Das, U. and Sarkhel, P. (2021) Birds of a feather flock together? Diversity and the spread of Covid cases in India. GDI Working Paper 2021-053. Manchester: The University of Manchester.

Abstract

Arresting Covid-19 infections requires collective action that is difficult to achieve in a socially and economically diverse setting. Using district-level data from India, we examined the effects of caste and religious fragmentation, along with economic inequality, on the growth rate of reported Covid-19 cases. Our findings indicate the positive effects of caste homogeneity, while demonstrating a limited impact of economic inequality and religious homogeneity. The gains from higher caste homogeneity eroded gradually with the unlocking procedure after the nationwide lockdown but community cohesion through caste remained dominant in rural areas even when mobility restrictions were withdrawn. Our findings indicate that planners should prioritise public health interventions in areas that are heterogeneous in terms of caste to compensate for the absence of community cohesion. The importance of our study lies in empirically validating the causal pathway between homogeneity and infection growth, thereby providing a basis for zoning infection-prone areas and advocating a differentiated policy response.

Keywords

Caste, Covid-19, diversity, unlock, social cohesion, India

JEL Codes

D70, I12, I18

Acknowledgements

The authors acknowledge Sushmita Chakraborty for the assistance received in generating the maps and Ritabrata Bose for helping with initial collation of the data.

1 Introduction

Adherence to social distancing is widely recommended across the globe as a protective measure against novel coronavirus (Covid-19). Ensuring the practice of compliance norms requires coordinated community action, the more so because private protective efforts involve positive externality. Unless compliance protocols are followed by other members of the community, the effectiveness of an individual effort is negligible. However, a large body of literature on community collective action, albeit in a different context like common property resources, has shown that major challenges to consensus building and norm enforcement may arise from differences in social class and ethnicity (Blair, 1996; Poteete & Ostrom, 2004). In such a setting, does diversity in the community pose a significant obstacle to arresting the spatio-temporal spread of infection? What is the nature of the heterogeneities or diversity that matter? Are social, religious and economic differences equally important or do some homogeneities matter more than others?

We posed these interrelated questions in the context of India, which had the third highest Covid-19 case-load at the time of writing,[1] and is one of the most diverse countries in the world in terms of social groups, religions and economic class (Munshi, 2019). In particular, we studied the implications of social, religious and economic diversity on the growth in the number of reported infections, starting from the beginning of the nationwide lockdown (25 March 2020), which was implemented in four phases across about two months: Phase 1: 25 March to 14 April; Phase 2: 15 April to 3 May; Phase 3: 4–17 May; Phase 4: 18–31 May. Thereafter, the unlocking process began and accordingly we considered four additional phases for our analysis, each consisting of 14 days and spanning two months: Phase 5: 1–14 June; Phase 6: 15–28 June; Phase 7: 29 June to 12 July; Phase 8: 13–26 July. This time period also allows us to examine the effects of the unlocking of the economy on the growth in the number of cases, and the associated implications of caste- and religion-based diversity along with economic inequality.

To estimate the implications of heterogeneity on the infection rate, we made use of a novel district-level dataset that matched high-frequency Covid-19 infection and mortality data with covariates of social groups and economic class, along with crucial information on the practices considered to be important transmission vectors for the virus. Our dataset also accounts for households' past mobility patterns. Besides household characteristics, we included supply-side data on health infrastructure and other amenities. The district-level data on Covid-19 cases (outcome) for the period of the study came from the Development Data Lab's (DDL) Covid-19 India database (Asher et al, 2020). The data for district-level controls like household characteristics and health facilities were obtained from multiple recent rounds of the National Sample Survey Office (NSSO), which are arguably among the most credible official surveys for India. To the

[1] As of April 8, 2021, close to 13 million Covid-19 cases have been reported in India, with more than 162,000 deaths.

best of our knowledge the dataset is one of the first in the Covid-19 context to combine multiple sets of information at the district level in India.

The idea that diversity could undermine voluntary cooperation for the public good, which in our case is adherence to compliance protocols, has been a persistent theme across developing countries (Banerjee et al, 2005; Habyarimana et al, 2007). The underlying mechanism that drives the relationship is understood through a number of channels. First, the preference link, where individuals of a particular group care more for in-group members than for outsiders, may become pertinent (Tajfel et al, 1971; Vigdor, 2004; Alesina & Ferrara, 2005). Second, homogeneous societies, by virtue of higher social interactions, often possess greater levels of social capital, which influences the expected reciprocity from others, reflecting higher levels of interpersonal trust. Accordingly, stronger social ties and networks ensure homogeneous communities can monitor violations more efficiently and hence are able to impose social sanctions effectively, which can potentially lead to improved collective action (Banerjee et al, 2005; Miguel & Gugerty, 2005). Finally, greater diversity is also associated with divergence in preferences and varied perceptions of risk that may lead to sub-optimal levels of compliance (Alesina et al, 1999; Munshi & Rosenzweig, 2018). However, given the virus's highly infectious nature, it is also possible that there are fewer interpersonal interactions in diverse communities, as a result of which the chances of the virus spreading may be lower in more heterogeneous communities (Bosancianu et al, 2020). This countervailing possibility turns the potential relationship between diversity and the spread of infection into an ambiguous one and thus necessitates empirical investigation.

One of the main dimensions of community diversity in this paper comes from caste, the importance of which lies in its stratification of the social architecture in India (Deshpande, 2000). Dating back to 1500–500 BCE, the Indian caste system comprises over 4000 social categories, making it one of the most socially diverse countries in the world (Munshi, 2019). In this paper, we define heterogeneity in terms of four broad caste groups: the Upper Castes (Others); the Scheduled Castes (SC, formerly 'untouchables'); the ethnic and tribal groups belonging to the Scheduled Tribes (ST); and Other Backward Classes (OBC). As Munshi (2019) describes it, these groups are delineated by a distinct spatial segregation that extends over them in terms of networks for information sharing and social identity, and is likely to play an important role in the protective effort against the spread of infection.

Besides caste, another important social dimension likely to guide group behaviour is religion (Bandiera et al, 2005). Ever since India gained independence in 1947, entailing a violent partition along religious lines, conflicts, especially those between Hindus and Muslims, have also been viewed as a source of social cohesion or fragmentation (Field et al, 2008; Susewind, 2017; Mitra & Ray, 2014). Incidentally, in the context of Covid-19, with Hindus being the dominant religious group in India, Muslim minorities have often been held responsible for violating the social distancing protocols.[2] Therefore, examining the effects of religious heterogeneity, along with the diversity associated with caste, both of which are deeply embedded in societal affairs, is likely to have important social implications regarding the spread of the disease, especially in terms of providing important policy pointers for accelerating medical interventions. Further, we have also considered economic diversity, manifest through inequalities in household consumption expenditure, and explored its relationship with the spread of the infection. Higher economic inequality is likely to reduce group solidarity and act as a deterrent to collective provisioning (Wade, 1994; Dayton-Johnson, 2000). In other words, we posit that notions of trust and coordinated community action can go beyond social and religious affiliations and might be aligned along economic lines with respect to adherence to the Covid-19 compliance protocols.

While estimating the relationship of interest, we can negate the possibility of bias from reverse causation entailed by residential self-selection because of the exogenous and unprecedented nature of the pandemic. Instead, the central challenge in ascertaining a causal relationship comes from the possibility of endogeneity, which may stem from Omitted Variable Bias (OVB).[3] This might happen, for instance, if caste and religion are occupation-specific and influence contact behaviours as well as the nature of social interaction. Despite controlling for various confounders, it is possible that we are unable to account for such unobservables simultaneously correlated with both the variable of interest and the outcomes, thereby inducing bias in our findings. To account for such possibilities, we undertook the following measures. First, we controlled for the Covid-19 cases reported in a district before the first phase of the lockdown (25 March 2020). This is critical for addressing the selection on unobservables as it subsumes the cumulative influence of initial hotspots, superior international connectivity, and integration of certain districts into the global supply chain, among other things, all of which are likely to be associated with residential patterns that may shape measures of heterogeneity at the district level. For example, note that Covid-19 did not originate in India,[4] but entered the country through districts and cities that facilitate international travel and handle large volumes of passengers, especially from China, Europe and the Middle East.[5] These districts are also likely to be better integrated into global supply chains and thereby to be favoured employment destinations across social, religious and economic lines. Such factors may result in higher or lower heterogeneity, depending on socioeconomic or socio-religious mobility. Such an effect would confound with the relationship of interest, thereby leading to biased estimates. Second, to account for the dynamic supply-side responses for each phase of the lockdown, we controlled for number of deaths in the

---

[2] https://www.orfonline.org/expert-speak/covid19-indian-muslims-69519.

[3] Given that Covid-19 was an exogenous shock, endogeneity via reverse causality is less of a concern.

[4] The virus first appeared in Wuhan, China, with the Chinese authorities confirming the possibility of human-to--human transmission in late January. See https://www.nytimes.com/2020/01/20/world/asia/coronavirus-china-symptoms.html.

[5] The first few cases in India were recorded in the state of Kerala and the cities of Delhi and Mumbai, among others with high numbers of international passengers and transnational migrants.

district for the previous period in all the regressions, in addition to the district-level physical and human health infrastructure. Here, a plausible assumption is that the district and other levels of administration are likely to be more sensitive to the number of Covid-19 deaths in the previous period when formulating or reviewing a policy response. Third, for all the regressions, we took intra-state and region fixed effects to account for any region-level unobserved characteristics that might be correlated with both the variables of interest and the outcome. Specifically, this variable and the time/period dummies should control for the variation in Covid-19 testing that is likely to be similar within but vary across these regions with time. Finally, we took the strategy used by Altonji et al (2005) and then Oster (2019) to examine the bounds of our coefficients with a sizeable and standard assumption of selection on unobservables. These steps ensured that the influence of unobservable confounders was accounted for and the coefficients of homogeneity derived from the main regression estimations could be viewed via a robust causal lens, which aids the devising of important policy instruments.

Our findings indicate, *ceteris paribus*, that districts with higher caste-group homogeneity are likely to experience lower growth in the number of Covid-19 infections on average. The relationship becomes stronger through extended phases of the lockdown but weakens with the beginning of the unlocking process. The overall relationship remains statistically significant for the first month of unlocking and dissipates thereafter. We also found a modest but positive effect of economic equality on the spread of infection in the first three phases of the lockdown, thereby underscoring the importance of social ties based on economic lines. We found no evidence of a significant relationship with religious homogeneity in a similar context. These results are statistically valid for a range of robustness checks. We also examined the implications of the unlocking procedure on the relationships of interest. Our findings indicate a significant influence of the unlocking process, with the gains of caste-group homogeneity attenuating once the lockdown was lifted. Interestingly, however, for the districts that were predominantly rural, we found that these gains increased in the unlocking phase. This heterogeneous finding potentially underscores the importance of social cooperation and ties, which are likely to be much stronger in rural areas than in the more urbanised ones.

Our main contribution through the paper is two-fold. First, we contribute to the growing literature on the relationship between social fragmentation and the provisioning of public goods. Importantly, we document this relationship in the context of a unique public good that is tied to public health and compliance with Covid-19 protocols. This has far-reaching implications for preserving both lives and livelihoods in the context of the post-pandemic world order. Second, from a policy perspective, our findings imply a need to identify socially heterogeneous geographical pockets as potential hotspots and ensure early administrative allocation of scarce medical facilities and services in these pockets, until the supply of these can be scaled up. Also, our findings underscore the importance of social protection and economic support, especially during lockdowns, to ensure that the gains made in arresting the growth of infection during lockdown remain intact even after the economy has been unlocked. Our research underlines the importance of social cohesion and community ties and builds a case against divisive political agendas that deepen the existing fissures across communities. Instead, we outline the potential gains from promoting harmony across different social groups, especially in those pockets that are fragmented.

The structure of the paper is as follows. Section 2 introduces the data used and presents summary statistics. Section 3 discusses the empirical framework by outlining the estimation and identification strategies. Section 4 discusses the results alongside some robustness checks. Section 5 concludes with a summary of findings, limitations of the study and its policy implications.

2 Data and descriptive statistics

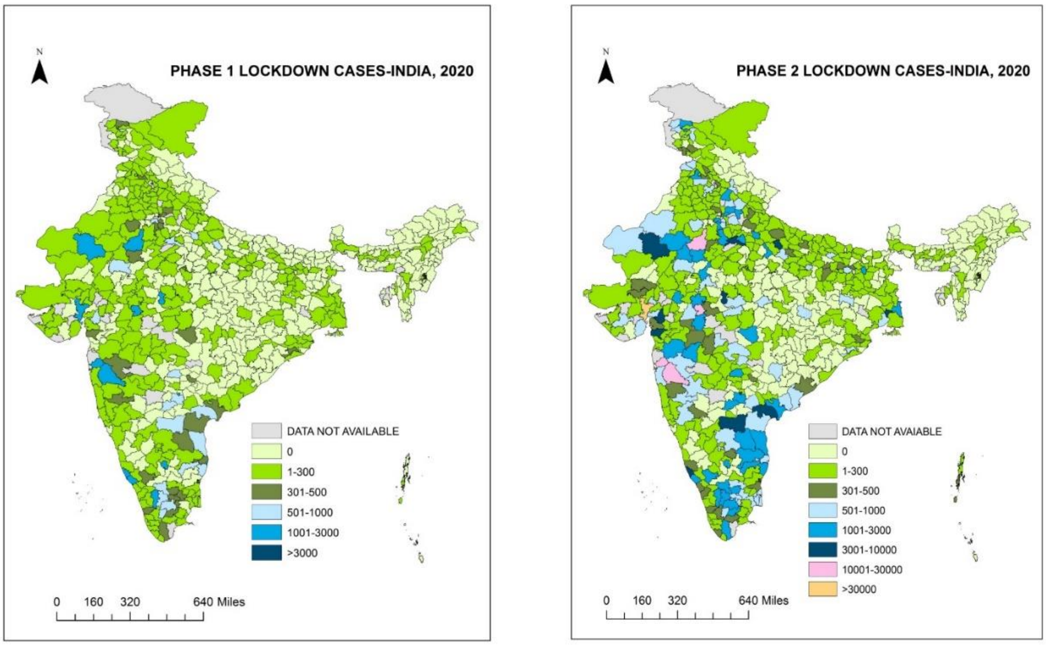
District day-level data on Covid-19 cases and deaths comes from the DDL Covid India database, which is available for 719 (98%) of the total 733 districts in India.[6] The period of analysis ranges from 25 March 2020 to 26 July 2020 and covers the period of nationwide lockdown, which ended on 31 May, as well as the subsequent two-month period of unlocking.[7] Figure 1, which provides a spatial view of the spread of Covid-19 infections in India through subsequent phases of the lockdown, indicates a shrinking of the green area, which represents the low infection band. Post-lockdown we observed a rising number of reported cases, as indicated in

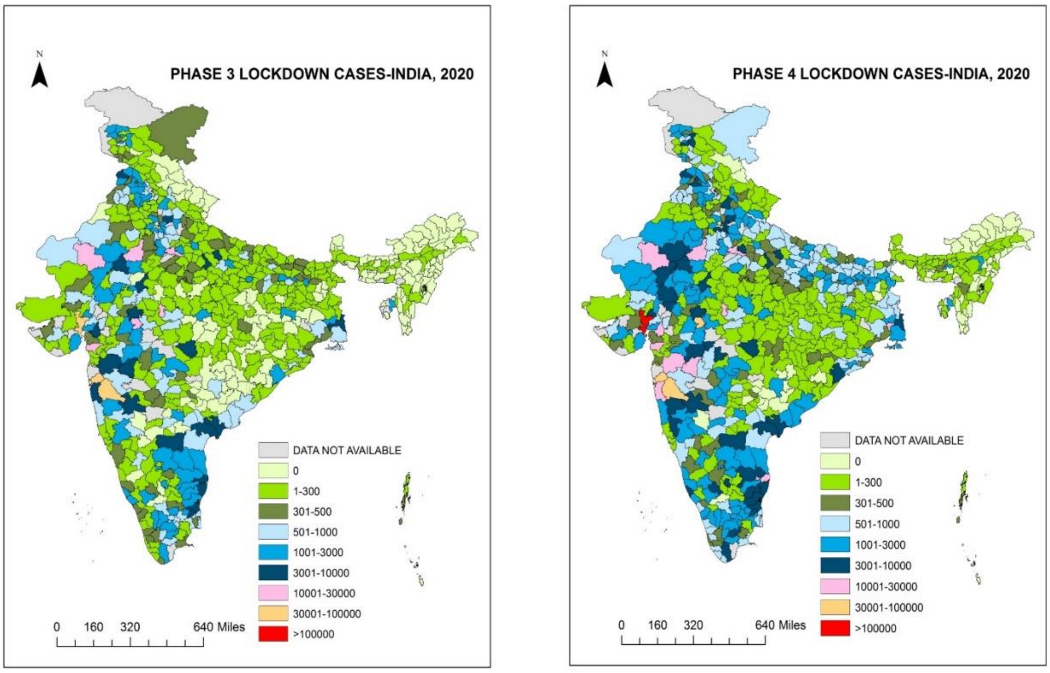
[6] The Socioeconomic High-resolution Rural–Urban Geographic Platform for India (SHRUG) is an open source platform to facilitate data sharing between researchers working on the country. It can be accessed at http://www.devdatalab.org/shrug. These numbers are taken from covidindia.com, a crowd-sourced project.

[7] The Indian government announced a nationwide lockdown in four continuous phases, with detailed guidelines issued on measures such as social distancing and hygiene practices. Thereafter, the respective state governments were given more freedom to decide on further restrictions, outside a countrywide negative list including, eg international travel and the opening of educational institutions. See https://timesofindia.indiatimes.com/india/lockdown-4-0-after-may-31-states-to-have-own-curbs-lists/articleshow/76024524.cms.

Table 1.

Figure 1: Spread of Covid-19 cases across districts of India for each phase of the lockdown (up to 31 May 2020)





Source: Authors' representation from the DDL Covid-19 India database (Asher et al, 2020).

To assess district-level social, religious and economic diversity, we used the Herfindahl Hirschman Index (HHI) (Rhoades, 1993),[8] and the Gini Index (GI), respectively. As discussed above, for caste diversity, we used the broad caste groups that include the historically disadvantaged groups of SCs, STs and OBCs in addition to the privileged castes aggregated under 'Others'.[9] For religion, representation across eight major religious groups was accounted for.[10] The HHI varies on a scale of 0 to 1, with a higher value implying increased homogeneity or concentration. The GI is based on monthly household consumption expenditure, which is also computed on a scale of 0 to 1, with a higher value implying higher inequality.[11] Figure 2 provides a spatial overview of the social group, religion and economic diversity across India's districts.

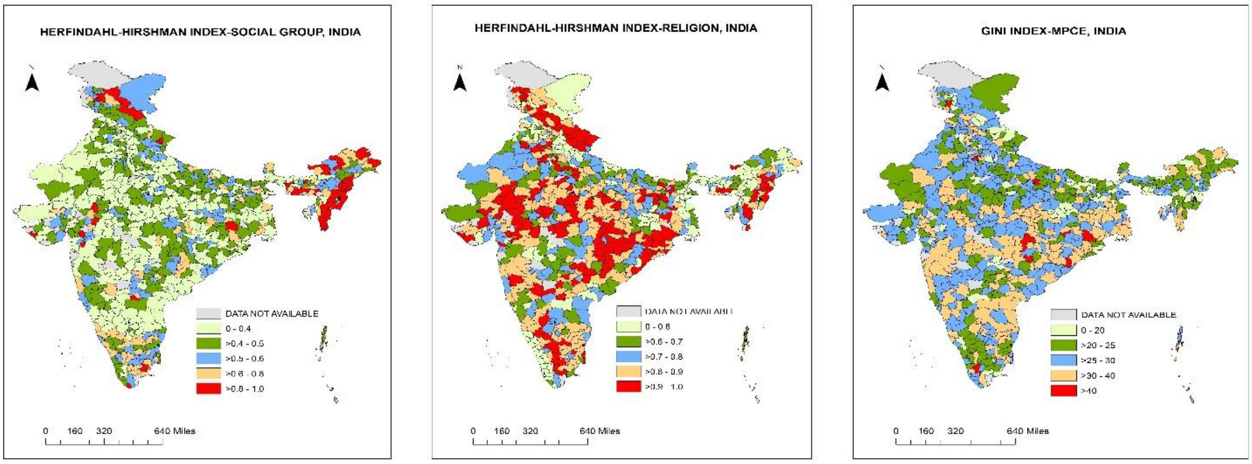
Figure 2: Social, religious and economic diversity across Indian districts (2018)

[8] The HHI index is a measure of the size of a firm with respect to the industry and is typically used as a measure of competition, with a higher value implying higher monopolisation.

[9] The social networks within the Indian caste system are splintered into various sub-castes or *jatis*. In this paper, we restrict ourselves to the broad caste groupings, as district-level representative data are only available for these groups for the period up to 2018.

[10] These include Hindu, Muslim, Christian, Sikh, Jain, Buddhist, Zoroastrian and Others.

[11] We assumed that intra-household as against inter-household coordination and cooperation is relatively much easier to achieve. Thus, in this paper we have computed the variables of interest (HHI and GI) using data at the household level, which were then aggregated at the district level using weights. We did, however, check for the robustness of our findings using individual-level variables which were qualitatively similar.



Source: Authors' calculations using the NSSO 76th round.

For analysis, the variables of interest were taken as standardised Z-scores, which are standard deviation units of difference from all-India means.[12] Data for this came from the 76th round of the NSSO survey on drinking water, sanitation, hygiene and housing conditions collected between July and December 2018.[13] These are to the best of our knowledge the latest available data to provide representative estimates of the socioeconomic profile at the district level by using appropriate sampling weights. Notably, data from NSSO are often used by the Indian government to estimate the prevalence of poverty and unemployment, alongside the generation of estimates on educational, agricultural and health indicators, among others, which gives credence to our main variables of interest as well as the controls.

We also controlled for a range of socioeconomic, demographic and mobility controls using the 76th round of the NSSO survey. This includes share of households with access to basic amenities, such as improved latrines and water sources, as per the World Health Organization (WHO) guidelines. Access to these amenities is likely to facilitate reduced movement and exposure to infections from others in the vicinity. As proximity to other households is critical in achieving compliance with social distancing measures, we controlled for density of population,[14] as well as share of households in the district living in slum or squatter settlements. Given the importance of hygiene, especially hand-washing habits, in preventing the spread of Covid-19, we accounted for the share of households with human faeces spotted around the premises during the survey, as well as the share of households whose members were washing their hands regularly with

[12] Standard deviation units of difference from the median were also considered as a robustness check.
[13] These district-wise estimates were computed from the household-level data for 106,838 households across India.

[14] The total area for each district in square kilometres is taken from DDL's Covid India database, which is discussed below.

soap after defecating.[15] The emerging literature suggests that the Covid-19 pandemic has an inherently urban character, with cities playing a critical role in disease transmission (Mishra et al, 2020). To account for this, the share of households in the district residing in urban areas was controlled for. Given the vulnerability of the elderly population (>=60 years) to infections (Daoust, 2020), the share of this population was also accounted for. For the economic profile, the share of households with at least one member with a graduate degree and regular employment were considered. We controlled for mobility in various ways, since it is a critical channel for transmission: first, we accounted for the family mobility trend by including the share of households in the district having lived in their current place of residence for at least 10 years. Second, we checked for the frequency of mobility patterns by including the share of households where someone was travelling more than 15 km for work.

Additionally, to account for the health profile of the district, information on the share of the population in each district with a chronic ailment was taken from the 75th NSSO round on Health Consumption.[16] For the availability of health infrastructure, both physical and human, the number of hospitals as well as the number of doctors per million households in a district was considered.[17] This information was taken from the DDL's Covid India database, which sources it from the Population Census and the Socio-Economic Caste Census for 2011 and 2012 (Asher et al, 2020). The above list of controls is available for 670 (91%) of the 733 (91%) districts of India and we were able to map these variables with information on Covid-19 cases and deaths for 661 (92%) of the 719 districts.

To explore the relationship between diversity and the spread of Covid-19 infection, we first present a scatter of social group homogeneity and the natural logarithm of cases across the eight periods considered in our analysis. These periods comprise the four phases of the lockdown and unlocking, respectively. We also used administrative data from the Government of India (GoI) notification dated 30 April 2020, issued by the Ministry of Health and Family Welfare (MoHFW). As per the notification, districts were classified into red, orange and green zones (dangerous, mildly dangerous, no apparent danger, respectively) for the week following the end of the second phase of lockdown (3 May 2020) using a multi-factorial approach.[18] Here, the movement of the districts coded

---

[15] In the survey, the household is considered as one where hand washing takes place if a majority of its members clean both their palms, front and back with all the fingers included, with soap or detergent, whether liquid soaps, hard soaps, hand sanitiser, or wet and dry tissue paper, etc.

[16] These data were collected between July 2017 and June 2018 and covered 113,823 households for each district; they provide a representative account for each district.

[17] A detailed description of the independent variables used in the analysis can be found in Appendix Table 1A.

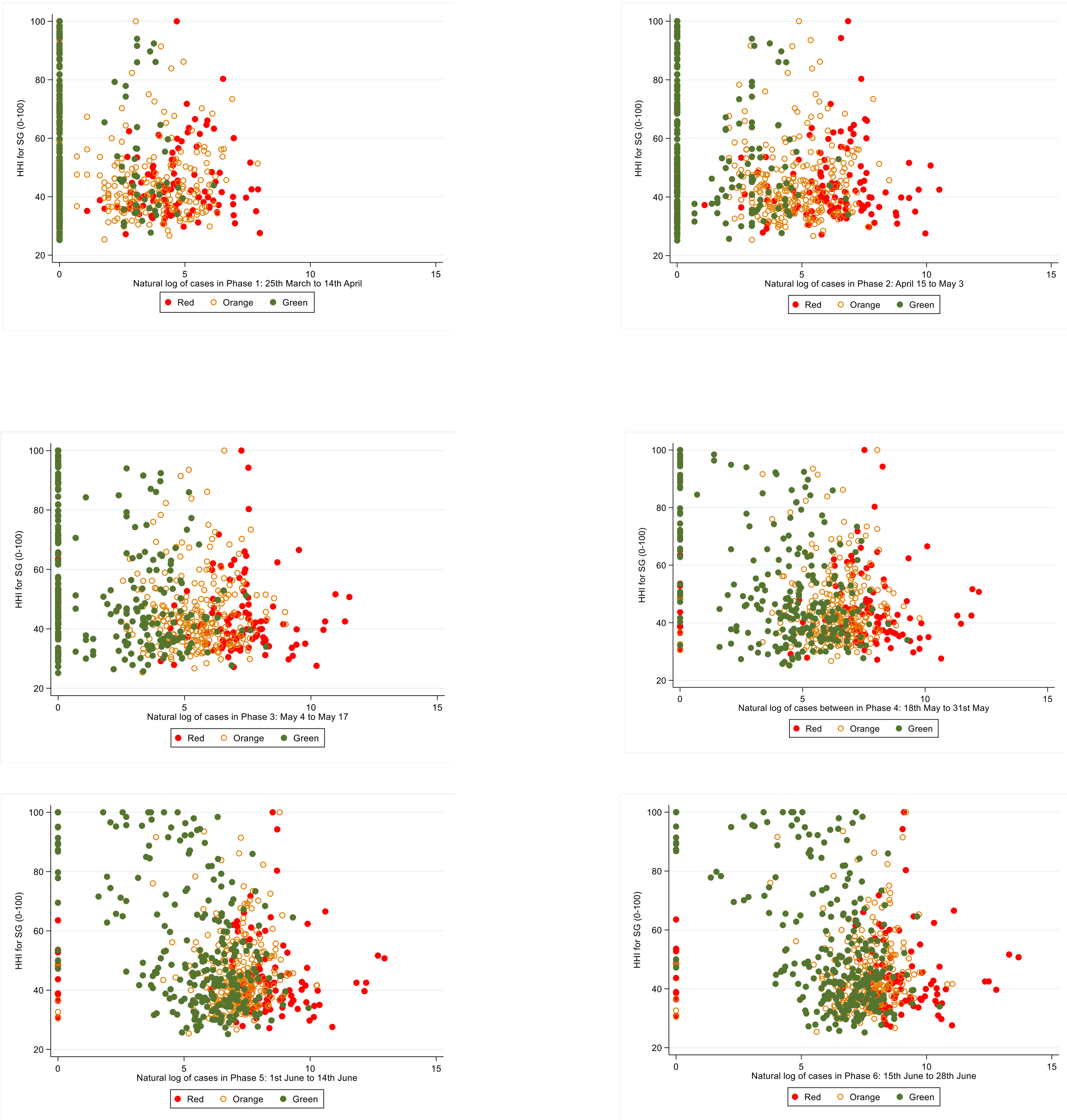
[18] This approach accounted for number of cases, extent of testing, surveillance feedback and doubling of rates in respective districts. As per these criteria, a district was classified as a green zone if there had been no confirmed cases of Covid-19 in the past 21 days. No subsequent notifications with such categorisations were disseminated by the Ministry to the best of our knowledge.

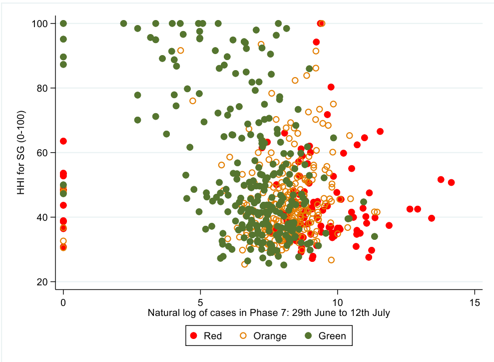
as green across the distribution of HHI for different phases suggests that the districts at the lower end of the distribution (lower social group homogeneity) moved further to the right compared with the green districts at the upper end of the distribution, which were relatively more homogeneous (

Figure 3). The trend is similar for consumption inequality, with cases appearing to rise faster in green-coded districts with higher economic inequality.[19]

[19] The scatter plots for GI from household consumption inequality can be found in Appendix Figure 3A.

Figure 3: Progression of cases and social group homogeneity across districts classified as red, orange and green by MoHFW via notification on 30 April 2020





Source: Author's calculations based on the DDL Covid-19 India database (Asher et al. 2020).

We now turn to a discussion of the descriptive statistics and provide summary measures of outcomes, variables of interest and other control variables (Table 1). It must be noted that, across the districts of India, close to 85% had no Covid-19 cases reported until the beginning of the lockdown. For each of the four lockdown phases, which were considered to be among the most stringent in the world,[20] the share of districts with no cases reduced to 47%, 31%, 26% and 15%, respectively. The average number of cases (deaths) per district in this period increased from about 4 (0) before the first phase of the lockdown to 2669 (80) for the period ending 31 May 2020. For the last period of analysis, ending 26 July 2020, the average cases and deaths were as high as 21,279 and 537, respectively.

Among the variables of interest, given that the vast majority of the country is Hindu, the degree of religious homogeneity was significantly larger than social group homogeneity on average. Religious homogeneity was found to be highest in the districts of Odisha, Uttarakhand and Gujarat with 11, six and five districts featuring in the top decile of the distribution. For social group homogeneity, the states of Nagaland, Arunachal Pradesh and Mizoram had the highest figures, with 10, eight and seven districts featuring in the top decile. The corresponding numbers for consumption inequality were highest for Chhattisgarh, Odisha and Madhya Pradesh, with eight, seven and six districts featuring in the top decile, respectively (as shown in Figure 2).

[20] https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker.

Table 1: Summary statistics

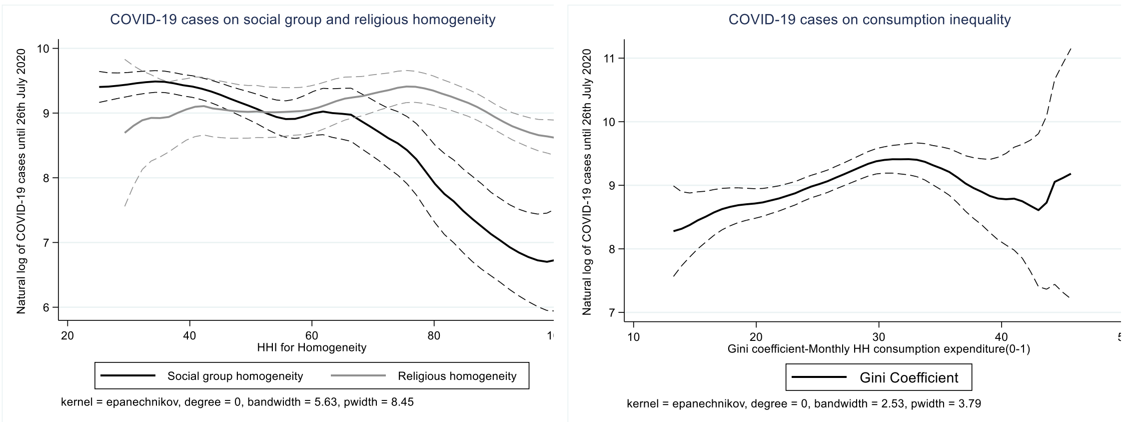
COVID-19 cases for the following periods	**Mean**	**Standard deviation**	**No of districts**
Pre-lockdown (before 25 March 2020)	4.25	20.21	719
1st phase of lockdown (25 March-–4 April)	111.91	521.86	719
2nd phase of lockdown (14 April–3 May)	597.46	3785.95	719
3rd phase of lockdown (4 May–17 May)	1321.92	9455.72	719
4th phase of lockdown (18 May–31 May)	2669.26	19949.42	719
1st phase of unlocking (1–14 June-)	4853.02	34572.44	719
2nd phase of unlocking (15–28 June)	8085.33	55819.61	719
3rd phase of unlocking (29 June–12 July)	13107.08	83636.75	719
4th phase of unlocking (13–26 July)	21279.03	109903.30	719
COVID-19 related deaths for the following periods	**Mean**	**Standard deviation**	**No of districts**
End of first lockdown (14 March 2020)	2.81	26.05	719
End of second lockdown (3 May)	18.46	161.38	719
End of third lockdown (17 May)	43.38	368.68	719
End of fourth lockdown (31 May)	79.72	689.19	719
1st phase of unlocking	141.35	1189.75	719
2nd phase of unlocking	264.03	2407.39	719
3rd phase of unlocking	381.64	3285.10	719
4th phase of unlocking	536.53	3985.15	719
Variables of interest (0–100)	**Mean**	**Standard deviation**	**No of districts**
HHI social group	48.85	17.25	670
HHI religion group	77.45	16.71	670
Gini coefficient	27.16	5.74	670
Share of households in a district (scale 0 to 100)	**Mean**	**Standard deviation**	**No of districts**
Using improved latrines in household (HH)	78.71	18.13	670
HH with improved water source	89.37	15.36	670
Living in slum/squatter settlement	1.36	4.08	670
HH washing hands with soap after defecating	72.82	22.70	670
Human faeces spotted around HH premises	8.48	11.97	670
With chronic ailments	2.88	4.16	670

HH in urban areas	25.32	21.82	670
Elderly (>=60 years) in HH	10.22	4.50	670
At least one member with regular employment in HH	8.17	6.42	670
Share of graduates in the HH	8.86	5.13	670
No of households in the district	399922	388013.10	670
Density of households per square km	1664.16	32063.58	648
Duration of stay at current place >=10 years	81.13	15.99	670
HH with any member travelling more than 15 km for work	8.06	8.03	670
Allopathic hospitals per million HH in district	56.46	76.57	670
Allopathic doctors per million HH in district	400.51	811.55	670

Note: From the third phase of lockdown, each lockdown lasted for two weeks beginning on a Monday and ending on a Sunday. The unlocking period has been broken down into four phases in a corresponding manner, as discussed in the introduction. For details, see Appendix Table 1A.

To further understand how the outcome of interest varied with measures of homogeneity, we used the kernel weighted local polynomial smoothened plots with a 95% confidence interval, which makes no restrictive distributional assumptions on the error term (Fan & Gijbels, 1996). Figure 4 shows that, for the duration of the analysis, the rate of growth of Covid-19 cases fell as social group homogeneity rose along the x-axis. The infection growth rate also appeared to be rising with higher income inequality, as measured by the higher values of the Gini coefficient along the x-axis. As this analysis is devoid of potential confounders, we now turn to a regression-based approach to validate the relationships of interest. For this, we first present the empirical framework and then present the corresponding findings over the next two sections.

Figure 4: Local polynomial plots for log of Covid cases and caste, religious and economic homogeneity across districts



Note: Results generated on Stata 16 using lpoly variant with 95% CI and default bandwidth using Rule to Thumb (RoT) estimator with degree 0.
Source: Authors' calculations.

3 Empirical framework

3. 1 Estimation strategy

First, we estimated the relationship between the indices associating homogeneity with the number of Covid-19 infections across districts at the end of the eighth period (on 26 July 2020) considered. To test our hypothesis, we regressed the logarithm of the number of Covid-19 cases reported in each district (d) of administrative region (r) on the variables of interest that capture homogeneity along caste $(HISG_{rd})$, religious $(HIRel_{rd})$ and economic lines $(GINI_{rd})$. To ensure comparability, we standardised the three indices and used them directly in the regression. To estimate the association, we first used the following model:

$$ln\bar{Y}_{rd(t=8)} = \beta_0 + \beta_1 lnY_{rd(t=0)} + \beta_2 ln\bar{D}_{rd(t=7)} + \beta_3 HISG_{rd} + \beta_4 HIRel_{rd} + \beta_5 GINI_{rd} + \beta_6 C_{rd} + AR_r + \epsilon_{rdt} \dots\dots\dots (1)$$

Here $\bar{Y}_{rd(t=8)}$ denotes 1 added to the number of infections reported in a district, *d* located in the administrative region, *r* at the end of the eighth period. Given that the natural logarithm of zero is undefined, we added 1 to each of the variables before logarithmic transformation to ensure that there was no sample selection bias (MaCurdy & Pencavel 1986). $Y_{rd(t=0)}$ and $\bar{D}_{rd(t=7)}$ indicate the number of Covid-19 cases reported before the announcement of the lockdown $(t = 0)$ and the number of reported fatalities caused by the virus at the district level in the previous period $(t = 7)$. For both these controls, 1 was added for the reasons stated previously. Importantly, these controls would account for the potential supply-side short-term response for the authorities to arrest the growth in infection, including timely medical interventions, strategic planning to contain the growth in infection and awareness campaigns, among many other things. $HISG_{rd}$ and $HIRel_{rd}$ represent the district-level standardised values of the HHI index for caste and religion, respectively and $GINI_{rd}$ is the standardised Gini coefficient for consumption expenditure. The vector of time-invariant control variables at the district level is given by C_{rd}. The administrative region fixed effects are denoted by AR_r, which account for inter-regional heterogeneities that are time invariant and may confound with our relationship of interest, while ϵ_{rdt} is the error term. The regression estimation was done through the normal Ordinary Least Squares (OLS), and robust standard errors that control for potential heteroscedasticity were used.[21] Here, β_3 to β_5 are the coefficients of interest.

Next, as discussed, we divided the time since the commencement of the first lockdown into eight periods, each consisting of about 15 days.[22] The first four periods correspond to the four lockdown phases, the next four represent the unlocking phases. To examine

[21] The National Sample Survey Office divides all the districts into a number of mutually exclusive and exhaustive administrative regions, which were used as fixed effects in our regressions.
[22] The initial phases (1 and 2) were longer as they lasted more than two weeks but later the lockdown was extended by intervals of two weeks.

the implications of homogeneity across these time periods, we estimated the following pooled OLS regression equation:

$$ln\bar{Y}_{rdt} = \beta_0 + \beta_1 lnY_{rd(t=0)} + \beta_2 ln\bar{D}_{rdt-1} + \beta_3 HISG_{rd} + \beta_4 HIRel_{rd} + \beta_5 GINI_{rd} + \beta_6 (HISG_{rd} * t) + \beta_7 (HIRel_{rd} * t) + \beta_8 (GINI_{rd} * t) + \beta_9 C_{rd} + AR_r + t + \epsilon_{rdt} \dots\dots\dots(2)$$

Here t denotes the period dummies. $\bar{Y}_{rdt}$ represent 1 added to the number of reported cases in period, t and $\bar{D}_{rdt-1}$ is the reported number of deaths up to the lagged period, $t-1$ in district, *d.* To assess the relationship between homogeneity and the changes in infections over time, we examined the coefficients, β_6 to β_8. The standard errors here are clustered at the district level.

Our central hypothesis links caste and religious homogeneity, alongside lower economic inequality, with a potentially slower growth of infection in a spatial unit. The underlying pathway for this, as argued, is that of within-group cooperation and trust, which facilitates coordinated community action to tackle the spread of a pandemic. To establish the causality of homogeneity or inequality in the growth in reported infection, we need to account for potential endogeneity, the two main concerns of which are reverse causality and the Omitted Variable Bias (OVB).

In the context of our study, reverse causality is of negligible concern. Given that the Covid-19-induced lockdown in India was an unanticipated shock and was introduced with just four hours' notice, the chances of residential clustering on social, religious and economic lines across districts as a response to the outbreak would be remote. This seems especially true since studies have documented the low levels of inter-district or inter-state migration in India even during non-Covid times (Munshi & Rosenzweig, 2009; Munshi, 2019; Roychowdhury, 2019). Post-lockdown, although mobility increased, it is still highly unlikely that households changed their residence amid the pandemic within or across districts. Hence, the chances of reverse causality restricting us to viewing the estimates through a causal lens are almost negligible.

To control for the OVB, which can be a major concern, we introduced a host of control variables, ranging from district-level economic conditions and urbanisation to sanitation and hygiene practice to health indicators and the availability of physical infrastructure and health workers, among many others. Note that the initial hotspots where the chances of the spread of infection remained high have been controlled for through the number of cases reported across districts before the start of the lockdown. This in our view is critical to address the issue of endogeneity, as it controls for various characteristics of the districts that became epicentres of the epidemic in India. For example, the major metropolitan cities of Delhi, Mumbai, Kolkata and Chennai accounted for more than half the total tally of cases;[23] they are also major centres for urban jobs. The time variant potential health measures in response to Covid-19 deaths were controlled through the district-level number of deaths in the lagged period. Given India's administrative structure, where each district is under the stewardship of a bureaucrat (district collector) reporting directly to the Chief Minister of a state, allocation of scarce administrative resources in a pandemic is likely to be most sensitive to Covid-19-related casualties. In addition, the region fixed effects were introduced to control for all the time-invariant characteristics that might be common to all the districts lying within the respective regions.

Despite the elaborate set of control variables that we incorporated, one could still argue that some potential confounders were omitted from the regressions, which could yield biased estimates. To ensure that this was not the case, we examined the potential variation in the estimates after accounting for the OVB through a strategy developed by Oster (2019) and based on Altonji et al (2005). This rests on the assumption that the selection on unobservable variables can be measured through the extent of selection on observables. For further elaboration, consider the following regression equation:

$$Y = \beta Z + \gamma X + u$$

Here Y is the dependent variable, Z denotes the primary variable of interest, X is the vector of control variables and u is the vector of all unobserved components omitted from the regression. Our aim was to examine the corresponding changes in β or whether it changed its sign when u is high. Here the primary assumption is:

$$\frac{cov(Z,u)}{Var(u)} = \delta \frac{cov(Z,\gamma X)}{Var(\gamma X)}$$

The above equation indicates that the relationship between Z and the unobserved component, is proportional to the correlation between Z and the observed component (γX) and the degree of proportionality is given by δ. With this, Oster (2019) showed that it is possible to derive a possible range of β through two parameters: δ and R_{max}. R_{max} is the *R*-squared value of a hypothetical regression that incorporates all the observables along with the unobserved components and hence is the maximum value of *R*-square it can achieve. While there can be various options to compute the bound on R_{max}, using data from a wide sample of randomised experiments, Oster (2019) proposed $R_{max} = \pi * R_0$, where R_0 is the *R*-squared value of the full model with observed control variables and $\pi = 1.3$.[24] For parameter δ, assuming that the extent of association of the observables is as much as the effect of omitted variables, then the δ lies between [-1,1] (Altonji et al, 2005; Mukhopadhyay & Sahoo, 2016). Oster (2019) proposed two equivalent approaches to test if the effect becomes statistically insignificant after

[23] https://www.financialexpress.com/lifestyle/health/covid-19-four-metros-account-for-half-of-all-cases-nationwide-tally-nears-2-4-lakh/1983731/.
[24] According to Oster (2019), the bounding value of the cut-off at 1.3 allows effects from at least 90% of the random experiments to survive.

accounting for the OVB. First, we check if δ, for which the coefficient of interest turns 0 with $R_{max} = 1.3 * R_0$, exceeds a threshold of 1. The argument is that, if δ goes beyond 1 or falls below -1, the extent of the association with unobservables has to be massively high in either direction, which may not be valid, especially after controlling for the robust set of observed covariates. Second, we examined whether β changes its sign if in the interval $\delta \epsilon$ [-1, 1] and $R_{max} = 1.3 * R_0$. If the sign does not change then, even with the very stringent condition of the extent of association with the unobservables being as high as that with the observables and in both the directions as well, the null hypothesis ($\beta = 0$) can be rejected. Our estimation exercise (described below) satisfies these conditions allowing us to interpret our findings in a causal framework.

4 Results

4.1 Regressions

We first present the estimates of the regression of the natural logarithmic value of the reported cases on the three different indices of homogeneity as outlined in equation (1) in Table 2. To check the level of stability of the coefficient of interest, we used four different specifications. In the first, apart from the variables of interest, we incorporated the natural log of cases reported before the start of lockdown in the district and the number of deaths in the previous period, alongside regional specific fixed effects. To this, in specification 2, we added indicators on access to amenities, such as an improved toilet and primary source of water. Further, we also added variables on the share of households (HH) in the district residing in slums or squatter settlements and general hygiene and cleanliness practices around their premises. In specification 3, we subsequently controlled for economic and demographic variables. In this specification, in addition, we accounted for the share of individuals in the district reporting chronic ailments. In the final specification, we controlled for mobility indicators, as specified previously, alongside average density of population in the district, as well as the existing physical and health infrastructure of the district. Because it controls for a host of observable covariates – spanning economic to demographic characteristics along with sanitary and health indicators and also the supply-side response to the outbreak – this is our preferred specification, and is used in the subsequent regressions as well.

The results appear to be fairly stable across specifications, with findings from the final specification (column 4) suggesting that a one standard deviation increase in homogeneity across caste groups, *ceteris paribus*, is associated with a decrease in the logarithmic value of the reported cases by about 0.18 on average. The results for the income GINI are qualitatively similar in the sense that an increase in inequality is related to a higher number of cases, albeit with a lower magnitude with an average effect size of about 0.13 for every standard deviation increase, thereby underscoring the importance of equity. Notably, the effect of religious homogeneity remains statistically indistinguishable from zero. The findings from other sets of covariates indicated the expected results. As an example, we found that the number of reported cases and deaths in the pre-lockdown phase was significantly associated with the current number of reported cases.

Table 2: OLS regression for the growth (natural logarithm) of Covid-19 cases from the beginning of the lockdown to the end of July 2020 on caste, religious and economic diversity across districts of India

	(1) model1 b/se	(2) model2 b/se	(3) model3 b/se	(4) model4 b/se
Logarithm of cases before the lockdown	0.302***	0.285***	0.205***	0.205***
	(0.056)	(0.053)	(0.073)	(0.075)
Logarithm of cumulative deaths in the previous period	0.315***	0.316***	0.265***	0.265***
	(0.042)	(0.043)	(0.063)	(0.059)
Z-score HHI SG	-0.241***	-0.249***	-0.176**	-0.175**
	(0.077)	(0.078)	(0.076)	(0.075)
Z-score HHI religion	-0.051	-0.061	-0.024	-0.023
	(0.070)	(0.063)	(0.054)	(0.053)
Z-score Gini	0.134**	0.162***	0.106	0.129**
	(0.065)	(0.060)	(0.067)	(0.065)
Additional district level controls				
Amenities, sanitation and hygiene	×	✓	✓	✓
Demographic, economic, long-term health outcomes	×	×	✓	✓
Mobility, density and health infrastructure	×	×	×	✓
Region fixed effects	✓	✓	✓	✓
Constant	8.258***	8.091***	3.309	3.382*
	(0.450)	(0.578)	(2.137)	(1.913)
Observations	661	661	661	661
R^2	0.785	0.793	0.805	0.807

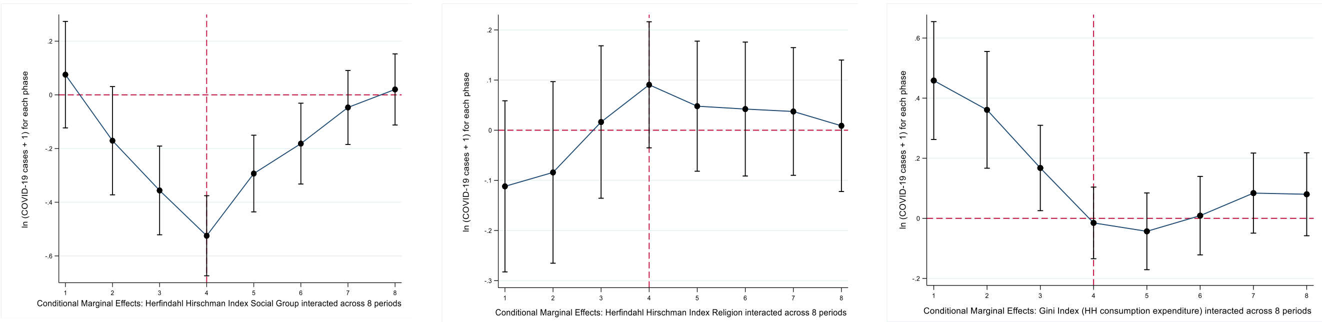
*Note: *** p-value < 0.01, ** p-value < 0.05, * p-value <0.10, standard errors are robust to heteroscedasticity. To ensure that districts with no cases in any period did not drop out of the analysis, unity was added to cases before taking the natural log. Results are robust to adding a smaller positive number. Results are robust with three- and seven-day rolling averages replacing daily reported cases for each period. Detailed tables are available on request.*

Source: Authors' calculations.

Next, we ran a pooled OLS regression using the number of cases reported at the district level at the end of each of the eight periods we considered, as outlined in equation (2). Through the interaction effects of the homogeneity indices and the period dummies, this allows us to examine the heterogeneous implications of our primary variables of interest on the changes in the reported number of infectious cases over time. Figure 5 presents the marginal effects of these interaction terms, controlling for the potential confounders discussed earlier. Note that at the end of period 4, the government-mandated lockdown ended, and the unlocking procedure commenced. The findings from the regressions indicate an interesting pattern in which we observe incremental marginal gains of higher homogeneity on the caste lines over each of the four lockdown phases. By the end of lockdown Phase 4, we find that one standard deviation increase in homogeneity in terms of caste groups through the HHI results in an average drop of 0.53 in the logarithmic value of the reported cases. However, post-lockdown, the relationship starts to weaken and by the end of the seventh period, we found no additional gains from higher caste group homogeneity (see panel (a), Figure 5).

As observed through panel (b) of Figure 5, no significant relationship between religious homogeneity and number of infections across time was found. For economic inequality, more equal districts were found to be associated with a lower number of reported cases; this relationship remained strong until the end of the third period of the lockdown but had eroded completely by end of the fourth phase. Over the unlocking period, the marginal gains from higher economic equality remained statistically indistinguishable from zero. Additionally, to account for any reporting errors, we also considered daily cases as rolling averages for three and seven days. Our findings remain statistically robust to this specification. For detailed results, see Appendix Table 2A.

Figure 5: Coefficients of variables of interest from OLS regression for the growth (natural logarithm) of Covid-19 cases across the four periods of lockdown and four of unlocking, up to 26 July 2020



(a) Period-wise coefficients for HHI caste group

(b) Period-wise coefficients for HHI religion

(c) Period-wise coefficients for Gini Index (HH consumption)

Note: The coefficients come from the full control scheme, as described in model2 above, with confidence bands of 95%. The vertical line at x=4 signifies the end of the nationwide lockdown (31 May 2020). To ensure that districts with no cases in any period did not drop out of the analysis, unity was added to the cases before taking the natural log. The results are robust to adding a smaller positive number. The results are robust with three- and seven-day rolling averages replacing daily reported cases for each period. Detailed tables available on request. Corresponding coefficients can be found in Appendix Table 2A.

4.2 Accounting for OVB

How effective are our variables of interest in arresting the growth in the reported cases of Covid-19? Does diversity in terms of caste groups or along economic lines lead to a higher number of cases? As argued earlier, we included a host of control covariates in our regression, which account for most of the variations in the number of cases across districts. Information about the number of tests carried out at the district level is unavailable but the region fixed effects and the period dummies have been incorporated into the model. To the extent that there are no systematic changes in the testing across time for districts within a region, the fixed-effect dummies are expected to account for a large part of the effect through differential testing numbers. Note that the associated R-square value from regressions corresponding with Figure 5 is observed to be 0.75, which indicates that 75% of the variation in the dependent variable has already been accounted for through the set of control variables used in the regression. To that extent, we argue that we can view the inferences drawn through a causal lens.

Nevertheless, because it could still be argued that some omitted variables have biased our estimations, we used the strategy developed by Oster (2019), based on Altonji et al (2005), which estimates the bounds of the main coefficient of interest, assuming that the extent of selection on unobservables can be measured through the selection on observables as explained earlier. The results from this exercise are presented in Table 3. For interaction coefficients which are found to be statistically significant at the 10% level or less, the δ needs to be well over 1 to bring them to zero (Oster, 2019). This, as argued earlier, is unlikely, especially since we are already controlling for a host of covariates, which potentially account for most of the unobservable factors as well. However, even if we assume that the extent of the remaining unobservable components confounding the dependent variables is proportional to that which is explained through the observed covariates, the significant effects of homogeneity with regard to caste and economic inequality through Gini remain preserved.

Table 3: Accounting for potential omitted variable bias

	HHI caste group			Gini (HH consumption)		
	Coefficient (controlled)	Coefficient (uncontrolled)	δ for β=0 with $R^2_{MAX} = 1.3 * R^2_o$	Coefficient (controlled)	Coefficient (uncontrolled)	δ for β=0 with $R^2_{MAX} = 1.3 * R^2_o$
25 March–14 April	0.075	-0.379	-0.453	0.459***	0.478	16.376
15 April–3 May	-0.171*	-0.644	0.885	0.361***	0.447	7.769
4–17 May	-0.356***	-0.838	1.691	0.168**	0.324	2.707
18–31 May	-0.525***	-1.008	2.325	-0.015	0.187	-0.220
1–14 June	-0.293***	-0.805	1.356	-0.043	0.124	-0.774
15–28 June	-0.182**	-0.734	0.816	0.009	0.146	0.188

29 June–12 July	-0.047	-0.640	0.206	0.084	0.193	2.021
13–26 July	0.020	-0.608	-0.085	0.080	0.174	2.218

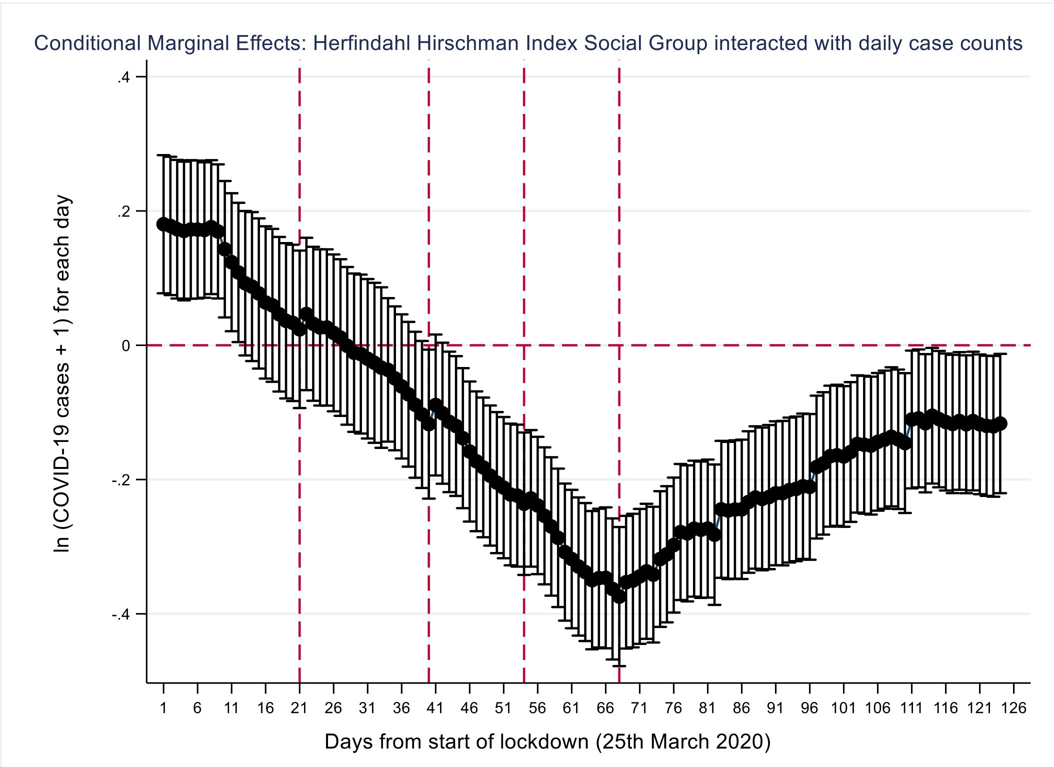
*Notes: *** p-value < 0.01, ** p-value < 0.05, * p-value <0.10. The uncontrolled coefficient is from regression of outcome exclusively on the variable of interest. Controlled coefficients are from the hypothetical regression with $\delta = 1$ and $R^2_{MAX} = 1.3 * R^2_o$, where R^2_o is from the final regression with all observed variables. Results obtained via user-written Stata command psacalc (Oster, 2016).*

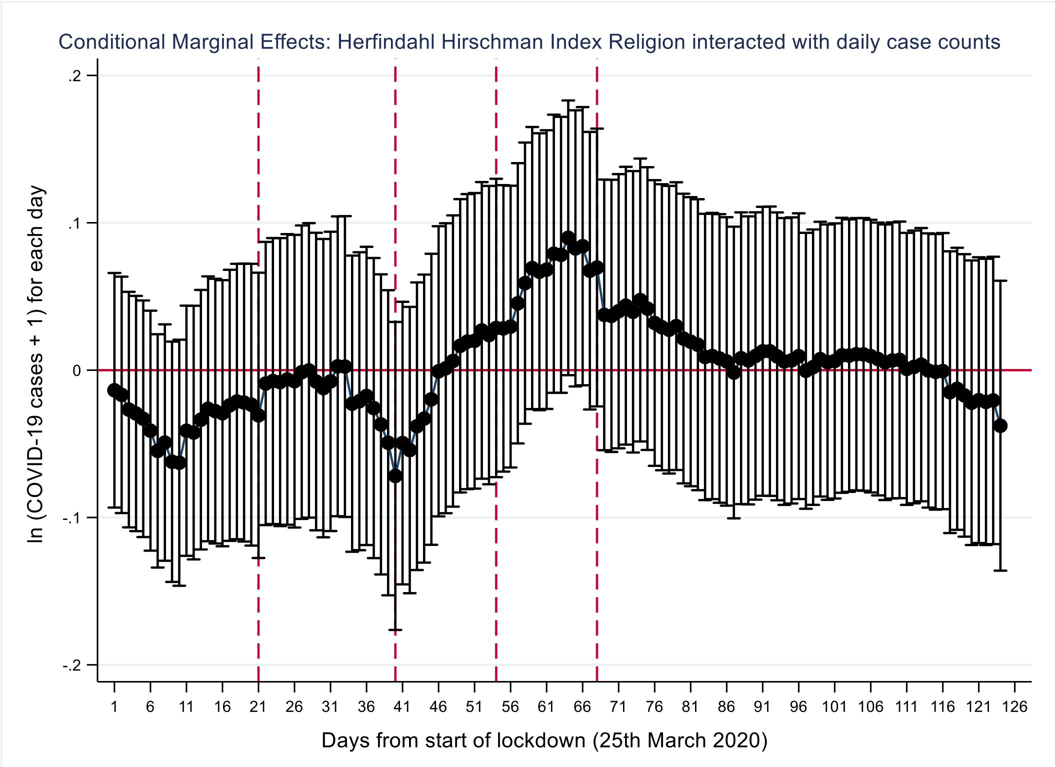
Source: Authors' calculations.

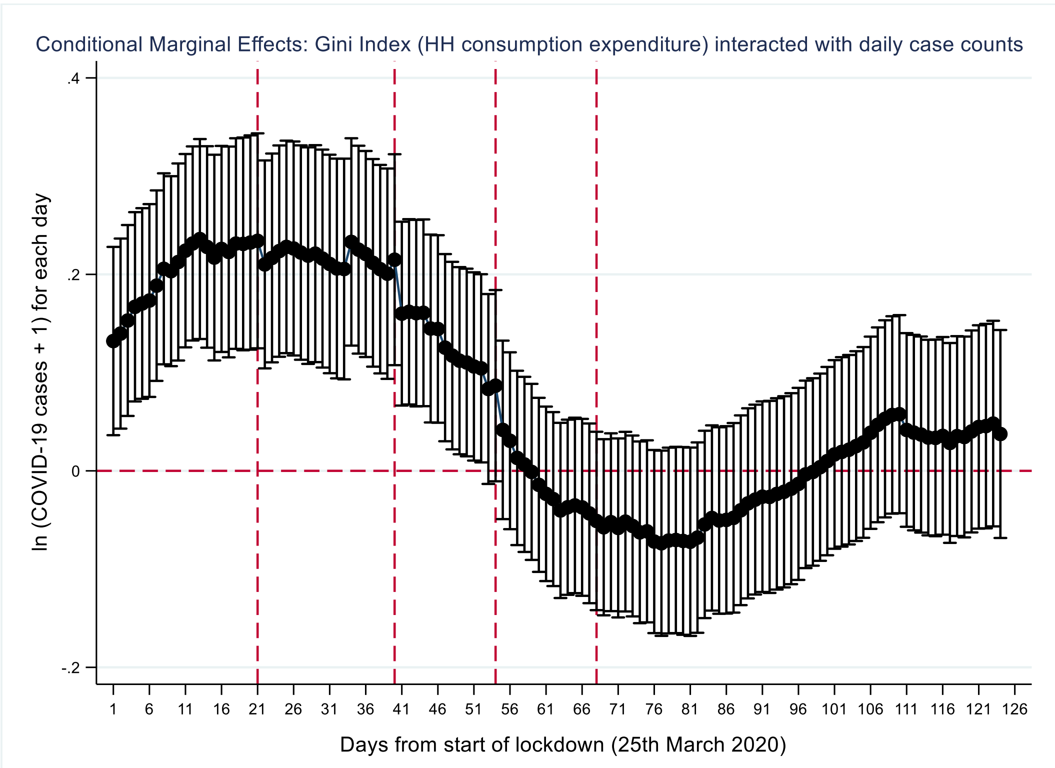
The findings, we argue, have considerable implications in terms of identifying the target group for scaling up compliance-inducing actions. Among all measures of social cohesion, our findings show, caste heterogeneity remained a persistently negative predictor of infection spread during lockdowns relative to religion and income. Further, the impact of higher caste homogeneity seemed to be multiplied with each subsequent phase of the lockdown.[25] It was only after the lockdown was lifted that the gains from caste group cohesion started to decline. Here one may argue that, because of the continuous stretch of nationwide lockdowns for more than two months, and despite the benefits that accrued to collective action, the associated private costs in terms of disruptions to livelihoods and declining economic well-being led to a weakening of the social ties among people of similar caste groups (Ahmed et al, 2020; *Lancet*, 2020). Incidentally, studies have indicated a decline over time in private compliance with non-pharmaceutical Covid-19 protocols like maintaining social distancing and wearing facemasks, possibly because of fatigue and economic costs (Bosancianu et al, 2020; Das et al, 2020). To complement our findings using period-related data, we also ran pooled OLS regressions taking daily data on the number of infection cases. Here, instead of period dummies, we incorporated daily dummies and examined the daily interaction coefficients on our main variables of interest. Figure 6 presents the estimations from the regression. The findings indicate similar inferences to those we obtained earlier. The gains from caste homogeneity seemed to increase over each day of the lockdown on average, but post-lockdown the size of the effect was reduced over time. Economic equality was also found to produce some discernible gains in the initial phases of the lockdown but the effect seems to have withered with time.

[25] According to the Blavatnik School of Governance study, India was found to have the most stringent lockdown among all the major countries in the world. See https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker. Accessed: 16 November 2020.

Figure 6: Coefficients of variables of interest from OLS regression for growth (natural logarithm) of Covid-19 cases using daily data from 25 March to 26 July 2020







Notes: The specification is same as in model2, the difference being that instead of the interaction of period dummies with variables of interest, daily dummies have been used. Results are robust to using three and seven day rolling averages instead of daily case counts.

Source Based on authors' calculations.

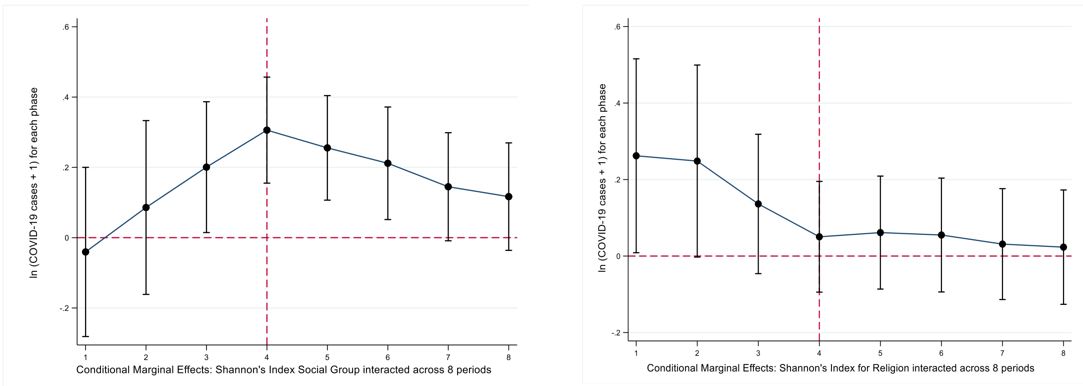
4.3 Robustness checks

We ran a battery of robustness checks to ensure our inferences were robust. We discuss these in turn below.

Shannon's entropy measure

First, we considered Shannon's entropy measure (SEM) of diversity instead of the HHI indices for caste group and religious homogeneity. The value of this measure ranges from zero to infinity with a higher value implying greater diversity (Frenken, 2004). Mathematically, this can be presented as $SEM = \sum_{i=1}^{k} P_i \log_2(1/P_i)$, where P_i is the share of each group in the overall population, which can be divided into k mutually exclusive groups. The results for this are presented in Figure 7. Unlike the HHI, which ranges between 0 and 1 with a higher value implying higher homogeneity, SEM ranges from 0 to infinity, with higher values representing greater diversity. Thus, although the HHI and SEM are inversely related, the coefficients are not directly comparable. However, the results suggest the relationships of interest are robust to the measure of diversity. Note that in some districts many of the potential groups were absent, which would make their share in the population zero. As a result, these districts dropped out of the analysis. To validate that this was not driving our findings, we re-estimated the model using HHI of heterogeneity for the sub-set of districts for which SEM could be computed. We were able to ascertain that the relationship of interest was robust; the findings are available in Appendix Table 3A.

Figure 7: Coefficients of variables of interest from OLS regression for the growth (natural logarithm) of Covid-19 cases across the four periods of lockdown and the four of unlocking, up to26 July 2020 (using Shannon's Index of Diversity)

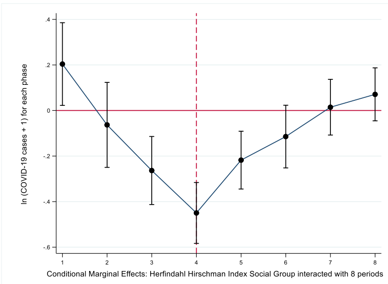


Notes: The coefficients come from the full control scheme, as described in model2, with confidence bands of 95%. The vertical line at x=4 signifies the end of nationwide lockdown (31 May 2020). To ensure that districts with no cases in any period did not drop out of the analysis, unity was added to cases before taking the natural log. The results are robust to adding a smaller positive number. The results are also robust for three- and seven-day rolling averages for cases instead of daily reported cases for each period. For detailed tables, see Appendix Table 3A.

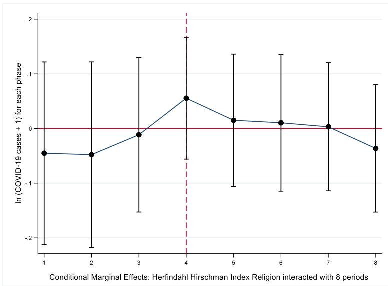
Regression with non-capital districts

Data indicate that the bigger cities were clusters with higher economic activity and a larger in-flow of migrants, making them likely to be the major hotspots within a state. Thus, they could potentially have driven our results. Although we adequately controlled for these characteristics in our specifications, as a robustness check we removed the districts that contained the capital of a state and re-estimated the model with non-capital districts across all the Indian states.

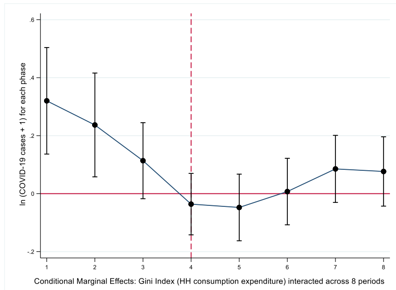
Figure 8, which presents the findings, indicates that we were able to replicate our results with this sub-sample of districts as well.



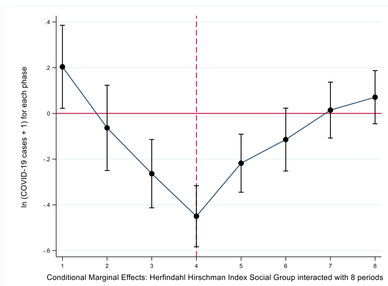
(a) Period-wise coefficients for HHI caste group

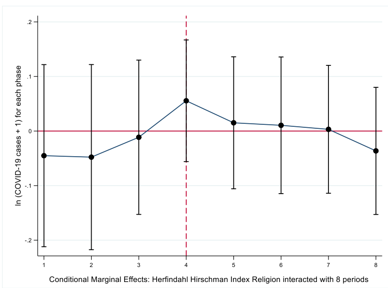


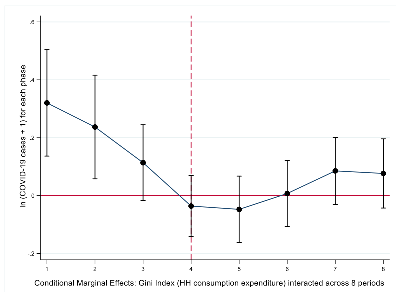
(b) Period-wise coefficients for HHI religion



(c) Period-wise coefficients for Gini Index (HH consumption)







(a) Period-wise coefficients for HHI caste group

(b) Period-wise coefficients for HHI religion

(c) Period-wise coefficients for Gini Index (HH consumption)

Figure 8: Coefficients of variables of interest from OLS regression for the growth (natural logarithm) of Covid-19 cases across the four periods of lockdown and four of unlocking, up to 26 July 2020 for non-capital districts across all Indian states

Notes: The coefficients come from the full control scheme, as described in model2, with confidence bands of 95%. The vertical line at x=4 signifies the end of the nationwide lockdown (31 May 2020). To ensure that districts with no cases in any period did not drop out of the analysis, unity was added to cases before taking the natural log. The results are robust to adding a smaller positive number before taking the logarithm. Also, the results are robust to using three- and seven-day rolling averages for cases instead of daily reported cases in each period. Detailed results are shown in Appendix Table 3A.

4.4 Further analysis

Effects of unlocking

We have argued that social cohesion along caste lines and, to a lesser extent economic equality, among other factors, can be the key to arresting the growth in Covid-19 cases over time. This was evident especially during the lockdown phases, where we found increasing gains of homogeneity across caste over time and higher returns of economic equality, at least during the early phases of the nationwide lockdown. Once the unlocking started, as indicated in Figure 5, there was no discernible gain from economic equality or, where there was a gain, which was statistically distinguishable from zero, the effect size was lower. So, is it the case that the government's unlocking of the economy after the mandated lockdown played a role in altering the influence of homogeneity on caste lines or that of economic equality? Did the returns from lower diversity erode with the unlocking of the economy?

We examined this using the following regression model:

$$\begin{aligned} ln\bar{Y}_{rd(unlock)} = \beta_0 + \ \beta_1 lnY_{rd(pre-lockdown)} + \beta_2 ln\bar{D}_{rd(lagged)} + \ \beta_3 HISG_{rd} \\ + \beta_4 HIRel_{rd} + \beta_5 GINI_{rd} + \beta_6(HISG_{rd} * unlock) + \beta_7(HIRel_{rd} * unlock) \\ + \beta_8(GINI_{rd} * unlock) + \beta_9 C_{rd} + AR_r + unlock + \epsilon_{rdt} \ldots\ldots\ldots (3) \end{aligned}$$

This looks the same as equation (2), the only difference being the introduction of an unlock dummy instead of period dummies. The 'unlock' dummy took the value of 1 for periods 5 to 8 and 0 for the periods 1 to 4. Note that, as in the earlier case, this regression controls for the number of fatalities in the lagged period along with the other confounders, as included earlier.

The findings from the regression, presented in

Table ***4***, suggest a positive and significant rise in the cases post-unlocking in areas with higher caste homogeneity. In other words, the marginal gains of fewer cases of Covid as a result of higher caste homogeneity seemed to become diluted once the unlocking of the economy started, leading to a roughly 0.61 increase in the logarithmic value of the number of reported cases with each additional standard deviation increase in caste homogeneity. Note that the relationship between this homogeneity and the reported cases remained negative and statistically significant even during the initial phases of the unlocking period. The positive sign on the interaction term merely indicates that the influence of social cohesion working through caste was significantly reduced after the government started the unlocking process.

Interestingly, this relationship does not seem to hold for religious homogeneity or economic inequality, as we observed their corresponding interaction coefficients to be statistically insignificant. The larger implication of the findings, as indicated earlier, is that the caste-related social cohesion seemed to have worked when there was a mandatory deterrence measure in the form of a nationwide stringent lockdown. Nevertheless, with the unlocking process, this cohesion weakened, perhaps in the wake of the massive economic disruptions leading to a rise in cases even in areas with higher homogeneity. Further research is needed to pinpoint the exact reasons, however.

Table 4: OLS regression for logarithm of Covid-19 cases across two periods (lockdown and unlocking) up to the end of July 2020 on measures of heterogeneity and consumption inequality

	(1) model1 b/se	(2) model2 b/se	(3) model3 b/se	(4) model4 b/se
Log of cases before lockdown	0.231***	0.237***	0.237***	0.231***

	(0.065)	(0.066)	(0.066)	(0.065)
Past period cumulative deaths	0.331***	0.307***	0.307***	0.332***
	(0.054)	(0.055)	(0.055)	(0.054)
Z-score HHI SG	-0.477***	-0.173**	-0.173**	-0.480***
	(0.095)	(0.069)	(0.069)	(0.095)
Z-score HHI religion	0.010	0.009	0.012	0.039
	(0.056)	(0.076)	(0.056)	(0.073)
Z-score Gini	0.115**	0.116**	0.164**	0.110
	(0.056)	(0.056)	(0.077)	(0.076)
Unlocking (base: lockdown)	2.007***	2.053***	2.052***	2.005***
	(0.168)	(0.174)	(0.174)	(0.169)
Unlocking # Z-score HHI SG	0.606***	NA	NA	0.614***
	(0.105)	NA	NA	(0.105)
Unlocking # Z-score HHI religion	NA	0.004	NA	-0.058
	NA	(0.095)	NA	(0.089)
Unlocking # Z-score Gini	NA	NA	-0.094	0.008
	NA	NA	(0.108)	(0.105)
Additional district level controls				
Amenities, sanitation and hygiene	✓	✓	✓	✓
Demographic, economic, long-term health outcomes	✓	✓	✓	✓
Mobility, density and health infrastructure	✓	✓	✓	✓
Region fixed effects	✓	✓	✓	✓
Constant	0.041	-0.201	-0.200	0.047
	(1.756)	(1.778)	(1.778)	(1.757)
Observations	1322	1322	1322	1322
R^2	0.760	0.750	0.750	0.760

*Notes: *** p-value < 0.01, ** p-value < 0.05, * p-value <0.10, standard errors are clustered at the district level.*
Source: Authors calculations

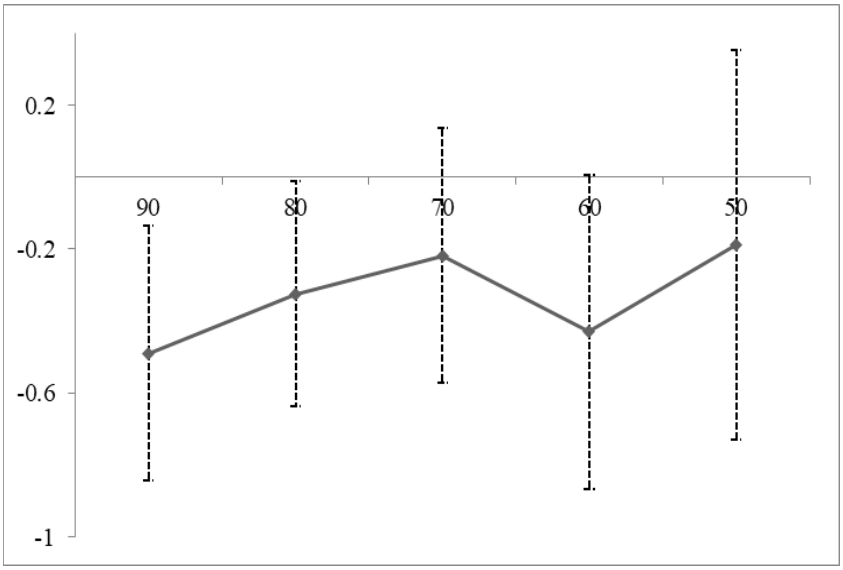
Nonetheless, the implications of caste homogeneity in areas more likely to be associated with higher social ties and coordination with stronger inter-personal relationship are likely to offer some insights into isolating the economic considerations from social cohesion. Since rural India is more likely to be characterised by these features, we examined whether the gains through caste homogeneity held in the less urbanised districts in the post-unlocking period (Banerjee, 1986). To provide evidence of this, we looked at the differential effects of unlocking in areas with varying levels of urbanisation. To understand this, we created a dummy variable, urb_k, which took the value of 1 for districts with less than k percent population living in urban areas and 0 for those with k or more than k percent urban population. We used this in our regression, given by the following:

$$\begin{aligned} ln\bar{Y}_{rd(unlock)} = \beta_0 + \ \beta_1 lnY_{rd(pre-lockdown)} + \beta_2 ln\bar{D}_{rd(lagged)} + \ \beta_3 HISG_{rd} \\ + \beta_4 HIRel_{rd} + \beta_5 GINI_{rd} + \beta_6 (HISG_{rd} * unlock) \\ + \beta_7 (HISG_{rd} * unlock * urb_k) + \beta_8 C_{rd} + \beta_9 (unlock * urb_k) + AR_r \\ + unlock + urb_k + \epsilon_{rdt} \dots \dots \dots (4) \end{aligned}$$

Here β_7 is our coefficient of interest.

Figure 9, which presents the findings with various levels of k, indicates exactly what we hypothesised. In areas with low urbanisation, the marginal gains from higher caste homogeneity seemed not only to remain intact even after the unlock process but also show reducing reduction in the growth in reported Covid-19 cases. Interestingly, for districts with a more than 80% or 90% rural population, the implications of caste homogeneity on the unlocking process were found to bear a significant and negative relationship with the number of reported cases, with the effect size being larger for the latter than the former. This underscores the importance of higher social cohesion and inter-personal ties. Coordinated action through strong caste networks is likely to play an important role in lowering the growth in infections. This was observed to be true even during the unlock period, when mobility was likely to increase and the likelihood of maintaining pandemic protocols like social distancing was lower. No such effect was observed for religion or economic inequality (see Appendix Figure 4A).

Figure 9: Coefficients of the three-way interaction to examine implications of rural population share on the caste homogeneity effects during the locking period



Notes: The coefficients come from the full control scheme, as described in model4, with confidence bands of 95%. The unlocking dummy takes the value 0 if data are from the period before the end of the nationwide lockdown and takes the value 1 if after unlocking. Detailed coefficients are available on request.

5 Policy implications

The key findings of the paper can be summarised via three important results. First, we have found that caste-group homogeneity has a significant influence in curtailing the spread of Covid-19 in India. Religious fragmentation, however, has a limited role. In addition, we found that income inequality exerted a modest negative influence on the growth of Covid-19 cases, one that was restricted to the initial phases of the lockdown. Accordingly, our results emphasise the importance of strong community cohesion – mainly working through caste-groups and to a lesser extent through economic affinity – in ensuring socially beneficial compliance with lockdown restrictions. A major challenge to sustaining the community norms here arises from the economic hardship caused by the subsequent lockdown phases. These are likely to intensify economic needs to violate compliance norms. Evidently, the gains through cohesion seemed to attenuate after the easing of the phases of mandatory lockdown. In this context, the paper first emphasises the importance of robust social protection and redistributive measures, which are vital links that need to be strengthened during the pandemic, not only to mitigate economic hardship but also to arrest the spread of infection through sustained community coordination.

Further, the effect of caste-group homogeneity was found to be more prominent in areas that were less urbanised; this effect continued to operate even after the mobility restrictions were withdrawn. It is our surmise that the relative importance of different dimensions of social fragmentation varies with the degree of urbanisation. In less urbanised districts, community cohesion is likely to be dominated by social identity. People in rural areas have a relatively restricted social network but one that entails strong ties relative to urbanised households, which have more contacts but weaker social ties (Sato & Zenou, 2015). In other words, urban interactions are more frequent but ties are voluntary in nature, while in rural areas they are mostly based on social norms and kinship (White & Guest, 2003). Studies have also found that people living in urban areas are less likely to participate in collective action (Bovaird et al, 2014). We believe that, among other factors, the nature of interactions and community structure in rural areas are the enabling reasons for a stronger relationship between homogeneity and spread of infections through better social cohesion. Importantly we have also observed that the community effect continued to display significant resilience in rural areas even in the unlock period when movement restrictions were partially lifted, even though its effect was found to disappear in more urbanised areas. Since Covid-19 cases are continuing to spread in India these possibilities open up a number of policy avenues for health interventions.

First, the government should prioritise socially heterogeneous areas relative to homogeneous ones, as the former might be more vulnerable to disease spread and possess lower social capital to tackle infection growth. Our findings provide a basis for identifying and zoning local areas into socially homogenous and heterogeneous blocks. The heterogeneous blocks should be prioritised for medical intervention as well as income support to compensate for the already weak community cohesion as a short and medium response to the pandemic. Second, in the relatively homogenous areas, as existing social ties are weakened with the gradual opening up of the economy, strategies to strengthen community networks might be undertaken. One possible channel could be decentralising health interventions through community channels, involving greater community participation. Research indicates that positive perceptions about one's community are found to enhance private compliance behaviour with respect to the pandemic (Das et al, 2020). Here, targeted intervention and educational messages catering to the relevance of the community through workers belonging to the same community can prove important. The effectiveness of these policies, if implemented, would then be an important area of future research.

Conclusion

Curtailing growth in infection requires social cohesion and strong community ties, something that might be difficult in a diverse and heterogeneous context. Such a setting places obstacles in the way of achieving the trust and cooperation required for coordinated community action during pandemics. Using daily data on Covid-19 infections at the district level, we found that social homogeneity through caste played a significant role in ensuring lower growth in Covid-19 cases in the Indian context. During the nationwide lockdown in particular, social cohesion played an important role in arresting the growth of infection. However, the marginal gains seemed to be reduced after the start of the unlocking process that potentially allowed for greater mobility. Importantly, the community effect continued to display significant resilience in rural areas even during the unlock period, although its effect was found to disappear in more urbanised areas. As we have accounted for a wide range of potential confounders, we have reasons to believe that our estimates reveal the impact of community cohesion on infection spread via coordinated compliance. This becomes even more robust after we account for the potential selection on unobservables, which might nullify the positive effects of caste-group homogeneity. Our analysis suggests that caste-wise heterogeneous areas, and to some extent economically unequal areas, are the most likely to be vulnerable to infection growth and that the government must prioritise a response in such areas while expanding its emergency capacity. In areas which are relatively homogenous, we advocate community strengthening efforts. Despite our robust findings, the exact mechanisms through which caste-based heterogeneity becomes important in India remain unexplored in our paper. We propose this as an important agenda for future research.

References

Ahmed, M.Z., Ahmed, O. Aibao, A., Hanbin, S., Siyu, L. and Ahmad, A. (2020). 'Epidemic of Covid-19 in China and associated psychological problems'. *Asian Journal of Psychiatry* 102092.

Alesina, A. and Ferrara, E.L. (2005). 'Ethnic diversity and economic performance'. *Journal of Economic Literature 43*, 762–800.

Alesina, A., Baqir, R. and Easterly, W. (1999). 'Public goods and ethnic divisions'. *Quarterly Journal of Economics 114*, 1243–1284.

Altonji, J.G., Elder, T.E. and Taber, C.R. (2005). 'Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools'. *Journal of Political Economy 113*, 151–184.

Asher, S., Lunt, T., Matsuura, R., & Novosad, P. (2019). The socioeconomic high-resolution rural-urban geographic dataset on India (SHRUG). *URL:* https://doi*.org/10.7910/DVN/DPESAK*

Bandiera, O., Barankay, I. and Rasul, I. (2005). 'Cooperation in collective action'. *Economics of Transition 13*, 473–498.

Banerjee, A., Iyer, L. and Somanathan, R. (2005). 'History, social divisions, and public goods in rural India'. *Journal of the European Economic Association 3*, 639–647.

Banerjee, B. (1986). *Rural to Urban Migration and the Urban Labour Market: A Case Study of Delhi.* New Delhi: Himalaya Publishing House.

Blair, H.W. (1996). 'Democracy, equity and common property resource management in the Indian subcontinent'. *Development and Change 27*, 475–499.

Bosancianu, C.M., Dionne, K.Y., Hilbig, H., Humphreys, M., Sampada, K.C., Lieber, N. and Scacco, A. (2020). *Political and Social Correlates of Covid-19 Mortality.* SocArXiv preprint. doi:10.31235/osf.io/ub3zd.

Bovaird, A., Van Ryzin, G.G., Loeffler, E. and Parrado, S. (2014). 'Activating citizens to participate in collective coproduction of public services'. *Journal of Social Policy* 44(1), pp. 1–23.

Daoust J-F (2020) Elderly people and responses to COVID-19 in 27 Countries. PLoS ONE 15(7): e0235590. https://doi.org/10.1371/journal.pone.0235590.

Das, U., Sarkhel, P. and Ashraf, S. (2020). 'Love thy neighbor? Perceived community abidance and private compliance to COVID-19 norms in India'. arXiv preprint. arXiv:2010.12350.

Dayton-Johnson, J. (2000). 'Determinants of collective action on the local commons: a model with evidence from Mexico'. *Journal of Development Economics 62*, 181–208.
Deshpande, A. (2000). Does caste still define disparity? A look at inequality in Kerala, India. *American Economic Review*, *90*(2), 322-325.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications: Monographs on Statistics and Applied Probability.* Boca Raton FL: CRC Press.

Field, E., Levinson, M., Pande, R. and Visaria, S. (2008). 'Segregation, rent control, and riots: the economics of religious conflict in an Indian city'. *American Economic Review 98*, 505–510.

Frenken, K. (2004). 'The Elgar companion to neo-Schumpeterian economics'. In Horst, J. and Pyke, A. (eds), *Entropy and Information Theory* (pp. 544–555). Cheltenham: Edward Elgar.

Habyarimana, J., Humphreys, M., Posner, D.N. and Weinstein, J.M. (2007). 'Why does ethnic diversity undermine public goods provision?'. *American Political Science Review,* 101(4)*,* 709–725.

Lancet. (2020). 'India under COVID-19 lockdown'. *Lancet.* doi:10.1016/S0140-6736(20)30938-7.

MaCurdy, T.E. and Pencavel, J.H. (1986). 'Testing between competing models of wage and employment determination in unionized markets'. *Journal of Political Economy 94*, S3–S39.

Miguel, E. and Gugerty, M. (2005). 'Ethnic diversity, social sanctions, and public goods in Kenya'. *Journal of Public Economics 89*, 2325–2368.

Mishra, S., Gayen, A. and Haque, Sk M. (2020). 'COVID-19 and urban vulnerability in India'. *Habitat International 103*. doi: 10.1016/jhabitatint.2020.102230.

Mitra, A. and Ray, D. (2014). 'Implications of an economic theory of conflict: Hindu–Muslim violence in India'. *Journal of Political Economy 122*, 719–765.

Mukhopadhyay, A. and Sahoo, S. (2016). 'Does access to secondary education affect primary schooling? Evidence from India'. *Economics of Education Review 54*, 124–142.

Munshi, K. (2019). 'Caste and the Indian economy'. *Journal of Economic Literature 57*, 781–834.

Munshi, K. and Rosenzweig, M. (2009). *Why is Mobility in India so Low? Social Insurance, Inequality, and Growth*. National Bureau of Economic Research (NBER) Working Paper w14850. Washington DC: NBER.

Munshi, K. and Rosenzweig, M. (2018). *Ethnic Diversity and the Under-supply of Local Public Goods*. Working Paper, University of Cambridge.

Oster, E. (2019). 'Unobservable selection and coefficient stability: theory and evidence'. *Journal of Business & Economic Statistics 37*, 187–204.

Poteete, A.R. and Ostrom, E. (2004). 'Heterogeneity, group size and collective action: the role of institutions in forest management'. *Development and Change 35*, 435–461.

Rhoades, S.A. (1993). 'The Herfindahl–Hirschman Index'. *Federal Reserve Buletin 188*.

Roychowdhury, P. (2019). 'Peer effects in consumption in India: an instrumental variables approach using negative idiosyncratic shocks'. *World Development 114*, 122–137.

Sato, Y. and Zenou, Y. (2015). 'How urbanization affects employment and social interactions'. *European Economic Review 75*, 131155.

Susewind, R. (2017). 'Muslims in Indian cities: degrees of segregation and the elusive ghetto'. *Environment and Planning A 49*, 1286–1307.

Tajfel, H., Billig, M.G., Bundy, R.P. and Flament, C. (1971). 'Social categorization and intergroup behaviour'. *European Journal of Social Psychology 1*, 149–178.

Vigdor, J.L. (2004). 'Community composition and collective action: analyzing initial mail response to the 2000 census'. *Review of Economics and Statistics 86*, 303–312.

Wade, R. (1994). *Village Republics: Economic Conditions for Collective Action in South India.* Cambridge: Cambridge University Press.

White, K.J. and Guest, A.M. (2003). 'Community lost or transformed? Urbanization and social ties'. *City & Community 2*, 239–259.