Determinants of Covid-19 Infections and fatalities in India: An Investigation using State and District Level Panel Data[1]

[Work in progress, Comments welcome]

Parantap Basu (Durham University Business School, UK)
(Email: parantap.basu@durham.ac.uk)
Phone +447872040275

Susmita Das (Dept. of Economics, Assam University, Silchar India)
(Email: dsusmita521@gmail.com)

Arnab Dutta Choudhury (Dept. of Economics, Assam University, Silchar India)
(Email: arnabdutta.duttachoudhury@gmail.com)

Ritwik Mazumder (Dept. of Economics, Assam University, Silchar, India)
(Email: ritwiksmazumder@gmail.com, ritwik.mazumder@aus.ac.in)

JEL Classification Code: I18, Government Policy, Regulation, Public Health; I31: General Welfare, Well Being.

Abstract

Using weekly covid infection and fatality data for all states in India, we investigate the determinants of covid cases focusing more on the second wave of infections. We include several macroeconomic indicators for states namely, per capita net state domestic product, the degree of urbanization, population density, percentage of aged population, share of agriculture amongst others. Our findings suggest that infections and fatalities have a clear rural-urban divide. Rural states which have higher percent of people living below the poverty line have experienced less cases and fatalities. Fatalities are more clustered in prosperous, urbanized and dense industrial states and it spread from urban to rural areas as suggested by block exogeneity tests involving rural and urban states as well as rural and urban district of an epicentre, Maharashtra. Unlocking significantly raised covid fatalities in all states. In general, aged populations suffered more fatalities.

[1] Hong Il Yu is duly acknowledged for answering econometric questions. The usual disclaimer applies.

1. Introduction

A staggering feature of Covid infections in India is its regional disparity. Basu and Mazumder (2021) document that confirmed cases are more concentrated in prosperous and urbanized regions with high population density. On the other hand, poor and less developed regions in India suffer fewer infections. In this present paper, we do a more comprehensive analysis extending Basu and Mazumder (2021) to understand the deep-rooted factors behind this regional variation of infections and fatalities. Basu and Mazumder offer a tentative hypothesis about low infections in poor regions; it is due to a combination of herd immunity and lack of tests. They did not include case fatalities and the second wave of data which we do in this paper.

We find a clear rural-urban divide of Covid cases and fatalities in India. Rural states classified by the urbanization ratio are predominantly poor as per the poverty indicator and infant mortality rate. These poor rural states experienced less Covid cases. Fatalities are more clustered in prosperous, urbanized and denser areas with lower poverty. However, from the end of the first wave, infections started spreading from urban industrial states to the rural states. This is confirmed by the block exogeneity tests involving rural and urban states as well rural and urban districts of Maharashtra which is a major epicentre of infections. This experience stands in sharp contrast with the experiences of US where poor black, Hispanic and indigenous population were exposed more to Covid infections than whites (Stafford, Hoyer and Morrison, 2020, Abedi et al. 2020). Similar pattern is also experienced in UK where black, Asian and Middle Eastern (BAME) groups suffered more infections (ONS 2020).

In addition, we find that removal of lockdown during the end of the first wave (8th June, 2020) contributed to a surge of cases. Another significant finding of our study is that aging has a significantly positive impact on covid deaths.

Our data source is the real time database available in www.covid19india.org.which is the most comprehensive dataset for Covid infections and fatalities in different regions of India widely used by researchers. We use state level data for all-India analysis and district level data for our case study on Maharashtra, one of the worst victims of covid-19 irrespective of the waves of infections in India. Our principal variables of interest are confirmed cases per

million state population, deaths per million, tests conducted per million, and several regional macro development indicators such as per capita net state domestic product, the degree of urbanization, population density, state-level head count poverty, percentage of aged population (60 years plus), percentage state GDP from agriculture and allied activities, a wealth inequality GINI. Most of these state level socio-economic and demographic features are drawn from the Census of India 2011, and other sources.

The paper is organized as follows. In section 2, we review the related literature. Section 3 discusses data, measurement and econometric issues. Section 4 reports the results of state level panel regressions. Section 5 presents the district level analysis using Maharashtra as a case study. Section 6 concludes.

2. Literature

Several recent studies have examined the trend in Covid-19 infections in India and its regional disparity. The most notable study is by Jalan and Sen (2020a) who point out using district level data that not all of India has been impacted uniformly by Covid-19, and that there was a strong case in favour of implementation of a more selective lockdown. They find that the heavily affected districts are in metro areas of Delhi, Mumbai, Indore, Jaipur, Chennai, Pune amongst others. The 20 significantly-affected districts included Agra, Ahmedabad, Bengaluru, Coimbatore, and Thane. Next, 42 districts were moderately-affected, and finally, 188 districts were mildly-affected. The unevenness in the spread of Covid-19 is also borne out by the fact that Jammu and Kashmir, Telangana, and an additional 16 per cent of all the districts in the rest of India together reported 86 per cent of all Covid cases in the country.[2] In a later study for the state of Kerala, Jalan and Sen (2020b) find that the state government controlled the first Covid wave by pre-emptively formulating a comprehensive set of *public actions* which were complied by state citizens. This was achieved by leveraging and reinforcing the citizen's *public trust* in the state. Specifically, the state's pandemic response contained supportive measures to ensure protection of all lives

[2]For a discussion on the methodology behind categorization of districts into heavily affected, significantly affected, moderately affected and mildly affected see Jalan and Sen (2020**a**, available at https://theprint.in/opinion/understand-the-method-in-covid-19s-madness-india-doesnt-need-complete-lockdown/398925/ and also in *The Wire,* May 5, 2020).

and livelihoods. The regional disparity in COVID infections in India has been also observed by Mandi, *et al* (2020) where they *construct a multi-dimensional vulnerability Index for Indian districts that can provide guidance for lifting the lockdown sequentially.* Further, Ray and Subramanian (2020) also noted this regional disparity in infections although their primary goal was to provide an interim report on the Indian lockdown provoked by the COVID-19 pandemic. Basu and Mazumder (2021) take the leads given by these studies and investigate the role of socio-economic determinants in determining the regional disparity in cases. However, their work was based on first wave of data and did not include fatalities.

In the backdrop of these studies, it is worth noting that our central research question is: how socio-economic and macro development indicators explain the overwhelming covid 19 death differentials across Indian states? Although these extant studies look at various ramifications of regional disparity of infections, this particular research question has largely remained unanswered.

Data

Our key data source for covid-19 related statistics is the national covid-19 portal for India https://www.covid19india.org/ which has been daily updated across states and districts of India since the onset of the pandemic in 2020. We take weekly cumulative total covid-19 figures for both confirmed cases per million (we say CASES) as well as fatalities per million (we call DEATHS in this paper) across 33 states and union territories of India. Our start date is 23-03-2020 and end data is 27-09-2021. We cover both waves of infections in India. Since our series comprises of weekly cumulative total confirmed cases and deaths (per million populations), the first difference yields the weekly new cases or deaths per million. Apart from covid infections and deaths, we compile state level development and socio-economic indicators primarily from Census of India, 2011 and few other sources (listed in Appendix 4 along with precise definition of each). These state level indicators are time invariant or fixed factors that vary only across states but not over time. The state level macroeconomic indicators are PCNSDP (per capita state net domestic product), URBAN (percentage of urban population at the state level), DENSITY (state level density of population), AGRI (percentage state domestic product from agriculture and allied activities), BPL (percentage of state

population lying below the poverty line), IMR (the infant mortality rate) and WFA (percentage of total work force engaged in agriculture and allied activities). Finally the variable AGED represents the percentage of 60 years plus population at the state level which we take as the state level old age population.

Underreporting of fatalities

For Covid research about India, one encounters a formidable problem of underreporting of cases, particularly fatalities. This may quite legitimately raise doubts about the reliability of our regression results. Given that our research focuses on determinants of fatalities, the underreporting typically gives rise to a measurement error issue for the dependent variable. To see it clearly, define $\tilde{y}_{it}$ = reported fatalities at date t in the i^{th} state, y_{it}=actual fatalities and v_{it} is a positive measurement error representing the underreporting. In other words, $\tilde{y}_{it} = y_{it} - v_{it}$. Let x_{it} be the vector explanatory variables. Our true regression equation is then: $y_{it} = \alpha + \beta x_{it} + u_{it}$ where u_{it} is the underlying error term which captures all omitted variables. In the actual regression with observed fatalities as the dependent variable is: $\tilde{y}_{it} = \alpha + \beta x_{it} + e_{it}$ where the composite error term is $e_{it} = u_{it} - v_{it}$. If u_{it} has zero conditional mean assumption then $E(e_{it}|u_{it})) = -E(v_{it}|x_{it})$. The bias then depends on the property of the measurement error. If the error does not depend on the independent variable, and we assume that $E(v_{it}|x_{it}) = \mu$ for some positive constant c for all i an t, then the estimator α is biased because it is $\alpha - \mu$ but the estimator of β is unbiased and consistent.

However, if the measurement error depends on the independent variables, in the sense that $E(v_{it}|x_{it})$ changes with x_{it} then we have the usual omitted variable bias problem affecting our estimator of β is then inconsistent. We need to find a suitable set of instrumental variables (IV) to rectify this bias. In this paper we report both OLS and IV.

3. Empirical Analysis

In figure 1 we plot the per capita NSDP, confirmed cases per million state population and the covid deaths per million across states after expressing each variable in a 0 to 1 scale for the sake of comparability. The near perfect synchronization between cases and fatalities suggests that cases and deaths are concentrated in the richer states. Motivated by this plot we compute the ordinary correlations between variables of interest. The pairwise ordinary correlation coefficients of "Deaths "in the cross-section of 33 states and Union Territories of

India are in presented in Table 1. Clearly deaths have a strong and positive (statistically significant) association with PCNSDP, URBAN and DENSITY implying that the covid deaths in India during the first wave are concentrated in the richer, more urbanised and densely populated states, which is very consistent with our findings on confirmed cases (CASES to be precise).

[Figure 1 and Table 1 come here]

Few observations are in order. First, Poor states classified by BPL are predominantly rural as suggested by the significant negative correlation between BPL and urban. The median BPL of urban states is 9.91 while rural states have a median of 29.43. Cases and fatalities are lower om poor states as indicated by the significant positive correlations of URBAN with CASES and DEATHS. Second, agriculturally dominant states have suffered less as is evident from the high negative correlation between AGRI and DEATHS (correlation touching almost -0.8). Agriculturally dominant states are relatively more rural in nature and thus have lower degrees of urbanization and lower population densities. Third, DEATHS and CASES correlate negatively wth IMR implying that states with poorer health infrastructures and poorer health status have had lesser covid deaths. Poor health infrastructure as well as low purchasing power, among several other associated factors can lead to high IMR at the state level. Surprisingly, these states also experienced less Covid cases and fatalities.

Motivated by these correlations, we next run a log-linear cross-state regression to focus on various developmental determinants of infections across India. We choose PCNSDP, URBAN, DENSITY, BPL and AGRI as explanatory developmental variables. Five model specifications are estimated which are reported in Table 2. Although AGRI is negative and significant throughout, URBAN, BPL and the interaction between URBAN and BPL are insignificant when we try to explain LOG(DEATHS). Furthermore, WFA is negative and significant implying that other things unchanged, a higher work force engaged in agricultural and allied activities lowers covid related deaths. In other words, states largely engaged in agrarian activities have suffered less during the first wave of covid-19. Finally, PCNSDP is highly significant and positive across models 4 and 5 implying that other things unchanged, richer states have experienced more covid fatalities.

[Table 2 comes here]

4.2 Dynamic panel regression

The dynamic panel regression results of weekly cumulative total reported deaths due to covid-19 during the first wave of covid infections in India are presented in Table 3. All models have a first order autoregressive term in the form of LOG(DEATHS(-1)) and this ensures no serial correlation in the residuals (all Durbin-Watson statistic values are close to 2.00 in Table 3). In model 2 AGRI is significant and negative implying that non-industrial states have had lesser deaths. The first phase of reopening in India during the first wave was termed Unlock 1.0 and the process started on 8th June 2020. We insert an UNLOCK dummy (which simply assigns score 0 to pre-June 8, 2020 observations and 1 to others) in model 2 and find that it is statistically significant and positive for the first wave of deaths which obviously means that deaths shot up after the lockdowns were relaxed. URBAN and the URBAN-BPL interactions do not explain deaths as seen in model 3. This is very different from what we find for CASES over the same period reported in Basu and Mazumder (2021). In fact apart from the lagged dependent variable none of the regressors in model 3 are significant. Finally PCNSDP and DENSITY have significant and positive coefficients which are consistent with our regressions for CASES as in Basu and Mazumder (2021). The richer and more densely populated states have a higher chance of covid related fatalities which is perfectly consistent with our earlier findings regarding the LOG(CASES) regressions.

[Table 3 comes here]

The panel regression results of weekly (new) covid deaths per million on state level factors during the first wave of infections in India are presented in Table 4. The new covid deaths are measured by the first difference of log cumulative cases. Due to the non-linear (parabolic to be precise) nature of the weekly death count (similar to the treatment of confirmed weekly covid-positive case count, Basu and Mazumder, 2021) time and time-squared terms are introduced along with the an AR(1) term for one period lagged weekly deaths [i.e., D log(DEATHS(-1)). The first 4 models in Table 4 are the same as Basu and Mazumder (2021) while the 5th model is an addition keeping in mind the key objective of the

study. There are few features that suggest robustness and consistency. The lagged weekly deaths are around 0.8 suggesting high persistence of new cases.

Time and time-squared terms are significant with expected signs, clearly suggesting a parabolic growth pattern of deaths (in fact very similar to confirmed cases) which means deaths peak and then levels off. The Durbin –Watson statistic is close to 2.0 which suggests that the resulting residuals are serially correlated errors are not serially correlated affirming consistency of the estimates.

What is noteworthy are the signs of the coefficients of URBAN and AGRI in Model 1 are expected but both are statistically insignificant. In fact, BPL which had a significantly negative influence on CASES is also insignificant although it still has a negative sign. Across models 1 to 4 in fact, BPL is found to be insignificant. Density turns out to be a very significant variable in explaining weekly deaths (Model 2) and so is PCNSDP. Model 5 is a new addition where an UNLOCK dummy is introduced, BPL is replaced by IMR[3], and AGRI is kept as the only state level structural factor. AGRI is statistically significant and negative implying that other things unchanged relatively more agricultural states would tend to have lower covid deaths. The big change in model 5 is that the coefficient of POOR dummy is now significant at 10 per cent whereas the BPL coefficient was insignificant throughout models 1 to 4. The UNLOCK dummy is significant and positive (reinforcing our model 2 results in Table 3) which confirms that that unlocking during the first wave of covid-19 infections in India significantly raised covid related deaths. However inserting IMR as a proxy for BPL in model 5 we get a negative coefficient but even this coefficient turns out to be insignificant. On the whole it is evident that poverty fails to explain DEATHS during the first wave.

[Table 4 comes here]

Table 5 presents a similar family of models on the second wave of deaths in India (roughly March to September, 2021). Several observations are common for the second wave estimates. First, BPL is insignificant across models. In Model 5, BPL is replaced by IMR and

[3] Infant mortality rate (IMR) is seen to be significantly positively correlated with BPL in Table 1 implying that poorer states have higher IMR. Since backward and disadvantaged states in terms of heath parameters are mostly poorer states, we take IMR as a proxy for BPL and verify the consistency of our estimates.

its coefficient turns out to be negative and highly significant implying that backward states with poorer overall health status have suffered less deaths during the second wave. Note that IMR is high in poor and underdeveloped Indian states. High IMR is associated with poor health status, and the negative sign of the coefficient of IMR in Model 5 implies that other things unchanged states with poor overall health status have suffered less in terms of deaths per million. This is consistent with the significant negative correlation (i.e. -0.58 to be precise) found between deaths per million and IMR shown in table 1. Second, URBAN and DENSITY are both significant but have opposite signs in model 1 which is in sharp contrast to the Table 3 results for the first wave. Controlling for DENSITY, the degree of urbanization has a positive impact on DEATHS. Third, other things unchanged, agrarian states have fewer deaths during second wave which is very similar to the first wave findings in Table 3.

[Table 5 comes here]

In Table 6 we report the results of both waves of data on weekly cumulative total DEATHS along with our usual time invariant state specific socio-economic and structural factors and introduce a SECONDWAVE dummy that assigns 1 to all second wave observations and 0 otherwise (noting that we have 32 weekly observations for second wave and 47 for first wave thereby giving us 79 weekly observations for 33 states and UTs). The SECONDWAVE dummy is also allowed to interact with the linear TIME trend term besides our usual TIME – SQUARED term that captures the non-linear nature of total covid deaths over the weeks.

[Table 6 comes here]

In Table 6 we finally have a glimpse of the relative impact of the second wave on covid-19 fatalities in India. Across models BPL is once again insignificant in explaining DEATHS. However, Model 1 shows that controlling for density, relatively more urbanised states have had more deaths. Further the SECONDWAVE dummy is positive and highly significant implying that on the whole India recorded significantly more covid-19 deaths during the second wave as compared to the first. Arguably this could be due to that lack of any stringent lockdown and monitoring during the second wave (unlike the one done during the first wave right from the end of March 2020) when most restrictions by respective states across the country were relaxed to a great extent. More importantly the

TIME*SECONDWAVE coefficient is significant and consistently positive across the models suggesting a small structural break in the trend of fatalities after the second wave. The URBAN*BPL interaction term is insignificant implying that the poverty after all is not significantly associated with deaths although our correlation matrix suggests a weak negative association. However this finding is on the basis of the pooled first and second wave data. This is in sharp contrast to our findings on confirmed cases during first wave (in Basu and Mazumder, 2021) where confirmed cases (i.e., covid-19 infections) were found to be significantly lower in the poorer states and poverty had a dampening impact on infections.

Estimates based on the combined data for the first and second waves reveal that DENSITY is insignificant in explaining deaths although this is not the case when the first and second waves are analysed separately (see Tables 4 and 5). Finally, in Model 4, after having controlled for IMR, the coefficient of PCNSDP turns out to be significant. Thus overall, our findings seem to suggest that richer, densely populated and relatively more urbanised states have suffered more covid-19 deaths in India while agricultural states have suffered less. Deaths have little to do with state level poverty. The unlocking during the first wave of infections in India significantly raised deaths and the second-wave weekly death counts were clearly higher than that observed during the first wave.

Age composition and Covid fatalities

Of particular interest is the impact of the percentage of old age population (at the state level) on state level deaths in India. We run two pooled regressions where DEATHS are explained on the basis percentage of 60 years plus population (we call AGED), AGRI, LEB (life expectancy at birth) controlling for the state level inequality measure (the GINI coefficient capturing wealth inequality, taken from Pandey and Gautam, 2020, for 31 states and UTs of India). For the sake of consistency checking we further introduce the positivity ratio (i.e., the ratio of CASES to TESTS) as an additional regressor in the second model. Old age population proportion, i.e., AGED is consistently statistically significant and positive across models implying that everything else equal, the higher the percentage of old age population at the state level, the higher the covid deaths per million. The simple correlation between DEATHS and AGED turns out to be 0.586 (significant at 1%; not reported in the correlation matrix) implying that states with higher old age populations have had more covid deaths.

LEB has a significant and negative coefficient in model 2. So, better life expectancy states have had lower deaths (although the correlation between the two is just -0.275, which is not significant at 10%). Our results are robust even after including the positivity ratio or CASES/TESTS ratio as a control factor in model 2.

[Table 7 comes here]

A pertinent econometric question at this juncture is whether our key regressors are exogenous in the LOG(DEATHS) regression models, as endogeneity leads to biased and inconsistent parameter estimates. In order to test this we conduct a series of 2SLS – IV regressions where exogeneity of our key regressors are tested using suitable instrumental variables under a 2 stage least squares set-up. The results are serially presented in Appendix 1. In particular four pivotal variables (state level macro and structural indicators) are tested for exogeneity here, i.e., PCNSDP, URBAN, DENSITY and AGRI. The Sargan-Hansen J-statistic is routinely reported and its statistical insignificance suggests that the orthogonality condition cannot be rejected.

4. Did infections spread from Urban to Rural areas in India?

A topical question is whether covid-19 infections (CASES here) spread from the urban to rural areas in India. To answer this question we first provide a brief anecdote of how exactly the covid-19 started its spread in India. The first cases of COVID-19 in India were reported on 30 January 2020 in three towns of Kerala, among three Indian medical students who had returned from Wuhan, the epicentre of the pandemic. Lockdowns were announced in Kerala on 23 March and in the rest of the country on 25 March[4]. The covid 19 is relatively more concentrated in the urbanized states of Maharashtra, Kerala and Goa. In fact, Maharashtra which saw its first covid positive case on 9th March[5], accounted for nearly 22.35 % of the total cases in India as well as about 30.55 % of all deaths till May 2021. Besides these three states, Delhi National Capital Region and Chandigarh are highly urbanized, rich and densely populated cities that are internationally well connected.

[4] Narasimhan, T. E. (30 January 2020). "India's first coronavirus case: Kerala student in Wuhan tested positive". *Business Standard India*

[5] "Covid-19 state tally: Cases soar to 33,053 in Maharashtra, nearly one-third of national total". *Hindustan Times*. 18 May 2020.

In case of Delhi the first case of COVID-19 infection was reported on 2 March 2020 (same as the first detected case in Hyderabad, Telengana)[6]. The city of Delhi alone has the sixth-highest number of confirmed cases of COVID-19 in India (till May 2021), after Maharashtra and Tamil Nadu which is a staggering statistic. The first case for Chandigarh was recorded on 19 March 2020[7], and by 27th September 2021 it had 55,315 confirmed cases per million population making it the 7th worst state in terms of infections per million. Currently, Kerala, Goa, Puducherry, Chandigarh, Maharashtra and Delhi are amongst the worst states in terms of confirmed cases as well as deaths per million. Thus most states with the highest covid numbers in India saw their first confirmed cases several weeks before the onset of the infamous lockdown on 24-04-2020. Clearly, the initial infections in India were all concentrated mostly in the relatively more urbanized and developed regions and that too in metro and million cities that are internationally well connected.

This motivates us to ask whether infections spread more dense urban states of India to predominantly rural states. Our working hypothesis in this case is "infections or cases (per million) in rural areas have been caused by cases in urban areas". We test this first using the infection data for all states in India by segregating states into more urbanised and less urbanised states. Although India predominantly has a rural dominance with around 35 present of its population living in urban areas, we still segregate 34 states and union territories on the basis of median level of degree of urbanisation (i.e., on the basis of median value of URBAN defined here as the percentage of urban population at the state level). The 17 states above the median level of urbanisation (35.78% in this case) are stacked under urban states while the remaining 17 are kept under rural states. Adding up infections per million or CASES across the 17 urban states for each time point yields the time series of urban cases per million at the all-India level (we call it UCASES). An analogous summation of cases per million across all rural states for each time point over our study period yields RCASES. This generates 79 time points (covering both waves of infections in India) of aggregate urban cases per million (UCASES) and aggregate rural cases per million (RCASES).

[6] As reported on the online covid portal available at https://www.covid19india.org/state/DL.
[7] "23-yr-old woman tests positive, 1st coronavirus case in Chandigarh". *The Economic Times*. 19 March 2020.

Our empirical problem boils down to the testing for econometric causality between LOG(UCASES) and LOG(RCASES). Before running any formal block exogeneity test, we first test for unit roots of each series in the presence of structural breaks. We employ the Zivot-Andrews test which tests for unit roots while detecting a structural break in the series. Results are reported in Tables 8 and 9. Both series are found stationary after a successful identification of the breakpoint. The break points for both rural and urban cases occur during the second wave. However, it is noteworthy that the urban break date (23rd March 2021) precedes the rural break date (6th April, 2021) by about a fortnight. In other words the urban infection spike occurred a couple of weeks earlier compared to the rural spike.

Taking this precedence of outburst of cases for urban areas as our stepping stone, we next perform a Granger causality/block exogeneity test for rural vs urban infections. An optimum lag length as 4 was determined by the standard Akaike-Schwartz test. A Vector Auto-regression (VAR) model for LOG(UCASES) and LOG(RCASES) in first difference as suggested by the Augmented Dickey-Fuller test (reported in Appendix 2 (Tables A2.1 and A2.2). For the sake of model fit we insert our usual UNLOCK dummy along with the SECONDWAVE dummy as exogenous factors in our 4 period VAR. We then run a pair of block-exogeneity tests where two hypotheses are simultaneously tested.

In the top half of table 10, we present the test results of the hypothesis that lagged LOG(RCASES) do not explain LOG(UCASES) in first difference. That is, current urban cases are not statistically explained by past rural cases. The p value suggests that this hypothesis can be rejected only at 14.7% significance. In view of this, we infer that rural infections did not cause the urban infections. But is the converse also true? In the second half of the table we test for the second hypothesis that lagged LOG(UCASES) do not explain LOG(RCASES) in first difference. The chi-square value is highly significant in this case implying that the second null hypothesis is rejected. Thus lagged LOG(UCASES) explain LOG(RCASES) implying that the urban infections have contributed to rural infections in India and the unidirectional of causality runs from UCASES to RCASES.

[Tables 8, 9 and 10 come here]

A Case study of Maharashtra

Is this causal sequence of urban rural infections true for district level data for major states? To this end, we pick Maharashtra which is an epicentre of covid infections as a case study. We perform the unit root tests for structural breaks and block exogeneity tests for all 36 districts of Maharashtra segregating districts as usual using the median level of urbanisation as a cut-off point. The median degree of urbanisation is 45.23% for Maharashtra based district wise figures of Census 2011. We conduct the breakpoint and causality tests for both waves.

For the first wave, the Zivot-Andrews breakpoint unit root test (not reported here) does not yield any significant break dates for LOG(UCASES) or LOG(RCASES). Following the same procedure as the all-India UCASES – RCASES causality we run a 4 period lagged VAR on the same variables for Maharashtra with only the UNLOCK dummy as an exogenous factor. The block exogeneity test results are presented in table 11. The hypothesis that lagged LOG(RCASES) do not explain LOG(UCASES) is accepted at 8.6% while the hypothesis that lagged LOG(UCASES) do not explain LOG(RCASES) is rejected at less than 0.01% (chi-square value being very highly significant). Thus for the state of Maharashtra district level infections data during the first wave suggests that urban infections have caused rural infections.

[Tables 11 comes here]

For the second wave of infections for Maharashtra we conduct the same structural break point unit root tests (results are in table 12) and find that for urban cases the break date is 29th of March, 2021 whereas the rural infections the break date is 5 weeks later, i.e., 3rd of May, 2021. This indicates that infections in rural Maharashtra have peaked at a much later date (more than a month to be precise) compared to urban Maharashtra.

[Tables 12 and 13 come here]

To test the Urban – Rural infections causality during second wave on the basis of district level data for Maharashtra, we once again conduct our post–VAR Block exogeneity/Granger causality tests keeping the UNLOCK dummy variable as an exogenous factor.

[Tables 14 comes here]

As the results in table 14 show, RCASES do not cause UCASES is accepted at 18.3 percent while UCASES do not cause RCASES is rejected at 1 percent implying that urban infections have caused rural infections even during second wave in Maharashtra and not the other way round. Thus the consistent finding for the state of Maharashtra seems to be that the direction of transmission of the covid was from urban to rural districts during both waves of infections in India. It is rather surprising that we do not find bi-directional causality during the second wave on the basis of district data, especially keeping in mind that at the end of the first wave, the covid had already spread to the rural areas and rural to urban transmission was also a practical possibility.

6. Summary and Conclusions

In this paper we explain the inter-state variations in covid-19 deaths in India on the basis of state level socio-economic, demographic and development indicators using panel data. We cover both the first and second waves of infections. Further, we test whether the transmission of the covid has been from urban to rural India. The transmission mechanism is studied using econometric causality analysis on the basis of time series data for urban and rural infections in the aggregate. The current study is an extension of Basu and Mazumder (2021) where the large interstate variations in covid-19 infections were explained on the basis of a similar set of all-India panel data only for the first wave of infections in India. However, this paper extends the analysis by incorporating second wave data for infections and deaths and further by investigating the transmission of the infection across urban and rural areas using econometric causality. In line with our previous work on confirmed cases we find the covid-19 fatalities in India are concentrated in the richer, urbanized, and densely populated states of India. The static and dynamic panel analysis suggests that agricultural states and states with poor health status have not suffered as much as the developed states. Remarkably, deaths are higher in states with higher proportions of old-age population implying that covid deaths have something to do with population aging even in a young age dominated, low income country like India.

Since covid-19 entered India purely through its metro and million cities with major international airports (being gateways to the country) there is clear evidence on the basis of state level data that the direction of transmission of the infection was from urban to rural India and not the converse. The unlocking had a significant role to play both for infections as well as fatalities since it aggravated both. Structural break tests in our time series suggests that the urban deaths and infections peak seem to have been reached slightly earlier compared to the rural peak. This is consistent with our empirical claim regarding the direction of transmission of the covid. We repeat the urban-rural causality exercise on district level infections data for the first and second waves taking the state of Maharashtra as a test case. We empirically establish once again that even at the district level for Maharashtra the transmission of infections during both waves was from urban to rural and not the other way round. Remarkably the all-India state level and district level findings (only for Maharashtra though) are consistent in the sense that bidirectional causality of urban-rural transmission is entirely absent.

Our exercise gives rise to a few vital policy directions. First, given that richer, and denser regions of India have primarily suffered from covid deaths, there is a clear case for targeted interventions and lockdowns in case of a future outbreak. Second, since aging clearly is a factor explaining deaths, there is a case for separation of the young and working from the old across households, particularly in case of a future outbreak. Third, if transmission has been from urban to rural areas irrespective of the periods of lockdown, it speaks poorly of the covid monitoring mechanism in India. Better monitoring and selected lockdowns are imperative.

Appendix 1

A1. Testing for Exogeneity of Regressors

For the sake of statistical robustness we need to check whether our key regressors explaining deaths throughout the analysis are indeed exogenous regressors. We test for the exogeneity of four of our key explanatory variables namely, PCNSDP, URBAN, DENSITY and AGRI. Here in Table A.1, we take PCNSDP as the only variable that explains deaths. We take 2 instruments, DENSITY and AGRI and run the 2SLS which yields a J-stat value of 1.59 which is insignificant even at 20%. The null hypothesis here is that PCNSDP is exogenous. In a similar fashion we run the 2SLS-IV models for

URBAN, DENSITY and AGRI. The tests show that our explanatory variables of interest are indeed exogenous once the instruments are judiciously chosen.

Table A1.1 The 2SLS – IV Regression for exogeneity of LOG(PCNSDP) Instrumental Variables: LOG(DENSITY), LOG(AGRI)				
Dependent Variable: LOG(DEATHS) Variables	Coefficient	Std. Error	t-Statistic	Prob.
C	-15.721	4.982	-3.156	0.0035
LOG(PCNSDP)	1.746	0.431	4.046	0.0003
R-squared	0.329	Mean dependent var		4.425
Adjusted R-squared	0.308	S.D. dependent var		1.205
S.E. of regression	1.003	Sum squared resid		32.164
F-statistic with p-value	16.373 (0.000)	Durbin-Watson		2.062
J-statistic	1.599	Instrument rank		3
Prob(J-statistic)	0.206	Inference: LOG(PCNSDP) is exogenous in the LOG(DEATHS) regression		

Source: Computed by the authors with secondary data.

Note: The figures in the table are EVIEWS 10 generated during second stage regression of LOG(DEATHS) on first stage estimates of LOG(PCNSDP). At every stage standard errors are HAC adjusted.

Table A1.2 The 2SLS – IV Regression for exogeneity of LOG(URBAN) Instrumental Variables: LOG(DENSITY), LOG(IMR)				
Dependent Variable: LOG(DEATHS) Variables	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.132	1.933	-0.586	0.562
LOG(URBAN)	1.599	0.553	2.891	0.007
R-squared	0.077	Mean dependent var		4.425
Adjusted R-squared	0.048	S.D. dependent var		1.205
S.E. of regression	1.176	Sum squared resid		44.219
F-statistic with p-value	8.358 (0.000)	Durbin-Watson		2.170
J-statistic	0.004	Instrument rank		3
Prob(J-statistic)	0.950	Inference: LOG(URBAN) is exogenous in the LOG(DEATHS) regression		

Source: Computed by the authors on the basis of secondary data. **Note**: The figures in the table are EVIEWS 10 generated during second stage regression of LOG(DEATHS) on first stage OLS estimates of LOG(URBAN). At every stage standard errors are HAC adjusted.

Table A1.3 The 2SLS – IV Regression for exogeneity of LOG(DENSITY)

Instrumental Variables: LOG(URBAN), LOG(WFA)				
Dependent Variable: LOG(DEATHS) Variables	Coefficient	Std. Error	t-Statistic	Prob.
C	1.058	0.530	1.996	0.054
LOG(DENSITY)	0.560	0.079	7.092	0.000
R-squared	0.028	Mean dependent var		4.425
Adjusted R-squared	-0.002	S.D. dependent var		1.205
S.E. of regression	1.206	Sum squared resid		47.013
F-statistic with p-value	7.052 (0.000)	Durbin-Watson		2.160
J-statistic	0.344	Instrument rank		3
Prob(J-statistic)	0.557	Inference: LOG(DENSITY) is exogenous in the LOG(DEATHS) regression		

Source: Computed by the authors on the basis of secondary data. **Note**: The figures in the table are EVIEWS 10 generated during second stage regression of LOG(DEATHS) on first stage OLS estimates of LOG(DENSITY). At every stage standard errors are HAC adjusted. WFA is the percentage work force in agriculture and allied activities.

Table A1.4 The 2SLS – IV Regression for exogeneity of LOG(AGRI) Instrumental Variables: LOG(IMR), LOG(LITGAP), LOG(BPL)				
Dependent Variable: LOG(DEATHS) Variables	Coefficient	Std. Error	t-Statistic	Prob.
C	6.744	0.534	12.692	0.000
LOG(AGRI)	-1.001	0.167	-6.002	0.000
R-squared	0.343	Mean dependent var		4.425
Adjusted R-squared	0.322	S.D. dependent var		1.205
S.E. of regression	0.992	Sum squared resid		39.333
F-statistic with p-value	9.059 (0.000)	Durbin-Watson		1.850
J-statistic	0.702	Instrument rank		4
Prob(J-statistic)	0.704	Inference: LOG(AGRI) is exogenous in the LOG(DEATHS) regression		

Source: Computed by the authors on the basis of secondary data. **Note**: The figures in the table are EVIEWS 10 generated during second stage regression of LOG(DEATHS) on first stage OLS estimates of LOG(AGRI). At every stage standard errors are HAC adjusted. LITGAP is the male–female gap in state level adult literacy rate simply computed as male adult literacy rate minus female adult literacy rate reflecting the gender inequality in basic education.

Appendix 2: Unit Root Tests of aggregated Urban and Rural CASES and VAR-lag Selection

Unit Root Tests of Natural Log of Urban and Rural Cases

Table A2.1 Augmented Dickey-Fuller test of LOG(UCASES)		
Null hypothesis: D(LOG(UCASES)) has a unit root	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-5.043	0.000

Critical values	1% level	-2.597	
	5% level	-1.945	
	10% level	-1.614	

Source: Estimated by the authors on the basis of secondary data.

Notes: 1. *EVIEWS-10 generated MacKinnon (1996) one-sided p-values. 2. UCASES is the time series of total confirmed cases per million populations for all urban states of India taken together. 3. No trend or intercept included in model. 4. No. of lags taken is 2 (Automatic based on SIC; max=4) and included time points =79.

Table A2.2 Augmented Dickey-Fuller test of LOG(RCASES)			
Null hypothesis: D(LOG(RCASES)) has a unit root		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-4.321	0.000
Critical values	1% level	-2.597	
	5% level	-1.945	
	10% level	-1.614	

Source: Estimated by the authors on the basis of secondary data.

Notes: 1. *MacKinnon (1996) one-sided p-values. 2. RCASES is the time series of total confirmed cases per million populations for all rural states of India taken together. 3. No trend or intercept included in model. 4. No. of lags taken is 3 (Automatic based on SIC; max=4) and included time points = 79.

Table A2.3 VAR Optimum Lag Order Selection Criteria

Endogenous variables: D(LOG(UCASES)) D(LOG(RCASES))
Exogenous variables: C SECONDWAVE UNLOCK
Sample size: 79
Included observations: 70

Lag	LogL	LR	FPE	AIC	SC	HQ
0	328.5053	16.26893	4.82e-07	-8.871579	-8.293395	-8.641917
1	338.2250	16.38474	4.11e-07	-9.035001	-8.328331	-8.754303
2	359.7285	35.01995	2.50e-07	-9.535100	-8.699945	-9.203367
3	370.4520	16.85120	2.07e-07	-9.727200	-8.763559	-9.344430
4	379.2699	13.35282*	1.82e-07	-9.864855*	-8.772728*	-9.431049*
5	384.6464	7.834270	1.76e-07*	-9.904182	-8.763570	-9.419340

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Source: Computed by the authors using EVIEWS-10.

Notes: 4 period lagged terms are selected based on minimized values of LR, SC and HQ statistics.

Table A2.4 Vector Auto regression Estimates of LOG(UCASES) and LOG(RCASES) in first difference

Sample (adjusted): 06 79
Included observations: 74 after adjustments
Standard errors in ( ) & t-statistics in [ ]

	D(LOG(UCASES))	D(LOG(RCASES))
D(LOG(UCASES(-1)))	0.475635	0.300165
	(0.14683)	(0.19905)
	[3.23936]	[1.50800]
D(LOG(UCASES(-2)))	0.400227	0.548689
	(0.12898)	(0.17485)
	[3.10303]	[3.13806]
D(LOG(UCASES(-3)))	0.058783	0.124290
	(0.07714)	(0.10457)
	[0.76206]	[1.18858]
D(LOG(UCASES(-4)))	-0.113699	-0.087303
	(0.03368)	(0.04565)
	[-3.37637]	[-1.91239]
D(LOG(RCASES(-1)))	0.114994	0.701945
	(0.09995)	(0.13549)
	[1.15056]	[5.18073]
D(LOG(RCASES(-2)))	0.033172	-0.271237
	(0.09231)	(0.12514)
	[0.35936]	[-2.16748]
D(LOG(RCASES(-3)))	-0.070917	-0.239521
	(0.05763)	(0.07813)
	[-1.23053]	[-3.06581]
D(LOG(RCASES(-4)))	0.048542	-0.032712
	(0.03522)	(0.04774)
	[1.37835]	[-0.68518]
C	-0.017900	-0.013328
	(0.01202)	(0.01630)
	[-1.48881]	[-0.81773]
Exogenous Factors		
SECONDWAVE	0.013173	0.011019
	(0.00912)	(0.01237)
	[1.44369]	[0.89078]
UNLOCK	0.013653	0.014964
	(0.00868)	(0.01177)
	[1.57305]	[1.27177]
R-squared	0.972420	0.960162
Adj. R-squared	0.968042	0.953839
Sum sq. resids	0.044522	0.081821
S.E. equation	0.026584	0.036038
F-statistic	222.1265	151.8421
Log likelihood	169.3847	146.8683
Akaike AIC	-4.280668	-3.672117
Schwarz SC	-3.938171	-3.329621

Source: Estimated by the authors. Results are EVIEWS 10 generated.

Notes: Exogenous factors are not dropped in the Block exogeneity Wald Tests in Table A2.6. The VAR has 4 period lags in line with VAR optimum lag length criteria presented in Table A2.5. The results are for India.

Appendix 3: District level Urban-Rural causality of infections during first wave for the state of Maharashtra

Unit Root tests of district level Urban and Rural Cases per million for Maharashtra

Table A3.1 Augmented Dickey-Fuller test of LOG(UCASES) for Maharashtra during first wave			
Null hypothesis: D(LOG(UCASES)) has a unit root		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.159	0.031
Critical values	1% level	-2.597	
	5% level	-1.945	
	10% level	-1.614	

Source: Estimated by the authors on the basis of secondary data.

Notes: 1. *EVIEWS-10 generated MacKinnon (1996) one-sided p-values. 2. UCASES is the time series of total confirmed cases per million populations for all urban districts of Maharashtra taken together. 3. No trend or intercept included in model. 4. No. of lags taken is 3 (Automatic based on SIC; max=4) and included time points =47 (first wave only). 5. Breakpoint unit root tests are conducted but significant structural breaks in LOG(UCASES) are not found.

Table A3.2 Augmented Dickey-Fuller test of LOG(RCASES) for Maharashtra during first wave			
Null hypothesis: D(LOG(RCASES)) has a unit root		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-4.858	0.000
Critical values	1% level	-2.597	
	5% level	-1.945	
	10% level	-1.614	

Source: Estimated by the authors on the basis of secondary data.

Notes: 1. *MacKinnon (1996) one-sided p-values. 2. RCASES is the time series of total confirmed cases per million populations for all rural districts of Maharashtra taken together. 3. No trend or intercept included in model. 4. No. of lags taken is 2 (Automatic based on SIC; max=4) and included time points = 47 (first wave only). 5. Breakpoint unit root tests are conducted but significant structural breaks in LOG(RCASES) are not found.

Appendix 4: Variable Definitions and Data Sources

AGED – percentage of state level population aged 60 years and above taken from Census 2011, available at https://www.censusindia.gov.in/vital_statistics/srs_report/9chap%202%20-%202011.pdf.

AGRI – Percentage contribution of State Domestic Product from agriculture and allied activities, compiled from RBI Handbook of Statistics on Indian Economy available at https://www.rbi.org.in/scripts/AnnualPublications.aspx? (Table 8: Net State Value Added by Economic Activity at Constant Prices, Base: 2011-2012)

BPL - Percentage of population below poverty line at the state level (2011-12) based on Tendulkar Methodology. State level figures for combined poverty estimates were obtained from https://niti.gov.in/state-statistics (Data Source: Planning Commission).

CASES – confirmed cumulative total covid-19 Infections per million state populations as on 14/02/2021. (Source: https://www.covid19india.org/for India.

DENSITY - Population density per sq.km as per 2011 Census, compiled from https://www.census2011.co.in/density.phpfor India (Source: Census of India, 2011), and 2010

GINI - State Wise Gini Coefficient for Household Asset Scores, NFHS -4, 2015-16 (taken from Pandey and Gautam, 2020, Table 1, pp23).

IMR – Infant Mortality Rate(per 1000 live births) for 2016 obtained from the NitiAyog, Government of India, available at, https://niti.gov.in/content/infant-mortality-rate-imr-1000-live-births (Source: Sample Registration System).

PCNSDP – Per capita NSDP for 2018-19, at 2011-12 prices, compiled from RBI Handbook of Statistics on Indian Economy available at https://www.rbi.org.in/scripts/PublicationsView.aspx?id=19743. [Source: National Statistical Office (NSO)].

TESTS – Cumulative total Tests conducted per million at the state level as on 21/02/2021(Source: https://www.covid19india.org/).

TCR – Total Crime Rate (reported) compiled from 'Crime in India 2018', National Crime Records Bureau (Ministry of Home Affairs, Government of India), page 9, TABLE 1A.1 IPC Crimes (State/UT-wise) - 2016-2018 available at https://ncrb.gov.in/sites/default/files/Crime .

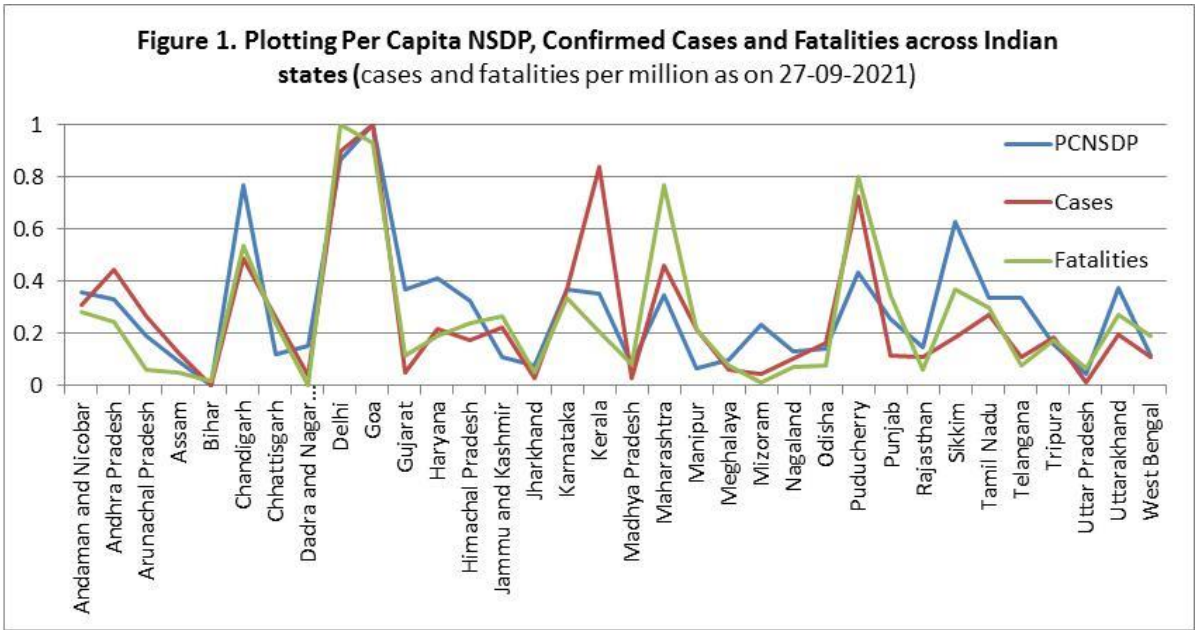
URBAN – urban population as a percentage of state population based on 2011 Census. For each state it is compiled from https://www.census2011.co.in/census/state/(Source: Census of India, 2011).

WFA – Work force in Agriculture and allied activities expressed as percentage of total work force at the state level. Source: Census of India 2011, Table T 00-009: Distribution of workers by category of workers (https://censusindia.gov.in/tables_published/a-series/a-series_links/t_00_009.aspx).

UCASES–sum of all the confirmed cases per million for the urbanized states/districts (above median level of urbanization) at each time point, compiled from https://www.covid19india.org/.

RCASES –sum of all the confirmed cases per million for the rural states/districts (below median level of urbanization) at each time point, compiled from https://www.covid19india.org/.

Appendix 5: Figures and Tables



Source: Plotted by the authors on the basis of secondary data. Covid statistics are drawn from covid19india.org. **Notes**: All variable are expressed in a 0 to 1 scale following the HDI-type attainment index formula.

Table 1. Ordinary correlation coefficients between all pairs of variables of interest									
Variables	DEATHS	CASES	TESTS	PCNSDP	URBAN	DENSITY	BPL	AGRI	IMR
DEATHS	1.0								
CASES	0.837 (0.000)	1.0							
TESTS	0.528 (0.001)	0.610 (0.000)	1.0						
PCNSDP	0.758 (0.000)	0.720 (0.000)	0.353 (0.041)	1.0					
URBAN	0.713 (0.000)	0.684 (0.000)	0.488 (0.003)	0.679 (0.000)	1.0				
DENSITY	0.528 (0.001)	0.468 (0.005)	0.728 (0.000)	0.371 (0.031)	0.713 (0.000)	1.0			
BPL	-0.364 (0.034)	-0.362 (0.035)	-0.316 (0.069)	-0.510 (0.002)	-0.348 (0.043)	-0.206 (0.241)	1.0		
AGRI	-0.775 (0.000)	-0.744 (0.000)	-0.541 (0.001)	-0.717 (0.000)	-0.688 (0.000)	-0.596 (0.000)	0.337 (0.051)	1.0	
IMR	-0.581 (0.000)	-0.585 (0.000)	-0.374 (0.029)	-0.569 (0.000)	-0.611 (0.000)	-0.336 (0.052)	0.541 (0.000)	0.569 (0.000)	1.0

Source: Computed by the authors on the basis of secondary data (www.covid19india.org) on 33 states and Union Territories of India. **Notes:** p-values are given in parentheses. Cumulative state-wise total deaths and cases per million state populations are for 27-09-2021.

Table 2. Log-linear regression of Deaths per million across the cross-section of 33 States and Union Territories of India					
Dependent Variable: Log(DEATHS)	Model 1	Model 2	Model 3	Model 4	Model 5
Explanatory Variables					
Constant	7.797** (0.009)	7.133* (2.593)	6.703* (2.375)	-3.524 (-1.088)	-7.328* (-2.693)
LOG(PCNSDP)				0.785** (3.168)	1.088** (5.337)
LOG(URBAN)	-0.235 (-0.386)	-0.064 (-0.103)	-0.102 (-0.163)		
LOG(AGRI)	-0.834* (2.475)	-0.834** (-2.734)	-0.820 (-2.591)	-0.473* (-2.654)	
LOG(BPL)	-0.228 (-1.519)				
LOG(WFA)					-0.216* (-2.011)
LOG(URBAN)*LOG(BPL)		-0.059 (-1.464)			
R square	0.40	0.39	0.37	0.45	0.40
Adjusted R square	0.34	0.33	0.33	0.41	0.36
F statistics	6.58**	6.47**	9.23**	12.63**	10.19**
Durbin-Watson	2.03	2.01	1.91	2.06	2.08

Source: Estimated by the authors on the basis of secondary data.

Notes: 1. Numbers in the parentheses are t ratios where HAC adjusted standard errors are used in all cases. ** means significant at 1% level and ** means significant at 5% level. 2. The total state level covid-19 deaths per million population are for 21–02 –2021.

Table 3. The Log-linear Panel Regression of Cumulative Weekly Total Deaths During first wave on State-level factors				
Depended variable: log(DEATHS) Explanatory Variables	Model 1	Model 2	Model 3	Model 4
CONSTANT	0.058 (0.319)	0.391** (3.598)	0.301 (1.230)	-0.755* (-2.096)
LOG(DEATHS(-1))	0.948** (91.908)	0.938** (114.71)	0.947** (62.099)	0.934** (60.755)
LOG(PCNSDP)				0.067* (2.337)
LOG(URBAN)	0.029 (0.659)		0.033 (0.591)	
LOG(DENSITY)	0.027 (1.636)			0.036* (2.320)
LOG(BPL)	-0.005 (-0.187)	-0.019 (-0.507)		
LOG(AGRI)		-0.043 (-1.955)	-0.030 (-1.007)	
UNLOCK		0.122 (1.929)		0.149 (1.201)
LOG(URBAN)*LOG(BPL)			-0.004 (0.622)	
R-squared	0.97	0.97	0.97	0.97
Adjusted R-squared	0.97	0.97	0.97	0.97
F-statistic	11115.20**	11130.69**	11102.29**	11174.53**
Durbin-Watson	1.93	1.95	1.91	1.94

Source: Estimated by the authors on the basis of secondary data.

Notes: 1. Numbers in the parentheses are t ratios where White's diagonally corrected standard errors are used throughout. ** means significant at 1% level and * means significant at 5% level. 2. These pooled estimates use White's diagonally corrected standard errors throughout. 3. Number of states and UTs = 33, number of weeks = 47; panel includes 1551 pooled observations for the first wave only.

Table 4. The Log-linear Panel Regression of weekly new Covid Deaths on State-level factors during the first wave in India					
Explanatory Variables Depended variable: D(log(DEATHS))	Model 1	Model 2	Model 3	Model 4	Model 5
CONSTANT	-1.330** (-2.855)	-1.602** (-5.326)	-0.298 (-0.580)	-3.915** (-4.690)	-0.455** (-2.669)
D(LOG(DEATHS(-1))	0.824** (53.938)	0.823** (54.601)	0.831** (32.025)	0.812** (29.236)	0.823** (31.511)
LOG(PCNSDP)				0.193** (3.353)	
LOG(URBAN)	0.122 (1.346)		0.118 (1.131)		
LOG(DENSITY)		0.085** (3.375)		0.077** (2.810)	
LOG(AGRI)	-0.070 (-1.425)		-0.067 (-1.290)		-0.096* (-2.531)
LOG(BPL)	-0.019 (-0.411)	-0.014 (-0.309)	-0.018 (-0.371)		
LOG(IMR)					-0.122 (-1.231)
UNLOCK					0.366* 2.112
TIME	0.107** (7.704)	0.108** (7.725)		0.115** (5.036)	0.085** (3.648)
TIME-SQUARED	-0.002** (-8.195)	-0.002** (-8.216)		-0.002** (-5.462)	-0.002** (-5.731)
R-squared	0.85	0.85	0.86	0.85	0.85
Adjusted R-squared	0.85	0.85	0.86	0.85	0.85
F-statistic	1452.97**	1743.78**	184.44**	1758.41	1459.33
Durbin-Watson	2.18	2.18	2.19	2.17	2.18

Source: Estimated by the authors on the basis of secondary data.

Notes: 1. Numbers in the parentheses are t ratios which use White's diagonally corrected standard errors throughout. ** means significant at 1% level and * means significant at 5% level. 2. Number of states and UTs = 33, number of weeks = 47; panel includes 1551 pooled observations. 4. D(DEATHS) implies the first difference of cumulative weekly total deaths, which is tantamount to weekly death count per million or weekly new deaths per million. 5. Model 3 is estimated under fixed time effects (periods fixed) supressing time and time-squared. 6. White's diagonally adjusted standard errors are used throughout.

Table 5. The Log-linear Panel Regression of Cumulative Weekly total Deaths on State-level factors during Second wave					
Depended variable: log(DEATHS) Explanatory Variables	Model 1	Model 2	Model 3	Model 4	Model 5
CONSTANT	0.027 (0.956)	0.065* (0.956)	0.055 (1.618)	-0.069 (-1.165)	0.021 (0.132)
LOG(DEATHS(-1))	0.986** (272.696)	0.984** (211.928)	0.983** (235.074)	0.985** (233.940)	0.978** (189.735)
LOG(PCNSDP)				0.011* (1.920)	0.022 (1.822)
LOG(URBAN)	0.014** (2.632)		0.002 (0.320)		
LOG(DENSITY)	-0.005** (-2.564)			-0.003 (-1.584)	
LOG(BPL)	-0.001 (-0.363)	-4.290 (-0.010)			
LOG(AGRI)		-0.006 (-1.703)	-0.005 (-1.316)		
LOG(IMR)					-0.031** (-2.795)
TIME	0.013** (4.779)	0.013** (10.198)	0.013** (13.150)	0.013** (4.815)	0.013** (4.473)
TIME-SQUARED	-0.003** (-4.626)	-0.0004** (-13.149)	-0.0004** (-14.147)	-0.0004** (-4.644)	-0.0004** (-4.403)
LOG(URBAN)*LOG(BPL)			5.110 (0.446)		
R-squared	0.994	0.994	0.994	0.994	0.994
Adjusted R-squared	0.994	0.994	0.994	0.994	0.994
F-statistic	30135.95**	36016.94**	29988.41**	36146.61**	30401.15**
Durbin-Watson	1.895	1.890	1.890	1.899	1.896

Source: Estimated by the authors on the basis of secondary data.

Notes: 1. Numbers in the parentheses are t ratios where White's diagonally corrected standard errors are used throughout. 2. ** means significant at 1% level and * means significant at 5% level. 3. Number of states and UTs = 33, number of weeks covering second wave = 32; panel includes 1056 pooled observations for second wave only.

Table 6. The Log-linear Panel Regression of Cumulative Weekly total Deaths on State-level factors during the first and second waves combined				
Depended variable: LOG(DEATHS) Explanatory Variables	Model 1	Model 2	Model 3	Model 4
CONSTANT	0.223** (3.225)	0.472** (8.086)	0.315** (3.184)	-0.141 (-0.504)
LOG(DEATHS(-1))	0.954** (124.683)	0.944** (145.244)	0.944** (145.765)	0.960** (85.644)
LOG(PCNSDP)				0.053** (2.556)
LOG(URBAN)	0.034* (1.942)		0.037 (1.673)	
LOG(DENSITY)	0.009 (1.124)			
LOG(BPL)	-0.012 (-0.771)	-0.119 (-1.201)		
LOG(AGRI)		-0.036** (-2.865)	-0.027 (-1.875)	
LOG(IMR)				0.0003 (0.016)
SECONDWAVE	0.549** (6.074)			
TIME*SECONDWAVE		0.001** (4.017)	0.001** (3.955)	0.002** (4.708)
TIME-SQUARED	-0.00008** (-4.556)	-0.00002** (-2.256)	-0.00002* (-2.195)	-0.00006* (-1.922)
LOG(URBAN)*LOG(BPL)			-0.004 (-0.995)	
R-squared	0.986	0.986	0.986	0.987
Adjusted R-squared	0.986	0.986	0.986	0.987
F-statistic	26236.97**	36539.66**	30448.75**	33479.23**
Durbin-Watson	1.899	1.881	1.882	1.882

Source: Estimated by the authors on the basis of secondary data.

Notes: 1. Numbers in the parentheses are t ratios where White's diagonally corrected standard errors are used throughout. 2. ** means significant at 1% level and * means significant at 5% level. 3. Number of states and UTs = 33, number of weeks covering first and second waves = 47+32 = 79; panel includes 2607 pooled observations for both waves of Covid-19 infections in India.

Table 7. Impact of Old Age population (%)on Deaths per million for 31 States and Union Territories of India		
Dependent Variable: LOG(DEATHS)	Model 1	Model 2
Explanatory Variables		
Constant	22.458 (1.545)	29.985* (2.306)
LOG(AGED)	1.954** (2.989)	1.676* (2.198)
LOG(AGRI)	-0.537** (-6.435)	-0.495** (-4.470)
LOG(LEB)	-5.412 (-1.467)	-6.684* (-2.012)
LOG(GINI)	-1.634** (-2.969)	-1.460** (-2.761)
LOG(CASES/TESTS)		0.500 (1.773)
R square	0.658	0.698
Adjusted R square	0.611	0.644
F statistics	13.962**	12.965**
Durbin-Watson	2.305	2.306

Source: Estimated by the authors on the basis of secondary data.

Notes: 1. Numbers in the parentheses are t ratios where HAC adjusted standard errors are used in all cases. ** means significant at 1% level and * means significant at 5% level. 2. The total state level covid-19 deaths per million population are for 27–09 –2021. 4. The State-wise GINI for 31 states and UTs is compiled from Pandey and Gautam (2020).

Table 8. Zivot-Andrews Breakpoint Unit Root Test for LOG(UCASES)			
Null hypothesis: LOG(UCASES) has a unit root with structural breaks in both intercept and trend Chosen Lag length: 10 (Max. lags 11) Included observations 79 Chosen break point is time point 53 (23-03-2021)			
		t-Stat	Prob.
Augmented Dickey-Fuller test statistic		-5.715	0.000
Test critical values	1% level	-5.57	
	5% level	-5.08	
	10% level	-4.82	

Source: Estimated by authors on the basis of secondary data.

Note: The 53rd time point starting 24-03-2020 is 23rd March 2021 which is the break date. LOG(UCASES) is in level and is not differenced.

Table 9. Zivot-Andrews Breakpoint Unit Root Test for LOG(RCASES)			
Null hypothesis: LOG(RCASES) has a unit root with structural breaks in both intercept and trend Chosen Lag length: 9 (Max. lags 11) Included observations 79 Chosen break point is time point 55 (06-04-2021)			
		t-Stat	Prob.*
Augmented Dickey-Fuller test statistic		-5.481	0.005
Test critical values	1% level	-5.57	
	5% level	-5.08	
	10% level	-4.82	

Source: Estimated by authors on the basis of secondary data.

Note: The 55th time point starting 24-03-2020 is 6th April 2021 which is the break date. LOG(RCASES) is in level and is not differenced.

Table 10. VAR Granger Causality/Block Exogeneity Wald Tests			
Included observations: 74; sample size is 79			
Dependent variable: D(LOG(UCASES))			
Excluded	Chi-square	df	Prob.
D(LOG(RCASES))	6.797882	4	0.1470
All	6.797882	4	0.1470
Inference	Lagged D(LOG(RCASES)) do not explain D(LOG(UCASES))		
Dependent variable: D(LOG(RCASES))			
Excluded	Chi-square	df	Prob.
D(LOG(UCASES))	44.30754	4	0.000
All	44.30754	4	0.000
Inference	Lagged D(LOG(UCASES)) explain D(LOG(RCASES))		

Source: Estimated by the authors on the basis of secondary data. **Notes**: The first null hypothesis is that the lagged D(log(RCASES)) do not explain D(log(UCASES)). The second null hypothesis is that D(log(UCASES)) do not explain D(log(RCASES)). The exogenous factors SECONDWAVE and UNLOCK are not dropped during the block exogeneity Wald tests.

Table11. Post VAR Granger Causality/Block Exogeneity Wald Tests on district level data for Maharashtra during the first wave of Covid-19 infections in India			
Included observations: 41; sample size is 47			
Dependent variable: D(LOG(UCASES))			
Excluded	Chi-square	df	Prob.
D(LOG(RCASES))	8.156	4	0.086
All	8.156	4	0.086
Inference	Lagged D(LOG(RCASES)) do not explain D(LOG(UCASES))		
Dependent variable: D(LOG(RCASES))			
Excluded	Chi-square	df	Prob.
D(LOG(UCASES))	30.115	4	0.000
All	30.115	4	0.000
Inference	Lagged D(LOG(UCASES)) explain D(LOG(RCASES))		

Source: Estimated by the authors on the basis of secondary data. **Notes**: The first null hypothesis is that the lagged D(log(RCASES)) do not explain D(log(UCASES)). The second null hypothesis is that lagged D(log(UCASES)) do not explain D(log(RCASES)). The exogenous factor "UNLOCK" is not dropped during the block exogeneity Wald tests. 4 lagged regressors are taken where lag order selection follows the same method as in the all-India Urban to Rural causality tests.

Table 12. Zivot-Andrews Breakpoint Unit Root Test for LOG(UCASES) for Second Wave of covid-19 Infections in Maharashtra			
Null hypothesis: D[LOG(UCASES)] has a unit root with structural breaks in both intercept and trend Chosen Lag length: 10 (Max. lags 11) Included observations:33 (22-02-2021 till 27-09-2021) Chosen break point is time point 07 (29-03-2021)			
		t-Stat	Prob.*
Augmented Dickey-Fuller test statistic		-5.754	0.000
Test critical values	1% level	-5.348	
	5% level	-4.859	
	10% level	-4.607	

Source: Estimated by authors on the basis of secondary data.

Note: The 7thtime point starting 22-02-2021 is 29thMarch 2021, which is the break date for LOG(UCASES) series for urban districts of Maharashtra. LOG(UCASES) is in first difference.

Table 13. Zivot-Andrews Breakpoint Unit Root Test for LOG(RCASES) for Second Wave of covid-19 Infections in Maharashtra			
Null hypothesis: D[LOG(RCASES)] has a unit root with structural breaks in intercept only Chosen Lag length: 10 (Max. lags 11) Included observations:33 (22-02-2021 till 27-09-2021) Chosen break point is time point 12 (03-05-2021)			
		t-Stat	Prob.*
Augmented Dickey-Fuller test statistic		-4.952	0.000
Test critical values	1% level	-4.949	
	5% level	-4.443	
	10% level	-4.194	

Source: Estimated by authors on the basis of secondary data.

Note: The12thtime point starting 22-02-2021 is 3rdMay 2021, which is the break date for LOG(RCASES) series for rural districts of Maharashtra. LOG(RCASES) is in first difference. For rural cases intercept break only provides better fit.

Table14. Post VAR Granger Causality/Block Exogeneity Wald Tests on district level data for Maharashtra during the second wave of Covid-19 infections in India			
Included observations: 28; sample size is 32			
Dependent variable: D(LOG(UCASES))			
Excluded	Chi-square	df	Prob.
D(LOG(RCASES))	4.849	4	0.183
All	4.849	4	0.183
Inference	Lagged D(LOG(RCASES)) do not explain D(LOG(UCASES))		
Dependent variable: D(LOG(RCASES))			
Excluded	Chi-square	df	Prob.
D(LOG(UCASES))	15.222	4	0.002
All	15.222	4	0.002
Inference	Lagged D(LOG(UCASES)) explain D(LOG(RCASES))		

Source: Estimated by the authors on the basis of secondary data. **Notes**: The first null hypothesis is that the lagged D(log(RCASES)) do not explain D(log(UCASES)). The second null hypothesis is that lagged D(log(UCASES)) do not explain D(log(RCASES)). The exogenous dummy variable "UNLOCK" is not dropped during the block exogeneity Wald tests. 3 lagged regressors are taken where lag order selection follows the same method as the all-India Urban to Rural causality tests.

References

Abedi, V., Olulana, O., Avula, V., Chaudhary, D., Khan, A., Shahjouei, S., ...&Zand, R. (2020). Racial, economic, and health inequality and COVID-19 infection in the United States. *Journal of racial and ethnic health disparities*, 1-11.

Basu, P. & Mazumder, R. (2020). Where are the Covid victims in India and where should vaccines go? *The Economic Times*, Dec 29 (7:41 IST).

Basu, P., & Mazumder, R. (2021). Regional disparity of covid-19 infections: an investigation using state-level Indian data. *Indian Economic Review*, 1-18.

Basu, P., Bell, C., & Edwards, T. H. (2021). COVID Social Distancing and the Poor: An Analysis of the Evidence for England, B.E Journal of Economics, https://doi.org/10.1515/bejm-2020-0250.

Brotherhood, L., Cavalcanti, T., Da Mata, D., & Santos, C. (2020).Slums and pandemics. *CEPR Discussion Paper No. DP15131*

Brown, C. S., & Ravallion, M. (2020). Inequality and the coronavirus: Socioeconomic covariates of behavioural responses and viral outcomes across US counties (No. w27549). *National Bureau of Economic Research*.

Chelliah, R. J., Rao, M. G., & Sen, T. K. (1992).Issues before Tenth Finance Commission. *Economic and Political Weekly*, pp. 2539-2550.

Davies, J.B. (2021). Economic Inequality and Covid-19 death rates in the first wave: A cross country analysis. *Covid Economics*. (73), 23 March.

Deaton, A. (2013). *The great escape: health, wealth, and the origins of inequality*. Princeton University Press.

Finch, W. H., & Hernández Finch, M. E. (2020). Poverty and Covid-19: rates of incidence and deaths in the United States during the first 10 weeks of the pandemic. *Frontiers in Sociology*, (*5*), 47.

Jalan, J., & Sen, A. (2020a). Understand the method in Covid-19's madness. India doesn't need complete lockdown. *The Print.*10 April, 2020 12:34 pm IST.

Jalan, J., & Sen, A. (2020b). Containing a pandemic with public actions and public trust: the Kerala story. *Indian Economic Review*, *55*(1), 105-124.

Jalan, J., & Sen, A. (2020c). Spread of Covid-19 in India: Across Space and Over Time. https://indiacovidspread.weebly.com.

Kermack, W. O., &McKendrick, A. G. (1927).A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of London.Series A, Containing papers of a mathematical and physical character*, *115*(772), 700-721.

Kletzer, K., & Singh, N. (1997).The political economy of Indian fiscal federalism. *Public Finance: Policy Issues for India*, pp. 259-298.

Mandi, J., Chakrabarty, M., & Mukherjee, S. (2020).How to ease Covid-19 lockdown?Forward guidance using a multi-dimensional vulnerability index.*Ideas for India* May 21, 2020.

Mishra, U.S. & Joe, W. (2020).Household Assets and Wealth Quintiles, India 2006–16Insights on Economic Inequalities.*Economic and Political Weekly*, Vol. 55 (6).

Pandey, A., & Gautam, R. (2020). Regional inequality in India: A state level analysis. *Journal of Community Positive Practices*, (4), 56-85.

Rao, M. G., & Singh, N. (2006). *The political economy of federalism in India*.Oxford University Press.

Ray, D., & Subramanian, S. (2020). India's lockdown: An Interim Report. *Indian Economic Review*, *55*(1), 31-79.

Stafford, K., Hoyer, M., & Morrison, A. (2020). Racial Toll of Virus grows even starker as more data emerges. *AP News*, April 19 (https://apnews.com/article/8a3430dd37e7c44290c7621f5af96d6b).

Strachan, D. P. (1989). Hay fever, hygiene, and household size. *BMJ: British Medical Journal*, *299*(6710), 1259.