A CLUSTERING ANALYSIS OF IMPACTS OF GOVERNMENT POLICY AND LITERACY RATES ON COVID-19 CASES IN INDIA

Thorid Wagenblast, Marya El Malki, Jochem Feenstra, Emily Ryan

ABSTRACT

At the start of November 2020, there were over eight million confirmed cases of COVID-19 in India. In addition to a nationwide lockdown, the 28 states that make up India implemented their own policies to stop the virus. Research on connections between socio-economic factors and COVID-19 has focused on age, economic factors, or health infrastructure. There has not been significant literature on the relation between the literacy rate and COVID-19 cases. In this paper, we focus on how the literacy rate impacts COVID-19 in Indian states and union territories. The reason for this is the importance of communication during a crisis. Illiteracy can form a barrier to clear communication on measures taken by state governments, which could lead to an increase of cases. We will be comparing data on literacy and COVID, in addition we will consider the role of internal migration as well, given the high illiteracy rate among this demographic, and finally we will analyze the policies taken by different states. We used data from the SHRUG India project, the 2011 India Census, and the COVID-19 India API for this research. Through clustering of the data, we found that COVID cases are generally higher in states with high literacy rates. Although contrary to our expectations, these findings are not solid proof, considering the uncertainty of how well represented COVID-patients are among the illiterate. This paper is best considered as a set-up for future policy analysis to improve communication and serves to put focus on the issue of illiteracy in this pandemic as well.

INTRODUCTION

GAPS IN THE LITERATURE

- Although it had been predicted by some (Obama White House, 2014; Norman et al, 2020) the SARS-CoV-2 pandemic took most countries in the world by storm (Adhamon, 2020). In terms of registered cases and deaths, India was one of the hardest hit countries by this pandemic. At the start of November, India had registered over 8 million cases and over 120.000 deaths (COVID-19 Dashboard).
- The first established case of COVID-19, the disease caused by the SARS-CoV-2 virus, in India was on 30 January. On 25 March the national government declared a lockdown, which was relaxed on 8 June (De, 2020). In the meantime states took their own measures and implemented their own strategies. India consists of 28 states and 8 union territories, which can be found in Appendix D. Different states struggled with different problems, for example

the mobility of migrant workers back to their home state was an issue of concern (Mondkar, 2020), but the effectiveness of communication as well. Several strategies were based on spreading health information through WhatsApp (WHO, 2 July 2020), which relies on the literacy of targeted communities.

- It is unknown how effective these strategies have been in regards to case numbers and deaths, and how much of a role population literacy plays in this. Search engine use on "Literacy" and "COVID" keywords has not yielded clear, peer reviewed papers on this subject.

RESEARCH QUESTION

- We will be looking at data concerning socio-economic factors, such as literacy, and COVID-19 data (up to October 25, 2020), at a state level in India, to try and establish a pattern. We will also analyze policies in several states and try and determine their efficiency in relation to the literacy rate of the state. In addition, we will consider the role of interstate labor migration in this, since illiteracy is high among migrants (NSS, 2010, Pg. 45) and state policies seemed to focus on the mobility and living conditions of migrants (WHO, 2 July 2020).
- The question we will focus on is: What role does literacy rate play in the impact of COVID-19 in India?
- To do this analysis we will be using COVID data from Covid-19 India API and data from the 2011 Census (COI, 2020). All factors can be found in Appendix A. We will try to establish a connection between literacy and COVID data and examine the policies of selected states to get a better insight into what policies certain states have implemented. This allows for a more complete picture and can give context to results based on data.

IMPORTANCE OF FINDINGS

- After clustering the data, we found that COVID cases are generally higher in states with high literacy rates, which is contrary to our assumption. We believe this is due to the higher rate of COVID cases in states with a higher GDP (Kumar et al, 2020), which were hit earliest in the pandemic and are assumed to have better testing capacities, leading to a warped source of data.
- Based on our established results, there is a positive connection between literacy and COVID cases, however this result should not be taken as a definitive conclusion on the relation between these factors. A crucial caveat when considering this result is the uncertainty of data on which it was based. Considering the limited testing, it is uncertain how well represented the illiterate are in the general case numbers, but also, how representative state data is compared to other states, given that some states have reduced testing (Majid, 2020). Unfortunately this means that no meaningful conclusion can be based on our results. It can be useful to future research however, in that it can provide a framework for researchers to use when the quality of the data has improved. In addition, this paper puts the spotlight on the role of literacy and communication in the pandemic, factors that so far have not been well represented. We believe it is important that in a crisis situation, policy responses should be inclusive of illiterate members of society.

Related Work

- Measures that counter COVID-19 are dependent on communication and communication is dependent on people's ability to consume information. Illiteracy limits people's means to absorb information. There has been limited available research into how COVID-19 and literacy are related. This paper will address the relation between those factors.
- India has a large population (Census 2011) and it is important that the pandemic is under control in India to be able to keep it under control in the rest of the world (Abraham, 2020). To achieve this, Indian states have implemented different policies (WHO, 2 July 2020; Bhagat et al, 2020). An essential part of policy is communication (Reynolds, 2014), at the same time part of the communication from the government was through "bureaucratically worded government orders" (Abraham, 2020), which we assume can cause a gap between the message and the illiterate part of the population. The spread of this virus can be reduced through simple measures (CDC, 2020), but it does require people to be aware of these measures. We want to check whether literacy seems to influence the effectiveness of policy communication, because if this is the case, governments should take this into account to improve the effectiveness of their policy and to protect the illiterate part of their population.
- Relations between COVID-19 and other socio-economic factors have been established in previous research papers. These papers show positive relations between COVID-19 and GDP (Kumar et al, 2020; Chaudhry), age distribution (Laxminarayan et al. 2020; Chaudhry et al. 2020), and urbanization (Kumar et al, 2020; Sarkar & Choucan, 2020). The papers that included literacy did not establish a positive connection between literacy and COVID-19 (Kumar et al, 2020; Sarkar & Choucan, 2020). What we know is that communication is important in a crisis (Reynolds, 2014) and we assume that a higher literacy rate helps with the communication. We also know that illiteracy is relatively high among internal migrants in India (NSS, 2010, Pg. 45), which makes migration an interesting factor to consider as well.
- There has been limited research on the relation between COVID-19 and the literacy rate in states in India and no clear connection between these factors have been established. In addition, published work is based on older COVID-19 data, as these numbers develop every day, but we do not know whether this influences the role of literacy or not. Finally, we do not know how well represented the illiterate are in the current data.
- This article will focus on the question: What role does literacy rate play in the impact of COVID-19 in India?
- Given that there seems to be a lack in the literature that specifically focuses on the connection between literacy and COVID-19 in India, the results of our work can provide more insight in this matter. At the same time it is worthwhile to reexamine previously analyzed correlations, since COVID-19 data keeps developing with new cases every day. The virus continues to spread, possibly to new demographics, while an increase in testing can lead to the illiterate getting tested more often. This could change previously established connections. A more up-to-date examination is useful because of this.

EXPLORATORY DATA ANALYSIS

OVERVIEW OF ANALYSIS APPROACH:

- For this project, data was collected from three main sources, the Development Data Lab SHRUG India project (Asher, 2019), the 2011 India Census (Census Organisation of India, 2011), and the Covid-19 India API (COVID19 India - API, 2020). Shapefiles for mapping of the data were sourced from the Database of Global Administrative Areas (GADM, 2020). Data was collected and analysed at the level of individual states and unions, due to the level of completeness of available data sets. The data was cleaned in accordance with tidy data standards (Wickham, 2014) and exploratory data analysis performed to identify preferred variables for clustering. Data was scaled and clustering was performed using the Scikit-Learn k-means clustering method.

DESCRIPTION OF DATA COLLECTION AND CLEANING:

- The following data sets were pulled in .dta and .csv formats from the Development Data Lab COVID India page:
    - Migration
    - NFHS (health survey)
    - Demographic
    - Hospitals
- COVID Case data was taken in .csv format from the COVID-19 India API (the state_wise_daily file). This data was downloaded on 26 October 2020, and all reported cases and deaths were aggregated up until this date.
- Selected missing data points from the Development Data Lab data sets were replaced with data from the 2011 India Census.
- The project team chose to work with state-wise data to account for completeness and consistency in key demographic data across the years, as well as availability of appropriate shapefiles for district/sub-districts. At a district or sub-district level, it was found that inconsistencies in the number and names of districts had changed significantly in the years between the 2011 census and the collection of COVID case data in 2020. As such, for this study it was determined that state-level data was more appropriate. Much of the Development Data Lab data was split out to a district level, so data was aggregated up to the state level for these data sets. Using state-wise data was considered to be a suitable compromise, especially given that COVID policies were largely implemented at a state (rather than district) level, making comparison at this scale more appropriate.
- The availability of COVID data was the leading factor in how the data was cleaned and processed. Two states/unions were removed after initial analysis, as they had no reported COVID cases (Daman & Diu and Lakshadweep) and it was unknown whether this was due to lack of testing capability or lack of actual cases. Given the relatively low populations of these states, we were comfortable that removing them would not unreasonably affect the outcomes of the analysis.

- Following cleaning and preliminary EDA, it was also determined that the states of Telangana and Ladakh would need to be modified. These two states were not established at the time of the 2011 census, and were part of other states (Andhra Pradesh and Jammu & Kashmir respectively). To allow for 2011 census data to be used for these states, the data from the original states were replicated for Ladakh and Telangana (where there were counts of data, these were added, where there were rates, these were population-weighted averaged).
- There were, after cleaning, several missing data points for population and literacy rates. For these states, the project team referred back to the India Census website to replace individual values, and validate against the existing data.
- Following EDA and cleaning, it was also found that there was no hospital infrastructure data available for any of the union territories. This represented a significant proportion of the population, and hospital infrastructure was not a feature of our original research question, so it was decided to exclude hospital infrastructure data (e.g. number of beds, number of staff) from the clustering analysis. Future work may look at excluding the unions from the analysis and looking at clustering only for states, in order to include the hospital infrastructure data.
- The final list of variables used, including their original source, can be found in a table in Appendix A of this report. They were chosen because of completeness and considered potentially relevant factors in the spread of COVID-19 in India. For migration, the long-term migration data was used on a state level. As this data was from the 2011 census, short-term migration was considered more fluctuant while the long-term migration data was assumed to change slower and still be correct to a certain extent today.

PROCESS FOR ANALYSIS:

- The analysis of the data was primarily done using Scikit-Learn's k-means clustering algorithm. All data was scaled so that values were between 0 and 1 using Scikit-Learn's MinMaxScaler in order to prevent data sets with larger numbers (e.g. population) from having an outsized influence on the clustering. MinMaxScaler was considered appropriate for the data sets as there were no significant outliers in any of the variables. K-means clustering was chosen over other methods as it is a familiar, general-purpose clustering algorithm that works for smaller numbers of clusters (suitable given a relatively small n) (Scikit-Learn, n.d.). Although the number of clusters wasn't known, the project team ran the algorithm for between 2 and 10 clusters and visually inspected the choropleths and KDE plots for the clustering outputs, to determine the number of clusters that gave the most meaningful output for the research question.

KEY LIMITATIONS IN THE DATA:

- Relying on 2011 Census Data for demographic information, besides this data being almost ten years old, the data did not reflect the current split of states and districts across India. The project team recommends that the work be revisited following the 2021 census, wherein state and district data for COVID-19 cases should reliably match state and district data for census collection. This will allow for analysis of a much richer dataset.
- Only analysing 34 states and union territories meant that there was not significant variation in the data. This was apparent after producing pair plots and histograms of all of the selected variables, where no discernible correlations could be observed. It’s also

acknowledged that by taking this approach, this imposes some serious limitations on the ability to analyse for certain demographics, for instance, rural/urban divides. In the future, once data sets are more consistent, and once shapefiles are available for the SHRUG India project, this work should be revisited.

- The reliability of COVID-19 case data is highly connected to rates of testing. At the time of writing, COVID-19 testing rates in India were increasing, but still below global averages. Further to this, there are differences even between states in the rates and regimes for testing, which will influence the data. Unfortunately, the project team was unable to identify a reliable source of information for rates of testing/number of tests performed. Future studies could consider using population antibody/serology tests to get a more realistic picture of case numbers (Menon, 2020).

GRAPHICS AND EDA:

- In order to take a look at the spread of COVID-19 across states in India, population adjusted numbers for total confirmed cases and total deaths were computed (per 100,000 population). The top ten states with highest number of confirmed cases, and the top ten states with highest number of deaths are plotted in figures 1 and 2.

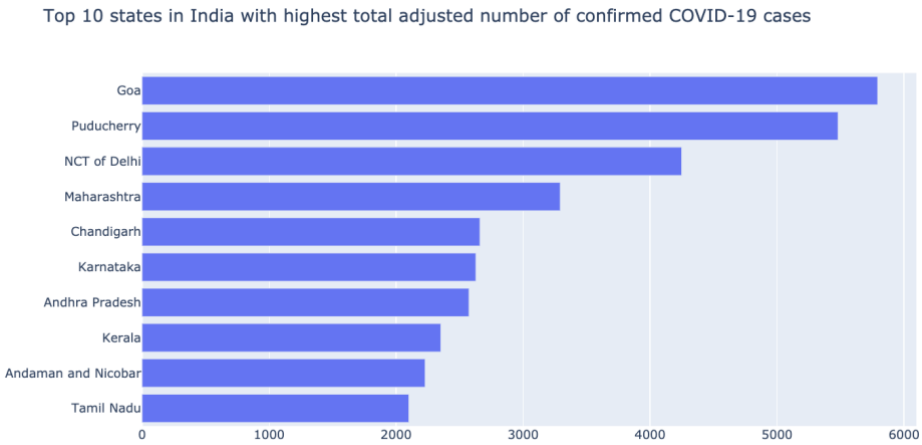

Figure 1: Bar plot representing the adjusted number of confirmed COVID-19 cases in the top 10 most affected states

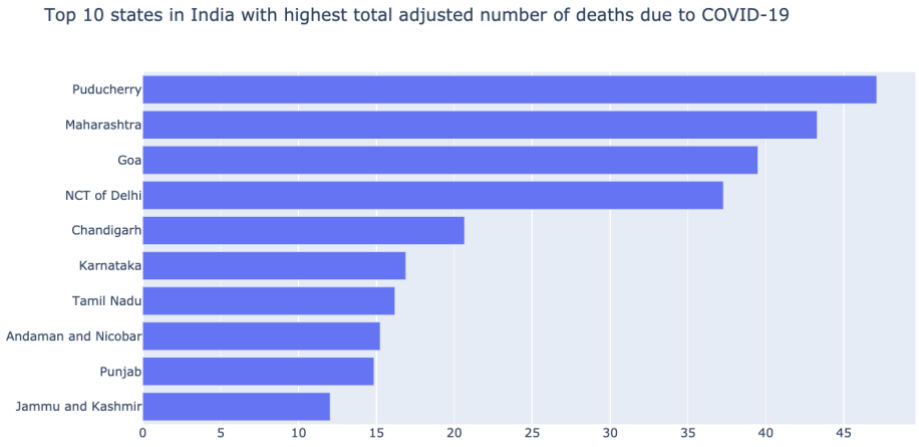

Figure 2: Bar plot representing the adjusted number of deaths due to COVID-19 in the top 10 most affected states

- To get a better sense of the distribution of the data across variables, box plots were constructed. As they will be of focal importance to this analysis, the plots of the adjusted number of COVID-19 cases and the percentage of literacy are presented below.

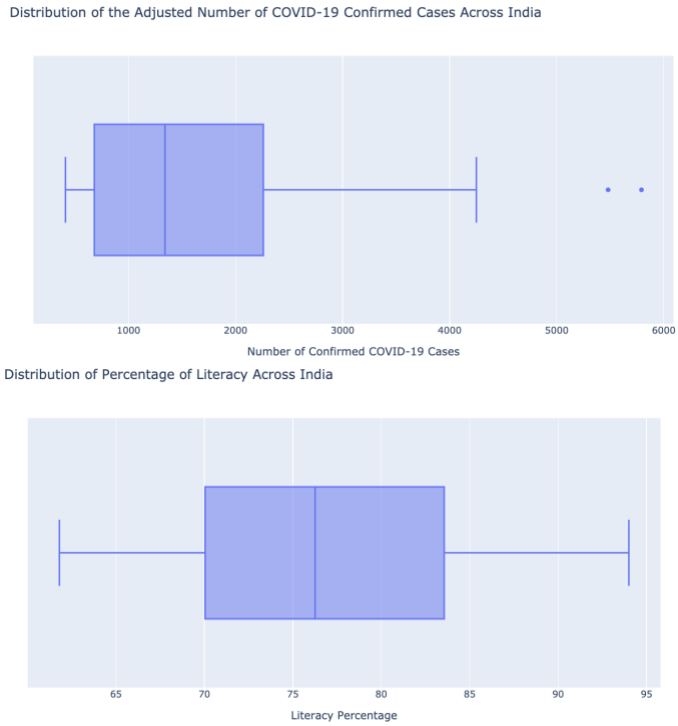

Figure 3: Box plots representing the distribution of the adjusted number of COVID-19 cases across states in India (a) and the distribution of the percentage of literacy across states in India (b)

- It is observed that the confirmed cases data possesses two outliers corresponding to Goa and Puducherry. The literacy data on the other hand shows no outliers, but highlights a large range of different literacy levels across the country, with a median at around 76%, ten percentage points below the global adult literacy rate (UNESCO, 2017). The highest rates can be found in Kerala, the lowest in Bihar (figure 4).
- The research question aims to investigate whether higher literacy rates are connected to lower case numbers. Figure 4 shows the spatial distribution of the three key variables concerning this question. The number of confirmed COVID cases goes from 410 per 100 000 in Bihar to over 5790 per 100 000 in Goa. The highest number of deaths is found in Puducherry with 47 of 100 000 deceased in relation with a COVID infection, while in Dadra and Nagar Haveli less than 1 person have reportedly died per 100 000 inhabitants in relation to COVID (see also figure 1 and figure 2).

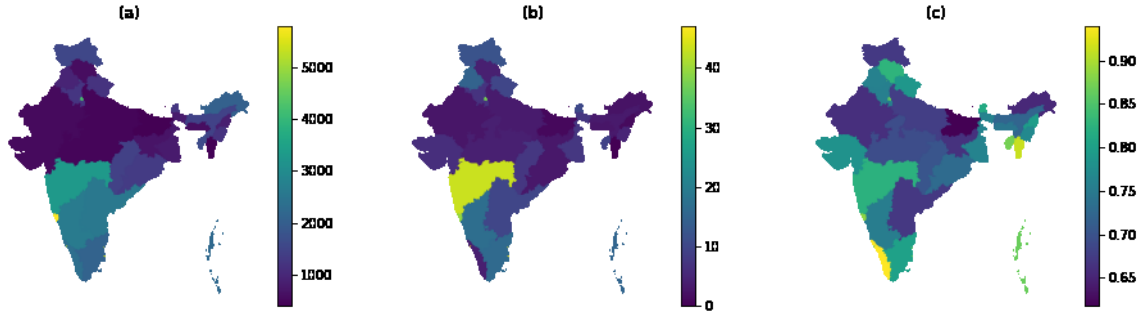

Figure 4: Choropleths for number of confirmed COVID-cases per 100 000 inhabitants (a), number of COVID-related deaths per 100 000 inhabitants (b), and literacy rate (c) .

ANALYSIS

- Our argument is that there is a connection between literacy rates (and COVID responses that included written communications) and COVID cases and fatalities.
- These arguments were checked using k-means clustering methods across a range of different variables. Exploratory data analysis of the variables considered interesting, and a preliminary scan of state policies showed that spatial clustering of the data wouldn't be appropriate, as in a number of cases, neighbouring states had quite different approaches to their COVID-19 policy for example Goa and Maharashtra. As such, clustering was performed over the variables independent of their location, but later mapped to visualise the clustering geographically (Figure 5). Taking into account the allocation of the variables to the clusters, the number of states in each cluster (Appendix C) and the variance of the variables mentioned above, it was focused on the clustering into three groups.

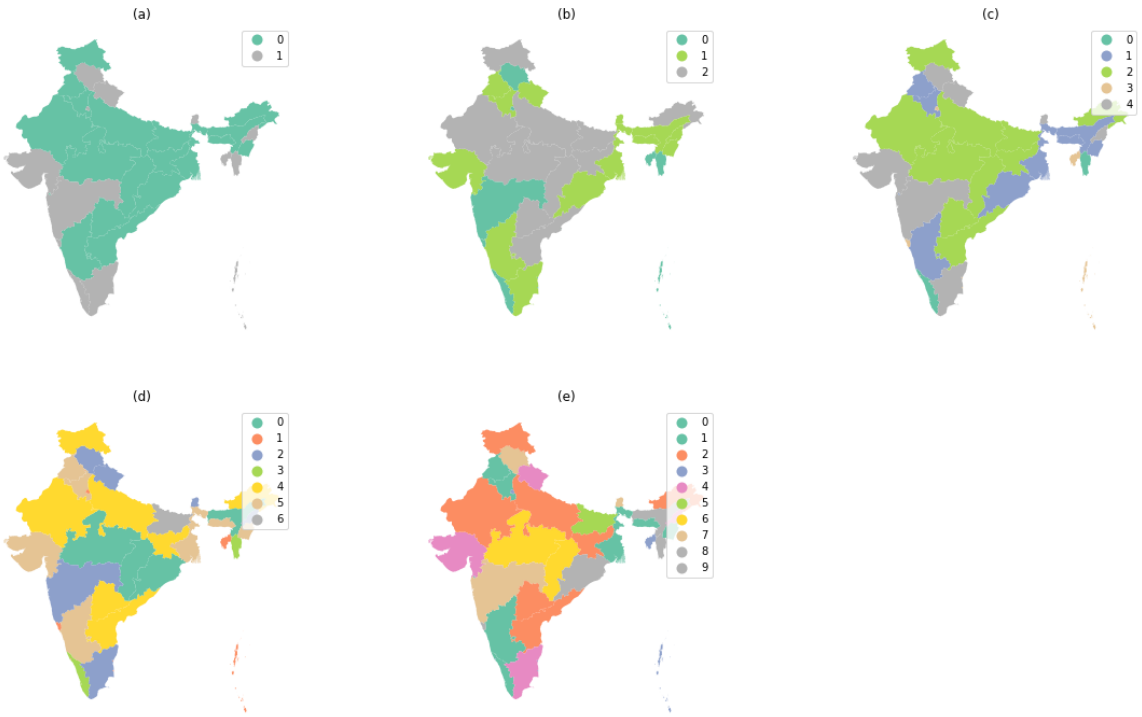

Figure 5: Clustering over the variables 3 to 13 and 15 to 18 (for further information see Appendix B) with (a) two clusters, (b) three clusters, (c) five clusters, (d) seven clusters, (e) ten clusters.

- Arguably, we might have considered the impacts of interstate migration on the spatial distribution of COVID cases, however we didn't have access to data on the migration of individuals in 2020.

SUPPORTING POINTS:

- Certain states were persistently placed in the same cluster, regardless of the number of clusters like the NCT of Delhi and Goa or Kerala and Mizoram as can be observed in Figure 5 (A map with the name of the states can be found in Appendix D). This shows that those states are similar in regard to the variables examined in this work.
- Consistent with earlier assumptions, states with higher rates of literacy had lower rates of long term migration (both in and out) according to the clustering.

SUPPORTING PATTERNS:

- As expected, literacy was a strong driver in the clustering of states (Figure 6). This is partly due to the stronger variance observed in the literacy variable used than what was observed in other variables (Figure 3).

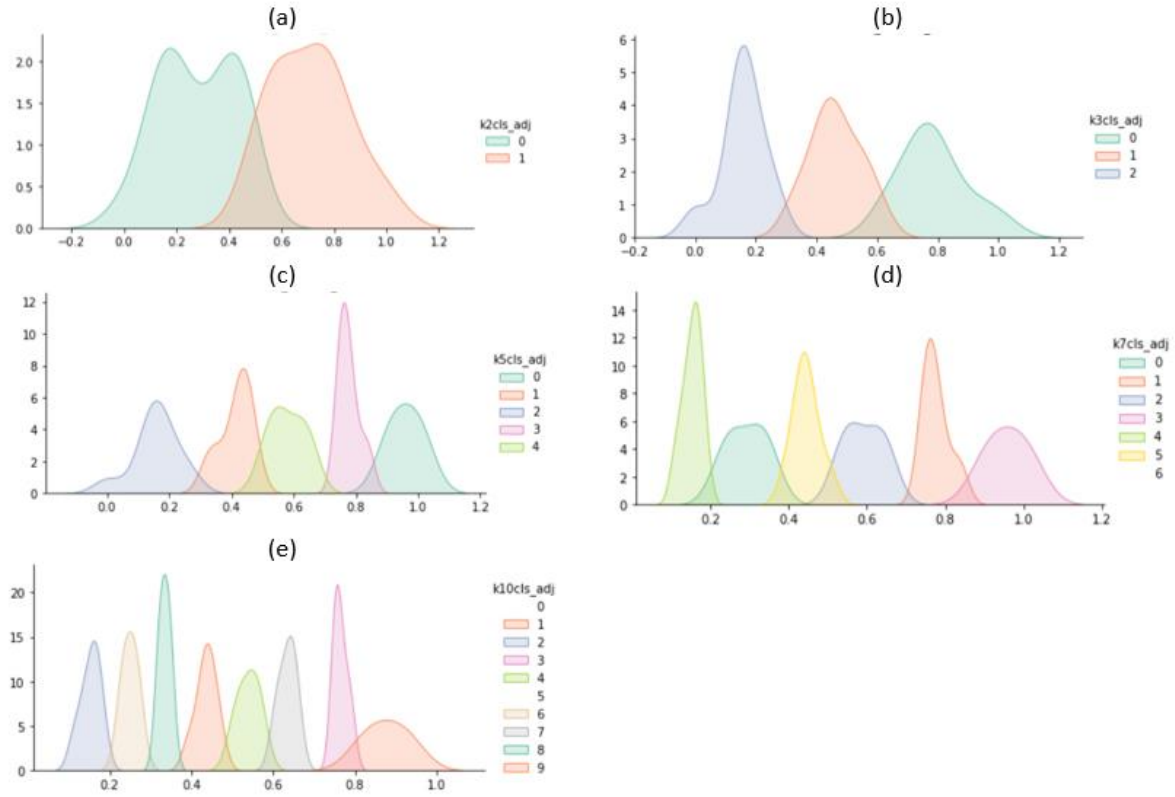

Figure 6: KDE-plot for literacy in the clustering for (a) two clusters, (b) three clusters, (c) five clusters, (d) seven clusters, (e) ten clusters.

- Another observation was made in connection to the literacy rate: as visualised in Figure 7, states in the cluster with the lowest literacy rates (group 2) are the ones with higher long-term migration rates both in and out of the state, whereas the cluster with higher literacy have lower migration rates in general. This backs previous findings that migrants in India are mostly from socioeconomically disadvantaged backgrounds, where education and literacy rates are lower (NSS, 2010) .

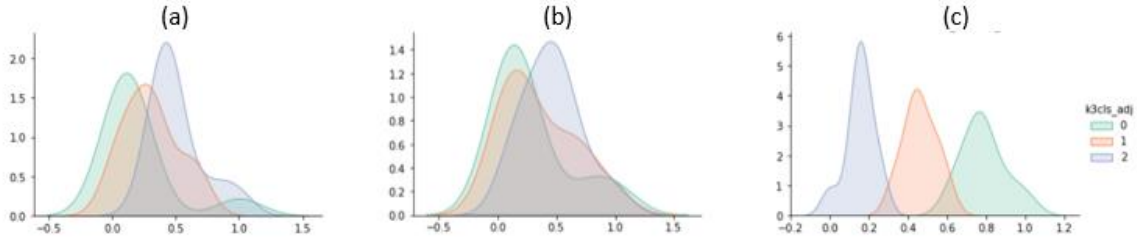

Figure 7: clustering with three clusters. KDE-plots for (a) long-term migration out of state, (b) long-term migration into state, (c) literacy rate.

- Contrary to the proposed argument, figure 7 shows that the states in a cluster with higher rates of literacy (group 0) also tended to have higher numbers of COVID cases, and a higher number of COVID-related deaths per 100 000 people, whereas states placed in group 1 and 2, so states with lower literacy rates are found to have lower numbers both among COVID-cases and -deaths in 100 000 people. The proposed argument that lower literacy is

correlated to higher rates of COVID-19 is hereby not confirmed, indeed the results seem to show the opposite.

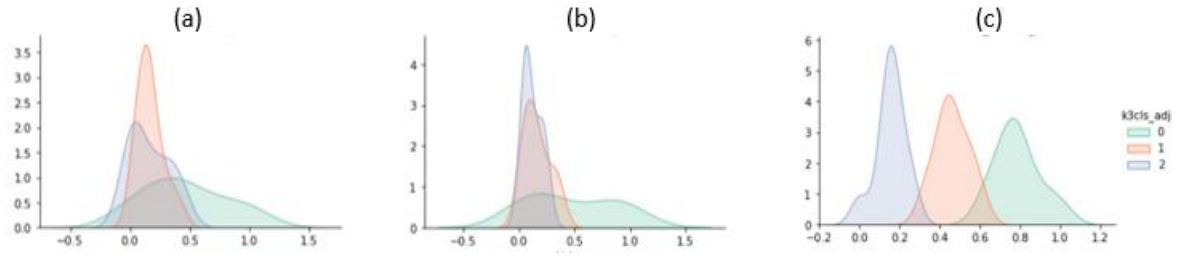

Figure 8: KDE-plot with three clusters for (a) covid cases per 100 000 people, (b) covid deaths per 100 000 people, (c) literacy rate.

- The main findings from the clusters were presented here. For more information on the clusters see Appendix C. The entire KDE-plots for clustering into three and five can also be found there.

INDIVIDUAL STATE POLICIES:

- These findings from the data demonstrate a connection between literacy rates and COVID-19 cases in India's states and union territories, however they are yet to be grounded in the policies of the individual states. In the following paragraphs, the impacts and reception of COVID-19 policies relevant to literacy and migrant workers is examined.
- **Kerala** has been commended for its swift, proactive and committed response to addressing the global pandemic (Menon et al., 2020). The Kerala government invested in emergency preparedness for a couple of years, as it faced floods in 2018, and the Nipah virus outbreak in 2019. This pushed the state to declare a state of emergency within three confirmed cases of COVID-19. Kerala's response to the pandemic included raising awareness through hand-washing campaigns, community engagement, as well as disseminating information through WhatsApp groups. In fact, despite low income per capita, Kerala possesses good development indicators (Menon et al., 2020), such as having the third highest percentage of water hand-washing access to households with 93%, and the highest rate of literacy across the country with 94%, as was observed in the data. It is also worth noting that 97% of Keralans speak the same language, facilitating intra-state communication (Kawoosa, 2018). In terms of social distancing measures and testing, the state followed WHO guidelines and conducted thorough excessive testing and contact tracing (Srivastav, 2020), as well as mandated a longer quarantine to isolate the virus. They also provided their residents with psychological support to help them throughout quarantine, especially for children, the elderly and migrant workers, and built shelters for the latter, right on time to control the movement of migrants for the nationwide lockdown (Learning from Kerala, 2020).
- **The NCT of Delhi**, the country's capital, currently stands as one of the top states affected by COVID-19 (OneIndia, 2020). Almost a month after Kerala, the chief minister declared COVID-19 an epidemic in the city. Delhi instituted a three-weeks national lockdown at the end of March, that keeps getting extended. This was followed by the launch of a COVID-19 helpline on WhatsApp to provide factual and up-to-date information to the public (Chakravarti,

2020). The city instituted a generic 14-day quarantine for citizens testing negative, as well as a requirement to move to a hospital in case of positive results. In June, due to a rapid surge in cases, the world's largest temporary COVID-19 hospital was built in Delhi to meet the needs of an influx of patients (ET Online, 2020). Finally, the city provided temporary relief centers and hunger relief centers to jobless migrant workers. However, those were not enough to meet demand, causing outrage in the region as hundreds of labourers were found to be living under a bridge with no access to proper sanitation (Pandey, 2020).

- **Maharashtra** (home to Mumbai and the Dharavi slum, as well as the tech hub city of Pune) has throughout the pandemic had the highest case numbers in the country. COVID policy in the state has been both celebrated and maligned for the handling of case numbers. Early reports on progress in the state suggest that the setup of screening and isolation centres for the Dharavi slum helped to minimise the impacts of COVID among Mumbai's most vulnerable residents (Pandya, 2020). Heavy police presence in Mumbai and Dharavi were credited with effectively controlling the virus (Mukhopadhyay, 2020). However, more recently, the state has seen a surge in cases, leading the state government to extend lockdown orders in the state (Banerjea, 2020). Early in the first lockdown, the state supported close to two million migrant labourers to return to their home state, and have since organised registration systems to allow these workers to return (ETGovernment, 2020). The impact of these policies on transmission dynamics in the state is uncertain.
- **Jammu and Kashmir** had a communications blackout for much of the pandemic, with even hospitals and healthcare staff unable to access high speed internet to get information on treatment best practice, while members of the public were deprived of basic information on measures like social distancing and handwashing (British Medical Journal, 2020). The communications blackout also meant that most of the union territory's residents could not access services like the Indian governments coronavirus tracking app, WhatsApp and Facebook messenger bots, and COVID-19 Twitter accounts - important sources of up to date information for residents of most other states. Since September, the state has seen a decline in cases, though this is thought to be due to a decline in testing and screening efforts (Majid, 2020). The actual number of cases in recent months is unknown.
- **Uttar Pradesh** had some of the most progressive policies regarding migrant labour. The state offered transport for both migrants living in UP to return to their home states, and for those living in other states to return to UP in the weeks following the lockdown measures. Further to this, payments of Rs 1000 were paid to migrant workers in quarantine. In May, a Migrant Commission was established to provide ongoing social security for these workers, although this has drawn criticism for being simply a band-aid solution to a larger problem (Singh, 2020). Further, the state has coordinated food and rationing systems to supply food to the unemployed and those in quarantine to discourage movement and further transmission (Srivastav, 2020).
- **Tamil Nadu** is among the 10 worst-hit states in India (Business Standard, 2020). The government began preparing for the pandemic a month prior to identifying the first case in the state by creating isolation segments in hospitals for their largest cities, conducting awareness campaigns as well as training essential personnel. This was followed by restrictions to movement such as closure of educational institutions, businesses and public transport, a ban on inter-state travel, as well as a curfew from 9 pm to 7 am (Ramakrishnan, 2020). Welfare measures were also put into place to help the most marginalized members of

the community, with cash assistance for ration card holders and public transport drivers and construction labourers, as well as free supply of essential foods and (Tamil Nadu Bureau, 2020). Moreover, concession concession packages were offered to companies that provide COVID-19 medical equipment to encourage production (Ramakrishnan, 2020). More recently, experts in Tamil Nadu opted for awareness campaigns focusing on hand-washing practices, mask usage and social distancing tips, instead of reinforcing a lock-down (TNN, 2020). As for migrant workers, the state encouraged migrant workers to take trains back to their home states, however, once the lockdown was instituted, the Tamil Nadu government's labour department opened up more shelters (Times Now, 2020). Many, however, remained stranded, requiring help from NGOs (Gilon, 2020).

- Many states implemented similar policies in terms of communications, payments in migrant labourers, transportation and food rations, but information on the specifics was difficult to source. Regardless, this is a chaotic problem domain, where cause and effect aren't clearly linked. The authors found it challenging to substantiate a link between the policies of a state and numbers of COVID cases under current conditions, and a more exhaustive review of actual policies settings and their implementation will need to be examined when more complete data is available.

Discussion and Conclusions

Summary of findings

- Little research was carried out so far on socio-economic factors and their relation to the spread of the COVID-pandemic. The research carried out mainly focuses on economic factors like the Gross Domestic Product (GDP) (Kumar et al, 2020; Chaudry) and the age distribution of the population (Laxminarayan et al. 2020; Chaudhry et al. 2020).
- Different approaches on how the states try to control the situation have been presented. Impacts of the measures taken were difficult to observe, as mostly long-term data was available. The measures are short term and therefore no clear connection to the results from the data was found.
- Exploratory Data Analysis was done showing the states that are most affected by the pandemic (Goa and Puducherry), giving a spatial overview of the three main variables - COVID-cases, COVID related deaths, literacy rate - and showing their distribution.
- Clustering was carried out for different numbers of clusters. It was observed that states were placed into the same cluster continuously, and the grouping was mainly driven by the literacy rate. An expected connection between literacy and migration was observed.
- The research question assumed that higher literacy rates in a state would lead to lower COVID case numbers. This relation was not observed. However, a clear, unexpected observation was made regarding the connection between case numbers and literacy: higher case numbers seem to be found with higher literacy rates and lower case numbers with lower literacy rates. This also holds for the number of COVID-deaths.

Implications of the findings:

- The results indicate a connection between literacy and COVID cases, however availability, granularity, and quality of data means that it is too soon to tell whether this connection is

meaningful. Further, there is some relationship between rates of migration and COVID. These findings do indeed reveal a potential gap in the current literature, and should form a starting point for future research.

- This work fills a gap in the current research, by identifying that there are some potential connections between rates of literacy and COVID cases. The work sets out a framework for how future researchers, with more complete datasets and updated census data, might look to revisit this topic. Further, this work expands existing theories around which policy responses might be most appropriate in regions with lower levels of literacy in local languages (e.g. areas with reported high levels of literacy, but large migrant labour workforces). It reveals that more work needs to be done to understand the impacts of pandemic or even natural disaster response communications that focus on the written word.
- The lack of a clear connection between the demographic data looked at and COVID cases in India was unsurprising given the current problem domain. At present, the effects of chosen policies are highly uncertain, while the reliability of data is questionable (Vogel, 2020). Creating and implementing public health policy in such a domain is consistently challenging and will require more intensive analysis.

POSSIBLE EXPLANATIONS FOR THE FINDINGS:

- The most likely explanation for our results was the fact that we weren't able to control for all variables that might influence numbers of COVID cases. Other factors at play (e.g. income, health infrastructure, age distribution) in certain states will drive COVID cases and deaths more than literacy or migration dynamics (Chaudhry et al., 2020).
- One alternative explanation for the findings is the fact that literacy of migrant workers is mixed (Zacharias, 2018). The census data for literacy accounts for literacy in any language, not just the language of the state in which you live. This is a plausible alternative, which will require deeper investigation, including the use of more detailed metrics on literacy rates.
- Another alternative explanation for the higher rates of COVID-19 cases amongst those with high literacy rates is the connection between income level and rates of testing. Recent work has found that lower income groups in the United States were less likely to have access to testing and screening (Finch, 2020). It is reasonable to assume that this trend might also hold for India.

CONCLUSION OF FINDINGS

- Although these results do not provide definitive answers, we believe this paper can serve as a useful framework for future research. The focus on the role of literacy in India in this pandemic has not been considered extensively and what little research was done on it did not have its primary focus on this factor. It is unlikely that literacy is the primary relevant factor for the spreading of the virus, but it would be short sighted not to consider it. We speculate that it is probably a sum of factors that promote the spreading of SARS-CoV-2. Migration seemed to increase COVID-19 cases, however how much this is influenced by illiteracy remains unclear.
- Future research can build on our analysis when taking literacy into consideration. This paper also shines a light on a more vulnerable part of society that should be included in the fight against this pandemic. Migrants and illiterate people are a significant part of Indian society:

Policy should work for everyone, as this makes measures more just and increases the efficiency of policy.

LIMITATIONS AND OPPORTUNITIES FOR FUTURE WORK:

The authors recognise a number of limitations to the work produced here, and would like to highlight these, and the opportunities for future research work.

- Firstly, the authors are unfamiliar with details of political and social context in India that might underpin correlations or causal effects. Further work would ideally incorporate perspectives from those with lived experiences in the affected regions, and the clusters identified should be validated by the people represented in these groups. The analysis also didn't take into account information on the caste system in India and how this influences access to healthcare. The authors of this article, having no experience with the sociopolitical environment in India did not have the ability to comment on these factors. Future work will need to investigate this.
- Secondly, future work should look at long term trends in variables like school attendance/completion and migration dynamics to gain a more detailed picture of relationships between these variables.
- Thirdly, the 2011 Census is the source of most of the demographic data - while this information is useful, it is close to a decade old, and may not reflect more recent trends and development of regions. In 2021, another census will be conducted, this analysis should be reviewed in light of updated demographics data from this census.
- Finally, COVID Case data was used for the analysis. While the data itself was of a good quality, the numbers represented were highly reliant on testing. All states have different testing protocols, and differing rates of testing per population. For this work, cases were normalised relative to population, but significant differences in testing rates means this might be less accurate. There are other measures for estimating COVID cases in an area, like comparing deaths with the known case fatality rate (CFR), however given we are still in the (relatively) early days of the pandemic, it's hard to understand the accuracy of the CFR, so this is a similarly unreliable measure. Antibody surveys may also be a useful way of estimating the true number of cases. However this data isn't currently available for India (Vogel, 2020).

REFERENCES

Abraham, T. (2020) COVID-19 communication in India, Journal of Communication in Healthcare, 13:1, 10-12, DOI: 10.1080/17538068.2020.1758428

Adhamon, T. (2020). WHO director-general's opening remarks at the media briefing on COVID-19 - 11 March 2020. Retrieved from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

Asher, S., Lunt, T., Matsuura, R., & Novosad, P. (2019). The Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG) [Dataset]. http://www.devdatalab.org/covid

Banerjea, A. (2020, October 29). Maharashtra govt extends Covid-19 lockdown till 30 November. Mint. https://www.livemint.com/news/india/maharashtra-govt-extends-covid-19-lockdown-till-30-november-11603971462595.html

Bhagat, R., Sahoo, H., Archana, S., Roy, K., Govil, D. (2020). The COVID-19, Migration and Livelihood in India A Background Paper for Policy Makers. International Institute for Population Sciences, Mumbai The COVID-19, Migration and Livelihood in India.

British Medical Journal. (2020, April 7). Communications blackouts in Jammu and Kashmir are making coronavirus lockdown doubly frightening | BMJ. BMJ. https://www.bmj.com/company/newsroom/communications-blackouts-in-jammu-and-kashmir-are-making-coronavirus-lockdown-doubly-frightening/

Business Standard. (2020, September 30). DATA STORY: Karnataka overtakes Tamil Nadu to become third-worst-hit state. Retrieved from https://www.business-standard.com/article/current-affairs/data-story-karnataka-overtakes-tamil-nadu-to-become-third-worst-hit-state-120093000390_1.html

Census Organisation of India. (2011). Census 2011 India [Dataset]. https://www.census2011.co.in/

Centers for Disease Control and Prevention. (2020, 4 November). How to Protect Yourself & Others. CDC. Retrieved 5 November, 2020, from https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html

Chakravarti, A. (2020, April 03). Delhi Government launches COVID-19 helpline on WhatsApp to provide credible information to citizens . Retrieved from https://www.indiatoday.in/technology/news/story/delhi-government-launches-covid-19-helpline-on-whatsapp-to-provide-credible-information-to-citizens-1662749-2020-04-03

Chaudhry, R., Dranitsaris, G., Mubashir, T., Bartoszko, J., & Riazi, S. (2020). A country level analysis measuring the impact of government actions, country preparedness and socioeconomic factors on COVID-19 mortality and related health outcomes. EClinicalMedicine, 25, 100464.

COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). Johns Hopkins University. Retrieved 5 November 2020.

COVID19-India API. (2020). [Dataset]. https://api.covid19india.org/

Database of Global Administrative Areas. (2018). GADM [Dataset]. Center for Spatial Sciences, University of California, Davis. https://gadm.org/download_country_v3.html

De, A. (2020, 1 October). Coronavirus India timeline: Tracking crucial moments of Covid-19 pandemic in the country. Indianexpress. https://indianexpress.com/article/india/coronavirus-covid-19-pandemic-india-timeline-6596832/

ET Government. (2020, June 9). Maharashtra to unveil new policy for migrant workers. ETGovernment.Com. https://government.economictimes.indiatimes.com/news/governance/maharashtra-to-unveil-new-policy-for-migrant-workers/76284682

ET Online. (2020, June 29). Delhi ramps up its Covid response with the biggest treatment centre in the world - Covid fight intensifies. Retrieved from https://economictimes.indiatimes.com/news/politics-and-nation/delhi-ramps-up-its-covid-response-with-biggest-treatment-centre-in-the-world/covid-fight-intensifies/slideshow/76683512.cms

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.

Gilon, C. (2020, May 27). In Tamil Nadu, NGOs and volunteers pave way for migrant workers as govt machinery struggles to cope with COVID-19 crisis - Health News , Firstpost. Retrieved November, from https://www.firstpost.com/health/in-tamil-nadu-ngos-and-volunteers-pave-way-for-migrant-workers-as-govt-machinery-struggles-to-cope-with-covid-19-crisis-8415291.html

Ioannidis, J.P.A. (2020). A fiasco in the making? As the coronavirus pandemic takes hold, we are making decisions without reliable data. Retrieved from STAT: https://www.statnews.com/2020/03/17/a-fiasco-in-the-making-as-the-coronavirus-pandemic-takes-hold-we-are-making-decisions-without-reliable-data/

Kawoosa, V. M. (2018, November 22). How languages intersect in India. Retrieved from https://www.hindustantimes.com/india-news/how-languagesintersect-in-india/story-g3nzNwFppYV7XvCumRzlYL.html

Kumar, A., Rani, P., Kumar, R., Sharma, V., & Purohit, S. R. (2020). Data-driven modelling and prediction of COVID-19 infection in India and correlation analysis of the virus transmission with socio-economic factors. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(5), 1231-1240.

Laxminarayan, R., Wahl, B., Dudala, S. R., Gopal, K., Mohan, C., Neelima, S., & Lewnard, J. A. (2020). Epidemiology and transmission dynamics of COVID-19 in two Indian states. Science.

The Lancet (2020). India under COVID-19 lockdown. Lancet (London, England), 395(10233), 1315. https://doi.org/10.1016/S0140-6736(20)30938-7

Learnings from Kerala. (2020, June 2). Retrieved November 06, 2020, from https://www.who.int/india/news/feature-stories/detail/responding-to-covid-19---learnings-from-kerala

Majid, Z. (2020, October 11). 'Mysterious' decline in Jammu & Kashmir's Covid-19 cases, explained. Deccan Herald. https://www.deccanherald.com/national/north-and-central/mysterious-decline-in-jammu-kashmir-s-covid-19-cases-explained-900354.html

Menon, B. S. (2020, October 29). Coronavirus: India tries new type of tests to tackle virus. BBC News. https://www.bbc.com/news/world-asia-india-53609404

Menon, J. C., Rakesh, P. S., John, D., Thachathodiyl, R., & Banerjee, A. (2020). What was right about Kerala's response to the COVID-19 pandemic?. BMJ Global Health, 5(7), e003212.

Mondkar, A. (2020, 4 April). COVID-19, India and crisis communication. Orfonline. https://www.orfonline.org/expert-speak/covid-19-india-and-crisis-communication-64102/

Mukhopadhyay, A. (2020, September 3). How Indian police contained coronavirus in Mumbai. DW.COM. https://www.dw.com/en/how-indian-police-contained-coronavirus-in-mumbai/a-54597021

Norman J., Bar-Yam, Y., Taleb, N. M. (2020) Systemic risk of pandemic via novel pathogens – Coronavirus: A note, New England Complex Systems Institute.

National Sample Service Office. (2010). Migration in India 2007-2008. NSS Report No.533 (64/10.2/2)

Obama White House. (2014). Remarks by the President on Research for Potential Ebola Vaccines. Retrieved November 5, 2020, from https://obamawhitehouse.archives.gov/the-press-office/2014/12/02/remarks-president-research-potential-ebola-vaccines

OneIndia. (2020, November 06). Coronavirus LIVE: India reports 47,638 new COVID-19 cases and 670 deaths in the last 24 hours. Retrieved from https://www.oneindia.com/india/coronavirus-live-india-reports-47638-new-covid-19-cases-and-670-deaths-in-last-24-hours-3043504.html

Pandey, G. (2020, April 21). Coronavirus in India: Desperate migrant workers trapped in lockdown. Retrieved from https://www.bbc.com/news/world-asia-india-52360757

Pandya, D. (2020, June 13). How Asia's Densest Slum Chased the Virus Has Lessons for Others. Bloomberg News. https://www.bloomberg.com/news/articles/2020-06-13/how-asia-s-densest-slum-chased-the-virus-has-lessons-for-others

Ramakrishnan, A. (2020, April 29). Tamil Nadu Government's Response to COVID-19. Retrieved from https://www.prsindia.org/theprsblog/tamil-nadu-government%E2%80%99s-response-covid-19

Reynolds, B. (2014).Crisis and Emergency Risk Communication (Manual). Chapter 1. https://emergency.cdc.gov/cerc/resources/pdf/cerc_2014edition.pdf

Sarkar, A., Chouhan, P. (2020). COVID-19: District level vulnerability assessment in India. Clinical epidemiology and global health. 10.1016/j.cegh.2020.08.017.

Scikit-learn developers. (n.d.). Clustering. Scikit-Learn. Retrieved 1 November 2020, from https://scikit-learn.org/stable/modules/clustering.html

Scikit-learn developers. (n.d.). Compare the effect of different scalers on data with outliers. Scikit-Learn. Retrieved 1 November 2020, from https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html

Singh, J. (2020, May 29). Why Adityanath's Simplistic Migration Commission Is a Non-Starter. The Wire. https://thewire.in/labour/uttar-pradesh-migration-commission

Srivastav, T. (2020, April 17). Why India has the upper hand against COVID-19. World Economic Forum. https://www.weforum.org/agenda/2020/04/india-covid19-coronavirus-response-kerala-uttar-pradesh/

Tamil Nadu Bureau. (2020, August 10). Free rations, coverage of COVID-19 treatment: Here's how Tamil Nadu is fighting the virus. Retrieved from https://www.thehindu.com/news/national/tamil-nadu/free-rations-coverage-of-covid-19-treatment-heres-how-tamil-nadu-is-fighting-the-virus/article32314037.ece

Times Now. (2020, March 29). Lockdown 21: No train home, Tamil Nadu government arranges shelter home for migrant workers: City - Times of India Videos. Retrieved from https://timesofindia.indiatimes.com/videos/city/chennai/lockdown-21-no-train-home-tamil-nadu-government-arranges-shelter-home-for-migrant-workers/videoshow/74875812.cms

TNN, 2. (2020, June). Tamil Nadu: Experts choose awareness over another lockdown: Chennai News - Times of India. Retrieved from https://timesofindia.indiatimes.com/city/chennai/experts-choose-awareness-over-another-lockdown/articleshow/76698766.cms

UNESCO. (2017). Literacy Rates Continue to Rise from One Generation to the Next (Fact Sheet No.45).

Vogel, G. (2020, April 22). Antibody surveys suggesting vast undercount of coronavirus infections may be unreliable. Science | AAAS. https://www.sciencemag.org/news/2020/04/antibody-surveys-suggesting-vast-undercount-coronavirus-infections-may-be-unreliable

Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1 - 23. doi:http://dx.doi.org/10.18637/jss.v059.i10

Zacharias, S. (2018). The Human Rights Issues Related to Right to Education of the Children of Migrant Labourers in Kerala.

APPENDIX A - VARIABLES AND DATA SOURCE

Table 1 details the variables used in clustering of the data, their meaning, and their source.

Table 1 Appendix A: Variables, description and source

	Variable name	Variable description	Variable source
1	cases_conf	Total confirmed COVID cases by state from February 2020 to 25 October 2020	https://api.covid19india.org/csv/latest/state_wise.csv Confirmed summed by state
2	deceased	Total confirmed COVID deaths by state from February 2020 to 25 October 2020	https://api.covid19india.org/csv/latest/state_wise.csv Deceased summed by state
3	hand_wash_water	Percentage/ratio of households with water available for hand washing	http://www.devdatalab.org/covid NFHS data
4	hand_wash_soap	Percentage/ratio of households with soap available for handwashing	http://www.devdatalab.org/covid NFHS data
5	mem_per_room	Average number of people per room in households	http://www.devdatalab.org/covid NFHS data
6	pc_pop_over65	Percentage/ratio of population in the state aged over 65	http://www.devdatalab.org/covid NFHS data
7	pc_ob_fem	Percentage/ratio of females in the state classified as obese	http://www.devdatalab.org/covid NFHS data
8	smoke_fem	Percentage/ratio of females in the state who smoke tobacco	http://www.devdatalab.org/covid NFHS data
9	inltmigrationrate	Rate of long term migration into the state	http://www.devdatalab.org/covid Migration data
10	outltmigrationrate	Rate of long term migration out of the state	http://www.devdatalab.org/covid Migration data
11	2011_pop	Population of each state from the 2011 Indian Census	http://www.devdatalab.org/covid Demographics data
12	2011_urban_pop_share	Proportion of state's population living in urban areas in 2011	http://www.devdatalab.org/covid Demographics data

13	2011_slum_pc	Percentage/ratio of state's population living in slums/informal settlements in 2011	http://www.devdatalab.org/covid slum_pop/2011_pop
14	2011_slum_pop	Total population living in slums in 2011	http://www.devdatalab.org/covid Demographic data
15	2011_pop_dens	Population density of each state in 2011 (population per square kilometre)	http://www.devdatalab.org/covid Demographic data
16	2011_literacy	Literacy rate of population in 2011	http://www.devdatalab.org/covid Demographic data
17	cases_adj	Population adjusted number of COVID cases per 100 000 people per state	cases_conf/2011_pop
18	deaths_adj	Population adjusted number of COVID deaths per 100 000 people per state	deceased/2011_pop

APPENDIX B - CHOROPLETHS

Figure 9 Appendix B: Choropleths for all the variables investigated

APPENDIX C - CLUSTERS AND SIZE, KDE-PLOTS

Table 2 Appendix C: clusters and cluster size.

cluster	2		3			5					7							10									
cluster name	0	1	0	1	2	0	1	2	3	4	0	1	2	3	4	5	6	0	1	2	3	4	5	6	7	8	9
cluster size	19	15	10	14	10	2	9	10	6	7	4	6	6	2	7	8	1	1	7	7	5	4	1	2	3	2	2

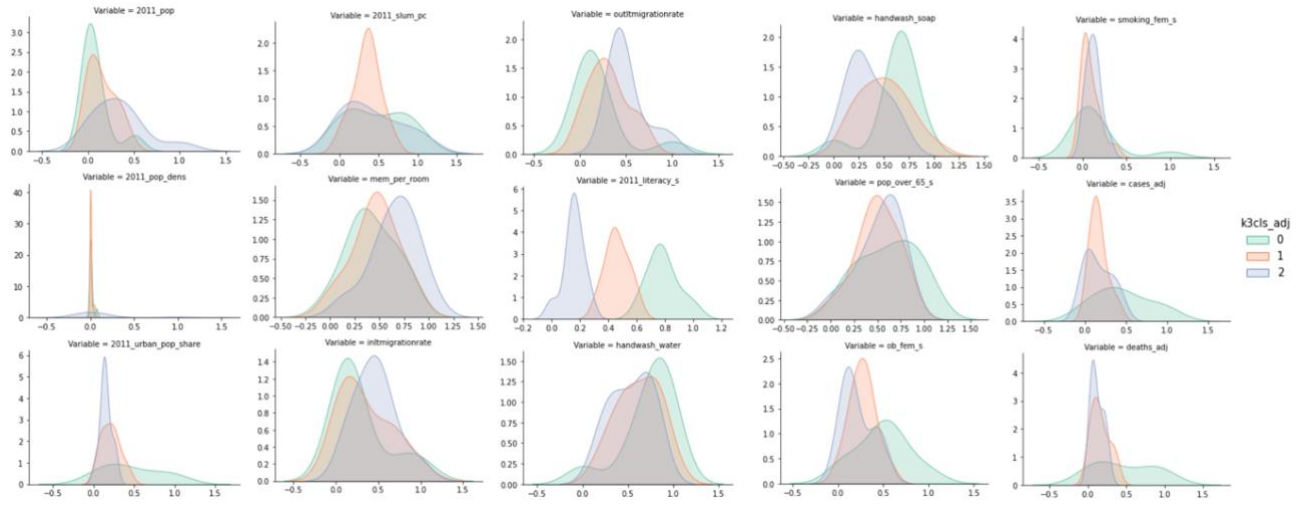

Figure 10 Appendix C1: KDE-plots for three clusters

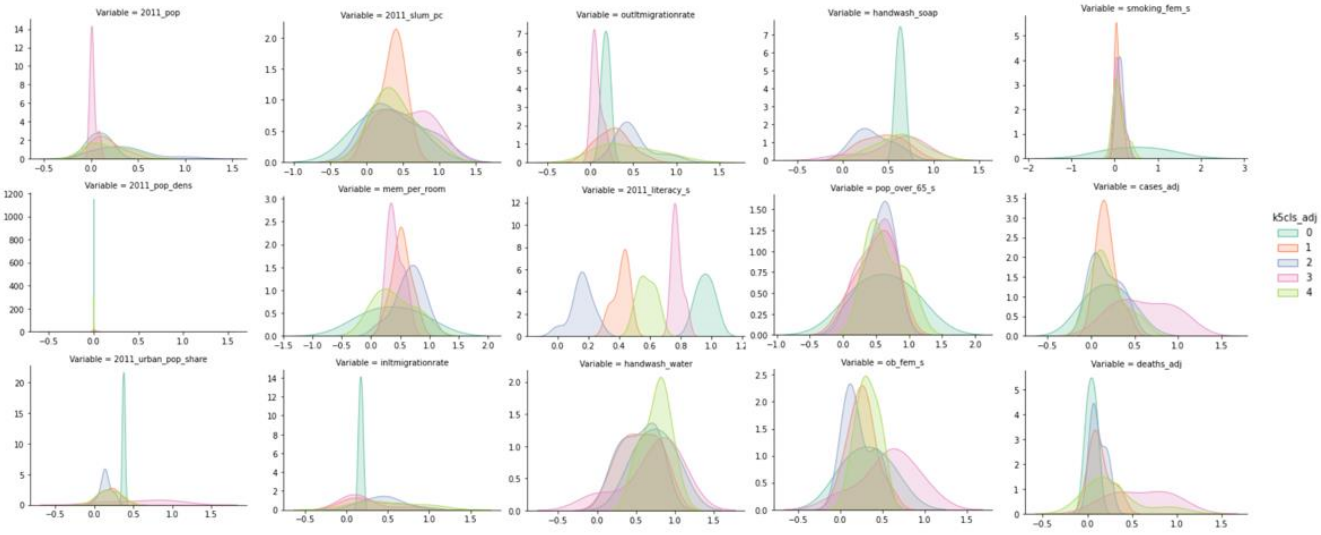

Figure 11 Appendix C2: KDE-plots for five clusters

APPENDIX D - STATES AND UNIONS, MAPPED

Figure 12 Appendix D: Map of India with the states and unions examined in this work