Journal of the Indian Society for Probability and Statistics (2021) 22:319–342
https://doi.org/10.1007/s41096-021-00106-1

RESEARCH ARTICLE

Effective Reproduction Number and Dispersion under Contact Tracing and Lockdown on COVID-19 in Karnataka

Siva Athreya[1] · Nitya Gadhiwala [1] · Abhiti Mishra [1]

Accepted: 30 August 2021 / Published online: 24 September 2021

Abstract

We study the effectiveness and limitations of contact-tracing, quarantine, and lockdown measures used in India to control the spread of COVID-19 infections. Using data provided in the media bulletins of Government of Karnataka we observe that the so called $20-80$ rule holds for secondary infections and classify them into clusters. Using a mixture of Poisson with Gamma model we establish that clusters show variation in deceased rates ($0\%-17.31\%$), low reproduction numbers ($0.21-0.77$), small dispersion($0.06-0.18$), and that super-spreading events can occur. Further, migration due to relaxation in lockdown is unlikely to be the sole cause of recent surge. The methodology presented is universal in nature and can be applied whenever such precise data is available.

Keywords Variation · Individual infectiousness · Maximum likelihood · Negative binomial · Superspreading event.

1 Introduction

For COVID-19, in the absence of a vaccine, key measures to contain infection spread have been lockdowns, contact tracing, quarantine, testing along with wide publicity of social distancing norms, hygiene guidelines, awareness of the symptoms of the disease and treatment. There are many efforts to understand control measures such as lockdowns, contact tracing and quarantine with respect to

✉ Siva Athreya
athreya@isibang.ac.in

Nitya Gadhiwala
nitya.g20@gmail.com

Abhiti Mishra
abhitimishra16@gmail.com

[1] Indian Statistical Institute, 8th Mile Mysore Road, Bangalore 560059, India

COVID-19 spread using stochastic models, see for e.g. (Joel et al. 2020; Ferretti et al. 2020). Contact tracing and other control measures were also used by countries during the Severe Acute Respiratory Syndrome (SARS) epidemic, see (Lipsitch et al. 2003; Steven et al. 2003) for a detailed analysis on control efforts and clusters initiated by "super-spread" events (SSEs) and community transmission. In Ramanan et al. (2020), the authors study the epidemiology and transmission of COVID-19 in two states of India namely, Tamil Nadu and Andhra Pradesh, using testing and contact-tracing data.

In Lloyd-Smith James et al. (2005), they argue that studying only the basic reproduction number can obscure the individual variation in infectiousness. Their motivation being 'super-spreading events' in which certain individuals had infected unusually large numbers of secondary cases (5–10 in the SARS epidemic). They studied contact tracing data from eight directly transmitted diseases, and showed that the distribution of individual infectiousness around the basic reproduction number is skewed. Using various models they then proceed to compare the effect of individual-specific control measures versus population-wide measures. They conclude that super-spreading events are a normal feature of disease spread and give a formal definition of the same.

To contain the spread of COVID-19 infections in India, the Union Government started a strict lockdown on 25th March and relaxed it over 5 phases as follows: Lockdown Phase 1 (25th March–14th April) and Lockdown Phase 2 (15th April–3rd May) were the strictest in terms of mobility; Lockdown Phase 3 (4th May – 17th May) and Lockdown Phase 4 (18th May – 31st May) included relaxations in travel between states; and Unlock 1.0 (1st–30th June), Unlock 2.0 (1st–31st July) had considerable relaxations.

In Karnataka, a state of India with a population of approximately 70 million, from the very beginning quarantine measures and contact tracing were put in place for all tested positive patients. Since 9th March 2020, the Government of Karnataka has been providing detailed media bulletins (Novel Coronavirus (COVID-19) 2020) containing specific guidelines on the virus and information on each patient who was tested positive in the state.

In this article we study the trace history provided in the media bulletins and try to understand the spread of the disease in the period from 9th March till 21st July 2020 in the state of Karnataka. From the trace history (see Covid19 india-timeline an understanding across states and union territories 2020) we classify the patients who tested positive into several clusters. We analyse each cluster and the spread of disease within them. We also comment on the reasons for the possible spurt in cases from 27th June, 2020 on-wards.

2 Materials and Methods

The COVID-19 media bulletins of the State of Karnataka, from 7th March to 26th June, provided detailed information on the tested positive patients. In particular there was data on how each one of them contracted the virus (either due to travel history or by being a contact of someone who has already tested positive for COVID-19) or what led to them being tested (either as a Severe Acute Respiratory Infection patient or someone with Influenza like symptoms).

2.1 Clusters

We first classify the tested positive cases into clusters based on the source of infection, for example “From Europe” or “Pharmaceutical Company Nanjangud”. Then in each cluster we place all the patients who contracted the virus independently from the place of origin, and then recursively add the patients to whom they passed the infection.

Before Phase 1 (25th March - 14th April) of the lockdown began, almost all the COVID-19 cases that were confirmed in Karnataka were either individuals who had some form of international travel history (from Middle East, USA, South America, United Kingdom and the rest of Europe) or those who were contacts of such individuals. Phase 1 and Phase 2 (15th April - 3rd May) of the lockdown in Karnataka saw heavy restrictions on travel and nearly all services and factories were suspended.

During Phase 1 and Phase 2 of the lockdown, a Pharmaceutical company in Nanjangud, Mysore, saw a sudden increase in the COVID-19 cases. Although the exact reason for the infection to have reached the company is unknown, the first patient to be infected (35 year old male, was confirmed to be infected on 26th March) came in contact with health care workers treating COVID-19 patients. Another cluster that began during this period was the “TJ Congregation”, which contained those who attended the Tablighi Jamaat Congregation from 13th to 18th March in Delhi. The first patient in this cluster was confirmed as a COVID-19 case on 2nd April. Both these clusters were very well contained and the last patients to be attributed to these clusters tested positive on 29th April and 21st May respectively. No more patients were attributed to these clusters since then. Phase 3 and 4 of the lockdown loosened restrictions on Domestic Travel and many infected individuals had some domestic travel history. The state saw a large influx of infected individuals from states like Maharashtra, Gujarat, Rajasthan and the Southern States (Tamil Nadu, Telangana and Andhra Pradesh). There were also patients whose source of infection was listed as inter-district travel in Karnataka, travel to foreign countries or other states, healthcare workers and policemen on COVID-19 duty and their contacts. The cases due to these reasons were too few to form separate clusters. We placed all these patients in a cluster called “Others”.

Testing strategy in India is governed by ICMR guidelines. The guidelines on 20th March mandated that all Severe Acute Respiratory Illness patients (i.e., patients with fever AND cough and/or shortness of breath) should be tested for COVID-19, while the guidelines on 4th April mandated the same for all symptomatic patients with Influenza like Illness (fever, cough, sore throat, runny nose). Thus two other clusters that began during Phase 1 and Phase 2 of the lockdown were the Severe Acute Respiratory Infection (“SARI”) (first infection 7th April) and Influenza Like Illness (“ILI”) (first infection 15th April) clusters. These clusters contain those patients who have a history of SARI(and ILI), and those who can be traced back as contacts of such patients. It should be noted that only the first generation of the patients in this cluster are those with a history of SARI (and ILI), but the subsequent contacts of these patients need not be. In the media bulletins, patients whose contact tracing was incomplete were mentioned as ‘Contact Under Tracing’. We have assumed that these patients did not fall under SARI or ILI and placed them in a cluster called “Unknown”, along with their contacts who tested positive. An initiative taken by the government was to create Containment Zones in certain regions. The guidelines for these zones were clearly specified. The first case in contact with a containment zone was reported on 24th April. Since then a large fraction of the increase in this cluster occurred during Phase 3 (4th May–17th May) and Phase 4 (18th May–31st May) of the lockdown. For all these clusters, there was no information provided on the source of infection for the ‘parents’.

Our consolidated list of clusters are then given by

From Middle East, From USA, From United Kingdom, From Rest of Europe,
From South America, From Maharashtra, From Rajasthan, From Southern States,
From Gujarat, Influenza like illness(ILI), Severe Acute Respiratory Infections(SARI),
Unknown, Pharmaceutical Company-Nanjangud, T.J. Congregation in Delhi,
Containment Zones, Others.

(2.1)

2.2 Reproduction number and Dispersion

In epidemiology, the “basic reproduction number” of an infection, denoted by R_0, can be thought of as the expected number of cases to have contracted the infection directly from one case. Thus on an average, each infected person passes on the infection to R_0 many healthy individuals. As mentioned earlier, in Karnataka during the period 9th March - 26th June we have observed the COVID-19 infection spread in a controlled environment. So whenever we calculate basic reproduction numbers we are actually calculating the short term effective reproduction number of the disease during this period. To be cognizant of this we shall use the notation R_{eff} to denote the basic reproduction number for a cluster instead of the usual notation R_0.

We will examine Reproduction number and dispersion for “The 8 clusters” in this section, namely:

From Southern States, Influenza like illness, Severe Acute Respiratory Infections,
Containment Zones, Unknown, Others, TJ Congregation in Delhi, and
Pharmaceutical Company Nanjangud.

(2.2)

These began before 3rd May 2020 and have more than 50 individuals. There are ten clusters that satisfy these criteria from (2.1). We have omitted two clusters from analysis which satisfy these criteria, namely: “From Maharashtra” and “From Middle East”. We will analyse them in a later section. In Fig. 3 we present a summary distribution of parents, children, grandchildren, and great grandchildren in each of “The 8 clusters”.

For each individual i in the cluster we will denote the number of children (or the number of tested positive cases) assigned to patient i by y_i. This means that there were y_i many positive infections whom the media bulletins listed as ‘Contact of Patient-i’. The mean of y_i is the basic reproduction number R_{eff}. In Table 1 we present a comparison of the summary distribution parameters (Maximum, Zeroes, Size, etc.) across clusters and we see that the variance does not match the mean. Further, as noted in Fig. 1, heterogeneity in the infectiousness of each individual implies that R_{eff} by itself is not a good measure of the infection spread. To account for the large variance, we now consider the standard method of mixture of Poisson distributions to model the data set. For each cluster, using the Negative Binomial with mean R_{eff} and dispersion k (see Lloyd-Smith James et al. 2005 and Section A for details) as the offspring distribution, we will use the Maximum Likelihood method for estimating R_{eff} and k (see Section B for details). Using the methods developed in Saha and Paul Sudhir (2005) we provide 95% confidence interval for k and conditional on the estimates we perform the χ^2-goodness of fit test. The details of the above can be found in Sect. B, C, and D of the Appendix.

2.3 Cases due to Migration in Phase 3,4 and Unlockdown 1.0

As mentioned earlier we had omitted two clusters from analysis, namely: From Maharashtra and From Middle East. Phase 3 and 4 of the lockdown, along with Unlockdown 1.0 in June loosened restrictions on Domestic Travel and International travel. The state saw a large influx of infected individuals from within India and abroad. During Phase-3 of the lockdown, the “From Maharashtra” cluster saw the most growth and dominated the test positive counts by a significant margin. The “From Maharashtra” cluster accounted for approximately 52.5% of cases in the stipulated period. The “From Middle East” cluster seems to have two phases. The first occurred before the lockdown was enforced during which international travel was suspended. The second, more recent, was due to the repatriation flights from the region. We provide the Maximum Likelihood estimators for R_{eff} and k, along with

Table 1 This table considers the different generations of infections as seen in Karnataka for “The 8 clusters”. For each generation, the table contains the number of individuals in that generation, the number of patients causing zero secondary infections, the maximum number of infections caused by an individual in that generation and the mean number of infections caused by an individual in that generation

Generation	Size	Zero	Maximum	Mean
Parent	2925	2528	30	0.4499
Child	1314	1176	51	0.3067
Grandchild	403	344	14	0.3524
Great Grandchild	142	112	45	0.7042
Great Great Grandchild	100	94	5	0.11

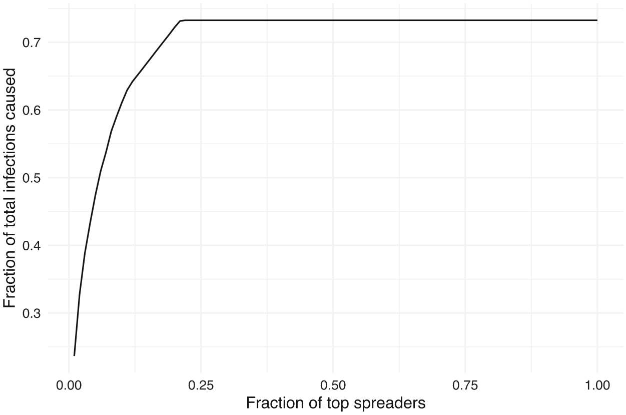


Fig. 1 The plot considers those individuals who were infected before 3rd May along with all those cases that can be traced back as contacts of them. The infected individuals have been ranked in terms of the number of secondary infections caused by them. Then the top x fraction of them are considered. In this graph, x has been plotted on the x-axis and the fraction of the total infections infected by them on the y-axis

their summary in Table 1. During this period domestic and international travellers were quarantined/tested on arrival. To make any meaningful inferences using reproduction numbers and dispersion one would have take into account a more detailed tracing history procedure from their origin of travel.

To understand cases due to Migration (6871 out of 10391) in this period we reorganized our clusters from (2.1) into four groups. Namely

Inter-District Travel: consisting of 429 patients who belong to ''Others'' cluster
whose testing positive is attributed to inter-district travel within Karnataka;

Inter-State Travel group: consisting of 582 patients who belong
to ''From Gujarat'', ''From Rajasthan'', ''From the Southern States''
(Kerala, Tamil Nadu, Telangana and Andhra Pradesh) and ''Others'' cluster
who had traveled to Delhi.;

Foreign group: consisting of 379 patients who belong to
the ''From Middle East'', ''From United Kingdom''
and ''From the rest of Europe'' cluster as well as a few cases
which originated from Nepal, Indonesia, Philippines and Malaysia; and

From Maharashtra cluster: consisting of 5481 patients in that cluster as in (2.1).
(2.3)

2.4 Data

We have sourced all our data from the Daily Media Bulletins of Government of Karnataka: https://karnataka.gov.in/common-10/en (till 27th April, 2020) and https://covid19.karnataka.gov.in/govt_bulletin/en (post 27th April, 2020). The media bulletins were very detailed and contained the following information till 21st, July 2020. We have converted them from their pdf format into usable CSV format and made them publicly available for use at our Data Repository at https://www.isibang.ac.in/~athreya/incovid19/.

3 Results

One thumb rule for disease spread, including COVID-19 anecdotally, is the 20/80 rule. The rule states that 80% of the secondary infections arise from 20% of the primary infections. From Fig. 1, it can be observed that for Karnataka, almost 20% of the individuals with the highest infectiousness are responsible for 70% of the total infections. The large deviation from the $y = x$ straight line represents the heterogeneity in the infected individual population.

If we consider the entire data as one Karnataka cluster then we find that its effective reproduction number is 0.2021 and dispersion is 0.0358 with a 95%

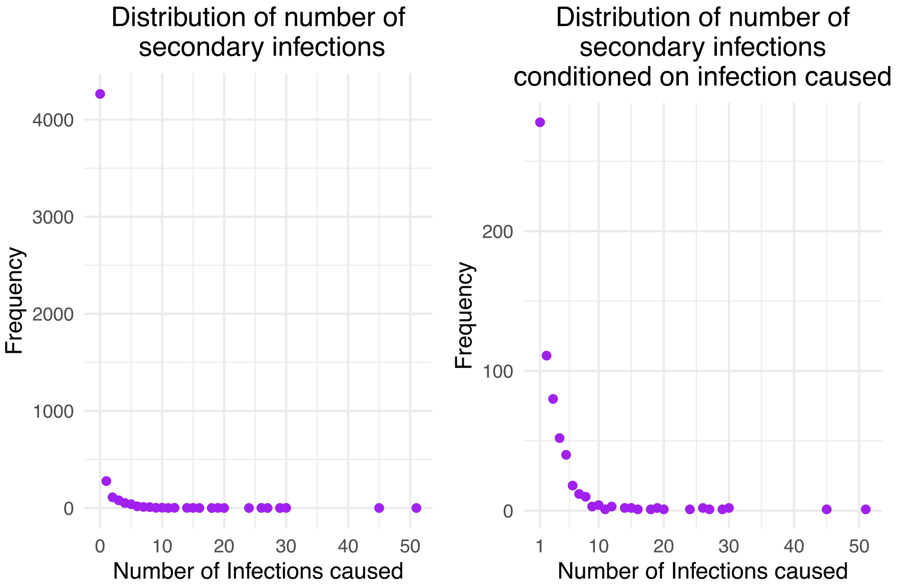


Fig. 2 This scatter plot considers the COVID-19 patients in Karnataka and represents the distribution of the number of infections caused by each patient. The patients belonging to “The 8 clusters” in study are considered here. The plot on the left shows the frequency distribution of the number of infections assigned to each infected individual as their contact. The plots have the number of infections caused on the x-axis and the number of patients that have caused x many infections on the y-axis. The graph on the right is the same as the one on the left without the point at 0. It is the distribution of number of infections caused conditioned on at least one infection caused

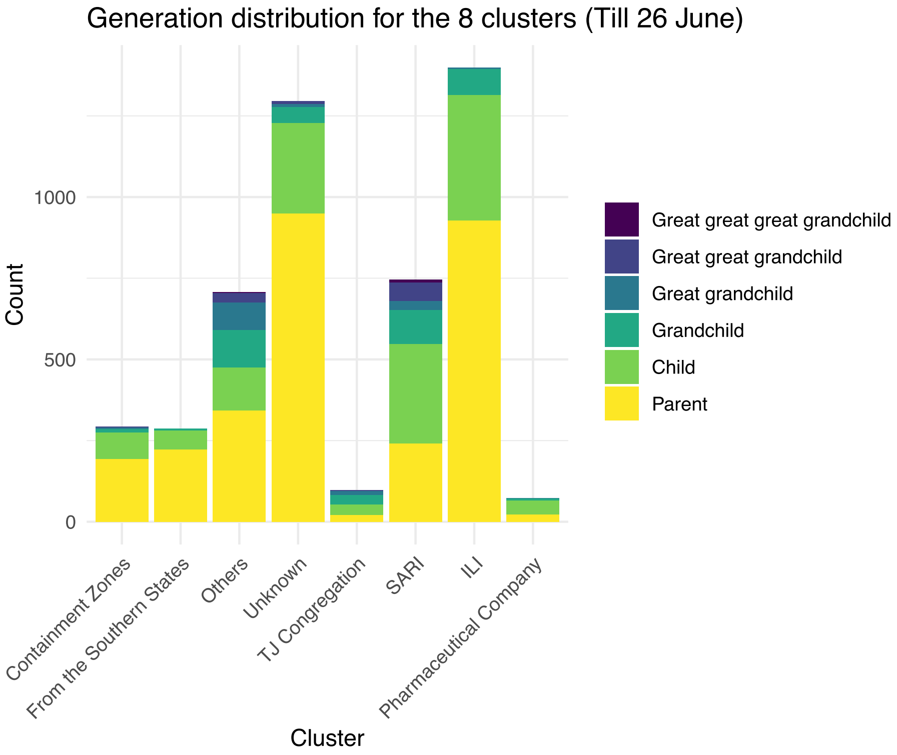


Fig. 3 This plot is a stacked histogram displaying the distribution of generations for “The 8 clusters” we consider till 26th’ June. The histogram represents the number of infections that belong to each of these clusters and each bar has been further filled with different colors to denote the number of primary infections, secondary infections and so on

confidence interval given by (0.033, 0.039). However to better understand the variations in the spread of infection we will present findings from each of the eight clusters. (Figs. 2, 3)

“The 8 clusters”

- *Heterogeneity and Variation:* In Table 2, we have computed the Maximum Likelihood estimates for R_{eff} and k for “The 8 clusters” and also performed the χ^2-goodness of fit test (see Section C for details regarding the goodness of fit). In Fig. 5 we have plotted the histogram from the derived Negative Binomial probabilities for each of “The 8 clusters” along with the observed relative frequencies of the number of infections caused. We have marked the 95th and 99th percentile for these distributions in the plot. In Table 2 and Fig. 4 we provide the confidence intervals for dispersion parameter with respect to “The 8 clusters”.

  The “TJ Congregation” and the “Pharmaceutical Company Nanjangud” clusters both have higher R_{eff} among all clusters. The p-values provide in Table 2 are not small for all clusters except for the cluster “Pharmaceutical Company Nanjangud”. This cluster has a very high variation, a maximum data point at

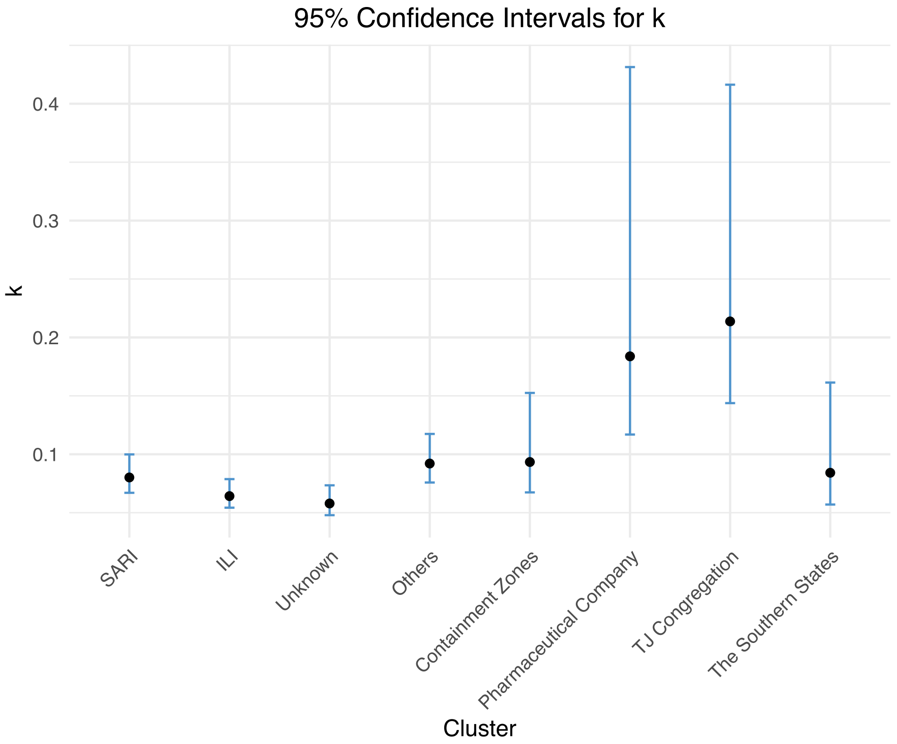


Fig. 4 This plot contains the calculated values of the dispersion parameter, k, along with its 95% confidence intervals for each of "The 8 clusters" in observation

24 (i.e., one person who has been assigned to 24 secondary infections) and also a significant proportion at 1 secondary infection caused. One can also see that the confidence interval for the dispersion for "Pharmaceutical Company Nanjangud" cluster is quiet large as well, as seen in Fig. 4 and that the histograms differ with the Negative Binomial model in Fig. 5 as noted.

For each cluster in "The 8 clusters", we found that basic reproduction number is less than 1 but the variance is larger than the mean. However, the distribution of secondary infections across all clusters is very skewed, with a significant mass at 0 due to the control measures taken. From the Negative Binomial model, we note that for most clusters their dispersion is low and is contained in a small confidence interval. Thus, though the clusters will most likely die out under the controlled environment, there is a reasonable chance of super-spreading events occurring.

- *Super-spreading events:* In Fig. 2, we examine the distribution of the number of infections designated as contacts of infected indiviuals patients. A large peak is seen at 0 infections caused. It can be seen that only 9 individuals in the population of 4895 have passed the infection on to more than 20 people. This could be the result of a super-spreader phenomenon or perhaps an effect of how the contact tracing and testing is performed. Assigning them as definitely arising

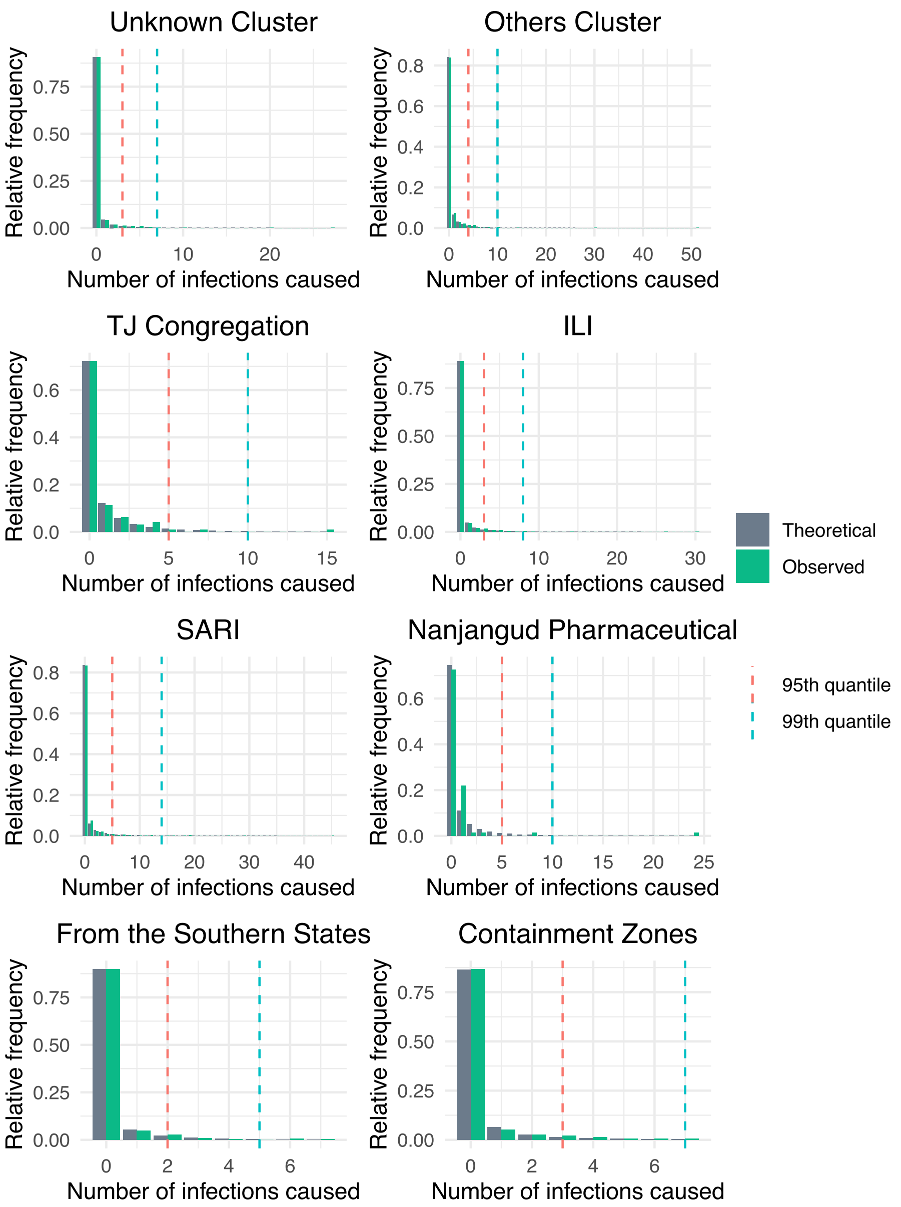


Fig. 5 This graph plots for each cluster the observed relative frequencies of the offspring distribution along with the theoretical negative binomial probabilities. The green bars represent the observed relative frequencies and the grey bars represent the probabilities as calculated from the negative binomial distribution

from one particular individual will need a more careful understanding of the latter. One can further note that due to effective quarantine measures there are 4265 infected individuals who have not passed the infection on to anyone else.

Note that in Table 2, the largest number of secondary infections assigned to an individual is quite high for some clusters. This might be indicative of the super-spreading phenomenon. From Lloyd-Smith James et al. (2005), a general protocol for defining a super-spreading event is as follows: (1) estimate the effective reproductive number, R_{eff}, for the disease and population in question; (2) construct a Poisson distribution with mean R_{eff}, representing the expected range of Z (without individual variation); (3) define a Super-spreading event as any infected individual who infects more than Z_n others, where Z_n is the nth percentile of the Poisson(R_{eff}) distribution.

If R_{eff} and k have been estimated then one can use the definition and the Negative Binomial model to understand the probability with which such events will occur.

If we were to consider a $99th$ percentile event with the above $R_{\text{eff}} = 0.3447$, then an event causing more than 2 secondary infections would be considered a super-spreading event. In the "Containment Zone" cluster, there is a person who has been assigned 7 secondary infections, this would be considered a super-spreading event. Under the Negative Binomial model the probability of observing 7 secondary infections is 0.0027. This may indicate one of two possibilities, either a very, very rare event has occurred or it is just an effect of the testing and contact tracing method that was followed.

The relative frequency of super-spreading events within "The 8 clusters" can be calculated using Table 2 and Fig. 5. The above indicates that the infection can be stemmed quicker by containing these super-spreading events by using effective contact tracing.

- *Variation over time:* If we consider 7th April and Descendants till 21st April, then there were 290 patients who tested positive and out of them 219 did not pass the infection to anyone else. There was one person who had been assigned 24 secondary infections and the mean number of secondary infections was at 0.6793 with a variance of 4.482. In contrast, if we consider the period 7th April to 3rd May and Descendants till 17th May, then there were 615 patients who tested positive and out of them 491 did not pass the infection to anyone else. There was one person who had been assigned 45 secondary infections and the mean number of secondary infections was at 0.7512 with a variance of 9.946.

  The basic reproduction number is by no means a unique number for a disease or for that matter within a cluster. It greatly varies: with time from beginning to end; within a region due to its population density; and with interventions put in place to curb the spread of the infection. In Fig. 6 we compute the reproduction number for each of "The 8 clusters" studied and note that there is a significant variation over time. The "Pharmaceutical Company" cluster seems to have a reproduction number of 4 during the first week and then tapers off to 0 in five weeks. The "TJ Congregation" cluster also has a reproduction number that has variation over time but eventually due to tracing and testing tapers off to 0. "SARI" and "ILI" clusters have fluctuations throughout the period, due to new parents being added to the cluster.
- *Variation over Generations:* Table 1 contains information on "The 8 clusters" with respect to generations within them. The maximum number of infections

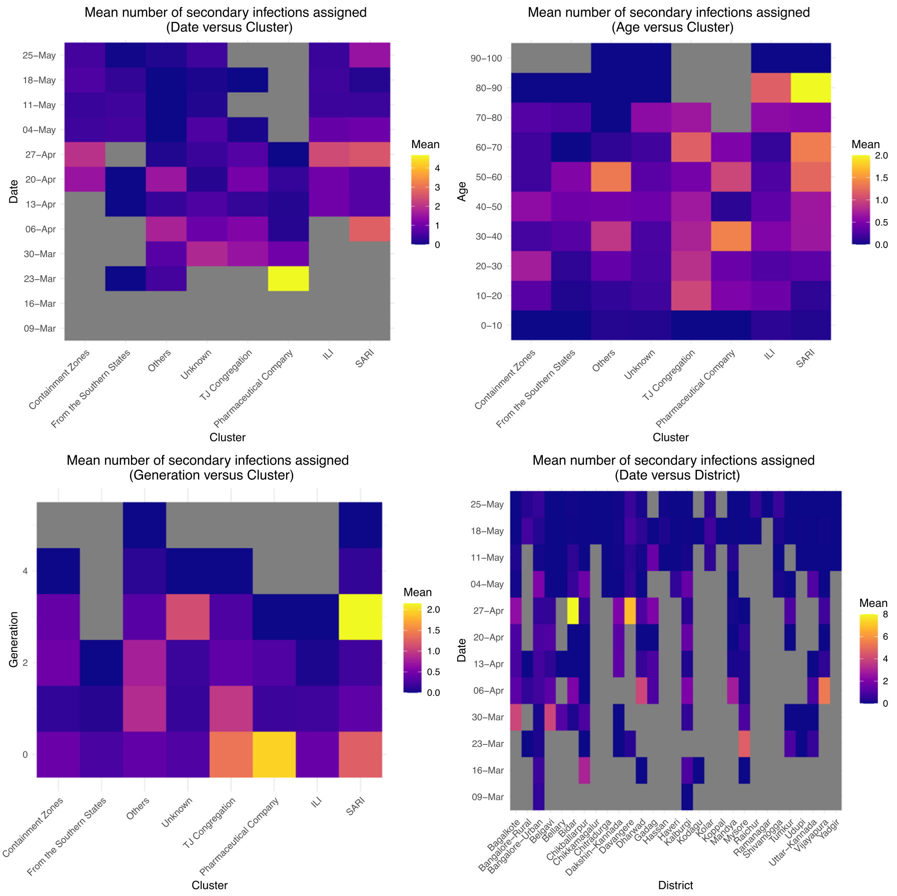


Fig. 6 This is a heatmap of the effective reproduction numbers of the individuals across diffferent criteria. These are the Dates of their infection and Cluster in the first graph, Age versus Cluster in the second graph, Generation versus Cluster in the third graph, and Date of their infection versus the District in the fourth graph. Each tile represents the mean number of secondary infections caused by the patients

caused by an individual in the first generation is 30. An individual in the “Influenza like illness” cluster and another in the “Others” cluster have caused 30 secondary infections each. Among the individuals in the second generation the one to have caused 51 infections belongs to the “Others” cluster. It is observed that the mean secondary infections of patients belonging to the Generation-4 (Great Grandchild) is 0.7042 and is significantly higher than the remaining generations. This is because of the small size of the generation (142) and one of the patients being assigned 45 secondary infections. While the highest generation that can be observed is Generation-6 (Great great great grandchild), they haven’t been included in Table 1 as there isn’t a Generation-7 for any cluster resulting in all the individuals in Generation-6 being assigned 0 infections.

A heat map representing the mean infections caused, as studied across clusters

Table 2 This table contains the following information about "The 8 clusters" that we have considered- the 'Size' row represents the number of infected individuals that belong to the cluster. The 'Zeros' row denotes the number of patients who haven't been assigned any secondary infection. The 'Maximum' row represents the maximum of the number of secondary infections assigned to any individual. The 'R_{eff}' and 'Variance' row represents the mean and variance (respectively) of the secondary infections assigned to individuals. The row 'k' contains the Maximum Likelihood estimates for the dispersion parameter, k, and the row 'p-value' contains p-value from the χ^2 goodness of fit test of the Negative Binomial Distribution fitted to the Data (See Sections B and C for details). This table computes the 90th, 95th and 99th quantiles of the Poisson(R_{eff}) distribution. The quantiles will define 90, 95 and 99th-Super-spreading events respectively

	Containment zones	ILI	SARI	TJ Congregation	Others	From the southern states	Pharmaceutical company	Unknown	Karnataka
Size	293	1398	746	97	707	286	73	1295	11005
Zeros	254	257	1243	593	53	622	70	1173	10282
Maximum	7	30	45	15	51	7	24	27	51
Variance	1.199	2.355	8.521	3.823	6.502	0.6897	8.757	1.637	1.867
R_0	0.3447	0.3369	0.6743	0.7732	0.5191	0.2168	0.726	0.2625	0.2022
k	0.09345	0.06428	0.09214	0.1839	0.08023	0.08424	0.2138	0.05797	0.0358
p-value	0.3744	0.618	0.3467	0.06031	0.7339	0.3245	0.001089	0.07156	0.01184
90th Percentile	1	1	2	2	1	1	2	1	1
95th Percentile	1	1	2	3	2	1	2	1	1
99th Percentile	2	2	3	4	3	2	4	2	2
Recovery Rate	96.06	91.21	82.68	98.75	98.06	95.91	100	93.08	97.58
Deceased Rate	3.94	8.79	17.32	1.25	1.94	4.09	0	6.92	2.42

and generations, is seen in Fig. 6. All clusters have been contained within 5 generations, this is seen by the fact the mean is 0 for the final generation of the cluster which is at most the fifth. One can see that the “TJ Congregation” and “Pharmaceutical Company Nanjangud” clusters have variation in mean across generation with the mean number of infections decreasing across each generation. The clusters closed out and did not added any new patients as per the bulletins. Generation 3 in the “SARI” cluster shows a very high mean. This is because there were only 17 individuals there out of which 1 person had been assigned 45 secondary infections. Similarly, the “Pharmaceutical Company Nanjangud” had one person among 22 parents who was assigned 25 secondary infections.

- *Variation with Age* We consider the age distribution across “The 8 clusters” in Fig. 8a. It is seen that the distribution of the coronavirus patients has a higher fraction of patients in the age group 25 and above, whereas not too many in the range $0-25$, when compared to the actual demographic distribution in Karnataka. A possible reason for this might be that many cases were restricted to travel of working professionals. The state also took steps quite early on to lock down schools and universities to prevent the younger segment of the population from being affected. The patients in the age group of $0-15$ are either primary or secondary contacts of someone in their respective cluster.

  In Fig. 6, we consider a heat map of ages across “The 8 clusters”. Patients below the age of 10 and those whose ages are greater than 90 have very very low mean of number of secondary infections caused. Most secondary infections are caused by middle aged people who are the most socially active ones. For both “SARI” and “ILI” the age group $70-90$ have higher means. This could be because of care takers and close family contracting the infection before the patient tested positive. The “TJ Congregation” has a higher mean across all groups from $10-80$ and the “Pharmaceutical Company Nanjangud” has similar features. This is perhaps due to the fact that “TJ Congregation” cluster arose from a meeting in Delhi and the “Pharmaceutical Company Nanjangud” consisted solely of company employees and their contacts.
- *Deceased and Recovery Rates:* Table 2 contains the observed recovery and deceased rates of patients in each of “The 8 clusters”. It can be seen in Table 2 that the recovery rates are much higher than the deceased rates. The “Pharmaceutical Company Nanjangud” cluster had no one above the age of 70 and consequently perhaps has highest rate of recovery with 0 deaths. The highest deceased rate is seen in the “SARI” cluster where the deceased rate is around 17%. The “ILI” cluster also has a higher death rate than the remaining clusters. This again is perhaps due to the fact that the parents in “SARI”/“ILI” cluster had higher viral load. The remaining clusters have death rates between 1–5%.

  Before 26th June, 95% of the cases and 59% of the deaths occurred in individuals less than 65 years old. Case fatality rate is 2.414%. Among the patients in Karnataka who were deceased, 66% did so before they tested positive. Among the deceased patients who tested positive while hospitalized, the median number of days before they passed away was 3. The highest number of days a patient was treated before passing away was 36. From the detailed information on deceased patents, it is also known that around 70% of them had comorbidities.

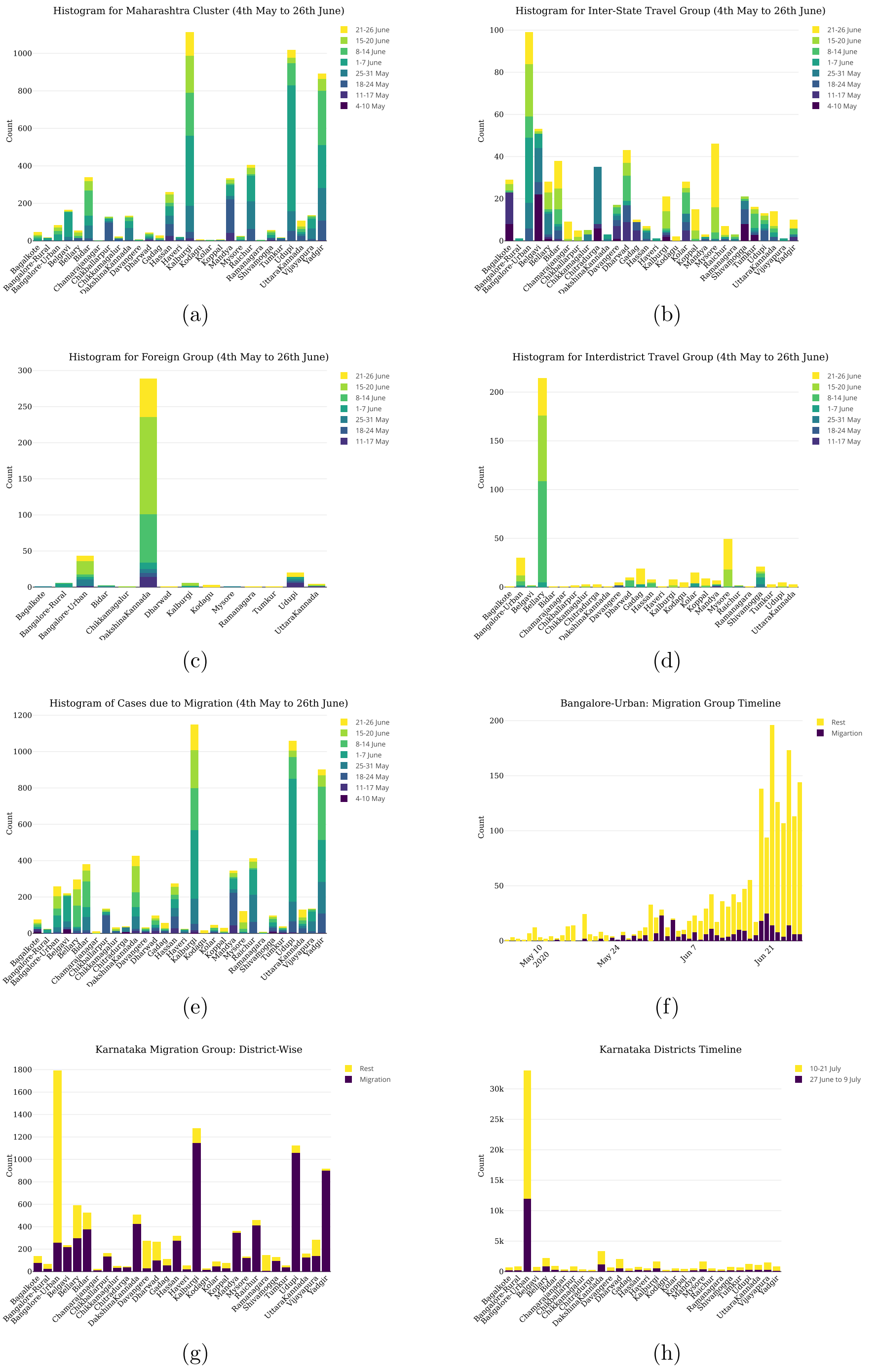
We also plot the days to recover (in Fig. 8c) and days to decease (in Fig. 8e) among patients who tested positive before 26th June belonging to “The 8 clusters”. It is seen that many patients who have passed away, do so on Day 0. This is because their samples, which result in positives, were sent for testing after their passing. It can be observed that the bulk of the deceased patients are between 45 – 75 years. There does not seem to be any observable correlation between days to recover and age. We caution against making significant inferences from this graph as the “recovery policy” has changed with time (See for e.g. 1st April and 8th May Guidelines).

- *Variation across Districts:* In Fig. 6 we have plotted a heat map of the mean number of secondary infections in each week for the different districts. This provides a framework for the time evolution of the reproduction number across districts as done in compartmental models. Most districts in Karnataka started having their COVID-19 cases quite late, during early May. Bangalore-Rural, Bellary, Davangere, Dharwad, Karwar, Kodagu, Tumkur and Udupi have several weeks where no one tested positive, as earlier outbreaks were well contained. The Pharmaceutical in Nanjangud is in the Mysore district and the end of the outbreak is visible. In Davangere District the week 27th-April to 3rd-May has a large mean because of a patient who was infected on 29th-April and had been assigned 30 secondary infections and one on 30th-April who was assigned 18 secondary infections. This is typical when there is a large mean. Most of the cases in the districts have low mean number of secondary infections in May. This was mainly due to the fact that those tested positive in this month had migrated from other states and caused very few recorded secondary infections.

Migration in Phase 3, 4 and Unlockdown 1.0

- *Variation across Districts:* The “From Maharashtra” group (or cluster) affected the districts of Kalaburagi (1113 cases), Udupi (1020 cases) and Yadgir (891 cases) the most. Bangalore-Urban received only 85 cases from Maharashtra (See Fig. 7a). The “Inter-State Travel” group affected Bangalore-Urban and Mysore the most, though the absolute numbers were very low at 99 and 46 cases respectively (See Fig. 7b). The “Foreign group” contributed 379 cases (via airports in Bangalore and Mangalore) with 76% were detected in the Dakshina Kannada District (See Fig. 7c) and 43 of them were assigned to Bangalore-Urban district. The “Inter-District Travel” group affected Ballari the most, with 214 cases. Mysore and Bangalore-Urban ranked next, but their absolute counts were quite low at 49 and 30 respectively (See Fig. 7d).

  As seen from above, overall, Kalaburagi, Udupi and Yadgir were the worst affected districts. The three districts together received 45.2% of all infections due to Migration and Bangalore-Urban received 257 cases due to migration (See Fig. 7e).



(a)

(b)

(c)

(d)

(e)

(f)

(g)

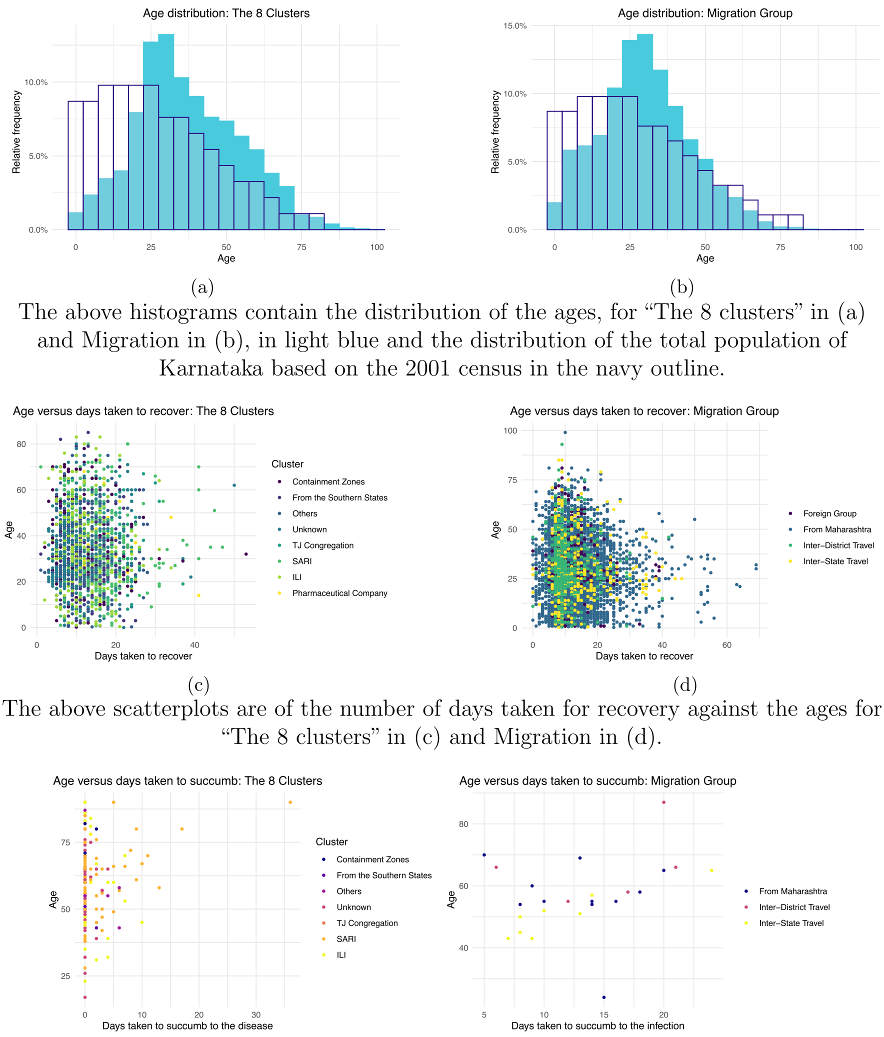
(h)

Fig. 7 The first seven graphs represent the data in the migration period from 4th May to 26th June. The last graph has number of cases from 27th June to 21st July

Table 3 In this table, the first column represents (cases due to migration)/(total cases in district in the period 4th May to 26th June) as a percentage. The second and third column give absolute counts of cases in each district from 27th June to 9th July and from 10th July to 21st July respectively

Name of district	Percentage of migration group (4th May- 26th June)	Number of cases (27th June- 9th July)	Number of cases (10th July- 21st July)
Bagalkote	57.04	188	459
Bangalore-Rural	37.88	266	601
Bangalore-Urban	14.34	11947	21061
Belgavi	92.83	133	646
Bellary	50.42	886	1367
Bidar	71.86	356	575
Chamarajanagar	55	112	232
Chikballarpur	82.82	150	649
Chikkamagalur	70.21	82	274
Chitradurga	82.61	46	175
DakshinaKannada	84.19	1182	2128
Davangere	11.36	140	552
Dharwad	36.74	558	1495
Gadag	52.25	127	387
Hassan	86.48	300	445
Haveri	41.07	175	302
Kalburgi	89.83	570	1064
Kodagu	58.62	66	185
Kolar	51.69	150	390
Koppal	39.24	77	401
Mandya	95.29	242	293
Mysore	91.04	416	1270
Raichur	90.17	174	372
Ramanagara	5.41	184	184
Shivamogga	75.97	243	500
Tumkur	69.09	259	426
Udupi	94.3	317	963
UttaraKannada	80.12	339	760
Vijayapura	48.58	288	1137
Yadgir	98.36	127	670

- *Migration versus Total:* In Table 3 we compute the percentage of cases due to the migration group across districts during this period. We observe that Dakshina Kannada (84%), Kalaburagi (89.8%), Mandya (95.3%), Raichur (90%), Udupi (94.3%) and Yadgir(98%) had very large proportion of their total cases due to the migration group. In contrast, in Bangalore-Urban migration accounted for 14.3% of the total cases (See Fig. 7f).



(a) (b)

The above histograms contain the distribution of the ages, for “The 8 clusters” in (a) and Migration in (b), in light blue and the distribution of the total population of Karnataka based on the 2001 census in the navy outline.

(c) (d)

The above scatterplots are of the number of days taken for recovery against the ages for “The 8 clusters” in (c) and Migration in (d).

(e) (f)

The above scatterplots are of the number of days taken for the patient to be deceased plotted against the ages for “The 8 clusters” in (e) and Migration in (f).

Fig. 8 Age distribution across clusters

- *Variation in Age, Recovery and Deceased:* The histogram of age distribution of the migration cases shows that the distribution is concentrated around 20-40 years as seen in Fig. 8b. There are also a higher proportion of cases having 0–20 age as compared to the histogram of all cases and a lesser proportion of elderly people as seen in Fig. 8a (which has the age distribution for the infected individuals belonging to the eight clusters studied earlier), indicating that most of the migrating individuals were families. This is probably because more children migrated along with parents, but very few elderly people did. Out of the

6871 migration cases, only 25 people have succumbed to the disease (as seen till 21*st* July). This is perhaps due to the fact that the elderly were in fewer proportion than in the 8 clusters that we analysed earlier. There were no casualities in the "Foreign group". All but one person who passed away were more than 40 years of age as seen in Fig. 8f. There was also a high recovery rate with 6657 people recovering (as seen till 21*st* July). Most people recovered within 20 days of testing positive as seen in Fig. 8d. Again we caution against making significant inferences from this graph as the "recovery policy" has changed with time.

Surge in July

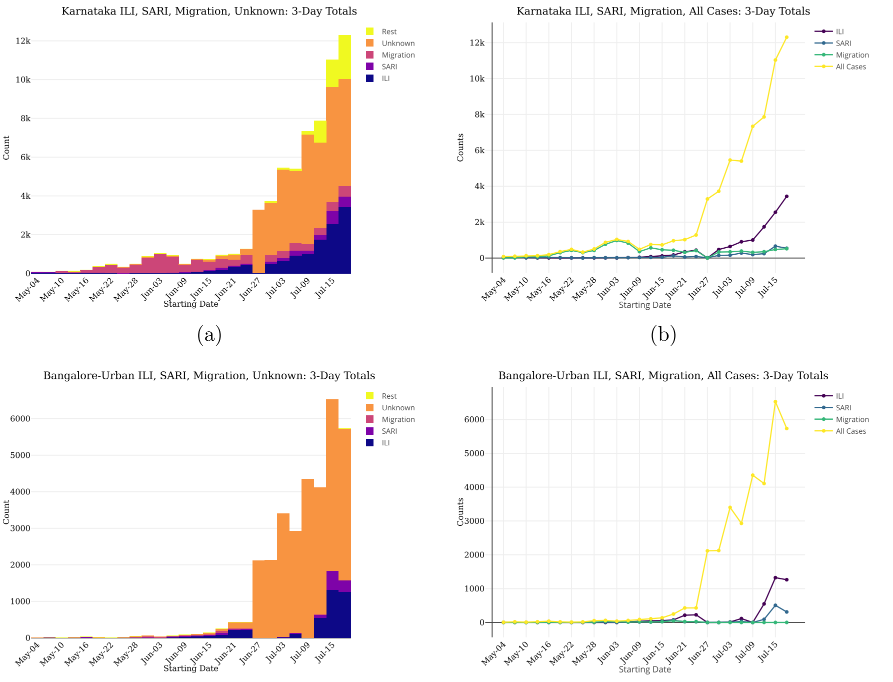
There was a sudden surge in cases in Karnataka after the migration period (4th May to 26th June). On 26th June the total cases in the state stood at 11005, which doubled on 9th July (31105 cases) and became four times on 21st July (71068 cases). We will try to outline the possible reasons for this surge.

- *ILI/SARI:* From middle of June, the "ILI" cluster cases in Karnataka have been increasing and there was a sharp rise in the first half of July. They also formed a significant proportion of total cases. In Bangalore-Urban district, the "ILI" cluster cases have been increasing since the middle of June, a sharp rise in the first half of July and also a significant proportion of total cases with over 50% on some days. The "SARI" cluster also shows an increase but the proportion fluctuates and is low, around 5%.
- *Variation across districts:* Bangalore-Urban accounted for approximately 50% of the surge in July with the count being 1953 on 26th June and rising to 34691 by 21st July. In Kalaburagi, there were 1339 cases on 26th June and 2973 cases on 21st July. In Udupi and Yadgir, the cases doubled from 26th June to 21st July (Udupi- 1126 cases on 26th June, 2406 cases on 21st July; Yadgir- 916 cases on 26th June, 1713 cases on 21st July).

4 Discussion

From 27th June to 29th June the media bulletins did not provide any description for the patients who tested positive and from 30th June onward the description was not as detailed as before. Post 21st July the media bulletins did not contain any individual information on those who tested positive. A disproportionately large number of cases were designated as contact under tracing and thus fell in the "Unknown" cluster (see Fig. 9a, c), making it impossible to proceed on a precise analysis for "The 8 clusters".

Our cluster classification was based on the trace history which is a measure of how contact tracing was done and how infected individuals are being identified for testing. It is important to note that the parent to child relationship in the trace history is indicative of the testing policy and contact tracing that was followed and need not be a definitive indicator of the genealogy of the infection spread. Among the four clusters where source of infection of the parents is not known ("Containment



(a)

(b)

(c)

(d)

Fig. 9 In the graphs on the left, we have considered the sum of daily counts of cases every three days-starting from 4th May and ending on 20th July for Karnataka (Fig. 9a and b)and Bangalore-Urban (Fig. 9c and d). The graphs are stacked histograms depicting the proportions of ILI, SARI, Migration, Unknown and the rest of the cases when taken as three day- sums. The bar representing 27th, 28th and 29th June is filled with Unknown cases entirely due to the absence of patient description. Hence the reader must note the trend instead of believing the absolute counts after this date. We see that the proportion of ILI and Unknown cases has been increasing in July. The counts of ILI reached 14,000 for Karnataka (4700 for Bangalore-Urban) towards the end, while the SARI counts reached 3000 for Karnataka (1400 for Bangalore-Urban). In the graphs on the right, we have again considered the sum of daily counts of cases every three days for Karnataka and Bangalore-Urban. For Karnataka, it can be seen that from 4th May to 14th June the proportion of Migration cases is quite high, but recently it has decreased. This is clearly not the case for Bangalore-Urban

Zone”, “SARI”, “ILI” and “Unknown” clusters) the mean of secondary infections in the first generation is highest for “SARI”, followed by “ILI”. This is because, the first generation in “SARI”/“ILI” cluster are those with a history of SARI/ILI, displaying symptoms and having a high viral load of the infection. Also “SARI”/ “ILI” clusters indicate some local transmission in the state making complete accounting of secondary infections via manual contact tracing a big challenge. We did notice that in the analysis of “The 8 clusters” that “SARI” and “ILI” clusters had R_{eff} less than 1 but there was a regular addition of new parents in these clusters. The continuing growth of these clusters indicates presence of viral load in the population. This could be due to one of many reasons. Patients in “SARI”, “ILI and “Unknown” clusters were not entirely contact traced or as their infection source was unknown, there were significantly many silent spreaders who did not fall into the contact tracing network.

Further, one could infer that severe restrictions by definition in the “Containment Zone” are proving effective with mean of secondary infections across all generations being less that one. Finally, the parents in the “Unknown” cluster presumably consist of patients, who at the time of testing, had mild symptoms or were asymptomatic patients (being part of random testing conducted routinely). If this is definitive, then one could conclude that the effective mean reproduction number for patients in this category is given by that of the “Unknown” cluster.

Another aspect to be considered is the Testing policy that was followed in May and June. There have been variations over time such as: non-uniform testing of the population across districts (e.g. testing only on migrants in Phase 3 and Phase 4 due to capacity constraints); and COVID-19 contact-workers [Health, Law and order, Sanitation] in earlier months were not being tested enough that they inadvertently were spreading the virus. The fraction of positive tests is around 6.77% on $21st$ July and it is the highest fraction recorded upto this period. On $15th$ May, 2020 it reached an all-time low of 0.7%. The number of total tests conducted up to $21st$ July is 1049982 which includes RAT, RT-PCR and other testing techniques. The details as to the amount of tests done using each technique was not mentioned before 17th July. These provide a comprehensive count of testing numbers in the state but not cluster-wise testing data. The number of infected individuals in the population differs from the number of positive test results. So equating the number of those tested positive to the number of infected individuals may be an error, because every individual in the population has not been tested. The State of Karnataka conducted a serological survey recently which provided insight on missed cases with a case to infection ratio of 1:40 (Babu Giridhara et al. 2021).

Finally, it seems unlikely that the Migration group in Phase 3, Phase 4 and Unlockdown 1.0 is the reason for the surge. We have already noted that the districts affected most by migration are Kalaburagi, Udupi and Yadgir (see Fig. 7g and above). From Fig. 9 we also note that the Migration group during the end of June and July did not account for a significant proportion of cases and the current surge was driven sharply by the cases in Bangalore-Urban district.

Appendix A Model

Let the random variable v represent the number of infections caused by a particular infected individual, called the individual infectiousness. We will model v coming from a probability distribution with mean R_{eff}. In particular we will assume v is Gamma distributed with mean R_0 and dispersion parameter k for some $k > 0$ and $Z \sim \text{Poisson}(v)$, allowing Z to represent the number of secondary infections caused by each infected individual. A standard calculation shows that for $z = 0, 1, 2, 3, \ldots$

[1] In the Epidemiological literature k is referred to as Dispersion and $k > 0$ is assumed, while in the Statistics literature $\frac{1}{k}$ is referred to as Dispersion given the connection with the Gamma distribution and is allowed to take negative values up to $-\frac{1}{R_{\text{eff}}}$.

$$P(Z = z) = \frac{\Gamma(k+z)}{z!\,\Gamma(k)} \left(\frac{k}{k+R_{\mathrm{eff}}}\right)^{k} \left(\frac{R_{\mathrm{eff}}}{k+R_{\mathrm{eff}}}\right)^{z}. \tag{A.1}$$

Thus one interprets Z as having Negative Binomial distribution with mean R_{eff} and Dispersion k.[1]It can also be seen that Z has variance $R_{\mathrm{eff}}\left(1 + \frac{R_{\mathrm{eff}}}{k}\right)$. Thus smaller values of k indicate larger variance. Depending on the heterogeneity different models can also be chosen. If one assumes $v = R_{\mathrm{eff}}$, then we are assuming a homogeneous population where each individual has the same infectiousness. This will imply $Z \sim \mathrm{Poisson}(R_{\mathrm{eff}})$ for $k = \infty$ and if we set $k = 1$ then $v \sim \mathrm{Exponential}(R_{\mathrm{eff}})$, (which arises from mean field models assuming uniform infection and recovery rates), and this implies $Z \sim \mathrm{Geometric}(R_{\mathrm{eff}})$.

Appendix B Maximum Likelihood Estimate

Given Data $\mathbf{y} := \{y_i\}_{i=0}^{n}$, the log-likelihood (modulo constant terms) is

$$L(R_{\mathrm{eff}}, k \mid \mathbf{y}) = \sum_{i=1}^{n} \left[y_i \log(R_{\mathrm{eff}}) - (y_i - k)\log\left(1 + \frac{R_{\mathrm{eff}}}{k}\right) + \sum_{j=0}^{y_i - 1} \log\left(1 + \frac{j}{k}\right)\right].$$

We follow (Lloyd-Smith James et al. 2005) to estimate $c = \frac{1}{k}$. First we rewrite the (conventionally accepted) log-likelihood as a function of R_{eff} and $c = \frac{1}{k}$.

$$L(R_{\mathrm{eff}}, c \mid \mathbf{y}) = \sum_{i=1}^{n} \left[y_i \log(R_{\mathrm{eff}}) - \left(y_i - \frac{1}{c}\right)\log(1 + R_{\mathrm{eff}}c) + \sum_{j=0}^{y_i - 1} \log\left(1 + cj\right)\right].$$

It is then standard (See Saha and Paul Sudhir 2005) that the Maximum Likelihood Estimator for R_{eff} is the sample mean, i.e. $R_{\mathrm{eff}} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and Maximum Likelihood Estimator for c is a solution to

$$\sum_{i=1}^{n} \left[\frac{1}{c^2}\log(1 + cR_{\mathrm{eff}}) - \frac{y_1 - R_{\mathrm{eff}}}{c(1 + cR_{\mathrm{eff}})} - \sum_{j=0}^{y_i - 1} \frac{1}{c(1 + cj)}\right] = 0. \tag{B.1}$$

Using (B.1) it is not possible to solve for c explicitly. A numerical approximation scheme is used to obtain an approximate value of c. We use the `uniroot` function in `R`.

Appendix C χ^2-goodness of fit test

Given Data $\mathbf{y} := \{y_i\}_{i=0}^{n}$. Let $\hat{R_0}$ and dispersion $\hat{k}$ be Maximum likelihood estimators. To see if Negative Binomial with mean $\hat{R_0}$ and dispersion $\hat{k}$ is a good fit for the data $\mathbf{y}$ we shall perform the χ^2-goodness of fit test. We will consider the range to $\{0, 1, \ldots, B\}$ with $B = \min\{n+1, 20\}$. Let $y_1, y_2, \ldots, y_n$ be the offspring data from a given cluster and let

$$p_j = \begin{cases} P(Z = j) & \text{for } 0 \leq j \leq B - 1, \\ P(Z \geq B) & \text{for } j = B. \end{cases}$$

and

$$Z_j = \begin{cases} \#\{k : y_k = j\} & \text{for } 0 \leq j \leq B - 1, \\ \#\{k : y_k \geq B\} & \text{for } j = B. \end{cases}$$

Then consider the statistic

$$\mathbf{X}^2 := \sum_{j=0}^{B} \frac{(Z_j - np_j)^2}{np_j} \equiv \sum_{j=0}^{B} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}.$$

As we have estimated two parameters, it is known that $\mathbf{X}^2$- has χ^2_{B-2} degrees of freedom, asymptotically as $n \to \infty$. One way to test if Z is the correct fit for the cluster is to compute the

$$\text{p-value} := \mathbb{P}(\chi^2_{B-2} \geq X^2).$$

There is strong evidence against the possibility that data arose from that model if *p*-value is very small.

Appendix D Confidence Intervals

To compute the confidence interval for the Negative binomial dispersion parameter k, we compute it for its reciprocal c and then invert it. We noted earlier that the maximum likelihood estimate for c had to be solved numerically and it is known that the asymptotic sampling variance is given by a series expansion (See Saha and Paul Sudhir 2005). Let $\hat{c}$ and $\hat{R_0}$ be the M.L.E. obtained. Then let

$$b = \frac{\hat{c}\hat{R_0}}{1 + \hat{c}\hat{R_0}} \text{ and } d_i = \prod_{j=0}^{i} (1 + j\hat{c}).$$

Then the variance of $\hat{c}$ is given by

$$\sigma^2(\hat{c}) = \left(\frac{n}{\hat{c}^4} \sum_{i=1}^{\infty} \frac{i!(\hat{c}b)^{i+1}}{(i+1)d_i} \right)^{-1}. \tag{D.1}$$

The 95% confidence interval for c is then given by

$$(\hat{c} - z_{0.95}\sigma^2(\hat{c}), \hat{c} + z_{0.95}\sigma^2(\hat{c})),$$

with $z_{0.95}$ being the 95*th* percentile of the standard normal distribution. The 95% confidence interval for k is then given by

$$\left(\frac{1}{\hat{c}+z_{0.95}\sigma^2(\hat{c})},\frac{1}{\hat{c}-z_{0.95}\sigma^2(\hat{c})}\right).$$

Note that the above interval will not be symmetric around k due to the inversion. For the computation of Variance in (D.1) we use a tolerance of 10^{-10}.

Acknowledgements We would like to thank Gautam Menon for introducing us to the question on Dispersion and also for pointing us to Lloyd-Smith James et al. (2005). We would like to thank Rajesh Sundaresan for providing us detailed suggestions that improved the presentation of the paper. Further we would like to thank Biswadeep Karmakar, P. Shankar, Deepayan Sarkar and Mohan Delampady for feedback and discussions. We also thank the anonymous referee for a careful reading of the paper.

References

Babu Giridhara R, Rajesh S, Siva A, Jawaid A, Kumar PP, Maroor PS, Rajagopal PM, Lalitha R, Mohammed S, Lalitha K et al (2021) The burden of active infection and anti-SARS-COV-2 IGg antibodies in the general population: results from a statewide sentinel-based population survey in karnataka, india. Int J Infect Dis 108:27–36

Covid-19 India-timeline an understanding across states and union territories (2020). http://www.isibang.ac.in/~athreya/incovid19. Accessed Mar 2020

Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L, Parker M, Bonsall D, Fraser C (2020) Quantifying SARS-COV-2 transmission suggests epidemic control with digital contact tracing. Science 368:6491

Joel H, Sam A, Amy G, Bosse NI, Jarvis CI, Russell TW, Munday JD, Kucharski AJ, John EW, Fiona S et al (2020) Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health 8(4):e488–e496

Lipsitch M, Cohen T, Cooper B, Robins James M, Stefan M, James L, Gopalakrishna G, Chew Suok K, Tan Chorh C, Samore Matthew H et al (2003) Transmission dynamics and control of severe acute respiratory syndrome. Science 300(5627):1966–1970

Lloyd-Smith James O, Schreiber Sebastian J, Ekkehard KP, Getz Wayne M (2005) Superspreading and the effect of individual variation on disease emergence. Nature 438(7066):355–359

Novel Coronavirus (COVID-19) (2020) Media bulletin, Government of Karnataka, Department of Health and Family Welfare, Bengaluru. https://covid19.karnataka.gov.in/govt_bulletin/en. Accessed Mar 2020

Ramanan L, Brian W, Reddy DS, Gopal K, Neelima S, Jawahar RKS, Radhakrishnan J, Lewnard JA et al (2020) Epidemiology and transmission dynamics of COVID-19 in two indian states. Science 370(6517):691–697

Saha KK, Paul Sudhir R (2005) Bias-corrected maximum likelihood estimator of the intraclass correlation parameter for binary data. Stat Med 24(22):3497–3512

Steven R, Christophe F, Donnelly Christl A, Ghani Azra C, Abu-Raddad Laith J, Hedley Anthony J, Leung Gabriel M, Lai-Ming Ho, Tai-Hing L, Thach Thuan Q et al (2003) Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science 300(5627):1961–1966