

1 **Into the thirteenth Month: A Case Study on the Outbreak Analytics and Modeling the spread of**
2 **SARS-CoV-2 Infection in Pune City, India**

3 Joy Monteiro^{1,2*}, Bhalchandra Pujari^{2,3}, Sarika Maitra Bhattacharrya⁴, Anu Raghunathan⁴, Ashwini
4 Keskar², Arsh Shaikh¹, Prasad Bogam⁵, Shweta Kadu⁶, Nikita Raut⁶, Devendra Vavale³, Rupa Mishra²,
5 Ajit Kembhavi^{2,7}, L.S. Shashidhara^{1,2,8*} and Vidya Mave^{2,5*}

6 ¹ Indian Institute of Science Education and Research Pune, Pune, India

7 ²Pune Knowledge Cluster, Pune, India

8 ³ Centre for Modeling and Simulation, Savitribai Phule Pune University, Pune, India

9 ⁴ CSIR-National Chemical Laboratory, Pune, India

10 ⁵ Johns Hopkins India, Pune, India

11 ⁶ Abasaheb Garware College, Pune, India

12 ⁷ Interuniversity Centre for Astronomy and Astrophysics, Pune, India

13 ⁸ Ashoka University, Sonapat, India

14 Corresponding authors: Dr. Joy Merwin Monteiro, Dr. L.S. Shashidhara, Indian Institute of Science

15 Education and Research, Doctor Homi Bhabha Road, NCL, Colony, Pashan, Pune 411008

16 Email: joy@iiserpune.ac.in; ls.shashidhara@ashoka.edu.in

17 Text count: 3144

18 Abstract text count: 242

19 Keywords: COVID-19, collaboration, projection modeling, resource allocation, lessons learned

20 *Equal Contribution

21 **Abstract:**

22 With more than 140 million people infected globally and 3 million deaths, the COVID 19 pandemic has
23 left a lasting impact. A modern response to a pandemic of such proportions needs to focus on exploiting
24 all available data to inform the response in real-time and allow evidence-based decision-making. The
25 intermittent lockdowns in the last 13 months have created economic adversity to prevent anticipated
26 large-scale mortality and relax the lockdowns have been an attempt at recovering and balancing economic
27 needs and public health realities. This article is a comprehensive case study of the outbreak in the city
28 limits of Pune, Maharashtra, India, to understand the evolution of the disease and transmission dynamics
29 starting from the first case on March 9, 2020. A unique collaborative effort between the Pune Municipal
30 Corporation (PMC), a government entity, and the Pune knowledge Cluster (PKC) allowed us to layout a
31 context for outbreak response and intervention. We report here how access to granular data for a
32 metropolitan city with pockets of very high-density populations will help analyze, in real-time, the
33 dynamics of the pandemic and forecasts for better management and control of SARS-CoV-2. Outbreak
34 data analytics resulted in a real-time data visualization dashboard for accurate information dissemination
35 for public access on the epidemic's progress. As government agencies craft testing and vaccination
36 policies and implement intervention strategies to mitigate a second wave, our case study underscores the
37 criticality of data quality and analytics to decode community transmission of COVID-19.

38

39

40

41 INTRODUCTION

42 Following a cluster of pneumonia cases recorded in Wuhan, China, in December 2019, caused by a novel
43 coronavirus SARS-CoV-2 (1, 2), COVID-19 subsequently spread to other cities and countries (3, 4).
44 India tops the charts of people at risk of SARS-CoV-2 infection exceeding 1.4 billion cases. The first
45 documented case in India was from an Indian national evacuated from China on January 30, 2020. Pune, a
46 city with about 5 million population (5), recorded its first case on March 9, 2020, with a national flying
47 from Dubai. Severe public health measures were implemented in Pune and across India to prevent the
48 spreading of the outbreak, including a complete lockdown starting from March 25 and dragging on for a
49 little more than two months into the beginning of June 2020 (6). At the end of April 2021, Pune recorded
50 a staggering >400,000 cases (7). Surveillance and contact tracing are critical components of effective
51 public health response to COVID-19 as has been showcased worldwide, including two states of the Indian
52 Subcontinent, Tamil Nadu and Andhra Pradesh (6).

53 In response to public health crises such as this pandemic, healthcare preparedness, rapid action from local
54 authorities for procuring essential supplies, and appropriate resource allocation require projecting the
55 trajectory of cases for coming weeks and months. Real-time advanced data analytics and computational
56 modeling using granular data are central to such an exercise (8-11). The Pune Knowledge Cluster (PKC),
57 a consortium of Pune-based scientists, engineers, technologists, academicians, students that believed in
58 the power of data science analytics and modeling to curb the spread of the disease, collaborated with Pune
59 Municipal Corporation (PMC) to provide data-driven management of the pandemic. Here, we present a
60 perspective of the PMC-PKC collaborative efforts that include centralized data abstraction, curation,
61 analysis, data visualization, and modeling of the early COVID-19 data to project the pandemic curve that
62 directed policy decisions on resource procurements and allocation (12).

63

64

65 ANALYTIC APPROACH

66 Data Sources

67 Case data from the PMC included a description of reported cases gathered in flat files where
68 each row is a case and each column a recorded variable (e.g., dates of onset and admission, gender, age,
69 location). This data was extensively curated using a combination of manual input and state-of-the-art data
70 science tools to fulfill the definition of 'tidy data' in the data science community (13). Regional level
71 demographic information (example. age stratification, sex, residential addresses, maps of population
72 densities) was collected independently from the smart city public data portal to delineate the underlying
73 characteristics of the affected populations. Our primary data was obtained directly from the Pune
74 Municipal Corporation and official press releases as part of an effort to build an epidemiological
75 surveillance dashboard for Pune city (7). The details of the number of records and the duration for which
76 they are available is present in **Supplementary Table 1**.

77 Geocoding Methodology

78 Due to the unstructured nature of addresses in India, assigning an address to a geographical location is a
79 challenging task (14). To understand the spatial evolution of the pandemic within PMC limits, we
80 developed a machine learning model that processed addresses of each record and assigned each address to
81 a prabhag, an administrative unit in PMC. To train the model, we used a database of 48000 addresses
82 which experts had manually assigned in PMC to their respective prabhags. We used 80 percent of the data
83 for training and 20 percent for validation.

84 We first created a database of localities around Pune, which were then assigned to their respective
85 prabhags to eliminate the addresses' most "noisy" parts. We first simplified and normalized the addresses
86 using the following rules. The address was tokenized into individual words. All "small" words -- tokens
87 of length lesser than 3, were discarded. Minimal replacement of abbreviations was done. For example:

88 Soc -> Society, Apt -> Apartment etc. All non-alphabets are eliminated. Fuzzy matching (15) was used to
89 replace identified tokens (1 and/or 2-grams) with a standardized form. For example, yerwada, yerawada,
90 yerwada were all mapped to yerawada. Compound words were replaced by concatenated versions (e.g.,
91 Laxmi Nagar replaced by Laxminagar). These addresses were converted to a vector using a term
92 frequency vectorizer. To emphasize certain unique localities within a prabhag, their frequency was
93 multiplied by three. In the term frequency vectorization, we omit those tokens which rarely appear in the
94 dataset. The threshold was chosen based on when the number of tokens on either side of the threshold is
95 dramatically different. The threshold identified was 0.00003, which gives us a vocabulary of ~5000
96 tokens. If the threshold was changed to 0.00004 and higher, the size of the vocabulary changes
97 proportionally. However, if the threshold was decreased to 0.00002, the vocabulary more than doubles to
98 ~12000 tokens, indicating the presence of a large number of rare, "noisy" tokens.

99 **Classification and Validation**

100 The resulting address vectors were used to train a variety of classifiers. On examining their performance,
101 we observed that the multinomial Naive Bayes classifier provided better performance in certain prabhags
102 than tree-based classifiers (Decision Tree, Random Forest, XGBoost). Thus, an ensemble architecture was
103 chosen for the classifier. We used an ensemble of Multinomial Naive Bayes, XGBoost, and Random
104 Forest classifiers with a Decision Tree classifier acting as a meta-classifier. With this architecture, we
105 achieved an accuracy of 87%. Hyperparameters for Random Forest and XGBoost were chosen to reduce
106 overfitting. The accuracy on the training set was ~91.5%, indicating slight overfitting. We also performed
107 10-fold cross-validation, which showed minimal variability in the accuracy, suggesting that the model
108 was robust.

109 **Outbreak Analytics: An Overview**

110 Data type priorities were critically defined by what actionable information can be predicted for use in
111 ground-level decision-making during the pandemic. As seen worldwide, the resultant outputs of the

112 outbreak data analytics for Pune were limited due to operational constraints suggesting an urgent need for
113 resources and capacities to ensure data availability and quality. The aspects of outbreak analytics (**Figure**
114 **1**) allowed a systematic understanding of the situation in Pune. 'Outbreak analytics' in this article refers to
115 several tools and methods used to collect, curate, visualize, analyze, model Pune pandemic data. An
116 integrated workflow represented describing the tools and their cross-talk are summarized below.

117 **Tracking the epidemic curve**

118 The first type of graphics needed for rapid assessment of ongoing dynamics is the epidemic curve (epi
119 curve), which shows case incidence time series as a histogram of new-onset dates for a given time interval
120 (16). **Figure 2a** shows the trend of cumulative cases ward-wise in Pune and their outcomes. Cumulative
121 cases, the fraction of active cases, i.e., the ratio of active cases and total cases, is another suitable
122 parameter to understand the state of the epidemic. In the beginning, when all cases are active, it will start
123 from unity, and at the end of the epidemic, when all the cases are resolved, then this will become zero.
124 Thus, this curve can vary between one and zero, and if it shows a decline, it implies that the epidemic is
125 fading out. Note that when there are multiple waves, then this curve will be oscillatory.

126 **Geo-mapping of epidemics**

127 Maps have been at the core of infectious disease epidemiology from a very early stage. They can be
128 typically used to visualize the distribution of disease, represent the 'ecological niche' of infectious diseases
129 at large scales, assess an outbreak's spatial dynamics, and strategize interventions (17). The changing
130 scenario month-wise for Pune is delineated through a scale that allows understanding geographical spread
131 (**Figure 3**). The changing geography indicates the spread of the disease from a highly infected ward with
132 primary infection to adjacent wards.

133 **Trend in test positivity and transmissibility assessment.**

134 The identification of infected cases in the population is dependent on the testing for the disease. The
135 number of tests done in phase I of the epidemic was low and indicated that it was mainly used as a
136 diagnostic for symptomatic people. **Figure 2b** illustrates the test positivity rates, which declined over
137 time, indicating an increase in the number of tests. The 'transmissibility' of a disease can be quantified and
138 used to refer to the rate at which new cases arise in the population, resulting either in epidemic growth or
139 decline. Rather than an intrinsic property of a specific disease, transmissibility thus quantifies the
140 propagation of a pathogen in a given epidemic setting and is impacted by multiple factors, including
141 population demographics, mixing, and levels of pre-existing immunity. The first measure of
142 transmissibility is the growth rate (r), which is estimated from a simple model where case incidence is
143 either exponentially growing ($r > 0$) or declining ($r < 0$). Typically, r is estimated directly from epidemic
144 curves using a log-linear model, where r is defined as the slope of linear regression on log-transformed
145 incidence.

146 **Figure 2c** shows test and case doubling time throughout the pandemic so far. While r quantifies the speed
147 at which a disease spreads, it does not contain information on the intervention level necessary to control
148 the disease. This is better characterized by the reproduction number (generically noted 'R'), which
149 measures the average number of secondary cases caused by each primary case. Researchers typically
150 distinguish the basic reproduction number (R_0), which applies in a large, fully susceptible population,
151 without any control measures, from the effective reproduction number (R_{eff}), which is the number of
152 secondary cases after accounting for behavioral changes, interventions, and declines in susceptibility.
153 Different methodological approaches have been developed to estimate the reproduction number. To
154 estimate R_{eff} , we used the Bayesian approach developed by Bettencourt and Ribeiro, and modified by K.
155 Systrom (18, 19).

156 R can be approximated using estimates of the growth rate r combined with knowledge of the generation
157 time distribution. A more reliable method is to derive $R(t)$ from compartmental models. **Figure 2d**
158 represents the change in R_0 values ward-wise over time. The prevalence of the disease can be calculated

159 through understanding IFR as opposed to CFR. This can be calculated by serological testing to detect the
160 %age population infected. A serosurvey with five prabhags done in the July-August 2020 time frame is
161 discussed here in connection with the cases detected and case fatality ratios.

162 **Tracking Case Fatality Rates**

163 Mortality in infectious diseases has a lag defined by the time interval between case detection and
164 outcome. The outcome can be represented as death or recovery. **Figure 2e** shows the overall cumulative
165 deaths. The case fatality ratios go up during phase 1 and early part of phase 2 of the epidemic, while the
166 case fatality rates (CFR) are coming down with time. CFR calculations calculated from publicly available
167 data face lag between the date of admission of patients and the date of death. Accounting for this lag is a
168 major source of uncertainty in CFR calculations. Due to the availability of date of isolation in the data
169 made available by PMC, we have calculated CFR, which reduces this uncertainty significantly, and
170 allows for a cohort-based (i.e., all people who were isolated in the same period) calculation of CFR.
171 Calculating CFR based on the date of isolation also allowed us to compute a "real-time" CFR which
172 accounts for currently active cases.

173 **Dashboard and real-time visualization of data**

174 In order to disseminate information about the pandemic state, an online dashboard was created and
175 operated by the team depicting the trajectories of the epidemic (7). Dashboard depicts various quantities
176 and figures developed by the team members such as daily and active number of cases as well as tests,
177 real-time effective reproduction rate ($R(t)$), case doubling time, test doubling time, active fractions,
178 cumulative test positivity, CFR and the forecasting. Such quantities were made available for the entire
179 city; whenever possible, more granular data based on wards and sub-wards (Prabhags) was also shown.
180 An example plot of the dashboard is shown in **Figure 4**.

181 **Forecasting**

182 Forecasting plays a vital role in policymaking by offering reasonable estimates of the upcoming burden
183 on health infrastructure. We used the state-of-the-art INDSCI-SIM model, an extended SEIR model (20,
184 21) that employs a nine-compartmental description of an epidemic, tailored specifically to COVID-19.
185 Because INDSCI-SIM supports various non-pharmaceutical interventions, one can utilize the model to
186 incorporate the effects of multiple measures such as lockdown, adherence to masks, social distancing, and
187 quarantining. For forecasting, we chose the number of patients in critical condition for two primary
188 reasons: 1) this data is less prone to errors of underreporting as hospitals typically test all the critical care
189 patients for COVID-19 and 2) this number directly affects the resource requirements to be made by the
190 municipal corporation. Our first forecast was made in June 2020, which was subsequently updated after
191 the second lockdown in July. The resulting forecast was made available on an online dashboard until the
192 end of the year. The reported cases were added every day as shown in **Figure 5** that shows a worst-case
193 and a best-case scenario.

194 **How much should we test?**

195 The way to control the infection is to increase the number of tests. However, can we quantify how much
196 should be good enough? We proposed that if the time doubling value of the test is lower than that of the
197 time doubling the value of infection, it is possible to test a larger population beyond just symptomatic
198 people and their contacts. In the worst-case scenario, the two-time doubling values should be at least the
199 same. This was done when PMC ramped up their tests from about 700-800 per day at the end of May to
200 approximately 2500 in mid-June. Was that enough?

201 Like the time doubling of cases, we can define the time doubling of tests as

$$202 \quad T(t) = T_0 2^{t/td}$$

203 T_0 -- the cumulative number of tests on June 16 = 76206.

204 td -- time doubling value

205 $T(t)$ -- Cumulative number of tests on a day starting from June 16.

206 We projected values of tests required to bring down the test doubling time (td). Note that the case
207 doubling time during that period was varying between 22-28 days. The projected tests required for
208 keeping the td value at a certain number over the next 19 days are given in the table. These numbers were
209 much higher than the number of tests done per day. This projection gave an idea to PMC about the
210 requirement of an increase in the number of tests per day (**Supplementary Table 2**).

211

212 **Estimation of undetected cases and distribution of the test using CFR**

213 We asked here, if we have a limited number of tests how do we distribute across the wards? For this, we
214 take the help of the ward wise CFR

$$215 \quad U(\text{ward}) = D(\text{ward}) * \left(\frac{1}{IFR} - \frac{1}{CFR(\text{ward})} \right)$$

216 **U(ward)** = estimate of undetected cases in a ward

217 **D(ward)** = the deceased in that ward

218 **CFR(ward)** = The CFR calculated for that ward

219 **IFR** - can be equated to the CFR of a country or state where the epidemic is under control

220 Total Number of undetected cases

$$221 \quad U = \sum U(\text{ward})$$

$$222 \quad T(\text{ward}) = T * \frac{U(\text{ward})}{U}$$

223 **T** --- Total per day tests for Pune

224 **T(ward)**--Tests per day that should be assigned to a ward

225 **Supplementary Table 3** shows the estimate of the ward-wise undetected cases and the number of tests
226 required to keep the test positivity ratio - 0.2.

227 **Other Important Parameters**

228 **Test Positivity**

229 The test positivity (TP) is defined as the ratio between the total number of positive outcomes(P) and the
230 total number of tests (T) that may represent the testing efficacy. In an ideal case, as the number of tests is
231 increased, this parameter should decrease or remain constant at a very low value, suggesting uniform
232 testing and can indicate that the testing is neither targeted nor the disease is exploding undetected. High
233 values of TP could imply the testing is targeted; that is, most tests are done only for the symptomatic
234 patients and their contacts, enhancing the TP values artificially. A potential drawback of high TP numbers
235 is the distinct possibility of a lower number of tests performed and, more crucially, missing out on
236 potentially infective yet asymptomatic sections of the population, leading to a silent explosion of the
237 disease without detection. On the other hand, a small TP value could imply higher random testing or, in
238 the best cases, more expansive and successful testing of the population through intense contact tracing.
239 Given the large population and limited testing capabilities, a correct balance has to be struck between
240 these two approaches: intensive and focused only target testing and ineffective and unplanned random
241 testing. Thus, apart from CFR, we need to keep track of the TP. We plotted the ward-wise CFR
242 (**Supplementary Figure 1**). Depending on the high and low values of CFR and TP, the values can be in
243 the four-quadrant. We discuss below what each of the quadrants means and how the testing strategy can
244 be modified.

- 245 a) Low CFR and Low TP- Low TP implies that there is not much sampling error and that the
246 number of detected cases is close to the actual cases. When combined with low CFR values, it
247 can be deduced that wards that come under this category have implemented good testing
248 strategies covering most of the population. As long as the wards manage to stay in this quadrant,
249 they can be expected to bring down the number of cases using the same strategy.
- 250 b) High CFR and High TP- This quadrant implies that testing is too targeted, and thus the sampling
251 efficiency is poor. The actual number of cases is many-fold higher than the detected ones.
- 252 c) Low CFR and High TP- This combination suggests that though the testing is targeted, the number
253 of actual cases is likely to be close to the number of detected cases. The wards in this quadrant
254 might have mixed features in terms of rate and number of cases. However, the targeted testing

255 strategy seems to be working for these wards, and as long as they can keep the CFR low, they can
256 follow the current testing strategy. Increasing values of CFR for these wards are an early
257 signature of community spread. The ward should immediately ramp up the number of tests and
258 change the strategy to include random testing over a larger population.

259 d) High CFR and Low TPR: Though low values of TP alone can indicate a wider testing strategy
260 spread over a larger population when considering the high values of CFR, the conclusion about
261 the testing strategy becomes flawed. The actual cases are much larger for the ward belonging to
262 this most undesirable quadrant than the detected ones. The wards with these values of TP and
263 CFR need to ramp up their testing strategy by many folds and must include random testing over a
264 larger population.

265

266 **SUMMARY**

267 In this report, we summarize the Pune Knowledge Cluster consortium activities that included real-time
268 data visualization for identifications of hotspots, provision of a snapshot of outbreak trends, and
269 forecasting of the epidemic over the next several months (7, 22). Our case study on the outbreak analytics
270 and modeling the spread of SARS-CoV-2 infection in Pune city, India, provides a feasible and scalable
271 proof-of-concept to facilitate recommendations for public health policies to local officials and help
272 forecast outbreaks.

273

274

275 **ACKNOWLEDGEMENTS**

276 We would like to thank Rubal Agrawal, Shantanu Goel and Saurabh Rao (Pune Municipal Corporation),
277 Srinivasa Bharath Padavala (Pricewaterhouse Coopers), external advisors: Gautam Menon, Dibyendu
278 Nandi (IISER Kolkata), Sneha Bhogale (PKC), Hemant Joshi (TCS), Rahul Jagtap (PMC), Shivik Garg,
279 Shrutika Lokapure, Collins Assissi (IISER Pune).

280

281 **AUTHOR CONTRIBUTIONS**

282 JM & LS conceived the study and BP, JM, SB, AR, AK1, NR, DV performed outbreak analytics and
283 modelling. All authors BP, JM, SB, AR, AK1, AS, PB, SK, NR, DV, RM, AK2, LS & VM contributed in
284 respective section of the manuscript. All authors declare no competing interests

285

286 **Figure List**

287 **Figure 1:** Schematic representation of tools and cross talk in the integrated pipeline for outbreak data
288 analytics in Pune, India

289 **Figure 2:** a. Presents the trend of cumulative cases ward-wise in Pune (representative wards) b. Shows
290 the ward positivity for representative wards c. Demonstrates the test and case doubling time throughout
291 the pandemic d. Represents the City wise $R(t)$ e. Represents the overall cumulative deaths.

292 **Figure 3:** Prabhag-level incidence of COVID-19 (per 1000 persons) in Pune city for the months of May,
293 September, December 2020, and March 2021. The incidence shows a systematic shift from the densely
294 populated interior regions of the city towards the more sparsely populated exterior regions over time.

295 **Figure 4:** Dashboard - Example plot for Prabhag-wise cases per 1000 population

296 **Figure 5:** Modelling/Forecasting - Shows a worst case, a best-case scenario along with the likely path the
297 data may take (region shaded with blue).

298 **Supplementary Figure 1:** Prabhag-level Case Fatality Rates (in percent) in Pune city between February
299 1 - April 15, 2021. The relatively high CFR in some prabhags with low incidence suggests a higher
300 number of undetected cases (see text for details).

301 **Supplementary Table 1:** Summary of characteristics available from the data sources

302 **Supplementary Table 2:** Forecasting the number of SARS-CoV-2 tests for Pune city, based on the
303 targeted time doubling 'td' of tests.

304 **Supplementary Table 3:** Estimation of undetected cases in the different ward offices $U(\text{ward})$ in Pune
305 city based on the ward-wise CFR values and assuming the IFR to be 1%. The distribution of tests across
306 the wards using the information of the Total vs. ward wise undetected cases $T(\text{ward}) = T * \frac{U(\text{ward})}{U}$

307 **Supplementary Table 1.** Summary of characteristics available from the data sources

308

Type of data	Number of Records	Total Records (from press release)	Duration	Remarks
Testing	309742	521249	2020-03-01 to 2020-09-15	Missing records are primarily from private testing agencies
Patients	114669	115770	2020-03-01 to 2020-09-11	
Mortality	2683	2706	2020-03-01 to 2020-09-11	Collated from official press releases, includes those who were declared COVID-positive post-mortem
Contact Tracing	52292	N/A	2020-03-01 to 2020-06-12	

309

310

311

312 **Supplementary Table 2.** Forecasting the number of SARS-CoV-2 tests for Pune city, based on the
313 targeted time doubling 'td' of tests.

Date	td = 20	td =15	td = 10
6/17/2020	2782	3775	5862
6/18/2020	2880	3953	6283
6/19/2020	2982	4140	6734
6/20/2020	3087	4336	7217
6/21/2020	3195	4541	7735
6/22/2020	3309	4756	8290
6/23/2020	3425	4981	8885
6/24/2020	3546	5216	9523
6/25/2020	3671	5463	10207
6/26/2020	3800	5721	10939
6/27/2020	3934	5992	11724
6/28/2020	4073	6275	12566
6/29/2020	4216	6572	13468
6/30/2020	4365	6883	14434
7/1/2020	4519	7208	15470
7/2/2020	4679	7549	16581
7/3/2020	4844	7906	17771
7/4/2020	5015	8280	19046
7/5/2020	5192	8672	20413

314

315

316

317

318

319

320

321 **Supplementary Table 3.** Estimation of undetected cases in the different ward offices U(ward) in Pune
 322 city based on the ward wise CFR values and assuming the IFR to be 1%. The distribution of tests across
 323 the wards using the information of the Total vs. ward wise undetected cases $T(ward) = T * \frac{U(ward)}{U}$
 324

Ward Office	Undetected	Tests	
		Max	Min
AB	42	210	31
BP	1573	7867	1166
BI	711	3553	527
DS	339	1695	251
DP	222	1110	165
HM	344	1722	255
KV	624	3120	462
KY	169	845	125
KB	178	890	132
NRV	255	1275	189
OP	635	3177	471
SG	44	218	32
SR	230	1150	170
WR	713	3563	528
WK	67	335	50
YKD	1227	6136	909
Grand Total	7373	36865	5464

TPR = 0.2

Td =15

325

326

327

328

329

330

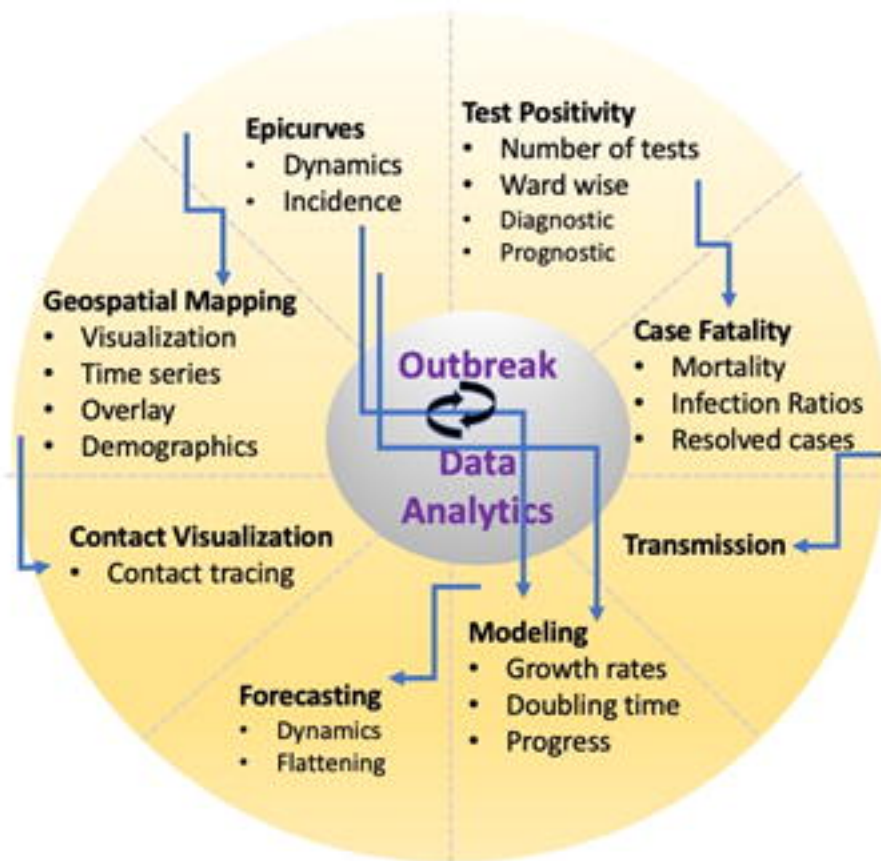
331

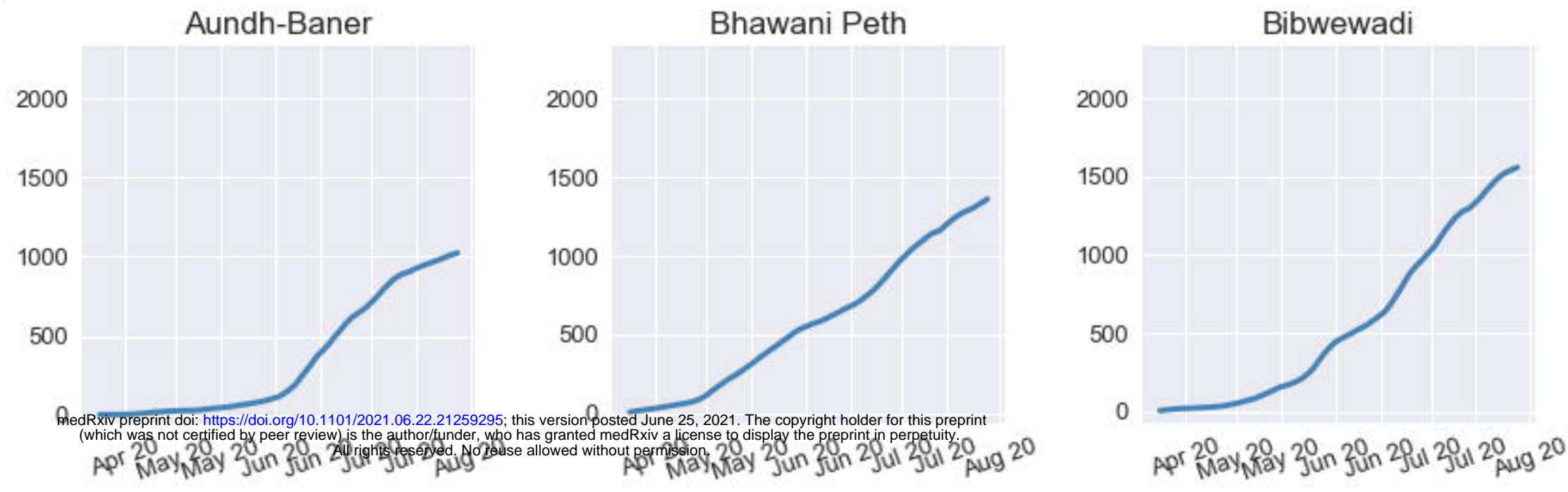
332

333 References

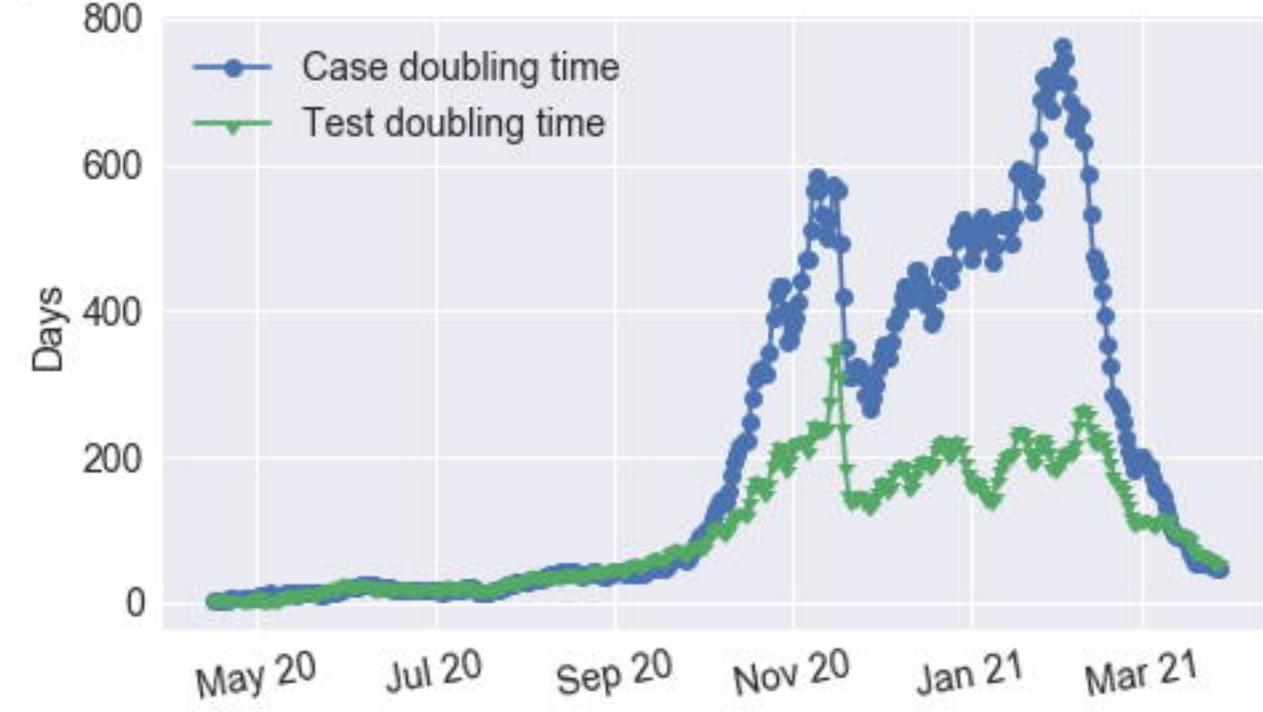
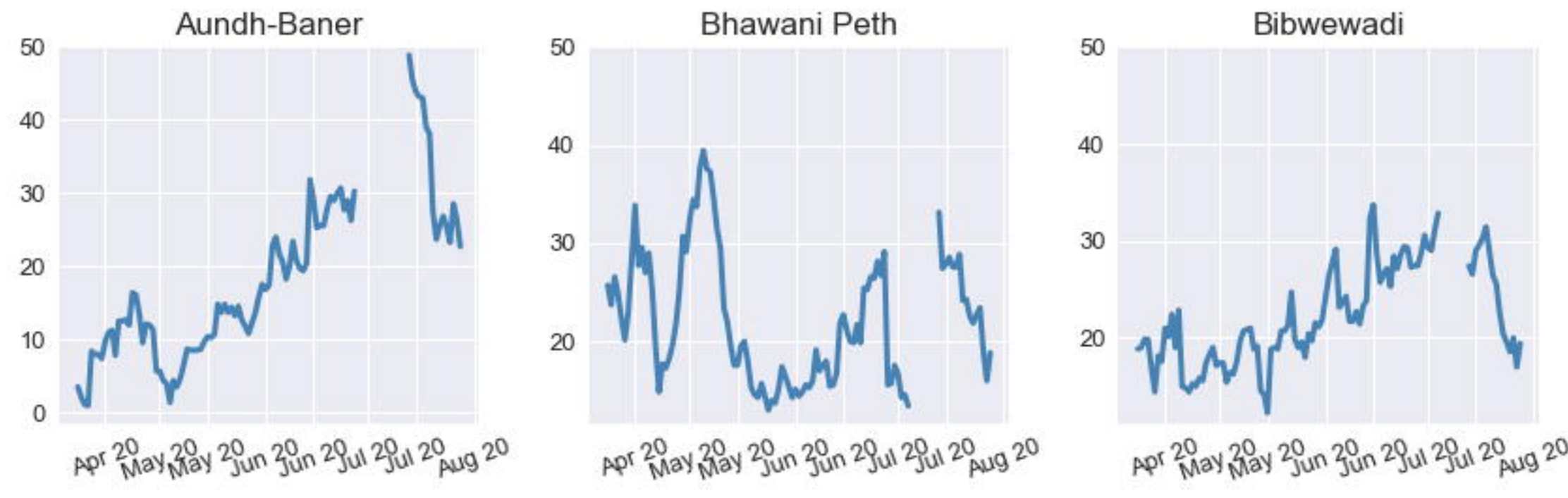
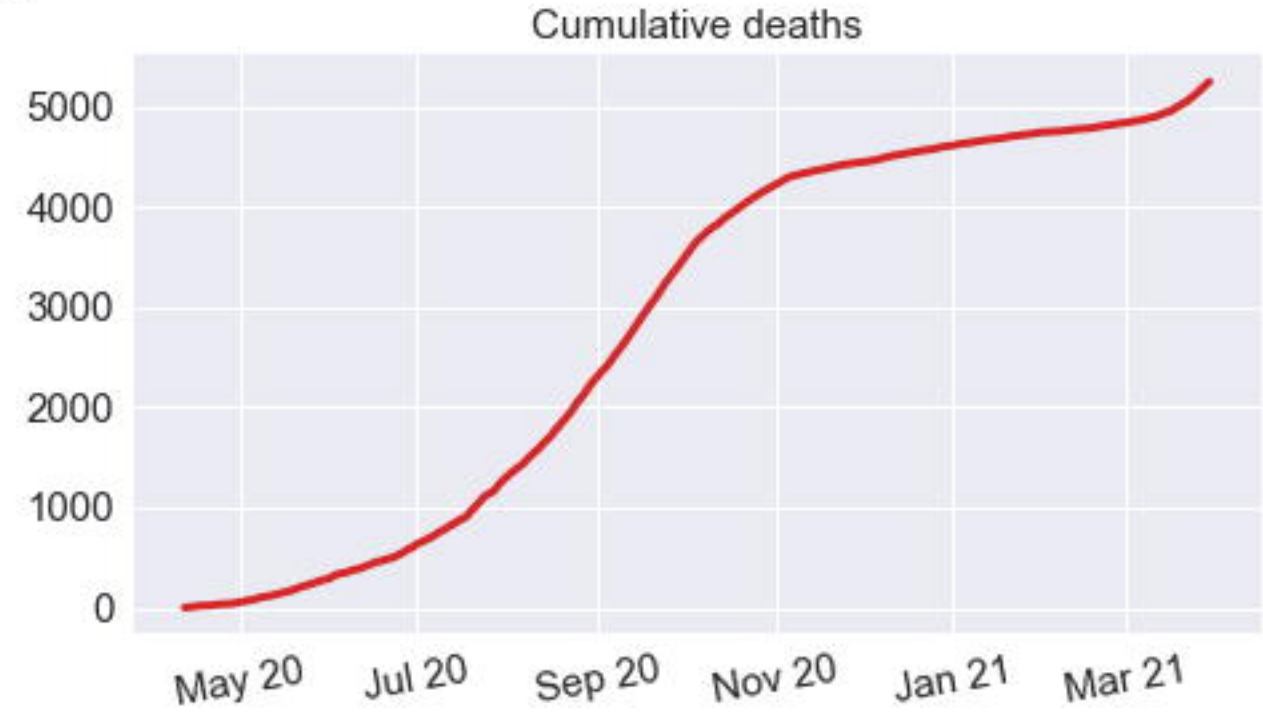
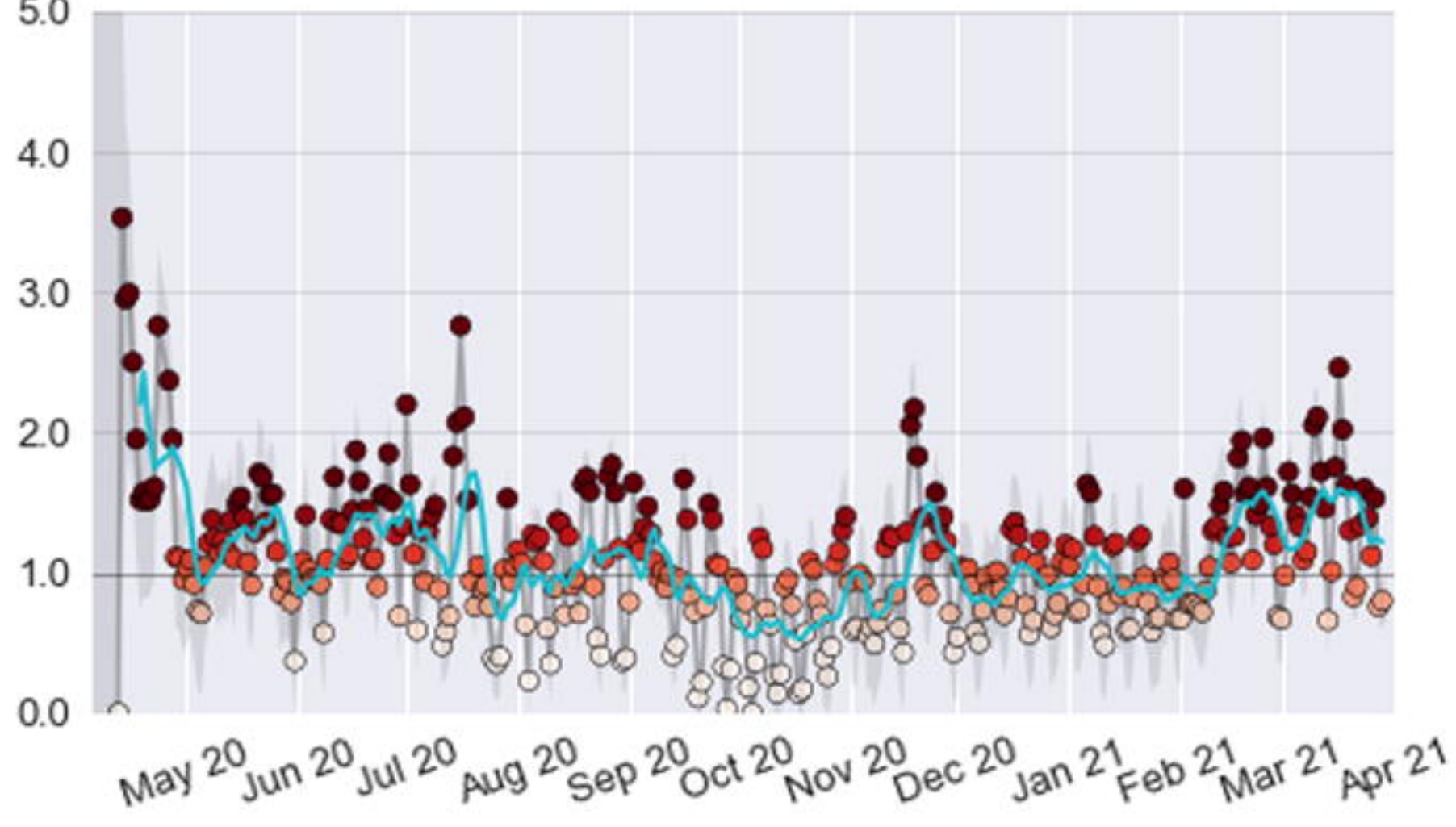
- 334 1. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for
335 mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study.
336 *Lancet*. 2020;395(10229):1054-62.
- 337 2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from
338 Patients with Pneumonia in China, 2019. *The New England journal of medicine*.
339 2020;382(8):727-33.
- 340 3. <https://coronavirus.jhu.edu/map.html> JHUaMC-DbtCfSSaECaJH. COVID-19 Dashboard
341 by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins. [Accessed April
342 15th 2021.]. <https://coronavirus.jhu.edu/map.html>.].
- 343 4. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in
344 real time. *The Lancet Infectious diseases*. 2020;20(5):533-4.
- 345 5. Census 2021. Pune City Population 2011 – 2021.
346 <https://www.census2011.co.in/census/city/375-pune.html>.
- 347 6. Laxminarayan R, Wahl B, Dudala SR, Gopal K, Mohan BC, Neelima S, et al.
348 Epidemiology and transmission dynamics of COVID-19 in two Indian states. *Science*.
349 2020;370(6517):691-7.
- 350 7. dashboard. PMCC-iP. <http://cms.unipune.ac.in/~bspujari/Covid19/Pune2/> [15 December
351 2020.].
- 352 8. Correction for modelling the COVID-19 pandemic in context: an international
353 participatory approach. *BMJ global health*. 2021;6(2).
- 354 9. Aguas R, White L, Hupert N, Shretta R, Pan-Ngum W, Celhay O, et al. Modelling the
355 COVID-19 pandemic in context: an international participatory approach. *BMJ global health*.
356 2020;5(12).
- 357 10. Mehta M, Julaiti J, Griffin P, Kumara S. Early Stage Machine Learning-Based Prediction
358 of US County Vulnerability to the COVID-19 Pandemic: Machine Learning Approach. *JMIR*
359 *public health and surveillance*. 2020;6(3):e19446.
- 360 11. Santosh KC. COVID-19 Prediction Models and Unexploited Data. *Journal of medical*
361 *systems*. 2020;44(9):170.
- 362 12. COVID-19 in Pune | DSCMS, PKC. Data Analysis and Forecasts of Covid-19 in Pune.
363 <http://cms.unipune.ac.in/%7Ebspujari/Covid19/Pune2/>.
- 364 13. science. Td.What's Tidy Data? How to organize messy datasets in Python with Melt and
365 Pivotal functions [https://towardsdatascience.com/whats-tidy-data-how-to-organize-messy-](https://towardsdatascience.com/whats-tidy-data-how-to-organize-messy-datasets-in-python-with-melt-and-pivotal-functions-5d52daa996c9)
366 [datasets-in-python-with-melt-and-pivotal-functions-5d52daa996c9](https://towardsdatascience.com/whats-tidy-data-how-to-organize-messy-datasets-in-python-with-melt-and-pivotal-functions-5d52daa996c9).
- 367 14. Chatterjee A JA, Sourav Roy, Arnab Ganguli, and Krishanu Seal “SAGEL: Smart
368 Address Geocoding Engine for Supply-Chain Logistics.” In Proceedings of the 24th ACM
369 SIGSPATIAL International Conference on Advances in Geographic Information Systems, 1–10.
370 SIGSPATIAL '16. New York, NY, USA: Association for Computing Machinery.
371 <https://doi.org/10.1145/2996913.2996917>. 2016.
- 372 15. GeeksforGeeks. FuzzyWuzzy Python library.
373 <https://www.geeksforgeeks.org/fuzzywuzzy-python-library/>.
- 374 16. Center for Disease Control. Interpretation of Epidemic (Epi) Curves during Ongoing
375 Outbreak Investigations. [https://www.cdc.gov/foodsafety/outbreaks/investigating-outbreaks/epi-](https://www.cdc.gov/foodsafety/outbreaks/investigating-outbreaks/epi-curves.html)
376 [curves.html](https://www.cdc.gov/foodsafety/outbreaks/investigating-outbreaks/epi-curves.html).
- 377 17. Moore DA, Carpenter TE. Spatial analytical methods and geographic information
378 systems: use in health research and epidemiology. *Epidemiologic reviews*. 1999;21(2):143-61.
- 379 18. K. Systrom, Estimating COVID-19's R_t in Real-Time. [https://github.com/k-sys/covid-](https://github.com/k-sys/covid-19/blob/master/Realtime%20R0.ipynb)
380 [19/blob/master/Realtime%20R0.ipynb](https://github.com/k-sys/covid-19/blob/master/Realtime%20R0.ipynb).
- 381 19. Bettencourt LM, Ribeiro RM. Real time bayesian estimation of the epidemic potential of
382 emerging infectious diseases. *PLoS one*. 2008;3(5):e2185.

- 383 20. Hazra et. al. "INDSCI-SIM A state-level epidemiological model for India", Private
384 Communication [2021] Note: <https://indscicov.in/indscisim>.
385 21. Pujari BS, Shekatkar S. Multi-city modeling of epidemics using spatial networks:
386 Application to 2019-nCov (COVID-19) coronavirus in India. medRxiv. 2020.
387 22. Mave V, Shaikh A, Monteiro JM, Bogam P, Pujari BS, Gupte N. Impact of National and
388 Regional Lockdowns on Growth of COVID-19 Cases in COVID-Hotspot City of Pune in Western
389 India: A Real-World Data Analysis. medRxiv. 2021.
390



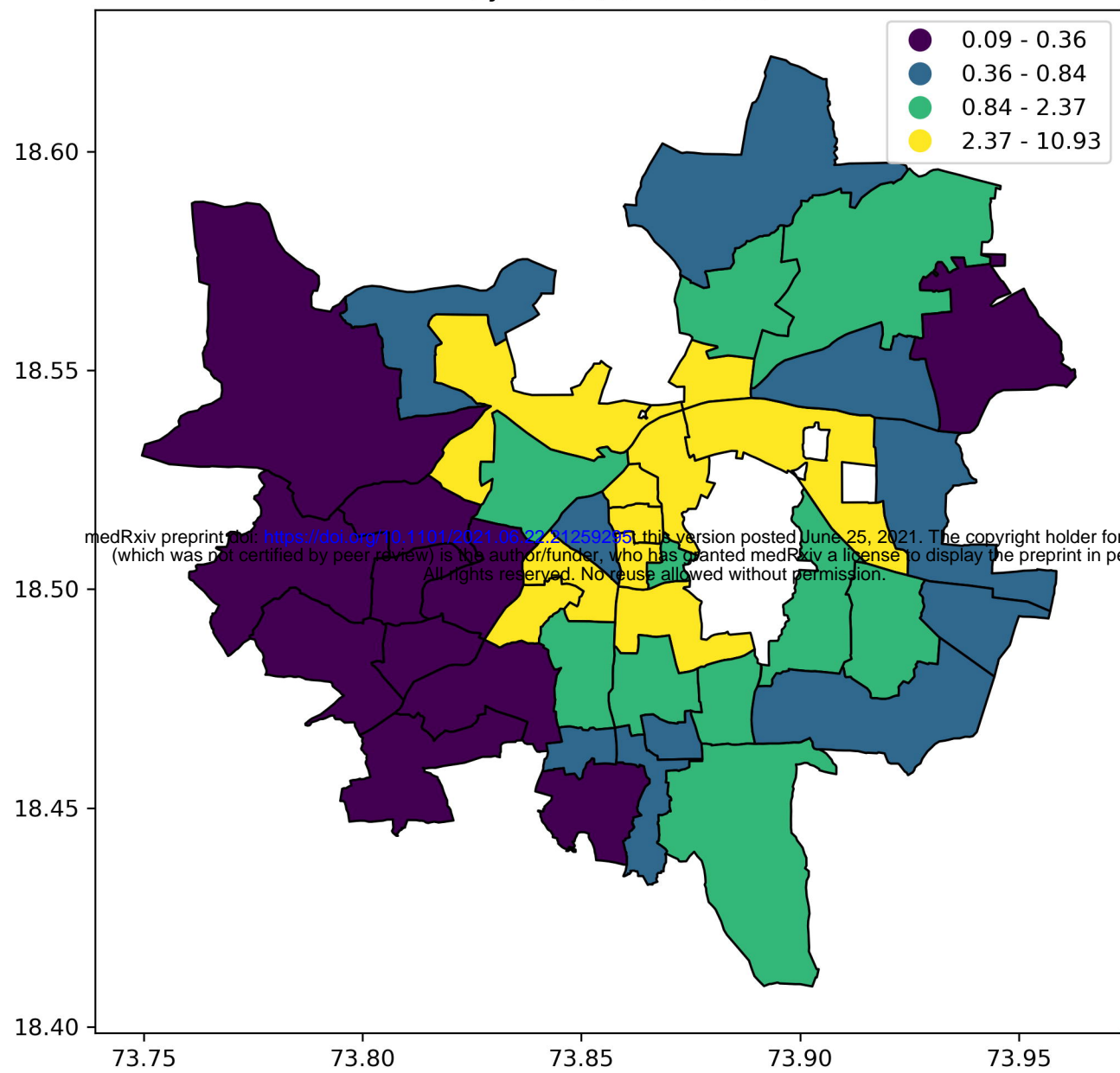
a

MedRxiv preprint doi: <https://doi.org/10.1101/2021.06.22.21259295>; this version posted June 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

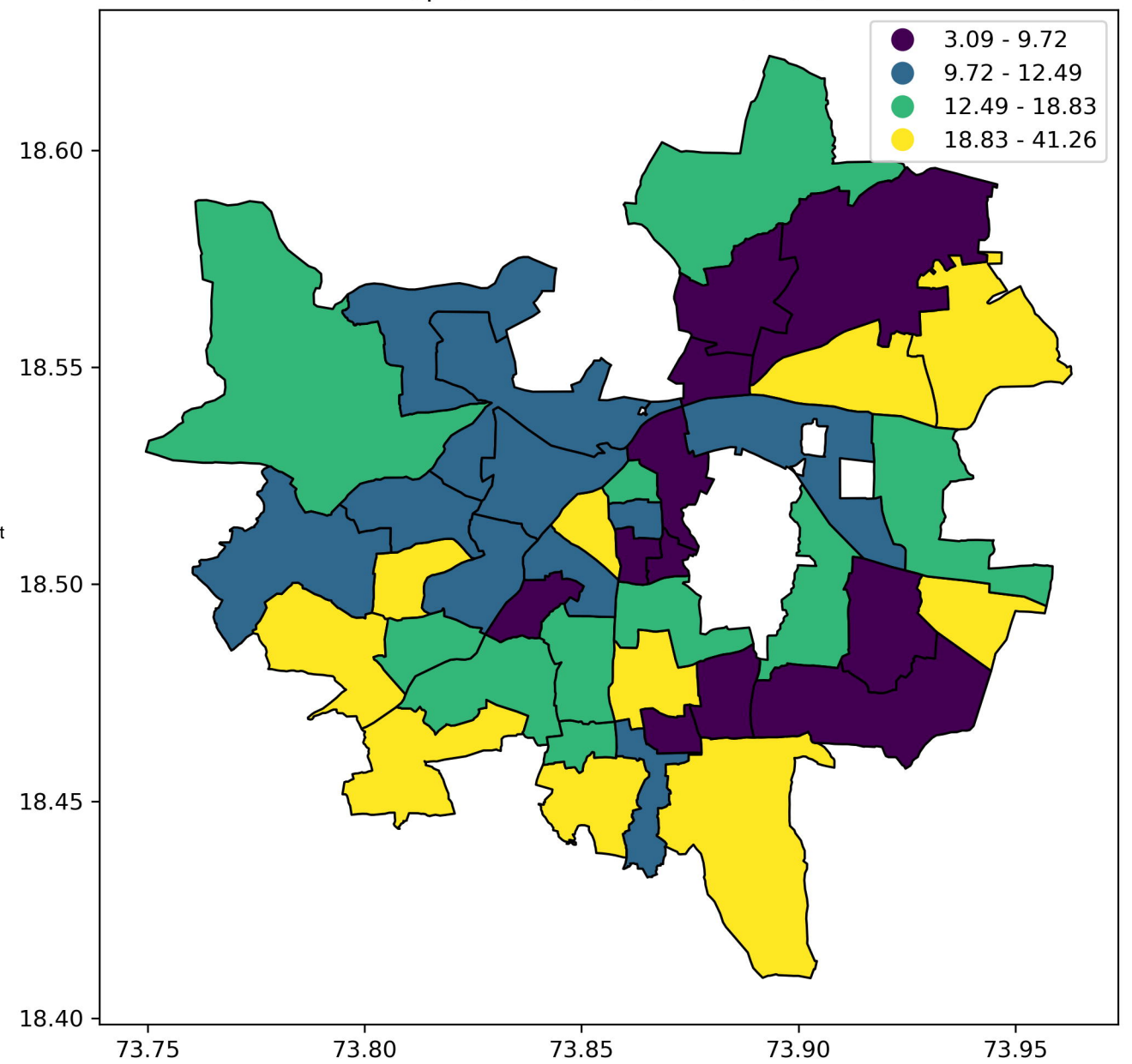
c**b****e****d**

a

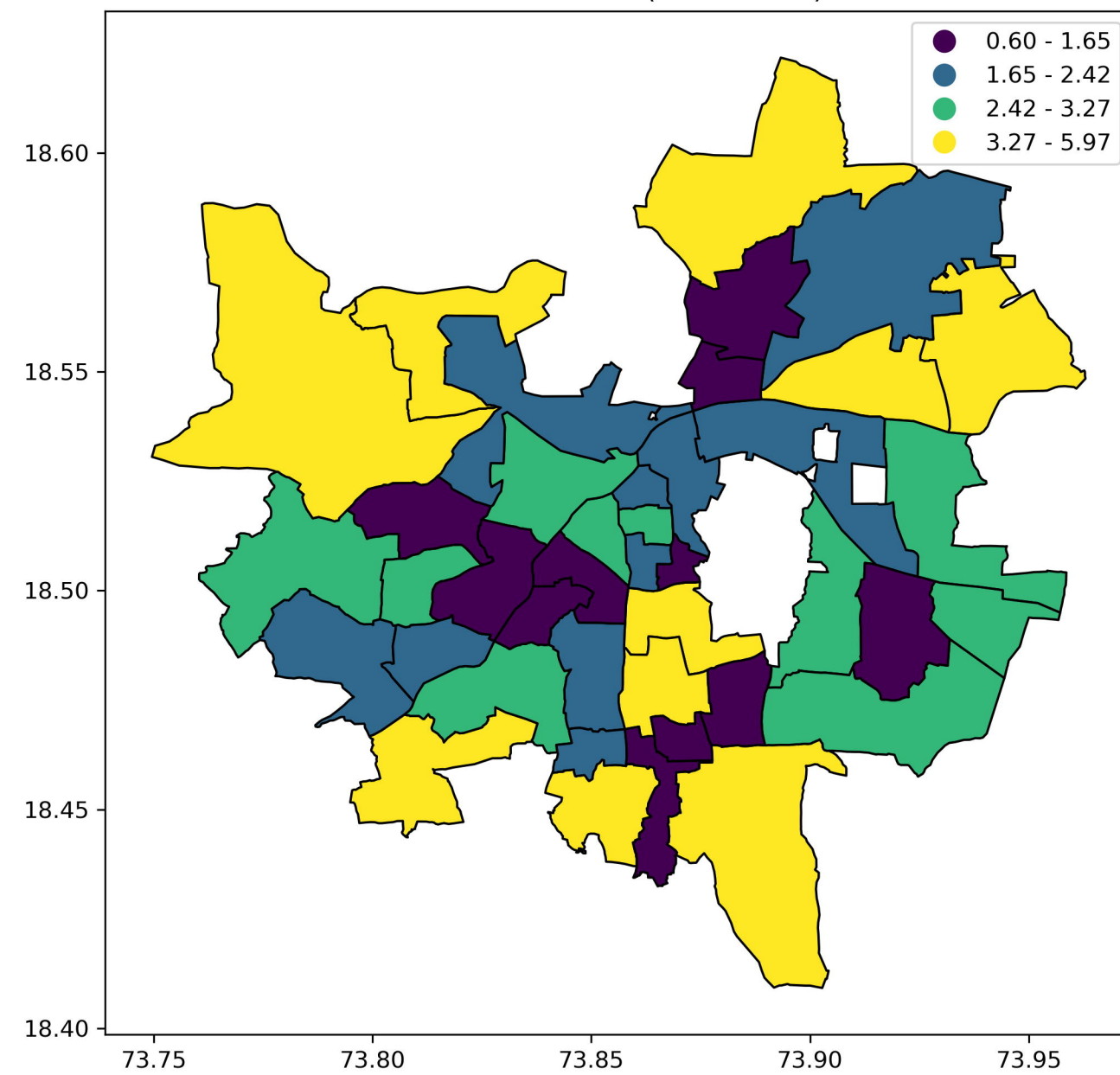
May 2020 (5363 cases)

**b**

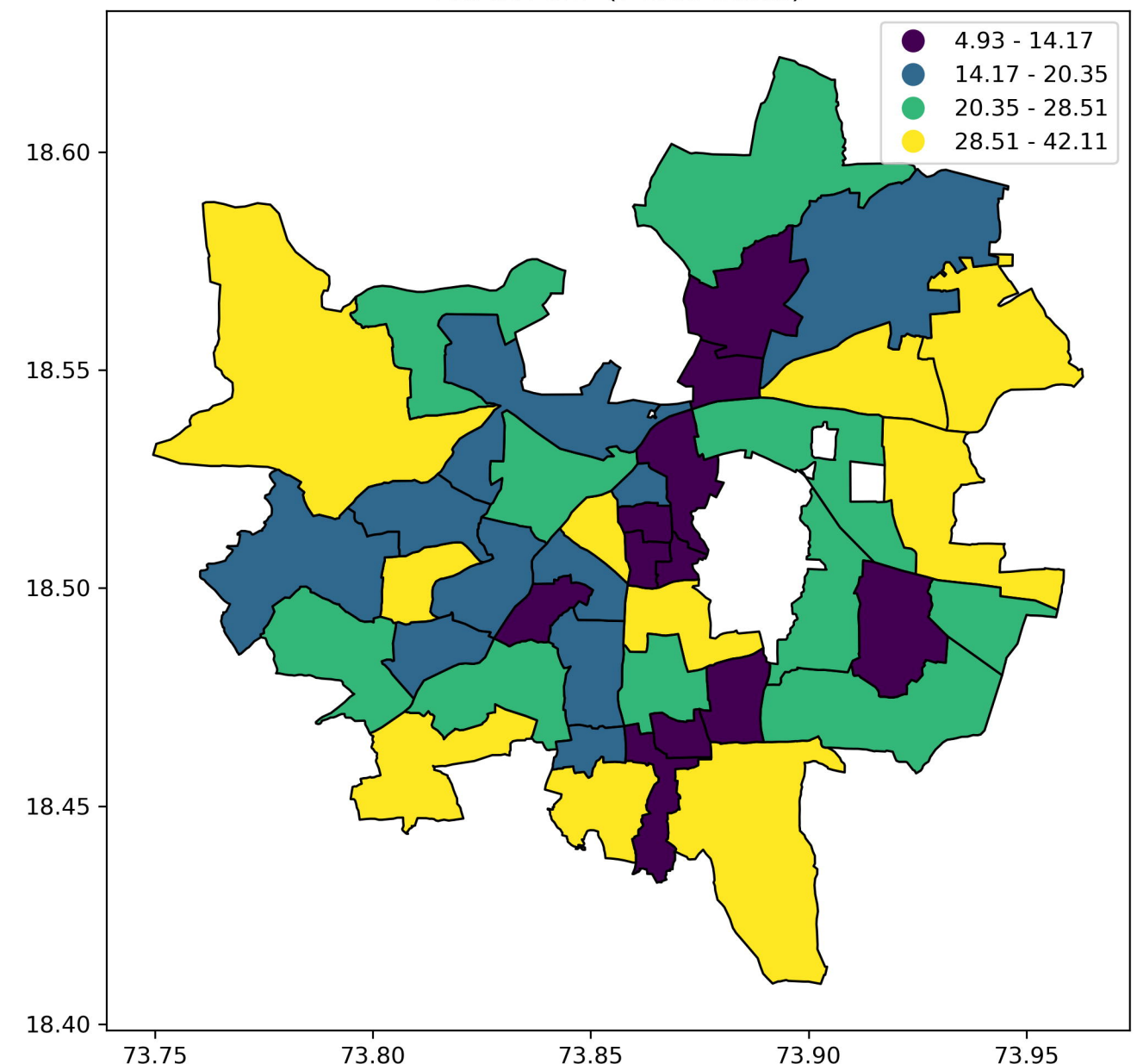
September 2020 (52077 cases)

**c**

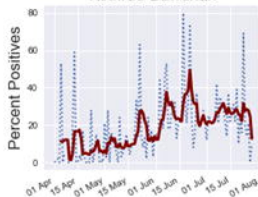
December 2020 (8578 cases)

**d**

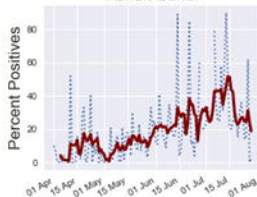
March 2021 (74358 cases)



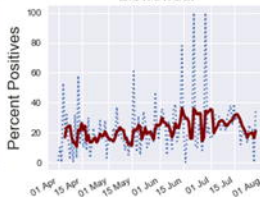
Kothrud-Bawdhan



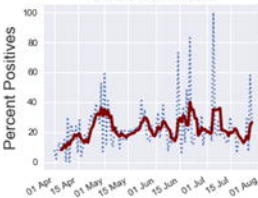
Aundh-Baner



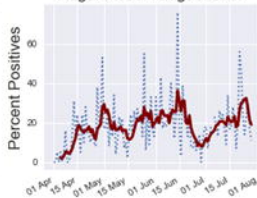
Bibwewadi



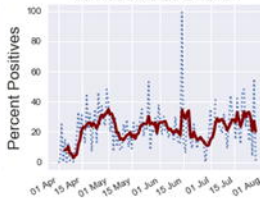
Dhole Patil Road



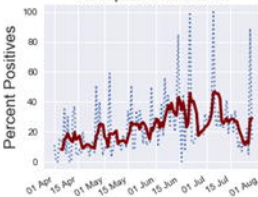
Nagar Road Vadgaonsheri



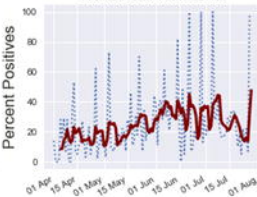
Yerwada-Kalas-Dhanori



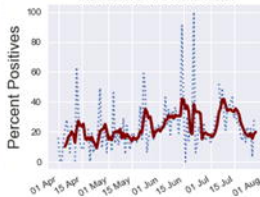
Hadapsar-Mundhawa



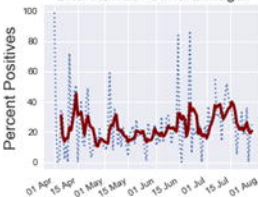
Wanawadi-Ramtekdi



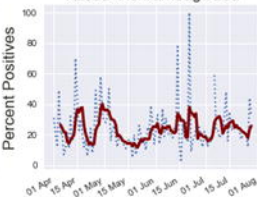
Kondhwa-Yewalewadi



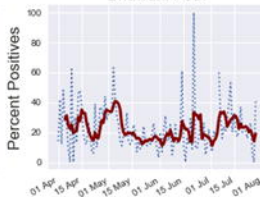
Dhankawadi-Sahakarnagar



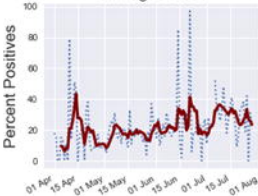
Kasba-Vishrambagwada



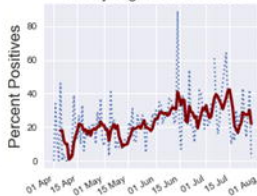
Bhawani Peth



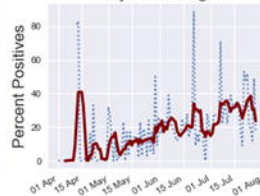
Sinhgad Road



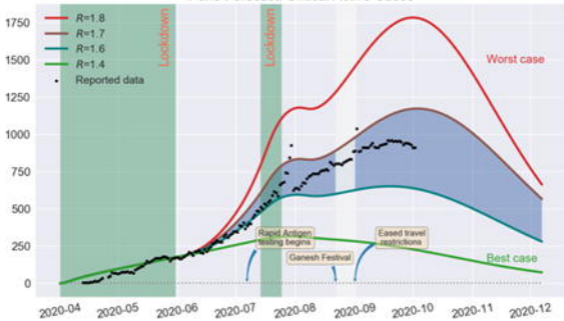
Shivajinagar-Ghole Road



Warje-Karvenagar



Pune Forecast: Critical Active Cases



Pune Forecast: Critical Active Cases

