Covid Data Analytics : Case Study on India

Bankar Prasad Vilas Subhasis Panda Vaibhav Anand Vineet Kumar

Indian Statistical Institute, Kolkata

https://covid-isical.tech contact@covid-isical.tech

SUMMER OF MACHINE LEARNING AT SKOLTECH

Introduction

We provide a macro-level India specific analysis on COVID related data as obtained from [1]. We built a dashboard available at [2] by giving multi-dimensional view and risk summary for India.

Spread Assessment Metrics

Increasing trend of occurrence of new cases is likely to indicate that the virus is spreading.

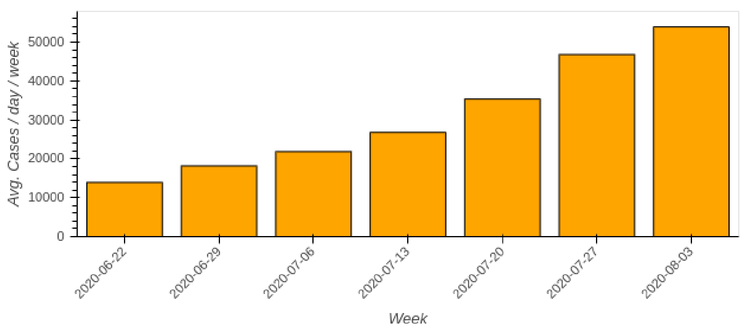


Figure 1: Average confirmed cases / day in India for each week (last 7 weeks)

Increasing trend of active cases indicates the possibility of increased stress on resources at present or in the near future.

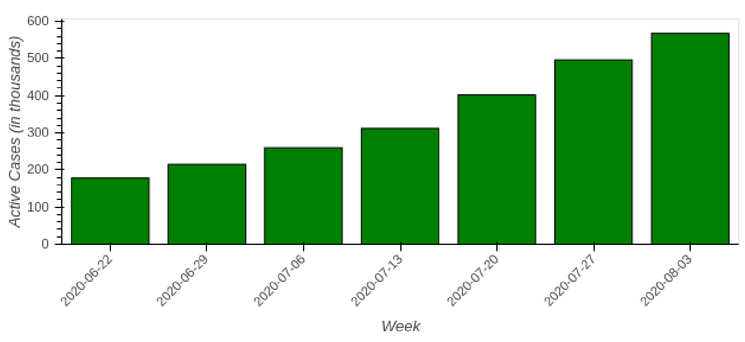


Figure 2: Active cases in India for the last 7 weeks

Risk Assessment Metrics

Avg. Confirmed Cases to Avg. Recoveries in Consecutive Weeks Traffic Intensity (term from queuing theory) —

- Leading indicator of stress on resources
- Values > 1 implies faster arrival of new cases compared to recovery

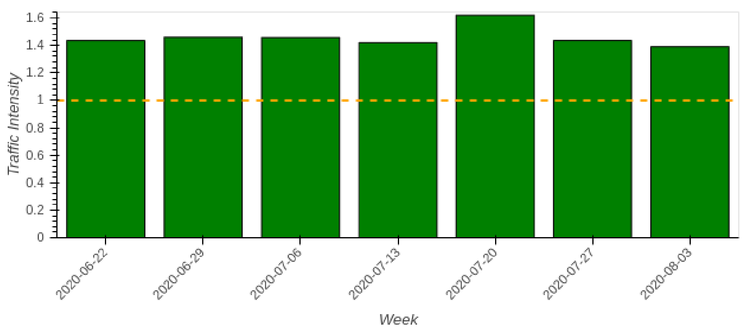


Figure 3: Traffic Intensity for India for the last 7 weeks

Points plot of Test in Each Week and % positive in each week Higher proportion positive test results on a larger test population (possibly less targeting) are anomalous and need investigation.

There are at least three possibilities:

- Rapid increase of rate of infection
- Carrying out tests in new (hitherto untested) areas with high but unknown rate of infection
- Improper testing leading to many incorrectly positive results

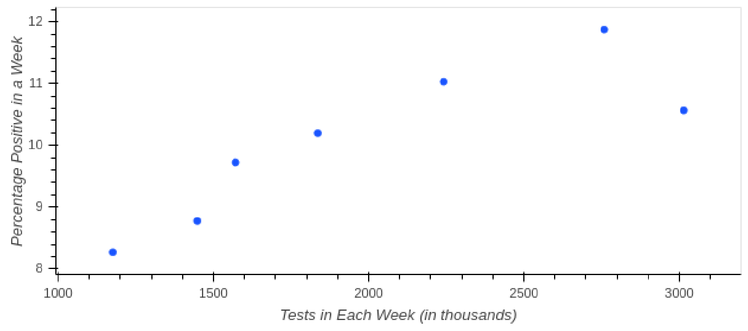
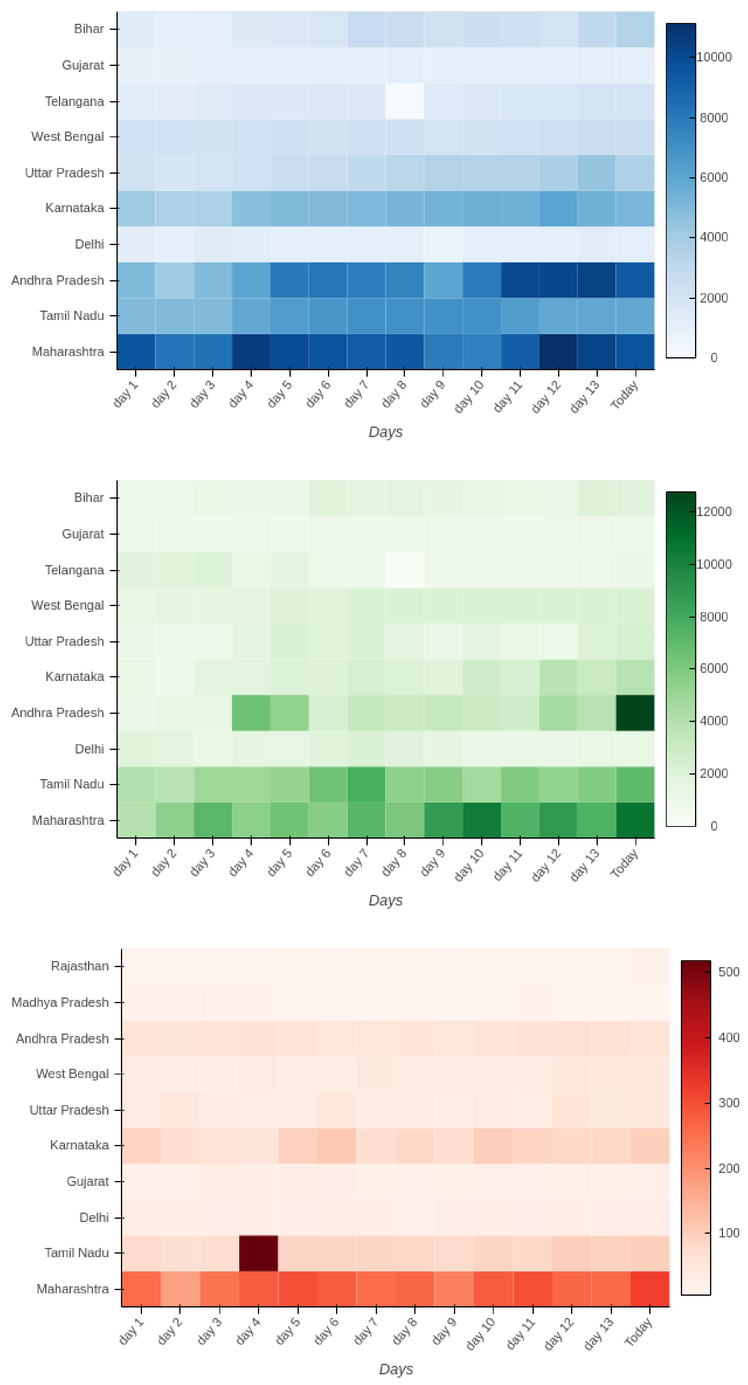


Figure 4: Assessing adequacy of testing in India

Comparative Assessment

Heat maps showing the top 10 States in India w.r.t confirmed (blue), recovered (green), and deceased (red), respectively.



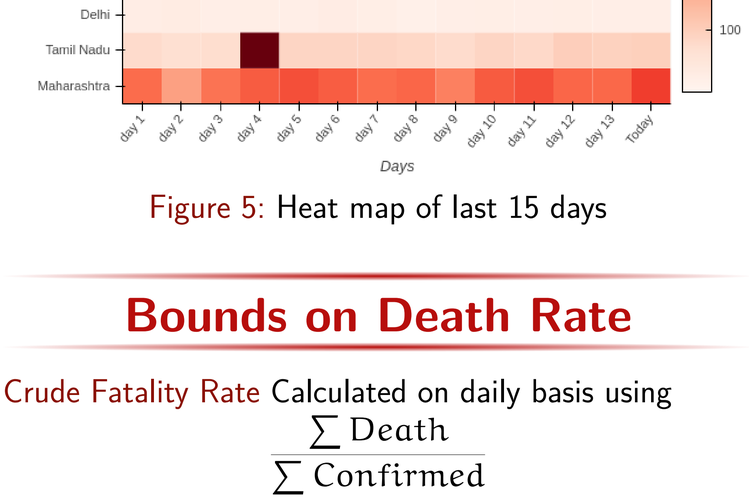


Figure 5: Heat map of last 15 days

Bounds on Death Rate

Crude Fatality Rate Calculated on daily basis using

$$\frac{\sum \text{Death}}{\sum \text{Confirmed}}$$

.

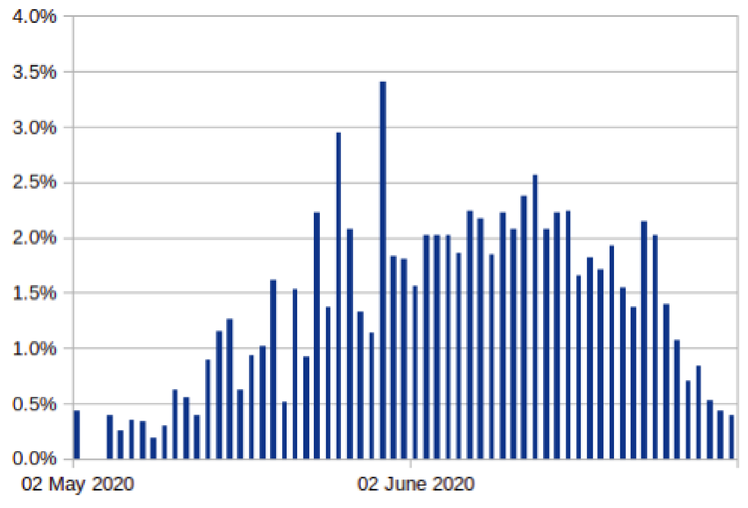


Figure 6: Crude Fatality Rate

Kaplan Meier Analysis The estimator of the survival function $S(t)$ (the probability that life is longer than t is given by:

$$\widehat{S}(t) = \prod_{i:\ t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

with t_i a time when at least one event happened, d_i the number of events (e.g., deaths) that happened at time t_i, and n_i the individuals known to have survived (have not yet had an event or been censored) up to time t_i.

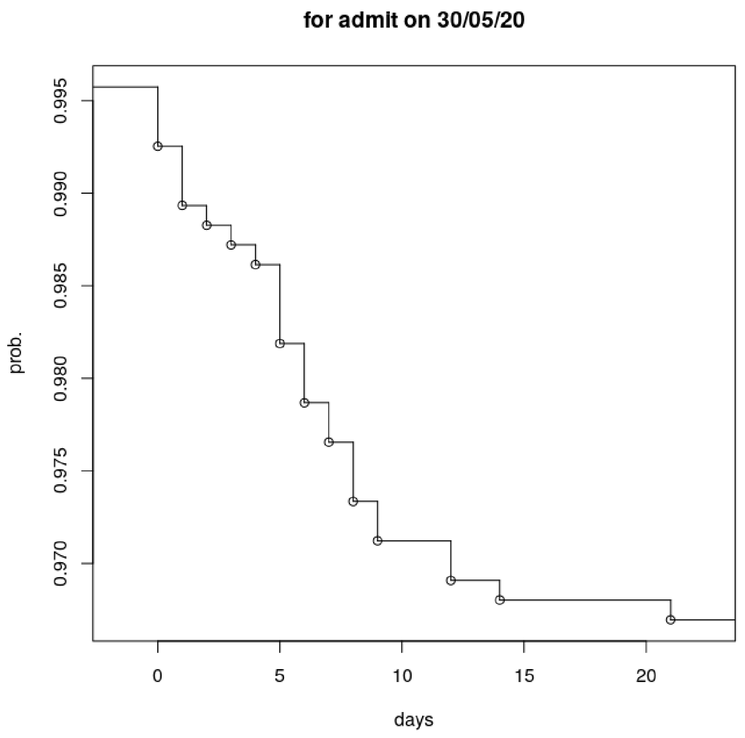
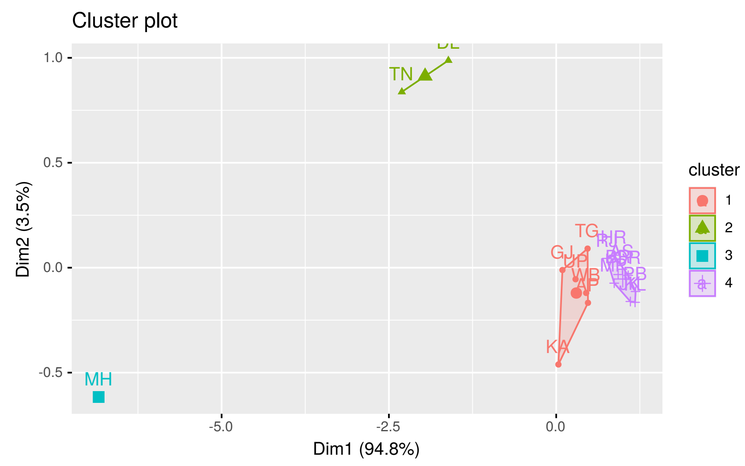


Figure 7: KM Curve for Survival Estimation

Spread-based Clustering

Clustered Districts based on number of active cases and deaths using k means and agglomerate clustering techniques.



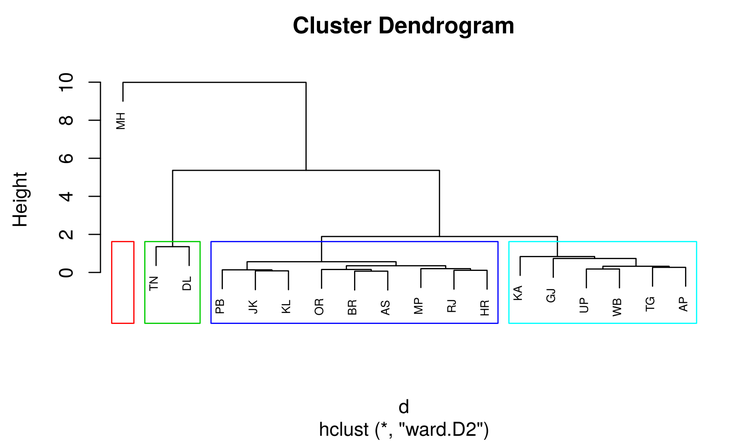


Figure 8: Clustering states based on infection spread

References

[1] https://api.covid19india.org/

[2] https://covid-isical.tech/