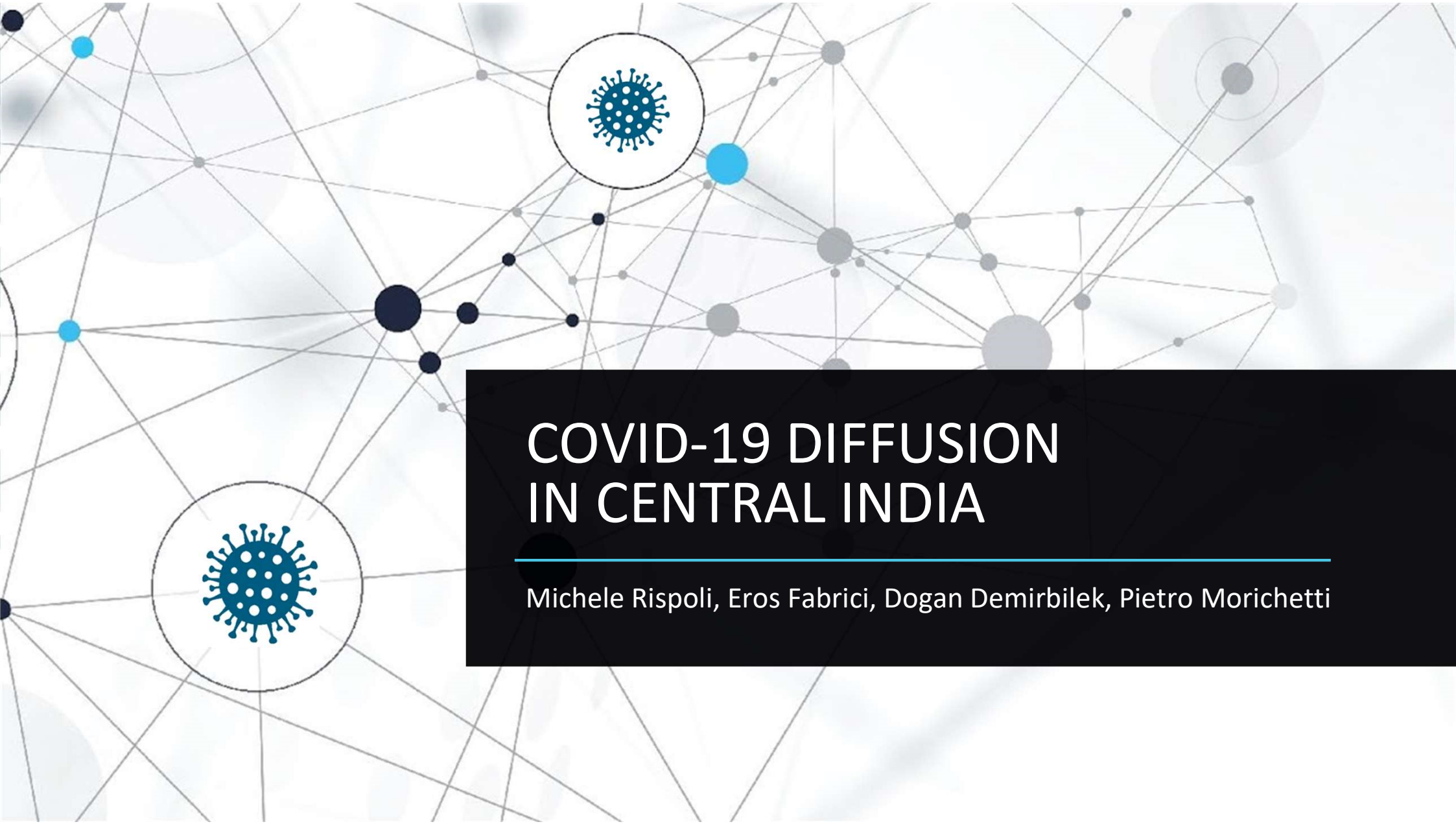
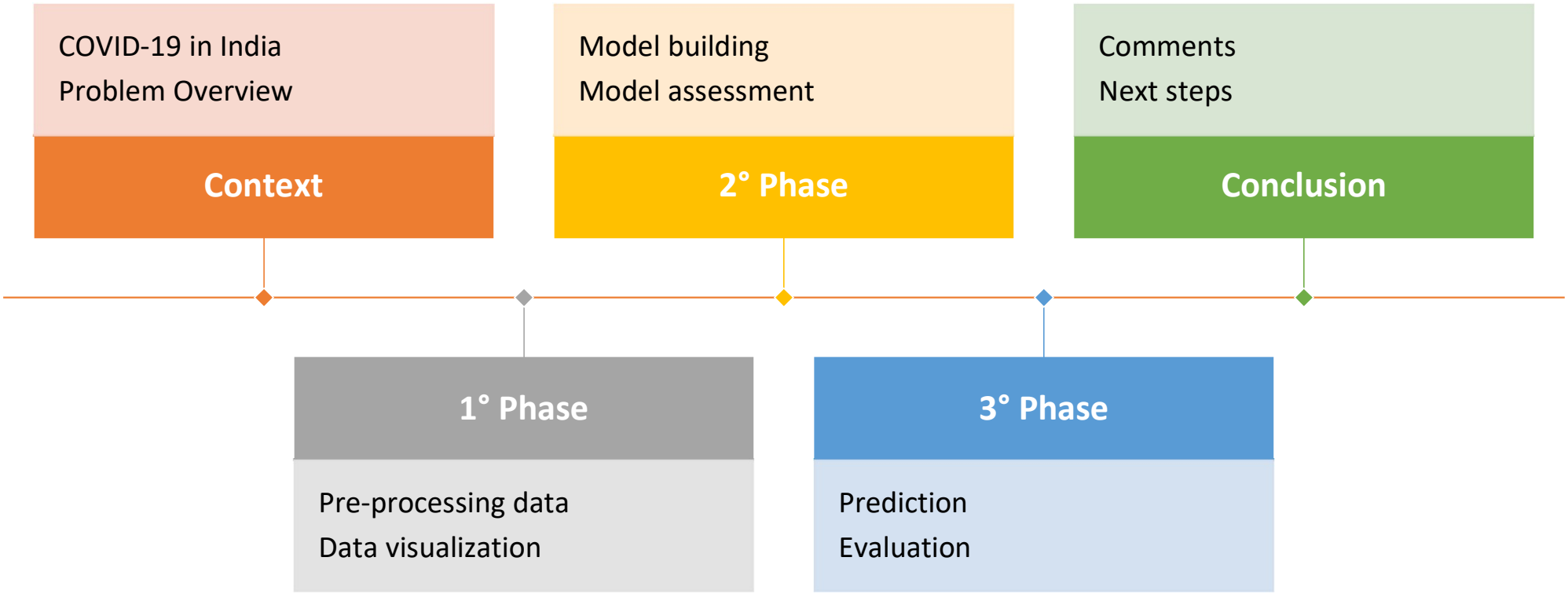


COVID-19 DIFFUSION IN CENTRAL INDIA

Michele Rispoli, Eros Fabrici, Dogan Demirbilek, Pietro Morichetti

Road Map of our Project



CONTEXT

- Covid-19 in Central India
- Problem Overview

Context

Covid-19 in India

- India is composed of 28 states, with 1.45 billion of people (24 times of italian population)
- First case reported on 30 January 2020, at this moment 4-th place in world
- OxCGRT gave a score of 100 at the end of June for the swift emergency policy

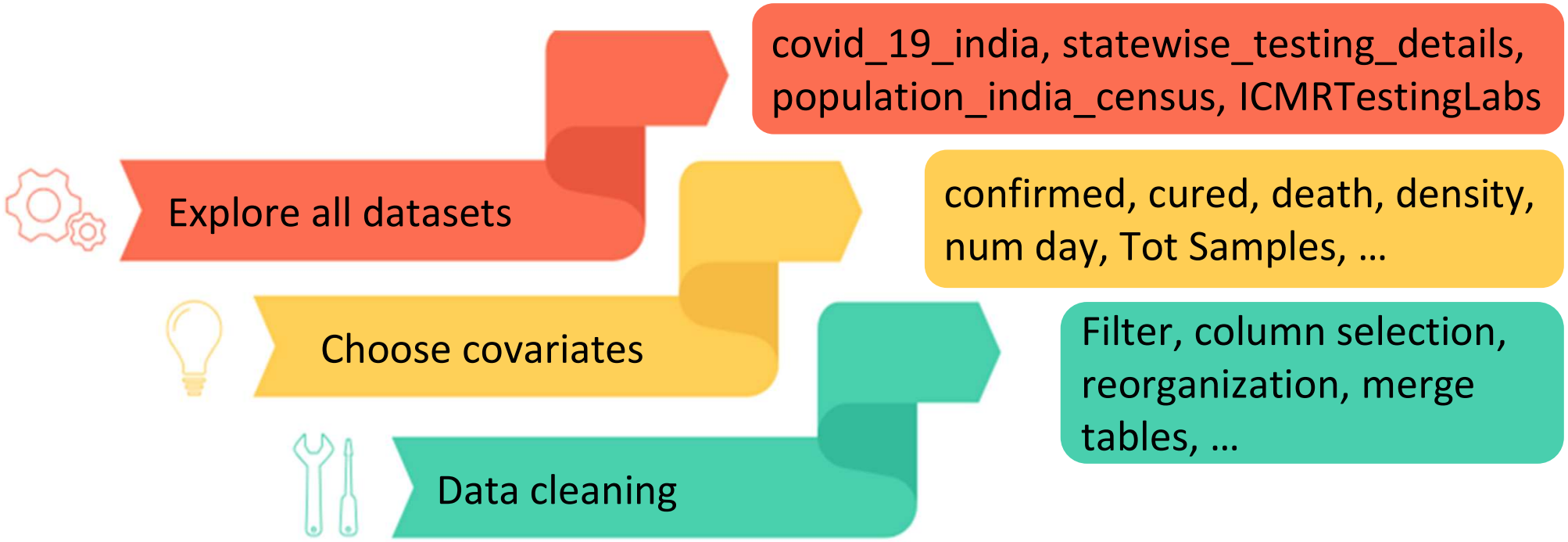
Problem Overview

- Focus on central states of India
- Build a predictive model over the daily confirmed cases of the pandemic
- Build a predictive model over the cumulative confirmed cases of the pandemic

1° PHASE

- Pre-processing data
- Data visualization

1° Phase: Pre-processing data

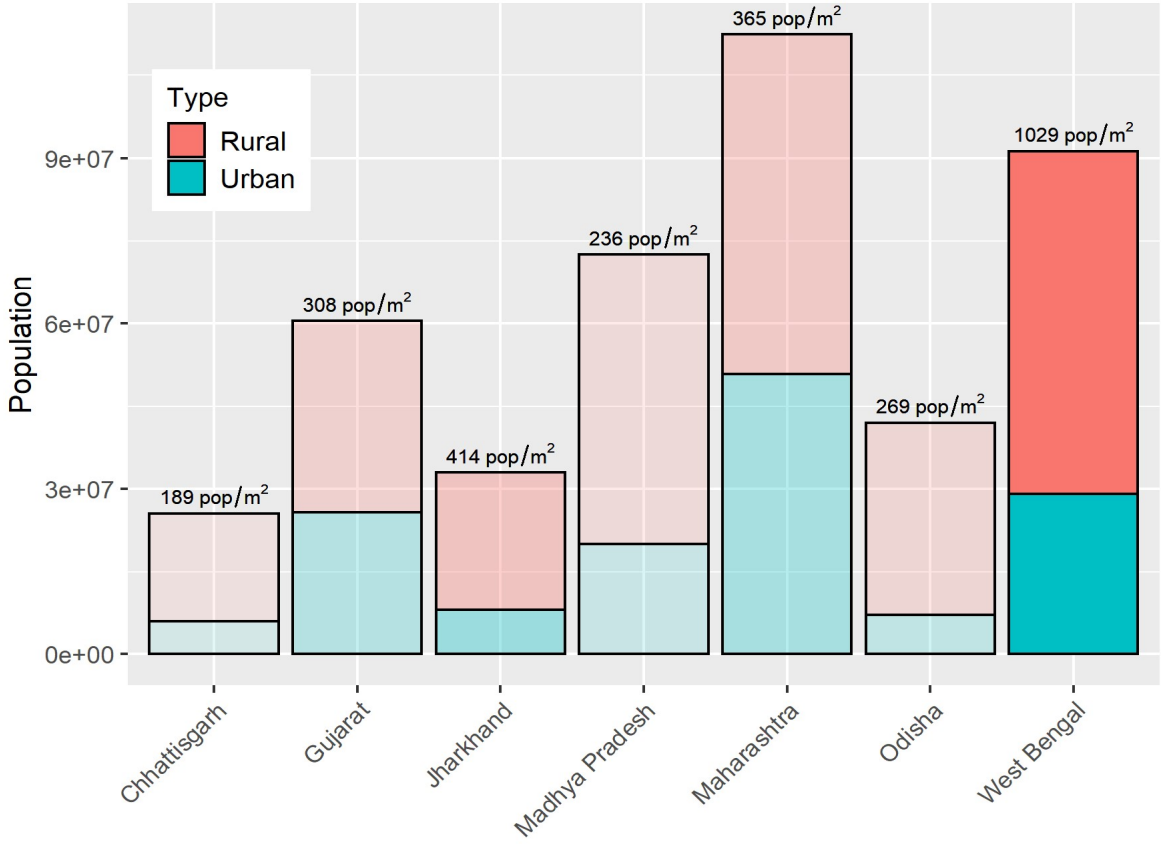


** These datasets are taken from Kaggle website*
** Some chosen covariates are used for data visualization and prediction*
** Last phase is valid for every kind of analysis*

1° Phase: Data visualization

- Total population divided by rural and urban and density for each state

- Focus on Maharashtra and West Bengal, they are the most populous

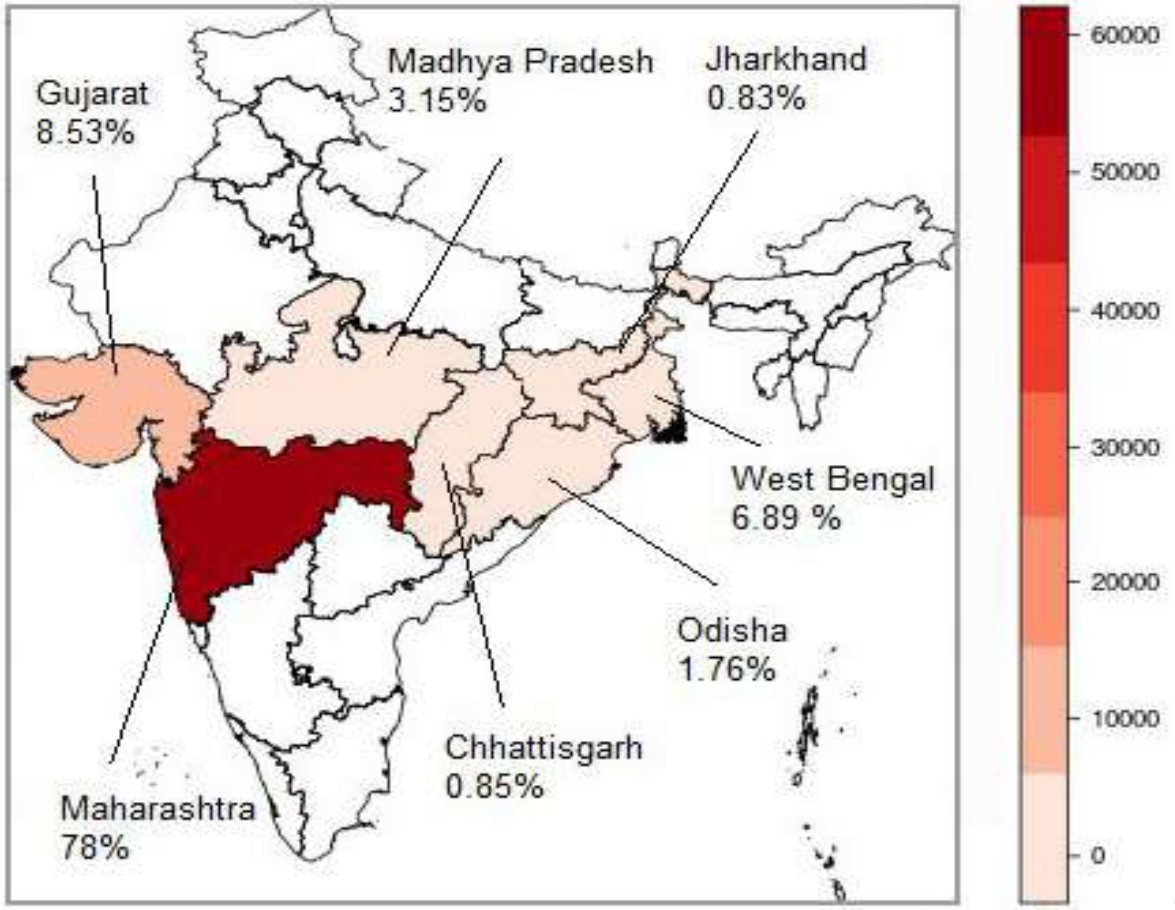


1° Phase: Data visualization

Cumulative Analysis

- Confirmed cases in each region, they are concentrated on Maharashtra.

- Also according the percentage scale, it is the most critical state

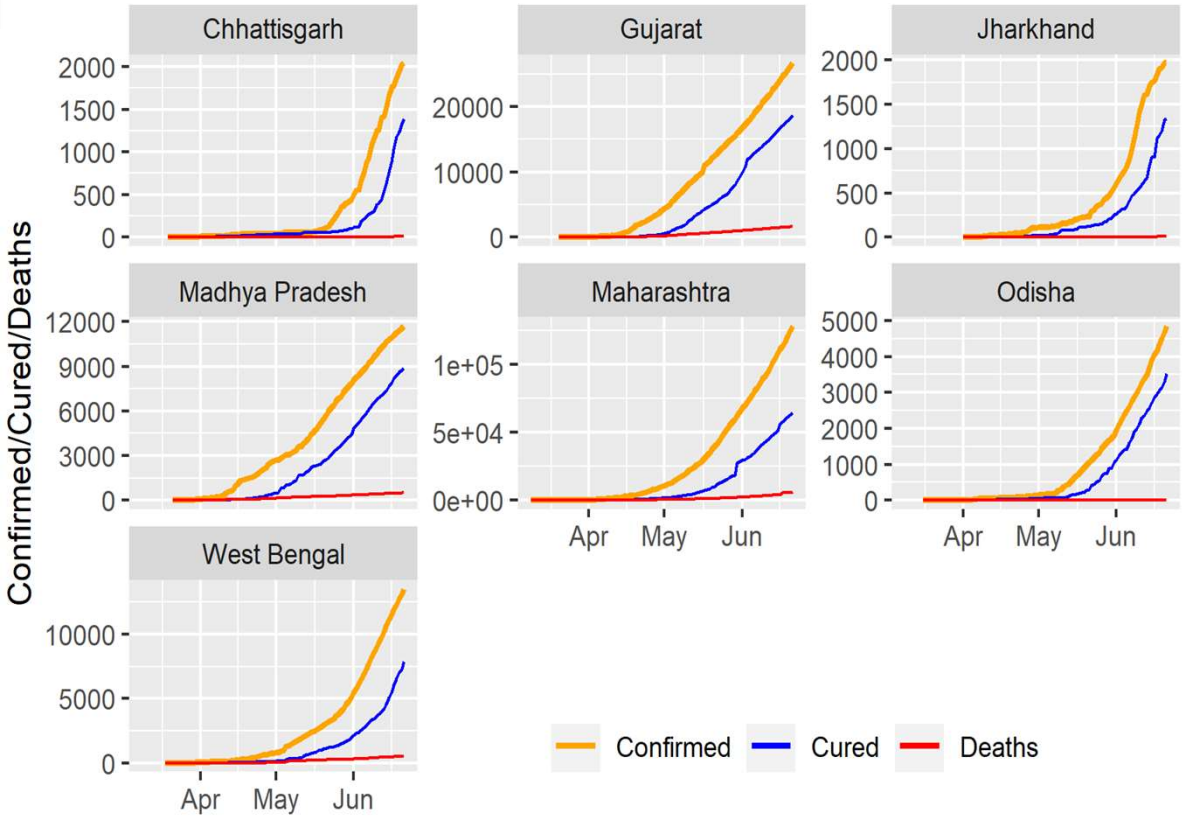


1° Phase: Data visualization

Cumulative Analysis

- Plot shows cumulative confirmed cases, cured and death trends

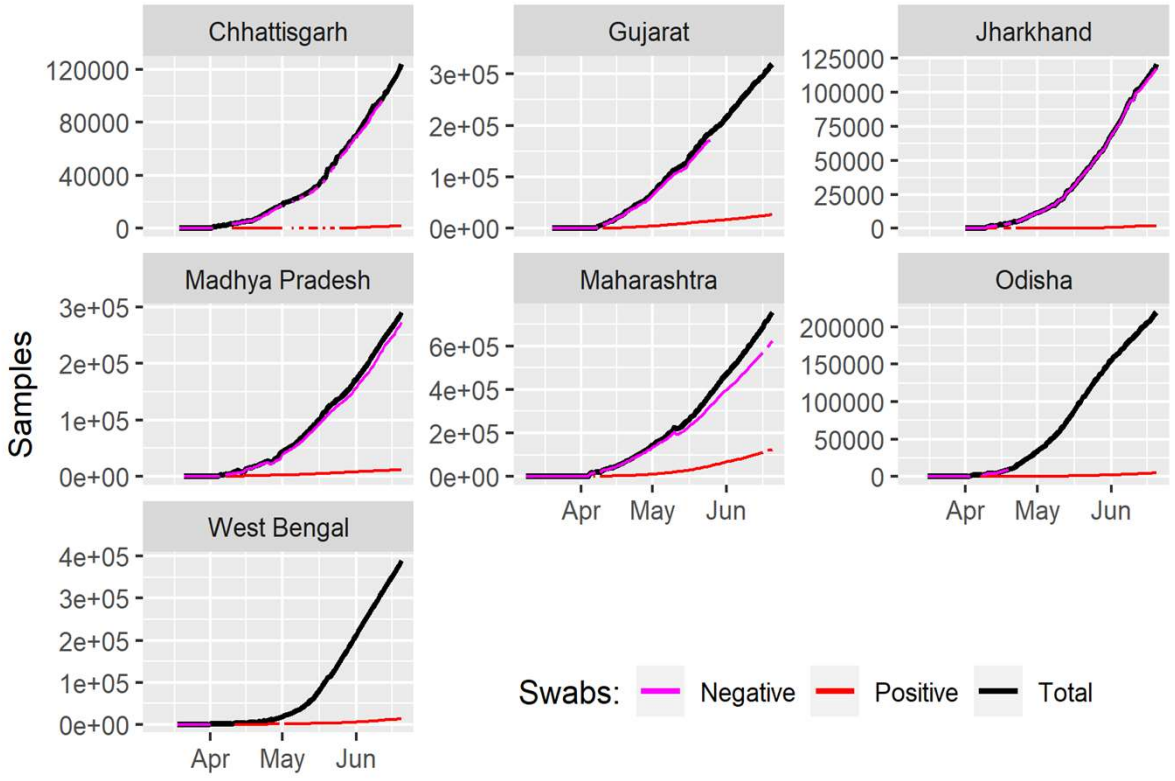
- Interesting note: apparently Cured+Deaths lags behind confirmed of about 14 days



1° Phase: Data visualization

Cumulative Analysis

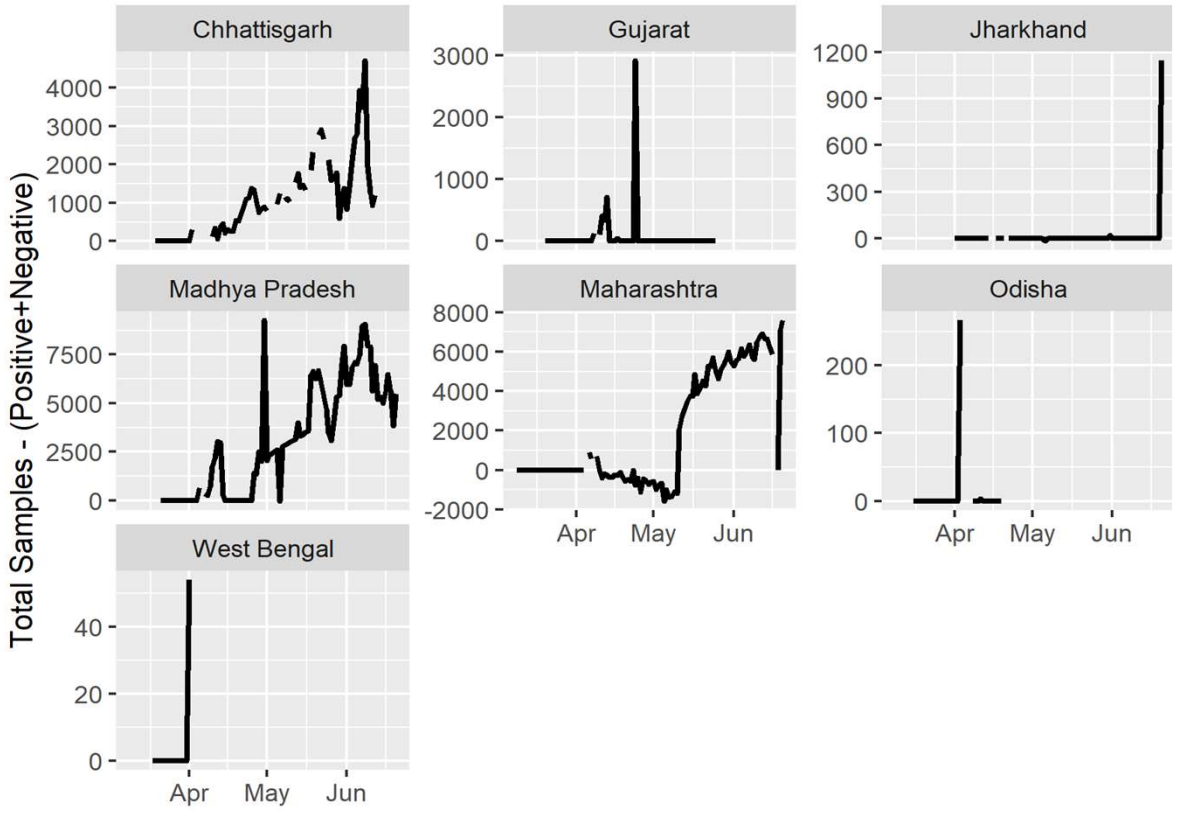
- Counting of test responses for each state
- Lots of missing data for positive/negative swabs



1° Phase: Data visualization

Cumulative Analysis

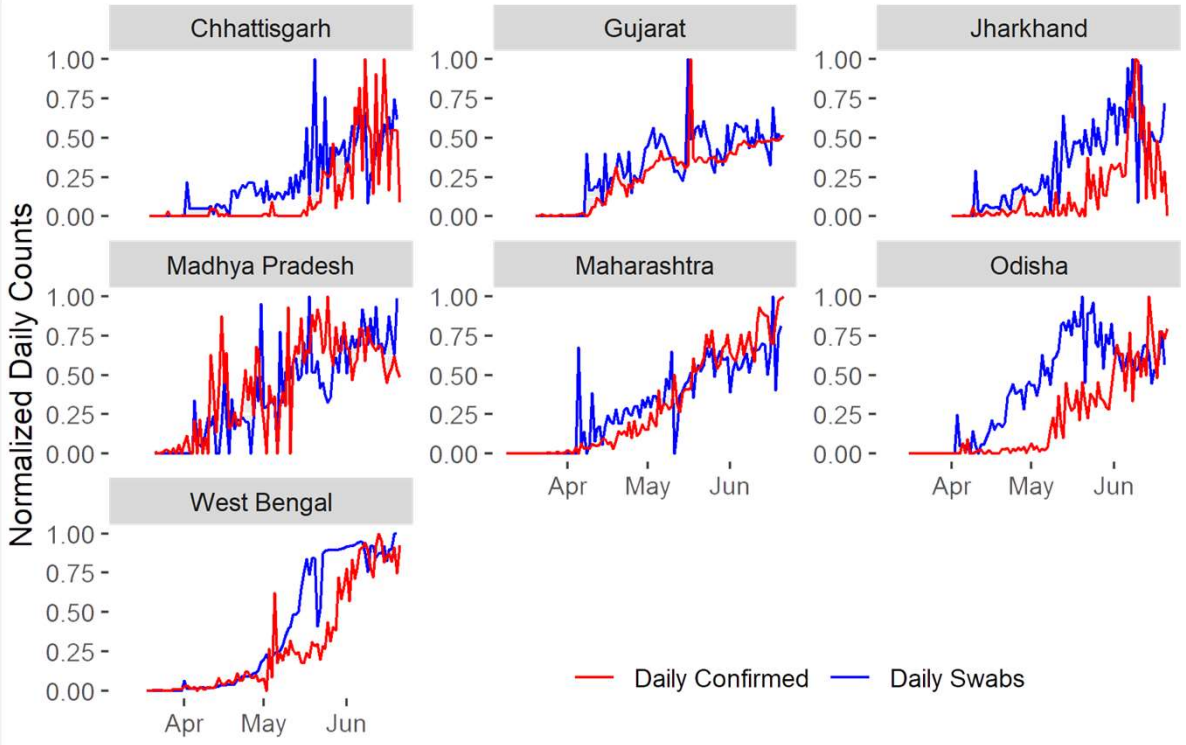
- Difference between the total swabs, and number of positive and negative
- Difference should be zero, so there is noise in data
- For some points we observed negative values (Maharashtra)



1° Phase: Data visualization

Daily Analysis

- Daily swabs vs daily confirmed
- Few irregular points where found in daily swabs count and also in daily confirmed cases
- Those points were handled setting them to zero



2° PHASE

- Model building
- Model assessment

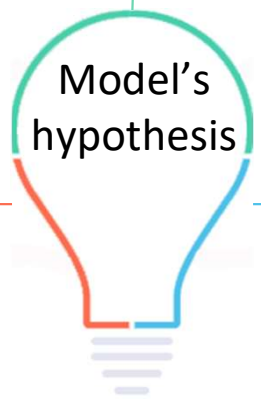
2° Phase: Model Building

Daily Analysis

Excluded
- State specific constants (Population, Healthcare and Testing facilities)
- Positive and Negative swabs
- Cured and Deaths

GLM poisson
- Integer values
- Range $[0, +\infty]$

Covariates
- Yesterday Tot samples
- Yesterday confirmed
- Num day
- Num day^2



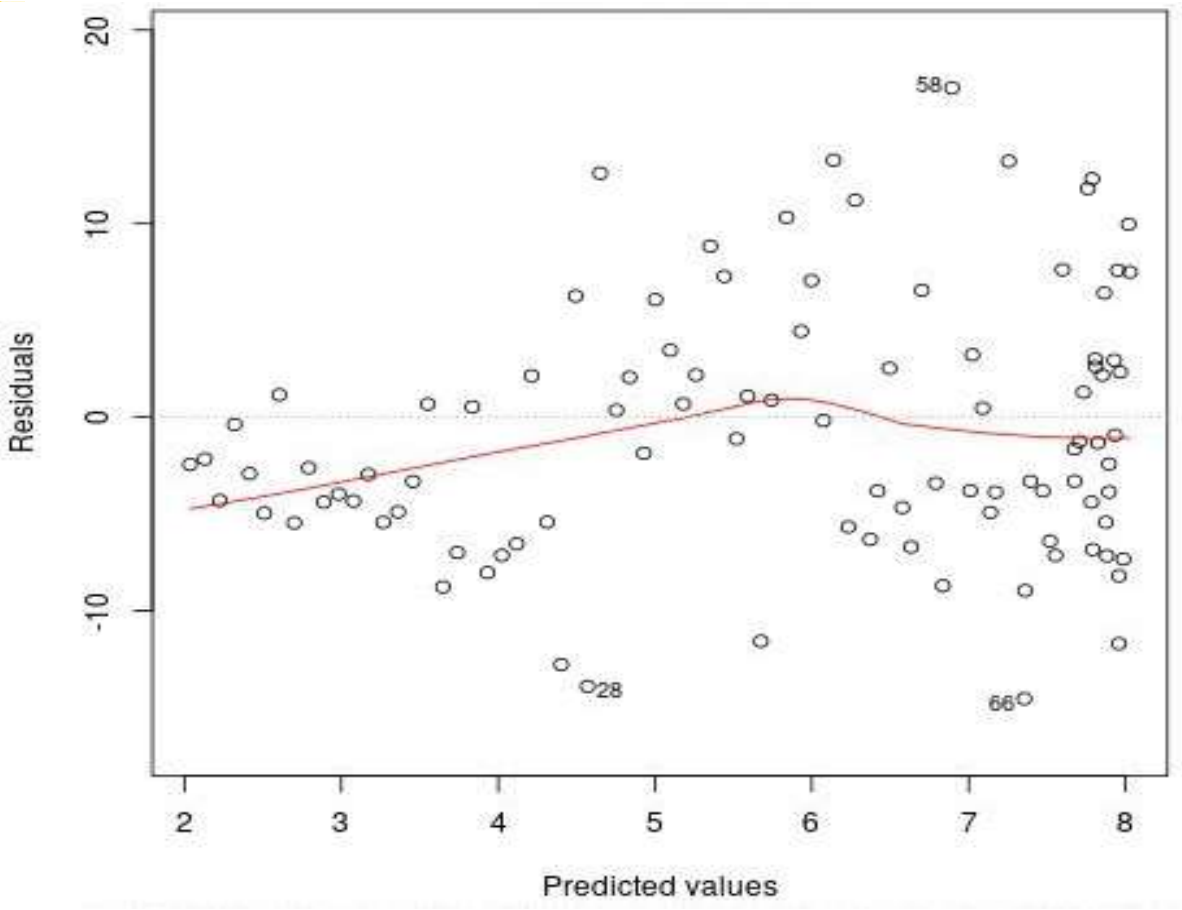
$$Daily\ Confirmed_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 \cdot Num\ Day_i + \beta_2 \cdot Num\ Day_i^2 + \beta_3 \cdot Yest\ Confirmed_i + \beta_4 \cdot Yest\ Tot\ Samples_i$$

2° Phase: Model Assessment

Daily Analysis

- Scaled residuals vs fitted values plot (Maharashtra)
- We used the AER package for dispersion test over poisson model
- P-value is very close to zero and we reject the equidispersion null hiphotesys
- We switched to **quasi-poisson** model



2° Phase: Model Assessment

Daily Analysis

- *Backward selection:* we discard the covariates with p-value over 0,05
- *F-Test:* p-value over 0,05 means pick the simplest model otherwise reject it

	MODELS (Maharashtra)	RES DEV	F-TEST
M1	$Num\ Day + Num\ Day^2 + Yest\ Confirmed + Tot\ Samples$	4217	0,68
M2	$Num\ Day + Num\ Day^2 + Yest\ Confirmed$	4224	0,19
M3	$Num\ Day + Yest\ Confirmed$	4299	<<
M4	$Yest\ Confirmed$	42000	-

** F-test was performed considering sequential models (eg. M1~M2, M2~M3, …)*

2° Phase: Model Assessment

Daily Analysis

	Intercept	Num Day	Num Day^2	Yest Confirmed	Yest Tot Samples	RES DEV
Gujarat	X	X	X	X	X	1311
Chhattisgarh	-	-	-	X	X	650,07
Jharkhand	-	X	X	X	X	521,52
Odisha	-	-	X	X	-	573
Maharashtra	X	X	-	X	-	4373
West Bengal	-	X	X	-	X	1084,8
Madhya Pradesh	X	X	--	X	-	2948,9

2° Phase: Model Building

Cumulative Analysis

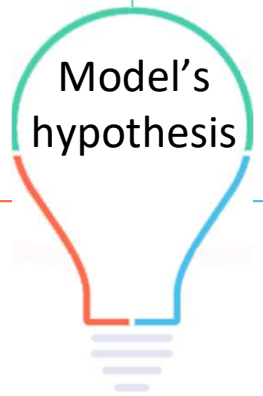
- One idea is extending daily models to cumulative cases, simply by summing them
- Second idea is using *tscount* package (i.e. modeling of count time series following glm), because it is a flexible class of models which can describe serial correlation in a parsimonious way

2° Phase: Model Building

Cumulative Analysis 2

Excluded
- State specific constants (Population, Healthcare and Testing facilities)
- Positive and Negative swabs

GLM poisson
- Cumulative count values
- Range $[0, +\infty]$



Covariates
- Previous confirmed cases
- Mean of past 14 confirmed cases
- Total Sample
- Previous daily confirmed cases
- Previous Cured and Deaths

Legend
- λ_t : estimated mean condition on previous process history

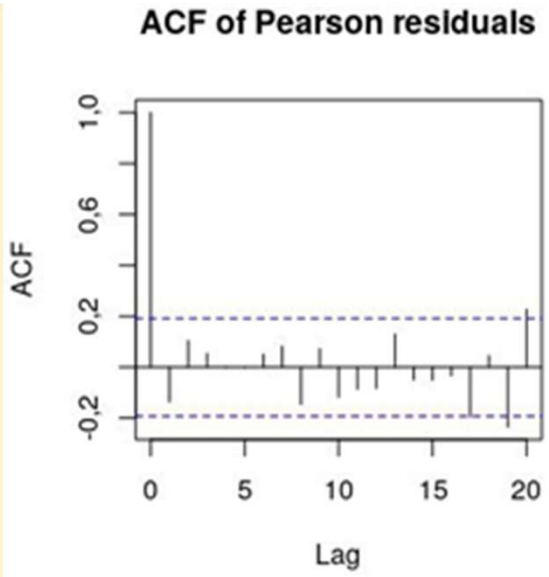
$$Cumulative\ Confirmed_t \sim Poisson(\lambda_t)$$
$$\log(\lambda_t) = \beta_0 + \beta_1 \cdot Cumulative\ Confirmed_{t-1} + \alpha \cdot \log(\lambda_{t-14}) + \eta_i \cdot Cov_{i,t}$$

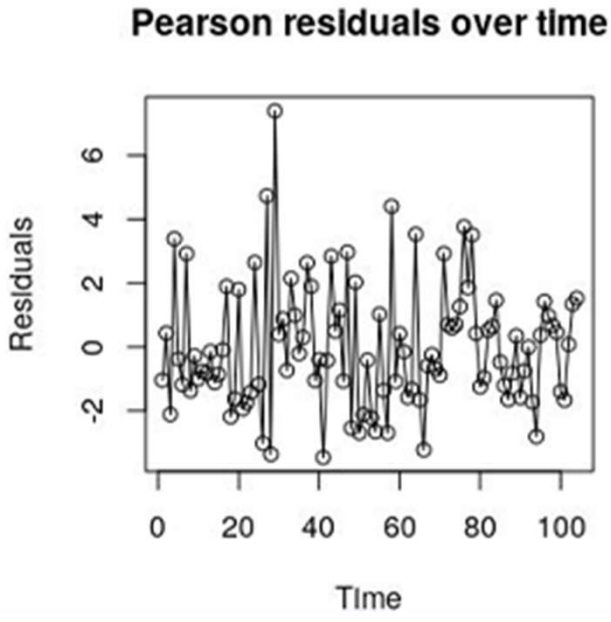
2° Phase: Model Assessment

Cumulative Analysis 2

- Diagnostics ACF and Residuals (Maharashtra)

- To assess a model we checked the residuals and we tried to observe not auto-correlated, random residuals





2° Phase: Model Assessment

Cumulative Analysis 2

- We choose the model which has lowest MSE on both train and test sets
- Information criteria were excluded because of small and discording differences

	MODELS (Maharashtra)	TOT MSE
M1	$\beta_0 + \beta_1 \cdot y_{t-1} + \alpha \cdot \lambda_{t-14} + \eta_1 \cdot Tot\ Samples + \eta_2 \cdot Yest\ Daily\ Confirmed + \eta_3 \cdot \text{Yest}\ Cured + \eta_4 \cdot Yest\ Death$	1247406
M2	$\beta_0 + \beta_1 . y_{t-1} + \alpha\ \cdot \lambda_{t-14} + \eta_1 \cdot Tot\ Samples\ + \eta_2 \cdot Yest\ Daily\ Confirmed + \eta_3 \cdot Yest\ Cured$	151067
M3	$\beta_0 + \beta_1 \cdot y_{t-1} + \alpha \cdot \lambda_{t-14} + \eta_1 \cdot Tot\ Samples + \eta_2 \cdot Yest\ Daily\ Confirmed$	98448
M4	$\beta_0 + \beta_1 \cdot y_{t-1} + \alpha \cdot \lambda_{t-14} + \eta_1 \cdot \text{Tot}\ Samples$	86735

2° Phase: Model Assessment

Cumulative Analysis 2

	Yest Cured	Yest Death	Yest Daily Confirmed	Tot Samples	TOT MSE
Gujarat	-	-	X	X	9576
Chhattisgarh	-	-	-	X	13179
Jharkhand	X	-	X	X	2613
Odisha	-	-	X	X	425
Maharashtra	-	-	-	X	86736
West Bengal	X	X	X	X	9274
Madhya Pradesh	-	-	X	X	19415

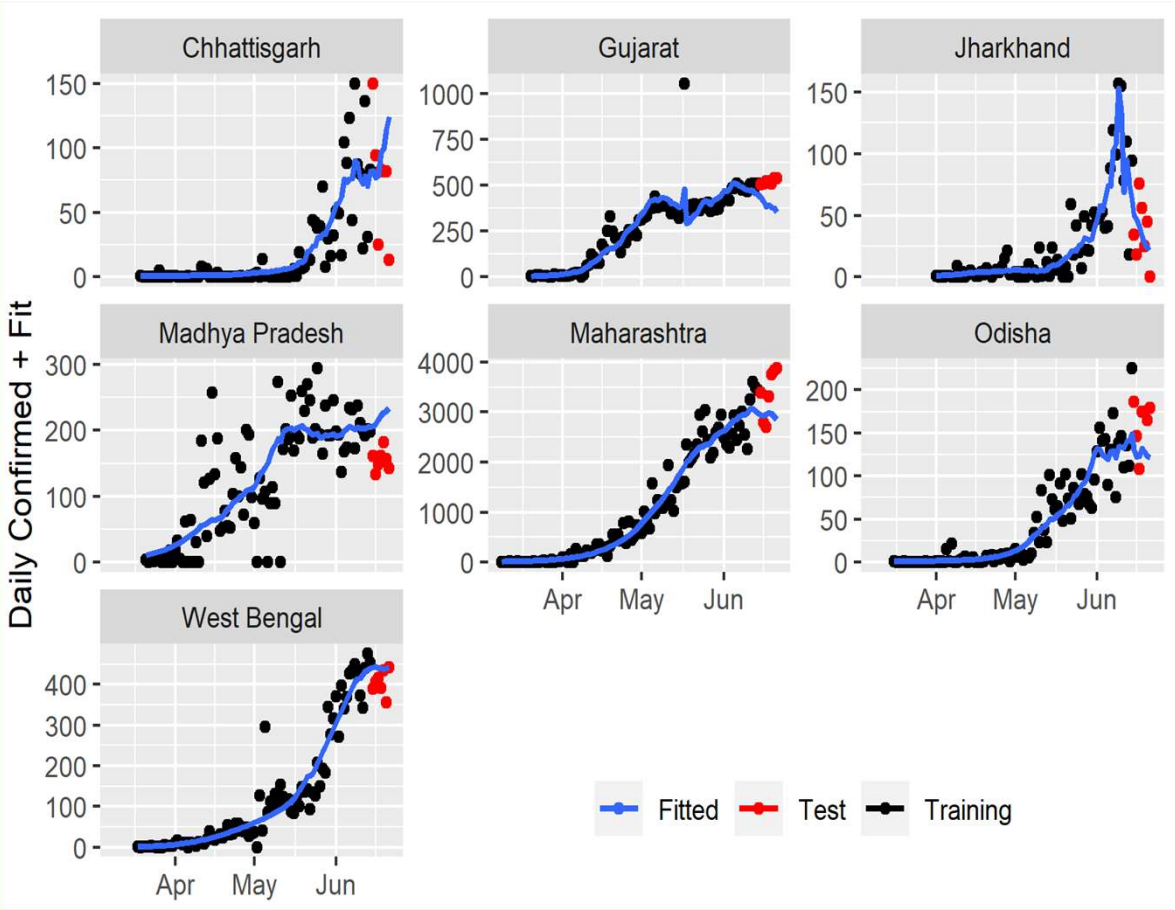
3° PHASE

- Prediction
- Evaluation

3° Phase: Prediction

Daily Analysis

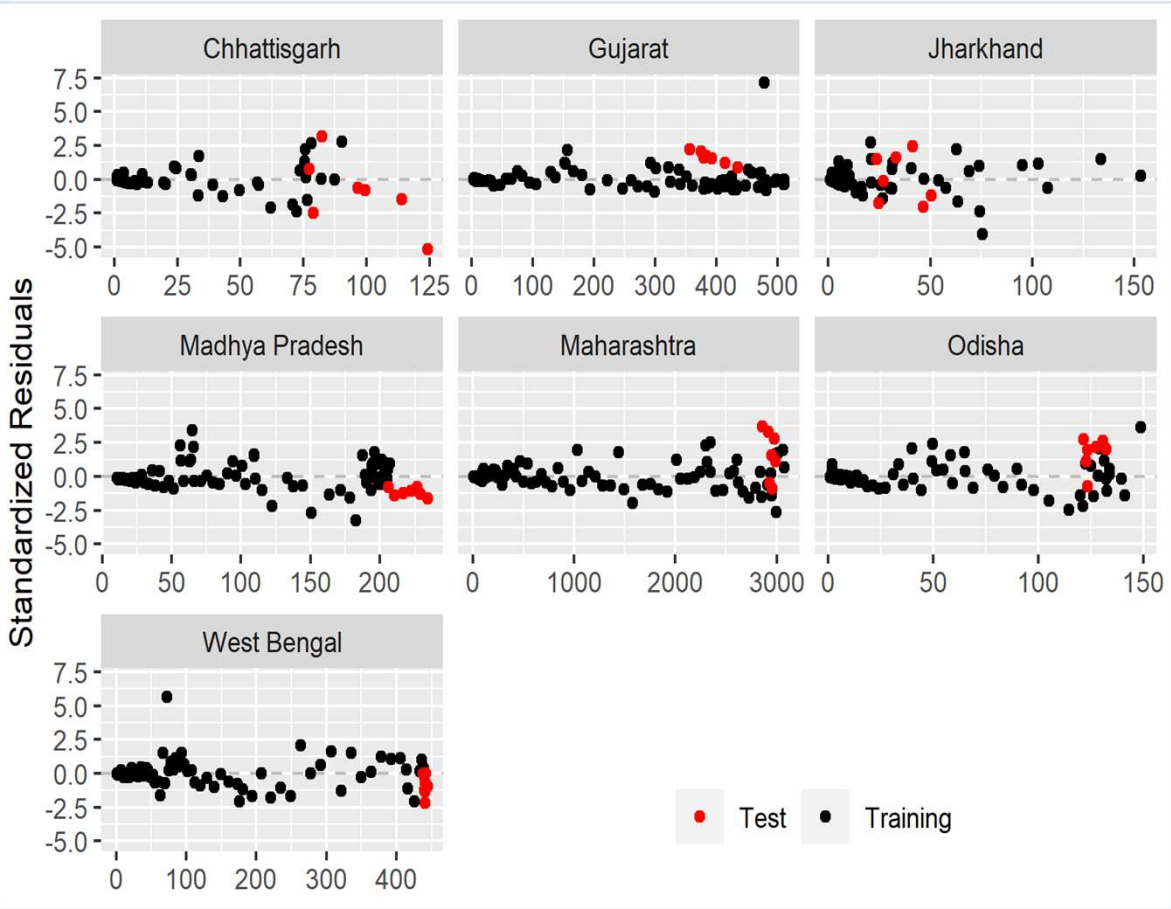
- Final predictive plot by using the selected models
- For some states the observations are very noisy
- Models fit well less noisy data



3° Phase: Evaluation

Daily Analysis

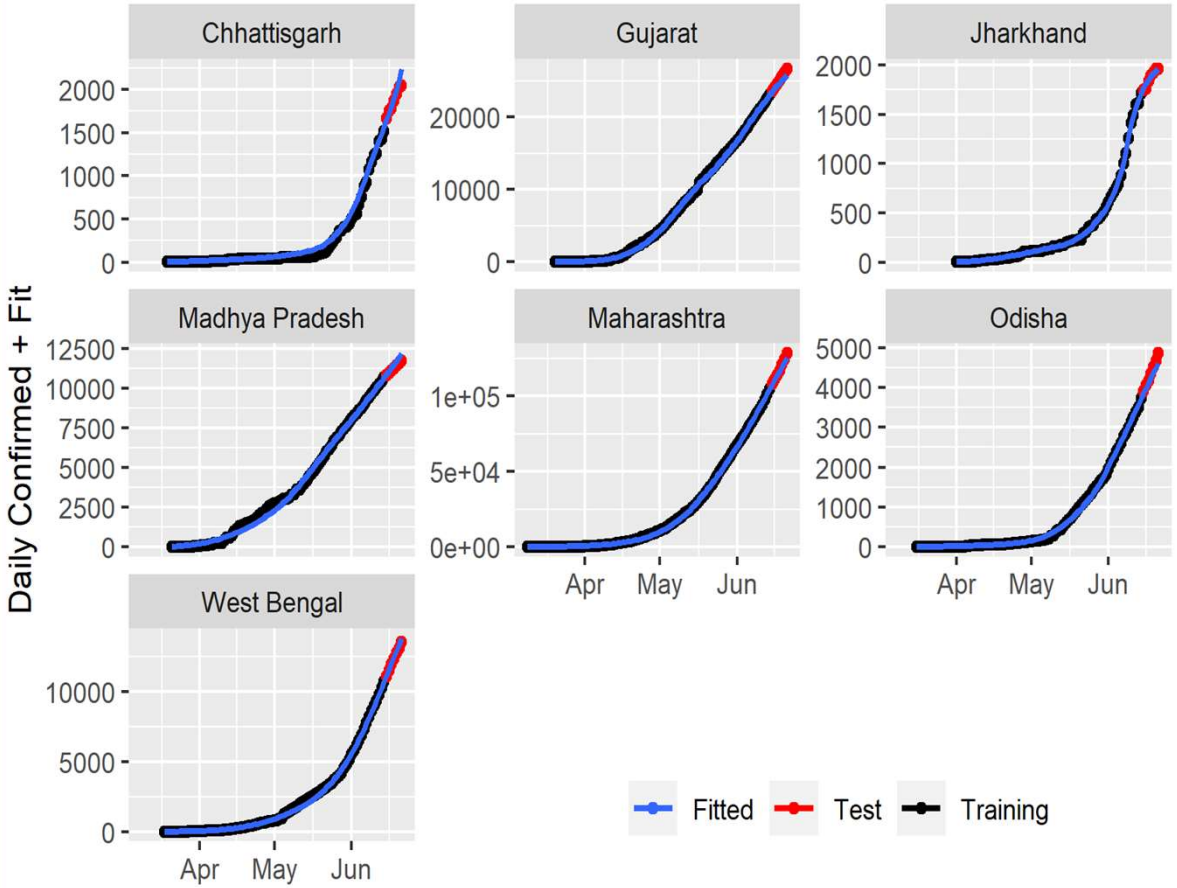
- Residuals plot for each state
- Residuals distributions show no systematic pattern
- Heteroskedasticity is observed but less than in Poisson case



3° Phase: Prediction

Cumulative Analysis 1

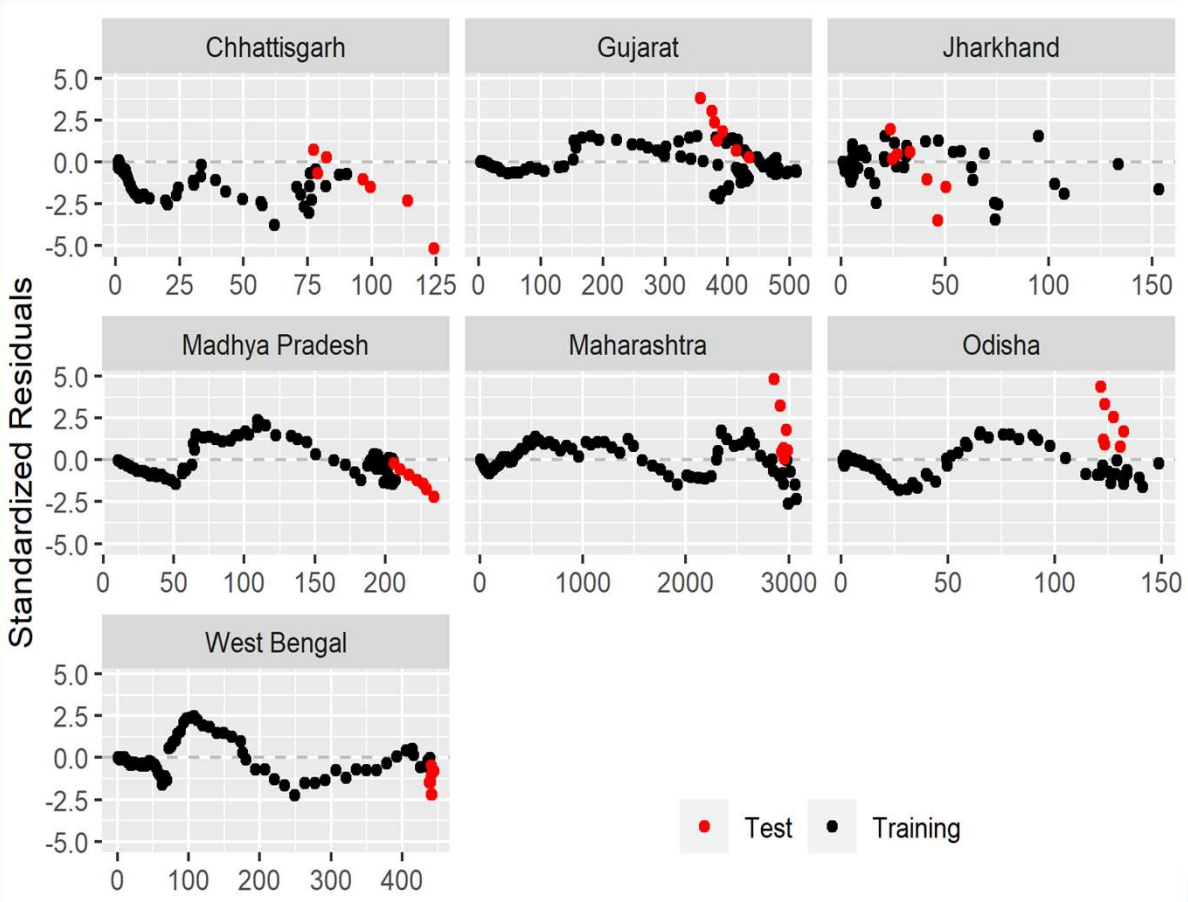
- Final cumulative predictive plot by using the Daily Model
- Apparently good fit on data



3° Phase: Evaluation

Cumulative Analysis 1

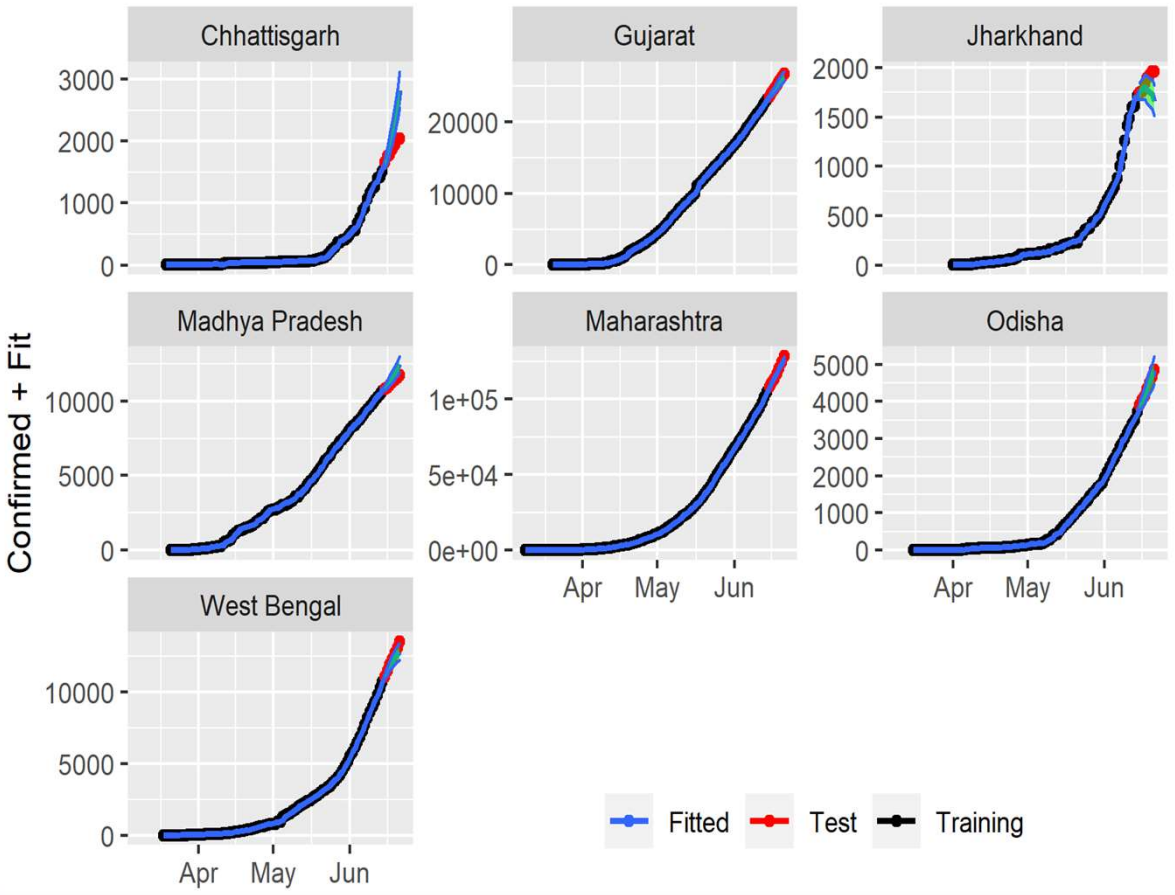
- Standardized residuals vs fitted values
- Residuals performance is not satisfying and shows systematic behaviour



3° Phase: Prediction

Cumulative Analysis 2

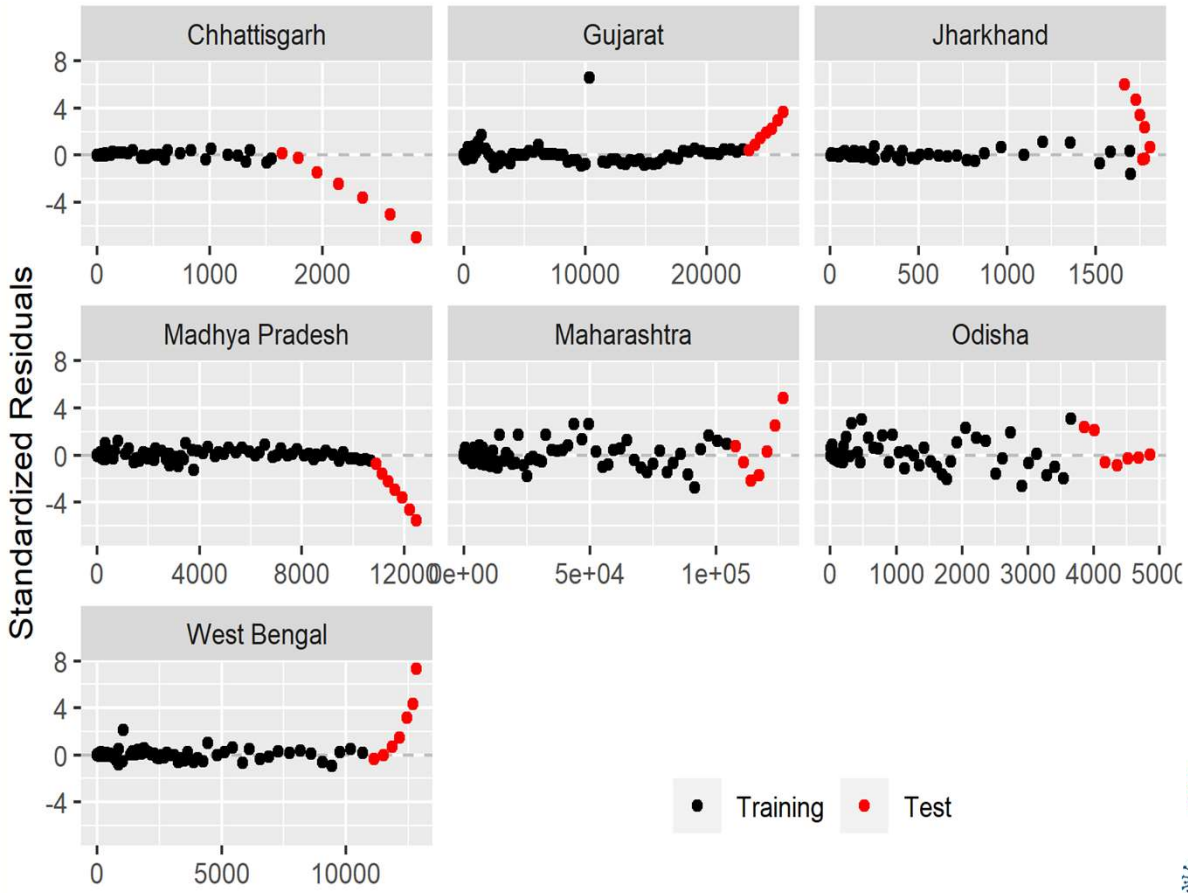
- Final predictive plot by using the selected models
- Chhattisgarh and Jharkhand show an aberrant behaviour on test data



3° Phase: Evaluation

Cumulative Analysis 2

- Standardized residuals vs predicted values
- Most of the states show a trending behaviour on test



CONCLUSION

- Comments
- Next steps

Conclusion: Comments

PROS

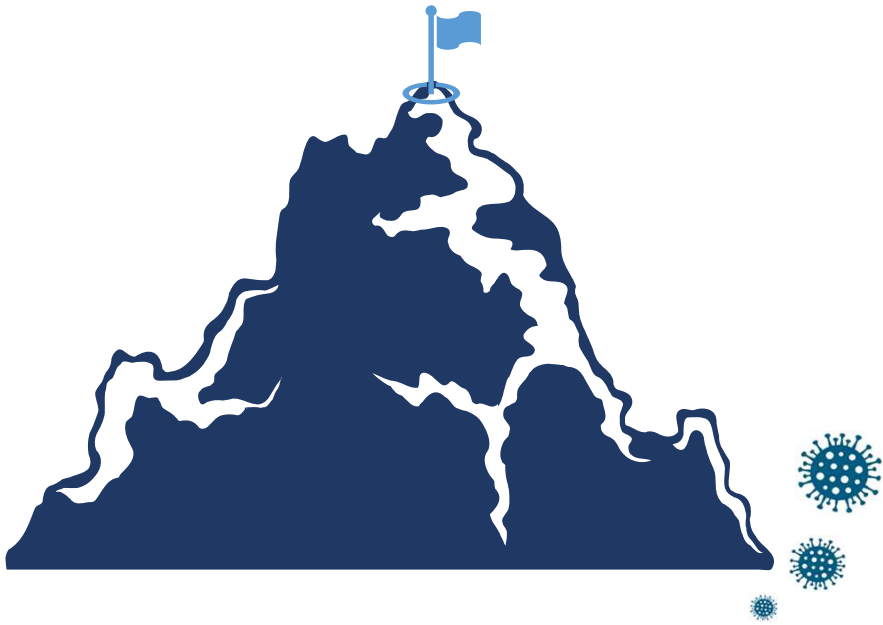
- Residuals distributions in daily models show no systematic pattern
- Gives acceptable predictions for spread and noisy data

CONS

- Need to predict covariates for future predictions
- Cumulative prediction from daily model shows deterministic residuals pattern

Conclusion: Next steps

- Study can be extended by using multilevel / hierarchical model instead of choosing different models for each state
- Bayesian models can be considered to combine prior information with data
- Other models such as logistic or Gompertz Curve can be tried especially for cumulative part



References

- Leonardo Egidi, “Covid-19 spreading outbreak – Italy”, https://www.leonardoegidi.com/covid-19

- Tobias Liboschik, Konstantinos Fokianos and Roland Fried, “tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models”, https://cran.r-project.org/web/packages/tscount/vignettes/tsglm.pdf

- Kaggle, “COVID-19 in India”, https://www.kaggle.com/sudalairajkumar/covid19-in-india?select=covid_19_india.csv

THANK YOU
FOR YOU ATTENTION

Michele Rispoli, Eros Fabrici, Dogan Demirbilek, Pietro Morichetti