CoViD-19 daily positive-case count prediction for Northern India

Emanuele Ballarin
Roberto Corti
Arianna Tasciotti

Department of Mathematics and Geosciences, University of Trieste

Statistical Methods for Data Science
Final Project $\sim$ July 2020

We are here...

CoViD-19 in India: What and Why

A global pandemic hitting hard the World's most populous democracy

- **CoViD-19**: clinical expression of the **human-host** infectious disease caused by *SARS-CoV-2*. Probably resulted from *animal spillover*.
- Identified during **Dec 2019**, in Wuhan (Hubei, China).
  As of Jul, 1, 2020: $>$ 10.4 mln cases, $>$ 511k deaths worldwide.

- **India**:
  Outbreak: ongoing since January 2020;
  1st $\in$ Asia and 4th $\in$ World country w.r.t. number of confirmed cases.
- Further focus on **Northern India**, i.e. the 9 regions of:
  Uttar Pradesh, Chandigarh, Haryana, Delhi, Himachal Pradesh, Jammu and Kashmir, Punjab, Rajasthan, Uttarakhand.

Formal Problem Statement

Understanding aims and limits of present work

Predict the number of new **daily** confirmed cases of *CoViD-19* for **Northern India**, on a *per-state* basis.

- Up to **10 days** since the last known daily *data-point*;
- While also providing **predictive uncertainty** information;
- Trying to favor **parameter parsimony**, **theoretically-sound** methods, **explainability** and **interpretability** of the results;
- By using only publicly-available data from the Internet.

Medical data $\rightarrow$ *extra-care required!*

We are here...

Data Gathering

Finding the source(s) of knowledge

- Kaggle (Sudalai Rajkumar):
  - Age grouping, early-outbreak ($<$ Mar 2020) [IMHW];
  - Per-state/per-date individual counting data [IMHW];
  - Confirmed individuals, early-outbreak [*covid19india*, volunt.];
  - Statewise geopolitical / hospital information [Indian Census 2011].

- *covid19india*'s API:
  - Per-state/per-date individual counting data; additional [volunt.];
  - Confirmed individuals, early-outbreak; additional [volunt.].

In a pipelined, incremental, **reproducible** manner (bash, binutils, R).

Self-contained *data arrangement*: **KF-CV** and **train/test** split.

Data Cleaning

Making *human world* more *machine-friendly*

Principled drop of data:

- Unnecessary data for the task considered (e.g. *lat/long* of labs);
- Technical-only information (e.g. IDs, index keys);
- Non-easily-processable data (with *permitted tools*; e.g. raw text).

Removal of to-be-**duplicated** data:

- To avoid *by-design* duplication and variance inflation.
  More on this in the forthcoming!

Dealing with **NA**s:

- Simple removal ($\rightarrow$ upfront information loss);
- Data-carry-on ($\rightarrow$ information-theoretical justification for T.S.s).

Data pre-processing

Making *machine-friendly* more *model-friendly*

Data transformations:

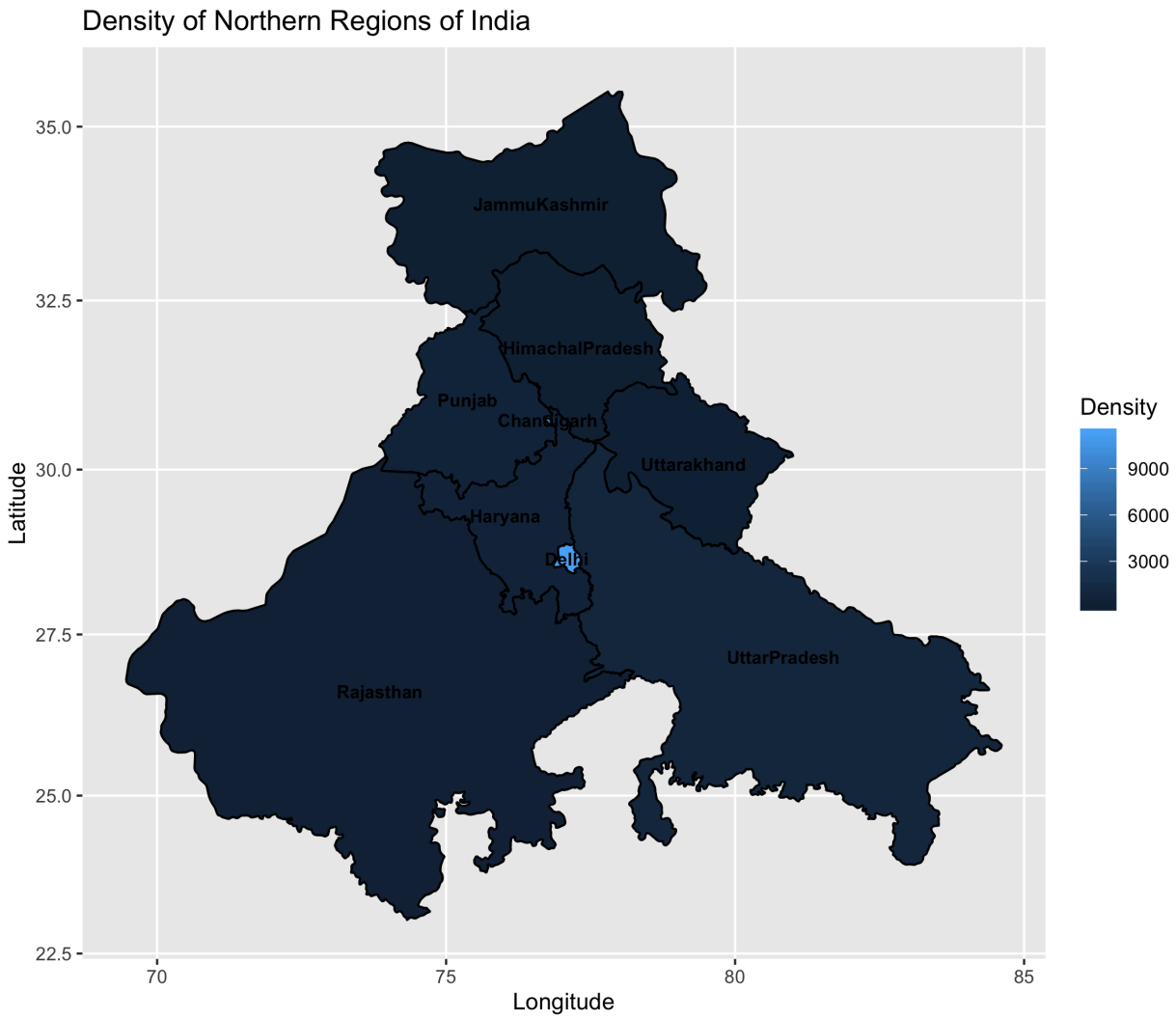
- Numerical encoding of categorical or string variables (e.g. date, sex);
- Daily averaging of patient information;
- Compressed, still informative quantities
  (e.g. urban population, rural population $\rightarrow$ urbanization).

Introduction of new variables:

- Daily metrics (e.g. daily positives, daily swabs);
- 3-, 4- lagged components (daily swabs), 10- lagged components (daily positive and average daily patient's age and sex).

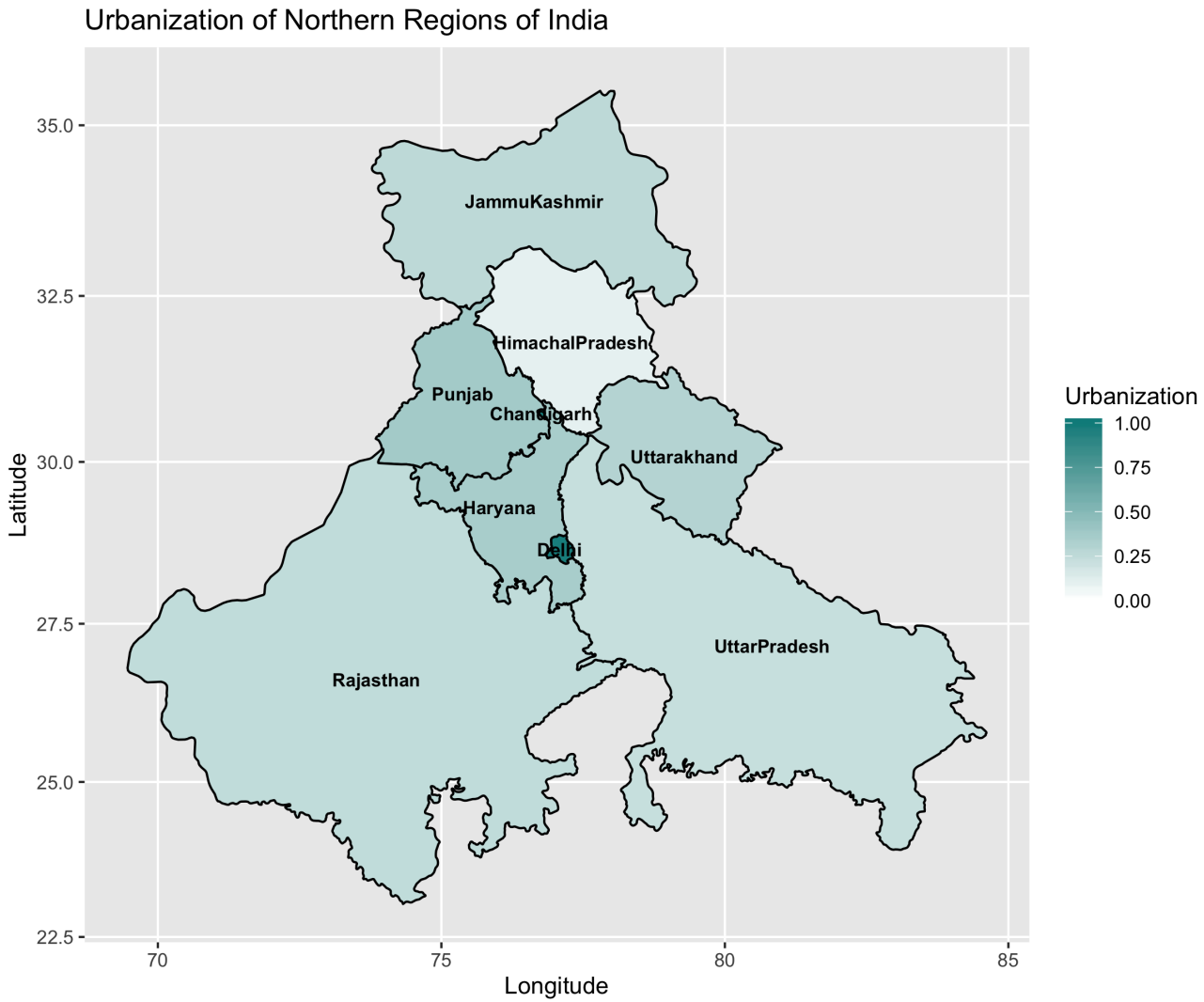
Exploratory Data Analysis: Demographic features (1)

A country of contrasts



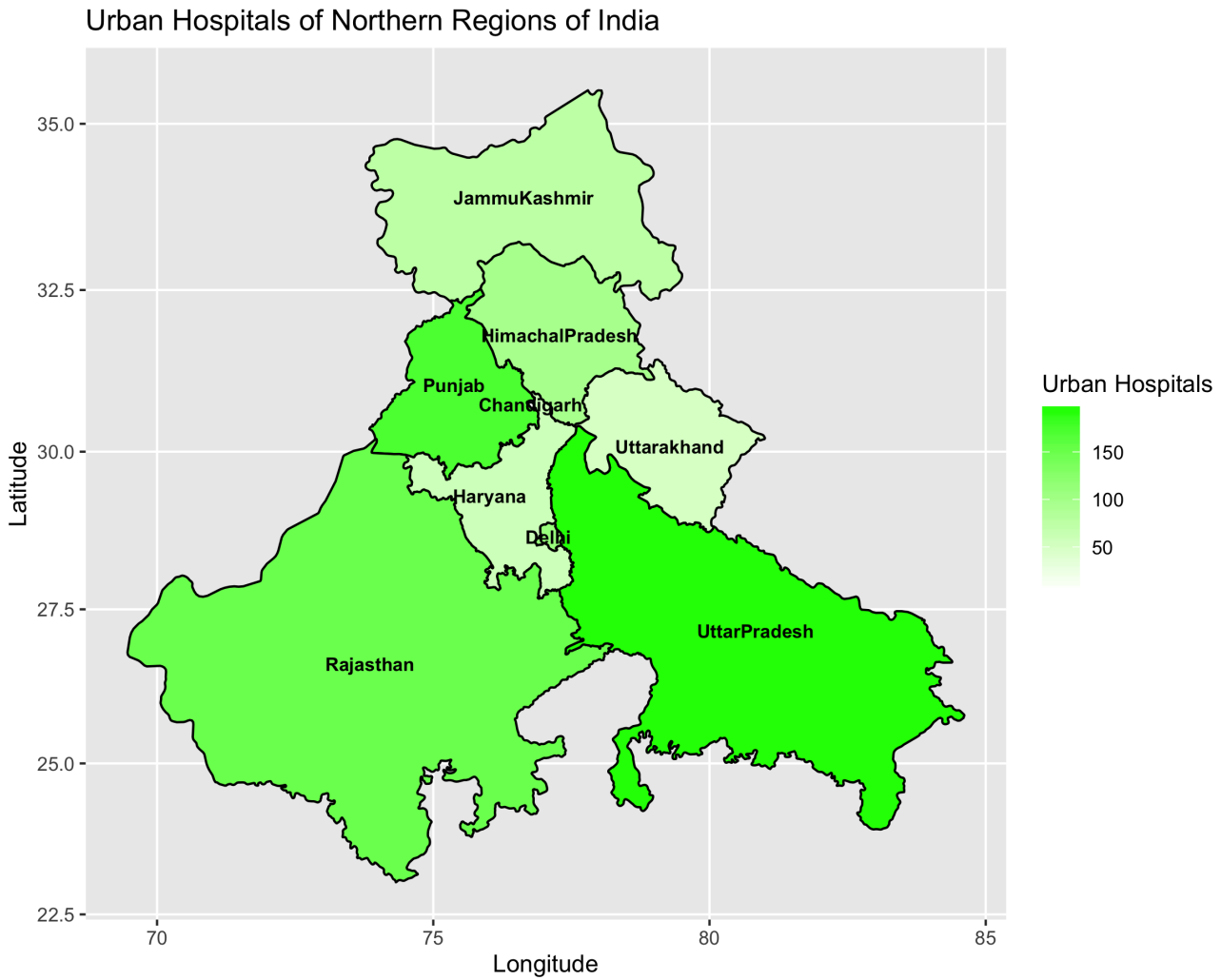
Exploratory Data Analysis: Demographic features (2)

Large rural, small urban states



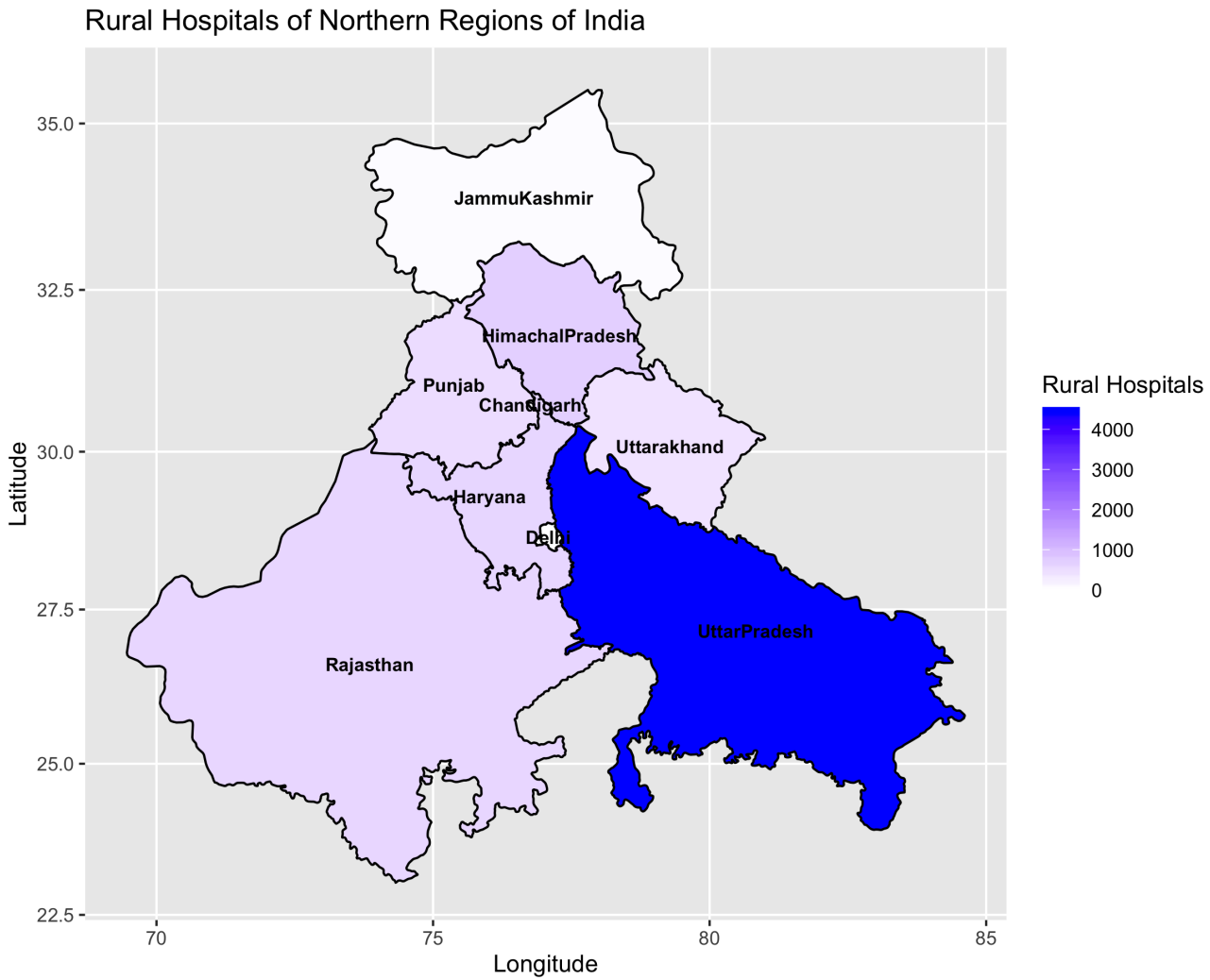
Exploratory Data Analysis: Health system organization (1)

The city is still the center of healthcare



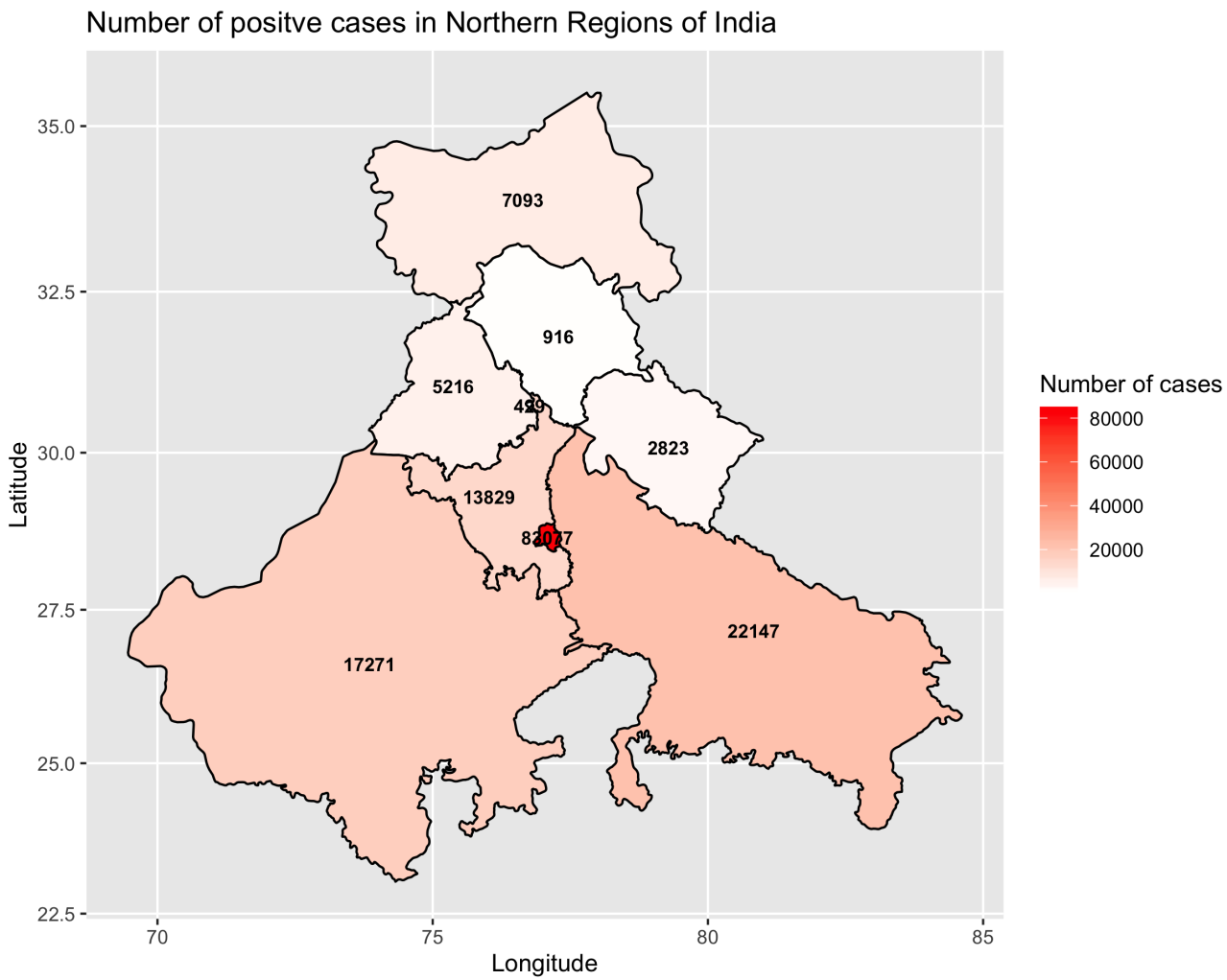
Exploratory Data Analysis: Health system organization (2)

The city is still the center for healthcare



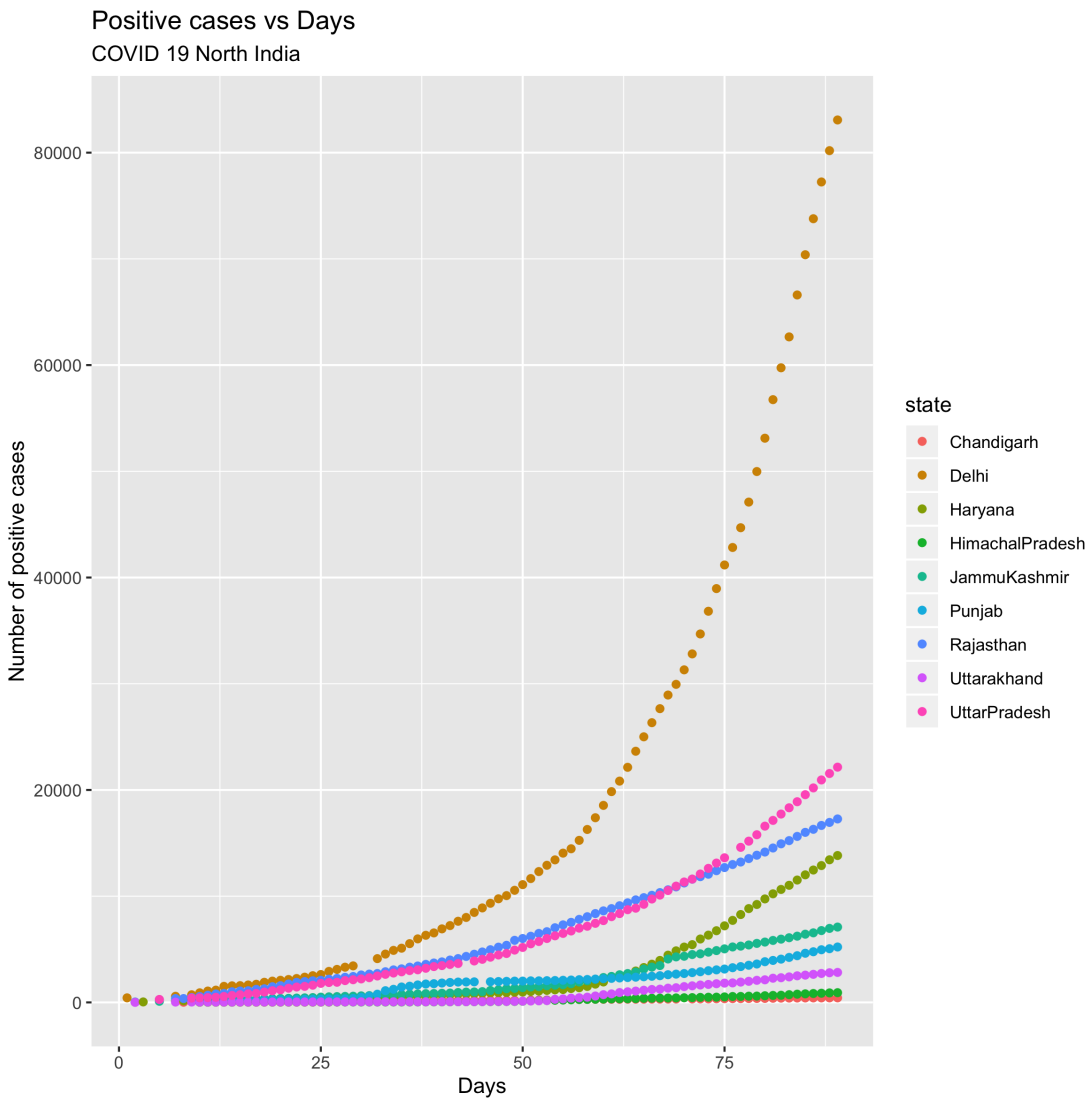
Exploratory Data Analysis: # positive cases per region

Not always a matter of population / landsize



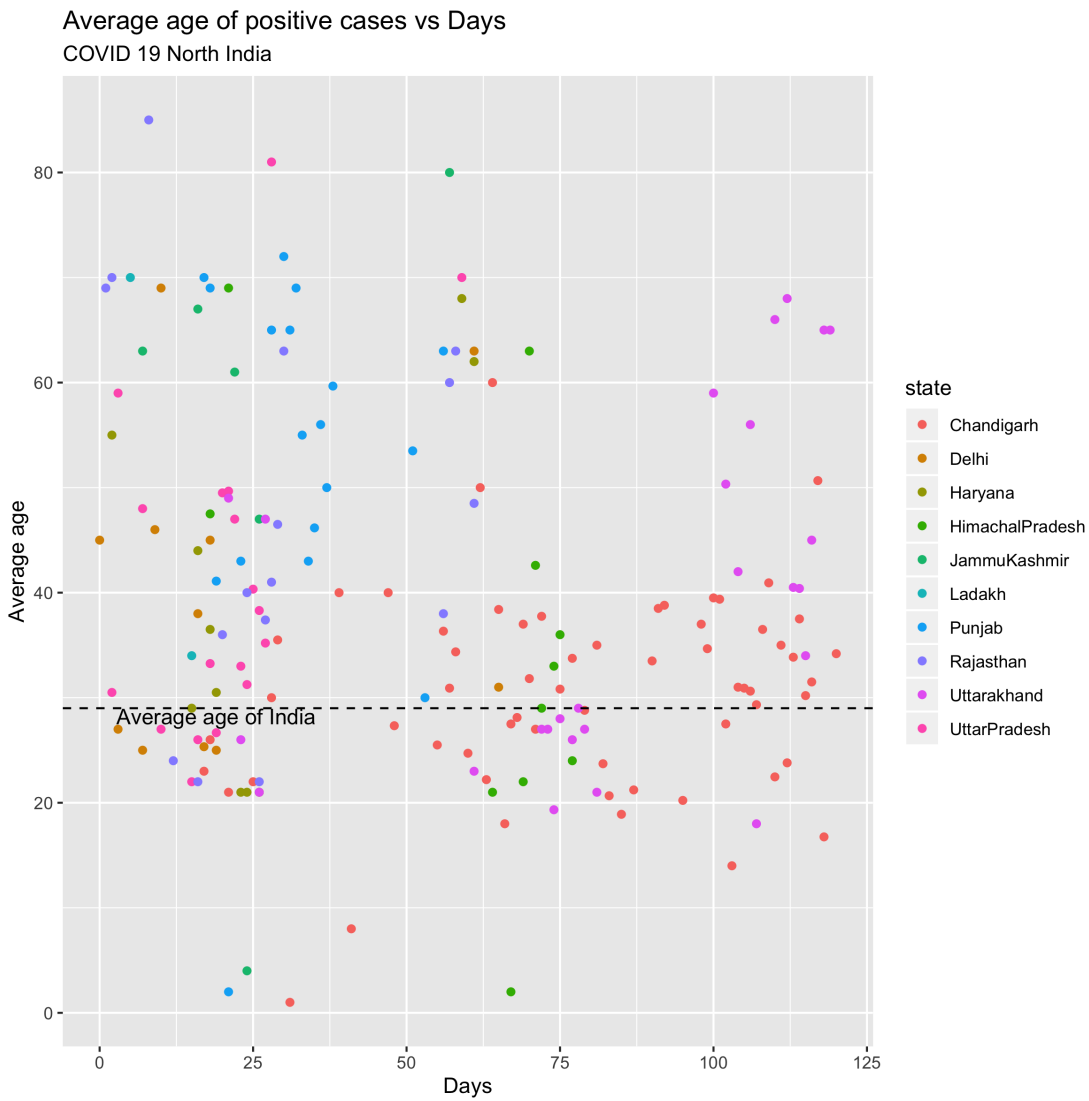
Exploratory Data Analysis: Trend of positive cases

An exponential trend, maybe?



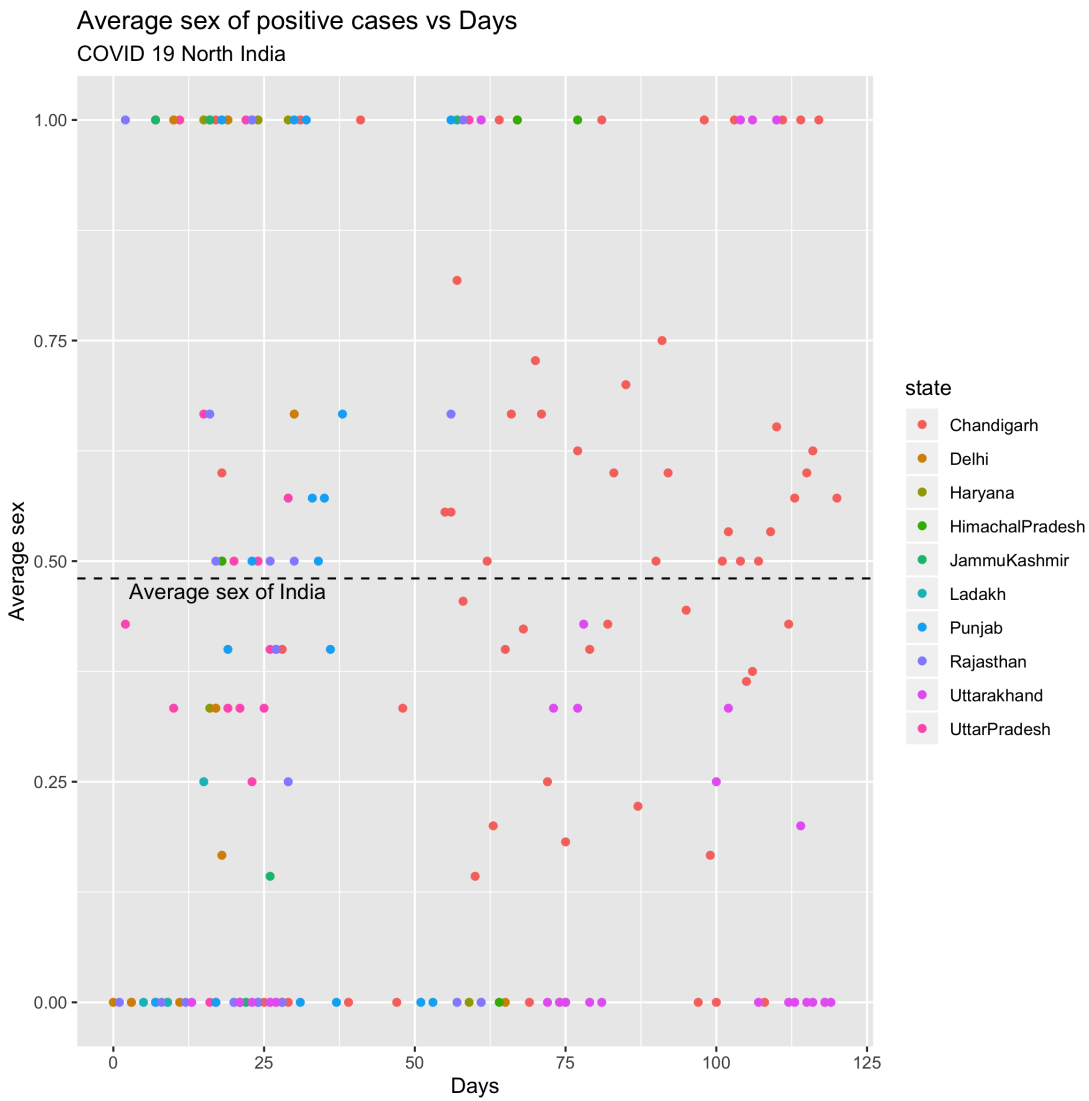
Exploratory Data Analysis: Trend of average age

The more vulnerable suffer the most



Exploratory Data Analysis: Trend of average sex

No evident sex prevalence



We are here...

GLM: *A recap*

Why not just linear?

- Difficult modeling of **epidemiological dynamics** with (mean) linear responses;
- Nonlinear trends confirmed by visual exploratory analysis.

GLM: *A recap*

Why not just linear?

- Difficult modeling of **epidemiological dynamics** with (mean) linear responses;
- Nonlinear trends confirmed by visual exploratory analysis.

A **Linear Model** for regression:

$$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\beta$$

GLM: *A recap*

Why not just linear?

- Difficult modeling of **epidemiological dynamics** with (mean) linear responses;
- Nonlinear trends confirmed by visual exploratory analysis.

A **Linear Model** for regression:

$$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\beta$$

A **Generalized Linear Model** for regression:

$$\mathrm{E}[\mathbf{Y}] = \mathrm{g}^{-1}(\mathbf{X}\beta)$$

$$\mathrm{Var}[\mathbf{Y}] = \phi V(\mathrm{E}[\mathbf{Y}])$$

GLM for count data:

Choice of family and link function

Poisson:

$$\mathbf{Y} \sim \text{Poisson}(\lambda)$$

$$f(\mathbf{Y}; \lambda) = \frac{e^{-\lambda}\lambda^{\mathbf{Y}}}{\mathbf{Y}!}$$

Negative Binomial:

$$\mathbf{Y} \sim \text{NBi}(k, \alpha)$$

$$f(\mathbf{Y}; k, \alpha) = \binom{\mathbf{Y} + k - 1}{k - 1} \frac{\alpha^{\mathbf{Y}}}{(1+\alpha)^{\mathbf{Y}+k}}$$

GLM for count data:

Choice of family and link function

Poisson:

$$\mathbf{Y} \sim \text{Poisson}(\lambda)$$

$$f(\mathbf{Y}; \lambda) = \frac{e^{-\lambda} \lambda^{\mathbf{Y}}}{\mathbf{Y}!}$$

Negative Binomial:

$$\mathbf{Y} \sim \text{NBi}(k, \alpha)$$

$$f(\mathbf{Y}; k, \alpha) = \binom{\mathbf{Y} + k - 1}{k - 1} \frac{\alpha^{\mathbf{Y}}}{(1 + \alpha)^{\mathbf{Y}+k}}$$

Log link function $\rightarrow$ *exponential* dynamics.

GLM: *Predictors and Interactions* (1)

A matter of choosing **what** to see

General criterion ($\leftarrow$ *NB, P, QP*):

- Add one/some predictor(s);
- *5-fold CV cumulative MAE* and *validation cumulative MAE;*
- Significance analysis on the linear coefficients.

Starting with no *lagged* component, all *non-interacting* predictors:
$\rightarrow$ **sequential** predictor addition always produces $r\mathcal{D}ev$ decrease;
$\rightarrow$ max. **overall** significance for all parameters, for every model family.

GLM: *Predictors and Interactions* (2)

A matter of choosing **how** to see

General criterion, but also:

- Max. 2^{nd}-order interaction;
- *Non sunt multiplicanda entia sine necessitate*.

Chosen interactions ($\leftarrow$ *P*, *QP*):

- Density:Urbanization $\rightarrow$ the *vertical city* hypothesis;
- Males:Urbanization $\rightarrow$ the *male migration* hypothesis in India.

The case of Urbanization: *significance absorption* and *predictive accuracy*.

GLM: *Predictors and Interactions* (3)

A matter of choosing **when** to see

Introduction of *lagged* components:

- Justified theoretically by phenomenon structure;
- *But will they really work?*

General criterion ($\leftarrow$ *NB*, *P*, *QP*):

- Add one/some *lagged* component(s);
- *Validation cumulative MAE*;
- Significance analysis on the *lagged*-component coefficients.

Results:

$\rightarrow$ No *cross-regressive lagged* predictor is useful;

$\rightarrow$ Autoregressing on response (lagged at 10 days) helps P/QP.

GLM: *Predictors and Interactions* (4)

A picture is worth a thousand words: the final model (QP)

```
## Call:
## glm(formula = deltapos ~ date + urbanh + ruralh + Urbanization +
##     Density + Females + Density:Urbanization + Males:Urbanization +
##     lagged_deltapos + Males, family = quasipoisson(link = log),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -13.944   -4.032   -1.447    2.046   33.278
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.012e+00  3.721e-01   8.095 3.47e-15 ***
## date                   3.838e-02  1.810e-03  21.203  < 2e-16 ***
## urbanh                -1.231e-02  1.420e-03  -8.670  < 2e-16 ***
## ruralh                -1.281e-03  2.408e-04  -5.319 1.50e-07 ***
## Urbanization          -1.845e+00  1.587e+00  -1.163  0.24539
## Density               -3.614e-03  1.204e-03  -3.000  0.00281 **
## Females               -1.264e-06  2.231e-07  -5.664 2.34e-08 ***
## lagged_deltapos        2.717e-04  8.976e-05   3.027  0.00258 **
## Males                  1.084e-06  2.121e-07   5.109 4.42e-07 ***
## Urbanization:Density   3.497e-03  1.111e-03   3.149  0.00173 **
## Urbanization:Males     7.899e-07  8.667e-08   9.114  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 43.24238)
##
##     Null deviance: 200886  on 581  degrees of freedom
## Residual deviance:  18930  on 571  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

GLM: *Predictors and Interactions* (5)

A picture is worth a thousand words: the MLE model (P)

```
## Call:
## glm(formula = deltapos ~ date + urbanh + ruralh + Urbanization +
##     Density + Females + Density:Urbanization + Males:Urbanization +
##     lagged_deltapos + Males, family = poisson(link = log), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -13.944   -4.032   -1.447    2.046   33.278
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           3.012e+00  5.658e-02  53.231  < 2e-16 ***
## date                  3.838e-02  2.753e-04 139.428  < 2e-16 ***
## urbanh               -1.231e-02  2.160e-04 -57.014  < 2e-16 ***
## ruralh               -1.281e-03  3.661e-05 -34.975  < 2e-16 ***
## Urbanization         -1.845e+00  2.414e-01  -7.646 2.07e-14 ***
## Density              -3.614e-03  1.832e-04 -19.731  < 2e-16 ***
## Females              -1.264e-06  3.393e-08 -37.247  < 2e-16 ***
## lagged_deltapos       2.717e-04  1.365e-05  19.905  < 2e-16 ***
## Males                 1.084e-06  3.226e-08  33.596  < 2e-16 ***
## Urbanization:Density  3.497e-03  1.689e-04  20.706  < 2e-16 ***
## Urbanization:Males    7.899e-07  1.318e-08  59.934  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 200886  on 581  degrees of freedom
## Residual deviance:  18930  on 571  degrees of freedom
## AIC: 22039
##
## Number of Fisher Scoring iterations: 5
```

GLM: *Predictors and Interactions* (6)

A picture is worth a thousand words: the case of Urbanization

```
## Call:
## glm(formula = deltapos ~ date + urbanh + ruralh + Urbanization +
##     Density + Females + Density:Urbanization + lagged_deltapos +
##     Males, family = quasipoisson(link = log), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -13.921   -4.729   -1.894    1.524   32.076
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.367e+00  3.326e-01  13.132  < 2e-16 ***
## date                  3.686e-02  2.088e-03  17.650  < 2e-16 ***
## urbanh               -3.505e-03  1.281e-03  -2.736 0.006407 **
## ruralh               -3.598e-03  2.152e-04 -16.719  < 2e-16 ***
## Urbanization         -1.174e+01  1.297e+00  -9.051  < 2e-16 ***
## Density               4.991e-03  9.375e-04   5.324 1.46e-07 ***
## Females              -2.409e-06  2.451e-07  -9.827  < 2e-16 ***
## lagged_deltapos       3.829e-04  1.042e-04   3.676 0.000259 ***
## Males                 2.340e-06  2.294e-07  10.204  < 2e-16 ***
## Urbanization:Density -4.250e-03  8.845e-04  -4.805 1.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 59.32441)
##
##     Null deviance: 200886  on 581  degrees of freedom
## Residual deviance:  23189  on 572  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

GLM: *Determining prediction intervals*

A *bootstrap* way

Generally, prediction intervals for GLMs are not easily computable...

Bootsrap-based solution:

- Fit the GLM, and collect the regression statistics $\hat{\beta}$ and $\text{Cov}(\hat{\beta})$:
- Simulate M draws of $\hat{\beta}^* \sim N(\hat{\beta}, \hat{\text{Cov}}(\hat{\beta}))$;
- Simulate $y^*|x$ from response $g^{-1}(x\hat{\beta}^*)$ and a variance determined by the response distribution;
- Determine the $\alpha/2$ and $1-\alpha/2$ empirical quantiles of the simulated response $y^*|x$ for each x.

Possibility to reparametrize a Quasipoisson model into a Negative Binomial using $\text{QP}(\mu, \theta) = \text{NegBin}(\mu, \theta = \frac{\mu}{\phi - 1})$.

MARS: *A recap* (1)

A *nice* piecewise linear model (Friedman, 1991)

Multivariate Adaptive Regression Splines (MARS) model response as a weighted sum of basis functions

$$\widehat{f}(x) = \sum_{i=1}^{k} c_i B_i(x).$$

The basis functions $B_i(x)$ takes one of the three forms:

- a constant 1;
- hinge function: $\max(0, x - k)$ or $\max(0, k - x)$, k : "knot";
- product of two or more hinge functions.

MARS: *A recap* (2)

The model building process

Two-stage approach:

- *Forward pass*: incrementally adding basis function that gives maximum reduction in sum-of-squares residual error;
- *Backward pass*: delete least effective term using GCV criterion.

HP-tuning:
$\rightarrow$ grid-search, 5-fold CV minimizing the cumulative MAE:

- degree: 1 by 1, from 1.

MARS: *A recap* (3)

Why that?

A non parametric regression method.

Pros:

- More flexible than linear models;
- Simple to understand and interpret compared to other ML-based solutions;
- Automated modelling of interactions.

Cons:

- The resulting fitted function is not smooth (not differentiable along predictors);
- *Greedy optimization vs. global optimization*.

MARS: *Predictors and Interactions*

A summary

```
## Call: earth(formula=deltapos~date+urbanh+ruralh+Urbanization+Density+F...),
##             data=train, degree=3)
##
##                                                                    coefficients
## (Intercept)                                                          281.615115
## h(lagged_deltapos-12)                                                 -0.681952
## h(272-lagged_deltapos)                                                -1.015568
## h(Urbanization:Density-8997.87)                                        0.230962
## h(57-date) * h(lagged_deltapos-272)                                    0.059451
## h(date-57) * h(lagged_deltapos-272)                                    0.051180
## h(date-25) * h(8997.87-Urbanization:Density)                           0.000563
## h(48-date) * h(Urbanization:Density-8997.87)                          -0.006039
## h(date-48) * h(Urbanization:Density-8997.87)                           0.032620
## h(lagged_deltapos-12) * h(Urbanization:Density-49.989)                -0.000041
## h(date-25) * h(Urbanization-0.348781) * h(8997.87-Urbanization:Density) -0.018028
## h(date-25) * h(8997.87-Urbanization:Density) * h(Urbanization:Males-4.70575e+06)  0.000000
## h(date-25) * h(8997.87-Urbanization:Density) * h(4.70575e+06-Urbanization:Males)  0.000000
##
## Selected 13 of 18 terms, and 5 of 10 predictors
## Termination condition: Reached maximum RSq 0.9990 at 18 terms
## Importance: date, Urbanization:Density, lagged_deltapos, Urbanization, ...
## Number of terms at each degree of interaction: 1 3 6 3
## GCV 7082.384    RSS 3694534    GRSq 0.9347155    RSq 0.9412833
```

MARS: *Determining prediction intervals*

Residual model (Stephen Milborrow, 2019)

We do the following:

- Estimate the mean absolute error at each point using a *residual model*;
- Assuming normality, re-scale the error to an estimated standard deviation: $sd \simeq \sqrt{\frac{\pi}{2}}\mathrm{E}[|\epsilon|]$;
- Convert the standard deviation to an estimated prediction interval for a given level α:

$$[\hat{Y} - z_{\alpha/2}sd, \hat{Y} + z_{\alpha/2}sd].$$

RF: *A recap* (1)

Harnessing the *wisdom of the crowd*

Based on the **decision tree** (regression) model and learning algorithm.

Enhancements:

- **Bootstrap** sampling on *training set*;
- **Random** feature selection;
- **Ensembling** of aggregates.

HP-tuning:
$\rightarrow$ grid-search, *5-fold CV* test *cumulative MAE*:

- mtry: 1 by 1, from 1 $\rightarrow$ 5;
- num.trees: 10 by 10, from 500 $\rightarrow$ 850.

RF: *A recap* (2)

Why that?

On the opposite side of the spectrum w.r.t. GLMs.

Pros:

- Typical *fire and forget* **ML**-based solution;
- Often requires minimal *HP-tuning*;
- *Automated* modeling of **interactions**.

Cons:

- Only averaged or modal **stepwise** regression capability;
- Difficult to interpret;
- May sometimes overfit or fit noise;
- **Non-deterministic** behavior and seed-dependence.

RF: *Determining prediction intervals*

Always beware of absolute certainty! (Zhang et al., 2019)

A well-acknowledged problem for *tree-based* models...

...and a new proposal:

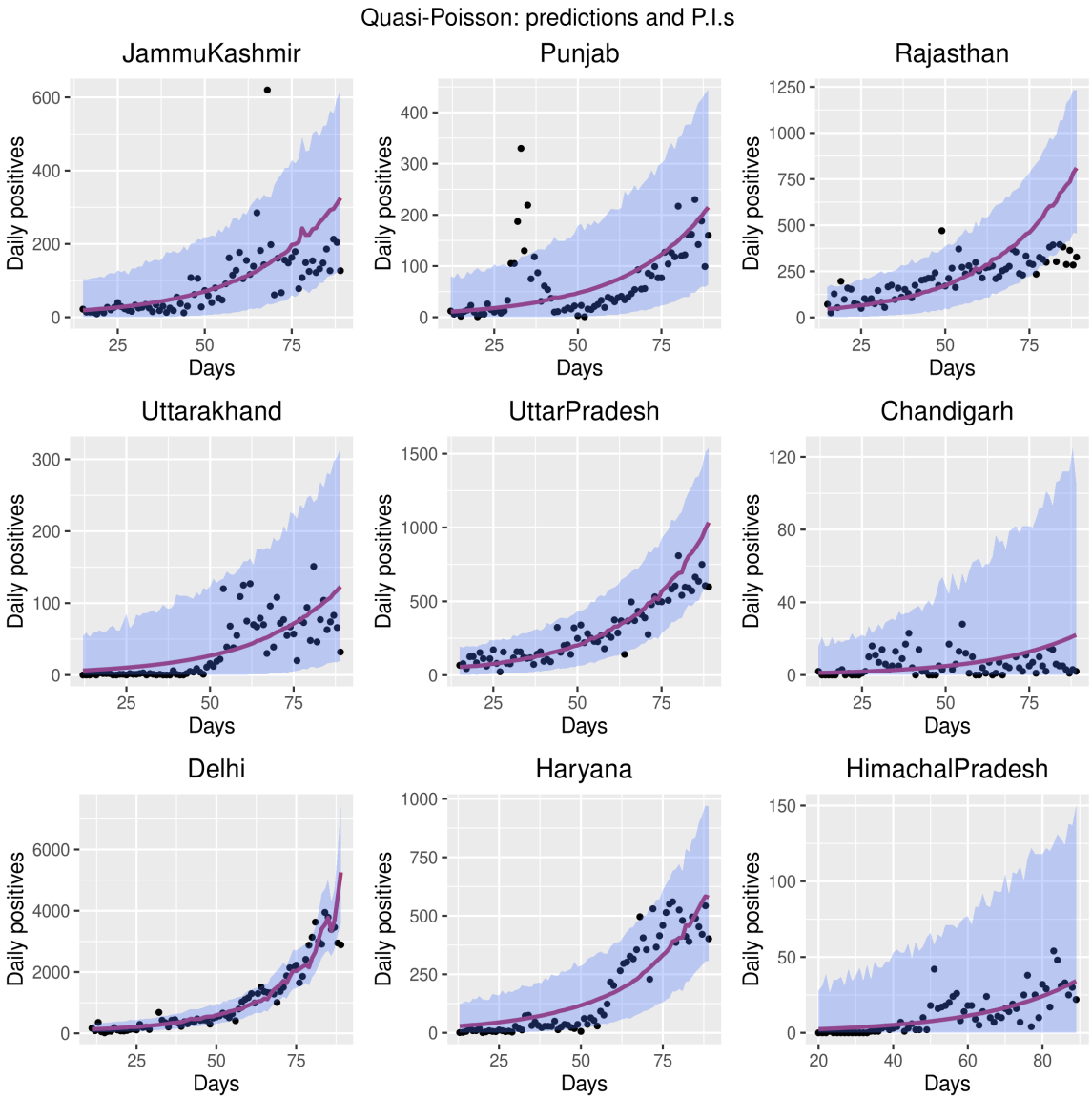
$$\left[\hat{Y} + D_{[n,\alpha/2]}, \hat{Y} + D_{[n,1-\alpha/2]}\right]$$

- Uses the *empirical quantile distribution* of **OOB prediction** errors;
- Well-grounded from the **asymptotic convergence** p.o.v.;
- Easy to compute, often partially pre-computed *for free*;
- R implementation already available!

We are here...

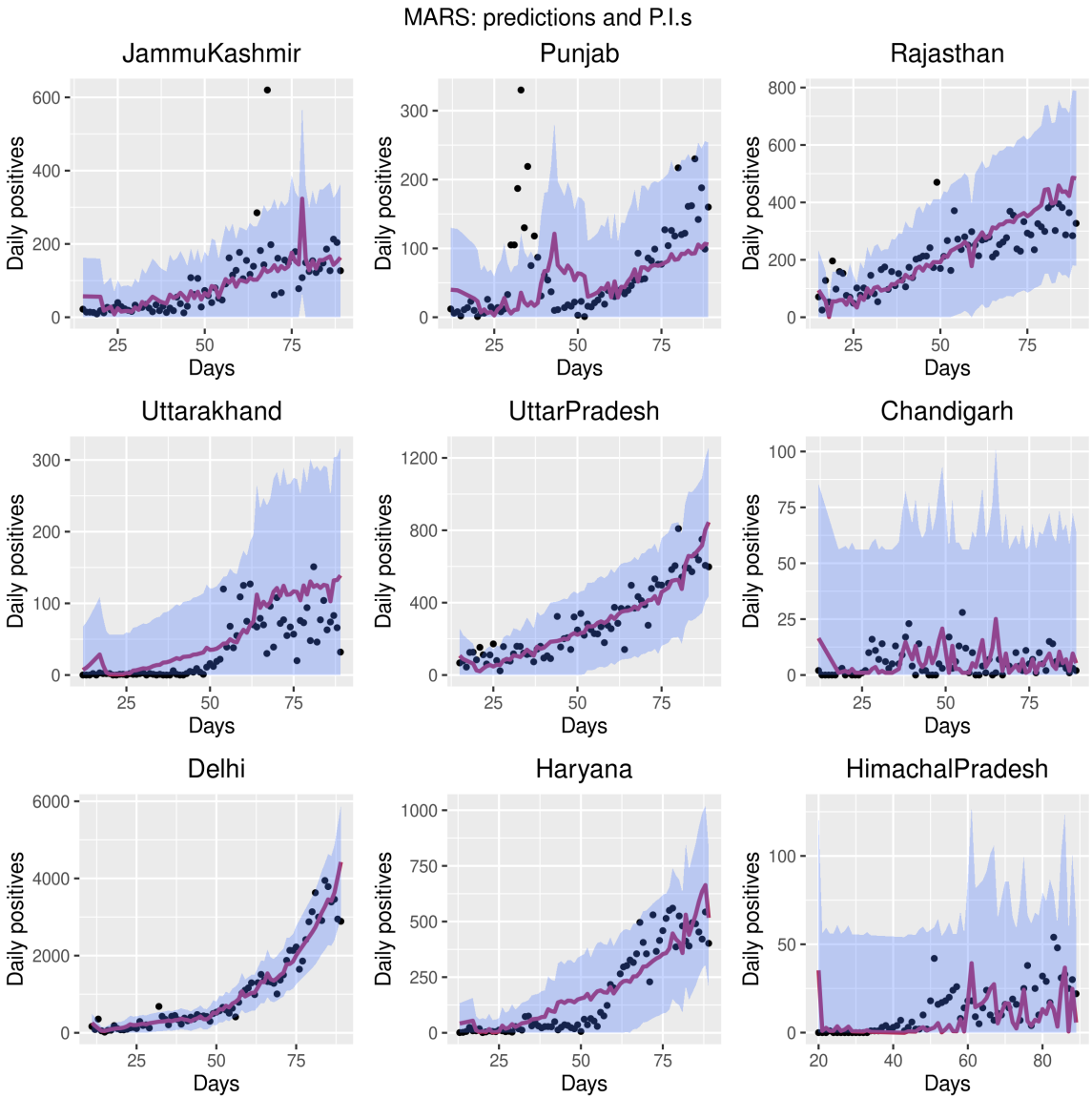
QuasiPoisson model

Predictions and Prediction Intervals



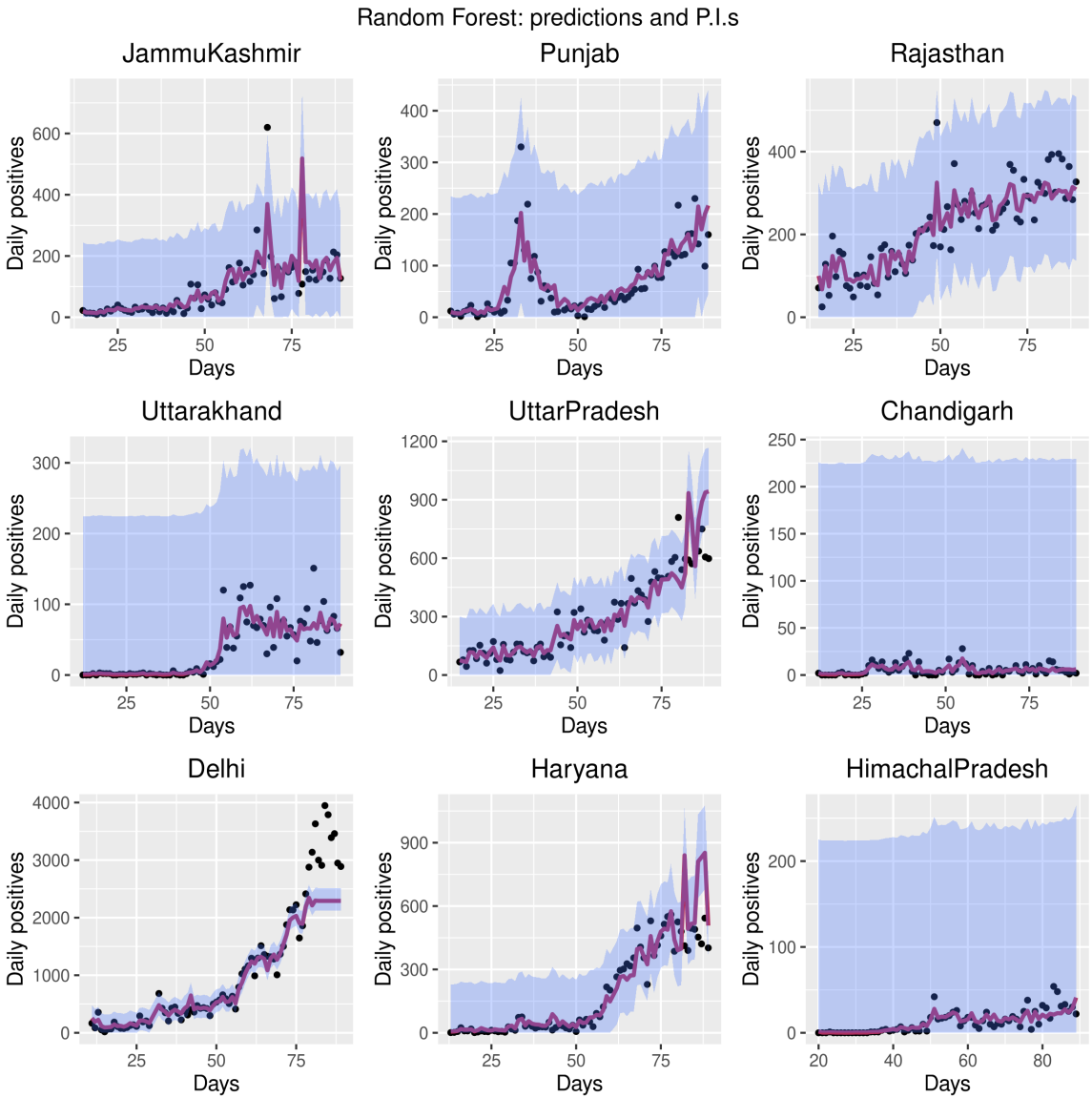
MARS model

Predictions and Prediction Intervals



Random Forest model

Predictions and Prediction Intervals



Accuracy measure: RMSE

Accuracy is not everything

RMSE of each model

COVID 19 North India

State	QuasiPoisson	MARS	RandomForest
Chandigarh	14.43	5.08	4.89
Delhi	986.90	725.12	1106.27
Haryana	92.15	118.03	251.48
Himachal Pradesh	12.17	22.98	15.61
Jammu and Kashmir	121.39	33.06	29.83
Punjab	50.47	73.98	55.57
Rajasthan	345.16	112.52	62.04
Uttarakhand	45.54	60.02	31.38
Uttar Pradesh	244.91	137.00	251.15

We are here...

Final remarks

ML vs. Stats doesn't belong to the real world

Parametric model:

- QuasiPoisson model correctly captures the pandemic trend;
- On average, it has a higher RMSE than non parametric models, partly due to noisy data.

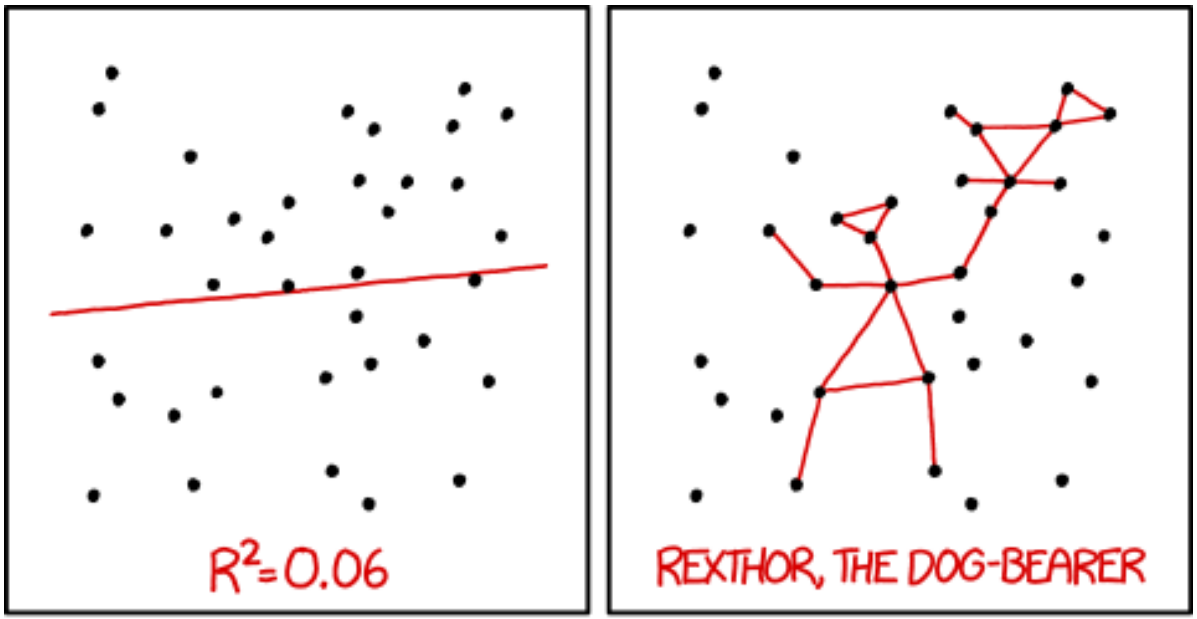
Non parametric models:

- Sometimes MARS and Random Forest favor local over global trends;
- Random Forest predictions are unstable.

Trade-off between analyst's *cognitive burden* and model ability to grasp deep inter-connections about reality.

Greetings

Thank you for your attention!



[https://github.com/emaballarin/northindia-covid-smds]