
COVID-19:
The north colombia outbreak

Introduction

1) The COVID timeline

2) Colombia, South America and Italy

3) The north of Colombia

4) Infographics and EDA

5) The dataset(s)

Methods & Motivations

6) Evaluation procedure

7) The Generalized Linear model

8) Beyond linear models

9) The Generalized Additive Model

Results & Conclusions

10) Comments

11) How we could have improved if we had more time

The COVID-19 timeline

Huanan Seafood

31/12/2019

Wholesale Market closes after Chinese Health officials registered a cluster of 41 patients with a mysterious pneumonia.

First death caused by the nCOV

11/01/2020

Chinese authorities identify a new type of coronavirus (called novel coronavirus). Few days after, China records its first death.

Global emergency

30/01/2020

The WHO declares a global public-health emergency 1 week after Wuhan is placed under quarantine.

Italy

21/02/2020

Italy outbreak begins. Cases are spiking in both Iran and South Korea.

Colombia

06/03/2020

First case is reported in Colombia

The IT lockdown

08/03/2020

Italy places all 60 million residents on lockdown.

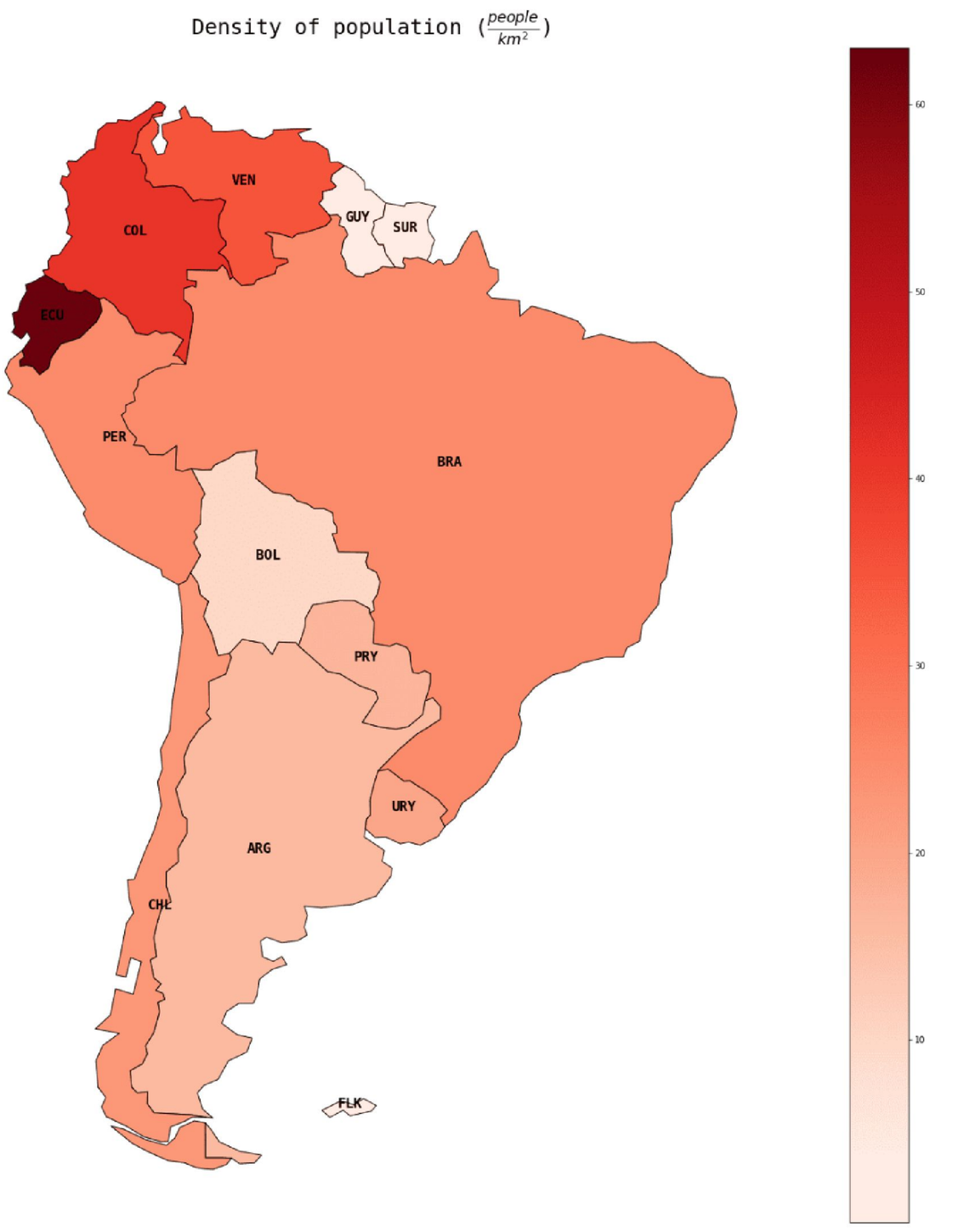
Colombia lockdown

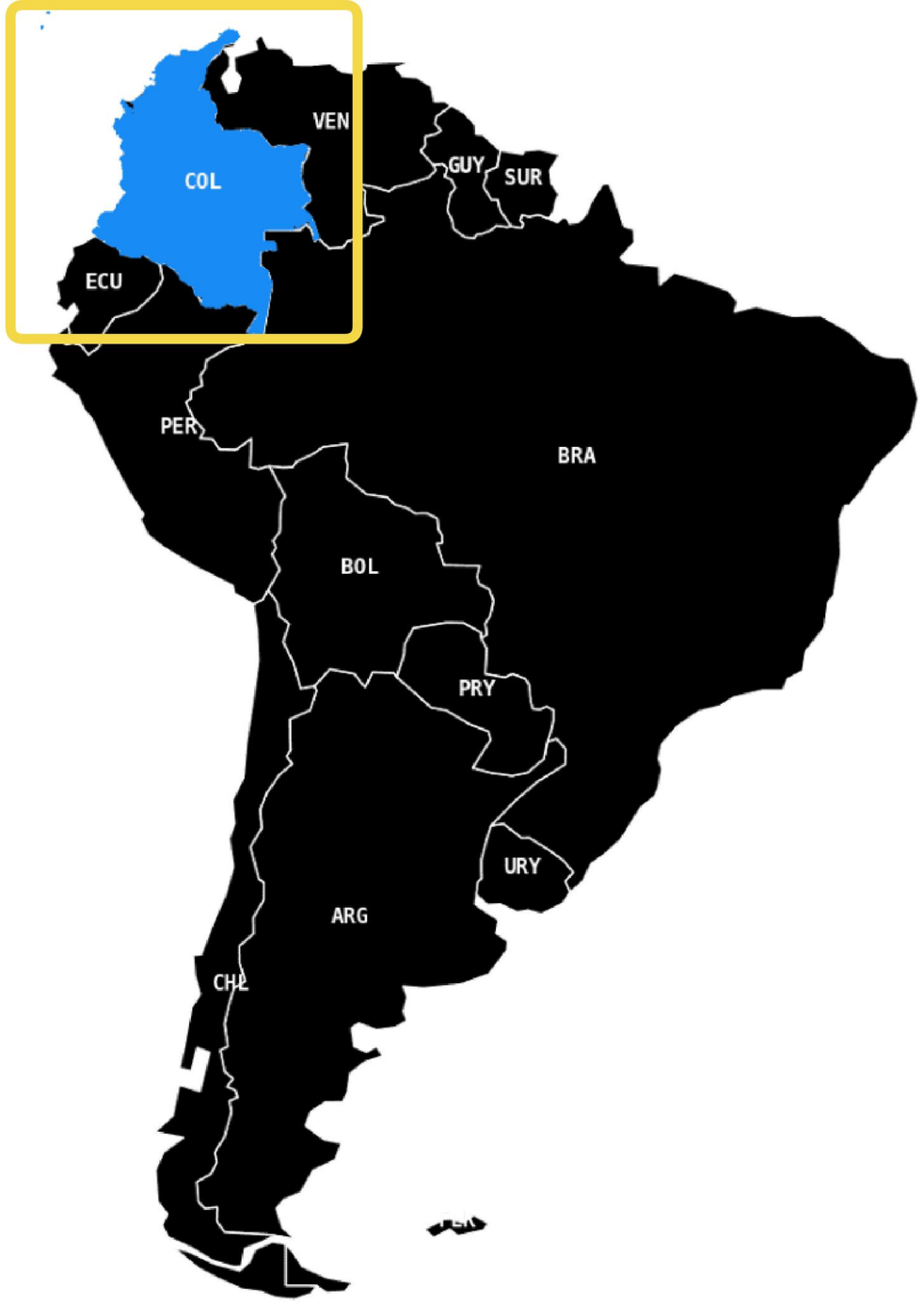
24/03/2019

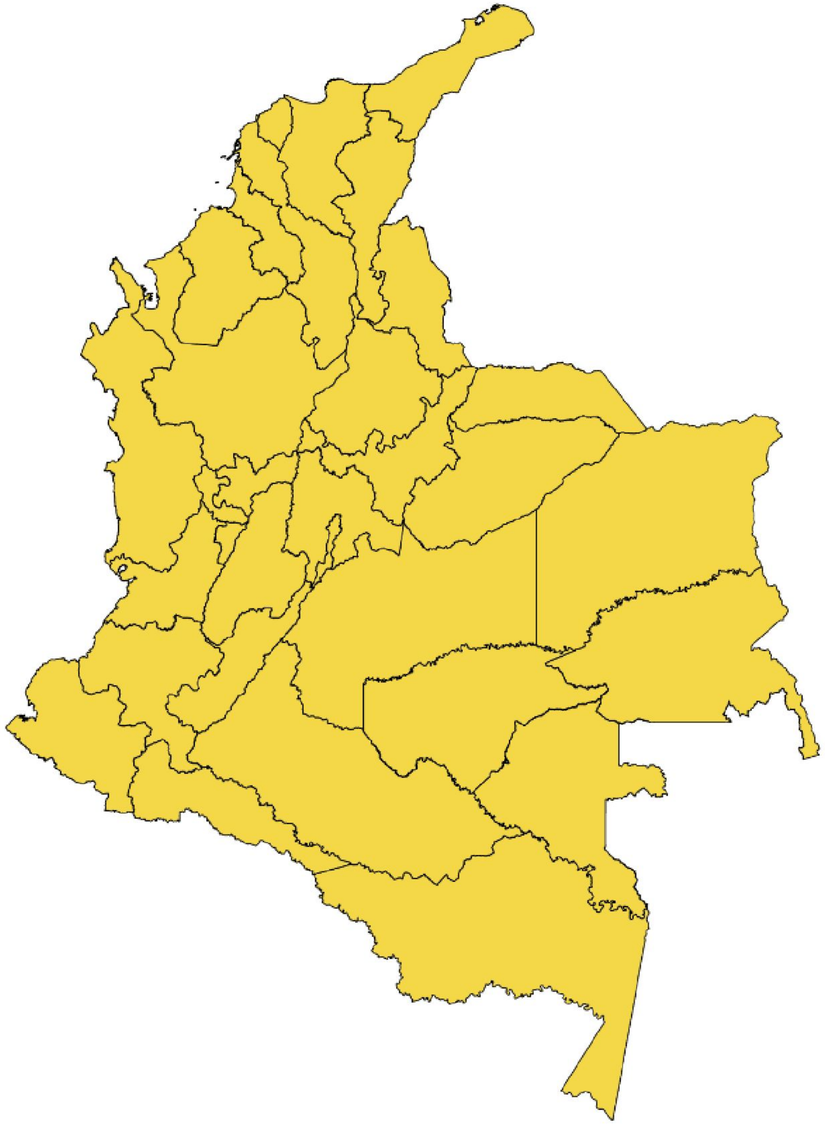
The Colombian president announce a 19-day nationwide quarantine, starting on 24 March.

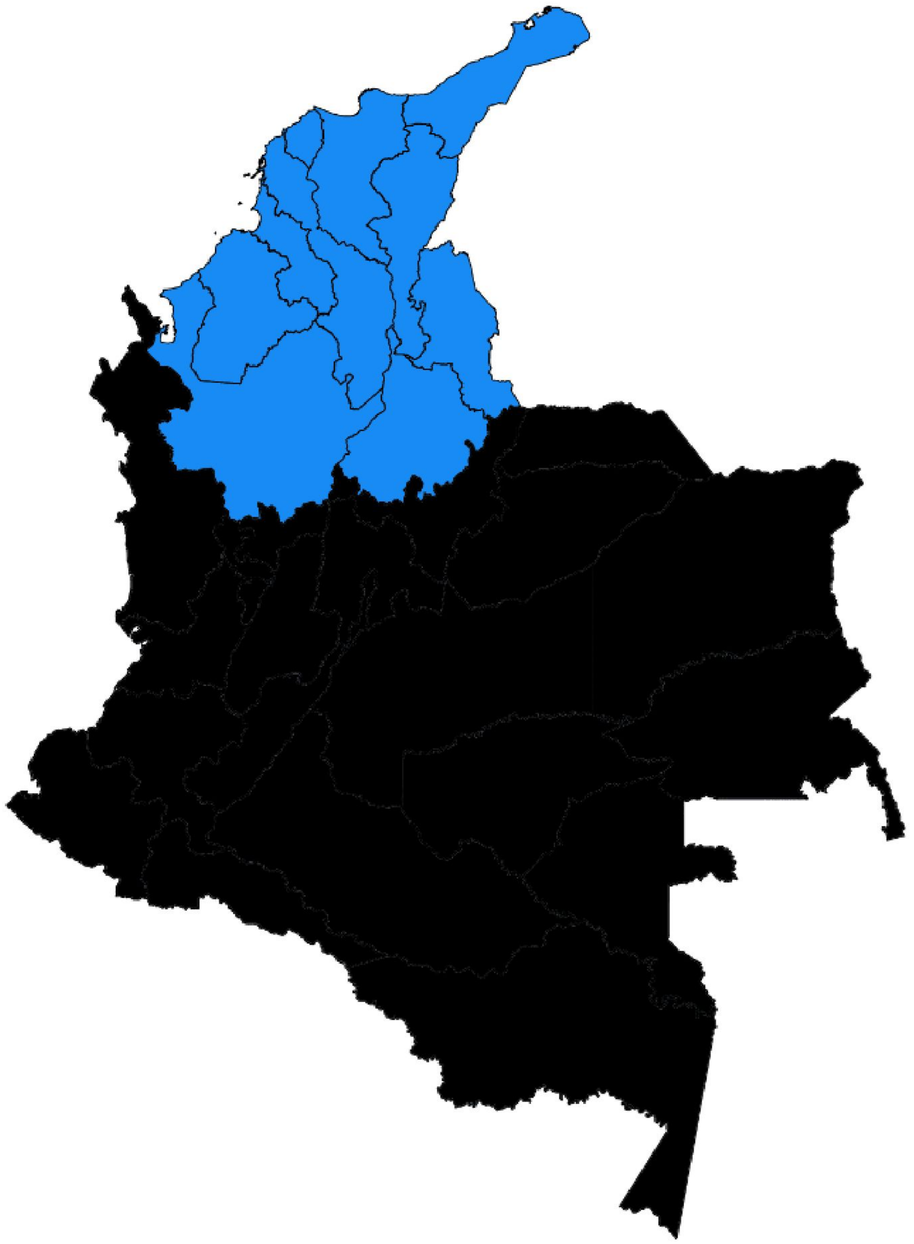
Source:
Wikipedia

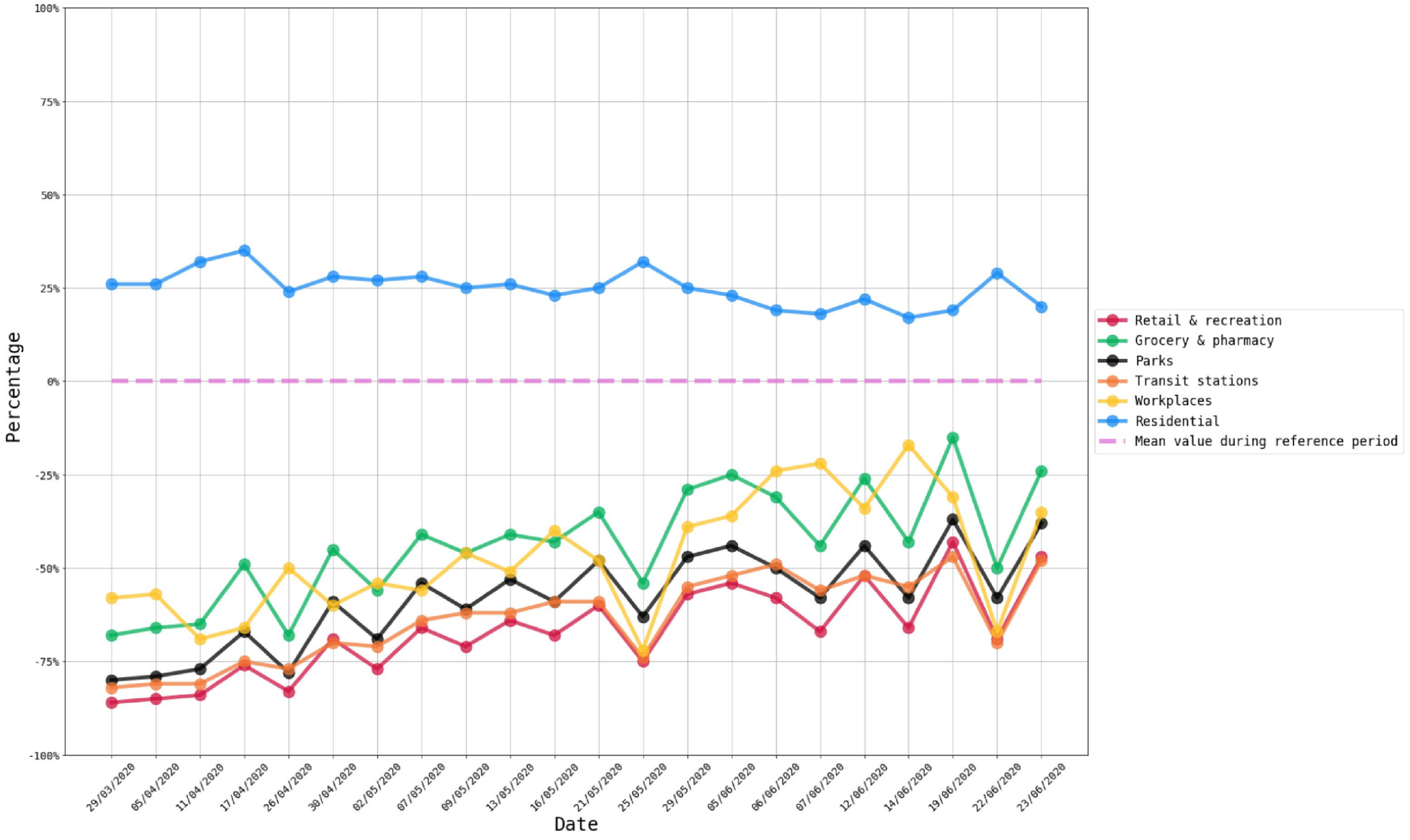

COL
VEN
GUY
SUR
ECU
PER
BRA
BOL
PRY
URY
ARG
CHL


Density of population ($\frac{people}{km^2}$)
COL
VEN
GUY
SUR
ECU
PER
BRA
BOL
PRY
URY
ARG
CHL
FLK
60
50
40
30
20
10

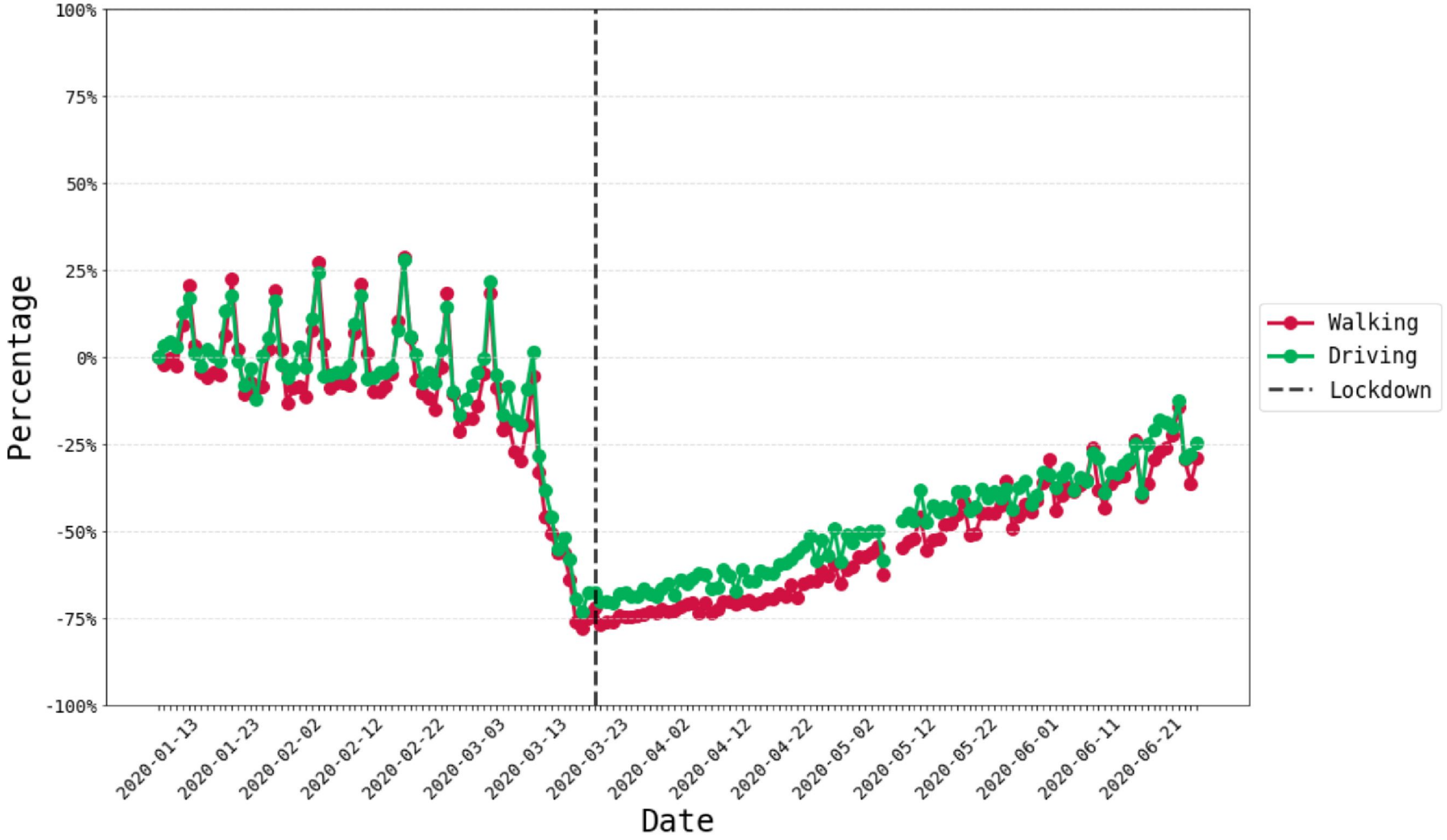

VEN
COL
GUY
SUR
ECU
PER
BRA
BOL
PRY
URY
ARG
CHL







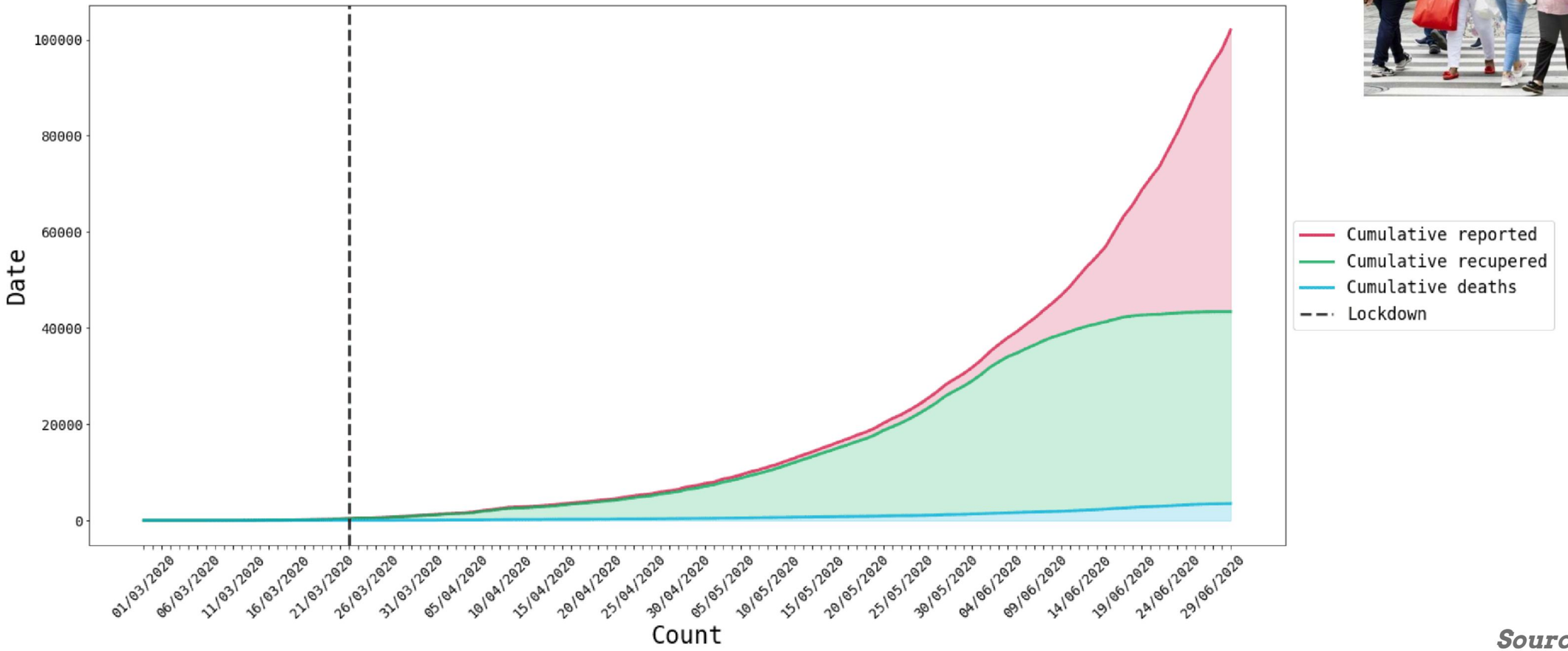
Source:
GOOGLE COVID-19 Community Mobility Report



Source:
APPLE Mobility Trends Reports

At which point are we now?

100.000 casos de covid,un llamado al autocuidado



Sources:
Instituto Nacional de Salud
elCOLOMBIANO

ITALY

SURFACE: 301,346 square km

POPULATION: 60 million

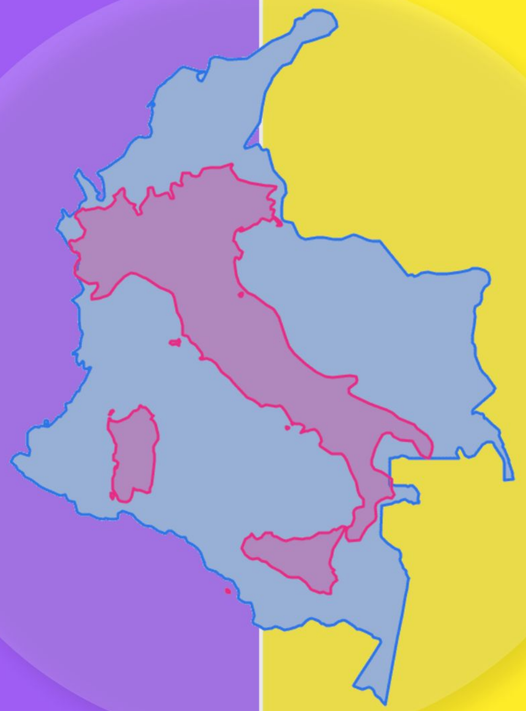
AVERAGE AGE: 45

DEMOGRAPHIC:

0-14 years: 13.6%
15-24 years: 9.61%
25-54 years: 41.82%
55-64 years: 13.29%
65 years and over: 21.69%

URBANIZATION: 60%

CITIES WITH MORE THAN
500'000 INHABITANTS : 6



COLOMBIA

SURFACE: 1,1417,8 square km

POPULATION: 50 million

AVERAGE AGE: 31

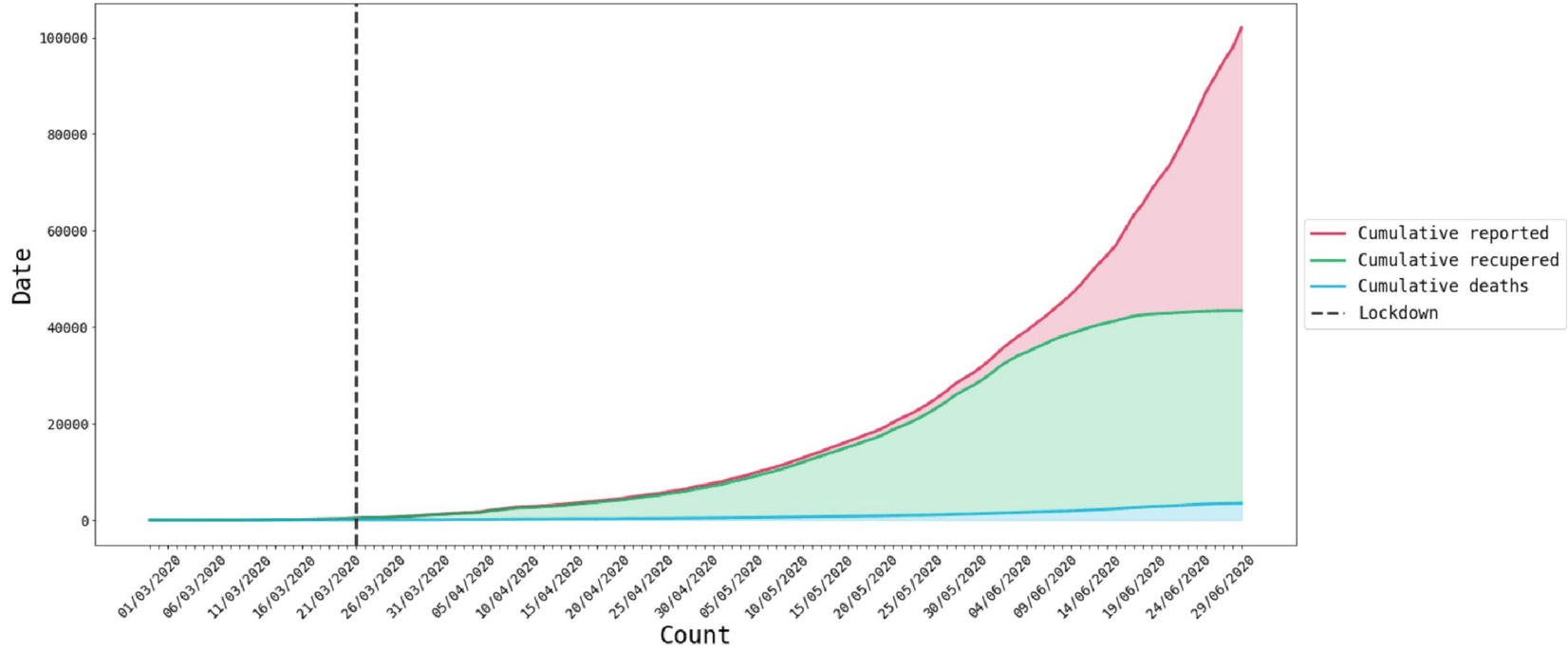
DEMOGRAPHIC:

0-14 years: 22.2%
15-64 years: 68.8%
65 years and over: 9.1%

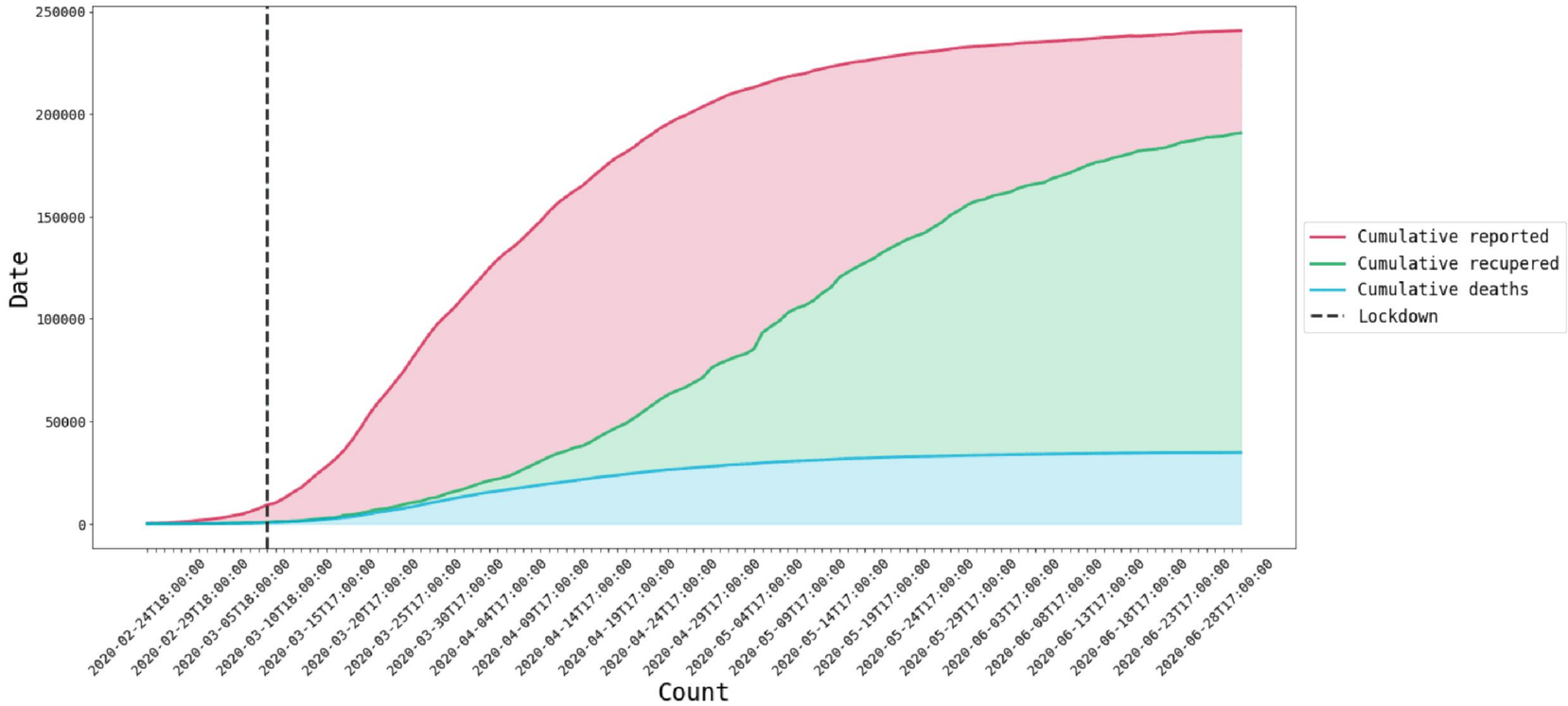
URBANIZATION: 71%

CITIES WITH MORE THAN
500'000 INHABITANTS : 12

COLOMBIA



ITALY



Speaking about the dataset ...

covid-19-colombia-confirmed.csv (2.61 KB)

Detail | Compact | Column

10 of 29 columns

state	2020-03-...	2020-03-...	2020-03-...	2020-03-...	2020-03-...	2020-03-...	2020-03-
Bogota	1	1	1	1	1	3	3
Amazonas	0	0	0	0	0	0	0
Antioquia	0	0	0	1	1	4	4
Arauca	0	0	0	0	0	0	0
Atlantico	0	0	0	0	0	0	0
Bolivar	0	0	0	0	0	1	1
Boyaca	0	0	0	0	0	0	0
Caldas	0	0	0	0	0	0	0
Caqueta	0	0	0	0	0	0	0
Casanare	0	0	0	0	0	0	0
Cauca	0	0	0	0	0	0	0
Cesar	0	0	0	0	0	0	0

“It is a capital mistake to theorize before one has data.”

Sherlock Holmes

Data finding, cleaning and exploration:

DATA SOURCES:

GitHub

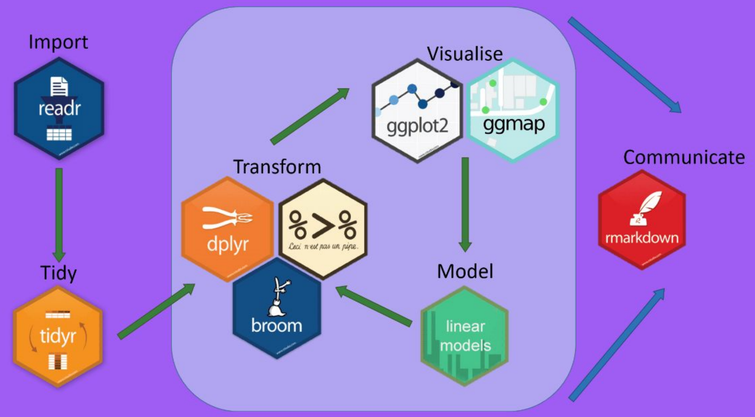
github.com/sebaxtian

github.com/yammadev

infogram

Instituto Nacional de Salud

TOOLS:



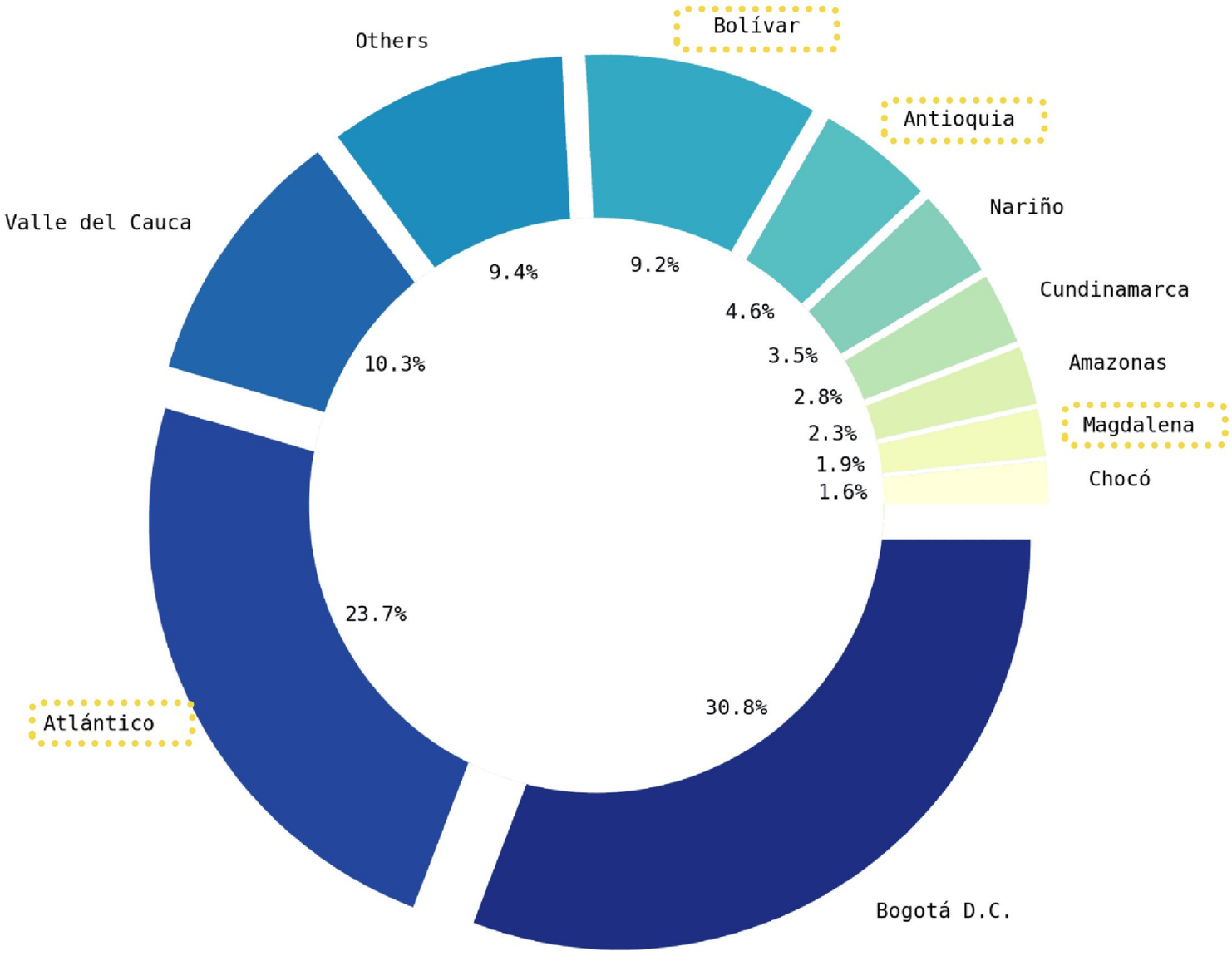
GeoPy

TROUBLES:

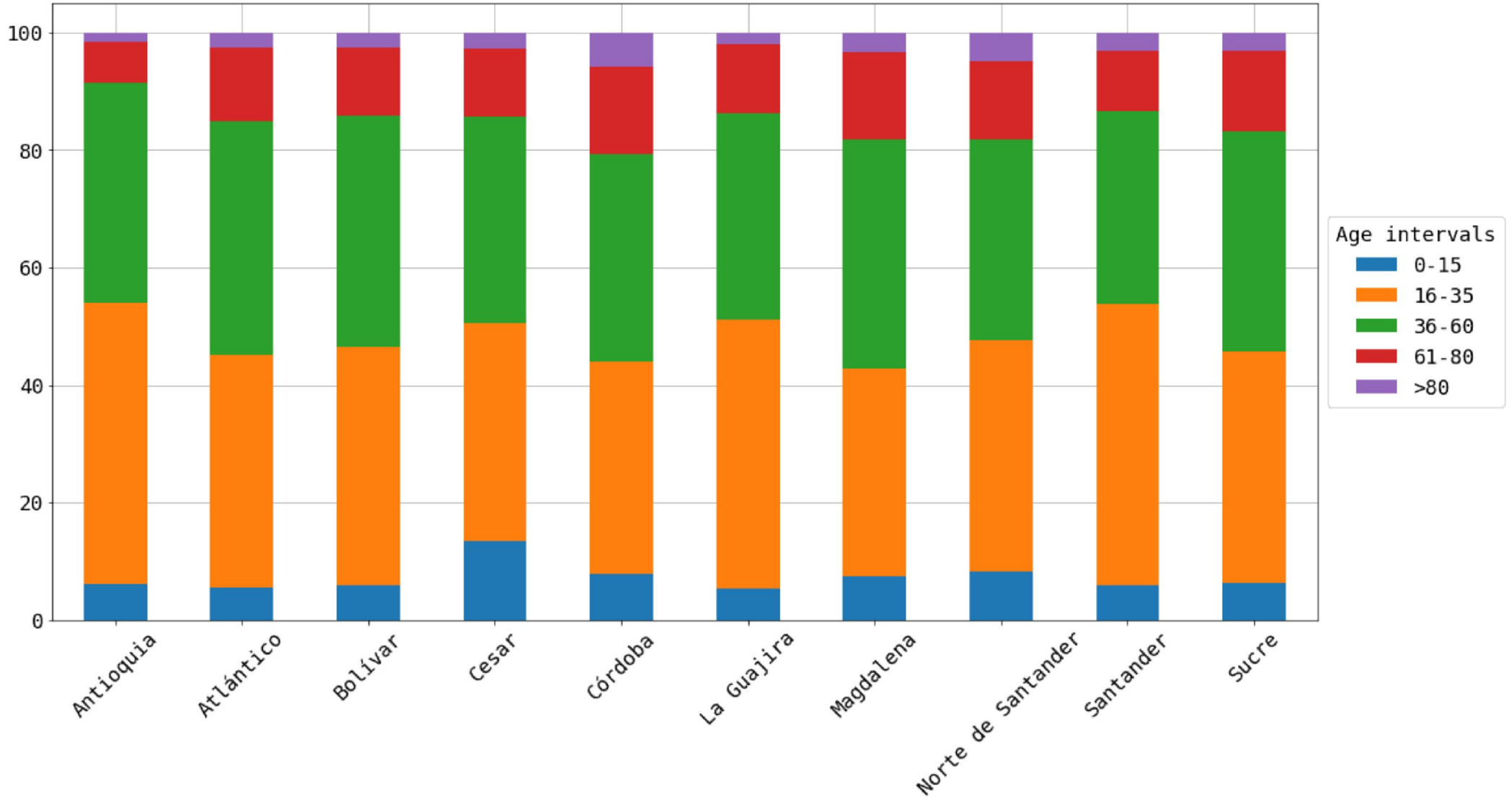
Inconsistency of data

Lack of data

Departments with higher number of cases.



Demographics of the dataset

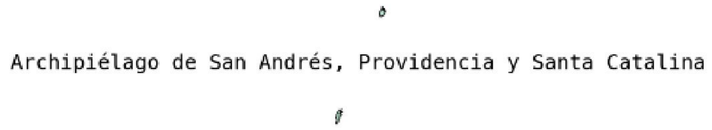


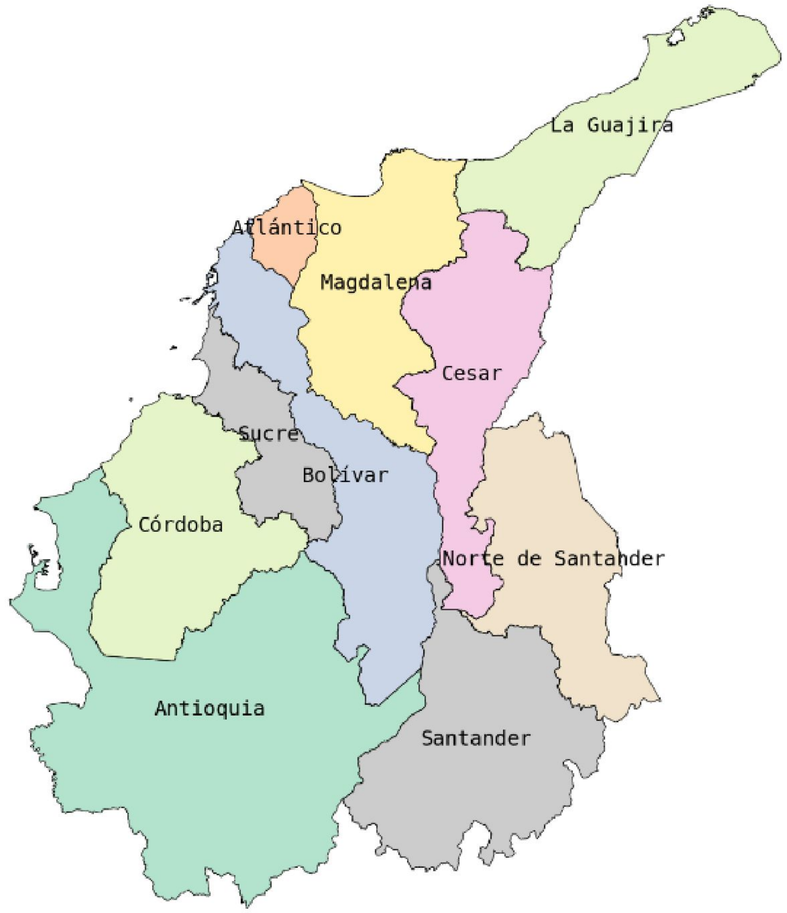
11 regions with a total population
of more than 18 million.

5 cities with more than 500'000 inhabitants:

- Medellin
- Barranquilla
- Cartagena
- Cucuta
- Soledade

Great diversity between the departments





Evaluation:

INDEXES

Scale dependant:

- MAE
- RMSE

Percentage error:

- MAPE

PROCEDURE

Time series cross validation

The forecast errors

$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T}$$

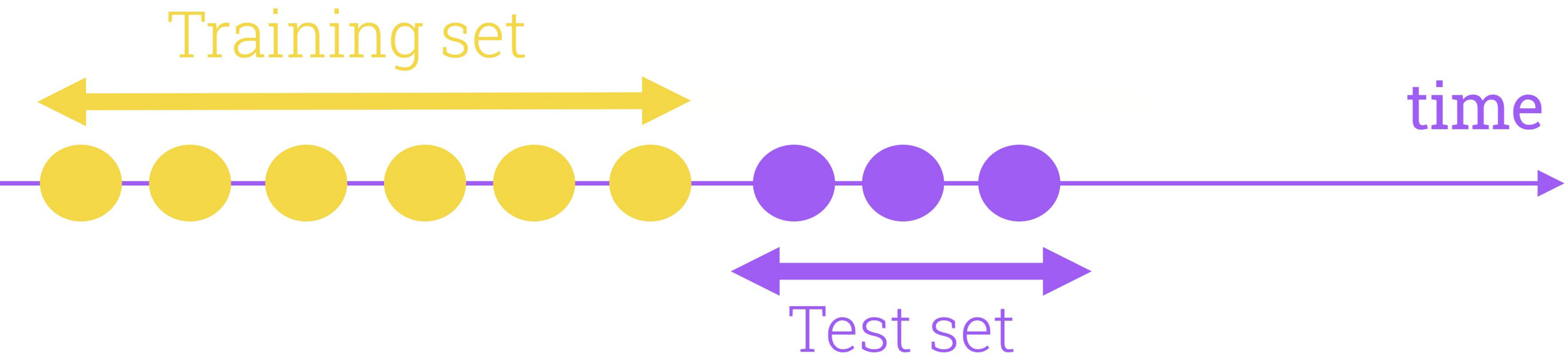
Where:

Training: $\{y_1, \ldots, y_T\}$

Test: $\{y_{T+1}, y_{T+2}, \ldots\}$

Source:
Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice

The forecast errors (2)



$$\text{Mean absolute error: } \text{MAE} = \text{mean}(|e_t|)$$

$$\text{Root mean squared error: } \text{RMSE} = \sqrt{\text{mean}(e_t^2)}.$$

Given: $p_t = 100e_t/y_t$

$$\text{Mean absolute percentage error: } = \text{mean}(|p_t|).$$

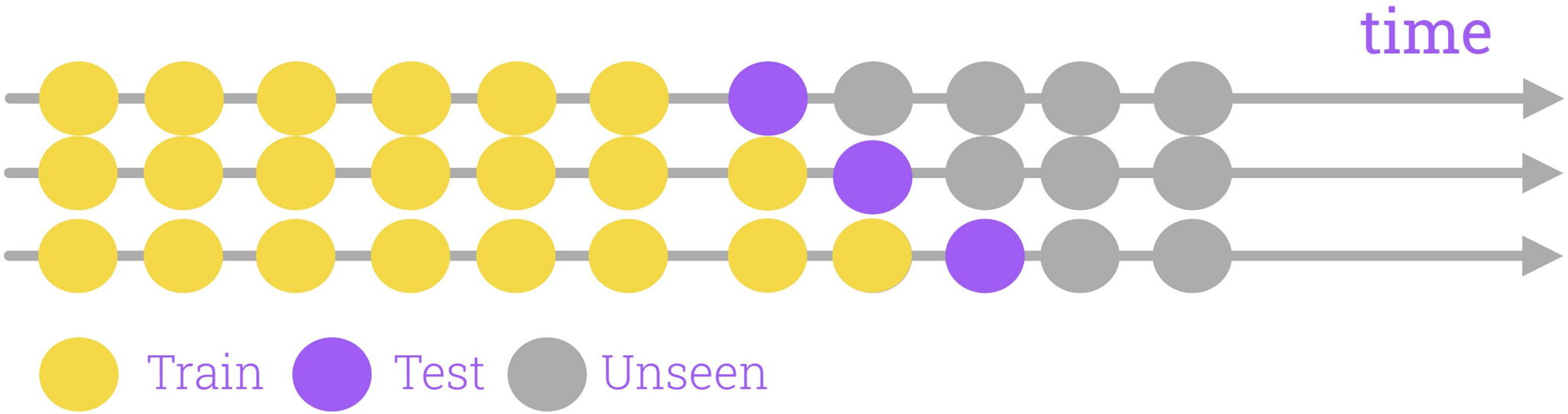
Time series Cross-Validation

INGREDIENTS:

1) Series of test sets, each consisting of a single observation.

2) Training set consists only of observations that occurred prior to the observation that forms the test set. of test sets, each consisting of a single observation.

3) A LOOP



Which kind of data we have?

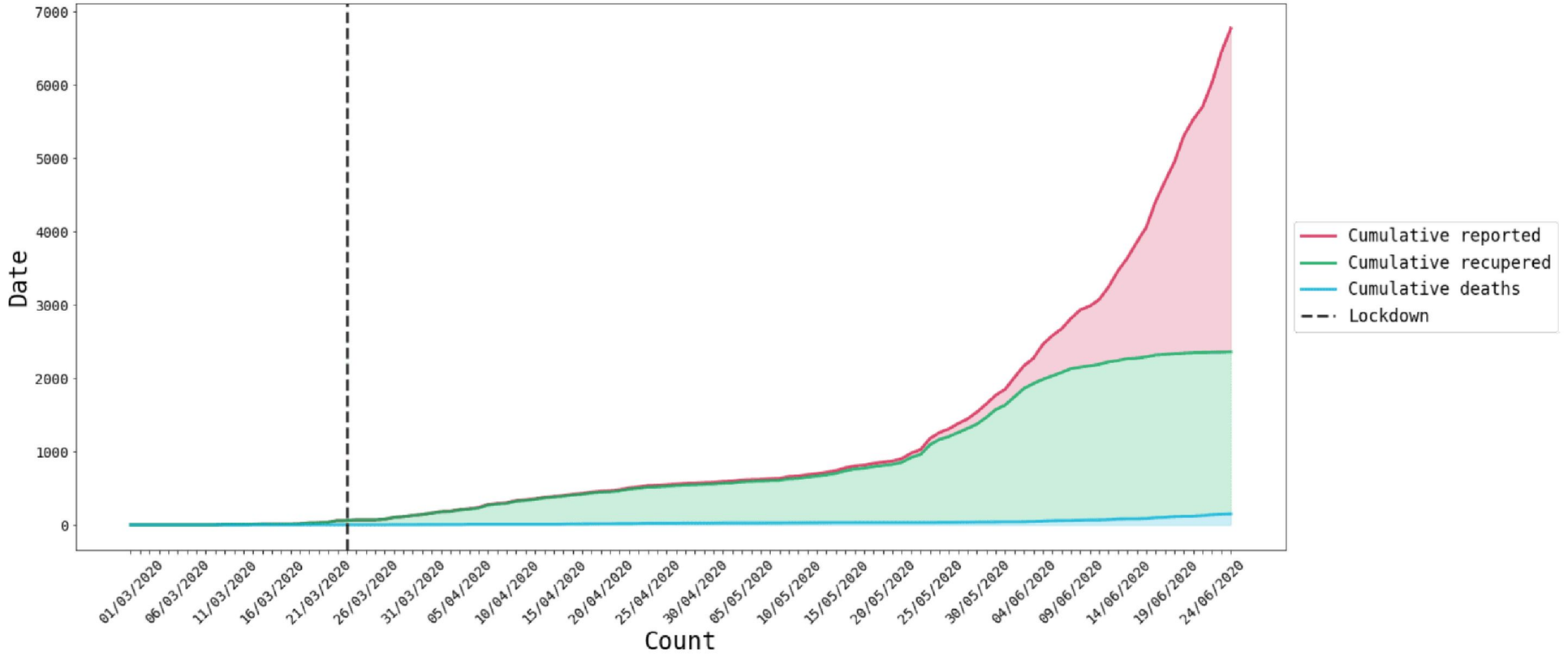
Positive count of COVID-19 cases in a specific area.

Different ways for modelling count data:

POISSON REGRESSION

QUASI POISSON REGRESSION

NEGATIVE BINOMIAL REGRESSION



Generalized Linear Model

GLM relax the normality assumption: response variables can be assumed non-normal

$$f(y;\theta,\phi) = exp\left(\frac{y\cdot\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

θ is the canonical parameter that depends on the regressors via a linear predictor and ϕ is a dispersion parameter.

Functions b(), c() are known and determine which member of the family is used (normal, binomial or Poisson)

$$g(E(Y_i)) = g(\mu_i) = \eta_i = x_i^T\beta$$

GLM: Poisson Regression

Probability function:

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

Appropriate to describe count data which are referred to specific time geographical area.

Parameter estimation based on ML

$$\phi = 1 \quad \theta = log\lambda \quad b(\theta) = \lambda = e^\theta \quad c(Y,\phi) = -logY!$$

Assumption: data are equidispersed ϕ = 1 and thus variance is identical to the mean

Canonical link is $g(\mu) = log(\mu)$, resulting in a log-linear relationship between mean and linear predictor provided by the model fitting functions glm(..., family = "poisson") in the stats package

GLM: Quasi Poisson model

Model that deal with over dispersed count data

Estimates of the coefficients are the same since the estimating equation do not change

The parameter is left unrestricted and estimated from the data

The standard estimation change accordingly to the parameter ϕ

Provided by the model fitting functions **glm(..., family = "quasipoisson")** in the **stats** package

Beyond GLM: Negative Binomial Model

An alternative model that can be considered when data exhibits overdispersion. The probability function is:

$$Pr(Z = z) = \binom{z-1}{k-1} p^k (1-p)^{z-k} \qquad z = k, k+1, \ldots$$

Mixture of Poisson when each unit Y is Poisson with mean drown from a Gamma distribution.

Parameter estimation based on ML including additional shape parameter than poisson and so it proves to be more flexible

Provided by the model fitting functions **glm.nb()** in the **MASS** package, which estimate both the shape parameter and the coefficients

Antioquia: A case study

Are data overdispersed?

First simpler Poisson model

$$log\mu_i = \beta_0 + \beta_1 \cdot date_i$$

AIC:
1014

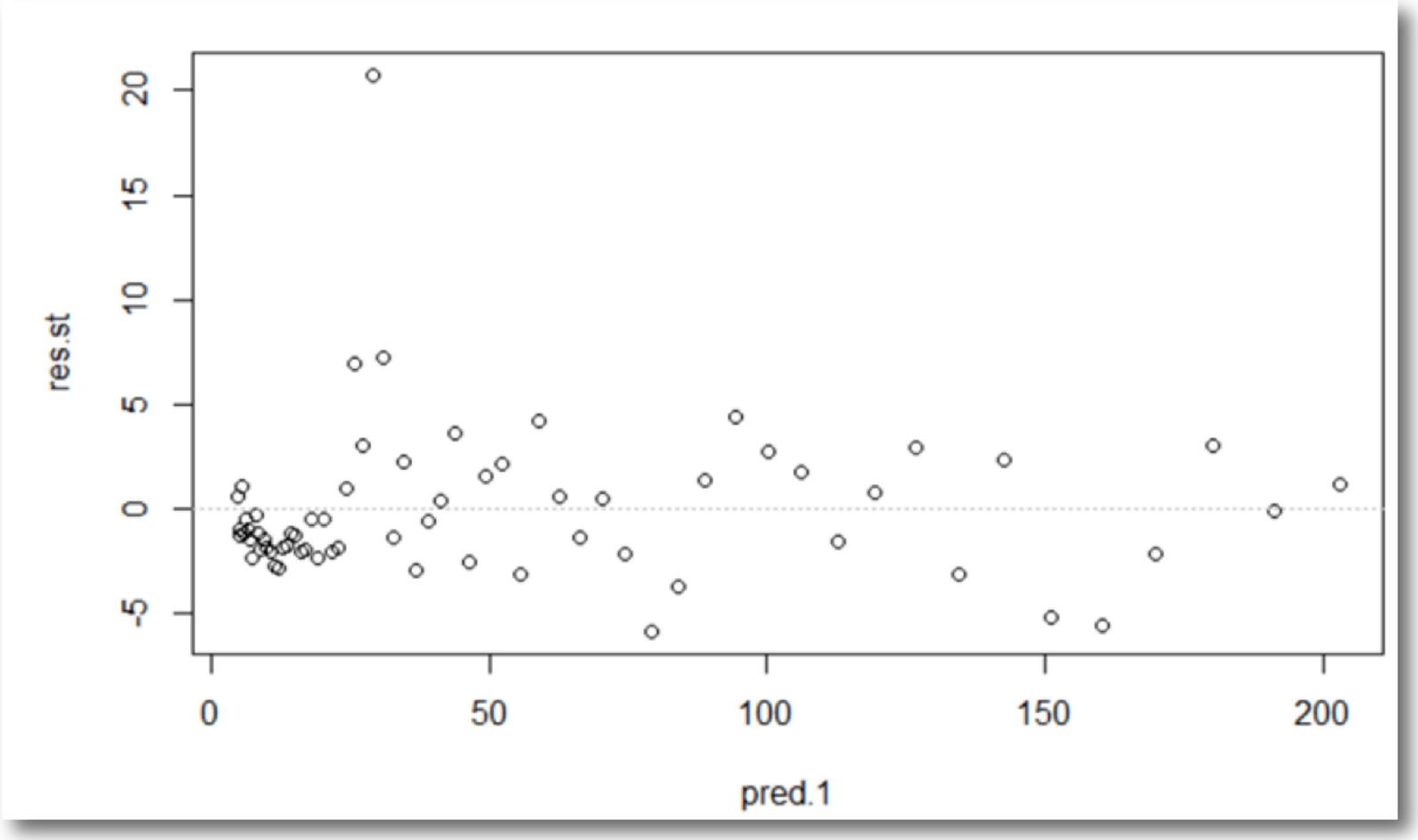
RESIDUAL DEVIANCE:
Null Deviance: 3804
Residual Deviance: 683.97

COEFFICENTS
Both significant
Date: 0.05

Antioquia: A case study

Poisson distribution: is a suitable model?

Check for overdispersion: plot standardized residual vs fitted



$$\hat{\phi} = \frac{\sum z_i^2}{n-p} = 13.91056$$

Our strategy for the model selection

1) Forward search:

From a null model add covariate if the p-value is small. At each step we consider:

AIC:

the lower the better

Occam's Razor principle

Residual deviance reduction

Log-Likelihood ratio test using ANOVA:

The null hypothesis is that the coefficient is 0

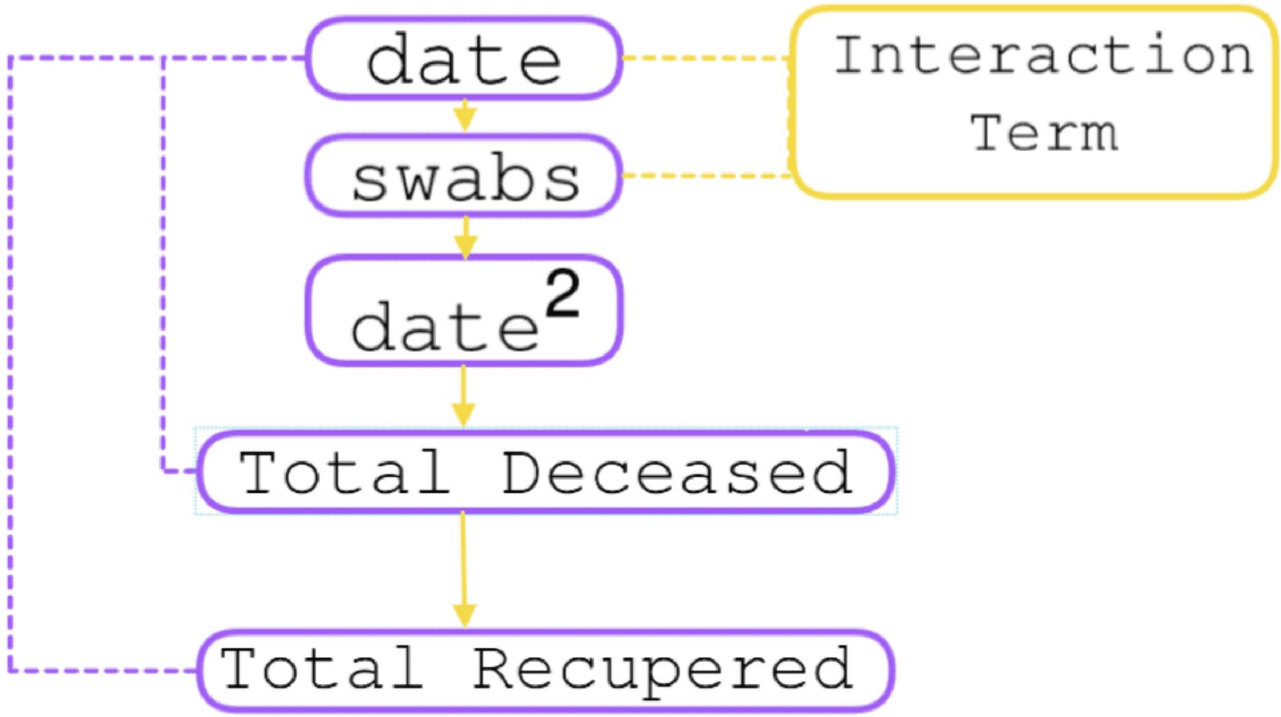
If the p-value is small enough, the null hypotheis is rejected

2) Perfomance Analysis

RMSE, MAE with Cross Validation

Antioquia: A case study

Covariate selection



Antioquia: A case study

Prediction daily cases of best model selected

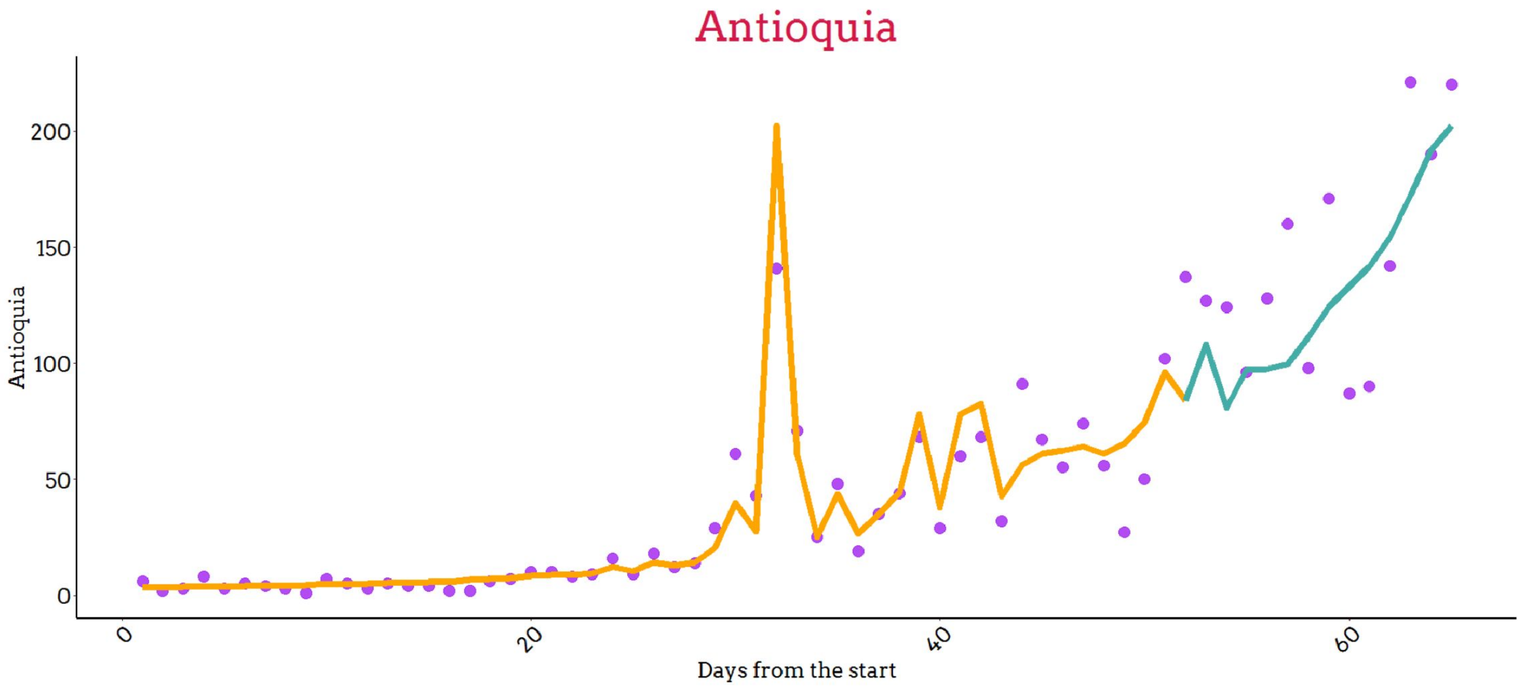
MASS::glm.nb(formula = total_reported ~ times + I(times^2) + daily_swabs + total_recupered, data = train_data, init.theta = 15.53248731, link = log)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.042e+00	2.382e-01	4.375	1.21e-05	***
times	3.273e-02	1.825e-02	1.793	0.0729	.
I(times^2)	4.857e-04	2.965e-04	1.638	0.1014	
daily_swabs	4.386e-05	9.078e-05	0.483	0.6290	
total_recupered	2.117e-02	2.309e-03	9.167	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 616.878 on 51 degrees of freedom
Residual deviance: 46.246 on 47 degrees of freedom
AIC: 342.62



Performance Indexes

##MAE: 51.2
##RMSE: 65.4

Antioquia: A case study

Call:
MASS::glm.nb(formula = total_reported ~ times +
daily_swabs *times+
    daily_swabs, data = train_data, init.theta = 3.846935133,
    link = log)

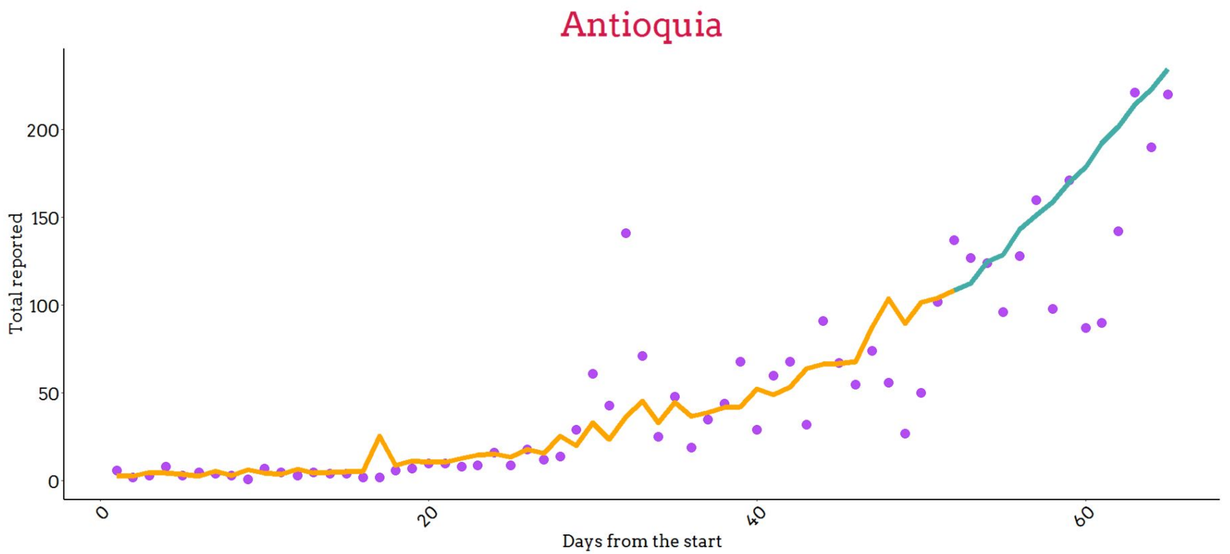
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	4.812e-01	3.085e-01	1.560	0.1188	
times	7.897e-02	1.169e-02	6.757	1.41e-11	***
daily_swabs	7.335e-04	3.031e-04	2.420	0.0155	*
times:daily_swabs	-1.248e-05	9.523e-06	-1.311	0.0190	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 227.426 on 51 degrees of freedom
Residual deviance: 50.939 on 48 degrees of freedom
AIC: 391.75

Best daily cases prediction



Performance Indexes

##MAE: 33.1
##RMSE: 44.25

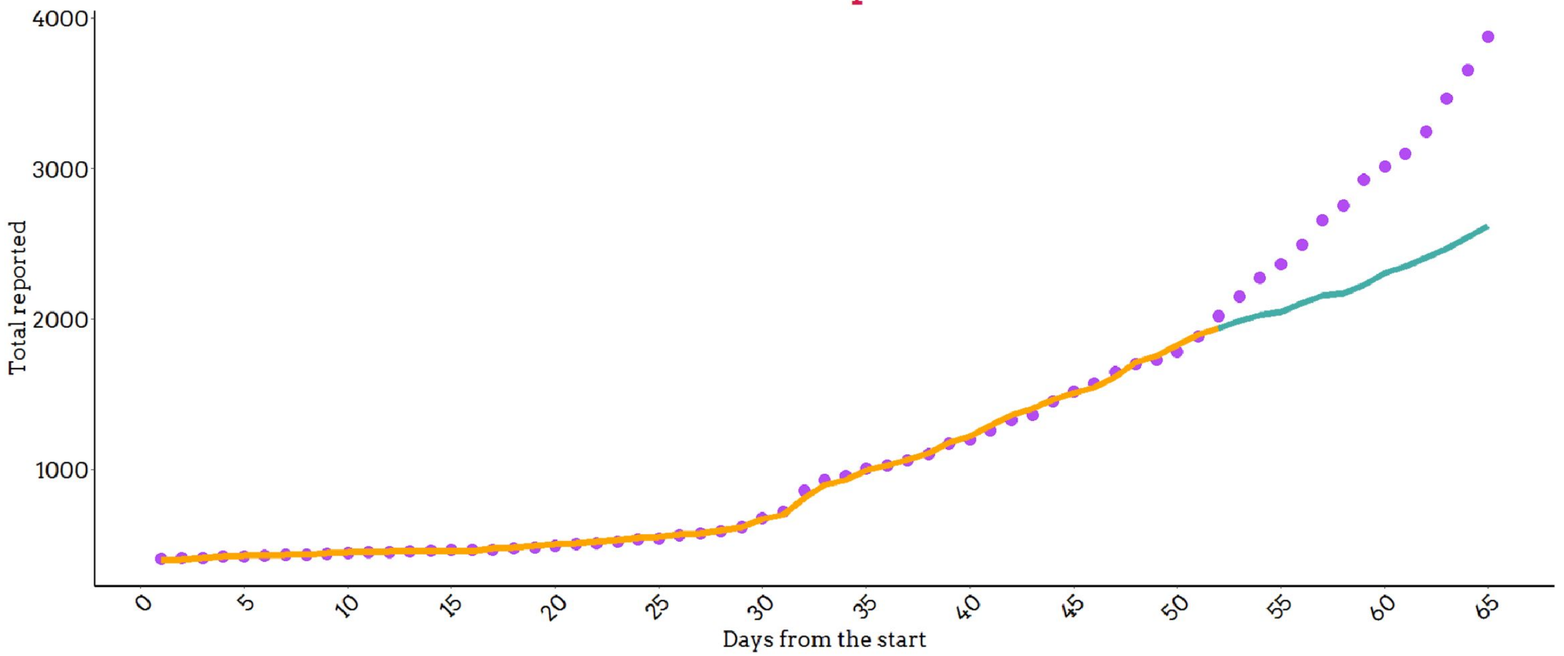
Antioquia: A case study

Prediction cumulative cases of best model selected

```
Call:
MASS::glm.nb(formula = accum_reported ~ times + test +
times *
    accum_recupered + accum_recupered, data = train_data,
init.theta = 43662431.02,
    link = log)


Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       5.262e+00  8.094e-02  65.008  < 2e-16 ***
times            -2.011e-03  2.640e-03  -0.762  0.44623
test              2.122e-05  5.355e-06   3.963  7.4e-05 ***
accum_recupered   1.191e-03  1.092e-04  10.902  < 2e-16
***
times:accum_recupered -7.469e-06  2.767e-06  -2.699
0.00695 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Null deviance: 13031.594  on 51  degrees of freedom
Residual deviance:    18.748  on 47  degrees of freedom
AIC: 469.01
```



Performance Indexes

##MAE: 438.8
##RMSE: 591.4814

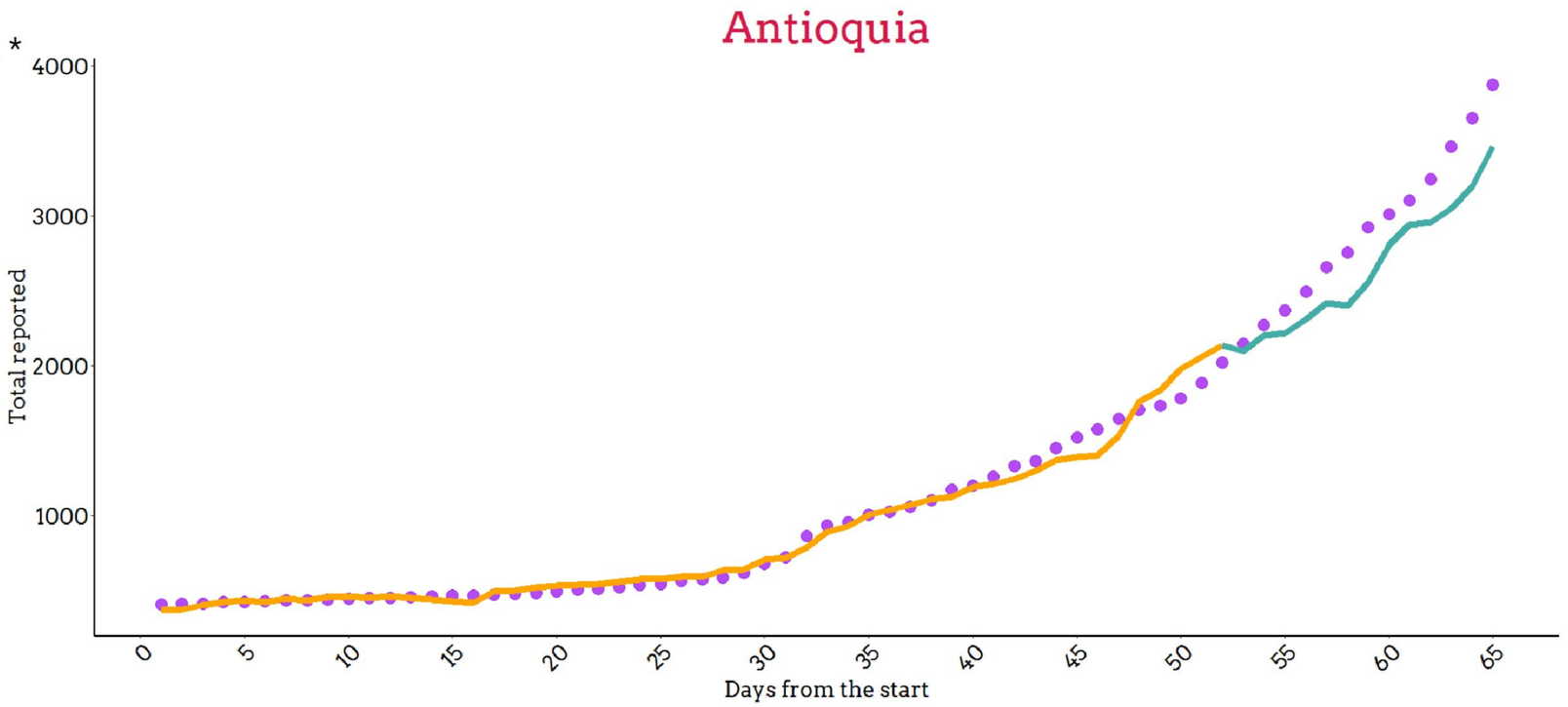
Antioquia: A case study

Best cumulative cases prediction

```
MASS::glm.nb(formula = accum_reported ~ times + test + test *
    times, data = train_data, init.theta = 438.8317357,
    link = log)


Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.130e+00  2.551e-01  20.111  < 2e-16 ***
times         -2.064e-02  4.869e-03  -4.239 2.24e-05 ***
test           7.594e-05  1.275e-05   5.958 2.55e-09 ***
times:test    -2.560e-07  1.563e-07  -1.638    0.101
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 4348.457  on 51  degrees of freedom
Residual deviance:   48.566  on 47  degrees of freedom
AIC: 551.3
```



Performance Indexes

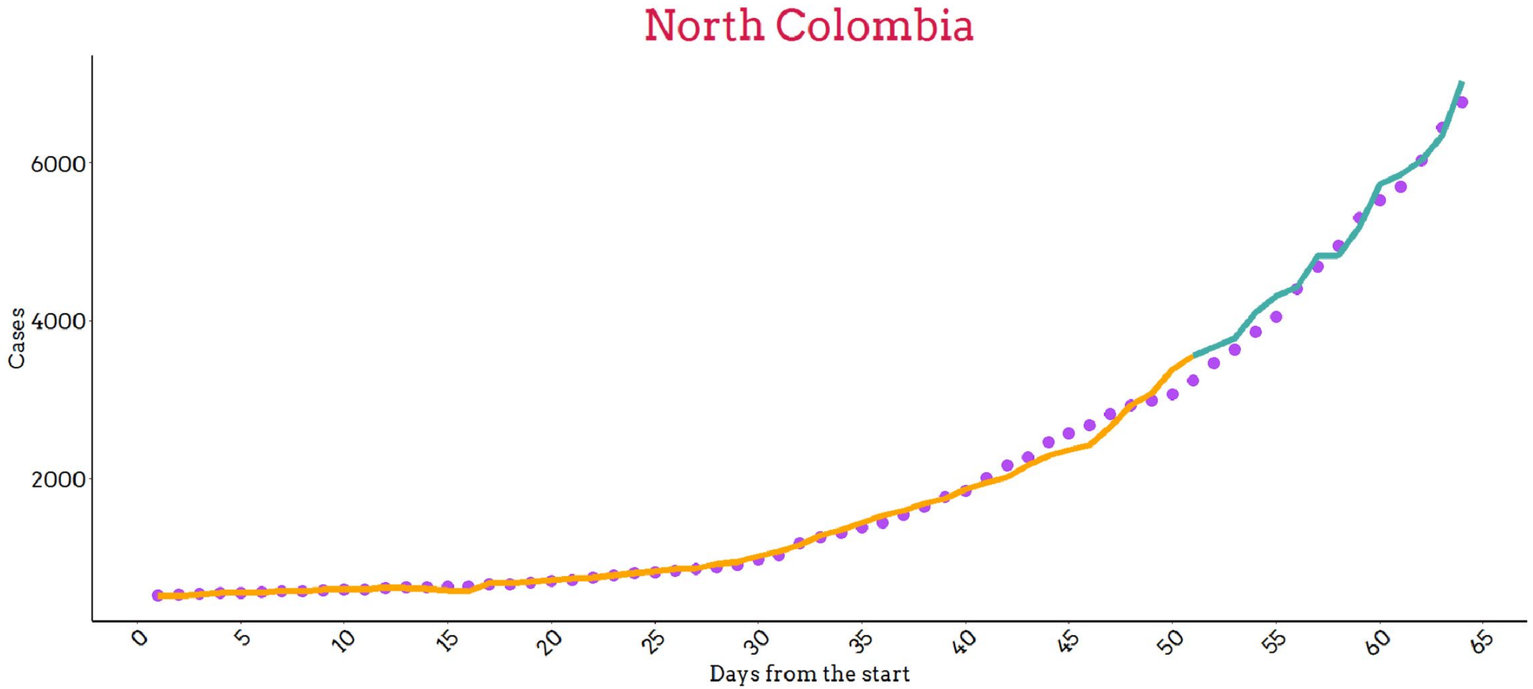
##MAE: 159.6
##RMSE: 205.0977

North Colombia cumulative cases prediction

```
MASS::glm.nb(formula = accum_reported ~ times + test
+ accum_deceased,
    data = train_data, init.theta = 643.3729735, link = log)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     5.729e+00  2.183e-02 262.476  < 2e-16 ***
times          -2.018e-02  3.305e-03  -6.105 1.03e-09 ***
test           3.861e-05  2.546e-06  15.166  < 2e-16 ***
accum_deceased -6.523e-03  1.872e-03  -3.485 0.000492
***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 8040.440  on 50  degrees of freedom
Residual deviance:   43.069  on 47  degrees of freedom
AIC: 552
```



Performance Indexes

##MAE: 206.632
##RMSE: 379.897

Generalized Additive Models

GAM is a generalized linear model in which the linear predictor depends linearly on unknown smooth functions of some predictor variables.

$$Y_i = \beta_0 + \sum_{k=1}^{K} b_K B_K(Z_i) + \texttt{other variables} + \epsilon_i$$

$$s(z) = \beta_0 + \sum_{k=1}^{K} b_K B_K(z)$$

Parameter estimation

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{k=1}^{K} b_k B_k(z_i) + \epsilon_i$$

$$\tilde{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} & B_1(z_1) & \dots & B_K(z_1) \\ . & & . & . & & & . \\ . & & . & . & & & . \\ . & & . & . & & & . \\ 1 & x_{n1} & \dots & x_{np} & B_1(z_n) & \dots & B_K(z_n) \end{bmatrix}$$

$$\tilde{\beta} = \begin{bmatrix} \beta \\ \mathbf{b} \end{bmatrix}$$

$$(\mathbf{y} - \tilde{\beta}\tilde{X})^T (\mathbf{y} - \tilde{\beta}\tilde{X})$$

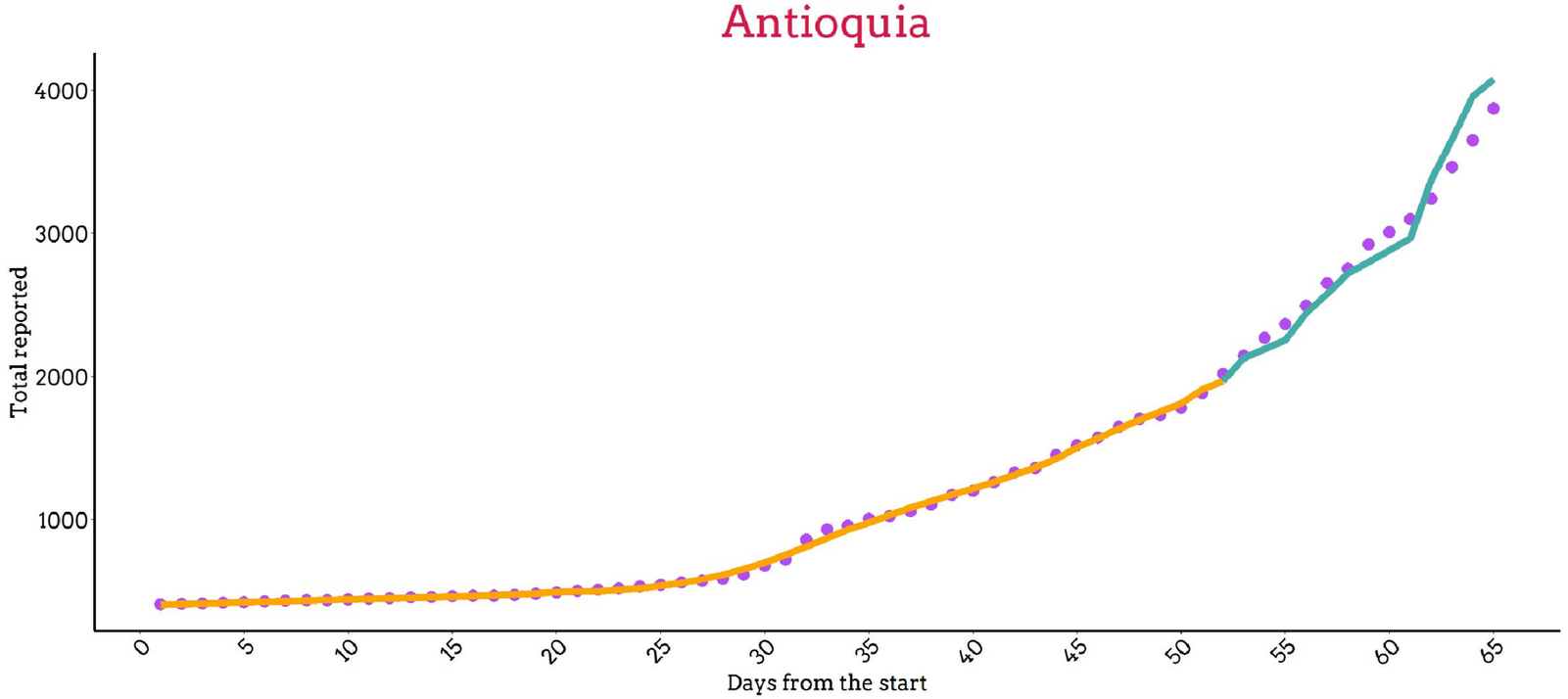
Roughness penalty

$$R(\mathbf{b}) = \mathbf{b}^T \mathbf{G} \mathbf{b}$$

$$(\mathbf{y} - \tilde{\beta}\tilde{X})^T (\mathbf{y} - \tilde{\beta}\tilde{X}) + \lambda \mathbf{b}^T \mathbf{G} \mathbf{b}$$

Cumulative Antioquia

```
##
## Family: poisson
## Link function: log
##
## Formula:
## accum_reported ~ s(times) + s(accum_deceased, k = 4)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 6.590278   0.005464    1206   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                    edf Ref.df  Chi.sq p-value
## s(times)         7.432  8.299 688.453  <2e-16 ***
## s(accum_deceased) 1.691  2.016   0.962   0.589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.998   Deviance explained = 99.8%
## -REML = 256.98  Scale est. = 1         n = 52
```

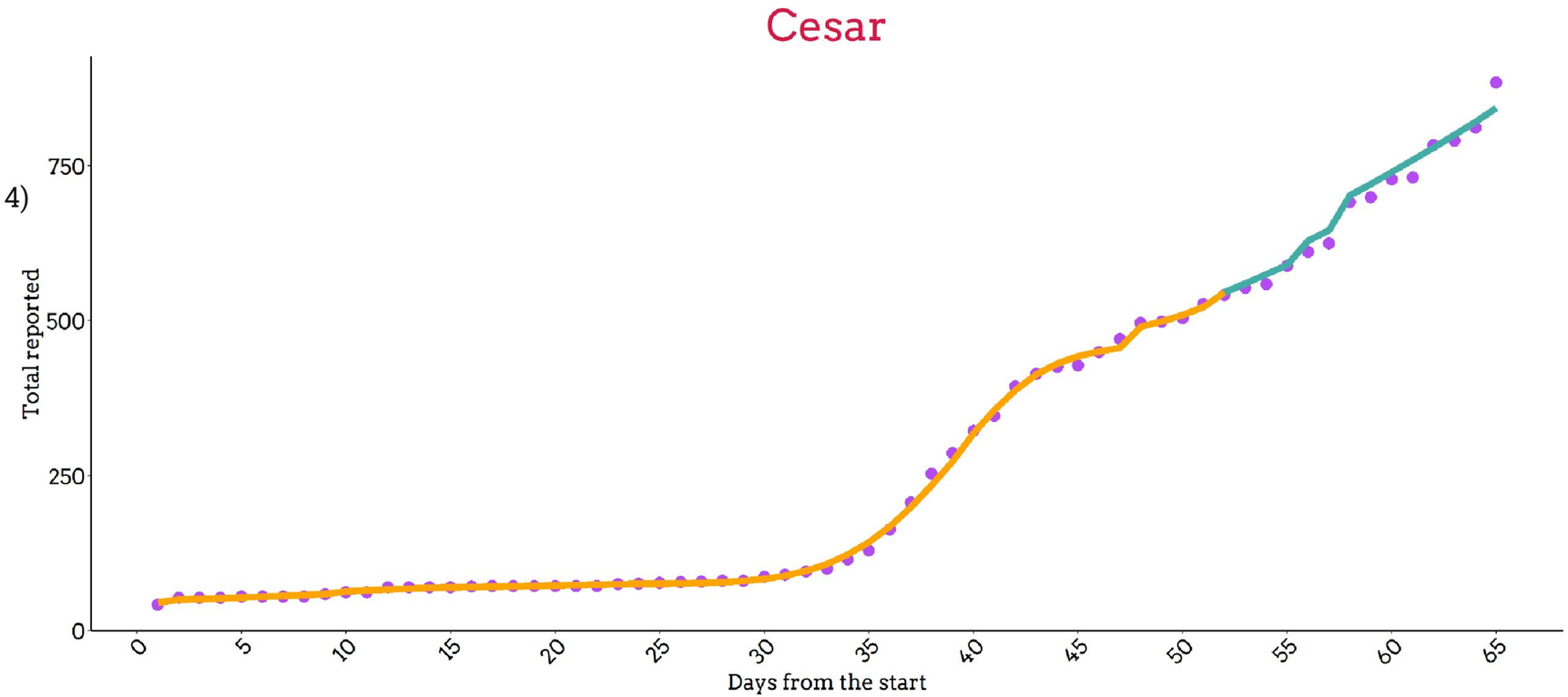


Performance Indexes

```
## MAE:  63.3
## RMSE:  78.56208
## MAPE:  0.4603641
```

Cumulative Cesar

```
##
## Family: poisson
## Link function: log
##
## Formula:
## accum_reported ~ s(times) + s(accum_deceased, k = 4)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.8081     0.0143   336.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                    edf Ref.df   Chi.sq p-value
## s(times)         7.889  8.653 2661.339  <2e-16 ***
## s(accum_deceased) 1.000  1.001    1.138   0.286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.999   Deviance explained = 99.9%
## -REML = 203.17  Scale est. = 1         n = 52
```

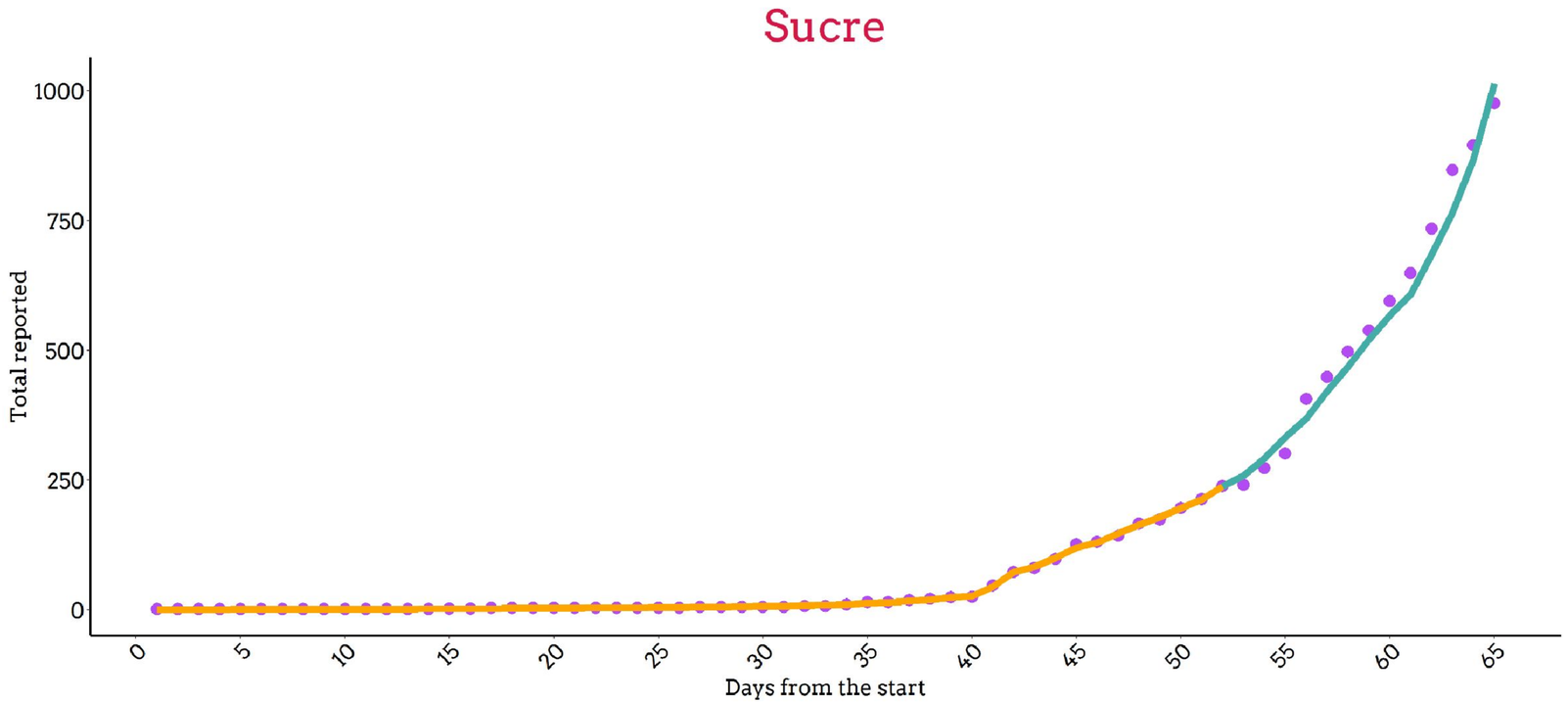


Performance Indexes

```
## MAE:  12
## RMSE:  15.54
## MAPE:  00.805
```

Cumulative Sucre

```
##
## Family: poisson
## Link function: log
##
## Formula:
## accum_reported ~ s(test) + s(times) +
s(accum_recupered) + s(accum_deceased,
##     k = 4)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.05769    0.08019   25.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                    edf Ref.df Chi.sq  p-value
## s(test)           1.00  1.000  0.330   0.5656
## s(times)          1.00  1.000  6.180   0.0129 *
## s(accum_recupered) 2.87  3.578 50.325 4.88e-08 ***
## s(accum_deceased)  1.00  1.000  0.064   0.8000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.999   Deviance explained = 99.8%
## -REML = 118.14  Scale est. = 1         n = 52
```

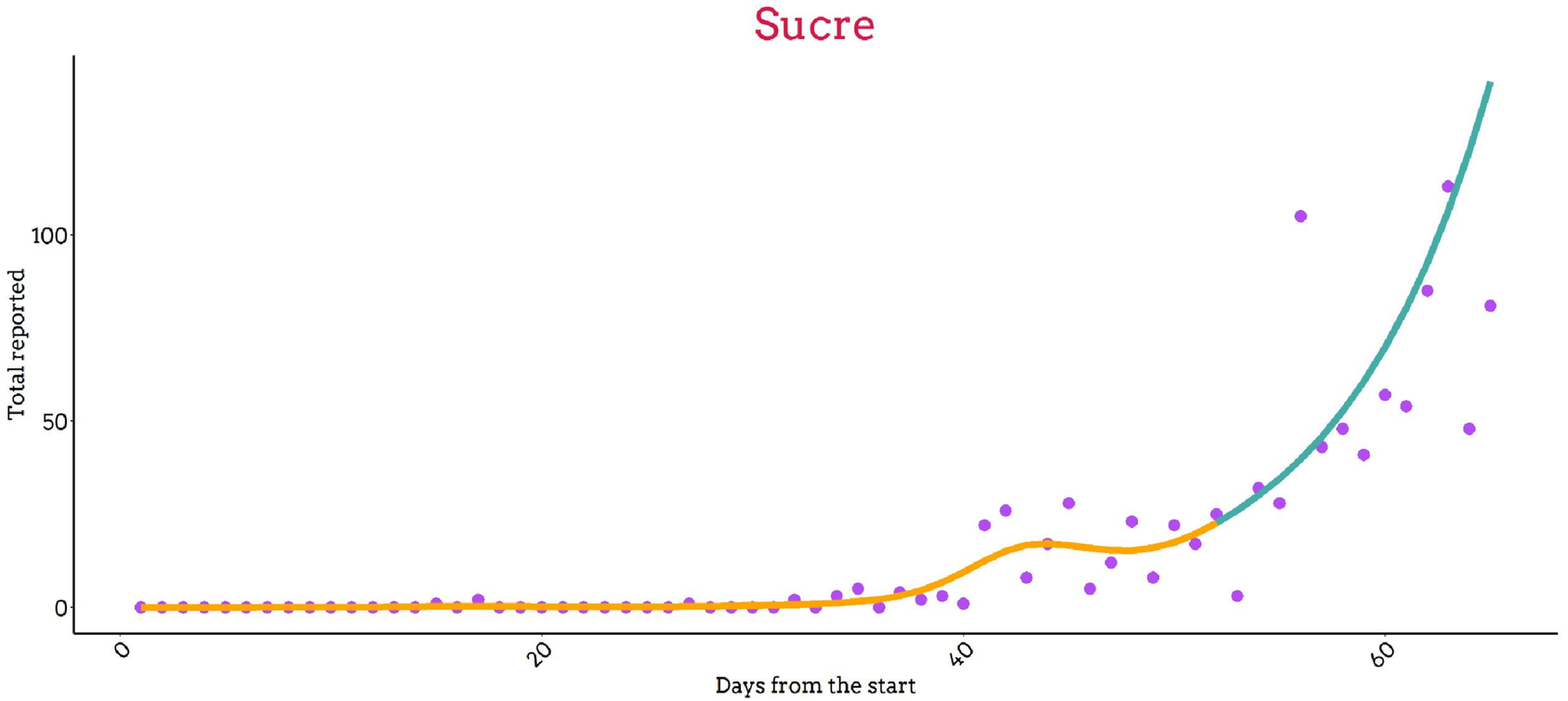


Performance Indexes

```
## MAE:  12
## RMSE:  15.54
## MAPE:  00.805
```

Daily Cases Sucre

```
##
## Family: poisson
## Link function: log
##
## Formula:
## total_reported ~ s(times)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.6045     0.4792  -1.262    0.207
##
## Approximate significance of smooth terms:
##          edf Ref.df Chi.sq p-value
## s(times) 5.72  6.587  156.5  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.732   Deviance explained = 84.4%
## -REML = 99.701  Scale est. = 1         n = 52
```

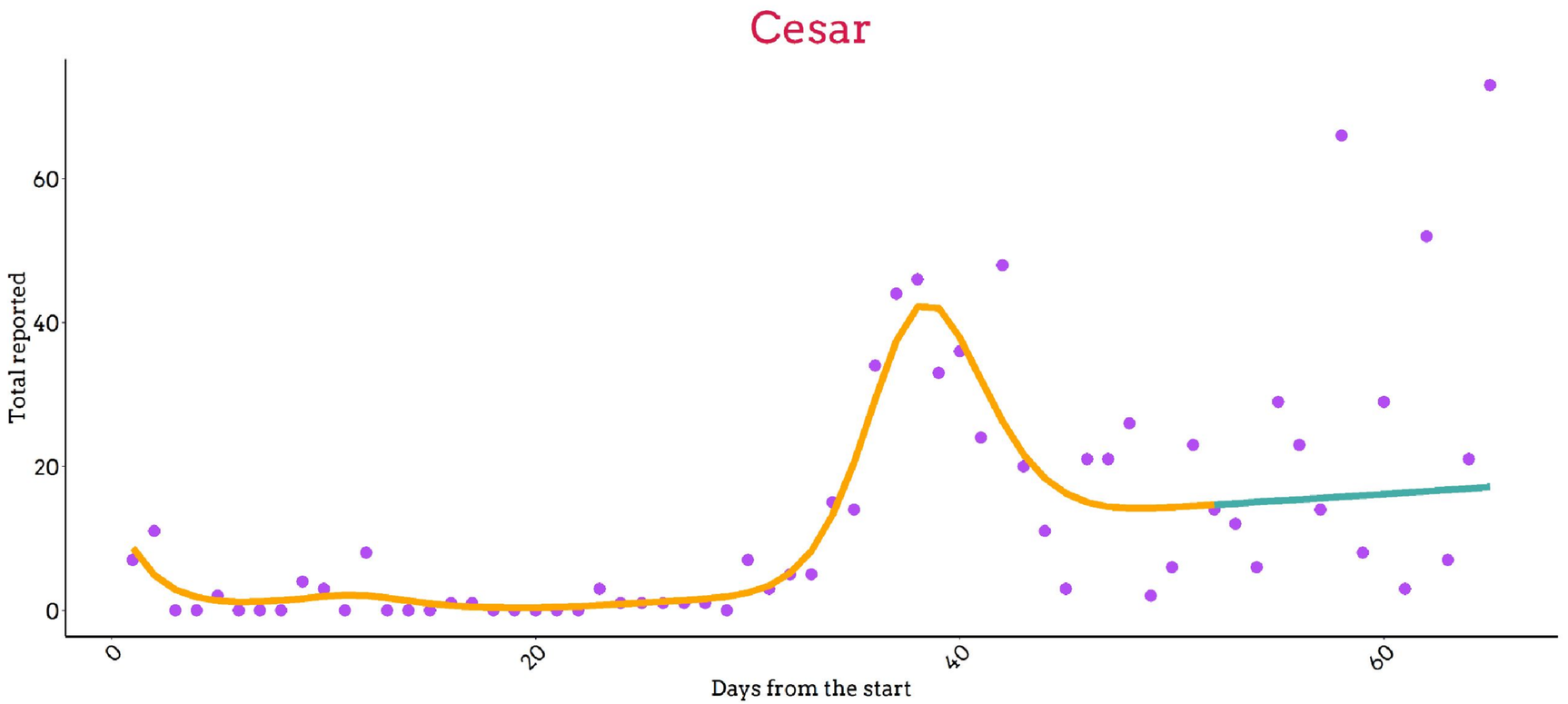


Performance Indexes

```
## MAE:  17.6
## RMSE:  28.01607
## MAPE:  49.805
```

Daily Cases Cesar

```
## Family: poisson
## Link function: log
##
## Formula:
## total_reported ~ s(times)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.3437     0.1043   12.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df Chi.sq p-value
## s(times) 8.133   8.75  407.6  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.788   Deviance explained =   82%
## -REML = 169.39  Scale est. = 1         n = 52
```

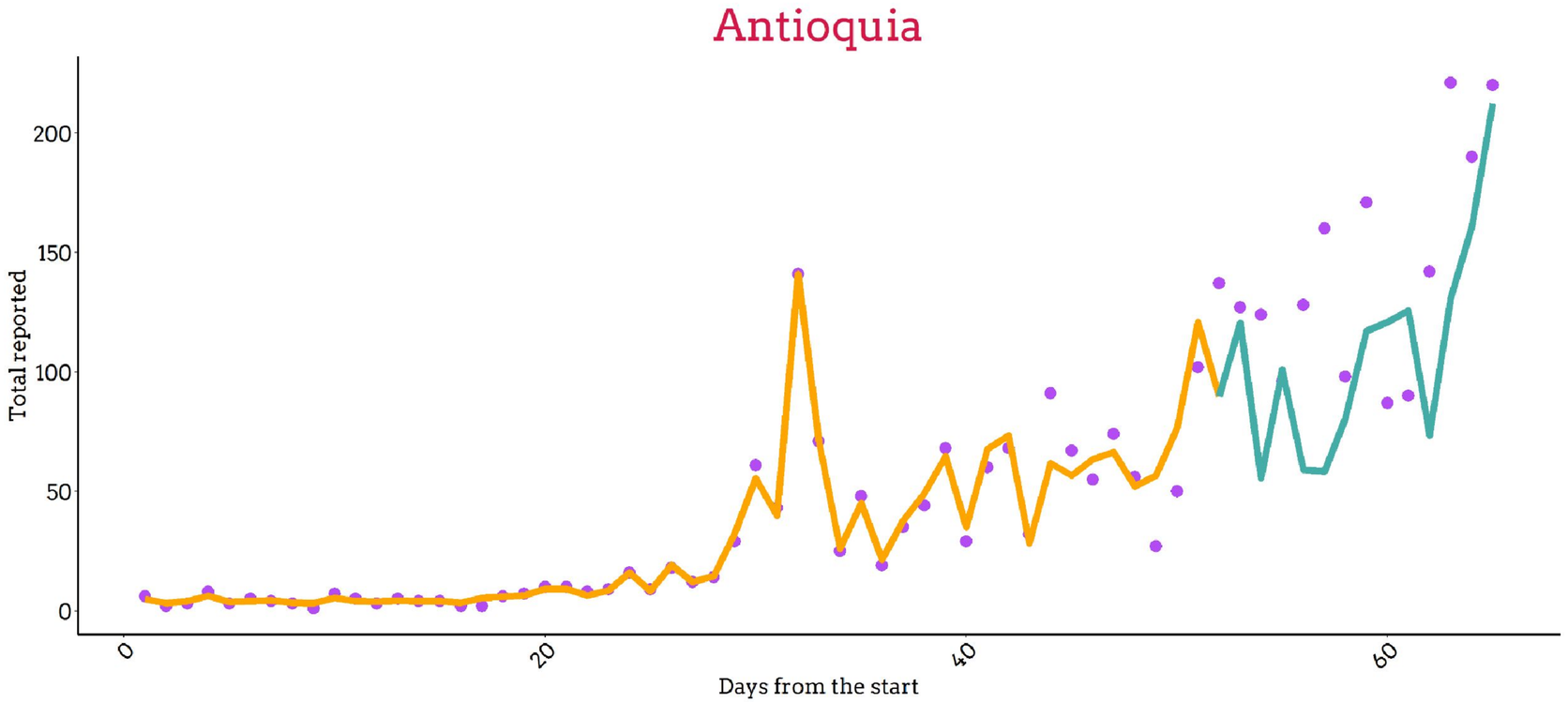


Performance Indexes

```
## MAE:  13.95
## RMSE:  20.29
## MAPE:  54.81
```

Daily Cases Antioquia

```
##
## Family: poisson
## Link function: log
##
## Formula:
## total_reported ~ s(daily_swabs) + s(times) +
s(total_recupered) +
##     total_deceased
##
## Parametric coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.79845    0.04646  60.234   <2e-16 ***
## total_deceased -0.16144    0.08178  -1.974   0.0484 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df  Chi.sq p-value
## s(daily_swabs)    1.000  1.000   7.962 0.00478 **
## s(times)          5.027  6.056 275.347 < 2e-16 ***
## s(total_recupered) 4.837  5.731 259.427 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.881   Deviance explained = 95.1%
## -REML = 189.73  Scale est. = 1         n = 52
```



Performance Indexes

```
## MAE: 35.15
## RMSE:  48.50
## MAPE:  57.14
```

Conclusions

We analyzed many models and we took in consideration some effects that could explain data.

We couldn't see effects of some political maneuvers(lockdown) and other times we couldn't explain some data(high variability after some date)(certain regions had a strange count of swabs)

Our predictions are pretty much satisfying for the regions with more reported cases. Even if the models we considered to be the best have an high AIC score, we chose them because they performed better on test data.

However there are some improvement that we could make.
- We didn't investigate if there was autocorrelation within residuals
- For some models we considered only days for which we had the corresponding cumulative number of swabs. This way we left out many days of data including the day of lockdown.

We have also some further ideas that we didn't had the time to test:

Zero inflated model for regions with small number of cases.

A bayesian model using Jeffrey's prior and a poisson likelihood so that then we could sample from the approximation of the posterior distribution using the markov chain montecarlo method.

More suitable models for timeseries: autoregression models (ARIMA)

Add covariates such density of population, range of age and other region-discriminant factors.

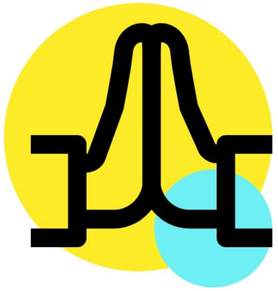
References

[1] A Winnerís Curse for Econometric Models: On the Joint Distribution of In-Sample Fit and Out-of-Sample Fit and its Implications for Model Selection , Peter Reinhard Hansen

[2] Regression Models for Count Data in R, Achim Zeileis, Christian Kleiber,Simon Jackman

[3] Data Analysis and Graphics Using R, An example-based approach, Third edition, John Maldonald and W. John Braun

Thank you



Paolo Pulcini

Lorenzo Taroni

Matilde Castelli