UNIVERSITÀ DEGLI STUDI DI TRIESTE

DATA SCIENCE & SCIENTIFIC COMPUTING

SISSA

CoViD-19 daily positive-case count prediction for Southern Mexico

Statistical Methods for Data Science

Leonardo Arrighi
Sebastiano Zagatti

University of Trieste

July 28, 2020

Overview

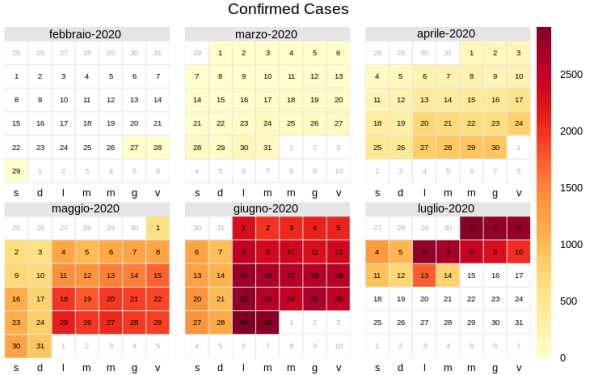
- CoViD19 is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
- First identified in December 2019 in Wuhan (Hubei, China)
- Global pandemic with over 16 million cases and 644 000 deaths

Problem Statement (1)

- Build a model to predict the number of new daily confirmed cases
- We have analyzed 11 states of Southern Mexico, including Mexico City, where the first appearance was on the 27^{th} of February 2020.

Problem Statement (2)

- Considered: 139 days



- Covariates: `days`, `age`, `sex`, `state`, `density of population`

Structure

1. Dataset
2. Data Visualization
3. About Models
4. Generalized Linear Model
5. Auto-Regressive Integrated Moving Average
6. Multivariate Adaptive Regression Splines
7. Regression Random Forest
8. Results
9. Conclusions

Dataset

- https://www.kaggle.com/carloslira/covid19-mexico
- 324 041 observations, 5 variables

	State	Sex	Age	Date	Density
	<chr>	<chr>	<dbl>	<date>	<chr>
1	CIUDAD DE MÉXICO	MASCULINO	65	2020-02-27	5965.6542
2	CIUDAD DE MÉXICO	MASCULINO	36	2020-02-27	5965.6542
3	CIUDAD DE MÉXICO	MASCULINO	59	2020-02-27	5965.6542
4	CHIAPAS	FEMENINO	19	2020-02-29	71.1750
5	CIUDAD DE MÉXICO	MASCULINO	46	2020-03-05	5965.6542
6	CIUDAD DE MÉXICO	MASCULINO	17	2020-03-07	5965.6542
7	CIUDAD DE MÉXICO	FEMENINO	31	2020-03-08	5965.6542
8	CAMPECHE	MASCULINO	39	2020-03-08	15.6491
9	CIUDAD DE MÉXICO	MASCULINO	72	2020-03-10	5965.6542
10	CIUDAD DE MÉXICO	FEMENINO	50	2020-03-10	5965.6542
11	CIUDAD DE MÉXICO	MASCULINO	42	2020-03-10	5965.6542
12	CIUDAD DE MÉXICO	MASCULINO	26	2020-03-11	5965.6542
13	QUINTANA ROO	MASCULINO	49	2020-03-11	33.5882
14	QUINTANA ROO	MASCULINO	39	2020-03-11	33.5882
15	CIUDAD DE MÉXICO	MASCULINO	50	2020-03-11	5965.6542

Pre-processing

- Select interested regions
- Drop unnecessary data
- Manage NA
- Select data until 14^{th} July
- Create time series
- Checked weekends' data
- Split the data in training and test sets (90/10):
  - Training set: data up to the 30^{th} of June
  - Test set: data from the 1^{st} of July to the 14^{th} of July

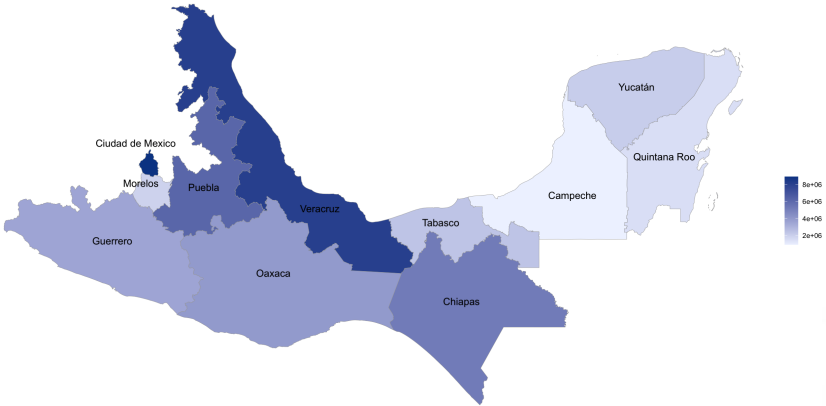
Structure

1. Dataset
2. Data Visualization
3. About Models
4. Generalized Linear Model
5. Auto-Regressive Integrated Moving Average
6. Multivariate Adaptive Regression Splines
7. Regression Random Forest
8. Results
9. Conclusions

Demographic Features

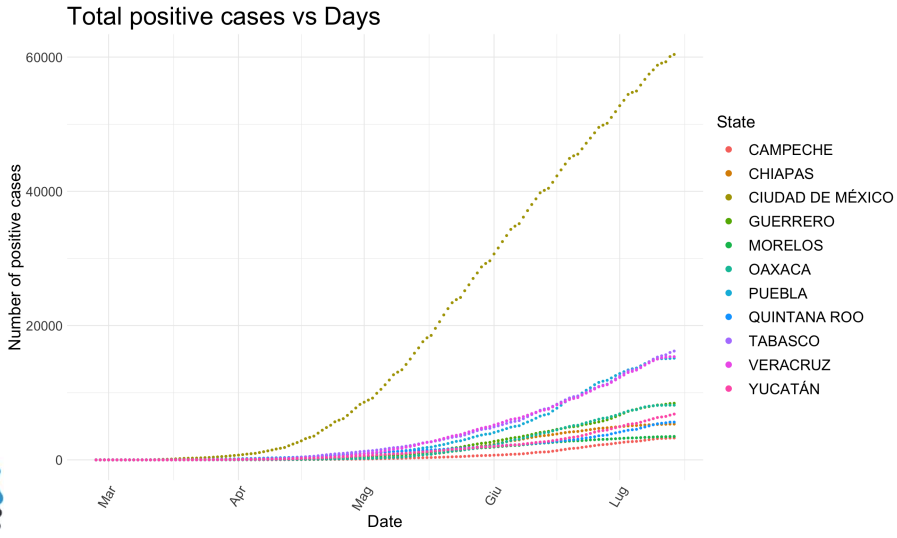
Total population of the considered regions: 45 926 900

Distribution of the population by state

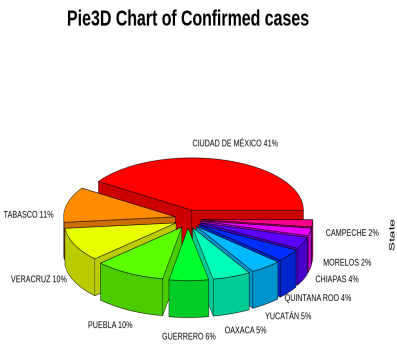


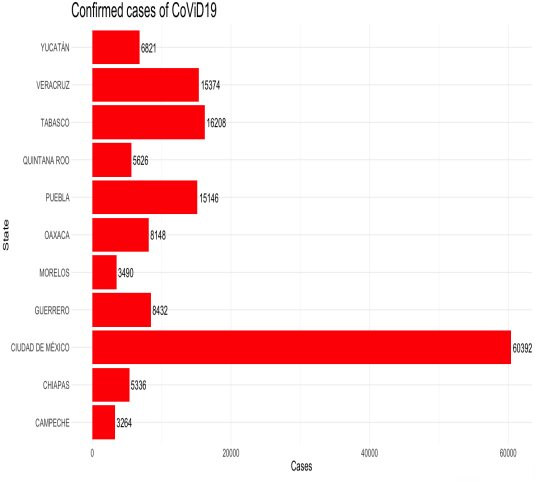
Trend of Positive Cases

Total confirmed cases on the 14^{th} July 2020: 148 237

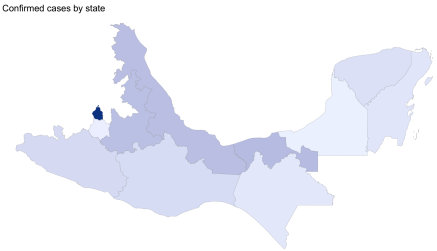


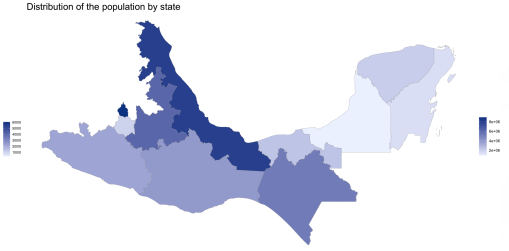
Relations with Population (1)



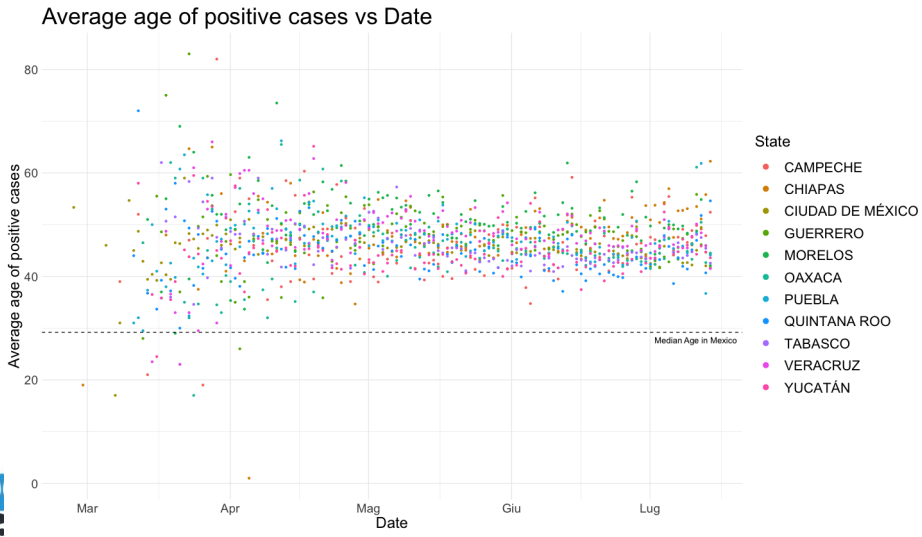


Relations with Population (2)



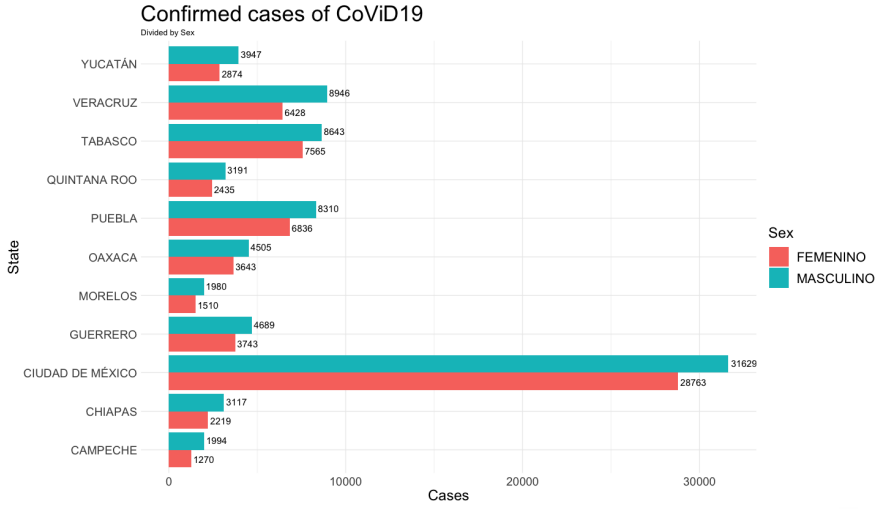


Relations with Demographic Data (1)

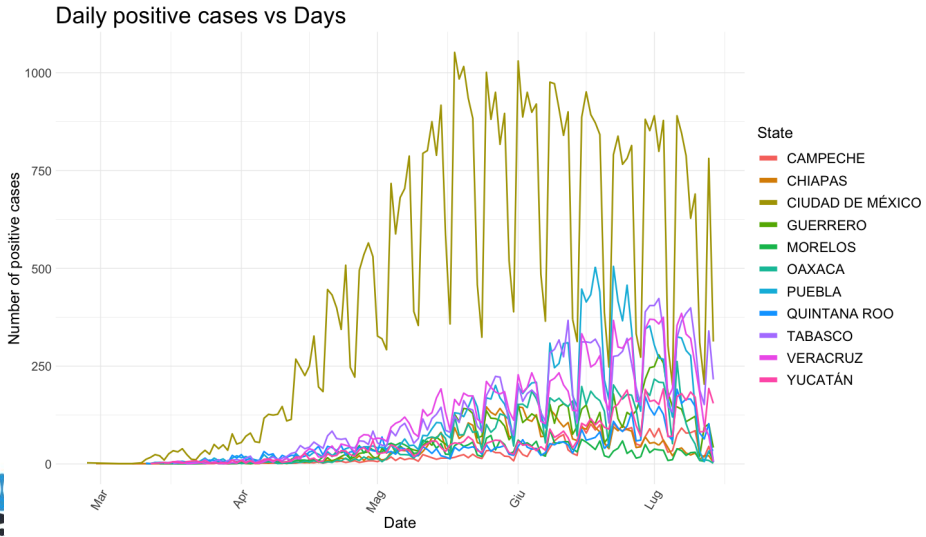


Relations with Demographic Data (2)

Total population sex ratio: 0.96 male(s)/female(s)

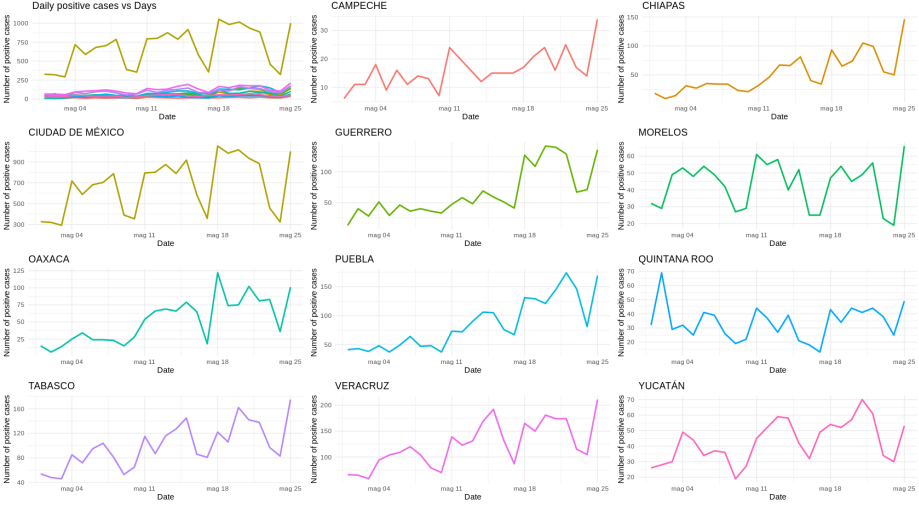


Trend of Daily Positive Cases



Trend of Daily Positive Cases

Zoom on May



Structure

1. Dataset
2. Data Visualization
3. About Models
4. Generalized Linear Model
5. Auto-Regressive Integrated Moving Average
6. Multivariate Adaptive Regression Splines
7. Regression Random Forest
8. Results
9. Conclusions

About Models: Specific Features

Problem: is there a single model for all regions?

- Geographical boundaries
- Federational republic
- Different outbreak's date
- Different responses to the emergency

Structure

1. Dataset
2. Data Visualization
3. About Models
4. Generalized Linear Model
5. Auto-Regressive Integrated Moving Average
6. Multivariate Adaptive Regression Splines
7. Regression Random Forest
8. Results
9. Conclusions

GLM: Model Selection

- Linear Model for regression:

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

- Generalized Linear Model for regression:

$$E[\mathbf{Y}] = g^{-1}(\mathbf{X}\beta)$$
$$\text{Var}[\mathbf{Y}] = \phi V(E[\mathbf{Y}])$$

- Epidemiological dynamics of CoViD-19 can not be modeled with a linear model
- Nonlinear trends can be confirmed by visual analysis

GLM for Count Data[1]

- Poisson:

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

- Negative Binomial:

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

- Exponential dynamics $\rightarrow$ log link function
- How to deal with over-dispersed data?
  - Quasi-Likelihood Poisson $\rightarrow$ $\mathrm{quasipoisson}$
  - Negative Binomial

GLM: Model with Quasi-Poisson

Yucatán Example

```
Call:
glm(formula = DailyConfirmed ~ I(Days) + I(Days^2) + Avg_Age +
    Avg_Sex, family = quasipoisson(link = "log"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.0290  -1.8367  -0.1519   1.2865   6.2292

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.163e-01  3.459e-01   0.336 0.737304
I(Days)      5.379e-02  8.635e-03   6.229 7.19e-09 ***
I(Days^2)   -1.408e-04  5.059e-05  -2.783 0.006251 **
Avg_Age     -1.476e-02  7.647e-03  -1.931 0.055891 .
Avg_Sex      1.466e+00  4.312e-01   3.400 0.000916 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 5.316233)

    Null deviance: 5412.04  on 124  degrees of freedom
Residual deviance:  671.44  on 120  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

GLM: Model with Negative Binomial

Yucatán Example

```
Call:
glm.nb(formula = DailyConfirmed ~ I(Days) + I(Days^2) + Avg_Age +
    Avg_Sex, data = train, link = log, init.theta = 6.994790046)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2073  -0.9454  -0.2258   0.5147   2.6053

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.378e+00  3.028e-01  -4.550 5.37e-06 ***
I(Days)      6.930e-02  8.181e-03   8.471  < 2e-16 ***
I(Days^2)   -2.403e-04  5.308e-05  -4.526 6.01e-06 ***
Avg_Age      6.410e-03  6.893e-03   0.930    0.352
Avg_Sex      1.572e+00  3.727e-01   4.217 2.48e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(6.9948) family taken to be 1)

    Null deviance: 1127.94  on 124  degrees of freedom
Residual deviance:  127.11  on 120  degrees of freedom
AIC: 861.63

Number of Fisher Scoring iterations: 1


              Theta:  6.99
          Std. Err.:  1.23

 2 x log-likelihood:  -849.633
```

GLM: Variables' importance

	CAMPECHE	CHIAPAS	CIUDAD DE MÉXICO	GUERRERO	MORELOS	OAXACA
I(Days)	7.346686	11.6166946	18.386675	11.50853	14.0849772	12.603114
I(Days^2)	3.904368	10.5908366	16.4748834	9.204122	12.7714811	10.597516
Avg_Age	2.648809	3.1264626	4.009676	1.874433	0.3255007	3.789534
Avg_Sex	1.58021	0.1405732	0.1516596	1.245475	0.8631409	1.175461

	PUEBLA	QUINTANA ROO	TABASCO	VERACRUZ	YUCATÁN	MEAN
I(Days)	11.90032508	3.96482	10.4162536	14.561295	6.229021	11.147126
I(Days^2)	7.83828887	1.234229	6.9439227	11.31785	2.783414	8.514628
Avg_Age	6.05049256	1.616365	0.4799553	4.839925	1.930607	2.79016
Avg_Sex	0.06616445	1.25349	1.0553259	2.685541	3.399832	1.237897

- Cases depends on time elapsed
- Presence of a quadratic behaviour
- `Avg_Sex` is a relevant variable only for certain regions
- `Avg_Age` could be correlated to `Days`

GLM: Prediction Intervals

Bootstrap based solution: `add_Pi.glm` {`ciTools`}

- Fit the GLM and collect the regression statistics $\hat{\beta}$ and Cov$\hat{\beta}$
- Simulate M draws of $\hat{\beta}^* \sim N(\hat{\beta}, \hat{\text{Cov}}(\hat{\beta}))$
- Simulate $y^*|x$ from response distribution with mean $g^{-1}(x\hat{\beta}^*)$ 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

Structure

1. Dataset
2. Data Visualization
3. About Models
4. Generalized Linear Model
5. Auto-Regressive Integrated Moving Average
6. Multivariate Adaptive Regression Splines
7. Regression Random Forest
8. Results
9. Conclusions

ARIMA: Auto-Regressive Integrated Moving Average[2]

- Three parameters: (p, d, q)
- Auto-Regressive (**AR(p)**) component, p is the number of lags used
- Integrated (**I(d)**) component, d is the degree of differencing
- Moving Average (**MA(q)**) component, q is the number of previous terms included in the model's error
- Data are stationary (tested with ACF)
- Use MLE to estimate the model for given parameters
- Use AIC and BIC to select the best model

ARIMA: Model

Yucatán Example

```
ARIMA(3,1,2) with drift

Coefficients:
         ar1      ar2      ar3      ma1     ma2   drift
      0.8501  -0.4219  -0.3760  -1.4819  0.8583  1.1635
s.e.  0.1219   0.1318   0.1264   0.0901  0.0698  0.5067

sigma^2 estimated as 209.3:  log likelihood=-506.03
AIC=1026.06   AICc=1027.03   BIC=1045.81

Training set error measures:
                     ME     RMSE      MAE  MPE MAPE      MASE        ACF1
Training set 0.006030542 14.05638 9.351789 -Inf  Inf 0.8798345 -0.01872235
```

ARIMA: Prediction Intervals

- Fitted values are in-sample one-step forecasts
- Assumed uncorrelated and normally distributed residuals
- 95% prediction intervals for the first step

$$\left[\hat{y}_{T+1|T} - 1.96\,\hat{\sigma},\ \hat{y}_{T+1|T} + 1.96\,\hat{\sigma}\right]$$

- This result is true for all ARIMA models regardless of their parameters and orders
- 95% prediction intervals for multi-step

$$\left[\hat{y}_{T+h|T} - 1.96\,\sqrt{\hat{\sigma}_h},\ \hat{y}_{T+h|T} + 1.96\,\sqrt{\hat{\sigma}_h}\right]$$

- where $\hat{\sigma}_h = \hat{\sigma}^2\left[1 + \sum_{i=1}^{h-1}\hat{\theta}_i^2\right]$, $h = 2, 3, \ldots$

Structure

1. Dataset
2. Data Visualization
3. About Models
4. Generalized Linear Model
5. Auto-Regressive Integrated Moving Average
6. Multivariate Adaptive Regression Splines
7. Regression Random Forest
8. Results
9. Conclusions

MARS: Multivariate Adaptive Regression Splines[3]

- Semi-parametric model
- Builds models in the form:

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

- Three forms of basis functions $B_i(x)$:
  - Constant
  - Hinge function: $\max(0, x - c)$ or $\max(0, c - x)$, c = constant
  - Product of two or more hinge functions

Hyperparameters tuning:

- `nfold = 10, ncross = 30`
- `degree = 1`

MARS: Model

Yucatán Example

```
Call: earth(formula=DailyConfirmed~I(Days)+Avg_Age+Avg_Sex, data=train, degree=1, nfold=10, ncross=30,
            varmod.method="earth")

                    coefficients
(Intercept)            0.2515766
h(I(Days)-35)          0.8480897
h(I(Days)-90)         -5.5200594
h(I(Days)-95)          8.6501952
h(Avg_Age-37.4433)     6.2605914
h(Avg_Age-40.5132)    -9.0820643
h(Avg_Age-49.85)       3.5752121

Selected 7 of 14 terms, and 2 of 3 predictors
Termination condition: Reached nk 21
Importance: I(Days), Avg_Age, Avg_Sex-unused
Number of terms at each degree of interaction: 1 6 (additive model)
GCV 230.0053  RSS 23081.49  GRSq 0.8742305  RSq 0.8973952  CVRSq 0.7590279

Note: the cross-validation sd's below are standard deviations across folds

Cross validation:   nterms 7.12 sd 1.25    nvars 2.08 sd 0.27

     CVRSq    sd     MaxErr   sd
     0.759 0.275      -96.7 44.3

varmod: method "earth"    min.sd 1.31    iter.rsq 0.444

stddev of predictions:
                          coefficients iter.stderr iter.stderr%
(Intercept)                 14.3497435     1.79792            13
h(50.8927-DailyConfirmed)   -0.1805278   0.0523045            29
h(DailyConfirmed-50.8927)    0.2719154   0.0426825            16

                               mean   smallest    largest      ratio
95% prediction interval   51.31913   16.11893   169.3512   10.50635

                                      68%    80%    90%    95%
response values in prediction interval   76     86     93     98
```

MARS: Prediction Intervals

Computed using `predict.earth` with `interval` argument:

- Estimate the mean absolute error at each point using a residual model
- Assuming normality, re-scale the error to an estimated standard deviation:

$$sd = \sqrt{\frac{\pi}{2}} E\left[|\epsilon|\right]$$

- Convert the standard deviation to an estimated prediction interval for a given level α:

$$\left[\hat{Y} - z_{\frac{\alpha}{2}} sd, \hat{Y} + z_{\frac{\alpha}{2}} sd\right]$$

Structure

1. Dataset
2. Data Visualization
3. About Models
4. Generalized Linear Model
5. Auto-Regressive Integrated Moving Average
6. Multivariate Adaptive Regression Splines
7. Regression Random Forest
8. Results
9. Conclusions

RF: Random Forest[4]

- Based on the decision tree (regression) model
- Each tree is grown on a bootstrap sample of the training set
- Bagging + random selection of features for splitting

Hyperparameters tuning:

- `mtry = 1`
- `num.trees = 100`

RF: Prediction Intervals

Computed using `RFIntervals`:

- Exploits the empirical quantile distribution of OOB prediction errors
- As n and B increase, the difference between the distribution of D and the empirical distribution of $D_1, \ldots, D_n$, becomes negligible, so it's reasonable to assume:

$$1 - \alpha \approx P\left[D_{[n,\alpha/2]} \leq D \leq D_{[n,1-\alpha/2]}\right]$$

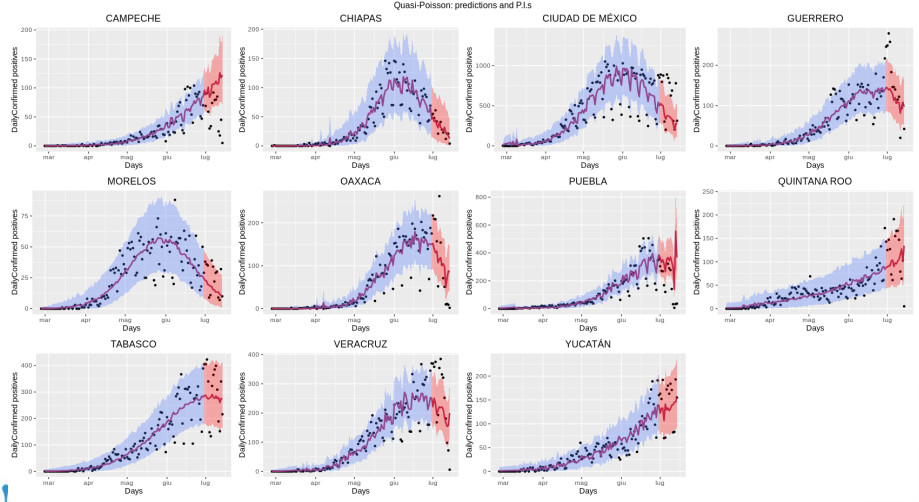
- Being $D = Y - \hat{Y}$ the prediction interval is given by:

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

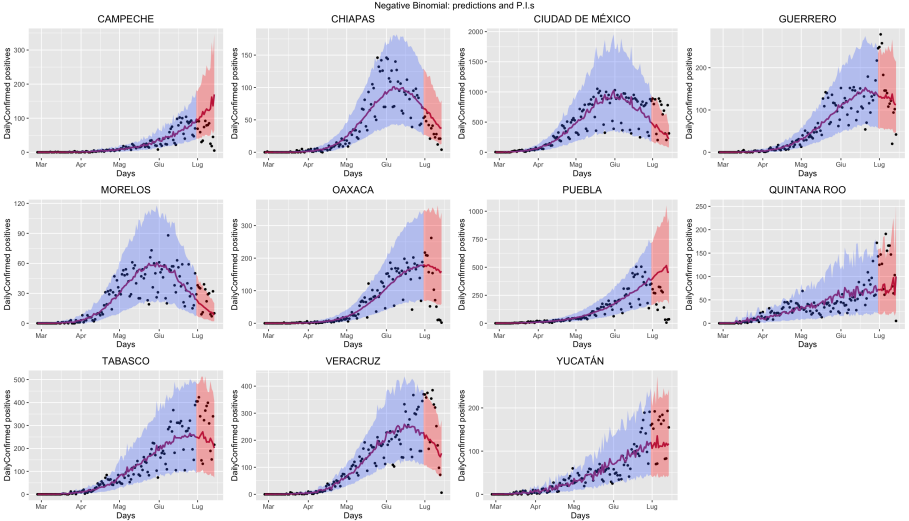
Structure

1. Dataset
2. Data Visualization
3. About Models
4. Generalized Linear Model
5. Auto-Regressive Integrated Moving Average
6. Multivariate Adaptive Regression Splines
7. Regression Random Forest
8. Results
9. Conclusions

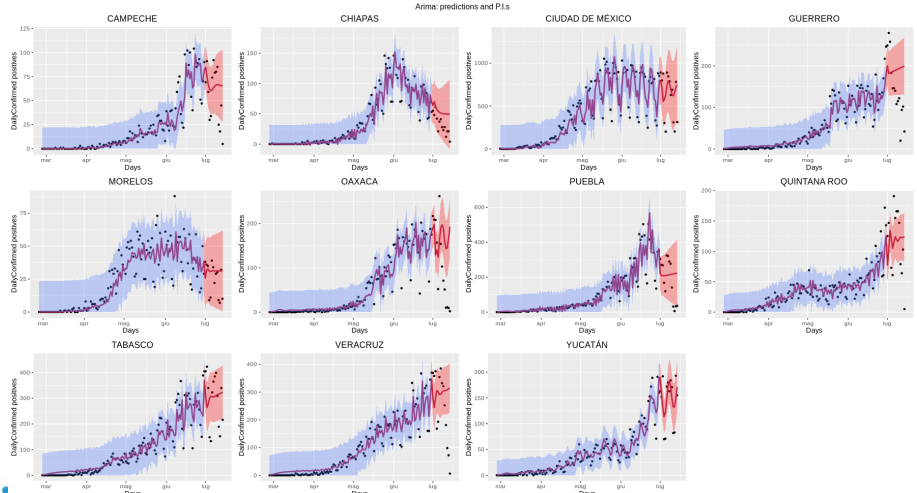
QuasiPoisson Results



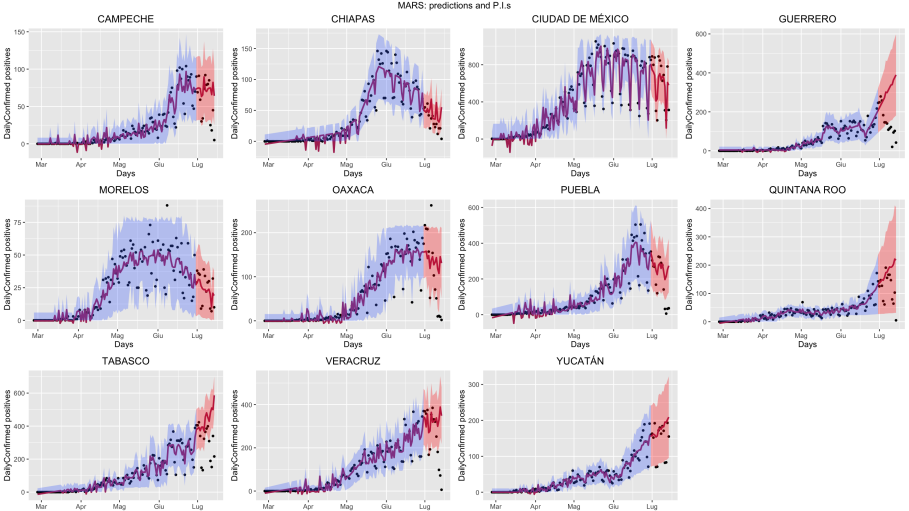
Negative Binomial Results



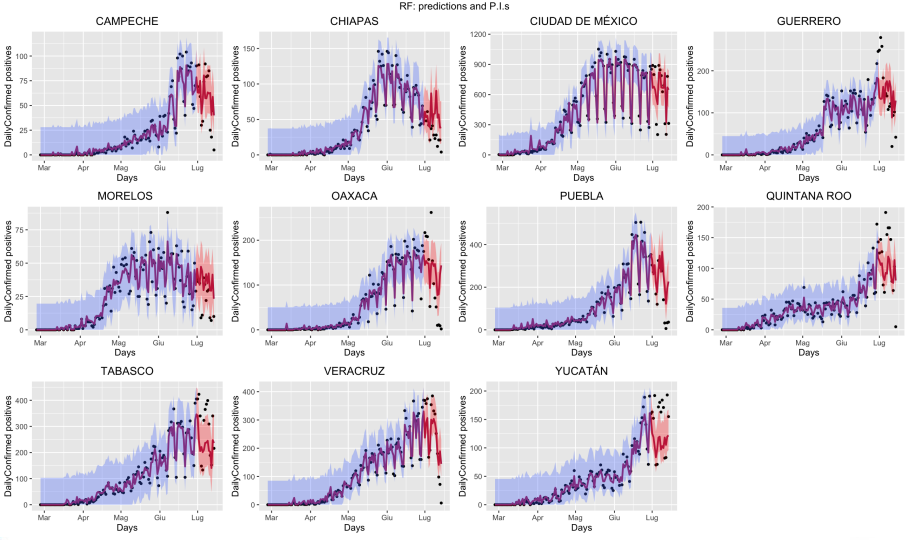
ARIMA Results



MARS Results



RF Results


RF: predictions and P.I.s
CAMPECHE
CHIAPAS
CIUDAD DE MÉXICO
GUERRERO
MORELOS
OAXACA
PUEBLA
QUINTANA ROO
TABASCO
VERACRUZ
YUCATÁN
DailyConfirmed positives
Days
Mar
Apr
Mag
Giu
Lug

Accuracy measure: RMSE

RMSE of each model

COVID 19 South Mexico

State	QuasiPoisson	NegBin	MARS	RF	ARIMA
CAMPECHE	59.20	80.21	30.67	24.97	29.30
CHIAPAS	7.69	22.03	22.43	31.11	25.46
CIUDAD DE MÉXICO	327.95	363.24	168.17	242.07	256.66
GUERRERO	62.97	73.85	203.00	73.80	91.20
MORELOS	13.77	14.48	9.89	17.81	13.99
OAXACA	67.27	102.77	84.50	74.28	105.49
PUEBLA	204.08	304.60	121.89	98.65	124.12
QUINTANA ROO	62.47	71.99	89.11	43.43	54.08
TABASCO	99.88	119.88	197.29	98.30	108.98
VERACRUZ	112.42	122.79	151.22	81.12	143.88
YUCATÁN	46.55	56.44	56.13	55.52	35.44

Structure

1. Dataset
2. Data Visualization
3. About Models
4. Generalized Linear Model
5. Auto-Regressive Integrated Moving Average
6. Multivariate Adaptive Regression Splines
7. Regression Random Forest
8. Results
9. Conclusions

Conclusions

- Analysis of 11 different states: No Free Lunch Theorem
- Most of the times RF has the lowest RMSE
- Counter-examples: Ciudad de México, Chiapas, Yucatán
- Looking at the results RF seems the best model to predict the decreasing behaviour of daily cases, so, even if there is no Free Lunch, RF is a good option to analyse and predict the number of daily positive cases in South Mexico

References

1. Haman J. (2020). *Generalized Linear Models with ciTools* R package. version 0.5.2, URL `https://cran.r-project.org/web/packages/ciTools/index.html`
2. Rob J Hyndman and George Athanasopoulos (2018). *Forecasting: principles and practice*, OTexts: Melbourne, Australia.
3. Milborrow S. (2019). *earth: Multivariate Adaptive Regression Spline Models.* R package. version 5.1.2, URL `https://cran.r-project.org/web/packages/earth/index.html`
4. Haozhe Zhang et al. (2019). *Random Forest Prediction Intervals*, The American Statistician