

# Filling the gaps: trying to predict cases in the weekend of March 7-8th in Spain

*J. J. Merelo*

*17 de marzo de 2020*

## Abstract

The Spanish Government declared a national emergency on March 14th, 2020 due to the coronavirus outbreak. That weekend, daily updates of the Ministry of Health official Bulletins started to be produced on a daily basis. The previous week, during which a government-sponsored parade took place, there were no bulletins. In this report we will try to predict, on the basis of previously published data, the actual numbers of infected and deceased people that could have been published during that two-days blackout. We check the model by using it to predict the data when it was finally updated on March 9th.

## Introduction

One of the essential weapons to fight pandemics is transparency. Nothing hurts more than out-guessing your government, or simply finding that the government is hiding information (Nishiura et al. 2020); uncertainty is the worst symptom for a regular citizen, but it might be even worse if you feel that information is trickled by political motivation.

That is one of the accusations that have been leveraged against the Spanish government. After a responsible person for the COVID-19 outbreak had been appointed, and even after the Italian government had taken harsh measures to stop the pandemic, from the power citizens were encouraged to attend one of the several political parades that had been organized on March 8th. No other public demonstrations were banned, either.

Those messages were accompanied by lack of information from the Health Ministry. The Ministry started publishing daily bulletins in [its web page](#). No bulletins were issued on weekends before that, and that was the main reason put forward by officials for not doing it this precise weekend. The very next day, two bulletins were published: [number 39 by 12:00](#) and [number 39b by 18:00](#). There was a 300 cases difference between them, which couldn't possibly have happened in the 6 hours that separated them. The case tally exceeded 1000, and government officials started to take measures, starting not by the national government, but by the regional government in Madrid. This was followed by other measures.

It's very likely the case that the number of cases exceeded 1000 during the weekend. In this report we are going to examine the published time series, and create a model that computes the number of cases during the weekend. We will try to find if the numbers reported on Monday could actually been reached somewhere during the weekend. We'll double-check the model against actual number reported after the blackout for sanity.

## Methodology

Data has been extracted from the [Datadista repository](#), in CSV format. This is the table of data available until March 6th:

| ##   | fecha      | casos | altas | fallecimientos | ingresos_uci |
|------|------------|-------|-------|----------------|--------------|
| ## 1 | 2020-02-25 | 3     | NA    | NA             | NA           |
| ## 2 | 2020-02-26 | 10    | NA    | NA             | NA           |
| ## 3 | 2020-02-27 | 16    | NA    | NA             | NA           |
| ## 4 | 2020-02-28 | 32    | NA    | NA             | NA           |
| ## 5 | 2020-02-29 | 44    | NA    | NA             | NA           |
| ## 6 | 2020-03-01 | 66    | NA    | NA             | NA           |

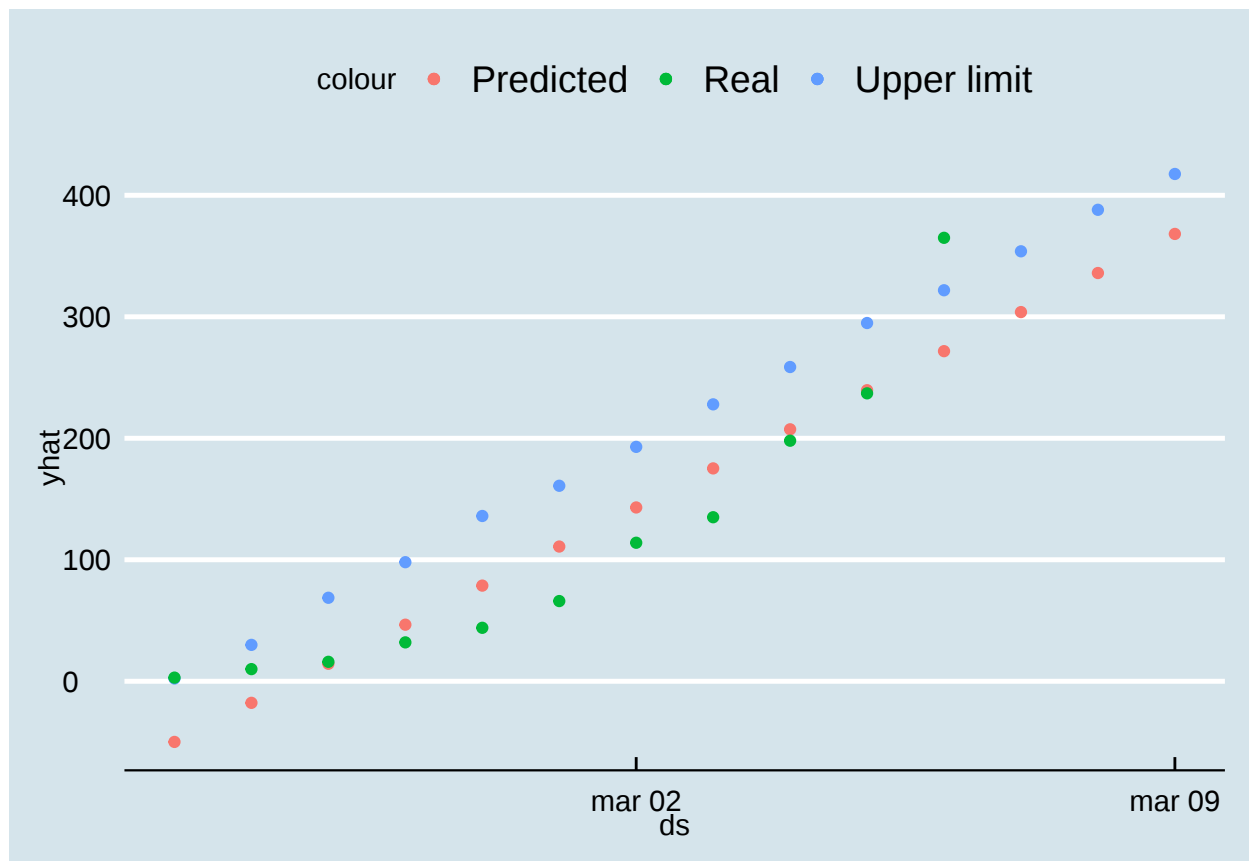
```
## 7 2020-03-02 114 NA NA NA
## 8 2020-03-03 135 NA NA NA
## 9 2020-03-04 198 NA 1 7
## 10 2020-03-05 237 NA 3 9
## 11 2020-03-06 365 NA 5 11
```

As it can be seen, there is a gap on February 29th and March 1st, which was the previous weekend. This is the meaning of the columns of the data frame

| Column         | Translation    |
|----------------|----------------|
| fecha          | date           |
| casos          | infected       |
| ingresos_uci   | treated in ICU |
| fallecimientos | deceased       |
| altas          | discharged     |

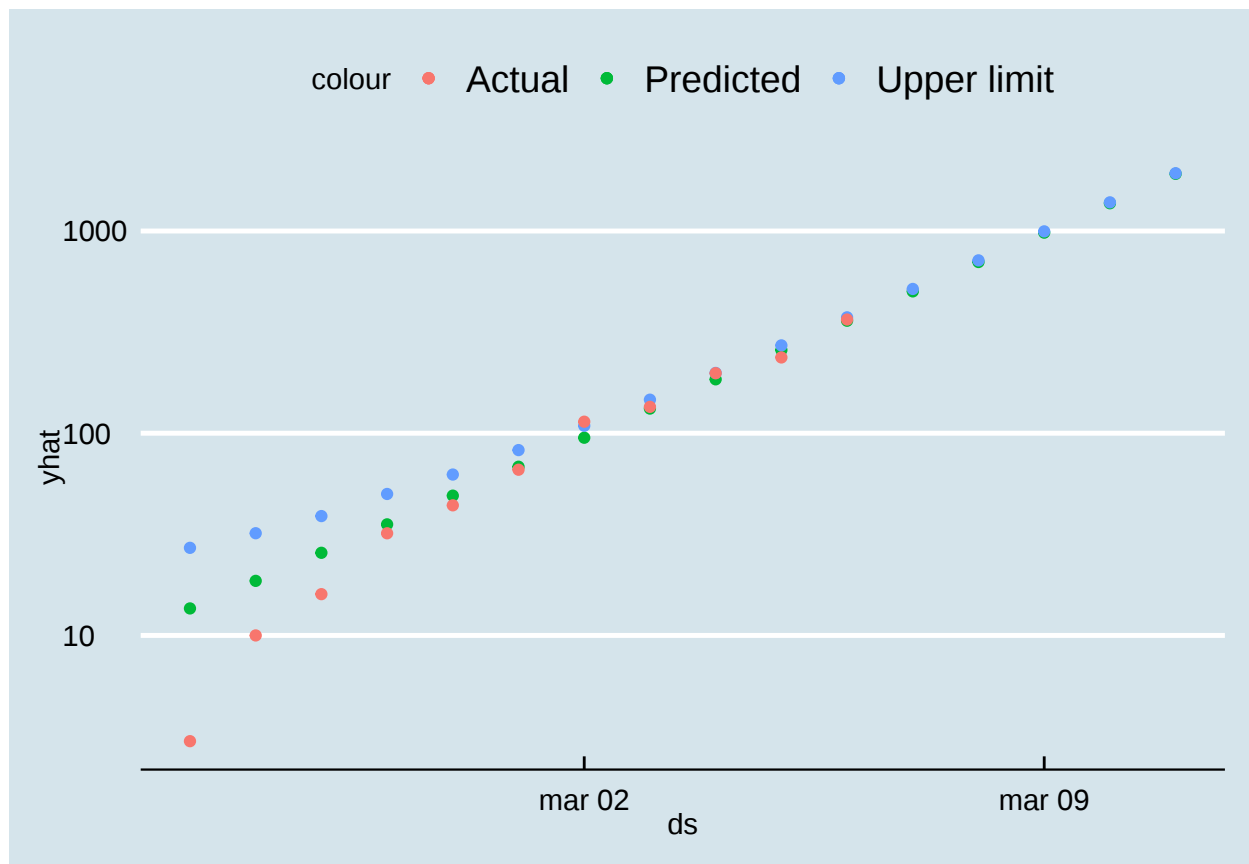
However, the rise in the previous days had been worrying, so it might have been sensible to publish data on that Saturday so that citizens could assess their risk when attending crowded environments.

We will use [Prophet](#) to create a model and compute these numbers. Number of cases predicted are below, along with the higher and lower bound.



This linear model already predicted between 350 and 400 infected on the Saturday, and closer to 400 on the Sunday; however, this model predicts values that are already under known cases (in green) on Friday, so it is a very optimistic one. We need to try another, non-linear, model.

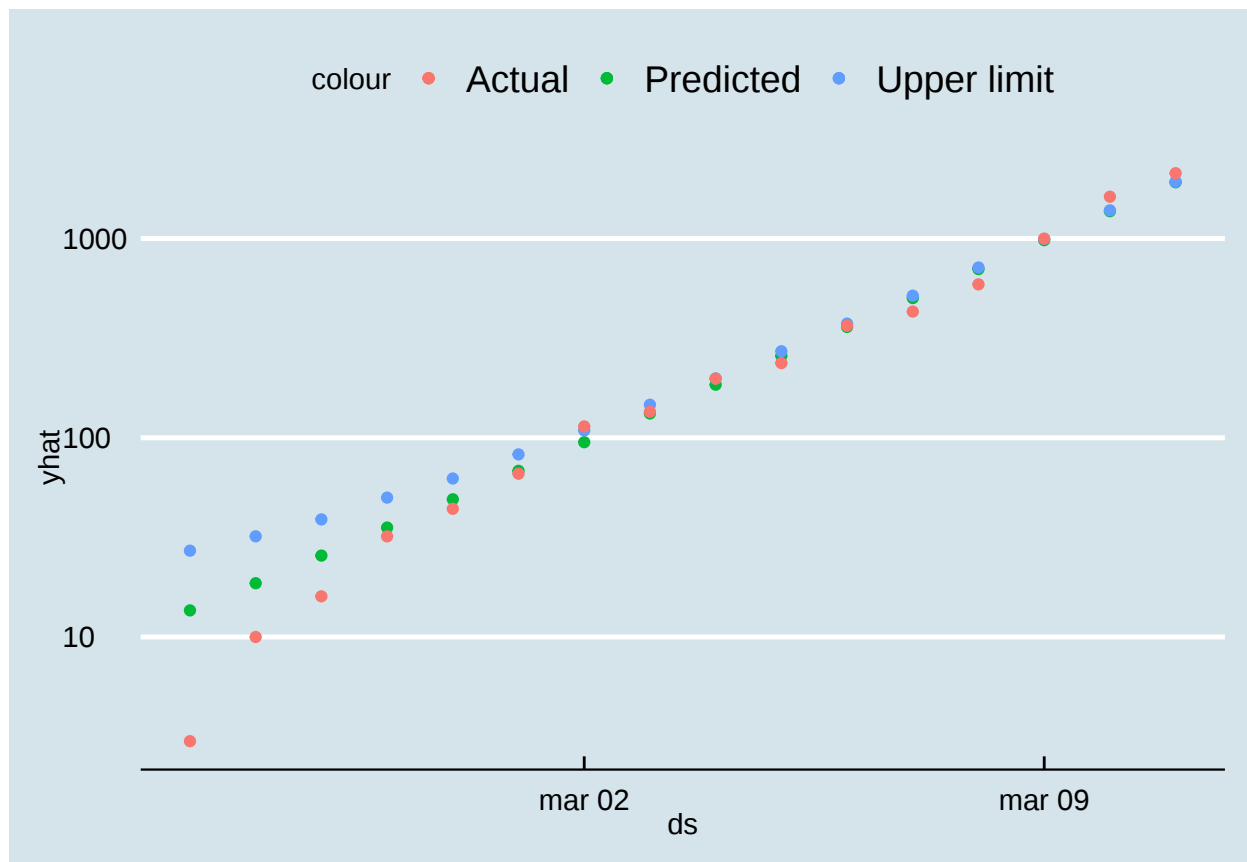
Prophet offers two options of growth. Linear is the default used above. Let's try the other version, logistic.



Quite obviously, the logistic curve reflects quite well what happened, at least in the last few days, with upper limit and predicted values becoming increasingly closer.

## Reality check against actual values

As a matter of fact, this blackout during the weekend was not general, since the Ministry of Health maintains a channel open with journalists, and the values for these dates were revealed in that channel. These values were also [published under request by El Datadista](#), as a matter of fact while I was writing this report. Let's then check predicted against real values next.



The predicted number of cases slightly leads the real number of cases on the days of the (general public, not journalists) blackout. Let's see the real numbers in a table:

| ##    | date       | actual | predicted  | difference  |
|-------|------------|--------|------------|-------------|
| ## 1  | 2020-02-25 | 3      | 13.60169   | -10.601688  |
| ## 2  | 2020-02-26 | 10     | 18.61550   | -8.615499   |
| ## 3  | 2020-02-27 | 16     | 25.62415   | -9.624146   |
| ## 4  | 2020-02-28 | 32     | 35.42131   | -3.421308   |
| ## 5  | 2020-02-29 | 44     | 49.11644   | -5.116444   |
| ## 6  | 2020-03-01 | 66     | 68.26043   | -2.260428   |
| ## 7  | 2020-03-02 | 114    | 95.02117   | 18.978829   |
| ## 8  | 2020-03-03 | 135    | 132.42912  | 2.570885    |
| ## 9  | 2020-03-04 | 198    | 184.72039  | 13.279606   |
| ## 10 | 2020-03-05 | 237    | 257.81652  | -20.816517  |
| ## 11 | 2020-03-06 | 365    | 359.99488  | 5.005121    |
| ## 12 | 2020-03-07 | 430    | 502.82600  | -72.826005  |
| ## 13 | 2020-03-08 | 589    | 702.48360  | -113.483603 |
| ## 14 | 2020-03-09 | 999    | 981.57568  | 17.424319   |
| ## 15 | 2020-03-10 | 1622   | 1371.70388 | 250.296122  |
| ## 16 | 2020-03-11 | 2128   | 1917.04042 | 210.959581  |

This check comes with a small twist, too. While during March 6th and 7th the prediction yields higher values than those reported (although only to the press), the actual cases when the general public blackout ended are *higher* than the prediction, to the point that two days into the week of March 9th reported cases exceed predictions by more than 200.

Either taking into account the actual (430) number of cases on Saturday 7th or predicted (702) on Sunday 8th, authorities should probably have taken other kind of decisions, or at least not actively encouraged

participation in mass parades. In either scenario, 1000 cases would have been reached any time during March 9th.

The fact that there are more reported cases, by a factor of almost 15th, than those that could be predicted with a model computed with cases up to March 6th, remains to be explained. Several explanations are possible, including under-reporting of cases before active measures started to be taken on Monday 9th. But this is left as future work.

## Conclusions

Transparency is of the utmost importance when dealing with a pandemic, and reporting accurate values, as well as putting in place the mechanism to be able to report them to the public, is very important.

While some people might claim that the direction of the outbreak couldn't be predicted, this report shows that a relatively simple script could, in fact, do that. The code is available in the source of this article at [this GitHub repository](#), with sources of data embedded in them.

For the time being, the fact that the model predicted less cases than the actual ones in the days following March 8th remains unexplained; reported cases are around 15% higher of what the model forecasted for those dates. One reason could be that, before the regional and national governments started to take measures, infections were undertested and thus underreported. Testing other hypotheses is left as future work.

## Acknowledgements

This report has been generated from data extracted from the Ministry of Health database and published by [Datadista](#) in GitHub under a free License. This paper uses RMarkdown, includes all code needed to generate it, and can be downloaded from [this repository](#).

## References

Nishiura, Hiroshi, Tetsuro Kobayashi, Yichi Yang, Katsuma Hayashi, Takeshi Miyama, Ryo Kinoshita, Natalie M Linton, et al. 2020. "The Rate of Underascertainment of Novel Coronavirus (2019-nCoV) Infection: Estimation Using Japanese Passengers Data on Evacuation Flights." *JCM*. Multidisciplinary Digital Publishing Institute.