Supplementary materials: Forecasting COVID-19 spreading through an ensemble of classical and machine learning models: Spain’s case study

Ignacio Heredia Cacha¹, Judith Sáinz-Pardo Díaz¹, María Castrillo Melguizo¹, and Álvaro López García¹ *

¹Instituto de Física de Cantabria (IFCA), CSIC-UC, Avda. los Castros s/n. 39005 - Santander (Spain)
*aloga@ifca.unican.es

ABSTRACT

This document contains the supplementary materials for the manuscript “Forecasting COVID-19 spreading through an ensemble of classical and machine learning models: Spain's case study”.
Analysis by Autonomous Community

In the following figure the MAPE obtained considering scenario 4 (the most complete) for the machine learning models, and the 3 forms of aggregation: mean, median and weighted mean, is shown for the 17 Spanish autonomous communities.

| Autonomous Community | ML (mean) | ML (median) | ML (WW/G) | Pop (mean) | Pop (median) | Pop (WW/G) | ML + Pop (mean) | ML + Pop (median) | ML + Pop (WW/G) |
|----------------------|-----------|-------------|-----------|------------|-------------|------------|----------------|------------------|-----------------|
| ANDALUCÍA            |           |             |           |            |             |            |                 |                   |                 |
| ARAGÓN               |           |             |           |            |             |            |                 |                   |                 |
| PRINCIPADO DE ASTURIAS|           |             |           |            |             |            |                 |                   |                 |
| CANTABRIA            |           |             |           |            |             |            |                 |                   |                 |
| CASTILLA Y LEÓN      |           |             |           |            |             |            |                 |                   |                 |
| CASTILLA-LA MANCHA   |           |             |           |            |             |            |                 |                   |                 |
| CANARIAS             |           |             |           |            |             |            |                 |                   |                 |
| CATALUÑA             |           |             |           |            |             |            |                 |                   |                 |
| EXTREMADURA          |           |             |           |            |             |            |                 |                   |                 |
| GALICIA              |           |             |           |            |             |            |                 |                   |                 |
| ISLAS BALEARES       |           |             |           |            |             |            |                 |                   |                 |
| REGIÓN DE MURCIA     |           |             |           |            |             |            |                 |                   |                 |
| COMUNIDAD DE MADRID  |           |             |           |            |             |            |                 |                   |                 |
| COMUNIDAD FORAL DE NAVARRA | |           |           |            |             |            |                 |                   |                 |
| PAÍS VASCO           |           |             |           |            |             |            |                 |                   |                 |
| LA RIOJA             |           |             |           |            |             |            |                 |                   |                 |
| COMUNIDAD VALENCIANA |           |             |           |            |             |            |                 |                   |                 |

**Supplementary figure 1.** Summary of the MAPE obtained for the cases of the 17 Spanish autonomous communities with the different models. ML models are trained in Scenario 4.
The main aspect to highlight in the previous figure is that in most cases it is still true that the ensemble of models (with any of the three forms of aggregation), manages to improve the MAPE results of the two families of models individually. This particularly clear in Autonomous Communities like Aragón, Cantabria, Castilla y León, Castilla La Mancha, Cataluña or País Vasco.

Regarding whether Machine Learning or population models obtain better results, it is observed that this changes depending on the autonomous community. In some of them, ML models always obtain better MAPE than population ones (e.g. Madrid, Castilla La Mancha, and La Rioja), while in others cases population models obtain better results than ML ones (e.g. Cantabria, Castilla y León and Cataluña).

**Supplementary materials**

*Supplementary figure 2.* Performance of some of the additional ML configuration we tried, for the Spain case in the test split. Models are aggregated with mean aggregation.
### Supplementary figure 3.

MAPE and RMSE obtained in each scenario (1, 2, 3, 4) according to each form of aggregation (mean, median, weighted) for different subsets of models (ML, Pop, All). We show averages for three different splits: the full test split, the test split before Omicron variant started (mid November), and the test split after Omicron variant started. All predictions are made for Spain.
Supplementary figure 4. Mean Percentage Error for each time step of the forecast, grouped by model family. Shades show the standard deviation between models of the same family. ML models are trained in Scenario 4. We show results for three different splits: the full test split, the test split before Omicron variant started (mid November), and the test split after Omicron variant started. All predictions are made for Spain. The no-omicron plot does not exactly reproduce the degradation of ML at large timesteps seen in Figure 9 because the split is not as stable as validation (i.e. Omicron is slowly appearing, driving ML models to underestimation).

Supplementary figure 5. Mean MAPE and RMSE, for each model, for the Spain case in the test split. ML models are trained in Scenario 4.

Supplementary figure 6. SHAP dependence plot for non-cases features. For each feature we plot the raw value of the feature vs its associated SHAP value. We average SHAP values across all ML models. All models are trained in Scenario 4 for the Spain case. We display values for all the dataset (train + val + test).
Supplementary figure 7. SHAP dependence plot for cases lags features. For each feature we plot the raw value of the feature vs its associated SHAP value. We average SHAP values across all ML models. All models are trained in Scenario 4 for the Spain case. We display values for all the dataset (train + val + test).

Explicit solution of the ODE of the Gompertz model and estimation of the initial parameters

Remember that the ODE which defines the Gompertz model is given by:

\[
\frac{dp}{dt} = ap(t) - bp(t) \log(p(t)),
\]

being \(p(t)\) the population at time \(t\), and \(a\) and \(b\) two parameters to determine.
Taking $y(t) = \log(p(t))$, we obtain its explicit solution as follows (note that we consider $y'(t) := \frac{dy}{dt}$ in order to simplify the notation):

\[
\frac{\partial p}{\partial t} = ap(t) - bp(t)\log(p(t)) \implies \frac{\partial y}{\partial t} = a - by(t) \implies y'(t) + by(t) = a \implies e^{bt}y'(t) + e^{bt}by(t) = e^{bt}a \implies e^{bt}y(t) = a \int e^{bt} dt \implies\]

\[
\implies y(t) = \frac{a}{b} + ce^{-bt} \implies p(t) = e^{\frac{a}{b} + ce^{-bt}}.
\]

In order to estimate the parameters $a$, $b$, and $c$ we fix three time instants $t_i$, $t_j$, and $t_k$ verifying: $h = t_j - t_i$ and $2h = t_k - t_i$. Be \( \alpha = \frac{\log(p(t_j)) - \log(p(t_i))}{\log(p(t_k)) - \log(p(t_i))} \) we get:

- $\alpha = \frac{e^{-bt_j} - e^{-bt_i}}{e^{-bt_k} - e^{-bt_i}} \implies \alpha = \frac{1}{1 + e^{-bh}} \implies b = -\frac{1}{h} \log \left( \frac{1 - \alpha}{\alpha} \right)$

- $\frac{p(t_j)}{p(t_i)} = e^{(e^{-bt_j} - e^{-bt_i})} \implies c = \frac{\log(p(t_j)) - \log(p(t_i))}{e^{-bt_j} - e^{-bt_i}}$

- $p(t_i) = e^{\frac{a}{b} + ce^{-bt_i}} \implies a = b \left( \log(p(t_i)) - ce^{-bt_i} \right)$

The process to be followed to obtain the initial parameters for the Logistic and Bertalanffy models is analogous to the previous one presented for the Gompertz case.