Four fitting model comparison for SARS-nCoV-2 total cumulative cases curve in 185 countries.

Max Pierini*

May 1, 2020

1 Abstract

Four different models have been compared for fitting SARS-nCoV-2 total cumulative cases curves in 185 countries over a period of 97 days. Evaluated models have been: *Simple Logistic Function* (**SLF**), *Simple Gompertz Function* (**SGF**), *Double Logistic Function* (**DLF**) and *Double Gompertz Function* (**DGF**). **DGF** model showed lower MSE, RMSE, NRMSE, MAE, NAE and higher Pearson R compared to the others.

SGF χ^2 scores have been found higher in: 170 countries compared with **SLF** ($\mu = 0.989$, $\sigma = 0.050$); 174 countries compared with **SGF** ($\mu = 0.969$, $\sigma = 0.110$); 154 countries compared with **DLF** ($\mu = 0.992$, $\sigma = 0.046$).

SGF χ^2_ν scores have been found higher: 165 countries compared with **SLF** ($\mu = 0.902$, $\sigma = 0.176$); 155 countries compared with **SGF** ($\mu = 0.889$, $\sigma = 0.185$); 143 countries compared with **DLF** ($\mu = 0.899$, $\sigma = 0.180$).

SGF AIC scores have been found higher: 158 countries compared with **SLF** ($\mu = 0.985$, $\sigma = 0.068$); 153 countries compared with **SGF** ($\mu = 0.991$, $\sigma = 0.049$); 137 countries compared with **DLF** ($\mu = 0.991$, $\sigma = 0.044$).

SGF BIC scores have been found higher in: 149 countries compared with **SLF** ($\mu = 0.985$, $\sigma = 0.063$); 143 countries compared with **SGF** ($\mu = 0.990$, $\sigma = 0.046$); 137 countries compared with **DLF** ($\mu = 0.992$, $\sigma = 0.044$).

Results suggest that *Double Gompertz Function* may be a good fitting model for SARS-nCoV-2 analysis and forecasting.

2 Methods

2.1 Data

SARS-nCoV-2 total cumulative cases data have been gathered from Johns Hopkins University GitHub repository [REF] and summed into single countries where regional level was provided [REF]. Data have been used "as is" without rejecting any outlier and/or error (negative daily Δ). Data and results have been stored in a `pandas` n-dimensional `DataFrame`.

*info@maxpierini.it

Raw data contained 185 countries and daily cumulative confirmed cases for 97 days, from 2020-01-22 to 2020-04-29.

2.2 Models

Models have been defined with `lmfit` (implementation of classical `curve_fit` in `scipy`) using Nelder-Mead method for fitting [REF].

Total residual from each function have been initially compared (unsorted, sorted, gaussian distribution) to find the model with residual μ closer to 0 and shorter σ. *Akaike Information Criterion* (**AIC**) mean, *Bayesian Information Criterion* (**BIC**) mean and χ^2 scores mean have been used to find the likely better fitting model that has been finally compared, country by country, with *AIC scores* probability (**AICp**), *BIC scores* probability (**BICp**) and χ^2 scores probability in relative probability distribuition space. See Appendix [REF] for formulas.

Models have been defined as follow:

- Simple Logisic Function (**SLF**):

```
def logit_function(x, a, b, k, e):
    d = k * (b - np.array(x))
    return (a / (1 + np.exp(d))) + e
```

$$f(t) = \frac{a}{1 + e^{k(b-t)}} + \varepsilon$$

- Double Logisic Function (**DLF**):

```
def double_logit_function(x, a1, b1, k1, a2, b2, k2, e):
    d1 = k1 * (b1 - np.array(x))
    g1 = a1 / (1 + np.exp(d1))
    d2 = k2 * (b2 - np.array(x))
    g2 = (a2 - a1) / (1 + np.exp(d2))
    return g1 + g2 + e
```

$$f(t) = \frac{a_1}{1 + e^{k_1(b_1-t)}} + \frac{a_2 - a_1}{1 + e^{k_2(b_2-t)}} + \varepsilon$$

- Simple Gompertz Function (**SGF**):

```
def gompertz_function(x, a, b, k, e):
    exp = - np.exp(k * (b - x))
    return a * np.exp(exp) + e
```

$$f(t) = a \cdot e^{-e^{k(b-t)}} + \varepsilon$$

- Double Gompertz Function (**DGF**):

```
def double_gompertz_function(x, a1, b1, k1, a2, b2, k2, e):
    exp1 = - np.exp(k1 * (b1 - x))
    g1 = a1 * np.exp(exp1)
    exp2 = - np.exp(k2 * (b2 - x))
```

```
    g2 = (a2 - a1) * np.exp(exp2)
    return g1 + g2 + e
```

$$f(t) = a_1 \cdot e^{-e^{k_1(b_1-t)}} + (a_2 - a_1) \cdot e^{-e^{k_2(b_2-t)}} + \varepsilon$$

3 Model fitting

Fitting has been performed with `lmfit` using Nelder-Mead method

```
model = lmfit.Model(function)
result = model.fit(data=y, params=p, x=x, method='Nelder', nan_policy='omit')
```

initial parameters `p` have been guessed as follows (where $\hat{y}$ are measured values):

- **SLF** and **SGF**

```
p = model.make_params(
    a=y[-1],
    b=max_y_i,
    k=.1,
    e=y[0]
)
```

$$a = \hat{y}_{-1}$$
$$b = x_{\mathbf{max}(\hat{y})}$$
$$k = 0.1$$
$$\varepsilon = \hat{y}_0$$

- **DLF** and **DGF**

```
p = model.make_params(
    a1=y[max_y_i] * 2,
    b1=max_y_i,
    k1=.1,
    a2=max(y),
    b2=len(y),
    k2=.1,
    e=y[0]
)
```

$$a_1 = 2\hat{y}_{\mathbf{max}(d\hat{y})}$$
$$b_1 = x_{\mathbf{max}(\hat{y})}$$
$$k_1 = 0.1$$
$$a_2 = \mathbf{max}(\hat{y})$$
$$b_2 = x_{\hat{y}_{-1}}$$

$$k_2 = 0.1$$

$$\varepsilon = \hat{y}_0$$

Fitting failed for one country only (Yemen) returning best fit information from 184 countries, for a total of 17,848 points for each of the four models.

Complete `python` backend for data gathering, fitting and analysis along with a `pickle` saved dataframe of all measured data and results is online avalaible at [REF].

Fitting examples are reported in figures [REF] [REF] [REF].

4 Analysis

Several skill score have been to average skil and skill interpolating extreme values (outliers).

Mean absolute error (**MAE**) is a natural, unambiguous, measure of average error [Willmott and Matsuura, 2006]. It shows the errors in the same unit as variables themselves. MAE is bounded below by 0 (best case) and unbounded above.

Normalized Absolute Error (**NAE**) …

Mean Squared Error (**MSE**) …

Root Mean Squared Error (**RMSE**) is very commonly used as a measure of deviation from the observed value. Although is has been criticized as being ambiguous [Willmott and Matsuura, 2006] and its dependence on the squared error means that it is not resistant to outliers deviating from a Gaussian distribution. We include it because of its sensitivity to large outliers.

Normalized Root Mean Squared Error (**NRMSE**) …

Pearson Correlation coefficient (**Pearson R**) depends on squared deviations and so is not a resistant measure. However, this statistic removes the effect of any bias in the interpolated data. Problems with correctly capturing the variance will not be highlighted as the measure normalizes the observed and modeled values by their standard deviations. The statistic is standardized. However, because of its insensitivity to biases and errors in variance, the correlation coefficient should be considered as a measure of potential skill [Murphy and Epstein, 1989; Wilks, 2006].

Chi-Square (χ^2) and Chi-Square score (or weight) and its relative probability distribution space … Within the probability distribution space, bounded from -1 to 1, a p value greater than 0.5 means model H_0 has more chances to be better fitting than alternative model H_1.

Reduced Chi-Square (χ^2_ν) and Reduced Chi-Square score (or weight) and its relative probability distribution space … Within the probability distribution space, bounded from -1 to 1, a p value greater than 0.5 means model H_0 has more chances to be better fitting than alternative model H_1.

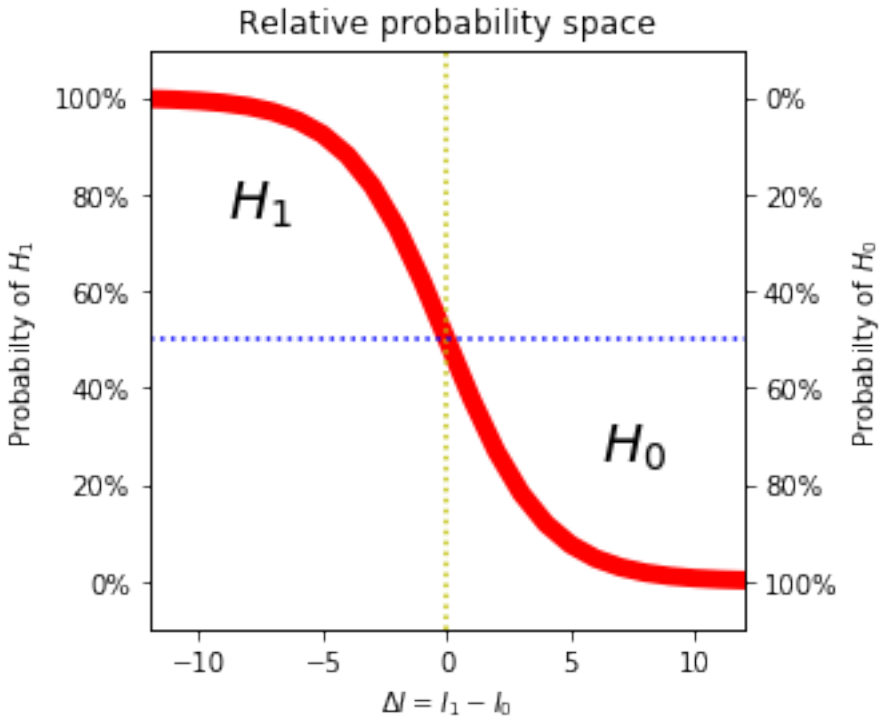
Akaike Information Criterion (**AIC**), its score (or weight) and its relative probability distribution space … Within the probability distribution space, bounded from -1 to 1, a p value greater than 0.5 means model H_0 has more chances to be better fitting than alternative model H_1.

Bayesian Information Criterion (**BIC**), its score (or weight) and its relative probability distribution space … Within the probability distribution space, bounded from -1 to 1, a p value greater than 0.5 means model H_0 has more chances to be better fitting than alternative model H_1.

Relative probability distribution spaces have been used to evaluate information criteria differences.

$$p = \frac{e^{-\frac{1}{2}(I_1 - I_0)}}{1 + e^{-\frac{1}{2}(I_1 - I_0)}}$$

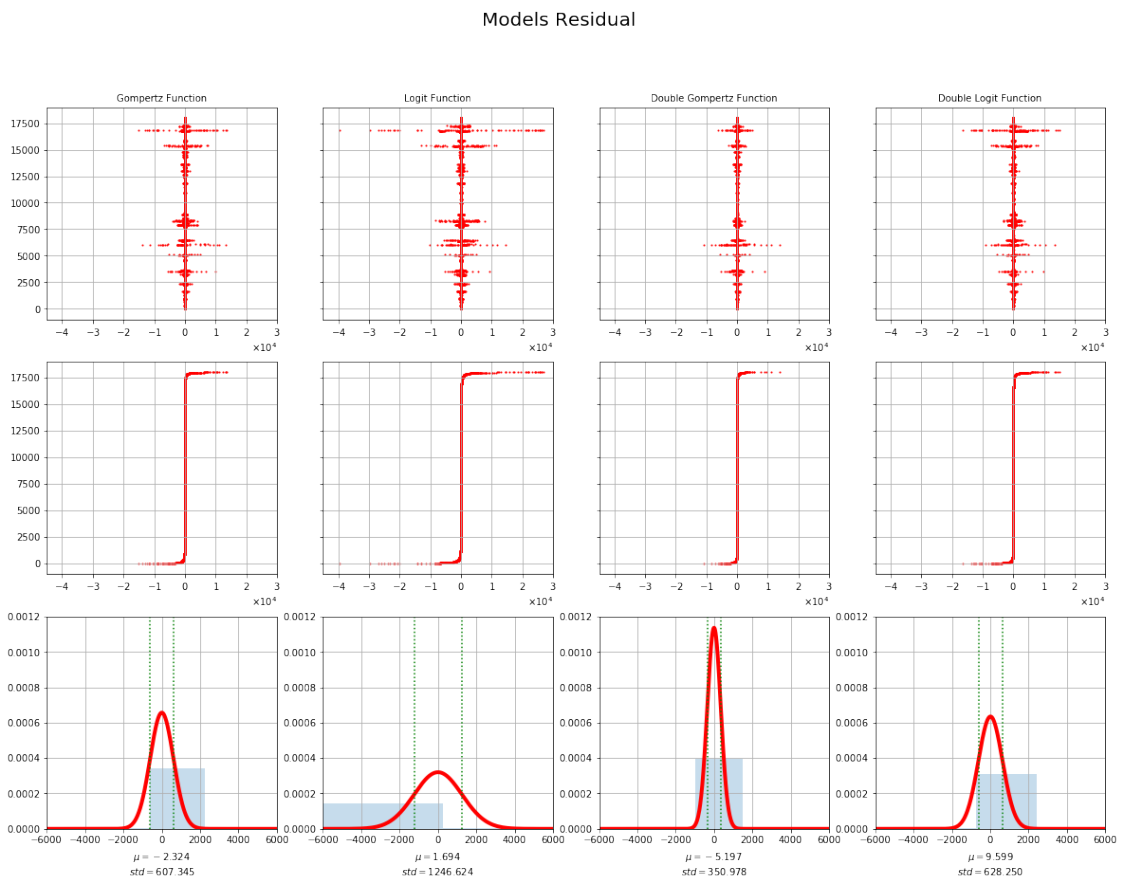
where I_i is information criterion of model i and $I_0 < I_1$.



Information Criteria difference relative probabilty densities have been computed and plotted to summarize H_0 model test against the others.

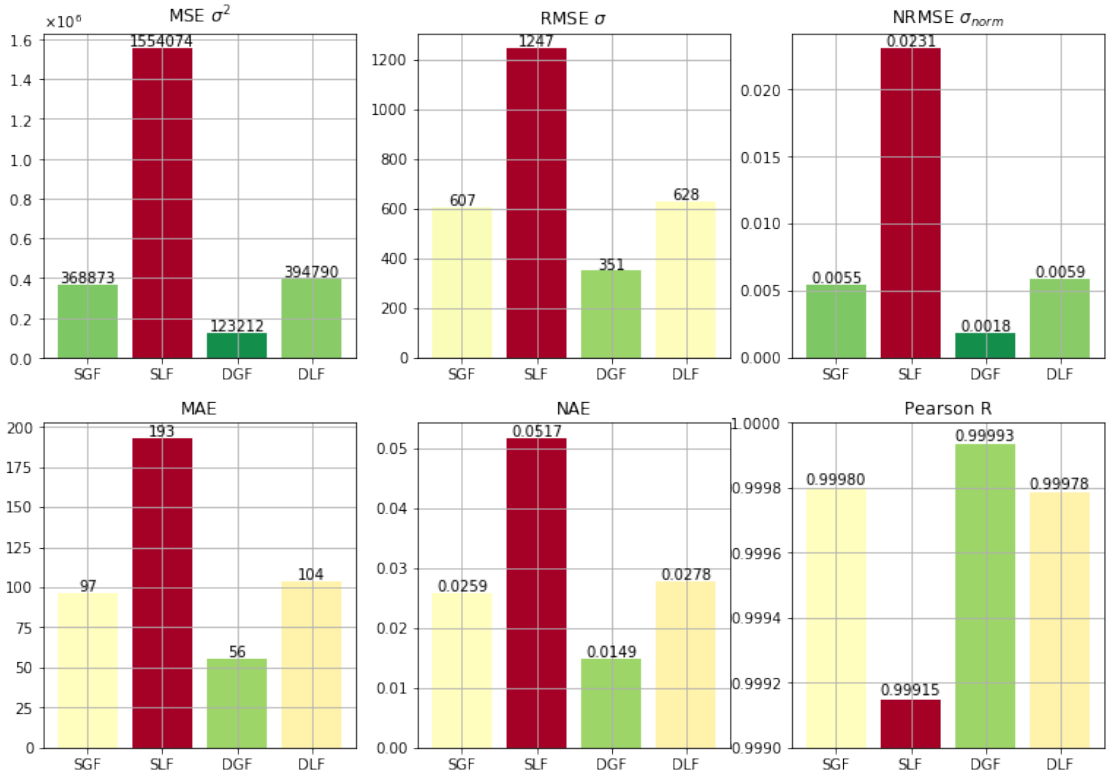
4.1 Residual

Total residual from each model have been collected and compared to get a first "rough" evidence of the most likely better fitting model [FIG].



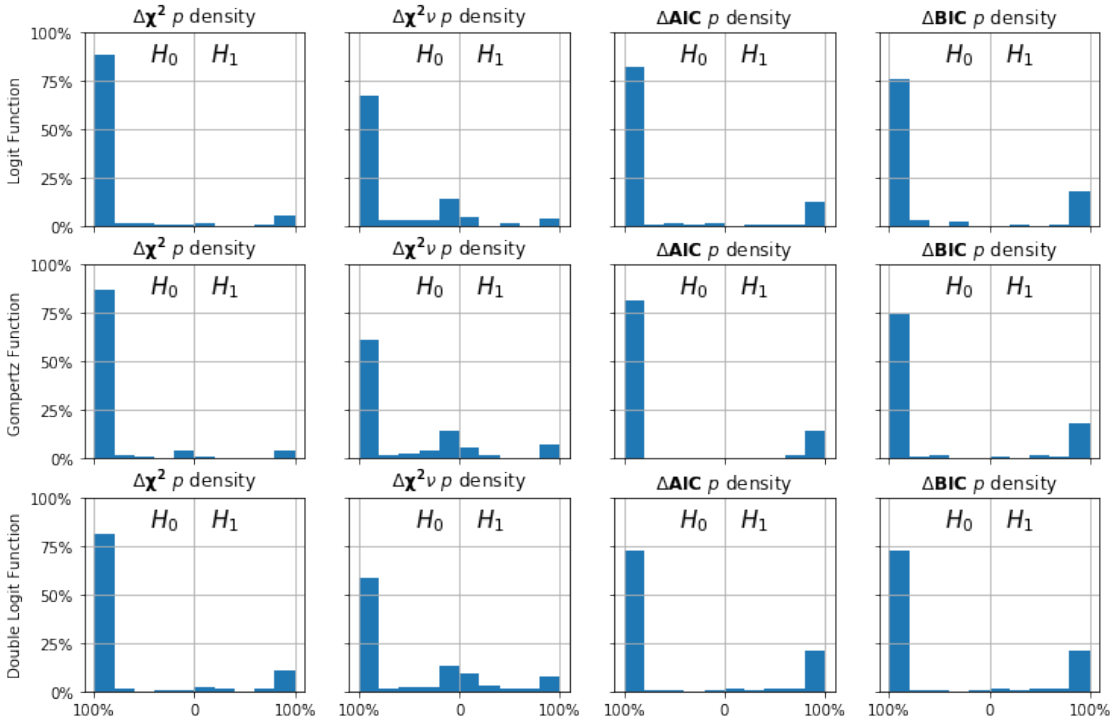
SLF showed a residual mean $\mu = 1.694$ closer to 0 but the wider standard deviation $\sigma = 1247$. **DGF** showed the lower residual standard deviation $\sigma = 351$ and a mean $\mu = -5.197$.

Mean Squared Error (**MSE**, Variance), Root Mean Squared Error (**RMSE**, Standard Deviation), Normalize Root Mean Squared Error (**NRMSE**, Normalized Standard Deviation), Mean Absolute Error (**MAE**), Normalized Absolute Error (**NAE**) and Pearson Correlation (**Pearson R**) have been calculated for all models residual [FIG] (see Appendix for formulas [REF]). **DGF** showed the best results for all indices confirming the first hypothesis that could have been the best fitting model among the chosen ones.



Since compared models are nested (**SLF** and **SGF** have 4 parameters; **DLF** and **DGF** have 7 parameters) Chi-Squared (χ^2), Reduced Chi Square ($\chi^2\nu$), Akaike Information Criterion (**AIC**) scores and Bayesian Information Criterion (**BIC**) scores have been used instead of classical H_0 null hypothesis. χ^2s, $\chi^2\nu$, **AIC**s and **BIC**s have been collected from all fits and compared with each other in relative probability distribution space, country by country, using values returned by `lmfit.minimize` [REF].

Information Criteria Δs have been evaluated in relative probabilty space and gathered in density plots (10 bins) [FIG].



χ^2 scores

	DGF	*F
SLF $p > .5$	170 countries	14 countries
SLF μ	0.98938	0.88512
SLF σ	0.05097	0.19146
SGF $p > .5$	174 countries	10 countries
SGF μ	0.96891	0.90397
SGF σ	0.11040	0.18984
DLF $p > .5$	154 countries	30 countries
DLF μ	0.99247	0.88473
DLF σ	0.04575	0.17539

χ^2_ν scores

	DGF	*F
SLF $p > .5$	165 countries	19 countries
SLF μ	0.90171	0.72851
SLF σ	0.17621	0.21372
SGF $p > .5$	155 countries	29 countries
SGF μ	0.88883	0.78190
SGF σ	0.18519	0.22068
DLF $p > .5$	143 countries	41 countries
DLF μ	0.89924	0.72938
DLF σ	0.18000	0.21070

AIC scores

	DGF	*F
SLF $p > .5$	158 countries	26 countries
SLF μ	0.98515	0.96308
SLF σ	0.06753	0.08659
SGF $p > .5$	153 countries	31 countries
SGF μ	0.99102	0.94953
SGF σ	0.04908	0.06415
DLF $p > .5$	137 countries	47 countries
DLF μ	0.99165	0.94011
DLF σ	0.04377	0.13275

BIC scores

	DGF	*F
SLF $p > .5$	149 countries	35 countries
SLF μ	0.98490	0.97480
SLF σ	0.06340	0.06107
SGF $p > .5$	143 countries	41 countries
SGF μ	0.99003	0.94497
SGF σ	0.04568	0.11687
DLF $p > .5$	137 countries	47 countries
DLF μ	0.99165	0.94011
DLF σ	0.04377	0.13275

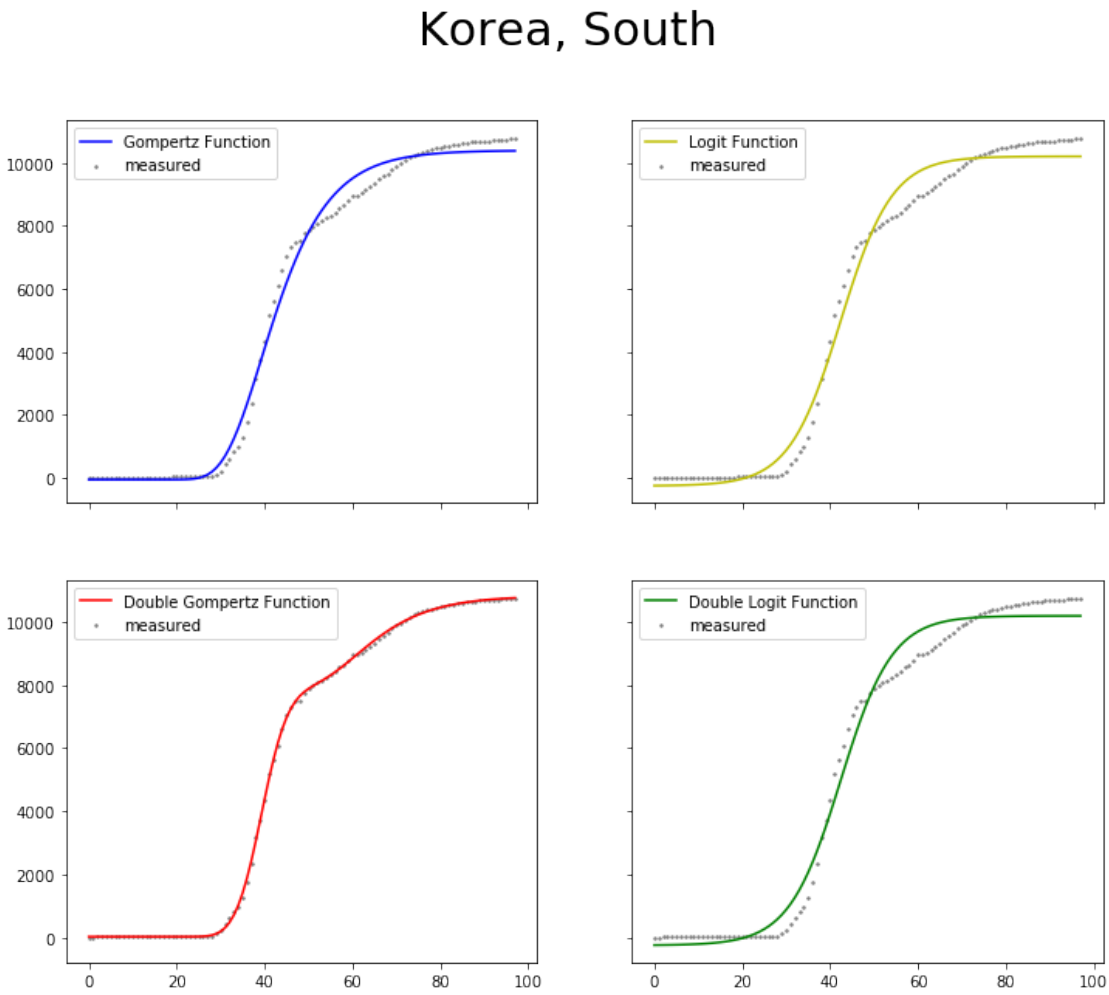
Calcuting mean μ and standard deviation σ for χ^2, χ^2_ν, **AIC** and **BIC** scores [ADD DIFFERENCE] strongly confirmed *Double Gompertz Function* as the better fitting model for SARS-nCoV-2 cumulative cases curve fitting. Results also showed that **DGF** is not only much more fitting than models with less parameters (**SLF** and **SGF**) as expected but also compared to *Double Logistic Function* with the same degrees of freedom.

5 Conclusions

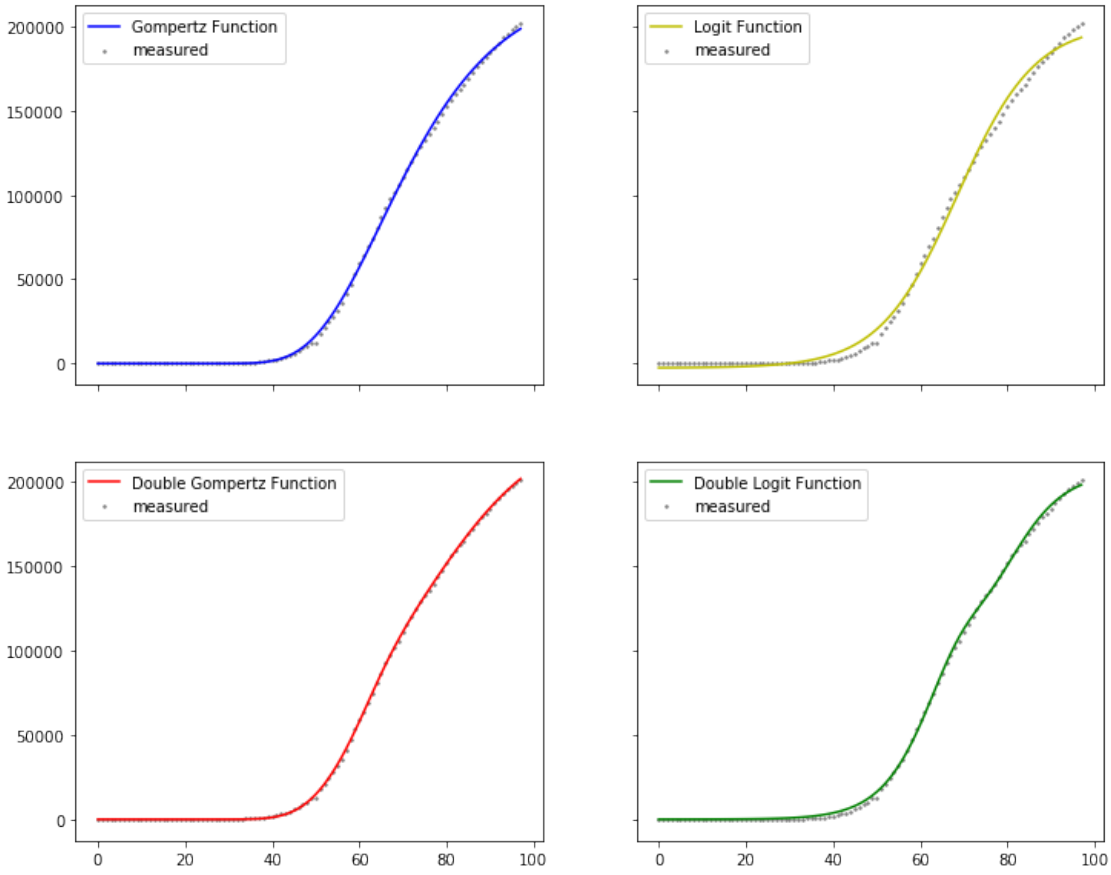
Among the compared models *Double Gompertz Function* has showed the best results and scores fitting data of SARS-nCoV-2 cumulative cases, suggesting that this model should be studied more deeply (possibly improved) and compared to other existing models for further analysis, including forecasting capabilities.

6 Plots

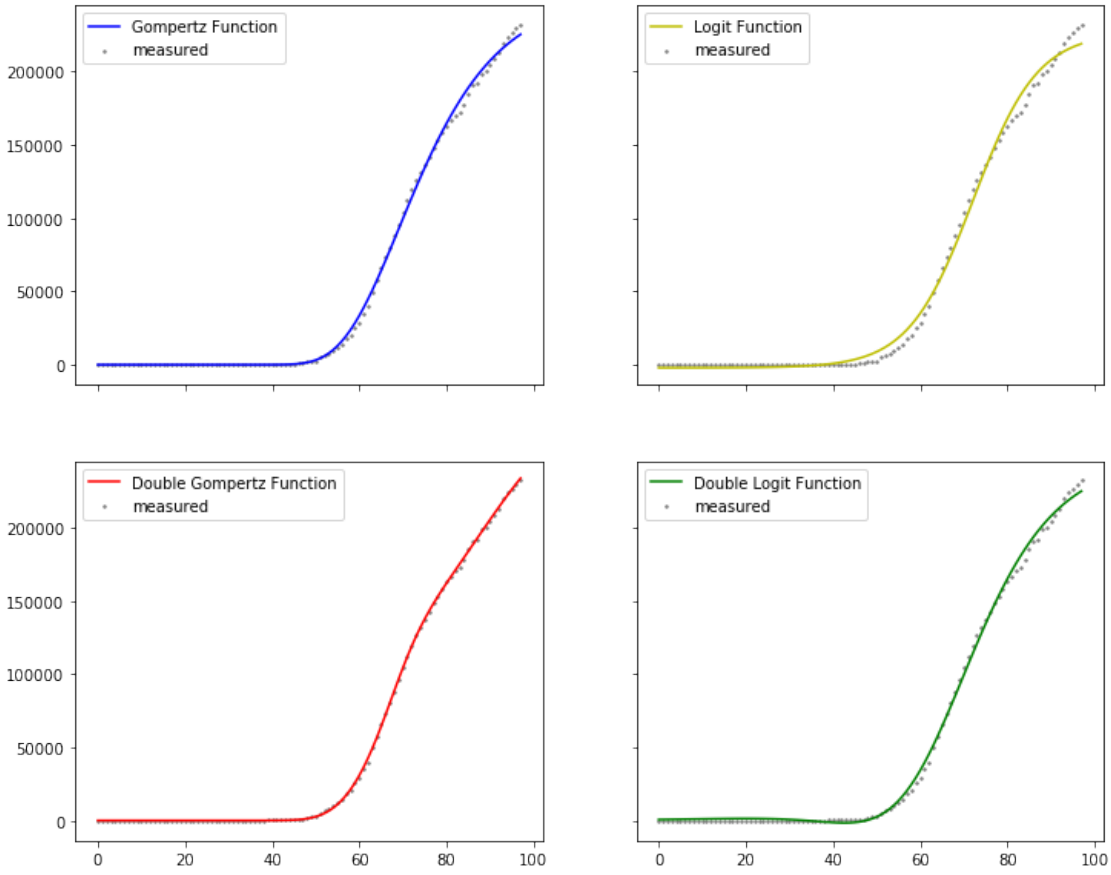
6.1 Fit examples



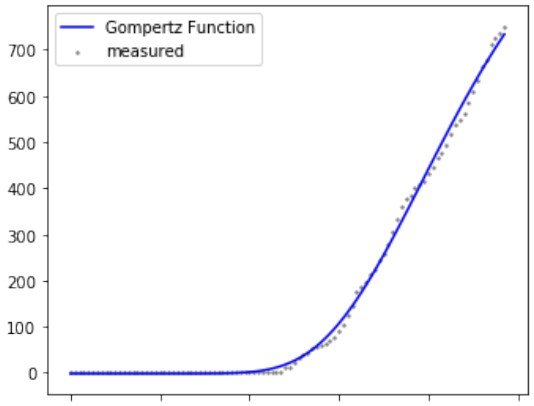
Italy

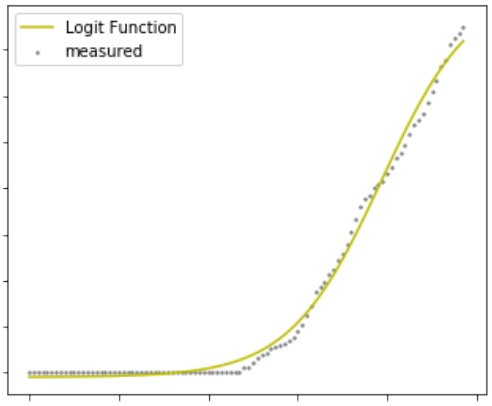

Gompertz Function
measured
Logit Function
measured
Double Gompertz Function
measured
Double Logit Function
measured
200000
150000
100000
50000
0
0
20
40
60
80
100

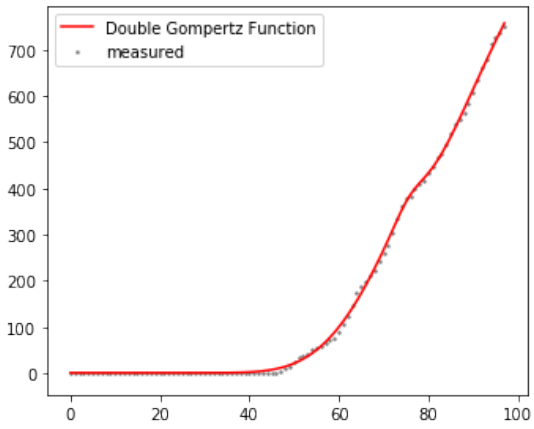
Spain

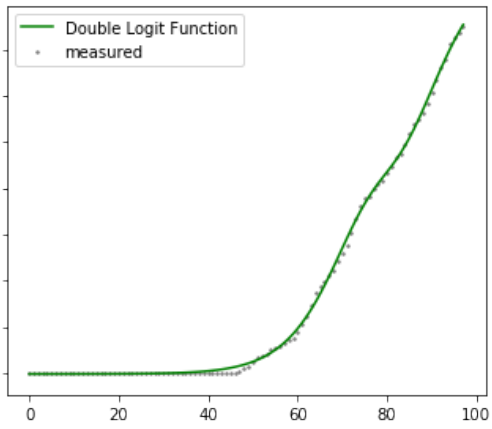

Gompertz Function
measured
Logit Function
measured
Double Gompertz Function
measured
Double Logit Function
measured
200000
150000
100000
50000
0
0
20
40
60
80
100

Albania

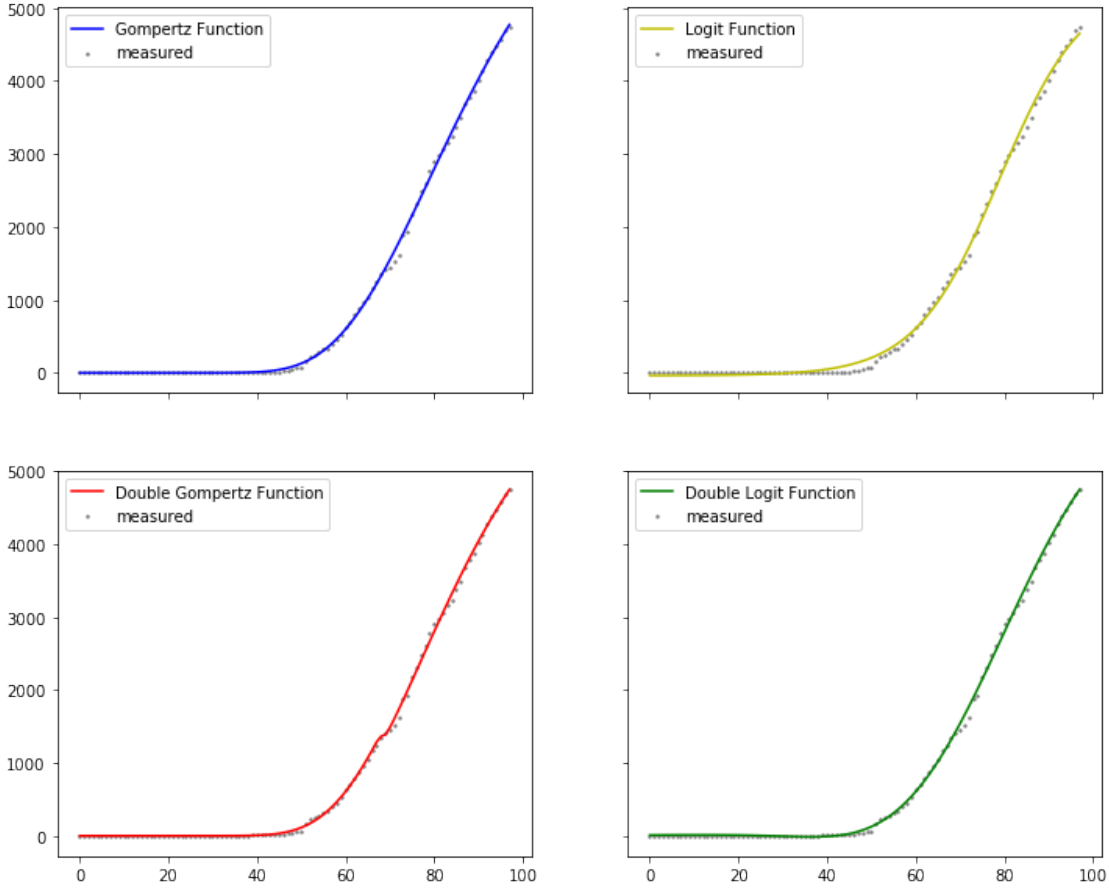

Gompertz Function
measured
700
600
500
400
300
200
100
0

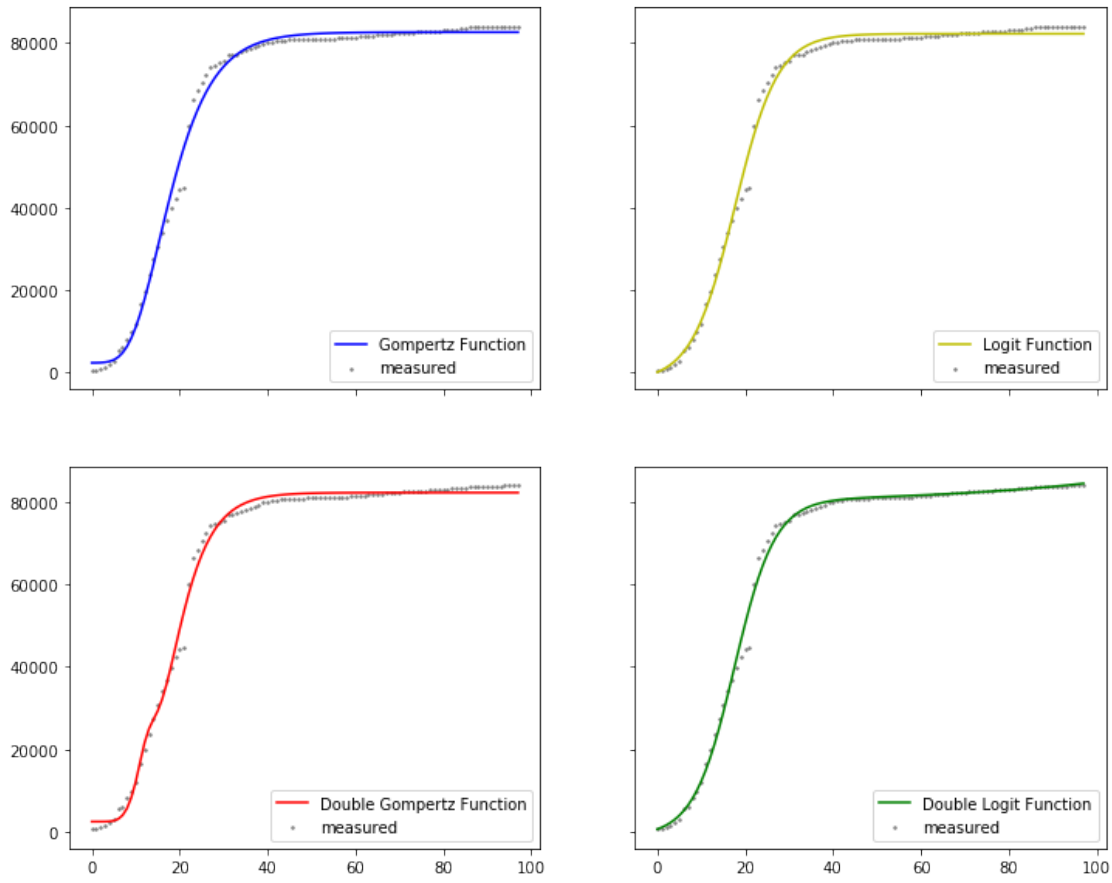

Logit Function
measured


Double Gompertz Function
measured
700
600
500
400
300
200
100
0
0
20
40
60
80
100


Double Logit Function
measured
0
20
40
60
80
100

Finland


Gompertz Function
measured
Logit Function
measured
Double Gompertz Function
measured
Double Logit Function
measured
5000
4000
3000
2000
1000
0
0
20
40
60
80
100



---

7 Appendix

7.1 Statistics formulas

In all formulas we assume: $\hat{y}$ as the measured (true) value and y as the predicted value; n is the number of values and n_{var} the number of variable parameters of the model.

- **MAE**: Mean absolute error (aka Mean Deviation)

$$\mathbf{MAE} = \frac{\sum |y - \hat{y}|}{n}$$

- **NAE**: Normalized absolute error (aka Normalized Mean Deviation)

$$\mathbf{NAE} = \frac{\sum |y - \hat{y}|}{\sum \hat{y}}$$

- **MSE**: Mean Squared Error (aka Variance)

$$\textbf{MSE} = \sigma^2 = \frac{1}{n}\sum (y - \hat{y})^2$$

- **RMSE**: Root Mean Squared Error (aka Standard Deviation)

$$\textbf{RMSE} = \sigma = \sqrt{\frac{1}{n}\sum (y - \hat{y})^2}$$

- **NRMSE**: Normalizer Root Mean Squared Error (aka Normalized Standard Deviation)

$$\textbf{NRMSE} = \sigma_\nu = \frac{\sqrt{\frac{1}{n}\sum (y - \hat{y})^2}}{\sum \hat{y}}$$

- **Pearson R**: Pearson Correlation Coefficient

$$\textbf{R} = \frac{\sum y\hat{y} - \frac{1}{n}\sum y \sum \hat{y}}{\sqrt{\sum y^2 - \frac{1}{n}\left(\sum y\right)^2}\sqrt{\sum \hat{y}^2 - \frac{1}{n}\left(\sum \hat{y}\right)^2}}$$

- χ^2: Chi-Square

$$\chi^2 = \sum (y - \hat{y})^2$$

- χ^2 **p**: Chi-Square score (or weight) in χ^2 relative probability distribution space:

$$\chi^{\mathbf{2}}_{\mathbf{p}} = \frac{e^{-0.5\cdot(\chi^{\mathbf{2}}_{\mathbf{1}} - \chi^{\mathbf{2}}_{\mathbf{0}})}}{1 + e^{-0.5\cdot(\chi^{\mathbf{2}}_{\mathbf{1}} - \chi^{\mathbf{2}}_{\mathbf{0}})}}$$

$$\chi^{\mathbf{2}}_{\mathbf{0}} \leq \chi^{\mathbf{2}}_{\mathbf{1}}$$

- χ^2_ν: Reduced Chi-Square

$$\chi^2 = \frac{\chi^2}{n - n_\nu}$$

- χ^2_ν **p**: Chi-Square score (or weight) in χ^2 relative probability distribution space:

$$\chi^{\mathbf{2}}_{\nu\mathbf{p}} = \frac{e^{-0.5\cdot(\chi^{\mathbf{2}}_{\nu\mathbf{1}} - \chi^{\mathbf{2}}_{\nu\mathbf{0}})}}{1 + e^{-0.5\cdot(\chi^{\mathbf{2}}_{\nu\mathbf{1}} - \chi^{\mathbf{2}}_{\nu\mathbf{0}})}}$$

$$\chi^{\mathbf{2}}_{\nu\mathbf{0}} \leq \chi^{\mathbf{2}}_{\nu\mathbf{1}}$$

- **AIC**: Aikake Information Criterion

$$\mathbf{AIC} = n \log\left(\frac{\chi^2}{n}\right) + 2n_{var}$$

- **AIC p**: Aikake Information Criterion score (or weight) in **AIC** relative probability distribution space:

$$\mathbf{AIC_p} = \frac{e^{-0.5\cdot(\mathbf{AIC_1}-\mathbf{AIC_0})}}{1+e^{-0.5\cdot(\mathbf{AIC_1}-\mathbf{AIC_0})}}$$

$$\mathbf{AIC_0} \leq \mathbf{AIC_1}$$

- **BIC**: Bayesian Information Criterion

$$\mathbf{BIC} = n \log\left(\frac{\chi^2}{n}\right) + \log(n) n_{var}$$

- **BIC p**: Bayesian Information Criterion score (or weight) in **AIC** relative probability distribution space:

$$\mathbf{BIC_p} = \frac{e^{-0.5\cdot(\mathbf{BIC_1}-\mathbf{BIC_0})}}{1+e^{-0.5\cdot(\mathbf{BIC_1}-\mathbf{BIC_0})}}$$

$$\mathbf{BIC_0} \leq \mathbf{BIC_1}$$