COVID-19 FORECAST: COMPARISON AND COMBINATION OF AUTOREGRESSIVE AND COMPARTMENTAL MODELS

Fabiano A.S. Ferrari[1,2], Evelly C.J. Silva[3], Marineide A. Rocha[3], Rogério A. Santana [1]

[1]*Instituto de Engenharia, Ciência e Tecnologia, Universidade Federal dos Vales do Jequitinhonha e Mucuri*
Avenida Um, 4050, Cidade Universitária, Janaúba, 39447-814, MG, Brasil
fabiano.ferrari@ufvjm.edu.br
[2]*Programa de Pós-Graduação em Modelagem Matemática e Sistemas, Universidade Estadual de MOntes Claros*
Av. Prof. Rui Braga, Vila Mauriceia, Montes Claros, 39401-089, MG, Brasil
[3]*Campus Janaúba, Universidade Federal dos Vales do Jequitinhonha e Mucuri*
Av. Brasil, 334, Centro, Janaúba, 39442-010, MG, Brasil

Abstract. The Imperial College was responsible for the firsts forecasts for covid-19 propagation, in the early stages of the pandemic. Their study had a positive effect in the British Lockdown and it may have saved many lives. Besides the Imperial College study, many other studies were proposed based on different approaches. Some models have projected much more deaths than we have observed, while others have projected much less. There are plenty reasons to explain why the model can fail, it can be caused by bad data, virus mutation, health care system collapse, and others. One strategy to reduce the error in the propagation forecast is to combine different approaches. In this work, we are going to present error analysis for the forecast of covid-19 propagation based on autoregressive models (such as ARIMA and SARIMA models) and compartimental models (such as SIR model and variations) and how the forecast can be improved by the combination of these two approaches.

Keywords: Covid-19, SARIMA, SIR, covsirphy.

1 Introduction

Since the emergence of covid-19 many models for the pandemic forecast have been proposed. In the fist month of the pandemic, one Imperial College's study predicted thousand hundreds deaths could happen, if nothing would made to decrease the epidemic effectiveness[1]. Nowadays, we have seen all sort of models, based on the most varied approaches. Among them there are models based on compartimental models, like SIR (Susceptible - Infected - Recovered) and variations [2]. That consists of a set of differential equations to describe the epidemic evolution. Another approach that can be used for covid-19 forecast is autoregressive models, such as ARIMA and SARIMA [3]. Autoregressive models are representations to describe future events of random variables based on time varying process, including for example, the effect of the moving average.

In this work, we are going to explore some features related to covid-19 forecast, using autoregressive models and compartimental models. In Section 2, we present the forecast models that we used. In Section 3, we present our data set of analysis, that is based on the number of infected people in the first year of the pandemic in Brazil. In Section 4 we present our results, where we evaluate the mean absolute percentage error (MAPE) for different situations. In Section 5, we present our final considerations.

2 Methodology

The covid-19 forecast simulation were made using SARIMA (Subsection 2.1) and SIRF (Subsection 2.2). And to evaluate the forecast error we used the mean absolute percentage error (MAPE) (Subsection 2.3). The simulations performed in this manuscript were made using codes developed in R and Python.

2.1 SARIMA

Most of the time series are nonstationary, i.e., the average is not constant through the process. Most of the models are develeped for stationary time series, we can work with a nonstationary time series by differentiating it up until the point the derivative becomes stationary [4, 5]. One example is the ARIMA model:

$$\phi(B)\Delta^d Z_t = \theta(B)\epsilon_t \qquad t = 1, 2, \ldots, n \tag{1}$$

where B is $B^j Z_t = Z_{t-j}$; $\phi(B) = (1 - \phi_1 B_1 - \ldots - \phi_p B^p)$ is an autoregressive polynomial of order p, $\theta(B) = (1 - \theta_1 B_1 - \ldots - \theta_p B^p)$ is an moving average polynomial of order q, $\Delta^d = (1 - B)^d$ indicates the number of derivatives for the time series to become stationary [4, 5].

If the time series present seasonality, then it is necessary to make some modifications in the original ARIMA model. Consider a seasonal time series $Z_(t)$ that is observed by a period of time s, the SARIMA$(p, d, q)(P, D, Q)_s$ model is going to be describe by

$$\phi(B)\Phi(B^s)\Delta^d\Delta_s^D z_t = \theta(B)\Theta(B^s)\epsilon_t, \qquad t = 1, 2, \ldots, n \tag{2}$$

where B is $B^j Z_t = Z_{t-j}$; $\phi(B) = (1 - \phi_1 B_1 - \ldots - \phi_p B^p)$ is an autoregressive polynomial of order p; $\Phi(B^s) = (1 - \phi_1 B^s - \ldots - \phi_p B^{Ps})$ is a seasonal autoregressive polynomial of order P; $\Delta^d = (1 - B)^d$ is the difference operator and d indicates the number of derivatis according with the average; $\Delta_s^D = (1 - B^s)^D$ is the seasonal difference operator and D indicates the number of derivatis to remove the seasonality; $\theta(B) = (1 - \theta_1 B_1 - \ldots - \theta_p B^p)$ is an moving average polynomial of order q; $\Theta(B^s) = (1 - \theta_1 B^s - \ldots - \theta_p B^{Qs})$ is an moving average polynomial of order q, and ϵ_t is a white noite with zero mean and variance $\sigma^2_{\epsilon_t}$.

SARIMA simulations were performed using the standard 'Statsmodel' Python's package. The parameters were obtained using the Maximum Likelihood Optimization.

2.2 SIRF

The SIRF (Susceptible - Infected - Recovered - Fatal) model is very similar to the SIR (Susceptible - Infected - Recovered) model, with the addition of the number of deaths term. In this case, the set of differential equations is going to be described as

$$\frac{dS}{dt} = -\frac{\beta SI}{N}, \tag{3}$$

$$\frac{dI}{dt} = (1 - \alpha 1)\frac{\beta SI}{N} - (\gamma + \alpha_2)I, \tag{4}$$

$$\frac{dR}{dt} = \gamma I, \tag{5}$$

$$\frac{dF}{dt} = \alpha_1 \frac{\beta SI}{N} + \alpha_2 I, \tag{6}$$

where β is the effective contact rate, γ recovered rate, α_1 natural death rate and α_2 death by the disease.

SIRF simulations were developed using the package "Covsirphy", version 2.19.0 [6]. The parameters were estimated using the Maximum Likelihood Optimization.

2.3 MAPE

To evaluate our forecast error we used Mean Absolute Percentage Error (MAPE) [7]. It can be described as

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|, \tag{7}$$

where n is the time series size, A_t is the actual value and F_t is the forecast value.

3 Data Set

To analyze our methodology we are going to analyze the number of infected people in Brazil from February 2020 up to February 2021. We divide the time series in three parts:

Serie 1: From February 25th up to July 30th.
Serie 2: From July 31st up to October 10th.
Serie 3: From October 11st up to February 10th.

We separate the time series to analyze the effect of the error in different times and at different tendencies.

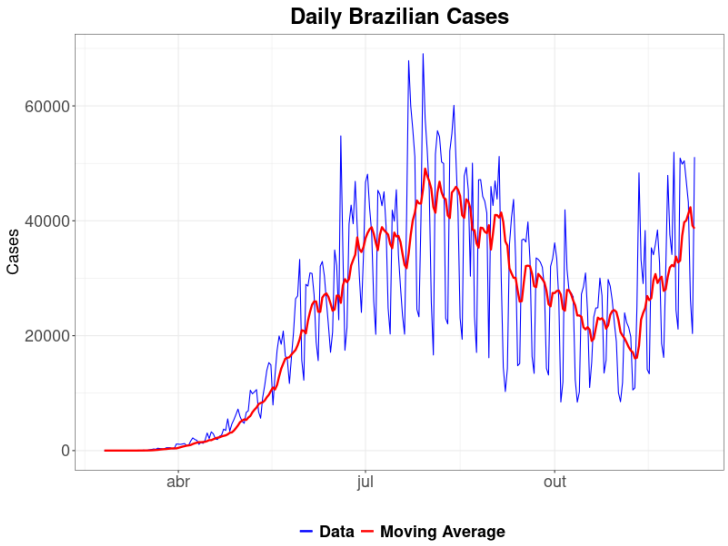

Figure 1. Daily cases of covid19 in Brazil from February 2020 up to February 2021, according to the Ministery of Health (*Ministério da Saúde*).

4 Analysis

In this section we present our analysis based on three approaches. First, we analyze the error provided by only applying the SIRF method. Then, we analyze the error provided by only applying SARIMA. In the last subsection, we provide the effect of the combination.

4.1 Forecast using Optimized SIRF

The use of SIRF model with optimized parameters provide good results when the training set is small. As shown in Figure 2, as we increase the training set the MAPE error increases.

4.2 Forecast using Auto SARIMA

The forecast results using SARIMA are the inverse of what we observe for SIRF. As shown in Figure 3, as the training set is increased the MAPE error decreases. Due to its statistics properties, SARIMA method requires a large data set to provide good results.

4.3 Forecast combining SIR and SARIMA

Due to the fact that SIRF works well for small training sets and SARIMA works well for large training sets, the combination of SIRF and SARIMA methods can optimize the forecast. We presented some strategies in Table 1.

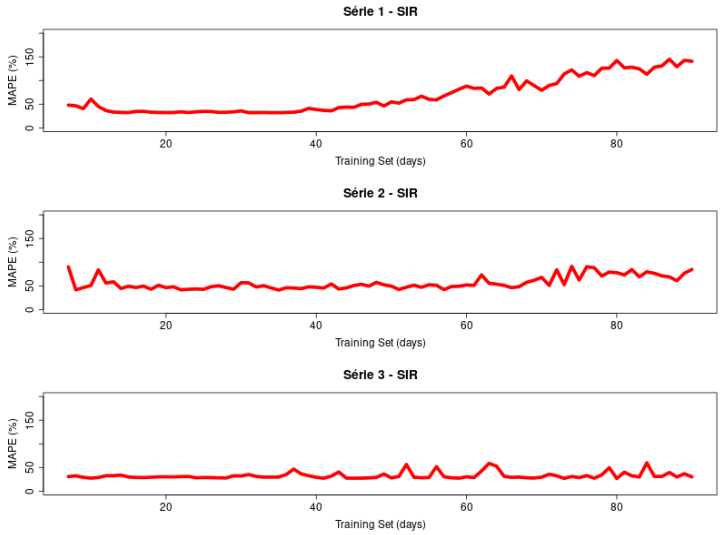

Figure 2. MAPE Error for 30 days prediction using SIRF with optimized parameters and varying the size of the training set.

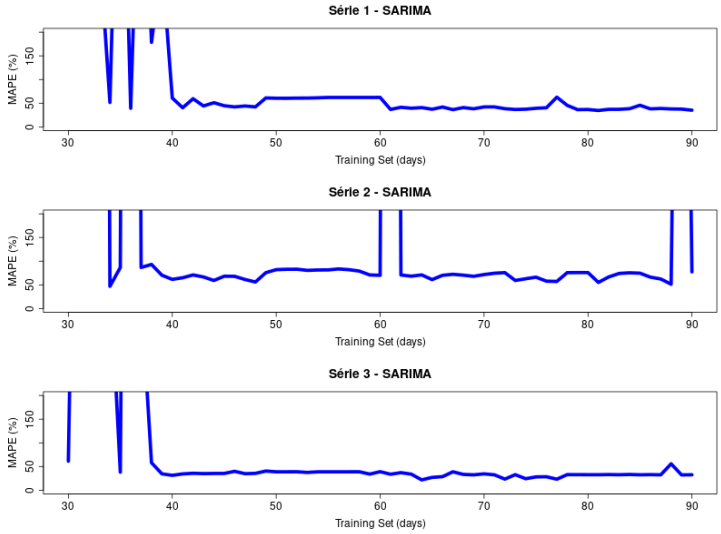

Figure 3. MAPE Error for 30 days prediction using autoSARIMA and varying the size of the training set.

Table 1. Some protocols for a combined SARIMA+SIR Covid-19 forecast

SARIMA	SIR	MAPE (%)
75 days training set + 30 days forecast	+ 30 days forecast	46
75 days training set + 45 days forecast	+ 30 days forecast	73
90 days training set + 30 days forecast	+ 30 days forecast	42
90 days training set + 45 days forecast	+ 30 days forecast	74

5 Conclusions

SARIMA model provides good forecast when we have a large training set, however, the Maximum likelihood optimization fail if we use a small data set. On the other hand, the SIRF model provides good forecast if we use a small training set, but the error increases fast with the size of the training set. The forecast efficacy can be increased by combining both methods. We show that this strategy can be efficient to increase the forecast efficacy. The idea proposed here is an ongoing research, and our results are presented as a concept.

Acknowledgements. We thanks all colleagues that helped us in the discussions about the methodologies for covid19 forecast. We thanks UFVJM for partial financial support.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] A. T. Biggs and L. F. Littlejohn. Revisiting the initial covid-19 pandemic projections. *The Lancet Microbe*, vol. 2, n. 3, pp. e91–e92, 2021.
[2] Y.-C. Chen, P.-E. Lu, C.-S. Chang, and T.-H. Liu. A time-dependent sir model for covid-19 with undetectable infected persons. *IEEE Transactions on Network Science and Engineering*, vol. 7, n. 4, pp. 3279–3294, 2020.
[3] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, and M. Ciccozzi. Application of the arima model on the covid-2019 epidemic dataset. *Data in brief*, vol. 29, pp. 105340, 2020.
[4] P. A. Morettin and C. Toloi. *Análise de séries temporais: modelos lineares univariados*. Editora Blucher, 2018.
[5] R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.
[6] H. Takaya. *Kaggle Notebook, COVID-19 data with SIR model*. , 2020.
[7] I. Svetunkov and F. Petropoulos. Old dog, new tricks: a modelling view of simple moving averages. *International Journal of Production Research*, vol. 56, n. 18, pp. 6034–6047, 2018.