Forecasting Time Series with Non Parametric Regression approach

Nicolás Rengifo Campo
nrengifoc@eafit.edu.co

Henry Laniado
hlaniado@eafit.edu.co

January 30, 2021

Abstract

In this research we propose a non-parametric forecasting method. In order to show its versatility different time series are considered, epidemic and economic ones. In most of the cases the model proposed shows a better performance than the arima family of models. Great results are guaranteed in one period forecast, for longer periods a good performance is also reached. The methodology employed by the model allows a more real analogy with the past of the series, since the most probable value depends of the whole past of the serie unlike the auto regressive processes. Our contribution is based on working with the increment variable instead of the raw one.

1 Introduction

The time series analysis provides models that pretend to represent how a variable will evolve in the future. As exposed in [Brooks, 2019], auto-regressive AR(p) models, auto-regressive integrated moving average ARIMA(n,p,q) are one of the most used models. These models are highly used for academical purposes, but since they have strong assumptions in a practical context, they might not give the ideal results. The forecasting and prediction problem have a great importance in economic area, nonetheless, since we are analyzing from the science perspective it should be understood that the mathematical model behind the estimations is not limited to a single problem.

In the last years the non-parametric statistics have been used for a lot of research purposes in many different areas, its objective of having as less as possible assumptions make them really useful in practice. In order to offer better results this will be the core of this research, more specifically the non-parametric regression which as said before can be used in many different areas.

Coronavirus disease is being a phenomenon that is changing the way we live. Many countries had past the epidemic peak, and others are still waiting for it. In an uncertainty phenomenon having scientific support studies is the most important information source in order to reduce the negative consequences that the pandemic can bring with it. It is time to offer different tools that help to foresee most of the troubles that COVID 19 causes in areas like the economic, health, social and others. Time series theory will be used in order to estimate interest variables in the pandemic context, in this work reported cases and deaths will be our target.

The COVID 19 pandemic has caused a lot of issues in many different areas, many of them result of the uncertainty of how the problematic will evolve. For example, the health systems around the world has been getting prepared since the crisis began. The problem has been defined as a pandemic, as told in [Cucinotta & Vanelli, 2020]. An area that has been close following the behavior of the virus is the scientific one, it is also needed when a crisis like the one we are living occur, every study area must help since each one special scope.

We will compare the performance of methods from arima family and the one proposed in this research on the COVID-19 evolving problematic. The development of the project will be mainly based in the mathematical theoretical component exposed in [Wasserman, 2006] and in the research shown in [Rodríguez & Siado, 2003]. The interest variables will be modeled as time series.

The mathematical and data research have to show how useful they are in a situation like this, in epidemic context, a well example is [Meltzer *et al.*, 2014], where the Ebola pandemic was modeled. The main idea with this research is to help the decisions that a government can take by offering them the enough scientific evidence of how the pandemic will evolve, otherwise many wrong decisions might be taken as a result of the quickness needed to mitigate the negative effects of the pandemic.

The main idea of this research will be based on the work proposed by [Rodríguez & Siado, 2003], since our core will be the non-parametric statistics another relevant source of information will be the theory and formal statistics concepts exposed in [Wasserman, 2006]. The main reason to use the non-parametric method over the classical ones, is that in stochastic prediction context, it offers better results as was exposed by [Bosq, 2012].

In non-parametric regression the most used model Nadaraya-Watson kernel-type smoother estimation will be the one to be used, [Ferraty & Vieu, 2004] supports the usefulness of this model such as its great performance. A chase where non-parametric regression was used not so much ago, was in the research done by [Kindzerske & Ni, 2007] where a forecasting for a traffic flow was proposed by using this method.

Other implementations for time series analysis will be mainly based on the works shown in [Brooks, 2019] and [Davidson *et al.*, 2004]. The problem of trying to predict the behavior of a variable in COVID 19 context has been proposed based in SEIR models like the work done by [Yang *et al.*, 2020]. Other approach is given from the Q exponential model developed in [Vasconcelos *et al.*, 2020], where deaths were estimated. Also, time series analysis had been used as exposed in [Maleki *et al.*, 2020].

On this research we will try to get advantage of the versatility of the model, and include it for the compliance of our objectives, trying to help with scientific foundation to the global and local public health issues by offering a close idea of how the pandemic will evolve in a near future.
In order to show the versatility of the proposed model, economic variables such as an index and a sale historic are considered in the analysis. We would see that forecast can be done with a low prediction error. In the case of the sale historic, thanks to a goodness of fit in the prediction we can obtain an optimal number of product, similar to the problem worked in [Petruzzi & Dada, 1999].

This paper is structured as follows, in the section 2 we define the mathematical models and the theory that supports our research, this is the arima traditional methodology and the non-parametric model based on regression that allows us to perform forecast. Later, on section 3 we explain the variables that we are considering to analyze and an exploratory to introduce them. Then the section 4 shows the results of our experiments and finally in section **??** we give some conclusions and future work derived from this research.

2 Mathematical Models

2.1 ARIMA Methodology

Based on the time series theory, and most specifically the ARIMA family of models as defined in [Brooks, 2019]. We can define as shown below the most general case $arima(p, d, q)$ and also some graphical approximations that usually help us to choice a proper parameter combination.

2.1.1 ARIMA(p,d,q) Model

Define the Auto Regressive Integrated Moving Average model with order p, d, q represented with lag operator:

$$\phi(L)(1-L)^d X_t = \theta(L)\varepsilon_t \tag{1}$$
$$(1-\phi_1 L - \cdots - \phi_p L^p)(1-L)^d X_t = (1-\theta_1 L - \cdots - \theta_q L^q)\varepsilon_t$$

Where:

- p: order of $AR(p)$ model.

- d: number of differences required to make the time series stationary.
- q: order of $MA(q)$ model.

Them, the stationary process W_t with the d differences apply to X_t, the result is a $ARMA(p, d)$:

$$\phi(L)W_t = \theta(L)\varepsilon_t \tag{2}$$

2.1.2 Auto Correlation Function

The Auto Correlation Function or ACF gives values of auto-correlation of the series, it describe how well the actual value of the data is related with the past values. The plot of the ACF plots the values of the auto-correlations -between the data- and the significance level, to considered statistically significant of the auto-correlation values. The plot sows trends, seasons, cyclic, etc. Where the ACF values are calculated by equation (3)

$$\hat{\rho_j} = \frac{C\hat{O}V(X_t, X_{t-j})}{V\hat{A}R(X_t, X_{t-j})} \tag{3}$$

2.1.3 Partial Auto Correlation Function

The Partial Auto Correlation Function or $PACF$ is similar to ACF (section 2.1.2), but this finds correlation of the residuals with the next lag value. The plot of the $PACF$ plots the values of the partial auto-correlations -between the residuals of the data- and the significance level, to considered statistically significant of the partial auto-correlation values. The plot sows hidden information in the residuals.

2.2 Non Parametric Methodology

2.2.1 Kernel Regression Estimator

In almost every cases considering a non parametric estimator assure as a better performance that using a traditional statistic [Ferraty & Vieu, 2004]. We define a bi-variate non parametric regression Nadaraya Watson regression, as exposed on [Wasserman, 2006].We wish to find the conditional hope $E(y|x)$, then according to the probability density function definition we can proceed as

$$\begin{aligned}
\hat{E}(y|x) &= \hat{r}_n(x) \\
&= \frac{\int y\hat{p}(x,y)dy}{\hat{p}(x)} \\
&= \frac{\frac{1}{nh_n}\sum_{i=1}^{n} y_i K(\frac{x_i - x}{h})}{\frac{1}{nh_n}\sum_{i=1}^{n} K(\frac{x_i - x}{h})} \\
&= \sum_{i=1}^{n} y_i w_i(x)
\end{aligned} \tag{4}$$

$$\text{Where } w_i(x) = \frac{K(\frac{x_i - x}{h})}{\sum_{i=1}^{n} K(\frac{x_i - x}{h})} \text{ y } h = \hat{\sigma}_x n^{\frac{-1}{5}}$$

$K(\cdot)$ is a Kernel function which integral is one, this function is used to control the weights of a data set, depending of its definition it assigns a weight to every observation and so estimate a density [Wasserman, 2006]. We introduce some of the most used Kernels that we will use in this research.

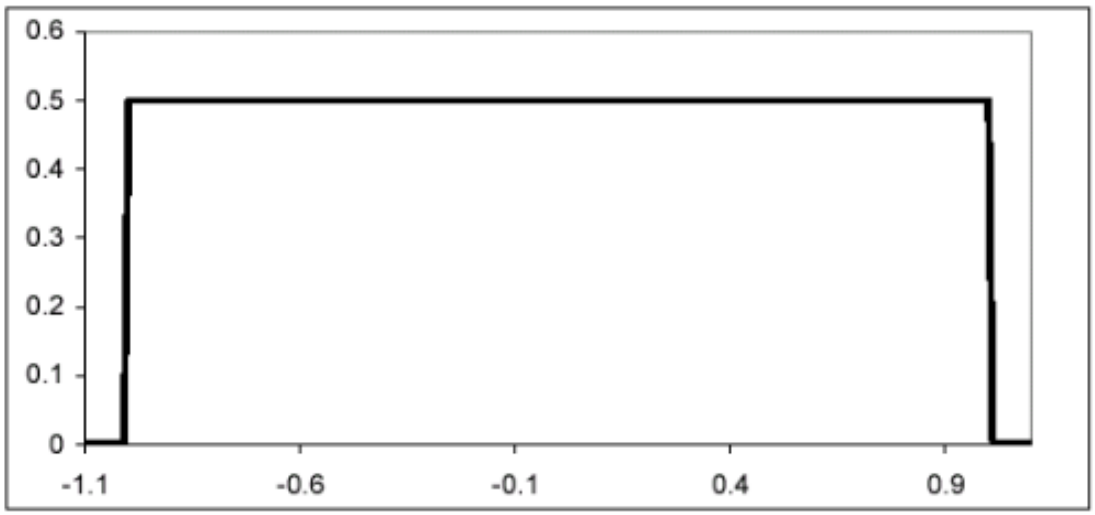

Figure 1: Uniform Kernel. $K(u) = \frac{1}{2} I_{[-1,1]}(u)$

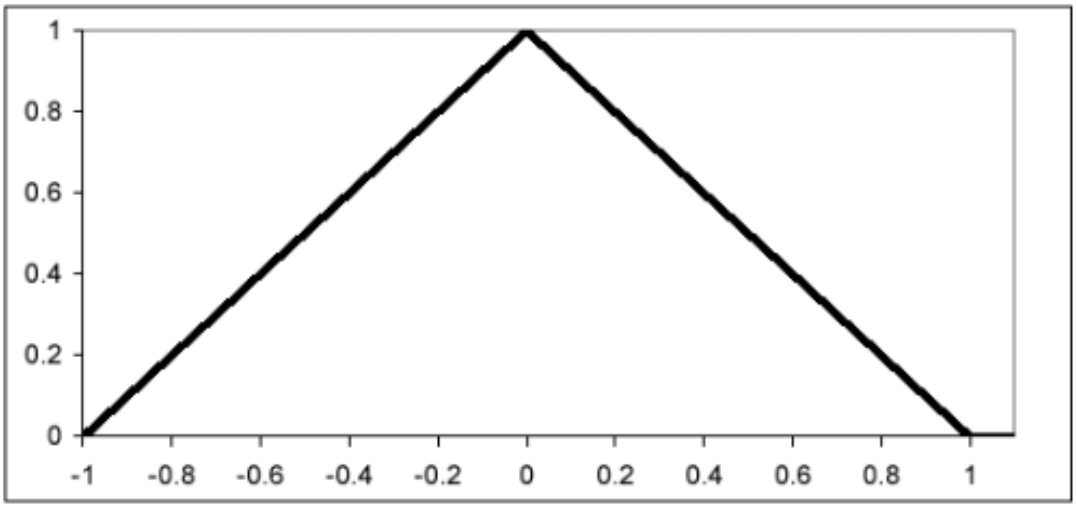

Figure 2: Triangular Kernel. $K(u) = (1 - |u|) I_{[-1,1]}(u)$

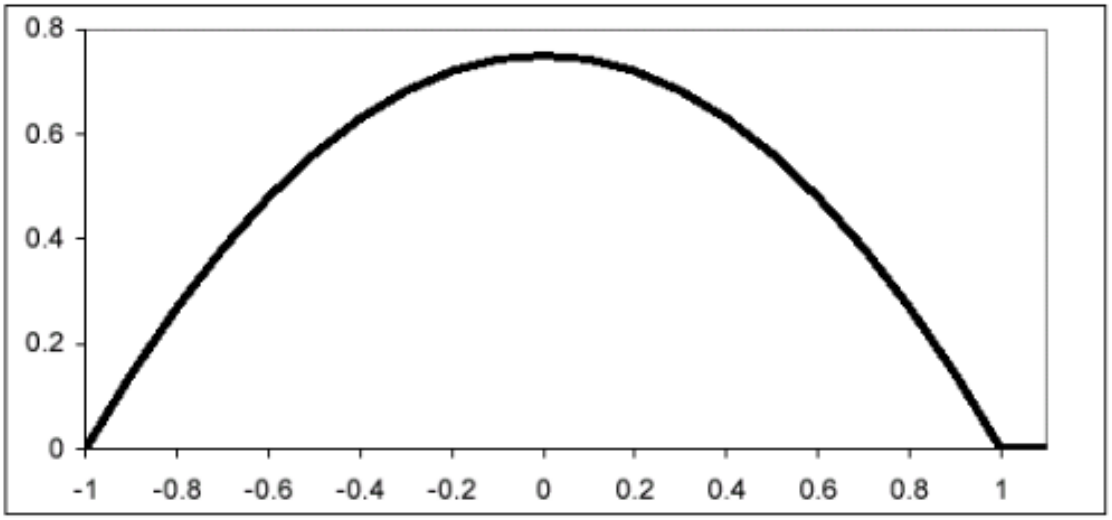

Figure 3: Bi-squared Kernel. $K(u) = \frac{15}{16}(1 - 2u^2 + u^4)I_{[-1,1]}(u)$

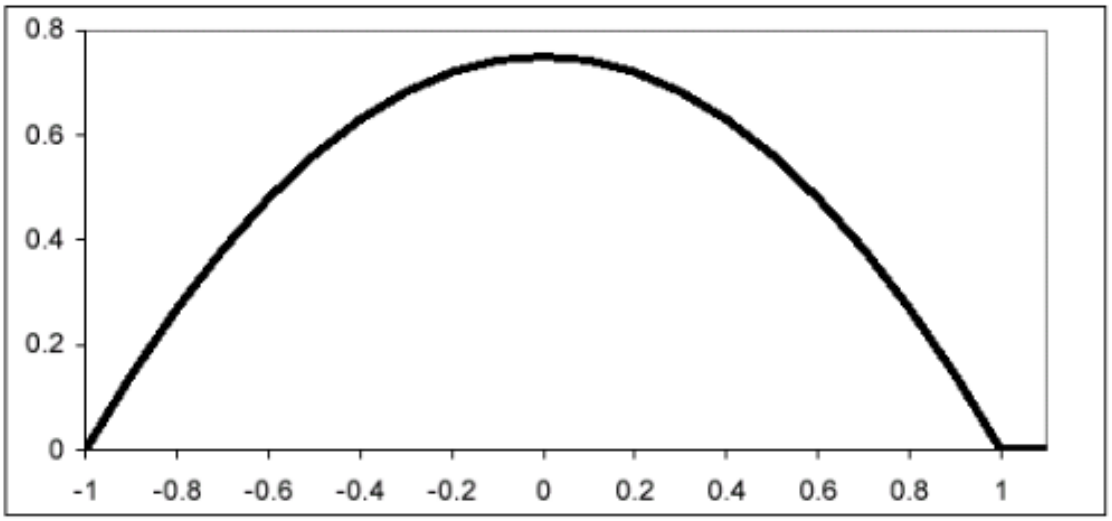

Figure 4: Epanechnikov Kernel. $K(u) = \frac{3}{4}(1 - u^2)I_{[-1,1]}(u)$

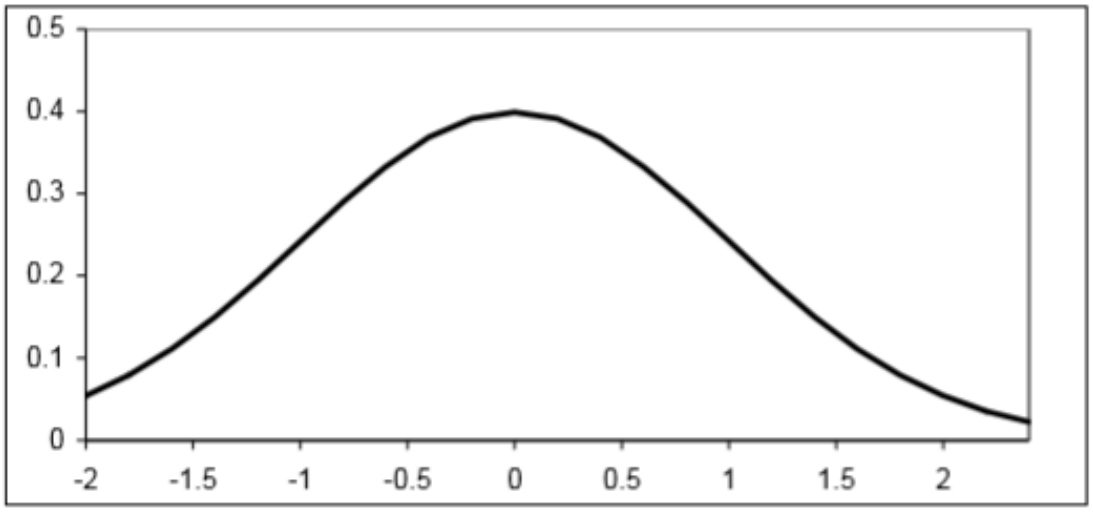

Figure 5: Gaussian Kernel. $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$

2.2.2 Non Parametric Forecast

First we must to clarify our assumptions for the model. The model is supposed to work with both stationary and no stationary time series, in both cases the non parametric estimation will have a good performance [Bosq, 1998]. Another assumption is that the time serie to be worked, follows a d Markov's process, where d is a markov coefficient. The methodology to work is taken from [Rodríguez & Siado, 2003].

Now we are going to define an associated process in order to build our non parametric forecasting method. This process is to be composed by explicating variables X_t and explicated ones Y_t, the first is understood as a window that is associated with a punctual estimation, the explicated variable. The bi-variate regression idea can be used to allow a better comprehension of how the model works.

Suppose that we have an uni-variate time series $\{Z_t\}$, with $1 \leq t \leq n$. We want to predict the $Z_{n+m}, m \geq 1.$, where m is the forecast horizon. Then, the whole process can be formulated as $\{X_t, Y_t\}$, where:

$$X_t = [Z_t, ..., Z_{t-d+1}], \quad Y_t = Z_{t+m}, \quad t \in \{d, ..., n\} \tag{5}$$

Considering the Kernel regression estimator, exposed on Section 2.2.1, based in the data of $\{Z_t\}$, for $E(Y_n|X_n) = E(Z_{n+m}|Z_n, ...Z_{n-d+1})$ we have:

$$\hat{r}_n(x) = \hat{E}(Y_n|X_n) = \sum_{t=d}^{n-m} W_t^m(x) Y(t), \tag{6}$$

$$where, \quad W_t^m(x) = \frac{K_d(\frac{x-X_t}{h_n})}{\sum_{t=d}^{n-m} K_d(\frac{x-X}{h_n})} \quad , x \in \mathbf{R}^d \tag{7}$$

Notice that the bandwidth h_n controls the size of the local neighbourhood it also must me a non negative number that converge to zero when n tends to infinity. $K_d(\cdot)$ is a d-variate kernel function, with multiple integral on one, it controls the shape of the weights. Moreover $x = X_n = Z_n, ...Z_{n-d+1}$ is a block of references which will be compared with other blocks, so a most similar block in distance to the reference block will have a higher weight. An interesting analogy is that the model considers the experience of the serie $\{Z_t\}$ and so the most similar block will be the one with the highest weight.

Finally, the non parametric forecasting estimation in conditional mean, m periods after the last

realization is:

$$\hat{Z}_{n+m} = \sum_{t=d}^{n-m} W_t^m(X_n) Y_t. \tag{8}$$

Under some regularity conditions, [Collomb, 1984] could prove:

$$|\hat{Z}_{n+m} - Z_{n+m}| \xrightarrow{c.s.} 0$$

In order to validate our model we cut the final q observations of the time series. The target is that the forecast is as most similar as possible to the q final observations.

2.2.3 Parameter Selection

We need a $K_d(\cdot)$ function to estimate density probability functions, we will work with the most used multivariate kernel which is the kernel product:

$$K_d(x1, ..., x_d) = \Pi_{j=1}^{d} K(x_j)$$

Since the bandwidth depends of an estimation of the variance we can work with a robust estimator of it, like the MAD is [Falk, 1997]. Then the bandwidth to use in the non parametric forecasting model is:

$$\hat{h_n} = \hat{MAD}(X)[n]^{\frac{-1}{d+4}}, \quad where \quad \hat{MAD}(X) = med(|X_i - med(X)|)$$

For the Markov coefficient election there are different options. [Auestad & Tjøstheim, 1990] propose an empirical methodology to find the best coefficient for the time serie. Another idea is worked by [Matzner-Løfber *et al.*, 1998], where it is said that a right choice could be $n/4$ or $n/5$ according to the length of the historic data n. In this work we will work with the previous idea and also considering PACF, which allows us to find the number of the lag that influence the most recent observation of the time serie $\{Z_t\}$ [Chen & Tsay, 1993].

A differential factor on this work is that we build our forecast in the difference of the original time series, this has showed us a better performance as will be exposed in next sections. In order to find the forecast, we feed the model with the real data if we want a daily forecast, making a sum of the last observation plus the forecast increment.

$$\Delta Z_t = Z_t - Z_{t-1}, \quad 1 \leq t \leq n. \tag{9}$$

For daily forecast.

$$\hat{Z_{t+m}} = Z_{t+m-1} + \Delta \hat{Z_{t+m}}, \quad 1 \leq m \leq q.$$

3 Data Analysis

The data sets that we consider are the official reports of the Colombian Ministry of Health through the National Institute of Health [Colombian Ministry of Health, 2020]. We are considering the accumulated of daily reported cases and also the accumulated of deaths by COVID-19. A first exploratory analysis is worked in this section. For every graphic, the day 1 is taken as the 3 march of 2020.

In figures 6 and 7 we plot the reported cases and deaths respectively. In the cumulative case, left figure, an exponential behaviour can be identified. Since our proposed model work on the difference of the variable we also plot the difference of each variable, right figure. In order to choose an adequate Markov's coefficient, we also expose the partial auto correlation function for each one of the time series.

On this work we will be making the forecast experiment daily. This means that once we predict one period for the next estimation the model will be fed with the immediately past real observation.

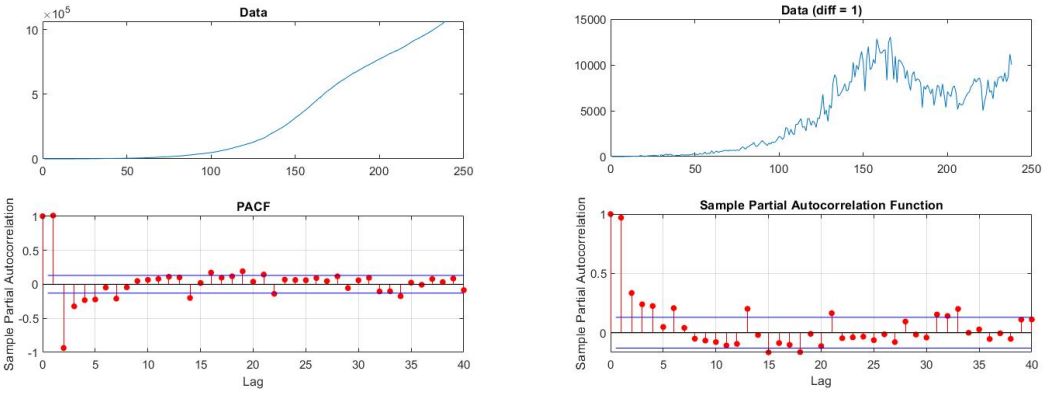

Figure 6: Accumulated and daily reported cases of COVID-19.

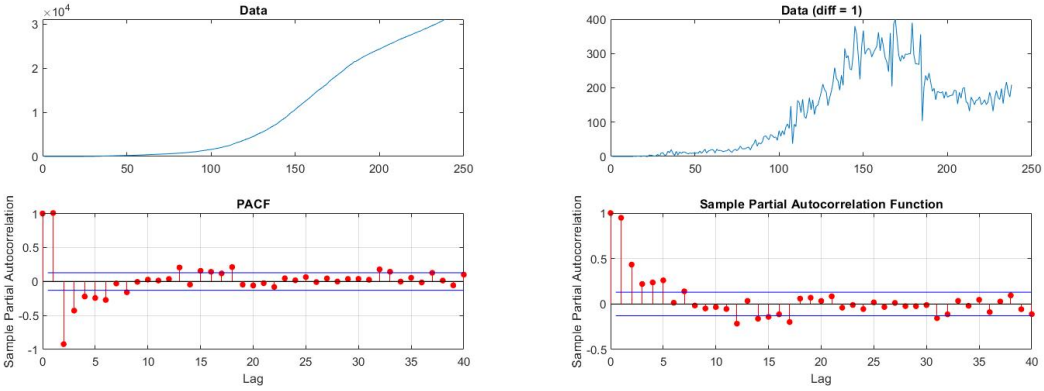

Figure 7: Accumulated and daily reported deaths of COVID-19.

The idea of making this approach is given by the uncertainty of the pandemics, in each day we will desire to know how the behaviour would be tomorrow. A critical case where this model can be used is in the prevention strategies.

Other variables are included and exposed below. We wanted to consider some economic variables, a historic data set of sales per week and also the IPC index of Colombia. The sales data shown in figure 8 is provided by an enterprise. The IPC index is taken from [Departamento Administrativo Nacional de Estadística: www.dane.gov.co", 2020], its evolution through the time can be observed in figure 9.

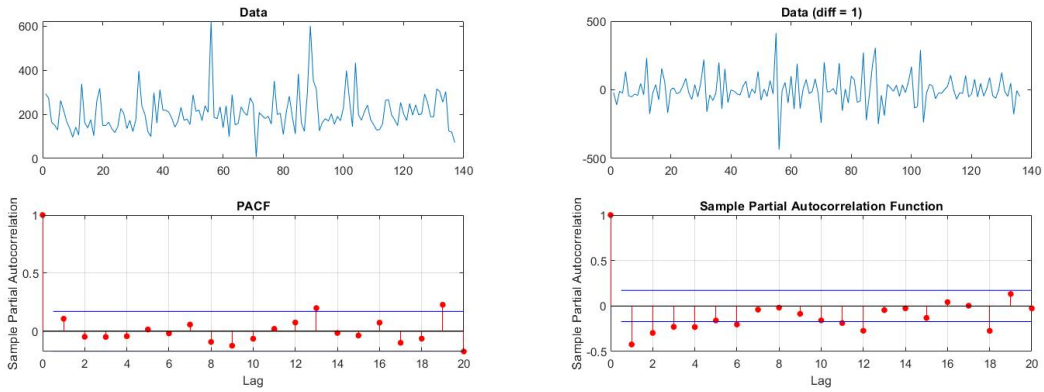

Figure 8: Historic sales of a product per week.

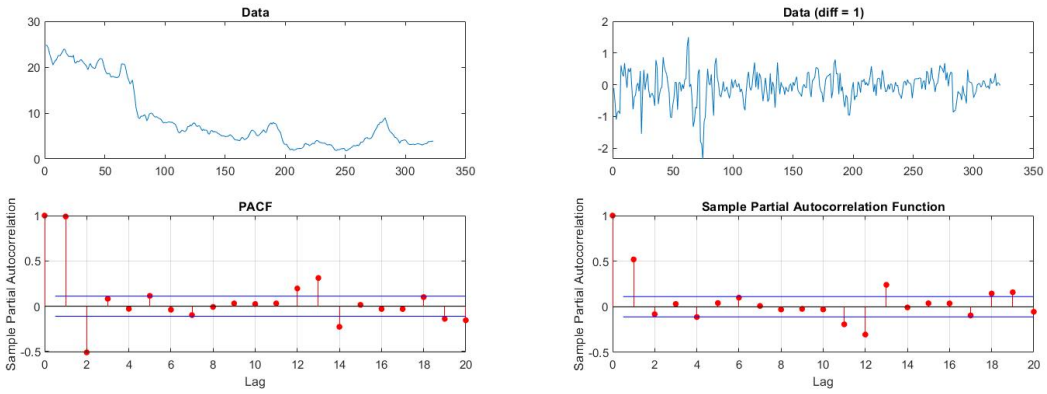

Figure 9: IPC index per month in Colombia.

4 Results

In this section we will show the comparison between the non parametric proposed model and a classical arima model. The arima model is selected according to the order that minimizes BIC and AIC information criteria. Every figure has explicitly defined the combination of parameters used to make the performance. It is important to mention that the arima models considered are only a pure arima process, seasonal and drift models are not considered to be chosen as the models that minimizes information criteria.

We consider the absolute percentage error defined as $\frac{abs(z-\hat{z})}{z}$, where $\hat{z}$ is an estimation of the real observed value in the time serie z. The forecast experiments are exposed below.

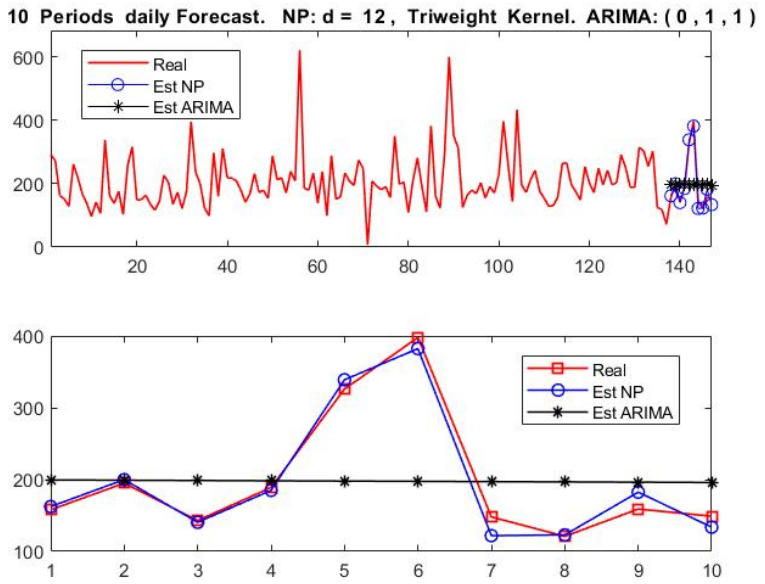

Figure 10: Ten weeks weekly Forecast of sales.

As can be observed in figures 10 and 12 the non parametric forecasting method is able to follow really close the behaviour of the time series.The non parametric model has a better forecasting performance under the absolute percentage error metric. This is illustrated in the figures 11 and 13, where the prediction error is plotted for each forecast day.

Depending of each variable the one period forecast can be more useful. For example the IPC

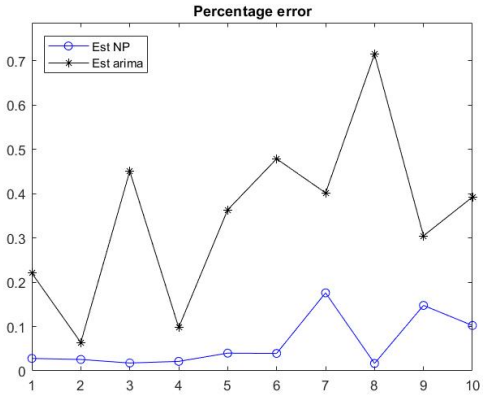

Figure 11: Percentage error of sales forecast.

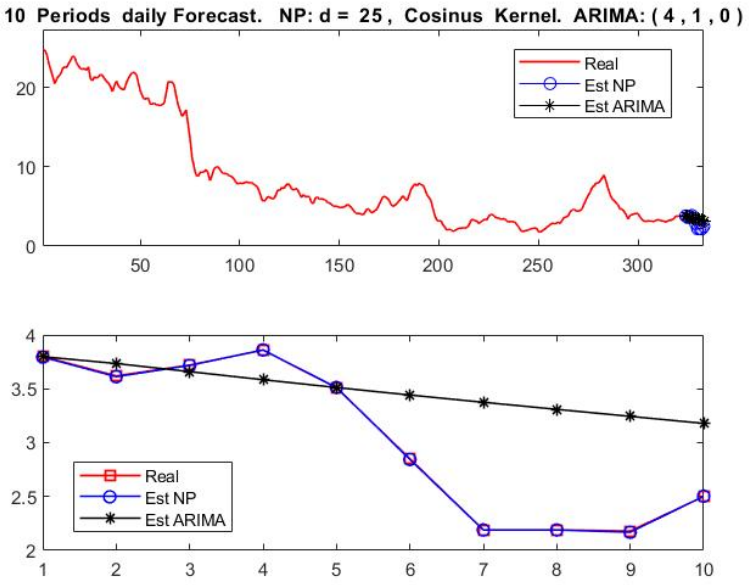

Figure 12: Ten months monthly Forecast of IPC.

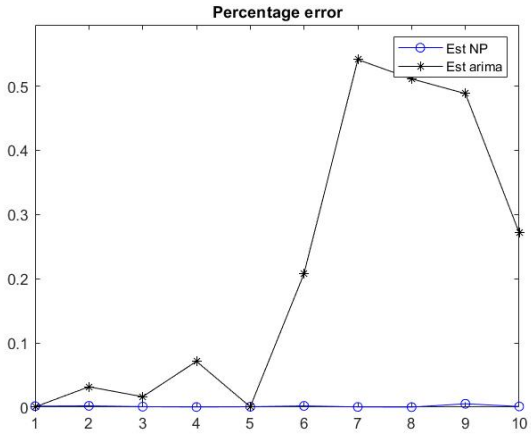

Figure 13: Percentage error of IPC forecast.

index which is a monthly frequency variable could be modeled thanks to the proposed model, allowing to understand how the inflation is evolving in the current month. For the sales variable a most interesting interpretation could be done, since it is a weekly variable through this model an expected demand can be estimated as same as the inventory to have of a product, which allows to have an optimal quantity of a product; similar to the problem studied in [Petruzzi & Dada, 1999].

Now we show the results of making daily forecast of reported cases and deaths of COVID-19 in Colombia. The figures 16 and 14 shows the evolution of the time series, cases and deaths respectively as well as its forecast by the non-parametric method and by the arima traditional methodology. The model proposed in this research has a lower absolute percentage error, this can be observed in the figures 17 and 15, where the daily prediction error is presented for each forecast model.

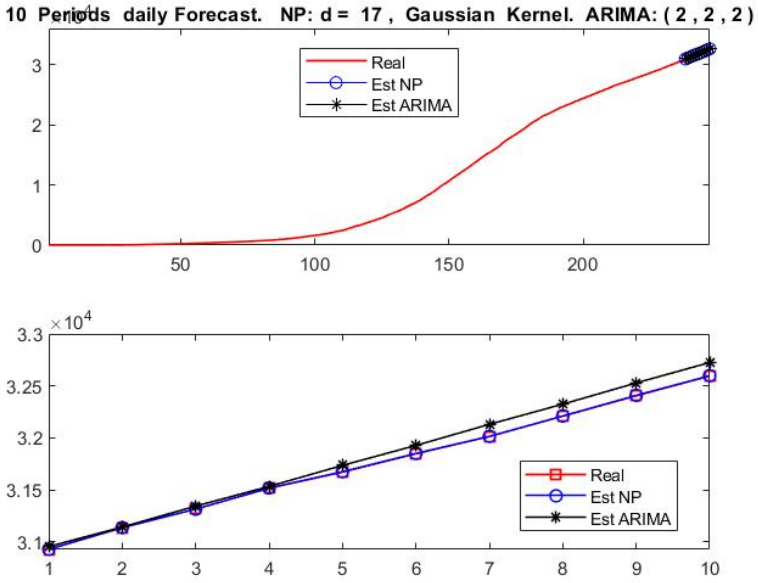

Figure 14: Ten days daily Forecast of deaths.

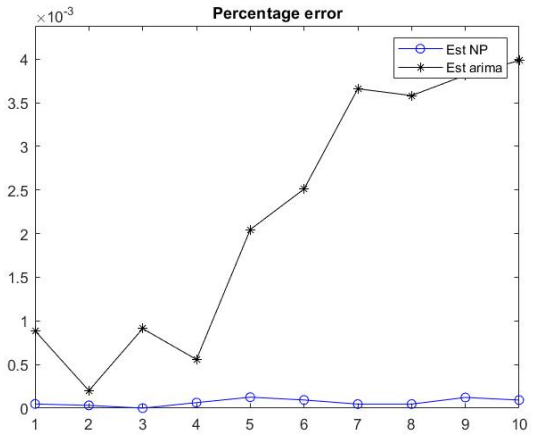

Figure 15: Percentage error of deaths forecast.

The COVID-19 variables of reported cases and deaths can be forecasted daily with a very low error, but it can not be so useful for prevention politics. Then we propose a seven days forecast, where each estimation is done based on the previous forecast value found. This decision is taken in order to represent a most real situation of a long period forecast where the model most recent

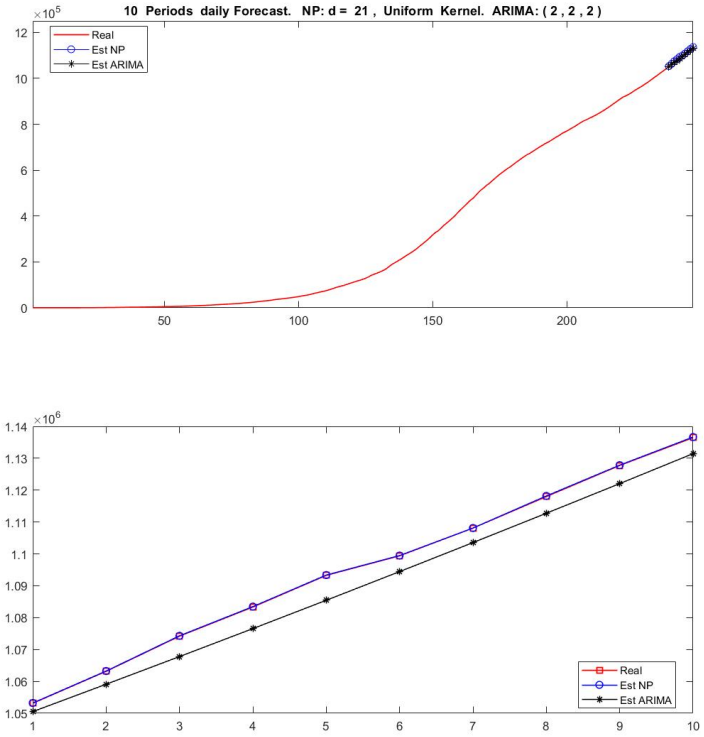

Figure 16: Ten days daily Forecast of reported cases.

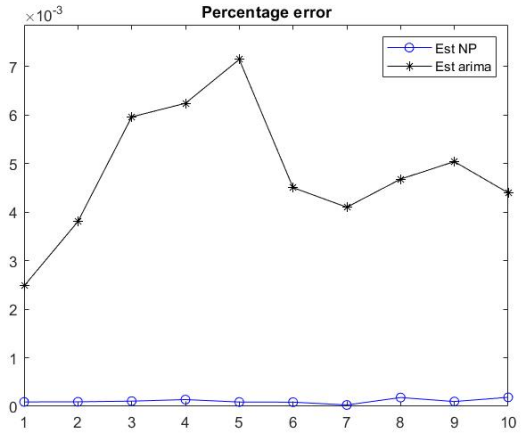

Figure 17: Percentage error of reported cases forecast.

predictions depends of past predictions. These estimations with its respective prediction error can be found respectively in figures 18 and 19 for reported cases, figures 20 and 21 for deaths.

It is important to mention that for the last two experiments, seven day forecast for reported cases and deaths, the original time series are cut. This cut is done because the non parametric model is comparing every past process, and the beginning of the serie is composed by many zeros and little values that are nothing but a problem to the proposed model.

These values can take to wrong suppositions in the weights used to define the forecast estimator exposed in the equation 6, since they can be considered like outliers for the recent behaviour of the variable. Then these outliers in the early part of the serie are omitted to improve the accuracy of the

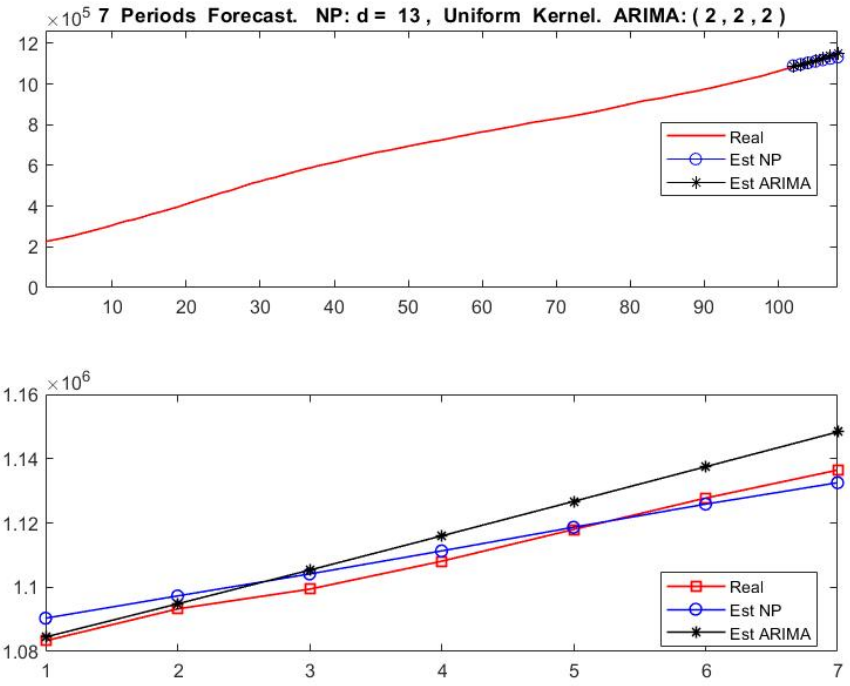

Figure 18: Seven days Forecast of reported cases.

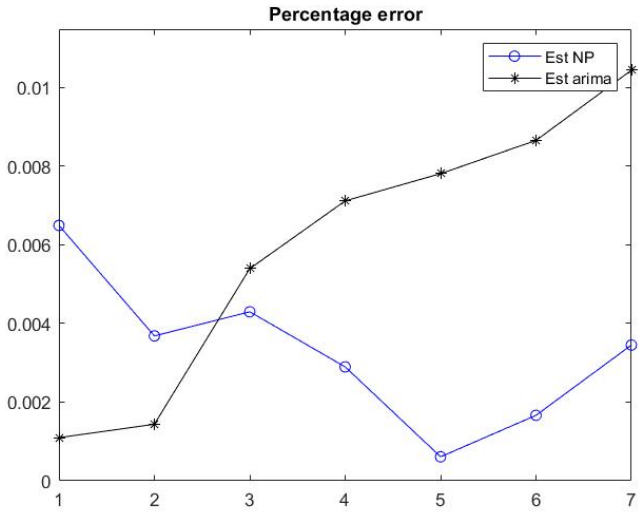

Figure 19: Percentage error, Seven days forecast of reported cases.

non parametric forecasting method. Finally the time series are taken since the 140th observation for the reported cases and for the confirmed deaths since the 70th day. These analysis are exposed on the figures 18 and 20, for cases and deaths respectively.

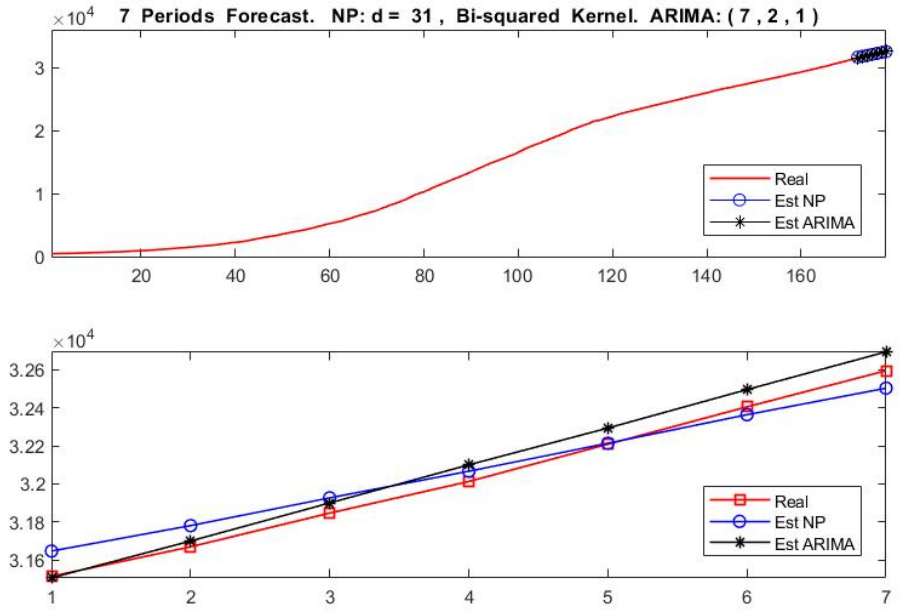

Figure 20: Seven days Forecast of deaths.

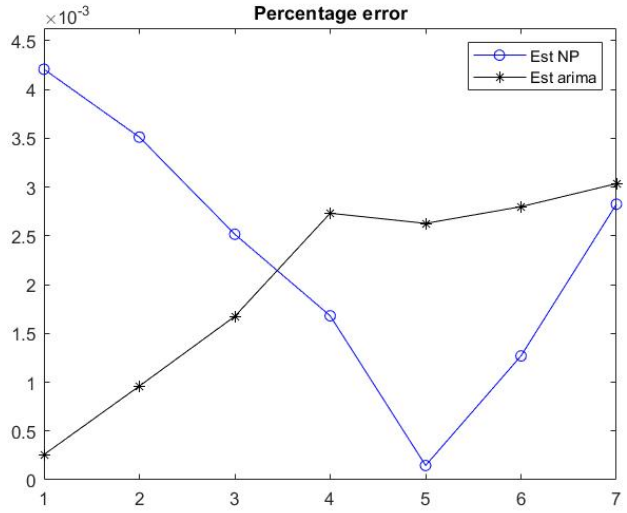

Figure 21: Percentage error, Seven days forecast of deaths.

5 Conclusions

In the section 4 we could evidence the great performance of the non parametric method proposed in this research to forecast some variables. As was observed, in almost every cases it has a lower prediction error than the arima models. The model is not limited to a certain area, we could find great estimations of very different context time series.

This model can be extrapolated to different study areas. An interesting case to study could be the financial area, where one period forecast can be really useful. We could witness very low error in predictions for dissimilar time series like epidemic, economic and inventory ones.

A great advantage that the proposed model offers is in the prediction horizon. Since we can make a prediction for a punctual date after the last observation without estimating the intermediate values. These allow to avoid the estimation error than can be accumulated in long periods forecast.

For future work, we could suggest to define a multi variate forecast model. There are some exogenous variables that can help to describe in a most accuracy way the behaviour of a variable through the time. Since our model is highly dependent of statistical moment estimators, any contribution to improve these estimators can lead to a better forecasting model. A final great advance that could be done is to design confidence intervals of the punctual forecast developed in this research.

References

[Auestad & Tjøstheim, 1990] Auestad, Bjørn, & Tjøstheim, Dag. 1990. Identification of nonlinear time series: first order characterization and order determination. *Biometrika*, **77**(4), 669–687.

[Bosq, 1998] Bosq, D. 1998. *Nonparametric statistics for stochastic processes, volume 110 of Lecture Notes in Statistics*.

[Bosq, 2012] Bosq, Denis. 2012. *Nonparametric statistics for stochastic processes: estimation and prediction*. Vol. 110. Springer Science & Business Media.

[Brooks, 2019] Brooks, Chris. 2019. *Introductory econometrics for finance*. Cambridge university press.

[Chen & Tsay, 1993] Chen, Rong, & Tsay, Ruey S. 1993. Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, **88**(421), 298–308.

[Collomb, 1984] Collomb, Gérard. 1984. Propriétés de convergence presque complète du prédicteur à noyau. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **66**(3), 441–460.

[Cucinotta & Vanelli, 2020] Cucinotta, Domenico, & Vanelli, Maurizio. 2020. WHO declares COVID-19 a pandemic. *Acta bio-medica: Atenei Parmensis*, **91**(1), 157–160.

[Davidson *et al.*, 2004] Davidson, Russell, MacKinnon, James G, *et al.* 2004. *Econometric theory and methods*. Vol. 5. Oxford University Press New York.

[Falk, 1997] Falk, Michael. 1997. On MAD and comedians. *Annals of the Institute of Statistical Mathematics*, **49**(4), 615–644.

[Ferraty & Vieu, 2004] Ferraty, F, & Vieu, Ph. 2004. Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. *Nonparametric Statistics*, **16**(1-2), 111–125.

[Kindzerske & Ni, 2007] Kindzerske, Matthew D, & Ni, Daiheng. 2007. Composite nearest neighbor nonparametric regression to improve traffic prediction. *Transportation research record*, **1993**(1), 30–35.

[Maleki *et al.*, 2020] Maleki, Mohsen, Mahmoudi, Mohammad Reza, Wraith, Darren, & Pho, Kim-Hung. 2020. Time series modelling to forecast the confirmed and recovered cases of COVID-19. *Travel Medicine and Infectious Disease*, 101742.

[Matzner-Løfber *et al.*, 1998] Matzner-Løfber, Eric, Gannoun, Ali, & De Gooijer, Jan G. 1998. Nonparametric forecasting: a comparison of three kernel-based methods. *Communications in Statistics-Theory and Methods*, **27**(7), 1593–1617.

[Meltzer *et al.*, 2014] Meltzer, Martin I, Atkins, Charisma Y, Santibanez, Scott, Knust, Barbara, Petersen, Brett W, Ervin, Elizabeth D, Nichol, Stuart T, Damon, Inger K, & Washington, Michael L. 2014. Estimating the future number of cases in the Ebola epidemic–Liberia and Sierra Leone, 2014–2015.

[Petruzzi & Dada, 1999] Petruzzi, Nicholas C, & Dada, Maqbool. 1999. Pricing and the newsvendor problem: A review with extensions. *Operations research*, **47**(2), 183–194.

[Rodríguez & Siado, 2003] Rodríguez, Norberto, & Siado, Patricia. 2003. Un pronóstico no paramétrico de la inflación colombiana. *Revista Colombiana de Estadística*, **26**(2), 89–128.

[Vasconcelos *et al.*, 2020] Vasconcelos, Giovani L, Macêdo, Antônio MS, Ospina, Raydonal, Almeida, Francisco AG, Duarte-Filho, Gerson C, Brum, Arthur A, & Souza, Inês CL. 2020. Modelling fatality curves of COVID-19 and the effectiveness of intervention strategies. *PeerJ*, **8**, e9421.

[Wasserman, 2006] Wasserman, Larry. 2006. *All of nonparametric statistics*. Springer Science & Business Media.

[Yang *et al.*, 2020] Yang, Zifeng, Zeng, Zhiqi, Wang, Ke, Wong, Sook-San, Liang, Wenhua, Zanin, Mark, Liu, Peng, Cao, Xudong, Gao, Zhongqiang, Mai, Zhitong, *et al.* 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of Thoracic Disease*, **12**(3), 165.