State of the Art Report on Predictive Analysis of COVID-19 Evolution in the Canary Islands

By Ginevra Iorio

Abstract

This report aims at giving a general framework on the use of Deep Learning methods for time series forecasting and conveying their major applications. After an initial research on relative works, it will be proven how and why these methods can be used to predict the spread of COVID-19 pandemic in the Canary Islands based on past observed data. The aim of this project is to give support to the Spanish public health authorities and to the hospitals in case of a new high peak of infection and avoid possible hospital overcrowdings.

Introduction

Forecasting is a Data Mining task that, starting from a known data model, aims at finding possible patterns and predicting future results. On the strength of its outcomes, forecasting has proven to be useful in many fields of research and consequently, predictive analysis with Deep Learning methods is widely applied nowadays. The methodologies with which it can be exerted are numerous, and the majority of them makes use of neural networks, between which Long Short-Term Memory recurrent neural networks (RRNs) provide great accuracy for forecasting and subsequently are the most efficient to use. Typically, RRNs have a short-term memory because they employ earlier knowledge that is persistent to inform the current neural network and they can experience a loss of information, referred to as the *vanishing gradient problem,* caused by the frequent use of the recurrent weight matrix. In a LSTM model, an identify function that is managed by a set of gates replaces this matrix and solves the problem. Thence, LSTM will be used in this project to build the predictive model.

Time Series Analysis and Forecasting Applications

As aforementioned, Deep Learning methods for forecasting systems have many different applications, the main of which happen to be the following:

- Demand Forecasting for Businesses

A crucial challenge for companies that handle supply and procurement is predicting the client demand. According to this, another typical application is for

the dynamic change of prices and rates for goods and services in response to demand and revenue targets [1]. The purpose of an accurate demand forecast is to minimize the deviation between the actual demand and forecasting and using time series for this purpose is appropriate when the demand pattern doesn't vary significally every year.

The advantages of demand forecasting are more effective production scheduling, inventory management and reduction, cost reduction, optimized transport logistics and increased customer satisfaction [2].

- Price Prediction for Apps

The aim of price prediction in time series forecasting is to generate occasions to improve and personalize the user experience when referring to apps. One relevant case is a flight booking business called Fareboom.com, that is successful in locating the most affordable tickets for its clients. Because they vary quickly and for no apparent reason, airfares are an issue and having some future price information would be useful for a user. Giving clients the knowledge about whether prices would drop or increase in the future would make customers more likely to make repeat purchases and choose Fareboom as their preferred travel tool [1].

- Forecasting Pandemic Spread

The primary method used in healthcare to forecast the spread of Covid-19 is time series analysis and forecasting. It was used to predict transmission, calculate mortality rates, track the epidemic's spread, and more [1]. Being a recent event, these techniques have been widely applied lately to gain the backing of the government and the community in order for them to get ready for the upcoming days of the pandemic.

On these premises, time series analysis and forecasting can be applied to almost all healthcare fields, from genetics, to medications expenditures, to diagnosis and treatment.

- Anomaly Detection for Cyber Security

One of the most frequent machine learning jobs is anomaly detection, which searches for outliers on the distribution of the data points. Time series data is very handy for this purpose, while it is not a requirement. Detecting outliers is very beneficial when it comes to fraud detection and for instance PayPal applies time series analysis to monitor each account's typical operational frequency in order to spot any unusual spikes in the volume of transactions.

Subsequently, Cyber Security takes advantage of these techniques to capture and report malware attacks through sensors. These tools are also applied for *predictive maintenance,* the practice of anticipating equipment failure to cut down on repair and downtime costs [1].

Seemingly, time series analysis for forecasting systems has many relevant applications in disparate fields. Nonetheless, this project will focus on one of the aforementioned practices, which is expressly time series analysis for pandemic prediction.

Why it is Important to Make Predictions on COVID-19

The new virus called Corona-virus Disease 2019 (COVID-19), or SARS-CoV-2, has made its first apparition in China, in the Wuhan city, at the end of 2019. Coronaviruses are positive single-stranded, enclosed, big RNA viruses that can infect a variety of animals in addition to humans. On the Huanan seafood market in Wuhan, China, the virus appears to have successfully crossed over from animals to people. Promptly after, it has started spreading globally and, in the first months of 2020, it has caused the entire world to shut down in order to fight against it. Indeed, COVID-19 uses receptor-mediated endocytosis to infect lung alveolar epithelial cells. Clinical indications of the illness, which include fever, coughing, nasal congestion, exhaustion, and other symptoms of upper respiratory tract infections, typically begin less than a week after the onset of symptoms in symptomatic patients. In almost 75% of patients, the infection can proceed to a severe condition with dyspnea and severe chest symptoms similar to pneumonia [3].

Accordingly, especially in the first months of the pandemic, hospitals were quickly overcrowding, the medications not always effective and the whole population threatened by this highly contagious virus. The healthcare systems of numerous nations, including Italy, Spain, France, and the United States, have been severely harmed by this disease's exponential global expansion. The COVID-19 epidemic, which has a significant detrimental impact on public health, is still currently one of the most pressing issues facing our society [4].

In an effort to fighting against COVID-19, particularly during the the first months of the spread, there were not so useful tools and the most ideal and convenient choice to curb the pandemic was predicting its trend in advance using Deep Learning methodologies. Several modeling, estimation, and forecasting practices have been proposed in order to comprehend and control this epidemic. A variety of mathematical models are now used to gauge and predict the progression of the confirmed infected cases and of the hospitals' crowding. The use of hospital resources and the development of treatment strategies to best manage infected patients have now become dependent on accurate forecasting of the number of COVID-19 cases.

Relative Works

Having numerous applications, time series analysis and forecasting has been widely applied in many different research projects and, in the past few years, being COVID-19 a major issue for society, a great deal of them has focused on forecasting the pandemic spread. Every epidemic of an infectious disease has a certain pattern, and these patterns are required to be discovered based on the dynamics of the outbreaks' propagation.

At an early stage of SARS-CoV-2, Zhao *et al.* [5] have immediately started estimating the transmissibility of COVID-19 by the basic reproduction number, which estimates the ability of a new pathogen to spread and is defined as the average number of secondary transmissions from one infected person. This has given the opportunity to promptly understand the trend of the new pandemic and immediately start acting for its containment. The same method has been applied by numerous other researches, among which Shim *et al.* study on the transmission of SARS-CoV-2 in South Korea [6]. The goal was to study the evolution of the reproduction number for COVID-19 in Korea using the Korean incidence curves for imported and local cases. The estimations of the reproduction number clearly show that the novel coronavirus was being transmitted continuously in Korea, and the case fatality rate seemed to be higher in males and older populations.

Regardless, any pandemic in a country or province typically happens at varying levels of magnitude throughout time, due to seasonal variations and the virus's ability to adapt. A study conducted in 2020 on the confirmed cases of different countries using the ARIMA model has proven that, indeed, while the confirmed cases' trends in China were going towards a stabilization in the upcoming weeks, exactly like South Korea and Thailand, Iran and Italy still had very unstable trends [7]. Suitably, different studies have needed to be conducted in different countries using Deep Learning methodologies to help the governments and the public health authorities contain the epidemic.

In [8], LSTM networks have been used to predict the COVID-19 outbreak in Canada based on the past transmission data. To discover the best model that can predict future infections with the least amount of errors, after choosing the network's main parameters, multiple tests were undertaken. The patterns retrieved have then shown that the Canadian public health authorities had been capable to minimize the human exposure, obtaining positive results compared to other countries such as USA and Italy. While transmission rates in the USA were growing exponentially, they were following a linear pattern in Canada. It has also been discovered that provinces that adopted social distance policies prior to the pandemic had fewer confirmed cases than other provinces.

This is not the case of Italy though, where the excess mortality during the COVID-19 outbreak was studied in [9]. Despite the strict lockdown measures implemented soon after the start of the pandemic, it put a tremendous amount of strain on the healthcare system, and it was especially bad in Lombardy and a few other specific provinces in other northern regions, where hospitals were overburdened. General practices and outpatient activities were frequently suspended to stop the spread of the virus. Emergency and intensive care units concentrated almost entirely on COVID-19 patients, with admissions for other causes significantly reduced. Based on the database released by ISTAT, the authors of [9] applied a two-stage time-series model, comparing the risk for mortality during the outbreak and the one of the pre-outbreak. The results revealed a death rate spike of up to 400% in a few particular northern provinces. The mortality rate proved to be higher in men compared to women and substantially lower in people <60 years old. It can be seen that strict lockdown measures, such as interregional travel restrictions, have probably helped to contain the infection's spread and the ensuing excess mortality in other areas, particularly in the South.

The retrieved data was useful to explain how risks vary among locations and subgroups, as well as the contribution of local or national policies put in place at the time.

Additionally, some studies used time series analysis in order to estimate the impact of lockdowns on reducing the number of cases. In [10], for instance, the quarantine measures adopted in Brazil by the president Bolsonaro were studied to evaluate their effectiveness. The data was collected from the Brazilian Ministry of Health website and used to conduct an interrupted time series (ITS) analysis to estimate the impact of social distance measures on reducing the number of COVID-19 cases and deaths in Brazil. It was proved that, in order to stop the spread of contagious diseases, social distancing is a well-established non-pharmaceutical strategy. As a result, social distance policies have shown to have a noticeable impact on the SARS-CoV-2 outbreak, according to studies utilizing the ITS design. According to Tobías [11], who examined how lockdown measures affected the SARS-CoV-2 pandemic in different countries, such Spain and Italy, the first lockdown had decreased incidence in both nations while maintaining favorable trends. The slope of the number of case occurrence shifted and turned negative after the second, harsher lockdown. After these studies it is then possible to conclude that social distancing and restrictions have been very useful to contain the pandemic and give a brake to the exponentially growing number of COVID-19 cases. The less the cases, the more hospitals can be able to tame the spread and avoid overcrowdings, hence, predicting the number of future hospitalization cases could also be a way of helping the public health authorities.

Ever since the end of 2019, COVID-19 has caused overwhelming disruptions to healthcare systems around the globe. Traditional epidemiological models, such as the SIR model, do not give health system officials enough time to plan effectively. The emergency medical services (EMS) demand and emergency department visits increased significantly after the pandemic spread, resulting in a lack of pre-hospital patient assessment, medical resources, and personnel equipment. The seek of [12] is to first look at how the pandemic has affected emergency calls over time and how long it takes an ambulance to be assigned, sent, and arrive on average and then design a predictive model to forecast the daily number of COVID-19 EMS calls. The study was conducted based on two datasets retrieved on the City of Austin Open Data Portal. In the modeling procedure, time series regression with the change point detection was used to forecast the daily frequency of pandemic EMS events, being the daily frequency of COVID-19 hospitalization the regressor. The change point detection method was applied and then, to reduce the noise and improve the predictive power, the time series of daily hospitalization was smoothed by an average of a period of seven days. Later on, autoregressive integrated moving average (ARIMA) models and binomial thinning have been applied after having divided the dataset in training (80%) and testing (20%). As a result, it was seen that since the start of the Covid-19 pandemic and the proclamation of the local state of emergency, the number of non-pandemic EMS incidents per day has sharply decreased. What's more intriguing though is that once the Covid-19 pandemic broke out, the percentage of non-pandemic dead calls increased. Some earlier conducted studies in the United States suggested that many people refused to be transported to hospitals out of fear of contagion. Another observation made was a significant prolongation of the EMS response time, indeed, according to local Austin news reports, the EMS department did not get enough funding to keep enough staff and ambulances on hand throughout the epidemic, which caused these delays. The study carried out in [12] is useful to open up a discussion amongst EMS

organizations, government officials, and healthcare partners about how to modify EMS demand shifts and reduce response times during upcoming pandemics.

In general, models fitted to hospitalized patients are more trustworthy and accurate than models suited to cases that have been confirmed [13]. The initial objective of [14] is thus to give short-term and medium-term projections for the number of COVID-19 patients who will be admitted to hospitals during the second wave of infections in Italy, or after October 13, 2020. Hospitalization patterns associated with COVID-19 paint a vivid picture of the overall strain on the national healthcare system. The data used in this project referred to the real-time number of COVID-19 hospitalizations of patients with mild symptoms and patients assigned to the intensive care unit in Italy from February 21, 2020, to October 13, 2020, extracted from the Italian Ministry of Health's website. According to the data, the number of COVID-19 patients admitted to hospitals for minor symptoms and the number of COVID-19 patients placed in the intensive care unit both achieved a first peak on April 4, 2020. After that, they had a declining trend until the middle of August, before picking up speed once more from the end of September to the middle of October 2020. The author in [14] computed forecasts using four different statistical techniques and their combination, such as linear ARIMA models, ETS models, linear and nonlinear NNAR models and TBATS models. To compare the performances of the single and hybrid prediction models, MAE, MAPE, MASE, and RMSE were used as metrics. In conclusion, the most accurate model for both individuals hospitalized with minor symptoms and patients admitted to ICU between those used, ended up being neural network autoregression (NNAR).

Given the fact that Long Short-Term Memory recurrent neural networks take significantly less time to train, it is really convenient to use it for time series analysis. This has been proved by [15], where hospitalizations in French regions and in Belgium were forecast using LSTM approach and obtaining good and truthful results with low efforts. In this project, Google Trends has been used to retrieve the data by its web interface. Then it was divided in training set (seven regions), validation set (three regions) and testing set (four regions). LSTM-based models were experimented on the training dataset and produced as output one single number: the estimated hospitalizations seven days later. For the optimization, the Mean Squared Error (MSE) has been used as error metric and a forecast on the COVID-19 hospitalized cases has been produced, giving support to the government to adjust the measures and restrictions, to hospitals to prepare for the incoming flow of patients, and serving as a warning in case of detection of a second future peak.

COVID-19 Threat to the Canary Islands

The SARS-CoV-2 pandemic has been a major global threat since the end of 2019 and, since then, taming its quick spread has been a significant concern. The Canary Islands, nonetheless, have accumulated 323,329 cases of Coronavirus as of the 22nd of March 2022 and 1,617 people have died. Important restrictions have been actuated in the first months of the spread, following four phases [16]:

1. Phase Zero: residents were allowed to go out once a day, with restrictions to the times and distance from home.
2. Phase One: from May 11, only in the provinces that met the requirements, small retails, bars and restaurants, outdoor markets and hotels could open with reduced capacity and social distancing rules.
3. Phase Two: from May 25, only in the provinces that met the requirements, restaurants could offer inside tables, shopping centers reopened, and small cultural centers could take place.
4. Phase three: from June 22, if the requirements were met, travel would have been allowed within Spain for residents.

More than two years later, COVID-19 keeps threatening the islands as the pandemic refuses to go away. Between July 1st and 5th, hospitalized cases increased by 13.7% and in Gran Canaria the percentage of occupancy of hospital beds by COVID-19 patients was 10.06%, a data that, accordingly to the Public Health Committee, was considered high risk [17].

Currently there are no predictions made on COVID-19 spread in the Canary Islands and, given the fact that the pandemic still doesn't seem to come to an end, they would turn out to be very helpful to contain the spread and help organize the hospitals. As seen previously, time series prediction analysis and forecasting conducted on Coronavirus cases has come in handy for many countries, such as China, Italy, or Canada. Being able to forecast a prediction on the future hospitalized cases would help the Canary Islands' public health authorities to prepare the hospitals receive the expected amount of patients. Lowering the risk of coming unprepared to a new possible peak would also be useful to learn how to behave in such situations and slowly go back to normality. What was then considered "high risk" would be prevented and there wouldn't be no similar situations.

In conclusion, the Canary Islands are in the need of a prediction method that, as seen after its use in other countries, could aid in case of new COVID-19 peaks and hospitals overflow.

Time Series Analysis of COVID-19 in the Canary Islands

Between all the time series forecasting methods, it has already been demonstrated that Long Short-Term Memory Recurrent Neural Networks (RNNs) is the most efficient one, and it will be applied in this project for COVID-19 hospitalization cases analysis and forecasting.

Starting from datasets on the daily patients of the Canary Islands' hospitals, both the ones that are in intensive care units and those that are not, these will be divided into *training set* (80%), that will train and test the models, and *testing set* (20%), that will test our model's performance. Moreover, the input data will have a temporal component, since time series need to collect data that is collected over regular intervals of time. Time series datasets usually can be broken down into trend, seasonability and error. When a specific pattern appears repeatedly over an extended period of time, trends can be identified, however, both trends and seasonality are often absent in real-world situations. The series also needs to be classified into stationary, if it doesn't depend on the time components and mean and variance

are constant, or non-stationary, if it has trend, seasonality effects and it changes with respect to time. A useful method to distinguish these two types of series is the Augmented Dickey Fuller (ADF) test [18], which finds the impact of trends on the data and the results are then interpreted observing the p-values of the test [8].

Recurrent LSTM networks may adapt the nonlinearities of a given COVID-19 dataset to overcome the constraints of conventional time series forecasting methods. The output of each LSTM block is passed on to the following block, which runs at a different time step, and so on until the last LSTM block produces the sequential output. This will allow to build a time series sequential model. Memory blocks, the core element of LSTM networks, were developed to combat vanishing gradients by remembering network parameters for extended periods of time. The three gates of LSTM help then to process the information and give an output between 0 and 1.

Finally, the model predictions will be examined using two different methodologies: the root mean square error (RMSE) and the mean absolute percentage error (MAPE). These will be used during the testing phase to measure the distance between the predicted and the actual values.

The patterns in the future data will then demonstrate that fast and effective measures taken by Canarian public health authorities to improve COVID-19 containment and obstruct hospitals' overcrowding will have a beneficial impact if they follow the retrieved trend.

Conclusions

To sum up, after having reviewed the state of the art and studied recent works conducted on the use of forecasting methods for predicting COVID-19 future outbreaks, it can be seen that there is a vacuum on the Canary Islands' pandemic spread's study. Accordingly, in order to solve this issue, it will be necessary to adopt the same Deep Learning methods that have been applied in many other countries, such as LSTM, which has been proven to be the most effective, being able to overcome the vanishing gradient problem. In fact, it has been made clear that every epidemic of an infectious disease has a certain pattern, and these patterns are required to be discovered in order to evaluate the outbreak and develop effective intervention strategies.

Bibliography

[1] AltexSoft (2022). Time Series Analysis and Forecasting: Examples, Approaches, and Tools. *AltexSoft,* accessed on October 2nd, 2022.
https://www.altexsoft.com/blog/business/time-series-analysis-and-forecasting-novel-business-perspectives/

[2] Business Science (2020). Time Series Demand Forecasting. *R-bloggers*, accessed on October 2nd, 2022.
https://www.r-bloggers.com/2020/11/time-series-demand-forecasting/

[3] Velavan TP, Meyer CG (2020). The COVID-19 epidemic. *Trop Med Int Health*.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7169770/

[4] Abdelhafid Zeroual, Fouzi Harrou, Abdelkader Dairi, Ying Sun (2020). Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study, Chaos, Solitons & Fractals, Volume 140, 110121, ISSN 0960-0779.
https://doi.org/10.1016/j.chaos.2020.110121

[5] Shi Zhao, Qianyin Lin, Jinjun Ran, Salihu S. Musa, Guangpu Yang, Weiming Wang, Yijun Lou, Daozhou Gao, Lin Yang, Daihai He, Maggie H. Wang (2020). Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases*, Pages 214-217, ISSN 1201-9712.
https://doi.org/10.1016/j.ijid.2020.01.050

[6] Eunha Shim, Amna Tariq, Wongyeong Choi, Yiseul Lee, Gerardo Chowell (2020). Transmission potential and severity of COVID-19 in South Korea. *International Journal of Infectious Diseases*, Volume 93, Pages 339-344, ISSN 1201-9712.
https://doi.org/10.1016/j.ijid.2020.03.031

[7] Tania Dehesh, H.A. Mardani-Fard, Paria Dehesh (2020). Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models. *medRxiv*.
https://www.medrxiv.org/content/10.1101/2020.03.13.20035345v1.full-text

[8] Vinay Kumar Reddy Chimmula, Lei Zhang (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals*, Volume 135, 109864, ISSN 0960-0779.
https://doi.org/10.1016/j.chaos.2020.109864

[9] Matteo Scortichini, Rochelle Schneider dos Santos, Francesca De' Donato, Manuela De Sario, Paola Michelozzi, Marina Davoli, Pierre Masselot, Francesco Sera, Antonio Gasparrini (2020). Excess mortality during the COVID-19 outbreak in Italy: a two-stage interrupted time-series analysis. *International Journal of Epidemiology*, Volume 49, Issue 6, Pages 1909–1917.
https://doi.org/10.1093/ije/dyaa169

[10] Silva, Lucas, Figueiredo, Dalson and Fernandes, Antônio (2020). The effect of lockdown on the COVID-19 epidemic in Brazil: evidence from an interrupted time series design. *Cadernos de Saúde Pública*, v. 36, n. 10, e00213920.
https://doi.org/10.1590/0102-311X00213920

[11] Tobías A (2020). Evaluation of the lockdowns for the SARS-CoV-2 epidemic in Italy and Spain after one month follow up. *Sci Total Environ*, 725:138539.
https://pubmed.ncbi.nlm.nih.gov/32304973/

[12] Xie, Y, Kulpanowski, D, Ong, J, Nikolova, E, Tran, NM (2021). Predicting Covid-19 emergency medical service incidents from daily hospitalization trends. *Int J Clin Pract*. 2021; 75:e14920.
https://doi.org/10.1111/ijcp.14920

[13] Holmdahl, I., Buckee, C. (2020). Wrong but useful — What COVID-19 Epidemiologic Models Can and Cannot Tell Us. *N. Engl. J. Med*. 383(4), 303–305.
https://www.nejm.org/doi/full/10.1056/NEJMp2016822

[14] Perone, G. (2022). Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy. *Eur J Health Econ* 23, 917–940.
https://doi.org/10.1007/s10198-021-01347-4

[15] Derval G., François-Lavet V., Schaus P. (2020). Nowcasting COVID-19 hospitalizations using Google Trends and LSTM. *UCLouvain*.
https://dial.uclouvain.be/pr/boreal/object/boreal%3A235257/datastream/PDF_01/view

[16] CamelTravel (2022). Coronavirus in the Canary Islands. *ABTA, Travel with confidence*.
https://cameltravel.co.uk/coronavirus-in-the-canary-islands/

[17] Ministry of Health (2022). Five Canary Islands Now Have Incidence Rates at High Risk for COVID. *CanarianWeekly*.
https://www.canarianweekly.com/posts/Five-Canary-Islands-now-have-incidence-rates-at-high-risk-for-Covid

[18] Y.-W. Cheung, K.S. Lai (1995). Lag order and critical values of the augmented dickey-fuller test. *J Bus Econ Stat*, 13 (3), pp. 277-280.
https://www.tandfonline.com/doi/abs/10.1080/07350015.1995.10524601?cookieSet=1