engineering proceedings

MDPI

Proceeding Paper

An Application of Neural Networks to Predict COVID-19 Cases in Italy †

Lorena Saliaj * and Eugenia Nissi

Department of Economic Studies, School of Economic, Business, Legal and Sociological Sciences, Università degli Studi "G. d'Annunzio", 65127 Pescara, Italy; eugenia.nissi@unich.it

* Correspondence: lorena.saliaj@unich.it

† Presented at the 8th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 27–30 June 2022.

Abstract: COVID-19 pandemic has become the greatest worldwide threat, as it has spread rapidly among individuals in most countries around the world. This study concerns the problem of weekly prediction of new COVID-19 cases in Italy, aiming to find the best predictive model for daily infection number in countries with a large number of confirmed cases. We compare the forecasting performance of linear and nonlinear forecasting models using weekly COVID-19 data for the period between 24 February 2020 until 16 May 2022. We discuss various forecasting approaches, including a Nonlinear Autoregressive Neural Network (NARNN) model, an Autoregressive Integrated Moving Average (ARIMA) model, a TBATS model, and Exponential Smoothing on the collected data and compared their accuracy using the data collected from 23 March 2020 to 20 April 2020, choosing the model with the lowest Mean Absolute Percentage Error (MAPE) value. Since the linear models seem to not easily follow the nonlinear patterns of daily confirmed COVID-19 cases, Artificial Neural Network (ANN) have been successfully applied to solve problems of forecasting nonlinear models. The model has been used for weekly prediction of COVID-19 cases for the next 4 weeks without any additional intervention. The prediction model can be applied to other countries struggling with the COVID-19 pandemic, to any possible future pandemics, and also help make better decisions in future.

Keywords: COVID-19; time series forecasting; NARNN; ARIMA

Citation: Saliaj, L.; Nissi, E. An Application of Neural Networks to Predict COVID-19 Cases in Italy. *Eng. Proc.* **2022**, *18*, 11. https://doi.org/10.3390/engproc2022018011

Academic Editors: Ignacio Rojas, Hector Pomares, Olga Valenzuela, Fernando Rojas and Luis Javier Herrera

Published: 21 June 2022

1. Introduction

The World Health Organization has recognized the COVID-19 virus as a global threat, declaring it a universal epidemic. Predicting COVID-19 infections' future trend is very important, as it has been having a significant worldwide negative impact on economics, medicine, finance, life expectancy, etc. The chance of having data in advance on its spread may enhance public health decision-making, allowing countries to avoid possible future crises by better allocating health resources.

Different forecasting models have been proposed for predicting the global or local spread of the pandemic, since 2020.

In our work, we provide forecasts for the confirmed Italian regions' new COVID-19 cases, using linear and nonlinear time series forecasting models and comparing their accuracy to analyze their advancement based on the daily reported data. Our aim is to forecast new confirmed COVID-19 cases through a comparison of the performance of these models, with the aim to have clear expectations of future new cases.

The purpose of our work is to determine the best COVID-19 new cases forecasting model.

Several studies try to predict the evolution of the COVID-19 pandemic. Batista [1] predicted the number of cases in China, South Korea, and the rest of the world during the first semester of 2020 using a logistic model. Safi and Sanusi [2] applied an ARIMA model on data collected during the first and second pandemic wave. Khan and Gupta [3]

chose an ARIMA (1,1,0) model for predicting Indian COVID-19 infection cases considering data that followed a linear trend. Abotaleb and Makarovskikh [4] proposed a combined ARIMA, Exponential Smoothing, BATS, and TBATS hybrid model for data collected until March 2021 in Russia. Gecili et al. [5] proposed an ARIMA model for American and Italian data collected from February 2020 until April 2020. Salaheldin and Abotaleb [6] chose the exponential growth model for daily COVID-19 forecasting in China, Italy, and USA.

In this paper, we aim to choose the best model among the most known time series forecasting models. Since the COVID-19 new-cases curve follows a nonlinear trend and considering that we have collected more recent pandemic data, this work emphasizes the importance of using nonlinear methods for modeling these time series, as classical linear models would not be able to identify the traits of nonlinear time series and, subsequently, would not give reliable predicted values. We have considered data from the beginning of the spread of the pandemic in Italy (24 February 2020) to 16 May 2022 in the Italian regions, months which were thought, according to previous proposed forecasting models, would correspond to quiet months from the point of the spread of the pandemic.

2. Materials and Methods

In this work, we considered data published online from Superior Health Institute on Epidemiology for Public Health related to COVID-19 infection cases in Italian regions for the period between 24 February 2020 and 16 May 2022 considering:

- new daily regional infections from 24 February 2020 to 16 May 2022;
- the last 8 days for testing daily cases (11 May 2022 to 16 May 2022);
- the last 30 days for testing the forecasting accuracy of the third wave.

The forecasting was conducted through the R package "forecasting", which provides methods and tools for forecasting univariate time series. We implemented an ARIMA model, a NNAR model, as well as a TBATS and Holt's linear model, and chose the best model considering the Mean Average Percentage Error (MAPE) for each of them as follows:

$$\mathrm{MAPE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{t}=1}^{\mathrm{n}} \left| \frac{\mathrm{A_t} - \mathrm{F_t}}{\mathrm{A_t}} \right| \quad (1)$$

where, n is the total number of observations, $\mathrm{A_t}$ is the actual value at time t, and $\mathrm{F_t}$ is the forecast value at time t.

2.1. ARIMA Model

The first model is ARIMA (Auto-Regressive Integrated Moving Average), which is the most common linear model for time series forecasting. It represents a time series as a function of its past values, its own lags, and the lagged errors, to forecast future values. An ARIMA model is compound by three terms: p, d, q:

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_p \varepsilon_{t-q} \quad (2)$$

where, p is the order of the Auto-Regressive (AR) term and refers to the number of y lags, which should be used as predictors; q is the order of the Moving Average (MA) term and it refers to the number of lagged errors used as predictors; while d is the number of differentiating required to make the time series stationary.

Although ARIMA is widely used for time series analysis, it is not easy to choose appropriate orders for its components, so we proceeded to determine them automatically, using the *auto.arima* function to obtain the best ARIMA model for each region (Table 1).

Table 1. The chosen ARIMA model for each region.

Region	ARIMA
Abruzzo	ARIMA (3,1,2)
Basilicata	ARIMA (0,1,5)
Calabria	ARIMA (3,1,2)
Campania	ARIMA (0,1,5)
Emilia-Romagna	ARIMA (2,1,3)
Fiuli-Venezia Giulia	ARIMA (2,1,3)
Lazio	ARIMA (0,1,5)
Liguria	ARIMA (5,0,0)
Lombardia	ARIMA (0,1,5)
Marche	ARIMA (0,1,5)
Molise	ARIMA (2,1,3)
Piemonte	ARIMA (0,1,5)
Puglia	ARIMA (3,1,2)
Sardegna	ARIMA (2,1,3)
Sicilia	ARIMA (0,1,5)
Toscana	ARIMA (2,1,3)
Trentino Alto-Adige	ARIMA (3,1,2)
Umbria	ARIMA (0,1,5)
Valle D'Aosta	ARIMA (3,1,2)
Veneto	ARIMA (2,1,3)

The model estimation concerns the use of statistical techniques to derive the coefficients that better fit the chosen ARIMA model. Once the model was identified and the parameters were estimated, it was used for forecasting. It is checked using statistical tests and residual plots that can be used to analyze the suitability of various models to historical data.

2.2. TBATS Model

The TBATS (Trigonometric Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend, and Seasonal component) model uses a combination of Fourier terms with an exponential smoothing state space model and a Box-Cox transformation, in an automated manner. The unit of time used in modeling was day.

2.3. Holt's Linear Trend

This model includes a prediction equation and two smoothing equations. It uses double exponential smoothing parameters to forecast future values: The first parameter is used for the overall smoothing, while the other for the trend smoothing equation. We obtained the current value considering the adjusted last smoothed value for the last period's trend and updated the trend over time, expressing it as the difference between the last two smoothed values.

Holt's forecast equation:

$$\hat{y}_{t+h|t} = l_t + hb_t \tag{3}$$

where

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{4}$$

indicates the first equation (level equation), while

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \tag{5}$$

indicates the trend equation, where:

$0 \leq \alpha \leq 1$ is the smoothing parameter for the trend, $0 \leq \beta \leq 1$; l_t indicates the time series value at time t; b_t is the time series trend at time t.

2.4. ANN Model

Artificial Neural Networks forecasting models are nonlinear models inspired by biological neural networks that identify and model nonlinear relationships between the variables. They are compounds of a collection of neurons, grouped in input, hidden, and output layers, and map a set of inputs into a set of output variables, through hidden layers of neurons. Their ability to learn from a training procedure and previous examples makes them a powerful forecasting tool. They have the ability to analyze new data based on previous results.

An ANN is composed of several layers:

- The first layer, known as the input layer, is the one that takes the data in input.
- The last layer, called the output layer, gives the results of the analysis or the solution to the problem.
- The hidden layers, through which data flows from the input layer to the output. This is where the data is analyzed and the outputs are taken. The nodes of the hidden layers detect the features in the pattern of the data and the relationships between them. Then, the requested output is sent from the hidden layer to the output layer.

In our study, the NAR network was developed using the "nnetar" function of R software "caret" package that fits a neural network model to a time series [7] developed by Hyndman, O'Hara, and Wang. A NNAR (p,k), where p indicates the number of non-seasonal lags used as inputs and k the number of nodes in the hidden layer, can be described as an AR process with nonlinear functions. Considering the traits of the new COVID-19 cases trend for Italian regions, we chose a (28-5-1) network, with 28 lags as input nodes and 1 hidden layer with 5 nodes. It has the form of a feedforward three-layer ANN, where neurons have a one-way connection with the neurons of the next layers. The data set was divided into training set (70%) and testing set (15%), while the last 8 days data were used for the validation.

The forecasting performance of all these models was evaluated using the Mean Absolute Percentage Error (MAPE), while the model fits were evaluated using AIC (Akaike Information Criterion), reported in Table 2.

Table 2. MAPE (%) for forecasting models' accuracy.

Region	ARIMA	TBATS	Holt's	ANN
Abruzzo	47.07	65.68	58.68	14.56
Basilicata	45.13	60.02	71.24	12.28
Calabria	40.23	54.10	46.53	28.07
Campania	62.11	69.36	55.81	25.64
Emilia-Romagna	39.62	50.08	49.31	34.15
Fiuli-Venezia Giulia	29.56	44.37	40.65	20.18
Lazio	33.85	43.78	53.12	24.52
Liguria	39.41	52.44	44.18	22.46
Lombardia	47.54	65.13	46.24	32.14
Marche	44.25	41.28	51.23	13.54
Molise	40.68	44.76	50.04	11.37
Piemonte	38.72	60.81	65.13	30.16
Puglia	33.15	47.85	44.32	26.45
Sardegna	39.87	44.65	49.16	27.09
Sicilia	22.45	42.11	45.02	20.33
Toscana	24.03	47.16	60.87	24.18
Trentino Alto-Adige	25.39	52.64	55.71	27.89
Umbria	20.48	54.54	54.39	18.52
Valle D'Aosta	32.18	53.07	47.85	19.15
Veneto	33.30	48.79	50.69	20.05

3. Results

Selection and accuracy measures for the forecasting models are reported in Table 2. MAPE was used to measure the performance of the models. We chose the best forecasting model according to the MAPE value, as it is recommended as an accuracy comparing unit when using different methods on a time series, considering as the most accurate model the one with the lowest MAPE value.

In addition to the graph, where it can be clearly seen, the above values of the table show that the ANN model has given more accurate forecasting values than the other linear forecasting models, for every region. According to MAPE, ANN improved the forecasting accuracy compared with ARIMA, TBATS, and Holt's.

The NARNN model gives better results in almost all the considered regions, with a considerable difference from the indicators of the other models. ANN model has the lowest MAPE for the considered period for all the regions, improving the forecasting performance up to 36.47%, considering Campania. TBATS model has the highest MAPE values for the considered period, indicating that it cannot appropriately follow our data's traits.

In Table 3 we present the MAPE value for the last 6 days data, considered as testing data. Once again, we can observe that the ANN model is the best for forecasting COVID-19 new cases in Italian regions. This fact confirms once again our assumption about choosing the best model for our time series, considering the nonlinear trend our data follow.

Table 3. MAPE (%) for 6 days' accuracy of ANN forecasting model for Italian regions.

Region	11 MAY 2022	12 MAY 2022	13 MAY 2022	14 MAY 2022	15 MAY 2022	16 MAY 2022
Abruzzo	13.54	14.42	10.28	12.96	13.08	12.84
Basilicata	12.52	10.85	12.27	13.63	11.05	12.71
Calabria	25.32	24.12	16.05	22.41	15.33	20.82
Campania	23.45	18.96	17.05	21.21	26.84	20.36
Emilia-Romagna	28.97	32.12	30.54	33	29.06	31.80
Fiuli-Venezia Giulia	20.45	17.33	19.56	17.84	19.07	21.32
Lazio	21.46	22.30	27.18	19.36	20.08	22.18
Liguria	20.97	21.54	19.75	21.13	16.33	19.54
Lombardia	31.84	30.46	32.56	28.72	31.88	32.99
Marche	12.72	14.56	16.22	15.02	13.21	11.88
Molise	10.76	11.64	12.12	10.98	12.54	10.35
Piemonte	28.12	25.12	22.31	20.64	29.22	31.60
Puglia	23.40	21.44	20.16	25.39	22.48	24.56
Sardegna	21.39	17.96	17.75	23.41	24.84	20.36
Sicilia	22.83	21.78	20.05	21.44	26.45	22.02
Toscana	22.18	18.56	19.34	21.12	20.88	20.52
Trentino Alto-Adige	26.59	23.17	16.75	24.31	15.42	20.78
Umbria	18.52	21.56	16.45	19.02	23.21	20.82
Valle D'Aosta	19.15	17.33	26.02	15.25	13.54	15.65
Veneto	20.05	24.06	26.69	18.43	18.28	17.98

We performed the forecasting for new COVID-19 cases in Italian regions using the above models. We conducted a 30-days-ahead forecast (until 16 June 2022) and compared the forecasting data with the testing data for 6 days (11 May 2022–16 May 2022), applying the forecasting models to the confirmed cases for the last 8 days data and compared the results with the actual COVID-19 data. We calculated the MAPE values as the difference between actual data and forecast values. The MAPE values for ANN forecasting model are represented in Table 3. Based on our analysis, we concluded that the prediction performance of the models was similar to the real data. In particular, the ANN model gave more accurate predictions, as its MAPE values were lower compared to the other models. We observed decreasing MAPE values, in particular for the last 6 days' testing values, as its values decreased from about 7% to 1%. Higher MAPE values were observed for the other predictive models. ARIMA had a worse predicting performance for the first 3 days and

the last day, while TBATS was the worst forecasting model when comparing the 6 days' training data MAPE values.

Figure 1 presents the forecasting results of ANN model for the following 30 days for COVID-19 new confirmed cases in Italian regions. The NARNN model values follow very well the time series' trend thanks to the training process, which enables the model to better understand the time series' features. Once trained, the ANN decides itself on the importance of the variables, as it keeps learning continuously, performing quite well with unfamiliar data, thanks to its ability to work with multiple parallel inputs, as well as nonlinearity and plasticity in finding the most suitable model for time-series forecasting [8].

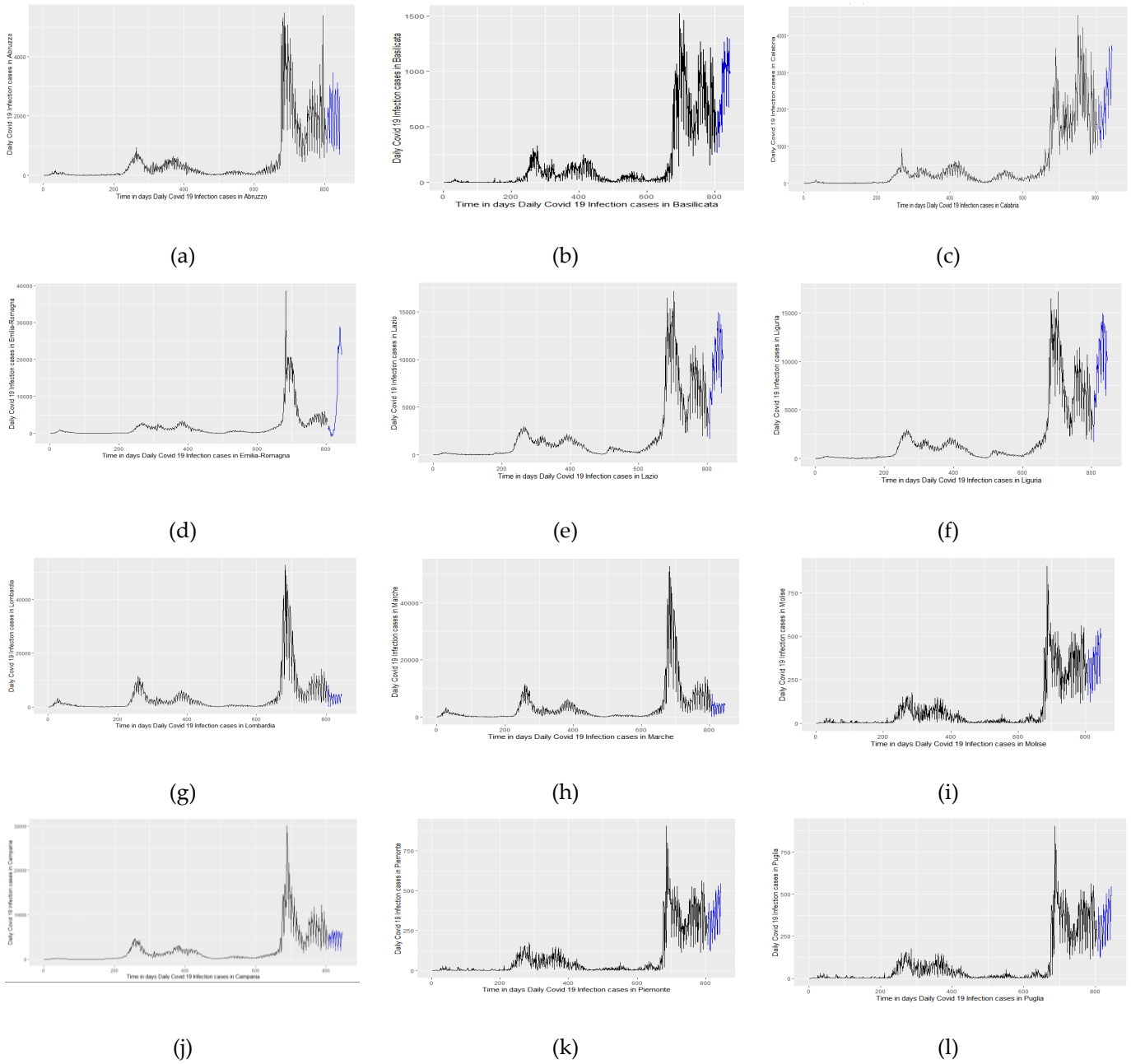

Figure 1. *Cont.*

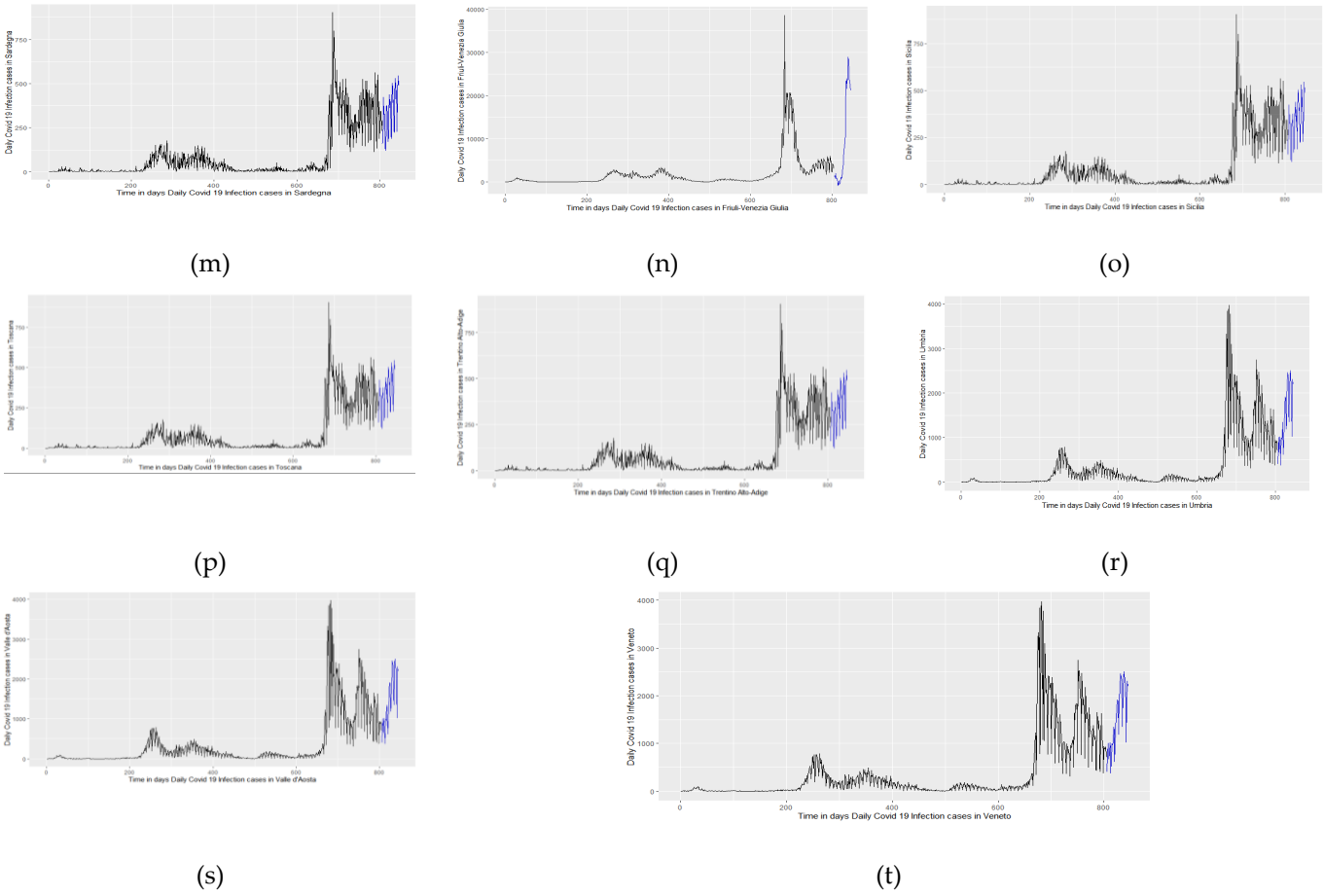

Figure 1. Daily COVID-19 Italian regions' new cases prediction with ANN model: (**a**) Abruzzo; (**b**) Basilicata; (**c**) Calabria; (**d**) Emilia-Romagna,(**e**) Lazio; (**f**) Liguria; (**g**) Lombardia; (**h**) Marche; (**i**) Molise; (**j**) Campania; (**k**) Piemonte; (**l**) Puglia; (**m**) Sardegna; (**n**) Friuli-Venezia Giulia; (**o**) Sicilia; (**p**) Toscana; (**q**)Trentino Alto-Adige; (**r**) Umbria; (**s**) Valle D'Aosta; (**t**) Veneto.

Figure 1 shows the trend of the number of new cases predicted by the ANN model for each region, obtained considering 28 lags as inputs and 5 nodes in the hidden layer. From the results obtained by the predictions of NARNN model, we can say that this model's predictions of the new COVID-19 confirmed cases are closer to the observed time series values. This is also emphasized by the value of MAPE for the test set, which is much lower than other forecasting models' MAPE values. According to the ANN (28-5-1) model, there will be an increasing trend in the number of new COVID-19 infections by the end of May, until 16 June in the following regions: Abruzzo, Basilicata, Calabria, Lazio, Liguria, Molise, Piemonte, Trentino Alto-Adige, and Veneto, while for the rest of them the ANN model predicted a constant to decreasing trend for the next 30 days.

4. Discussion

In this work, we evaluated four different time series forecasting models for predicting daily Italian regions' COVID-19 confirmed new cases. Using various models let us compare their forecasting accuracy and make an optimal selection. For our time series, the ANN model was preferred over the other linear forecasting models. It was chosen based on MAPE value, as it had the lowest value among all the forecasting models. The ANN (28-5) model gives better results in all the considered indicators with a considerable difference from the indicators of the other linear models. It predicted an increase in the number of new COVID-19 infections by the end of May 2022, in almost all the Italian regions. The results are valid for a short period of time because in the long run they can be influenced

by other factors such as vaccination, immunization of the population, and measures taken by government authorities to limit the spread of the infection, etc.

The above-considered models can be implemented on new data as they become available for possible future COVID-19 new confirmed cases forecasting, in order to improve forecasting accuracy, maybe taking into consideration other patients' parameters as possible inputs for the ANN model, since additional data would improve forecasting performance. Predictions about possible future new cases would be very helpful for the allocation of medical resources, handling the spread of the pandemic, and getting more prepared in terms of health care systems. People that deal with decision-making could find it very helpful for future projections regarding intervention for reducing and controlling the spread of the infection.

Author Contributions: Conceptualization, E.N. and L.S.; methodology, E.N.; software, L.S.; validation E.N.; formal analysis, E.N. and L.S.; investigation, L.S.; resources, L.S.; data curation, E.N.; writing—original draft preparation, L.S.; writing—review and editing, E.N.; visualization, E.N.; supervision, E.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available at: https://www.iss.it/ (accessed on 5 May 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Batista, M. Estimation of the final size of the COVID-19 epidemic. *MedRxiv* 2020, *Preprint*. [CrossRef]
2. Safi, S.K.; Sanusi, O.I. A hybrid of artificial neural network, exponential smoothing, and ARIMA models for COVID-19 time series forecasting. *Model Assist. Stat. Appl.* **2021**, *16*, 25–35. [CrossRef]
3. Khan, F.M.; Gupta, R. ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. *J. Saf. Sci. Resil.* **2020**, *1*, 12–18. [CrossRef]
4. Abotaleb, M.; Makarovskikh, T. System for Forecasting COVID-19 Cases Using Time-Series and Neural Networks Models. *Eng. Proc.* **2021**, *5*, 46. [CrossRef]
5. Gecili, E.; Ziady, A.; Szczesniak, R.D. Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy. *PLoS ONE* **2021**, *16*, e0244173. [CrossRef] [PubMed]
6. Abotaleb, M.S.A. Predicting COVID-19 Cases using Some Statistical Models: An Application to the Cases Reported in China Italy and USA. *Acad. J. Appl. Math. Sci.* **2020**, *6*, 32–40. [CrossRef]
7. Reilly, D.L.; Cooper, L.N. An overview of Neural Networks: Early models to real world systems. In *How We Learn; How We Remember: Toward an Understanding of Brain and Neural Systems*; World Scientific Series in 20th Century Physics; World Scientific Publishers: Singapore, 1990; Volume 10, pp. 300–321. [CrossRef]
8. Zhang, G.; Patuwo, B.E.; Hu, M.Y. Forecasting with artificial neural networks:: The state of the art. *Int. J. Forecast.* **1998**, *14*, 35–62. [CrossRef]