Society for Mathematical Biology

ORIGINAL ARTICLE

Comparing Short-Term Univariate and Multivariate Time-Series Forecasting Models in Infectious Disease Outbreak

Daniel Bouzon Nagem Assad[1,2] (ORCID) **· Javier Cara**[1] **· Miguel Ortega-Mier**[1]

Abstract

Predicting infectious disease outbreak impacts on population, healthcare resources and economics and has received a special academic focus during coronavirus (COVID-19) pandemic. Focus on human disease outbreak prediction techniques in current literature, Marques et al. (Predictive models for decision support in the COVID-19 crisis. Springer, Switzerland, 2021) state that there are four main methods to address forecasting problem: compartmental models, classic statistical models, space-state models and machine learning models. We adopt their framework to compare our research with previous works. Besides being divided by methods, forecasting problems can also be divided by the number of variables that are considered to make predictions. Considering this number of variables, forecasting problems can be classified as univariate, causal and multivariate models. Multivariate approaches have been applied in less than 10% of research found. This research is the first attempt to evaluate, over real time-series data of 3 different countries with univariate and multivariate methods to provide a short-term prediction. In literature we found no research with that scope and aim. A comparison of univariate and multivariate methods has been conducted and we concluded that besides the strong potential of multivariate methods, in our research univariate models presented best results in almost all regions' predictions.

Keywords Outbreak disease · COVID-19 · Forecasting · Univariate approach · Multivariate approach

Daniel Bouzon Nagem Assad, Javier Cara and Miguel Ortega-Mier contributed equally to this work.

✉ Daniel Bouzon Nagem Assad
daniel.nagem.bouzon@alumnos.upm.es

1 Universidad Politécnica de Madrid, Department of Organization Engineering, Business Administration and Statistics, Escuela Técnica Superior de Ingenieros Industriales, José Gutiérrez Abascal, 2, 28006 Madrid, Spain

2 Universidade do Estado do Rio de Janeiro, Rua São Francisco Xavier, 524, Maracanã, 20550-900 Rio de Janeiro, Brazil

1 Introduction

Infectious diseases can rapidly spread because it caused by breathing in an airborne virus, insect bite, sexual intercourse, skin contact by patient who is already suffering with that disease (Kaur et al. 2020).

When the disease spread runs out of control and infect a community or region with specified health behavior, or other health-related events more than normal expectancy it is called epidemic (Porta 2014). The term pandemic is commonly taken to refer to a widespread epidemic of contagious disease throughout a whole country or one or more continents at the same time (Honigsbaum 2009).

Although personal measures should be taken to avoid the infection and therefore their spread, for instance not to share personal things, to clean hands properly, to always take good and safe food, to get vaccinated or to cover month when sneezing or coughing (Kaur et al. 2020), health systems and governments of all countries must be able to develop and improve nonpharmacological measures like animal source containment, early detection and diagnosis, rigorous infection control, timely case report and rapid information dissemination, quarantines, mask obligation, lockdown and pharmacological measures like vaccine development (Yang et al. 2020).

Over the last few decades, mathematical models applied over infectious diseases growth have been helpful to gain insights into the transmission dynamics (Chowell et al. 2016) allowing scientists to forecast new cases and deaths as well as evaluate the interventions' impact (Metcalf and Lessler 2017).

Although still showing numerous limitations and pitfalls often driven by data scarcity and delay, Smirnova and Chowell (2017) state that the integration of mathematical models' prediction results with public health practice has the potential to increase the timeliness and quality of health care unit responses.

In addition, Chen's et al. (2021) research investigates the temporal and spatial distribution characteristics of the COVID-19 outbreak in China such as the influence of different meteorological factors, the proportion of the population flow entered from Wuhan on other regions and the effects of nonpharmaceutical interventions. As a result of dealing with different factors, the authors were able to predict the number of infected cases under different controlling scenarios and conditions.

In this context, during COVID-19 pandemic, Marques et al. (2021) applied four univariate forecasting approaches using real COVID-19 data from 5 countries. These approaches are classical statistical models, compartmental models, state-space model and machine learning models and will be presented in Sect. 2.

After evaluating and comparing 66 previous works (see Table 15), we conclude that less than 10% of previous research applied multivariate techniques and none of them used more than one country or region. Thus, this *research contributes* to forecasting methods application over human infections diseases outbreaks by being the first attempt to evaluate, over real time-series data:

- Of three different countries (Brazil, Italy and USA);
- Using six univariate and two multivariate methods;
- Providing a short-term prediction of 28 days ahead which is two or four times longer than similar previous research.

In Sect. 3, we present the all time series evaluated in this research and their features, how we choose the data range for all time series and how we split these data in data training and data test.

In Sect. 4, we detailed explain all forecasting methods used and how the error criterion was selected. Thereafter, in Sect. 5 we applied these methods for all time series, specify how the results are obtained and compared, choose the best model for each time-series and make a short-term prediction of 28 days.

Finally, in Sect. 6 we present research's conclusions, address limitations and make proposals for further research.

2 Theoretical Background

Epidemics or pandemics disease outbreak have been devastating populations worldwide all over the years (Hays 2005; White 2006) and Kaur et al. (2020). From Athens epidemics ("Plague of Athens") in 430–427 B.C (see Hays 2005 for more details) to coronavirus (SARS-CoV 2) also known as COVID-19 on going pandemic, the civilizations have lived with epidemics or pandemics caused mainly by virus and bacteria.

Kaur et al. (2020) summarized the most relevant disease outbreaks in human history like blackdeath (black plague), cholera, malaria and influenzas virus (Spanish, Hong Kong and Russian Flu). In addition, Hays (2005), White (2006) and Yamey et al. (2017) point out many others like the smallpox, blackdeath (black plague), cholera, influenza, HIV/AIDS, measles, dengue, Ebola, Zika virus.

Table 1 summarizes in a nonexhaustive list of worldwide human outbreaks diseases (epidemics or pandemics) by year, impact in number of deaths and where each one occurred.

Besides the number of human deaths caused by epidemics and pandemics, Kaur et al. (2020) state that it will not disappear in future if we do not find efficient ways to stop before spreading any disease to other population or countries.

Many authors use time-series approach (Chen et al. 2021; ArunKumar et al. 2021; Katris 2021; Benítez et al. 2020) to explain, evaluate and estimate further values (forecast) the behavior of some variable like outbreak disease cases, deaths, or transmission rate all over the time.

A time series is a set of data points arranged in time and its analysis intends to reveal reliable and meaningful statistics (Marques et al. 2021) that can be used to evaluate some patterns and forecast future values (Hyndman and Athanasopoulos 2018). It drew the attention of the scientific community when Yule introduced a general approach for time-series analysis in 1927 (Yule 1927).

In the same year, one deterministic compartimental model widely applied in epidemiology science was proposed by Kermack and McKendrick (1927), the susceptible–infectious–removed (SIR) model.

Almost three decades later (1950s) classical time-series statistical models started to appear (Holt 1957; Brown 1959; Winters 1960; Box and Jenkins 1970) as well as machine learning (Samuel 1959) and space-state model (Kalman 1960).

Bring forecasting methods to human infectious disease outbreak context, Chretien et al. (2014) proposed a framework to classify research as follows: Population-based

Table 1 Majority of worldwide epidemics or pandemic. Source: The authors adapted from Yamey et al. (2017), Yang et al. (2020), Kaur et al. (2020)

Outbreak disease	Year	Impact	Countries
Athens Plague	430–427 B.C	75,000–100,000 deaths (25% of population)	Greece
Antonine Plague	165–180	5 million deaths	Asia Minor, Egypt, Greece, Italy
Justinian Plague	541–542	35 million deaths	Eastern roman empire
Black Plague	1346–1353	50 million deaths (60% of Europe population). 75–200 million deaths worldwide	Europe, Africa and Asia
Fifth Cholera	1881–1896	981,899 deaths	Asia, Africa, France, Germany, Russia and South America
Russian Flu	1889–1890	1 million deaths	Russia, Canada and Greenland
Modern Plague	1894–1903	10 million deaths	India and China mostly
Sixth Cholera	1899–1923	1.5 million deaths	North Africa, Middle East, India, Eastern Europe, Russia and America
Spanish Flu	1918–1919	20–50 million deaths	Australia, Canada, UK, USA, France and Spain
Asian Flu	1956–1958	2 million deaths	China, Singapore, USA and Hong Kong
Hong Kong Flu	1968–1969	1 million deaths	Hong Kong and spread Singapore, Vietnam, USA, India, Australia and Europe
HIV/AIDS	1980-present	39 million deaths	Worldwide
Severe Acute Respiratory Syndrome (SARS CoV)	2002–2003	8422 people infected in 32 countries and 919 (11%) died	China, Vietnam, Singapore, Canada and others
Swine Flu	2009–2010	150,000–750,000 deaths	Mexico, USA, Africa and Southeast Asia
Middle East respiratory coronavirus (MERS-CoV)	2012	858	Saudi Arabia
Ebola	2013	10,600 deaths	Guinea, Liberia, Sierra Leone
Zika virus	2015–2016	20 deaths	Latin America and Caribbean
Coronavirus SARS CoV 2 (COVID-19)	2019-present (until Dec 22, 2021)	4.94 million deaths	Worldwide

forecasting studies (seasonal or pandemic), forecast type (temporal or spatial–temporal) and forecasting method (mechanistic, Statistical).

To the same authors, the forecasting method were divided into compartmental model, regression tree, generalized linear model, agent-based model, survival analysis, Bayesian network and time series model.

Focusing on forecasting method, during ongoing COVID-19 pandemic, Marques et al. (2021) presented four different univariate approaches for epidemiological time-series prediction, which will be able to provide support for Governments and Healthcare decision-makers. They worked with five countries real data: China, USA, Brazil, Italy and Singapore.

In this research, we adopt the framework proposed by Marques et al. (2021) that divided epidemiological time-series prediction in: classical statistical models (Sect. 2.1), compartmental models (Sect. 2.2), state-space models (Sect. 2.3) and machine learning models (Sect. 2.4).

In the following sections, we do not aim to present a exhaustive list of forecasting methods, but we present all methods applied over human disease outbreak prediction (summarized in "Appendix A," Table 15). These methods were obtained after an extensive literature review which steps are presented in "Appendix D."

2.1 Classical Statistical Models (CSM)

In this section, we present CSM methods found in literature that are divided into:

- Exponential smoothing (ES) or their generalization error, trend and seasonal (ETS);
- Autoregressive integrated moving average (ARIMA);
- Vector autoregressive (VAR);
- Vector error correction (VEC);
- Vector autoregressive moving average (VARMA).

ES was proposed in the late 1950s (Holt 1957; Brown 1959; Winters 1960), and has motivated some of the most successful forecasting methods.

ARIMA was introduced by Box and Jenkins (1970) in the 1970 and takes into consideration changing disturbances in time and tendencies.

Hyndman and Athanasopoulos (2018) state that ES, or their generalization ETS, and ARIMA models are the two most widely used approaches to time-series forecasting and provide complementary approaches to the problem. While ES models are based on a description of the trend and seasonality in the data, ARIMA models aim to describe the autocorrelations in the data.

VAR, VEC and VARMA are the most used models the prediction of multivariate time series in econometric research. But these models can also be applied to predict human disease outbreaks (for more details, see Wu et al. 2018; Khan et al. 2020).

For instance, Kiang et al. (2021), ArunKumar et al. (2021), Talkhi et al. (2021), Katris (2021), Khan et al. (2020), Bomfim et al. (2020), Liang et al. (2020), Ramos et al. (2020), Zhang et al. (2019), Wang et al. (2019), Li et al. (2019), Choi et al. (2019), Chakraborty et al. (2019), Chumachenko et al. (2019), Haddawy et al. (2018), Wu et al. (2018), Wu et al. (2018) Zhao et al. (2018), Jerónimo-Martínez et al. (2017), Ray et al.

(2017), Anggraeni and Aristiani (2016), Ke et al. (2016), Li et al. (2016), Johansson et al. (2016), Pradhan et al. (2016), Wu et al. (2015) Mekparyup and Saithanu (2015), Kane et al. (2014), Feng et al. (2014), Soebiyanto et al. (2010), Shen et al. (2008), Medina et al. (2007), Burkom et al. (2007) and Nobre et al. (2001) research applied these models to several human disease outbreaks like COVID-19, Ebola, Zika virus, dengue hemorraric fever (DHF), scarlet fever (SF), tuberculosis, malaria, leprosy, hemorragic fever with renal syndrome (HFRS), hand, foot and mouth disease (HFMD), HIV/AIDS, tuberculosis, malaria, influenza-like illness (ILI) and others acute respiratory infection (ARI). Predicted variables (daily cases, reproduction number, among others), prediction range and other methods applied for each mentioned research are summarized in Table 15.

2.2 Compartmental Models (CM)

In this section, we present CM methods found in literature that are divided into:

- Susceptible–infectious–removed (SIR);
- Susceptible-exposed-infectious-removed (SEIR);
- Susceptible-infectious-susceptible (SIS);
- Cellular automation (CA);
- Growth models (GM).

One deterministic model widely considered in epidemiology is the SIR model, which is based on the classification of the individuals into three stages of infection and was introduced almost one hundred years ago by Kermack and McKendrick (1927).

All over the years SIR model was improved, and other stages were added (Krause et al. 2018), for instance: SEIR with or without intervention and SIS among others.

Considering single variables, GM like Richards (GMR), Gompertz (GMG), Logistic (GML) and Cellular Automation (CA) are widely used (Gerardi and Monteiro 2011) to describe and predict infectious diseases spread cases and deaths.

Research such as Chen et al. (2021), Katris (2021), Paul et al. (2021), Benítez et al. (2020), Wang et al. (2020), Smirnova et al. (2019), Eilertson et al. (2019), Suparit et al. (2018), Basile et al. (2018), Li et al. (2018), Valeri et al. (2016), Yang et al. (2014), Wang et al. (2013), Towers and Chowell (2012), Aguiar et al. (2011), Gerardi and Monteiro (2011), Laneri et al. (2010), Santos et al. (2009), Finkenstädt et al. (2005) and Gamerman and Migon (1991) applied these models to several human disease outbreaks like COVID-19, Measles, ILI, dengue, DHF, Skin and Soft Tissue Infections (SSTIS). Predicted variables (daily cases, reproduction number, among others), prediction range and other methods applied for each mentioned research are summarized in Table 15.

2.3 State-Space Models (SSM)

In this section, we present SSM methods found in literature that are divided into:

- Hidden Markov Model (HMM);
- Monte Carlo Markov Chain (MCMC);

- Kalman filter (KF);
- Exponential smoothing state-space model with Trigonometric, Box-Cox transformation, ARMA errors, trend and seasonal components (TBATS).

A SSM, also known in the technical literature as HMM, can be defined as a class of probabilistic models that describes the dependence between a latent state variable and an observed measurement (Koller and Friedman 2009). The term "state space" originated in control engineering subject (Kalman 1960). HMM can also be combined with simulation approach like Monte Carlo. It is called, according to Wang et al. (2013), MCMC.

SSM is a general framework for ES, ARMA and Trend and Seasonal component where TBATS, according to Talkhi et al. (2021), is widely applied to univariate time series.

The KF is a state-space model provides estimates of the unknown variables given the measurements observed over time using only the previous estimate for calculation which reduces the need for saving the whole data from previous iterations (Haykin 2004).

Research like Talkhi et al. (2021), Han et al. (2021), Benítez et al. (2020), Eilertson et al. (2019), Yang et al. (2014), Wang et al. (2013), Nunes et al. (2013), Mode et al. (1991) applied these models to several human disease outbreaks like COVID-19, Ebola, Zika virus, ILI, SSTIS, HIV/AIDS. Predicted variables (daily cases, reproduction number, among others), prediction range and other methods applied for each mentioned research are summarized in Table 15.

2.4 Machine Learning Models (MLM)

In this section, we present MLM methods found in literature that are divided into:

- Multilayer perceptron (MLP);
- Artificial recurrent neural network (RNN);
- Long short-term memory (LSTM);
- Convolutional neural network (CNN);
- Feed-forward neural networks with a single hidden layer and lagged inputs (NNETAR) that is also divided into neural network autoregressive (NNAR) and nonlinear auto-regressive neural network (NARNN);
- Extreme learning machine algorithm (ELM);
- Automated machine learning (AutoML);
- Ensemble empirical mode decomposition (EEMD);
- Cross-location attention graph neural network (CLAGNN);
- Support vector machine (SVM);
- Bayesian model averaging (BMA);
- Kernel conditional density estimation (KCDE);
- Kernel ridge regression Gausian process network (KRRGPN);
- Neural fuzzy inference system (NFIS);
- Random forest (RF);
- Generalized regression neural network (GRNN);
- Genetic algorithm (GA);

- Wavelet neural network (WNN).

First defined by computer scientists at the Dartmouth Conferences in 1956, artificial intelligence (AI) field draws upon computer science, mathematics, psychology, linguistics, neuroscience and many others (Ongsulee 2017).

Evolved from the study of pattern recognition and computational learning theory in AI, machine learning (ML) appears in 1959 (Samuel 1959) to explore the study and build algorithms that can learn from and make predictions on data (Kohavi 1998). Deep learning (DL) and neuro networks (NN) are subfields of ML (for more details, see Alzubi et al. 2018; Ongsulee 2017).

Some ML algorithms widely applied to forecast are MLP, RNN like LSTM with or without false nearest neighbors (FNN), CNN, NNETAR, ELM, AutoML, EEMD, CLAGNN, SVM, BMA, NARNN, NNAR, KCDE, KRRGPN, NFIS, RF, GRNN, GA and WNN.

All algorithms mentioned above were applied, for example, in ArunKumar et al. (2021), Talkhi et al. (2021), Katris (2021), Han et al. (2021), Ribeiro et al. (2020), Deng et al. (2020), Wang et al. (2020), Bomfim et al. (2020), Liang et al. (2020), Zhang et al. (2019), Wang et al. (2019), Choi et al. (2019), Chakraborty et al. (2019), Stolerman et al. (2019), Wu et al. (2018), Ray et al. (2017), Caicedo-Torres et al. (2017), Nguyen et al. (2017), Chau and Ngoc Anh (2016), Wu et al. (2015), Kane et al. (2014), Gerardi and Monteiro (2011) and Peng et al. (2008) research. Predicted variables (daily cases, reproduction number, among others), prediction range and other methods applied for each mentioned research are summarized in Table 15.

2.5 Research Synthesis

Table 15 summarizes forecast research applied in human infectious disease outbreak (pandemic or epidemic) context considering the "general methods" pointed out by Marques et al. (2021) the results. Only research that in fact make predictions were considered.

Columns 3 to 7 address the approach used in each research mentioned in last section. In columns 8 to 9 is presented the range of time windows used as well as the prediction range. Time windows found on previous research were day (d), week (w), month (m) or year (y). The prediction range are expressed in time windows mentioned, but some models proposed to forecast the whole pandemic period (wpp).

The variable measured/evaluated and forecasted in each research (column 10) on those time windows can be number of patient cases (ca), deaths (de) and recovered (re), admitted and discharged from hospital or intensive care unit (adhosp and dishosp) and transmission rate (rt). Excluding transmission rate, all measures mentioned can be counted in two different ways: by time window or cumulative. For example, daily cases (dca), monthly deaths (mde), yearly patients admitted in hospital (yadhosp), cumulative cases (Cca), cumulative patients discharged from hospital (Cdishosp).

Columns 11 to 13 presents in which countries each research applied the methods specified in columns 3 to 7, the type of forecast approach divided into univariate (Uni), causal or multivariate (Mul) and the disease outbreak studied.

After comparing sixty six works in Table 15 we can conclude that:

- Approach: only 7 (10.61%) research apply multivariate methods to make predictions. 23 (34.84%) apply causal methods to make predictions combining number of specific queries all over the time on web search engines (Google, Baidu index), climate variables (temperature, air pollution, rainfall) or another seasonality infectious diseases. 36 (54.55%) apply only univariate methods to make predictions. The number of publications by approach is presented in Fig. 1.
- Disease outbreaks: 14 research forecasted Dengue, Chikungunya or DHF, fourteen ILI, nine COVID-19, six HFRS, five Malaria, four Measles three Ebola and three HIV/AIDS.
- Countries: 15 research were applied in disease outbreak in China, nine in Brazil, nine in USA, 4 in Thailand and three in Japan. Only 4 research worked with African countries.
- Data range and prediction: only 4 research proposed to predict the whole pandemic period (wpp). Considering different time windows, twenty six research forecasted less or equal to six steps ahead.
- Time window: twenty eight research worked with monthly cases and twenty four with weekly cases.
- Variables: number of patient cases was studied in 63 research (95.45%) while deaths, recovered and hospital admission or discharge or transmission rate are not much explored (deaths appears in second place with only five research).
- Epidemiological time-series prediction: 34 research applied CSM, twenty three applied MLM, twenty applied CM and only eight applied SSM. We found no research in which all approaches were applied. The number of publications by type of epidemiological time-series prediction is presented in Fig. 2. Only two research used three univariate approaches (Talkhi et al. 2021; Katris 2021) in a single country (Iran and Greece, respectively), sixteen research used two approaches, forty five research used only one approach and three reseach used approaches not mentioned by Marques et al. (2021).
- On twenty CM models only five worked with growth models and basically applied three models: Richards (GMR), Gompertz (GMG), Logistic (GML). But we point out that there is other fourteen GM models (Fekedulegn et al. 1999; Kaps et al. 2000; Tsoularis and Wallace 2002; Khamis 2005) that were not explored in human disease outbreak context.

Although the current review shows benefits of using CM models including to provide mid- and long-term predictions and mostly uses susceptible–infectious–removed models or their variations, many assumptions over them must be made before obtaining all parameters (Smirnova et al. 2019), results of all stages and then a prediction of a whole pandemic period.

In addition, the real-time data COVID-19 showed us that new stages are necessary to be considered like immunity period and rate of reinfection, vaccination, period of strong nonpharmacological measures (quarantine and lockdown), among others.

The current research is the first attempt to evaluate over real data of three different countries (Brazil, Italy and USA) using three univariate approaches (CSM, SSM and MLM) proposed by Marques et al. (2021). We apply the same univariate methods proposed in Talkhi et al. (2021) and add KF.
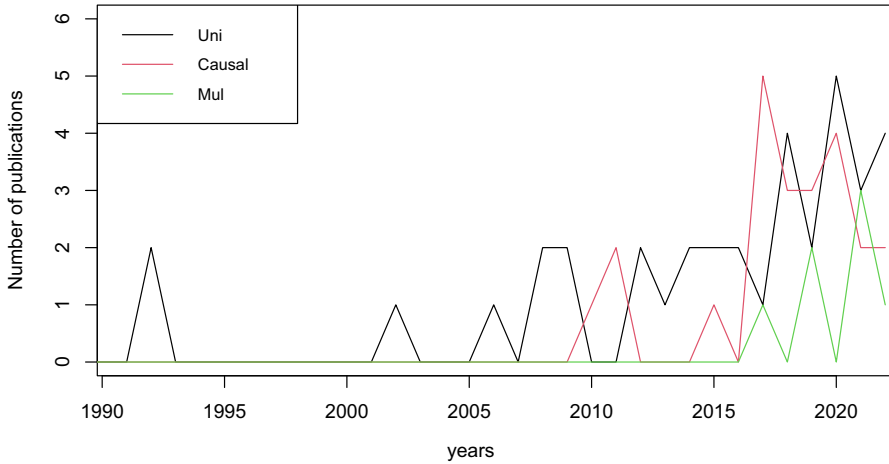
Fig. 1 (Color figure online) Research publication by approach by year Source: The authors
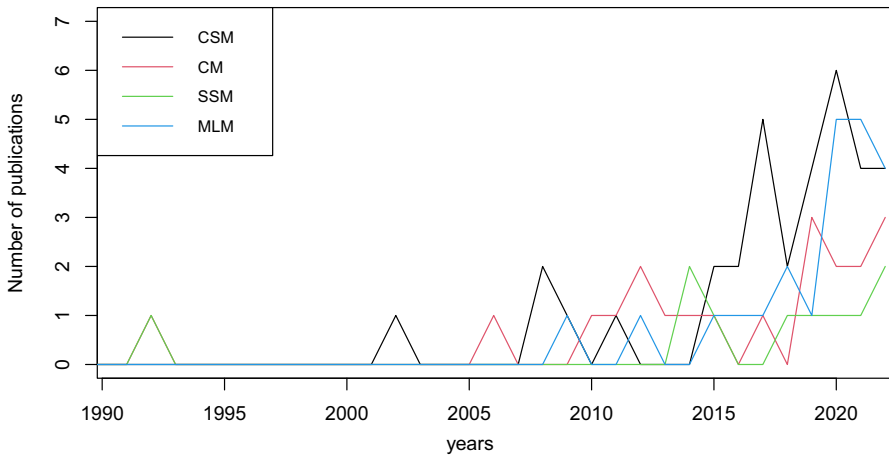

Fig. 2 (Color figure online) Research publication by model by year. Source: The authors

Finally, we apply two multivariate approaches and compare their results with pre-vious mentioned univariate methods to find which approach can better fit real data and give us a reliable short-term prediction to each region/country.

3 Data Sets Selection and Problem Statement

In this research, we work with real COVID-19 data of Rio de Janeiro (RJ) (Assad 2022) city health regions in Brazil, Italy (IT) regions (Krispin 2021) and US states (Dobbyn 2020). All time series used are presented in figures below.

We select these data sets because Brazil, Italy and US population were highly affected by COVID-19 pandemic and adopted different rules to fight against COVID-19 dissemination.

After first COVID-19 wave starts to spread, Italian government establish hard common measures for all country regions. Could we expect that the number of daily new cases from one region helps to explain and predict this number in other region or only the past data of the same region is better enough?

In other words, given common government rules which approach best fits and predict daily number of cases: univariate or multivariate methods?

Regional divisions of each Italian time series used in this research are quickly presented in Table 2. For more details, see supplementary material.

In USA, each state has autonomy to establish measures as long as they consider necessary to fight against COVID-19 dissemination. As a result many states adopted different measures, but the question proposed in Italy remains: Could we expect that the number of daily new cases from one State helps to explain and predict this number in other State or only the past data of the same region is enough?

USA has 51 states and work with this number of time series would be useless and time consuming considering the scope of this research. Thus, we choose the state with highest number of positive cases (California) and its surrounding states (Oregon, Nevada, Arizona) also presented in Table 2.

Closer to US policy, in Brazil each state were in charge of defining necessary measures to avoid COVID-19 dissemination. Here we bring the RJ city health regions time series with the same question, but we want to evaluate if the lower distance betwwen these health regions (comparing to US states and IT regions distances) could bring us a different result comparing to IT and US time series. RJ health regions are also presented in Table 2.

In Figs. 3, 4 and 5, we can see that COVID-19 pandemic started at different dates (presented below) and the data set range of each country or region can also vary according to data set source. To establish comparisons between forecasting techniques, we work with the same time-series range to all regions. Thus, in this research we work with time-series range of 369 days.

- Rio de Janeiro city time-series range available: from January 13, 2020, to December 22, 2021, but we decide to start on March 12, 2020 (when cases start to appear in every day). 651 days;
- Italy regions time-series range available: from February 24, 2020, to July 27, 2021. 520 days;
- US time-series range available: from March 4, 2020, to July 3, 2021. 369 days.

In this research, we evaluate all mentioned time series presented above using univariate and multivariate approach. Apply multivariate approach can potentially provide us reliable predictions given the high correlation that each time series has with the others in the same region at the same lag (correlation) and at different lags (autocorrelation) as we can see in Fig. 6. All correlation plots are available in section C.

All data sets are divided into training and test data and their lengths are 341 and 28 days, respectively. Our short-term forecasting is 28 days ahead. The reasons for choosing forecasting range of 28 days are presented below.
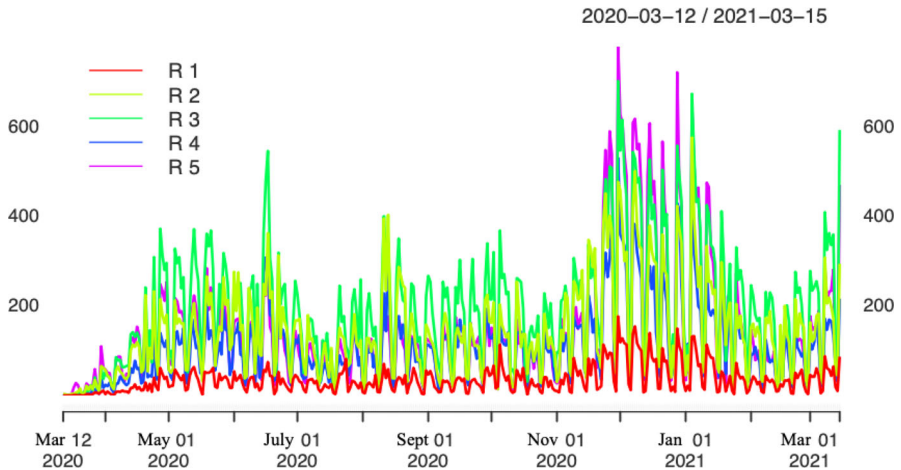
Fig. 3 (Color figure online) Rio de Janeiro city COVID-19 daily cases per health region. Source: The authors
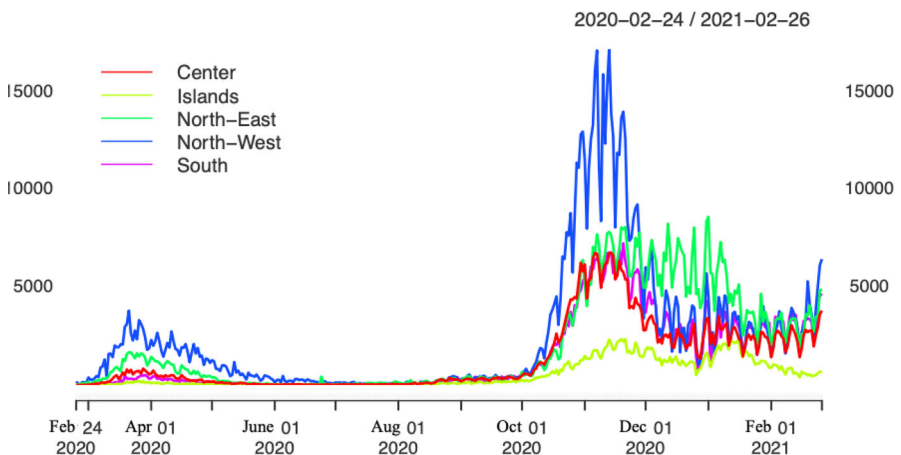
Fig. 4 (Color figure online) Italy regions COVID-19 data per region. Source: The authors

- We work with 341 past observations which is more than 10 times the prediction length. It does not mean that our past data are larger enough to well train some models and give reliable predictions;
- Forecasting daily new cases four weeks ahead allow decision makers in health departments to better plain resource availability or governments to choose adequate measures. At least, better than in most of the previous research daily predictions worked with shorter forecasting range (seven or fourteen days);
- Considering that resource availability depends on health departments resource allocation (doctors, beds, among others), the smaller time window unit (until this moment is daily data) we are able to work with and provide reliable predictions

Fig. 5 (Color figure online) US states COVID-19 data per region. Source: The authors
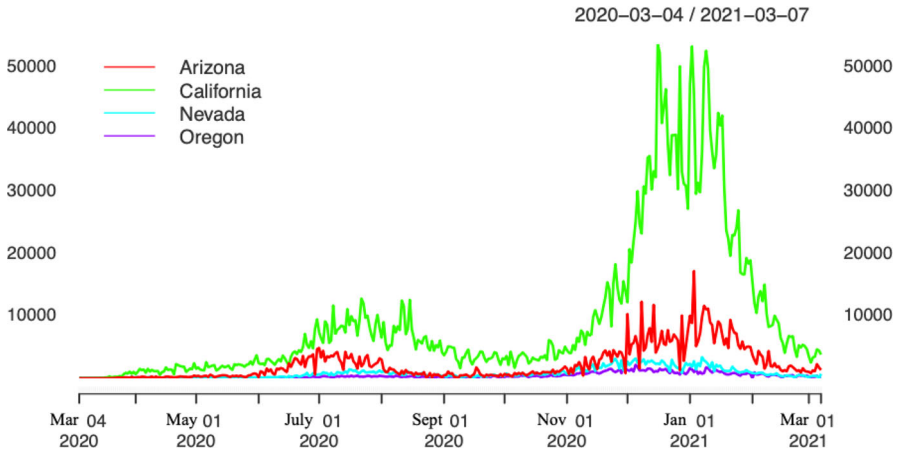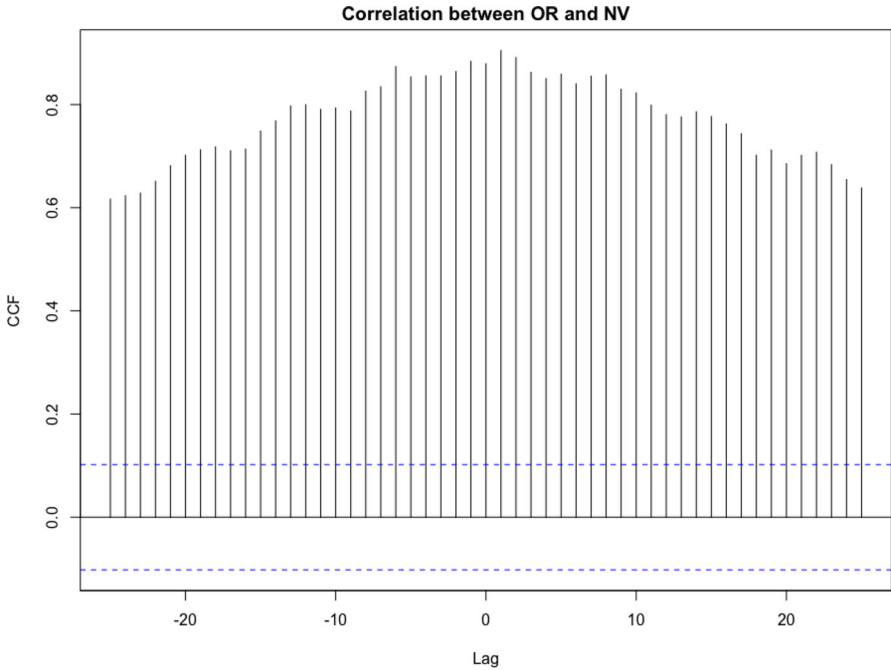

Fig. 6 (Color figure online) Auto-correlation plot between NV and OR. Source: The authors

will result in the more useful information to help decision makers to meet resources requirements that ensure an adequately treatment to patient demand.

Table 2 Regions or states division (for more details see "Appendix B")

RJ (Fig. 12)	IT (Fig. 13)	US (Fig. 14)
Region (R) 1	Center (CEN)	Arizona (AZ)
R 2	Islands (ISL)	California (CA)
R 3	North-West (NOW)	Nevada (NV)
R 4	North-East (NOE)	Oregon (OR)
R 5	South (SOT)	–

Source: The authors

Table 3 Forecasting models applied in this research

Models	Univariate	Multivariate
CSM	ES, ARIMA	VAR
CM	–	–
SSM	TBATS, SSM Univariate (SSM-U)	SSM Multivariate (SSM-M)
MLM	MLP, NNETAR	–

Source: The authors

4 Forecasting Models Applied

In this research, we expand the framework proposed by Marques et al. (2021) by using more univariate approaches and adding multivariate approach. Models applied in next sections are presented in Table 3.

To build each model is necessary estimate some times more than 10 parameters and present all of them in 1 or more table is not the aim of this research. Thus, we explain in Sect. 4.1 the main features of each model.

4.1 Applied Models Description

In this section, we provide a detailed explanation of forecasting methods summarized in Table 3.

- ES: ETS is a class of models that essentially works with 2 components equations trend and season that can be added or multiplied to the remainder. In each model these components can not be significant, also known as none (N) or can be significant and better describe original time-series features as follows: additive (A) or additive damped (Ad) or multiplicative (M). This class of models can be combined in 18 different ways (A, N, A; M, Ad, M; for instance). Equations of each model are presented in Fig. 7. For more details, see Hyndman and Athanasopoulos (2018);
- ARIMA: ARIMA or seasonal ARIMA (SARIMA) is a class of models that combine autoregressive (AR) and moving average (MA) with differenced values. The AR part of ARIMA (p) shows that the time series is regressed on its own past data. The MA part of ARIMA (q) indicates that the forecast error is a linear combination of past respective errors. The I part of ARIMA (d) shows that the data values

have been replaced with differenced values of d order to obtain stationary data, which is the requirement of the ARIMA model approach (Kotu and Deshpande 2019). When we work with SARIMA the same components appears lagged by the length of seasonal time window (frequency) as P, D and Q. For instance, ARIMA $(p = 5, d = 0, q = 3)(P = 0, D = 1, Q = 1)$ [*frequency* = 7]. For more details, see Hyndman and Athanasopoulos (2018) and Kotu and Deshpande (2019);

- Space-state model univariate (SSM-U): The state of a deterministic dynamic system is the smallest vector that summarises the past of the system in full (Haykin 2004). The linearity of state dynamics and observation process and the normal distribution of noise in state dynamics and measurements are the assumptions of SSM. A linear autoregressive equation $x(t) = A * x(t) + W(t)$ where $W(t) \approx N(0, Q)$ with a measurement that is $y(t) = C * y(t) + V(t)$ where $V(t) \approx N(0, R)$, define the linearized process in which $y(t) \in \mathbb{R}$. The random variables $W(t)$ and $V(t)$ represent the process and measurement noise, respectively, and are assumed to be independent of each other and with normal distributions. In our case we will work with a vector length ($n = 2$ for linear model and $n = 3$ for order 2 polynomial model) which means a $n * n$ dimensions A matrix. We select the best approach for each time series based on the lowest Akaike information criterion (AIC) criteria.

- MLP: MLP is a supplement of feed-forward neural network. It consists of three types of layers—the input layer, output layer and hidden layer. The input layer receives the input signal to be processed. An arbitrary number of hidden layers that are placed in between the input and output layer are the true computational engine of the MLP. Similar to a feed-forward network in a MLP the data flows in the forward direction from input to output layer. The neurons in the MLP are trained with the backpropagation learning algorithm. MLPs are designed to approximate any continuous function and can solve problems which are not linearly separable. In time-series problem the input layer is past observations and we set then to choose between 1 and 28 (prediction length) according to Mean Square Error the optimal number of lags used and which lags will be used. The same criteria were used to define number of hidden nodes in each hidden layer;

- NNETAR: NNETAR is a feed-forward neural networks with a single hidden layer and lagged inputs. This model works with 2 (for nonseasonal time-series) or 3 (for seasonal time-series) parameters: the number of past observations used as input layers (p), the number of past observations lagged by the length of seasonal time window used as input layers (P) and the number of neurons (k) in the single layer. For instance, ($p = 21$, $P = 1$, $k = 11$)[7]. For more details, see Hyndman and Athanasopoulos (2018);

- TBATS: BATS model is Exponential Smoothing Method + Box-Cox Transformation + ARMA model for residuals. Aiming to reduced the parameters of model when the frequencies of seasonalities are high and giving more flexibility to deal with complex seasonality, De Livera et al. (2011) propsed TBATS model which is BATS model + Trigonometric Seasonal. Equations of the TBATS model are presented in equations below where ω and ϕ are Box-Cox and the damping parameters, respectively, ARMA(p, q) process model the error and m_1 to m_J list the seasonal periods used (in our case there is only m_1 always equal to 7) while k_1 to k_J are the corresponding number of Fourier terms used (in our case there is only k_1). For

instance, TBATS ($\omega = 0.21$, $[p = 0, q = 0]$, $\phi = 0.96$, $[\langle m_1 = 7, k_1 = 3 \rangle]$).

$$y_t^{(\omega)} = \frac{y_t^{(\omega)} - 1}{\omega}, \quad \omega \neq 0,$$

$$y_t^{(\omega)} = \log y_t, \quad \omega = 0,$$

$$y_t^{(\omega)} = l_{t-1} + \phi * b_{t-1} + \sum_{i=1}^{t} s_{t-m_i}^i + d_t,$$

$$l_t = l_{t-1} + \phi * b_{t-1} + \alpha * d_t,$$

$$b_t = (1 - \phi) * b_t + \phi * b_{t-1} + \beta * d_t,$$

$$s_t^i = s_{t-m_i}^i + \gamma_i * d_t,$$

$$d_t = \sum_{i=1}^{p} \phi_i * d_{t-i} + \sum_{i=1}^{q} \theta_i * \epsilon_{t-i} + \epsilon_t,$$

$$s_t^i = \sum_{j=1}^{k_j} s_{j,t}^i,$$

$$s_t^i = s_{j,t-1}^i * \cos \lambda_j^i + s_{j,t-1}^{*i} * \sin \lambda_j^i + \gamma_1^i * d_t,$$

$$s_t^{*i} = s_{j,t-1}^i * \sin \lambda_j^i + s_{j,t-1}^{*i} * \cos \lambda_j^i + \gamma_2^i * d_t$$

- VAR: A VAR(p) model is a generalization of the univariate autoregressive (AR) where (p) shows that the time series is regressed on past data of all time series for forecasting a vector of time series (Hyndman and Athanasopoulos 2018). Each variable has one equation that includes a constant and lags of all of the variables in the system.
- Space-state model multivariate (SSM-M): A SSM-M model is a generalization of SSM-U and works similarly, but with $y(t) \in \mathbb{R}^m$ where m is the number of time series considered.

4.2 Error Evaluation

To compare models results a error criterion must be selected, but choosing the right forecasting metric is not straightforward (Vandeput 2021) because each error criterion has shortcomings (for more details, see Shcherbakov et al. 2013).

For instance, Vandeput (2021) states that although the mean absolute percentage error (MAPE) is one of the most used KPIs to measure forecast accuracy it is considered a poor-accuracy indicator as long as it divides each error individually by, in our research, the daily cases, so it is skewed: high errors during low-demand periods will significantly impact MAPE.

Shcherbakov et al. (2013) provides an analysis of existing and quite common forecast error measures that are used in forecasting and divide them in:

ADDITIVE ERROR MODELS

Trend	Seasonal		
	N	A	M
N	$y_t = \ell_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$	$y_t = \ell_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/\ell_{t-1}$
A	$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$
A_d	$y_t = \ell_{t-1} + \phi b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + \phi b_{t-1})$

MULTIPLICATIVE ERROR MODELS

Trend	Seasonal		
	N	A	M
N	$y_t = \ell_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$	$y_t = (\ell_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
A	$y_t = (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
A_d	$y_t = (\ell_{t-1} + \phi b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$

Fig. 7 ETS equations. Source: Hyndman and Athanasopoulos (2018)

- Measures based on absolute forecast error: mean absolute error (MAE), median absolute error (MdMAE), mean square error (MSE) and root mean square error (RMSE);
- Measures based on percentage errors: mean absolute percentage error (MAPE), median absolute percentage error (MdAPE), root mean square percentage error (RMSPE) and median percentage error of the quadratic (RMdSPE);
- Measures based on symmetric errors: symmetric mean absolute percentage error (sMAPE) and median mean absolute percentage error (sMdAPE);
- Measures based on relative errors: mean relative absolute error (MRAE), median relative absolute error (MdRAE) and geometric mean relative absolute error (GMRAE);
- Measures based on scaled error: mean absolute scaled error (MASE), root mean square scaled error (RMSSE).

The same authors (Shcherbakov et al. 2013) states the following shortcomings for each type of error measures

- Measures based on absolute forecast error:

  1. The scale dependency. Does not work with objects in different scales or magnitudes;

2. The high influence of outliers in data on the forecast performance evaluation. If data contain an outliers with maximal value then absolute error measures provide conservative values;
3. RMSE, MSE have a low reliability: the results could be different depending on different fraction of data.

- Measures based on percentage errors:

   1. Appearance division by zero when the actual value is equal to zero;
   2. Nonsymmetrical issue—the error values differ whether the predicted value is bigger or smaller than the actual;
   3. Outliers have significant impact on the result, particularly if outlier has a value much bigger then the maximal value of the regular cases;
   4. The error measures are biased. This can lead to an incorrect evaluation of the forecasting models performance.

- Measures based on symmetric errors:

   1. If the actual value is equal to forecasted value, but with opposite sign, or both of these values are zero, then a divide by zero error occurs;
   2. These criteria are affected by outliers in analogous with the percentage errors;
   3. If more complex estimations have been used, the problem of interpretability of results occurs and this fact slows their spread in practice;
   4. In fact, they do not solve the nonsymmetrical issue problem.

- Measures based on relative errors:

   1. division by zero error still occurs when predicted value obtained by reference model is equal to the actual value;
   2. If naive model has been chosen then division by zero error occurs in case of continuous sequence of identical values of the time series.

- Measures based on scaled error:

   1. If the forecast horizon real values are equal to each other, then division by zero occurs;
   2. Besides it is possible to observe a weak bias estimates.

Thus, considering that all time series are in the same scale and we want to minimize the amount of number with scientific notation, we choose the root mean square error (RMSE) accuracy criteria to compare all models presented in Table 3. The results of each model is presented in Tables 4, 5 and 6. The error evaluation is divided into 3 parts and was applied to using each model:

- In-sample (RMSE IN): comparing training data with fitted values obtained;
- Out-sample all (RMSE OUT-ALL): comparing all test data with predicted values obtained;
- Out-sample mean (RMSE OUT-MEAN): comparing a piece of test data (7 days ahead) with predicted values obtained 4 times and calculate the average error. We run the same model without parameters re-estimation, but we add a new week data (7 days).
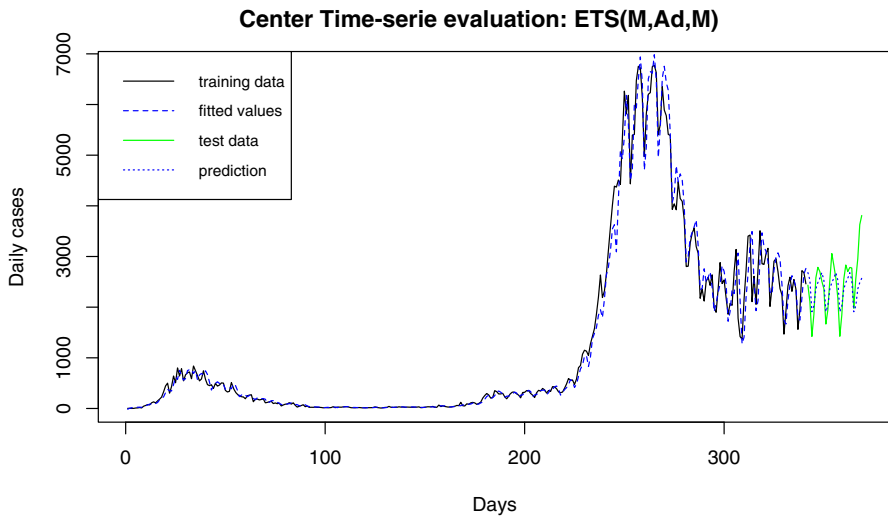
Fig. 8 (Color figure online) ETS model applied in Center. Source: The authors

The reasons why we define our forecast range as 28 days ahead were presented in the end of Sect. 3.

5 Experimentation

In this section, we apply the methods presented in the second (Sect. 5.1) and third (Sect. 5.2) columns of Table 4 to each health region (Rio de Janeiro city), region (Italy) and states (USA).

5.1 Univariate Approaches

Tables 4, 5 and 6 summarize the models obtained by each method and in Tables 7, 8 and 9 the RMSE IN, RMSE OUT-ALL and RMSE OUT-MEAN are presented for each model. These methods only consider the previous values of the same variable to make predictions. In all models the seasonality time window is 7 (weekly).

All plots of time-series approach applied are available in supplementary material. In Fig. 8 an example is provided showing the results of ETS model application in Center (Italy Region). The best type model (ETS(M, Ad, M)) in each class of model (ETS) for Fig. 8, for instance is chosen by the lowest AIC criteria.

From Tables 7, 8 and 9, we can conclude that the best error in-sample considering RMSE criteria are NNETAR (with thirteen) and MLP (with one) for all time series which is not surprising since neural networks work better the more data we give them.

Although outperforming on in-sample comparison, ML models do not obtained the same result by evaluating RMSE OUT-ALL and RMSE OUT-MEAN in which they got lowest RMSE in only five and two time series, respectively.

Table 4 Forecasting parameters per model per region for Rio de Janeiro health region

Models	R 1	R 2	R 3	R 4	R 5
ES	A, N, A	A, N, A	A, N, A	A, N, A	A, N, A
ARIMA	(5, 0, 3)(0, 1, 1)[7]	(2, 0, 1)(0, 1, 2)[7]	(2, 0, 1)(0, 1, 2)[7]	(2, 0, 1)(0, 1, 2)[7]	(2, 0, 1)(0, 1, 2)[7]
SSM-U	3	3	3	3	3
MLP	12 inputs, 8 hidden nodes	11 inputs, 8 hidden nodes	12 inputs, 9 hidden nodes	10 inputs, 8 hidden nodes	10 inputs, 8 hidden nodes
NNETAR	(21, 1, 11)[7]	(14, 1, 8)[7]	(15, 1, 8)[7]	(6, 1, 4)[7]	(15, 1, 8)[7]
TBATS	1, {4, 2}, −, {⟨7, 3⟩}	1, {4, 2}, −, {⟨7, 3⟩}	1, {2, 2}, −, {⟨7, 3⟩}	1, {4, 4}, −, {⟨7, 2⟩}	1, {4, 2}, 0.865, {⟨7, 3⟩}

Source: The authors

Table 5 Forecasting parameters per model per region for Italy regions

Models	CEN	ISL	NOW	NOE	SOT
ES	M, Ad, M	A, N, A	M, A, M	M, N, M	A, Ad, A
ARIMA	(1, 0, 1)(0, 1, 2)[7]	(2, 1, 2)(0, 0, 0)[7]	(1, 0, 1)(0, 1, 2)[7]	(4, 0, 3)(0, 1, 2)[7]	(5, 1, 2)(1, 0, 0)[7]
SSM-U	3	3	3	3	3
MLP	16 inputs, 12 hidden nodes	12 inputs, 10 hidden nodes	17 inputs, 12 hidden nodes	16 inputs, 12 hidden nodes	13 inputs, 10 hidden nodes
NNETAR	(23, 1, 12)[7]	(25, 1, 13)[7]	(21, 1, 11)[7]	(25, 1, 13)[7]	(23, 1, 12)[7]
TBATS	0.21, {0, 0}, 0.96, {⟨7, 2⟩}	1, {0, 0}, −, {⟨7, 2⟩}	0.25, {0, 0}, 0.96, {⟨7, 3⟩}	0.07, {0, 0}, 1,{⟨7, 2⟩}	1, {4, 3}, 0.89, {⟨7, 2⟩}

Source: The authors

Table 6 Forecasting parameters per model per state for US

Models	AZ	CA	NV	OR
ES	A, Ad, A	A, N, A	A, N, A	A, N, A
ARIMA	(0, 1, 2)(0, 0, 2)[7]	(0, 1, 2)(0, 0, 1)[7]	(0, 1, 1)(0, 0, 2)[7]	(2, 1, 3)(0, 0, 2)[7]
SSM-U	2	3	3	3
MLP	21 inputs, 14 hidden nodes Difference order:1	12 inputs, 10 hidden nodes	13 inputs, 9 hidden nodes	17 inputs, 10 hidden nodes
NNETAR	(22, 1, 12)[7]	(22, 1, 12)[7]	(25, 1, 13)[7]	(25, 1, 13)[7]
TBATS	1, {0, 0}, 0.8, {(7, 3)}	1, {3, 3}, –, –	1, {0, 0}, –, {(7, 2)}	1, {0, 0}, –, {(7, 1)}

Source: The authors

Table 7 Forecasting results per model per region for Rio de Janeiro health region

Models	RMSE IN					RMSE OUT-ALL					RMSE OUT-MEAN				
	R 1	R 2	R 3	R 4	R 5	R 1	R 2	R 3	R 4	R 5	R 1	R 2	R 3	R 4	R 5
ES	18.6	62.5	77.2	52.7	73.2	19.9	57.2	119.4	47	104.9	14.4	43.5	65.8	33.9	64
ARIMA	17	59.5	73.1	49.7	67.4	16.6	51.5	99.5	37.9	112	13.5	46.4	69.9	38.1	63.2
SSM-U	20.4	72.5	89.5	57.4	83.5	24.8	75.3	137.7	58.9	118.5	19.5	66.7	107.3	48.1	82.1
MLP	8.8	31.3	31.7	26.2	29.8	17.4	53	92.5	34.3	91.7	15.2	45.9	79.8	38.5	62.6
NNETAR	2.1	21.8	23.4	38.8	21	16.5	65.4	99.4	46.4	101.4	13.2	66.1	96	48	66.5
TBATS	17.2	59.5	72	51.7	72.7	20.4	56.3	118.1	49	104.1	14.1	46.6	78.4	39.5	54.9

Source: The authors

Table 8 Forecasting results per model per region for Italy regions

Models	RMSE IN					RMSE OUT-ALL					RMSE OUT-MEAN				
	CEN	ISL	NOW	NOE	SOT	CEN	ISL	NOW	NOE	SOT	CEN	ISL	NOW	NOE	SOT
ES	256.3	128.4	1085.9	757.8	364.7	436.7	454.3	827.7	991.9	614	414.2	143.4	764.6	673.7	368.8
ARIMA	242.8	127.1	429.9	761	348.8	418.2	489.8	611.4	974.9	632.7	382.1	155.5	676.1	914.1	541.1
SSM-U	165.2	54	397	587.4	238.8	551.4	420.4	493.1	1237.5	841.7	390.3	132.1	500.1	713.3	550.4
MLP	73.9	38.6	103.5	242.7	81.1	471.2	879.7	1064.5	1424.7	1002.3	438.9	406.8	951.2	938.4	681.7
NNETAR	36.9	17.9	71	77.9	45.7	544.3	426.4	2717.5	1720.3	1386.8	404.9	153.9	1474.5	1086.9	609.3
TBATS	238.4	127	427.9	782.7	329.5	364.9	463.2	542.1	674.7	488.4	355.3	145.1	498.3	739.3	494.2

Source: The authors

Table 9 Forecasting results per model per state for US

Models	RMSE IN				RMSE OUT-ALL				RMSE OUT-MEAN			
	AZ	CA	NV	OR	AZ	CA	NV	OR	AZ	CA	NV	OR
ES	1265.8	3167.4	274.3	152.3	662.5	7774.3	389.8	304.5	557.2	1996.3	194.5	152.5
ARIMA	1286.4	2980.6	276.1	153.5	1195.4	6838	334.4	324.1	669.8	1959.1	190.1	148.9
SSM-U	1165.6	1616.6	193.2	113.3	1568.3	7609.9	381.1	274.3	779.8	2301.4	219.8	145.5
MLP	229	1210.5	102.7	48.7	1749.9	9550.3	150.7	490.7	843.9	2507.3	149	255.7
NNETAR	152.1	500.8	37.1	18.5	948.4	29.3×10^3	234	694	699	7023.6	182.1	318.1
TBATS	1251.7	2913.2	275.3	152.1	617.7	9332.4	395.3	301	590	2617	197.2	153.6

Source: The authors

⚫ Springer

Table 10 Best performance models frequency

Models	RMSE IN				RMSE OUT-ALL				RMSE OUT-MEAN			
	RJ	IT	US	Total	RJ	IT	US	Total	RJ	IT	US	Total
ES	0	0	0	0	0	0	0	0	3	2	1	6
ARIMA	0	0	0	0	0	0	1	1	0	0	1	1
SMM-U	0	0	0	0	0	2	1	3	0	1	1	2
MLP	1	0	0	1	4	0	1	5	0	0	1	1
NNETAR	4	5	4	13	1	0	0	1	1	0	0	1
TBATS	0	0	0	0	0	3	1	4	1	2	0	3

Source: The authors

Trying to predict daily cases 28 days ahead without adding new data or parameter re-estimation (OUT-ALL), MLP showed better results for four RJ health regions and one US state. In the second place, TBATS showed better results for three IT regions and one US state. SSM-U appeared in the third position being chosen in two IT regions and one US state.

However, when we predict daily cases 28 days ahead adding new data weekly without parameter re-estimation (OUT-MEAN) we conclude that ES models give us better predictions for six time series while TBATS models and SMM-U were chosen for three and two time series, respectively. All these results are summarized in Table 10.

SSM-U best approach considering the lowest AIC criteria were order two polynomial model ($n = 3$) in thirteen time series. Only in AZ time-series linear model ($n = 2$) was chosen.

After comparing 6 different class of univariate forecasting models and point out which class of model according to lowest RMSE criteria, in next section we present two multivariate forecasting models.

5.2 Multivariate Approaches

SSM-M and VAR methods consider previous values of all variables available to make predictions. In Table 11, we summarize the forecasting error results.

VAR models are divided into four types of deterministic regressors: none, constant, trend or both (constant and trend). We select the deterministic regressors type to each multivariate time series using the lowest AIC criteria. In addition, to select the VAR model order (p) we adopt the Schwarz Criterion ($SC(n)$) and obtained $p = 1$ to RJ with constant and trend deterministic regressors (18 parameters), $p = 23$ to IT with trend deterministic regressors (113 parameters) and $p = 2$ to US with constant and trend deterministic regressors (20 parameters).

In SSM-M, we select a vector length (nx) that gave us the lowest error considering Akaike information criterion (AIC). The nx can be 8 (linear model) or 12 (polynomial order 2 model) to USA and 10 (linear model) or 15 (polynomial order 2 model) to RJ and IT (two or three times the number of univariate time series).

From Table 11, we can conclude that:

Table 11 Forecasting results per multivariate model per region for Rio de Janeiro, Italy and USA

Models	RMSE IN					RMSE OUT-ALL					RMSE OUT-MEAN				
RJ	R 1	R 2	R3	R 4	R 5	R 1	R 2	R 3	R 4	R 5	R 1	R 2	R 3	R 4	R 5
VAR(1)	102.8	67.7	110.3	55.7	83.0	63.3	100.3	166.8	66.7	118.5	29.4	58.7	94.9	46.8	71.1
SSM-M(10)	21.2	73	91.3	59.5	86.4	83.1	107.7	184.2	78.4	123.8	20.0	68.9	112.3	45.5	79.5
IT	CEN	ISL	NOW	NOE	SOT	CEN	ISL	NOW	NOE	SOT	CEN	ISL	NOW	NOE	SOT
VAR(23)	2331.1	3686.2	2292.1	308.3	2322.3	4749.5	1533.5	4522	11×10^3	4502.2	1026.5	1024.2	1326.4	2163.3	1062.1
SSM-M(10)	216.6	94.1	477.7	742.8	335.2	2080.8	3389.4	1808	2022.5	1824.3	589.7	262.5	833.9	1231.9	565.0
US	AZ	CA	NV	OR	–	AZ	CA	NV	OR	–	AZ	CA	NV	OR	–
VAR(2)	2801.1	14.9×10^3	559.5	142.9	–	2908.8	13.1×10^3	992.9	461.4	–	959.9	3519.2	394.1	185.0	–
SSM-M(8)	1341.2	2047.9	229.8	131.6	–	3860.8	4676.9	4645.5	4674.9	–	991.7	785.1	344.8	115.7	–

Source: The authors

Table 12 Univariate and multivariate selected models and error in-sample

Data	IN Regions	Univariate		Multivariate	
		Model	RMSE	Model	RMSE
RJ	R 1	NNETAR	2.1	SSM-M	21.2
	R 2	NNETAR	21.8	VAR	70.1
	R 3	NNETAR	23.4	SSM-M	91.3
	R 4	MLP	26.2	VAR	59.4
	R 5	NNETAR	21	VAR	85.6
Italy	CEN	NNETAR	37.6	SSM-M	216.6
	ISL	NNETAR	17.9	SSM-M	94.1
	NOW	NNETAR	71	SSM-M	477.7
	NOE	NNETAR	77.9	VAR	308.3
	SOT	NNETAR	45.7	SSM-M	335.2
US	AZ	NNETAR	152.1	SSM-M	1341.2
	CA	NNETAR	500.8	SSM-M	2047.9
	NV	NNETAR	37.1	SSM-M	229.8
	OR	NNETAR	18.5	SSM-M	131.6

Source: The authors

- Linear models were chosen for all RJ ($n = 10$), IT ($n = 10$) and US ($n = 10$) data considering the lowest AIC criteria. In univariate time series, we obtained the opposite (almost all models obtained lowest AIC with polynomial order 2 models).
- VAR(1) model obtained best in-sample (IN) error in four RJ health regions while in three US states and four IT regions SSM-M(8) and SSM-M(10) outperform VAR approach considering in-sample error. In other words, SSM-M models better fitted training data in ten time-series training while VAR models better fitted the other four time-series;
- Predicting 28 days ahead without add new data or parameter re-estimation (OUT-ALL) VAR models tied with SSM-M models. VAR(2) achieve better results in three US states while SSM-M(10) fitted better four IT regions.
- predicting 28 days ahead adding new data weekly without parameter re-estimation (OUT-MEAN) SSM-M models better fitted all IT regions and two US states while VAR(1) better fitted four RJ health regions. In other words, SSM-M models showed better results for eight time series while VAR models better fitted the other six time series.

In the next section, we compare results obtained with all approaches mentioned in Table 3 and detailed presented in Tables 4 to 9 and 11.

5.3 Comparing Results of Univariate and Multivariate Methods

In Tables 12 and 13 we compare best model (univariate and multivariate) for all time series considering IN, OUT-ALL and OUT-MEAN RMSE results. This comparison combines Tables 7, 8, 9 and 11 presented in previous sections.

Table 12 reinforces the flexibility of neural networks to fit training data (IN) when working with a large number of observations. Results obtained by NNETAR (thirteen times) and MLP (one time) models outperform all univariate and multivariate models applied in this research.

Table 13 shows us that besides NNETAR not present the same performance taking into account out-sample results, we see another neural network method (MLP) providing the lowest RMSE OUT-ALL to 4 of 5 RJ health regions. It suggests that to RJ data, working with a large number of observations, neural networks methods can also give us a reliable short-term prediction (OUT-ALL).

However, to Italy five regions and USA four states, neural networks short-term prediction (OUT-ALL) only presented better results for NV and CA in US while TBATS models outperform in four Italy Regions (CEN, NOW, NOE and SOT) and in AZ (US).

Despite of high correlation between variables of RJ, IT and US (see Fig. 6 and supplementary data) time-series data, we see multivariate approach outperform only in R3 from RJ and in OR from US (OUT-ALL) using VAR models and OR from US (OUT-MEAN) using SSM-M.

It is important to emphasize that, although univariate models obtained the lowest RMSE in 39 of 42 time series, the difference of results between univariate and multivariate best approaches is lower in RJ than in IT and USA. It may occur because health regions in RJ city are close comparing to US states or IT regions.

The univariate methods could also outperform multivariate because we chose pure simpler models (SSM-M and VAR) and we did not combine them or propose to include more complex models on this analysis like VARMA or some neural network multivariate method. Finally, comparing RMSE OUT-ALL and OUT-MEAN results we can observe that:

- The best class of Univariate models only remains the same in 5 time series (R1, CEN, ISL and NOW, CA). In all these predictions, as expected, OUT-MEAN results were lower than OUT-ALL;
- Even changing the model selected OUT-MEAN results are lower than OUT-ALL in both approaches (Univariate and Multivariate);

Then, we can conclude that although we can make a reliable forecast 28 days ahead, updating the new daily cases weekly allows us to reduce the expected mean error of the forecast in all time series used.

5.4 Forecasting 28 Days Ahead

In Tables 12 and 13, we compared the error results between univariate and multivariate approach which provide us many useful insights.

In this section, we summarize the results presented in Tables 12 and 13 to select the best model for each time series evaluated and then apply it considering all data available (training and test data) to predict daily new cases 28 days ahead. The reasons for choosing forecasting range of 28 days were presented the end of Sect. 3.

Table 13 Univariate and multivariate selected models and error out-sample

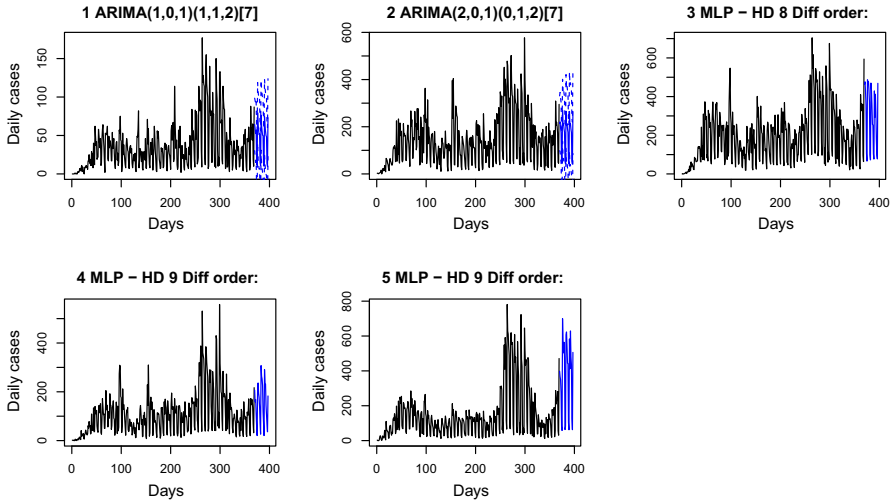| Data	Regions	OUT-ALL Univariate Model	RMSE	Multivariate Model	RMSE	OUT-MEAN Univariate Model	RMSE	Multivariate Model	RMSE
RJ	R 1	NNETAR	16.5	SSM-M	93.3	NNETAR	13.2	SSM-M	20
	R 2	ARIMA	51.5	VAR	90.9	ES	43.5	VAR	56.9
	R 3	MLP	92.5	VAR	92.2	ES	65.8	VAR	86
	R 4	MLP	34.3	SSM-M	84.1	ES	33.9	VAR	42.2
	R 5	MLP	91.7	VAR	92.1	TBATS	54.9	VAR	69.6
Italy	CEN	TBATS	364.9	SSM-M	2104.3	TBATS	355.3	SSM-M	589.7
	ISL	ES	454.3	VAR	1496.7	ES	143.4	SSM-M	262.5
	NOW	TBATS	542.1	SSM-M	1843.5	TBATS	498.3	SSM-M	833.9
	NOE	TBATS	674.7	SSM-M	2049.1	ES	673.7	SSM-M	1231.9
	SOT	TBATS	488.4	SSM-M	1860.5	ES	368.7	SSM-M	565.0
US	AZ	TBATS	617.7	VAR	1986	ES	557.2	VAR	900
	CA	ARIMA	6838	SSM-M	5643.6	ARIMA	1959.1	SSM-M	785.1
	NV	NNETAR	246.7	VAR	406.5	MLP	157.2	VAR	298.1
	OR	SSM-U	268.6	VAR	117.4	ARIMA	148.9	SSM-M	115.7

Source: The authors

Fig. 9 (Color figure online) RJ health regions forecasting. Source: The authors
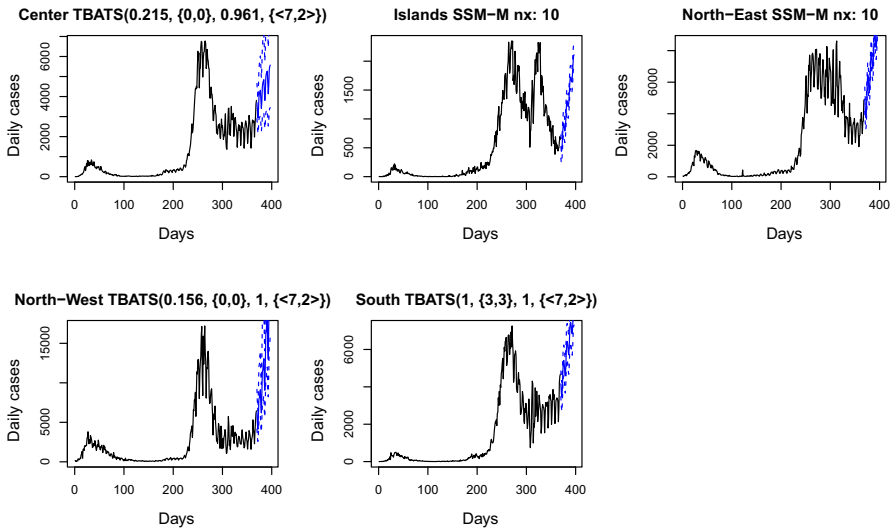

Fig. 10 (Color figure online) IT regions forecasting. Source: The authors

To provide the daily new cases prediction proposed, we re-estimate all parameters of models selected in third column of Table 14. Finally, in Figs. 9, 10 and 11, we present the forecasting values with confidence interval of 0.95.
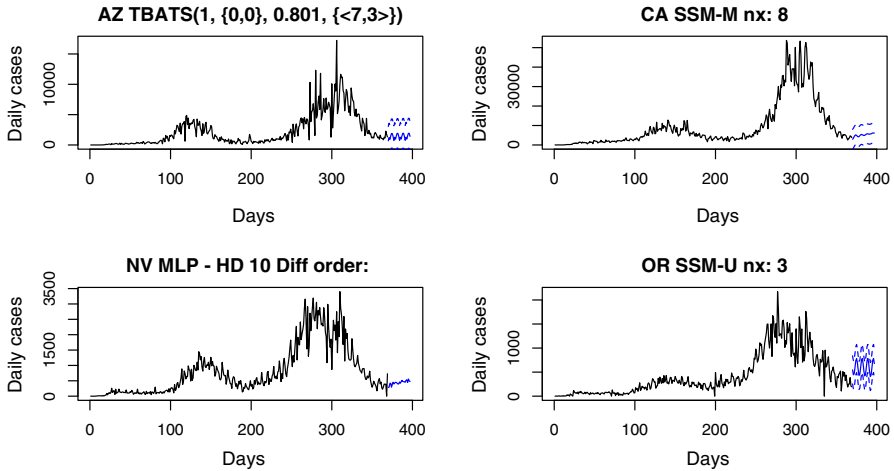
AZ TBATS(1, {0,0}, 0.801, {<7,3>})

CA SSM-M nx: 8

NV MLP - HD 10 Diff order:

OR SSM-U nx: 3

Fig. 11 (Color figure online) US states forecasting. Source: The authors

Table 14 Best model selection to each time series

Region/State	RMSE IN	RMSE OUT-ALL	RMSE OUT-MEAN
R1	NNETAR	NNETAR	ARIMA
R2	NNETAR	ARIMA	ES
R3	NNETAR	MLP	ES
R4	MLP	MLP	ES
R5	NNETAR	MLP	TBATS
CEN	NNETAR	TBATS	TBATS
ISL	NNETAR	SSM-U	SSM-U
NOW	NNETAR	SSM-U	TBATS
NOE	NNETAR	TBATS	ES
SOT	NNETAR	TBATS	ES
AZ	NNETAR	TBATS	ES
CA	NNETAR	SSM-M	SSM-M
NV	NNETAR	MLP	MLP
OR	NNETAR	SSM-U	SSM-M

Source: The authors

6 Conclusions

In this research, we apply 6 univariate and 2 multivariate models to evaluate 14 time series from a Brazilian city (RJ), all Italian regions and 4 US states. For each time series, we pointed out the best approach considering the lowest RMSE criteria.

An extensive literature review (for more details, see "Appendix D") were conducted to find forecasting models applied to human infectious disease outbreak (research's scope) presented in Sects. 2.1 to 2.4.

In mentioned sections, we only pointed out forecasting models applied to the scope of this research which are summarized in Sect. 4. Thus, it is suggested to explore forecasting methods used in other subjects or knowledge area. An extensive list of forecasting methods can be seen at Petropoulos et al. (2022).

Although unusual in current literature of human infectious disease outbreak prediction or forecasting (less then 10% of research we found), we apply multivariate methods because of the high correlation and auto-correlation between different time series from the same region in many lags as we saw in Fig. 6.

In "Appendix C," all auto-correlation plots are presented where we see a significant correlation between regions data until lag 15 to RJ and in all lags to Italy regions and US states.

In-sample (RMSE IN) results obtained best results using univariate MLM for all time series which is expected considering that these types of models usually provide better results the more data are available for training.

However, the same pattern was not observed in both out-sample (RMSE OUT-ALL and RMSE OUT-MEAN) results evaluation. In RMSE OUT-ALL univariate MLM outperform 4 times, TBATS 4 times and SSM-U 3 times. In RMSE OUT-MEAN ES outperform 6 times and TBATS 3 times.

Besides the strong potential of multivariate methods, we did not observe them outperforming univariate methods. It only happens 3 times (RMSE OUT-ALL and RMSE OUT-MEAN for CA and RMSE OUT-MEAN for OR). For this three time-series SSM-M have got the most reliable predictions.

Our prediction presented in Figs. 9, 10 and 11, suggests that in the next 28 days:

- 4 RJ health regions will remain on the same level of daily new cases, but in R5 is expected to face a considerable increasing of daily COVID-19 new cases. However, it will be at least lower than levels observed in previous data;
- IT regions will face a exponential increasing of daily COVID-19 new cases, excluding CEN Region;
- In US states, we can expect different behaviours of daily COVID-19 new cases. To AZ, it is expected a tiny decreasing while in CA and NV will increase. In OR it is expected that daily cases remains in the same level of 600 new daily cases.

As further research, we suggest the application of multivariate MLM techniques like multivariate MLP or LSTM (largely and successfully applied in literature for univariate time-series approach). Another possible way is to combine the mentioned multivariate methods with VARMA.

Causal models are largely applied in current literature and should be also explored. However, this type of approach also depends from collecting data from other sources that sometimes unavailable.

Finally, it is important to emphasize that the set of models and data collection that should be applied to any forecast human disease outbreaks depends on the type of disease transmission.

In airborne infectious diseases transmission like COVID-19, influenza, among others, we observe interesting applications combining daily/weekly or monthly cases with search engine of Google or Baidu (in China) or mobility data to find better predictions. On the other hand, diseases transmitted by vectors such as mosquitoes like dengue,

Zika virus among others are typically combined with temperature and rainfalls, for instance.

Declarations

Conflict of interest The authors have declared that no competing interests exist.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Appendix A: Research Synthesis

In this section, we present all similar research that we found in literature. As mentioned in Sect. 2.5, columns 3 to 7 address the approach used in each research between CSM, CM, SSM and MLM.

In columns 8 to 9 is presented the range of time windows used as well as the prediction range. Time windows found on previous research were day (d), week (w), month (m) or year (y). The prediction range is expressed in time windows mentioned, but some models proposed to forecast the whole pandemic period (wpp).

The variable measured/evaluated and forecasted in each research (column 10) on those time windows is number of patient cases (ca), deaths (de) and recovered (re), admitted and discharged from hospital or intensive care unit (adhosp and dishosp) and transmission rate (rt).

Excluding rt, all measures mentioned can be counted in two different ways: by time window or cumulative. For example, daily cases (dca), monthly deaths (mde), yearly patients admitted in hospital (yadhosp), cumulative cases (Cca), cumulative patients discharged from hospital (Cdishosp).

Columns 11 to 13 present in which countries each research applied the methods specified in columns 3 to 7, the type of forecast approach divided into univariate (Uni), causal or multivariate (Mul) and the disease outbreak studied.

Appendix B: Geographical Regions

In this section, we present the map of regions mentioned in Table 2 for each time series evaluated in this research (Figs. 12, 13 and 14).

Table 15 Literature review synthesis

Authors	Year	CSM	CM	SSM	MLM	Others	Data range	Data prediction	Measure	Country	Approach	Disease outbreak
Current	2022	ES, ARIMA	–	KF, TBATS	MLP, NNETAR	–	24/02/20 to 11/07/21	28 d	dca	Brazil, Italy, US	Uni and Mul	COVID-19
Chen et al. (2021)	2021	–	SEIR	–	–	Spatial autocorrelation	21/01/20 to 12/03/20	wpp	dca	China	Causal	COVID-19
Kiang et al. (2021)	2021	ARIMA	–	–	–	–	01/01/10 to 12/12/16	6 m	mca	Thailand	Causal	Dengue
ArunKumar et al. (2021)	2021	ARIMA	–	–	LSTM	–	22/01/20 to 12/03/20	60 d	Cca, Cde, Cre	16 countries	Uni	COVID-19
Talkhi et al. (2021)	2021	ES, ARIMA	–	TBATS	NNETAR, ELM, MLP	Prophet	20/02/20 to 15/08/20	30 d	Cca, Cde	Iran	Uni	COVID-19
Katris (2021)	2021	ES, ARIMA	SIR	–	RNN	Newbolt–Granger, Bates–Granger	21/01/20 to 14/08/20	4 w	rt	Greece	Uni	COVID-19
Paul et al. (2021)	2021	–	GML, GMR	–	–	–	22/01/20 to 14/09/20	wpp	dca, rt	10 countries	Uni	COVID-19
Han et al. (2021)	2021	–	–	KF	AutoML	–	15/03/20 to 17/05/20	15 d	dca, dde	Brazil (Ceará)	Mul	COVID-19

Table 15 continued

Authors	Year	CSM	CM	SSM	MLM	Others	Data range	Data prediction	Measure	Country	Approach	Disease outbreak
Ribeiro et al. (2020)	2020	–	–	–	EEMD	–	01/01/07 to 31/12/18	3 m	mca	Brazil (5 states)	Mul	Meningitis
Benítez et al. (2020)	2020	–	GML, GMR	Gaussian model	–	–	65 to 104 d	10 d	dca, dde, dad-hosp, dishosp	Spain, Sierra Leone and Colombia	Uni	COVID-19, Ebola and zika virus
Khan et al. (2020)	2020	VAR	–	–	–	–	08/03/20 to 27/06/20	50 d	dca, dde, dre	Iran	Mul	COVID-19
Deng et al. (2020)	2020	–	–	–	CLAGNN	–	01/03/10 to 30/03/19	2 to 15 w	wca	Japan and US	Mul	ILI
Wang et al. (2020)	2020	–	GML	–	RNN	Prophet	22/01/20 to 16/07/20	6 m	Cca	5 countries	Uni	COVID-19
Bomfim et al. (2020)	2020	ARIMA, ARIMAX	–	–	RNN	–	2007 to 2015	18 w	wca	Brazil (Fortaleza)	Causal	Dengue
Liang et al. (2020)	2020	ARIMA	–	–	SVM, BMA	–	2004 to 2016	24 m	mca	China (31 cities)	Causal	Tuberculosis
Ramos et al. (2020)	2020	ARIMA	–	–	–	–	2006 and 2016	36 m	mca	Brazil (Maranhão)	Uni	Leprosy

Table 15 continued

Authors	Year	CSM	CM	SSM	MLM	Others	Data range	Data prediction	Measure	Country	Approach	Disease outbreak
Zhang et al. (2019)	2019	ARIMA	–	–	SVM	–	2004 to 2017	6 m	mca	China	Causal	HFRS
Wang et al. (2019)	2019	ARIMA	–	–	LSTM, NARNN	–	01/06/08 to 30/06/18	18 m	mca	China	Uni	HFMD
Li et al. (2019)	2019	ARIMAX	–	–	–	–	01/01/11 to 30/01/16	15 m	mca	China	Causal	HIV/AIDS
Choi et al. (2019)	2019	ARIMA	–	–	RNN	Linear Regression	2010 to 2018	26 w	wca	US	Causal	ILI
Smirnova et al. (2019)	2019	–	SEIR	–	–	–	52 weeks	52 w	wca and rt	UK (New Castle)	Uni	Measles
Eilertson et al. (2019)	2019	–	SIR	KF	–	–	1980 to 2014	1 to 50 y	yca	193 countries	Uni	Measles
Chakraborty et al. (2019)	2019	ARIMA	–	–	NNAR	–	01/05/90 to 31/10/11	6 w and 1 y	mca	Philippines	Uni	Dengue
Stolerman et al. (2019)	2019	–	–	–	SVM	–	2001 to 2012	4 y	wca	Brazil (7 cities)	Causal	Dengue
Chumachenko et al. (2019)	2019	ARIMA, ES	–	–	–	–	2010 to 2017	12 m	mca	Ukraine	Uni	Measles
Haddawy et al. (2018)	2018	ARIMAX	–	–	–	Cluster	2012 to 2013 (99 w)	4 w	wca	Thailand	Causal	Malaria

Table 15 continued

Authors	Year	CSM	CM	SSM	MLM	Others	Data range	Data prediction	Measure	Country	Approach	Disease outbreak
Wu et al. (2018)	2018	ARIMA	–	–	–	Cluster	01/01/07 to 31/12/16	6 m	mca	China	Causal	HFRS
Suparit et al. (2018)	2018	–	SEIR	–	–	–	1 w 15 to 22 w 16	6 m	wca	Brazil (Bahia)	Uni	Zika virus
Wu et al. (2018)	2018	ARIMA,VAR	–	–	RNN, CNN	–	2009 to 2016	8 w	wca	USA and Japan	Mul	ILI
Basile et al. (2018)	2018	–	SEIR	–	–	–	2010 to 2014	2 w	wca	Spain (Catalonia)	Causal	ILI
Zhao et al. (2018)	2018	ARIMA	–	–	–	Space-time Autoregressive Integrated Moving Average (STARIMA)	2005 to 2014	12 m	mca	China (Hubei)	Mul	HFRS

Table 15 continued

Authors	Year	CSM	CM	SSM	MLM	Others	Data range	Data prediction	Measure	Country	Approach	Disease outbreak
Li et al. (2018)	2018	–	SEIR	–	–	Sequential Monte Carlo approach of particle filtering	1921 to 1956	1 m	mca	Canada (Saskatchewan)	Uni	Measles
Jerónimo-Martínez et al. (2017)	2017	ARIMA	–	–	–	–	2009 to 2012	18 w	wca	Colombia (Bogota)	Uni	ARI
Ray et al. (2017)	2017	ARIMA	–	–	KCDE	–	1990 to 2010	26 w	wca	USA and Japan	Uni	Dengue and ILI
Tabataba et al. (2017)	2017	–	–	–	–	Agent based simulation	2014	4 w	wca	Liberia	Uni	Ebola
Chuang et al. (2017)	2017	–	–	–	–	Distributed lag non-linear model (DLNM)	1998 to 2015	6 w	wca	Taiwan	Causal	Dengue

Table 15 continued

Authors	Year	CSM	CM	SSM	MLM	Others	Data range	Data prediction	Measure	Country	Approach	Disease outbreak
Guo et al. (2017)	2017	–	–	–	–	Bootstrap Aggregating (Bagging) and Rank Aggregation	2011 to 2014	12 m	mca	China	Causal	ILI
Caicedo-Torres et al. (2017)	2017	–	–	–	KRRGPN	–	2007, 2014 to 2016	1 to 4 w	wca	Colombia	Uni	Dengue and Chikungunya
Nguyen et al. (2017)	2017	–	–	–	NFIS	Cluster	2008 to 2014	12 m	mca	China (10 locations)	Causal	Dengue
Anggraeni and Aristiani (2016)	2016	ARIMA, ARIMAX	–	–	–	–	2010 to 2014	26 w	wca	Indonesia (Surabaya)	Causal	Dengue
Ke et al. (2016)	2016	ES	–	–	–	–	2006 to 2016	6 m	mca	China	Uni	HFRS
Valeri et al. (2016)	2016	–	GM	–	–	–	2014 to 2015	wpp	wca	Guinea, Sierra Leone and Liberia	Causal	Ebola

Table 15 continued

Authors	Year	CSM	CM	SSM	MLM	Others	Data range	Data prediction	Measure	Country	Approach	Disease outbreak
Li et al. (2016)	2016	ARIMA	–	–	–	–	1992 to 2013	12 m	mca	China (Jiaonan)	Causal	HFRS
Johansson et al. (2016)	2016	ARIMA	–	–	–	–	1985 to 2012	1 to 6 m	mca	Mexico (all states)	Causal	Dengue
Pradhan et al. (2016)	2016	ES	–	–	–	Segmented regression	2014 to 2016	24 m	mca	India (Odisha)	Causal	Malaria
Chau and Ngoc Anh (2016)	2016	–	–	–	RF	–	2001 to 2012	30 d	dca	Vietnam (Hanoi – 29 districts)	Mul	Cholera
Wu et al. (2015)	2015	ARIMA	–	–	NARNN, GRNN	–	2004 to 2012	12 m	mca	China (Jiangsu)	Uni	HFRS
Mekparyup and Saithanu (2015)	2015	ARIMA	–	–	–	–	2007 to 2013	24 m	mca	Thailand (Chonburi)	Uni	DHF
Kane et al. (2014)	2014	ARIMA	–	–	RF	–	2006	1 w	wca	Egypt	Uni	ILI
Feng et al. (2014)	2014	ARIMA	–	–	–	–	2006 to 2012	26 w	wadhosp	China	Causal	HFMD
Yang et al. (2014)	2014	.	SIRS	KF	–	–	2003 to 2012	6 w	wca	US (115 cities)	Uni	ILI

Table 15 continued

Authors	Year	CSM	CM	SSM	MLM	Others	Data range	Data prediction	Measure	Country	Approach	Disease outbreak
Wang et al. (2013)	2013	–	SEIR	MCMC	–	–	2004 to 2006	24 m	mca	US (Arizona)	Uni	SSTIS
Nunes et al. (2013)	2013	–	–	HMM	–	–	2008 to 2012	26 w	wca	Portugal	Uni	ILI
Towers and Chowell (2012)	2012	–	SEIR	–	–	–	2009	wpp	dadhosp	Chile	Uni	ILI
Aguiar et al. (2011)	2011	–	SEIR	–	–	–	1983 to 2006	19 y	mca	Thailand (Chiang Mai)	Uni	Dengue and DHF
Gerardi and Monteiro (2011)	2011	–	SIS, SIR, CA	–	GA	–	2007	52 w	wca	Brazil (Rio de Janeiro)	Uni	Dengue
Laneri et al. (2010)	2010	–	SEIR	–	–	–	1987 to 2006	3 m	mca	India (Gujarat and Rajasthan states)	Causal	Malaria
Soebiyanto et al. (2010)	2010	ARIMA	–	–	–	–	2005 to 2008	1 w	wca	China (Hong Kong)	Causal	ILI
Santos et al. (2009)	2009	–	CA	–	–	–	1999	5 y	wca	Brazil (Salvador and Mossoró)	Causal	Dengue

Table 15 continued

Authors	Year	CSM	CM	SSM	MLM	Others	Data range	Data prediction	Measure	Country	Approach	Disease outbreak
Peng et al. (2008)	2008	–	–	–	WNN	–	2002 to 2004	52 w	wca	US	Uni	ILI
Shen et al. (2008)	2008	ARIMA	–	–	–	–	1957 to 2001	24 m	mca	China (Bei-jing)	Uni	SF
Medina et al. (2007)	2007	ES, ARIMA	–	–	–	–	1996 to 2004	3 m	mca	Mali (Niono)	Uni	ARI and Malaria
Burkom et al. (2007)	2007	ES	–	–	–	–	2002 to 2003	1 d	dca	US	Uni	ILI
Finkenstädt et al. (2005)	2005	–	SEIR	–	–	–	1984 to 2002	4 w	wca	France	Uni	ILI
Nobre et al. (2001)	2001	ARIMA	–	–	–	–	1980 to 1995	6 m	mca	US	Uni	Malaria
Mode et al. (1991)	1991	–	–	MCMC	–	–	1987 to 1988	36 m	mca	US	Uni	HIV/AIDS
Gamerman and Migon (1991)	1991	–	GM	–	–	–	1985 to 1988	6 m	mca	Brazil	Uni	HIV/AIDS

Source: The authors
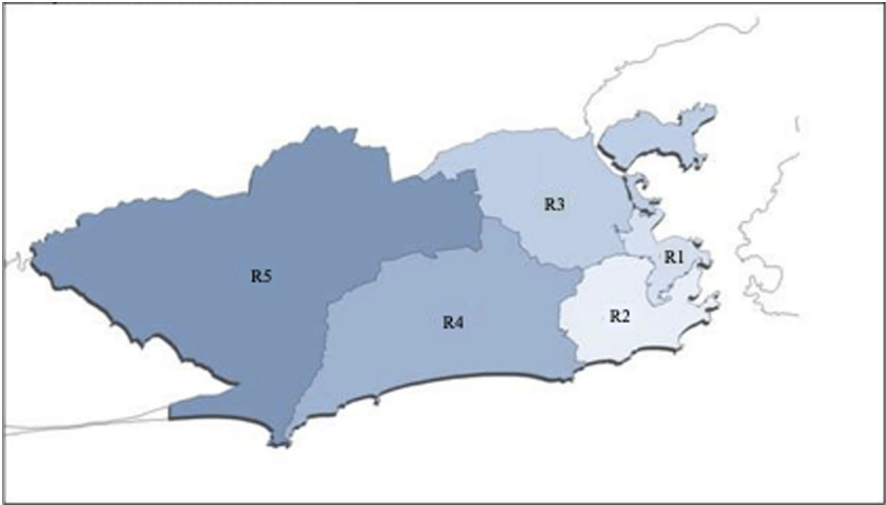
Fig. 12 (Color figure online) Rio de Janeiro health regions. Source: The authors

Fig. 13 (Color figure online) Italy regions. Source: The authors
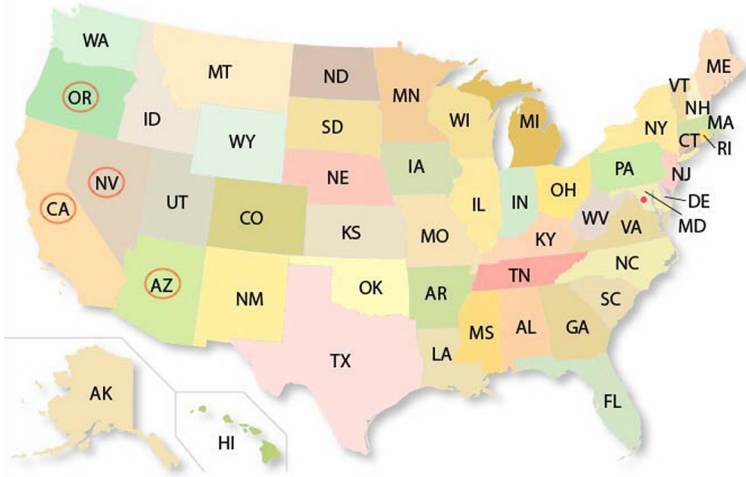
Fig. 14 (Color figure online) US states. Source: The authors
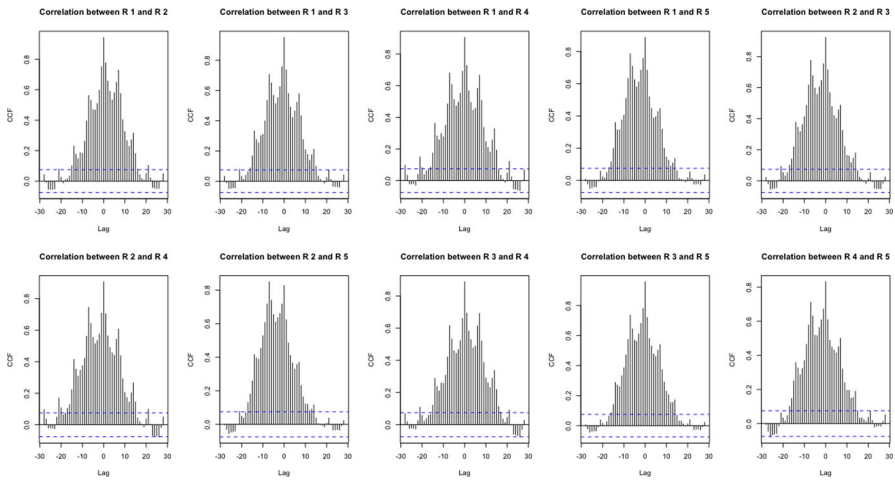
Fig. 15 (Color figure online) Rio de Janeiro Health Regions correlation plot. Source: The authors

Appendix C: Correlation Plots

In this section, we present the correlation plots between all variables from each country or region evaluated in this research. All plots are also available at GitHub (Assad 2022). In Sect. 3, we present a single plot in order to show why working with multivariate approach could be worth (Figs. 15, 16 and 17).
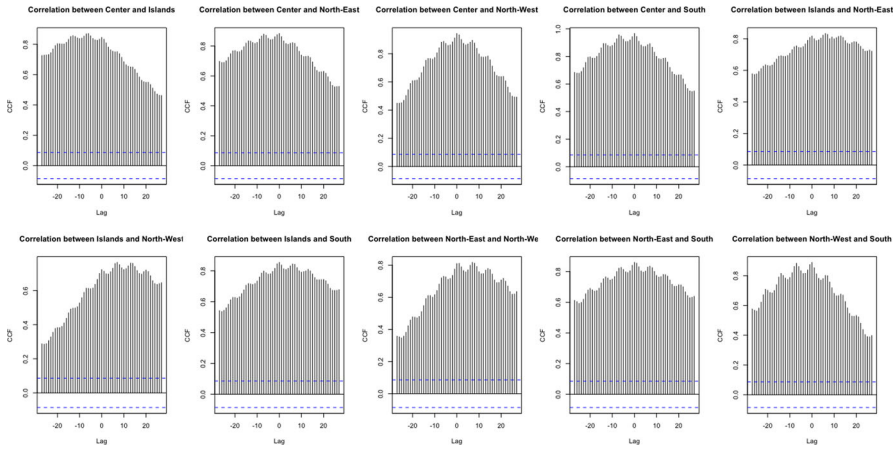
Fig. 16 (Color figure online) Italy regions correlation plot. Source: The authors
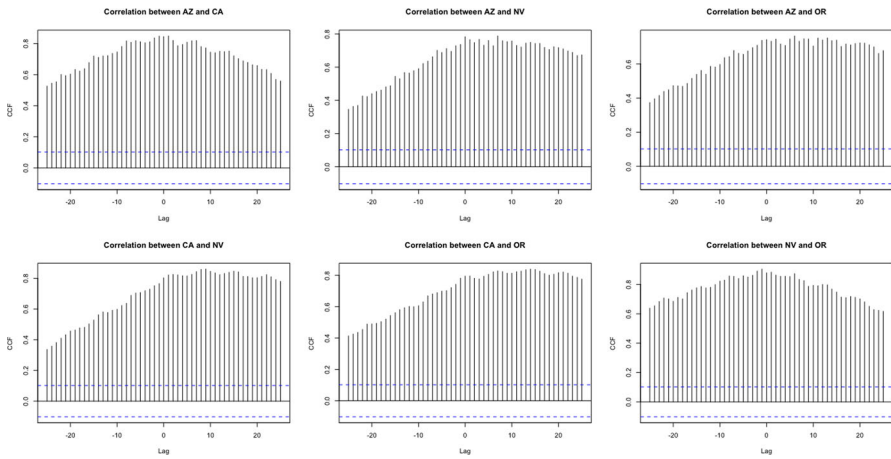
Fig. 17 (Color figure online) US states correlation plot. Source: The authors

Appendix D: Literature Review Steps and Results

In this research, we conduced a extensive literature review over forecasting methods applied to human diseases outbreaks. We retrieved articles from three different scientific databases: Web of Science, SCOPUS and PubMED.

We used the following keywords: pandemic*, epidemic*, corona*, covid*, diseas*, outbreak*, predict*, forecast*, model*, techniq*, approach*, method*, time*, serie*. Keywords combination presented in the Fig. 18. Research metadata was retrieved in April 30, 2022.

In the third search, after removing duplicate results, we obtained 654 research including 10 reviews. Then, we evaluate all results obtained we select only research that properly make predictions which were 66.

Fig. 18 (Color figure online) Literature review keywords and steps. Source: The authors

All 66 research are summarized in Table 15 in order to provide a comparison between our research contribution with current literature.

References

Aguiar M, Ballesteros S, Kooi BW, Stollenwerk N (2011) The role of seasonality and import in a minimalistic multi-strain dengue model capturing differences between primary and secondary infections: complex dynamics and its implications for data analysis. J Theor Biol 289:181–196

Alzubi J, Nayyar A, Kumar A (2018) Machine learning from theory to algorithms: an overview. J Phys Conf Ser 1142:012012

Anggraeni W, Aristiani L (2016) Using google trend data in forecasting number of dengue fever cases with ARIMAX method case study: Surabaya, Indonesia. In: 2016 International conference on information & communication technology and systems (ICTS). IEEE, pp 114–118

ArunKumar K, Kalaga DV, Kumar CMS, Chilkoor G, Kawaji M, Brenza TM (2021) Forecasting the dynamics of cumulative Covid-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: auto-regressive integrated moving average (ARIMA) and seasonal auto-regressive integrated moving average (SARIMA). Appl Soft Comput 103:107161

Assad DBN (2022) Short-term-forecasting. GitHub. https://github.com/DanielAssad/Short-term-forecasting.git

Basile L, Oviedo de la Fuente M, Torner N, Martínez A, Jané M (2018) Real-time predictive seasonal influenza model in Catalonia, Spain. PLoS ONE 13(3):0193651

Benítez D, Montero G, Rodríguez E, Greiner D, Oliver A, González L, Montenegro R (2020) A phenomenological epidemic model based on the spatio–temporal evolution of a gaussian probability density function. Mathematics 8(11):2000

Bomfim R, Pei S, Shaman J, Yamana T, Makse HA, Andrade JS Jr, Lima Neto AS, Furtado V (2020) Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas. J R Soc Interface 17(171):20200691

Box G, Jenkins G (1970) Control. Halden-Day, San Francisco

Brown RG (1959) Statistical forecasting for inventory control. McGraw/Hill, New York

Burkom HS, Murphy SP, Shmueli G (2007) Automated time series forecasting for biosurveillance. Stat Med 26(22):4202–4218

Caicedo-Torres W, Montes-Grajales D, Miranda-Castro W, Fennix-Agudelo M, Agudelo-Herrera N (2017) Kernel-based machine learning models for the prediction of dengue and chikungunya morbidity in Colombia. In: Colombian conference on computing. Springer, Berlin, pp 472–484

Chakraborty T, Chattopadhyay S, Ghosh I (2019) Forecasting dengue epidemics using a hybrid methodology. Physica A 527:121266

Chau NH, Ngoc Anh LT (2016) Using local weather and geographical information to predict cholera outbreaks in Hanoi, Vietnam. In: Advanced computational methods for knowledge engineering. Springer, Austria, pp 195–212

Chen Y, Li Q, Karimian H, Chen X, Li X (2021) Spatio–temporal distribution characteristics and influencing factors of Covid-19 in China. Sci Rep 11(1):1–12

Choi SB, Kim J, Ahn I (2019) Forecasting type-specific seasonal influenza after 26 weeks in the united states using influenza activities in other countries. PLoS ONE 14(11):0220423

Chowell G, Sattenspiel L, Bansal S, Viboud C (2016) Mathematical models to characterize early epidemic growth: a review. Phys Life Rev 18:66–97

Chretien J-P, George D, Shaman J, Chitale RA, McKenzie FE (2014) Influenza forecasting in human populations: a scoping review. PLoS ONE 9(4):94130

Chuang T-W, Chaves LF, Chen P-J (2017) Effects of local and regional climatic fluctuations on dengue outbreaks in southern Taiwan. PLoS ONE 12(6):0178698

Chumachenko D, Turiy A, Chukhray A (2019) Application of statistical simulation for measles epidemic process forecasting. In: 2019 IEEE 2nd Ukraine conference on electrical and computer engineering (UKRCON). IEEE, pp 1086–1090

De Livera AM, Hyndman RJ, Snyder RD (2011) Forecasting time series with complex seasonal patterns using exponential smoothing. J Am Stat Assoc 106(496):1513–1527

Deng S, Wang S, Rangwala H, Wang L, Ning Y (2020) Cola-GNN: cross-location attention based graph neural networks for long-term ILI prediction. In: Proceedings of the 29th ACM international conference on information & knowledge management, pp 245–254

Dobbyn A (2020) Covid19 US: cases of COVID-19 in the United States. R package version 0.1.7. https://CRAN.R-project.org/package=covid19us

Eilertson KE, Fricks J, Ferrari MJ (2019) Estimation and prediction for a mechanistic model of measles transmission using particle filtering and maximum likelihood estimation. Stat Med 38(21):4146–4158

Fekedulegn D, Mac Siúrtáin MP, Colbert JJ (1999) Parameter estimation of nonlinear models in forestry. Silva Fennica 33(4):327–336

Feng H, Duan G, Zhang R, Zhang W (2014) Time series analysis of hand-foot-mouth disease hospitalization in Zhengzhou: establishment of forecasting models using climate variables as predictors. PLoS ONE 9(1):87916

Finkenstädt B, Morton A, Rand D (2005) Modelling antigenic drift in weekly flu incidence. Stat Med 24(22):3447–3461

Gamerman D, Migon HS (1991) Forecasting the number of aids cases in Brazil. J R Stat Soc Ser D (The Statistician) 40(4):427–442

Gerardi D, Monteiro L (2011) System identification and prediction of dengue fever incidence in Rio de Janeiro. Math Probl Eng 2011

Guo P, Zhang J, Wang L, Yang S, Luo G, Deng C, Wen Y, Zhang Q (2017) Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model. Sci Rep 7(1):1–11

Haddawy P, Yin MS, Wisanrakkit T, Limsupavanich R, Promrat P, Lawpoolsri S, Sa-angchai P (2018) Complexity-based spatial hierarchical clustering for malaria prediction. Jo Healthc Inform Res 2(4):423–447

Han T, Gois FNB, Oliveira R, Prates LR, de Almeida Porto MM (2021) Modeling the progression of COVID-19 deaths using Kalman filter and AutoML. Soft Comput 1–16

Haykin S (2004) Kalman filtering and neural networks, vol 47. John Wiley & Sons, New York

Hays JN (2005) Epidemics and pandemics: their impacts on human history. Abc-clio, United States of America

Holt C (1957) Forecasting seasonals and trends by exponentially weighted averages (onr memorandum no. 52). Carnegie Institute of Technology, Pittsburgh, USA, 10

Honigsbaum M (2009) Pandemic. Lancet 373(9679):1939

Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice. OTexts, Australia

Jerónimo-Martínez LE, Menéndez-Mora RE, Bolívar H (2017) Forecasting acute respiratory infection cases in Southern Bogota: EARS vs. ARIMA and SARIMA. In: 2017 Congreso Internacional de Innovacion Y Tendencias en Ingenieria (CONIITI). IEEE, pp 1–6

Johansson MA, Reich NG, Hota A, Brownstein JS, Santillana M (2016) Evaluating the performance of infectious disease forecasts: a comparison of climate-driven and seasonal dengue forecasts for Mexico. Sci Rep 6(1):1–11

Kalman RE et al (1960) Contributions to the theory of optimal control. Bol Soc Mat Mexicana 5(2):102–119

Kane MJ, Price N, Scotch M, Rabinowitz P (2014) Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinform 15(1):1–9

Kaps M, Herring W, Lamberson W (2000) Genetic and environmental parameters for traits derived from the Brody growth curve and their relationships with weaning weight in Angus cattle. J Anim Sci 78(6):1436–1442

Katris C (2021) A time series-based statistical approach for outbreak spread forecasting: application of COVID-19 in Greece. Expert Syst Appl 166:114077

Kaur H, Garg S, Joshi H, Ayaz S, Sharma S, Bhandari M (2020) A review: epidemics and pandemics in human history. Int J Pharma Res Health Sci 8:3139–3142. https://doi.org/10.21276/ijprhs.2020.02.01

Ke G, Hu Y, Huang X, Peng X, Lei M, Huang C, Gu L, Xian P, Yang D (2016) Epidemiological analysis of hemorrhagic fever with renal syndrome in China with the seasonal-trend decomposition method and the exponential smoothing model. Sci Rep 6(1):1–7

Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc R Soc Lond Ser A Contain Pap Math Phys Character 115(772):700–721

Khamis A (2005) Nonlinear growth models for modeling oil palm yield growth. J Math Stat 1(3):225–233

Khan F, Saeed A, Ali S (2020) Modelling and forecasting of new cases, deaths and recover cases of COVID-19 by using vector autoregressive model in Pakistan. Chaos Solitons Fractals 140:110189

Kiang MV, Santillana M, Chen JT, Onnela J-P, Krieger N, Engø-Monsen K, Ekapirat N, Areechokchai D, Prempree P, Maude RJ et al (2021) Incorporating human mobility data improves forecasts of dengue fever in Thailand. Sci Rep 11(1):1–12

Kohavi R (1998) Glossary of terms. Spec Issue Appl Mach Learn Knowl Discov Process 30(271):127–132

Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT press, United States of America

Kotu V, Deshpande B (2019) Time series forecasting. Data Science; Elsevier, Amsterdam, pp 395–445

Krause AL, Kurowski L, Yawar K, Van Gorder RA (2018) Stochastic epidemic metapopulation models on networks: SIS dynamics and control strategies. J Theor Biol 449:35–52

Krispin R (2021) Covid19italy: The 2019 Novel Coronavirus COVID-19 (2019-nCoV) Italy Dataset. R package version 0.3.1. https://CRAN.R-project.org/package=covid19italy

Laneri K, Bhadra A, Ionides EL, Bouma M, Dhiman RC, Yadav RS, Pascual M (2010) Forcing versus feedback: epidemic malaria and monsoon rains in Northwest India. PLoS Comput Biol 6(9):1000898

Li S, Cao W, Ren H, Lu L, Zhuang D, Liu Q (2016) Time series analysis of hemorrhagic fever with renal syndrome: a case study in Jiaonan County, China. PLoS ONE 11(10):0163771

Li X, Doroshenko A, Osgood ND (2018) Applying particle filtering in both aggregated and age-structured population compartmental models of pre-vaccination measles. PLoS ONE 13(11):0206529

Li K, Liu M, Feng Y, Ning C, Ou W, Sun J, Wei W, Liang H, Shao Y (2019) Using Baidu search engine to monitor aids epidemics inform for targeted intervention of HIV/AIDS in China. Sci Rep 9(1):1–12

Liang X, Xu Q, Guan R, Zhao Y (2020) Forecasting tuberculosis incidence in china using Baidu index: a comparative study. In: Proceedings of the 4th international conference on medical and health informatics, pp 22–29

Marques JAL, Gois FNB, Xavier-Neto J, Fong SJ (2021) Predictive models for decision support in the COVID-19 crisis. Springer, Switzerland

Medina DC, Findley SE, Guindo B, Doumbia S (2007) Forecasting non-stationary diarrhea, acute respiratory infection, and malaria time-series in Niono, Mali. PLoS ONE 2(11):1181

Mekparyup J, Saithanu K (2015) Forecasting the dengue hemorrhagic fever cases using seasonal ARIMA model in Chonburi, Thailand. Global J Pure Appl Math 11:401–407

Metcalf CJE, Lessler J (2017) Opportunities and challenges in modeling emerging infectious diseases. Science 357(6347):149–152

Mode CJ, Fife D, Troy SM (1991) Stochastic methods for short term projections of symptomatic HIV disease. Stat Med 10(9):1427–1440

Nguyen HL, Duong TH, Nguyen CP, Nguyen DC, Chiem TP, Nguyen MH, Nguyen TNM, Nguyen HV (2017) Specific k-mean clustering-based perceptron for dengue prediction. Int J Intell Inf Database Syst 10(3–4):269–288

Nobre FF, Monteiro ABS, Telles PR, Williamson GD (2001) Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology. Stat Med 20(20):3051–3069

Nunes B, Natário I, Lucília Carvalho M (2013) Nowcasting influenza epidemics using non-homogeneous hidden Markov models. Stat Med 32(15):2643–2660

Ongsulee P (2017) Artificial intelligence, machine learning and deep learning. In: 2017 15th International conference on ICT and knowledge engineering (ICT&KE). IEEE, pp 1–6

Paul A, Reja S, Kundu S, Bhattacharya S (2021) COVID-19 pandemic models revisited with a new proposal: plenty of epidemiological models outcast the simple population dynamics solution. Chaos Solitons Fractals 144:110697

Peng L-Z, Yi L-X, Hua S-Y (2008) A new epidemic disease predicting method. In: 2008 International conference on intelligent computation technology and automation (ICICTA), vol 1. IEEE, pp 550–553

Petropoulos F, Apiletti D, Assimakopoulos V, Babai MZ, Barrow DK, Ben Taieb S, Bergmeir C, Bessa RJ, Bijak J, Boylan JE, Browell J, Carnevale C, Castle JL, Cirillo P, Clements MP, Cordeiro C, Cyrino Oliveira FL, De Baets S, Dokumentov A, Ellison J, Fiszeder P, Franses PH, Frazier DT, Gilliland M, Gönül MS, Goodwin P, Grossi L, Grushka-Cockayne Y, Guidolin M, Guidolin M, Gunter U, Guo X, Guseo R, Harvey N, Hendry DF, Hollyman R, Januschowski T, Jeon J, Jose VRR, Kang Y, Koehler AB, Kolassa S, Kourentzes N, Leva S, Li F, Litsiou K, Makridakis S, Martin GM, Martinez AB, Meeran S, Modis T, Nikolopoulos K, Önkal D, Paccagnini A, Panagiotelis A, Panapakidis I, Pavía JM, Pedio M, Pedregal DJ, Pinson P, Ramos P, Rapach DE, Reade JJ, Rostami-Tabar B, Rubaszek M, Sermpinis G, Shang HL, Spiliotis E, Syntetos AA, Talagala PD, Talagala TS, Tashman L, Thomakos D, Thorarinsdottir T, Todini E, Trapero Arenas JR, Wang X, Winkler RL, Yusupova A, Ziel F (2022) Forecasting: theory and practice. Int J Forecast 38(3):705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001

Porta M (2014) A dictionary of epidemiology. Oxford University Press, United States of America

Pradhan A, Anasuya A, Pradhan MM, Ak K, Kar P, Sahoo KC, Panigrahi P, Dutta A (2016) Trends in malaria in Odisha, India—an analysis of the 2003–2013 time-series data from the national vector borne disease control program. PLoS ONE 11(2):0149126

Ramos ACV, Gomes D, Santos Neto M, Berra TZ, Yamamura M, Crispim JdA, Martoreli Junior JF, Bruce ATI, Dos Santos FL (2020) Trends and forecasts of leprosy for a hyperendemic city from Brazil's northeast: evidence from an eleven-year time-series analysis. PLoS ONE 15(8):0237165

Ray EL, Sakrejda K, Lauer SA, Johansson MA, Reich NG (2017) Infectious disease prediction with kernel conditional density estimation. Stat Med 36(30):4908–4929

Ribeiro MHDM, Mariani VC, dos Santos Coelho L (2020) Multi-step ahead meningitis case forecasting based on decomposition and multi-objective optimization methods. J Biomed Inform 111:103575

Samuel AL (1959) Some studies in machine learning using the game of checkers. IBM J Res Dev 3(3):210–229

Santos L, Costa M, Pinho STRd, Andrade RFS, Barreto FR, Teixeira M, Barreto ML (2009) Periodic forcing in a three-level cellular automata model for a vector-transmitted disease. Phys Rev E 80(1):016102

Shcherbakov MV, Brebels A, Shcherbakova NL, Tyukov AP, Janovsky TA, Kamaev VA et al (2013) A survey of forecast error measures. World Appl Sci J 24(24):171–176

Shen Y, Jiang C, Dun Z (2008) Analysis and prediction of epidemiological trend of scarlet fever from 1957 to 2004 in the downtown area of Beijing. In: International workshop on biosurveillance and biosecurity. Springer, Berlin, pp 164–168

Smirnova A, Chowell G (2017) A primer on stable parameter estimation and forecasting in epidemiology by a problem-oriented regularized least squares algorithm. Infect Dis Model 2(2):268–275

Smirnova A, Sirb B, Chowell G (2019) On stable parameter estimation and forecasting in epidemiology by the Levenberg–Marquardt algorithm with Broyden's rank-one updates for the Jacobian operator. Bull Math Biol 81(10):4210–4232

Soebiyanto RP, Adimi F, Kiang RK (2010) Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. PLoS ONE 5(3):9450

Stolerman LM, Maia PD, Kutz JN (2019) Forecasting dengue fever in Brazil: an assessment of climate conditions. PLoS ONE 14(8):0220106

Suparit P, Wiratsudakul A, Modchang C (2018) A mathematical model for zika virus transmission dynamics with a time-dependent mosquito biting rate. Theor Biol Med Model 15(1):1–11

Tabataba FS, Lewis B, Hosseinipour M, Tabataba FS, Venkatramanan S, Chen J, Higdon D, Marathe M (2017) Epidemic forecasting framework combining agent-based models and smart beam particle filtering. In: 2017 IEEE International conference on data mining (ICDM). IEEE, pp 1099–1104

Talkhi N, Fatemi NA, Ataei Z, Nooghabi MJ (2021) Modeling and forecasting number of confirmed and death caused COVID-19 in Iran: a comparison of time series forecasting methods. Biomed Signal Process Control 66:102494

Towers S, Chowell G (2012) Impact of weekday social contact patterns on the modeling of influenza transmission, and determination of the influenza latent period. J Theor Biol 312:87–95

Tsoularis A, Wallace J (2002) Analysis of logistic growth models. Math Biosci 179(1):21–55

Valeri L, Patterson-Lomba O, Gurmu Y, Ablorh A, Bobb J, Townes FW, Harling G (2016) Predicting subnational Ebola virus disease epidemic dynamics from sociodemographic indicators. PLoS ONE 11(10):0163544

Vandeput N (2021) Data science for supply chain forecasting. De Gruyter, Berlin. https://doi.org/10.1515/9783110671124

Wang X, Panchanathan S, Chowell G (2013) A data-driven mathematical model of CA-MRSA transmission among age groups: evaluating the effect of control interventions. PLoS Comput Biol 9(11):1003328

Wang Y, Xu C, Zhang S, Yang L, Wang Z, Zhu Y, Yuan J (2019) Development and evaluation of a deep learning approach for modeling seasonality and trends in hand-foot-mouth disease incidence in mainland China. Sci Rep 9(1):1–15

Wang P, Zheng X, Li J, Zhu B (2020) Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos Solitons Fractals 139:110058

White P (2006) Epidemics and pandemics: their impacts on human history. Reference Reviews

Winters PR (1960) Forecasting sales by exponentially weighted moving averages. Manag Sci 6(3):324–342

Wu W, Guo J, An S, Guan P, Ren Y, Xia L, Zhou B (2015) Comparison of two hybrid models for forecasting the incidence of hemorrhagic fever with renal syndrome in Jiangsu Province, China. PLoS ONE 10(8):0135492

Wu H, Wang X, Xue M, Wu C, Lu Q, Ding Z, Zhai Y, Lin J (2018) Spatial-temporal characteristics and the epidemiology of haemorrhagic fever with renal syndrome from 2007 to 2016 in Zhejiang Province, China. Sci Rep 8(1):1–14

Wu Y, Yang Y, Nishiura H, Saitoh M (2018) Deep learning for epidemiological predictions. In: The 41st international ACM SIGIR conference on research & development in information retrieval, pp 1085–1088

Yamey G, Schäferhoff M, Aars OK, Bloom B, Carroll D, Chawla M, Dzau V, Echalar R, Gill IS, Godal T et al (2017) Financing of international collective action for epidemic and pandemic preparedness. Lancet Glob Health 5(8):742–744

Yang W, Karspeck A, Shaman J (2014) Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. PLoS Comput Biol 10(4):1003583

Yang Y, Peng F, Wang R, Guan K, Jiang T, Xu G, Sun J, Chang C (2020) The deadly coronaviruses: the 2003 SARS pandemic and the 2020 novel coronavirus epidemic in China. J Autoimmun 109:102434

Yule GU (1927) Vii. On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers. Philos Trans R Soc Lond Ser A Contain Pap Math Phys Character 226(636–646):267–298

Zhang C, Fu X, Zhang Y, Nie C, Li L, Cao H, Wang J, Wang B, Yi S, Ye Z (2019) Epidemiological and time series analysis of haemorrhagic fever with renal syndrome from 2004 to 2017 in Shandong Province, China. Sci Rep 9(1):1–9

Zhao Y, Ge L, Zhou Y, Sun Z, Zheng E, Wang X, Huang Y, Cheng H (2018) A new seasonal difference space-time autoregressive integrated moving average (SD-STARIMA) model and spatiotemporal trend prediction analysis for hemorrhagic fever with renal syndrome (HFRS). PLoS ONE 13(11):0207518