Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(5): 9080–9100.
DOI: 10.3934/mbe.2023399
Received: 24 November 2022
Revised: 21 February 2023
Accepted: 01 March 2023
Published: 13 March 2023

Research article

Modeling COVID-19 pandemic with financial markets models: The case of Jaén (Spain)

Julio Guerrero[1], **María del Carmen Galiano**[1] **and Giuseppe Orlando**[1,2,3,*]

[1] Department of Mathematics, University of Jaén, Campus de las Lagunillas s/n, Jaén 23071, Spain

[2] Department of Mathematics, University of Bari, via Edoardo Orabona 4, Bari 70125, Italy

[3] Department of Economics, HSE University, 16 Soyuza Pechatnikov Street, St Petersburg 190121, Russia

* **Correspondence:** Email: giuseppe.orlando@uniba.it; Tel: +39 080 5049218.

Abstract: The main objective of this work is to test whether some stochastic models typically used in financial markets could be applied to the COVID-19 pandemic. To this end, we have implemented the ARIMAX and Cox-Ingersoll-Ross (CIR) models originally designed for interest rate pricing but transformed by us into a forecasting tool. For the latter, which we denoted CIR*, both the Euler-Maruyama method and the Milstein method were used. Forecasts obtained with the maximum likelihood method have been validated with 95% confidence intervals and with statistical measures of goodness of fit, such as the root mean square error (RMSE). We demonstrate that the accuracy of the obtained results is consistent with the observations and sufficiently accurate to the point that the proposed CIR* framework could be considered a valid alternative to the classical ARIMAX for modelling pandemics.

Keywords: COVID-19; forecasting; Cox-Ingersoll-Ross model; ARIMAX; Milstein method

1. Introduction

Coronavirus disease 2019 (COVID-19) is a lung disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). In December 2019, the Chinese authorities reported different cases of this virus in Wuhan. This disease spread rapidly throughout the world from less than 30 cases at the end of December 2019 to more than 8,455,738 confirmed cases on June 20, 2021.

The first case in Spain was a German tourist, on January 31, 2021. Since then, several cases were confirmed throughout the country. In Andalusia, the first positive case was reported in Seville on February 26, 2020. Two days later the first case was confirmed in the province of Jaén, and on March 6 the first case of coronavirus in the city of Jaén.

As the days went by, due to the high number of infections in the country, the state of alarm was decreed and a national lockdown was imposed on March 14, which was extended until June 21. The application of measures such as the use of a mask, perimeter confinements, the closure of non-essential services, etc., reduced the infection rate two weeks after the declaration of quarantine. During these years, many researchers from various disciplines have used various modelling tools to analyze the impact of the pandemic at the global and local levels. In our case, we are going to focus on modelling the contagion in Jaén in two different ways. The first approach is based on the Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) model [1] which, we found, provides better performances than the Autoregressive Integrated Moving Average (ARIMA) (on the same line, see also [2–6]). This is a common model in time series forecasting and is often adopted in finance [7–10].

The second approach is based on the Cox, Ingersoll and Ross (CIR) model. This is a model designed for interest rate pricing that we turn into a forecasting tool. The advantage of this model lies in its simplicity (being a single factor model) and analytical tractability [11]. The CIR model is mean reverting and is a squared Ornstein–Uhlenbeck process [12]. Notice that the "Ornstein-Uhlenbeck process is a natural model to consider in a biological context because it stabilizes around some equilibrium point. This corresponds to the homeostasis often observed in biology" [13], and the quasi-stationarity is relevant with reference to mortality plateaus [13]. Furthermore, we point out a connection to models for short-term interest rates in financial modelling.

We prove that the proposed transformation of the CIR model, which we have denoted CIR*, outperforms the classical ARIMAX. For this purpose both the Euler-Maruyama method and the Milstein method were used, and this is one of the main contributions of the present study. Notice that the suggested approach not only extends the models available to scholars to model pandemics but, also, paves the way for similar approaches where financial models can be converted into econometric models.

For the implementation of the models on real-world data, we use the data from the moving averages for 14 days of the daily cases of the city of Jaén, that is, each value represents the average number of people infected in each of the previous 14 days. This is because, due to the weekend effect and occasional misreporting, we found that the moving average is a more reliable target. This is due to two reasons: a) the relatively small size of the city of Jaén, which affects the number of cases, and b) the effect of the weekend, when reporting is altered. The two effects lead to highly irregular behaviour in the time series considered. Also, the 14-day moving average is the standard way of reporting data in Spain during the pandemic, both for publishing data and deciding the different measures (lockdowns, mobility restrictions, etc.). Over that time series, we estimate the parameters of the considered models, in such a way that they best fit the data. The forecasts have been validated with 95% confidence intervals and with statistical measures of goodness of fit, such as the RMSE.

This article is organized as follows. Section 2 briefly summarizes the literature. Section 3 reports the data and describes the CIR model, as well as the methodology on which it is based. That is followed by an explanation of the suggested adaptation to forecasting, its calibration and a sketch of the ARIMAX model. Section 4 shows the obtained results of the two models by comparing them. Section 5 concludes.

2. Literature review

Among those works that adopted the ARIMA (and the like models) to estimate the cases of the COVID-19 pandemic, we mention Ekinci [14], who considered data from USA, India, Brazil, France, Russia, UK, Italy, Spain and Germany. When comparing ARMA-GARCH, ARMA-TGARCH and ARMA-EGARCH models, it was found that while considering the conditional variance effect improves the forecasting power, the asymmetric effect (such as asymmetric GARCH models) has mixed results. Sahai et al. [15] adopted the ARIMA for analyzing the trend of COVID-19 cases in Spain, Italy, France, Germany and the US. The authors claim that their model provides considerable forecast accuracy and could be a useful tool for governments to ramp up their healthcare preparations. Subramaniam et al. [8] drew a parallel between forecasting stock prices and cases of the pandemic by means of the ARIMA model. Then, they explored the correlation between the predictive efficiency of the ARIMA model and variation in the data. Katoch et al. [16] adopted an ARIMA model to analyzed the temporal dynamics of the COVID-19 outbreak in India from 30 January 2020 to 16 September 2020. Their approach suggests "varying epidemic's inflexion point and final size for underlying states and the mainland, India". Regarding the alternative between ARIMA and ARIMAX, in the literature, it has been found that the latter may yield better forecast compared to the seasonal ARIMA (SARIMA) model and Neural Networks (e.g., see Suhartono [17]). This is because the ARIMAX is most suited to deal with calendar effects [2–5, 18].

As regards the Cox, Ingersoll and Ross (CIR) model [19, 20], as already mentioned, it has been proposed for the pricing of interest rates. At the time of its introduction, it quickly gained popularity in finance because it was perceived as "an improvement on the Vasicek model [21], not allowing for negative rates and introducing rate-dependent volatility, as well as for its relatively handy implementation and analytical tractability" [11]. The CIR process is a variant of a squared Bessel process and, in the physical community, its analogues were used both to enrich the list of modified stochastic processes used for the description of various time series [22, 23] as well as to model some financial historical data [24].

Other applications of the CIR model include stochastic volatility modelling in option pricing problems [25, 26], FX [27] or default intensities in credit risk [28, 29]. As mentioned, the Ornstein–Uhlenbeck (OU) and the Cox–Ingersoll–Ross (CIR) processes are strongly linked, and the reflected OU (ROU) [30] is used in queuing theory [30, 31], in population dynamics [32], catastrophes [33, 34], etc. In this study, similar to what has been done by Orlando et al. [11, 35–38] when developing the CIR# model, we transform the original CIR model into a forecasting tool and compare its performance with that of the well-known ARIMAX model.

3. Materials and methods

3.1. Data

The available data are the moving averages of 14 days of infections in the city of Jaén, from February 2, 2020 to October 8, 2021 and have been provided by the Health and Family Council of the Andalusian Regional Government (see Figure 1). Jaén is a relatively small city (around 110,000 inhabitants), which represents a challenge since the number of COVID-19 daily cases is small and with large fluctuations, even after the moving averages are computed. Therefore, we shall have the

opportunity of testing different methods in unfavourable circumstances.

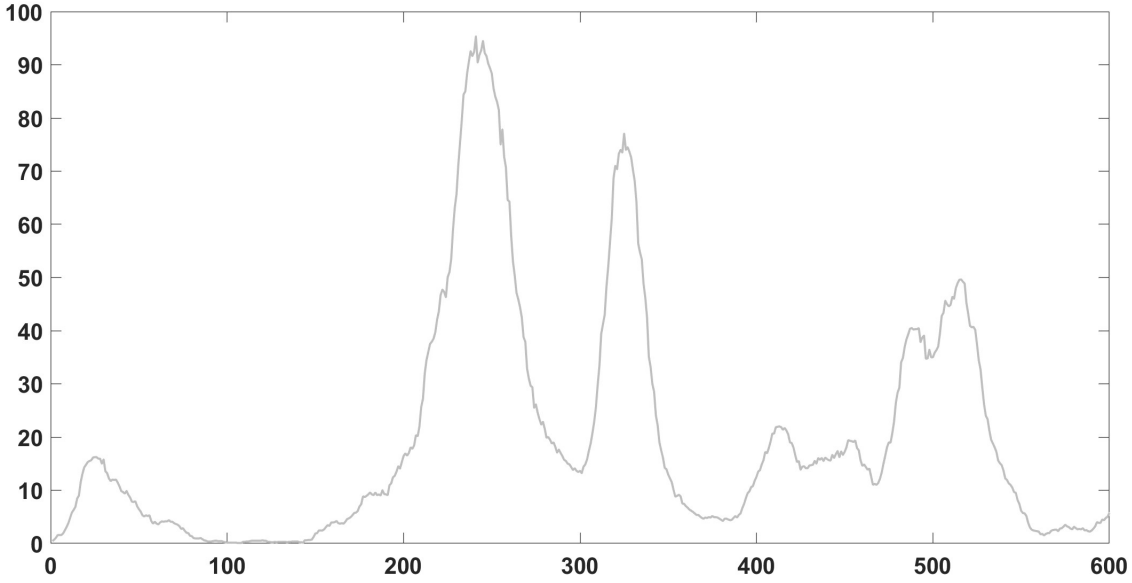

Figure 1. 14-day moving average of the daily cases of COVID-19 in the city of Jaén (ordinate). The abscissa represents the days elapsed since February 2, 2020.

A visual inspection employing the partial autocorrelation function (PACF) confirms that the data is correlated (see Figure 2).

3.2. ARIMAX model

Leaving aside the cases where data show evidence of non-stationarity where an initial differencing removes the integrated (I) part, the ARMAX model can be described as:

$$\begin{aligned} y(t) + a_1 y(t-1) + \ldots + a_{n_a} y(t-n_a) = \\ b_1 x(t-n_k) + \ldots + b_{n_b} x(t-n_k-n_b+1) + \\ c_1 \varepsilon(t-1) + \ldots + c_{n_c} \varepsilon(t-n_c) + \epsilon(t) \end{aligned} \tag{3.1}$$

with $y(t)$, dependent/output variable at time t; n_a, number of poles; n_b number of zeroes plus 1; n_c, number of c coefficients; n_k, dead time in the system. Moreover, $y(t-1)\ldots y(t-n_a)$ denotes the dependence between the current output and the previous outputs, $x(t-n_k)\ldots x(t-n_k-n_b+1)$ indicates the dependence between the current output and both the previous and delayed inputs, and $\epsilon(t)$ expresses a white-noise error.

The orders of the ARMAX model are given by the parameters n_a, n_b and n_c. n_k is the delay, and q is the delay operator. The ARMAX in compact form can be written as

$$A(q)y(t) = B(q)x(t-n_k) + C(q)\varepsilon(t) \tag{3.2}$$

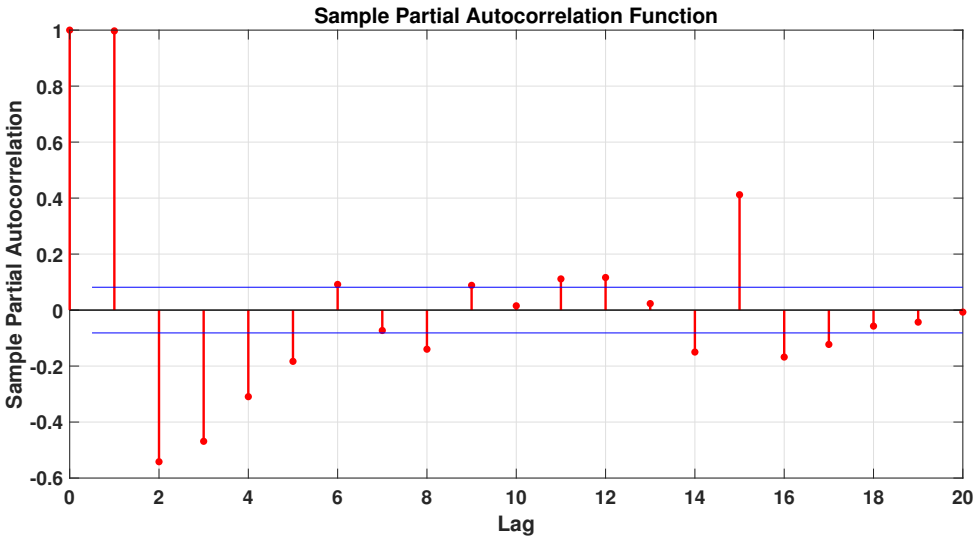

Figure 2. Partial autocorrelation function (PACF) over a 14-day moving average of the daily cases of COVID-19 in the city of Jaén: ordinate PACF, abscissa lags. Notice the autocorrelation at lags 1, 2, 3, 4, 5, 14, 15 and 16.

such that,

$$A(q) = 1 + a_1 q^{-1} + \ldots + a_{n_a} q^{-n_a}$$
$$B(q) = b_1 + b_2 q^{-1} + \ldots + b_{n_b} q^{-n_b+1}$$
$$C(q) = 1 + c_1 q^{-1} + \ldots + c_{n_c} q^{-n_c}.$$

The ARIMAX model can be seen as a generalization of the ARIMA because adds to the structure above described an integrator in the white noise $\varepsilon(t)$ as follows:

$$A(q)y(t) = B(q)x(t - nk) + \frac{C(q)}{(1 - q^{-1})}\varepsilon(t). \tag{3.3}$$

3.2.1. Estimation of the ARIMAX model

To estimate the ARIMAX model the following steps have been performed.

1) Ensure stationarity of the time series by conducting Augmented Dicky Fuller (ADF) Test.

2) Model identification, i.e., specification of the autoregressive (AR) and moving average (MA) terms with the help of the autocorrelation function (ACF) and partial autocorrelation function (PACF).

3) Parameter estimation according to Ljung [39] and related implementation in Matlab [40]. The best model is selected based on Akaike information criterion (AIC) values [41].

From now on, we refer to the ARMAX model (and not to the ARIMAX) assuming that the integration (I) has been removed.

3.3. CIR* model

As mentioned, the CIR model emerged in 1985 from the hand of John C. Cox, Jonathan E. Ingersoll and Stephen A. Ross [19,20] as an improvement of the Vasicek model to prevent negative interest rates.

The CIR model is defined by the following stochastic differential equation (SDE):

$$\begin{cases} dX_t = \alpha\,(\mu - X_t)\,dt + \sigma\,\sqrt{X_t}dW_t \\ X_0 = x_0 \end{cases} \tag{3.4}$$

Here, α, μ and σ are positive constants, $X(t)$ is the interest rate, t is time, and W_t denotes the standard Wiener process.

The parameters include the following:

- $\alpha(\mu - X_t)$ is the same factor as in Vasicek's model and represents a mean reversion term.
- The standard deviation factor $\sigma\,\sqrt{X_t}$ removes negative rates.
- $\sqrt{X_t}$ increases the standard deviation as the short-term rate increases.

This model can only have positive solutions since when the interest rate is 0 it ends up being positive later on. Also, when it is low or close to 0, the standard deviation is close to 0.

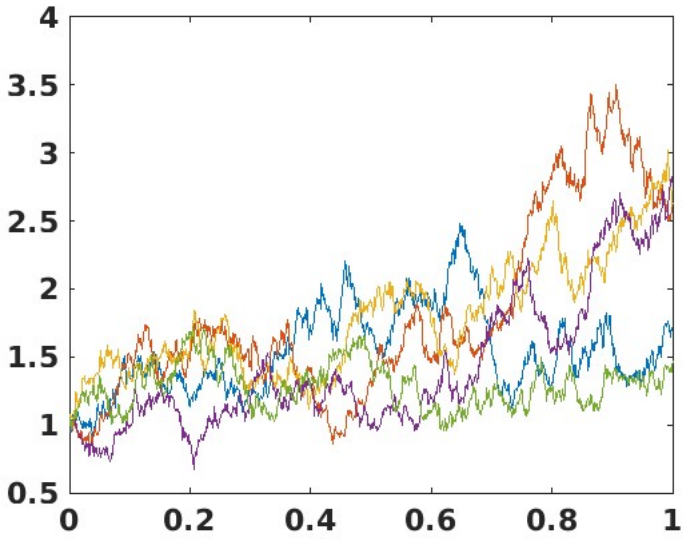

Figure 3. Simulated paths of the CIR model. Ordinate simulated variable X_t, abscissa t.

The only solution to (3.4) is what is known as the CIR process. Integrating Eq (3.4):

$$X_t = X_s + \alpha \int_s^t (\mu - X_u)du + \sigma \int_s^t \sqrt{X_u}dW_u, \qquad s < t\,, \tag{3.5}$$

and therefore

$$E[X_t|X_s] = X_s + \alpha \int_s^t (\mu - E[X_u|X_s])du, \qquad s < t\,.$$

If we call $m_t = E[X_t|X_s]$ we have

$$\frac{d}{dt}m_t = \alpha(\mu - m_t), \qquad s < t\,,$$

whose solution is

$$m_t = X_s e^{-\alpha(t-s)} + \mu(1 - e^{-\alpha(t-s)})\,.$$

So

$$E[X_t|X_s] = X_s e^{-\alpha(t-s)} + \mu(1 - e^{-\alpha(t-s)}), \qquad s < t\,, \tag{3.6}$$

and therefore

$$E[X_t|X_s] - \mu = (X_s - \mu)e^{-\alpha(t-s)}, \qquad s < t\,. \tag{3.7}$$

Thus, $E[X_t|X_s] - \mu$ has the same sign as $X_s - \mu$. In addition, if $\mu > 0$ and $\alpha > 0$, starting with $X_s > 0$ we conclude that $X_t > 0$.

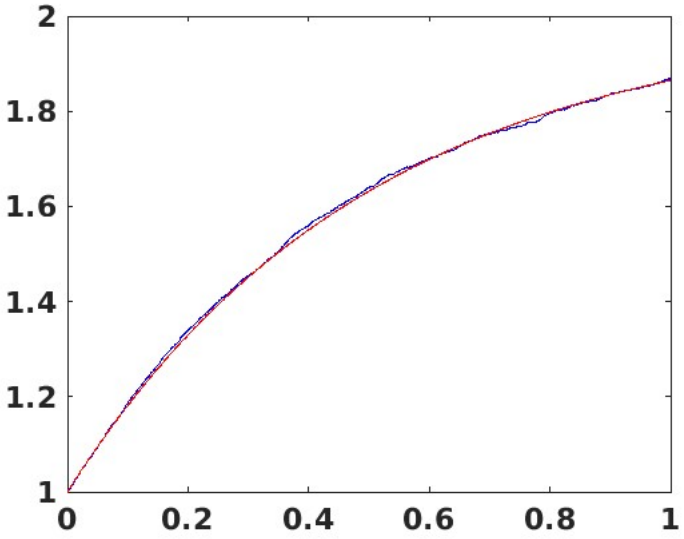

Figure 4. Numerical (blue) and theoretical (red) mean comparison of the CIR model with Euler's method for 1000 simulated paths, with $x_0 = 1$, $\alpha = 2$, $\mu = 2$ and $\sigma = 1$ and 5000 subintervals. Ordinate m_t, abscissa t.

Similarly, the variance is:

$$Var[X_t|X_s] = \frac{X_s\sigma^2}{\alpha}(e^{-\alpha(t-s)} - e^{-2\alpha(t-s)}) + \frac{\mu\sigma^2}{2\alpha}(1 - e^{-\alpha(t-s)})^2 \tag{3.8}$$

As stated at the beginning, the fundamental advantage of this model is that the solutions are non-negative. However, the distribution of the CIR model is not Gaussian, which makes it difficult to analyze.

The density function is given by:

$$f(X_s, s, X_t, t) = ce^{-(u+v)}\left(\frac{v}{u}\right)^{\frac{q}{2}} I_q(2\sqrt{uv})$$

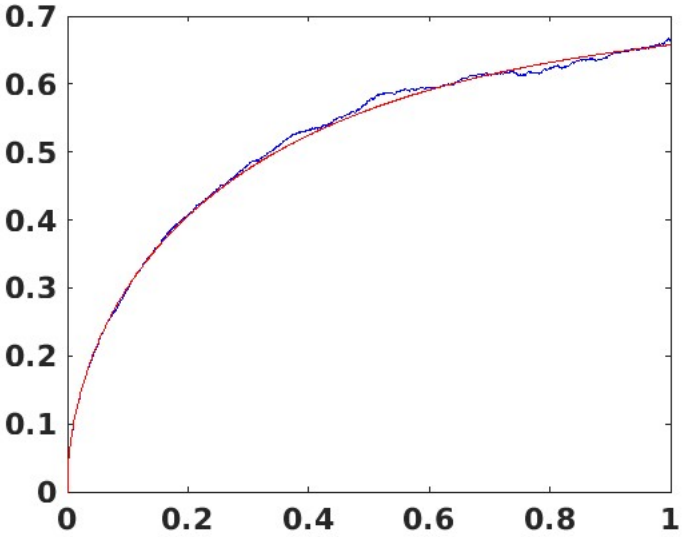

Figure 5. Comparison of the numerical (blue) and theoretical (red) standard deviation of the CIR model with the Euler method for 1000 simulated paths, with $x_0 = 1$, $\alpha = 2$, $\mu = 2$ and $\sigma = 1$ and 5000 subintervals. Ordinate σ_t, abscissa t.

where

$$c = \frac{2\alpha}{\sigma^2(1 - e^{-\alpha\Delta t})}$$
$$u = cX_s e^{-\alpha\Delta t}$$
$$v = cX_t$$
$$q = \frac{2\alpha\mu}{\sigma^2} - 1$$
$$\Delta t = t - s$$

$I_q(\cdot)$ is a Bessel function of first type and order q:

$$I_q(x) = \sum_{j=0}^{\infty} \left(\frac{x}{2}\right)^{2j+q} \frac{1}{k!\Gamma(j + q + 1)}$$

where Γ is the gamma function.

Let $z_t = 2cX_t$. Then, the conditional distribution of z_t given z_s is a non-central *chi*-squared distribution $\chi^2_d(2u)$, with $d = \frac{4\alpha\mu}{\sigma^2}$ degrees of freedom, and the non-centrality parameter is $\lambda = 2u$.

Therefore,

$$z_t|z_s \sim \chi^2(d, \lambda)$$

where:

$$d = \frac{4\alpha\mu}{\sigma^2}$$
$$\lambda = \frac{4\alpha}{\sigma^2(1 - e^{-\alpha\Delta t})} e^{-\alpha\Delta t} X_s$$

Since $z_t = 2cX_t$, X_t conditional on X_s has the same distribution as $z_t/2c$ conditional on $z_s/2c$. So,

$$X_t|X_s \sim \frac{z_t}{2c}|\frac{z_t}{2c} \sim \frac{1}{2c}\chi^2(d,\lambda)$$

At this point, we are going to analyze the different behaviours that the deterministic solution of the CIR model equation can present in terms of the relations among the different parameters, which will be useful later to give an interpretation of the model parameters, although we know that the stochastic part would give oscillations with respect to said behaviour. We shall allow in this analysis for negative values of α since in certain regions of the data the calibrated values of α result in negative values. According to Eq (3.7) we distinguish two cases, depending on whether X_t is greater or less than μ and in each case two subcases, depending on whether α is positive or negative:

1) If $X_t < \mu$:

- If $\alpha > 0$, X_t approaches μ from below.
- If $\alpha < 0$, X_t moves away from μ downwards.

2) If $X_t > \mu$:

- If $\alpha > 0$, X_t approaches μ from above.
- If $\alpha < 0$, X_t moves away from μ upwards.

3.3.1. Estimation of the parameters

To approximate the data well, it is necessary to give a good adjustment of the parameters. In the case of the CIR* model, we must estimate three parameters, α, μ and σ. We will generally refer to them as the parameter vector $\theta \equiv (\alpha, \mu, \sigma)$. The procedure that will be followed to estimate the parameters is the one shown in [42], and it is the maximum likelihood method (MLE), which is based on maximizing the objective function under consideration. For the maximum likelihood estimation of the parameter vector $\theta \equiv (\alpha, \mu, \sigma)$, the transition densities are required. The CIR process is one of the processes for which we know its density function explicitly. Given X_t at time t, the density of $X_{t+\Delta t}$ at time $t + \Delta t$ is:

$$p(X_{t+\Delta t}|X_t;\theta,\Delta t) = ce^{-(u+v)}\left(\frac{v}{u}\right)^{\frac{q}{2}} I_q(2\sqrt{uv})$$

where

$$\begin{aligned} c &= \frac{2\alpha}{\sigma^2(1-e^{-\alpha\Delta t})} \\ u &= cX_t e^{-\alpha\Delta t} \\ v &= cX_{t+\Delta t} \\ q &= \frac{2\alpha\mu}{\sigma^2} - 1 \end{aligned}$$

where $I_q(2\sqrt{uv})$ is a Bessel function.

The likelihood function for time series with N observations is:

$$L(\theta) = \prod_{i=1}^{N-1} p(X_{t_{i+1}}|X_{t_i}; \theta, \Delta t) \tag{3.9}$$

To simplify the calculations, it is usual to work with the log-likelihood expression, which consists of taking logarithms in the Eq (3.9).

$$\ln L(\theta) = \sum_{i=1}^{N-1} \ln p(X_{t_{i+1}}|X_{t_i}; \theta, \Delta t) \tag{3.10}$$

from which the log-likelihood function of the CIR process can be easily derived.

$$\ln L(\theta) = (N-1)\ln c + \sum_{i=1}^{N-1}\left[-u_{t_i} - v_{t_{i+1}} + \frac{1}{2}q\ln\left(\frac{v_{t_{i+1}}}{u_{t_i}}\right) + \ln\left(I_q\left(2\sqrt{u_{t_i}v_{t_{i+1}}}\right)\right)\right] \tag{3.11}$$

where $u_{t_i} = cX_{t_i}e^{-\alpha\Delta t}$ y $v_{t_{i+1}} = cX_{t_{i+1}}$

To find the maximum likelihood estimate $\hat{\theta}$ of the parameter vector θ, we have to maximize the function (3.11) over its parameter space.

$$\hat{\theta} = (\hat{\alpha}, \hat{\mu}, \hat{\sigma}) = \arg\max_{\theta} \ln L(\theta) \tag{3.12}$$

Since the logarithm function is monotonically increasing, maximizing the log-likelihood function is equivalent to maximizing the likelihood function.

To solve the problem (3.12), we resort to numerical computation. For global optimal convergence, the initial optimization points are essential, for which the method of least squares will be used. We first write the equation of the discretized CIR*:

$$X_{t+\Delta t} - X_t = \alpha(\mu - X_t)\Delta t + \sigma\sqrt{X_t}W_t \tag{3.13}$$

where W_t is distributed with zero mean and variance Δt.

Dividing Eq (3.13) by $\sqrt{X_t}$, we get

$$\frac{X_{t+\Delta t} - X_t}{\sqrt{X_t}} = \frac{\alpha\mu\Delta t}{\sqrt{X_t}} - \alpha\sqrt{X_t}\Delta t + \sigma W_t \tag{3.14}$$

Based on Eq (3.14), the initial values of $\hat{\alpha}$ and $\hat{\mu}$ are found by minimizing the function:

$$(\hat{\alpha}, \hat{\mu}) = \arg\min_{\alpha,\mu} \sum_{i=1}^{N-1}\left(\frac{X_{t_{i+1}} - X_{t_i}}{\sqrt{X_{t_i}}} - \frac{\alpha\mu\Delta t}{\sqrt{X_{t_i}}} + \alpha\sqrt{X_{t_i}}\Delta t\right)^2$$

The exact expression of the solution is on page 3 of [42]. The estimate of $\hat{\sigma}$ is the standard deviation of the residuals.

To optimize the objective function (3.11), we need to evaluate the Bessel function $I_q(2\sqrt{uv})$. The function *besseli* is implemented in Matlab, but this usually causes problems, because the function $I_q = (2\sqrt{uv})$ approaches infinity very quickly. Fortunately, Matlab allows us to give another scaled version, which we will call $I_q^1(2\sqrt{uv})$, which solves the divergence problem in such a way that

$$I_q^1(2\sqrt{uv}) = I_q(2\sqrt{uv})\exp(-2\sqrt{uv}),$$

and therefore

$$I_q(2\sqrt{uv}) = \frac{I_q^1(2\sqrt{uv})}{\exp(-2\sqrt{uv})}.$$

Rewriting the expression (3.11), we get

$$\ln L(\theta) = (N-1)\ln c + \sum_{i=1}^{N-1}(-u_{t_i} - v_{t_{i+1}} + \frac{1}{2}\ln\left(\frac{v_{t_{i+1}}}{u_{t_i}}\right) + \\ + \ln\left(I_q^1(2\sqrt{u_{t_i}v_{t_{i+1}}})\right) + 2\sqrt{u_{t_i}v_{t_{i+1}}})$$

3.3.2. Numerical methods

To obtain an approximation of the exact solution of the CIR SDE we need to establish a numerical scheme. In our case, we will implement the Euler-Maruyama and Milstein numerical schemes, to see later if there are notable differences between them.

The Euler-Maruyama scheme or method is an extension of Euler's method for ordinary differential equations to SDEs.

Let $\{X_t, 0 \le t \le T\}$ be an Itô process solution of the following SDE:

$$\begin{cases} dX_t = f(t, X_t)dt + g(t, X_t)dW_t \\ X_0 = x_0 \end{cases} \tag{3.15}$$

where $W(t)$ represents the Wiener process, and suppose we want to solve this SDE in the time interval $[0, T]$.

The Euler-Maruyama approximation Y_i to the true solution of X is defined as follows:

- Divide the interval $[0, T]$ in N subintervals of size $\Delta t = T/N$ being $0 = t_0 < t_1 < ... < t_N = T$.
- Set the initial condition $Y_0 = x_0$.
- Define recursively Y_i for $1 \le i \le N$,

$$Y_{i+1} = Y_i + f(t_i, Y_i)\Delta t + g(t_i, Y_i)\Delta W_i \tag{3.16}$$

where $\Delta W_i = W_{t_{i+1}} - W_{t_i}$.

The variables ΔW_i are independent and identically distributed normal random variables, that is, $\Delta W_i \sim \sqrt{\Delta t}Z$ with $Z \sim N(0, 1)$.

The Milstein method [43] is used to increase the accuracy of the Euler-Maruyama method. This is achieved by introducing a term of order 2 by using the partial derivative with respect to x of $g(t, x)$.

Given an Itô process $\{X_t, 0 \le t \le T\}$ which is a solution of the SDE (3.15), the approximation of the Milstein method Y_i to the true solution of X is given by the following:

- Divide the interval $[0, T]$ into subintervals of size $\Delta t = T/N$ with $0 = t_0 < t_1 < ... < t_n = N$.
- Take as initial condition $Y_0 = x_0$.

- Recursively define Y_i for $1 \leq i \leq N$ by:

$$Y_{i+1} = Y_i + f(t_i, Y_i)\Delta t + g(t_i, Y_i)\Delta W_i + \frac{1}{2}g(t_i, Y_i)\frac{\partial g(t_i, Y_i)}{\partial x}[(\Delta W_i)^2 - \Delta t] \tag{3.17}$$

where $\Delta W_i = W_{t_{i+1}} - W_{t_i}$.

The variables ΔW_i are independent and identically distributed normal random variables, that is, $\Delta W_i \sim \sqrt{\Delta t}Z$ with $Z \sim N(0, 1)$. As we can see, the expression of the scheme is the same as that of the Euler-Maruyama method, except that the following term is added:

$$\frac{1}{2}g(t_i, Y_i)\frac{\partial g(t_i, Y_i)}{\partial x}[(\Delta W_i)^2 - \Delta t]$$

Therefore, if $\frac{\partial g(t,x)}{\partial x}$ turns out to be 0, this method is equivalent to the Euler-Maruyama method. When a method satisfies $E(|Y_i - X(t_i)|) \leq K(\Delta t)^\gamma$ for some γ, that method is said to be a strong approximation of order γ. Applying this, the Euler-Maruyama method is a strong approximation of order $\gamma = 1/2$, while the Milstein method is a strong approximation of order $\gamma = 1$ if $f(t, X_t)$ and $g(t, X_t)$ are C^1 functions. The functions with which we are working in these models comply with this, so the order of convergence of the Milstein method will always be greater than that of Euler-Maruyama.

4. Results

Since we have daily data, we set the time step $\Delta t = 1$. To compare both the Euler-Maruyama and the Milstein methods, we chose a specific time window and a number of forecasts. A window of 100 data will be taken from the first real data available, and the next 500 days will be estimated. The prediction is made for the day following the last one of the windows. Then the window is moved one unit to the right, and the process is repeated until all the forecasts are performed.

Figure 6 shows a global view of the three estimates that we want to compare. Next, Figures 7 and 8 represent, for a given time window, the real data (in grey), the ARMAX model estimates (in magenta) and the ones calculated with the CIR* model (blue) with the Milstein method. Note that since the Euler-Maruyama and Milstein methods nearly overlap, the differences between the plotted curves cannot be seen unless zoomed in further. For this reason, we refer the reader to Table 1.

It can be seen that the blue curve generally fits the real data better, so it stands to reason that the CIR* model is better than the ARMAX model. To verify this rigorously, the root mean square error (RMSE) of the CIR* model has been calculated with the Euler-Maruyama and Milstein method, and the RMSE of the ARMAX model. These errors are collected in Table 1. As shown, the CIR* model gives better results than the ARMAX model.

Table 1. Comparison of CIR* model errors (Euler-Maruyama and Milstein) and ARMAX in the given time window.

	CIR* with Euler	CIR* with Milstein	ARMAX
RMSE	1.0556	1.0558	2.2157

To verify that the ARMAX filters correctly the data, the PACF depicted in Figure 9 shows that the errors/residuals are not correlated. This is further confirmed by the Durbin-Watson (DW) test. In

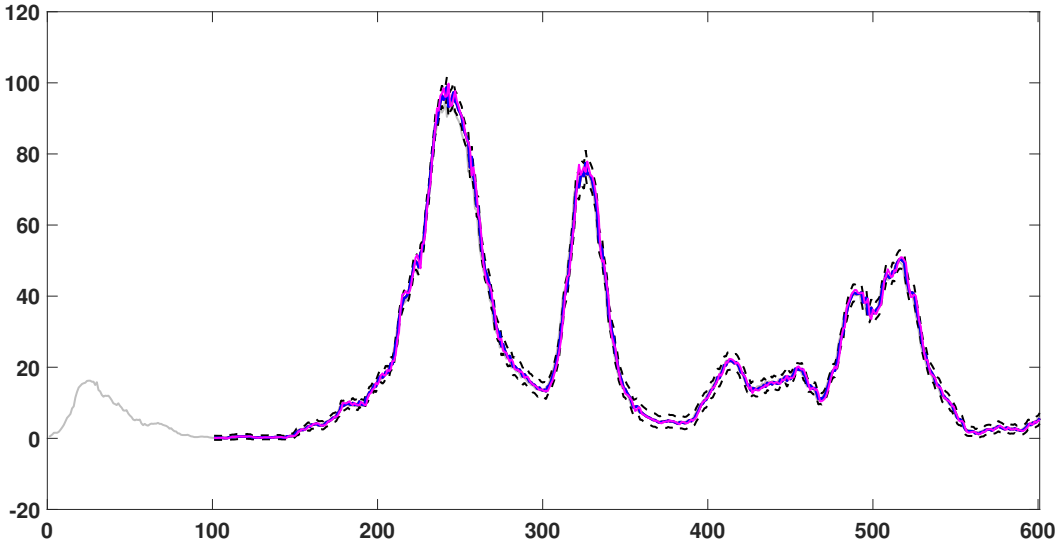

Figure 6. Approximations of the CIR* with Milstein method and ARMAX. Window = 100, time interval = 1, and number of simulations = 1000. Real cases (light grey), CIR* with Milstein method (blue), ARMAX (magenta), 95% confidence interval (dashed lines). Ordinate 14-day moving average of real and modeled COVID-19 cases, abscissa number of days elapsed since February 2, 2020.

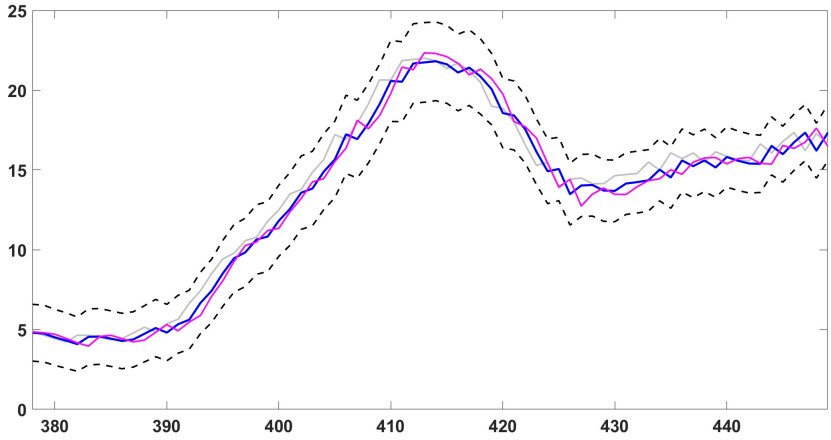

Figure 7. Zoomed-in comparison between CIR* and ARMAX model of increasing and decreasing number of infections. Real cases (light grey), CIR* with Milstein method (blue), ARMAX (magenta), 95% confidence interval (dashed lines). Ordinate 14-day moving average of real and modeled COVID-19 cases, abscissa number of days elapsed since February 2, 2020. Notice how the prediction stays within the dashed band.

fact, the DW statistic is 1.9868, and the p-value of 0.8753 suggests absence of autocorrelation in the residuals.

Finally, we are going to give an interpretation of the parameters of the CIR* model at different

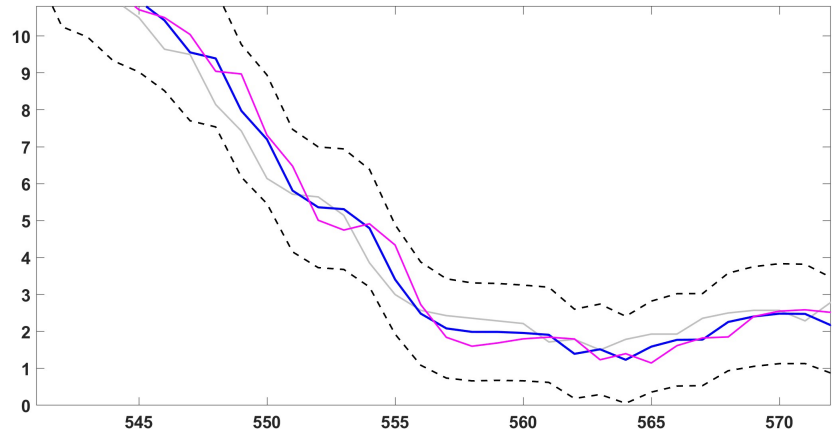

Figure 8. Zoomed-in comparison between CIR* and ARMAX model of a decreasing number of infections. Real cases (light grey), CIR* with Milstein method (blue), ARMAX (magenta), 95% confidence interval (dashed lines). Ordinate 14-day moving average of real and modeled COVID-19 cases, abscissa number of days elapsed since February 2, 2020. Notice how the prediction stays within the dashed band.

stages of the pandemic, based on the analysis of the deterministic solution of the CIR* model given in Section 3.3.1.

4.1. Time frame in which COVID-19 cases increase

We begin by studying a stage in which infections are increasing. For the estimated parameters to have the same trend, the window from which they are estimated must also be in the growth range, which will force us to take a small window and estimate a few values, since if we look at Figure 1, we see that the periods in which the infections grow are very short. Taking this into account, we are going to focus on the section that goes from day 210 to 229, that is, 20 forecasts, taking a window of 30 days. Based on the results obtained, we observe that both the mean of α and μ are negative. Since the estimated data values are greater than the mean, then X_t would move away from the mean upwards.

To corroborate that this is true, another stage of increase in cases of the pandemic has been taken, specifically from day 310 to 319, that is, 10 days, and a window of 30 days. The number of forecasts had to be reduced because, as previously mentioned, the window of days must be in the growth range for good analysis. In Figure 10 one can see both sections and in Table 2, the exact values of the mean of the parameters in the two stages.

Table 2. Growing stages of COVID-19 infections.

	Stage 1	Stage 2
Average α	-0.0562	-0.0990
Average μ	-3.6346	-3.5105
Average σ	0.2253	0.2007

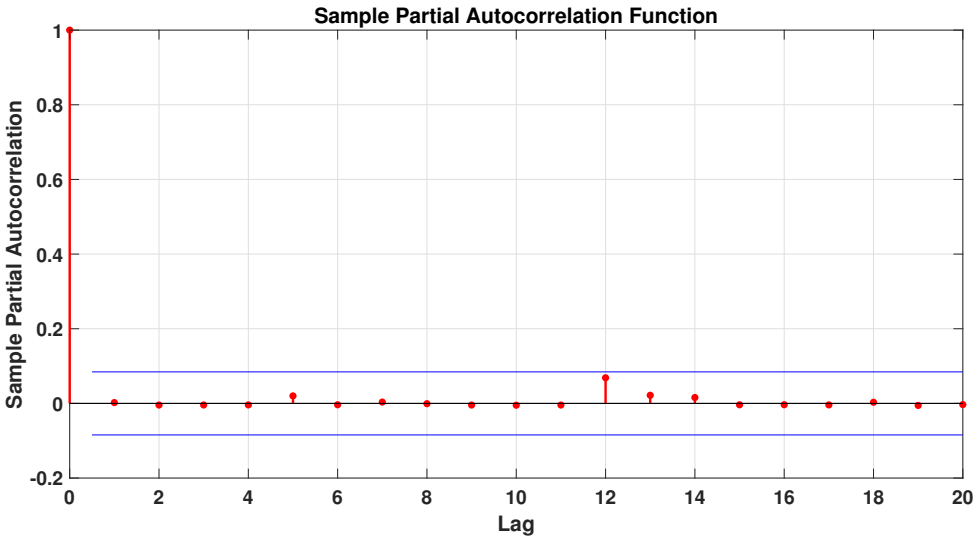

Figure 9. PACF over the error of the ARMAX model. Ordinate PACF, abscissa lags. Notice the absence of autocorrelation due to the ARMAX filtering.

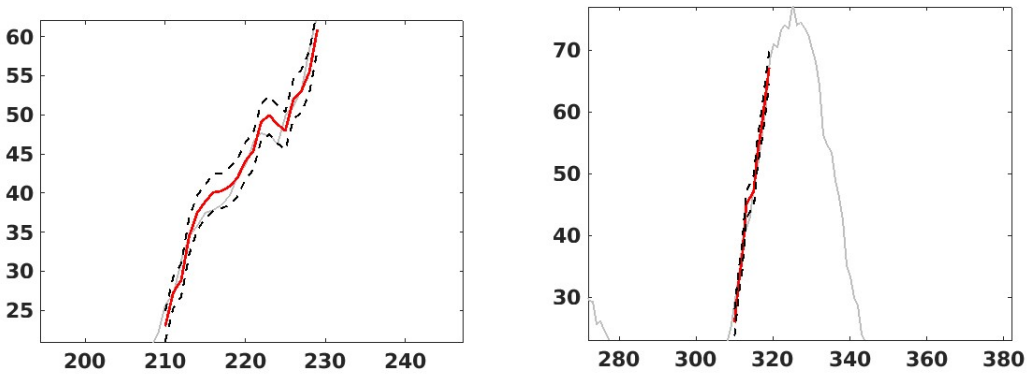

Figure 10. Stage 1 (left) and stage 2 (right) of a steep growth of the COVID-19 pandemic daily cases. Real cases (light grey), CIR* with Milstein method (red), 95% confidence interval (dashed lines). Ordinate 14-day moving average of real and modeled COVID-19 cases, abscissa number of days elapsed since February 2, 2020. Notice how the prediction stays within the dashed band.

4.2. Tme frame in which COVID-19 cases decrease

Keeping the ideas we used from the previous case, we now take a period in which the cases decrease. We estimate the interval that goes from day 265 to 284. In this case, we again obtain a negative mean of α, which makes sense, since in periods of strong growth or decrease the values move away from the mean. On the contrary, now the mean of the parameter μ is very large and exceeds the mean of infections in that section, therefore, the values X are far from the mean, but this time below.

As in the previous case, another period has been taken to verify the results. The estimates of both stages can be seen in Figure 11, and the comparison of the mean values of the parameters shown in Table 3.

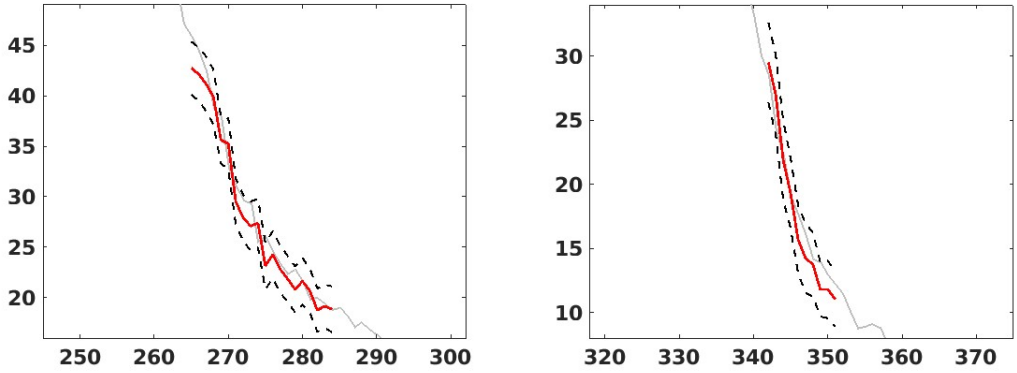

Figure 11. Stage 3 (left) and stage 4 (right) of a decisive decrease in the daily cases of the COVID-19 pandemic. Real cases (light grey), CIR* with Milstein method (red), 95% confidence interval (dashed lines). Ordinate 14-day moving average of real and modeled COVID-19 cases, abscissa number of days elapsed since February 2, 2020. Notice how the prediction stays within the dashed band.

Table 3. Decreasing of COVID-19 infections.

	Stage 3	Stage 4
Average α	-0.01664	-0.0305
Average μ	190.5401	95.5082
Average σ	0.2644	0.2925

4.3. Time frame in which COVID-19 cases are stable

Finally, we are going to interpret the parameters in a section where the values are relatively constant. As can be seen in Figure 1, there are only a few sections where this occurs. We are going to take days between 110 and 139, in which it is observed that the COVID cases are close to the 0 value. This may seem surprising, but it makes sense because it precisely coincides with the end of the state of alarm. In fact, on the right side of the graph, from day 150 the cases begin to increase, coinciding with the de-escalation process and the summer of 2020 when restrictions were relaxed.

At the aforementioned stage, the mean of the parameters α is 0.1397, that is, positive, unlike the previous cases. The mean of the parameters μ is 0.3303, and that of σ is 0.1448, so we have a smaller deviation than the previously analysed stages.

Table 4. Relatively constant stage of COVID-19 infections.

	Stage 5
Average α	0.1397
Average μ	0.3303
Average σ	0.1448

According to the analysis of the deterministic solution, X_t will oscillate around the mean, which makes sense, since the real mean of those days is around 0.33, and the value of the data varies between 0.25 and 0.55 cases.

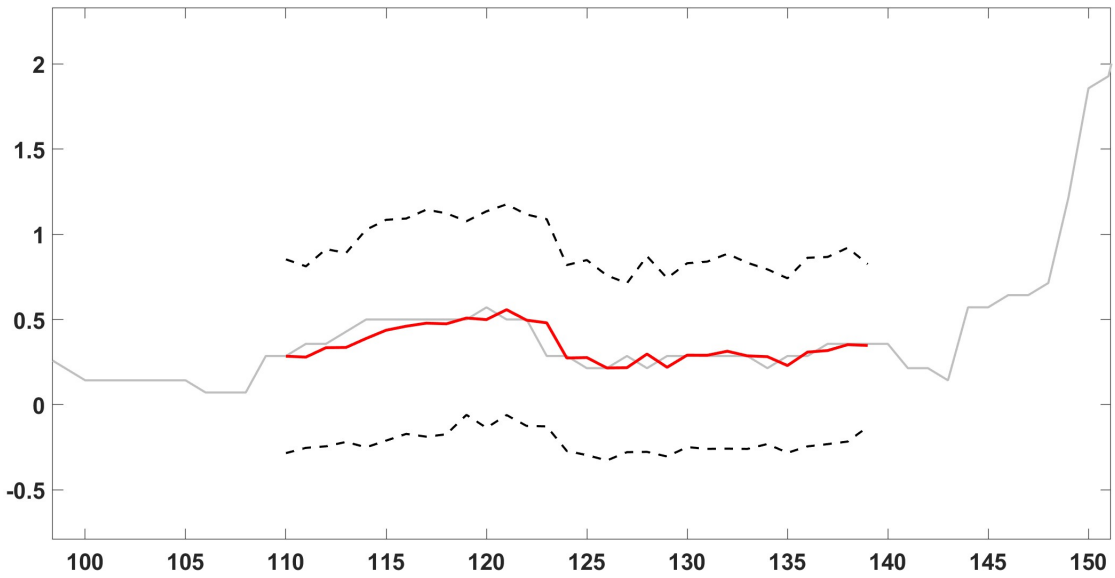

Figure 12. Example of a stage of the COVID-19 pandemic where the number of cases is relatively flat. Real cases (light grey), CIR* with Milstein method (red), 95% confidence interval (dashed lines). Ordinate 14-day moving average of real and modeled COVID-19 cases, abscissa number of days elapsed since February 2, 2020. Notice how the prediction stays within the dashed bands.

5. Conclusions

The main objective of this work is the study and development of some stochastic models typically used in financial markets applied to the COVID-19 pandemic in the city of Jaén.

For solving stochastic equations, both the Euler-Maruyama method and the Milstein method were used with reference to the CIR* stochastic process. Over the reported COVID-19 daily cases of the pandemic in the city of Jaén (Spain), the maximum likelihood method was used for parameter calibration. The forecasts given with this model have been validated with 95% confidence intervals and with statistical measures of goodness of fit, such as the RMSE (root mean square error). The results obtained are consistent with the observations and quite accurate. For comparison, the classical ARIMAX model has been used, resulting in more accurate predictions for the suggested CIR* model. The reason could be the relatively small size of the city of Jaén, causing large fluctuations in the number of cases that are not sufficiently softened by the moving averages, resulting in a worse behaviour of ARIMAX in comparison with CIR*. The importance of the suggested approach is twofold because it not only extends the models available to scholars to model pandemics but, also, paves the way for similar approaches in which financial models can be converted into econometric models.

Future research could be aimed at enlarging the scope to the whole province of Jaén and to other cities and/or provinces of Andalusia. In such a case, we could expect that a greater number of data could imply a longer time for the trend to change. In addition, the number of healed and deceased could also be studied, although the latter, being much smaller, will present the aforementioned problems. In terms

of considered models, future research could include a comparison with the more advanced CIR# by Orlando et al. [11, 36–38].

In addition, although our work has only been done for one equation, it could also be generalized to systems of equations to discover the interrelation between different cities or health districts. Finally, note that stochastic differential equations are not only a very powerful tool for modelling economic-financial variables, but also in the epidemiological field, being proven from this practical point of view [32].

Acknowledgments

G.O. is a member of the research group of GNAMPA - INdAM (Italy) and acknowledges the project on anomalous diffusion and its applications to fractal domains. J.G. acknowledges Junta de Andalucía through the project FEDER-UJA-1381026. The Consejería de Salud y Familia of Junta de Andalucía is acknowledged for providing the COVID-19 data.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. T. C. Mills, *Time Series Techniques for Economists*, Cambridge University Press, Cambridge, England, UK, 1990.
2. M. H. Lee, N. Hamzah, Calendar variation model based on ARIMAX for forecasting sales data with Ramadhan effect, in *Proceedings of the Regional Conference on Statistical Sciences 2010 (RCSS'10)*, **10** (2010), 349–361.
3. S. Chadsuthi, C. Modchang, Y. Lenbury, S. Iamsirithaworn, W. Triampo, Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time–series and ARIMAX analyses, *Asian Pac. J. Trop. Med.*, **5** (2012), 539–546. https://doi.org/10.1016/S1995-7645(12)60095-9
4. A. Suharsono, Suhartono, A. Masyitha, A. Anuravega, Time series regression and ARIMAX for forecasting currency flow at Bank Indonesia in Sulawesi region, *AIP Conf. Proc.*, **1691** (2015). https://doi.org/10.1063/1.4937107
5. W. Anggraeni, R. A. Vinarti, Y. D. Kurniawati, Performance comparisons between Arima and Arimax Method in Moslem kids clothes demand forecasting: Case study, *Procedia Comput. Sci.*, **72** (2015), 630–637. https://doi.org/10.1016/j.procs.2015.12.172
6. G. Shilpa, G. Sheshadri, ARIMAX model for short-term electrical load forecasting, *Int. J. Recent Technol. Eng. (IJRTE)*, **8** (2019), 2786–2790. https://doi.org/10.35940/ijrte.D7950.118419
7. A. A. Ariyo, A. O. Adewumi, C. K. Ayo, Stock Price Prediction Using the ARIMA Model, in *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, IEEE, (2014), 106–112.

8. G. Subramaniam, I. Muthukumar, Efficacy of time series forecasting (ARIMA) in post-COVID econometric analysis, *Int. J. Stat. Appl. Math.*, 2020. https://doi.org/10.22271/maths.2020.v5.i6a.609
9. G. Orlando, M. Bufalo, Modelling bursts and chaos regularization in credit risk with a deterministic nonlinear model, *Finance Res. Lett.*, **47** (2022), 102599. https://doi.org/10.1016/j.frl.2021.102599
10. G. Orlando, M. Bufalo, R. Stoop, Financial markets' deterministic aspects modeled by a low-dimensional equation, *Sci. Rep.*, **12** (2022), 1–13. https://doi.org/10.1038/s41598-022-05765-z
11. G. Orlando, R. M. Mininni, M. Bufalo, A new approach to CIR short-term rates modelling, in *New Methods in Fixed Income Modeling*, Springer, (2018), 35–43. https://doi.org/10.1007/978-3-319-95285-7_2
12. Y. Mishura, A. Yurchenko-Tytarenko, Standard and fractional reflected Ornstein–Uhlenbeck processes as the limits of square roots of Cox–Ingersoll–Ross processes. Stochastics, **95** (2022), 99–117. https://doi.org/10.1080/17442508.2022.2047188
13. O. O. Aalen, H. K. Gjessing, Survival models based on the Ornstein-Uhlenbeck process, *Lifetime Data Anal.*, **10** (2004), 407–423. https://doi.org/10.1007/s10985-004-4775-9
14. A. Ekinci, Modelling and forecasting of growth rate of new COVID-19 cases in top nine affected countries: Considering conditional variance and asymmetric effect, *Chaos, Solitons Fractals*, **151** (2021), 111227. https://doi.org/10.1016/j.chaos.2021.111227
15. A. K. Sahai, N. Rath, V. Sood, M. P. Singh, ARIMA modelling & forecasting of COVID-19 in top five affected countries, *Diabetes Metab. Syndr. Clin. Res. Rev.*, **14** (2020), 1419–1427. https://doi.org/10.1016/j.dsx.2020.07.042
16. R. Katoch, A. Sidhu, An application of ARIMA model to forecast the dynamics of COVID-19 epidemic in India, *Global Bus. Rev.*, **2021** (2021). https://doi.org/10.1177/0972150920988653
17. Suhartono, M. H. Lee, D. D. Prastyo, Two levels ARIMAX and regression models for forecasting time series data with calendar variation effects, *AIP Conf. Proc.*, **1691** (2015). https://doi.org/10.1063/1.4937108
18. A. Tanyavutti, U. Tanlamai, ARIMAX versus Holt Winter methods: the case of blood demand prediction in Thailand, *Int. J. Environ. Sci. Educ.*, **13** (2018), 519–525.
19. J. C. Cox, J. E. Ingersoll Jr, S. A. Ross, A theory of the term structure of interest rates, *Econometrica*, **53** (1985), 385–407. https://doi.org/10.2307/1911242
20. J. C. Cox, J. E. Ingersoll Jr, S. A. Ross, A theory of the term structure of interest rates, in *Theory of Valuation*, 2nd Edition, Singapore, World Scientific, (2005), 129–164. https://doi.org/10.1142/9789812701022_0005
21. O. Vasicek, An equilibrium characterization of the term structure, *J. Financ. Econ.*, **5** (1977), 177–188. https://doi.org/10.1016/0304-405X(77)90016-2
22. A. V. Chechkin, F. Seno, R. Metzler, I. M. Sokolov, Brownian yet non-gaussian diffusion: From superstatistics to subordination of diffusing diffusivities, *Phys. Rev. X*, **7** (2017), 021002. https://doi.org/10.1103/PhysRevX.7.021002

23. W. Wang, A. G. Cherstvy, A. V. Chechkin, S. Thapa, F. Seno, X. Liu, et al. Fractional Brownian motion with random diffusivity: emerging residual nonergodicity below the correlation time, *J. Phys. A: Math. Theor.*, **53** (2020), 474001. https://doi.org/10.1088/1751-8121/aba467

24. S. Ritschel, A. G. Cherstvy, R. Metzler, Universality of delay-time averages for financial time series: analytical results, computer simulations, and analysis of historical stock-market prices, *J. Phys.: Complexity*, **2** (2021), 045003. https://doi.org/10.1088/2632-072X/ac2220

25. A. Canale, R. M. Mininni, A. Rhandi, Analytic approach to solve a degenerate parabolic PDE for the Heston model, *Math. Methods Appl. Sci.*, **40** (2017), 4982–4992. https://doi.org/10.1002/mma.4363

26. G. Orlando, G. Taglialatela, A review on implied volatility calculation, *J. Comput. Appl. Math.*, **320** (2017), 202–220. https://doi.org/10.1016/j.cam.2017.02.002

27. G. Ascione, F. Mehrdoust, G. Orlando, O. Samimi, Foreign exchange options on Heston-CIR model under Lévy process framework, *Appl. Math. Comput.*, **446** (2023), 127851. https://doi.org/10.1016/j.amc.2023.127851

28. D. Duffie, Credit risk modeling with affine processes, *J. Banking Finance*, **29** (2005), 2751–2802. https://doi.org/10.1016/j.jbankfin.2005.02.006

29. G. Orlando, M. Bufalo, H. Penikas, C. Zurlo, *Modern Financial Engineering | Topics in Systems Engineering*, World Scientific Publishing Company, Singapore, 2021.

30. A. R. Ward, W. Glynn, Properties of the reflected Ornstein–Uhlenbeck process, *Queueing Syst.*, **44** (2003), 109–123. https://doi.org/10.1023/A:1024403704190

31. V. Giorno, A. G. Nobile, L. M. Ricciardi, On some diffusion approximations to queueing systems, *Adv. Appl. Probab.*, **18** (1986), 991–1014. https://doi.org/10.2307/1427259

32. L. M. Ricciardi, Stochastic population theory: Diffusion processes, in *Mathematical Ecology*, Springer, Berlin, Germany, (1986), 191–238. https://doi.org/10.1007/978-3-642-69888-0_9

33. V. Giorno, A. G. Nobile, R. di Cesare, On the reflected Ornstein–Uhlenbeck process with catastrophes, *Appl. Math. Comput.*, **218** (2012), 11570–11582. https://doi.org/10.1016/j.amc.2012.04.086

34. G. Orlando, G. Zimatore, Business cycle modeling between financial crises and black swans: Ornstein–Uhlenbeck stochastic process vs Kaldor deterministic chaotic model, *Chaos: Interdiscip. J. Nonlinear Sci.*, **30** (2020), 083129. https://doi.org/10.1063/5.0015916

35. G. Orlando, R. M. Mininni, M. Bufalo, A new approach to forecast market interest rates through the CIR model, *Stud. Econ. Finance*, **37** (2019), 267–292. https://doi.org/10.1108/SEF-03-2019-0116

36. G. Orlando, R. M. Mininni, M. Bufalo, Interest rates calibration with a CIR model, *J. Risk Finance*, **20** (2019), 370–387. https://doi.org/10.1108/JRF-05-2019-0080

37. G. Orlando, R. M. Mininni, M. Bufalo, Forecasting interest rates through Vasicek and CIR models: A partitioning approach, *J. Forecasting*, **39** (2020), 569–579. https://doi.org/10.1002/for.2642

38. G. Orlando, M. Bufalo, Interest rates forecasting: Between Hull and White and the CIR#—How to make a single-factor model work, *J. Forecasting*, **40** (2021), 1566–1580. https://doi.org/10.1002/for.2783

39. L. Ljung, System identification, in *Signal Analysis and Prediction*, Birkhäuser, Boston, MA, (1998), 163–173. https://doi.org/10.1007/978-1-4612-1768-8_11

40. MathWorks, *Estimate ARMAX Model*, 2022. Accessed date: 18 November 2022. Available from: https://www.mathworks.com/help/ident/ref/armax.html.

41. P. Stoica, Y. Selen, Model-order selection: a review of information criterion rules, *IEEE Signal Process. Mag.*, **21** (2004), 36–47.

42. K. Kladívko, Maximum likelihood estimation of the Cox-Ingersoll-Ross process: the MATLAB implementation, *Tech. Comput. Prague*, **7** (2007).

43. G. N. Milstein, Approximate integration of stochastic differential equations, *Theory Probab. Appl.*, **19** (1975), 557–562.