
*Research article***Building the yield curve in low-liquidity markets****Tsvetelin S Zaeovski**^{1,2,*}¹ Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. Georgi Bonchev Str., Block 8, 1113 Sofia, Bulgaria² Centre of Excellence in Informatics and Information and Communication Technologies, Sofia, Bulgaria* **Correspondence:** Email: t_s_zaeovski@math.bas.bg, t_s_zaeovski@abv.bg.

Abstract: This paper proposes a semi-statistical framework for constructing yield curves from primary-market government securities data in illiquid bond markets. Because government debt issuances are infrequent and sparse, the yield curve cannot be directly estimated from available market information. To address this limitation, the term structure of interest rates is characterized by classical affine stochastic models, calibrated through a transition-density-based estimation. For practical application, we employ the models of Vašíček and Cox–Ingersoll–Ross. The proposed framework combines historical observations with stochastic modeling to handle incomplete market data. The methodology is empirically evaluated using U.S. and Bulgarian Treasury securities. Results suggest that the approach produces stable and economically consistent yield curve estimates, particularly in environments where secondary-market data are limited or unreliable. The framework is especially relevant for emerging markets and may also support backtesting, forecasting, model validation, and structural analysis of bond markets.

Keywords: yield curve; term structure of interest rates; semi-statistical approach; Vašíček model; CIR model; American treasury securities

JEL Codes: E43, E47, C02

1. Introduction

The yield curve, also known as term structure of interest rates, is one of the most important financial factors that describes economic realities and indicates different future scenarios. It presents the rate of return that investors require to buy a debt. Typically, this debt is issued by the government, and thus, the default risk is not considered. Alternatively, we may assume that it is incorporated in investors' preferences. If the expected rate of return is independent of the debt's term, then the yield curve would be flat. Of course, this assumption is not realistic, as long maturities lead to more market uncertainty.

Many methods have been proposed for exploring the yield curve. Some of them use deterministic functions that capture its shape, including Nelson and Siegel (1987), Svensson (1994), and Gürkaynak et al. (2007). Later, some authors generalized the Nelson–Siegel–Svensson family in a stochastic way that guarantees the absence of arbitrage; see Christensen et al. (2009) and Christensen et al. (2011). Another important class of models featuring a stochastic element includes the classical affine short-rate models of Vašíček (1977), Dothan (1978), Brennan and Schwartz (1982), Cox et al. (1985), Ho and Lee (1986), Black and Karasinski (1991), Constantinides (1992), and Pearson and Sun (1994) as well as the forward-rate model of Heath et al. (1992). All these models assume that a set of parameters governs the market dynamics. Alternatively, Liu and Wu (2021) provided a nonparametric approach based on the Gaussian kernel. We refer also to the comprehensive investigation of Brigo and Mercurio (2006).

Usually, the investigation of the term structures of interest rates relies on the assumption that sufficient market information is available. On the other hand, governments relatively rarely issue new debt, and thus, researchers have to rely mainly on information from the secondary markets. If they are deep enough, then the prices for different maturities should be available, and the yields can be obtained at sufficient points. Based on these, the whole curve can be approximated; for example, McCulloch (1971) and Vašíček and Fong (1982) applied splines. Unfortunately, this is often not the case, and we have to use only information from primary markets. However, it is almost always insufficient; thus, we need to impose some additional assumptions to fill this gap. There are two main investigations in the field of yield curve analysis under infrequent trading and missing data; Cortazar et al. (2007) and Nagy (2020), both based on the Kalman filter. The main difference between them is that Cortazar et al. (2007) assumed that the short rate is described by a multifactor Vašíček model, whereas Nagy (2020) applied the Nelson and Siegel (1987) model to the yield curve.

The traditional methods (parametric or not) based on the current information are not applicable in the case of sparse data. A natural solution to this problem is the assumption that some parametric model drives the market and to estimate it later based on past observations. As we mentioned above, Cortazar et al. (2007) assumed that the short rate is driven by several Ornstein–Uhlenbeck (OU) processes as in Vašíček and Fong (1982). Note that negative rates are admissible. The authors motivated this choice by the large part of the so-called inflation-protected bonds in the Chilean government debt. Moreover, the market realities indicate that negative rates are not so unrealistic. On the other hand, Nagy (2020) applied the deterministic model of Nelson and Siegel, which is flexible enough to capture all possible shapes of the yield curve. Both of these important works relied on Kalman filter technique for practical implementation. The first advantage of this approach is its ability to filter out possible noise from the empirical data. Second, the Kalman filter can overcome the problem that neither Ornstein–Uhlenbeck processes nor yield curves are directly observable in the markets; available information is only for the bond prices. Third, this algorithm does not necessitate full data for properly estimating the short rate in the Vašíček model in Cortazar et al. (2007) or the Nelson–Siegel curve in Nagy (2020). Without providing details, the Kalman algorithm is based on the dynamic adjustment of the unobservable variables with respect to (w.r.t.) incoming new information about the observable bond prices. The deviations between the observed values and the predicted ones are first obtained. Next, having in mind that the noise in the observations is assumed to be Gaussian, we seek those parameters that maximize the likelihood w.r.t. the normal distribution. The parameter set includes the model parameters as well as the standard deviation of the normal distribution for the errors. Thus, both methods, those of Cortazar et al. (2007) and Nagy (2020), minimize the resulting errors in some sense.

Let us mention that the models of Cortazar et al. (2007) and Nagy (2020) were established with a focus on the Chilean and Hungarian financial realities. Despite some similarities with the Bulgarian specifics which motivate the current investigation, there are many differences. Bulgaria is in a currency board from July, 1, 1997 until January, 1, 2026, when Bulgaria enters the eurozone. The Bulgarian currency was first fixed to the Deutsche Mark with exchange rate 1:1000 and later to the Euro with rate 1:1.95583. This currency system imposes significant restrictions on the Bulgarian government, resulting in the infrequent issuance of new debt. Moreover, on dates when this occurs, the new debt consists of a single instrument. On the other hand, the information from the secondary market in Bulgaria is not reliable; there are many dates with reports with quite different prices of one and the same emissions. Various reasons for this exist: The secondary bond trading is underdeveloped, many of the deals are arranged previously as futures contracts but are reported on their transaction dates, there are many repurchase agreements (the so-called repos), and so on. All this necessitates the use of the information from the primary market.

Our suggestion is to assume that some affine term structure model is valid. Next, we extract the short rates at the issue dates using the available bond prices. Then, we maximize a criterion based on the transition densities of the short-rate stochastic process to calibrate the particular model and, consequently, derive the yield curve.

There are several main differences with the methods of Cortazar et al. (2007) and Nagy (2020). First, the use of the Kalman filter is not imperative because we have a one-to-one mapping between bond prices and the short rate. Also, one can say that the noise is incorporated in the Brownian motion. On the other hand, other authors argue that the various types of noise are very likely in low-liquidity markets. This important question needs an additional in-depth investigation. Although this is out of the scope of this study, we discuss how our approach can be generalized by involving more complicated short rate models as well as by noise removal through the Kalman filter.

The most important difference comes from the calibration methodology. Whereas Cortazar et al. (2007) and Nagy (2020) approximated the bond prices by minimizing the resulting errors, we estimate through the transition densities. Thus, we evaluate how likely the next value of the short rate is given the previous one. This way, we take into account not only the prices themselves but also the intrinsic relations within their temporal structure.

It should be noted that the approach we suggest can be generalized to capture the specifics of data that is sparse but not necessarily consisting only of a single per-date observations. Moreover, it can be modified in a way that captures more complicated term structures than those generated by a single-factor affine model. All this is discussed in Section 7.

As we mentioned above, many short-rate models lead to affine term structures. We chose two of them to implement our approach, namely the Vašíček and Cox–Ingersoll–Ross (CIR) models. Both assume that the short rate is driven by a mean reverting stochastic process. In the Vašíček model, the short rate is presented by an Ornstein–Uhlenbeck process, and thus, its transition distributions are Gaussian. Alternatively, Cox et al. (1985) suggested the use of the so-called CIR process that leads to noncentered gamma transition distributions. The main difference is that the normal distribution admits negative values, whereas the noncentered gamma one is defined on the positive real half-line. In general, the mathematical models of financial assets rely on two probability measures. The real-world (or physical) measure is important for underlying assets, whereas the risk-neutral measure describes derivative instruments. The difference between them is interpreted as a market price of risk. For the

models of Vašíček and Cox–Ingersoll–Ross, the risk premium can be chosen in a way that changes both parameters related to the mean reversion: the speed of reversion and the long term mean.* On the other hand, bonds are a very specific kind of derivative because their payoff does not depend directly on some underlying asset. Accordingly, the risk-neutral measure is more important for our study, and we will work within its framework. Nevertheless, we will discuss the parameters under both measures as well as the relations between them.

Historically, negative risk-free rates seem very exotic. A decade ago, negative values were observed only at the Japanese exchange. On the other hand, this phenomenon has been widespread across all markets for most of the last decade, excluding the recent several years. Thus, the pointed-out main drawback of the Vašíček model turns into an advantage. Our empirical investigation supports this hypothesis. On the other hand, we can use a shifted version of the CIR model when negative rates are very likely or to rely only on the Vašíček model in such situations. On the other hand, negative interest rates are rare in the emerging markets. Unlike advanced economies (such as the US, Eurozone, Switzerland, or Japan) that have experimented with this policy, emerging markets typically maintain higher interest rates to control inflation and attract capital. Furthermore, emerging economies are considered riskier by investors. To offset this risk, their central banks need to offer positive and competitive interest rates. If interest rates become too low or negative, the local currency would collapse due to massive capital flight. However, there are exceptions from this theoretical rule. Such an exception is clearly observed in the Bulgarian treasury securities. This is mainly due to the long-standing currency board, the subsequent conservative financial policy, and the good state of public finances.

Another advantage of the proposed approach is the possibility of forecasting the future expected returns. Once we calibrate the model, we may use the transition densities of the underlying processes to predict the future bond prices as well as other market descriptors such as the yield curve, the short rate, different forwards, and so on. Also, the confidence intervals can be derived.

After stating the approach for constructing the yield curve, we need to check its efficiency and reliability. To do this, we use the comprehensive historical data provided by Liu and Wu (2021). This data contains daily observations for American treasury securities (bills, notes, and bonds) for maturities between one and 360 months. The choice of empirical data from the U.S. Treasury market, which is highly liquid, has a significant advantage. First, we can extract the empirical yield curves using the whole data. After this, we can choose randomly only one observation per date and then apply our approach to this restricted data. Thus, we validate the suggested method, checking how near our approximation is to the real yield curve. More precisely, the backtesting approach is based on the following steps. We choose a hundred-day period and only one random observation for each of these days – thus, we collect data consisting of hundred nonhomogeneous observations. We apply the suggested approach to this data. Once we calibrate the corresponding affine stochastic model (Vašíček or CIR), we can build the yield curve for the next day and compare it to the term structure reported by Liu and Wu (2021). We prepare this validation test several times, including high- and low-volatility periods. The results of this backtesting strongly indicate that the produced curves capture well enough the true term structures, even for this highly restricted data.

We can mention several areas where the approach we suggest is applicable. First, the secondary markets in emerging economies are not so deep, and thus, the available information is not reliable. On the other hand, the available observations at the primary markets are quite insufficient. Also, their

*This must be handled carefully for the CIR model to prevent the occurrence of a local martingality.

liquidity is sometimes very low, leading to not credible enough signals. These circumstances make the use of the traditional methods impossible, whereas our approach is still applicable. Also, this method allows various tests to be conducted for detecting different structural properties in rich enough data sets. For example, we can search for stability by choosing different random subsets of the main data. Alternatively, data restrictions could be implemented in a nonrandom fashion to better identify some structural properties. Last but not least, the suggested approach can be helpful for different debt markets with shallow structures.

In light of the previous conclusions, the theoretical results obtained in this paper are applied to the historical data for the Bulgarian government debt reported on the official website of the Bulgarian National Bank. We use a moving window to derive the yield curve during the time. This way, the impact of the recent observations is larger. We provide many numerical experiments to analyze the derived results for the short rates, realized returns, bond prices, yield curves, and more.

Last but not least, let us mention the works of Xie et al. (2006), Khakimzhanov et al. (2019), Makushkin and Lapshin (2021), Chen et al. (2022), and Payne et al. (2023), that are more or less related to the investigated field.

The plan of the paper is as follows. We provide the necessary theoretical base in Section 2. Our approach for building the yield curve is presented in Section 3. The calibration methodology is established in Section 4. The proposed model is validated and backtested via the American treasury securities in Section 5. We apply the derived theoretical results to the Bulgarian government debt in Section 6. We conclude in Section 7.

2. Theoretical base

2.1. Risk-neutral bond pricing

Let $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{Q})$ be a filtered probability space. The measure $\mathbb{Q}$ is assumed to be risk-neutral; that is, all discounted bond price processes are $\mathbb{Q}$ -martingales. Let B_t be a Brownian motion in this space. We will specify below the bond structure under these terms. Two main kinds of bonds are issued in the debt markets. The first ones, known as zero-coupon bonds, do not pay any amount until the maturity date. Alternatively, the coupon bonds provide periodically some payments in addition to the face value at maturity. We will denote by $P(t, T)$ the price at the moment $t \leq T$ of a zero-coupon bond that matures at moment T with face value one. Note that it is a stochastic process w.r.t. the variable t . The *yield curve*, also known as *term structure of interest rates*, is defined as the process

$$y(t, T) = -\frac{\log(P(t, T))}{T - t}. \quad (1)$$

Suppose now that a bond pays in addition to its face value amounts of C_i , $i = 1 \dots N$, at moments $T_1 < T_2 < \dots < T_N < T$. We can view each of these payments as C_i shares of zero-coupon T_i -bonds. Thus, the price of the coupon bond can be written as

$$P^C(t, T) = \sum_{i=1}^N C_i P(t, T_i) + P(t, T), \quad (2)$$

where $P(t, T_i)$ and $P(t, T)$ are the prices of the corresponding zero-coupon bonds.

Next we present the main results for the Vašíček and Cox–Ingersoll–Ross models. For more details, we refer to Sections 3.2.1 and 3.2.3 of Brigo and Mercurio (2006).

2.2. Vašíček model

In the Vašíček model, the short rate follows the Ornstein–Uhlenbeck process

$$dr_t = \xi(\eta - r_t)dt + \theta dB_t \quad (3)$$

with an initial condition r_0 . The constant ξ presents the speed of reversion to the long term mean level η . The constant θ is the volatility of the short rate. Process (3) is Gaussian, with transition densities

$$N\left(\eta + (r_t - \eta) \exp(-\xi(T - t)), \frac{\theta^2(1 - \exp(-2\xi(T - t)))}{2\xi}\right) \quad (4)$$

when $t < T$. The price of a zero-coupon bond is

$$P(t, T) = e^{-A(T-t)r_t + B(T-t)} \quad (5)$$

for

$$\begin{aligned} A(s) &= -a(s; 1) = \frac{1 - e^{-\xi s}}{\xi} \\ B(s) &= b(s; 1) = \left(\eta - \frac{\theta^2}{2\xi^2}\right)(A(s) - s) - \frac{\theta^2 A^2(s)}{4\xi}. \end{aligned} \quad (6)$$

2.3. CIR model

In the CIR model, the short rate follows a square-root diffusion process given by the stochastic differential equation

$$dr_t = \xi(\eta - r_t)dt + \theta \sqrt{r_t} dB_t. \quad (7)$$

Defined in that way, the short rate process is again mean-reverting with speed of reversion ξ , long term mean η , and volatility θ . The main difference between the OU and CIR processes is that the second does not admit negative values, because the diffusion term vanishes when the process is near zero, and the drift directs the process upward. Process (7) can reach the zero when $2\xi\eta < \theta^2$. Otherwise, if $2\xi\eta \geq \theta^2$, then the volatility is small enough, and the process stays always positive. For more detail, see Feller (1951) or Pitman and Yor (1982). The bond price is again given by Formula (5). The difference comes from the functions (6), which now are

$$\begin{aligned} A(s) &= -a(s; 1) = \frac{2(\exp(s\gamma) - 1)}{2\gamma + (\xi + \gamma)(\exp(s\gamma) - 1)} \\ B(s) &= b(s; 1) = \frac{2\xi\eta}{\theta^2} \log\left(\frac{2\gamma \exp((\xi + \gamma)s/2)}{2\gamma + (\xi + \gamma)(\exp(s\gamma) - 1)}\right) \\ \gamma &= \sqrt{\xi^2 + 2\sigma^2}. \end{aligned} \quad (8)$$

The transition distribution of r_T conditional on r_t , $t < T$ is a noncentered gamma with parameters

$$\left(\frac{2\xi\eta}{\theta^2}, c(T-t), r_t e^{-\xi(T-t)} \right), \quad (9)$$

where the function $c(\cdot)$ is defined as

$$c(s) = \frac{2\xi}{\theta^2 (1 - \exp(-\xi s))}. \quad (10)$$

We use the following parametrization for the noncentered gamma probability density function (PDF) w.r.t the set $\{\kappa, c, \lambda\}$:

$$f(x) = c e^{-c(x+\lambda)} \left(\frac{x}{\lambda} \right)^{\frac{\kappa-1}{2}} I_{\kappa-1} \left(2c \sqrt{x\lambda} \right), \quad (11)$$

where $I_{\kappa-1}(\cdot)$ is the modified Bessel function of the first kind. Note that $2c(T-t)r_T$ is distributed under the noncentered χ^2 -law with parameters

$$\left(\frac{4\xi\eta}{\theta^2}, 2r_t e^{-\xi(T-t)} c(T-t) \right). \quad (12)$$

Having in mind Formulas (1) and (5), we obtain for the Vašíček and CIR yield curves

$$y(t, s) = \frac{A(s)r_t - B(s)}{s}, \quad (13)$$

where the functions $A(\cdot)$ and $B(\cdot)$ are given by Formulas (6) or (8).

3. Building the yield curve

As we already mentioned, if we know the prices of a rich enough set of bonds, then we can obtain the corresponding points of the yield curve and then approximate it. In practice, however, this is often not the case: Each day, we know at most one price, and therefore, we can not use any trivial method. Alternatively, we assume that the bond market follows some of the above-mentioned stochastic models.

Let the time horizon be T , and suppose that n observations (bond prices) are available in the interval $[0, T]$. Let these bonds are issued at dates t_i , $i = 1, 2, \dots, n$ and mature at T_i , $t_i < T_i$. We can think that the bonds are ordered w.r.t. to the issue date, $t_1 < t_2 < \dots < t_n$. We know at the moment t_i only the prices $P_j = P(t_j, T_j)$ for $j \leq i$. Note that these prices are recorded at the dates of issuance, not at the current time. Thus, the available information at the moment t consists of the prices P_j for $t_j \leq t$.

Suppose that one of the presented above affine models (Vašíček or CIR) describes the market. Let the parameters of the underlying process, (3) or (7), be ξ , η , and θ . We need now the following statements to continue our investigation:

Proposition 3.1. *Let a zero-coupon bond be issued at time t and mature at date T , $t < T$. If the market is described by the Vašíček or CIR model, then the bond price, considered as a function w.r.t. the current short-rate value $P(r) = P(t, T; r)$, decreases continuously. Moreover, $P(-\infty) = +\infty$, and $P(+\infty) = 0$. Furthermore, the statement holds for the coupon bonds, too.*

Note that even though the CIR model cannot produce negative interest rates, Formula (5) still holds for negative values of r .

Proof. The proposition holds because the bond price for both models is given by Formula (5). Note that the function $A(s)$ from this formula is positive; see Equations (6) and (8). The statement holds for the coupon bonds because they can be presented as a weighted sum of zero-coupon bonds; see Equation (2). $\square$

Remark 3.1. *The OU process is normally distributed, and thus, the support of the short rate in the Vašíček model is the whole real line $(-\infty, \infty)$. On the opposite, the domain for the CIR model is $(0, \infty)$ or $[0, \infty)$. Thus, some values for the bond price $P(r)$ may lead to impossible (negative) values for the CIR short rate. If this happens, we will consider the parameter set $\{\xi, \eta, \theta\}$ as inadmissible. Thus, the PDF at this point still exists but is zero.*

We can formulate and prove now a proposition that is the base of our forecasting algorithm:

Proposition 3.2. *Let us denote by P_{obs} an observed bond price. The equation that gives the relation between the theoretical and real prices (see Formula (2)),*

$$P^C(t, T) \equiv \sum_{j=1}^N C_j P(t, T_j) + P(t, T) = P_{obs}, \quad (14)$$

has a unique solution w.r.t. the initial short rate value. We will denote it by r_{t_i} . If the bond is without coupon payments, then

$$r_{t_i} = \frac{B(T_i - t_i) - \log(P_{obs})}{A(T_i - t_i)}, \quad (15)$$

where the functions $A(\cdot)$ and $B(\cdot)$ are given by Formulas (6) or (8) (for the Vašíček or CIR model).

Proof. Proposition 3.1 leads to the existence and uniqueness of the solution. Note that the admissibility condition for the CIR model (from Remark 3.1) guarantees that the solution is non-negative. Equation (15) holds due to Formula (5). $\square$

Remark 3.2. *The meaning of Proposition 3.2 is in extracting the current short rate from the bond price under the assumptions of one of the considered models. This is similar to deriving the yield to maturity but without the assumption of a constant rate during the period.*

Our algorithm for building the yield curve is as follows:

1. Let i be the largest number such that $t_i \leq t$. We will use the bond prices P_j , $j = 1, 2, \dots, i$, and their coupon payments (if they exist) as well as their issue and maturity dates.
2. We extract the short rate values r_{t_i} for a fixed parameter set $\{\xi, \eta, \theta\}$ as the solution of Equation (14). If we have a zero-coupon bond, then we use Formula (15). Note that if we work under the CIR model, we must restrict the set of possible triples $\{\xi, \eta, \theta\}$ to those which return positive short rates.
3. We derive the transition densities of the short rate process based on the already-obtained values of r_{t_j} through the results given in Sections 2.2 or 2.3 for the Vašíček and CIR model, respectively. Let us denote them by $p(x | r_{t_{j-1}}, t_{j-1}, t_j; \xi, \eta, \theta)$. The meaning of this notation is the probability density w.r.t. the variable x at the moment t_j conditional on the assumption that the short rate value at the moment t_{j-1} is $r_{t_{j-1}}$. We also mark the dependence on the parameters $\{\xi, \eta, \theta\}$.

4. We calibrate the parameters $\{\xi, \eta, \theta\}$ maximizing a criterion dependent on the transition densities via the values $p(r_{t_j} | r_{t_{j-1}}, t_{j-1}, t_j; \xi, \eta, \theta)$, $j = 1, 2, \dots, i$. We discuss an appropriate criterion in the following section. We base our calibration on the available to date t historical data – P_j , $j = 1, 2, \dots, i$. Alternatively, we can use only some last $k < i$ observations instead of the whole data, namely P_j , $j = i - k + 1, \dots, i$. Thus, the calibration is made in a way that makes the short rate transition densities most plausible to the recent historical data.
5. We derive the yield curve at the moment t using Equation (13).

4. Calibration methodology

Let Γ be the set of the admissible parameters for the corresponding model, and $\{\xi, \eta, \theta\} \in \Gamma$. Note that all positive triples are admissible for the Vašíček model, whereas some restrictions are imposed for the CIR model (see Remark 3.1).

First, we need to identify the observations that we will use to calibrate the theoretical models: Suppose that they are $P_j = P(t_j, T_j)$, $j = 1, 2, \dots, i$. Note that $t_1 < t_2 < \dots < t_i$. Using Proposition 3.2, we derive the values r_{t_j} , $j = 1, 2, \dots, i$, for the short rate process. If we have a coupon bond, then r_{t_j} is the solution of Equation (14); otherwise, we use Formula (15) for the non-coupon bonds. Thus, we extract a sequence for the short rate process $r_{t_1}, r_{t_2}, \dots, r_{t_i}$ related to the moments $t_1, t_2, \dots, t_i$. Let us note the importance of the strong inequality $t_j < t_{j+1}$. In the opposite case, $t_j = t_{j+1}$ for some j , we may derive two different values for the short rate; this will lead to a serious confusion. Using this approach, we extract a semi-statistical sample for the short rate process from the observed bond prices. Because we know the theoretical transition densities, it is natural to use a maximum likelihood estimation to find the values of $\{\xi, \eta, \theta\}$ that lead to the best fit. This way, the criterion for maximizing is

$$F(\xi, \eta, \theta) = \prod_{j=2}^i p(r_{t_j} | r_{t_{j-1}}, t_{j-1}, t_j; \xi, \eta, \theta), \quad (16)$$

where $p(\cdot | r_{t_{j-1}}, t_{j-1}, t_j; \xi, \eta, \theta)$ is the transition density of the short rate process between the moments t_{j-1} and t_j conditional on the value $r_{t_{j-1}}$. Note that if Proposition 3.2 returns a negative CIR interest rate, then our approach will reject the current triple $\{\xi, \eta, \theta\}$, as the PDF at this negative point is zero. Recall that we have a normal distribution for the Vašíček model and a noncentered χ^2 distribution for the CIR one; see Sections 2.2 and 2.3. Thus,

$$p(x | r_{t_{j-1}}, t_{j-1}, t_j; \xi, \eta, \theta) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (17)$$

for the Vašíček model, where

$$\begin{aligned} \mu &= \eta + (r_{t_{j-1}} - \eta) \exp(-\xi(t_j - t_{j-1})) \\ \sigma &= \frac{\theta^2 (1 - \exp(-2\xi(t_j - t_{j-1})))}{2\xi}. \end{aligned} \quad (18)$$

For the CIR model, we have

$$\begin{aligned}
p\left(x|r_{t_{j-1}}, t_{j-1}, t_j; \xi, \eta, \theta\right) &= ce^{-u-v}\left(\frac{v}{u}\right)^{\frac{q}{2}} I_q\left(2\sqrt{uv}\right) \\
c &= \frac{2\xi}{\theta^2\left(1-e^{-\xi(t_j-t_{j-1})}\right)} \\
u &= cr_{t_{j-1}}e^{-\xi(t_j-t_{j-1})} \\
v &= cx \\
q &= \frac{2\xi\eta}{\theta^2} - 1.
\end{aligned} \tag{19}$$

To facilitate computation, we can apply a logarithmic transformation to Criterion (16). Furthermore, we need a regularization factor because the used data is not rich enough. Some empirical tests show that the product $\xi\eta$ stays relatively stable. On the contrary, the terms ξ and η themselves are very variable: the pure maximum likelihood estimation (MLE) criterion (16) produces values close to zero for ξ and very large ones for η without a significant improvement in the objective function. For this, we suggest an L_2 regularization, modifying Criterion (16) into

$$F(\xi, \eta, \theta) = \sum_{j=2}^i \ln p\left(r_{t_j}|r_{t_{j-1}}, t_{j-1}, t_j; \xi, \eta, \theta\right) + \lambda_n \left(r_{t_j} - \eta\right)^2. \tag{20}$$

The weight λ_n depends on the sample size n as c/n . This way, we have a lower impact when n is larger, and this reduction is suitable from a statistical point of view. Also, the term $\left(r_{t_j} - \eta\right)^2$ leads to values for r_t that are near to the long term mean value η . Without any explicit theoretical obligation, such a choice for η will lead to realistic values for the risk premium. It turns out that a suitable level for the constant c is $c = 40000$; it ensures that the regularization term has enough influence without dominating.

5. Validation and backtesting

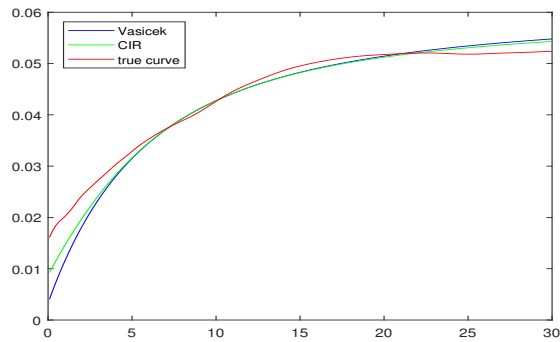
We will validate the proposed approach for constructing the yield curves using the empirical data provided by Liu and Wu (2021).[†] We use this data because it is very detailed and allows extracting the empirical curves with a relatively high precision. After that, we can choose only one observation per date and then apply our approach to the so-restricted data. Thus, we can compare our approximations with the real curves. The dataset contains two parts. The first one consists of daily observations for maturities between 1 and 85 months, whereas the second part is for maturities between 1 and 360 months. We use only the second part of the sample; all reported numbers of rows and columns in the current paper refer to the restricted table. Because Liu and Wu (2021) reported the processed values for the yield to maturity, we need to reconstruct the bond prices through this information. We do this assuming that the bonds are without coupon payments, and thus, their prices can be obtained as $P = 100e^{-yT}$, where T is the time to maturity, y is the corresponding rate of return, and the face value is assumed to be \$100. Our backtesting algorithm is as follows:

[†]See the website <https://sites.google.com/view/jingcynthiawu/yield-data>.

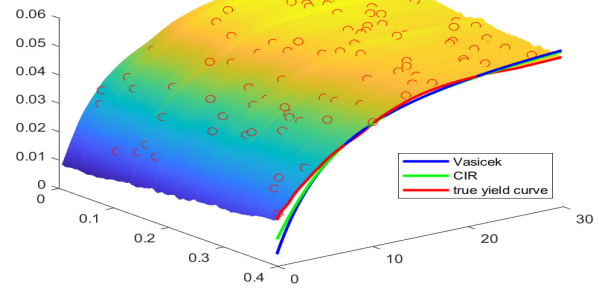
1. We choose from the statistical sample provided in Liu and Wu (2021) one hundred rows between N and $N + 99$ with daily observations. The value of N will be specified later. Note that each row consists of 360 observations covering maturities from 1 to 360 months.
2. We reconstruct the bond prices as $P_{ij} = 100e^{-jy_{ij}/12}$, where i is the row's number, and j is the number of the months before the maturity.
3. We generate randomly one hundred integers in the range of $\{1, 2, \dots, 360\}$. We denote them by n_i , $i = 1, 2, \dots, 360$. This is made through the MATLAB command `randi`. The particular realization is reported in Appendix A.
4. We restrict the statistical sample to the values P_{i,n_i} , $i = 1, 2, \dots, 360$. We will use only these observations.
5. We apply the calibrating approach provided in Section 4 to the above-restricted data. The used minimization technique is explained in detail in Section 6.2.
6. We construct the yield curve using the calibrated parameters via Formula (13).
7. We compare this curve with the true one generated by the $(N + 100)$ -th row of the statistical sample.

We apply the above-defined algorithm, choosing N amongst $N \in \{4600, 5479, 5610, 5672, 5900, 7900, 8609, 9031\}$. Below, we describe the market specifics during these periods:

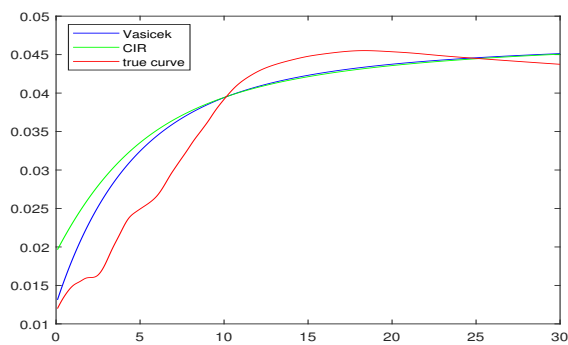
- (A) 27.04.2004–19.09.2004; validation date 19.09.2004 ($N = 4600$): A medium-volatility period
- (B) 29.10.2007–25.03.2008; validation date 26.03.2008 ($N = 5479$): A high-volatility period marked by the beginning of the global financial crisis of 2008.
- (C) 08.05.2008–29.09.2008; validation date 30.09.2008 ($N = 5610$): An extremely high-volatility period following the Lehman Brothers default. On September 29 occurs one of the biggest declines of the main stock market indices.
- (D) 06.08.2008 – 29.12.2008 ; validation date 30.12.2008 ($N = 5672$): One of the most-turbulent periods ever.
- (E) 22.06.2009–27.11.2009; validation date 30.11.2009 ($N = 5900$): A relatively volatile period, with a tendency towards stabilization.
- (F) 3.07.2017 – 23.11.2017; validation date 24.11.2017 ($N = 7900$): A very calm period.
- (G) 05.05.2020 – 24.09.2020; validation date 25.09.2020 ($N = 8609$): A relatively stable period after the beginning of the the COVID 19 pandemic. It is characterized by significantly low interest rates.
- (H) 7.01.2022 – 1.06.2022; validation date 2.06.2022 ($N = 9031$): A period with significant volatility characterized by very high inflation.



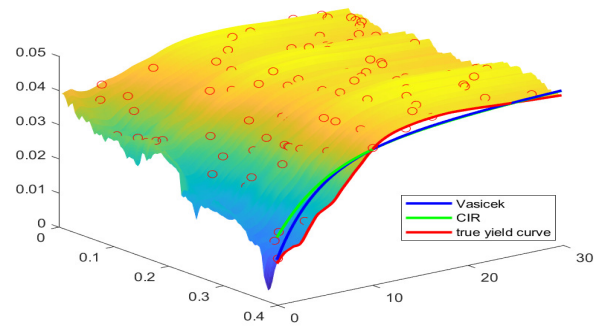
(a) yield curves, Sept 19, 2004



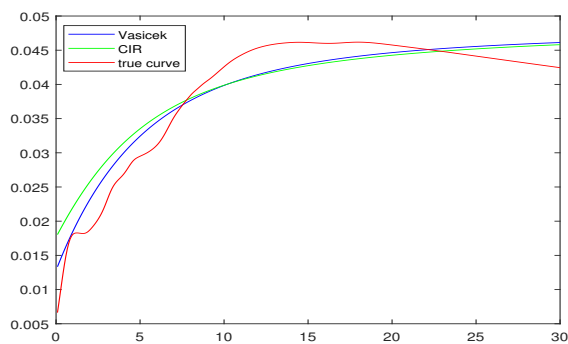
(b) surface, Apr 27, 2004–Sept 19, 2004



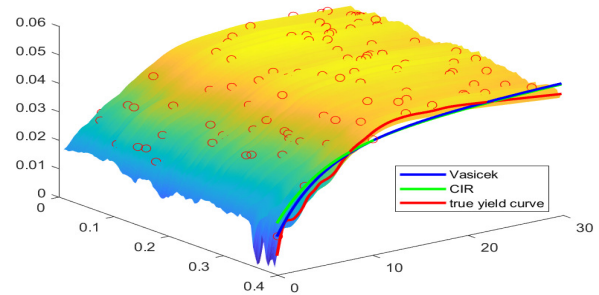
(c) yield curves, March 26, 2008



(d) surface, Oct 29, 2007–March 26, 2008

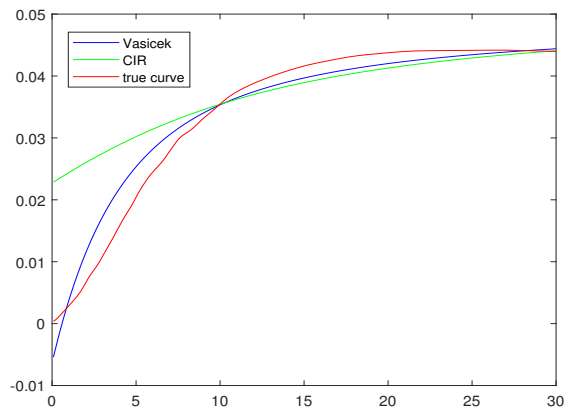


(e) yield curves, Sept 30, 2009

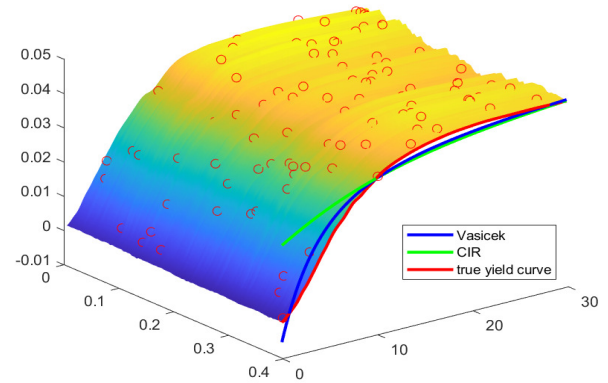


(f) surface, May 8, 2008–Sept 30, 2009

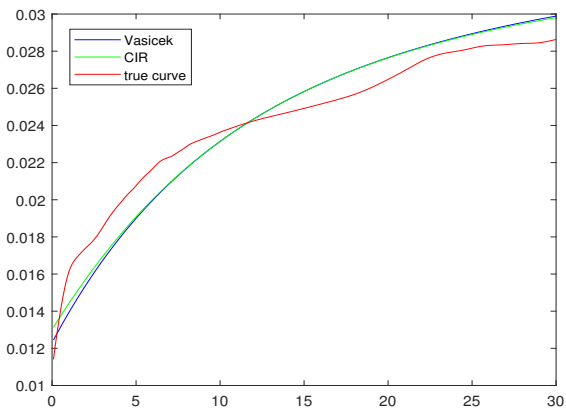
Figure 1. Backesting through US Treasury instruments. The figure presents the forecasts for the yield curve based on restricted data for US treasury securities. The first column is for the yield curves themselves, whereas the second one presents the whole historical surface. The dots give the points that are used for the calibration.



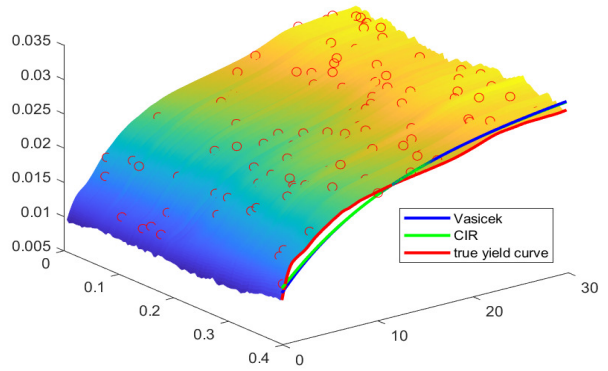
(a) yield curves, Nov 30, 2009



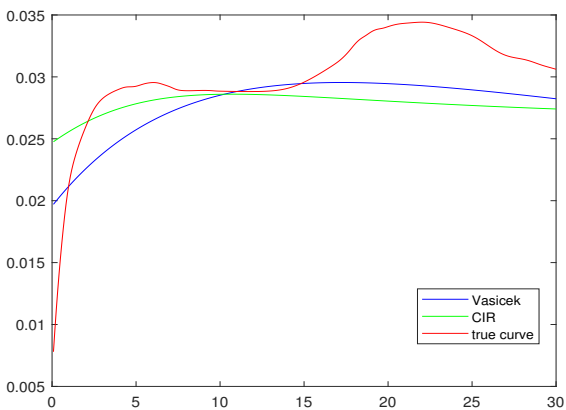
(b) surface, June 7, 2009–Nov 30, 2009



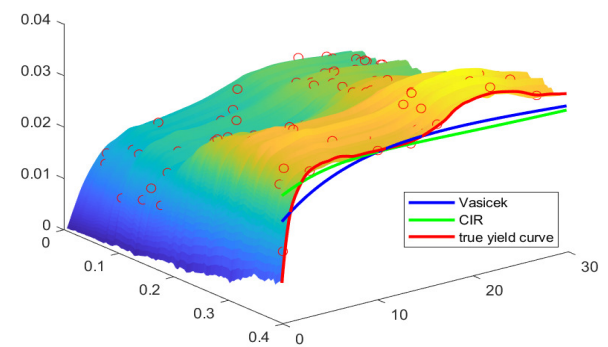
(c) yield curves, Nov 24, 2017



(d) surface, July 3, 2017–Nov 24, 2017



(e) yield curves, Jun 2, 2022



(f) surface, Jan 7, 2022–Jun 2, 2022

Figure 2. Basketsing through US Treasury instruments. The figure continues the presentation of the results for the yield curve given in Figure 1.

The yield curves obtained through the Vašíček and CIR models are presented in the left graphics of Figures 1 and 2 by blue and green lines, respectively. The true curves are plotted in red. We can see that the Vašíček model accurately forecasts the realized term structures, whereas CIR model produces relatively larger deviations. The possible reason is that the domain of its underlying distribution (noncentered gamma) is the positive real half-line. This hypothesis is supported by the results of the next section, as well. The whole yield surface during the related hundred-day periods is presented in the right graphics. The red circles mark the unique daily observations that we use to calibrate the models in our algorithm. They are chosen through the random numbers n_i , $i = 1, 2, \dots, 360$ reported in Appendix A. Note that no other information is used for building the yield curves. The realized term structures as well as the Vašíček and CIR forecasts are presented by the same colors. Let us pay attention to Examples (C) and (H); see Figures 1e and 2e. These periods are marked by a significant volatility. The true curves are very specific, especially the second one. Although to a lesser extent for the second, both models capture the curve shapes. Two other very specific examples are (D) and (G). The CIR model returns too unrealistic curves characterized by a very high initial value; see the first column of Figure 6. In Section 7, we present a possible approach for avoiding this problem.

Table 1. Backtesting. The table presents the calibration results for eight examples based on restricted US historical data. We also report the root mean square deviations as well as the results of the Diebold–Mariano statistical test.

Vašíček	ξ	η	$\xi\eta$	θ	MLE value	RMSE	DM
(A)	0.1950	0.0974	0.0190	0.0521	-40.7062	0.0028	5.4337
(B)	0.1691	0.0916	0.0155	0.0501	-38.8291	0.0038	-9.3990
(C)	0.1507	0.0967	0.0146	0.0466	10.8965	0.0026	-7.8344
(D)	0.1049	0.1248	0.0131	0.0445	-626.3954	0.0048	-5.4341
(E)	0.2406	0.0907	0.0218	0.0693	-202.7635	0.0026	-8.1671
(F)	0.0852	0.0525	0.0045	0.0157	409.9504	0.0012	8.2582
(G)	0.0437	0.0542	0.0024	0.0108	213.1041	9.5706×10^{-4}	-7.8410
(H)	0.0664	0.0715	0.0047	0.0220	86.6914	0.0033	-6.1251

CIR	p-value						
(A)	0.1093	0.1202	0.0131	0.2164	-830.7121	0.0020	1.0206×10^{-7}
(B)	0.0903	0.1117	0.0101	0.2243	-928.7758	0.0048	6.6253×10^{-19}
(C)	0.1019	0.1097	0.0112	0.2407	-794.9703	0.0034	5.3861×10^{-14}
(D)	0.0278	0.1653	0.0046	0.1923	-7681.6	0.0071	1.0183×10^{-7}
(E)	0.0202	0.1945	0.0039	0.0950	-11104	0.0077	5.4373×10^{-15}
(F)	0.0564	0.0644	0.0036	0.1006	229.7305	0.0011	2.8711×10^{-15}
(G)	0.0113	0.1752	0.0020	1.4835	-2563.8	0.0132	5.1494×10^{-14}
(H)	0.0309	0.0874	0.0027	0.1243	-796.9049	0.0041	2.3815×10^{-9}

The calibrated parameters as well as the MLE values are reported in Table 1. We also present the root mean square deviations (RMSD). Furthermore, we perform the Diebold and Mariano (2002) statistical test to compare the forecasts of both models. The results are also provided in Table 1. The statistics

are given in the first part of this table because the test is conducted in the order Vašíček vs. CIR. The related p -values are given in the second part of the table because they are symmetric w.r.t. the models. There are two examples, (A) and (F), with positive statistics and low enough p -value which predict an advantage for the CIR model. On the other hand, the RMSDs are relatively close (0.0028 vs. 0.0020 and 0.0012 vs. 0.0011). All other cases indicate an advantage for the Vašíček model.

6. Evidence from Bulgaria

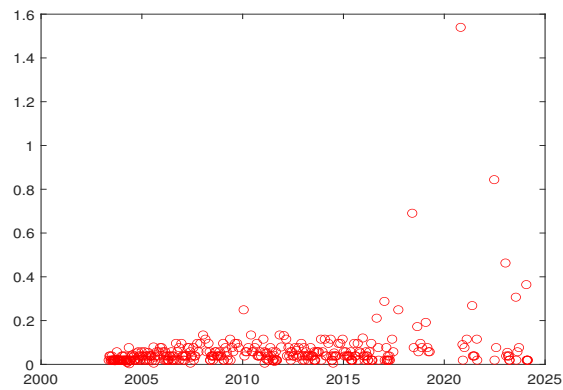
6.1. Data description

We use historical data from the primary market consisting of different bond prices reported in the official website of the Bulgarian National Bank[‡] between January 2002 and December 2022. There is information about the minimal, maximal, and average-weighted prices of the bids approved per 100 units nominal value; we use the average-weighted ones in our investigation. The total number of observations is 356 for 21 years or an approximate average of 17 observations per year. The largest period without observations is approximately a year and a half, between December 6, 2017 and June 21, 2019. There are only two consecutive days at which the government issued a new debt. The periods between two dates averages 21 days. The length of the periods between the observations measured in years is depicted in Figure 3a. The time to maturity varies between 3 months and 20 years and they are depicted in Figure 3b. Between the blue vertical lines is the above-mentioned largest period without observations. We can clearly see that the life lengths of the emissions are mainly three months, a year, three years, five years, and ten years. There is also one twenty-year emission. Intermediate points are for reopened emissions. The issued securities are two kinds, treasury bills (total 76 observations) and treasury bonds (total 280). The first ones are without coupon payments. On the contrary, the larger part of the treasury bonds pay coupons twice per year; only nine issued bonds are without coupons. There is an important feature of the first coupon: It does not consist of the whole amount but only of the proportion for the days between the first coupon date. The unique date on which the Bulgarian government issues more than one security is June 21, 2019; the government sells ten- and twenty-year emissions. We remove from the statistical sample the twenty-year bond, but we use it later to check the results of the forecasting algorithm.

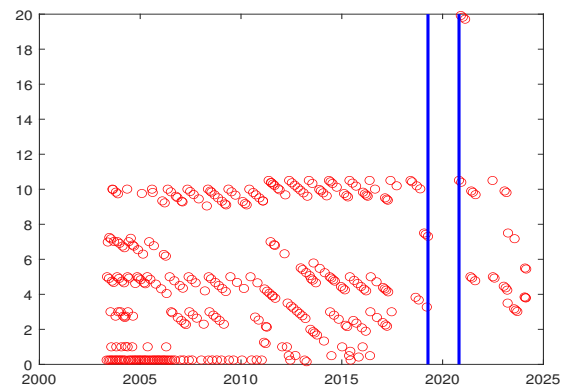
On the other hand, in its official website,[§] the Bulgarian National Bank presents information about the secondary market. One can easily see that most of the reports are about the repurchase agreements (repos). Also, the bid–ask spreads are not available. In addition, there are no reports for many dates and no information at all before 2019. Moreover, there are many reports for one and the same bonds at one and the same dates but with different prices. All these make the information from the secondary market unsuitable for extracting the yield curve.

[‡]<https://www.bnb.bg/FiscalAgent/FABankInformation/FAGSMarket/index.htm>

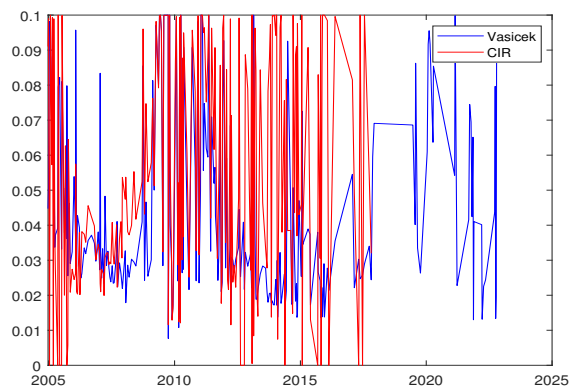
[§]see, <https://www.bnb.bg/FiscalAgent/FABankInformation/FADealsDCK/index.htm>



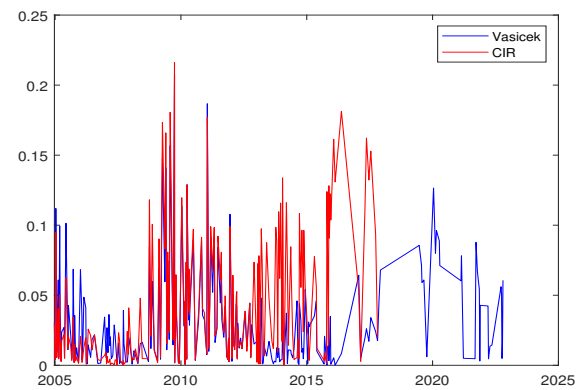
(a) Periods between the observations



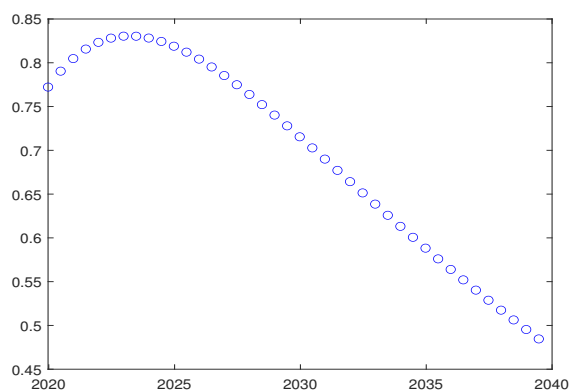
(b) Frequence of the observations and time to maturity



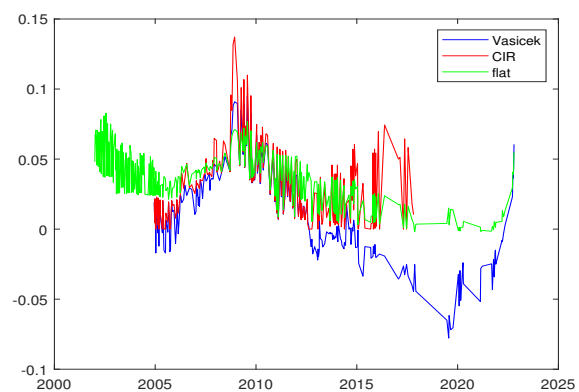
(c) Forecasting interest rates



(d) Forecasting bond prices



(e) Bond structure



(f) Short rates

Figure 3. Figure 3a illustrates the length of the periods between two observations; Figure 3b is for the time to maturities; Figure 3c presents the goodness-of-fit obtained through Equation (25); Figure 3d presents the deviations between the forecasted and actual prices; Figure 3e gives the present value of the coupons for a twenty-year treasury bond issued on June 21, 2019; Figure 3f depicts the short rates obtained from the calibrated models as well as the realized interest rate.

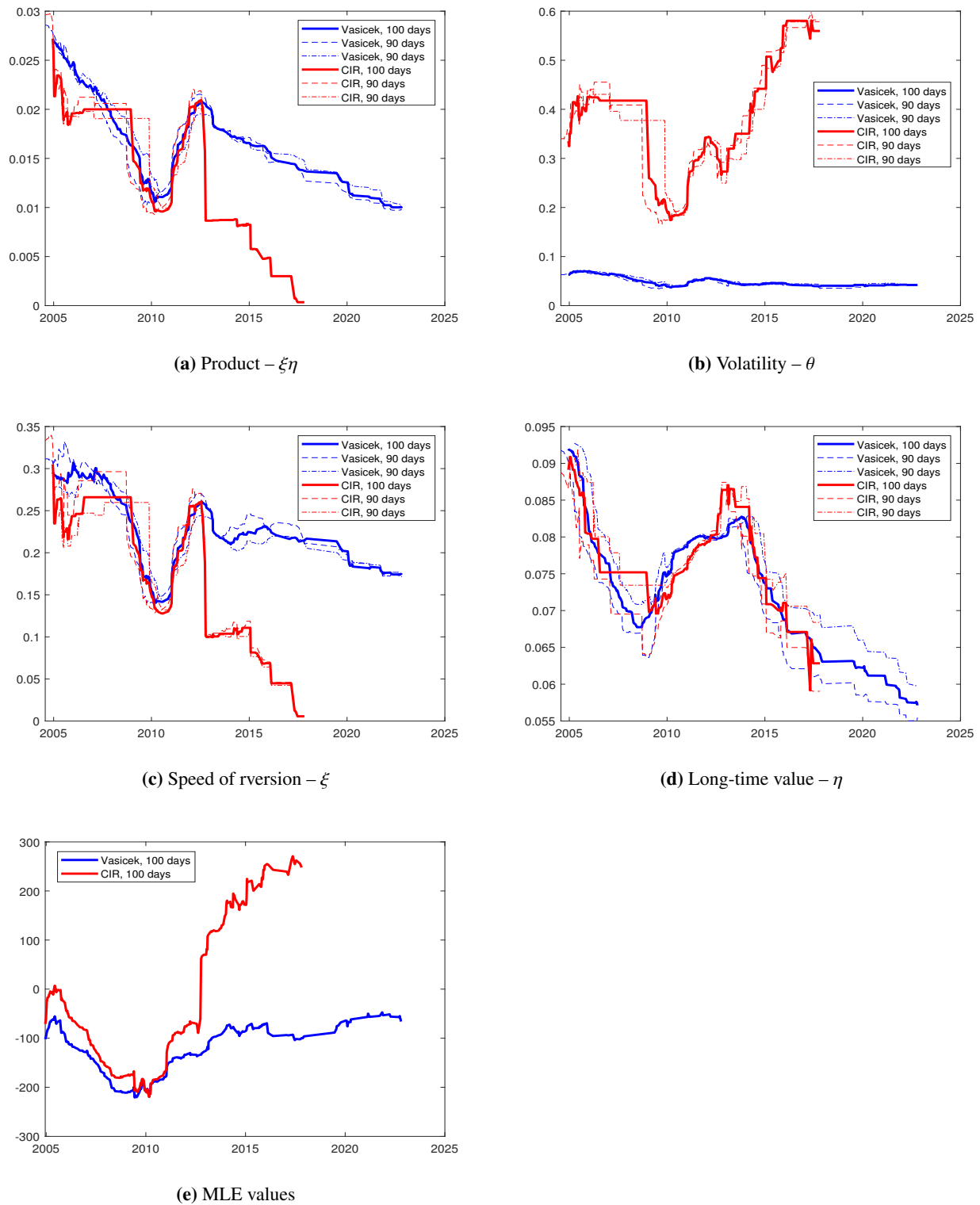


Figure 4. Values of the parameters during the time. The figure presents the calibrated parameters and the value of the optimization criterion over time. The bold lines are for the calibration based on a 100-step window. The dotted lines are for 90- and 110-step windows.

6.2. Values of the parameters during the time

We calibrate both models using a moving window with size $n = 100$. The maximization of criterion (20) is prepared by the use of the standard MATLAB functions `fminsearch` and `fminbnd` as well as of nonlinear programming based on simulated annealing. Having in mind that the initial point is of outstanding importance in the multivariate optimization, we use the derived parameters in the previous iteration as an initial value for the next one. We compare the obtained result with another, for which we do not impose an initial condition. This approach is effective because usually, the new parameters are close to the previous ones. However, sometimes, the market realities may change significantly, leading to a very different new estimation. It turns out that after October 25, 2017, the minimization algorithm cannot find even one triple $\{\xi, \eta, \theta\}$ for the CIR model which leads to positive values for the short-rate. Even though such a triple exists, it would be useless due to the large deviations from the empirical observations. For this reason, we restrict the investigation of the CIR model until this date. The results are presented in Figures 4a–4e. Recall that the bond prices are obtained under the risk-neutral measure, not under the real one. In this light, the risk premium that connects both measures is usually chosen in a way that keeps the term $\xi\eta$. Also, the Girsanov theorem leads to one and the same value of θ under both measures. These joint terms are presented in Figures 4a and 4b, whereas Figures 4c and 4d are for ξ and η themselves. The produced values for the MLE criterion can be seen in Figure 4e. The values of $\xi\eta$ and θ for the Vašíček model vary in relatively closed diapasons; $\xi\eta$ are between 0.0099 and 0.0272, whereas θ is between 0.0371 and 0.0705. The parameters ξ and η have their own significance, as they vary in the intervals (0.1406, 0.3071) and (0.0571, 0.0920), respectively.

Despite the restrictions of the historical sample for the CIR model, we can again distinguish clearly two periods. During the first one, $\xi\eta \in (0.0096, 0.0272)$, $\theta \in (0.1741, 0.4269)$, $\xi \in (0.1277, 0.3052)$, and $\eta \in (0.0696, 0.0909)$. Note that the values of $\xi\eta$ and θ are typical for both models. On the other hand, the parameters of the CIR model are highly volatile during the second period. Furthermore, the value of the MLE criterion increases to very unusual values. All these can be explained by the negative rates that are usual in this period and impossible in the CIR framework. ¶

The parameter values based on the full data are reported in the first part of Table 2. Recall that for the CIR model, we use only the first 323 observations because the dataset is restricted. The MLE values are 567.2231 and 176.3192 for the Vašíček and CIR estimations, respectively. The values of $\xi\eta$ and θ are 0.0161, 0.0535 (Vašíček) vs. 0.0160, 0.3401 (CIR). The values of ξ and η themselves are {0.1541, 0.1046} (Vašíček) and {0.1645, 0.0973} (CIR). We provide an additional test based only on the first period. The results are reported in the second part of Table 2: 150.2931 and 64.7517 for the MLE, 0.0223 and 0.0177 for $\xi\eta$, 0.0681 and 0.3854 for θ , 0.2384 and 0.1977 for ξ , 0.0936 and 0.0894 for η . We can conclude that the results are typical for both models.

It should be noted that all obtained parameters for the CIR process lead to $2\xi\eta < \theta^2$, which strongly indicates that negative interest rates are not uncommon. Also, this means that the CIR process can reach zero, and there is a relatively high mass close to zero. Regardless that the condition $2\xi\eta < \theta^2$ is not unrealistic from the practical point of view, one must treat it very carefully if working jointly under the risk-neutral and real-world probability measures. This is due to the possible appearance of a local martingality instead of true one. This is not the case in the current study, as we work mainly under the risk-neutral measure.

¶All parameter values are available and can be provided upon request.

Table 2. Parameters and MLE values for the calibrations based on the entire Bulgarian dataset, as well as on its initial subsample.

Total data	ξ	η	$\xi\eta$	θ	MLE value
Vašíček model	0.1541	0.1046	0.0161	0.0535	567.2231
CIR model	0.1645	0.0973	0.0160	0.3401	176.3192
first period					
Vašíček model	0.2384	0.0936	0.0223	0.0681	150.2931
CIR model	0.1977	0.0894	0.0177	0.3854	64.7517

6.3. Bond structure

Given that the current bond price $P(0, T)$ is the sum of the present values of all future cash flows (all coupons and the face value), it follows that the bond structure, which is formed by them, is closely related to the yield curve.

Definition 6.1. Let the bond pay N coupons with the amount of C_i , and each of them is due after T_i , $i = 1, 2, \dots, N$, time. Let also F be the face value (we abandon the restriction $F = 1$). The bond structure is formed by the following $N + 1$ -dimensional vector $\{\alpha_1, \alpha_2, \dots, \alpha_n; \beta\}$:

$$\begin{aligned}\alpha_i &= C_i e^{-y(0, T_i)T_i} \\ \beta &= F e^{-y(0, T)T}.\end{aligned}\tag{21}$$

The following presentation for the bond price stands:

$$P(0, T) = \beta + \sum_{i=1}^N \alpha_i.\tag{22}$$

We have mentioned above that the Bulgarian government issued debt maturing at different dates only on June 21, 2019, namely ten- and twenty-year emissions. Their coupon rates were 0.5 and 1.5, respectively. The twenty-year bond has been removed from the statistical sample, but we can use it for checking the performance of the Vašíček model; this date does not fall in the area for the CIR model. The obtained parameters are

$$\{\xi_{Vas} = 0.2136, \eta_{Vas} = 0.0632, \theta_{Vas} = 0.0404\}.\tag{23}$$

Based on them, we derive the current short rate, $r_{Vas} = -0.0650$. This value leads to the twenty-year bond price $P_{Vas} = 92.4675$. The realized price is $P = 98.46$. The relatively large deviation can be explained by the fact that the previous debt was issued a year and a half ago, December 6, 2017. The coupon part of the bond structure (the present value of all coupons) is presented in Figure 3e, and the impact of the face value is $\beta_{Vas} = 64.6122$.

6.4. Realized rate of return

Let us first extract the realized return for all issued bonds. Although the Bulgarian National Bank reports these rates in its official website, we derive them ourselves because the bank uses a methodology based on the EXCEL function `yield`. It calculates the interest rate under the assumption of a discrete-time. However, we work in a continuous time, and thus, some little deviations appear. We use the following method. Let F be again the bond face value, $C_1, C_2, \dots, C_n$ be the coupon payments, τ be the length of the bond life measured in years, $\tau_1, \tau_2, \dots, \tau_n$ be the remaining time to maturity after the date of the corresponding coupon, and P be the current bond price. Let us consider the function

$$f(r) = Fe^{-r\tau} + \sum_{i=1}^n C_i e^{-r(\tau-\tau_i)}. \quad (24)$$

Obviously, function (24) decreases between $+\infty$ and 0 for $r \in \mathbb{R}$ because $F > 0$ and $C_i > 0$. Hence, the equation $f(r) = P$ has a unique solution. Namely, this solution is the realized rate of return because Equation (24) gives the present value of all cash flows. Let us comment on the specifics of the Bulgarian treasury bonds. They pay coupons twice per year, and the last one is at maturity. Hence, $n = [2\tau] + 1$, where we denote by $[x]$ the integer part of the positive number x . Hence, $\tau_i = (n - i) / 2$, which means that $\tau_1 = [2\tau]/2, \dots, \tau_{n-1} = 1/2, \tau_n = 0$. Note that the last coupon is payable together with the face value at the maturity date. The coupon amount $C_i, i = 2, \dots, n$ is obtained as $C_i = 50d$, where d is the coupon interest rate. As we mentioned above, the first coupon is specific. It pays only a part of the whole amount proportional to the period before it. Hence, $C_1 = 50kd$, where $k = 2\tau - [2\tau]$.

The realized returns together with the generated by the models of Vašíček and Cox–Ingersoll–Ross yields can be seen in Figure 3f. We can see that in the first above-mentioned period, all rates follow one and the same trend. On the contrary, there are significant deviations during the second period.

6.5. Validation and forecasting

We perform now several experiments to validate our results. We use the already-calibrated parameters on the moving window to check the forecasts that both models produce. To do this, we need a suitable criterion based on the PDFs. We suggest the following formula:

$$c_j = \frac{1}{p(r_{t_{j+1}} | r_{t_j}, t_j, t_{j+1}; \xi_j, \eta_j, \theta_j) + C}, \quad (25)$$

where C is a positive constant. Thus, the possible values are reduced to the interval $(0, 1/C)$. The value r_{t_j} is the rate of return that the corresponding model produces at the end of the j th window, that is, at the moment t_{j+99} . Note that the value $r_{t_{j+1}}$ is still not known at time t_j . This way, the term (25) measures the models' forecasts: The lower the value of c_j , the better the forecast. The results for $C = 10$ are presented in Figure 3c, in blue for the Vašíček model and in red for the CIR one. Note that some extremely low CIR values are due to a feature of the noncentered gamma distribution; some parameters make its initial value infinitely large.

The second validation test we provide is related to the bond prices. Once we have the calibrated parameters at the moment t_j , we can check what prices these parameters forecast for the next moment t_{j+1} . The percent deviations from the realized prices can be seen in Figure 3d. The average deviation is 2.98% for the Vašíček forecast, whereas the CIR model deviates averagely by 4.01%.

6.6. Yield curves

We present and discuss now some yield curves, in fact, forecasts for the yield curves, that our approach produces. We chose the dates December 16, 2004, July 29, 2009, March 2, 2010, October 25, 2017, February 24, 2021, and October 19, 2022 due to the following reasons: Dec 16, 2004 is the 100th date of our dataset, and thus, it is the first date for which we have a calibration; July 29, 2009 is the date with the highest realized return (7.69%); March 2, 2010 saw both models producing very specific curves; Oct 25, 2017 is the last date for which we have calibrations for both models; on Feb 24, 2021 the issued debt is without coupon payments and with largest maturity (five years); Oct 19, 2022 is the last date in our data. Note that for the last two dates, we have calibrations only for the Vašíček model.

Table 3. Short rates extracted from the dataset of the Bulgarian Treasury securities.

	16-Dec-2004	29-Jul-2009	02-Mar-2010	25-Oct-2017	24-Feb-2021	19-Oct-2022
Vašíček model	1.94%	9.51%	5.69%	−4.46%	−2.81%	6.05%
CIR model	1.93%	11.00%	6.18%	1.03%	—	—
Realized return	2.86%	7.69%	5.65%	0.42%	−0.17%	5.49%

The yield curves that both models generate are presented in Figures 5a–5f. The Vašíček curves are in blue, and the CIR ones are in red. We mark by the corresponding circles the short rates and report them in Table 3. In fact, they are the initial values of the yield curves.

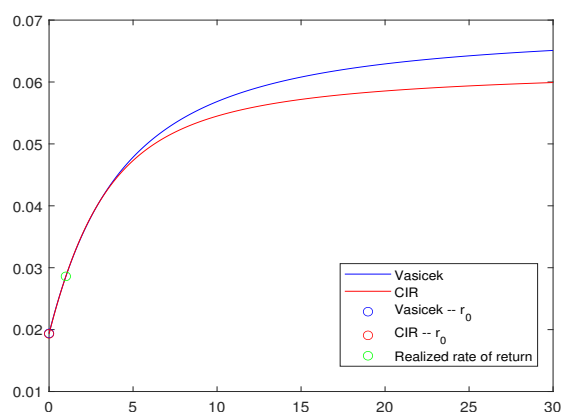
Let us discuss the realized returns marked by green points. They are close to the yield curves at the related maturities. In other words, the green points are near the intersections of both curves. Furthermore, these points coincide when the bond is without coupon payments, which is due to Formula (1) and the way we derive the realized rate in Section 6.4. The results for Feb 24, 2021 generate such an example. As one might expect, the curves are closer during the first period, when the interest rates are positive, and more distant otherwise.

A question that arises during the periods of permanent positive rates is which model to choose. A possible solution is to use a Nelson–Siegel approximation to combine the Vašíček and CIR curves. Recall that this is a four-parameter model that interpolates the yield curve through functions of the form

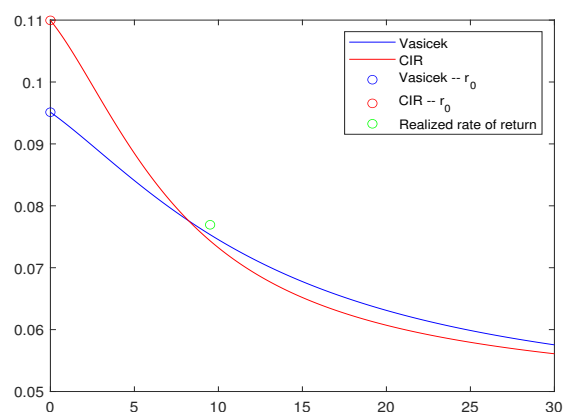
$$y(t) = \beta_0 + \beta_1 \left(\frac{1 - e^{-\lambda t}}{\lambda t} \right) + \beta_2 \left(\frac{1 - e^{-\lambda t}}{\lambda t} - e^{-\lambda t} \right). \quad (26)$$

The results of this approach for March 2, 2010 are depicted in cyan in Figure 5c. The calibrated parameters are $\beta_0 = 0.0442$, $\beta_1 = 0.0153$, $\beta_2 = 0.0225$, and $\lambda = 0.2485$.

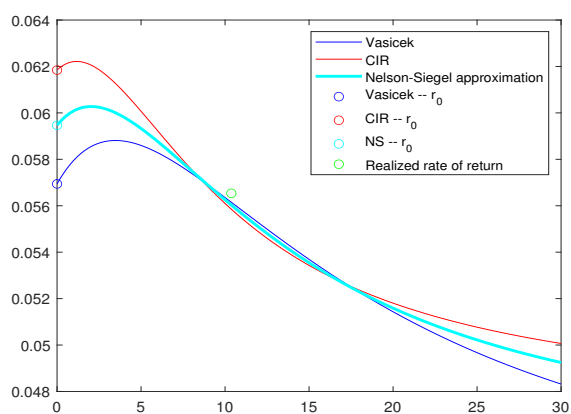
A recent application of the Nelson–Siegel model covering the Canadian yield curves during the COVID–19 period can be found in Severino et al. (2022).



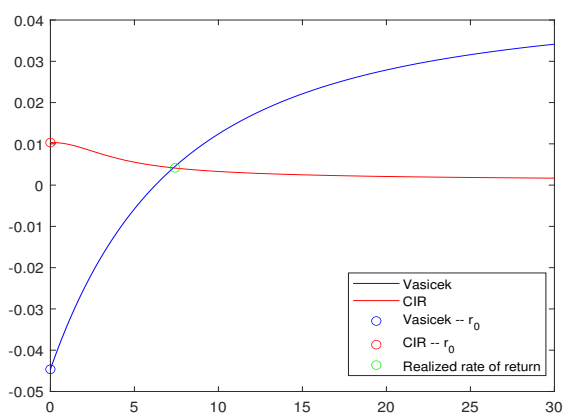
(a) yield curves, Dec 16, 2004



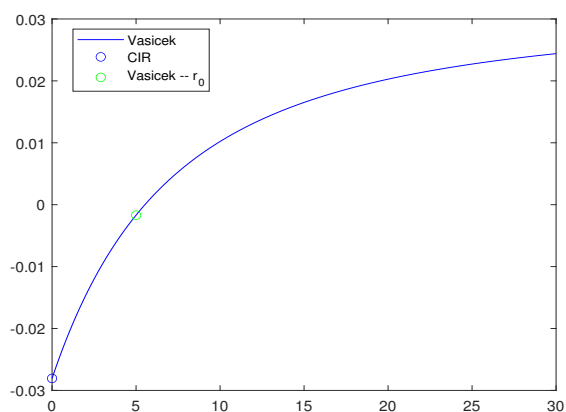
(b) yield curves, July 29, 2009



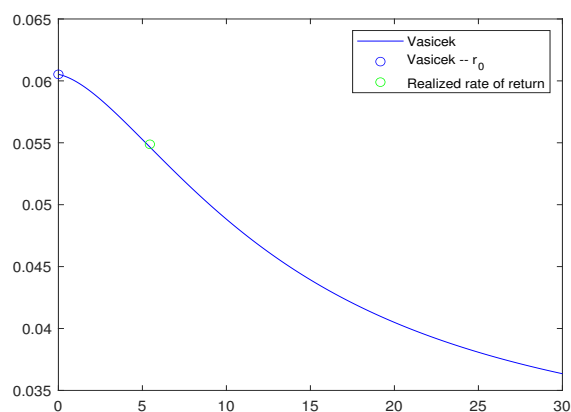
(c) yield curves, March 2, 2010



(d) yield curves, Oct 25, 2017



(e) yield curves, Feb 24, 2021



(f) yield curves, Oct 19, 2022

Figure 5. Forecasts and yield curves. The figure presents the yield curves extracted from Bulgarian historical data.

7. Concluding remarks and further works

In this paper, we propose a new method for constructing the yield curve based only on the historical prices realized at the primary market of government debt. We have incorporated a stochastic process to compensate for the limited information assuming that the bond market is described by either the Vašíček or the Cox–Ingersoll–Ross model. Our method has a semi-statistical feature, as it is based on extracting historical short-rate values through theoretical models. The maximum likelihood estimation method is adapted to the specifics of the examined problem. We validate and backtest the proposed method using the information for US Treasury bills, notes, and bonds. In addition, we have applied the method to a historical dataset for the Bulgarian sovereign debt. Several practical experiments have been provided. We conclude that both models are applicable to our task. However, the results strongly suggest that the Vašíček model and the related Ornstein–Uhlenbeck process are more suitable for the available data, especially when interest rates are negative.

Let us discuss some possible generalizations of our approach that are a task for further works. First, suppose that at some date, we have more than one observation. This would generate more than one potential value for the short rate at this date. We can approximate the actual value averaging the obtained ones. Moreover, different weights can be used taking into account different factors, such as maturities, emission volumes, and so on. Furthermore, we can use multifactor affine models. This would lead to at least two significant advantages. We will discuss briefly how the model may be calibrated. Let the number of factors be n . Suppose that for some date, we have k observations. If $k = n$, then we reach an $n \times n$ linear system. After solving it, we derive the values of all factors at that date. If at some date, $k > n$, then we have $\binom{k}{n}$ combinations of order n . Each of them leads to an $n \times n$ system, and thus, we reach $\binom{k}{n}$ solutions for the factors. We can proceed as above by averaging possibly with proper weights. If $k < n$, then we can solve the corresponding $k \times k$ -system to obtain only k of the factors and ignore the rest. We can apply again the above-described MLE algorithm to calibrate the parameters through the transition densities. Dates for which $k < n$ should be excluded from the time series for the ignored factors. Alternatively, we can use the most likely w.r.t. the previous date term values that in addition solves the $k \times n$ -system. In addition to that enlarged applicability of the multifactor models, they can capture more precisely the yield curve features.

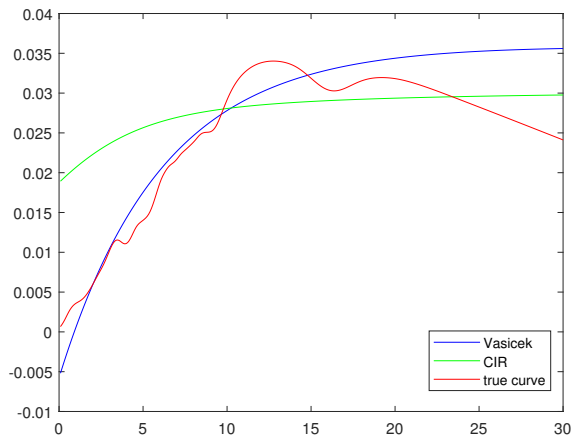
As noted earlier, the CIR model produces very unrealistic estimations in the examples (D) and (G) due to its positive domain. If even one price observation is close enough to one, then the corresponding short rate value will be negative, and thus, the probability density function taken there will be zero. Hence, Criterion (16) will be zero, too. Criterion (20) will be infinitely small. That way, we will reject some values for the triple $\{\xi, \eta, \theta\}$ that are not so bad. A natural way to solve this problem is to add some positive constant to the corresponding density. A natural choice is one, as a negative rate will not affect the MLE criterion and thus will not automatically reject the corresponding triple $\{\xi, \eta, \theta\}$.

Another problem stands with the CIR model. If $2\xi\eta < \theta^2$, then the degrees of freedom in the noncentral χ^2 distribution is less than two.^{||} Hence, its density function tends to infinity in zero. If even one observation leads to a positive but small enough value for the short rate, then the corresponding term in the criterion (no matter (16) or (20)) will be large enough to increase too much the whole criterion value. Thus, just one observation may lead to an arbitrarily large likelihood. This is an inappropriate feature. To avoid it, we can transform the PDF through a decreasing, positive, and bounded function

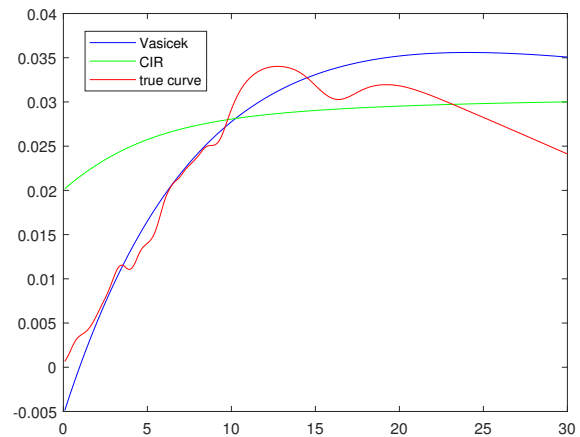
^{||}Note that in this case, the short-rate process can touch zero

defined on the positive half line. One possible choice is $1/(1+x)$. That way, we have a minimizing problem. Thus, the criterion would turn into

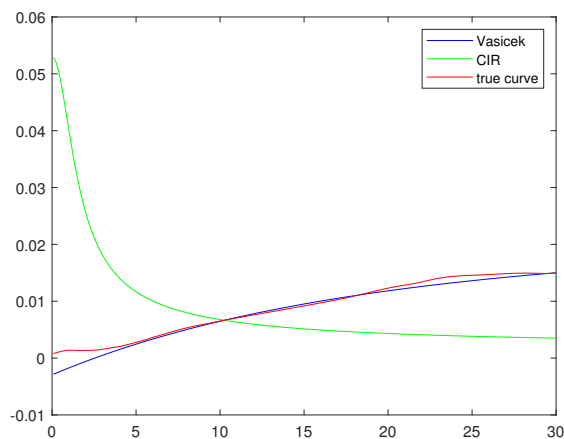
$$F(\xi, \eta, \theta) = \sum_{j=2}^i \left(\frac{1}{p(r_{t_j} | r_{t_{j-1}}, t_{j-1}, t_j; \xi, \eta, \theta) + 1} + \lambda (r_{t_j} - \eta)^2 \right). \quad (27)$$



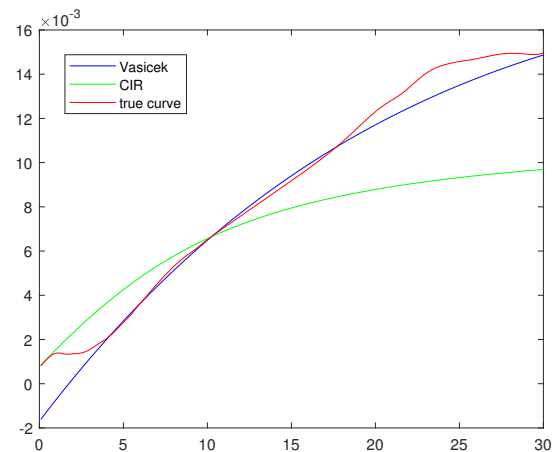
(a) yield curves, Dec 31, 2008



(b) modified yield curves, Dec 31, 2008



(c) yield curves, Sept 25, 2020



(d) modified yield curves, Sept 25, 2020

Figure 6. Modified criterion. The figure presents the yield curves obtained through a modified optimization criterion.

The values obtained through Criterion (27) for $\lambda = 4$ are reported in Table 4 and visualized in the second column Figure 6. The first column is for the original criterion (20). These results indicate that the modification (27) has its own importance. Of course, it can be used only after a detailed investigation whether and when it satisfies the "nice" properties of the classical MLE method, such as consistency,

asymptotic efficiency and normality, sufficiency, and so forth. This goes beyond the scope of the current study and is left for a future research.

Table 4. Backtesting results through the modified criterion.

Vašíček	ξ	η	$\xi\eta$	θ	MLE	RMSE	DM
(D)	0.0686	0.1673	0.0115	0.0377	−30.1794	0.0048	−13.1900
(G)	0.0206	0.0962	0.0114	0.0092	−14.4405	7.3926×10^{-4}	−5.4836
CIR							p-value
(D)	0.0526	0.0825	0.0043	0.1565	−63.6374	0.0073	7.8760×10^{-8}
(G)	0.0287	0.0558	0.0016	0.1742	−75.0398	0.0031	1.2815×10^{-40}

From a practical point of view, it is interesting to be seen how Bulgarian government debt will behave after the country enters the Eurozone on January 1, 2026. We can expect that the interest rates will converge to the rates in the Eurozone, more precisely, to those of the countries with similar profiles. On other hand, the Bulgarian currency has been pegged to the euro** for nearly thirty years, and the effects of this convergence may already be exhausted. Future observations and investigations can confirm or reject this hypothesis.

Alternatively, if there are indications of negative interest rates, we can consider the so-called *shifted* CIR-model. Its main characteristic is a deterministic shift added to the original CIR dynamics. There are several approaches to achieve this. First, a suitable negative constant will move the PDF support to the left, allowing for negative rates. Consequently, the number of model parameters increases from three to four. On the other hand, a suitable deterministic function can be used for the shift; we refer to Chapter 3.9 of the classic work by Brigo and Mercurio (2006) as well as to some novel investigations such as Di Francesco and Kamm (2022). In this light, the CIR model can be useful, but it must be applied with caution when the interest rates are negative or close to zero. This is beyond the scope of this paper and is left for future research.

A. The used random numbers

The generated 100 random numbers for the backtest purposes of Section 5 (forming the vector n) are:

293, 326, 46, 328, 228, 36, 100, 197, 344, 347, 57, 349, 344, 175, 288, 51, 152, 329, 285, 345, 236, 13, 305, 336, 244, 273, 267, 141, 236, 62, 254, 12, 100, 17, 35, 296, 250, 114, 342, 13, 158, 137, 275, 286, 68, 176, 160, 233, 255, 271, 100, 245, 236, 59, 43, 179, 345, 123, 211, 81, 270, 92, 182, 251, 320, 345, 197, 50, 54, 93, 302, 92, 293, 88, 334, 126, 71, 91, 222, 170, 127, 299, 211, 198, 330, 103, 272, 271, 137, 204, 28, 20, 191, 280, 336, 47, 205, 169, 5, 122.

Use of AI tools declaration

The author declares that he has not used Artificial Intelligence (AI) tools in the creation of this article.

**To the DM in the beginning.

Data availability

The data used during the work on this article can be provided upon request.

Acknowledgments

The author would like to express sincere gratitude to the anonymous reviewers for the helpful and constructive comments which substantially improve the quality of this paper.

The work was supported by the Centre of Excellence in Informatics and ICT under Grant No BG16RFPR002-1.014-0018-C01, financed by the Research, Innovation and Digitalization for Smart Transformation Programme 2021-2027 and co-financed by the European Union.

Conflict of interest

The author declares that he has no conflicts of interest.

References

- Black F, Karasinski P (1991) Bond and option pricing when short rates are lognormal. *Financ Anal J* 47: 52–59. <https://doi.org/10.2469/faj.v47.n4.52>
- Brennan MJ, Schwartz ES (1982) An equilibrium model of bond pricing and a test of market efficiency. *J Financ Quant Anal* 17: 301–329. <https://doi.org/10.2307/2330832>
- Brigo D, Mercurio F (2006) *Interest Rate Models - Theory and Practice: With Smile, Inflation and Credit*. Springer Finance. Springer Science & Business Media, 2nd, <https://doi.org/10.1007/978-3-540-34604-3>
- Chen YT, Wui C, Yeh CY (2022) Asset pricing tests of infrequently traded securities: The case of municipal bonds. *Rev Asset Pricing Stud* 12: 754–807. <https://doi.org/10.1093/rapstu/raac002>
- Christensen JHE, Diebold FX, Rudebusch GD (2009) An arbitrage-free generalized Nelson–Siegel term structure model. *Econom J* 119: C33–C64. <https://doi.org/10.1111/j.1368-423X.2008.00267.x>
- Christensen JHE, Diebold FX, Rudebusch GD (2011) The affine arbitrage-free class of Nelson–Siegel term structure models. *J Econom* 164: 4–20. <https://doi.org/10.1016/j.jeconom.2011.02.011>
- Constantinides GM (1992) A theory of the nominal term structure of interest rates. *Rev Financ Stud* 5: 531–552. <https://doi.org/10.1093/rfs/5.4.531>
- Cortazar G, Schwartz ES, Naranjo LF (2007) Term-structure estimation in markets with infrequent trading. *Int J Financ Econ* 12: 353–369. <https://doi.org/10.1002/ijfe.317>
- Cox JC, Ingersoll JE Jr, Ross SA (1985) A theory of the term structure of interest rates. *Econometrica* 53: 385–407. <https://doi.org/10.2307/1911242>
- Di Francesco M, Kamm K (2022) On the deterministic-shift extended CIR model in a negative interest rate framework. *Int J Financ Stud* 10: 38. <https://doi.org/10.3390/ijfs10020038>

- Diebold FX, Mariano RS (2002) Comparing predictive accuracy. *J Bus Econ Stat* 20: 134–144. <https://doi.org/10.1198/073500102753410444>
- Dothan LU (1978) On the term structure of interest rates. *J Financ Econ* 6: 59–69. [https://doi.org/10.1016/0304-405X\(78\)90020-X](https://doi.org/10.1016/0304-405X(78)90020-X)
- Feller W (1951) Two singular diffusion problems. *Ann Math* 54: 173–182. <https://doi.org/10.2307/1969318>
- Gürkaynak RS, Sack B, Wright JH (2007) The US treasury yield curve: 1961 to the present. *J Monetary Econ* 54: 2291–2304. <https://doi.org/10.1016/j.jmoneco.2007.06.029>
- Heath D, Jarrow R, Morton A (1992) Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. *Econometrica* 60: 77–105. <https://doi.org/10.2307/2951677>.
- Ho TSY, Lee SB (1986) Term structure movements and pricing interest rate contingent claims. *J Financ* 41: 1011–1029. <https://doi.org/10.1111/j.1540-6261.1986.tb02528.x>
- Khakimzhanov S, Mustafin Y, Kubenbayev O, et al. (2019) Constructing a yield curve in a market with low liquidity. *Russ J Money Financ* 78: 71–98. <https://doi.org/10.31477/rjmf.201904.71>
- Liu Y, Wu JC (2021) Reconstructing the yield curve. *J Financ Econ* 142:1395–1425. <https://doi.org/10.1016/j.jfineco.2021.05.059>
- Makushkin M, Lapshin V (2021) Yield curve estimation in illiquid bond markets. *HSE Econ J* 25: 177–195. <https://doi.org/10.17323/1813-8691-2021-25-2-177-195>
- McCulloch JH (1971) Measuring the term structure of interest rates. *J Bus* 44: 19–31. <https://doi.org/10.1086/295329>
- Nagy K (2020) Term structure estimation with missing data: Application for emerging markets. *Q Rev Econ Financ* 75: 347–360. <https://doi.org/10.1016/j.qref.2019.04.002>
- Nelson CR, Siegel AF (1987) Parsimonious modeling of yield curves. *J Bus* 60: 473–489. <https://doi.org/10.1086/296409>
- Payne J, Szőke B, Hall GJ, et al. (2023) Estimating historical yield curves with sparse data. *Work. Pap.*, Princeton University. Available from: https://www.balintszoke.com/wp-content/uploads/2023/07/PSHS_method_07052023_website.pdf.
- Pearson ND, Sun TS (1994) Exploiting the conditional density in estimating the term structure: An application to the Cox, Ingersoll, and Ross model. *J Financ* 49: 1279–1304. <https://doi.org/10.1111/j.1540-6261.1994.tb02454.x>
- Pitman J, Yor M (1982) A decomposition of Bessel bridges. *Probab Theory Relat Field* 59: 425–457. <https://doi.org/10.1007/BF00532802>
- Severino F, Cremona MA, Dadié É (2022) COVID-19 effects on the Canadian term structure of interest rates. *Rev Econ Anal* 14: 471–502. <https://doi.org/10.15353/rea.v14i4.4962>
- Svensson LEO (1994) Estimating and interpreting forward interest rates: Sweden, 1992–1994. <https://doi.org/10.3386/w4871>

- Vašíček (1977) An equilibrium characterisation of the term structure. *J Financ Econ* 5: 177–188. [https://doi.org/10.1016/0304-405X\(77\)90016-2](https://doi.org/10.1016/0304-405X(77)90016-2)
- Vašíček OA, Fong HG (1982) Term structure modeling using exponential splines. *J Financ* 37: 339–348. <https://doi.org/10.1111/j.1540-6261.1982.tb03555.x>
- Xie C, Chen H, Yu X (2006) Yield curve estimation in the illiquid market: Framework, models and empirical study. *Int J Inf Technol Decis Mak* 5: 467–481. <https://doi.org/10.1142/S0219622006002064>



AIMS Press

©2026 Tsvetelin S. Zaevski, licensee AIMS Press.
This is an open access article distributed under the
terms of the Creative Commons Attribution License
(<https://creativecommons.org/licenses/by/4.0>)