A Mean Reverting Stochastic Process (MRSP) using an AR(N) Model and A Kalman Filter For Generating Intravalues for The Daily DJIA Time Series

Apostolos P. Leros, Athina P. Bougioukou, And Theodoros I. Maris

General Department, National Kapodistrian University of Athens, 34400 Psachna, Evia, GREECE

lerosapostolos@gmail.com, athina.mpougioukou@gmail.com, theodorosmaris@gmail.com

Abstract

In this paper a mean reverting stochastic process (MRSP) model is presented for generating intravalues of time series. The deterministic or mean part of the process is forecasted by an autoregressive of order n, AR(n), model. The unobservable AR(n) coefficients are calculated by a Kalman Filter using n time series observations. The stochastic part of the process is a Brownian motion multiplied by a volatility term. Measures of the Kalman filter covariance matrix along with the process itself are used to capture the volatility dynamics for the intravalues of the time-series. The MRSP model also provides for the evolution of the intravalues of the time series. The applicability of the model is demonstrated using the daily Dow Jones Industrial Average (DJIA) time series.

Keywords: -Mean reverting stochastic process, Autoregressive model, Kalman filter, Time-series, Intravalues, Process evolution, Dow Jones Industrial Average

1 Introduction

A time series is a set of data points or observations $\{x_t,\ t=0,1,2,...,\mathrm{N}\}$ measured generally at equally spaced time intervals. The values of the time series x_t fluctuate up-and-down as time $t\in[t_0,t_N]$ progresses. These up-and-down fluctuations is termed the high frequency data of the time series. A center value of these fluctuations is termed the volatility of the time series. Within the last decades the estimation of volatility is of major concern. It is of enormous value since it provides information for the future dispersion of the time series.

One volatility measure, regarding say a time series of a stock, is the "breakout volatility", that is, the difference between the high and low values of the previous day. A 70% of this value is considered a volatility estimate. Another measure, is to look into a number of past high and low values along with open and close values and calculate a deterministic value for the volatility [1, 2, 3]. In addition, it is well known that the distribution of the volatility has fat tails, that is, it has very high or low values with high probability [4, 5]. Also, it may have long range dependency, meaning that a value in the long past, say 20 days ago for a stock, may still have an impact in future values. Thus, an additional approach for volatility estimation is to use heterogeneous autoregressive models [4] utilizing past data. Yet, in the daily volatility forecasts higher statistics such as skewness and kurtosis may also be considered based say on a per minute time series. Other approaches consider volatility clustering; the phenomenon of calm periods (relatively no change in values) and periods of high volatility of the time series. Here the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models [5] are very prominent. These approaches since they are autoregressive depend on past square observations and past variances to model current and future variances.

The intraday volatility aspects of time series have been studied extensively [6,7, 8, 9, 10]. In [6] a time-inhomogeneous diffusion model was adopted and using log penalized splines the volatility was estimated for the high-frequency intraday five-minute Shanghai Stock Exchange Composite Index (SHCI). In [7] the long memory and periodicity in intraday volatility was considered through the parameterization of the Fractionally Integrated Periodic EGARCH and the Seasonal Fractional Integrated Periodic EGARCH. In [8] new empirical evidence was provided for intraday scaling behaviour of stock market returns utilizing a 5-minute stock market index (the Dow Jones Industrial Average) from the New York Stock Exchange, showing that the return time series has a multifractal nature during the day. In [9] the volatility in the Japanese stock market based on a 4-year

CC BY

sample of 5-minute Nikkei 225 returns from 1994 through 1997 was considered, showing that the intraday volatility exhibits a doubly U-shaped pattern associated with the opening and closing of the separate morning and afternoon trading sessions on the Tokyo Stock Exchange. In [10] the intraday behaviour of stock returns and trading volume using the data on OMXS 30 stocks was analyzed. They found that returns follow a reverse J-shaped pattern with the peak at the beginning of the trading day, while trading volume attains its maximum towards the market closure. The highest volatility and kurtosis are observed at 09:30-10:00 and 11:30-12:00 daily periods, when the macroeconomic news is released.

Volatility though does not provide the evolution of the time series but the dispersion of the up-and-down changes around a level value. In addition, this paper is concerned with the evolution of the intravalues of a time series. It focuses on a model to provide intravalues or high frequency (tick) forecasts of a time-series using a combination of four algorithmic techniques: (a) A mean reverting stochastic process, (b) a volatility measure, (c) an AR(n) autoregressive model, and (d) a Kalman filter. The mean reverting stochastic model consists of two parts: (i) The deterministic part, giving the evolution of the time series. It is the difference between a mean value the process reverts to and the current value of the process itself. In addition, this difference is multiplied by a positive constant or a positive time dependent function called speed of reversion. (ii) The other part, called stochastic, consists of a Brownian motion multiplied by some volatility term of the time series. This volatility term may be a positive constant, a positive time dependent function or envelop, and/or even a positive function of the process itself. For the estimation of the process mean reverting value, an autoregressive model of order n, [AR(n)], is used based on a sample of n past values. The AR(n) model is formalized in state space. Its unobservable coefficients are realized using a Kalman filter. The trace and/or the determinant of the Kalman filter covariance matrix is used along with the evolution of the process itself to capture the evolution dynamics of the volatility of the time series. Experimental results are presented demonstrating the applicability of the model using the Dow Jones Industrial Average (DJIA) time series.

This paper is organized as follows: Section 2 gives the mathematical preliminaries for the mean reverting stochastic process, the state space formulation of the autoregressive model and the discrete Kalman Filter equations. Section 3 provides the framework for the discretization of the mean reverting stochastic process, and gives possible deterministic and/or stochastic volatility functions which could be used in the model to forecast high frequency values (intravalues, intraday or tick values) of the time-series. Section 4 gives simulation results for the intravalues of the time-series data of Dow Jones Industrial Average (DJIA) along with their assessment. Finally, section 5 provides the conclusions of this work and gives possible further directions.

2 Mathematical Preliminaries

2.1 Description of the Mean Reverting Stochastic Process

The differential equation of the general mean reverting stochastic process S_t, formally, has the form [11, 12]:

$$\begin{aligned} dS_t &= A(t)\left(\mu(t) - S_t\right)dt + G\left(t, S_t\right)dB_t \\ S_0 &= initial\ condition \end{aligned} \tag{1}$$

The parameter $A(t)$ is a constant or a deterministic function of time t but always a positive quantity and denotes the rate (speed or strength) of reversion. The parameter $\mu(t)$ is allowed to be a constant or a deterministic function of t or even a stochastic process and denotes the mean value (also called long term mean) around which the process tends to oscillate. The stochastic term $G\left(t, S_t\right)dB_t$ consists of two parts. The diffusion-coefficient $G\left(t, S_t\right) > 0$, which is a deterministic or even a stochastic positive function of time, and the continuous scalar constant-diffusion $B_t\left(\cdot,\cdot\right)$, being the Brownian motion (or Wiener or Wiener-Levy) process having the following properties:

Property 1: It is a process with independent increments, that is, the set of N random variables $\delta_i(\cdot)=\left[B(t_i,\cdot)-B(t_{i-1},\cdot)\right]$, for $i=0,1,...,N$, are mutually independent for any partition $t_0<t_1<...<t_N$ of the time interval $t_0=0$ and $t_N=T$.

Property 2: At time t_0, $B(t_0,\omega_i)=0$, for all $\omega_i\in\Omega$ (Ω is the probability sample space consisting of all possible outcomes of an experiment and ω_i is an outcome), except possibly at a set ω_i of probability zero (by convention).

Property 3: The process independent increments are Gaussian random variables, such that, for any time instants t_i and t_{i-1} in T, the mean is $E\left[B(t_i)-B(t_{i-1})\right]=0$, and the variance is $E\left\{\left[B(t_i)-B(t_{i-1})\right]^2\right\}=q\left[t_i-t_{i-1}\right]$. The parameter $q\geq 0$ is the diffusion of the process $B(\cdot,\cdot)$.

Usually, the scalar Brownian motion process $\{B_t, t\geq t_0\}$, is assumed to be of unit variance $q=1$. This assumption is without loss of generality, since $q\neq 1$ can be absorbed in the diffusion-coefficient $G(t,S_t)$, that is, $G(t,S_t)$ can be replaced with $\sqrt{q}G(t,S_t)$.

The behaviour of the deterministic part of the process is that when $S_t>\mu(t)$, the so called drift term $A(t)(\mu(t)-S_t)$ is negative, which results in a pull of the process S_t back down toward the equilibrium level. Conversely, if $S_t<\mu(t)$, the drift term $A(t)(\mu(t)-S_t)$ is positive, which pulls the process S_t back up to a higher equilibrium value. For the stochastic term $G(t,S_t)dB_t$, since $B(t_i,\cdot)$ for a given $t_i\in T$ is a random variable composed of a sum of independent Gaussian increments, it is also Gaussian with mean value $m_{B(t_i)}=E\left[B(t_i)\right]=0$ and variance $V_{B(t_i)}=E\left[B(t_i)^2\right]=q\left[t_i-t_{i-1}\right]$. Thus, q indicates how fast the mean square value of $B(\cdot,\cdot)$ diverges from its initial zero value at t_0.

Because the coefficients of the linear equation (1) are measurable (in other words, the random process is completely determined by the realization of S_t) and bounded on the interval $[t_0,T]$, there exists a unique continuous solution for $t_0\leq t\leq T$ taking the form [12]:

$$S_t=e^{-A(t)[t-t_0]}S_0+\int_{t_0}^{t}e^{-A(s)[t-s]}\left[\mu(t)ds+\sigma(s)dB_s\right] \tag{2}$$

The solution (2) has the characteristic that fluctuates randomly, but tends to revert to some fundamental level $\mu(t)$ with some reversion behavior, which depends upon the choices of the speed of the reversion parameter $A(t)>0$, and the nonrandom or random but continuous function $G(t,S_t)>0$.

2.2 State space - Autoregressive model - Discrete-time Kalman Filter

A linear, discrete-time, finite-dimensional system with noisy input and noisy output is described with the following state-space equations [13, 14]:

$$x_{k+1}=F_k x_k+w_k \tag{3}$$

$$z_k=y_k+v_k=H_k{}^T x_k+v_k \tag{4}$$

In equations (3) and (4) the unobserved variables of interest are the system state x_k at discrete-time $k \geq 0$ and y_k is the system output. Since this usually is noisy, a noise process $\{v_k\}$ is added to it resulting in the observed measurement process $\{z_k\}$. The matrices F_k and H_k are of proper dimensions and known. The system input noise process $\{w_k\}$ and the measurement output noise process $\{v_k\}$, are independent of each other and individually each being a Gaussian white (uncorrelated from instant to instant and stationary) noise with zero mean and known covariance, i.e., $\{w_k\} \sim WN(0, Q_k \delta_{ks})$ and $\{v_k\} \sim WN(0, R_k \delta_{ks})$, where δ_{ks} denotes the Kronecker delta which is 1 for $k = s$ and zero otherwise. We assume that the initial state x_0 is a Gaussian random variable with known mean $E[x_0] = \bar{x}_0$ and known covariance $E\{[x_0 - \bar{x}_0][x_0 - \bar{x}_0]^2\} = \bar{P}_0$, and also that it is independent of w_k and v_k, for any k.

Now a model that expresses a univariate time-series system output y_k as a linear combination of past observations y_{k-n} and white noise v_k (which is the observed measurements z_k) is referred to as an autoregressive model of order n [or $AR(n)$] model and is given by the following equation:

$$y_k = -a_k^{(1)} y_{k-1} - a_k^{(2)} y_{k-2} - \ldots - a_k^{(n)} y_{k-n} + v_k \tag{5}$$

In (5) the unknown parameters $a_k^{(1)}, \ldots, a_k^{(n)}$ are referred to as the $AR(n)$ coefficients, which in case they are constant, with sufficient y_k measurements, can be found by solving a set of linear equations. Due to random errors in y_k though, it is more realistic to consider the coefficients $a_k^{(i)}$, $i = 1, \ldots, n$, as being noisy. Thus, it is assumed that they are of the form $a_{k+1}^{(i)} = a_k^{(i)} + w_k^{(i)}$, where each $\{w_k^{(i)}\}$ is a zero mean, white, Gaussian random process, independent of $\{w_k^{(j)}\}$ for $i \neq j$, and also independent of $\{v_k\}$. Defining now all these unknown noisy $AR(n)$ coefficients as an n-dimensional state vector $x_k = [x_k^{(1)} \quad x_k^{(2)} \quad \ldots \quad x_k^{(n)}]^T @ [a_k^{(1)} \quad a_k^{(2)} \quad \ldots \quad a_k^{(n)}]^T$ and defining an n-dimensional, white, zero mean, Gaussian process $\{w_k\}$ as the vector process formed from the $\{w_k^{(i)}\}$, we get the following system state equation:

$$x_{k+1} = x_k + w_k \tag{6}$$

Also, by defining the row vector of observations:

$$H_k^T = [-y_{k-1} \quad -y_{k-2} \quad \ldots \quad -y_{k-n}] \tag{7}$$

and the process $\{z_k\}$ by $z_k = y_k$, then Eq. (5) using (7) becomes the observed state measurements equation:

$$z_k = y_k = H_k^T x_k + v_k \tag{8}$$

Thus, with the above definitions, the $AR(n)$ model has been transformed into a linear, discrete-time, finite-dimensional noisy input state space Eq. (6) with Eq. (8) being the noisy output.

Remark 1: In this state space formulation of the $AR(n)$ model, it is not required for the time series data to be stationary. This is in contrast to the classical formulation of ARMA and ARIMA models, where it is necessary for the time series data to be transformed by differencing or by removing trend and seasonal components before

processing [15, 16]. Also, ARMA and ARIMA models need a lot of past (historical) data (usually more than 50) in order to produce predictions. A third reason for not considering the classical ARMA and ARIMA models is that the more than 50 past observations in no way can indicate the time series behavior in the next time interval. With these in mind, a reasonable value of n = 3 to 5 historical data values are sufficient.

The one-step prediction problem now is to produce an estimate at time $k+1$ of the system states $\hat{x}_{k+1/k+1}$ (which are the $AR(n)$ coefficients) using n noisy measured time-series data $z_{k-1}, z_{k-2}, ..., z_{k-n}$, and from (8) the predicted $z_{k+1} = \hat{y}_{k+1}$ can be calculated. In this case, the solution to the problem, is given by the discrete-time Kalman filter recursive equations [13, 14] as follows:

(a) Time update (or prediction equations):

$$\hat{x}_{k+1/k} = \hat{x}_{k/k} \tag{9}$$

$$P_{k+1/k} = P_{k/k} + Q_k \tag{10}$$

(b) Measurement update (or correction) equations:

$$K_{k+1} = P_{k+1/k} H_{k+1} \left[H_{k+1}^T P_{k+1/k} H_{k+1} + R_{k+1} \right]^{-1} \tag{11}$$

$$\hat{x}_{k+1/k+1} = \hat{x}_{k+1/k} + K_{k+1} \left[z_{k+1} - H_{k+1}^T \hat{x}_{k+1/k} \right] \tag{12}$$

$$P_{k+1/k+1} = \left(I^{nxn} - K_{k+1} H_{k+1}^T \right) P_{k+1/k} \left(I^{nxn} - K_{k+1} H_{k+1}^T \right)^T + K_{k+1} R_{k+1} K_{k+1}^T \tag{13}$$

with the matrices being $R_k = E\left[v_k^2 \right]$, $Q_k = E\left[w_k w_k^T \right]$, and I^{nxn} is the nxn identity matrix (all 1's in the main diagonal and zeros elsewhere).

Equation (12) is initialized with $\hat{x}_{1/0}$ set equal to the vector of a priori estimates of the coefficients. The term $H_{k+1}^T \hat{x}_{k+1/k}$ is one step predicted output $\hat{z}_k$ and the quantity $\left[z_{k+1} - H_{k+1}^T \hat{x}_{k+1/k} \right]$ is one step prediction sequence, usually called innovation or residual, $r_{k+1} = \left[z_{k+1} - H_{k+1}^T \hat{x}_{k+1/k} \right]$.

Eq. (13) is initialized with $P_{1/0}$ set equal to the a priori covariance matrix of the error in the estimate of these coefficients.

The matrix K_{k+1} is called the Kalman filter gain. Notice that the gain matrix K_{k+1} depends inversely on R_{k+1} - the larger the variance of the measurement error, the lower the weight is given to the measurement in making the forecast for the next period, given the current information set.

The matrix Q_k describes the confidence in the system state equation (3) or (6); an increase in this matrix means that we trust less the process model and more the measurements. Q_k can be estimated using the Maximum Likelihood Estimation method [13, 14], but often is picked by simulations to be $Q_k = \gamma I^{nxn}$, with $\gamma > 0$ being a scalar.

The error covariance matrix $P_{k+1/k+1}$ in (13) depends on the measurements via K_{k+1}. As the k measurements are processed, it is desirable for the covariance to be $P_{k+1/k+1} \le \rho_k I^{nxn}$, where the scalar $\rho_k > 0$ approaches

zero or a small number ρ, as $k \to \infty$. Then, for almost all the measurements the mean square parameter estimation error will approach zero, or some small quantity [13].

3 Forecasting high frequency (intravalues) time-series data using Mean Reverting Stochastic Process and Kalman filter daily closing predictions

3.1 Discrete time Mean Reverting Stochastic Process

Since the time-series data is available at discrete time intervals, we need to discretize the mean reverting stochastic process (1). For the discretization, we consider the interval $[0,\ T]$. The required step in the interval dt, can then be $dt = T/N$, for some integer N. Now, for $i = 1, 2, ..., N$, with B_i denoting B_{t_i}, where $t_i = idt$, from Property 2 (see Subsection 2.1), the Brownian motion gives $B_0 = 0$ with probability 1, and from Properties 2 and 3, we get $B_i = B_{i-1} + dB_i$, where each dB_i is an independent random variable of the form $\sqrt{qdt}N(0,1)$, where $N(0,1)$ is the normal distribution with mean zero and variance 1.

Now, in order to solve numerically Eq. (1) in the interval $[0,\ T]$, we apply the method of Euler–Maruyama [17], and obtain an approximation for the proposed mean reverting stochastic process S_{t_i}, denoted by S_i, for $i = 1, 2, ..., N$, as follows:

$$\begin{aligned} S_{i+1} &= A(t_i)\left[\mu(T) - S_i\right]dt + G\left(t_i, S_i\right)dB_i; \\ dB_i &= \sqrt{qdt}N(0,1),\ S_{i0} = initial\ condition \end{aligned} \tag{14}$$

The behavior of the discretized mean reverting stochastic process S_i remains the same as for the continuous case S_t, described in Subsection 2.1. That is, once the speed of the reversion parameter $A(t_i) > 0$, and the random or nonrandom discrete function $G\left(t_i, S_i\right) > 0$ have been chosen, the process would evolve randomly to reach the specified long term mean $\mu(T)$.

We can allow now the long term mean $\mu(T)$ to be $\mu(T) = \hat{y}_k = \hat{z}_k = H_k^T \hat{x}_k$ (8), with H_k^T known from the measurements and $\hat{x}_k$ as being estimated by the discrete Kalman filter (Eqs. (9)-(13)). Choosing, for example, $dt = 1\ \text{minute}$, then within a daily interval of $T = 6.5\ \text{hours}$, the solution of (15) would provide $N = 390$ estimated intravalues. For the random evolution (or trajectory) of these intravalues though, we have to specify the volatility function $G\left(t_i, S_i\right) > 0$ at each instant of time $i = 1, 2, ..., N = 390$.

3.2 Volatility choices for the mean reverting stochastic process

The choices for the volatility function, as described in Section 2.1, are deterministic and/or stochastic and vary according to different behaviours one might select for the process to evolve throughout the interval instances $i = 1, 2, ..., N = 390$. That is, one might select a specific discrete determinist or a specific discrete stochastic or even both deterministic and stochastic discrete volatility function $G\left(t_i, S_i\right) > 0$, for all the time instances, or segment the time interval instances $i = 1, 2, ..., N = 390$ within the day and choose different volatility functions for each segment.

A few possible choices for the volatility function $\sigma = G\left(t_i, S_i\right) > 0$ are as follows:

- With H_i, L_i, C_i, and O_i denoting the high, low, close, and open values of the time series, respectively, possible deterministic volatilities for the whole interval are:

  i) Average of high and low values of the previous day, or n previous days.

  ii) Previous day closing value, or average of n previous closing values.

  iii) Variance of n previous closing values.

  iv) Parkinson [1]: $$\sigma_{Parkinson} = \sqrt{\frac{1}{4n\ln(2)}\sum_{i=1}^{n}\left(H_i - L_i\right)^2}$$

  v) Garman Klass [2]: $$\sigma_{GK} = \sqrt{\frac{1}{2n}\sum_{i=1}^{n}\left(H_i - L_i\right)^2 - \left(2\ln(2)-1\right)\left(C_i - O_i\right)^2}$$

  vi) Rogers and Satchell [3]: $$\sigma_{RS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\left(H_i - C_i\right)\left(H_i - O_i\right) + \left(L_i - C_i\right)\left(L_i - O_i\right)\right]}$$

  vii) Yang-Zhang [18]: $\sigma = \sqrt{\sigma^2_{Close-to-Open} + k\sigma^2_{Open-to-Close} + (1-k)\sigma^2_{RS}}$ with $k = 0.34\left(1.34 + \frac{n+1}{n-1}\right)^{-1}$

- Another choice used in this work is to use the Kalman filter covariance matrix which provides information for the spread of error in the estimation of the $AR(n)$ coefficients for the one day ahead predicted measurement. That is, for the estimate $\hat{x}_{k+1/k}$ of the unobservable system state $x_{k+1/k}$, the error covariance is $P_{k+1/k} = E\left\{\left(x_{k+1/k} - \hat{x}_{k+1/k}\right)\left(x_{k+1/k} - \hat{x}_{k+1/k}\right)^T\right\}$. The rows of this error covariance matrix span the error space of the $AR(n)$ coefficients. Since this matrix is positive definite, different possibilities of volatility measures may include $\sqrt{\det\left[P_{k+1/k}\right]} > 0$, $\sqrt{eigenvalues\left[P_{k+1/k}\right]} > 0$, $\sqrt{trace\left[P_{k+1/k}\right]} > 0$, $\sqrt{norm\left[P_{k+1/k}\right]} > 0$, or combinations of these. As the order of the autoregressive model increases though, so are these measures of the Kalman filter covariance matrix (which all give constant volatility for the whole interval).

  The determinant of the Kalman filter covariance matrix, $\det\left[P_{k+1/k}\right] > 0$, can be thought as an estimate of volatility spanning the error space, since from linear algebra it is known that for a set of linearly independent vectors $u_1, u_2, \ldots, u_n$ in R^n, the absolute value of the determinant of the matrix M with rows $u_1, u_2, \ldots, u_n$ indicates the volume $V(\Pi) = \left|det(M)\right|$ of the parallelepiped $\Pi = \left\{a_1u_1 + a_2u_2 + \ldots + a_nu_n : 0 \le a_i \le 1\right\}$, $i = 1, 2, \ldots, n$ formed by these vectors. When $n = 2$, Π is a parallelogram and $V(\Pi)$ denotes the area of Π. In general, $V(\Pi) = 0$ if and only if the vectors $u_1, u_2, \ldots, u_n$ are linearly dependent (i.e., if and only if the vectors do not form a coordinate system in R^n).

  The trace of the Kalman filter covariance matrix, $trace\left[P_{k+1/k}\right] > 0$, gives another estimate of volatility, since the trace is the sum of the diagonal elements of a matrix, and for the error covariance the trace is the sum of the mean square errors, which is a performance index for the estimation.

Similar volatility measures are provided by the real part of eigenvalues and the norm of the Kalman filter covariance matrix since the eigenvalues and the norm are interrelated for a positive definite matrix.

- Another choice is to use the stochastic process $S_i > 0$ itself, at each instance $i = 1, 2, \ldots, N = 390$, and/or any of the above constants multiplied by the process $S_i > 0$ at each instance $i = 1, 2, \ldots, N = 390$ (gives stochastic volatility not constant but varying throughout the interval).
- Additionally, another choice is to use any of the above constant volatilities multiplied by a uniform distribution of some level at each instance $i = 1, 2, \ldots, N = 390$ (gives stochastic volatility varying throughout the interval).
- Yet another choice is to use any of the above volatilities multiplied by a binomial distribution at each instance $i = 1, 2, \ldots, N = 390$ to model possible instances of inactivities (gives stochastic volatility varying throughout the interval).

4 MRSP structure, parameters, simulations, and assessment of results

The general structure of the time series mean reverting stochastic process (MRSP) predictions and evolution of intravalues is presented in Fig. 1.

In Fig. 1 the bottom part represents the time series data values $\{z_t, t = 0, 1, 2, \ldots, \mathrm{N}\}$. The part above it represents the AR(n)-Kalman filter predicted values y_k based on a set of n previous time series data values $\{z_1, z_2, \ldots z_n\}$. The top part represents the mean reverting stochastic process (MRSP) intravalues S_i generated while the process evolves to reach the predicted values y_k.

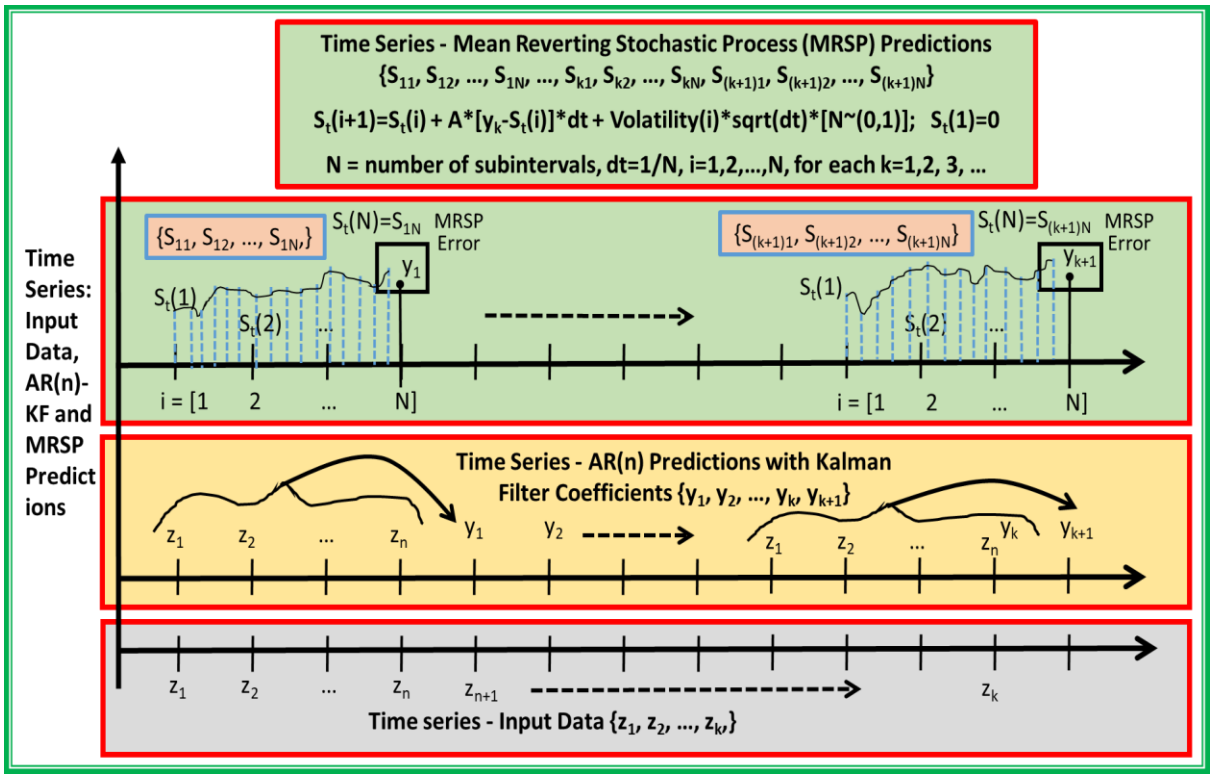


Fig. 1: Time Series Mean Reverting Stochastic Process (MRSP) Predictions and evolution of intravalues

The parameters of the mean reverting stochastic model chosen for simulation are as follows:

- The strength of the mean reversion is calculated adaptively as $A(i) = i \cdot dt \cdot \left|\mu(T) - S_i\right| > 0$ for $i = 1, 2, ..., N = 390$ with $A(1)$ being some small positive number, e.g., $A(1) = 0.001$. This adaptive calculation of $A(i)$ keeps track of the deviations of S_i as it evolves toward its end AR(n)-Kalman filter estimated value $\mu(T)$. As this deviation increases the $A(i)$ values increase as well to bring the process as close as possible to $\mu(T)$.

- The volatility term as previously stated can be both deterministic and stochastic as well as a function of the process itself. Here it is chosen as $G\left(k_{i+1/k}, S_i\right) = \sqrt{\left(tr\left[P_{k+1/k}\right] + \det\left[P_{k+1/k}\right]\right) S_i}$. This choice stems from the Cox-Ingersoll-Ross model [19] which captures the mean reverting phenomenon and avoids the possibility of negative values for all values of $A(t) > 0$ and $\mu(T) > 0$ once the condition $2A(t)\mu(T) > \left(tr\left[P_{k+1/k}\right] + \det\left[P_{k+1/k}\right]\right)$ is satisfied. The stochastic part of this MRSP has the standard deviation $\sqrt{\left(tr\left[P_{k+1/k}\right] + \det\left[P_{k+1/k}\right]\right)}$ and is proportional to $\sqrt{S_i}$. According to [19] this is significant because it states that as the short-rate increases, the standard deviation will decrease.

- The mean reverting term $\mu(T)$ is the AR(n)-Kalman filter estimate $\mu(T) = \hat{y}_k = \hat{z}_k = H_k^T \hat{x}_k$. The daily time interval is $T = 6.5\ hours$ (Dow Jones Industrial Average (DJIA) trading hours per day), which gives the per minute intraday time instances $i = 1, 2, ..., N = 390$. The order of the autoregressive model is chosen as $AR(3)$, indicating regression over the past three ($n = 3$) daily closing values.

- The initial conditions, S_1, of the mean reverting stochastic model is chosen to be the current daily Open value.

- The Kalman filter initial covariance matrix is picked arbitrarily as $P_{1/0} = 1.5 * n^{(-n)} * I^{nxn}$.

- The covariance matrix Q_k describes the confidence in the system state equation (2.3); an increase in this matrix means that we trust less the process model and more the measurements. The traditional adaptation method proposed [in 20] is used to adaptively calculate Q_k using the residuals and the Kalman filter gain, $Q_k = K_{k+1}\left(\frac{1}{n} r_{k+1} r^T_{k+1}\right) K^T_{k+1}$.

- The measurement covariance R_k is set equal to the variance of the n observations vector $H_k^T = \left[-y_{k-1} \quad -y_{k-2} \quad \text{L} \quad -y_{k-n}\right]$.

- The initial condition for the state is chosen $\hat{x}_{1/0} = -0.35 * \left(\left[1 \quad 1 \quad \text{L} \quad 1\right]^{nx1}\right)^T$.

For the simulation of mean reverting stochastic model the DJIA daily data is used. Specifically, a sample of 21 trading days from 4 April, 2017 until 4 May, 2017.

Fig. 2 presents this time series data sample with the corresponding High, Low, and Close values. This sample is considered to be adequate for our purposes as it possesses enough variability in the values as shown.

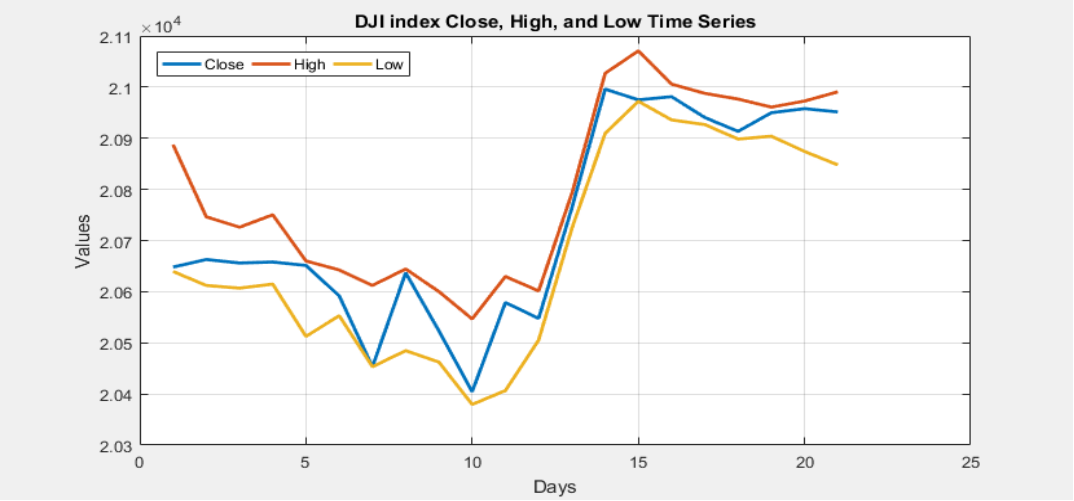


Fig. 2: DJIA index values for Close, High, and Low from 4 April, 2017 until 4 May, 2017

The simulation results are presented and described in the following figures:

Fig.3 presents the Root-Mean-Square (RMS) errors between the $AR(3)$-Kalman filter predictions and the corresponding daily close values (top), between the $AR(3)$-Kalman and the mean reverting stochastic process (MRSP) predictions (middle),and between the MRSP predictions and the corresponding daily close values (bottom). It is observed that these RMS errors are relatively small, considering the data values which are in the order of 20 thousand. Mean Square Error (MSE) measures are scale dependent and address model performance, i.e., $MSE = \frac{1}{n}\sum_{k=1}^{n}(z_k - y_k)^2$. Thus, the MSE can only provide a relative comparison between different models. These measures are zero if and only if the values are identical.

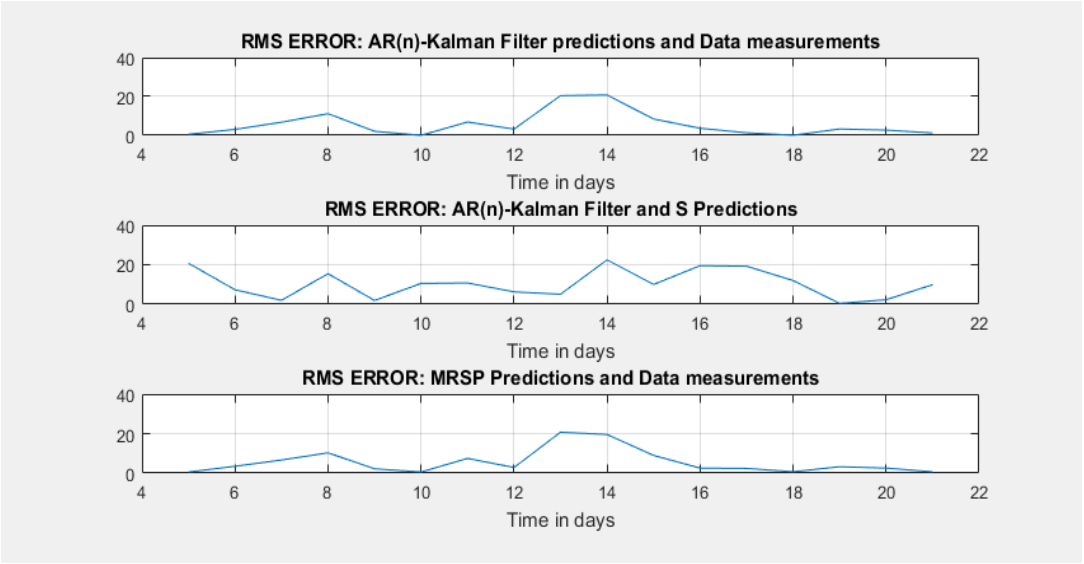


Fig. 3: RMS errors between: AR(3)-Kalman filter predictions and DJIA Close values (top), AR(3)-Kalman filter and MRSP predictions (middle), and MRSP predictions and data measurements (bottom)

Fig.4 presents the Mean-Absolute-Percentage-Error (MAPE) errors between the $AR(3)$-Kalman filter predictions and the corresponding daily close values (top), between the $AR(3)$-Kalman and the mean reverting stochastic process (MRSP) predictions (middle), and between the MRSP predictions and the corresponding daily close values (bottom).

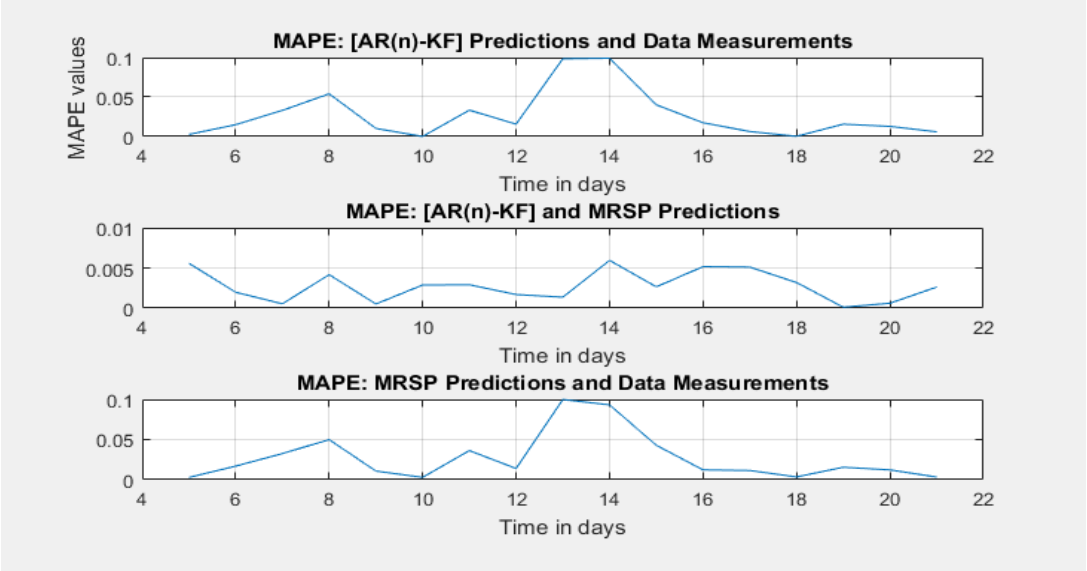


Fig. 4: MAPE errors between: AR(3)-Kalman filter predictions and DJIA Close values (top), AR(3)-Kalman filter and MRSP predictions (middle), and MRSP predictions and data measurements (bottom)

It is observed that these MAPE errors are extremely small, considering the data values which are in the order of 20 thousand. MAPE is the average value of the absolute values of errors expressed in percentage terms. We consider data to be in a relative scale if they are strictly positive and the importance of the difference is given by the ratio and not by the arithmetic itself, i.e., $MAPE = \frac{100}{n}\sum_{t=1}^{n}\frac{|z_k - y_k|}{|z_k|}$. The MAPE cannot be determined if the measured values are equal to zero and it tends to infinity if measurements are small or near to zero. This is a typical behavior, when relative errors are considered.

Next Fig. 5, Fig. 6, Fig. 7 and Fig. 8, present the MRSP intravalues in blue starting from the daily Open value indicated with the red square in the left-hand side of the plot. At the right-hand side of the plot the corresponding daily Close values are shown with a magenta square and the predicted values with the green square. The plots correspond to the DJIA trading days 28 April, 2017 until 3 May, 2017 (excluding the weekend), with corresponding Close values $2.0940x10^4$, $2.0913x10^4$, $2.0949x10^4$, and $2.0957x10^4$.

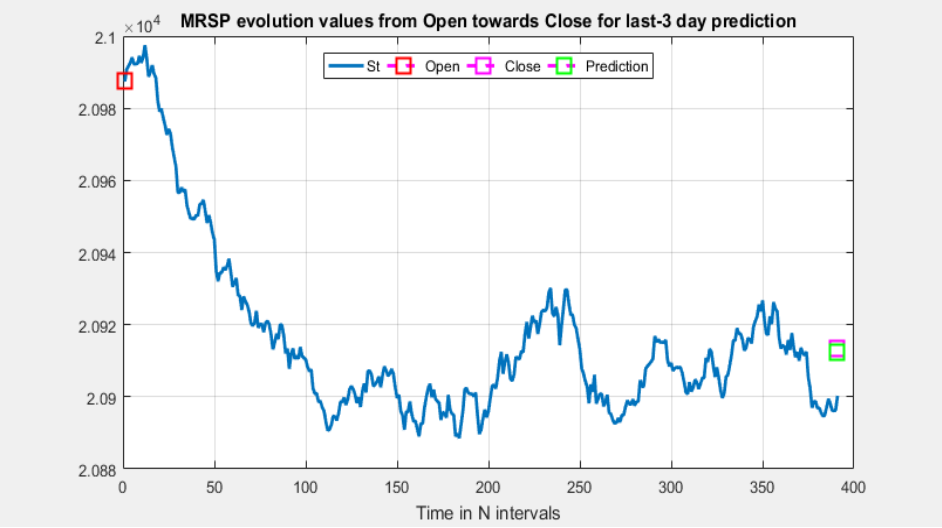


Fig. 5: MRSP Intravalues for day 28 April, 2017. Close value $2.0940x10^4$

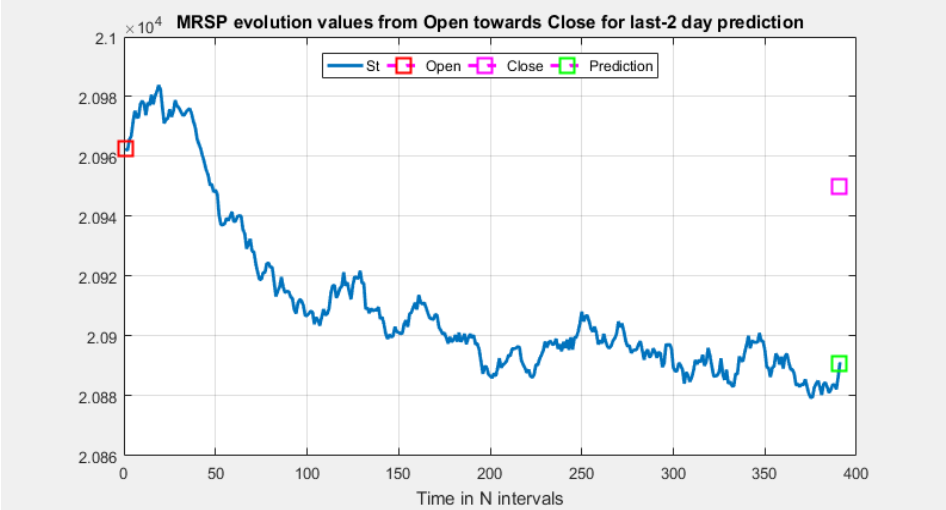


Fig. 6: MRSP Intravalues for day 1 May, 2017. Close value 2.0913x10^4

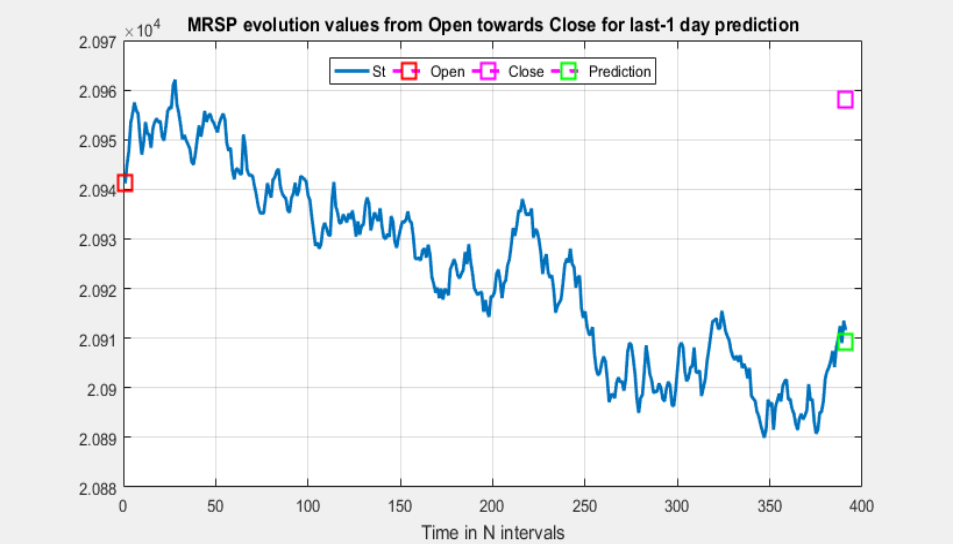


Fig. 7: MRSP Intravalues for day 2 May, 2017. Close value 2.0949x10^4

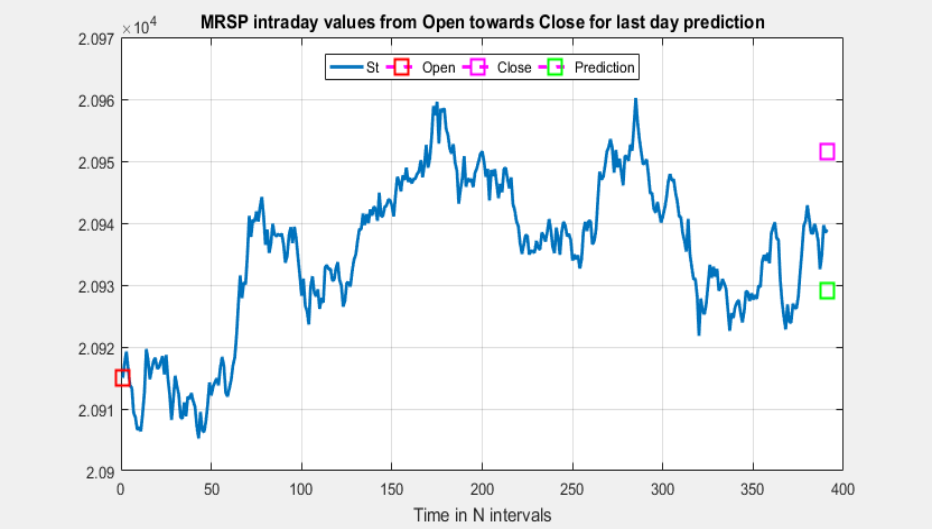


Fig. 8: MRSP Intravalues for day 3 May, 2017. Close value 2.0957x10^4

Fig.9 presents the MRSP intravalues for day 4 May, 2017 with Close value $2.0951x10^4$.

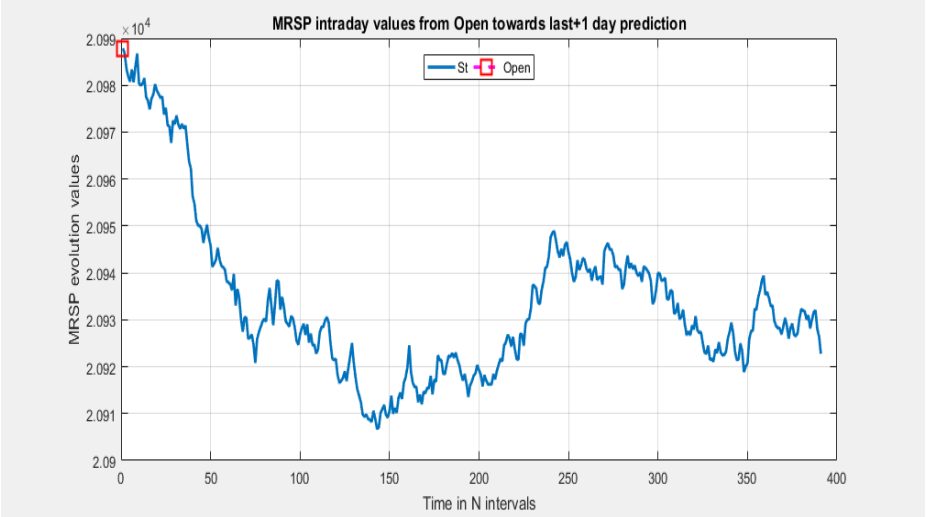


Fig. 9: MRSP Intravalues for day 4 May, 2017. Close value $2.0951x10^4$

Next Fig. 10 and Fig. 11 present the adaptive values of the MRSP strength coefficient A for the last day and the last+1 day, respectively, from the starting value 0.001 as the intravalues evolve.

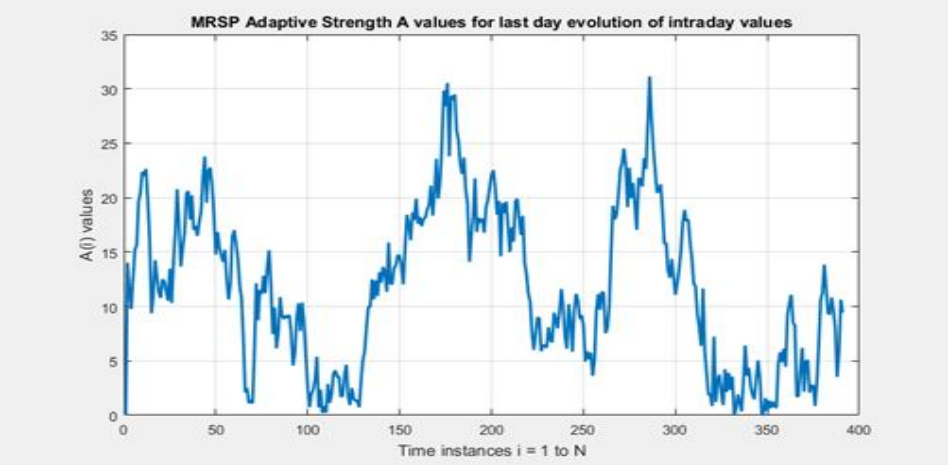


Fig.10: MRSP Adaptive Strength A values for last day as intravalues evolve

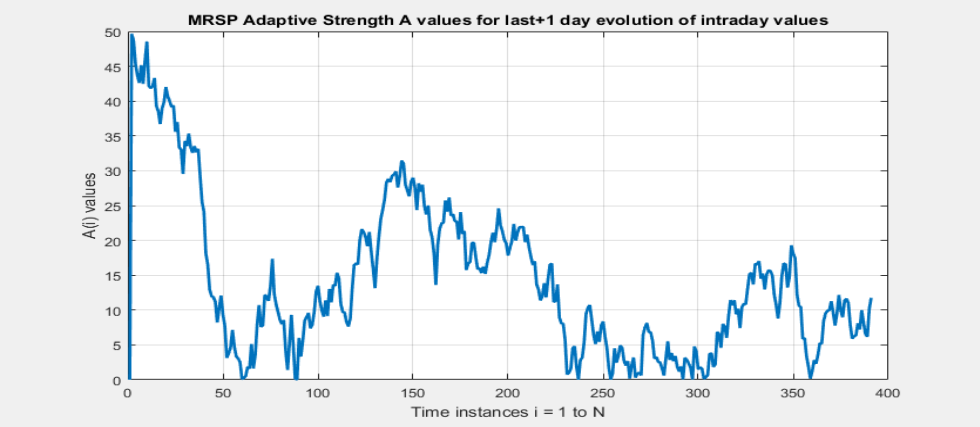


Fig. 11: MRSP Adaptive Strength A values for last+1 day as intravalues evolve

Next Fig. 12 and Fig.13 show the volatility values $G(t_i, S_i) = \sqrt{(tr[P_{k+1/k}] + \det[P_{k+1/k}])S_i}$ for last day and last+1 day, respectively, as MRSP intravalues evolve.

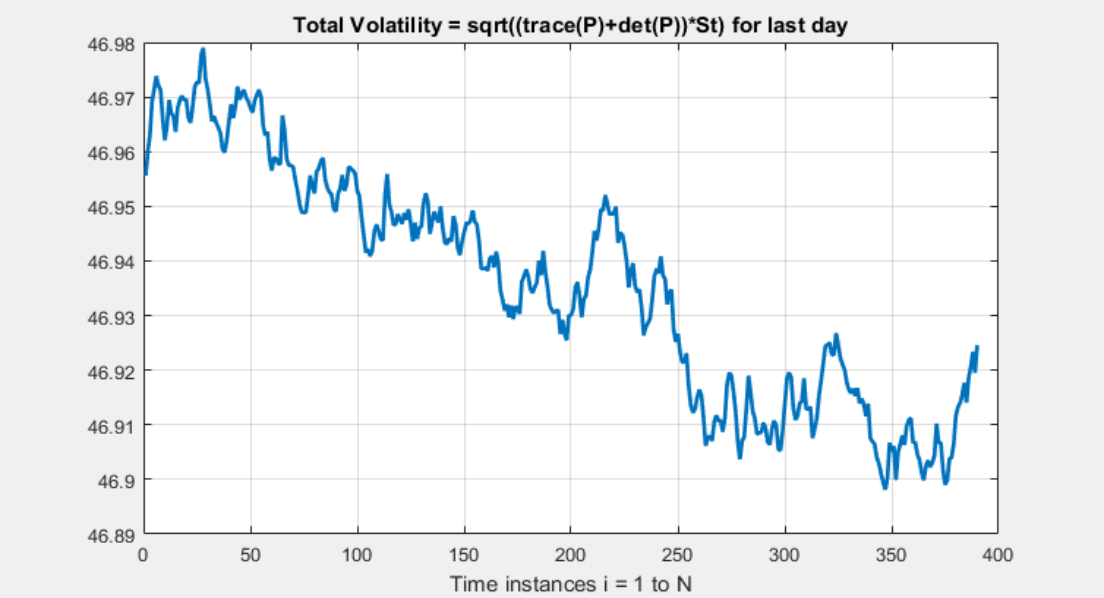


Fig. 12: MRSP Volatility values for last day as intravalues evolve

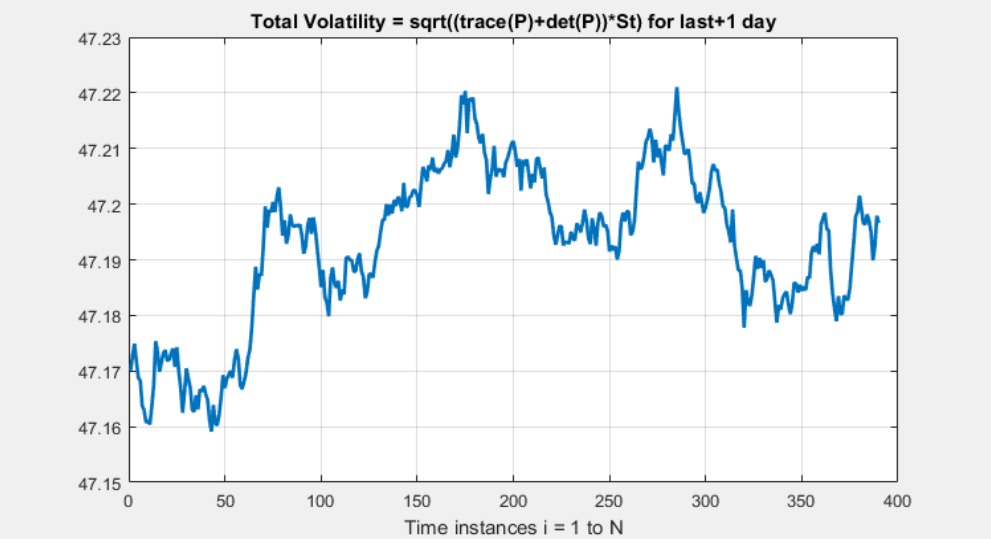


Fig. 13: MRSP Volatility values for last+1 day as intravalues evolve

Comments: From Fig. 5, Fig. 6, Fig. 7, and Fig. 8 it can be deduced that the intravalues produced by the proposed MRSP model starting from the Open values do not end far away from their corresponding actual Close values. Similar results were obtained for all the previous simulated DJIA data values (too many to be presented here). Also, from the same figures the following observations come out:

- When the Open value is high enough or low enough relative to the close value, the MRSP evolves quite fast in the beginning of the time interval in such a way to reach the estimated level of the Close value, and then from there, it "cruises" towards this Close value for the rest of the time interval (see Fig. 5, Fig. 6, and Fig. 7).

- On the other hand, when the Open and estimated Close values are relatively in the same level, then the MRSP wanders towards the Close value throughout the whole-time interval (see Fig. 8). This behavior of the MRSP model seems most often to be similar to the actual evolution of the time series during the day.

- In addition, since the MRSP model by construction ends up close to the daily Close value starting from an Open value, the evolution of the intravalues it produces, exhibits the U-shaped patterns [9] as well as the reverse J-shaped patterns [10] as reported in the literature (see Fig. 5, Fig. 6, Fig. 8 and Fig. 9).

Thus, based on these observations, the MRSP model can be used for trading purposes throughout the time interval from the actual Open and its estimated Close values, since it provides a stochastic evolution of intravalues of a time series, within of course, the chosen values for its parameters and the volatility measure implemented.

Remark 2: The sensitivity of the model depends on the "seed" value of the MATLAB random number generator routine rng(.). For the simulations here, using a few runs with the command rng('shuffle'), the seed value was chosen as rng(1111414785). The other parameter having an impact on the model's results is the Kalman filter initial error covariance $P_{1/0}$. This though did not have a large of an impact, since after a few iterations its value is regulated with the incoming covariance of the observations. Also, from simulations, the AR(n) model of order n=3 seemed to provide more acceptable results, since more past historical data does not necessarily indicate the time series behavior in the next time interval. The used in the simulations volatility term $G(t_i, S_i) = \sqrt{\left(tr\left[P_{k+1/k}\right] + \det\left[P_{k+1/k}\right]\right) S_i}$, consisting of the trace and the determinant of the Kalman filter covariance matrix multiplied by the process itself, for all time instants $i = 1, 2, ..., N = 390$, gave values which are not far apart from the corresponding values calculated with Parkinson's [1] volatility formula.

5 Conclusion and further research

In this paper a MRSP model has been presented for generating intravalues (or tick data) of a time series along with their evolution. The theoretical and practical aspects of the model have been described in details. The proposed MRSP uses a combination of four algorithmic techniques: (a) A deterministic part providing the directional evolution and a stochastic part giving the up-and-down fluctuations, (b) a volatility measure for the size of the fluctuations, (c) a state space formalized AR(n) model for estimating the final value of the process direction, and (d) a Kalman filter providing for the coefficients of the AR(n) model as well as contributing through the trace and/or determinant of its covariance matrix for the process volatility term.

Acceptable simulation results have been presented for the DJIA time series demonstrating the applicability of the model. That is, the proposed MRSP model starting from the daily Open values evolves within an acceptable error towards the predicted daily Close value, giving during the day high frequency (intravalues or every minute tick data) forecast values based on the chosen volatility measure for the time series and the values of its parameters.

Based on these results, some further research direction would be the examination of various advanced stochastic control theory techniques to adaptively determine the $A(i) > 0$ parameter for the evolution of the MRSP throughout the interval of interest.

A second direction is to examine the results of an additive Jump-Diffusion term to the mean reverting stochastic process to account for fat tails which are present in the tick data. Moreover, since the intravalues exhibit small and/or larger jumps, and for some instances do not change at all, it is realistic to consider any of the above deterministic and/or stochastic volatilities, but in addition, it is realistic to use an additive Jump-Diffusion term to the mean reverting stochastic process to account for fat tails. Then, such a model would be $dS_t = A(t)\left(\mu(t) - S_t\right)dt + G\left(t, S_t\right)dB_t + dJ_t$ [21, 22], where the jumps are defined as

$J(t)=\sum_{j=1}^{P(t)} Y_j$, $dJ(t)=Y_{P(t)}dP(t)$, with $P(t)$ being a Poisson process with intensity λ, and with Y_j being independent identically distributed (iid) random variables modeling the size of the $j\text{-}th$ jump, independent of P and B.

References:

[1] Parkinson, The Extreme Value Method for Estimating the Variance of the Rate of Return, Journal of Business,1980, vol. 53, no. 1, pp. 61-65.

[2] Garman and Klass, On the estimation of security price volatilities from historical data, Journal of Business,1980, vol. 53, pp. 67-78.

[3] Rogers and Satchell, Estimating variance from high, low and closing prices, Annals of Applied Probability,1991, vol 1, pp. 504-512.

[4] Corsi, A Simple Approximate Long-Memory Model of Realized Volatility, Journal of Financial Econometrics,2009, Vol. 7, No. 2, pp. 174–196.

[5] Tsay, Analysis of financial time series, Financial Econometrics, A Wiley-Interscience Publication, John Wiley & Sons, Inc, published simultaneously in Canada, 2002.

[6] Wang, Yu, and Li, On Intraday Shanghai Stock Exchange Index, Journal of Data Science,2010, vol 8, pp. 413-427.

[7] Rossi and Fantazzini, Long memory and Periodicity in Intraday Volatility, Department of Economics and Management DEM Working Paper Series, 2012.

[8] Gencay and Selcuk, Intraday dynamics of stock market returns and volatility, ESEVIER, Physica A,2006, vol, 367, pp. 375–387, Available online, www.sciencedirect.com).

[9] Andersen, Bollerslev, and Cai, Intraday and interday volatility in the Japanese stock market, ESEVIER, Journal of International Financial Markets, Institutions and Money,2000, vol 10, pp. 107–130.

[10] Lunina and Dzhumurat, *The Intraday Dynamics of Stock Returns and Trading Activity: Evidence from OMXS 30*, Master Essay II, Lund University, School of Economics and Management, June 2011, Available Online, http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=1973850&fileOId=1973852.

[11] Bougioukou P. Athina, Leros P. Apostolos, Papakonstantinou Vassilios, Modelling of non-stationary ground motion using the mean reverting stochastic process, Applied Mathematical Modelling,2008, vol 32, pp. 1912–1932.

[12] Ludwing Arnold, Stochastic Differential Equations: Theory and Applications, John Wiley and Sons, 1973.

[13] Anderson and Moore, Optimal Filtering, (Information and system sciences series, Thomas Kailath Editor), Prentice-Hall, Inc., Englewood Cliffs, N.J., 1979.

[14] Maybeck S. Peter, Stochastic Models, Estimation and Control, Vol I, Academic Press. 1979.

[15] Brockwell and Davis, Introduction to Time Series and Forecasting,2nd Edition, Springer-Verlag, Inc., New York, 2002.

[16] Commandeur and Koopman, Practical Econometrics: An Introduction to State Space Time Series Analysis, Oxford University Press, New York, 2007.

[17] Higham D.J., An algorithmic introduction to numerical simulation of stochastic differential equations, *Siam* Rev, 2001, vol 43 No 3, pp525–546.

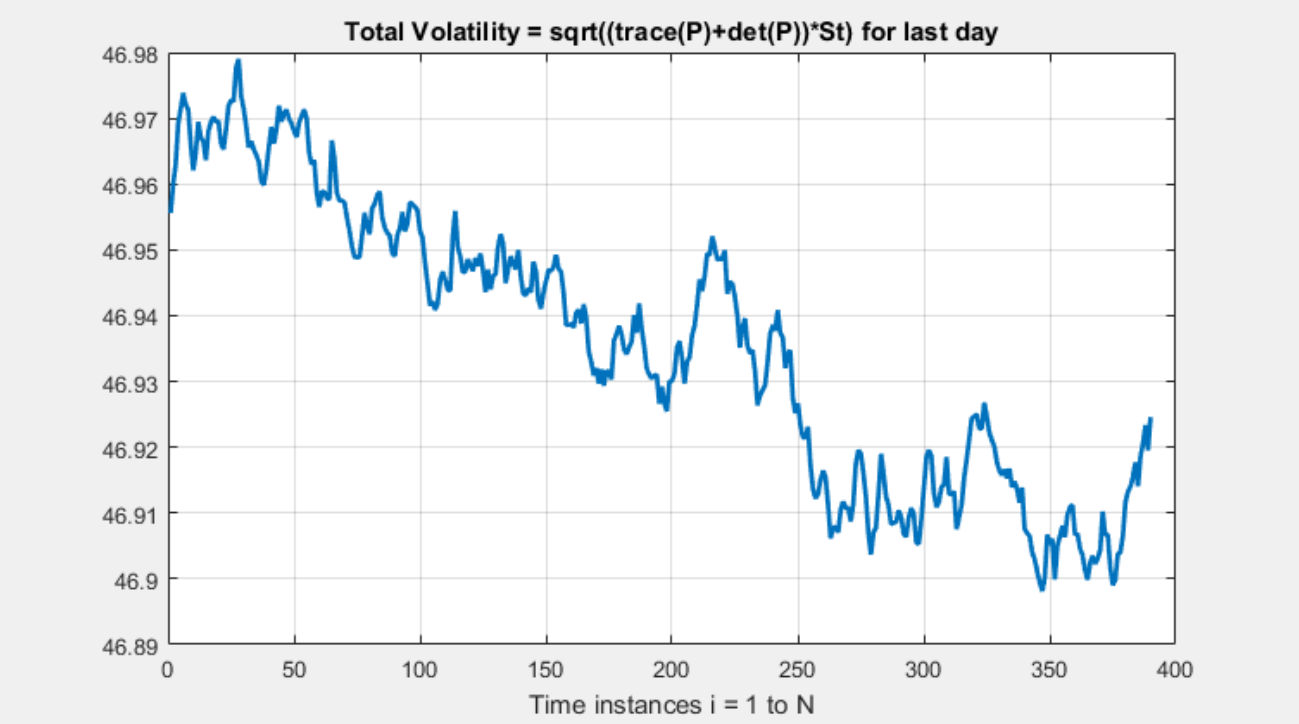
[18] Yang and Zhang, Drift-independent volatility estimation based on high, low, open, and closing Prices, *Journal of Business*, vol 73, 2000, pp. 477-491.

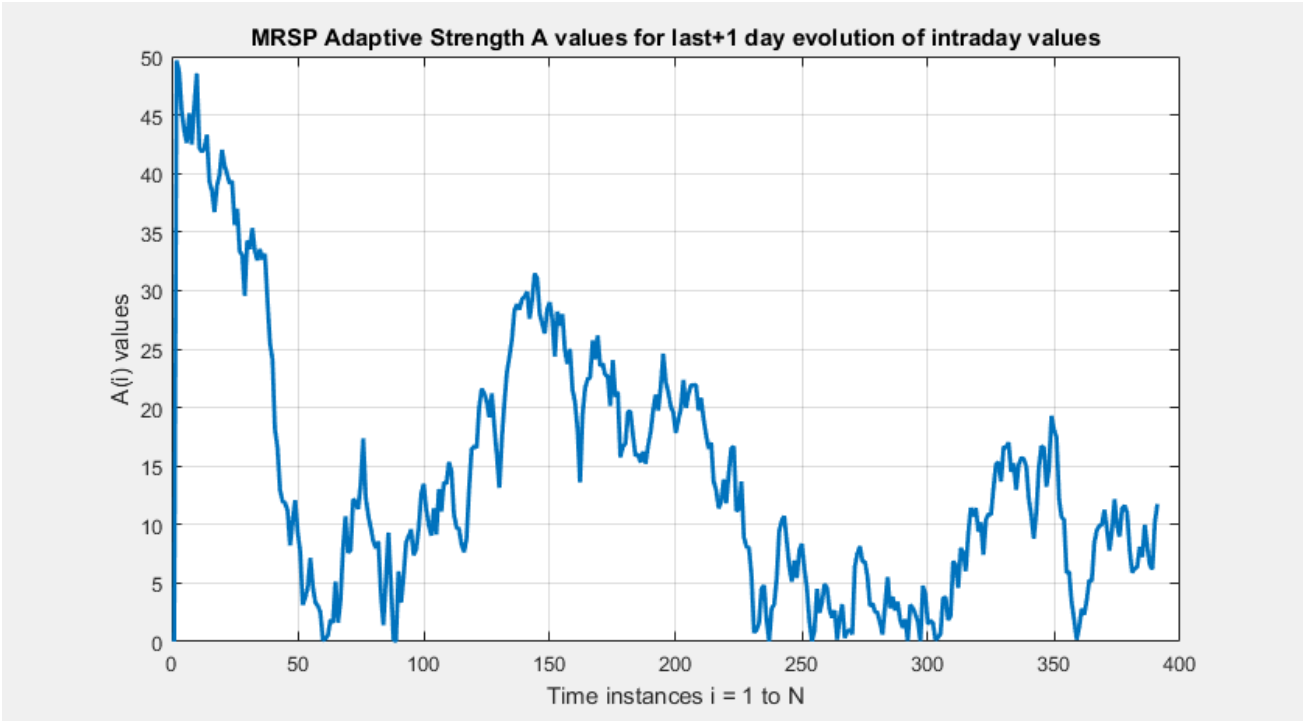
[19] Cox John C., Ingersoll Jonathan E., and Ross Stephen A., A Theory of the Term Structure of Interest Rates. Econometrica1985, vol 53, no. 2, pp.385-408.

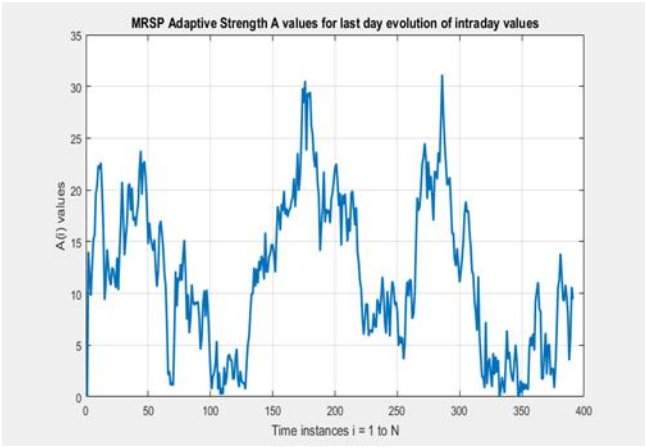
[20] Mohamed, A. H. and Shwarz, K. P. Adaptive Kalman filtering for INS/GPS, Journal of Geodesy, 1999, vol 73 No 4, pp. 193-203.

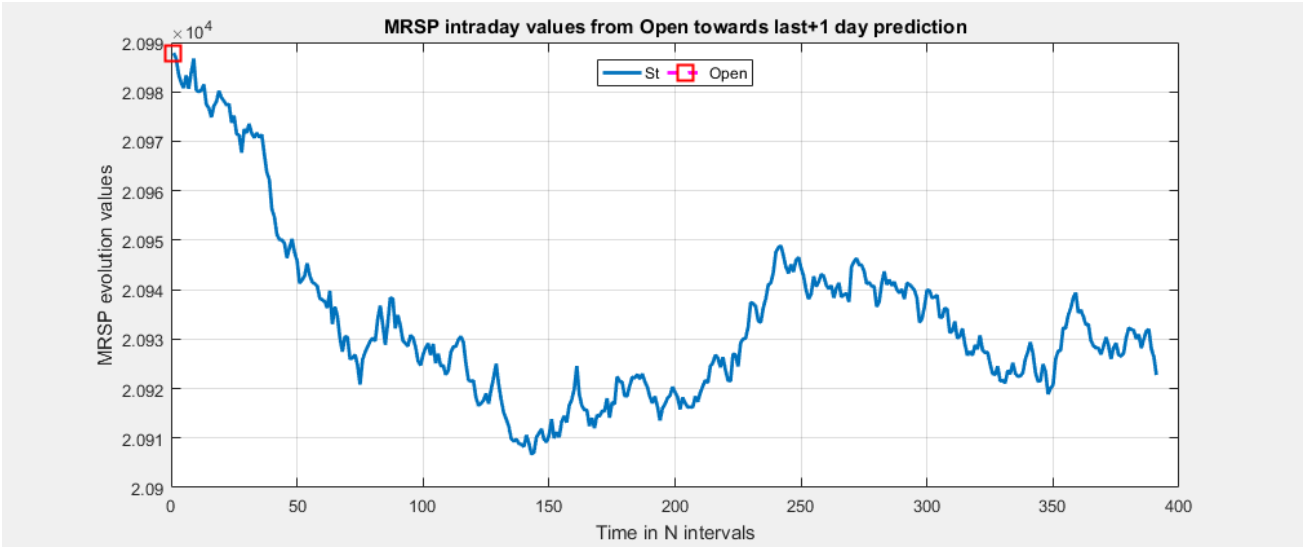
[21] Hanson, Applied Stochastic Processes and Control for Jump-Diffusions: Modeling, Analysis and Computation, Society for Industrial and Applied Mathematics (SIAM), 2007.

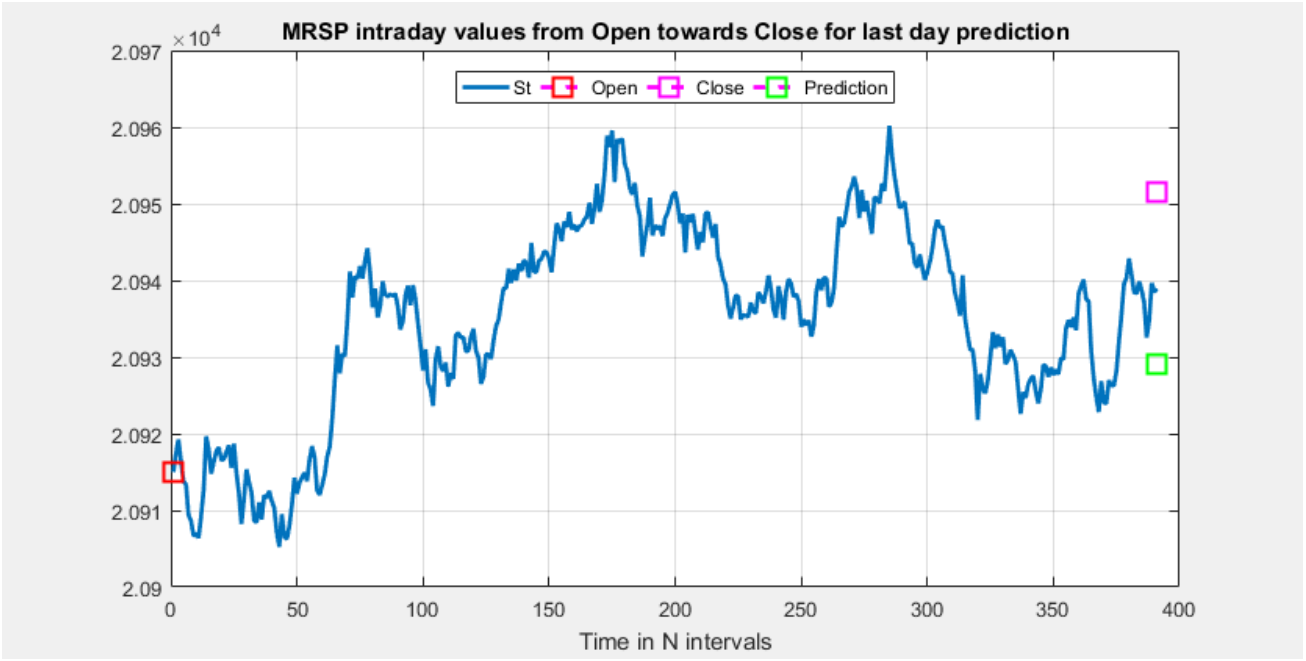
[22] Brigo D., Dalessandro A., Neugebauer M., Triki F., A Stochastic Processes Toolkit for Risk Management, 15 November 2007, Available Online, Accessed 20 May 2017, from the linkhttps://papers.ssrn.com/sol3/papers.cfm?abstract_id=1109160.

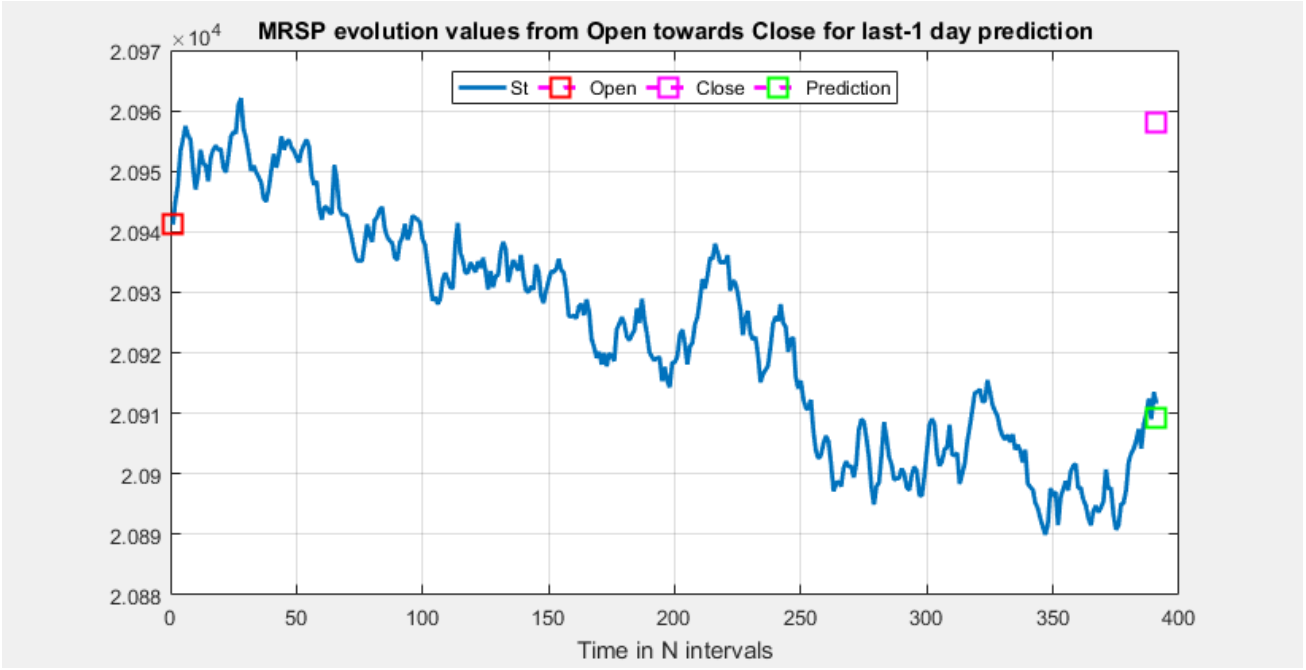


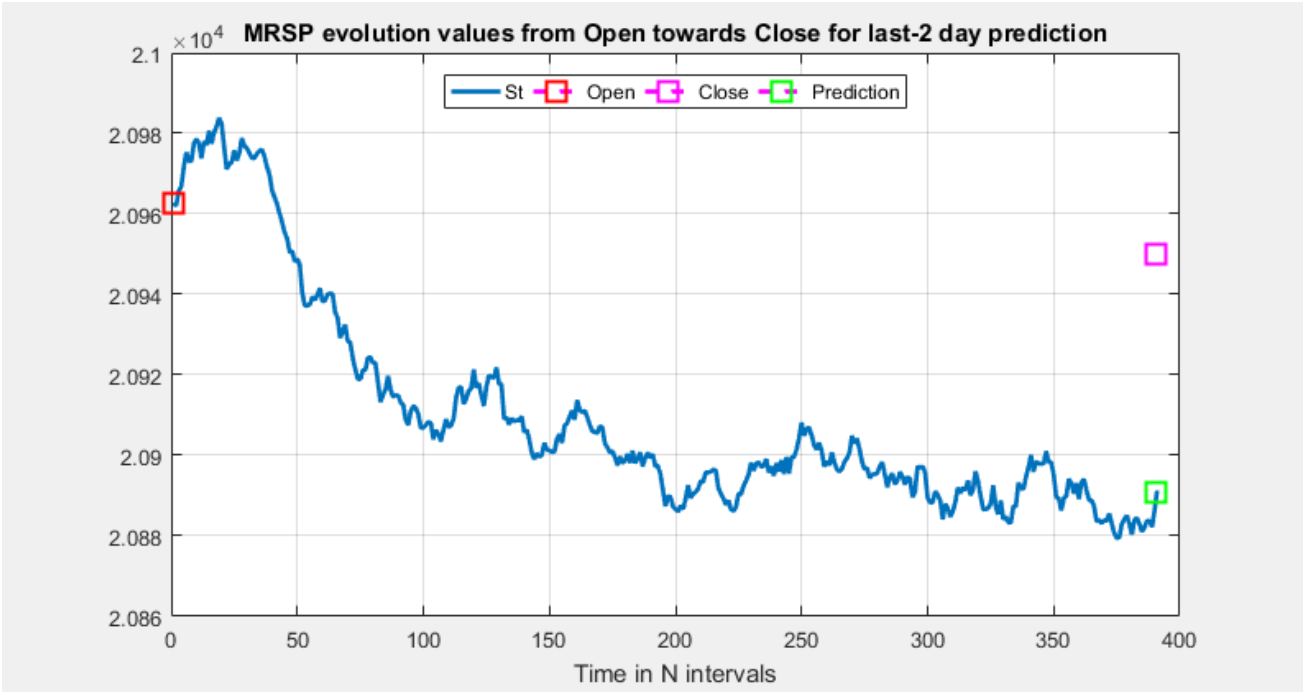

MRSP Adaptive Strength A values for last+1 day evolution of intraday values
A(i) values
Time instances i = 1 to N

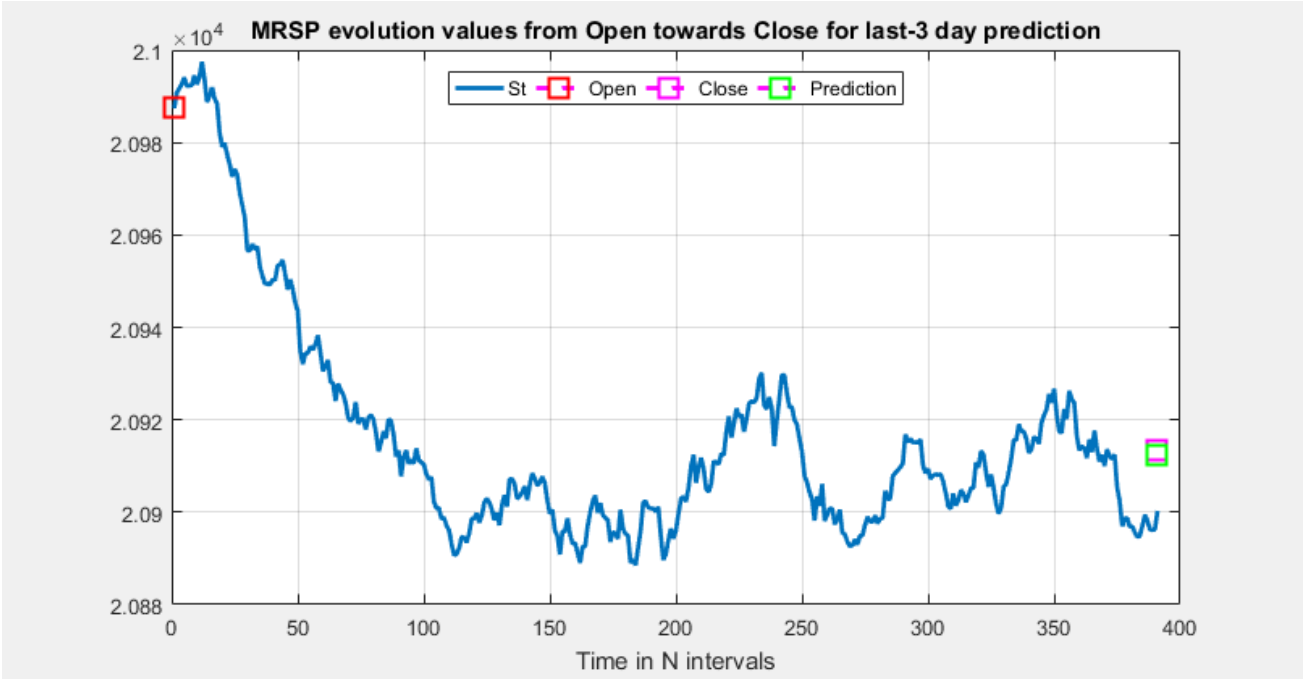

MRSP Adaptive Strength A values for last day evolution of intraday values
A(i) values
Time instances i = 1 to N

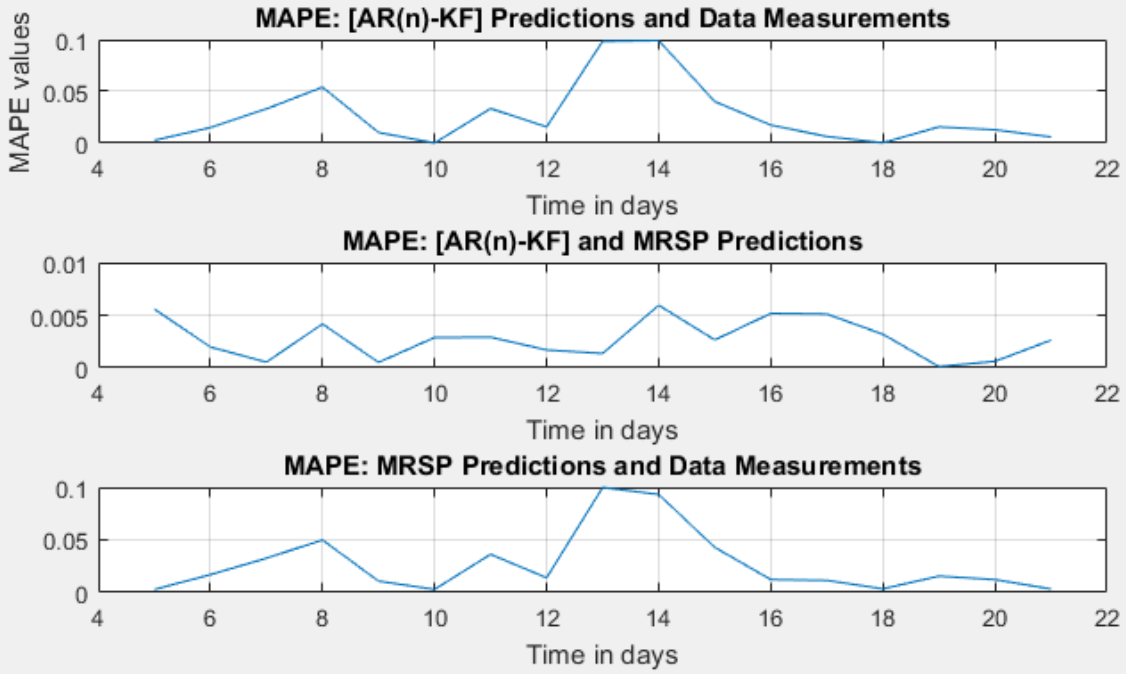

MRSP intraday values from Open towards last+1 day prediction
St
Open
MRSP evolution values
Time in N intervals

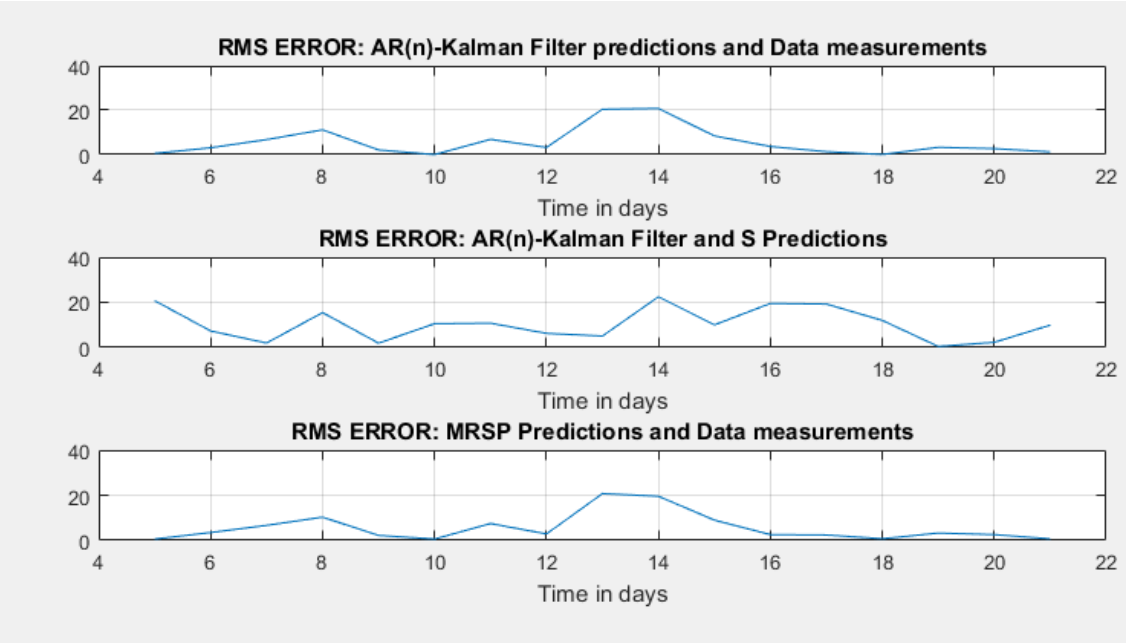

MRSP intraday values from Open towards Close for last day prediction
×10^4
2.097
2.096
2.095
2.094
2.093
2.092
2.091
2.09
St
Open
Close
Prediction
0
50
100
150
200
250
300
350
400
Time in N intervals

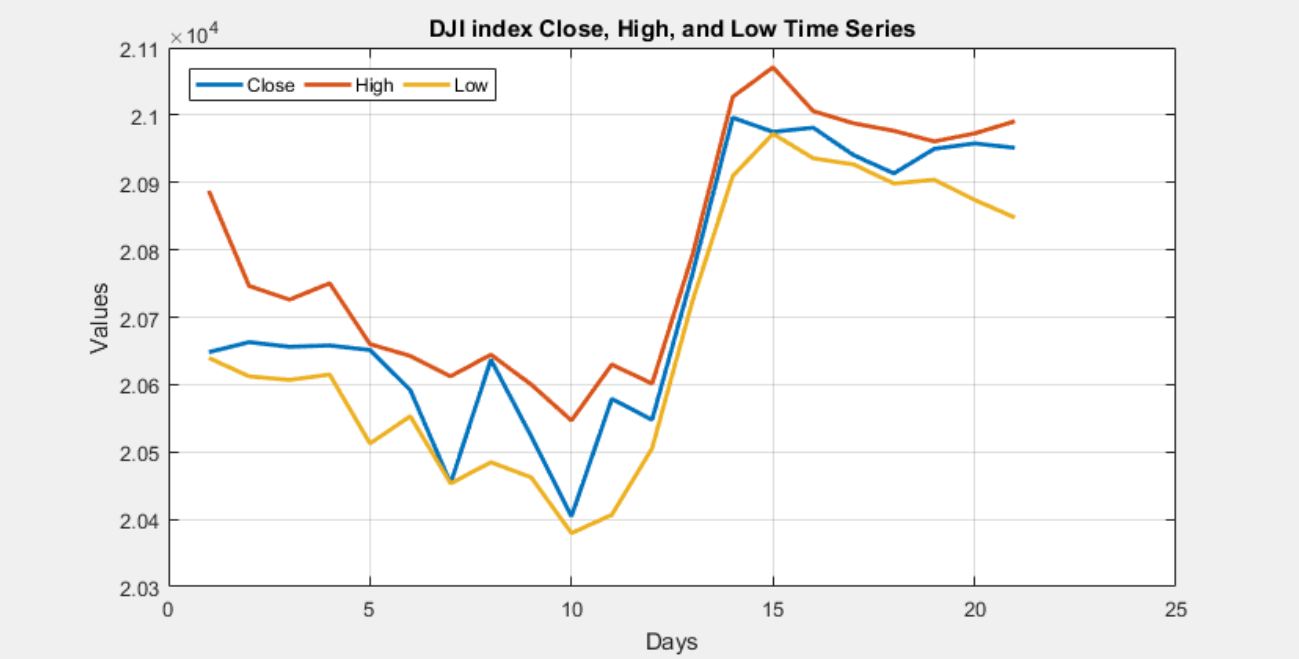

MRSP evolution values from Open towards Close for last-1 day prediction
×10^4
2.097
2.096
2.095
2.094
2.093
2.092
2.091
2.09
2.089
2.088
St
Open
Close
Prediction
0
50
100
150
200
250
300
350
400
Time in N intervals

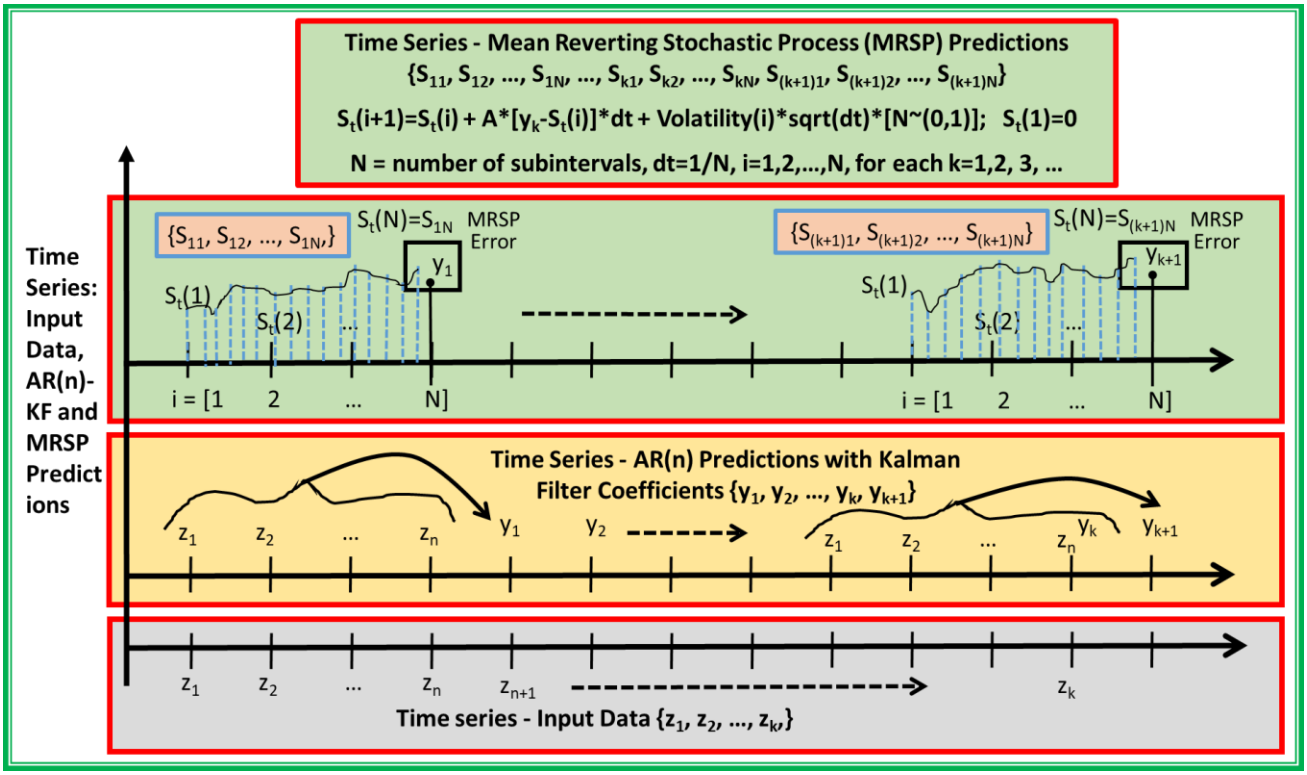

MRSP evolution values from Open towards Close for last-2 day prediction
×10^4
2.1
2.098
2.096
2.094
2.092
2.09
2.088
2.086
St
Open
Close
Prediction
0
50
100
150
200
250
300
350
400
Time in N intervals

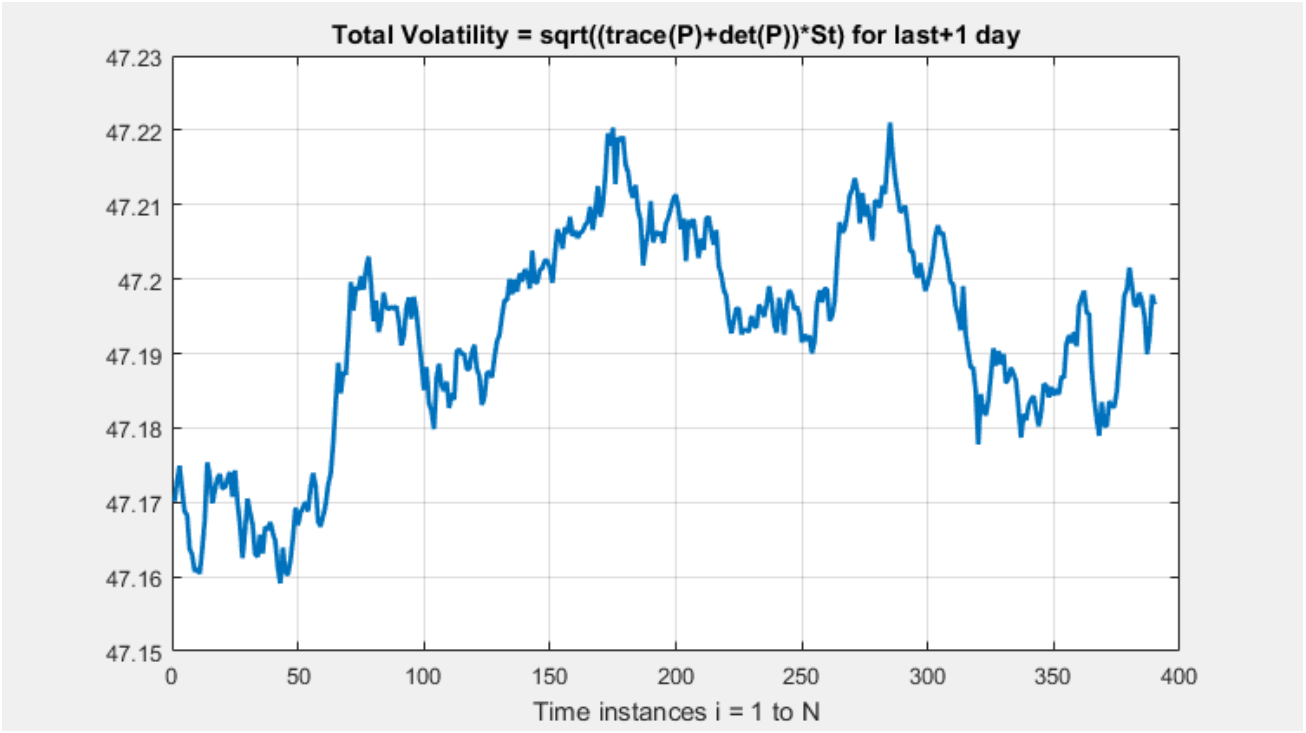

MRSP evolution values from Open towards Close for last-3 day prediction
×10^4
2.1
2.098
2.096
2.094
2.092
2.09
2.088
St
Open
Close
Prediction
0
50
100
150
200
250
300
350
400
Time in N intervals


MAPE: [AR(n)-KF] Predictions and Data Measurements
MAPE values
Time in days
MAPE: [AR(n)-KF] and MRSP Predictions
Time in days
MAPE: MRSP Predictions and Data Measurements
Time in days


RMS ERROR: AR(n)-Kalman Filter predictions and Data measurements
Time in days
RMS ERROR: AR(n)-Kalman Filter and S Predictions
Time in days
RMS ERROR: MRSP Predictions and Data measurements
Time in days


DJI index Close, High, and Low Time Series
×10^4
Close
High
Low
Values
Days
2.11
2.1
2.09
2.08
2.07
2.06
2.05
2.04
2.03
0
5
10
15
20
25


Time Series - Mean Reverting Stochastic Process (MRSP) Predictions
{S11, S12, ..., S1N, ..., Sk1, Sk2, ..., SkN, S(k+1)1, S(k+1)2, ..., S(k+1)N}
St(i+1)=St(i) + A*[yk-St(i)]*dt + Volatility(i)*sqrt(dt)*[N~(0,1)]; St(1)=0
N = number of subintervals, dt=1/N, i=1,2,...,N, for each k=1,2, 3, ...
Time Series: Input Data, AR(n)-KF and MRSP Predictions
{S11, S12, ..., S1N,}
St(N)=S1N
MRSP Error
y1
St(1)
St(2)
i = [1 2 ... N]
{S(k+1)1, S(k+1)2, ..., S(k+1)N}
St(N)=S(k+1)N
MRSP Error
yk+1
Time Series - AR(n) Predictions with Kalman Filter Coefficients {y1, y2, ..., yk, yk+1}
z1 z2 ... zn
Time series - Input Data {z1, z2, ..., zk,}
zn+1
zk


Total Volatility = sqrt((trace(P)+det(P))*St) for last+1 day
47.23
47.22
47.21
47.2
47.19
47.18
47.17
47.16
47.15
0
50
100
150
200
250
300
350
400
Time instances i = 1 to N