International Journal of Forecasting 35 (2019) 1250–1262

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting and trading on the VIX futures market: A neural network approach based on open to close returns and coincident indicators

Luca Vincenzo Ballestra, Andrea Guizzardi [*], Fabio Palladini

Department of Statistical Science, University of Bologna, Italy

ARTICLE INFO

Keywords:
VIX
VIX futures
Forecasting
Coincident indicators
Trading strategies
Weak efficiency

ABSTRACT

Previous work has highlighted the difficulty of obtaining accurate and economically significant predictions of VIX futures prices. We show that both low prediction errors and a significant amount of profitability can be obtained by using a neural network model to predict VIX futures returns. In particular, we focus on open-to-close returns (OTCRs) and consider intraday trading strategies, taking into account non-lagged exogenous variables that closely reflect the information possessed by traders at the time when they decide to invest. The neural network model with only the most recent exogenous variables (namely, the return on the Indian BSESN index) is superior to an unconstrained specification with ten lagged and coincident regressors, which is actually a form of weak efficiency involving markets of different countries. Moreover, the neural network turns out to be more profitable than either a logistic specification or heterogeneous autoregressive models.

1. Introduction

The VIX index represents the market's estimate of the future volatility of the S&P 500 over the next thirty days. It provides a benchmark for the short-term expected volatility, as futures and options contracts can be inscribed on (see Whaley, 2008). In fact, the implied volatility reflects the market makers' point of view about the expected volatility of the futures' underlying assets. Thus, since market makers are often among the most informed agents, the implied volatility should outperform the historical one in forecasting the realized volatility of the futures' underlying assets (Shu & Zhang, 2012).

Despite the importance of VIX and its common use as a volatility measure, very little attention has been paid to the problem of forecasting it. In particular, the few works on the subject have shown that the VIX is predictable to some extent. However, this finding, while of theoretical interest, is not necessarily helpful for traders, because VIX can be traded only as derivative contracts, the dynamics of which do not always follow those of the VIX index. For example, Asensio (2013), Degiannakis (2008), Konstantinidi and Skiadopoulos (2011) and Konstantinidi, Skiadopoulos, and Tzagkaraki (2008), who are among the few authors who have focused on VIX futures (henceforth referred to as VXFs), highlight only weak evidence of statistical predictability and experience a low level of profitability when implementing trading strategies based on VIX forecasts. The overall picture is not encouraging for investors: on the one hand, there is evidence that VIX is predictable; on the other, it seems to be very hard to trade VXFs by learning from the (predicted) VIX dynamics.

The present paper fills this "forecasting gap" by presenting a new approach for modelling VXF returns that provides a significant amount of predictability and allows us to build profitable trading strategies.

Specifically, we rely on a feed-forward neural network model that yields a very general form of non-linearity. Moreover, we consider exogenous variables that closely

[*] Corresponding author.
E-mail address: andrea.guizzardi@unibo.it (A. Guizzardi).

https://doi.org/10.1016/j.ijforecast.2019.03.022
0169-2070/© 2019 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

reflect the information possessed by traders at the time when they decide to invest. In particular, our information set includes non-lagged exogenous variables that are available only a few hours before the opening of the Chicago Board Options Exchange (CBOE).

Instead of forecasting VXF daily returns (DRs), we predict VXF open to close returns (OTCRs), which are free of spurious effects related to trading timing (Anderson, Eom, Hahn, & Park, 2012). No less importantly, considering intraday returns allows us to connect forecast performances easily to the profits earned by those investors who open and close a position on the same day, taking advantage of the fact that the stock volatility is substantially higher intraday than overnight (Muravyev & Ni, 2016).

The contributions of the present paper are fourfold. First, we show that accurate predictions of OTCRs on VFX can be obtained by using an appropriate modelling approach.

Second, we show that a neural network model with the most recent exogenous variable (namely the return on the BSESN index) as its only input variable is superior to an unconstrained model with ten lagged and coincident regressors. That is, VXF prices strongly reflect the most recent publicly-available information, which is actually a form of weak efficiency involving the markets of different countries.

Third, we compare the VXF OTCR predictions provided by the neural network model with those yielded by a logistic specification, a simple (Naïve) model that always forecasts negative VXF OTCRs, a heterogeneous autoregressive (HAR) model, and two augmented heterogeneous autoregressive (HAR_X) models. The results obtained reveal that the neural network outperforms all of the other models significantly as far as mean directional accuracy is concerned.

Fourth, we simulate and test various trading strategies, with different abilities to filter out false signals. Again, the predictions of the neural network model turn out to be more profitable than those of its rival models.

The remainder of the paper is organized as follows. Section 2 describes the main issues related to predicability and profitability in the VIX/VXF market. Section 3 presents the model specifications, as well as the measures of forecast accuracy and profitability. Section 4 shows and discusses the main estimation results, focussing on the comparative assessment of the models' predictions and on the profitability of the corresponding trading strategies (considering both VXF OTCRs and VFX DRs). Finally, Section 5 concludes.

2. Main issues related to predictability and profitability in the VIX/VXF market

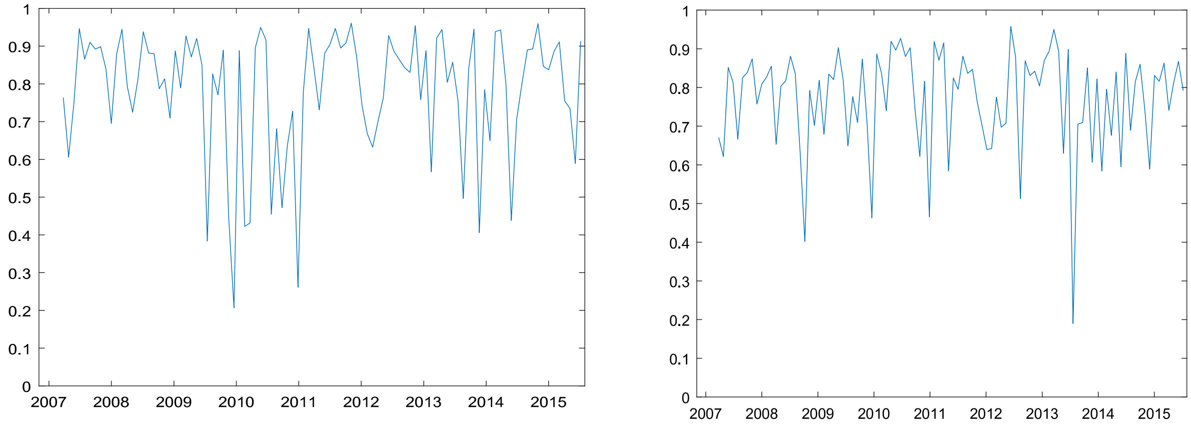
The research on VIX has been dominated largely by autoregressive conditional heteroscedasticity models that take into account non-linearity, long memory features and/or lagged exogenous variables. Ahoniemi (2006) tests and compares the predictive capabilities of probit, ARIMAX-GARCH and ARFIMA models. By considering a large set of U.S. financial and macroeconomic variables, she finds that the addition of exogenous regressors enhances the forecasting performance, whereas the improvements from adding GARCH errors or long-memory features are negligible. Degiannakis (2008) introduces a threshold effect for modelling asymmetry, but obtains no incremental information for forecasting VIX. Konstantinidi et al. (2008) model several implied volatility indices, including VIX, in a multivariate VAR framework; however, it does not yield any significant improvement in forecasting. Long memory features are also exploited by Fernandes, Medeiros, and Scharth (2014), among others, who apply a heterogeneous autoregressive (HAR) model coupled with a neural network in order to capture non-linearities better. Nevertheless, they find little evidence of non-linearity, since the HAR model augmented with the neural network performs as well as the linear HAR model with no neural network. Psaradellis and Sermpinis (2016) analyze three volatility indices including VIX, and find significant evidence of strong non-linearities by employing support vector regression models coupled with a genetic algorithm.

Other approaches look at both VIX and VXF prices, with the aim of studying the causality direction, and/or of investigating the forecast accuracy and the profitability of trading the VXFs. Shu and Zhang (2012) find that VXF prices drive the spot VIX if a linear model is employed. However, after searching for non-linear relationships, spot and futures prices react simultaneously to the arrival of new information. Jabłecki, Kokoszczyński, Sakowski, Ślepaczuk, and Wójcik (2014) and Luo and Zhang (2012) show that the shape of the implied volatility term structure and the volatility risk premium help in forecasting VIX.

A very important point to make is that predictability of the VIX index does not necessarily imply that VXF prices can be predicted too. This is pointed out clearly by Degiannakis (2008), who observes that the log-returns of VIX and of its futures have opposite signs on 26% of trading days. He concludes that "an agent cannot utilize VIX predictions in creating abnormal returns in implied volatility futures markets", and highlights the need for future work that focuses on the predictability of VXFs directly. Similarly, Degiannakis (2008), Konstantinidi et al. (2008) and Konstantinidi and Skiadopoulos (2011), despite finding some evidence of predictability of the VIX index, do not succeed in using VIX forecasts to construct profitable trading strategies.

Asensio (2013) addresses the topic from a more theoretical point of view, and talks about a "VIX-VFX puzzle" in order to stress the complexity of the VIX/VXF market. In particular, he identifies a number of factors that cause VXFs to be "consistently overpriced relative to the subsequent moves in the underlying VIX index".

We have investigated the VIX-VXF puzzle by computing the correlation between the DRs on VIX and the DRs on the VXF over the time period from March 26, 2007, to December 20, 2016. For the VXF, we use the continuous time series provided by Thomson Reuters Datastream (Type 0), which contains the prices of either the nearest contract month futures or the second-nearest contract month futures (for more details about time series



(a) DRs

(b) OTCRs

Fig. 1. Correlations between VIX and VXF (computed on monthly sub-samples).

construction, see Thomson Reuters Datastream, 2010). We found that the correlation is 0.823 when considering the whole dataset (2,454 observations), but falls to 0.783 when taking the average of the correlations computed on monthly sub-samples (the chosen time interval contains 116 monthly sub-samples). Furthermore, Fig. 1(a) shows how the monthly correlations vary over time. As can be seen, the profile is quite erratic, with many spikes around 0.4 and a minimum that is close to 0.2.

If we consider the correlation between the VIX OTCRs and VXF OTCRs, the trading gap between VIX and VXF is still more serious: the correlation is 0.767 on the whole data set, and falls to 0.771 if we compute the average of the correlations on the 116 monthly sub-samples. Moreover, the correlation dynamics are still very erratic, as Fig. 1(b) shows.

Therefore, since the dynamics of the VXFs do not reflect closely those of the VIX, even very accurate predictions of the VIX may not allow an investor to earn significant profits by trading VIX futures, as has been highlighted in the literature.

We should acknowledge that Psaradellis and Sermpinis (2016) claim that VIX forecasts can yield a "noteworthy prospect" of achieving economically significant profits in the VXFs market. In particular, they use a non-linear long-memory model to predict the VIX and then go long (short) in the VXF when the forecasted value of the index is greater (smaller) than its current value. Nevertheless, even if Psaradellis and Sermpinis succeeded in constructing profitable trading strategies, investors who trade VXFs based on VIX forecasts could still fail to obtain significant profits due to the poor and erratic correlation between the VIX and its futures.

Jabłecki et al. (2014) fill the trading gap between VIX and VXF by proposing an original and interesting approach that takes into account both the current level of VIX and the volatility term structure. However, even should they succeed in building a profitable trading strategy, their predictions of VXF levels do not outperform naïve forecasts.

Some authors have attempted to overcome the non-tradability of the VIX index, as well as to avoid the use of VIX futures (see Ahoniemi, 2006; Degiannakis, Filis, & Hassani, 2018), by proposing trading sessions based on buying/selling straddles of options on the S&P 500. The predictions of the VIX drive the decision to buy or sell the straddle; in particular, a long (short) position is taken if the VIX is expected to rise (fall). However, such an approach is guaranteed to be profitable only if the strikes of the traded straddles coincide with the S&P (i.e., the straddles are delta-neutral), which cannot always be the case, due to the limited number of straddle strikes that are available on the market.

Finally, while some authors have already focused on the predictability of DRs on VXFs, no one has ever tested either the predictability of opening-to-close returns on VXFs or the profitability of the related trading strategy that amounts to opening and closing a position on the same day.

3. Methodology and data

We bypass the complex, and at least to some extent non-predictable, relationship between the VIX and its futures by modeling the VXF dynamics directly. We do this using a neural network approach, which appears to be more successful than a (linear) time series approach for anticipating the evolution of the implied volatility, as has been suggested by the literature.

Furthermore, the choice of the exogenous variables is very important too. When modelling financial phenomena, it is common practice to take the information set from the same market to which the variables being explained belong. Nevertheless, as was suggested by Shen, Jiang, and Zhang (2012), under the weak efficiency hypothesis, the price dynamics in markets that close right before or at the very beginning of U.S. trading should incorporate much more information than lagged variables on the U.S. market. Thus, we augment the information set with some "coincident indicators" taken from Asian stock exchange markets, in order to take advantage of the time zone difference.

The VIX index is calculated using options with two consecutive expirations of more than 23 days and less than 37 days to expiration (for further information, see CBOE, 2015). Typically, once a week, some of the options used for the calculation start to have less than 24 days to expiration, and thus they are rolled over to new maturities. The VIX index usually experiences a jump when this happens, and as a consequence, some bias arises if returns are computed as the log closing price difference between two consecutive days.

We avoid this problem by setting our dependent variable as the VXF open-to-close return (OTCR), which is calculated as $\ln(close_t) - \ln(open_t)$. OTCRs offer the advantage of taking into account only the "genuine" autocorrelation that arises from partial price adjustment and time-varying risk premia (Anderson et al., 2012), and incorporating the relevant information about the variability of financial assets, since the stock volatility is substantially higher intraday than overnight (Muravyev & Ni, 2016). No less importantly for the purposes of the present work, the use of intraday returns allows the forecasting performances to be connected easily to the profits earned by a trader who opens and closes a position on the same day. In this respect, it is also worth observing that, from a practical standpoint, a trading strategy that consists of buying/selling a VXF is not affected by small liquidity issues, since the liquidity of the futures market has grown considerably over the years (Shu & Zhang, 2012).

With respect to the usual approach based on the log difference between subsequent closing prices, we acknowledge that we do not measure any overnight gap. However, this does not represent a limit because we consider a trading strategy that opens and closes the positions within the same day.

3.1. Model specification

Both the VFX and the VIX, as well as the variables that are used to explain them (see Section 3.3), are collected on a set of (consecutive) trading days 1, 2, ..., T_3 (so that T_3 denotes the size of our dataset). Then, we fix two positive integers T_1 and T_2, such that $T_1 < T_2 < T_3$, in order to form three sub-samples. Specifically, the training set, containing the data observed on days 1, 2, ..., T_1, is used to estimate the econometric models. The validation set, containing the data observed on days $T_1 + 1$, $T_1 + 2$, ..., T_2, is then used to train the neural network, so as to minimize overfitting problems (that is, to reach a trade-off between model complexity and expected forecasting accuracy). The first T_2 observations are also used for optimizing the parameters of the employed trading strategy.

Finally, a third subset, the test set, which contains the data observed at $T_2 + 1$, $T_2 + 2$, ..., T_3, is exploited to assess both the ex-post forecast performances of the models and the profitability of the resulting trading strategies. We model intraday returns on the VXF by means of a multilayer augmented feed-forward neural network, a black-box approach that has been shown to be able to approximate complex (non-linear) relationships (see (Thenmozhi, 2006).

We specify a single hidden layer neural network. The input layer is made by S nodes, or neurons, that correspond to the explanatory variables. A constant term $w_{\cdot,0}$ (the so-called bias) is also included. These input terms are first multiplied by a matrix W of weights, then transformed by a non-linear function (the so-called transfer or activation function, which we denote by f):

$$h_j(x_1, x_2, \ldots, x_S) = f\left(w_{j,0} + \sum_{i=1}^{S} w_{j,i}x_i\right), \quad j = 1, 2, \ldots, J, \tag{1}$$

where J is the number of neurons in the hidden layer, to be selected according to the parsimony principle (so as to achieve the best trade-off between complexity and forecasting accuracy). Following a common approach, the activation function is chosen to be the hyperbolic tangent function:

$$f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}. \tag{2}$$

Then, as happens for input nodes, the hidden neurons' output is sent to the output layer, multiplied by a second matrix V of weights and transformed by a non-linear function:

$$out_n\left(h_1(\cdot), h_2(\cdot), \ldots, h_J(\cdot)\right) = f\left(v_{n,0} + \sum_{j=1}^{J} v_{n,j}h_j(\cdot)\right), \quad n = 1, 2, \tag{3}$$

where out is a two-dimensional vector that represents the final neural network prediction. In particular, if $out_1 < out_2$, the OTCR VFX is forcasted down, whereas if $out_1 > out_2$, it is forcasted up. Normally, each node in a given layer is connected to all of the other nodes. For a given number of hidden layers (in our case, one), the model's complexity and ability to approximate the input depend on the number of neurons J. However, parsimony is usually seen as the leading principle for model specification, as additional complexity increases both the risk of overfitting and the computational time. Once the structure of the neural network has been created, the parameters are estimated by minimizing a suitable loss function, which is done by applying a numerical optimization algorithm (in network terminology, we say that the network is trained with a learning algorithm). The loss function that we choose, which is very common for our two-class problem, is the "softmax cross-entropy":

$$E = -\sum_{n=1}^{2} y_n \ln\left(\frac{e^{out_n}}{e^{out_1} + e^{out_2}}\right), \quad n = 1, 2, \tag{4}$$

and either $(y_1, y_2) = (0, 1)$ if the observed direction of the OTCR VFX is down or $(y_1, y_2) = (1, 0)$ if the observed direction of the OTCR VFX is up. The problem of bad local minima is dealt with by considering different starting points in the learning phase.

We evaluate whether the complexity and non-linearity implied by the neural network approach are worthwhile by using the following logistic regression as a benchmark for comparison:

$$\pi(x) = \frac{e^{\beta_0 + \sum_{j=1}^{S} \beta_j x_j}}{1 + e^{\beta_0 + \sum_{j=1}^{S} \beta_j x_j}}, \tag{5}$$

where $\pi(x)$ denotes the probability that the future observation is positive and x_1, x_2, ..., x_s are the same lagged endogenous and/or exogenous variables that we employ in the neural network model.

Moreover, still for comparison purposes, we consider a simple model (Naïve) that always predicts VXF OTCRs to be negative (since an empirical examination of the VXF time series shows that negative OTCRs are more likely to occur than positive OTCRs).

Finally, for the sake of comparison with a more conventional approach, we also try to forecast VXF movements based on VIX predictions. Specifically, we use a heterogeneous autoregressive model of the realized volatility (HAR) and an augmented HAR (HAR_X) to predict VIX. The HAR model, which was due originally to Corsi (2009), and the HAR_X model are often employed for predicting the volatility of financial time series, since they are parismonious approaches and can capture long-memory effects (see e.g. Busch, Christensen, & Nielsen, 2011; Degiannakis & Filis, 2017; Degiannakis et al., 2018; Fernandes et al., 2014).

According to the HAR model, the realized volatility (RV) is predicted as follows:

$$RV_t = b_0+b_1RV_{t-1}+b_2MA(RV)^5_{t-1}+b_3MA(RV)^{22}_{t-1}+\epsilon_t, \quad (6)$$

where

$$MA(RV)^5_{t-1} = \frac{1}{5}\sum_{k=1}^{5} RV_{t-k}, \quad MA(RV)^{22}_{t-1} = \frac{1}{22}\sum_{k=1}^{22} RV_{t-k}. \quad (7)$$

The HAR_X specification, on the other hand, predicts the realized volatility by also taking into account the current values of S exogenous regressors:

$$RV_t = b_0 + b_1RV_{t-1} + b_2MA(RV)^5_{t-1} + b_3MA\,(RV)^{22}_{t-1} + \sum_{i=1}^{S}\left(c_{1,i}x_{i,t} + c_{2,i}MA\,(x_i)^5_t + c_{3,i}MA\,(x_i)^{22}_t\right), \quad (8)$$

where $MA\,(x_i)^5_t$ and $MA\,(x_i)^{22}_t$ are defined analogously to Eq. (7). The index t in the exogenous variables does not imply the use of future data that do not belong to the current information set; in fact, the exogenous regressors refer to markets that close before the VFX opens (see Section 4.1), and thus, they become available at least four hours in advance.

The HAR and HAR_X models are designed specifically for forecasting the realized volatility, and thus, they cannot be applied to VXF returns directly. We therefore consider VIX in place of RV. By doing so, we predict the VIX based on either Eq. (6) or Eq. (8), and then forecast the future direction of the VXF OTCR as follows:

$$\begin{cases} VXF_t - VXF_t < 0 \textit{ if } VXF_t - VXF_t < 0 \\ VXF_t - VXF_t > 0 \textit{ if } VXF_t - VXF_t > 0 \\ VXF_t - VXF_t = 0 \textit{ if } VXF_t - VXF_t = 0. \end{cases} \quad (9)$$

3.2. Forecast accuracy

The forecast accuracy can be measured using either a loss function based on the magnitude of the forecasting error, such as the mean square forecasting error (MSFE) or the mean absolute forecasting error (MAFE), or a classification loss function (directional forecasting). While the former approach may be the most popular in the literature, the latter allows for a better assessement of the potential profitability. In fact, as Diebold and Mariano (1995), Granger and Pesaran (2000) and Leitch and Tanner (1995), among others, have showed, the directional accuracy (DA) is connected more closely with profits than standard accuracy measures such as MSFE. In addition, Blaskowitz and Herwartz (2011) emphasize the robustness of DA in the presence of signal bias and outliers. Finally, some papers (see e.g. Costantini, Cuaresma, & Hlouskova, 2016; Degiannakis, 2008) have combined DA with the profit/loss of a trading strategy and obtained decision-based loss functions that allow the accuracy to be assessed in economic terms.

Thus, the present paper employs the mean directional accuracy, which, for the kth model (we are going to compare six different models), is computed as follows:

$$MDA_k = \frac{1}{T_3 - T_2}\sum_{t=T_2+1}^{T_3} \mathbf{1}_{sign(OTCR_t)=sign(\widehat{OTCR}_{k,t})}, \quad (10)$$

where $\widehat{OTCR}_{k,t}$ denotes the OTCR at day t forecasted by the kth model, and $\mathbf{1}$ is an indicator function that is equal to one if the two signs coincide, and zero otherwise.

We test the directional forecasting accuracy using the market-timing test for predictive accuracy (Pesaran & Timmerman, 1992). The null hypothesis is that the predicted and realized signs are independent, i.e., the forecasted market directions are not informative regarding the sign of the realized returns. Granger and Pesaran (2000) provide the following version of the test:

$$PT = \frac{\sqrt{T_3 - T_2}\left(\frac{N_{pp}}{N_{pp}+N_{np}} - \frac{N_{pn}}{N_{pn}+N_{nn}}\right)}{\left(\frac{\hat{\pi}_f(1-\hat{\pi}_f)}{\hat{\pi}_o(1-\hat{\pi}_o)}\right)^{1/2}}, \quad (11)$$

where the subscripts p and n indicate positive and negative VXF returns, respectively, N_{pn} is the number of times that the VXF return was negative and the forecast was positive, and N_{pp}, N_{nn} and N_{np} are defined correspondingly. Moreover, $\hat{\pi}_o = \frac{N_{pp}+N_{np}}{T_3-T_2}$ is the probability that returns are positive and $\hat{\pi}_f = \frac{N_{pp}+N_{pn}}{T_3-T_2}$ is the probability that returns are forecast to be positive. As was shown by Granger and Pesaran (2000), under the null hypothesis that the predicted and the realized signs are independent, PT has a standard normal distribution (with a zero mean and unit variance). Thus, we can easily test whether the predicted and realized signs are independent by comparing them with the quantile of the standard normal distribution.

We devote special attention to data-snooping biases, which are a common problem in inference with non-linear models because of the many degrees of freedom that are lost. We therefore check whether the predictive superiority of the neural network is systematic rather than due to luck by assessing the predictive performance of this highly parametrized non-linear specification through the application of a Monte Carlo cross-validation

Table 1
Descriptive statistics on OTCR.

	VIX	VXF
Mean	−0.0056	−0.0010
Median	−0.0108	−0.0033
Minimum	−0.2844	−0.2030
Maximum	0.3270	0.1930
Standard deviation	0.0590	0.0352
Skewness	0.690	0.299
Kurtosis	5.97	6.27

technique. Specifically, we consider 1000 random permutations of the sequence of days 1, 2, …, T_3 on which data are observed. In each permuted sequence, we form the training set with the data at places 1, 2, …, T_1, the validation set with the data at places T_1+1, T_1+2, …, T_2, and the test set with the data at places T_2+1, T_2+2, …, T_3. By so doing, we end up with a different (random) distribution of the economic variables among the training, validation and test sets for each of the 1000 permutations. We then check the distribution of the mean directional accuracy of the simulated neural network model over the set of Monte Carlo permutations.

Furthermore, we also evaluate the superior predictive ability (SPA) of the rival models by applying the test developed by Hansen (2005), which is described briefly here. We consider each model as the benchmark in turn for computing the following relative performance of model k at time t:

$$d_{k,t} \equiv L\left(sign(OTCR_t), sign(\widehat{OTCR}_{0,t})\right) - L\left(sign(OTCR_t), sign(\widehat{OTCR}_{k,t})\right), \tag{12}$$

where $k = 0$ refers to the model that is chosen as the benchmark, whereas $k = 1, 2, \ldots$ refer to the rival models and $L(\cdot)$ is a given loss function. Let us consider the sample average of $d_{k,t}$:

$$\bar{d}_k = \frac{1}{T_3 - T_2} \sum_{t=T_2+1}^{T_3} d_{k,t}, \tag{13}$$

and define:

$$A_k = \sqrt{T_3 - T_2} \cdot \bar{d}_k. \tag{14}$$

Moreover, let $\hat{\omega}_k$ denote a consistent estimator of the standard deviation of A_k. Then, the null hypothesis of the test is that the predictive ability of the benchmark is superior to those of the other five models. Such a hypothesis is rejected based on the significance of the studentized test statistic

$$T^{SPA} \equiv \max\left[\max_{k=1,2,\ldots} \frac{A_k}{\hat{\omega}_k}, 0\right]. \tag{15}$$

Suitable p-values for the statistics T^{SPA} are calculated based on bootstrap resamples.

The trading strategy that we consider is as follows: on day t, depending on the forecasted value of $OTCR_t$, either we do nothing or we take a long/short position on the VXF when the market opens and liquidate it when the market closes. Accordingly, we measure the total profit in the time interval from day T_2+1 to day T_3 by using the cumulative directional value:

$$CDV = \sum_{t=T_2+1}^{T_3} DV_t, \tag{16}$$

where

$$DV_t = O_t \cdot OTCR_t, \tag{17}$$

with $O_t = 1, -1$, or 0 if we take a long, short or flat position, respectively, at day t. The Hansen test described above is also used to assess the superior abilities of the models to generate profits. In particular, we use the opposite of the directional value Eq. (17) as a loss function in Eq. (12).

3.3. Dataset

We take into account the VIX futures continuous time series constructed by Thomson Reuters Datastream (Type 0), which contains the prices of either the nearest contract month futures or the second-nearest contract month futures. We consider daily data from March 26, 2007, to September 30, 2017, removing days with no values (e.g. holidays). With this choice, we collect $T_3 = 2{,}639$ observations on the whole dataset. We form the training, validation and test sets introduced in Section 3.2 by setting T_1 =1,718, T_2 =2,086 (i.e., the training set contains data from March 26, 2007, to January 21, 2014; the validation set contains data from January 22, 2014, to July 8, 2015; and the test set contains data from July 9, 2015, to September 30, 2017).

We compute the logarithmic OTCR series of both VIX and VXF. Descriptive statistics (for the subsample consisting of the first T_2 observations) are provided in Table 1.

As is typical of many financial time series, the distributions of log-returns are slightly asymmetric and show kurtosis values greater than three. The Jarque–Bera (JB) normality test always rejects normality ($p < 5\%$). The time series of both VIX and VXF returns are stationary, as is indicated by the Augmented Dickey-Fuller (ADF), Philips-Perron (PP) and KPSS tests. However, the futures oscillate less than their underlying (i.e., the difference between the maximum and the minimum is smaller).

Thomson Reuters Datastream is also the source of the independent variables. We consider both lagged endogenous and exogenous variables selected from among the Asian world stock indices that close right before the opening of the U.S. market in order to account for possible market sentiment on the latest economic news or responses to progress in major world affairs (see e.g. Shen et al., 2012). Specifically, variables are selected by looking at their correlations with the VXF OTCR. The results suggest that lags 0 and 1 of the DRs of the following four indices be kept: Nikkei 225 (N225), Hang Seng (HSI), ASX 200 (ASX200) and SENSEX (BSESN). We also allow for possible autoregressive dependence by keeping the first two lags of the dependent variable, even if the (linear) autocorrelation function was not significant. The closure values for non-lagged indices are available from eight hours and a half (N225) to five hours (BSESN) prior to the opening of the CBOE. Thus, we try to exploit the information that is available to traders in their "nowcasting" activity as much as possible, assuming that they need a minimum time lag in order to estimate models and set up their investment strategies. We do not consider data from European markets.

The data also exhibit significant cross-correlations at higher orders for BSESN and N225, but we do not take this into account due to the parsimony principle. We also exclude macroeconomics, bonds and commodities, because the literature has often questioned how informative their content is (see e.g. Psaradellis & Sermpinis, 2016).

In summary, we work with four independent variables, namely the DRs of the Nikkei 225, Hang Seng, ASX 200 and SENSEX. Standard *t*, JB, ADF, PP, KPSS tests show that each of these variables has a zero mean and is normally distributed and mean-stationary (at the 95% confidence level). Overall, if we count both lagged and coincident variables, we perform an initial specification step by considering, besides the intercept, ten regressors (those reported in Table 2) for both the neural network and the logistic models, and fifteen regressors (those reported in Table 3) for the HAR_X model.

4. Results

4.1. Model specification and estimation

Let us consider the logistic regression in Eq. (5) with the 10 regressors above. This model, which we call $L10$, is estimated by maximum likelihood on the sub-sample of data that contains the first $T_2 = 2086$ observations. As can be seen from Table 2, not all of the variables turn out to be significant. Thus, only the current BSESN is retained (together with the intercept) after a backward stepwise selection process, since it is the only significant variable. We call this parsimonious model $L1$. It is worth noting that BSESN contains the most recent information, since India is the market that closes last among the four that we considered. $L10$ shows a better goodness-of-fit than $L1$ but has a worse BIC, and thus, the less-parametrized model turns out to be our choice.

We account for a more general form of non-linearity by also estimating a feed-forward single-hidden-layer neural network. Analogously to what was done for the logistic specification, we consider two sets of inputs, namely the 10 variables in Table 2 and the current BSESN only. The models are labelled $N10_J$ and $N1_J$, respectively, where J represents the number of neurons in the hidden layer and is chosen according to the procedure outlined below. We will find the best value of J based on the cross-entropy in Eq. (4) for the models with both one and ten regressors.

The neural network is trained with the scaled conjugate gradient algorithm on the training set that contains the first T_1 =1,718 observations, and we stop the training algorithm early based on the network performance achieved on the validation set containing T_2– T_1 = 368 observations. The maximum number of iterations of the conjugate gradient method is capped at 1000.

Table 2
Parameters' estimations (benchmark models on VXF).

	$L10$	$L1$
Intercept	-0.2570^{**}	-0.250^{***}
$VXF\ OTCR_{t-1}$	-0.0110	
$VFX\ OTCR_{t-2}$	-0.0060	
$ASX200\ DR_t$	-0.0882	
$ASX200\ DR_{t-1}$	-0.0773	
$BSESN\ DR_t$	-0.1053^{**}	-0.0942^{**}
$BSESN\ DR_{t-1}$	0.0462	
$HSI\ DR_t$	0.0385	
$HSI\ DR_{t-1}$	0.0297	
$N225\ DR_t$	0.0306	
$N225\ \ DR_{t-1}$	-0.0064	
BIC	2.9008×10^3	2.8635×10^3
Pseudo R^2	0.0112	0.0068

Note: ** $p < 0.01$.

For both the $N1_J$ and $N10_J$ architectures, the number J of neurons in the hidden layer is chosen as follows. First, we generate 10,000 randomly-selected sets of weights, which we use as initial weights for training the neural network. Then, we consider the following expected performance indices:

$$EP(N\cdot_J) = \frac{1}{10{,}000}\sum_{i=1}^{10{,}000} MDA_i(N\cdot_J), \tag{18}$$

where MDA_i is the mean directional accuracy (computed according to Eq. (10), with $OTCR_t$ for $t = 1, 2, \ldots, T_2$) that we obtain when we train the neural network starting from the ith set of weights.

We let J vary from 2 to 20 and find the value that maximizes the EP in Eq. (18). For both the $N1_J$ and $N10_J$ specifications, the maximum $EP(N\cdot_J)$ value is obtained with $J = 16$ neurons, which we thus identify as the best complexity for the two models. The expected performance of the network with only $BSESN_t$ as input, 16 hidden nodes and two output neurons is $EP(N1_{16}) = 59.7\%$, whereas the expected performance of the network with all the ten regressors is $EP(N10_{16}) = 58.4\%$. Moreover, once 16 neurons have been chosen in the hidden layer, we also compute the maximum of the expected performance over the set of 10,000 sets of initial weights:

$$EP_{max}(N\cdot_{16}) = \max_{i=1,2,\ldots,10{,}000} MDA_i(N\cdot_{16}), \tag{19}$$

Both the $N1_{16}$ and $N10_{16}$ networks reach the same maximum MDA, equal to 61.4%, while the maximum MDA values achieved by the logistic specifications $L1$ and $L10$ are lower, namely 59.2% and 60.2%, respectively. Moreover, the maximum MDA values achieved by HAR, HAR_X_6 and HAR_X_{15} are 49.4%, 53.0% and 52.8%, respectively.

Essentially, the gap in accuracy between the neural networks and the logistic models is due to the greater ability of the networks to grasp upwards movements (which are less common in the sample but more profitable, as their average OTCR is 2.64% vs. an average OTCR of –2.50% for negative returns). In particular, if we focus on upward returns, the maximum values of MDA that are reached by $N1$ and $N10$ are 37.8% and 41.0%, respectively, while the logistic specification achieves only 38 true positives on 910 positive observations. In contrast, the neural

Table 3
Parameters' estimations (benchmark models on VIX).

	HAR_X_{15}	HAR_X_6	HAR
Intercept	0.4524^{**}	0.3377^{**}	0.282^{**}
$VIX\ DR_{t-1}$	0.5898^{**}	0.7122^{**}	0.7913^{**}
$MA(VIX\ DR)^5_{t-1}$	0.3340^{**}	0.2368^{**}	0.1820^{**}
$MA(VIX\ DR)^{22}_{t-1}$	0.0564^{**}	0.0365^{*}	0.0140
$BSESN\ DR_t$	-0.3055^{**}	-0.4020^{**}	
$MA(BSESN\ DR)^5_t$	0.0137	−0.1151	
$MA(BSESN\ DR)^{22}_t$	0.0993	−0.1586	
$ASX200\ DR_t$	-0.2168^{**}		
$MA(ASX200\ DR)^5_t$	-0.3076^{*}		
$MA(ASX200\ DR)^{22}_t$	0.4116		
$HSI\ DR_t$	−0.0602		
$MA(HSI\ DR)^5_t$	0.0112		
$MA(HSI\ DR)^{22}_t$	-0.4622^{*}		
$N225\ DR_t$	−0.0387		
$MA(N225\ DR)^5_t$	-0.2180^{*}		
$MA(N225\ DR)^{22}_t$	−0.2569		
BIC	8.5321×10^3	8.5771×10^3	8.7977×10^3
R^2	0.9683	0.9665	0.9624

Notes: ** $p < 0.01$, * $p < 0.05$.

network is superior to the HAR class for forecasting the downward movements of the VXF.

In general, the neural networks yield predictive accuracy levels that are considerably higher than those achieved by other models, which indicates the existence of a marked non-linearity in the relationship between futures OCTRs and the exogenous variables.

If we agree that predictive performances constitute an effective measure of informative contents, then markets are capable of packing the information about past events into current information. That is, the dependence on the most recent indicator (the BSESN index) subsumes the information contained in all of the other (less recent) variables. This conclusion holds for both the logistic specification and the neural network model, so it is robust to different types of non-linearity in the relationships linking variables. Furthermore, our finding that all of the relevant information about current events is contained in the most recent past is consistent with the results of several empirical studies (see for example Ahoniemi, 2006; Degiannakis, 2008) that have shown the low predictive accuracy of time series models with long memory. For all of these reasons, only parsimonious specifications (with one input variable) of the neural network and logistic models are considered hereafter.

The estimation of the HAR and HAR_X models, which is shown in Table 3, provides a slightly different picture. The introduction of the exogenous variables still brings a clear improvement in the BIC, but now, unlike what we experienced for the logistic model, the coincident BSESN is not the only informative variable, as the lagged endogenous variables and indices from markets other than India are statistically significant too.

Moreover, the BIC of HAR_X_{15} is slightly better than that of HAR_X_6. Nevertheless, we consider not only the best performing HAR_X_{15}, but also HAR_X_6 with BSESN as the only exogenous regressor, and the HAR as a benchmark, in order to perform broader comparisons with the $L1$ and the $N1_{16}$ specifications.

4.2. Assessing forecasting performances

We then shift the focus to out-of-sample performances by considering the set of data from day $T_2 + 1$ to day T_3. We find that the out-of sample exact classification rates of models $L1$, $N1_{16}$, HAR_X_{15}, HAR_X_6 and HAR are 60.2%, 65.8%, 52.4%, 52.4% and 42.1% respectively.

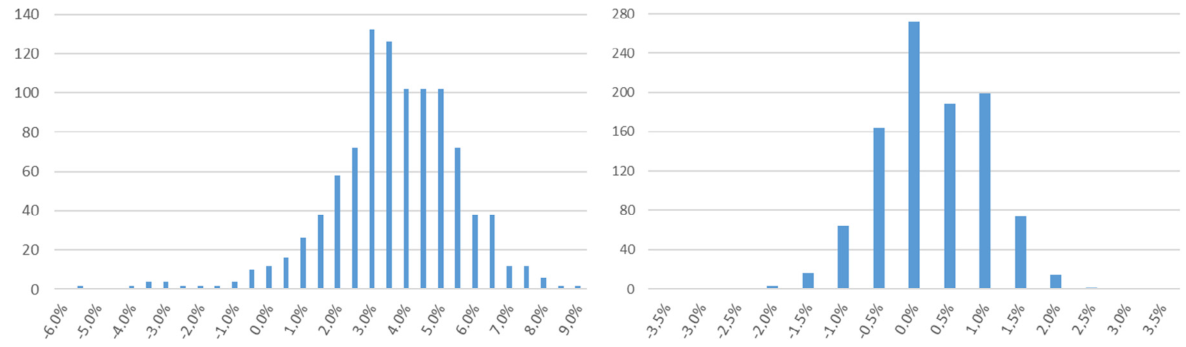
At first glance, the performances of the models that include exogenous variables look to be in line with the exact classification rates obtained in previous works (i.e., 61.9% in Ahoniemi, 2006; 55.4% in Konstantinidi & Skiadopoulos, 2011; 70% Degiannakis et al., 2018). However, it must be stressed that no straight comparison is possible unless we consider the frequency of observing negative OTCRs in the test sample.

This point becomes particularly important when working with VXFs, since negative returns are more frequent than positive returns. As a hypothetical example, consider the case where the frequency of negative returns in the test sample is 70%. Then, a model that reaches an exact classification rate of 70% does not perform any better than a Naïve model which always forecasts negative outcomes.

Accordingly, we should measure the models' performances in terms of their performances relative to the observed frequency of negative OTCRs in the test set, which we find to be equal to 59.8%. Noting that the observed frequency of negative OTCRs coincides with the MDA of the Naïve model, we obtain a measure of the relative performance (RP) by subtracting the MDA of the Naïve model from that of each of the competing specifications. In particular, the one-input neural network achieves a relative performance of $RP_{N1} = 6.0\%$ (i.e. 65.8% − 59.8%) and the logistic specification reaches a relative performance of $RP_{L1} = 0.4\%$ (i.e. 60.2% − 59.8%), while HAR_X_{15}, HAR_X_6 and HAR perform worse than the Naïve specification.

One could argue that forecasting performances (relative to the Naïve model) depend how data are allocated among the estimation, validation and test sets. We determine whether our results are robust to sample partition by performing a Monte Carlo cross-validation robustness check for the models with positive relative performances. We consider 1000 random permutations of the sequence of days $1, 2, \ldots, T_3$ on which data were observed. For each permuted sequence, we continue to form the training set with the data at places $1, 2, \ldots, T_1$, the validation set with the data at places $T_1 + 1, T_1 + 2, \ldots, T_2$, and the test set with the data at places $T_2 + 1, T_2 + 2, \ldots, T_3$. Then, we re-estimate both the $L1$ and $N1_{16}$ models by using the data at places $1, 2, \ldots, T_2$ (again, we select the neural network by considering 10,000 different random initial weights and choose the neural network that achieves the best performance on the validation set). Finally, for each of the 1000 Monte Carlo sequences, we compute the MDA performances of the models by using the data at places $T_2 + 1, T_2 + 2, \ldots, T_3$ as follows:

$$RP_{N1} = MDA_i(N1_{16}) - MDA(Na\ddot{\imath}ve), \qquad (20)$$



(a) RP_{N1} (b) RP_{L1}

Fig. 2. Out-of-sample relative performances, 1000 random permutations of all of the T_3 data.

$$RP_{L1} = MDA(L1) - MDA(\text{Naïve}). \qquad (21)$$

The distribution of the relative performance (over the 1000 random permutations) is reported in Fig. 2. As can be seen, RP_{N1} is rarely negative, reaches a maximum value of 8.7%, and has a median equal to 3.4%. Moreover, the relative performance (6.0%) which we experienced using the true (baseline) sequence of OTCR values corresponds to the 92.7th percentile of the distribution of RP_{N1}.

The above evidence suggests that $N1_{16}$ performs considerably better than the Naïve model (the relative performance of which is null), and that such a conclusion is robust to sample allocation.

Finally, let us perform a model comparison using both the PT and SPA tests described in Section 3.2. For the sake of brevity, we only consider the $L1$, $N1_{16}$ and Naïve specifications, since, as has been observed already, the HAR_X_{15}, HAR_X_6 and HAR specifications perform worse than the Naïve specification.

The Hansen test considers each of the $L1$, $N1_{16}$ and Naïve models as the benchmark in turn (so as to check, for each model, whether it achieves a better predictive accuracy than the other five models). Moreover, we compute the test statistic T^{SPA} (see Eq. (15)) by taking the opposite of the MDA as the loss function, and the p-values associated with T^{SPA} are obtained based on 1000 bootstrap resamples. A small p-value indicates that the predictive accuracy of the model that was chosen as the benchmark is inferior to that of at least one of the alternative specifications.

The results are reported in Table 4. The PT test higlights the fact that the models' market timing abilities differ. In fact, it is possible to reject the null hypothesis of independence between the true and forecasted directions of change only for model $N1_{16}$ ($p < 1\%$), whereas the probability of making a type-I error for model $L1$ is close to 10%.

Moreover, in accordance with our previous findings, the Hansen test confirms that $N1_{16}$ is the only model that provides systematic and sizeable improvements in forecast accuracy with respect to the other models. Then, if we agree that the predictive performance is an effective specification test (i.e., a measure of the informative value of the input variables and of the validity of the functional form of the models), the following conclusions hold. First, the most recent information (the current BSESN) encompasses the contributions of all of the less recent variables, which is actually a form of weak efficiency that involves markets of different countries; second, the functional relationship that links the VXF OTCR to the current BSESN is complex and non-linear.

Table 4
Model performance and superior predictive ability.

	PT (p-value)	SPA (p-value)
Naïve	//	0.007
$L1$	0.080	0.005
$N1_{16}$	1.35E–09	1

4.3. Trading simulation

Nevertheless, accuracy does not necessarily imply profitability. Thus, we also checked economic profits by simulating a simple trading strategy that amounts to either doing nothing or opening a long/short position when the market opens and liquidating it when the market closes.

In particular, when predicting using the $N1_{16}$ and $L1$ models, trading is done as follows: on day t, before the opening of the CBOE, we forecast the probability that the VXF OTCR will be positive for that day, which, for the sake of brevity, we denote by ${p_t}^+$. Moreover, let THR denote a given threshold (*filter*). If ${p_t}^+ \geq 0.5 + THR$, then on day t we take a long position on the VXF ($O_t = 1$) when the market opens and liquidate it when the market closes. If ${p_t}^+ \leq 0.5 - THR$, then on day t we take a short position on the VXF ($O_t = -1$) when the market opens and liquidate it when the market closes. Finally, if $0.5 - THR < {p_t}^+ < 0.5 + THR$, then on day t we stay flat ($O_t = 0$). The trading strategies applied to the HAR_X_{15}, HAR_X_6 and HAR outputs are analogous to that described above, with the exception that we take into account the magnitude of the forecasted DR on VIX. Specifically, THR is now the smallest magnitude of the forecasted DR on VIX that is required in order to open a long/short position on VXF. That is, if the magnitude of the forecasted VIX DR does not exceed THR, then on day t we stay flat ($O_t = 0$).

Note that the threshold allows us to optimize the trading strategy and, as a consequence, it is determined based on the first T_2 observations, once the final models have been selected/estimated. In particular, following a common practice, we attempt to avoid "false signals" by simply filtering out the weakest signals. Precisely as was done by Ahoniemi (2006), we consider six different threshold levels for each model (see Table 5), with the goal of determining which filter yields the highest level of profitability on the first T_2 observations. As far as profitability is concerned, we measure it by means of the cumulative directional value in Eq. (16) (where we replace T_2 and T_3 with 1 and T_2, respectively). The bid–ask spread is neglected, and, following a common approach (see, e.g., Psaradellis & Sermpinis, 2016), commissions are set to 50 cents per contract.

The results obtained are reported in Table 5. As can be seen, the neural network performs significantly better than the logistic model with every filter, and is also more profitable than the *HAR*, *HAR_X*$_{15}$, *HAR_X*$_6$ models, provided that $THR \leq 2.5\%$. The augmented HAR models are more profitable than the simple HAR model with any filter, which further confirms the importance of making use of the time zone difference when collecting exogenous information. The neural network is the only model that achieves the best trading performance with $THR = 0$, i.e., that can discriminate even the weakest signals from the BSESN correctly.

Furthermore, the HAR_X models achieve performances that are quite similar to those of the logistic specification. They produce large numbers of true positives, which is crucial for profitability, as the highest VXF OTCRs are usually experienced in correspondence with upward movements. On the other hand, the logistic is the specification that performs best for predicting true negatives, with an exact classification rate of greater than 92% for any $THR \leq 5\%$.

Once the filter (*THR*) that yields the optimal trading strategy has been selected (for each specification), we measure the out-of-sample profitability on the data observed on days $T_2 + 1, T_2 + 2, \ldots, T_3$. The goal is to check whether there are also significant differences in the economic performances of the best strategies that can be constructed based on the predictions of the neural network, logistic, Naïve, HAR and HAR_X models. This is accomplished by means of the SPA test, in which the opposite of *DV* (see Eq. (17)) is used as the loss function, and the *p*-values of the statistic T^{SPA} are computed based on 1000 bootstrap resamples.

The results obtained are reported in Table 6 (the maximum drawdown in the 5th column is actually the *CDV* in Eq. (16) that is computed by considering only losses reported on consecutive trading days). As can be seen, all of the models except for *HAR* generate relevant profits. However, the Naïve model exhibits an extremely large maximum drawdown, which makes it impossible for an investor who does not have substantial additional capital to compensate for such losses to match the performance reported in Table 5. It is interesting to observe that the neural network outperforms all of its competitors in terms of profitability, and also yields the smallest number of false signals.

If we consider the risk (measured by the standard deviation of returns and the maximum drawdown), the strategy built on the forecasts of the logistic regression yields the best performance, but the result depends on the filter applied; in particular, the smallest standard deviation and the smallest maximum drawdown are achieved if the percentage of effective trading days is 80% (see the second column of Table 6). Nevertheless, the strategy based on neural network forecasts has the highest Sharpe ratio, i.e., yields the best trade-off between expected profit and risk. Finally, as far as the HAR class is concerned, it is clear that the performances of *HAR_X*$_6$ and *HAR_X*$_{15}$ are very similar, and that they are much more profitable than *HAR*.

The SPA test confirms, on an inferential base, that the $N1_{16}$ neural network provides systematic improvements in economic performance over all of the other models. We conclude that we can achieve significantly higher profits than the logistic, Naïve, HAR and HAR_X models by taking into account information with a minimum time zone difference and using a very flexible non-linear specification.

4.4. Trading simulation based on daily returns (DRs)

We provide additional information that is more comparable with the existing literature by assessing the profitability of the previous models when the dependent variable is chosen to be the VXF DR, rather than the VXF OTCR.

Following the same procedure as in Section 4.1, we start by selecting the more informative exogenous variables in the logistic model, as well as the best neural network architecture. The results indicate that two coincident variables, namely the BSESN DR and the ASX200 DR, and two lagged variables, namely the VXF DR and the BSESN DR, are statistically significant in explaining the dynamics of the DRs. We denote the logistic model with those regressors as *L*4. We use the same information set to train the network, and find that the best architecture is now achieved by employing 10 nodes in the single hidden layer. We indicate this optimal architecture as $N4_{10}$. For the sake of comparison, we also calculate the performances of the *HAR*, *HAR_X*$_6$ and *HAR_X*$_{15}$ models, considering the same values of *THR* as in Table 5.

The results obtained (with the optimally chosen filter) are reported in Table 7. Again, the neural network turns out to be the best-performing model, yielding profits that are at least 1.7 times as high as those provided by any of the rival specifications. Moreover, the models that predict the VXF directly are more profitable if they are used with OTCRs rather DRs (compare with Table 6). The gap is even more evident if we consider the Sharpe ratio. On the other hand, models that forecast the VIX direction (the HAR class) yield higher profits when the trading strategies are based on DRs rather than OTCRs. However, if we focus on the risk, the standard deviation of profits is larger overall than that experienced in the case of OTCRs, which reflects the higher level of uncertainty that affects DRs.

Finally, on an inferential basis, the SPA test confirms that the neural network provides systematic improvements over all other models considered.

Table 5
Profitabilities for different probability thresholds in the time period from day 1 to day T_2.

	$N1_{16}$	$L1$		HAR	HAR_X_6	HAR_X_{15}
$THR = 0$	663.5% (100%)	308.0% (100%)	$THR = 0$	−33.9% (100%)	433.2% (100%)	420.6% (100%)
$THR = 0.5\%$	622.0% (95.4%)	312.6% (96.0%)	$THR = 0.25\%$	−78.3% (87.5%)	372.8% (93.7.%)	429.8% (94.2%)
$THR = 1\%$	641.1% (93.5%)	323.3% (94.4%)	$THR = 0.5\%$	−7.5% (75.6%)	374.9% (87.0%)	448.5% (87.7%)
$THR = 2.5\%$	639.9% (87.5%)	384.8% (89.2%)	$THR = 1\%$	65.0% (50.7%)	422.8% (72.4%)	462.9% (75.2%)
$THR = 5\%$	537.6% (77.6%)	458.6% (70.5%)	$THR = 1.5\%$	99.2% (32.2%)	433.6% (59.7%)	412.2% (65.1%)
$THR = 10\%$	362.4% (46.6%)	82.7% (10.0%)	$THR = 2\%$	−25.9% (18.4%)	349.9% (48.9%)	384.1% (54.0%)

Note: The fraction of trading days in the time period considered is given in parentheses.

Table 6
Trading strategies' performances and superior predictive ability in the time period from day $T_2 + 1$ to day T_3.

Models	Profitability (CDV)	# false signals ($1 - MDA$)	Returns' standard dev.	Maximum drawdown	SPA test (p-value)
Naïve	226.9% (100%)	40.1%	4.3%	53.4%	0.012
$L1$	466.8% (80%)	34.2%	3.8%	15.1%	0.028
$N1_{16}$	647.8% (100%)	34.2%	4.2%	24.7%	1
HAR	17.4% (49.4%)	59.0%	4.6%	49.0%	0
HAR_X_6	326.3% (62.7%)	47.6%	4.3%	16.4%	0.003
HAR_X_{15}	317.5% (78.3%)	47.6%	4.3%	19.3%	0.005

Note: The fraction of trading days in the time period considered is given in parentheses.

Table 7
Trading strategies' performances and superior predictive ability in the time period from day $T_2 + 1$ to day T_3.

Models on DRs	Profitability (CDV)	# false signals ($1 - MDA$)	Returns' standard dev.	Maximum drawdown	SPA test (p-value)
Naïve	47.4% (100%)	41.2%	4.7%	52.2%	0
$L4$	324.9% (85.9%)	40.4%	4.7%	21.2%	0
$N4_{10}$	632.9% (100%)	36.7%	4.5%	21.1%	1
HAR	78.6% (49.4%)	60.1%	4.9%	47.7%	0
HAR_X_6	336.6% (62.7%)	49.0%	4.7%	18.4%	0.016
HAR_X_{15}	349.5% (78.3%)	49.4%	4.6%	27.2%	0.020

Note: The fraction of trading days in the time period considered is given in parentheses.

Moreover, despite the fact that BSSEN remains the only variable that is significant at both the coincident and lagged levels, it is no longer capable of explaining DRs on its own (as the coincident ASX200 DR and the lagged VXF DR are found to be statistically significant). This is due to the fact that the dynamics of DRs are more complex than those of OTCRs, since they are also influenced by nonsynchronous trading effects and bid–ask bounces (see Anderson et al., 2012).

It is also interesting to note that the performance of the Naïve model deteriorates significantly if the trading strategy is based on VXF OTCRs rather than on VXF DRs. In fact, the high profitability of the Naïve specification that we obtained by considering OTCRs (226.9%, see Table 6) is due essentially to the fact that the volatility of financial markets (and hence also of the VIX index) is normally greater in the morning and smaller at night (see e.g. Daigler, 1998; Garcia, Martelli, Rona, & Ta, 2018), so that selling the VXF in the morning and buying it at night turns out to be a profitable trading strategy. When considering VXF DRs, such an intraday effect is no longer exploited, because trading is done only when the market closes, and the regularity of the volatility trend is somehow broken by the random information flow that arrives overnight (see Anderson et al., 2012).

Finally, even if the neural network achieves almost the same performance when considering DRs rather than OTCRs, the use of DRs implies that traders do not have a minimum time lag for forecasting the VXF, and set up their investment strategies accordingly. In fact, they should place the order before the market closes, when the closure price of the VXF (or of the VIX if they are forecasting with the HAR class of models), is still unknown. As a consequence, the profit assessment might not mirror the performance of the trading strategy that is followed in practice correctly when considering DRs. This does not occur if OTCRs are considered, because traders have all of the time (overnight) to forecast the VXF (or the VIX) and to place the order.

5. Conclusions

We have investigated several relevant aspects related to the predictability of the VIX future (VXF). The use of open-to-close returns (OTCRs) "... offer[s] the advantage of taking into account only the 'genuine' autocorrelation that arises from partial price adjustment and time-varying risk premia ..." (Anderson et al., 2012). Moreover, the focus on VXF is close to the perspective of investors who recognize that the implied volatility can be traded only as a futures contract.

The present paper contributes to the existing literature in several respects. First, we show that the dynamics of the VXF do not reflect those of the VIX index closely. This is in line with previous work, but our analysis also highlights the fact that the "VIX-VXF puzzle" is more serious when measured on OTCRs than when measured on DRs. In particular, the correlation between the intraday returns of VIX and VXF is only 0.767 if calculated on the whole sample (2,454 observations), but 0.771 if we take the average of the correlations computed on monthly subsamples. On the top of that, the correlation distribution is quite erratic, so that an investor who wants to trade the VXF based on some, even very accurate, prediction of VIX intraday returns might not be able to earn significant profits.

Second, we establish that the neural network and logistic models with only one input variable, namely the most recent exogenous one, are superior to the unconstrained models with ten (lagged and coincident) regressors. Specifically, the specification that includes only the BSESN index yields a lower value of BIC and a higher expected performance. That is, prices on the CBOE reflect the last publicly available data, subsumed by the Indian index, which is actually a form of weak efficiency that involves markets of different countries.

In addition, one-day lagged endogenous variables are less informative than the price dynamics in a market that closes right before the US market, even if we fit the data with a very flexible approximator such as a neural network. In fact, we find evidence that lagged variables increase the network's complexity and cause overfitting problems. Thus, we can conclude that limiting the number of input nodes in these "black-box" models is more effective at reducing overfitting than pruning the hidden layer units. Overall, our findings reinforce the scepticism that is widespread in the literature regarding the usefulness of a pure long memory time series approach for VXF (or VIX) forecasting.

Third, we compare the proposed neural network model with a logistic regression, a HAR and a HAR_X model, as well as with a Naïve model that always forecasts negative outcomes. Our results indicate that the mean directional accuracy achieved by the neural network specification is significantly higher than those achieved by any of the other models. Moreover, as is revealed by Monte Carlo cross-validation, the better performance of the neural network is robust to sample selection bias. Then, first, non-linearity matters. Furthermore, a considerable degree of predictability of VXF OTCR signs can be obtained if non-linearity is combined with information which is available on a market that closes right before the U.S. market. Specifically, we forecast directional changes correctly on 65.8% of trading days, which we consider represents a fairly good predictive performance, confirming the existence of strong non-linearities and supporting the use of a neural network.

Fourth, in line with the most recent literature, we assess the profitability of alternative trading strategies that rely on predictions of VXF directional changes. We find that the use of filters for limiting false signals (leaving out the weakest signals) enhances the profitability for all models employed except for the neural network, which is capable of exploiting all of the information contained in the coincident BSESN index. Overall, the neural network's performance was 647.8% in almost three years (553 trading days), which is significantly higher than those of the Naïve, logistic, HAR and HAR_X models.

Finally, the present paper does not use data from European markets, since we assume that traders need a minimum time lag to estimate models and set up investment strategies. Moreover, we focus only on the future with the nearest expiration date; i.e., we do not take into account the term structure of the VXF. However, it could be interesting to consider also some European market indexes and/or futures with longer maturities. This could be the subject of a future study.

References

Ahoniemi, K. (2006). Modeling and forecasting implied volatility: An econometric analysis of the VIX index. Working paper, Helsinki School of Economics.

Anderson, R. M., Eom, K. S., Hahn, S. B., & Park, J. H. (2012). Sources of stock return autocorrelation. Working paper, University of California at Berkeley.

Asensio, I. O. (2013). The VIX-VIX futures puzzle. Working paper, University of Victoria.

Blaskowitz, O., & Herwartz, H. (2011). On economic evaluation of directional forecasts. *International Journal of Forecasting*, *27*, 1058–1065.

Busch, T., Christensen, B. J., & Nielsen, M. Ø. (2011). The role of implied volatility in forecasting future realized volatility and jumps in foreign exchange, stock, and bond markets. *Journal of Econometrics*, *160*, 48–57.

CBOE, (2015). The CBOE Volatility Index – VIX. CBOE White Paper.

Corsi, A. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, *7*, 174–196.

Costantini, M., Cuaresma, J. C., & Hlouskova, J. (2016). Forecasting errors, directional accuracy and profitability of currency trading: The case of EUR/USD exchange rate. *Journal of Forecasting*, *35*, 652–668.

Daigler, R. T. (1998). Intraday futures volatility and theories of market behavior. *The Journal of Futures Markets*, *17*, 45–74.

Degiannakis, S. A. (2008). Forecasting VIX. *Journal of Money, Investment and Banking*, *4*, 5–19.

Degiannakis, S. A., & Filis, G. (2017). Forecasting oil price realized volatility using information channels from other asset classes. *Journal of International Money and Finance*, *76*, 28–49.

Degiannakis, S. A., Filis, G., & Hassani, H. (2018). Forecasting global stock market implied volatility indices. *Journal of Empirical Finance*, *46*, 111–129.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 134–144.

Fernandes, M., Medeiros, M. C., & Scharth, M. (2014). Modeling and predicting the CBOE market volatility index. *Journal of Banking and Finance*, *40*, 1–10.

Garcia, C., Martelli, A., Rona, L., & Ta, A. (2018). Intraday volatility prediction. Working paper.

Granger, C. W. J., & Pesaran, M. H. (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, *19*, 537–560.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics*, *23*, 365–380.

Jabłecki, J., Kokoszczyński, R., Sakowski, P., Ślepaczuk, R., & Wójcik, P. (2014). Does historical VIX term structure contain valuable information for predicting VIX futures? *Dynamic Econometric Models*, *14*, 5–28.

Konstantinidi, E., & Skiadopoulos, G. (2011). Are VIX futures prices predictable? an empirical investigation. *International Journal of Forecasting*, *27*, 543–560.

Konstantinidi, E., Skiadopoulos, G., & Tzagkaraki, E. (2008). Can the evolution of implied volatility be forecasted? evidence from european and US implied volatility indices. *Journal of Banking and Finance*, *32*, 2401–2411.

Leitch, G., & Tanner, J. E. (1995). Professional economic forecasts: Are they worth their costs? *Journal of Forecasting*, *14*, 143–157.

Luo, X., & Zhang, J. E. (2012). The term structure of VIX. *Journal of Futures Markets*, *32*, 1092–1123.

Muravyev, D., & Ni, X. (2016). Why do option returns change sign from day to night? Working paper.

Pesaran, M., & Timmerman, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business and Economic Statistics*, *10*, 461–465.

Psaradellis, I., & Sermpinis, G. (2016). Modelling and trading the U.S. implied volatility indices. evidence from the VIX, VXN and VXD indices. *International Journal of Forecasting*, *32*, 1268–1283.

Shen, S., Jiang, H., & Zhang, T. (2012). Stock market forecasting using machine learning algorithms. Working paper, Department of Electrical Engineering, Stanford University, Stanford, CA.

Shu, J., & Zhang, J. E. (2012). Causality in the VIX futures market. *The Journal of Futures Markets*, *32*, 24–46.

Thenmozhi, M. (2006). Forecasting stock index returns using neural networks. *Delhi Business Review*, *7*, 59–69.

Thomson Reuters Datastream, (2010). Futures continuous series – methodology and definitions. Working paper.

Whaley, R. E. (2008). Understanding the VIX. *Journal of Portfolio Management*, *35*, 98–110.

Luca Vincenzo Ballestra Ph.D. in Applied Mathematics from the University of Milan. In January 2016 he became Associate Professor of Mathematics for Economics at the University of Bologna (Department of Statistical Sciences), where he currently teaches Actuarial Mathematics. His research interests include: quantitative finance, option pricing, volatility modelling, and, more in general, the analysis and the application of stochastic models for finance.

Andrea Guizzardi Ph.D. in statistics from the University of Bologna, is actually associate professor at the University of Bologna (Department of Statistical Sciences), where he teaches Trading Algorithms. His research interests include the application of statistical modelling and forecasting evaluation to trading systems methods.

Fabio Palladini received a master's degree in Statistics from the University of Bologna. He is an adjunct Professor at the master in quantitative risk management finance (University of Bologna) and an options trader. His research interests include quantitative finance, option pricing, volatility modelling.