On the usefulness of cross-validation for directional forecast evaluation

Christoph Bergmeir[1], Mauro Costantini[2], and José M. Benítez[1*]

[1]Department of Computer Science and Artificial Intelligence, E.T.S. de Ingenierías Informática y de Telecomunicación, University of Granada, Spain.
[2]Department of Economics and Finance, Brunel University, United Kingdom

Abstract

The usefulness of a predictor evaluation framework which combines a blocked cross-validation scheme with directional accuracy measures is investigated. The advantage of using a blocked cross-validation scheme with respect to the standard out-of-sample procedure is that cross-validation yields more precise error estimates of the prediction error since it makes full use of the data. In order to quantify the gain in precision when directional accuracy measures are considered, a Monte Carlo analysis using univariate and multivariate models is provided. The experiments indicate that more precise estimates are obtained with the blocked cross-validation procedure. An application is carried out on forecasting UK interest rate for illustration purposes. The results show that in such a situation with small samples the cross-validation scheme may have considerable advantages over the standard out-of-sample evaluation procedure as it may help to overcome problems induced by the limited information the directional accuracy measures contain due to their binary nature.

KEY WORDS Blocked cross-validation; out-of-sample evaluation; forecast directional accuracy; Monte Carlo analysis; linear models.

*Corresponding author. DECSAI, ETSIIT, UGR, C/ Periodista Daniel Saucedo Aranda s/n, 18071 - Granada, Spain, E-mail address: j.m.benitez@decsai.ugr.es

1 Introduction

Assessing and evaluating the accuracy of forecasting models and forecasts is an important and long-standing problem which a forecaster always faces when choosing among various available forecasting methods. This paper aims to investigate the usefulness of a blocked cross-validation (BCV) scheme along with directional accuracy measures for forecast evaluation. Several forecast error measures such as scale-dependent, percentage and relative measures have been used largely for forecast evaluation (see Hyndman and Koehler (2006); Costantini and Pappalardo (2010), Costantini and Kunst (2011) among others).

However, Blaskowitz and Herwartz (2009) point out that directional forecasts can provide a useful framework for assessing the economic forecast value when loss functions (or success measures) are properly formulated to account for the realized signs and realized magnitudes of directional movements. In this regard, Blaskowitz and Herwartz (2009, 2011) propose several directional accuracy measures which assign a different loss to the forecast, depending on whether it correctly predicts the direction (rise/fall) of the time series or not. The idea behind this kind of measure is that there are many situations where the correct prediction of the direction of the time series can be very useful, even if the forecast is biased (an investor buys stock, if its price is expected to rise, Blaskowitz and Herwartz (2009); a central bank tends to increase the interest rate, if inflation is expected to rise, Milas and Naraidoo (2012)).
For purposes of out-of-sample (OOS) forecast evaluation, the sample is divided into two parts. A fraction of the sample is reserved for initial parameter estimation while the remaining fraction is used for evaluation. However, this procedure may fail to work well when the overall amount of data is limited and/or a lot of parameters are to be estimated. As the directional accuracy measures use the predictions in a binary way (correct/incorrect prediction of direction), the problems may be even more prominent when using such measures. Provided that the data used for forecasting are stationary (see Arlot and Celisse (2010)), the cross-validation scheme may help improve the estimation of the forecast directional accuracy, as it uses the data more efficiently by splitting them into k-folds. In this context, the use of the directional forecasting accuracy is also recommended since changes in the sign are frequent with stationary data (no increasing/decreasing trend).
This paper makes a contribution to the existing literature by investigating whether, and to what extent, the k-fold BCV procedure proposed by Bergmeir and Benítez (2012) may provide a better estimate of forecast directional accuracy than the standard OOS procedure. The use of the k-fold blocked scheme is suggested because it yields more precise error measures, in

the sense that the error measure calculated using BCV is a better estimate of the generalization error (the expected loss of the model on unknown future observations; see Blum et al. (1999); Hastie et al. (2009)).

This paper aims to evaluate if this benefit is also retained when the forecasts are tested for directional accuracy. To this end, we provide a Monte Carlo analysis using simple univariate and multivariate linear autoregressive models. The models are estimated and evaluated both with BCV and traditional OOS evaluation methods. Furthermore, the models are also evaluated on an additional validation set which consists of new unknown future data. This allows us to compare the directional accuracy obtained by the evaluation procedure and the directional accuracy obtained using the new future data. In this way, it is possible to ascertain how well the evaluation procedure is able to predict the future loss of a certain model. The Monte Carlo experiment results show that the advantage of using a BCV scheme is quite remarkable.

We use simple linear models as these models are likely to show a rather conservative behavior compared to more complex models regarding the differences in the outcome of the forecast evaluation, as more complex models require more data for parameter estimation, so the observed effects may be even stronger with complex models. Also, autoregressive models have been extensively used in the literature for directional forecasts, especially for the prediction of the exchange rate and the interest rate, as it is of primary importance for investors and policy makers to better understand the movements of these variables for the decision-making process (see Kong (2000); Sosvilla-Rivero and García (2005); Kim et al. (2008); Blaskowitz and Herwartz (2014); Blaskowitz and Herwartz (2009); Altavilla and De Grauwe (2010), among others). The use of these models has been justified on the basis of the potential correlation in the change of the exchange rate due to data measurement or aggregation (see Kong (2000)) and in the interest rate due to monetary policy of the central bank.

We also offer an empirical application to the UK interest rate data. The forecast results show that, when using directional accuracy measures in small sample sizes, it may happen that distinct forecast approaches reveal identical realized average loss/success. The BCV scheme uses additional information from other test sets and is less likely to obtain identical loss estimates, thus it is able to distinguish the performance of the models.

The rest of the paper is organized as follows. Section 2 reviews the BCV procedure. Section 3 describes the directional accuracy measures. Section 4 provides the Monte Carlo results. Section 5 discusses our empirical findings, and Section 6 concludes.

2 Blocked cross-validation

Cross-validation is an estimator widely used to evaluate prediction errors (Borra and Di Ciaccio, 2010; Khan et al., 2010). In k-fold cross-validation (see, e.g., Hastie et al. (2009)), the overall available data is randomly partitioned into k sets of equal size: each of the k sets is used once to measure the OOS forecast accuracy and the other $k-1$ sets are used to build the model. The k resulting error measures are averaged using the mean to calculate the final error measure. The advantage of cross-validation is that all the data is used both for training (initial estimation) and testing and the error measure can be computed k times instead of only one. Therefore, by averaging over the k measures, the error estimate using cross-validation has a lower variance compared to an error estimate using only one training and test set. In this way, a more accurate evaluation of the generalization error can be obtained (see Blum et al. (1999) for a theoretic result on this).

Since the cross-validation scheme requires the data to be i.i.d. (see Arlot and Celisse (2010)), modified versions of cross-validation for time series analysis have been proposed (for a large survey see Bergmeir and Benítez (2012)). Identical distribution translates to stationarity of the series (Racine, 2000). Independence can be assured by leaving a margin of a certain distance in time d between training and test values, after which the values are approximately independent (and no autocorrelation is present). The value d will be typically related to the order of the model, as we assume that all the autocorrelation is considered during model building. Therefore, the values in a neighborhood of d, around a value which is used for testing, cannot be used for training (see, e.g., McQuarrie and Tsai (1998) or Kunst (2008) for the procedure).

If the test set is chosen randomly from the data, removing a neighborhood of d values from the training set for each value in the test set may lead (depending on d and on the size of the test set) to a considerable loss of data and may even result in an insufficient amount of data for model estimation. A solution to this problem is to choose the test set as a block of sequential data (see also Racine (2000)), so that omission of dependent values is only necessary at the borders of this data block. This scheme, applied to k-fold cross-validation, yields k-fold BCV (Bergmeir and Benítez, 2012). Figure 1 shows a simple example how the BCV procedure works in practice (see also Bergmeir and Benítez (2011)).

The choice of k depends on the computational cost and the amount of available data. More specifically, the number of models to be estimated increases as k increases and the smaller k, the smaller is the training set (which may represent a problem if only few data is available). Typical choices

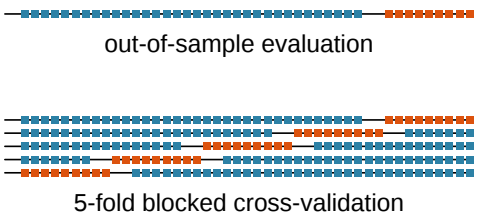

Figure 1: Training and test sets chosen for traditional OOS evaluation, and 5-fold BCV. Blue dots represent points from the time series in the training set and orange dots represent points in the test set. In the example, we assume that the model uses two lagged values for forecasting, which is why at the borders always two values are omitted.

for k are 5 or 10 (see Khan et al. (2010); Hastie et al. (2009)). In this work we use $k = 5$.

In the following, we offer a theoretical analysis to show the advantage of using BCV over OOS. Specifically, we will show that the BCV procedure yields an error measure with a lower variance than that of OOS.

Let x_t be a time series. Let D be a $(N, p + 1)$-matrix of time series data to which an autoregressive model is applied. The rows of D are of the form $((x_{t-p}), \ldots, x_{t-1}, x_t, x_{t+h})$, where h is the forecast horizon. Let $\kappa : \{1, \ldots, N\} \mapsto \{1, \ldots, k\}$ be an index function that indicates the partition to which row i belongs to. The index function is built according to the paradigm of BCV, so that k blocks of equal size of data are generated (we assume for simplicity that the length of the time series is a multiple of k).

Let us consider the kth cross-validation estimate, where the kth partition is used as the test set. Note that this is equivalent to OOS evaluation, using an OOS period with a length of $1/k$th the length of the training data. We estimate a model $\hat{f}$ on the data D using all rows i with $\kappa(i) \neq k$. Denote such a model by $\hat{f}^{-\kappa_k}$ in the following. Then, an error measure M for $\hat{f}^{-\kappa_k}$ is calculated on D using all rows j with $\kappa(j) = k$. We denote such an error measure $M(\hat{f}^{-\kappa_k}, \kappa_k)$. So, we define the OOS estimate as:

$$OOS(\hat{f}) = M(\hat{f}^{-\kappa_k}, \kappa_k).$$

From this, we can straightforwardly define the BCV estimate as:

$$BCV(\hat{f}) = \frac{1}{k} \sum_{i=1}^{k} M(\hat{f}^{-\kappa_i}, \kappa_i).$$

In this case, it is straightforward to see that BCV uses all the information available to OOS and additional information from other test sets. Let us

consider the variance of the error measure. Assuming that the error measures M calculated on different test sets are uncorrelated, we have:

$$Var(BCV(\hat{f})) = Var(\frac{1}{k}\sum_{i=1}^{k} M(\hat{f}^{-\kappa_i}, \kappa_i)) = \frac{1}{k^2}\sum_{i=1}^{k} Var(M(\hat{f}^{-\kappa_i}, \kappa_i)),$$

and for stationary data:

$$Var(BCV(\hat{f})) = \frac{1}{k^2}\sum_{i=1}^{k} Var(M(\hat{f}^{-\kappa_k}, \kappa_k)) = \frac{Var(OOS(\hat{f}))}{k}.$$

As the BCV estimate has a smaller variance than that of the OOS procedure, for unbiased estimates the BCV procedure yields a more precise estimate of the generalization error (in the sense of Blum et al. (1999); Hastie et al. (2009)). We investigate this in our Monte Carlo experiments in Section 4.

However, there are some cases in which the use of the BCV procedure may not be straightforward, or not be advisable. In the BCV procedure, for all the folds but the first and the last one, the test set interrupts the training set, then for some forecasting models (e.g., exponential smoothing methods or models with a moving average part) it may be difficult to handle training sets that consist of two non-continuous parts. Nevertheless, this is not an issue in the broad class of (linear or non-linear) pure autoregressive models of fixed order, as in the embedded form of the series only the respective rows have to be removed before estimating the model.

Also, the use of the BCV procedure is not straightforward when only the forecasts, but not the forecasting models, are available, as it then may not be possible to generate forecasts for the different test sets. This may be the case when one evaluates the forecasting record of an international organization (e.g. IMF, EC or OECD).

Finally, the use of BCV may also not be recommended when changes at a certain point in time (structural breaks) are present in the data. In this respect, it may be counterproductive to use data before the break as it does not provide valuable information for future values of the series, both for model estimation and evaluation.

All in all, the use of BCV is beneficial in the following cases. First, the model allows for non-continuous training periods. Second, the forecaster controls the model building steps and can produce the forecasts. Finally, full use of the data can be made. In many applications, this is the case, especially when the performance of (linear or non-linear) autoregressive models for stationary data is to be evaluated.

3 Directional accuracy measures

Conventional measures of forecasting accuracy are based on the idea of a quadratic loss function in that larger errors carry proportionally greater weights than smaller ones. Such measures respect the view that forecast evaluation should concentrate on all large disturbances whether or not they are associated with directional errors which are of no special interest in and of themselves. However, several studies argue that incorrectly predicted directions are among the most serious errors a forecast can make (see Chung and Hong (2007); Kim et al. (2008); Solferino and Waldmann (2010), Blaskowitz and Herwartz (2009, 2011) among others). In this respect, this study applies some directional accuracy measures (Blaskowitz and Herwartz (2009, 2011)) for forecast evaluation.
Using the indicator function $I[\ldots]$, the realized and predicted directions R_t and P_t are given by:

$$R_t = I\left[(y_{t+h} - y_t) > 0\right],$$

$$P_t = I\left[(\hat{y}_{t+h} - y_t) > 0\right],$$

where y_t is the current value of the series, $\hat{y}_{t+h}$ is the value of the forecast, and y_{t+h} is the true value of the series at time $t + h$.

Using R_t and P_t, the directional error (DE) for h-step-ahead forecasts can be defined as follows:

$$\mathrm{DE}_t = I[R_t = P_t].$$

Using DE, a general framework for the directional accuracy (DA) can be obtained:

$$\mathrm{DA}_t = \begin{cases} a & \text{for} \quad \mathrm{DE}_t = 1 \\ b & \text{for} \quad \mathrm{DE}_t = 0 \end{cases}$$

In this framework, a correct prediction of the direction takes a value a, which can be interpreted as a reward, and an incorrect prediction takes a value b, a penalty. Based on the DA, several directional accuracy measures can be defined.
The mean directional accuracy (MDA) is defined straightforwardly as the mean of the DA:

$$\mathrm{MDA} = \mathrm{mean}(\mathrm{DA}_t).$$

This measure acquires well the degree up to which the predictor is able to correctly predict the direction of the forecast, and it is robust to outliers. It should be noted that the following holds:

$$\text{MDA} = (a - b)\,\text{mean}(\text{DE}_t) + b,$$

so that MDA is a linear transformation of the mean of the DE, depending on a and b (this linear relationship can be derived from a contingency table of sums of correct/incorrect upward/downward predictions). As MDA does not take into account the actual size of the change, it does not measure the economic value of the forecast (the predictor can be able to forecast the direction in cases of low volatility quite well, but it can fail when the volatility is high). Therefore, we use the directional forecast value (DV), which multiplies DA by the absolute value of the real changes, thus assessing better the actual benefit/loss of a correct/incorrect direction of the prediction.
The mean DV (MDV) is defined as:

$$\text{MDV} = \text{mean}(|y_{t+h} - y_t| \cdot \text{DA}_t).$$

In order to have a scale-free measure, the absolute value of the change can be divided by the current value of the series (Blaskowitz and Herwartz, 2011). Then, the mean directional forecast percentage value (MDPV) can be defined as follows:

$$\text{MDPV} = \text{mean}\left(\left|\frac{y_{t+h} - y_t}{y_t}\right| \cdot \text{DA}_t\right).$$

According to Blaskowitz and Herwartz (2011), common values for a and b are $(a, b) = (1, 0)$, or $(a, b) = (1, -1)$. In the first case, DA_t is identical to DE_t. In the second case, b is actually a penalty. In our study, we use $(a, b) = (1, -1)$ to consider the more general case where penalties are involved.

4 Monte Carlo experiment

In this section, we provide a Monte Carlo analysis. We consider univariate and multivariate experiments. In the univariate experiment, we generate series from a stable AR(3) process, while in the multivariate experiments the data is generated from bivariate and trivariate VAR(2) models, respectively. The design of the first two experiments (univariate case and bivariate VAR(2) model) is stochastic, in the sense that for every Monte Carlo trial the model parameters are generated randomly and, as a result, we obtain different data

for every trial. The third experiment (trivariate VAR(2) model) is designed in line with the empirical application.

As for the stochastic design, our experiments allow to explore larger regions of the parameter space. In all experiments, one-step-ahead predictions are considered. Series with lengths of 50, 70, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, and 600 values are used. For each length, 1000 Monte Carlo experiments are conducted. The experiments are performed with the R programming language (R Development Core Team, 2009) in the following way. Series are first generated and partitioned into a data set which is available to the forecaster, the *in-set*, and a set of data from the end of the series as unknown future, the *out-set*. We use 70 percent of the data as in-set, and the rest of the data as out-set. Then, the in-set is partitioned into training and test sets using the OOS and 5-fold BCV procedures (20 percent of the in-set are used as test set, so that the OOS evaluation coincides with the last fold of the BCV). Models are then built and values of the test sets are predicted to compute the directional accuracy measures (see Section 3) and the root mean squared forecast error (RMSFE). In addition to the OOS evaluation and the 5-fold BCV, we also perform OOS evaluation with rolling and recursive schemes. In the recursive scheme, for every point in the test set, the model is re-estimated using the training set and all the values from the test set prior to the current value to be forecasted. The rolling scheme is similar to the recursive scheme, but for every value that is added to the end of the training set, a value from the beginning of the training set is discarded.

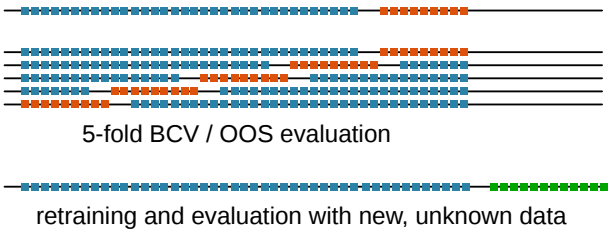

Figure 2: Illustration of the Monte Carlo experiments. The data are partitioned into an in-set, which is used for BCV and OOS evaluation, and an out-set (green), which is completely withheld. After model estimation and estimation of the directional accuracy measures, the models are estimated again using all available data in the in-set to forecast the unknown future (the out-set). This is a typical application scenario of forecasting. In our experiments, the directional accuracy measures are calculated on the out-set, and the error estimates given by BCV and OOS evaluation can be compared to the reference errors calculated on the out-set.

In this way, estimates are obtained only using the data of the in-set. Then, we build models using all data of the in-set, and predict the values of

the out-set and calculate the directional accuracy measures on the out-set (see also Figure 2). Thus, for each kind of model we obtain an error estimate using only the in-set data, and an error measure on future values of the series, the out-set data. This allows us to compare the in-set error estimates using BCV, OOS, rolling and the recursive scheme, and the out-set errors. To this end, we calculate the root mean squared error of the in-set estimates with respect to the out-set errors, and we call this measure the root mean squared predictive accuracy error (RMSPAE). It is defined as follows:

$$\text{RMSPAE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(M_i^{out-set} - M_i^{in-set})^2} \ ,$$

where n is the number of series, i.e., trials in the Monte Carlo simulation, and M is the accuracy measure in consideration, calculated for one model and one series. In the case of M^{in-set}, the BCV, OOS, rolling, or recursive procedure is used on the in-set, and in the case of $M^{out-set}$, the data of the in-set are used for training, and the data of the out-set are used for testing.

The RMSPAE is an appropriate measure to compare the performances of different evaluation procedures. Indeed, the RMSPAE does not assess the performance of a forecasting model, but it assesses the performance of an evaluation procedure (e.g., BCV) to predict the generalization error for a determined model, regardless of whether the model performs well or not.

4.1 Univariate case

Series are generated from a stable AR(3) process. Real-valued roots of the characteristic polynomial are chosen randomly from a uniform distribution in the interval $[-r_{max}, -1.1] \cup [1.1, r_{max}]$, with r_{max}=5.0. From these roots, the coefficients of the AR model are estimated (for a more detailed description of the procedure, see Bergmeir and Benítez (2012)). The first 100 time series observations are discarded to avoid possible initial value effects. The series are then normalized to zero mean and unit variance. As percentage measures such as the MAPE and the MDPV are heavily skewed when the series have values close to zero (see e.g. Hyndman and Koehler (2006); this is also confirmed in unreported preliminary experiments), for each series we subtract the overall minimum (calculated over all series) from all values to obtain a series of non-negative values, and then we increment all values by 1, to achieve a series which only contains values greater 1. In this way, new coefficients are estimated and a new series is generated for each iteration.

For forecasting purposes, we consider the data generating process, AR(3),

Table 1: Univariate results. OOS, 5-fold BCV, recursive, and rolling procedures.

	MDA	MDV	MDPV	RMSFE
RMSPAE 5-fold BCV				
AR(1)	0.1803	0.3576	0.0692	0.1595
AR(2)	0.1830	0.3579	0.0693	0.1457
AR(3)	0.1820	0.3593	0.0695	0.1481
RMSPAE OOS				
AR(1)	0.2697	0.5203	0.1013	0.2293
AR(2)	0.2775	0.5217	0.1014	0.2180
AR(3)	0.2756	0.5243	0.1018	0.2220
RMSPAE Recursive Scheme				
AR(1)	0.2738	0.5223	0.1016	0.2170
AR(2)	0.2775	0.5222	0.1017	0.2084
AR(3)	0.2804	0.5240	0.1019	0.2117
RMSPAE Rolling Scheme				
AR(1)	0.2687	0.5195	0.1012	0.2164
AR(2)	0.2748	0.5202	0.1012	0.2089
AR(3)	0.2780	0.5213	0.1014	0.2138

Notes: Series of length 100. The RMSPAE is calculated over 1000 trials.

and other two autoregressive processes, namely AR(1) and AR(2). Evaluation is performed using 5-fold BCV, OOS, rolling and recursive schemes.

Table 1 reports the RMSPAE results for the directional accuracy measures (see Section 3) and the RMSFE for BCV, OOS, rolling, and recursive schemes. A series length of 100 is considered in the table (to save space, results for other series lengths are not reported here; they are available upon request).

We clearly see that the values for BCV are consistently smaller than the respective values of OOS, rolling, and recursive evaluation. This occurs for all the models and different measures considered. As regards the RMSFE measure, the values for RMSPAE obtained with the BCV procedure are around 0.15 for all the models, whereas the other procedures provide values over 0.20. For the directional accuracy measures, the RMSPAE shows similar findings across the models. For example, when the model AR(3) is considered along the MDV measure, the BCV provides a value of 0.3593, where this value steps up to 0.5213, 0.5240 and 0.5243 for the other three schemes respectively. Among these schemes, similar results are found.

All in all, these results show that the measures calculated on the in-set using BCV estimate more precisely the out-set measures. Therefore, using

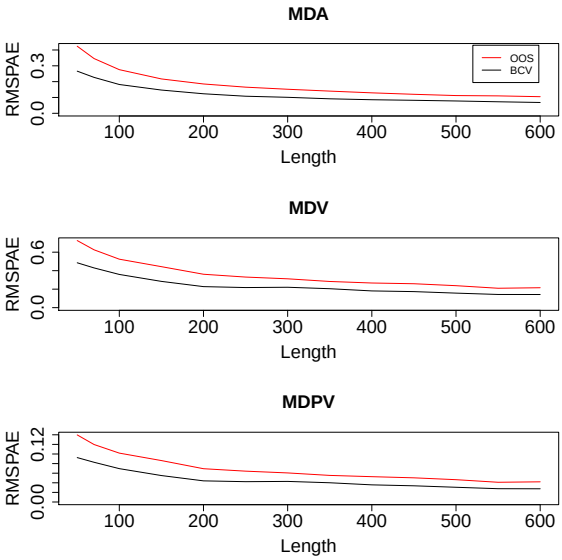

Figure 3: RMSPAE averaged over the Monte Carlo trials AR(3). Series of different lengths.

BCV, one is able to estimate the directional accuracy for a given method more precisely when predicting unknown future values of the series.

Another important result emerges from the experiment. Figure 3 shows the RMSPAEs for the series of all lengths when an AR(3) model is considered. The results indicate that in general the RMSPAE decreases with increasing length of the series, so that the directional accuracy is estimated more precisely if more data are available. Also, advantages of cross-validation are bigger if the series are shorter which is often the case in empirical applications.

4.2 Multivariate case

For the multivariate case, two studies are performed. The first study is in line with the univariate experiments discussed so far, while the second study is motivated by our application.

4.2.1 Multivariate experiment with stochastic design

The purpose of this multivariate Monte Carlo simulation study is to verify the robustness of the results in Section 4.1. The data generating process is a bivariate VAR(2) model. Series are generated in a similar way as in Section 4.1. Eigenvalues for the companion matrix of the VAR model are

generated, with an absolute value smaller than 1, in order to obtain a stable model (Lütkepohl, 2006). The companion matrix is generated from these eigenvalues by the procedure described by Boshnakov and Iqelan (2009). The covariance matrix is randomly chosen by generating an upper triangular matrix from a uniform distribution in the $[-1, 1]$ interval where the elements on the diagonal are set equal to 1. Therefore, a random symmetric matrix is built up. The random values for the VAR process are then drawn from a Gaussian distribution, and multiplied by the Cholesky form of the covariance matrix. As in the univariate experiment, the first 100 observations are discarded and the resulting series are normalized to have zero mean and unit variance. Then, the series are shifted to prevent problems with percentage measures (by incrementing each value by 1 and subtracting the overall minimum from the series).

In analogy to the application in Section 5, we use only the first component of the bivariate model for the evaluation. Along with the VAR(2) model, two other models are used for forecasting purposes, namely the bivariate VAR(1) and VAR(3) model.

Table 2: Multivariate results, using the stochastic design. OOS, 5-fold BCV, recursive, and rolling procedures.

	MDA	MDV	MDPV	RMSFE
RMSPAE 5-fold BCV				
VAR(1)	0.2203	0.1165	0.0212	0.1118
VAR(2)	0.2157	0.1264	0.0230	0.1219
VAR(3)	0.2212	0.1324	0.0241	0.1361
RMSPAE OOS				
VAR(1)	0.3220	0.1815	0.0340	0.1712
VAR(2)	0.3132	0.1922	0.0356	0.1812
VAR(3)	0.3299	0.2022	0.0374	0.2187
RMSPAE Recursive Scheme				
VAR(1)	0.3167	0.1782	0.0336	0.1563
VAR(2)	0.3091	0.1946	0.0361	0.1537
VAR(3)	0.3127	0.1981	0.0368	0.1662
RMSPAE Rolling Scheme				
VAR(1)	0.3218	0.1803	0.0339	0.1599
VAR(2)	0.3078	0.1917	0.0355	0.1566
VAR(3)	0.3128	0.1962	0.0361	0.1695

Notes: Series of length 100. The RMSPAE is calculated over 1000 trials.

In Table 2, the Monte Carlo results of RMSPAE for the directional ac-

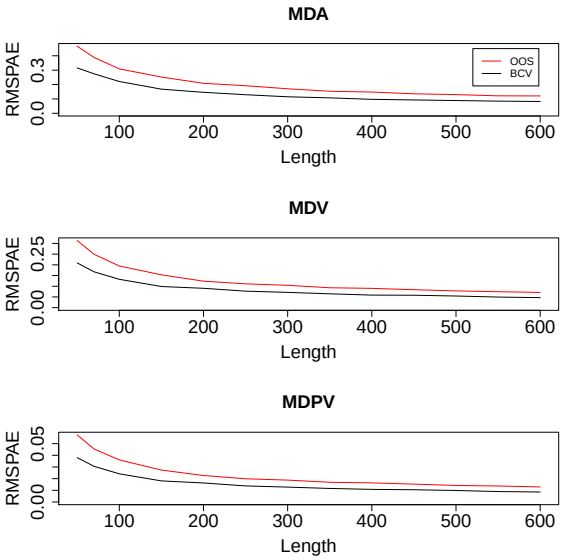

Figure 4: RMSPAE averaged over the Monte Carlo trials for VAR(2). Series of different lengths.

curacy measures and the RMSFE are reported, using BCV, OOS, rolling, and recursive procedures for series of length 100 (as in the univariate case, results for series of other lengths are not reported here to save space; they are available upon request). The results confirm those in the univariate case. From Table 2, it can be seen that the RMSPAE is consistently smaller using the BCV procedure.

As regards the RMSFE, the results show that for the VAR(2) model the value of this conventional accuracy measure is equal to 0.1219 for the BCV procedure and larger than 0.15 for the OOS procedures. With respect to the directional accuracy measures, we find that the values of MDV and MDPV are equal to 0.1264 and 0.0230 when using BCV, whereas these values are around 0.20 and 0.035 for the OOS schemes. Among OOS procedures, the results here confirm those in the univariate experiment.

Figure 4 shows the Monte Carlo results in terms of RMSPAE for series of all lengths when a VAR(2) model is considered. The findings observed in the univariate case are confirmed: the advantage of using the cross-validation scheme is preserved and it is also bigger with shorter series.

4.2.2 Multivariate experiment related to the empirical application

The stochastic design of the Monte Carlo experiments shown in Section 4.2.1 allows us to explore a broad class of different data generating processes. However, it may ignore potential local accuracy differences. Therefore, in this section we focus on a data generating process consistent with the data used for the empirical application. We follow the procedure in Costantini and Kunst (2011) and estimate a VAR(2) model using data from Section 5 for all the Monte-Carlo trials. We set the maximum lags to 8. The optimal number of lags selected by the AIC and BIC criteria are 2 and 1, respectively. We use a model with two lags as recommended model in macroeconomic systems, see Lütkepohl (2006). The estimated parameters of the model are reported in Appendix A. Apart from the different data generating process, the experiment is identical to those in Section 4.2.1. The results are shown in Table 3 and Figure 5. The BCV procedure shows a better performance in terms of RMSPAE than the OOS procedure for all the forecast measures. These results confirm those obtained in the previous experiments.

Table 3: Multivariate results, using a data generating process in line with the empirical application. OOS, 5-fold BCV, recursive, and rolling procedures.

	MDA	MDV	MDPV	RMSFE
RMSPAE 5-fold BCV				
VAR(1)	0.2222	0.0644	0.0035	0.0615
VAR(2)	0.2325	0.0697	0.0037	0.0695
VAR(3)	0.2314	0.0686	0.0037	0.0856
RMSPAE OOS				
VAR(1)	0.3429	0.0968	0.0051	0.0908
VAR(2)	0.3481	0.1033	0.0055	0.1037
VAR(3)	0.3549	0.1041	0.0055	0.1229
RMSPAE Recursive Scheme				
VAR(1)	0.3403	0.1003	0.0053	0.0769
VAR(2)	0.3454	0.1030	0.0055	0.0763
VAR(3)	0.3482	0.1049	0.0056	0.0805
RMSPAE Rolling Scheme				
VAR(1)	0.3447	0.1009	0.0054	0.0780
VAR(2)	0.3465	0.1058	0.0056	0.0770
VAR(3)	0.3481	0.1059	0.0056	0.0816

Notes: Series of length 100. The RMSPAE is calculated over 1000 trials.

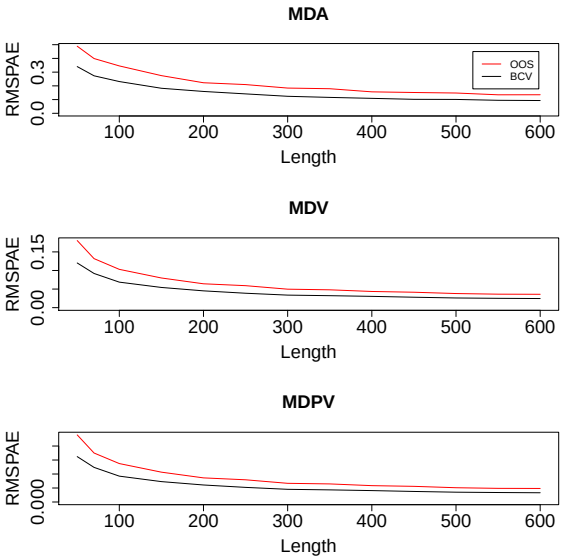

Figure 5: RMSPAE averaged over the Monte Carlo trials for a trivariate VAR(2). Series of different lengths.

5 Empirical application

In this section, we offer an application to UK interest rate as an example of the use of BCV in practice. While the findings of the Monte Carlo simulation show that in general the BCV performs better than the OOS, we now focus on the particularity of directional accuracy measures of binary output, which potentially leads to a loss of information. The main purpose of our study is neither to support or establish an economic theory nor to show the suitability of a particular method, but to investigate the usefulness of the cross-validation scheme along with directional accuracy measures. Therefore we consider simple linear models for forecasting the UK quarterly interest rate (for a forecasting exercise on UK interest data see also Barnett et al. (2012)).

In order to have a realistic setup with regard to the evaluation procedures, we do not perform a partition of the data into in-set and out-set, but just use BCV, OOS, rolling, and recursive evaluation in the way it would be used in an empirical application. Therefore, we provide directly the results of the directional accuracy measures.

The data set consists of quarterly annualized real GDP growth, quarterly annualized inflation rate and the three-month Treasury bill rate. The data is taken from the OECD Main Economic Indicators database and it covers

the period 1965:1-2011:1. The CPI data has been seasonally adjusted using Tramo seats. GDP growth is defined as 400 times the log difference of GDP and inflation is similarly defined using CPI. The interest rate is used without any change. The series are shown in Figure 6. All the series have been tested for stationarity using the DF-GLS unit root test of Elliott et al. (1996). The results show that the inflation and GDP growth rates are stationary at 5% level (the statistics are -3.102 and -5.129, respectively) while interest rate is stationary at 10% level (the statistics is -1.920).

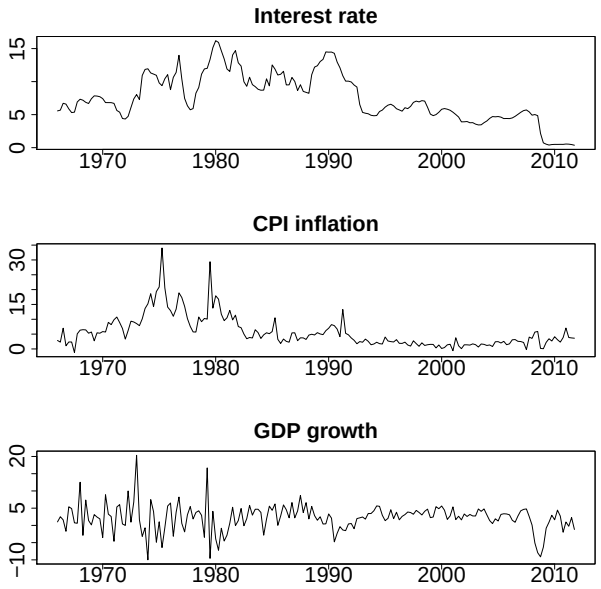

Figure 6: UK quarterly interest rate, CPI inflation rate, and GDP growth rate.

In the application, we consider a trivariate VAR model with interest rate, CPI inflation rate and GDP growth rate (VAR_3), a bivariate VAR model with interest rate and CPI inflation rate (VAR_{cpi}) and a bivariate VAR model with interest rate and GDP growth rate (VAR_{gdp}). All the VAR models are of order two. Furthermore, we use 5-fold BCV and for the OOS a period which coincides with the 5th fold. Therefore, the last 20% of the data are used for OOS evaluation.

Table 4 reports the results. It should be noted that the MDV measure yields negative values for the OOS procedure, and sometimes for the rolling and recursive schemes. This result may be due to the fact that the last part of the sample period of the series is fairly stable, so that forecasting of the direction is difficult and the models do not perform well. Furthermore,

Table 4: UK interest rate forecasting results. OOS, 5-fold BCV, recursive, and rolling procedures.

	MDA	MDV	MDPV	RMSFE
5-fold BCV				
VAR_3	0.1778	0.0819	0.0165	0.8990
VAR_{cpi}	0.2444	0.1122	0.0152	0.9209
VAR_{gdp}	0.1556	0.1013	0.0181	0.9009
OOS				
VAR_3	0.1667	-0.0048	0.0191	0.6146
VAR_{cpi}	0.1111	-0.0161	0.0036	0.6989
VAR_{gdp}	0.1667	-0.0048	0.0191	0.6147
Recursive				
VAR_3	0.2778	0.1521	0.0637	0.5271
VAR_{cpi}	0.1111	-0.0161	0.0036	0.6033
VAR_{gdp}	0.2778	0.1521	0.0637	0.5267
Rolling				
VAR_3	0.2778	0.1521	0.0637	0.5355
VAR_{cpi}	0.1111	-0.0161	0.0036	0.6077
VAR_{gdp}	0.2222	0.0024	0.0328	0.5352

Notes: The values for 5-fold BCV are averaged error measures over the 5 folds. VAR_3 includes interest, CPI inflation and GDP growth rates; VAR_{cpi} interest and CPI inflation rates; VAR_{gdp} interest and GDP growth rates. Note that the table does not report the RMSPAE as Tables 1 and 2, but it shows the values of the directional accuracy measures.

neither the OOS nor the recursive procedure are capable of distinguishing the performance of two models, VAR_3 and VAR_{gdp}, as the exactly same values for all the directional measures are found. For example, for the MDA, values of 0.1667 and 0.2778 are found for the OOS and recursive procedures, both for VAR_3 and VAR_{gdp}. A different situation is observed in case of the conventional RMSFE measure. For the recursive scheme, for example, the VAR_3 and VAR_{gdp} obtain different values, 0.5271 and 0.5267, respectively.

In contrast, the cross-validation procedure is able to distinguish the performance of the two models when all the directional accuracy measures are considered (e.g., for the MDA, values of 0.1778 and 0.1556 are found for the VAR_3 and VAR_{gdp}, respectively). It should be noted that the rolling scheme also performs well in this application, in the sense that it differentiates the performance of the models in terms of directional accuracy (e.g., for the MDA and the mentioned VAR models, the values are 0.2778 and 0.2222, respectively). However, as for the OOS and recursive scheme, the rolling scheme does not cope with the fact that the few information for evaluation may be

not enough to distinguish the performance of the models.

Furthermore, the results are examined in more detail in Figure 7. We focus on the BCV and OOS procedures (similar results to OOS are found for rolling and recursive schemes). It should be noticed that in the last fold of BCV, which is also used for OOS, all models yield identical results in terms of the directional accuracy, with the exception of the VAR_{cpi} which yields an incorrect directional forecast in one case (the other two models are able to predict the direction correctly). Using OOS, only the information from the 5th fold is used, and the VAR_3 and VAR_{gdp} models yield the same results on this fold. In contrast, BCV uses forecasts of all folds, so that it helps distinguish the forecasting performance of the models in terms of directional forecast accuracy. These results have important implications for macroeconomic applications where the amount of data available can be limited: the use of the blocked cross-validation is highly recommended for directional forecasts.

6 Conclusions

This paper investigates the usefulness of a predictor evaluation framework which combines a k-fold blocked cross-validation scheme with directional accuracy measures. The advantage of using a blocked cross-validation scheme over other procedures such as the standard out-of-sample procedure is that the blocked cross-validation allows one to obtain a more precise error estimate of the generalization error from the data as it uses all the available data both for training and testing.

In this paper we evaluate whether, and to what extent, the k-fold blocked cross-validation procedure may provide more precise results than the standard out-of-sample procedure even when dealing with directional forecast accuracy. To this end, a Monte Carlo analysis is performed using simple univariate and multivariate linear models. The results show that the blocked cross-validation is able to estimate the directional accuracy more precisely than the out-of-sample procedure when predicting unknown future values of the series. These results are obtained in both the univariate and multivariate design.

An empirical application is also carried out on forecasting UK interest rate data using three different simple VAR(2) models. The limited amount of available data, together with the use of directional accuracy measures, leads to identical realized average loss/success on some occasions when using the standard out-of-sample procedure. The blocked cross-validation is less likely to yield identical estimates for a given sample size, as it uses additional

information from other test sets. In particular, in our application the blocked cross-validation can distinguish the performance of the different forecasting models. This result has important implications for the macroeconomic applications where the amount of data is often limited: the use of the block cross-validation scheme is highly recommended when dealing with directional forecast evaluation.

Acknowledgements

This work was supported in part by the Spanish Ministry of Science and Innovation (MICINN) under Projects TIN2009-14575, TIN2011-28488, and P10-TIC-06858. This work was performed while C. Bergmeir holded a scholarship from the Spanish Ministry of Education (MEC) of the "Programa de Formación del Profesorado Universitario (FPU)", and he was visiting the Department of Economics, University of Vienna. Furthermore, we would like to thank Robert Kunst and Rob J Hyndman for useful advice.

Appendix A: Parameters of the VAR data generating process

The resulting trivariate VAR(2) model has the form:

$$Y_t = \mu + \sum_{j=1}^{2} \Phi_j Y_{t-j} + \varepsilon_t.$$

The parameters estimated from the data used in the empirical application are the following:

$$\mu = (0.102, 0.109, 2.959)',$$

$$\Phi_1 = \begin{pmatrix} 1.153 & 0.009 & 0.012 \\ 0.522 & 0.608 & 0.141 \\ 0.516 & -0.079 & 0.036 \end{pmatrix}$$

$$\Phi_2 = \begin{pmatrix} -0.200 & 0.009 & 0.047 \\ -0.415 & 0.220 & -0.107 \\ -0.571 & -0.035 & 0.110 \end{pmatrix}$$

All polynomial roots have absolute values smaller 1, so that the model is stable.

The variance-covariance matrix has the following form:

$$\Sigma = \begin{pmatrix} 0.882 & 0.530 & 0.319 \\ 0.530 & 8.100 & -2.134 \\ 0.319 & -2.134 & 14.145 \end{pmatrix}$$

References

C. Altavilla and P. De Grauwe. Forecasting and combining competing models of exchange rate determination. *Applied Economics*, 42:3455–3480, 2010.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

A. Barnett, H. Mumtaz, and K. Theodoridis. Forecasting UK GDP growth, inflation and interest rates under structural change: a comparison of models with time-varying parameters. Working Paper 450, Bank of England, 2012.

C. Bergmeir and J.M. Benítez. Forecaster performance evaluation with cross-validation and variants. In *International Conference on Intelligent Systems Design and Applications, ISDA*, pages 849–854, 2011.

C. Bergmeir and J.M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

O. Blaskowitz and H. Herwartz. Adaptive forecasting of the EURIBOR swap term structure. *Journal of Forecasting*, 28(7):575–594, 2009.

O. Blaskowitz and H. Herwartz. On economic evaluation of directional forecasts. *International Journal of Forecasting*, 27(4):1058–1065, 2011.

O. Blaskowitz and H. Herwartz. Testing directional forecast value in the presence of serial correlation. *International Journal of Forecasting*, 30(1): 30–42, 2014.

A. Blum, A. Kalai, and J. Langford. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proceedings of the International Conference on Computational Learning Theory*, pages 203–208, 1999.

S. Borra and A. Di Ciaccio. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Computational Statistics & Data Analysis*, 54(12):2976–2989, 2010.

G.N. Boshnakov and B.M. Iqelan. Generation of time series models with given spectral properties. *Journal of Time Series Analysis*, 30(3):349–368, 2009.

J. Chung and Y. Hong. Model-free evaluation of directional predictability in foreign exchange markets. *Journal of Applied Econometrics*, 22:855–889, 2007.

M. Costantini and R. Kunst. Combining forecasts based on multiple encompassing tests in a macroeconomic core system. *Journal of Forecasting*, 30: 579–596, 2011.

M. Costantini and C. Pappalardo. A hierarchical procedure for the combination of forecasts. *International Journal of Forecasting*, 26(4):725–743, 2010.

G. Elliott, T.J. Rothenberg, and J.H. Stock. Efficient tests for an autoregressive unit root. *Econometrica*, 64:813–836, 1996.

T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning*. Springer, 2009. ISBN 9780387848846.

R.J. Hyndman and A.B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

J.A. Khan, S. Van Aelst, and R.H. Zamar. Fast robust estimation of prediction error based on resampling. *Computational Statistics & Data Analysis*, 54(12):3121–3130, 2010.

T.-H. Kim, P. Mizen, and T. Chevapatrakul. Forecasting changes in UK interest rates. *Journal of Forecasting*, 27(1):53–74, 2008.

Q. Kong. Predictable movements in yen/dm exchange rates. IMF working paper 143, 2000.

R. Kunst. Cross validation of prediction models for seasonal time series by parametric bootstrapping. *Austrian Journal of Statistics*, 37:271–284, 2008.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2006. ISBN 9783540262398.

A.D.R. McQuarrie and C.-L. Tsai. *Regression and time series model selection*. World Scientific Publishing, 1998.

C. Milas and R. Naraidoo. Financial conditions and nonlinearities in the European Central Bank (ECB) reaction function: In-sample and out-of-sample assessment. *Computational Statistics & Data Analysis*, 56(1):173–189, 2012.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

J. Racine. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99(1):39–61, 2000.

N. Solferino and R. Waldmann. Predicting the signs of forecast errors. *Journal of Forecasting*, 29(5):476–485, 2010.

S. Sosvilla-Rivero and E. García. Forecasting the dollar/euro exchange rate: Are international parities useful? *Journal of Forecasting*, 24:369–377, 2005.

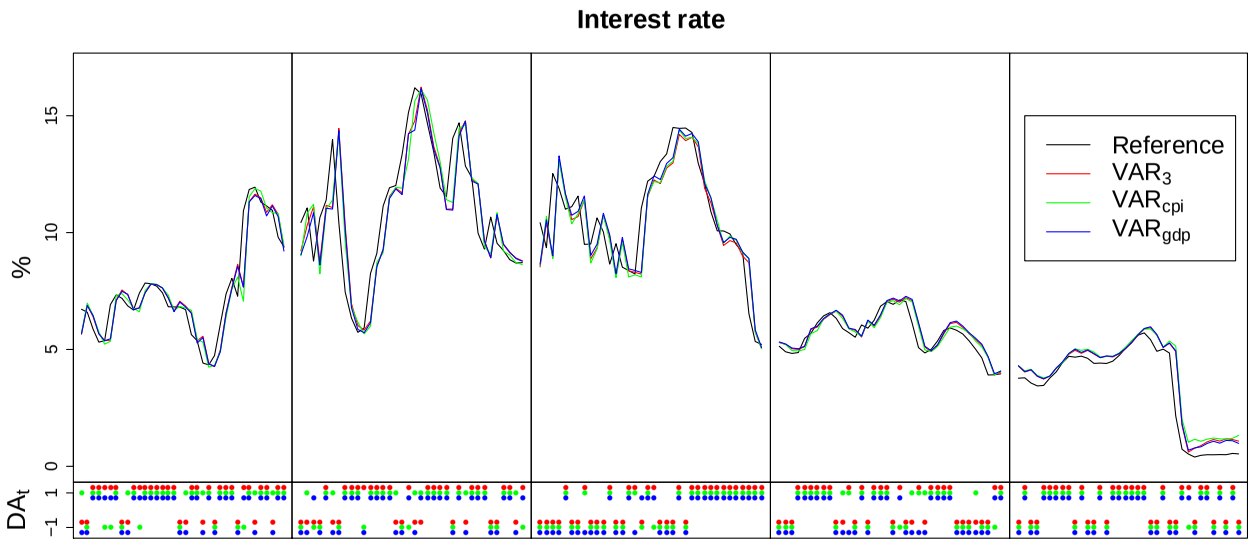

Figure 7: UK quarterly interest rate forecasts. Forecasts and directional accuracy of the three different VAR models are shown. The directional accuracy as defined in Section 3 only takes the values -1 and 1. As the models obtain very similar results in terms of forecasts, it may be difficult to distinguish them with respect to their directional accuracy performance.