Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
International Journal of Forecasting 21 (2005) 473–489
international journal of forecasting
www.elsevier.com/locate/ijforecast

Performance evaluation of judgemental directional exchange rate predictions☆

Andrew C. Pollock[a,*], Alex Macaulay[a,1], Mary E. Thomson[b,2], Dilek Önkal[c,3]

[a]*Division of Mathematics, School of Computing and Mathematical Sciences, Glasgow Caledonian University, Cowcaddens Road, Glasgow G4 0BA, UK*
[b]*Department of Psychology, Glasgow Caledonian University, Cowcaddens Road, Glasgow G4 0BA, UK*
[c]*Faculty of Business Administration, Bilkent University, 06800 Bilkent, Ankara, Turkey*

Abstract

A procedure is proposed for examining different aspects of performance for judgemental directional probability predictions of exchange rate movements. In particular, a range of new predictive performance measures is identified to highlight specific expressions of strengths and weaknesses in judgemental directional forecasts. Proposed performance qualifiers extend the existing accuracy measures, enabling detailed comparisons of probability forecasts with ex-post empirical probabilities that are derived from changes in the logarithms of the series. This provides a multi-faceted evaluation that is straightforward for practitioners to implement, while affording the flexibility of being used in situations where the time intervals between the predictions have variable lengths. The proposed procedure is illustrated via an application to a set of directional probability exchange rate forecasts for the US Dollar/Swiss Franc from 23/7/96 to 7/12/99 and the findings are discussed.

Keywords: Accuracy; Exchange rate; Forecasting; Judgement; Probability

1. Introduction

Exchange rate movements are primarily affected by expectational elements arising from market sentiment. These manifest themselves in optimism, pessimism and varying degrees of uncertainty in the minds of market participants. Analysts' judgements of perceived market sentiment, as well as their responses to the uncertainties attributable to political and economic events, play fundamental roles in their forecasts of currency movements (Larson & Madura, 2001). In particular, there exist significant "individual

☆ Earlier versions of this paper were presented at the 22nd International Symposium on Forecasting, Dublin, Ireland (June 2002), and the 30th Meeting of the EURO Working Group on Financial Modelling, Capri, Italy (May 2002). The authors are grateful to the participants, whose comments significantly contributed to the present paper.

* Corresponding author. Tel.: +44 141 331 3613; fax: +44 141 331 3608.
E-mail addresses: a.c.pollock@gcal.ac.uk (A.C. Pollock), abma@gcal.ac.uk (A. Macaulay), mwi@gcal.ac.uk (M.E. Thomson), onkal@bilkent.edu.tr (D. Önkal).
[1] Tel.: +44 141 331 3052; fax: +44 141 331 3608.
[2] Tel.: +44 141 331 3899; fax: +44 141 331 3636.
[3] Tel.: +90 312 290 1596; fax: +90 312 266 4958.

0169-2070/$ - see front matter © 2005 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2004.12.006

effects" in financial agents' expectation formation (Ito, 1990), due at least in part to private versus shared information/beliefs (Wang, 2001). Biased forecasts may be observed as a result of such behavioural dynamics (Daniel, Hirshleifer, & Teoh, 2002). In domains where the financial consequences of forecast errors are critical, profiling the predictive strengths and weaknesses of forecasters gains a special significance. This paper sets out a procedure that enables a detailed analysis of forecasting performance of directional probability predictions of foreign exchange rate movements.

The efficient use of judgement and extensive evaluation of predictive performance in financial domains require not only a prediction of the movement (direction or magnitude of change) but also a probability assessment (the probability of a rise or fall, or a confidence band) associated with the prediction. A probabilistic approach is essential for detecting the presence of biases. Biases that cause overconfidence and overreaction cannot adequately be examined if there exists no information on the analyst's assessment of the uncertainty that surrounds the prediction (Wilkie, Tuohy, & Pollock, 1993). Hence, procedures for examining probabilistic prediction accuracy are of critical importance from both the perspective of forecast users as well as from the viewpoint of the analysts preparing the forecasts (Önkal-Atay, Thomson, & Pollock, 2002).

Directional probability forecasts provide effective tools for analysts in their efforts to convey information to clients that incorporate assessments of uncertainty. While magnitude predictions are more appropriate for situations involving hedging decisions, directional predictions are more appropriate for speculative decisions when taking long or short positions is the key issue (Moosa, 2000). Furthermore, directional predictions play a fundamental role in the identification and timing of buy and sell actions in trading and investment decision support systems. Technical analysis, widely used by financial practitioners, provides forecasts that are essentially directional in nature (Murphy, 1999). The inclusion of probabilities along with directional predictions presents a powerful decision support tool that enables assessments of the confidence placed in the analyst's forecasts.

The value of analysts' directional probability predictions of exchange rate movements depends, however, on their accuracy. When probability forecasts are supplied, it is important for the decision makers to assess not only the overall quality of these predictions, but also the specific aspects of performance that highlight particular strengths and weaknesses (Wilkie et al., 1993). These also allow analysts to recognise their own limitations, permitting them to rectify specific biases and to use their expertise more effectively. It is therefore extremely important to have systems in place that provide this valuable feedback.

This paper extends the procedure previously developed by Wilkie and Pollock (1996) to allow detailed evaluation of the performance of probabilistic directional forecasts that are not constrained to fixed intervals between predictions. In currency management practice, it is common for analysts to advise their clients about reassessing their currency holdings in light of market developments, thus accentuating the significant need for proper procedures for evaluating rolling forecasts that are made at intervals that do not have fixed lengths. That is, the analyst can provide a forecast for specific fixed horizons, but events in the market can lead to the intervals between the forecast revision dates that are not fixed in length. This practice is consistent with motives to maximise profit opportunities where forecasts are used as a basis for action decisions regarding currency holdings. Currency positions can be changed very quickly; hence it is unrealistic to restrict actions for adjusting the composition of currency holdings to set dates.

Extending the performance measures of Wilkie and Pollock (1996), the current paper presents new dimensionless measures that are more straightforward for analysts to use and that facilitate comparisons over different periods and across differing exchange rates. Furthermore, the proposed procedure enables a refined derivation of the standard deviation assessment used to obtain the empirical directional probabilities (which are compared with the probability forecasts to examine performance). Previous work has used past movements of the exchange rate to obtain the empirical standard deviation. In practice, however, it is desirable to have standard deviation assessments obtained from the predictive horizon used to calculate the actual change. As a result, the derived empirical directional probabilities will only be dependent on the behaviour of the series in the prediction period and not on any estimates obtained prior to the prediction

period. This has important practical implications, since the analyst making the predictions is now prevented from manipulating the procedure to alter performance.

To illustrate the application of the procedure, an empirical analysis is applied to directional probability predictions of the US Dollar/Swiss Franc (USD/CHF) from 23/7/96 to 7/12/99. The procedure used to evaluate these predictions is based on the assumption that daily changes in the logarithms of the exchange rate follow a normal distribution with time-varying means and standard deviations. Over short horizons (e.g., 30 days) the means and standard deviations can, however, be considered to be approximately constant.

The remainder of the paper is structured as follows: Section 2 sets out methodological issues relating to the analysis of currency predictions. Section 3 examines the assumptions made about the statistical distribution of changes in logarithms of the exchange rate and the implications on obtaining empirical probabilities, as well as on the formation of probabilistic currency predictions. Section 4 presents details of the statistical performance measures. Section 5 offers an illustration of the application of the proposed procedure, while Section 6 provides some concluding remarks.

2. Methodological issues relating to the analysis of exchange rate predictions

The proposed procedure requires converting directional exchange rate predictions to a form amenable to performance analysis, as well as allowing for the ex-post adjustment of forecasts to achieve consistency with the intervals from which the empirical probabilities are derived. These issues are discussed below.

2.1. Converting recommendations to a form appropriate for performance analysis

To undertake the analysis of probabilistic predictions, it is first necessary to convert the information to an appropriate form. The directional currency-forecasting task can be viewed as a simple two-alternative (i.e., rise or fall) situation. Studies of probability judgement have tended to use a half-range method (Ronis & Yates, 1987), which requires predictions to be expressed by two components. Firstly, a choice is made between two alternative directions: rise or fall. Secondly, the level of confidence is indicated by assigning a probability (in the range 0.5 to 1.0) to the chosen direction. An assigned probability of 0.5 implies a no-change prediction, whereas a probability of 1.0 implies total confidence in the predicted direction's occurrence.

In practice, probability forecasts may also be made using the full-range method. That is, the probability for a designated direction (e.g., rise) in the exchange rate would be given on a scale from zero to unity. Accordingly, values below 0.5 would indicate a predicted change in the other direction (e.g., fall) in the rate and values above 0.5 would indicate a predicted change in the designated direction (e.g., rise) for the rate. This is consistent with the use of predictions made by analysts that are grouped into a number of categories. For example, an analyst could set bands associated with probability predictions for action decisions on the USD/CHF rate as follows: (i) 0 to 0.2, buy CHF assets and sell USD assets; (ii) 0.21 to 0.4, hold existing CHF assets but reduce holdings of USD assets; (iii) 0.41 to 0.59, attempt to balance holdings of CHF and USD assets; (iv) 0.6 to 0.79, hold USD assets and reduce holdings of CHF assets; and (v) 0.8 to 1, buy USD assets and sell CHF assets.

It is easy to convert full-range probability statements to half-range probability statements. A full-range probability above 0.5 would assign a half-range probability equal to the full-range probability with the direction assigned as a rise. A full-range probability below 0.5 would assign a half-range probability equal to unity less the full-range probability with the direction assigned as a fall. For example, if a full-range probability prediction of 0.73 is made, then the half-range probability would be 0.73 when a rise is predicted. If a full-range probability of 0.24 is made, then the half-range probability would be 0.76 with a fall predicted. The full-range 0.5 probability, i.e., no change prediction, could be arbitrarily assigned as a rise or fall with a half-range probability equal to 0.5.

2.2. Ex-post adjustment of predictions

Before examining the formation and evaluation of probabilistic currency predictions in a practical context, we need to address a critical issue regarding the

continuity of forecast modifications or "rolling forecasts." Specifically, analysts often make forecasts that can be revised before the end of the initially specified horizon or extended beyond the horizon. The evaluation of a forecaster's performance, therefore, needs to be made in a way that relates to the interval between the predictions rather than the fixed prediction intervals relating to the initial forecast. Therefore, the initial forecast horizon need not be the same as the interval between the dates when predictions are made. This is realistic given the volatility and liquidity of financial markets. Such news-initiated rolling forecasts reflect the changing expectations and are 'rational' in that they enable asset holdings in different currencies to be rapidly adjusted in the light of new information so that profits can be increased and losses reduced. For example, an analyst basing his recommendation to a client for a specified time interval of say 30 days may find that technical analysis indicators show a change in market conditions 20 days into the 30 day horizon. The analyst could then update his recommendation at 20 days rather than wait for the 30 days to elapse so that his client can take appropriate action immediately. The analytical procedure used to examine performance should, therefore, allow for the possibility of evaluating predictions over flexible horizons that may be different from the original predictive horizon.

The procedure proposed by the current study explicitly addresses this issue by using an adjusted empirical probability (under the normal distribution assumption) that is based on the prediction of a stable ratio of the mean to the standard deviation of the daily changes in logarithms of the exchange rate. Probability predictions, although originally set for a specific horizon length, are adjusted to the same horizon length used to compute the empirical probabilities. In this way, the subjective mean (μ), subjective standard deviation (σ) and subjective probability (α) for a specific prediction period can be adjusted so that they can be directly compared with the empirical mean, standard deviation and probability obtained from the series over the interval from when the forecast was made to when it was updated. The procedure used to adjust the subjective probabilities can be explained as follows: given the subjective probability (α) for a predictive horizon of n days via the cumulative distribution function of the standard normal (Φ), $\alpha = \Phi\{\sqrt{n}(\mu/\sigma)\}$, the adjusted subjective probability (α^*) for a period n^* days, with $n^* \neq n$, is given by $\alpha^* = \Phi\{\sqrt{n^*}(\mu/\sigma)\}$. Hence, given the subjective probability, α, for n days, the ratio, μ/σ, can be directly obtained from the inverse cumulative distribution function. This can be used to obtain the adjusted subjective probability, α^*, using n^*.

The procedure can be explained with reference to the following example. Consider an analyst who makes a subjective prediction, α, of 0.81 for a 30-day predictive horizon, n. The inverse cumulative distribution function of the standard normal gives a value of 0.878 {i.e., $\sqrt{n}(\mu/\sigma)=0.878$}. For $n=30$, $(\mu/\sigma)=0.878/\sqrt{30}=0.160$. For the adjusted horizon, n^*, of 15 days the cumulative distribution function of the standard normal gives an adjusted probability, α^*, of $\Phi\{\sqrt{15}(0.160)\}=\Phi\{0.621\}=0.732$.

3. The distribution of daily exchange rate changes and its implications

In applying accuracy analysis to currency series, it is necessary to derive empirical probabilities for daily exchange rate changes from the actual series. The procedure is summarised below.

3.1. The assumption of normally distributed movements of the logarithms of the exchange rate

It is desirable to derive the empirical probabilities based on first differences of the logarithms of the actual exchange rate. The transformation of actual rates to logarithmic values takes into account the fact that changes in the exchange rates are likely to be dependent on the level of the rate. That is, large changes tend to occur when the actual exchange rate is at high levels and small changes tend to occur at low levels of the rate. The use of first differences stems from the view that, in general, currency series are not stationary: the variance and autocovariance functions depend on time. In particular, the variance tends to increase over time and first order serial correlation is exhibited with a value close to unity. In other words, the series tend to follow what is described by Nelson and Plosser (1982) as a difference–stationary process. Evidence suggests that trends in exchange rate series tend to be associated with high

order positive serial correlation. Exchange rate series can, however, be made stationary via simple transformations. In particular, taking first differences of the logarithms of a difference–stationary series with a linear trend simultaneously takes out the effect of the trend and the first order serial correlation of unity, resulting in a differenced series with constant drift and zero first order serial correlation.

Given the difference–stationary form of currency series, it is therefore appropriate to examine the distribution of these daily changes in logarithms of the series. There have been a number of studies examining the statistical aspects of daily exchange rate movements (Boothe & Glassman, 1987; Coppes, 1995; Corporale, Hassapis, & Pittis, 1998; Corporale & Pittis, 1996; Hsieh, 1988; Rogalski & Vinso, 1978; Westerfield, 1977). This work has reported that the changes are symmetric but with fatter tails than the normal distribution. However, these studies have generally been based on horizons greater than 1 year. The normal distribution, on the other hand, is found to provide an appropriate approximation for the behaviour of daily changes in the logarithms for floating exchange rates from developed economies, if allowance is made for time-varying means and standard deviations (Friedman & Vandersteel, 1982; Zhou, 1996).

The departures from the normal distribution illustrated in studies using longer horizons can often be attributed to psychological factors influencing market participants—their optimism, pessimism, and uncertainty. These expectations are aggregated to form the market sentiment that prevails in a particular period (Tvede, 1990). The bullish and bearish sentiments in the market manifest themselves in a trend (a non-zero drift) which financial agents, whether fundamentalists or technicians, attempt to identify. Depending on contextual contingencies, however, market sentiment may change and a bull market may become a bear market and vice versa. In short, the parameters of the distribution may change over time. Primary trends may be viewed as lasting for more than 1 year and are perceived as reflecting the underlying sentiment of the market. They are, therefore, associated with a relatively stable distribution over time. On the other hand, secondary trends are much shorter term (i.e., 1 to 3 months) and basically mirror corrective actions of the financial players. For example, market participants may feel that short-term excessive bullish sentiment regarding a particular currency has been too strong in that the mean change has been excessively large, hence, they may review their positions. Such short-term sentiment changes may result in a lower mean exchange rate change or even a negative mean reflecting a short-term reversal. Secondary trends can, therefore, cause the location parameter of the daily distribution to change in relatively short periods. In addition, the market is also likely to be influenced by periods of stability and instability that are associated with collective uncertainty in the minds of the market participants, for instance on whether a primary trend is likely to continue or reverse. This can cause variability in the dispersion parameter of daily exchange rate changes over relatively short periods. Consequently, a normal distribution appropriate for daily changes in (logarithms of) the exchange rate is likely to be characterised by a distribution that exhibits frequent shifts in the location and dispersion parameters. Furthermore, as the parameters are inherently related to market sentiment (optimism, pessimism, and uncertainty), their behaviour is not likely to be captured by standard statistical techniques. In using the normality assumption for daily data, in practice, it is therefore more appropriate to use shorter horizons (e.g., less than 50 days) than longer horizons. Hence, the assumption of normality is, in general, approximately satisfied for short horizons for daily changes in the logarithms of exchange rates.

3.2. Obtaining the empirical probabilities

Empirical directional probabilities (obtained at the end of the adjusted prediction period) are used to examine various dimensions of accuracy of the probability forecasts (made at the beginning of the prediction period). The role of the empirical probabilities is, therefore, purely to evaluate the predictions and not to give an alternative statistical model to provide forecasts with which the original predictions can be compared. The empirical probabilities are then used in the performance analysis procedure set out in Section 4 to evaluate the predictive accuracy of the forecasts. Although this is not a concern of the current study, it would be possible to use the same procedure to evaluate the performance of statistical models and

compare them directly with the original predictions. Studies along these lines, using actual exchange rate series that compare judgemental predictions with statistical models, have been previously undertaken (Pollock & Wilkie, 1992; Thomson, Pollock, Henriksen, & Macaulay, 2004).

The dates when subjective probability predictions are made are used as boundaries to divide the whole period into a number of sub-periods. A sub-period is defined as the period elapsing between the first trading day after the prediction is made to the day that the prediction is updated. Sub-periods can, therefore, have differing lengths. It is, then, necessary to obtain empirical probabilities for the exchange rate changes in a form that is consistent with the method used to give subjective probability predictions for the sub-periods. To do this, estimates of the mean and standard deviation of the distribution of exchange rate changes can be obtained, ex-post, for each of the sub-periods. These mean and standard deviation estimates, under the assumption that daily changes follow independent normal distributions, can then be used to obtain empirical probabilities (EPs) for the sub-periods. The procedure used to obtain these empirical probabilities for the full-range method is summarised below.

(1) For day i, i=1, 2,. . ., n_j, within sub-period j of length n_j, let $\Delta x_{i,j}=x_{i,j}-x_{i-1,j}$ denote the change in the logarithm of the exchange rate. The mean of the daily changes, m_j, is then obtained.
(2) The standard deviation of the daily changes, s_j, is calculated.
(3) The quantity $t_j=\sqrt{n_j}\,(m_j/s_j)$ is obtained.
(4) The cumulative probability $F(t_j)=P(t\leq t_j)$ is calculated, where t has Student's t distribution with n_j-1 degrees of freedom. This quantity gives the empirical probability of a rise in the exchange rate between the beginning and end of the sub-period. Values greater than 0.5 indicate a predicted rise in the rate and values below 0.5 indicate a predicted fall in the rate.

To illustrate this procedure and the calculation of EPs, suppose that the USD/CHF exchange rate moves from an initial value of 1.60 in Day 0 to a value of 1.65 in Day 5 as given in Table 1.

The first row gives the day number and the second row the exchange rate. The third row gives the logarithms to base 10 of the exchange rate. The fourth row gives the first differences in the logarithms of the rate. It is this last row that provides the basic input data to derive the EPs.

The four stages used to derive the EPs for this series are as follows:

(1) Calculate the mean, m=0.00267.
(2) Calculate the standard deviation, s=0.00506.
(3) Obtain the t value, $t=\sqrt{5}\,(0.00267/0.00506)=1.182$.
(4) Obtain the cumulative probability, $F(1.182)=P(t<1.182)=0.849$,

using Student's t distribution with $k-1=4$ degrees of freedom.

The EP is thus 0.849, corresponding to a rise in the exchange rate.

Normality was examined by using the Lilliefors (1967) and Jarque and Bera (1980) tests. The Lilliefors test was used, in addition to the Jarque–Bera test, as it is often more appropriate when sample lengths are relatively short, since more powerful tests that rely on third and fourth moments are likely to be unstable (Harvey, 1993). To examine the assumption of no serial correlation of successive daily changes, Bartlett's (1946) test of serial independence is applied.

3.3. Implications of normality on the formation of probabilistic currency predictions

It is argued that effective judgemental prediction requires the consideration of the underlying probability distribution on which the series are perceived to be formed (Keren, 1991). Although Keren concedes

Table 1
Calculation of changes in the logarithms of the exchange rate

Day no. (t=0, 1, 2, 3, 4, 5)	0	1	2	3	4	5
Ex. rate (X_t)	1.60	1.61	1.59	1.62	1.64	1.65
Log. ex. rate (x_t)	0.20412	0.20683	0.20140	0.20952	0.21484	0.21748
Change log. ex. rate (Δx_t)		0.00271	−0.00543	0.00812	0.00532	0.00264

that there is no way of determining what a person making a prediction is actually assuming in terms of a particular distribution, he strongly suggests that evaluators should attempt to specify a distribution or, at least, be encouraged to think in that way. In accordance with the earlier discussions on normality, it may be concluded that it is desirable for judgemental directional predictions to be based on the assumption of normally distributed currency movements (Wilkie & Pollock, 1996).

4. Procedures for the statistical analysis of the probability predictions

Prior to the evaluation of the probability predictions, a number of adjustments are undertaken so that an effective performance analysis can be conducted. Following the adjustments, an evaluation of the probability forecasts is made using the statistical procedures (detailed below) designed to identify diverse aspects of performance.

4.1. Preliminary adjustments

Two preliminary adjustments are made to the data before the application of performance analysis. Firstly, weighting is necessary to take into account the effect of the varying sub-period length. For example, adjusted predictions evaluated over 30 days are given a weighting that is twice the weight used to evaluate the adjusted predictions over 15 days. Secondly, on the basis of technical correctness, it is appropriate to omit weekdays when the markets are closed (usually 8 days a year in the case of the London market) from the analysis.

4.2. Outcome indices

The proportion of correct directional forecasts is a commonly used measure of directional predictive performance. For a sub-period, j, of length n_j days, the simple outcome index, d_j, takes values of 1 or 0 depending on whether or not the predicted direction is correct. For a set of directional forecasts the proportion correct, $M(d)$, is the number of times the correct directional response is made (taking the different lengths of the sub-periods into account), divided by the total number of days over the whole period (i.e., $n = \sum_j n_j$). This is given in Eq. (1):

$$M(d) = \frac{1}{n} \sum_j n_j d_j \tag{1}$$

The simple outcome index (d_j) is refined to produce a weighted outcome index (c_j^*) that, in addition, takes into account the relative movement of the series. Like d_j, c_j^* has a maximum possible value of unity and a minimum possible value of zero, but unlike d_j, c_j* can take any value between these two extremes. Formally, c_j* is defined in Eq. (2):

$$c_j^* = 0.5 + p_j^* \tag{2}$$

where p_j^* is a weight that is related to the population mean and standard deviation of the daily changes in the logarithms of the exchange rate over sub-period j and takes a value between -0.5 and 0.5. The weighted outcome index c_j^* depends, therefore, on both the empirical full-range population probability between the beginning and end of the adjusted sub-period and on whether or not the predicted direction is correct.

To obtain p_j^*, it is assumed that daily changes in the logarithms of the exchange rate in sub-period j follow independent normal distributions with mean μ_j and standard deviation σ_j. The probability, q_j^*, that the sum of daily changes in the logarithms of the exchange rate over sub-period j is positive is given by Eq. (3):

$$q_j^* = \Phi\{\sqrt{n_j}(\mu_j/\sigma_j)\} \tag{3}$$

where Φ is the cumulative distribution function of the standard normal.

The quantity p_j* is defined by Eq. (4):

$$p_j^* = (2d_j - 1)\left|q_j^* - 0.5\right| \tag{4}$$

In practice the mean and standard deviation parameters (μ_j and σ_j) would be unknown, and hence estimates (m_j and s_j) of the mean and standard deviation of the daily changes for sub-period j need to be calculated. The empirical mean and standard deviation (m_j and s_j) are used in place of the unknown parameters (μ_j and σ_j) in Eqs. (2), (3) and (4) to give estimates of population values p_j^*,q_j^* and c_j^* which are

denoted p_j, q_j and c_j respectively. These estimates are defined in Eqs. (5), (6) and (7):

$$c_j = 0.5 + p_j \tag{5}$$

$$q_j = F\{\sqrt{n_j}(m_j/s_j)\} \tag{6}$$

$$p_j = (2d_j - 1)|q_j - 0.5| \tag{7}$$

where F denotes the cumulative distribution function of the t distribution with n_j–1 degrees of freedom. For example, suppose that over a specific sub-period, j, with a length of 25 days (n_j), the changes in the logarithms of the currency series gave a mean of −0.0004 (m_j) and standard deviation of 0.0025 (s_j). The empirical probability (q_j) would be the cumulative distribution of the t distribution with 24 degrees of freedom. That is, q_j=$F(\sqrt{25}\ (-0.0004/0.0025))$=$F(-0.80)$=0.2158. As this value is below 0.5, a fall occurred over the period with an empirical probability of 0.2158. If a fall was correctly predicted (d_j=1) then, p_j=0.2842; hence the weighted outcome index, c_j=0.7842. If, on the other hand, a rise was incorrectly predicted, p_j=−0.2842 and c_j=0.2158.

The quantity 0.5+|p_j| reflects the relative magnitude of a movement in the currency series over sub-period j. The sign of p_j reflects whether the forecasted direction is correct or incorrect. If the correct direction is predicted, p_j is positive and c_j is greater than 0.5. If the incorrect direction is predicted, p_j is negative and c_j is less than 0.5. In the extreme case where there is only a very small change in the series (exchange rate quotations used in this study were USD/CHF middle closing rates specified to five significant figures such that a zero change was highly unlikely), c_j takes a value very close to 0.5 (whether or not the correct direction is predicted). In the other extreme case where there is an exceptionally large change in the exchange rate, c_j takes a value close to zero when the incorrect direction is predicted and a value close to unity when the correct direction is predicted. Therefore, c_j can take any value between zero and unity and can be viewed as a continuous variable. The empirical weighted outcome index c_j is similar to that used by Wilkie and Pollock (1996), but in the current study the definition is modified to take into account the variable lengths of the sub-periods to allow c_j to be directly obtained from the empirical full-range probability q_j.

Extreme values of c_j, for example, 0.975 or more or 0.025 or less, can be viewed as particularly important. As c_j is derived from q_j, which is an empirical probability using the normal distribution assumption, values of c_j can be considered indicative of a change in the exchange rate over the sub-period that is significantly different from zero at the 5% level of significance. A value of 0.975 or more reflects the fact that the correct directional prediction was made and a value of 0.025 or less reflects the fact that an incorrect direction was predicted. That is, a value 0.975 for c_j is equivalent to a value of 0.975 for q_j when a rising series is correctly predicted to rise, and to a value of 0.025 for q_j when a falling series is correctly predicted to fall. Similarly, a value of c_j of 0.025 is equivalent to a value of 0.025 for q_j when a falling series is incorrectly predicted to rise, and to a value of 0.975 for q_j when a rising series is incorrectly predicted to fall.

A mean weighted outcome index, $M(c)$, can be derived as the mean of the c_j's adjusted by the length of the sub-period. $M(c)$ is defined in Eq. (8):

$$M(c) = \frac{1}{n}\sum_j n_j c_j \tag{8}$$

The measure $M(c)$ is directly related to the profitability of actions associated with a set of probability forecasts. Values of $M(c)$ above 0.5 would be consistent with profits being made from currency operations and values below 0.5 would be consistent with losses.

4.3. Hypothetical forecasters

When assessing judgement, it is informative to evaluate the relative performance displayed by a forecaster with that of two hypothetical participants: the "random walk forecaster" (RWF) and the "perfect forecaster" (PF). The RWF assigns all probabilities as 0.5 with an arbitrary direction, and hence provides a no-knowledge or equal-belief benchmark. The value of $M(c)$ for the RWF is 0.5. The PF, on the other hand, always predicts the correct direction of movement and assigns to that direction a probability equal to the empirical weighted outcome index c_j. This provides

an important benchmark at the other end of the performance scale: a subject could not possibly perform better than the PF. The value of $M(c)$ for the PF is $0.5 + \frac{1}{n} \sum_j n_j|p_j|$.

4.4. Overall accuracy measures

Denoting by r_j the forecaster's half range probability response for predictive sub-period j ($0.5 \leq r_j \leq 1$), the mean response across all sub-periods is defined as $M(r) = \frac{1}{n} \sum_j n_j r_j$. The forecaster's Mean Square Probability Score (MSPS) is computed using the probability response, r_j, and the empirical weighted outcome index, c_j. The MSPS is defined in Eq. (9):

$$\text{MSPS} = \frac{1}{n} \sum_j n_j (r_j - c_j)^2 \tag{9}$$

On the MSPS, the PF would have a value of zero and the RWF a value $\frac{1}{n} \sum_j n_j p_j^2$.

The Mean Absolute Probability Score (MAPS) may also be computed using c_j. The MAPS is defined in Eq. (10):

$$\text{MAPS} = \frac{1}{n} \sum_j n_j |r_j - c_j| \tag{10}$$

On the MAPS, the PF would have a value of zero and the RWF a value $\frac{1}{n} \sum_j n_j |p_j|$.

4.5. Relative accuracy measures

The interpretation of a subject's $M(c)$ is often problematical as its upper limit is constrained by the value for the PF. The value of $M(c)$ for the PF depends on the actual movements of the exchange rate and can, therefore, take different values for different series and different periods of time. It is desirable to have a relative measure that permits comparisons between predictions made for different series and for different sets of dates. This can be achieved by a simple transformation, via the percentage perfect forecaster adjusted mean weighted outcome index, PM(c), given in Eq. (11):

$$\text{PM}(c) = 100 \left\{ \frac{M(c) - 0.5}{\frac{1}{n} \sum_j n_j |p_j|} \right\} \tag{11}$$

On PM(c), the PF and RWF would have convenient values of 100 and 0 respectively. Values of PM(c) below zero indicate directional performance that is worse than the RWF {i.e., if $M(c)<0.5$, then PM(c)<0}.

The numerical values of MSPS and MAPS are similarly influenced by p_j. For instance, the MSPS value for the RWF will take various numerical values for different series, and this makes comparisons difficult. It is desirable, therefore, to construct measures that are dimensionless, giving the same values for the PF and the RWF in all situations. An approach, similar to Theil (1966), which was extended to probability measures by Wilkie and Pollock (1996), was used so that a relative measure of the MSPS can be obtained. The MSPS is divided by the MSPS value for the RWF $\left(\frac{1}{n} \sum_j n_j p_j^2\right)$ to give an expression UMSPS in the form of Eq. (12), which is analogous to Theil's U^2 statistic.

$$\text{UMSPS} = \frac{\text{MSPS}}{\frac{1}{n} \sum_j n_j p_j^2} \tag{12}$$

UMSPS has a value of zero for the perfect forecaster and unity for RWF. This value can be multiplied by 100 to give the Percentage Mean Squared Probability Score (PMSPS). The square root of UMSPS can be taken to give the relative Root Mean Square Probability Score, URMSPS (i.e., URMSPS= $\sqrt{}$UMSPS), which is turn may be multiplied by 100 to give the Percentage Root Mean Square Probability Score (PRMSPS). The PMSPS and PRMPS are defined in Eqs. (13) and (14):

$$\text{PMSPS} = \text{UMSPS}*100 \tag{13}$$

$$\text{PRMSPS} = \text{URMSPS}*100 \tag{14}$$

A similar procedure can be applied to the MAPS. The MAPS may be divided by the theoretical MAPS value for the RWF to give UMAPS as in Eq. (15):

$$\text{UMAPS} = \frac{\text{MAPS}}{\frac{1}{n} \sum_j n_j |p_j|} \tag{15}$$

The Percentage Mean Absolute Probability Score (PMAPS) may similarly be constructed as in Eq. (16):

$$\text{PMAPS} = \text{UMAPS}*100 \tag{16}$$

4.6. Accuracy components

The MSPS is essential as part of the overall evaluating procedure, since its various decompositions identify specific dimensions of forecasting performance which illustrate particular strengths and weaknesses in a forecaster's approach. This, of course, is vital for selecting appropriate training and debiasing techniques. The MSPS can be decomposed in a number of ways. The decomposition proposed here uses an extension of Yates' (1982) procedure, which was further modified in Wilkie and Pollock (1996). This is presented in Eq. (17):

$$\mathrm{MSPS} = \mathrm{RAV} + \mathrm{SC} + B^2 \tag{17}$$

where Resolution Adjusted Variability, RAV=$V(c)(1-\mathrm{SL})^2$, with $V(c)$ the variance of (c_j), i.e., $V(c) = \frac{1}{n}\sum_j n_j c_j^2 - [M(c)]^2$; SL is the slope of the fitted regression line of (r_j) on (c_j), i.e., SL=$C(r,c)/V(c)$, with $C(r,c)$ the covariance between (r_j) and (c_j), i.e., $C(r,c) = \frac{1}{n}\sum_j n_j r_j c_j - M(r)M(c)$; SC is the scatter, the variance about the fitted regression line of (r_j) on (c_j), i.e., SC=$V(r)$–$\mathrm{SL}^2V(c)$, with $V(r)$ the variance of (r_j), i.e., $V(r) = \frac{1}{n}\sum_j n_j r_j^2 - M(r)^2$; and B is bias, defined as $B=M(r)$–$M(c)$.

The first term on the right hand side of Eq. (17), (RAV) has, generally, the dominant effect on the MSPS. The RAV is a composite measure involving SL and $V(c)$. The lower the value on this measure the better. A key component of the RAV is the slope measure.

The slope (SL) is a measure of resolution or 'discrimination' and reflects the ability to group events into categories (Yates, 1990). In the present context, it measures the degree to which higher probabilities are assigned for correctly forecast large-scaled changes in the exchange rate. For the PF, $r_j=c_j$ for all forecasts and so that SL=1 and hence the closer SL is to unity the better the performance. SL has a value of zero for the RWF (since the fitted regression line is horizontal). According to Yates (1990), SL is a critical component of accuracy, reflecting an individual's level of expertise. SL is particularly important in currency speculation where there is a need to discriminate between periods when the exchange rate is likely to show a large movement in a particular direction from when it is not. Good resolution is essential to obtain profits from currency speculation.

The $V(c)$ measure also has an important impact on RAV. A low value of $V(c)$ associated with good performance on $M(c)$ can result in a low RAV value. RAV, therefore, is associated with the two most important aspects of exchange rate forecasting: good directional performance and resolution. A good performance on RAV is consistent with good profitability performance. In situations where $M(c)$ is negative, RAV has to be viewed with caution as a low value of $V(c)$ could be associated with low directional performance.

In practice the interpretation of RAV needs to be compared with the random walk forecaster whose RAV value varies according to actual movements of the logarithms of the exchange rate. It is desirable to have a relative measure that permits ready comparisons of predictions made for different sets of dates and for different series. This can be achieved by a transformation to give the Percentage Resolution Adjusted Variability (PRAV) as defined in Eq. (18):

$$\mathrm{PRAV} = \frac{100.\mathrm{RAV}}{\frac{1}{n}\sum_j n_j p_j^2} \tag{18}$$

On PRAV, the RWF would have a value 100 and the PF a value of zero.

The second term of Eq. (17), scatter (SC), reflects variation in the (r_j) values that is not explained by variation in the (c_j) values. SC is the variation about the fitted simple linear regression of (r_j) on (c_j) and reflects unexplained variation in the responses, i.e., variation in the forecaster's responses (r_j) that is not explained by variation in the weighted outcome index (c_j). The scatter term is zero for both the PF and the RWF who, of course, use no information in the assessment of probability. SC reflects variation in the probability responses that is not related to variation in the outcomes. This could arise from forecasters using inconsistent strategies in forming predictions or identifying patterns in the series that are not relevant. SC, however, has to be viewed in relation to $V(r)$. In the special case where SL is zero, SC equals $V(r)$. If $M(r)$ is close to 0.5, $V(r)$ and SC would be very small. SC tends to be

negatively related to SL and positively related to $M(r)$.

The last term on the left hand side of Eq. (17) is the Bias squared (B^2) term which reflects the under/overconfidence in predictions. The bias is positive in cases of overconfidence and negative in cases of underconfidence. On B, both the RWF and PF would have a value of zero. Under/overconfidence can be particularly a problem in the risk management of currency operations. Decisions as to whether or not to hedge against currency movements could be adversely affected by persistent under or overconfidence in probability forecasts.

The decomposition in Eq. (17) may also be expressed in terms of PMSPS by dividing Eq. (17) throughout by $\frac{1}{n}\sum_j n_j p_j^2$ and then multiplying by 100 to give Eq. (19):

$$\mathrm{PMSPS} = \mathrm{PRAV} + \mathrm{PSC} + \mathrm{PB} \tag{19}$$

where $\mathrm{PSC} = \frac{100.\mathrm{SC}}{\frac{1}{n}\sum_j n_j p_j^2}$ conveys the Percentage Scatter, and $\mathrm{PB} = \frac{100.B^2}{\frac{1}{n}\sum_j n_j p_j^2}$ gives the Percentage Bias.

Finally, for completeness, it is convenient to define the Percentage Slope as PSL=[100*SL] and the Percent Mean Response as PM(r)=$M(r)$*100.

The values for the accuracy measures for the RWF and PF are summarised in Table 2.

Table 2
Values of the performance measures for the random walk forecaster (RWF) and the perfect forecaster (PF)

Measure	RWF	PF
PM(c)	0	100
PRMSPS	100	0
PMSPS	100	0
PMAPS	100	0
PRAV	100	0
PSC	0	0
PB	0	0
PSL	0	100

Here: PM(c) is the Percentage Perfect Forecaster Adjusted Mean Weighted Outcome Index; PRMSPS is the Percentage Root Mean Squared Probability Score; PMSPS is the Percentage Mean Squared Probability Score; PMAPS is the Percent Mean Absolute Probability Score; PRAV is Percentage Resolution Adjusted Variability; PSC is the Percentage Scatter; PB is the Percentage Bias; and PSL is the Percentage Slope.

4.7. Statistical tests on the accuracy statistics

To consider whether the accuracy statistics indicated that forecast performance was significantly better than the RWF, tests were applied to the relevant measures for the whole period and the grouped 10 sub-periods. To apply the procedure it was, however, necessary to ignore the fact that the sub-periods had differing lengths. Due to non-normality of the performance measures, non-parametric tests were used. The Wilcoxon signed rank test was used in conjunction with the absolute probability and squared probability overall performance measures (MAPS and MSPS) following the procedure set out in Diebold and Mariano (1995). The test essentially involves comparing the medians of the differences in accuracy between the forecast performance and the RWF. For the absolute probability measure the null hypothesis was that the median $\{|r_j-c_j|-|0.5-c_j|\}=0$ against the alternative that the median $\{|r_j-c_j|-|0.5-c_j|\}<0$. Similarly for the squared probability measure the null hypothesis was that the median $\{(r_j-c_j)^2-(0.5-c_j)^2\}=0$ against the alternative that the median $\{(r_j-c_j)^2-(0.5-c_j)^2\}<0$.

Non-parametric tests were also applied to the component measures. The Wilcoxon test was used for the weighted outcome index and bias. The test on the weighted outcome index involved the null hypothesis that the median of $(c_j-0.5)=0$ against the alternative that the median of $(c_j-0.5)>0$. Bias was examined with the Wilcoxon test using the null hypothesis that the median of $(r_j-c_j)=0$ against the alternative that the median of $(r_j-c_j)\neq 0$. The statistical significance of the slope measure (SL) was also examined using Spearman's rank correlation test with the null hypothesis that the correlation =0 against the alternative that the correlation >0. The analysis for RAV and SC is much more complex as these are composite variables dependent to a large extent on $V(c)$ and SL. Hence, the Wilcoxon test on the weighted outcome index and the Spearman test on the slope can be used to examine these composite components. SC is also very dependent on $V(r)$. The results presented in the next section give $r_j>0.5$ for all j and hence scatter arises from the variation not explained by the weighted outcome index and slope. All the components of accuracy can, therefore, be

tested by reference to bias, slope and the weighted outcome index.

5. An application of the procedure

To illustrate how the procedure works in practice, daily USD/CHF quotations from Barclays Bank International (quoted at or near 17.00 hours UK time) were used to derive the empirical probabilities. A graph of the USSD/CHF series is shown in Fig. 1. Logarithms to base 10 were taken and the resulting series was first differenced. The period extended from 23/7/96 to 7/12/99. The period was split into thirty-five non-overlapping sub-periods, the boundaries of which were determined by the dates of revision of probability recommendations published in a market newsletter by an established financial advisory company. The newsletter gave tactical probability predictions on the USD/CHF as well as other background information. The numbers of days for each sub-period reflected trading days on the London market with weekends and 25 bank holiday days excluded. The currency predictions (provided by the company to its clients together with similar equity predictions for a large number of countries) were made to provide information to the newsletters' subscribers which they could use to avoid exposure in assets denominated in currencies likely to fall while increasing their holdings in assets denominated in currencies likely to rise. This is particularly relevant to the interests of many of the company's clients in relation to the management of their international equity portfolios such that gains from equity holdings were not offset by adverse currency movements. The company revised these probability forecasts at intervals that were not fixed in length but tended to reflect market conditions. This allowed its clients to have the opportunity to liquidate or increase holdings before the end of the forecast horizon when they were notified of possible changes in market conditions. The procedure was used to obtain the thirty-five USD/CHF probability predictions between 22/7/96 to 7/12/99.The sub-period numbers are associated with the dates given in Table 3.

The information and probability predictions contained in the newsletter, together with previous values of the probability estimates, were used to form judgemental predictions. The predictions were considered to relate to a 30-working-day horizon (i.e., excluding Saturdays, Sundays and non-trading weekdays). It was considered that this horizon was sufficient to allow the clients to consider tactical repositioning of their holdings at frequent intervals. The horizon was also deemed long enough to reflect movements in secondary trends, which manifest

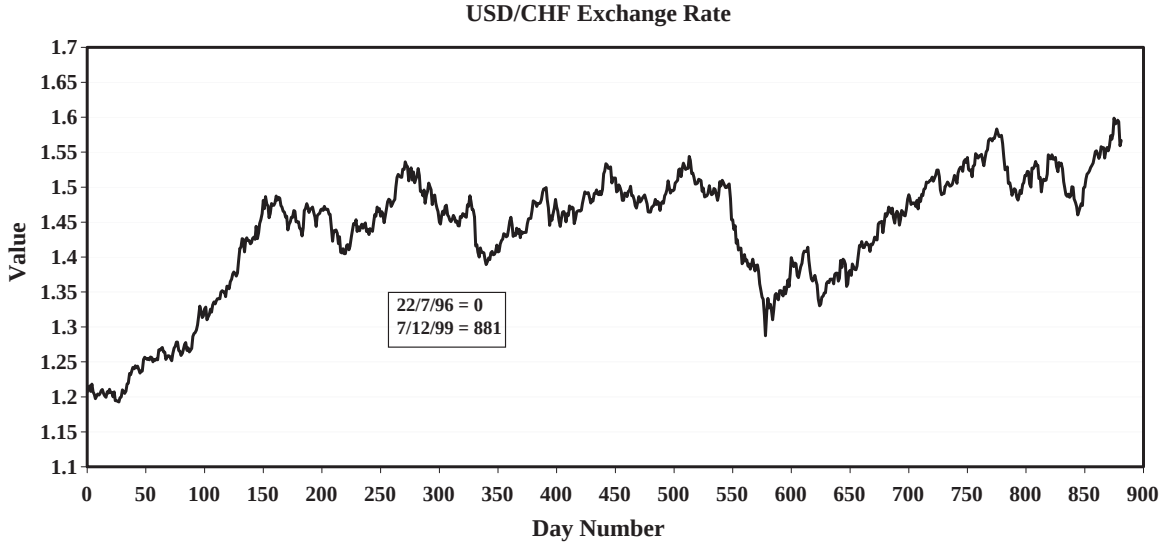

Fig. 1. USD/CHF exchange rate.

Table 3
Subjective predictions and empirical values for USD/CHF movements

Sub-period number	Start day number	Number of days	Dates of sub-period	Empirical probability	Subjective probability 30 days	Subjective probability adjusted
1	1	28	23/07/96–30/08/96	0.410	0.53	0.529
2	30	5	02/09/96–06/09/96	0.868	0.55	0.520
3	35	15	09/09/96–27/09/96	0.931	0.66	0.615
4	50	35	30/09/96–15/11/96	0.745	0.70	0.714
5	85	15	18/11/96–06/12/96	0.869	0.69	0.637
6	100	28	09/12/96–20/01/97	0.987	0.71	0.704
7	131	30	21/01/97–03/03/97	0.868	0.60	0.600
8	161	28	04/03/97–14/04/97	0.433	0.59	0.587
9	191	21	15/04/97–14/05/97	0.237	0.55	0.542
10	213	12	15/05/97–02/06/97	0.404	0.54	0.525
11	226	20	03/06/97–30/06/97	0.806	0.61	0.590
12	246	17	01/07/97–23/07/97	0.749	0.75	0.694
13	263	26	24/07/97–29/08/97	0.540	0.70	0.687
14	290	29	01/09/97–09/10/97	0.261	0.68	0.677
15	319	25	10/10/97–13/11/97	0.139	0.65	0.637
16	344	34	14/11/97–05/01/98	0.965	0.61	0.617
17	381	30	06/01/98–16/02/98	0.351	0.67	0.670
18	411	13	17/02/98–05/03/98	0.769	0.59	0.560
19	424	33	08/03/98–23/04/98	0.555	0.81	0.821
20	459	31	26/04/98–09/06/98	0.342	0.73	0.733
21	492	24	10/06/98–13/07/98	0.836	0.67	0.653
22	516	24	14/07/98–14/08/98	0.384	0.64	0.626
23	540	13	17/08/98–03/09/98	0.034	0.54	0.526
24	554	16	04/09/98–25/09/98	0.148	0.45	0.463
25	570	17	28/09/98–20/10/98	0.321	0.55	0.538
26	587	21	21/10/98–18/11/98	0.754	0.67	0.644
27	608	57	19/11/98–10/02/99	0.649	0.64	0.689
28	668	20	11/02/99–10/03/99	0.876	0.67	0.640
29	688	26	11/03/99–19/04/99	0.862	0.60	0.593
30	716	21	20/04/99–19/05/99	0.453	0.78	0.741
31	738	41	20/05/99–16/07/99	0.925	0.60	0.616
32	780	22	19/07/99–17/08/99	0.146	0.61	0.594
33	802	2	18/08/99–19/08/99	0.239	0.70	0.554
34	804	30	20/08/99–01/10/99	0.401	0.58	0.580
35	835	47	04/10/99–07/12/99	0.872	0.62	0.649

Omitted days no. 25, 112, 113, 117, 179, 180, 205, 220, 285, 373, 374, 378, 449, 450, 465, 480, 550, 634, 635, 639, 704, 705, 725, 745, 810.

themselves in short-term drift, as well as reflecting primary trends which manifest themselves in long-term drift. The resulting probability predictions were, therefore, continually reviewed with predictions being revised approximately every month. In addition, as changes in market conditions can occur at any time, updates were provided when the need occurred.

Table 3 summarises the results. Column 1 indicates the sub-period number, column 2 the start day number, column 3 the number of trading days in the sub-period, column 4 the dates of the sub-periods and column 5 the empirical probability for the sub-period. Lilliefors' test for non-normality indicated four significant values at the 5% level (sub-periods 15, 25, 30 and 35) but there were no significant values at the 1% level. The Jarque–Bera test indicated three significant values at the 5% and 1% levels (sub-periods 15, 30 and 35). The results for sub-periods 15, 30 and 35 reflected the presence of one clear negative outlier in each case (i.e., 28/10/97, 23/1/98 and 6/12/99). Removing the outlier in all three cases resulted in non-significant values of the

Jarque–Bera test. Non-normality can have implications on the resulting empirical probabilities. A negative outlier biases the mean downwards and the standard deviations upwards. As the overall effect is usually more pronounced on the mean, the empirical probability is biased downwards. In the three cases identified above the effect of removing the outlier would have caused the empirical probability to increase from 0.139 to 0.299 for sub-period 15, from 0.453 to 0.764 for sub-period 30 and 0.872 to 0.972 for sub-period 35. On balance, it can be considered that the changes in the logarithms of the exchange rate in the sub-periods are, therefore, approximately normally distributed and that instances of non-normality caused by outliers did not have a major effect on the accuracy analysis. Bartlett's test for serial correlation gave no significant values at the 5% level.

Probability predictions for a 30-day predictive horizon (column 6) are also presented in Table 3. As the sub-periods were of varying lengths the adjusted probability predictions are also given (column 7) which allows comparison with the empirical probabilities (column 5) for the full-range approach.

The performance statistics were calculated using 10 sub-periods on a moving basis to give statistics for 26 groups of 10 consecutive sub-periods which comprise the whole period. These are listed in Table 4 (column 1) with the number of trading days (column 2) which varied from one group to the next. In fact the total length of the groups in days varied from 209 (sub-periods 2–11) to 287 (sub-periods 26–35) with a total over the whole period of 856 days. The full-range probability predictions given in Table 3 were converted to half-range probability predictions and the results are

Table 4
Probability performance measures

Sub-periods	No. of days	PM(c)	PMAPS	PRMSPS	PM(r)	PB	BSGN	PSL	PSC	PRAV
1–10	217	61.9	73.1	72.2	61.3	3.0	Neg	18.9	2.7	46.4
2–11	209	72.3	70.4	71.2	62.2	8.0	Neg	15.7	2.5	40.1
3–12	221	73.4	66.5	68.8	63.0	6.9	Neg	16.8	2.4	38.1
4–13	232	70.9	71.1	70.1	63.7	2.1	Neg	15.8	3.3	43.7
5–14	226	45.8	92.5	87.5	63.0	0.3	Pos	7.1	3.6	72.6
6–15	256	29.9	98.4	95.4	63.0	2.4	Pos	4.1	3.0	85.6
7–16	242	21.9	104.6	102.0	62.0	4.7	Pos	−0.1	2.9	96.5
8–17	242	−3.2	119.6	113.7	62.9	25.6	Pos	−0.0	3.7	100.0
9–18	227	6.2	113.5	111.0	63.0	16.2	Pos	−1.8	3.6	103.4
10–19	239	20.2	125.7	116.3	66.4	19.8	Pos	−5.1	8.1	107.3
11–20	258	12.4	136.0	124.9	67.9	33.4	Pos	−8.3	6.6	116.0
12–21	262	15.2	131.7	121.7	68.4	31.3	Pos	−6.4	5.7	111.2
13–22	269	3.6	143.3	127.2	67.8	43.6	Pos	−5.8	6.4	111.8
14–23	256	−7.9	136.0	123.5	66.9	44.2	Pos	−0.8	7.2	101.1
15–24	243	12.2	127.6	115.7	65.9	19.1	Pos	−3.8	8.2	106.5
16–25	235	24.1	125.5	111.4	65.3	11.7	Pos	−3.0	10.7	101.5
17–26	222	6.7	133.5	123.8	66.8	37.6	Pos	−0.1	15.8	99.9
18–27	249	32.3	108.0	110.9	66.3	19.1	Pos	1.9	15.3	88.6
19–28	256	37.4	103.6	105.3	66.7	13.7	Pos	2.6	12.4	84.8
20–29	249	45.2	87.8	93.8	63.9	1.1	Pos	1.0	5.5	81.4
21–30	239	55.1	83.2	86.1	63.5	0.0	Pos	4.7	5.2	69.0
22–31	256	61.4	83.9	85.1	63.1	1.0	Neg	2.1	4.2	67.2
23–32	254	50.1	83.7	88.1	62.8	0.2	Neg	3.6	3.8	73.6
24–33	243	62.7	82.0	85.8	63.3	1.6	Neg	−0.6	3.8	68.2
25–34	257	51.8	85.9	87.6	63.3	0.0	Pos	3.9	3.7	72.9
26–35	287	68.5	78.7	80.6	64.1	2.2	Neg	1.1	2.3	60.5
1–35	856	44.8	94.7	93.2	64.0	0.7	Pos	2.3	5.5	80.6

Here: PM(c) is the Percentage Perfect Forecaster Adjusted Mean Weighted Outcome Index; PMAPS is the Percent Mean Absolute Probability Score; PRMSPS is the Percentage Root Mean Squared Probability Score; PM(r) is the Percent Mean Response; PB is the Percentage Bias; BSGN is the Bias Sign; PSL is the Percentage Slope; PSC is the Percentage Scatter; and PRAV is the Percentage Resolution Adjusted Variability.

presented in Table 4 (columns 3 to 11). The overall performance measures, PM(c), PMAPS and PRMSPS (columns 3 to 5), the mean responses, PM(r) (column 6) and component measures, PB, BSGN (the sign of B:+ve or −ve), PSL, PSC and PRAV (columns 7 to 11), were obtained using the twenty-six sets of sub-periods. The performance statistics were also calculated for the whole period (i.e., all 35 sub-periods).

On the accuracy measures, the PM(c) value of 44.8 for the whole period was considerably better than the RWF (i.e., value of zero), reflecting good overall directional performance. The Wilcoxon signed rank test on the weighted outcome index indicated significance at the 5% level, which supports the view that the overall directional performance was better than the RWF. In fact, 24 of the 26 sub-period groups showed values better than the RWF, with the best performance being shown for the sub-period groups at the beginning and end of the period. For the sets of sub-periods the Wilcoxon test gave significant values for two consecutive sets (i.e., 2–11 and 3–12). The PMAPS and PRMSPS values for the whole period of 94.7 and 93.2 respectively were slightly better than the RWF (who would score a value of 100 on each measure). The Wilcoxon signed ranks test did not indicate significance for either the absolute or squared measures, which indicates that probability forecast performance was not significantly better than the RWF. For the PMAPS and PRMSPS, 13 out of the 26 sub-period groups gave better values than the RWF. The best performance, again, was shown for the sub-period groups at the beginning and end of the period. For the sets of sub-periods, the Wilcoxon tests gave significant values for two consecutive sets for the absolute measure (i.e., 2–11 and 3–12) and three sets for the squared measure (i.e., 2–11, 3–12 and 4–13). The PM(r) measure was 64.0% for the whole period, but the responses were generally higher in the middle of the period, which coincided with the poorer overall performance.

On the accuracy components, PB and BSGN illustrated slight overconfidence over the whole period with respective values of 0.7 and ‘pos’ (a value of zero for PB reflects perfect confidence). There was almost perfect confidence for sub-period groups at the beginning and end of the whole period with clear overconfidence illustrated in the middle of the period. The Wilcoxon test did not indicate significant under/overconfidence over the period as a whole or for any of the sets of sub-periods. The PSL measure indicated that resolution over the whole period was somewhat better than the RWF with a value of 2.3 (whose PSL value would be zero). The Spearman test for the whole period was not significant, hence it can be concluded that the resolution was not significantly better than the RWF. The results generally, illustrated good resolution at the beginning of the period and towards the end of the period, with negative resolution in the middle part of the period. For the sets of sub-periods the Spearman test gave significant values for four consecutive sets (i.e., 1–10, 2–11, 3–12 and 4–13). PSC was reasonable for the whole period with a value of 5.5. PSC was, however, low in the first half of the period but increased up to the sub-period group 17–26, after which it fell back such that for the last sub-period groups of the period it was below the mean of the whole period. The value of the PRAV over the whole period (80.6) was better than the RWF (value 100). In fact, for the sub-period groups, 17 out of 26 were better than the RWF. Again the pattern showed the best performance for sub-period groups at the beginning and end of the period. For the sets of sub-periods, two of the three lowest values for SC and RAV coincided with significant values on the $M(c)$ and SL tests (i.e., 2–11 and 3–12).

The accuracy statistics, therefore, indicate that the probability predictions were considerably superior to the RWF for the sub-period groups at the beginning of the period and to a lesser extent at the end of the period, but, generally, poorer for sub-period groups in the middle of the period. The results suggest that this poor performance can be attributed to low directional performance, overconfidence, negative or low resolution and relatively high scatter. The explanation for this could lie in the time series characteristics of the USD/CHF over the period. The performance statistics clearly illustrated good performance when the series exhibited a clear upward trend (i.e., approximately day numbers 32 to 150). When there was no clear trend (i.e., approximately day numbers 151 to 522), performance was much poorer. The sharp decline (i.e., approximately day numbers 523 to 577) was not really identified in the

predictions, with the result that performance was poor when this period was included in the calculation of accuracy statistics.

As clearly observed from the illustration above, an evaluation of predictive performance using these accuracy measures provides useful insights into the strengths and weaknesses of the probability predictions in relation to the characteristics of the series. In addition, specific elements of accuracy (e.g., under/overconfidence, resolution and scatter) are identified, revealing factors that highlight the analyst's strengths and weaknesses in forecasting performance. Accordingly, the accuracy measures may easily be used as powerful feedback tools.

6. Conclusion

A procedure has been outlined that can be used to identify specific strengths and weaknesses of judgement in the context of directional probability currency forecasting. The aim was to provide a flexible examination of currency predictions that can be easily applied in a practical context using daily exchange rate data. The main strength of the procedure is that it can be used in situations where the length of time between predictions is variable, as in the case of rolling forecasts. The procedure also provides a selection of measures that are amenable to practitioners' everyday usage, while providing information on a spectrum of performance aspects that can be utilized in forecaster training (Benson & Önkal, 1992).

While the procedure has been applied to daily exchange rates, it could just as easily be applied to weekly or monthly data for longer predictive horizons. In addition, the procedure could incorporate the examination of point predictions with associated confidence bands, as well as predictions from quantitative models. Future extensions could involve detailed evaluations of composite predictions, which carry significant consequences for corporate forecasting practices. In a similar vein, the procedure further supports the formation of consistent probability predictions for related cross-exchange rates (Pollock, Macaulay, Önkal-Atay, & Thomson, 2002). Finally, the proposed procedure also pertains to financial price series other than currencies (e.g., share price indices and most equity series), thus presenting an effective tool for training and performance feedback on various financial platforms.

References

Bartlett, M. S. (1946). On the theoretical specification of sampling properties of autocorrelated time series. *Journal of the Royal Statistical Society*, *8*(Series B), 27–41.

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, *8*, 559–573.

Boothe, P., & Glassman, D. (1987). The statistical distribution of exchange rates: Empirical evidence and economic implications. *Journal of International Economics*, *2*, 297–319.

Coppes, R. C. (1995). Are exchange rate changes normally distributed. *Economics Letters*, *47*, 117–121.

Corporale, G. M., Hassapis, C., & Pittis, N. (1998). Conditional leptokurtosis and non-linear dependence in exchange rate returns. *Journal of Policy Modeling*, *20*, 518–601.

Corporale, G. M., & Pittis, N. (1996). Modelling the sterling–deuchmark exchange rate: Non-linear dependence and thick tails. *Economic Modelling*, *13*, 1–14.

Daniel, K., Hirshleifer, D., & Teoh, S. H. (2002). Investor psychology in capital markets: Evidence and policy implications. *Journal of Monetary Economics*, *49*, 139–209.

Diebold, F. X., & Mariano, R. S. (1955). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–263.

Friedman, D., & Vandersteel, S. (1982). Short run fluctuations in foreign exchange rates: Evidence from the data, 1973–79. *Journal of International Economics*, *13*, 171–186.

Harvey, A. C. (1993). *Time series models*. London: Harvester Wheatsheaf.

Hsieh, D. A. (1998). The statistical properties of daily foreign exchange rates: 1974–1993. *Journal of International Economics*, *24*, 129–145.

Ito, T. (1990). Foreign exchange rate expectations: Micro survey data. *The American Economic Review*, *80*, 434–449.

Jarque, C. M., & Bera, A. K. (1980). Efficiency tests for normality, homoscedasticity and serial independence in regression residuals. *Economic Letters*, *6*, 255–259.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, *77*, 213–217.

Larson, S. J., & Madura, J. (2001). Overreaction and underreaction in the foreign exchange market. *Global Finance Journal*, *12*, 153–177.

Lilliefors, H. W. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*, 399–402.

Moosa, I. A. (2000). *Exchange rate forecasting*. London: Macmillan.

Murphy, J. J. (1999). *The technical analysis of financial markets*. Paramus, NJ: New York Institute of Finance.

Nelson, C. R., & Plosser, C. I. (1982). Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics*, *10*, 139–162.

Önkal-Atay, D., Thomson, M. E., & Pollock, A. C. (2002). Judgemental forecasting. In M. P. Clements, & D. F. Hendry (Eds.), *A companion to economic forecasting* (pp. 133–151). Oxford: Blackwell Publishers.

Pollock, A. C., Macaulay, A., Önkal-Atay, D., & Thomson, M. E. (2002). Consistent judgmental directional probability exchange rate predictions. In K. D. Lawrence, M. D. Geurts, & J. G. Guerard Jr. (Eds.), *Advances in Business and Management Forecasting*, *vol. 3* (pp. 161–175). Oxford: JAI.

Pollock, A. C., & Wilkie, M. E. (1992). Currency forecasting: Human judgement of models. *VBA Journaal*, 21–29.

Rogalski, R. J., & Vinso, J. D. (1978). Empirical properties of foreign exchange rates. *Journal of International Business Studies*, *9*, 69–79.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, *40*, 193–218.

Theil, H. (1966). *Applied economic forecasting*. Amsterdam: North Holland.

Thomson, M. E., Pollock, A. C., Henriksen, K. B., Macaulay, A. (2004). The influence of forecast horizon on judgemental probability forecasts of exchange rate movements. *European Journal of Finance*, 10, 290–307.

Tvede, L. (1990). *The psychology of finance*. Oslo: Norwegian University Press.

Wang, J. -X. (2001). Quote revision and information flow among foreign exchange dealers. *Journal of International Financial Markets, Institutions, and Money*, *11*, 115–136.

Westerfield, J. M. (1977). An examination of foreign exchange risk under fixed and floating rate regimes. *Journal of International Economics*, *7*, 181–200.

Wilkie, M. E., & Pollock, A. C. (1996). An application of probability judgement accuracy measures to currency forecasting. *International Journal of Forecasting*, *12*, 25–40.

Wilkie, M. E., Tuohy, A. P., & Pollock, A. C. (1993, June). Examining heuristics and biases in judgemental currency forecasting. *VBA Journaal,* 12–17.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, *30*, 132–156.

Yates, J. F. (1990). *Judgment and decision making*. New Jersey: Prentice Hall.

Zhou, B. (1996). High frequency data and volatility in foreign exchange rates. *Journal of Business and Economic Statistics*, *14*, 45–52.

Biographies: Andrew C. POLLOCK is a Reader in the School of Computing and Mathematical Sciences, Glasgow Caledonian University. He completed a PhD on exchange rates and has published widely in this area. His particular research interest is the application of analytical techniques to the forecasting of exchange rates and, more generally, financial time series.

Alex MACAULAY is Senior Lecturer in Statistics in the School of Computing and Mathematical Sciences, Glasgow Caledonian University. He has a BSc in Mathematics and an MSc in Statistics (Stochastic Processes). His particular research interest is forecasting of financial time series and in the evaluation of predictive performance.

Mary E. THOMSON is a Reader in the Department of Psychology, Glasgow Caledonian University. She completed a PhD on judgement in currency forecasting and has published articles and papers in a variety of books and journals in this area. Her research interests focus on the role of judgement in financial forecasting and decision making.

Dilek ÖNKAL is a Professor of Decision Sciences and Dean of the Faculty of Business Administration at Bilkent University, Turkey. She received a PhD in Decision Sciences from the University of Minnesota, and is doing research on judgemental forecasting, decision analysis, risk perception, and risk communication.