



Effective return, risk aversion and drawdowns

Michel M. Dacorogna^a, Ramazan Gençay^{b,c,*,1}, Ulrich A. Müller^d,
Olivier V. Pictet^e

^aZurich Re, Zurich, Switzerland

^bDepartment of Economics, Mathematics and Statistics, University of Windsor, 401 Sunset Avenue,
Windsor, ONT, Canada

^cDepartment of Economics, Bilkent University, Bilkent, Ankara, Turkey

^dOlsen & Associates, Seefeldstrasse 233, CH-8008, Zurich, Switzerland

^eDynamic Asset Management, Geneva, Switzerland

Received 8 June 2000

Abstract

We derive two risk-adjusted performance measures for investors with risk averse preferences. Maximizing these measures is equivalent to maximizing the expected utility of an investor. The first measure, X_{eff} , is derived assuming a constant risk aversion while the second measure, R_{eff} , is based on a stronger risk aversion to clustering of losses than of gains. The clustering of returns is captured through a multi-horizon framework. The empirical properties of X_{eff} , R_{eff} are studied within the context of real-time trading models for foreign exchange rates and their properties are compared to those of more traditional measures like the annualized return, the Sharpe Ratio and the maximum drawdown. Our measures are shown to be more robust against clustering of losses and have the ability to fully characterize the dynamic behaviour of investment strategies. © 2001 Elsevier Science B.V. All rights reserved.

PACS: C52; C53

Keywords: Performance measures; Sharpe ratio; Effective return; Drawdown

1. Introduction

Evaluating the performance of an investment strategy gives rise to many debates. This is due to the fact that the performance of any financial investment is measured not only by the increase of capital but also by the risk incurred during the time to reach this increase in capital. Return and risk must be evaluated together. Already in

* Corresponding author. Tel.: +1-519-253-3000 ext. 2382; fax: +1-519-973-7096.

E-mail address: gencay@uwindsor.ca (R. Gençay).

¹ This paper was written while Gençay visited Olsen & Associates as a research scholar.

1966, William Sharpe [1] introduced a measure of mutual funds performance which was to become later an industry standard under the name of Sharpe Ratio, [2].

An appropriate performance measure is the most crucial determinant for judging the performance of investment strategies. A high performance should result if the total return is high, when the total return curve increases linearly over time and loss periods (if there are any) are not clustered. Unfortunately, the Sharpe Ratio does not entirely satisfy these requirements. First, the definition of the Sharpe Ratio² puts the variance of the return into the denominator which makes the ratio numerically unstable at extremely large values when the variance of the return is close to zero. This creates a lack of identification between the return and its volatility. Second, the Sharpe Ratio is unable to consider the clustering of profits and losses. An even mixture of profit and loss trades is usually preferred to clusters of losses and clusters of profits, provided the total set of profit and loss trades is the same in both cases. As shown later, the proposed performance measures solve this problem by observing returns over different time interval sizes. Third, the Sharpe Ratio treats the variability of profitable returns (which are unimportant to investors) the same way as the variability of losses (which are an investor's major concern).

In this paper, we introduce performance measures which are directly related to the *utility* of a strategy to a risk-averse investor. In such an approach, maximization of the performance measures implies the maximization of the expected utility of an investor. Unlike the Sharpe Ratio, the proposed performance measures also inform the investor on the best choice of leverage.

The proposed two new risk-adjusted performance measures are named as X_{eff} and R_{eff} . They are numerically more stable than the Sharpe Ratio and exhibit fewer deficiencies. As a return curve reflects the real risk of a trading strategy, both X_{eff} and R_{eff} are based on the analysis of the return curve: the sum of the accumulated total return and the current unrealized return of open positions. Accounting for unrealized returns is a means to avoid a bias in favor of strategies with a low transaction frequency. Both X_{eff} and R_{eff} reach the value of the annualized total return if the return curve is a straight line as a function of time. For all nonlinear equity curves, X_{eff} and R_{eff} are smaller than the annualized total return.

R_{eff} differs from X_{eff} such that it has a high risk aversion in the zone of negative returns and a low one for profits, whereas X_{eff} assumes constant risk aversion. This means that R_{eff} is dominated by the large drawdowns, or the ability of the model to prevent these. Moreover, R_{eff} does not need any distributional assumption on the returns which makes the measure better suited to fat tailed returns. The two measures give a slightly different ranking sequence to a set of investment strategies with

² The definition of the Sharpe Ratio is

$$S_I \equiv A_{\Delta t} \frac{\bar{r}}{\sqrt{\sigma_r^2}},$$

where \bar{r} is the average return and σ_r^2 is the variance of the return around its mean and $A_{\Delta t}$ is an annualization factor depending on the frequency at which the returns are measured, usually monthly, so $A_{\Delta t} = \sqrt{12}$ [2].

different properties. Trading models that frequently hold a neutral position and have weak drawdowns are valued higher by R_{eff} ; models with a steady profit flow, interrupted by some large drawdowns, are valued higher by X_{eff} .

In this paper, we explore the performance measures in the context of investment strategies in the foreign exchange (FX) market, called real-time trading models. These models use high-frequency live data feeds and their recommendations are transmitted to the traders through data feed lines instantaneously. A popular real-time trading model is used to evaluate the statistical properties of X_{eff} , R_{eff} and the Sharpe Ratio. The out-of-sample test period is seven years of 5 min series on three major foreign exchange rates and one cross rate. In addition to X_{eff} , R_{eff} and Sharpe Ratio, the annualized return and the maximum drawdown are also used as performance measures. The robustness of the underlying probability distributions of X_{eff} and R_{eff} relative to the distribution of the Sharpe Ratio is evaluated by means of distributions generated from simulated data using three traditional statistical processes, namely the random walk, the generalized autoregressive conditional heteroskedastic process (GARCH) and the autoregressive GARCH (AR-GARCH) process.

The results of the empirical study indicate that X_{eff} and R_{eff} are more stringent performance measures relative to the Sharpe Ratio, the annualized return and the maximum drawdown due to the fact that they utilize more points of the entire equity curve. These two performance measures leave no degrees of freedom for randomness when a simulated equity curve is compared with the actual equity curve of the foreign exchange returns. They are robust against randomness and fully characterize the strategies. X_{eff} and R_{eff} have proven their usefulness in optimizing trading model parameters [3] and can be used to rank investment strategies of different kinds, not just FX trading models.

The assumptions leading to the proposed performance measures are presented in Section 1.1. The rest of the paper is organized as follows. In Sections 2 and 3, the performance measures are derived. The simulation methodology is presented in Section 4, and the trading models are described in Section 5. The empirical results are presented in Section 6 and we conclude finally in Section 7.

1.1. Basic assumptions and concepts

The new performance measures of this paper are based on some assumptions that differ from those found in most of the literature. The Sharpe Ratio, as a conventional measure, stays (approximately) constant if the leverage of an investment is changed. Therefore, it cannot be used as a criterion to decide on the choice of leverage.

Real investors however care about the optimal choice of leverage because they do not have an infinite tolerance for losses. Unlike the Sharpe Ratio, the performance measures of this paper give the investor also a criterion of how to choose the leverage. If a certain strategy is leveraged, the same investor will have a different risk, reflected by different values of X_{eff} and R_{eff} .

Finding the maximum of X_{eff} or R_{eff} in a set of possible investment strategies is equivalent to portfolio optimization, where the allocation size (leverage) of different

financial instruments is determined. There is a strong relationship between our performance measures and classical portfolio theory [4]. The main goal of portfolio optimization is to find the maximum of the return \bar{X} for a given variance σ^2 or, equivalently, the maximum of a joint target function:

$$\bar{X} - \lambda\sigma^2 = \max, \quad (1)$$

where λ is a Lagrange multiplier. Although X_{eff} independently originates from utility theory, its definition is in perfect agreement with Eq. (1) and thus with classical portfolio theory. A comparison to Eq. (8) indicates this resemblance. The risk aversion γ plays a role analogous to the Lagrange multiplier λ . The second term $-\lambda\sigma^2$ of Eq. (1), as the corresponding term of X_{eff} , has a natural interpretation: it is the *risk premium* associated with the investment. The performance measure X_{eff} , although applied to simple investments here, can be used as a target function of arbitrarily complex investment portfolios. This is also the case for R_{eff} .

Our performance measures are based on the return history of a certain investment strategy, irrespective of the circumstances under which this strategy was designed. In general, we want to analyze the true history of returns of a complete investment strategy, including all its hidden parts. In practice, however, many investors measure their performance in terms of the deviation from a benchmark. To accommodate these investors, we can replace the return history by the tracking error which is measured by the history of return differences between the investment and a benchmark investment. The methodology with the trading error history is the same as before where the resulting performance should then be called outperformance.

In general, returns cannot be expected to be serially independent. Loss returns may be clustered to form so-called drawdowns (periods with a substantial total loss). The clustering of losses varies, it may be stronger for certain markets and trading strategies than for others. Our performance measures have been designed with special attention to drawdowns since these are the worst events for an investor. The badness of a drawdown is mainly determined by the size of the total loss. Local details of the return curve and the duration of the drawdown period are less important. If we use only one interval size to measure all returns, we would most likely miss drawdown periods because this interval size is either too small (for full clusters of losses) or too large (thus diluting the drawdown with surrounding profitable periods). This is why the performance measures of this paper use different interval sizes to analyze the return history. Owing to this *multi-horizon* feature, the worst drawdown periods cannot be entirely missed, whatever their duration. This is particularly important for R_{eff} , where drawdowns have a overproportional impact on the final result. Therefore, R_{eff} is not only multi-horizon but also has an overlapping scheme of observation intervals.

The perception of performance may also depend on the investor's wealth. The investors studied by Barberis et al. [5] are more risk-averse after a drawdown than after profitable trades. Stutzer [6] makes a distinction between constant absolute risk aversion (CARA) and constant relative risk aversion (CRRA). In CRRA, the risk aversion is relative to the investor's changing wealth which includes all accumulated profits

and losses of the particular investment. In absolute terms, a CRRA-type investor has a changing risk aversion as found by Barberis et al. [5].

We agree with Barberis et al. [5] and Stutzer [6] that absolute risk aversion levels often depend on an investor’s wealth. Nevertheless, we have two reasons to stay in a constant absolute risk aversion (CARA) framework. First, we do not know the true wealth, which depends on not only the past success of the investment but also the chosen leverage. Moreover, an investor’s true, total wealth may also result from many other investments with entirely different performance histories. In lack of full information, we approximately assume wealth to be constant. Similarly, the profits of the trading models used as the example of this paper are not thought to be re-invested. The allocated capital and the risk aversion levels are assumed to be constant.

The second reason for using constant absolute risk aversion is that the multi-horizon feature of our performance measures already causes a behavior similar to wealth-dependent risk aversion. Instead of measuring wealth, we measure returns over different time interval sizes. If a long drawdown occurs, our performance measures are low because returns are considered also over the long interval of this drawdown. In a one-horizon CRRA approach, the performance measure is low because the wealth shrinks during the drawdown period. The methods are different but the outcome is similar.

2. Effective return with constant risk aversion, X_{eff}

Let x be a random variable with mean \bar{x} and variance σ^2 over the testing interval ΔT . Let $u(x) = -e^{-\gamma x}$ be a utility function with the constant risk aversion³ level γ . As explained in Section 1.1, we mean constant absolute risk aversion (CARA) in the sense of Stutzer [6] rather than a risk aversion relative to the changing wealth of an investor. The expected utility of variable x is computed by

$$E[u(x)] = \int_{-\infty}^{\infty} u(x)P(x) dx, \tag{2}$$

where $P(x)$ is the probability distribution of the variable x . If a Gaussian distribution $\mathcal{N}(\bar{x}, \sigma^2)$ is used in Eq. (2), the expected utility becomes

$$E[u(x)] = u(\bar{x})e^{\gamma^2\sigma^2/2} = -\exp\left\{-\gamma\left(\bar{x} - \frac{\gamma\sigma^2}{2}\right)\right\}. \tag{3}$$

³The risk aversion is defined by

$$r(x) = -\frac{u''(x)}{u'(x)},$$

where u' and u'' are the first and the second derivatives of u with respect to x , respectively. We assume $r(x)$ to be constant for X_{eff} .

This form of the expected utility depends on the type of distribution⁴ chosen for $P(x)$ in Eq. (2). In the context of a trading model, x is defined to be the *total return* (or the wealth reached by unit investment) of the model

$$x(\Delta t) = \tilde{R}(t) - \tilde{R}(t - \Delta t) \quad \text{where } \tilde{R}(t) = R(t) + r_0(t), \tag{5}$$

where R is the total return of the past trades up to time t and r_0 is the unrealized return of the current trading model position. Δt is the time horizon on which this quantity is measured. For a test period ΔT , the variable x takes $N = \Delta T/\Delta t$ values. The choice of \tilde{R} as opposed to R is due to the continuity property of this quantity contrary to R , which abruptly changes its value after a trade.

A measure of trading model performance comparable to the yearly average return can be deduced from the expected utility framework [8]. Let X_{eff} be defined as the performance measure, which is in units of yearly return but includes a risk term reflecting volatility of returns. By inverting the expected utility, the variable X_{eff} is written as

$$X_{eff} = -\frac{\ln(-E[u(x)])}{\gamma}. \tag{6}$$

By using Eq. (3), a simple expression for X_{eff} is obtained as a function of the mean return \bar{x} , the variance σ^2 , and the risk aversion level γ

$$X_{eff} = \bar{x} - \frac{\gamma\sigma^2}{2}, \tag{7}$$

where σ^2 is the variance of x .

For a given time horizon Δt , X_{eff} has the units of an average return and is diminished by a factor proportional to $\gamma\sigma^2$. This measure, contrary to the Sharpe Ratio, is numerically stable and can differentiate between two trading models with a straight line behavior ($\sigma^2 = 0$) by choosing the one with the better average return.

The measure X_{eff} still depends on the time intervals Δt and does not permit the comparison of X_{eff} values for *different* intervals. The usual way to enable such a comparison across different intervals is through *annualization*, that is, multiplication with the annualization factor

$$x_{eff, ann} = \frac{1 \text{ year}}{\Delta t} x_{eff} = \frac{1 \text{ year}}{\Delta t} \left(\bar{x} - \frac{\gamma\sigma^2}{2} \right) = \bar{X} - \frac{\gamma\sigma^2(1 \text{ year})}{2\Delta t}, \tag{8}$$

⁴To explore this dependence further, Dacorogna et al. [7] have also used a rectangular distribution around the mean. With such a distribution, the expected utility has the form

$$E[u(x)] = u(\bar{x}) - \frac{\sinh \gamma\sigma\sqrt{3}}{\gamma\sigma\sqrt{3}}. \tag{4}$$

If the risk factor is expanded in both cases, the first two terms are 1 and $C^2\sigma^2/2$ for both Gaussian and rectangular distributions. They only differ in the $o(\sigma^4)$. The third term of the exponential expansion is a better approximation than 0 for Eq. (4). For the X_{eff} measure, we therefore use Eq. (3) as the expression for the expected utility keeping in mind that it depends on the type of distribution considered for the variable x .

where \bar{X} is the annualized return, no longer dependent on the size of Δt . Volatility is also annualized in a usual way: σ^2 is expressed as $[\sigma^2 (1 \text{ year})]/\Delta t$, where $\sigma^2(1 \text{ year})$ is the annualized volatility of returns about the mean. This volatility is approximately independent⁵ of the size of Δt . In the last form of Eq. (8), the risk aversion γ is thus multiplied by a factor that does not essentially vary as a function of Δt . In good approximation, γ turns out to be a parameter independent of the size of Δt . It is thus reasonable to assume the same γ for different interval sizes Δt . Annualized effective returns $x_{eff, ann}$, computed for different intervals Δt , can be directly compared.

A comparable approach has been followed by Hodges [9] to construct a generalized Sharpe Ratio but his measure needs the explicit probability distribution of the returns because it is based on a maximization of the utility function with respect to the exposure. Moreover, the question of the measurement frequency is not treated in his paper.

2.1. A multi-horizon performance measure

The measure in Eq. (8) still depends on the time horizon Δt and is insensitive to changes occurring with much longer or much shorter horizons, as discussed in Section 1.1. To remedy this problem, a weighted average of the $x_{eff, ann}$ is computed with n different time horizons Δt_i :

$$X_{eff} = \frac{\sum_{i=1}^n w_i x_{eff, ann}(\Delta t_i)}{\sum_{i=1}^n w_i}, \quad (9)$$

where the weights w_i can be chosen according to the relative importance of the time horizons Δt_i and may differ for trading models with different trading frequencies. This equation takes advantage of the fact that *annualized* $x_{eff, ann}$ values have no *systematic* dependence on the horizon Δt_i . Substituting $x_{eff, ann}$ from Eq. (8), X_{eff} becomes

$$X_{eff} = \bar{X} - \frac{1}{2} \frac{\sum_{i=1}^n \gamma_i w_i (1 \text{ year}/\Delta t_i) \sigma_i^2}{\sum_{i=1}^n w_i}, \quad (10)$$

where the variable \bar{X} is still the average yearly return over the whole sample,⁶ γ_i is the risk aversion constant for each horizon, and σ_i^2 is the variance computed for the time horizon Δt_i .

In the discussion of Eq. (8), it is shown that the risk aversion γ_i has no systematic dependence on the size of the horizon Δt_i . However, investors using a trading model might perceive the risks of different horizons differently, and γ_i can be set according to their trading preferences. For simplicity, we will assume a common risk aversion parameter, γ , and put a tilde to the weighting parameter \tilde{w}_i as it now reflects not only the weight but also the specific risk aversion of the horizon. Thus, the final form of

⁵ It is exact if x is a Gaussian random walk with linear drift.

⁶ $X_{eff} \leq \bar{X}$ as $\sigma^2 > 0$.

the performance indicator is

$$X_{eff} = \bar{X} - \frac{\gamma}{2} \frac{\sum_{i=1}^n \tilde{w}_i (1 \text{ year}) \sigma_i^2 / \Delta t_i}{\sum_{i=1}^n \tilde{w}_i} . \tag{11}$$

When computing X_{eff} , we need a reasonable value of the risk aversion γ . We determined γ by presenting various return curves of different kinds to market practitioners and asking them about their preference for future investments in these assets (assuming unchanged leverage in the future). The intuitive ranking of these return curves has to coincide with the ranking of the computed X_{eff} values. This is the case if γ is chosen between 0.08 and 0.15. Although we recommend such γ values, users may adapt γ to their own level of risk aversions which may differ from that of typical market practitioners. The weights \tilde{w}_i are determined with a weighting function which allows the selection of the relative importance of the different horizons. The time horizons Δt_i are assumed to be in a geometric sequence such as 1, 2, 4, 8, ... days. The weighting function used in the computation of X_{eff} is

$$\tilde{w}(\Delta t) = \frac{1}{2 + (\ln \Delta t / 90 \text{ days})^2} . \tag{12}$$

This choice is motivated as follows. First, this weighting function encompasses a wide range of interval sizes Δt . Drawdowns of many different sizes are captured, as explained in Section 1.1. Second, there are limits. The weighting function smoothly declines on both sides of a maximum at around three months as a function of $\ln \Delta t$. Extremely long intervals have a low weight, because the lifetime of the whole investment, as well as the period of available and relevant historical data, is limited to a few years. On the side of short intervals, there is also a limit: most investors do not regard tiny intra-day oscillations of their portfolio value as relevant for their investment. Third, the maximum of $\tilde{w}(\Delta t)$ is at $\Delta t = 90$ days (also counting weekends, $\Delta t \approx$ three months), which is a typical time horizon for many investors. This choice is reasonable, but special short-term or extremely long-term investors are free to shift the maximum weight to other interval sizes.

3. Effective return with variable risk aversion, R_{eff}

The X_{eff} is based on the assumption of constant risk aversion over the full $\Delta \tilde{R}$ axis. A natural generalization of this approach is to distinguish the clustering of losses from the clustering of gains. In X_{eff} , they both contribute to the same amount in the risk term. For R_{eff} , we shall assume that the investor has a stronger risk aversion to clustering of losses than gains, as also found by Benartzi and Thaler [10]. Thus, the R_{eff} algorithm includes two risk aversion levels: a low one, ϱ_+ , for positive $\Delta \tilde{R}$ (profit intervals) and a high one, ϱ_- , for negative $\Delta \tilde{R}$ (drawdowns):

$$\varrho = \begin{cases} \varrho_+ & \text{for } \Delta \tilde{R} \geq 0, \\ \varrho_- & \text{for } \Delta \tilde{R} < 0, \end{cases} \quad \varrho_+ < \varrho_- . \tag{13}$$

The high value of ϱ_- reflects the high risk aversion of typical market participants in the loss zone. Trading models may have some losses, but if the loss observations strongly vary in size, the risk of very large losses becomes unacceptably high. On the side of the positive profit observations, a certain regularity of profits is also better than a strong variation in size. However, this distribution of positive returns is never as vital for the future of a market participant as the distribution of losses (drawdowns). Therefore, ϱ_+ is much smaller than ϱ_- . In Müller et al. [11], ϱ_+ is set to be equal to $\varrho_-/4$ and the values are chosen⁷ to be $\varrho_- = 0.20$, $\varrho_+ = 0.05$.

The utility function is obtained by inserting Eq. (13) into the definition of constant risk aversion and integrating twice over ΔR :

$$u = u(\Delta\tilde{R}) = \begin{cases} -\frac{e^{-\varrho_+ \Delta\tilde{R}}}{\varrho_+} & \text{for } \Delta\tilde{R} \geq 0, \\ \frac{1}{\varrho_-} - \frac{1}{\varrho_+} - \frac{e^{-\varrho_- \Delta\tilde{R}}}{\varrho_-} & \text{for } \Delta\tilde{R} < 0. \end{cases} \tag{14}$$

The utility function $u(\Delta\tilde{R})$ is monotonically increasing and reaches its maximum 0 when $\Delta\tilde{R} \rightarrow \infty$ (infinite profit). All other utility values are negative. The utility function is assumed to be continuously differentiable although the risk is not continuous. The return is obtained by inverting the utility function so that

$$\Delta\tilde{R} = \Delta\tilde{R}(u) = \begin{cases} -\frac{\log(-\varrho_+ + u)}{\varrho_+} & \text{for } u \geq -\frac{1}{\varrho_+}, \\ -\frac{\log(1 - \frac{\varrho_-}{\varrho_+} - \varrho_- u)}{\varrho_-} & \text{for } u < -\frac{1}{\varrho_+}. \end{cases} \tag{15}$$

3.1. From single to multi-horizon performance measures

Similar to X_{eff} , R_{eff} has a multi-horizon weighting scheme. Several interval sizes Δt_j are considered. In addition to this, there is a scheme of overlapping intervals for each interval size. Overlapping intervals are a means to enhance the significance of statistical results [12] and the multi-horizon scheme is a means to capture the main drawdown periods as explained in Section 1.1.

The computation of the utility for one time horizon Δt_j is explained first. Many overlapping intervals, all of them of length Δt_j , are considered. In the i th such interval, we observe a return $\Delta\tilde{R}_{ji}$ which is defined as follows:

$$\Delta\tilde{R}_{ji} = \tilde{R}(t_{end\ ji} - \tilde{R}(t_{start\ ji}), \quad t_{end\ ji} - t_{start\ ji} = \Delta t_j, \tag{16}$$

where the overlapping intervals are phase-shifted,

$$t_{start\ ji} = \left(\frac{i}{m} - 1\right) \Delta t_j, \quad t_{end\ ji} = \frac{i}{m} \Delta t_j \tag{17}$$

with an overlapping factor m and

$$i = 1 \dots N_j, \quad N_j = \text{Max}(i | t_{start\ ji} < T). \tag{18}$$

⁷ These values are under the assumption that the return is measured in percentage. They have to be multiplied by 100 if the returns are not expressed in percentage.

In this formula, the number N_j of the observed intervals is chosen to include every interval that overlaps with the total sample from $t = 0$ to T . The utility of the i th observation is

$$u_{ji} = u(\Delta\tilde{R}_{ji}) \tag{19}$$

with the utility function $u(\Delta\tilde{R})$ of Eq. (14). The mean utility of the full sample is

$$u_j = \frac{\sum_{i=1}^{N_j} \Delta t_{ji} u_{ji}}{\sum_{i=1}^{N_j} \Delta t_{ji}}, \tag{20}$$

where the weight Δt_{ji} of the observation is the amount of time during which the i th analysis interval coincides with the sample. Observations inside the sample have a full weight of Δt_j and Δt_j is the maximum value that Δt_{ji} can take,

$$\Delta t_{ji} \leq \Delta t_j. \tag{21}$$

The mean utility u_j can be transformed back to an effective return value by applying Eq. (15) such that

$$\Delta\tilde{R}_{eff\ j} = \Delta\tilde{R}(u_j). \tag{22}$$

This $\Delta\tilde{R}_{eff\ j}$ is the typical, effective return for the horizon Δt_j , but it is not yet annualized. Annualization is necessary for a consistent, universal R_{eff} definition. It means correcting the resulting $\Delta\tilde{R}_{eff\ j}$ by multiplying the ratio of one year to the analysis interval size. The annualized effective return is defined as

$$R_{eff\ j} = \frac{1 \text{ year}}{\Delta t_{eff\ j}} \Delta\tilde{R}_{eff\ j}, \tag{23}$$

where

$$\Delta t_{eff\ j} = \frac{\sum_{i=1}^{N_j} (\Delta t_{ji})^2}{\sum_{i=1}^{N_j} \Delta t_{ji}}. \tag{24}$$

This is the final form of the effective return for one horizon, Δt_j .

This measure is now generalized to a multi-horizon measure, taking a sequence of different horizons. Similarly to X_{eff} , R_{eff} is defined as a weighted mean over n different horizons Δt_j (which are typically chosen as a geometric sequence such as 1, 2, 4, 8, ... days):

$$R_{eff} = \frac{\sum_{j=1}^n w_j R_{eff\ j}}{\sum_{j=1}^n w_j} \tag{25}$$

with the weights w_j depending on the horizon Δt_j as in Eq. (12).

4. Simulation methodology

The distributions of the performance measures under various null processes will be calculated by using a simulation methodology. The random walk process is defined by

$$r_t = \alpha + \varepsilon_t, \tag{26}$$

where $r_t = \log(p_t/p_{t-1})$ and $\varepsilon_t \sim N(0, \sigma^2)$. The random walk estimation involves the regression of the actual foreign exchange returns on a constant. A simulation sample for the random walk series with drift is obtained by sampling from the Gaussian random number generator with the mean and the standard deviation of the residual series. The simulated residuals are added to the conditional mean defined by the estimator $\hat{\alpha}$, to form a new series of returns. The new series of the returns has the same drift in prices and the same variance. From the new series of returns, the simulated price process is recovered recursively by setting the initial price to the true price at the beginning of the sample. The trading models use the bid and ask prices as inputs. Half of the average spread is subtracted (added) from the simulated price process to obtain the simulated bid and ask prices.

The GARCH(1,1) process is written as

$$r_t = \gamma_0 + \varepsilon_t, \quad (27)$$

where $\varepsilon_t = h_t^{1/2} z_t$, $z_t \sim N(0, 1)$ and $h_t = \alpha_0 + \alpha_1 h_{t-1} + \beta_1 \varepsilon_{t-1}^2$. GARCH specification [13] allows the conditional second moments of the return process to be serially correlated. This specification implies that periods of high (low) volatility are likely to be followed by periods of high (low) volatility. GARCH specification allows the volatility to change over time and the expected returns are a function of past returns as well as volatility. The parameters and the normalized residuals are estimated from the foreign exchange returns using the maximum likelihood procedure. The simulated returns for the GARCH(1,1) process are generated from the simulated normalized residuals and the estimated parameters. The AR(p)-GARCH(1,1) process is written as

$$r_t = \gamma_0 + \sum_{i=1}^p \gamma_i r_{t-i} + \varepsilon_t, \quad (28)$$

where $\varepsilon_t = h_t^{1/2} z_t$, $z_t \sim N(0, 1)$ and $h_t = \alpha_0 + \alpha_1 h_{t-1} + \beta_1 \varepsilon_{t-1}^2$. The estimated parameters of the AR(p)-GARCH(1,1) processes together with the simulated residuals are used to generate the simulated returns from these processes. As before, half of the average spread is subtracted (added) from the simulated price process to obtain the simulated bid (ask) prices.

The simulation method is applied to two trading models which represent different investment strategies to test the quality of the performance measures. They are described below.

5. Trading models

The purpose of the simulations is to test performance measures rather than trading model algorithms. However, since trading models are used in the simulations, a description of these models may be useful here.

A distinction should be made between a price change forecast and an actual trading recommendation. A trading recommendation naturally includes a price change forecast,

but it must also account for the specific constraints of the dealer of the respective trading model because a trading model is constrained by its past trading history and the positions to which it is committed. A price forecasting model, on the other hand, is not limited to similar types of constraints. A trading model thus goes beyond predicting a price change such that it must decide if and at what time a certain action has to be taken.

Trading models offer a real-time analysis of foreign exchange movements and generate explicit trading recommendations. These models are based on the continuous collection and treatment of foreign exchange quotes by market makers around the clock at the tick-by-tick frequency level. There are many important reasons to utilize high-frequency data in the real-time trading models. Among them, the model indicators acquire robustness by utilizing the intraday volatility behavior in their build-up. Another reason is that any position taken by the model may need to be reversed quickly although these position reversals may not need to be observed often. The stop-loss objectives need to be satisfied and the high-frequency data provides an appropriate platform for this requirement. Using high-frequency data reduces the risk of slippage by giving signals at unpredictable times (contrary to models based on daily closing data). More importantly, the customer's trading positions and strategies within a trading model can only be replicated with a high statistical degree of accuracy by utilizing high-frequency data in a real-time trading model.

The trading models imitate the trading conditions of the real foreign exchange market as closely as possible. They do not deal directly but instead instruct human foreign exchange dealers to make specific trades. In order to imitate real-world trading accurately, they take transaction costs into account in their return computation, they do not trade outside market working hours except for executing stop-loss and they avoid trading too frequently. In short, these models act realistically in a manner which a human dealer can easily follow.

Every trading model is associated with a local market that is identified with a corresponding geographical region. In turn, this is associated with generally accepted office hours and public holidays. The local market is defined to be open at any time during office hours provided it is neither a weekend nor a public holiday. The O&A trading models presently support the Zurich, London, Frankfurt, Vienna and New York markets. Typical opening hours for a model are between 8:00 and 17:30 local time, the exact times depending on the particular local market.

The central part of a trading model is the analysis of the past price movements which are summarized within a trading model in terms of indicators. The indicators are then mapped into actual trading positions by applying various rules. For instance, a model may enter a long position if an indicator exceeds a certain threshold. Other rules determine whether a deal may be made at all. Among various factors, these rules determine the timing of the recommendation. A trading model thus consists of a set of indicator computations combined with a collection of rules. The former are functions of the price history. The latter determine the applicability of the indicator computations to generating trading recommendations. The model gives a recommendation not only for

the direction but also for the amount of the exposure. The possible exposures (gearing) are $\pm\frac{1}{2}$, ± 1 or 0 (no exposure).

5.1. The real-time trading (RTT) model

The real-time trading model studied in this paper is classified as a one-horizon, high-risk/high-return model [3]. The RTT is a trend-following model and takes positions when an indicator crosses a threshold. The indicator is a momentum based on specially weighted moving averages with repeated application of the exponential moving average operator. In case of extreme foreign exchange movements, however, the model adopts an overbought/oversold (contrarian) behavior and recommends taking a position against the current trend. The contrarian strategy is governed by rules that take the recent trading history of the model into account. The RTT model goes neutral only to save profits or when a stop-loss is reached. Its profit objective is typically at 3%. When this objective is reached, a gliding stop-loss prevents the model from losing a large part of the profit already made by triggering its going neutral when the market reverses. The gearing function for the RTT is

$$g(I_p) = \text{sign}(I_p) f(|I_p|) c(I)$$

such that

$$I_p = p - MA_p^4 (\tau \approx 20)$$

and

$$f(|I_p|) = \begin{cases} 1, & \text{if } |I_p| > b, \\ 0.5, & \text{if } a < |I_p| < b, \\ 0, & \text{if } |I_p| < a, \end{cases}$$

and

$$c(I) = \begin{cases} +1 & \text{if } |I_p| < d, \\ -1 & \text{if } |I_p| > d, \end{cases}$$

where $a < b < d$. I_p is the indicator function where p is the logarithmic price and MA_p^4 is the fourth-order iterative exponential moving average (EMA) operator.⁸ $f(|I_p|)$ measures the size of the signal and $c(|I_p|)$ acts as a contrarian strategy. a and b are functions of *current position, volatility and trading frequency*. d is a function of *position in, previous position in, sign of the return of the previous position*.

Since X_{eff} and R_{eff} are implicit functions of the gearing function, the optimization of the RTT model is based on the X_{eff} and R_{eff} performance. The model is subject to the open-close and holiday closing hours. The model has maximum stop-loss and maximum gain limits set by the environment.

⁸ A detailed explanation of iterated exponential moving average kernels are presented in Ref. [14].

5.2. A simple exponential moving average (EMA) model

Another investment strategy used in this study is based on a very simple indicator which is popular among traders. The EMA model indicator is a momentum-based indicator consisting of a difference between two exponential moving averages of range $\tau = 0.5$ and 20 days. The gearing function for the EMA model is

$$g(I_p) = \text{sign}(I_p) f(|I_p|),$$

where

$$I_p = \text{EMA}(\tau = 0.5) - \text{EMA}(\tau = 20)$$

and

$$f(|I_p|) = \begin{cases} \text{if } |I_p| > a, 1, \\ \text{if } |I_p| < a, 0, \end{cases}$$

where $a > 0$. The model is subject to the open-close and holiday closing hours. The model has maximum stop-loss and profit objective which are set to the same values as in the RTT model.

6. Empirical results

The methodology of this section places a historical realization in the simulated distribution of the performance measure under the assumed process and calculates its p -value.⁹ This tells us whether the historical realization is likely to be generated from this particular distribution or not. More importantly, it would tell whether the historical performance is likely to occur in the future. A small p -value (less than 5%) indicates that the historical performance lies in the right tail (or the left tail) and the studied performance distribution is not representative of the data generating process assuming that the trading model is a good one. If the process which generated the performance distribution is close to the data generating process of the foreign exchange returns, the historical performance would lie within two standard deviations of the performance distribution indicating that the studied process may be retained as the representative of the data generating process.

⁹ The p -value represents a decreasing ratio of the reliability of a result. The higher the p -value, the less we can believe that the observed relation between variables in the samples is a reliable indicator of the relation between the respective variables in the population. Specifically, the p -value represents the probability of error that is involved in accepting our observed result as valid, that is, as *representative of the population*. For example, the p -value of 0.05 indicates that there is a 5% probability that the relation between the variables found in our sample is a *fluke*. In other words, assuming that in the population there was no relation between those variables whatsoever, and by repeating the experiment, we could expect that approximately every 20 replications of the experiment there would be one in which the relation between the variables in question would be equal or stronger than ours. In many areas of research, the p -value of 5% is treated as a *borderline acceptable* level.

Table 1
GARCH(1,1) parameter estimates, 1990–1996, 5 min frequency^a

	USD-DEM	USD-CHF	USD-FRF	DEM-JPY
α_0	4.95 (4.23)	0.11 (0.12)	9.38 (7.09)	2.97 (4.03)
α_1	0.1111 (0.0005)	0.1032 (0.0007)	0.1572 (0.0007)	0.0910 (0.0005)
β_1	0.8622 (0.0007)	0.8578 (0.0009)	0.8137 (0.0009)	0.8988 (0.0006)
LL	6.45	6.17	6.29	6.34
$Q(12)$	5.08	32.96	4.04	55.94
$\hat{\varepsilon}_{\sigma^2}$	1.04	1.03	1.07	1.05
$\hat{\varepsilon}_{sk}$	-0.07	-0.03	-0.05	0.16
$\hat{\varepsilon}_{ku}$	11.73	7.28	22.93	27.73

^aNotes: α_0 values are 10^{-9} . The numbers in parentheses are the standard errors. The standard errors of α_0 are 10^{-11} . LL is the average log likelihood value. $Q(12)$ is the Ljung and Box portmanteau test for serial correlation and distributed χ^2 with 12 degrees of freedom. The $\chi^2_{0.05}(12)$ is 21.03. $\hat{\varepsilon}_{\sigma^2}$, $\hat{\varepsilon}_{sk}$ and $\hat{\varepsilon}_{ku}$ are the variance, skewness and the excess kurtosis of the residuals.

The data is the 5 min¹⁰ ϑ -time series from January 1, 1990 to December 31, 1996 for the three major foreign exchange rates, USD-DEM, USD-CHF (Swiss Franc), USD-FRF (French Franc), and the cross-rate DEM-JPY (Deutsche Mark-Japanese Yen). The high-frequency data inherits intra-day seasonalities and requires deseasonalization. This paper uses the deseasonalization methodology advocated in Ref. [15] named as the ϑ -time seasonality correction method. The ϑ -time method uses the business time scale and utilizes the average volatility to represent the activity of the market. The ϑ -time method is based on three geographical markets namely the East Asia, Europe and the North America. A more detailed exposition of the ϑ methodology is presented in Ref. [15]. The optimization and the validation of the trading models are done with data prior to January 1, 1990. Therefore, our results here provide a complete *ex-ante* test for the trading performance measures under the studied processes with seven years of 5 min frequency data. The simulations for each process are done for 1000 replications.

The GARCH(1,1) estimation results are presented in Table 1. The numbers in parentheses are the robust standard errors and the GARCH(1,1) parameters are statistically significant at the 5% level for all currency pairs. The Ljung-Box statistic is calculated up to 12 lags for the standardized residuals and it is distributed with χ^2 with 12 degrees of freedom. The Ljung-Box statistic indicate no serial correlation for the USD-DEM and USD-FRF but the URD-FRF and DEM-JPY remains serially correlated. The variance of the normalized residuals are near one. There is no evidence of skewness but the excess kurtosis remains large for the residuals.

A further direction is to investigate whether a conditional mean dynamics with GARCH(1,1) innovations would be a more successful characterization of the dynamics

¹⁰ The real-time system uses tick-by-tick data for its trading recommendations. The historical realizations and the simulations in this paper are carried out with 5 min data as it is computationally expensive to use the tick-by-tick data for the simulations. Although, the data frequency used in this paper is slightly lower, the historical performance of the currency pairs from the 5 min series are exactly compatible with performance of the real-time trading model which utilize the tick-by-tick data.

Table 2
AR(4)-GARCH(1,1) parameter estimates, 1990–1996, 5 min frequency^a

	USD-DEM	USD-CHF	USD-FRF	DEM-JPY
α_0	3.90 (3.40)	8.19 (9.03)	7.28 (5.80)	2.92 (3.93)
α_1	0.099 (0.0005)	0.0874 (0.0006)	0.1349 (0.0007)	0.088 (0.0005)
β_1	0.8796 (0.0006)	0.8833 (0.0007)	0.8411 (0.0008)	0.9008 (0.0006)
γ_1	-0.176 (0.001)	-0.208 (0.001)	-0.200 (0.002)	-0.130 (0.002)
γ_2	-0.011 (0.001)	-0.031 (0.002)	-0.025 (0.002)	-0.090 (0.002)
γ_3	0.003 (0.001)	-0.001 (0.002)	-0.005 (0.002)	-0.005 (0.002)
γ_4	-0.004 (0.001)	-0.002 (0.001)	-0.008 (0.002)	-0.010 (0.002)
LL	6.46	6.19	6.30	6.35
$Q(12)$	0.21	0.69	0.10	0.81
$\hat{\epsilon}_{\sigma^2}$	1.04	1.03	1.07	1.05
$\hat{\epsilon}_{sk}$	-0.07	-0.04	-0.05	0.15
$\hat{\epsilon}_{ku}$	12.29	7.86	21.84	27.98

^aNotes: α_0 values are 10^{-9} . The numbers in parentheses are the standard errors. The standard errors of α_0 are 10^{-11} . LL is the average log likelihood value. $Q(12)$ refer to the Ljung-Box portmanteau test for serial correlation and it is distributed χ^2 with 12 degrees of freedom. The $\chi^2_{0.05}(12)$ is 21.03. $\hat{\epsilon}_{\sigma^2}$, $\hat{\epsilon}_{sk}$ and $\hat{\epsilon}_{ku}$ are the variance, skewness and the excess kurtosis of the residuals.

of the high-frequency foreign exchange returns. The conditional means of the foreign exchange returns are estimated with four lags of these returns. The additional lags did not lead to substantial increases in the likelihood value. The results of the AR(4)-GARCH(1,1) are presented in Table 2. The numbers in parentheses are the robust standard errors and all four lags are statistically significant at the 5% level. The negative autocorrelation is large and highly significant for the first lag of the returns. This is consistent with the high-frequency behavior of the foreign exchange returns and is also observed in Ref. [15]. The Ljung-Box statistic indicates no serial correlation in the normalized residuals. The variance of the normalized residuals are near one. There is no evidence of skewness but the excess kurtosis remains large for the residuals.

The simulations for the RTT model are reported in Table 3. Each column reports a performance measure which are the annualized return, X_{eff} , R_{eff} , Sharpe Ratio and the maximum drawdown.¹¹ There are four row panels corresponding to USD-DEM (US Dollar-Deutsche Mark), USD-CHF (US Dollar-Swiss Franc), USD-FRF (US Dollar-French Franc) and DEM-JPY (Deutsche Mark-US Dollar). In each row panel, the actual historical realizations of the foreign exchange rates and the p -values of the three processes are reported. All p -values are reported in percentage terms.

For the USD-CHF, the annualized return, X_{eff} , R_{eff} , Sharpe Ratio and the maximum drawdown are 3.65%, -1.77%, -4.42%, 0.29% and 16.08%, respectively. It is the worst strategy of all those presented here. The RTT model generates an annualized return of 3.65% after transactions costs. It is interesting to see that X_{eff} and R_{eff} are

¹¹ The maximum drawdown is a measure of risk often used by practitioners but it does not give an assessment of the performance per say and is sample dependent in that it will tend to increase with the sample.

Table 3
Real-time trading (RTT) model^{1a}

Description	Annual return	X_{eff}	R_{eff}	Sharpe	Max. drawdown
USD-DEM	9.63	3.78	4.43	0.87	11.02
p -value (Random walk)	0.0	0.0	0.0	0.0	0.0
p -value (GARCH(1,1))	0.4	0.1	0.0	0.3	0.0
p -value (AR(4)-GARCH(1,1))	0.1	0.1	0.0	0.1	0.0
USD-CHF	3.65	-1.77	-4.42	0.29	16.08
p -value (Random walk)	8.9	0.7	0.6	6.3	0.0
p -value (GARCH(1,1))	8.4	1.4	0.9	6.4	0.1
p -value (AR(4)-GARCH(1,1))	3.7	1.9	2.3	4.6	0.1
USD-FRF	8.20	4.80	4.95	0.94	11.36
p -value (Random walk)	1.2	0.2	0.1	0.2	0.0
p -value (GARCH(1,1))	0.9	0.1	0.1	0.1	0.0
p -value (AR(4)-GARCH(1,1))	0.3	0.2	0.1	0.0	0.0
DEM-JPY	6.43	3.81	3.45	0.69	12.03
p -value (Random walk)	2.1	0.2	0.1	0.6	0.0
p -value (GARCH(1,1))	0.5	0.1	0.1	0.4	0.0
p -value (AR(4)-GARCH(1,1))	0.5	0.1	0.1	0.4	0.0

^aNotes: The rows corresponding to USD-DEM, USD-CHF, USD-FRF and DEM-JPY present the performance of the trading model with the actual foreign exchange series from January 1, 1990 until December 31, 1996 with 5 min frequency. The p -values are calculated from 1000 replications of the corresponding stochastic process. The simulated returns for the corresponding stochastic processes are generated from the simulated residuals and the estimated parameters. From the new series of returns, the simulated price process is recovered recursively by setting the initial price to the true price at the beginning of the sample. The trading models use the bid and ask prices as inputs. Half of the average spread is subtracted (added) from the simulated price process to obtain the simulated bid and ask prices. p -values are reported in % (e.g. 8.4 refers to 8.4%).

both negative indicating that although the annualized return is positive after transaction costs it does not cover the *cost* of risk taking by the model. The fact that R_{eff} is much smaller than X_{eff} indicates the presence of a strong clustering of losses which reflects also the large maximum drawdown of 16.08%. The corresponding p -values are 8.9, 0.7, 0.6, 6.3 and 0.0. Based on the p -value of the annualized return, the random walk hypothesis cannot be rejected at the 5% level as the p -value remains at 8.9%. The p -value of the Sharpe Ratio is at 6.3% which also does not provide strong evidence against the random walk hypothesis. For the X_{eff} and R_{eff} , the p -values are 0.7% and 0.6%, respectively. Amongst the four performance measures which utilize return in their calculations, the R_{eff} and X_{eff} provide the smaller two p -values. The maximum drawdown provides the smallest p -value which is 0% but, as mentioned above, this is really not a performance measure but rather a pure risk measure. The ordering of the p -values for the GARCH(1,1) and the AR(4)-GARCH(1,1) processes remains the same as in the case of the random walk process.¹²

¹² A detailed analysis for the rejection of the random walk, GARCH(1,1) and AR(4)-GARCH(1,1) as possible data generating mechanisms for the foreign exchange dynamics is analyzed in Ref. [14].

Table 4
Exponential moving average (EMA) model^a

Description	Annual return	X_{eff}	R_{eff}	Sharpe	Max. drawdown
USD-DEM	3.33	-0.67	-2.48	0.30	17.73
<i>p</i> -value (Random walk)	12.6	2.8	2.3	10.8	1.5
<i>p</i> -value (GARCH(1,1))	14.1	2.3	1.8	11.3	0.8
<i>p</i> -value (AR(4)-GARCH(1,1))	10.0	3.8	3.7	5.8	3.1
USD-CHF	4.40	-0.46	-3.00	0.35	17.10
<i>p</i> -value (Random walk)	8.8	0.9	0.4	6.4	0.1
<i>p</i> -value (GARCH(1,1))	9.6	0.7	0.5	7.0	1.0
<i>p</i> -value (AR(4)-GARCH(1,1))	5.9	1.7	2.4	6.4	1.0
USD-FRF	6.01	2.25	2.05	0.61	14.52
<i>p</i> -value (Random walk)	6.4	0.6	0.4	2.6	0.4
<i>p</i> -value (GARCH(1,1))	4.6	0.6	0.4	1.7	0.0
<i>p</i> -value (AR(4)-GARCH(1,1))	3.2	0.7	0.6	2.1	0.2
DEM-JPY	7.09	2.63	2.97	0.69	13.06
<i>p</i> -value (Random walk)	6.6	1.4	1.0	2.8	0.1
<i>p</i> -value (GARCH(1,1))	2.8	0.5	0.2	1.5	0.2
<i>p</i> -value (AR(4)-GARCH(1,1))	1.8	0.5	0.4	1.2	0.2

^aNotes: The rows corresponding to USD-DEM, USD-CHF, USD-FRF and DEM-JPY present the performance of the trading model with the actual foreign exchange series from January 1, 1990 until December 31, 1996 with 5 min frequency. The *p*-values are calculated from 1000 replications of the corresponding stochastic process. The simulated returns for the corresponding stochastic processes are generated from the simulated residuals and the estimated parameters. From the new series of returns, the simulated price process is recovered recursively by setting the initial price to the true price at the beginning of the sample. The trading models use the bid and ask prices as inputs. Half of the average spread is subtracted (added) from the simulated price process to obtain the simulated bid and ask prices. *p*-values are reported in % (e.g. 8.4 refers to 8.4%).

In fact, for all four currencies that we have studied here, the ordering of the *p*-values for all three processes remains in similar order. Therefore, we may conclude that the ordering of the *p*-values are quite robust across different currencies and different stochastic processes. Those performance measures which utilize limited information from the entire equity path provide larger *p*-values whereas performance measures which utilize all information from the equity curve yield substantially smaller *p*-values.

A comparison between X_{eff} and R_{eff} for the real performance is interesting. It is not always true that X_{eff} is bigger than R_{eff} . In two of the best trading strategies (USD-DEM and USD-FRF) $R_{eff} > X_{eff}$ showing that penalizing clustering of returns regardless of their signs leads sometimes to an overestimation of the risk.

In Table 4, the results for the simple EMA model are presented. They indicate that a simple EMA model provides net positive annualized returns after taking the transaction costs into account for all four currencies although overall these results are worse than the results achieved by the RTT models.

The examination of the p -values with the same three processes indicates similar ordering of the p -values. For the USD-DEM, the p -values of the random walk process indicate 12.6% for the annualized return and 10.8% for the Sharpe Ratio. Based on these p -values, the null hypothesis that the random walk process is the data generating mechanism of the foreign exchange returns is retained. The X_{eff} , R_{eff} and maximum drawdown, on the other hand, provide substantially smaller p -values. For the X_{eff} , R_{eff} and maximum drawdown, the p -values are 2.8%, 2.3% and 2.5%, respectively. A similar discrepancy occurs between the p -values of the performance measures under the GARCH(1,1) and AR(4)-GARCH(1,1) processes for the USD-DEM. In particular, the p -value of the annualized return and the Sharpe Ratio remains over 10% whereas the p -values of the other remaining three performance measures are less than 3%.

Comparing the actual results of the EMA models and the RTT models, X_{eff} and R_{eff} are better for the RTT models (except for USD-CHF). This is confirmed by the larger drawdowns which the EMA models incurred. The ranking given by those two measures seems the most consistent looking at the different behavior of the equity curves.

7. Conclusions

We present in this paper two performance measures for investment strategies: X_{eff} and R_{eff} . Both are given in the form of risk-adjusted returns. Risk adjustment takes the intuitively well understood and numerically stable form of a risk premium deducted from the annual return. The use of the utility function approach allows us to take into account directly the risk aversion of the investors. By summing up those measures computed over different frequencies for the returns, the measures include an examination of the clustering behavior of the returns at different frequencies. Owing to also considering low frequency (long time intervals), long clusters of losses are directly captured, resulting in a behavior similar to those performance measures that model risk aversion as relative to wealth.

If X_{eff} gives a measure that is based on constant positive risk aversion, R_{eff} allows us to differentiate between clustering of positive and negative returns assigning a higher risk to these periods with negative returns. Moreover, R_{eff} does not need any distributional assumption on the returns which makes the measure better suited to fat tailed returns.

An empirical study of the behavior of these two measures to evaluate trading strategies in the FX market shows their robustness against randomness and their ability to fully characterize the strategies. They allow the true performance of our trading models to be revealed. Contrary to the Sharpe Ratio, they do not suffer from the limitations of Gaussian assumptions on the distribution of returns. These performance measures have proven their usefulness in optimizing trading model parameters [3] and can also be used to rank different trading strategies as shown in this paper. Since they are not linked to particular trading models they can be used in a more general context for evaluating the outcome of any investment strategy.

Acknowledgements

R. Gençay is grateful to Olsen & Associates for their generosity and for providing an excellent research platform for high frequency finance research. Gençay also thanks the Social Sciences and Humanities Research Council of Canada and the Natural Sciences and Engineering Research Council of Canada for financial support.

References

- [1] W.F. Sharpe, Mutual fund performance, *J. Bus.* 39 (1) (1966) 119–138.
- [2] W.F. Sharpe, The sharpe ratio, *J. Portfolio Manage.* 21 (1994) 49–59.
- [3] O.V. Pictet, M.M. Dacorogna, U.A. Müller, R.B. Olsen, J.R. Ward, Real-time trading models for foreign exchange rates, *Neural Network World* 6 (1992) 713–744.
- [4] H.M. Markowitz, *Mean–Variance Analysis in Portfolio Choice and Capital Markets*, Basil Blackwell, Oxford, Cambridge, 1987.
- [5] N. Barberis, M. Huang, T. Santos, Prospect theory and asset prices, Technical Report, University of Chicago, GSB, 1999, pp. 1–48.
- [6] M. Stutzer, A large deviations approach to portfolio analysis, Technical Report, University of Iowa, Iowa City, IA, 1998, pp. 1–34.
- [7] M.M. Dacorogna, U.A. Müller, O.V. Pictet, A measure of trading model performance with a risk component, Olsen Research Institute Discussion Paper, Zurich, Switzerland, 1991.
- [8] R.L. Keeney, H. Raiffa, *Decision with Multiple Objectives: Preferences and Value Tradeoffs*, Wiley, New York, 1976.
- [9] S. Hodges, A generalization of the sharpe ratio and its applications to valuation bounds and risk measures, preprint of the financial options research center of the University of Warwick, presented at the Newton Institute in Cambridge October 1998, pp. 1–17.
- [10] S. Benartzi, R.H. Thaler, Myopic loss aversion and the equity premium puzzle, *Quart. J. Econ.* 110 (1995) 73–92.
- [11] U.A. Müller, M.M. Dacorogna, O.V. Pictet, A trading model performance measure with strong risk aversion against drawdowns, Olsen Research Institute Discussion Paper, Zurich, Switzerland, 1993.
- [12] U.A. Müller, Statistics of variables observed over overlapping intervals, Olsen Research Institute Discussion Paper, Zurich, Switzerland, 1993.
- [13] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, *J. Econometrics* 31 (1986) 307–327.
- [14] R. Gençay, G. Ballocci, M.M. Dacorogna, R.B. Olsen, O. Pictet, Real-time trading models and the statistical properties of foreign exchange rates, *International Economic Review*, 2000, forthcoming.
- [15] M.M. Dacorogna, U.A. Müller, R.J. Nagler, R.B. Olsen, O.V. Pictet, A geographical model for the daily and weekly seasonal volatility in the foreign exchange market, *J. Int. Money Finance* 12 (1993) 413–438.