On the Consistency of Performance Measures for Hedge Funds

Huyen Nguyen-Thi-Thanh*

Abstract

This article presents a comparative study of ten measures documented as the most used by researchers and practitioners: Sharpe, Sortino, Calmar, Sterling, Burke, Stutzer, modified Sharpe, upside potential ratio, Omega and AIRAP. In examining the modifications of fund performance in terms of ranks and deciles when the performance measure changes, numerous significant modifications were observed despite strong positive correlations among fund rankings established by different measures. A clustering of these measures by means of the ascendant hierarchical clustering technique shows that Sharpe, Stutzer, Omega and Sortino are the most consistent whilst AIRAP index appears to be the most distinct from its peers. Besides, some measures are found to be more stable or persistent than the others in measuring hedge fund performance.

Keywords: hedge funds, investment funds, performance evaluation, performance measure, Sharpe ratio.

JEL Classification Codes: G2, G11, G15

Biographical notes: Huyen Nguyen-Thi-Thanh is currently an Assistant Professor in Finance at the University of Maine (France) and an affiliated researcher at La Rochelle Business School (France). She received her PhD in 2007. Her main research interests are hedge funds, mutual funds, performance evaluation and risk management. She has published articles related to hedge funds in academic journals.

*GAINS-ARGUMANS, University of Maine and La Rochelle Business School, France. Email: hnguyen@univ-lemans.fr

The author is indebted to Georges Gallais-Hammon, Georges Hübner and Pascal Grandin for their valuable suggestions. She also thanks two anonymous referees of the journal for helpful and constructive comments.
Introduction

Since the seminal work of Treynor (1965), Sharpe (1966), and Jensen (1968), performance measures have always been the focus of much attention from both researchers and practitioners. While researchers employ these measures to examine market efficiency, practitioners use them in at least two circumstances. First, they evaluate the past performance in the hope that it is a reliable indicator of future performance, particularly in order to choose the best funds for investment. Second, they measure performance to compare the results of one fund to its competitors or those of the indices representing the market from which are selected the assets held. From an internal viewpoint, directors of management companies rely on this evaluation to judge the efficacy of their portfolio managers and determine their appropriate compensation. On an external scale, investors allocate their capital and control the efficacy of these investments relative to their objectives based on fund performances.

As far as hedge funds are concerned, the evaluation of their performance is a complex task due to specific characteristics of their returns. On the one hand, the latter are often asymmetric and leptokurtic (with fat tails), which makes the use of traditional measures based on the so-called paradigm of "mean-variance" inappropriate. On the other hand, the opportunistic and dynamic nature of hedge fund strategies, usually coupled with short movements across multiple assets, and the absolute performance objective makes the application of usual multi-factor models inefficient. These elements explain essentially the recent development of new measures, theoretically more satisfactory but mathematically much more complex; each one uses a distinct approach with its advantage and its inconvenience. This abundance along with the absence of a back-testing mechanism makes the choice of performance measures quite problematic. On a practical scale, while it is widely recognized that hedge fund returns are not normally distributed, only two percent of European multi-managers pay attention to the third- and fourth-order moments (i.e., skewness and kurtosis) of the return distribution while measuring funds’ risks and performances (see Amenc, Giraud, Martellini and Vaissié (2004) and Amenc, Malaisie and Vaissié (2005)). While performance analyses have important implications, their results might be, a priori, dependent upon the measure(s) employed. Despite the importance of these issues, the literature on this subject is not only narrow but also offers little insight.

Eling and Schuhmacher (2007) found highly positive correlations between the rankings of hedge fund established by thirteen measures on a sample of 2,763 hedge funds. They confirmed that all the measures give virtually identical rankings. In the same spirit, Kooli, Morin, and Sedzo (2005) conducted a comparative study on 675 hedge funds and stated that the two Sharpe ratios rank funds in a similar order since the two rankings are highly and positively correlated with each other. The main limit of these studies is that they rely solely on rank correlation coefficients to study the consistency between various performance measures. The results all indicate that despite their different approaches adopted, these measures rank funds in a quasi-identical order. This finding raises the question of the raison d'être of recent new methods which claim to be more adequate theoretically than traditional measures. However,
are the correlations between the rankings established by different measures sufficient to draw the conclusion that they are consistent? In fact, performance measures are used to determine the rank of a given fund (or a group of funds) and/or to identify the best funds for investment. Whatever the objective is, investors are concerned with a small group of funds and not the whole sample. As a result, a high positive correlation (but not perfect) might lead to a different investment decision if incoherent elements are among the sub-sample concerned. On the contrary, a weak positive correlation can always give rise to similar final decisions if incoherent elements are absent from the sub-sample under investigation. From such point of view, rank correlations are simply informative and can not be conclusive. Given the important implications of performance evaluation, the study of the consistency of different performance measures needs further in-depth investigation.

This article contributes to the existing literature in several ways. After a refined analysis of the consistency between performance measures through a quantile-by-quantile study of fund (ordered) rankings, their consistency is assessed as a whole by using a mapping of all the measures provided by the ascendant hierarchical clustering technique. Furthermore, also investigated is the persistence or the stability of performance measures. The results of this study have valuable implications for investors as well as for fund managers. From an investor's viewpoint, a stable measure is expected to display some predictive power with respect to future performance of funds. From a fund manager's viewpoint, their natural incentive is to choose stable measures so as to sell their performance persistence if this one is relatively satisfactory.

The remainder of the article is organized as follows. Section 1 introduces the measures considered while Section 2 presents the sample used. Section 3 studies the impact of the choice of performance measures on hedge fund rankings. Section 4 examines the consistency between measures under consideration by means of the ascendant hierarchical clustering technique. Section 5 is devoted to a robustness check. Section 6 deals with the issue of stability of performance measures while Section 7 concludes.

1. Performance measures studied

There are generally two kinds of performance measures in the literature of portfolio performance evaluation. The first one, evaluates fund managers’ added values (their selectivity and their timing ability), including measures based on the Capital Asset Pricing Model (CAPM) such as Jensen, Treynor and other recently elaborated measures – multi-factor models with variable alpha(s) and beta(s), models with optional factors or regime switching models. The second stream relates to measures leading to a complete fund ranking, in the spirit of the Sharpe ratio. These measures are primarily used in the first stage of screening procedures to establish a shortlist of the best performing funds on which deeper quantitative and/or qualitative analysis will be applied latter, before any investment decision. This study focuses on performance measures that allow a complete ranking and not those designed to evaluate manager skills. To this end, ten measures documented as the most used by practitioners and researchers are selected: Sharpe,
Sortino, Calmar, Sterling, Burke, Stutzer, modified Sharpe, upside potential ratio, Omega and AIRAP. The following subsections discuss these measures.

1.1. Measure based on the mean-variance paradigm – the Sharpe ratio

Despite its age, the Sharpe ratio (Sharpe, 1966) has been up to now, without doubt, the most popular and the most used in portfolio performance evaluation, comprising hedge fund portfolios.

\[
\text{Sharpe}_p = \frac{\overline{R}_p - \overline{R}_f}{\sigma_p}
\]

where \(\overline{R}_p\) and \(\sigma_p\) are respectively the mean and the standard deviation of monthly returns on the portfolio to be evaluated, \(\overline{R}_f\) is the risk-free rate over the same period and represented here by the three-month US Treasury bill rate. The Sharpe ratio’s popularity certainly rests on the simplicity of its concept and the ease of its implementation. However, it suffers from two drawbacks. Not only it is based on the normality assumption of returns, which is obviously inappropriate for the hedge fund context, but also it gives illogical rankings when excess returns are negative. Yet, as reported a recent EDHEC’s survey in 2003, more than 80% of European funds of hedge funds’ managers use the Sharpe ratio in monitoring hedge funds’ performance (Amenc, Giraud, Martellini and Vaissié, 2004); 69% of them use it to report to their clients (Amenc, Malaisie and Vaissié, 2005).

1.2. Measures without return distributional assumptions – drawdown based ratios

The drawdown (DD) of a portfolio is the loss incurred over a given investment period and relative to a historical high. As risk indicator, the DD is used in three performance measures: Calmar (Young, 1991), Sterling (Kestner, 1996) and Burke (Burke, 1994).

\[
\text{Calmar}_p = \frac{\overline{R}_p - \overline{R}_f}{MDD_T}
\]

\[
\text{Sterling}_p = \frac{\overline{R}_p - \overline{R}_f}{\frac{1}{T} \sum_{t=1}^{T} MDD_t + 10\%}
\]

\[
\text{Burke}_p = \frac{\overline{R}_p - \overline{R}_f}{\sqrt{\sum_{i=1}^{n} (MDD_i)^2}}
\]

where \(MDD_T\) is the largest DD over a \(T\)-year period; \(\frac{1}{T} \sum_{t=1}^{T} MDD_t\) is the average of \(T\) annual \(MDD\) during the period. In practice, a three-year evaluation period is generally used. Regarding the Sterling ratio, a 10 percent is often added to the denominator, signifying an increase of 10 percent of the drawdown risk. With respect to the denominator of the Burke ratio, it is calculated as the sum of the \(n\) largest DD over the period with \(n\) is generally fixed at 5. Since historical losses in Sterling and Burke ratios involve several MDD, they are expected to be less dependent on aberrant values than the Calmar ratio. Always according to the EDHEC’s survey in 2003, 80 percent of European funds of funds’ managers employ DD-based indicators to monitor their constituent funds. On a technical scale, the implementation of these drawdown
based measures is simple without penalizing upside potential of portfolios, especially without assuming normal return distributions. Nevertheless, the DD only measures a physical fact of past realizations with little indication of future risk.

1.3. Measure taking into account the downside risk – the Sortino ratio

Elaborated by Sortino and Price (1994), the Sortino ratio is an extension of the Sharpe ratio in order to take into consideration the downside risk relative to a predetermined benchmark. The numerator is the excess return with respect to a Minimum Accepted Return (MAR) fixed by investors and proxied here by the three-month US Treasury bill rate. The risk in the denominator is measured by the lower semi-variance that is the variance of returns below the MAR:

$$\text{Sortino}_p = \frac{\overline{R}_p - \text{MAR}}{\sqrt{\frac{1}{T} \sum_{t=1}^{T} (R_{pt} - \text{MAR})^2}}$$

Despite its simplicity, the main inconvenience of the Sortino ratio rests in its foundation on the semi-variance which emanates from the quadratic utility function of investors. Yet, this form is commonly admitted for not describing the true behavior of investors. However, the EDHEC’s survey in 2003 states that 22 percent of funds of hedge funds’ managers employ the Sortino ratio for reporting to their clients.

1.4. Measures taking into account higher moments of return distributions

1.4.1. Stutzer index

Developed by Stutzer (2000), the Stutzer index assumes that investment fund managers have an aversion towards non-positive excess returns relative to a predetermined benchmark. As a result of this, managers choose portfolios with the probability of having non-positive excess returns converge as fast as possible. Stutzer suggests using this speed as the portfolio’s performance indicator.

$$\text{Stutzer}_p = \max_{\theta} \left[ -\ln \frac{1}{T} \sum_{t=1}^{T} e^{\theta r_{pt}} \right]$$

where $\theta$ is a negative value to be determined in order to maximize the objective function; $T$ is the number of elementary periods, $r_{pt}$ is the portfolio’s return in excess of a risk-free rate taken to be the three-month US Treasury bill rate in this study.

The merit of the Stutzer index is twofold. First, it provides a consistent performance measure with the traditional Sharpe ratio when excess returns are normal. Second, it "rewards" portfolios whose return distributions are desirable; that is, those with positive skewness, whilst penalizing those whose returns have undesirable negative skewness. However, this double advantage is accompanied by several limits. In addition to the underlying complex mathematical aspect, the Stutzer index presents two problematic technical issues: the constant absolute risk aversion utility function assumption and the necessarily long evaluation period for robust results. The first hypothesis is controversial because individuals are documented to display an absolute risk aversion which decreases in function of their wealth. The second condition implies that the Stutzer index is only efficient while to be used to evaluate long-horizon
investments such as those of pension fund investors. On the contrary, hedge fund history is often relatively short.

1.4.2. Modified Sharpe ratio

Based on a new measure of potential extreme loss called the modified Value-at-Risk (MVAR) developed by Favre and Galeano (2002), Gregoriou and Gueyie (2003) suggest replacing the risk measure in the denominator of the Sharpe ratio by the MVAR to form the modified Sharpe ratio:

$$M\text{-Sharpe}_p = \frac{\bar{R}_p - \bar{R}_f}{MVAR_p}$$

with

$$MVAR_p = W \left[ \mu - \left\{ z_c + \frac{1}{6} (z_c^2 - 1) S + \frac{1}{24} (z_c^3 - 3 z_c) K - \frac{1}{36} (2 z_c^3 - 5 z_c) S^2 \right\} \sigma \right]$$

where $W$ represents the portfolio’s value at risk, $\mu = \bar{R}$ is the mean return, $\sigma$, $S$ and $K$ are respectively the standard deviation, the skewness and the excess kurtosis of returns, $z_c$ is the critical value at the $1-\alpha$ confidence level (e.g., $z_c = -1.96$ when $\alpha = 95\%$). The modified Sharpe ratio is appealing since the risk measured by the MVAR takes into account higher moments of return distributions: the skewness and the kurtosis. Despite this interest, as stated by Mina and Ulmer (1999), the Cornish-Fisher method used to compute the MVAR is fast but less accurate and occasionally unstable.

1.5. Measures taking into account the whole return distribution

1.5.1. Upside potential ratio (UPR)

In the same spirit of Sortino and Price (1994), Sortino, Van der Meer, and Plantinga (1999a, 1999b) suggest an extension of the Sortino ratio in order to take into consideration the "upside potential" of portfolios. The idea is to replace the numerator of the Sortino ratio by the average of returns above the MAR supposed to be the three-month US Treasury bill rate:

$$UPR_p = \frac{1}{T} \sum_{R_{pt} > MAR} (R_{pt} - MAR) \sqrt{\frac{1}{T} \sum_{R_{pt} < MAR} (R_{pt} - MAR)^2}$$

The innovation of the UPR relative to the Sortino ratio is that the gain (the numerator) and the loss/risk are computed in the same logic, i.e. with respect to the same benchmark return. This common threshold assures consistency in the perception of gain as well as the perception of loss of individuals so far demonstrated in behavior finance. This measure is especially appealing when no assumption on return distribution form is required.

1.5.2. Omega index

The Omega index developed by Keating and Shadwick (2002) measures the ratio between the upside variations of a portfolio (its reward) and its downside variations (its risk), both of them relative to a common threshold (MAR).
\[ \Omega_p = \frac{I_g}{I_l} = \frac{\int_{-\infty}^{\infty} [1 - F(R)] dR}{\int_{-\infty}^{\infty} F(R) dR} \]

where \( F(R) \) is the cumulative probability function of returns. In the Omega formula, the numerator represents the accumulated reward and the denominator proxies the accumulated loss. As the UPR, the Omega index considers the whole distribution of returns and makes no assumption about investors’ utility function form.

1.5.3. AIRAP

Developed by Sharma (2004) exclusively to the hedge fund context, the merit of the AIRAP (Alternative Investments Risk Adjusted Performance) measure is twofold. First, not only does it require no assumption on portfolio returns but also it incorporates the whole return distribution. Second, it is elaborated on the basis of the Expected Utility Theorem. More precisely, investors are supposed to have a Constant Relative Risk Aversion (CRRA), that is a logarithmic function or a power utility function:

\[
AIRAP_p = \left[ \sum_{t=1}^{T} (1 + R_t)^{1-c} p_t \right]^{1/(1-c)} - 1 \text{ où } c > 0, c \neq 1 \text{ (Power utility function)}
\]

\[
AIRAP_p = \left[ \prod_{t=1}^{T} (1 + R_t)^{p_t} \right] - 1 \text{ où } c = 1 \text{ (Logarithmic function)}
\]

where \( p_t \) is the probability to obtain the return \( R_t \) and \( c \) is the coefficient of relative risk aversion (ARA) which is constrained to be greater than 0. The case where \( c = 1 \) corresponds to a risk neutrality of investors. Given that Osband (2002) suggests \( c = 2 \) and that Ait-Sahalia, Parker and Yogo (2004) empirically find \( c = 3.2 \) for wealthy individuals (who primarily form hedge fund clients), Sharma assumes that \( c = 4 \). Besides the utility function foundation, another interesting characteristic of AIRAP is the possibility to adjust the ARA coefficient (the \( c \) parameter) according to investors’ risk aversion profiles. Nevertheless, from a practical point of view, the choice of a value for \( c \) requires solid justification. Yet, the literature offers little insight on this issue.

2. The sample

The sample includes 149 Equity Long/Short hedge funds extracted from the Center for International Securities and Derivatives Markets (CISDM) database. The Equity Long/Short strategy consists of combining two simultaneous operations in the same portfolio: buying undervalued stocks and selling short overvalued stocks. To be included in the sample, each fund must have a complete monthly return history over the period January 2000 - December 2005.

Since the normality of the return distribution plays a fundamental role in the choice of performance measures, a normality test is indispensable. To this end, the Shapiro-Wilk test (Shapiro and Wilk, 1965), documented as the most appropriate for a short series, is conducted on the given sample of 149 hedge funds. The normality hypothesis is rejected, at the significance level of 0.05, in 59.7 percent of the funds.
Note that this distributional mix is a common situation in practice. As a result of this mix, performance measures which take into account the whole distribution of fund returns are *a priori* the most appropriate.

3. Hedge fund rankings: consequences of the choice of performance measures

Regarding the assessment of fund performance, investors and fund managers are essentially concerned about two things: the first is to know whether a fund over-performs the market and the second is whether it also does better than others. In what follows, only the second concern is analyzed. Specifically, the question that arises here is to determine whether fund rankings according to performance indicators are similar or different.

3.1. Highly positive correlation coefficients

As it is commonly done in the literature, a first consistency analysis of the predefined measures is conducted by calculating Spearman rank correlations. The summarized results are reported in Table 1ii. Expectedly, the obtained assessment is similar to that achieved by previous studies: fund rankings are all highly and positively correlated and all correlation coefficients are statistically significant at the one percent level. The global mean of correlation coefficients is particularly high, 0.957, with a maximum of 0.998 and a minimum of 0.869. This finding suggests that all measures lead to nearly identical rankings. Hence, one could conjecture that the choice of performance measures does not matter; sophisticated and *de facto* more efficient performance measures do not rank funds much differently from simpler ones. Yet, this assessment is somewhat intriguing because it is widely known that different performance measures use quite distinct approaches. Given the important implications of this analysis for investors and fund managers, further investigation on fund ranking consistency must be carried out.

3.2. Modifications in funds’ ranks

A case-by-case study of fund rankings shows that the percentage of funds that receive the same rank according two measures is only 11 percent on average. In many cases, this value is primarily equal or smaller than five percent. Two other calculations will corroborate the evidence that rank correlations might not be sufficiently conclusive of the consistency between performance measures: (i) the percentage of funds whose ranks are modified by at least five places (upwards or downwards) when the performance indicator is replaced by another (Panel A of Table 2); (ii) the percentage of funds whose change in ranks is at least fifteen places (Panel B of Table 2). Results indicate that on average, 54 percent of funds are subjected to a modification of at least five places. In the worst case, this proportion can attain a very high level of 92 percent as the case of the Sharpe ratio in comparison with the Stutzer ratio. Always in average terms, 16 percent of funds have their ranks modified by at least fifteen places because of changes in performance measures. This value signifies that 16 percent of funds are seriously overrated or underrated relative to their peers when performance indicators are modified. The highest percentage of funds with at least 15-place changes is attained when we compare the rankings provided by UPR and AIRAP with those established by other measures (21 percent and 29 percent respectively). In most cases, this proportion is higher than 20 percent. In the most extreme case (UPR *versus* AIRAP), the risk of a biased ordering affects 40 percent of the population. It is interesting to note that both indicators take into account the
whole distribution of fund returns. These findings provide a first evidence of a serious risk of erroneous rankings if the performance measure is not rigorously selected.

3.3. Modifications in funds’ performance classes

In order to better appraise the consequences of the choice of performance measures on fund rankings, we proceed to observe fund movements across deciles following the change in performance measures. This exercise is based on the fact that all investment fund rating agencies periodically attribute to each fund a certain number of stars (from five stars to one star like the systems of Morningstar and Europerformance) or a note (from one to five in the Lipper’s system). As far as the rating methods are concerned, it is well known that they greatly differ from one agency to another. According to common rating practices, all funds belonging to the same performance class will receive the same star number or the same note, regardless of their absolute performance. From such viewpoint, the first preoccupation of fund managers is to belong to a performance class that is as good as possible, rather than to be rated at a lower nth place.

The division of funds in deciles is conducted as follows: for each performance measure considered, all the funds are first classified in a descending order on the basis of their absolute performance. In this order, the first fund performs best and the last fund performs worst. Since the sample is composed of 149 funds, the attribution of funds in deciles on the basis of their absolute performance is conducted such that each decile includes 15 funds and the last decile contains the last 14 funds.

The percentages of funds that maintain their decile are reported in Table 3, while those that move to another decile appear in Tables 4 and 5. As indicated in Table 3, on average, only 58 percent of funds stay in the same decile when a performance indicator changes. In the best case, 85 percent of these funds remain in the decile while in the worst case, only 35 percent remain. This indicates that a significant population of funds has a modification of their performance class as a result of replacing the evaluation measure by another.

This evidence is then corroborated by the results appearing in Tables 4 and 5. Table 4 can be read as follows: the value at the intersection between the first line and the second column in Panel A signifies that when the Sharpe ratio is replaced by the Sortino ratio, 12 percent of funds receive a place in an upper decile. For the UPR ratio, this proportion is 24 percent (the third column). For the Calmar ratio, it is 20 percent (the fourth column) and so on. Panel B of Table 4 details the percentages of funds for which the difference between the new upper decile and the old decile is more than one (decile). For instance, a move from the third decile to the first decile, the difference is two classes. For illustration purposes, consider the value at the intersection between the first line and the third column of Panel B. It indicates that 5 percent of funds display a migration to a new (upper) decile where difference with the previous one is greater than one. Table 5 is organized identically but concerns descending movements; all movements towards a lower performance class, for example a shift from the second decile towards a lower decile such as the third or the fourth deciles.
According to Tables 4 and 5, a significant number of funds display ascending or descending movements after the performance measure changes. On average, these funds represent 21 percent of the population under consideration in terms of ascending movements (Panel A of Table 4) and also 21 percent in terms of descending movements (Panel A of Table 5). A closer examination of these cases shows that a category, comprising exactly 4 percent, is affected by significant upward or downward shifts in performance classes. In terms of star numbers, this signifies that 4 percent of the funds receive approximately two stars more or less than attained using a different method. Beyond the technical aspect of performance measures, what is at stake are, on the one hand, the right compensation for fund managers and, on the other hand, the selection of the right funds and thus the optimal allocation of investors’ capital.

4. Consistency of performance measures: an ascendant hierarchical clustering application

In this section, the consistency of these performance measures is examined by means of the ascendant hierarchical clustering technique. The principle of the ascendant hierarchical clustering technique consists of exploring the data’s structure so as to divide a set of individuals into clusters such that two individuals in the same cluster bear a stronger similarity to each other than two individuals belonging to two different clusters. The ascendant clustering (agglomerative clustering) builds a hierarchy in a bottom-to-top fashion by starting from smaller clusters and sequentially merging them into "parental" nodes.

The merit of this clustering technique is that it makes use of the entire information of fund rankings in order to provide a mapping of all the measures under investigation as a whole, under the form of a dendogram tree. Such a visual representation allows users to identify easily individuals having similar "behavior" and to determine the appropriate number of groups. The ascendant hierarchical clustering is applied to the 10 rankings by using the squared Euclidean distance and the Ward criteria (Ward, 1963) as cluster measures.

The dendogram tree in Figure 1 produces deeper insights into the consistency level between measures as a whole: the "behavior" of the 10 measures displays weak homogeneity. Only two clusters of measures are visibly formed. The first one contains Sharpe, Stutzer, Omega and Sortino (according to its order to join the group). The second one includes Calmar and Burke. To the extent that they are both based on the drawdown as measure of risk, this finding is not surprising. Reduced accuracy in the splitting of the tree might assign Sterling to the first cluster. The other measures are geometrically dispersed and form merely one-entity clusters. In other words, these measures are quite different from each other. Among them, AIRAP and M-Sharpe differ the most from their peers because they are at the bottom of the tree and the last to join the already formed cluster with a quite high distance. If the consistency between Calmar and Burke is somewhat expected, the proximity between Sharpe, Stutzer, Omega, and Sortino is surprising, especially in the case of Sharpe ratio and Omega index given that the former uses a "mean-variance" approach whilst the latter takes into account the whole distribution of returns. It is important to note that the sample is composed of both Gaussian and non-Gaussian distributions of returns.
5. Robustness check of inconsistency findings

In order to examine the impact of the evaluation horizon (which can also be considered as the impact of the sample choice) on obtained results, the same analyses are conducted on the same 149 hedge funds but over two other horizons: five years (January 2001 – December 2005) and three years (January 2003 – December 2005). The normality test’s results always show a mixed sample of gaussian and non-gaussian returns despite that the proportions of funds having gaussian returns and funds with non gaussian ones significantly vary over time: the normality hypothesis is rejected, at the significance level of 0.05, for 38.3 percent of funds over the 5-year period and only 11.4 percent over the 3-year period, comparatively to 59.7 percent over the 6-year period. This finding confirms again the necessity of the use of measures based on the whole return distribution.

In line with the previous 6-year horizon results, funds’ relative performances are found again to change considerably following the change in performance measures, whatever the horizon is, despite extremely high rank correlations. Detailed results are available upon request.

Concerning the clustering analysis, the partitions of the ten performance measures are slightly different according to periods. The only notable change is M-Sharpe. Being at the bottom of the tree over the 6-year period, it is the last measure to join the universal cluster at a distance of more than 70,000. Yet, over the other two periods, it is close to Sharpe, Omega, Stutzer, Sortino and forms a rather homogenous cluster with these measures.

6. Stability of performance measures

In this section, the study subject is performance measures while the performance persistence of hedge funds is used as an analysis instrument. To be more precise, by studying the persistence in the performance of the 149 previously defined hedge funds, we aim to identify which measures, among the ten measures under consideration, provide rankings that are more or less stable over time. As far as investors are concerned, a stable measure is expected to display some predictive power regarding the funds’ future performance. With respect to fund managers, they naturally employ measures that are in favor of their performance persistence. From an academic viewpoint, such a study highlights possible impacts of the choice of performance measures on the results of performance persistence.

6.1. Persistence tests

In general, measuring fund performance persistence consists of examining the relation between the relative performance of funds over a defined period and their relative performance over the next period. Funds are considered as displaying some performance persistence if this relation is positive and vice versa. To this end, we observe the performance evolution of funds across two three-year subperiods. Although the choice of subperiod length is constrained by the sample that we have, the three-year length is not completely arbitrary with regard to the lock-up period usually demanded by hedge fund managers. Often varying from one fund to another, this period is generally fixed from one year to three years. Nonetheless,
due to the increasingly volatile context of financial markets over the last years, alternative managers have
a tendency to demand longer lock-up time.

With respect to persistence tests, non-parametric tests are chosen as they are the most used, largely
because of their conceptual simplicity, their facility in application and the absence of econometric biases
which involve parametric tests. Consequently, two tests are used: the contingency table-based test and the
Spearman rank correlation test.

The first test aims to determine whether a fund obtains the same relative performance; that is, does it
stay in the same quantile in the second subperiod. The results allow conjecture that through this test, the
existence of a semi-strong persistence of performance measures is tested. In general, a 4x4 contingency
table is adopted to test the performance persistence in terms of quartiles. The idea consists of firstly
ordering funds in quartiles on the basis of their relative performance and then determining, in the second
step, the number of funds for different scenarios at the end of the two subperiods. From a financial
viewpoint, examining fund movements across quartiles from one subperiod to another is meaningful with
regard to fund rankings published by the financial press. Once the contingency table is formed, following
Kahn and Rudd (1995), the Chi-square test is used to test statistically the presence or the absence of
possible persistence. The null hypothesis of this test, which is an absence of performance persistence, is
accepted if and only if the significance level (or the p-value) is smaller than the error level of the test often
fixed at five percent.

With respect to the Spearman rank correlation test, it is used to determine whether a fund will receive
the same (absolute) rank in the following period. In this regard, it tests the existence of persistence in
strong form. Consequently, it can be viewed as more advanced than the previous test.

Regardless of the persistence tests used, since they are applied to the same sample of funds, one can
speculate that the measure of the conjecture concerning the funds’ performance persistence is accepted
displays some persistence or stability. This persistence or stability could be considered as the existence of
some predictive power of this measure with respect to funds’ future performance.

6.2. Results

Table 6 summarizes results of the two persistence tests. According to the Chi-square test, seven out of
ten measures confirm the presence of performance persistence in terms of quartiles: Sharpe, Sortino,
Calmar, Sterling, AIRAP, Omega and Burke. Among them, this persistence is statistically significant at a
one percent level for Sharpe, Sortino and Calmar ratios and a ten percent significance level for the Burke
ratio. This result implies some predictive power of these seven indicators regarding funds’ future
performance. The three measures in favor of an absence of persistence are UPR, Stutzer and M-Sharpe. It
is interesting to note that in these three cases, the null hypothesis (absence of persistence) is accepted at p-
values that are quite close to ten percent: 12.9 percent (UPR), 13.1 percent (Stutzer) and 15 percent (M-
Sharpe).
Unlike the Chi-square test's results, those of the Spearman rank correlation test (third and fourth columns) are not really conclusive. Eight out of ten measures are negative but not statistically significant. The coefficient associated with the Burke ratio is slightly positive but statistically insignificant. The coefficient for the AIRAP index is not only statistically negative at a five percent level but displays a non-neglected level of -0.213. From a financial viewpoint, these results suggest that for nine of the ten measures under consideration, no rank stability is detected; funds’ ranks in the previous period are unlikely to be indicative of their future ranks. On the contrary, according to the AIRAP index, the past ranks seem to have a certain predictive power of the future ranks which are likely to be the inverse of the past. A case-by-case examination reveals that a large number of highly ranked funds in the first period happen to be among the most poorly ranked in the second period and vice versa.

7. Concluding remarks

The necessity of having appropriate performance measures that take into account the characteristics of hedge fund returns has led to development of new measures that are more elaborate and theoretically more efficient than traditional ones. While the choice of measures is still challenging because of the lack of a mechanism to test the empirical robustness of these measures, it is essential to know whether this choice really matters and its consequences. This article proposes a comparative study of ten measures documented in research and practice as the most used: Sharpe, Sortino, Calmar, Sterling, Burke, Stutzer, modified Sharpe, upside potential ratio, Omega and AIRAP.

In examining the modifications of fund performance in terms of ranks and deciles when the performance measure changes, numerous significant modifications were observed despite strong positive correlations among fund rankings established by different measures. While 21 percent of funds are affected by upward shifts in performance classes, the same proportion is observed in terms of downward shifts. Among them, 4 percent suffer from a change of at least two performance classes, which implies a serious excess or lack of at least two stars relatively to ranks obtained from another evaluation method.

A mapping of the ten measures by means of the ascendant hierarchical clustering technique shows two clear-cut clusters including measures that have very close "behavior" in ranking funds. The first cluster contains Sharpe, Stutzer, Omega and Sortino whilst the second is composed of Calmar and Burke. On the contrary, AIRAP is highly distinct from the others.

With respect to the persistence of the ten measures, the results indicate that Sharpe, Sortino, Calmar, Sterling, Burke, AIRAP and Omega lead to stable rankings in quartiles over time while UPR, Stutzer and M-Sharpe do not. This implies that the first group of measures displays some predictive power regarding the funds’ future relative performance. In contrast, in terms of rank stability, except for the AIRAP index which seems to result in rank inversion, no measure displays predictive power; funds’ ranks in the previous period are unlikely to be indicative of their ranks in the following period.

These findings have valuable implications for hedge fund managers and investors. Not only does the evaluation of investment in hedge funds depend strongly on the choice of performance measures but also
on the fact that fund managers can earn higher compensation by playing upon some performance measures that are more stable than others. Given these findings, it is indispensable to develop an efficiency test guiding investors in the choice of appropriate performance measures or a mechanism considering simultaneously several measures in a complementary fashion.

Endnotes

i Consider two portfolios A and B with \( r_A = -5\% \), \( r_B = -5\% \), \( \sigma_A = 20\% \), \( \sigma_B = 25\% \). The Sharpe ratios for these portfolios are \( \text{Sharpe}_A = -5/20 = -0.25 \), \( \text{Sharpe}_B = -5/25 = -0.20 \). Since \( \text{Sharpe}_A < \text{Sharpe}_B \) portfolio B is preferred to portfolio A even though portfolio B has the higher standard deviation. A justification for this problem is to associate the higher standard deviation with the higher probability of obtaining positive returns in the future. Consequently, at the same negative return level, portfolio B is more interesting than portfolio A.

ii Detailed results are available upon request.

For instance, Morningstar attributes stars to funds on the basis of their relative performance according to the following mechanism: 5 stars to the first 10% of best funds; 4 stars to the following 22.5% of funds; 3 stars to 35% of funds that follow; 2 stars to the following 22.5%; 1 star to 10% of funds at the end of the rating list, those with the worst performance.

References


