Assessing Hedge Fund Performance: Does the Choice of Measures Matter?

Huyen Nguyen-Thi-Thanh

Assessing Hedge Fund Performance: Does the Choice of Measures Matter?

Huyen Nguyen-Thi-Thanh*

October 2007
(Draft only, do not quote without permission)

Abstract

In this paper, we conducted a comparative study of ten measures documented as the most used by researchers and practionners: Sharpe, Sortino, Calmar, Sterling, Burke, modified Stutzer, modified Sharpe, upside potential ratio, Omega and AIRAP. This study was carried out in two stages on a sample of 149 hedge funds. First, we examined the modifications of funds' relative performance in terms of ranks and deciles when the performance measure changes. Despite strong positive correlations between funds' rankings established by different measures, numerous significant modifications were observed. Second, we studied the stability/persistence of the ten measures in question. Our results show that some measures are more stable or persistent than the others in measuring hedge fund performance.

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

JEL Classification : G2, G11, G15

*Assistant Professor, La Rochelle Business School / CEREG (France) and Associate Researcher, Orleans Laboratory of Economics (University of Orleans, France), 102 rue de Coureilles – Les Minimes, 17024 La Rochelle cedex 1, France. E-mail: nguyenthithanhh@esc-larochelle.fr

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 practioners. While researchers employ these measures to examine market efficiency, practioners use them in at least two circumstances. First, they evaluate the past performance in the hope that it is a reliable indicator of the future performance, particularly in order to choose the best funds to invest in. Second, they measure the performance so as to compare the results of one funds to those of 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 thus the appropriate compensation for the latter. From an external viewpoint, investors base on funds' performances to allocate their capital and control later the efficacy of these investments relatively to their objectives.

Regarding hedge funds, the evaluation of their performance is a complexe 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 traditionnal measures based on the so-called paradigm of "mean-variance" inappropriate. On the other hand, the opportunistic and dynamique nature of hedge fund strategies, usually coupled with short mouvements across multiple assets, and the absolute performance objective make the application of usual multi-factor models inefficient. These elements explain essentially the recent development of new measures, theoritically more satisfactory but mathematically much more complexe, each one uses a distinct approach with its avantage and its inconvenience. This abondance along with the absence of a back-testing mechanisme makes the choice of performance measures quite difficult. 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 & Schuhmacher (2005) found highly positive correlations between the rankings of hedge fund indexes established by various measures. Eling & Schuhmacher (2006) enlarged the same analysis to thirteen measures (Sharpe, Treynor, Jensen, Omega, Sortino, Kappa3, upside potential ratio, Calmar, Sterling, Burke, excess return on value at risk, conditional Sharpe and modified Sharpe) and on a sample of 2 763 hedge funds. They confirmed that all the measures give virtually identical rankings. In the same spirit, Kooli, Morin & Sedzro (2005) conducted a comparative study on a sample of 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 the rank correlation coefficient to study the coherence 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 does beg naturally the ques-

tion of raison d'être of recent new methods which are supposed to be theoritically more adequate than traditionnal measures. However, are the correlations between the rankings established by different measures sufficient to draw the conclusion that they are coherent or not?

In fact, performance measures are used in two main objectives. The first one is to determine the rank of a pre-defined fund (or a group of funds). The second is to identify the best funds to invest in. Whatever the objective is, investors are concerned with a small group of funds and not to the whole sample. As a result, a high positive correlation (but non perfect) might lead to a different investment decision if incoherent elements are among the subsample concerned. On the contrary, a weak positive correlation can always give rise to similar final decisions if incoherent elements are absent from the subsample under question. From such point of view, rank correlations are simply informative and can not be conclusive. Given the important implications of performance evaluation, the study on the coherence of different performance measures need further in-depth investigation.

In this context, the contribution of this paper is double. The first contribution consists in conducting a refined coherence study between performance measures by extending the analysis to several quantiles of the (ordered) sample. The second contribution lies in the examination of the *persistence or the stability of performance measures via* a study on the persistence of funds over time. More precisely, the subject of this study is performance measures' persistence while fund persistence is used here as an evaluation tool. On the basis of the same sample, a measure that indicates a certain persistence in fund performance can be considered as more stable and thus more reliable; it is said to show a certain predictif power about the future performance of funds. The results of this study have valuable implications for investors as well as for fund managers. Whatever their objective, it is in their best interest to choose those measures that are persistent.

The rest of the paper is organized as follows. The following section (section 2) presents the measures considered and the sample used. Section 3 studies the impact of the choice of performance measures on hedge fund rankings. Section 4 examines the coherence between measures by means of the technique of descendant hierarchical classification. Section 5 deals with the issue of peristence or stability of performance measures in order to identify the most reliable ones for investors and fund managers. Section 6 concludes the paper.

1 Performance measures considered and data sample

1.1 Performance measures considered

In this study, we restrict ourself exclusively to performance measures that lead to a complete ranking and not those evaluating managers' skills. They are Sharpe, Sortino, Calmar, Sterling, Burke, modified Stutzer, modified Sharpe, upside potential ratio, Omega and AIRAP. Table 1 gives an outline of studied measures, each one being presented with its formula and some major characteristics. For many measures, the minimum acceptable return (MAR) fixed by investors are supposed to be the risk-free rate, which is represented here by the 3-month T-bill rate. The data was collected from Thomson Datastream.

1.2 Data sample

The sample includes 149 hedge funds belonging to the Equity Long/Short category. These funds are extracted from the CISDM (Center for International Securities and Derivatives Markets) database[1]. The Equity Long/Short strategy consists in combining two operations at the same time and in the same portfolio: buying under-valued stocks and selling short over-valued stocks. To be included in the sample, each fund must have a complete monthly return history over the study period — from january 2000 to december 2005 — which makes a total of 72 observations.

The 6-year horizon is chosen for two main reasons. On the one hand, it allows a relatively large sample (given the objective of this study) and quite representative, in terms of size, of a hedge fund style in practice. On the other hand, this horizon provides a history which is long enough to conduct analyses over the sub-periods.

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, documented as the most appropriate one for short series, is conducted on the sample of 149 pre-defined hedge funds. Table 2 summarizes the results of this test at the 5% confidence level. In order to examine the impact of evaluation horizon on obtained results, I also carried out the normality test over two other shorter horizons: 5 years (january 2001–december 2005) and 3 years(january 2003–december 2005). The detailed results over the three horizons are provided in table **??** in the appendix.

Over the 6-year horizon, the normality hypothesis is rejected in 57% of the funds. However, when the horizon is shortened, the number of the funds whose returns are gaussian substantially increases: from 40.3% over the 6-year horizon to 61.7% over the 5-year hori-

[1]The CISDM database is used in many academic studies. Among them, we can refer to Edwards & Caglayan (2001), Capocci & Hübner (2004), ou encore Kouwenberg (2003).

Table 1: Performance measures considered

Measure	Formula	Characteristics
Sharpe ratio	$\text{Sharpe} = \frac{\overline{R}-\overline{R}_f}{\sigma}$ where $\overline{R}, \sigma$ are respectively the mean and the standard deviation of returns on the funds under evaluation, $\overline{R}_f$ is the risk-free rate	– considers mean returns and standard deviation of returns.
Sortino ratio	$\text{Sortino} = \frac{\overline{R}-\tau}{\sqrt{\frac{1}{T}\sum_{t=1,R_{pt}<\tau}^{T}(R_{pt}-\tau)^2}}$ where τ is the minimum acceptable return fixed by the investor, represented here by the risk-free rate R_f, T is the number of return observations, $1/T$ stands for the occuring probability of the return R_{pt}	– considers the mean and the semi lower standard deviation of returns.
Measures without return distributional assumptions / *drawdown* ratios		
Calmar ratio	$\text{Calmar}_T = \frac{\overline{R}-\overline{R}_f}{MDD_T}$ where $MDD_T = \max[\max(\sum_{t=0}^{T} R_t) - R_T]$ where MDD_T denotes the maximum *drawdown* (the largest loss) realized over the period T	– the risque on the denominator is the largest loss realized over the study period. ⇒ very sensitive to extreme values and *a priori* little predictive of the future.
Sterling ratio	$\text{Sterling}_T = \frac{\overline{R}-\overline{R}_f}{\frac{1}{T}\sum_{t=1}^{T}(MDD_t)+10\%}$ where MDD_t is the maximum *drawdown* of the year t	– the risque is the **mean** of the most largest annual losses realized **augmented** an arbitrary 10%. ⇒ less sensitive to extreme values than Calmar ratio; stays little predictive of the future.
Burke ratio	$\text{Burke}_T = \frac{\overline{R}-\overline{R}_f}{\sqrt{\sum_{i=1}^{n}(MDD_i)^2}}$ where MDD_i are the largest *drawdown* (MDD) of the study period T; in practice, $n = 5$	– the square of the MDD_i penalizes large losses. – the risk is the sum of the largest annual losses ⇒ less sensitive to extreme values than the Calmar ratio; stays little predictive of the future.

* An exponential utitily function is formulated as $U(R) = -\exp^{-\alpha R}$ with $\alpha > 0$;

Table 1 (cont)

Measure	Formula	Characteristics
Measures taking into account higher moments of return distributions		
Modified Stutzer index	$\text{Stutzer}^{mod} = sign(\overline{R})\sqrt{2.\text{Stutzer}}$	– assumption: exponential utility function*, thus
	$\text{Stutzer} = \max_\theta \left[-\ln \frac{1}{T} \sum_{t=1}^{T} \exp^{\theta r_t} \right]$	constant absolute risk aversion (CARA)
	where θ is a negative value to be defined so as to maximize the Stutzer index; r_t is the excess return relatively to a predefined threshold represented here by the risk-free rate R_f; T is the number of return observations; $sign(\overline{R})$ is the sign of mean return	– takes into account the mean, the standard deviation and the skewness of rturns.
		– T must be large enough, i.e. long investment horizon.
Modified Sharpe ratio	$\text{M-Sharpe} = \frac{\overline{R} - \overline{R}_f}{MVAR}$	
	where $MVAR$ is the *Modified Value-at-Risk*†	– takes into account (*via* the $MVAR$) the mean, the standard deviation, the skewness and the kurtosis of returns.
Measures taking into account the whole distribution of returns		
Upside potential ratio (UPR)	$UPR = \frac{\text{Potential gain}(\tau)}{\text{Downside risk}(\tau)} = \frac{\frac{1}{T}\sum_{t=1,R_{pt}>\tau}^{T}(R_{pt}-\tau)}{\sqrt{\frac{1}{T}\sum_{t=1,R_{pt}<\tau}^{T}(R_{pt}-\tau)^2}}$	
	where τ is the minimum acceptable return defined by the investor, represented here by the risk-free rate R_f, T is the number of return observations, $1/T$ stands for the occuring probability of the return R_{pt}	– coherent measure of the ratio between the gain and the loss relatively to a predefined threshold.
Omega index	$\Omega(\tau) = \frac{\text{Gains}}{\text{Pertes}} = \frac{\int_{\tau}^{\infty}[1-F(R)]dR}{\int_{-\infty}^{\tau}F(R)dR}$	– no assumption on the utility function.
	where τ is the minimum acceptable return fixed by the investor, represented here by the risk-free rate R_f; $F(R)$ is the cumulative function of returns	– *ad-hoc* measure.
AIRAP	$AIRAP = \left[\sum_i (1+R_t)^{(1-c)} p_t\right]^{\frac{1}{(1-c)}} - 1$	– assumption: power utility function ‡, deacreasing absolute risk aversion with wealth(DARA).
	where c is the coefficient of relative risk aversion, p_t is the occuring probability associated to the observed return R_i; $c = 4, p_t = 1/T$ following Sharma (2004)	– inconvenience: choice of c and p_t

† $MVAR = W\left[\overline{R}_p - \left\{z_c + \frac{1}{6}\left(z_c^2-1\right)S + \frac{1}{24}\left(z_c^3-3z_c\right)K - \frac{1}{36}\left(2z_c^3-5z_c\right)S^2\right\}\sigma\right]$ where σ, S, K denote respectively the standard deviation, the skewness and the kurtosis of returns (Favre & Galeano 2002);

‡ power utility function is formulated as $U(1+R) = \frac{(1+R)^{1-c}-1}{1-c}$ with $c > 0, c \neq 1$.

Table 2: Results of the normality test

	Evaluation horizon		
	6 years (2000-2005)	5 years (2001-2005)	3 years (2003-2005)
Accept the normality hypothesis (in nbr.)	60	92	132
(en %)	(40.3)	(61.7)	(88.6)
Reject the normality hypothesis (in nbr.)	89	57	17
(en %)	(59.7)	(38.3)	(11.4)
Total (in nbr.)	149	149	149
(en %)	100	100	100
Normality over THREE horizons		55 (36.9%)	
NON-normality over THREE horizons		14 (9.4%)	
Normality over TWO horizons		39 (26.2%)	
NON-normality over TWO horizons		41 (27.5%)	

Results are presented at the 5% confidence level.

zon and to 88.6% over the 3-year horizon. When considering the three horizons at the same time, we notice that the percentage of funds for which the normality hypothesis is rejected is extremely small relatively to that for which this hypothesis is supported: 9.4% versus 36.9%. In contrast, the proportion of funds with gaussian returns over any two horizons (among the three studied) and that with non-gaussian returns over any two horizons are quite comparable because they are 26.2% and 27.5% respectively.

To sum up, whatever the evaluation horizon is, it is obvious that the sample under consideration is composed of gaussian return distributions and non-gaussian ones, a common scenario in practice. As a result of this, performance measures which take into account the whole distribution of fund's returns are *a priori* the most appropriate/adequate.

2 Hedge fund rankings: consequences of the choice of performance measures

Regarding the assessment of fund's performance, investors and fund managers are essentially concerned about two things: the first is to know if the fund overperforms the market and the second is that if the fund does also better than other funds. In what follows, I will focus only on the second concern. Specifically, the issue that arises here is to determine whether funds' rankings are similar to or different from their peers according to the performance indicator chosen.

Table 3: rank correlations obtained from various performance measures

	Sharpe	Sortino	UPR	Calmar	Sterling	Burke	M-Stutzer	M-Sharpe	AIRAP	Omega
Panel A: 6-year rank correlations (1/2000-12/2005)										
Sharpe	1	0.993	0.967	0.968	0.986	0.964	0.998	0.915	0.939	0.995
Sortino	0.993	1	0.980	0.973	0.990	0.970	0.996	0.941	0.932	0.995
UPR	0.967	0.980	1	0.955	0.964	0.950	0.971	0.933	0.905	0.971
Calmar	0.968	0.973	0.955	1	0.976	0.995	0.969	0.896	0.906	0.969
Sterling	0.986	0.990	0.964	0.976	1	0.978	0.987	0.927	0.928	0.988
Burke	0.964	0.970	0.950	0.995	0.978	1	0.965	0.893	0.903	0.967
M-Stutzer	0.998	0.996	0.971	0.969	0.987	0.965	1	0.924	0.940	0.997
M-Sharpe	0.915	0.941	0.933	0.896	0.927	0.893	0.924	1	0.869	0.919
AIRAP	0.939	0.932	0.905	0.906	0.928	0.903	0.940	0.869	1	0.934
Omega	0.995	0.995	0.971	0.969	0.988	0.967	0.997	0.919	0.934	1
Mean	*0.972*	*0.977*	*0.956*	*0.958*	*0.972*	*0.956*	*0.975*	*0.916*	*0.920*	*0.973*
Global mean						*0.957*				
Maximum						*0.998*				
Minimum						*0.869*				
Panel B: 5-year rank correlations (1/2001-12/2005)										
	Sharpe	Sortino	UPR	Calmar	Sterling	Burke	M-Stutzer	M-Sharpe	AIRAP	Omega
Sharpe	1	0.996	0.971	0.975	0.990	0.974	0.996	0.992	0.921	0.997
Sortino	0.996	1	0.980	0.983	0.994	0.982	0.995	0.998	0.917	0.997
UPR	0.971	0.980	1	0.969	0.974	0.967	0.973	0.983	0.885	0.977
Calmar	0.975	0.983	0.969	1	0.986	0.998	0.975	0.982	0.886	0.979
Sterling	0.990	0.994	0.974	0.986	1	0.986	0.989	0.992	0.906	0.993
Burke	0.974	0.982	0.967	0.998	0.986	1	0.974	0.981	0.885	0.979
M-Stutzer	0.996	0.995	0.973	0.975	0.989	0.974	1	0.991	0.918	0.995
M-Sharpe	0.992	0.998	0.983	0.982	0.992	0.981	0.991	1	0.914	0.993
AIRAP	0.921	0.917	0.885	0.886	0.906	0.885	0.918	0.914	1	0.917
Omega	0.997	0.997	0.977	0.979	0.993	0.979	0.995	0.993	0.917	1
Mean	*0.979*	*0.983*	*0.964*	*0.970*	*0.979*	*0.970*	*0.978*	*0.981*	*0.905*	*0.981*
Global mean						*0.969*				
Maximum						*0.998*				
Minimum						*0.885*				
Panel C: 3-year rank correlations (1/2003-12/2005)										
	Sharpe	Sortino	UPR	Calmar	Sterling	Burke	M-Stutzer	M-Sharpe	AIRAP	Omega
Sharpe	1	0.989	0.960	0.944	0.961	0.943	0.998	0.971	0.866	0.995
Sortino	0.989	1	0.985	0.953	0.972	0.950	0.994	0.992	0.860	0.994
UPR	0.960	0.985	1	0.944	0.952	0.938	0.969	0.990	0.848	0.968
Calmar	0.944	0.953	0.944	1	0.976	0.987	0.947	0.944	0.785	0.946
Sterling	0.961	0.972	0.952	0.976	1	0.974	0.963	0.955	0.784	0.967
Burke	0.943	0.950	0.938	0.987	0.974	1	0.944	0.935	0.767	0.944
M-Stutzer	0.998	0.994	0.969	0.947	0.963	0.944	1	0.981	0.874	0.997
M-Sharpe	0.971	0.992	0.990	0.944	0.955	0.935	0.981	1	0.879	0.981
AIRAP	0.866	0.860	0.848	0.785	0.784	0.767	0.874	0.879	1	0.867
Omega	0.995	0.994	0.968	0.946	0.967	0.944	0.997	0.981	0.867	1
Mean	*0.959*	*0.965*	*0.950*	*0.936*	*0.945*	*0.931*	*0.963*	*0.959*	*0.837*	*0.962*
Global mean						*0.941*				
Maximum						*0.998*				
Minimum						*0.767*				

All the correlation coefficients presented are statistically significant at the 5% confidence level.

2.1 Highly positive correlation coefficients

As it is commonly done in the literature, we began the coherence analysis of predefined measures by calculating Spearman rank correlations whose results are reported in table 3. As expected, we obtained the same assessement as that acheived by previous studies: fund rankings are all highly and positively correlated and all correlation values are statistically significant at the 1% confidence level, whatever the evaluation horizon is. Over the 6-year and 5-year horizons, the global means of correlation coefficients are particularly high, 0.957 and 0.969 respectively, with a minimum of 0.869 and 0.885 and with a maximum of 0.998 common to the two horizons. Regarding the 3-year horizon, the maximum is the same value 0.998 but the minimum is lower (0.767), which leads to a global mean slightly lower (0.941) than those of the other two horizons. This decrease is largely due to a slight diminution of the correlations between AIRAP ratio and other measures over the 3-year period.

Obviously, these correlation values are really high and suggest that all measures lead to the nearly identical ranking. And thus, we can conjecture that the choice of performance measures has no significant effect on fund ranking. Yet, this strong similarity is somewhat curious and calls for deeper analysis as it is well known that different performance measures use quite distinct approaches. For instance, the Sharpe ratio regards return distributions as gaussian and thus takes into account only the mean and the variance of returns. By contrast, drawdown-based measures like Calmar, Sterling and Burke rely on several extreme returns without any distributional assumption. More laborious indicators such as M-Stutzer and M-Sharpe consider the asymetry (M-Stutzer), or even fat tails (M-Sharpe) of return distributions. The most promising solution remains that suggested by UPR, AIRAP and Omega ratios which incorporate the whole distribution. Yet, in light of the correlation values, the theoritical substance of sophisticated methods, usually accompanied by mathematical and/or statistical complexity, do not modify visibly fund's rankings relatively to their peers. Given the important implications of this analysis for investors and fund managers, we will proceed with refined examination of fund rankings provided by different measures in the following sections.

2.2 Refined analysis of funds' rankings

2.2.1 Non-negligible modifications in funds' ranks

If the correlation values obtained in the previous section suggest that rankings are similar whatever the measure chosen is, we find that in detail, this similarity is not completely risk-free as it hides some serious bias in ranking/ordering funds / fund's hierarchical attribution. As shown in panel A of table 4, the percentage of funds that receive the same rank according two measures is only 11% in average. This proportion can be as high as

Table 4: Rank comparison for the 6-year horizon (in %)

	Sharpe	Sortino	UPR	Calmar	Sterling	Burke	M-Stutzer	M-Sharpe	AIRAP	Omega
Panel A: Equality of ranks (in %)										
Sharpe	100	24	3	13	21	11	32	11	3	19
Sortino	24	100	10	13	23	13	21	20	2	19
UPR	3	10	100	5	5	6	4	3	5	8
Calmar	13	13	5	100	18	26	10	11	1	12
Sterling	21	23	5	18	100	13	15	15	1	16
Burke	11	13	6	26	13	100	11	10	3	10
M-Stutzer	32	21	4	10	15	11	100	8	4	19
M-Sharpe	11	20	3	11	15	10	8	100	1	9
AIRAP	3	2	5	1	1	3	4	1	100	4
Omega	19	19	8	12	16	10	19	9	4	100
Mean	*15*	*16*	*5*	*12*	*14*	*12*	*14*	*10*	*3*	*13*
Global mean						*11*				
Maximum						*32*				
Minimum						*1*				
Panel B: Modification of MORE than 5 places (in %)										
	Sharpe	Sortino	UPR	Calmar	Sterling	Burke	M-Stutzer	M-Sharpe	AIRAP	Omega
Sharpe	100	77	40	54	65	52	92	50	40	88
Sortino	77	100	52	61	77	57	81	59	35	81
UPR	40	52	100	38	38	40	44	48	32	46
Calmar	54	61	38	100	60	87	52	48	29	52
Sterling	65	77	38	60	100	58	66	59	31	69
Burke	52	57	40	87	58	100	51	46	30	53
M-Stutzer	92	81	44	52	66	51	100	54	40	88
M-Sharpe	50	59	48	48	59	46	54	100	22	56
AIRAP	40	35	32	29	31	30	40	22	100	40
Omega	88	81	46	52	69	53	88	56	40	100
Mean	*62*	*64*	*42*	*53*	*58*	*53*	*63*	*49*	*33*	*64*
Global mean						*54*				
Maximum						*92*				
Minimum						*22*				
Panel C: Modification of MORE than 15 places (in %)										
	Sharpe	Sortino	UPR	Calmar	Sterling	Burke	M-Stutzer	M-Sharpe	AIRAP	Omega
Sharpe	100	2	18	19	8	18	0	20	21	1
Sortino	2	100	9	14	5	15	0	9	22	0
UPR	18	9	100	28	19	28	15	15	40	17
Calmar	19	14	28	100	15	2	15	28	34	17
Sterling	8	5	19	15	100	13	6	18	28	7
Burke	18	15	28	2	13	100	17	27	36	16
M-Stutzer	0	0	15	15	6	17	100	17	20	0
M-Sharpe	20	9	15	28	18	27	17	100	34	15
AIRAP	21	22	40	34	28	36	20	34	100	22
Omega	1	0	17	17	7	16	0	15	22	100
Mean	*12*	*9*	*21*	*19*	*13*	*19*	*10*	*20*	*29*	*11*
Global mean						*16*				
Maximum						*40*				
Minimum						*0*				

All calculations of one performance measure avec itself, i.e. all the 100% values, are not taken into account in calculating the means.

32% (Sharpe *versus* M-Stutzer) but also can be as low as 1% like in the case of two rankings established by AIRAP on the one hand, and by Calmar, or Sterling, or M-Sharpe on the other hand. This percentage is extremely low in the two cases of UPR and AIRAP. Theirs values are mostly equal or smaller than 5% (column 4 to 10 in panel A, table 4). However, when we refer to the correlation coefficients between the rankings of UPR and AIRAP with that of the other measures, they are almost largely higher than 0.9.

Two other calculations will show that high positive rank correlations might not sufficiently conclusive of the coherence between performance measures: (i) the percentage of funds whose ranks are modified by *more than* 5 places (upwards or downwards) when the performance indicator is replaced by another (panel B of table 4); (ii) the percentage of funds whose the change in ranks is of *at least* 15 places (panel C of table 4). Results indicate that on average, 54% of funds are subjected to a modification of at least 5 places. In the worst case, this proportion can attain a very high level of 92% as the case of Sharpe ratio in comparison with M-Stutzer ratio. Always in average terms, 16% of funds have their ranks modified by more than 15 places because of changes in performance measures. This value signifies that 16% of funds are seriously overrated or underrated relatively 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% and 29% respectively). In most cases, this proportion is higher than 20%. In the most extreme case (UPR *versus* AIRAP), the risk of a biased ordering affects 40% of the population (panel C, table 4). It is interesting to note that these two indicators all take into account the whole distribution of fund returns.

In sum, despite highly positive correlations between fund rankings established by different measures under consideration, comparing the ranks that a fund is attributed by these measures indicates a pronounced modification of ranks for an important population of funds. This finding provides a first evidence of a serious risk of erronous rankings if the performance measure is not rigorously and rightly selected.

2.2.2 Significant changes in performance classes

In order to better appraise the consequences of the choice of performance measures on funds' rankings, we proceed to observe funds' movements across deciles following the change in performance measures. This exercice is of a great interest for the following reason. Let us remind that all investment fund's rating agencies attribute periodically to each fund a certain star number (from 5 stars to 1 star like the systems of Morningstar and Europerformance) or a note (from 1 to 5 in the Lipper's system)[2]. As far as the rating methods

[2]For instance, Morningstar attribute stars to funds on the basis of their relative performance according to the following mechanism:

- 5 stars to the first 10% of best funds,

are concerned, it is well known that they differ from one agency to another. According to common rating practices, all funds beloning to the same performance classe will receive the same star number or the same note, regardless of their absolute performance. From such viewpoint, the first concern of fund managers is to belong to a performance class as good as possible, rather than to be rated in a n^{th} place.

Table 5 summarizes the principle according to which the attribution of funds in deciles is conducted. For each performance measure considered, all the funds are first classfied in a deacreasing order, on the basis of their absolute performance. In this order, the first fund is the most performant and the last fund is the less performant. Since the sample is made 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.

Table 5: Attribution of funds in deciles on the basis of their absolute performance

Deciles*	Ordered number of constituent funds	Number of funds in each decile
1^{st} decile	1 à 15	15
2^{nd} decile	16 à 30	15
3^{rd} decile	31 à 45	15
4^{th} decile	46 à 60	15
5^{th} decile	61 à 75	15
6^{th} decile	76 à 90	15
7^{th} decile	91 à 105	15
8^{th} decile	106 à 120	15
9^{th} decile	121 à 135	15
10^{th} decile	136 à 149	14

* Since the sample is made 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 first decile groups the top-performing funds while the last decile includes the less performant ones.

The percentages of funds that maintain their decile are reported in table 6, while those that move to another decile appear in tables 7 and 8. We state from table 6 that, on average, only 58% of funds stay in the same decile when the performance indicator changes. In the best case, this proportion attains 85% while in the worst case, it is situated at the level of only 35%. These values indicate that a significant population of funds suffer from a modification of their performance class as a result of replacing the evaluation measure by another.

- 4 stars to the following 22.5% of funds,
- 3 stars to 35% of funds that follow,
- 2 stars to the following 22.5% of funds,
- 1 star to 10% of funds at the end of the rating list, those with the worst performance.

Table 6: Percentage of funds that stay in the same decile

	Sharpe	Sortino	UPR	Calmar	Sterling	Burke	M-Stutzer	M-Sharpe	AIRAP	Omega
Sharpe	100	76	50	57	70	58	84	56	48	83
Sortino	76	100	64	59	74	61	81	66	50	76
UPR	50	64	100	48	47	48	56	52	35	53
Calmar	57	59	48	100	58	81	53	49	36	57
Sterling	70	74	47	58	100	62	68	58	44	66
Burke	58	61	48	81	62	100	54	48	36	54
M-Stutzer	84	81	56	53	68	54	100	55	48	85
M-Sharpe	56	66	52	49	58	48	55	100	35	57
AIRAP	48	50	35	36	44	36	48	35	100	48
Omega	83	76	53	57	66	54	85	57	48	100
Mean	*65*	*67*	*50*	*55*	*61*	*56*	*65*	*53*	*42*	*64*
Global mean						*58*				
Maximum						*85*				
Minimum						*35*				

All calculations of a performance measure with itself, i.e. all 100% values, are not taken into account when calculating the mean.

Table 7: Percentage of funds that move to a **superior** decile

	Panel A: Movement to a superior decile									
	Sharpe	Sortino	UPR	Calmar	Sterling	Burke	M-Stutzer	M-Sharpe	AIRAP	Omega
Sharpe	100	12	24	20	15	20	8	30	21	9
Sortino	12	100	18	19	13	18	9	24	19	12
UPR	26	18	100	24	27	25	22	31	31	25
Calmar	23	22	28	100	22	9	24	34	30	22
Sterling	15	13	26	19	100	17	15	28	24	16
Burke	21	21	28	9	20	100	23	34	28	24
M-Stutzer	8	9	21	23	17	22	100	30	20	7
M-Sharpe	14	9	17	17	14	18	15	100	25	13
AIRAP	31	31	34	35	32	36	32	40	100	32
Omega	9	12	22	21	18	22	7	30	19	100
Mean	*18*	*16*	*24*	*21*	*20*	*21*	*17*	*31*	*24*	*18*
Global mean						*21*				
Maximum						*36*				
Minimum						*7*				
	Panel B: Movement to a superior decile – a difference of more than one decile									
	Sharpe	Sortino	UPR	Calmar	Sterling	Burke	M-Stutzer	M-Sharpe	AIRAP	Omega
Sharpe	100	0	5	5	1	5	0	4	9	0
Sortino	0	100	1	6	0	6	0	0	9	0
UPR	3	1	100	8	3	8	3	1	9	1
Calmar	3	3	5	100	1	0	3	7	9	4
Sterling	1	1	3	4	100	3	1	3	9	2
Burke	4	3	5	0	1	100	3	7	10	3
M-Stutzer	0	0	3	5	0	5	100	2	9	0
M-Sharpe	9	6	6	10	8	9	8	100	14	7
AIRAP	2	3	10	9	5	9	1	11	100	1
Omega	0	0	4	5	0	5	0	3	8	100
Mean	*2*	*2*	*5*	*6*	*2*	*6*	*2*	*4*	*10*	*2*
Global mean						*4*				
Maximum						*11*				
Minimum						*0*				

All calculations of a performance measure with itself, i.e. all 100% values, are not taken into account when calculating the mean.

This evidence is then corroborated by the results that appear in tables 7 and 8. Table 7 can be read as follows: the value at the intersection between the 1^{st} line and the 2^{nd} column of the panel A signifies that when the Sharpe ratio is replaced by the Sortino ratio, 12% of funds receive a place in a superior decile. For the UPR ratio, this proportion is 24% (the $3^{r}d$ column). For the Calmar ratio, it is 20% (the 4^{th} column), etc. The panel B of this table details the percentages of funds whose the difference between the new superior decile and the old decile is greater than one (decile). For instance, for a move from the $3^{r}d$ decile to the 1^{st} decile, this difference is two classes. For illustration purpose, consider the value situated at the intersection between the 1^{st} line and the $3^{r}d$ column of the panel B. This one means that 5% of funds display a migration to a new (superior) decile whose difference with the older one is greater than 1. Table 8 has identical organisation but concerns solely descending movements, i.e. all movements towards an inferior performance class, e.g. a shift from the 2^{nd} decile towards one of inferior deciles such as the 3^{rd} or the 4^{th} decile, etc.

Table 8: Percentage of funds that move to an inferior decile

	Panel A: Movement to an inferior decile									
	Sharpe	Sortino	UPR	Calmar	Sterling	Burke	M-Stutzer	M-Sharpe	AIRAP	Omega
Sharpe	100	12	26	23	15	21	8	14	31	9
Sortino	12	100	18	22	13	21	9	9	31	12
UPR	24	18	100	28	26	28	21	17	34	22
Calmar	20	19	24	100	19	9	23	17	35	21
Sterling	15	13	27	22	100	20	17	14	32	18
Burke	20	18	25	9	17	100	22	18	36	22
M-Stutzer	8	9	22	24	15	23	100	15	32	7
M-Sharpe	30	24	31	34	28	34	30	100	40	30
AIRAP	21	19	31	30	24	28	20	25	100	19
Omega	9	12	25	22	16	24	7	13	32	100
Mean	*18*	*16*	*25*	*24*	*19*	*23*	*18*	*16*	*34*	*18*
Global mean						*21*				
Maximum						*36*				
Minimum						*7*				
	Panel B: Movement to an inferior decile – a difference of more than one decile									
	Sharpe	Sortino	UPR	Calmar	Sterling	Burke	M-Stutzer	M-Sharpe	AIRAP	Omega
Sharpe	100	0	3	3	1	4	0	9	2	0
Sortino	0	100	1	3	1	3	0	6	3	0
UPR	5	1	100	5	3	5	3	6	10	4
Calmar	5	6	8	100	4	0	5	10	9	5
Sterling	1	0	3	1	100	1	0	8	5	0
Burke	5	6	8	0	3	100	5	9	9	5
M-Stutzer	0	0	3	3	1	3	100	8	1	0
M-Sharpe	4	0	1	7	3	7	2	100	11	3
AIRAP	9	9	9	9	9	10	9	14	100	8
Omega	0	0	1	4	2	3	0	7	1	100
Mean	*3*	*3*	*4*	*4*	*3*	*4*	*3*	*9*	*6*	*3*
Global mean						*4*				
Maximum						*14*				
Minimum						*0*				

All calculations of a performance measure with itself, i.e. all 100% values, are not taken into account when calculating the mean.

In light of tables 7 and 8, we can confirm that a significant number of funds display ascending or descending movements after the performance measure change. On average, these funds represent 21% of the population considered in terms of ascending movements (panel A of table 7) and also 21% in terms of descending movements (panel A of table 8).

A closer examination of these cases shows that a category of them, exactly 4% of funds on average, are affected by significant upward or downward shifts in performance classes. In terms of star numbers, this signifies that 4% of funds receive at least more or less two stars than they deserve. 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.

3 Stability of performance measures

The problem of investment funds' performance persistence generally raises two different issues. The first one is of theoretical nature and refers to the market efficiency notion. The second one, which is more pragmatic than axiomatic, is related to fund selection and investors' capital allocation. It is the second issue that we deal with in this section. In the second direction, questions that arise are usually: "Are winners the same?"; "Does this performance comes from the manager's real ability or is it due to chance?". If it is due to chance, it can not last. Such questions are important because they pertain to the predictability of future performances from past performances and thus to the efficient capital allocation question faced by investors.

From an empirical viewpoint, results of the performance peristence of different kinds of investment funds are contradictory. Hence, they do not offer solid proof that past performances are good indicators of funds' future performances. The litterature on hedge funds is not an exception. Agarwal & Naik (2000) found significant persistence when working with quartely returns over the period 1990-1998. Edwards & Caglayan (2001) also stated a certain persistence of alphas (obtained from a 6-factor model) for 1-year and 2-year horizons. Similarly, the findings of Baquero, Horst & Verbeek (2005) indicate a strong and positive persistence for quarterly horizons and a weak positive persistence for annual horizons. In the same spirit, Capocci, Corhay & Hübner (2005) observed a persistence phenomenon among funds having average performances. In contrast, Brown, Goetzmann & Ibbotson (1999), Peskin, Urias, Anjilvel & Boudreau (2000), Schneeweis, Kazemi & Martin (2001) do not find any evidence of persistence of hedge fund performance. The study of Kat & Menexe (2003) goes further to examine the nature of the performance persistence and reveals interesting findings: there is no persistence in the mean returns. However, significant persistence is detected in standard-deviation of returns and in the correlation with stocks; some weak persistence is also found in skewness and kurtosis. It is important to note that these studies often use different performance measures and persistence measures as well as different study periods.

The issue that we deal with in this section is different from that of the above mentionned studies in the sense that instead of investigating the persistence of hedge fund performance,

we examine the stability or the persistence of performance measures. In other words, the study subject is performance measures while the persistence de la performance des 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 measures, among the 10 measures under consideration, which provide rankings that are more or less stable over time. As far as investors are concerned, they need measures that allow them to predict funds' future performances with certain degree of accuracy. Regarding fund managers, they naturally employ measures that are in favour of their performance persistence. From academical viewpoint, such a study highlights possible impacts of the choice of performance measures on the results of performance persistence.

Given this objective, this section is organized in two subsections. We first present the methodology of the analysis (subsection 3.1). Then, we discuss the obtained results (subsection 3.2).

3.1 Methodology

In general, measuring funds' performance persistence consists in examining the relation between the relative performance of funds over a defined period and their relative performance over the next period. Funds are regarded as displaying some performance persistence if this relation is positive and *vice versa*. This notion naturally leads us to the issue of the evaluation period. Since the longest period of our sample is 6 years, it seems appropriate to form two 3-year subperiods and then observe the performance evolution of funds across these subperiods. Although the choice of subperiod length is constrained by the sample that we have, the 3-year length is not completely arbitrary in connection with the lockup period demanded by hedge fund managers. Often varying from one fund to another, this period is generally fixed from 1 year to 3 years. Due to the increasingly volatile context of financial markets over the last years, alternative managers have tendency to demand rather long lockup periods.

Regarding persistence tests, there are two kinds of tests in the literature: parametric tests and non-parametric ones. For this study, we chose non-parametric tests as they are the most used because of their conceptual simplicity, their facility in application and the absence of econometric bias which involve parametric tests. Consequently, two tests are used: the contingency table-based test and Spearman rank correlation test.

3.1.1 Contingency table

Orginally, this test consists in ordering funds in two categories for each of the two periods under consideration: winners and loosers. The criteria used to determine the category to

which a fund belongs is the median performance of the whole sample. According to this mechanism, a fund with performance that is higher than the median performance will be classified in the group of winners and, conversely, a fund with a performance that is smaller than the median performance will be sorted in the group of loosers. After considering the two subperiods, we obtain four categories of funds: winner-winner (WW), winner-looser (WL), looser-winner (LW) and looser-looser (LL). These cateogories are arranged in a table called "contingency table" like the one presented below. In each case, we have the number of funds corresponding to each of the four predetermined categories.

Table 9: Contingency table 2×2 in performance persistence test

2^{nd} **subperiod** / 1^{st} **subperiod**	Superior performance relatively to the median	Inferior performance relatively to the median
Superior performance relatively to the median	n_{WW} Winners	n_{WL} Inconstancy of the performance
Inferior performance relatively to the median	n_{LW} Inconstancy of the performance	n_{LL} Loosers

Once the contingency table is formed, we need to conduct a non-parametric test in order to test statistically the presence ou the absence of a possible persistence. Three concurrent tests can be considered: the Z-test (Malkiel 1995), the Z statistique of the *Odd Ratio*, also known as *Cross Product Ratio* (Brown & Goetzmann 1995) and the Chi-square statistique (Kahn & Rudd 1995). We adopted the third test for this analysis. For a 2×2 contingency table, Chi-square statistique is computed following the formula (1):

$$\chi^2 = \frac{(n_{WW} + n_{WL} + n_{LW} + n_{LL})(n_{WW}.n_{LL} - n_{WL}.n_{LW})}{(n_{WW} + n_{WL})(n_{LP} + n_{LL})(n_{WW} + n_{LW})(n_{WL} + n_{LL})} \sim \chi^2_1 \tag{1}$$

The statistique follows the Chi-square law with one degree of liberty. For a contingency table of l lines and c columns, the Chi-square statistique is calculated according to the formula below:

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{\widehat{n}_{ij} - n_{ij}}{n_{ij}} \sim \chi^2_{(l-1)(c-1)} \tag{2}$$

where $\widehat{n}_{ij}$ denotes the observed frequency at the intersection of the a i^{th} line and the j^{th} column, n_{ij} represents the theoritical frequency (under the null hypothesis) at the intersection of the i^{th} line and the j^{th} column. The statistique hence follows the Chi-square law with $(l-1)(c-1)$ degrees of liberty. When the number of members in each category is smaller than 10, the Yates correction must be applied and it is equal to:

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{(|\widehat{n}_{ij} - n_{ij}| - \frac{1}{2})^2}{n_{ij}} \sim \chi^2_{(l-1)(c-1)} \tag{3}$$

In our analysis, we decided to work with a $4x4$ contigency table in order to test the

performance persistence in terms of quartiles. The idea consists in ordering funds in quartiles on the basis of their relative performance and then determine the number of funds for different scenarios at the end of the two subperiods. This principle is illustrated in table 10.

Table 10: 4×4 contingency table used to test performance persistence

	Sous-période 2			
Sous-période 1	Q1	Q2	Q3	Q4
Q1	n_{Q1Q1}	n_{Q1Q2}	n_{Q1Q3}	n_{Q1Q4}
Q2	n_{Q2Q1}	n_{Q2Q2}	n_{Q2Q3}	n_{Q2Q4}
Q3	n_{Q3Q1}	n_{Q3Q2}	n_{Q3Q3}	n_{Q3Q4}
Q4	n_{Q4Q1}	n_{Q4Q2}	n_{Q4Q3}	n_{Q4Q4}

Two reasons explain this choice. From statistical viewpoint, it allows having a reasonable number of funds in each scenario given our sample size. From financial viewpoint, examining funds' movements across quartiles from one subperiod to the other is meaningful with regard to funds' rankings published by financial press. It is necessary to note that in this case, funds are often classified in quartiles or in quintiles.

The null hypothesis of the Chi-square test is absence of performance persistence. If the calculated statistique value is higher than the value of the theoritical statistique (read from the Chi-square table), then there is performance persistence. In other words, we accept the null hypothesis of persistence absence if the significant level (or the p-value) is smaller than the error level of the test which is often fixed at 5%.

To summarize, the test procedure is as follows:

- For each fund, calculate its absolute performance over each subperiod in question by using the ten performance measures under consideration.
- For each performance measure, order the funds on the basis of their absolute performance and then divide them into quartiles such that each of the first three quartiles ($Q1, Q2, Q3$) contains 37 funds and the fourth and last quartile $Q4$ includes the 38 funds that remain becaus our sample is composed of 149 funds.
- Determine the quartiles to which belongs each fund for each subperiod in order to establish the 4×4 contingency table at the end of the second subperiod.
- Conduct the corrected Chi-square test by following the above described mechanisme.

The Chi-square test as described above aims to determine whether a fund obtains the same relative performance, i.e. stays in the same quartile, in the second subperiod. Consequently, we could say that through this test, the existence of a semi-strong persistence of performance measures is tested.

3.1.2 Spearman rank correlation

The objective of the Spearman rank correlation test is to determine whether a fund displaying a performance will receive the same rank in the following period. In this regard, this test can be viewed as more advanced than the previous one in order to test the existence of persistence in strong form. Applying this test on our sample of 149 hedge funds results in the following steps:

- For each fund, compute its absolute performance over the two predefined subperiods using the ten performance measures under consideration.
- For each performance measure, order the 149 hedge funds for the first and the second subperiods.
- For each performance measure, calculate the Spearman correlation coefficient between the ranking of the first subperiod and that of the second subperiod. Higher the coefficient, higher the similarity between the two rankings.
- Conduct the significativity test for each coefficient in order to valid statistically the similarity (persistence level) or the difference (non persistence) between the two subperiods' rankings.

Whatever the persistence test used, since it is applied to the same sample of funds, we can put forward/suggest that a measure for which the hypothesis of a performance persistence is accepted displays a certain persistence or stability and vice versa. This persistence or stability could be viewed as the existence of some predictability power of this measure on funds' future performance.

3.2 Results

Table 11 reports the ten $4x4$ contingency tables corresponding to the ten performance measures in question while table 12 summarizes results of two persistence tests (contingency table-based test and Spearman rank correlation test).

According to the contingency table-based test (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 signficant at 1% level for Sharpe, Sortino and Calmar ratios, 10% significant level for Burke ratio. This result implies some predictability power of these seven evaluation indicators on the future performance of funds. The three measures in favour of absence of persistence are UPR, M-Stutzer and M-Sharpe. It is interesting to note that in these three cases, the

Table 11: 4×4 contingency table for all performance measures

		2^{nd} subperiod						2^{nd} subperiod				
	Sharpe	Q1	Q2	Q3	Q4	Total	**Burke**	Q1	Q2	Q3	Q4	Total
1^{st}	Q1	7	4	14	12	37	Q1	7	7	12	11	37
sub-	Q2	14	5	9	9	37	Q2	14	10	5	8	37
period	Q3	9	16	3	9	37	Q3	13	9	6	9	37
	Q4	7	12	11	8	38	Q4	3	11	14	10	38
	Total	37	37	37	38	149	Total	37	37	37	38	149
		2^{nd} subperiod						2^{nd} subperiod				
	Sortino	Q1	Q2	Q3	Q4	Total	**M-Stutzer**	Q1	Q2	Q3	Q4	Total
1^{st}	Q1	6	3	16	12	37	Q1	7	4	14	12	37
sub-	Q2	12	10	8	7	37	Q2	13	7	8	9	37
period	Q3	8	16	3	10	37	Q3	9	14	6	8	37
	Q4	11	8	10	9	38	Q4	8	12	9	9	38
	Total	37	37	37	38	149	Total	37	37	37	38	149
		2^{nd} subperiod						2^{nd} subperiod				
	UPR	Q1	Q2	Q3	Q4	Total	**M-Sharpe**	Q1	Q2	Q3	Q4	Total
1^{st}	Q1	7	4	16	10	37	Q1	5	5	15	12	37
sub-	Q2	10	10	8	9	37	Q2	12	9	8	8	37
period	Q3	11	13	4	9	37	Q3	10	10	6	11	37
	Q4	9	10	9	10	38	Q4	10	13	8	7	38
	Total	37	37	37	38	149	Total	37	37	37	38	149
		2^{nd} subperiod						2^{nd} subperiod				
	Calmar	Q1	Q2	Q3	Q4	Total	**AIRAP**	Q1	Q2	Q3	Q4	Total
1^{st}	Q1	5	6	15	11	37	Q1	5	6	15	11	37
sub-	Q2	15	7	8	7	37	Q2	8	7	10	12	37
period	Q3	13	10	4	10	37	Q3	9	14	8	6	37
	Q4	4	14	10	10	38	Q4	15	10	4	9	38
	Total	37	37	37	38	149	Total	37	37	37	38	149
		2^{nd} subperiod						2^{nd} subperiod				
	Sterling	Q1	Q2	Q3	Q4	Total	**Omega**	Q1	Q2	Q3	Q4	Total
1^{st}	Q1	7	5	16	9	37	Q1	7	4	15	11	37
sub-	Q2	11	11	6	9	37	Q2	11	8	9	9	37
period	Q3	11	14	3	9	37	Q3	9	17	2	9	37
	Q4	8	7	12	11	38	Q4	10	8	11	9	38
	Total	37	37	37	38	149	Total	37	37	37	38	149

Table 12: Results of persistence tests

	Persistence in terms of quartiles		**Persistence in terms of ranks**	
	Conclusion	p-value (corrected Chi-square test)	Conclusion	Spearman Coef.
Sharpe	P***	0.008	NP	-0.064070
Sortino	P***	0.007	NP	-0.072338
UPR	NP	0.129	NP	-0.067317
Calmar	P***	0.009	NP	-0.013775
Sterling	P**	0.034	NP	-0.025855
Burke	P*	0.058	NP	0.003849
M-Stutzer	NP	0.131	NP	-0.055777
M-Sharpe	NP	0.150	NP	-0.129712
AIRAP	P**	0.027	**Reversion****	**-0.212741**
Omega	P**	0.014	NP	-0.082162

P indicates persistence while NP indicates the non-persistence. ***, **, * denote the significant level at respectively 1%, 5% and 10% of confidence level. The Chi-square statistique is systematically corrected following the Yates's formula.

null hypothesis (absence of persistence) is accepted at p-values that are quite close to 10%: 12.9% (UPR), 13.1% (M-Stutzer) and 15% (M-Sharpe).

Unlike the Chi-square test's results, those of the Spearman rank correlation test (3^{rd} and 4^{th} columns) are not really conclusive. Eight out of ten measures are negative but not statistically significant. The coefficient associated to Burke ratio is slightly positive but statistically insignificant. The coefficient for AIRAP index is not only statistically negative at 5% level but display an non-neglected level of -0.21. From a financial viewpoint, these results suggest that for nine out of the dix measures under consideration, no rank stability is detected and that AIRAP index seems lead to some reversion in performance. By extrapolating, we can say that except for AIRAP index, no performance measure shows a predictif power regarding funds' ranks.

Concluding remarks

The necessity to have appropriate performance measures that are apt to take into account the characteristics of hedge funds' returns has led to significant development of new measures that are more elaborated and theoretically more efficient that traditionnal ones. However, to the best of our knowledge, there has been no mechanism to test the empirical robustness of these measures, which makes the choice of such or such measure to use quite difficult. The very few studies in the literature observed that funds' rankings established by many performance measures are highly and positively correlated. Hence they concluded that there are no visible influences of the choice of the measures on funds' performance evaluation. Yet, being based solely on correlations between rankings, these results are insufficient to draw any clear-cut conclusion about the quasi-absolute coherence between predefined measures. Given important implications of the performance evaluation's results, we conducted in this paper a comparative study of ten measures documented as the most used by researchers and practionners: Sharpe, Sortino, Calmar, Sterling, Burke, modified Stutzer, modified Sharpe, upside potential ratio, Omega and AIRAP. This study is carried out in two stages on a sample of 149 hedge funds.

First, we examined the modifications of funds' relative performance in terms of ranks and deciles when the performance measure changes. Despite strong positive correlations between funds' rankings established by different measures, which is in perfect concordance with previous work in the literature, numerous significant modifications were observed. While only 11% of funds maintain their initial ranks after the change in performance indicator, 54% suffer from an increase or a decrease of more than 5 places, 16% display a modification of more than 15 places. Besides, many significant moves between deciles are also found: only 58% of funds stay in the same decile, 21% moves to a superior decile; among this 21%, 4% have a shift of more than one decile. Such percentages for descending

shifts are at comparable levels of ascending ones. These findings show how the choice of performance measure is crucial to the evaluation and the selection of hedge funds.

Second, we studied the stability/persistence of the ten measures in question. We defined the stability/persistence of a measure as its capacity to provide stable rankings of funds over time. Higher the ranking's stability, higher the measure's predictive power on funds' future performance. This criteria is assessed at two levels: quartile performance classes (*via* the contingency table-based test) and absolute ranks (*via* the rank correlation coefficient-based test). Our results show that Sharpe, Sortino, Calmar, Sterling, Burke, AIRAP and Omega seem lead to stable rankings in quartiles while UPR, M-Stutzer and M-Sharpe do not. In other words, the first ones display a certain predictive power of funds' future performance. In contrast, in terms of absolute rankings, except for AIRAP index which seemingly conducts to some performance reversion, no measure provides rank stability. This result suggests that no measure is capable to predict future ranks of funds in a quite precise manner.

Our findings have important implications for investors (to select funds and then follow them up) as well as for fund managers (to present their funds' performance to their clients). Whatever the role of the users, it is in their best interest to favour those measures that are as stable/persistent as possible.

References

Agarwal, V. & Naik, N. Y. (2000), 'On Taking the "Alternative" Route: The Risks, Rewards, and Performance Persistence of Hedge Funds', *Journal of Alternative Investments* **2**, 6–23.

Baquero, G., Horst, J. t. & Verbeek, M. (2005), 'Survival, Look-Ahead Bias, and Persistence in Hedge Fund Performance', *Journal of Financial & Quantitative Analysis* **40**(3), 493–517.

Brown, S. & Goetzmann, W. (1995), 'Performance Persistence', *Journal of Finance* **50**(2), 679–698.

Brown, S. J., Goetzmann, W. N. & Ibbotson, R. G. (1999), 'Offshore Hedge Funds: Survival and Performance, 1989-95', *Journal of Business* **72**(1), 91–117.

Capocci, D., Corhay, A. & Hübner, G. (2005), 'Hedge Fund Performance and Persistence in Bull and Bear Markets', *European Journal of Finance* **11**(5), 361–392.

Capocci, D. & Hübner, G. (2004), 'Analysis of Hedge Fund Performance', *Journal of Empirical Finance* **11**(1), 55–89.

Edwards, F. R. & Caglayan, M. O. (2001), 'Hedge Fund Performance and Manager Skill', *Journal of Futures Markets* **21**(11), 1003–1028.

Eling, M. & Schuhmacher, F. (2005), 'Hat die Wahl des PerformancemaBes einen Einfluss auf die Beurteilung von Hedgefonds-Indzes?', *Kredit und Kapital* **39**.

Eling, M. & Schuhmacher, F. (2006), Does the Choice of Performance Measure Influence the Evaluation of Hedge Funds?, Technical report. Paper presented at the 2006 FMA European Conference.

Favre, L. & Galeano, J.-A. (2002), 'Mean-Modified Value-at-Risk Optimization with Hedge Funds', *Journal of Alternative Investments* **5**(2), 21–25.

Kahn, R. N. & Rudd, A. (1995), 'Does Historical Performance Predict Future Performance?', *Financial Analysts Journal* **51**, 43–52.

Kat, H. M. & Menexe, F. (2003), 'Persistence in Hedge Fund Performance: The True Value of a Track Record', *Journal of Alternative Investments* **5**(4), 66–72.

Kooli, M., Morin, F. & Sedzro, K. (2005), Evaluation des Mesures de Performance des Hedge Funds. Paper presented at the Annual Conference of the French Association of Finance, June.

Kouwenberg, R. (2003), 'Do Hedge Funds Add Value to a Passive Portfolio?', *Journal of Asset Management* **3**(4), 361–382.

Malkiel, B. (1995), 'Returns from Investing in Equity Mutual Funds: 1971 to 1991', *Journal of Finance* **50**(2), 549–572.

Peskin, M., Urias, M., Anjilvel, S. & Boudreau, B. (2000), Why Hedge Funds Make Sense?, Quantitative strategies paper, Morgan Stanley.

Schneeweis, T., Kazemi, H. & Martin, G. (2001), Understanding Hedge fund performance: Research Results and Rules of Thumb for the Institutional Investors, Research paper, Lehman Brothers.

Sharma, M. (2004), 'AIRAP - Alternative RAPMs for Alternative Investments', *Journal of Investment Management* **2**(4), 106–129.