Journal of Statistical and Econometric Methods, vol.2, no.1, 2013, 81-114
ISSN: 1792-6602 (print), 1792-6939 (online)
Scienpress Ltd, 2013
array multirow

Tracking Hedge Fund Performance: A Balanced Sampling Strategy

Donatien Tafin Djoko[1]

Abstract

This paper discusses the application of statistical, survey sampling technique to hedge fund tracking problems. I describe a strategy that allows an investor or a fund of hedge funds manager, to construct a small tracking portfolio that replicates the time series changes of the total relative Net-Asset-Value (NAV) of a large basket of funds. The trackers are constructed using a method of balanced sampling, in which components are selected randomly with unequal inclusion probabilities. Empirical studies are performed on directional hedge fund styles: Commodities trading advisors, Global macro and Equity hedge. In all cases, empirical results show that the proposed strategy replicated efficiently the total fund's relative NAVs using only ten percent of the sample. The constructed portfolios are stable in the long run, allowing the investor to implement a passive investment strategy in the alternative investment universe. I also consider a larger sample of funds, mixing the aforementioned category of hedge funds. The market tracking ability of balanced sampling remains statistically significant.

Jel classification: G11, C42, C51, C63
Keywords: Hedge fund, Fund tracking, Survey sampling, Balanced sampling

[1] University of Neuchâtel, Faculty of Economics and Business, Institute of Statistics, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland, e-mail: donatien.tafin@unine.ch

Article Info: *Received* : December 6, 2012. *Revised* : January 14, 2013
Published online : March 1, 2013

1 Introduction

While investing in hedge funds (HF) yields several advantages such as double digit returns and diversification opportunities, the selection of funds in which to invest is a highly challenging issue. This challenge has important implications for the ex-post investor's risk-return structure. In this article, the attention will be directed solely to the selection problem. How to design a strategy to reproduce closely the behavior of a large universe of HF with a subset of his constituent. The goal hence is not to maximize the investor wealth, but to reproduce faithfully the time series changes of HF full sample performances.

Driven by the strong growth of indexing in the traditional equity universe, a large range of models have been implemented to address the market tracking issue. [9] formulated the market tracking problem, also known as dimension reduction problem, as a mean-variance optimization, making fundamental assumptions on the distribution of the replicated benchmark. However the main difficulty with this approach is the computational burden associated with estimating the variance-covariance matrix for the returns of all assets within the index. [4] extend the Markowitz-based formulations and derived a meta-heuristic decision support for portfolio management and optimization. Using a linear factor model based on macro-economic variables, they estimated the covariance matrix and then solved the tracking problem. In addition to the traditional tracking error minimization approach, [1] advocated a cointegration-based model. Basically, this implementation relies on the fact that the difference between the benchmark and the tracking portfolio is by construction supposed to be stationary. That is, in the long run the deviation of the tracker to his benchmark will be small. Furthermore, [6] illustrated a time-series clustering approach. They argue that clustering might be an optimal methodology for building a sparse tracking portfolio when full replication of an aggregate basket of assets is not feasible. By suitably defining Euclidean distances between the time series of asset prices available in a given market, a hierarchical clustering 'discovers' the correlation structure of the target benchmark.

The purpose of this manuscript is to extend the market tracking principle to the HF world. There are four main reasons to study HF indexing models. First, in practice an investment manager is exposed to a relatively small num-

ber of individual HF. According to [8], a portfolios of approximately 10 funds are sufficient to construct a diversified portfolio of HF. Second, a portfolio of HF presents specific characteristics such as very high transaction costs, liquidity constraint (notice periods, lockup periods and redemption restriction), and slow rebalancing cycles. As a consequence, an allocation in HF must be inspected as a long term investment. HF indexation may ensure that the tracking portfolio is tied to the fund universe in the long run. Thirdly, the selected tracker should result in a stable portfolio, filtering out the noisy part of the joint variation in the sample. Fourth, if the fund benchmark performances can be reconstructed, we considerably reduce the capital incentive and qualitative analysis burden associated with managing a large portfolio of funds.

Most previous work on tracking financial markets with a subset of its constituents focuses on hard-to-solve optimization techniques and econometric models. However, optimization routines tend to construct highly concentrated replicas, while econometric models such cointegration require consistent estimation of error term and/or structural loadings with direct specification of the data generating process. In this article, I adopt a different perspective and implement an automatic survey sampling approach to select small portfolio of funds in order to track the total HF universe relative NAVs. This small portfolio is selected randomly by means of a method of balanced sampling. [10] developed and implemented the balanced sampling replication strategy in the case of the S&P 500 total market capitalization. To not be mistaken, the term 'balanced' here does not allude to its financial connotation, but instead refers to the statistical sampling meaning. Statistically speaking, a sample is said to be balanced if the estimator of the mean matches or is approximately equal to the population mean. Hence, balanced sampling is implemented to construct a tracking portfolio whose relative NAV is proportional, at any trading period, to the total relative NAV of a large basket of funds. By balancing the randomly selected replication portfolios on principal components, the proposed sampling design ensures that the trackers constantly reconstruct the time series evolution of the entire basket of funds, while preserving its diversification property. The diversification principle is automatically incorporated in the efficient balanced sampling procedure without imposing additional constraints.

This article makes several contributions. I move from a hard to solve com-

plex optimization problem, to a simpler and more intuitive survey sampling framework. The issue of efficiently selecting a subset from a large population is a natural sampling topic. There are obviously various ways of randomly selecting a sample. I show that this method can easily generate random portfolios that efficiently track the total performance of HF and maintains the diversification property. In addition to the balanced sampling design, the article discusses three alternative random selection strategies and a simple heuristic for constructing a tracking portfolio consisting of funds with the biggest asset under management (AUM). One of the computational benefits of the method is the possibility to run a large number of simulations to evaluate the tracking strategy without increasing the computational difficulty. Furthermore, instead of tracking the return series, we focus on the relative NAV.

The outline of the manuscript is as follows. In Section 2, I define the notations and state the aim of the study. Section 3 formulates our approach for modeling relative NAVs within the statistical survey sampling framework and for investigating HF indexing. I discuss the computation underlying the inclusion probabilities, which is made proportional to individual fund relative NAV. Then, A short review of survey sampling theory is introduced. Section 4 presents the empirical results. I report both a multivariate analysis based on selecting $10,000$ portfolios, and a univariate characterization of the methodology according to the minimum variance portfolio. Further on, I compare the results of our sampling strategy with three alternative random sampling approaches and a simple heuristic involving the construction of a tracking portfolio consisting of funds with the biggest asset under management. Finally, I provide a robustness check of our main results. Section 5 concludes the article.

2 Background

To crystalize the context of the analysis, let denote by $R_t = NAV_t/NAV_{t-1}$ the vector of relatives net-asset-value at time t. We have available d funds, constituents of the HF universe with $R_t = (R_t^1, \cdots, R_t^d)$, $t = 1, 2, \cdots, T$, where the jth component $R_i^{(j)}$ of R_i denotes the relative NAV of fund jth on the ith investment period. If T denote the current time, we want to choose the

portfolio of HF to track the HF universe at time $T+1$. The total value of the HF universe at time t may be expressed as

$$r_t = \sum_{i=1}^{d} R_t^i \tag{1}$$

We assume the usual hypothesis that any fund is infinitely divisible. Any fraction between $[0, 1]$ can thus be selected in the portfolio.

The aim is to select a small portfolio of n funds at time τ that is a subset of a large sample of HF. We would like the portfolio to track the HF universe as closely as possible. To keep the model theoretically simple, we will allocate the same amount to each fund in the tracking portfolio. It is not difficult to generalize the following theory to situations where all the amounts allocated are not equal. In fact during the empirical investigation, the allocation is adjusted according to the inclusion probabilities. Additionally, in the sequel I use the term market to denote the the HF universe.

Let us assume that w_i is the level of capital invested in fund i within the tracking portfolio, and I_i be the indicator variable that takes value 1 if HF i is selected in the portfolio and 0 otherwise. We thus have

$$\sum_{i=1}^{d} I_i = n \tag{2}$$

Variables I_i are supposed to be random with a Bernoulli distribution. We suppose that π_i is the probability of selecting unit i in the tracking portfolio. In the following lines, R_t^i and the capital allocated in each fund w_i are not supposed to be random. The only source of randomness is I_i. The expectations and variances are then computed with respect to I_i. 0.7 cm

3 Methodology

This section presents the methodology for solving the fund selection problem in the context of market tracking. I first characterize the tracking portfolio value at any trading time t and the implications in the random selection process. I then describe how to compute the probability of selecting a fund i in our replicating portfolio. The selection probabilities are defined proportionally to the total market relative NAV. Next, I present the balanced sampling design.

For a complete description of sampling algorithms, see [11]. I approximate the cube method as in [5] to select portfolios that are unbiased, with small variances. Finally, I present the algorithm designed to select a balanced sampling portfolio.

3.1 Tracking portfolio value

Given our our interest in constructing a tracking portfolio of HF, we build a model that provides a clear description of the tracker values, V_t. If the portfolio is constructed at time τ, therefore its value is defined as

$$V_\tau = \sum_{i=1}^{d} I_i w_i$$

and for any other t, the factor by which every amount (w_i) invested in the i-th fund grows during the tracking period is given by R_t^i/R_τ^i. Therefore, the portfolio value at time t is expressed as

$$V_t = \sum_{i=1}^{d} w_i I_i \frac{R_t^i}{R_\tau^i}$$

If all the constituents within the portfolio have the same allocation, ie, $w_i = w$, then

$$V_\tau = w \sum_{i=1}^{d} I_i = wn$$

and for any other time t

$$V_t = w \sum_{i=1}^{d} I_i \frac{R_t^i}{R_\tau^i} \tag{3}$$

The portfolio will be constructed randomly by means of a sampling technique. To satisfy our tracking objective at time τ, that is, to build a portfolio that is as close as possible to the complete market during the period $\tau - q, \ldots, \tau$, we would like

$$\frac{V_t}{V_\tau} \approx \frac{r_t}{r_\tau}, \text{ for all } t = \tau - q, \ldots, \tau \tag{4}$$

Equations (4) imply q linear constraints on V_t. However, in the market tracking issue, the replicator is bound by the limited number of n securities out of the market d ($n \ll d$) that can be selected to track the target benchmark.

Since the number of observations q is larger than n, we can not satisfy more constraints than the number of units in the tracking portfolio. For this reason, we should reduce the number of constraints that must be satisfied. In what follow, period $\tau - q$ to τ is referred to as the portfolio estimation period, while the period that follows is referred to as the investment period.

3.2 Selection probability

The key and powerful characteristic of our tracking strategy is randomness. Each fund is selected randomly according to a given probability mass. Let π_i be the probability of selecting individual HF i. Thus, we can write

$$\pi_i = Pr(\text{selecting fund } i \text{ in portfolio } j) = \mathrm{E}(I_i),$$

and given Equation (2)

$$\sum_{i=1}^{d} \pi_i = \sum_{i=1}^{d} \mathrm{E}\left(I_i\right) = \mathrm{E}\left(\sum_{i=1}^{d} I_i\right) = \mathrm{E}(n) = n \tag{5}$$

the sum of the selection probabilities is thus equal to the number of funds in the tracking portfolio.

Since each unit in the portfolio have the same allocation w at time τ, a convenient alternative for computing the inclusion probabilities is to construct the vector of π_i proportional to the relative NAVs at time τ, that is,

$$\pi_i \propto R_\tau^i \ , \ i = 1, \ldots, d$$

Given Equation (5), the selection probabilities can easily be derived according to the following expression

$$\pi_i = \frac{nR_\tau^i}{r_\tau} \tag{6}$$

Consequently, it is evident that the portfolio value at any given trading period, is proportional to the total market relative NAV. Effectively, a short decomposition of Equations (3) and (6), gives

$$\begin{aligned} \mathrm{E}\left(V_t\right) &= \mathrm{E}\left(w \sum_{i=1}^{d} I_i \frac{R_t^i}{R_\tau^i}\right) = w \sum_{i=1}^{d} \frac{R_t^i}{R_\tau^i} \mathrm{E}\left(I_i\right) = w \sum_{i=1}^{d} \frac{R_t^i}{R_\tau^i} \pi_i \\ &= w \sum_{i=1}^{d} \frac{R_t^i}{R_\tau^i} \frac{nR_\tau^i}{r_\tau} = \frac{nw}{r_\tau} \sum_{i=1}^{d} R_t^i = nw \frac{r_t}{r_\tau} \end{aligned}$$

where r_τ is expressed as in Equation (1).

Indeed, if the allocation decision consists of investing the same amount to each fund, the selection probabilities are proportional to the total relative NAV in the HF universe. This strategy is unbiased under the sampling design and increases the likelihood that some funds with large relative NAV will be selected. The randomness of the selection scheme ensures that all constituents are given the same opportunity to enter the tracking portfolio and not only those that are 'tied' to the market. The diversification is thus automatically guarantied. This procedure also avoids selecting a disproportionate number of funds whose performs badly. The aim is thus not only to track the total market relative NAV but predominantly to keep its intrinsic diversification. In other words, keeping in balance the relative structure between highly and modestly performing funds in the tracking portfolio.

3.3 Balanced sampling design

Given my interest in tracking the total relative NAV of funds in the HF universe with a subset of its constituents, I build a model based on sampling methods. Sampling can be defined as a process of selecting statistical units from a population to estimate unknown quantities of the population. In this subsection, I present a set of results developed in the framework of survey sampling theory in order to provide a very efficient estimation of a total. [5] have proposed a sampling algorithm, called the cube method, which enables selection of a sample from a register with approximately the same means in the sample as in the population for all the variables of the register. This sample can be selected with equal or unequal selection probabilities. For more details on sampling algorithms see [11].

In survey sampling, statistical units are generally establishments, businesses, people or households. The aim is to estimate through a sample, a total of the variable of interest y_i,

$$Y = \sum_{i=1}^{d} y_i$$

The values y_i are not supposed to be random. The only source of randomness is the process of selecting the sample. A sample is a subset of the population that is selected randomly. Random samples can be selected by several sampling

designs such as simple random, stratified, and balanced sampling. Statistical units can be selected with equal or unequal selection probabilities. Let π_i be the probability of selecting a given unit and I_i be the indicator random variables that takes the value 1 if unit i is in the sample and 0 otherwise. The expectation of I_i is equal to π_i. If all the selection probabilities are positive, the [7] estimator given by

$$\widehat{Y} = \sum_{i=1}^{d} \frac{y_i}{\pi_i} I_i$$

is unbiased for Y.

In order to construct an efficient sampling design, one generally uses auxiliary information that is known for all the population units, for instance a register. Let $\mathbf{x}_i$ be a vector of $\mathbb{R}^p$ containing the p values taken by the p auxiliary variables on statistical unit i. The $\mathbf{x}_i$ are supposed to be known on all the units of the population. A sample is said to be approximately balanced on a set of auxiliary variables if

$$\sum_{i=1}^{d} \frac{\mathbf{x}_i}{\pi_i} I_i \approx \sum_{i=1}^{d} \mathbf{x}_i \tag{7}$$

This means that the Horvitz-Thompson estimator of the total is very close to the true population total for all the variables that are known at the population level. Unfortunately, a sample can rarely be exactly balanced, because the selection of a sample is an integer number problem (a unit is selected or not in the sample). It is thus rarely possible to exactly satisfy the p balancing equations given in Expression (7). This is called the 'rounding problem'. In this case, the objective is to find the best possible approximation. For expected large sample size, the rounding problem is negligible, but not in the current financial application.

In this paper, our aim is to select a portfolio of funds using this technique of balanced sampling. We will thus randomly select a subset of individual HF to reconstruct as close as possible, in the long run the time series changes total relative NAVs in the universe of funds. Since our principal objective is to implement a passive investment strategy in the alternative investment world, our tracker has to be manageable in terms of size, transaction costs, and rebalancing cycles. In other words, the balanced sampling replicator should result in a steady portfolio structure.

Under these specifications, statistical units are fund's relative NAV, and the auxiliary information is the set of values taken by each fund during a period. In order to construct the tracking portfolio, we refer to the following result.

Result 1. *For any vector of selection probabilities* $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_i, \ldots, \pi_d)^t$ *such that*

$$0 < \pi_i \leq 1 \text{ and } \sum_{i=1}^{d} \pi_i = n \in \mathbb{N}$$

there is a random vector $\mathbf{s} = (s_1, \ldots, s_i, \ldots, s_d)^t \in \mathbb{R}_N$ *such that*

(i) $\sum_{i=1}^{d} \frac{\mathbf{x}_i}{\pi_i} s_i = \sum_{i=1}^{d} \mathbf{x}_i$

(ii) $\mathrm{E}(s_i) = \pi_i$, *for* $i = 1, \ldots, d$,

(iii) $0 \leq s_i \leq 1$, *for all* $i = 1, \ldots, d$,

(iv) $\mathrm{card}\{I | 0 < s_i < 1\} \leq p$, *for all* $i = 1, \ldots, d$

where $\mathrm{card}(U)$ *is the cardinality of set* U *and* p *denotes the number of auxiliary variables* $\mathbf{x}_i$.

The proof of Result 1 directly follows from [5]'s Result 1 and Section 4 of the same paper, where an algorithm is proposed to generate a balanced sample. This algorithm is called the 'flight phase of the cube method'. A faster algorithm is proposed in [3] and a complete development on the methods of balanced sampling is given in [3]. A generator of balanced vectors is implemented in the R language, in the 'sampling' package (function fastflightcube). The cube denomination is related to the fact that each random vector $\mathbf{s}$ can be viewed as a vertex of a d-cube.

The interpretation of Result 1 is as follows. The strict equality of Expression (7) cannot be obtained by the I_i's that only take the values 0 and 1. Nevertheless, it is possible to obtain a strict equality given as Result 1 *(i)* if we accept that at most p values for the s_i are not equal to 0 or 1 but can take a value between 0 and 1. Consequently, the balancing equations will be more difficult to achieve if the number of auxiliary variables is large. In our alternative investment case study, the number of auxiliary variables is the number of observations during the period used to construct the portfolio. This number can thus be large. A dimensionality reduction technique is therefore necessary.

In the next sections, vector $\mathbf{s} = (s_1, \ldots, s_i, \ldots, s_d)^t$ is called a balanced vector on matrix

$$\mathbf{A} = \left(\frac{\mathbf{x}_1}{\pi_1} \ldots \frac{\mathbf{x}_i}{\pi_i} \ldots \frac{\mathbf{x}_d}{\pi_d} \right)^t$$

with probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_i, \ldots, \pi_d)^t$ and will be denoted by

$$\mathbf{s} \sim BV(\mathbf{A}, \boldsymbol{\pi})$$

Random vector $\mathbf{s}$ is such that $\mathrm{E}(\mathbf{s}) = \boldsymbol{\pi}$ and the number of values in $\mathbf{s}$ that are not equal to 1 and 0 is less or equal to the number p of columns of $\mathbf{X}$. Result 1 *(i)* can then be written with a matrix notation:

$$\mathbf{A}^t\mathbf{s} = \mathbf{A}^t\boldsymbol{\pi}$$

If the rows of matrix $\mathbf{A}$ correspond to the number of funds in HF universe, and the columns of matrix $\mathbf{A}$ correspond to the observations, random vector $\mathbf{s}$ is a portfolio that reproduces the characteristics of the funds described in matrix $\mathbf{A}$. Unfortunately, the number of observations is generally large. Vector $\mathbf{s}$ will thus contain a small number of integer values. In order to apply balanced sampling, we thus need to reduce the dimensionality of the data. This is achieved trough singular value decomposition of the matrix of relative NAVs. The decomposition shows that approximately few components are enough to reproduce the matrix relative NAVs. Consequently, instead of working on matrix $\mathbf{A}$, I used $\mathbf{A}^k$ where k are the first components from the singular value decomposition. See the Appendix for a detail explanation of the procedure.

3.4 Implementation details

The purpose of this subsection is to discuss an approximation of the balanced sampling design. This is important in the context of our analysis. It allows to implement the theoretical results of the methodology highlight in Subsection 3.3 to various alternative investment data. Accordingly, [10] develop Algorithm (3.1). Basically, the motivation behinds Algorithm (3.1) is to be able to apply a balanced sampling framework to a limited number of dimensions. At the beginning of the algorithm, the method is applied to the largest possible dimension that is equal to the minimum between the total number of individual HF in the data set minus one $(d-1)$, and the length of

the estimation period q. At each iteration, the dimension is reduced in order to progressively round all the components of $\mathbf{s}$ (the balanced vector) to 0 or 1. Finally a portfolio of n funds is selected in such a way that it is as balanced as possible. In market tracking terminology, this means that the selected (balanced) portfolio is tied in the long run, to the total fund's relative NAV.

Algorithm 3.1 (H). *Selecting the portfolio*

- *Initialize by setting* $\mathbf{s}^{(0)} = \boldsymbol{\pi}$
- *Define* $p = \min(d-1, q)$
- *For* $j = 0$ *to* p *do* $\mathbf{s}^{(j+1)} = BV(\mathbf{A}^{(p-j)}, \mathbf{s}^{(j)})$

Note that $\mathbf{s}_{(j)}$ is a martingale because $\mathrm{E}(\mathbf{s}^{(j+1)}|\mathbf{s}^{(j)}) = \mathbf{s}^{(j)}$. By using iterated conditioning, we obtain that $\mathrm{E}(\mathbf{s}^{(j)}) = \boldsymbol{\pi}$ for all $j = 1, \ldots, p$. Therefore, Algorithm (3.1) satisfies the given inclusion probabilities. At step 1, $\mathbf{s}^{(1)}$ contains p non integer values. At step 2, $\mathbf{s}^{(2)}$ contains $p-1$ non integer values. At step j, $\mathbf{s}^{(j)}$ contains $p-j$ non integer values. At step p, all the components of $\mathbf{s}^{(p)}$ are either equal to 0 or to 1. At the end of the algorithm, the final vector $\mathbf{s} = \mathbf{s}^{(p+1)}$ contains only 0 and 1 and is an approximately efficient balanced sample.

This model has several appealing features. First, it is distribution free. There is no assumption on the data generating process of HF's relative NAV and the optimal tracker is derived without specifying any structural form between the total market value and its individual constituents. Under the balanced sampling design, the selection technique is statistically efficient. In other words, the selected portfolios are guarantied to be unbiased with a small variance. Third, the model yields tracking portfolios that are automatically diversified without specific heuristic hypothesis, allowing the investor to perform allocation in an optimal way. Fourth, the proposed replication strategy is highly flexible, providing an alternative insight to the computational intensive constraint portfolio choice.

4 Empirical Analysis

4.1 Data description

To attest that the proposed methodology provides general results, I used three different HF categories to evaluate the statistical significance balanced sampling tracking portfolios. We focus on live funds from two main HF data providers: Hedge Fund Research (HFR), and Barclay Commodity Trading Advisors (CTAs) databases. The former data provider is composed of returns reported on different frequencies (mainly monthly) and additional qualitative/quantitative information such as main strategy, sub-strategy, asset under management, $\cdots$ etc. on 5230 individual funds and 2720 Funds-of-Funds that are still active on September 30, 2008, while the later data system is widely recognized by both practitioners and academicians as the largest, most comprehensive, available Commodity Trading Advisors sample. Generally, CTAs are funds primarily trading listed commodity and financial futures contracts. The database consists of 981 reporting funds as of September 2008. CTAs, also denominated Managed futures are by no means homogeneous investment vehicles. CTAs managers employ a large range of strategies and asset classes. Combined, the two databases provide a rather completed and detailed picture of the hedge funds universe.

The alternative investment industry data are subject to different biases resulting from the data collection process. Fund managers *voluntarily* provide information to databases and the industry lacks an uniform reporting standard. Since I confine the analysis to funds in the live database, I recognized that it suffers from survivorship bias. However, the relevance of such a bias for the current investigation limited by the fact that it will affect both the tracker and the full population of funds. The tracking portfolio is just a reconstruction of what we have in the large sample.

For the scope of our empirical analysis, we impose a set of filters on both databases. First we selected funds that reported in U.S. dollar net of all fees on a monthly basis. Then we required that each fund has at least 12 years of reported returns. Additionally, for each strategy, we extract the group of funds with the longest consecutive stretch of non-missing NAVs and AUM. The analysis is performed on HF directional strategies: CTAs funds from Barclay; Global Macro and Equity Hedge funds from HFR. After imposing these

Table 1: Summary statistics on hedge fund relative NAVs

	Barclay CTAs		Global Macro		Equity Hedge	
Sample Size	134		101		108	
	Mean	SD	Mean	SD	Mean	SD
Mean	1.011	0.006	1.008	0.004	1.012	0.005
Mean t-stat	2.421	1.132	2.085	1.047	3.738	2.939
SD	0.058	0.03	0.048	0.022	0.048	0.024
Skewness	0.447	0.754	0.303	1.235	0.025	0.879
Skew t-stat	2.165	3.655	1.372	5.592	0.125	4.379
Kurtosis	5.586	4.551	6.372	8.66	6.058	4.778
Kur t-stat	6.268	11.031	7.633	19.604	7.618	11.904
Min	0.845	0.109	0.865	0.116	0.844	0.092
Max	1.223	0.141	1.166	0.077	1.184	0.125
JB	178.024	826.616	471.629	3289.655	217.447	949.74
LB(6)	10.129	7.363	8.693	6.861	20.741	77.413
$\rho(R)$	7.922	4.734	6.866	4.391	6.934	4.822
LM(6)	7.879	7.856	6.171	5.385	14.29	11.043

This table reports monthly average summary statistics of HF relative NAVs for the entire sample period (estimation and investment periods). The mean and t-statistic value, the standard deviation, the standardized skewness and kurtosis with their respective t-statistics, the Jarque-Bera normality test statistic (JB), the Ljung-Box test statistic for no serial correlation [LB(6)], the first-order serial correlation of relative NAVs [$\rho(R)$], the Lagrangian multiplier test statistic for no serial correlation in relative NAVs [LM(6)]. The critical values at 5% are 5.99 for JB, 12.59 for LB(6), $\rho(R)$, and LM(6).

restrictions, the data comprises 134 CTAs covering the period from October 1996 to June 2008, for a total of 141 observations; 101 Global Macro funds from July 1997 to September 2007, giving 141 observations; and 108 Equity Hedge funds from May 1995 to September 2007 for a total of 149 monthly observations. To avoid in-sample spurious findings, these samples period are divided in two sub-periods: The first 60 months are used for backtesting the quality of the model (also called estimation period), while the second samples (longer or equal to the first one) are used for the out-of-sample investigation (called investment period).

Table 1 reports several summary statistics for the fund relative NAVs under study. The statistics are computed over the entire data period for each HF

and then averaged across funds within each HF category. The monthly average relative NAVs are all bigger than one and significant, ranging between 1.008 to 1.012 with small standard deviations amongst funds. The average monthly volatilities range between 6% and 5%. Equity hedge funds tends to exhibit high expected relative NAVs with low risk, whereas Global Macro funds have low expected relative NAVs with low risk. Barclay CTAs funds have high average relative NAVs and high risk.

Skewness measures are all positive across HF categories, suggesting that periods of booms are more present than periods of crashes. Kurtosis measures are equal to 5.6 for Barclay CTAs funds, 6.4 for Global Macro and 6.1 for Equity Hedge funds, values that are not in line with the normality assumption. Consequently, we reject the normality assumption with strong confidence for all HF categories in the sample. Except for Equity Hedge funds, in average, I don't find strong evidence for serial correlation and ARCH effect in fund's relative NAVs.

4.2 Balanced sampling tracking portfolio

We now turn to the analysis of the performance of the market tracking ability of the strategy described above. The goal here is to construct a long term and stable buy-and-hold strategy in the alternative investment world.

The balanced sampling portfolio selection strategy implemented for 3 HF categories under study is performed as follows. I divide the sample in two non-overlapping sub-periods. The first 60 months are used to select the tracking portfolios and perform a backtest of the method. The second sub-period is dedicated to the out-of-sample analysis. In the sequel I refer to the first sub-period as estimation period and to the second as investment period. In all the figures, the vertical line separates the two sub-periods. The number of funds selected in each tracking portfolio, $n = 10$. Figure 1 represents less than 10% of the population of interest across HF categories.

In finance, there is general consensus on how difficult is to separate skill from luck. This issue is even more stringent for hedge funds investors. The opacity inherent from the HF industry makes hard to distinguish knowledgeable fund picking strategies from lucky ones. The flexibility of our method allows us to avoid that by running a large number of simulations. This mul-

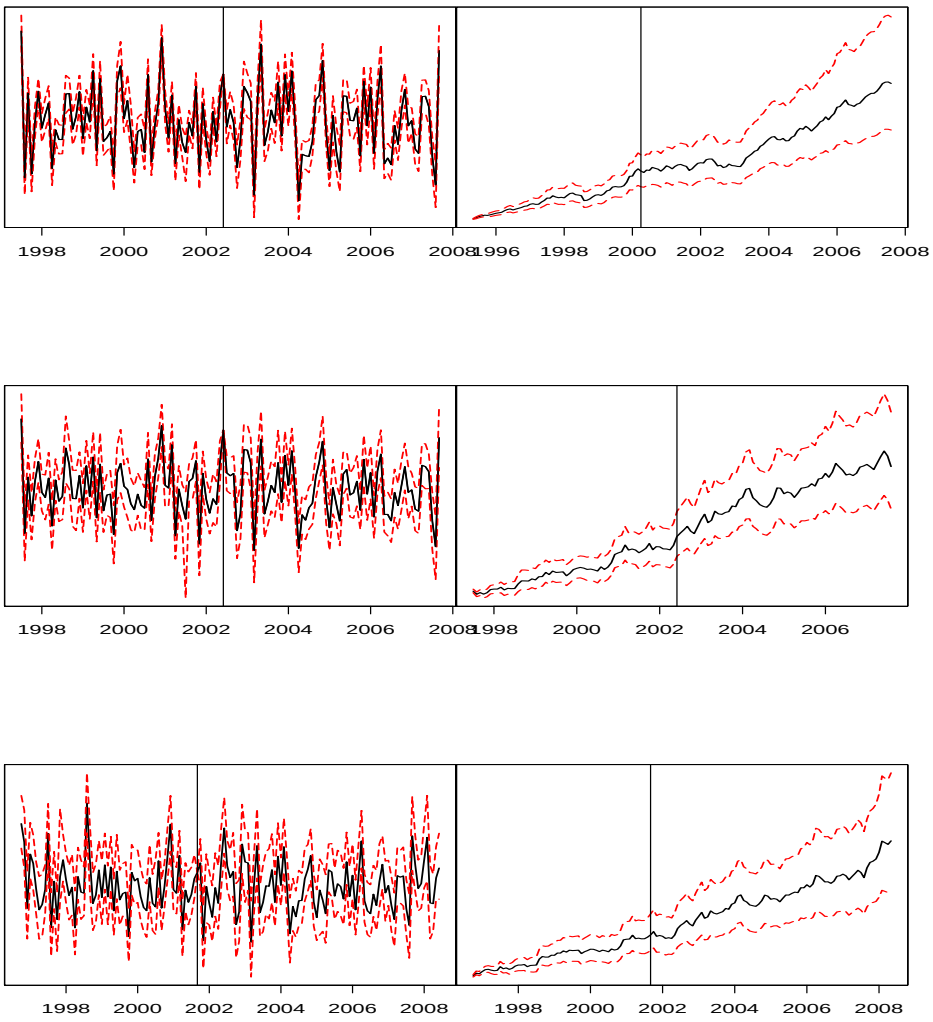

Figure 1: The first-column entries are the total HF relative NAVs (black line) embedded in the 5th and 95th percentiles (red line) of the corresponding values for a selection of 10,000 portfolios derived using the balanced sampling tracking strategy. The second-column entries represented the wealth achieved investing in an equal weighted portfolio of the full sample of funds (black line) embedded in the 5th and 95th percentiles (red line) of the corresponding values for a selection of 10,000 portfolios with the proposed strategy. From the first to the third row, we have respectively Barclay CTAs, Global Macro, and Equity Hedge funds. Each tracking portfolio is composed of $n = 10$ funds.

tivariate analysis is useful to assess the statistical quality of the sampling design. I fixed the number of simulations to $10,000$, that is, we construct $10,000$ tracking portfolios for each HF category. For each HF style, the inclusion probabilities are proportional to the total relative NAVs in the last month of the estimation period (60th month).

The first column-entries of Figure 1 display the dynamics of the total relative NAVs and, in dashed lines, the 5th and 95th percentiles relative NAVs of the $10,000$ portfolios selected according to Algorithm (3.1). The second column-entries of Figure 1 shows the evolution of the wealth achieved investing 1 USA \$ in an equal weighted portfolio constituted of the entire sample of funds with the corresponding 5th and 95th percentiles from $10,000$ balanced sampling trackers. Inspection of these figures highlight how efficiently the selected portfolios tracked the total relative NAVs. The performance is significant in both estimation and investment sub-periods. The inter-percentile range is narrowed in the two sub-periods and the gap is relatively stable in the long run.

The graphs in Figure 2 display the dynamics the deciles of the ratios between the $10,000$ simulated tracking portfolio values and total HF relative NAVs (first column) for the three categories of alternative investment under study. More precisely, the ratios are defined as:

$$Ratio_t = \frac{V_t}{V_\tau}\frac{r_\tau}{r_t} \tag{8}$$

where $\tau = 60$ refers to the portfolio estimation date. The corresponding ratio for the wealth achieved is displayed in the second column of Figure 2. Several comments are of interest. First, with only 10 funds, the sample of trackers remains very 'close' to the portfolio investing in all the funds in the universe. The method of balanced sampling works well in the sense that the deviation during the estimation period is very small. During the in-sample empirical analysis, the ratio dynamics of relative NAVs range between 0.97 and 1.04 (except for one large swift for Global Macro funds), with a deviation close to zero on average. At allocation time (60th month), all the portfolios have the same value and $Ratio_\tau = 1$. After this month, the portfolios start to deviate from the total market value, but stay in a narrow range. For the wealth achieved, we observe comparable dynamics. It is interesting to notice that in expectation, the tracking portfolio ratios stay well around 1, which indicates

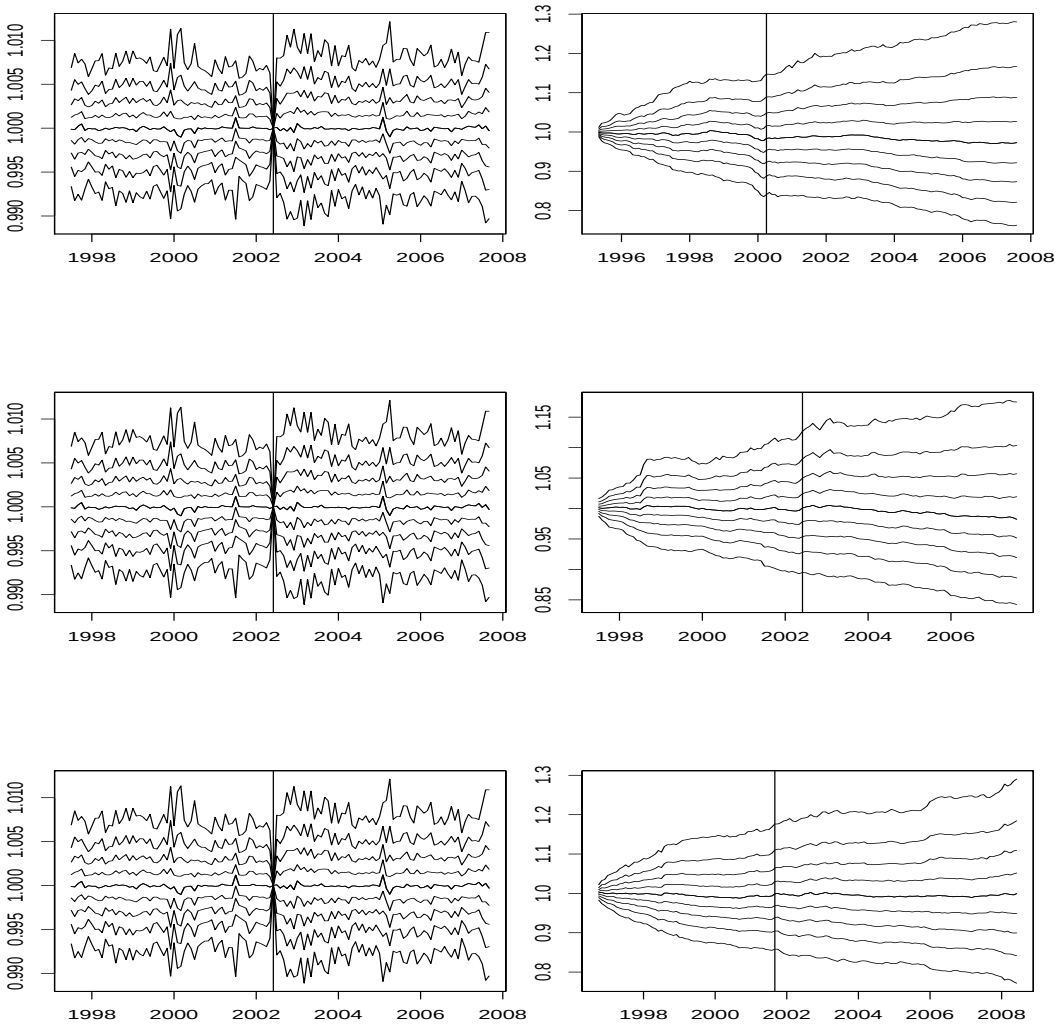

Figure 2: Deciles of the ratios from $10,000$ portfolios selected by balanced sampling over the portfolio constituted of the entire sample of funds. The first column are relative NAVs and the second are wealth achieved. From the first to the third row, we have respectively Barclay CTAs, Global Macro, and Equity Hedge funds. Each tracking portfolio is composed of $n = 10$ funds.

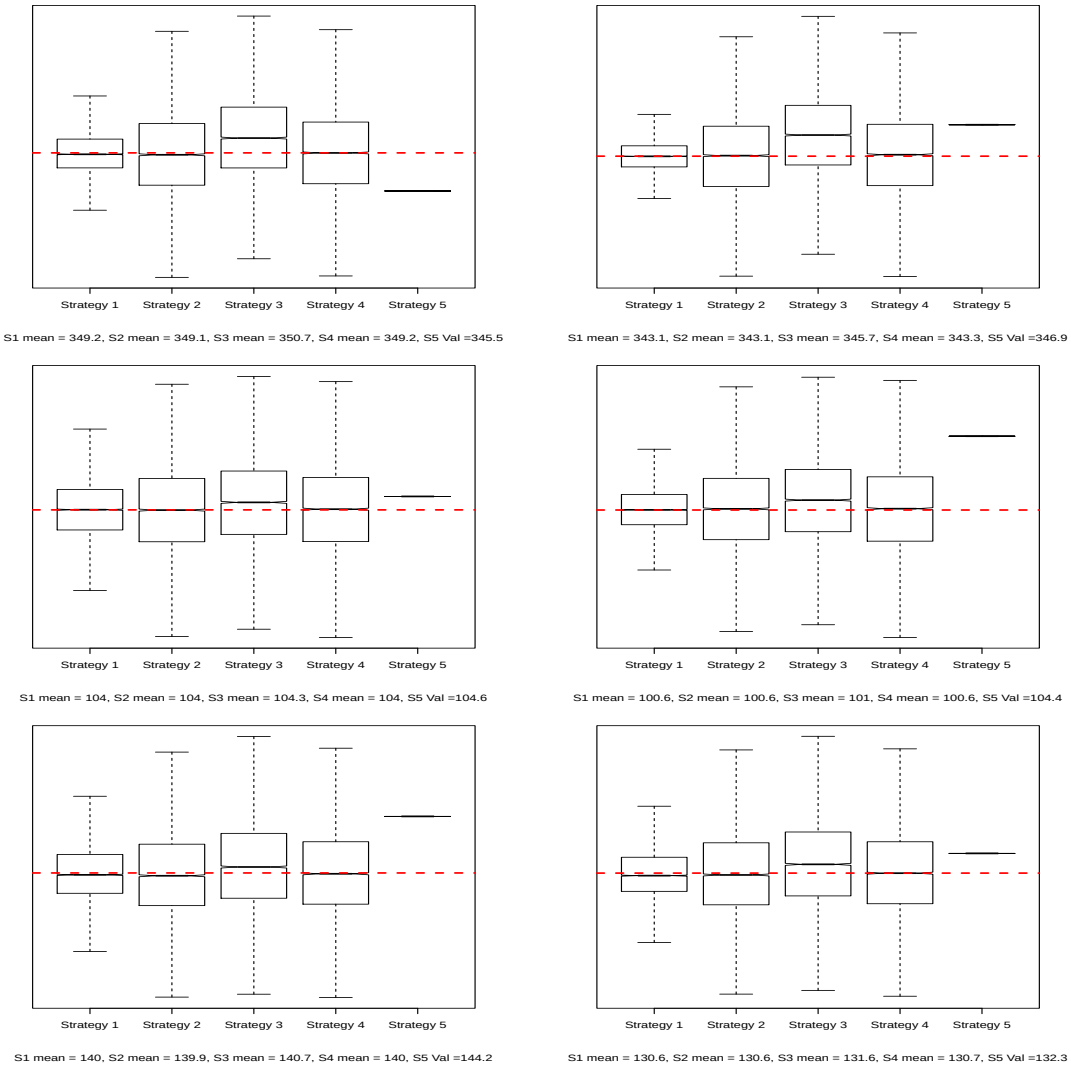

Figure 3: Boxplots of the replicated relative NAVs for the $10,000$ realizations of each tracking portfolio for strategies 1 to 4 and the realized value of strategy 5. The first-column entries highlight the values at a given trading day of the estimation period. The second-column entries are the numbers occurring in a random day of the investment period. The horizontal red dotted line indicates the raw value of the total relative NAVs as of the specified trading day. Mean values for each strategy appear in the subtitle below the figure. From the first to the third row, we have respectively Barclay CTAs, Global Macro, and Equity Hedge funds. Each tracking portfolio is composed of $n = 10$ funds.

that the portfolios are unbiased. Furthermore, by increasing the number of replication, I decrease the variance while increasing the power of the balanced tracking strategy. This important characteristic is further illustrated in the sequel from Figure 3.

4.3 Comparison with other methods of selection

We now turn to the analysis of the performance of our tracking strategy relative to several alternatives. This control phase is performed against three different random selection strategies and one 'do nothing' strategy consisting of a portfolio with the n biggest funds according to their AUM. Each strategy will provide some insights of the relative value of tracking strategy trough randomization over a simple asset based heuristic strategy. In this section, we compare the following four strategies:

- Strategy 1: The proposed method of balanced sampling described in this paper.

- Strategy 2: A method of unequal probability sampling with fixed sample size. The selection probabilities are the same as for Strategy 1. A large set of methods of unequal probability sampling are described for instance in [11]. Most of these methods give equivalent results in term of variance. We have chosen the random systematic method.

- Strategy 3: Simple random sampling without replacement where the same amount is allocated to each selected fund. This strategy is biased under the process of selecting the sampling design because it gives too much importance to poor performing funds (small relative NAVs).

- Strategy 4: Simple random sampling without replacement where an amount proportional to the relative NAVs is allocated to each selected HF. This strategy is unbiased under the sampling design but can provide a portfolio with very unequal weights.

- Strategy 5: This strategy would involve the construction of a tracking portfolio consisting of the $n = 10$ funds with the biggest AUM ('do

nothing' strategy). The funds will be included in the tracker with weights corresponding to their size in the HF universe.

Figure 3 presents boxplots for the realizations of the $10,000$ tracking portfolios for each of the 5 strategies on a given trading day during the estimation and investment periods, along with mean values for each. It can be seen from Figure 3 that a dominance relationship exists between our balance sampling design and the four alternative tracking strategies. Our replication strategy is unbiased and much less volatile in both the estimation and investment subsamples. While random sampling strategies 2, and 4 are effectively unbiased but more volatile than the proposed balanced sampling design, strategy 3 is biased and also less efficient. The beauty of a large number of simulations is the ability to discriminate among concurrent strategies. However, it is difficult to compare our proposed methodology to the 'do nothing' portfolio. The graphs in Figure 3 show its value is different to the true total relative NAVs. In what follow, I will try to solve this issue by moving from a multivariate analysis to a univariate one. I will compare the cumulative relative mean square error of the big funds portfolio to the best tracker amongst the $10,000$ constructed portfolios.

4.4 Univariate analysis

Statistically speaking, all the selected portfolios are equivalent, in the sense they reproduce the time varying evolution of the total HF relative NAVs and they are automatically diversified between highly and poorly performing funds. There is no particular reason to prefer one over another. It is up to the investor to choose a portfolio in line with his investment appetite. Moreover, since the portfolio is constructed during first the sub-sample (60 months length), there is no guaranty that the best tracking portfolio will keep performing well in the second sub-sample of data.

For practical reason, I select one of the $10,000$ constructed portfolios to directly assess its tracking performance relative to the portfolio consisting of the n biggest funds. Since in this manuscript I consider an investor who allocates his HF portfolio of n funds to minimize the distance to the total performance of funds in the universe, I select the fund with the minimum

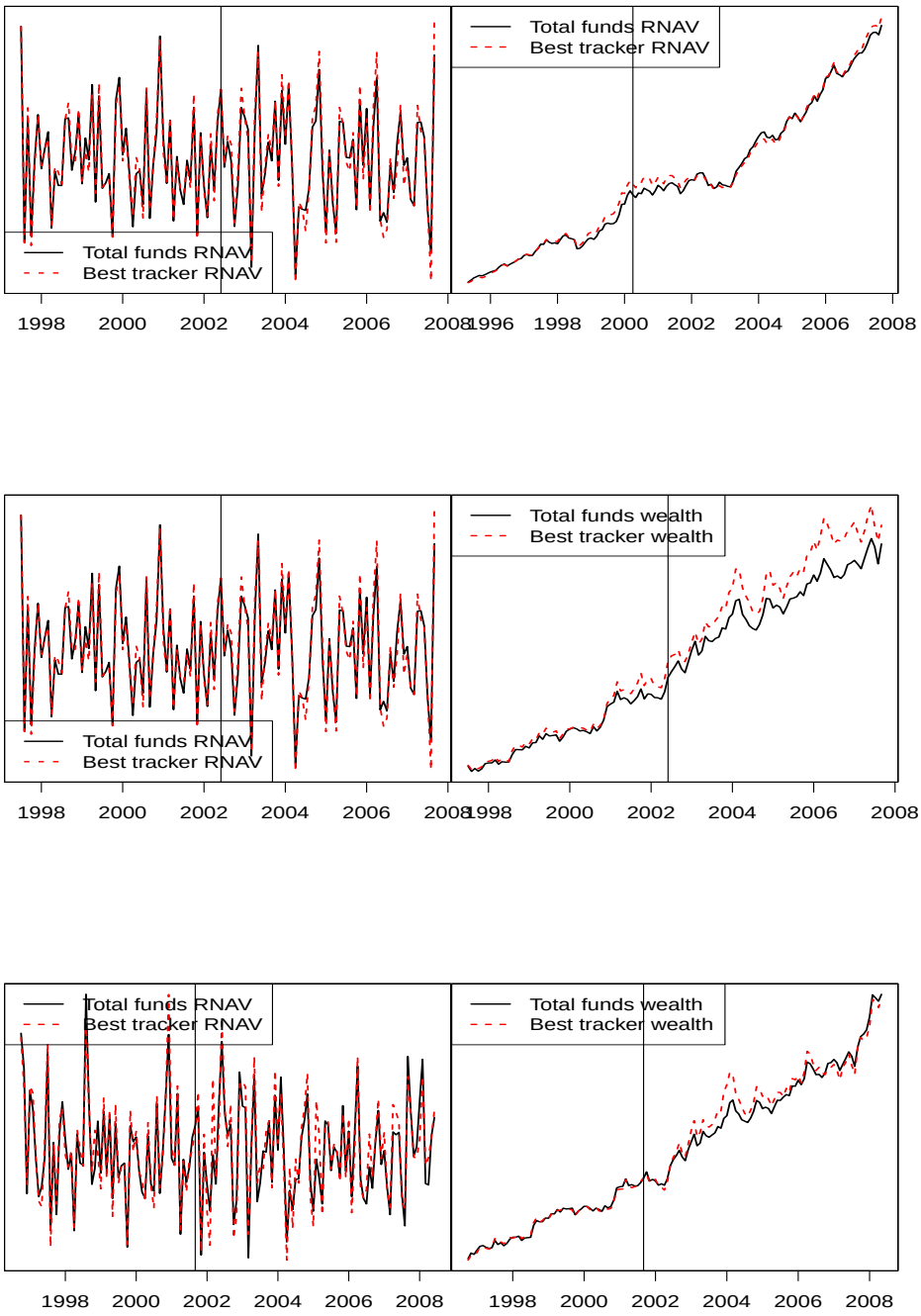

Figure 4: The evolution of relative NAVs (first-column) and wealth achieved (second-column) from the best tracking portfolio out of the 10,000 selections, as derived by the balanced sampling model. From the first to the third row, we have respectively Barclay CTAs, Global Macro, and Equity Hedge funds. The tracker is composed of $n = 10$ funds.

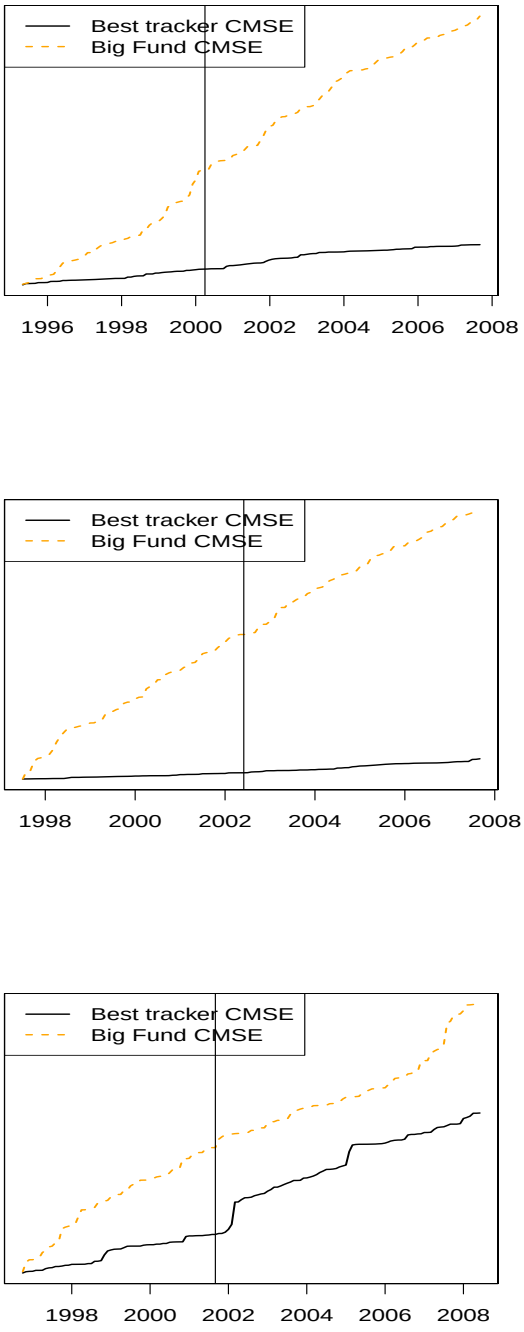

Figure 5: Cumulative value of the relative mean square errors. Portfolio consisting of the 10 funds with the largest AUM (in orange), balanced sampling (in black). From the up to down, we have respectively Barclay CTAs, Global Macro, and Equity Hedge funds. Each tracking portfolio is composed of $n = 10$ funds.

Relative Mean Square Error (RMSE) during the estimation period (first subsample). According to Equation 4, the RMSE for strategy 1 is defined as

$$\text{RMSE}(V_t) = \text{E}\left(\frac{r_\tau \frac{V_t}{V_\tau} - r_t}{r_t}\right)^2 \tag{9}$$

while the corresponding measure for strategy 5 is expressed as

$$\text{RMSE}(V_t) = \left(\frac{r_\tau \frac{r_t}{r_\tau} - r_t}{r_t}\right)^2 \tag{10}$$

Figure 4 shows the evolution of the total relative NAVs and wealth achieved of the minimum variance portfolio along the corresponding values for all the funds in the HF universe.

Several comments are of interest. First for the 3 HF categories under investigation, the best tracker is closely tied to the total market values. Second, it is of great interest to notice that the best in-sample selected portfolio performs very well in the out-of-sample analysis. Third the performance quality of the best tracking strategy is stable in the long term perspective. Unreported results of two-sample Wilcoxon test and paired samples t-test strongly reject the hypothesis that the two series are different at any confidence level. The second-column entries of Figure 4 show that our balanced sampling tracker and the portfolio consisting of all the funds in the market achieved approximately identical wealth.

Finally, I illustrate the tracking quality of the minimum variance portfolio relative to the big funds portfolio. Recording, the big funds portfolio is AUM-weighted portfolio. Figure 5 displays the time plots of the cumulative relative mean square errors for the two strategies as expressed in Equations 9 and 10. Not surprisingly, the balanced sampling strategy strongly over-performs the 'do nothing' strategy. As it appears, the gain of adopting the proposed tracking design is both statistically and economically significant, as it guaranties a diversify and unbiased portfolio, achieving approximately the true dynamic changes of the total relative NAVs in the HF universe.

4.5 Robustness analysis

To further evaluate the robustness of the balanced tracking strategy, I perform a last empirical analysis. The experiment is based on the same data. I

mixed the 3 samples of HF to have a consistent alternative investment universe. The final data set results in 343 funds covering the period from July, 1997 to September, 2007, which includes the dotcom bubble burst and the summer 2007 subprime crisis. I keep the same empirical framework using the first 60 months of data to construct the $10,000$ tracking portfolios and left the allocation constant for the rest of the period. The portfolios are not rebalanced, mainly because I am interested to the long term behavior of the selected portfolios.

The indexing portfolios contains $n = 35$ funds, which represents approximately 10 percent of the total number of funds in the benchmark. The graphs in Figure 6 show that the experiment with the new large benchmark performs well despite the low frequency associated to HF data. In particular, we notice from the second-row entry in Figure 6 that variability amongst the $10,000$ constructed portfolios is reduced. The deciles of the ratio defined in Equation 8 range between 0.99 and 1.01. Since the proposed tracking procedure is unbiased, a smaller variance means that the estimate tracker is closer to the true value of the benchmark. In this large sample context, the performance of the balanced sampling strategy improves. The superiority of our proposal is further highlight in the graphs of Figure 7, where we can see that the distribution values of the selected tracking portfolios is unbiased and less volatile in both sub-periods than the 4 alternative strategies. The evolution of the cumulative relative mean square error, as depicted in the lower graph of Figure 7 shows the dominance relationship among strategies 1 to 5. It can be seen that a stochastic dominance relationship exists again indicating that our balanced sampling tracking strategy is to be preferred on the basis of its performance.

We now turn to a comparison analysis between moments of the total relative NAVs of the balanced sampling and 'do nothing' tracking strategies and the original funds from which the tracking portfolios are derived. Table 2 presents the comparison summary statistics. The results are striking in both estimation and investment sub-periods. During the first sub-period, the average moments amongst the $10,000$ portfolios (mean, standard deviation, skewness and kurtosis) are similar to those of their benchmark. This is not the case for the big AUM funds, which figures differ significantly from the benchmark. Expect for the skewness, I have comparable numbers in the investment period.

In order to understand the relationship between the $10,000$ selected track-

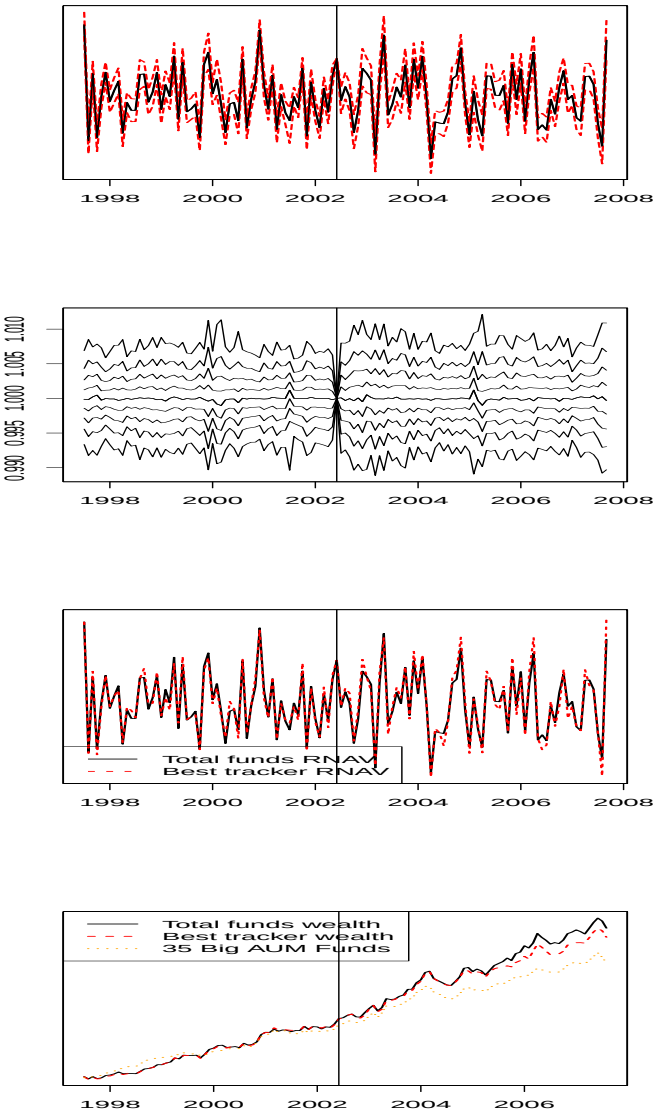

Figure 6: From upper to lower plot, the total HF relative NAVs (black line) embedded in the 5th and 95th percentiles (red line) of the corresponding values; Deciles of the ratios from $10,000$ selected portfolios; The min variance portfolio; The wealth achieved investing one dollar in the best tracking portfolio and the portfolio consisting of all the funds in HF universe. These results are derived from the large sample context, combining the three HF categories under study to obtain an universe of 343 funds. The number of funds in each tracker is $n = 35$.

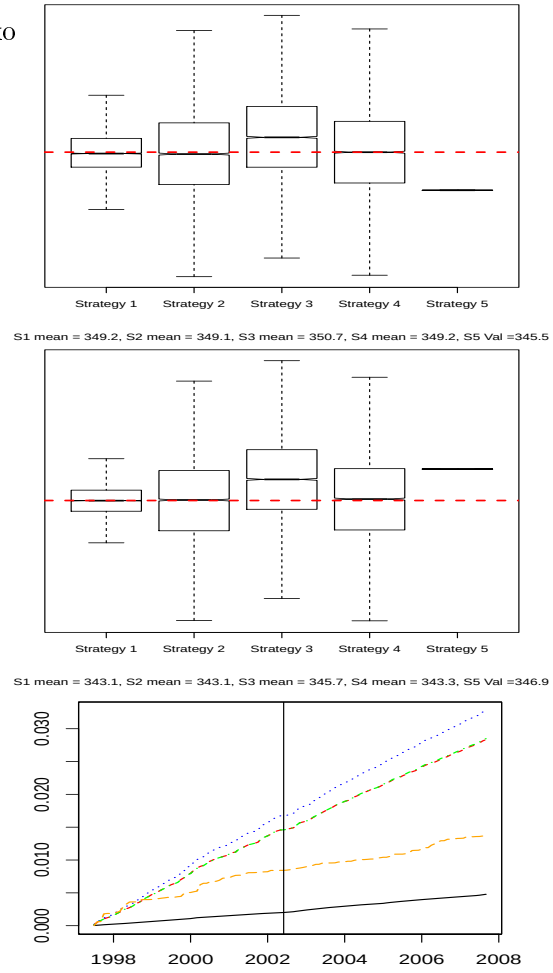

Figure 7: The first two row-entries display boxplots of the replicated relative NAVs for the $10,000$ realizations of each tracking portfolio for strategies 1 to 4 and the realized value of strategy 5. The first graph highlights the values at a given trading day of the estimation period and the second are the numbers occurring in a random day of the investment period. The horizontal red dotted line in both first plots indicates the raw value of the total relative NAVs as of the specified trading day. Mean values for each strategy appear in the subtitle below the figure. The lowest third graph displays the evolution of the cumulative value of the relative mean square errors for strategy 1 (in black), strategy 2 (in red), strategy 3 (in blue), strategy 4 (in green), strategy 5 (in orange), we have the portfolio consisting of the 35 funds with the largest AUM and the proposed balanced sampling portfolio is in black line. These results are derived from the large sample context, combining the three HF categories under study to obtain an universe of 343 funds. The number of funds in each tracker is $n = 35$.

Table 2: Moments and sensitivity analysis over the estimation and investment periods

Panel A: Allocation period

Panel A.1: Moments

	μ	σ	sk	ku
Benchmark	346.37	6.93	0.42	-0.06
$10,000$ trackers	346.42	7.15	0.39	-0.06
Best tracker	346.38	7.03	0.39	-0.19
Big AUM funds	348.27	5.58	0.19	0.13

Panel A.2: Beta Sensitivity

	$\hat{\beta}$	$t.test$	$Adj.Rsq$
	0.99	1535.42	0.99

Panel B: Investment period

Panel B.1: Moments

Benchmark	346.01	7.31	-0.01	-0.47
$10,000$ trackers	346.04	7.57	0.00	-0.36
Best tracker	346.20	8.36	0.08	-0.41
Big AUM funds	347.412	5.84	-0.081	-0.219

Panel B.2: Beta Sensitivity

	$\hat{\beta}$	$t.test$	$Adj.Rsq$
	0.98	1399.14	0.99

This table reports summary statistics on the total relative NAVs for balanced sampling $10,000$ trackers, the best tracker (in term of relative mean square error), the big AUM, and the benchmark (consisting of the entire sample of funds). The statistics corresponding to the $10,000$ selected portfolios are sample averages. Panels B.1 and B.2 report average statistics for linearly regressing each of the $10,000$ tracking portfolios to the benchmark.

ing portfolios according to the balanced sampling strategy and the target benchmark, a simple ordinary least squared regression analysis is implemented. I regress the total relative NAVs of each selected fund to their benchmark counterpart and the average betas estimated, t-statistics and adjusted-R^2 are reported in Table 2. If a tracker efficiently replicates the benchmark, the regression slope should be equal approximately to 1. According to the table, the average betas and adjusted-R^2 are close to 1.

4.6 The economic value of balanced sampling strategy: Diversification

The randomness of the selection frame gives equal opportunity to individual HF to be part of the tracker. All constituents have a strictly positive probability of being selected, ensuring that the population of interest is totally covered. Additionally, the balanced sampling design yields representative sample, since it allows estimating exactly the population (HF) total relative NAVs, that is, without bias. The key diversification property is then automatically incorporated in the efficient balanced sampling strategy.

To highlight the diversification characteristic of the balanced sampling strategy, I decomposed the $10,000$ trackers according to which HF category each fund belongs and their size (AUM). While the upper plot of Figure 8 depicts the proportion of funds relative to style, the lower plot shows the tracking portfolio distribution based on fund sizes[2] (Small Funds: AUM ≤ 10.5, Middle Funds: AUM $\in (10.5, 154)$, Big Funds: AUM ≥ 154).

One can clearly see that the randomness of the selection scheme guaranties a stable portfolio constitution across the $10,000$ trackers. The population of individual HF is well covered from Small to Big funds and amongst styles. On average[3], balanced sampling trackers are formed of 25.78%, 49.53%, and 24.68% for the Small, Middle, and Big funds, respectively; Additionally,

[2]Fund classifications by size are performed according to the first and third quartile of the population AUM. The figures are in millions US$ and correspond to 10.5, and 154 for the first and third quartile, respectively.

[3]The decomposition of the minimum variance portfolio is of equal magnitude to the average: 20%, 48.57%, and 31.42% for the Small, Middle, and Big funds, respectively; And, 40.43%, 29.35%, and 30.21% for CTAs, Equity Hedge, and Macro funds, respectively.

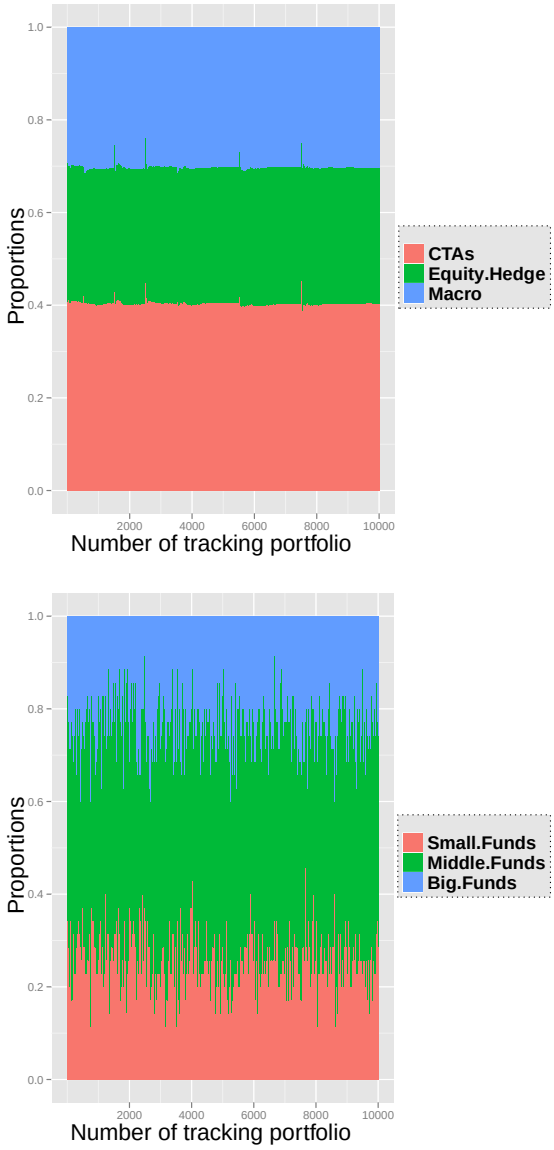

Figure 8: The upper plot shows the decomposition of each tracking portfolio relative to various HF category. The lower plot highlight the tracking portfolio distribution according to fund AUM. Fund classifications by size are performed according to the first and third quartile of the population AUM in millions of US$. Small Funds: AUM $\leq$ 10.5 (first quartile), Middle Funds: AUM $\in$ (10.5, 154), Big Funds: AUM $\geq$ 154 (third quartile). These results are derived from the large sample context, combining the three HF categories under study to obtain an universe of 343 funds. The number of funds in each tracker is $n = 35$.

40.35%, 29.42%, and 30.22% for CTAs, Equity Hedge, and Macro funds, respectively.

5 Conclusion

In this paper, I investigate the HF indexing problem with cardinality constraints in the survey sampling framework. Whereas most previous work has been devoted to the variance and tracking error minimization, and cointegration based index tracking, I formulate that the issue of selecting a small sample from a large one is a natural sampling problem. I solve the issue in a balanced sampling selection process and easily derive a large number of potential tracking portfolios. For a review on HF indexing, see [2]. The present study contributes to the market indexing literature by providing several additional insights. From a computational point of view, the proposed model enables researchers to considerably reduce the computational complexity inherent to the cardinality constraint portfolio choice, both in terms of processing time and model assumptions. Despite the necessary large number of simulations to statistically assess the long term quality of the tracker, the estimation of the model remains tractable. From a theoretical perspective, the methodology implements does rely on any assumption on the data generating process and it is free of heuristic constraints. This setting yields a highly consistent, data driven tracking portfolios.

We demonstrate that the portfolios constructed under the balanced sampling design closely track the total relative NAVs of individual funds in the HF universe. The performance of the tracker is robust both in-sample and out-of-sample, and it is stable in the long run. This characteristic is fundamental for a passive investment strategy - specially in the alternative investment world - since it avoids a costly rebalancing mechanism. I consider three different HF categories and a large sample consisting of merging the 3 main styles in an unique global benchmark, and I further confirmed in this case the relevance of constructing HF tracker under a statistical balanced sampling design. This development can be successfully applied for all asset classes and various auxiliary information, from traditional investments to complex funds of HF portfolios.

Appendix: Reduction of dimensionality

In order to construct the tracking portfolio, suppose we use the HF relative NAVs from time $\tau - q$ to time τ. Let $\mathbf{r}$ denotes the matrix of relative NAVs for d funds:

$$\mathbf{r} = \begin{pmatrix} R^1_{\tau-q} & \cdots & R^i_{\tau-q} & \cdots & R^d_{\tau-q} \\ \vdots & & \vdots & & \vdots \\ R^1_{\tau} & \cdots & R^i_{\tau} & \cdots & R^d_{\tau} \end{pmatrix}$$

matrix $\mathbf{r}$ has dimension $(q+1) \times d$. The length of the estimation period $(q+1)$ is generally large to apply a balanced sampling procedure. To solve this issue, a dimensionality reduction of matrix $\mathbf{r}$ is implemented.

Suppose we have a vector of selection probabilities whose components are strictly in the interval $(0, 1)$ and whose sum is equal to $n \in \mathbb{N}$. This vector is computed in order to be proportional to the vector of relative NAVs, $R^1_\tau, \cdots, R^i_\tau, \cdots, R^d_\tau$ during the last known time τ of the estimation period. Therefore, the reduction of dimensionality can be obtained as follows.

Firstly, we construct matrix $\mathbf{C} = (c_{jt})$ of dimension $d \times (q+1)$ where

$$c_{jt} = \frac{\pi_i}{n} \sum_{i=1}^{d} R^i_t = \frac{\pi_i r_t}{n}, \text{ for } t = \tau - q, \ldots, \tau, i = 1, \ldots, d$$

The sums of the rows of matrix $\mathbf{C}$ are thus the same as those of matrix $\mathbf{r}$. The procedure continues by computing matrix $\mathbf{F} = \mathbf{r} - \mathbf{C}$, and deriving the singular value decomposition of $\mathbf{F}$, ie,

$$\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^t$$

where $\mathbf{U} = (u_{\alpha t})$ is a $(q+1)\times(q+1)$ real matrix, $\mathbf{\Sigma}$ is a $(q+1)\times d$ diagonal matrix with nonnegative real numbers on the diagonal that are ordered decreasingly, and $\mathbf{V}^t = (v_{\alpha i})$ is the transposition of $\mathbf{V}$ that is an $d \times d$ real matrix. The relative NAVs can then be decomposed as follows.

$$R^i_t = \frac{\pi_i}{n} r_t + \sum_{\alpha=1}^{\min(d,q+1)} \sigma_\alpha u_{\alpha t} v_{\alpha i}$$

Finally, a restriction of the dimension can be obtained by applying the sum only of the first k components. One can construct a matrix

$$\mathbf{r}^{(k)} = \begin{pmatrix} R^{1,(k)}_{\tau-q} & \cdots & R^{i,(k)}_{\tau-q} & \cdots & R^{d,(k)}_{\tau-q} \\ \vdots & & \vdots & & \vdots \\ R^{1,(k)}_{\tau} & \cdots & R^{i,(k)}_{\tau} & \cdots & R^{d,(k)}_{\tau} \end{pmatrix}$$

where

$$R_t^{i,(k)} = \frac{\pi_i}{n} r_t + \sum_{\alpha=1}^{k} \sigma_\alpha u_{\alpha t} v_{\alpha i}$$

Fixing $k = n$ is enough to reproduce almost perfectly the original matrix of fund's relative NAVs.

Instead of selecting a balanced portfolio on matrix $\mathbf{r}$, we can thus use a matrix that contains $\boldsymbol{\pi}$ and the first k components of matrix $\mathbf{F}$

$$\mathbf{X}^{(k)} = \begin{pmatrix} \pi_1 & v_{11} & \cdots & v_{\alpha 1} & \cdots & v_{k1} \\ \vdots & \vdots & & \vdots & & \vdots \\ \pi_i & v_{1i} & \cdots & v_{\alpha i} & \cdots & v_{ki} \\ \vdots & \vdots & & \vdots & & \vdots \\ \pi_d & v_{1d} & \cdots & v_{\alpha d} & \cdots & v_{kd} \end{pmatrix}.$$

Next define matrix $\mathbf{A}^k$ by

$$\mathbf{A}^k = \mathrm{diag}^{-1}(\boldsymbol{\pi})\, \mathbf{X}^{(k)} = \begin{pmatrix} 1 & v_{11}/\pi_1 & \cdots & v_{\alpha 1}/\pi_1 & \cdots & v_{k1}/\pi_1 \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & v_{1i}/\pi_i & \cdots & v_{\alpha i}/\pi_i & \cdots & v_{ki}/\pi_i \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & v_{1d}/\pi_d & \cdots & v_{\alpha d}/\pi_d & \cdots & v_{kd}/\pi_d \end{pmatrix}$$

where $\mathrm{diag}(\boldsymbol{\pi})$ is the diagonal matrix containing vector $\boldsymbol{\pi}$ on its diagonal.

Acknowledgement

The author is very grateful to Professor, Yves Tillé for having introduce me to survey sampling. Thank you for your interesting comments and continuous encouragement.

References

[1] C. Alexander. Optimal hedging using cointegration. *Transactions of the Royal Society Series A*, 357(1758):2039–2058, 1999.

[2] C. Alexander and A. Dimitriu. *Hedge Fund Index Tracking. in Hedge Funds: Insights in Performance Measurement, Risk Analysis, and Portfolio Allocation, G .N. Gregoriou, G. Hbner, N. Papageorgiou, and F. Rouah (ed.), pp. 165-179.* 2005.

[3] G. Chauvet and Y. Tillé. A fast algorithm of balanced sampling. *Journal of Computational Statistics*, 21(1):53–61, 2006.

[4] U. Derigs and N. H. Nickel. Meta-heuristic based decision support for portfolio optimization with a case study on tracking error minimization in passive portfolio management. *OR Spectrum*, 25(3):345–378, 2003.

[5] J.-C. Deville and Y. Tillé. Efficient balanced sampling: The cube method. *Biometrika*, 91(4):893–912, 2004.

[6] S. M. Focardi and F. J. Fabozzi. A methodology for index tracking based on time-series clustering. *Quantitative Finance*, 4(4):417–425, 2004.

[7] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[8] F.-S. Lhabitant and M. Learned. Hedge fund diversification: How much is enough? *Journal of Alternative Investments*, 5(3):23–49, 2003.

[9] H. M. Markowitz. *Mean-variance analysis in portfolio choice and capital markets.* Blackwell, Oxford, 1987.

[10] D. Tafin Djoko and Y. Tillé. Selection of balanced portfolios to track the main properties of a large market. *UNINE/ISTAT Working Papers.*, 2012.

[11] Y. Tillé. *Sampling Algorithms.* Springer, New York, 2006.