Published in Proceedings of Workshop Axess Thinktank, 2018, Geneva, Switzerland, 27th April, which should be cited to refer to this work.
DOI: 10.2139/ssrn.3045057

Alternative Risk Premia: Is the Selection Process Important?

Francesc Naya[a] and Nils S. Tuchschmid[a]*

February 2018

[a] *Haute Ecole de Gestion Fribourg*

Chemin du Musée 4 CH-1700 Fribourg, Switzerland

**corresponding author: +41 26 429 63 50; nils.tuchschmid@hefr.ch*

First Version: June 2017

Abstract

Alternative Risk Premia (ARP) are rule-based strategies. They should reward investors exposed to non-traditional systematic risk factors. Yet, allocation to ARP is not straightforward. First, there are many ARP indices proposed by different providers that claim to capture the same underlying risk premia. Second, a proposed index may not automatically mimic an existing risk premium whose performance is sustainable or persistent. Our findings confirm these suspicions. If some categories of indices show risk-return characteristics that are rather homogeneous, others are highly heterogeneous. Stated otherwise, performance is provider dependent making the choice of an index an important component of the allocation process. Differences between simulated past results and live data are then calculated. Results are indisputable. There is a significant overfitting bias. Once launched, the performance of ARP indices dropped significantly. To summarize, this research paper shows that investors should take no short cuts. When it comes to allocating capital to ARP, an extensive due diligence process is required.

Alternative Risk Premia (ARP) are better understood when compared to traditional long-only risk factors. Everyone investing in financial securities as equities or bonds is indeed aware that capital is at risk and that it is this same risk that calls for the existence of "risk premia". The idea behind ARP is the same. Investors must be rewarded for the risk they take. The difference comes from the way the exposure is achieved. ARP are non-traditional risk factors and roughly speaking correspond to any long-short strategies or styles whose risk is *a priori* not diversifiable. Examples could be FX carry, Interest Rate (IR) carry or equity size.

ARP have experienced an increase in popularity recently, among both researchers and investors. Size-wise, the ARP market still looks small compared to the hedge fund industry whose assets under management (AuM) estimates are close to $3 trillion, but new indices are continuously created and AuM are increasing.[1] According to CitiGroup, the ARP market accounted for a tiny $15 billion in 2011. By the end of 2015 it had risen to $241 billion.[2] Yet, there is nothing new in the idea that returns of risk assets can be expressed with a set of limited factors, whatever the name we want to give them. Single-factor or multi-factor return generating processes can be found in articles published as early as the sixties and seventies. One can cite, for example, the paper of Sharpe [1964] and the use of a diagonal model or the seminal article of Ross [1976] and the contribution of his Arbitrage Pricing Theory to modern finance. However, when it comes to multi-factor pricing models, one often points to the article of Fama and French [1993] in which size and book-to-market were added to a "market" factor to explain returns on equities. Since then, other factors have been proposed. The work of Jegadeesh and Titman [1993] led Cahart [1997], for instance, to add "momentum" to the list and to show that this fourth factor helps make the persistence of alpha disappear. More recently, Pastor and Stambaugh [2003] suggested a "liquidity" factor. However, with the availability of databases and the development of computing power, empirical papers have flourished and it seems that new explanatory factors

are “discovered” endlessly. Hence, a question remains. How many factors should be considered? The true number of relevant factors in the return generating process of risk assets is indeed unknown and it is still the task of any investor attracted by the idea of risk premia and of their replicating factors to decide which ones are truly relevant.

With the recent addition of ARP, the problem is magnified and becomes more critical. How many of these products really meet the criteria of “risk premia”? And if some were to fail the test, would it imply that one should disqualify them or would it mean that there are “market anomalies” that can be both sustainable and profitable? Stated otherwise, among the dozens of risk premia categories and the hundreds or even thousands of products available, which ones are true risk premia, which ones are anomalies that are likely to persist over time and which ones just result from backtesting biases whose attractive risk-return profiles are likely to disappear as soon as they are adopted? Investors must answer not only these questions, but also whether the choice of an ARP provider is part of their selection process and whether they can trust the numerous backtest results they will face.

Our objectives in this article are thus threefold. We first propose to rigorously define what should be called ARP. We present a list of characteristics that any alternative risk premium should possess. Then, we look specifically at the current status of the industry. Based on the classification used by the different providers of ARP, we analyse the performance of the products and check their degree of homogeneity. Intuitively, within a given category, the performance of ARP should not be too heavily provider dependent. The choice of a specific provider should not alter the end-result too much. Classic correlation and drawdown analyses as well as clustering techniques show rather the opposite. In the majority of cases, products within the same category do not behave in the same way. The latter indicates that a thorough due diligence process is required not only to decide which of the ARP to select but also to which of the different providers money should be directed. Finally, we quantify how important the overfitting bias can be. Indeed, the performance differential among providers is obviously not

the only problem investors need to be aware of. The ARP industry is still young. Therefore, live data are limited and backtests still dominate. Our results clearly indicate that the bias is severe. As a rule of thumb, applying an 80% discount on any performance measure shown in a backtest seems to be a conservative approach that investors must follow if they want to avoid being disappointed once they decide to allocate capital.

The number of research papers focusing on ARP has recently grown. Broadly speaking, these papers fall into four categories. First, one can find articles focusing on ARP as a concept or investment opportunities. In that particular case, the question is thus to know if and how ARP exist, that is if the concept of ARP is itself relevant. For instance, Carhart et al. [2014], Hamdan et al. [2016], Blin et al. [2017], Israel and Maloney [2014], Lempérière et al. [2017] and Roncalli [2017] discuss ARP, present a list of possible candidates and their source of existence. Second, there are articles that tend to specifically focus on a limited set of ARP and analyse in detail their behaviour and risk-return characteristics. For example, Piotroski [2000], and Asness and Frazzini [2013] look at value premium; Correia et al. [2012] at value and value strategies in credit markets; Carhart [1997], and Daniel and Moskowitz [2016] at momentum; Asness et al. [2013] at both momentum and value; Asness et al. [2014] at quality premium; or Frazzini and Pedersen [2014] at low beta. The third group of research papers focuses on capital allocation and portfolio construction made up of ARP. One can cite for instance the article of Roncalli and Weisang [2016] that discusses factor investing and risk parity. One can think about the article of Bruder et al. [2016] who introduce skewness risk to risk parity solution, or that of Brandt and Santa-Clara [2009] who compute portfolio weights using the assets' characteristics. Finally, one can find articles that use the idea of ARP and factor investing to replicate hedge fund returns, such as Fung and Hsieh [2004], Maeso and Martellini [2017], or Tancar and Viebig [2008]. The list is obviously not exhaustive. Yet, to the best of our knowledge, papers that focus on due diligence and the selection process, that is papers that question the homogeneity of ARP among the different providers or that quantify the degree of data mining bias, remain extremely limited.

One could cite here the recently published article of Suhonen et al. [2017] that analyses the persistence of ARP performance after the live date. Our paper belongs to this field of research.

Defining Alternative Risk Premia

The concept of alternative risk premium stands for rule-based trading strategies that offer a premium to investors for carrying a specific type of systematic risk. Hence, as it should be for any risk premium, ARP are expected to yield positive returns. Moreover, ARP hold both long and short positions since their intention is to extract, or more precisely to isolate, a particular source of risk for which investors must be rewarded. Size is a good example. Buying small caps while shorting big companies should first eliminate exposure to market risk and second extract the "size" risk premium that is often justified on the basis of informational disadvantage or liquidity differential. As ARP are by construction both long and short, it thus implies that their correlation to markets and to equity markets in particular is low. The latter is the second component that characterizes any ARP. In that regard, one must also note that ARP are quite often spread related strategies. It is clearly the case when the intention is to profit from interest rate differential along the curve, from interest rate differentials among currencies, from calendar spreads within commodities or from performance differential between winners and losers within single equities. Names can vary; they will be called "carry strategies" in the first cases and named "momentum strategies" in the last example.

Positive expected returns and low correlation are still not enough for the industry to consider a strategy to belong to the ARP family. In general, these strategies must have the following nine characteristics:

1) Positive expected returns: Even though any strategy can experience periods of underperformance, investors must be paid for the risk they take. Thus, a strategy must yield positive (excess) returns to satisfy the risk premium criterion.

2) Low correlation with benchmark: The correlation of an alternative risk premium with traditional markets is low by construction. In some cases it can be zero or even negative. In addition, the correlations among ARP should be low as well. They are thus perceived as good instruments to build a well-diversified portfolio.

3) Transparency: ARP are rule-based strategies. As such, their construction process should be transparent and known to investors.[3] Their returns cannot be called "*alpha*". They do not stem from some specific managers' skills.

4) Persistency: Even if risk premia can be time-varying, one cannot expect them simply to disappear. Hence, ARP should be observable within different market conditions, within different regions or even different asset classes if the investment strategy is not asset-specific.

5) Identification: The reason why these premiums exist must be known or clearly identified. They should be justifiable, that is based on economic and rational criteria and supported by empirical research.

6) Low fees: Built upon mechanical and replicable rules, ARP are expected to have lower fees than hedge funds.

7) Rule-based: The rules used to create a strategy that isolates a specific risk factor or risk premium are known *(*the *transparency* argument), constant and replicable.

8) High liquidity: In most cases these products trade on a daily basis, thus providing investors with high liquidity and flexibility to rebalance their portfolios.

9) Long/short: Contrary to smart beta indices, which are long-only indices and vulnerable to the direction of the market, alternative risk premia indices are long/short strategies, neutralizing their market exposure with each of them capturing a particular factor exposure, e.g. FX carry or volatility carry.

Exhibit 1

Returns' Attribution Characteristics				
	Traditional Beta	Smart Beta	ARP	Alpha
Positive expected returns	✓	✓	✓	✓
Low correlation with benchmark			✓	✓
Transparency	✓	✓	✓	
Persistency	✓	✓	✓	
Identification	✓	✓	✓	
Low fees	✓	✓	✓	
Systematic rules	(✓)	✓	✓	
High liquidity	✓	✓	✓	(✓)
Long/Short			✓	✓

Characteristics and differences of traditional beta, smart beta, alternative risk premia, and alpha strategies.

Some of the ARP characteristics are shared with traditional passive strategies. Others are features of active strategies, and most of them are common with the so-called "smart beta" products. Indeed, smart beta and ARP are closely related as both concepts have emerged with the principle of factor investing and have grown significantly in recent years. They are neither pure passive investment strategies nor pure active strategies. Instead, smart beta and ARP fall somewhat in between. They are intended to outperform a benchmark or to "beat the market", but they do so by using systematic and replicable rules which are known to everyone, as opposed to "alpha generators". So, what does differentiate ARPs from smart beta indices? First, the universe for smart beta indices is mostly composed of equities and bonds, while the universe of ARP is extended to currencies and commodities.[4] Second and more importantly, a smart beta index attempts to beat its benchmark by overweighting and underweighting the assets to gain exposures to some specific factors. However, it remains as close as possible to its benchmark to which it will be correlated with the intention to maximize the information ratio. On the contrary, ARP strategies are not correlated to benchmarks. They attempt to isolate the premium that a specific risk factor offers from all the other risk factors, in particular market risk. To achieve market neutrality, ARP are thus long/short strategies.

Positive expected returns and low correlation are thus the two main attributes that should characterize any alternative risk premium. If one sticks with these two principles, some of the

indices that are available to investors will be automatically rejected. The “defensive” or “hedge” indices for example which end up having long volatility one way or another are indeed not meeting the positive expected return requirement, even if their correlation can be low under normal market conditions. The relevance of other indices in terms of risk premium can also be questioned since a long history of positive returns can also be explained by market imperfections or investors' behavioural biases. This approach is for instance advocated by Hamdan et al. [2016] or Roncalli [2017] who distinguish between what could be “pure” ARP such as carry long-short strategies and others that could stem from “market anomalies”. Yet, there should always be a rationale behind the existence and the persistence over time of any alternative risk premium. For that, they can rely on three main explanations. First, the theory of asset pricing tells us that in arbitrage-free and efficient markets, risk assets yield positive expected returns in excess of the risk-free rate because they expose investors to a series of systematic or non-diversifiable risks for which they want to be compensated. Applied to ARP, the premia exist as rewards to systematic risk and are commonly attributed to skewness risk. The second explanation is linked to the idea that markets are not efficient. The existence of structural constraints for example give rise to investment opportunities that ARP can exploit. The third explanation relies on behavioural biases such as the investors' tendency to be overconfident and to under-react to the latest market information which in turn can create pockets of opportunities that some ARP can profit from.[5]

The ARP Industry

There are at least three dimensions to classify ARP and their respective indices: asset class (equities, rates, credit, currencies and commodities), premium or “style” (carry, value, momentum, trend, quality, etc.),[6] and the region or “universe” (Global, US, Europe, Asia, emerging markets, G10 countries, etc.). With regard to commodities, the universe refers either

to single commodity indices, i.e. gold, cooper, WTI or Brent or to "baskets" of commodities or commodity indices.[7]

There is no consensus on what ARP are and what they are not. Yet, the number of indices made available within each of the categories should somehow help to assess the degree of "market acceptance". The rationale behind this is very simple: if a premium is well-accepted by the investors, there should be demand, and as a consequence the providers will be willing to offer it. Similarly, if an index is offered but there is no demand (meaning that investors do not see it as an alternative risk premium or at least that they do not allocate capital to it), it will tend to disappear. Exhibit 2 shows the number of providers (from a sample of seven major investment banks) offering each of the ARP indices.

Exhibit 2

Returns' Attribution Characteristics					
	Equities	**Rates**	**Credit**	**Commodities**	**Currencies**
Carry	6	7	2	7	7
Value	6	4	1	3	6
Momentum	6	5	2	5	3
Trend	3	4	2	4	6
Volatility Carry	7	7	3	6	6
Mean Reversion	4				2
Size	7				
Quality	7				
Low Vol/Beta	7				
Profitability	1				
Merger Arbitrage	1				

Number of providers offering an index by *style* and *asset class*, jointly forming each of the ARP. Carry, value, momentum, trend and volatility carry are found in all asset classes. Mean reversion in equities and currencies, size, quality, low vol/low beta, profitability and merger arbitrage are equity-specific.

It is clear which candidates have received a higher market acceptance or more attention, and which ones have not. The most widely offered ARP are *carry* and *volatility carry* first, followed by *value*, *momentum* and *trend* for all asset classes excluding *credit*. We then find *size*, *quality*, *low volatility / low beta*, and, to a lesser extent, *mean reversion*, as equity factors. Contrarily, *profitability* and surprisingly *merger arbitrage* have received a low market acceptance. For the

latter, it might stem from the relative complexity of its building process which could make providers reluctant to offer it.

Note that not all ARP indices are included in our sample. Some indices can be offered by other providers than the seven our sample is made of. One could cite for instance *growth*, *investment*, *skew* or *correlation* factors for equity markets*, liquidity* factor for all asset classes or *mean reversion* for rates and commodities.[8] Yet, their absence in Exhibit 2 tends to indicate that their market acceptance is still low for the time being.

Data

We have extracted from Bloomberg the daily prices of 233 ARP indices being currently offered by seven major investment banks. Each of these indices should capture an alternative risk premium, that is, it should be exposed to a well-defined risk factor for one asset class. From an initial bigger sample, we thus have excluded all cross-asset indices and "cross-style" indices (for instance, an index that combines equity value, equity size and equity momentum); indices that aim at maximizing returns (seeking alpha generation) rather than extract an alternative risk premium; indices that are long volatility; "duplicates" (i.e. the same index but expressed in different currencies) and the indices that did not have data available at least for the period from June 2010 to April 2017. From the 233 indices left, 213 are of "excess return" type of which 173 are denominated in USD, 39 in EUR and one in JPY. To ensure coherence in the analysis, we thus restrict the sample to the 173 excess return USD denominated indices. For each asset class, we then selected a few long-only benchmarks: the MSCI ACWI and S&P 500 for global and US equity indices respectively, Bloomberg Barclays US Aggregate Bond Index for IR and bond markets, and the S&P GSCI for the commodity market.

Descriptive statistics are shown in the Appendix in Exhibit 10 for equity premia and Exhibit 11 for rates, credit, commodities and FX premia. The results are estimated with historical data covering the period from 17 June 2010 to 5 April 2017. For each ARP category, Exhibits 10 and

11 give the number of indices available and calculate annualized average returns, annualized volatility, skewness, and kurtosis coefficients. We do so by creating equally-weighted (EW) indices made up of the individual indices provided by the different investment banks. With the exception of equity carry, value, trend, size, profitability and FX momentum, all EW indices have performed positively during the sample period. It might be worthwhile to mention that for all asset classes except FX, the best performance is to be found in volatility carry strategies. In Forex, it is the mean reversion strategy that performs best. It is also worthwhile noting that volatility carry is the strategy that exhibits the highest negative skewness coefficient.

Highest and lowest mean returns of ARP indices within each of the categories as well as the highest and lowest historical volatilities can also be seen in Exhibits 10 and 11. These results already give a flavour of how heterogeneous some of the categories can be. Indeed, statistics of EW indices might not provide very relevant information if the components themselves show different risk-return profiles. For instance, the EW equity size index shows negative annualized mean returns of -0.34%. Yet, its range varies between -4.33% and 2.01%. Stated otherwise, unless one were o naively pick everything available, the choice of a provider would clearly have had an impact.

The Importance of the Selection Process

Creating categories and populating them with different indices is one thing. Making sure they are indeed all the same or at least present very similar characteristics is something else. In this section, we intend to analyse how close the indices of a specific category really are. By definition, a premium is well-defined, rules-based and easily implementable. The indices that *a priori* capture the same risk factor must behave the same. If we observe significant differences among the different providers, a key question then naturally comes up: are we really getting exposure to an alternative risk factor?

To answer these questions, two different types of tests are carried out: first we use standard correlations and drawdown analyses. Second, we apply a clustering technique. Each approach uses the same set of data but it tackles the problem differently: for the correlations and drawdown analyses, groups are first created. We select the indices from the different providers of a specific premium (same asset class, same style, and if possible same region/universe) and study their behaviour. Doing so, we take the categories and the indices that compose them as they are, that is as granted. Conversely, the clustering technique is agnostic. It does not start from an existing classification.

Correlations and Drawdown Analyses

We first sort the indices according to the categories used by the providers included in our sample. Then, we compute and analyse cross-correlations and maximum drawdowns within each of the categories. We provide results for a subset of them: *FX carry, commodity trend, equity size, equity trend, equity value*, and *equity low volatility*.[9] *A priori,* the providers of the same alternative risk premium should capture the same underlying risk factor and that for the same asset class and the same region. Therefore, one should expect the correlation to be high and positive and one should also expect these indices to drawdown at the same time. If this is not the case, it can mean either that the factor is not well-defined, or that each provider applies its own "systematic rules" that differ enough from the others. We present the groups formed and the results of the correlations summarized for these six categories along with the usual set of descriptive statistics as historical mean returns, historical volatility, return-volatility ratio, skewness and kurtosis coefficients in Exhibits 3 and 4. Detailed cross-correlations and drawdown results are shown in Exhibit 12 and Exhibit 13 in the Appendix. When applicable, we include statistics of traditional benchmarks as well.[10]

FX carry is made of five indices that take G10 currencies as their reference universe. The data is available for all indices from 5 January 2006. The strategy intends to capture the yield spread between high yielding currencies vs. low yielding currencies and naturally bears the risk of

sudden currency devaluations/depreciations or reversal (skewness risk). It can be constructed by ranking the G10 currencies by their interest rate levels and then to go long the three highest yielding currencies, equally weighted, and short the three lowest yielding currencies, again equally weighted. As in all the following groups, the rebalancing period can differ among providers.

Commodities trend group is composed of four indices whose universe is described as "Global", which means that they are made of a basket of different commodities and refer in the majority of cases to the S&P GSCI components. The index is taken as the benchmark. The starting period is 3 January 2007. The strategy aims at capturing trends in the commodity market. One example of construction is to rank the commodities by their 12-month cumulative past returns and go long when cumulative performance is positive and short when it is negative.

For **equity size**, we compare three indices targeting US equities and the S&P 500 index is taken as the benchmark. The start date is 5 January 2006. Size is a well-established and extensively studied strategy. It is based on the principle that "*small caps outperform big caps*" over long enough investment horizons. Typically, one ranks equities by market capitalization and takes respectively long and short positions picking names at the bottom and the top of the list.

The **equity trend** group is made up of four indices, all targeting a Global universe. Start date is 4 January 2007 and here the MSCI ACWI index is used as a benchmark.[11] A combination of short and long term moving averages could serve as an example to illustrate an equity trend construction process. According to this simple trading rule, a long (short) position is initiated when the short-term moving average is higher (lower) than the long-term moving average.

The **equity value** premium contains three different indices targeting US equities. Start date is 4 January 2006 and the S&P 500 index serves as benchmark. Typically, the strategy leads to ranking the set of stocks by value metrics such as the book-to-market price ratio, earnings-to-price or dividend-to-price, among others. Then, it takes long positions in "undervalued" stocks and short positions in "overvalued" stocks.

Finally, the **low volatility** group contains three indices targeting global equities and with data available since 5 January 2006. The strategy ranks the stocks by their historical volatility and takes long and short positions in low and high past volatile stocks respectively. The rolling window length used to estimate past volatilities is at the providers' choice.[12]

First, one notices that all indices show very low or even negative correlation coefficients with their respective benchmarks. The latter was expected. The results confirm that ARP indices are not simply exposing investors to traditional risk factors. Focusing then on results shown in Exhibit 3, it can be seen that indices within each of the six categories present different risk-return characteristics. The latter observation is even striking if one restricts the analysis to the mean returns and to the Sharpe ratios. For equity size for example, an apparently well-defined risk premium, differentials amount to 2.11% and 0.47 points respectively with minimum mean return and Sharpe ratio equal to 1.46% and 0.22 whereas the maximums are 3.57% and 0.69.

The average cross-correlations for each of the six groups, as well as their minimum and maximum values, are given in Exhibit 4. Here again, the degree of heterogeneity varies notably among ARP. On one side, one has the FX carry and equity trend premia whose average cross-correlations are slightly higher than 0.75, with lowest coefficients equal to 0.58 and 0.64 respectively. On the other side, the indices for the commodity trend, equity size and equity value premia exhibit lower cross-correlation coefficients, averaging 0.65. 0.51 and 0.46 respectively. Finally, at the other extreme of the spectrum, there is the equity low volatility premia with extremely low average correlation of 0.18. For the latter group, the minimum cross-correlation is practically zero.

Results in Exhibit 13 given in the Appendix show that these indices might experience their drawdown at the same time but that the amplitude of their losses can vary significantly. For the commodity trend premium for instance the indices can be separated into two groups. The first one experiences its drawdown in 2007, the other one in 2013. The same observation can be made for equity value, but for one of the groups it also coincides with the drawdown period of

the benchmark. Note finally that the duration of the drawdown period of the equity low volatility premium is clearly not the same, even though the start or "peak" date coincides.

Overall, the results suggest that some premia, such as FX carry or equity trend, show some degree of homogeneity whereas for others, such as equity size, equity value and equity low volatility, results are highly provider dependent. In that case, it is hard enough to find commonalities among the indices to bypass the crucial problem linked to the selection of a provider.

Obviously, the degree of homogeneity is not constant over time. Simple rolling window analysis reveals that indices naturally tend to correlate more when their correlations with their benchmark increases. The latter is particularly apparent for the commodity trend indices late in 2009 or in 2016. Yet, the situation is not always so clear-cut. Here again, Equity Size and Equity Value indices do not show any clear pattern in the way their cross-correlations and correlations with their benchmarks vary over time.

Exhibit 3

Descriptive Statistics: Subset of Strategies

		FX Carry	Commodity Trend	Equity Size	Equity Trend	Equity Value	Equity Low Vol.
Mean Return	Average	-0.66%	3.74%	2.57%	-0.12%	-0.30%	1.32%
	Min	-1.32%	0.03%	1.46%	-2.61%	-0.81%	-0.52%
	Max	-0.01%	9.07%	3.57%	1.04%	0.62%	2.87%
Volatility	Average	8.65%	8.24%	7.07%	5.77%	6.59%	4.37%
	Min	5.27%	1.87%	5.21%	3.10%	6.08%	2.91%
	Max	11.86%	12.41%	9.41%	9.35%	6.83%	5.29%
Sharpe	Average	-0.10	0.35	0.40	0.04	-0.05	0.35
	Min	-0.21	0.02	0.22	-0.28	-0.13	-0.10
	Max	0.00	0.73	0.69	0.21	0.09	0.58
Skewness	Average	-0.77	-0.23	-0.07	-0.29	0.57	0.08
	Min	-1.14	-0.46	-0.13	-0.68	0.26	-0.27
	Max	-0.27	0.04	0.05	-0.08	1.07	0.26
Kurtosis	Average	9.49	9.16	10.45	9.03	12.56	9.39
	Min	4.42	5.11	8.64	7.21	6.22	7.57
	Max	14.71	12.74	12.56	12.21	22.67	10.91

Summary of the descriptive statistics for the six premia selected to carry out the correlations and drawdowns analyses. This table shows the annualized mean returns, annualized volatilities, Sharpe ratios, skewness and kurtosis coefficients. For each group and coefficient, we show the average,

minimum and maximum values. The sample periods begin on January 2006 for FX carry, equity size, value, and low volatility; and on January 2007 for commodity trend and equity trend premia; and end on April 2017 for the six groups.

Exhibit 4

Correlations Summary: Subset of Strategies

		FX Carry	Commodity Trend	Equity Size	Equity Trend	Equity Value	Equity Low Vol.
Cross-correlations	Average	0.77	0.65	0.51	0.76	0.46	0.18
	Min	0.58	0.37	0.33	0.64	0.37	0.02
	Max	0.96	0.82	0.67	0.86	0.52	0.34
Correlations with the benchmark	Average		-0.19	0.00	-0.04	0.14	-0.31
	Min		-0.32	-0.25	-0.16	0.04	-0.43
	Max		-0.10	0.24	0.06	0.25	-0.23

Summary of the indices' cross-correlation coefficients for the six selected risk premia and their respective benchmarks. The benchmark indices are the S&P GSCI for the commodity trend premium, the MSCI ACWI equities premia with a universe of global equities and the S&P 500 for the premia with a universe of US equities. The number of indices used is five for the FX carry premium (targeting G10 currencies), four for the equity trend and the commodity trend premium (targeting baskets of commodities, either the S&P GSPI components or an unspecified basket), and three for the equity size, equity value, and equity low volatility premia. The sample periods begin on January 2006 for FX carry, equity size, value, and low volatility; and on January 2007 for commodity trend and equity trend premia, and end on April 2017 for the six groups.

Cluster Analysis

The goal of clustering techniques is to classify and to group objects according to their degree of similarity, but without *a priori* making any assumptions on the characteristics of the objects themselves. In our case, we apply these techniques to check if the ARP indices tend to naturally group into the different categories that their providers have created for them or on the contrary, if they cluster in a way that does not match with the categories that they are supposed to populate.

We use the agglomerative hierarchical clustering algorithm and map the results in dendrogram plots (also called “hierarchical trees”). The mapping shows some essential information. First, it shows whether the indices that should capture the same risk factor tend to cluster together. Second, it reveals if all of the three dimensions previously mentioned above (i.e. asset class, risk factor and universe) are relevant for diversification purposes, and which one comes first. We cluster the 173 USD-denominated indices of “excess return” type and set a cutoff level of 90, for

which we obtain 12 groups summarized in Exhibit 5. Exhibit 6 focuses on the equity asset class only and we analyse whether the factor categorization of the hierarchical tree is the same as in the industry classification, to what extent, and whether the region has any impact. We set the cutoff level at 80 to get 10 groups. For the latter, we include EUR-denominated indices as well.[13]

ARP tend to be not uniformly well defined among researchers and investment professionals. In turn, this can lead to heterogeneities among indices that are still presented as belonging to the same categories. The same results appear to come out when clustering techniques are applied. As expected, some ARP are well-clustered. This is the case for many carry strategies, such as commodity carry and volatility carry, FX carry and volatility carry, or equity volatility carry. The equity trend family of indices also appears to be homogeneous. The homogeneity within the other groups is not so clear-cut. This is the case for commodity momentum and commodity trend indices, FX value, rates carry and equity momentum and size. Finally, equity quality and low volatility, as well as rates momentum and trend indices do not cluster well.

From Exhibits 15 and 16 in the Appendix we can also identify cases in which the asset class groups several factors, such as in Group 1, 2 and clearly 8 from Exhibit 5. We thus have clusters for commodities, FX and rates that leave the factor or investment style at a second stage. The contrary happens in Groups 10, 11 and 12. These groups merge together forming a big "volatility carry" cluster composed by indices from various asset classes.

Regarding the equities clusters, the well-grouped factors are trend and volatility carry. On the other side, the factors that do not cluster together are quality and low volatility. These indices all together create various groups, suggesting that the factors themselves are either too close to each other or not well defined to make each investment style sufficiently distinct from the others.

Exhibit 5

Cluster Results: All Asset Classes ARP Indices

Cluster	1	2	3	4	5	6
# Indices	17	9	5	42	7	11
Indices Grouped	Commodity momentum, carry, and trend	FX trend and momentum	4 equity and 1 credit trend indices	Equity low vol, carry, quality, size, momentum, mean reversion, FX value and mean reversion, commodity carry and volatility carry, rates value	4 equity value, 2 size and 1 profitability	5 equity momentum, 3 quality and 3 low volatility
Comments	Different providers. Zooming inside, the indices split in a group for each factor or strategies ("Commodity cluster")	Zooming inside, indices split according to region (G10 or EM) ("FX momentum/trend Cluster")	Equity indices from various providers. Credit index is the last to cluster ("Equity trend cluster")	Very diverse cluster. Indices targeting Japan or Asia cluster together, FX value forms a subgroup	Zooming inside we find two groups: 1) 3 value indices 2) 3 indices of a different group but from the same provider	All indices except one equity momentum are from the same provider

Cluster	7	8	9	10	11	12
# Indices	11	18	17	16	10	10
Indices Grouped	Commodity carry indices	Rates carry, momentum and trend	13 FX carry, 1 FX value, 2 credit carry, 1 commodity carry	6 rates vol carry, 9 equity vol carry, 1 credit volatility carry	10 FX volatility carry	10 commodity volatility carry
Comments	Various providers ("Commodity carry cluster")	All rates-based indices ("Rates cluster")	Zooming inside, FX carry form a subgroup and split further by region ("FX carry cluster")	Various providers. Zooming inside, indices split by asset class ("Volatility carry cluster")	Zooming outside, this group joins the previous one, creating a big volatility carry cluster ("FX volatility carry cluster")	Zooming outside, this group joins the previous two, making a big cluster that comprises the volatility carry indices for all the asset classes ("Commodity volatility carry")

Groups formed by clustering the 173 USD-denominated and "excess return" type ARP indices from various asset classes. The clustering technique used is hierarchical with the Euclidean distance and Ward method to form the groups. By setting the cutoff level at 90, 12 groups are formed. This table summarizes Exhibit 16.

Exhibit 6

Cluster Results: Equity ARP Indices

Cluster	1	2	3	4	5
# Indices	6	12	7	12	8
Indices Grouped	3 quality, 3 low volatility	3 momentum, 2 value, 2 size, 2 low volatility, 2 quality, 1 carry	4 low volatility, 1 carry, 2 quality	7 momentum, 3 quality, 1 low volatility, 1 profitability	6 mean reversion, 2 volatility carry
Comments	Zooming inside, factors split. All are from the same provider except 1 quality index	Except 1 momentum index, all indices are from the same provider and made of Japan or Asia universe	Low volatility and quality clustered together	Zooming inside, momentum indices form a subgroup and the remaining indices form another one	Except 1 mean reversion index, all are from the same provider

Cluster	6	7	8	9	10
# Indices	4	8	11	8	8
Indices Grouped	4 trend	8 volatility carry	7 value, 4 carry	4 size, 3 momentum, 1 value	5 size, 2 profitability, 1 value
Comments	Trend indices well grouped, diversity in providers	Except 1 index, all indices are from the same provider	Diversity in providers for value but not for carry. Value index of IB7 clusters first with carry indices (from the same provider)	Size indices form a subgroup except for one index, which clusters with different factors (momentum and value) from the same provider	Zooming inside, size indices do not cluster together

The groups are formed by clustering 84 equity-based USD and EUR denominated “excess return” type ARP indices. The clustering technique used is hierarchical with the Euclidean distance and Ward method to form the groups. By setting the cutoff level at 80, 10 groups are formed. This table summarizes Exhibit 17 in the Appendix.

Selection Biases

Biases when it comes to using historical observations can be numerous. Yet, we tend to group them under the usual expression of "*Survivorship Bias"*. In the case of hedge funds for example, "*survivorship bias"* impacts any risk-return analysis if the sample only contains funds that have survived over time. That is, the sample does not contain those funds which did wind down or which were forced to wind down following for example investors' withdrawal after a severe drawdown. As the latter tend to be also funds that performed badly, it is clear that their absence can significantly bias the results of any analysis.[14] To standard *"survivorship bias*", one should add a selection bias that matters even more in the case of the young ARP industry. Selection bias stems from the fact that databases available to researchers contain observations only from the funds that have decided to be part of them. As funds that do not perform too well are also the ones that have less incentive to show up, the samples and the results will once again be impacted. When it comes to the ARP industry, it should be clear that the providers backtest any investment strategy before launching it. The backtest's main objective is not only to show or to confirm that a "premium" exists but also to convince potential investors that such a strategy is relevant per se and useful within their existing portfolios. Stated otherwise, the index will only be presented and launched if the backtest is attractive enough. As such, biases can easily arise. Therefore, we are confronted with both a survivorship bias and a selection bias. Indeed, our sample is made up of the survivors, that is the ARP indices available yesterday and that are still available today. Samples of ARP, in that sense, suffer from these biases and performance analysis should be read keeping that in mind. However, more could be done. Assessing how important these biases can be should help investors in their decision-making process. Our objective here is to focus on the selection bias that mainly stem from data mining and from extensive testing where both time periods and investment rules are changed to maximize the likelihood that a strategy "*works*".

The "start date" (also called the "live date") of an index is the date at which the index begins to be truly traded and sold to investors. The performance of the strategy that the providers show to the investors begins much earlier than this date. The providers make a series of backtests and show their results, which they usually call "*historical results*", "*simulated past returns*" or "*backtesting*". If the live date is clearly identified, it is then possible to test whether there is any difference in performance before and after the launch of an ARP index. One could *a priori* argue that there will be differences that can be purely random. Yet, if the differences are significant and show clear patterns, it will indicate the presence of "*overfitting*": the provider can be backtesting various rules for one strategy or premium and offer only the most successful one. Similarly, it can decide to start selling the index to investors when the backtest shows the most attractive returns. Similarly, the provider might be showing results that favour the strategy the most, that is, selecting a backtest sample period that shows the best of the possible results.

Here we use a sample of 255 ARP indices provided by twelve anonymized investment banks. For each of the indices, we can clearly identify the backtest period, the live date or the launch date and therefore the "*live*" period, that is the period during which the ARP indices have been actually launched and actively traded. The data set covers the four asset classes and is categorized into four main groups: carry, momentum, value and hedge. Each group is further divided in multiple subgroups.[15]

We first look at differences in terms of average returns for the period from the first date at which an index is available until the day before the starting date, minus the average returns for the period from the starting date until the end of the period. Let $t_i = 1, \dots, T_i$, be the observation number for index $i = 1, \dots N$, $r_{t,i}$ the return of index i on observation number t, and $m_i \in (1, T_i)$ be the live date of index i:

$$r_i^{diff} = \left(\frac{1}{m_i - 1} \sum_{t=1}^{m_i - 1} r_{t,i} \right) - \left(\frac{1}{T_i - m_i} \sum_{t=m_i}^{T_i} r_{t,i} \right) \quad (1)$$

Exhibit 8 shows that average returns after the start date are clearly lower than the ones shown in the backtest. This is the case for 227 indices, or 89.4% of the indices that made up the sample. The average annualized return differential is 5.87%. Stated otherwise, a lower performance of an index after its live date compared to its backtest is the norm and not the exception. One might raise the concern that, in order to attract clients with lower risk appetite, live returns are lowered because the providers reduce risk and leverage when the indices are made investible. Indeed, volatility levels decrease after the start date in 79.6% of cases. Yet, is it the only reason behind the reduction in performance? It is hard to believe so. First, reduction in volatility affects our sample to a lesser extent. Second, the annualized risk reduction amounts to 1.29% compared to 5.87% average return contraction. The latter is clearly seen in the Sharpe ratio differentials. Close to 85% of the indices exhibit a lower Sharpe ratio after the start date.

The performance of an index is thus quasi systematically lower once the index is truly investible. In that regard, our results are in line with the ones obtained by Suhonen et al. [2017] and the haircut that should be applied does not differ significantly. In their paper, median annualized Sharpe ratios before and after the launch date amount respectively to 1.20 and 0.31, hence translating into a median haircut of 73%. In our sample, the median annualized "backtested" Sharpe ratio is 0.79 and the one after the launch date is 0.18. The median haircut is thus very close to the one reported by Suhonen et al. and amounts to 74.55%. In both cases, the ratio differential is huge and illustrates well how important the selection bias is.

Note that the previous results do not stem from the choice of the risk metric. Indeed, some could argue that downside risk is more appropriate to appreciate the risk-adjusted performance of these long-short indices. However, using either the Sharpe ratio or the Sortino ratio leads to the same conclusion. The median haircut to apply is still impressive. 74.55% for the first ratio, 76.90% for the second one.

One may also wonder if the differences in terms of sample size between the backtested strategies and their live data could not explain the performance differential. To simulate

strategies, researchers use historical observations that extend over decades and that cover different business cycles and market regimes. On the contrary, a majority of investible indices has been launched over the past recent years. To deal with this issue, we constrain the backtest sample period to have the same duration as the one of the live data. If an index for example has been launched 18 months ago, then only the last 18 months of the backtest is taken into consideration. Applying this procedure for each of the 255 ARP indices the sample is made up of, and calculating the different statistics lead to unchanged results. The haircuts tend even to slightly increase as it can be seen in Exhibit 7. It is however worthwhile noticing that imposing the same sample duration changes somehow the percentages of indices whose mean returns, volatilities and Sharpe ratios decreased after the launch date. 77.3% and 75.2% of the indices have lower mean returns and lower Sharpe ratios, compared to respectively 89% and 85% when the entire sample periods were considered. More important however is the fact that 54% only of the indices now displays lower volatility after the launch date. The latter result seems therefore to contradict or at least to weaken the argument that lower volatility was due to a conscious decision made by index managers to lower risk or to decrease leverage once investors had allocated capital. Lower volatility seems to rather stem from the recent market regime where both implied and realized volatilities have been known to be muted.

Exhibit 7

Median Sharpe Ratio and Sortino Ratio Haircuts

	Full Sample Period	Constrained Sample Period
Sharpe Ratio	74.55%	81.06%
Sortino Ratio	76.90%	82.41%

Median Sharpe and Sortino ratio haircuts of 255 ARP indices. The first column shows the results for the full sample period. The second column shows the results when the duration of the backtest period is set to be of the same as the one of the live data.

The overfitting effect appears to be present among categories, providers and asset classes, and regardless of the indices' currency denominations (see Exhibit 8 below and Exhibits 14 and 15 in

the Appendix). None of the categories, providers or asset classes appears to be free from this bias[16]. Finally, it is worthwhile mentioning that the performance differential tends to be positively related to the pre-launch average return. That is to say, the higher is the performance shown in the backtest, the higher will be the return differential once the index has been launched as shown in Exhibit 18.

Exhibit 8

Overfitting Evidence : Results by Category

Category	N	$\overline{\mu}_{backtest}$	$\overline{\mu}_{live}$	$\overline{\mu}_{diff}$	$\overline{\eta}_{diff}$	% $Indices$ $\overline{\mu}_{diff} < 0$	$\overline{\sigma}_{diff}$	% $Indices$ $\overline{\sigma}_{diff} < 0$	$SR\ haircut$	% $Indices$ $SR_{diff} < 0$	$\overline{m}$	$\overline{T} - \overline{m}$
Carry	112	7.32%	1.66%	-5.66%	-4.31%	91.07%	-1.11%	77.68%	70.74%	86.61%	2720	926
Momentum	72	5.48%	-1.32%	-6.80%	-4.74%	95.83%	-1.36%	76.39%	100.63%	90.28%	2972	749
Value	67	4.00%	0.74%	-3.26%	-3.27%	79.10%	-1.26%	85.07%	73.40%	74.63%	3076	638
Hedge	4	3.84%	-2.53%	-6.37%	-6.99%	100.00%	-5.48%	100.00%	28.79%	100.00%	1745	1144
Total	**255**	**5.87%**	**0.51%**	**-5.36%**	**-4.23%**	**89.41%**	**-1.29%**	**79.61%**	**74.55%**	**84.71%**	**2869**	**804**

Overfitting evidence classified by strategies' category. The backtest period and live period differ among the indices. N is the number of indices in the sample; $\overline{\mu}_{backtest}$ the annualized mean backtested returns; $\overline{\mu}_{live}$ the annualized mean live returns; $\overline{\mu}_{diff}$ the average difference between annualized mean backtested returns and annualized mean live returns; $\overline{\eta}_{diff}$ the median difference between annualized mean backtested returns and annualized mean live returns; % $Indices\ \overline{\mu}_{diff} < 0$ the percentage of indices with mean live returns lower than mean backtested returns; $\overline{\sigma}_{diff}$ the average difference between annualized backtested returns' volatilities and annualized live returns' volatilities; % $Indices\ \overline{\sigma}_{diff} < 0$ the percentage of indices with live returns' volatilities lower than backtested returns' volatilities; $SR\ haircut$ the median Sharpe ratio percentage decrease between backtested and live returns; % $Indices\ SR_{diff} < 0$ the percentage of indices in which the Sharpe ratio of the live period is lower than the Sharpe ratio of the backtested period; $\overline{m}$ the average duration of the backtested period (in number of days); and $\overline{T} - \overline{m}$ the average duration of the live period (in number of days), with $\overline{T}$ being the average total observation period.

Conclusions

This article stresses how relevant the selection process can be when investors allocate capital to Alternative Risk Premia. In theory, ARP are rule-based, transparent and replicable. Picking one index supplied by one specific provider among other indices coming from the same provider or from different providers should not matter. Stated otherwise, the selection process should not significantly alter performance of any ARP after controlling for leverage. Yet, selection does matter. The due diligence process appears to be of crucial importance. A sample of ARP belonging to the same category and representing the same investment strategy is not automatically homogeneous. A series of traditional correlations and drawdown analyses, first following, then by the application of clustering techniques, have shown indeed a fair degree of heterogeneity. Different indices for the same premium can lead to very diverse results. Of course, the degree of heterogeneity varies across the strategies. Some categories, such as FX carry, show a fair degree of homogeneity among their indices while others, such as the category called “equity low volatility”, are greatly heterogeneous, thus being highly provider-dependent. The second issue the paper looks at is the presence of selection bias and more precisely of an overfitting bias within the ARP industry. Due to its relatively young age, the historical performance of most ARP indices comes from simulated past returns rather than real live data. One might therefore question the relevance of these results as providers could show only the results that favour the strategy represented by their indices. Our findings support this suspicion. ARP indices quasi-systematically decrease in performance once an index has been launched. The latter is true for 89% of the indices in our sample and the average return differential between the pre-launch period and the live data amounts to an astonishing 5.87%. The same pattern is obtained in terms of risk-adjusted returns. The median backtest Sharpe ratio equals 0.79. Using live data, the same ratio decreases to 0.18 only, corresponding to a median haircut of 75%.

In a nutshell, our findings confirm that a due diligence process on ARP providers and a well-managed selection procedure are crucial parts of any allocation decision process to ARP. Neglecting them is a sure way to be disappointed.

Acknowledgements

We would like to thank Berouz Fatemi from Tages Capital LLP, London (UK) and Duncan Larraz for their continuous support to this project and for providing part of the data used in our analyses. We are also grateful to Jérôme Berset from EFG Asset Management, Geneva (Switzerland), for his valuable comments and for providing us access to market data. We also would like to thank Cedric Kohler from Fundana as well as Jerome Teiletche, Robert Kosowski and all the participants of the Unigestion Reseach Seminar for their comments and suggestions that helped to improve our manuscript. All remaining errors are our own.

References

Aiken A., C. Clifford and J. Ellis, "Out of the Dark: Hedge Fund Reporting Biases and Commercial Databases.", *Review of Financial Studies*, Vol 26, No 1, (2013), pp. 208- 243.

Asness, C., and A. Frazzini, "The devil in HML's details." *The Journal of Portfolio Management*. Vol 39, No. 4 (2013), pp. 49-68.

Asness, C., A. Frazzini, and L. H. Pedersen, "Quality minus junk." (2014).

Asness, C., T. J. Moskowitz, and L. H. Pedersen, "Value and momentum everywhere." *The Journal of Finance*. Vol. 68, No. 3 (2013), pp. 929-985.

Blin, O., J. Lee, and J. Teiletche, "Alternative risk premia investing: from theory to practice." *Unigestion paper*, (2017).

Boillat, P., N. De Skowronski, and N. S. Tuchschmid, "Cluster analysis: application to sector indices and empirical validation." *Financial Markets and Portfolio Management*. Vol. 16 (2002), pp. 467-486.

Brandt, M. W., and P. Santa-Clara, "Parametric portfolio policies: exploiting characteristics in the cross-section of equity returns." *The Review of Financial Studies*. Vol. 22, No. 9 (2009), pp. 3411-3447.

Brown, S., W. Goetzmann, R. Ibbotson, and S. Ross, "Survivorship Bias in Performance Studies." *The Review of Financial Studies*. Vol. 5, No. 4 (1992), pp. 553-580.

Bruder, B., N. Kostyuchyk, and T. Roncalli, "Risk parity portfolios with skewness risk: an application to factor investing and alternative risk premia." (2016).

Carhart, M. M. "On persistence in mutual fund performance." *The Journal of Finance*. Vol. 52, No. 1 (1997), pp. 57-82.

Carhart, M. M., U.W. Cheah, G. De Santis, H. Farrell, and R. Litterman, "Exotic beta revisited." *Financial Analysts Journal*. Vol. 70 (2014), No. 5.

Chordia, T. and L. Shimakuvar, "Momentum, Business Cycle and Time-varying Expected Return.", *The Journal of Finance*. Vol. 57, No. 2 (2002), pp. 985-1019.

Correia, M., R. Scott, and T. Irem, "Value investing in credit markets." *Review of Accounting Studies.* Vol. 13, No. 3 (2012) pp. 572-609.

Daniel, K., and T. J. Moskowitz, "Momentum crashes." *Journal of Financial Economics*. Vol. 122, No. 2 (2016), pp. 221-247.

Fama, E., and K. French, "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics*. Vol. 33, No. 1 (1993), pp. 3-56.

Fama, E., and K. French, "A five factor asset pricing model." *Journal of Financial Economics*. Vol. 116, No. 1 (2015), pp. 1-22

Frazzini, A., and L. H. Pedersen, "Betting against beta." *Journal of Financial Economics*. Vol. 111, No. 1 (2014), pp. 1-25.

Fung, W., and D. A. Hsieh, "Performance characteristics of hedge funds and commodity funds: natural vs. spurious biases." *Journal of Financial and Quantitative Analysis*. Vol. 35, No. 3 (2000), pp. 291-307.

– –. "Hedge fund benchmarks: a risk-based approach." *Financial Analysts Journal*. Vol. 60, No. 5 (2004), pp. 65-80.

Gibson, R., and S. Gyger, "The style consistency of hedge funds." *European Financial Management*. Vol. 13, No. 2 (2007), pp. 287-308.

Hamdan, R., F. Pavlowsky, T. Roncalli, and B. Zheng, "A primer on alternative risk premia." (2016).

Harvey, C. R., L. Yan, and Z. Heqing, "... and the cross-section of expected returns." *Review of Financial Studies.* Vol. 29, No. 1 (2016), pp. 5-68.

Israel, R., and T. Maloney, "Understanding style premia." *The Journal of Investing*. Vol. 23, No. 4 (2014), pp. 15-22.

Jegadeesh, N., and S. Titman, "Returns to buying winners and selling losers: implications for stock market efficiency." *The Journal of Finance*. Vol. 48, No. 1 (1993), pp. 65-91.

Lempérière, Y., C. Deremble, T. T. Nguyen, P. Seager, M. Potters, and J. P. Bouchaud, "Risk premia: asymmetric tail risks and excess returns." *Quantitative Finance*. Vol. 17, No. 1 (2017), pp. 1-14.

Liang, B. "Hedge Funds: The Living and the Dead." *Journal of Financial and Quantitative Analysis*. Vol. 35, No. 3 (2000), pp. 309-326.

Lintner, J. "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets." *The Review of Economics and Statistics*. Vol. 47, No. 1 (1965), pp. 13-37.

Maeso, J. M., and L. Martellini, "Factor investing and risk allocation: from traditional to alternative risk premia harvesting." *The Journal of Alternative Investments*. Vol. 20, No. 1 (2017), pp. 27-42.

Maillard, S., T. Roncalli, and J. Teiletche, "The properties of equally weighted risk contribution portfolios." *The Journal of Portfolio Management*. Vol. 36, No. 4 (2010), pp. 60-70.
McClean D. and J. Pontiff, "Does academic research destroy stock return predictability?", *Journal of Finance*, forthcoming.
Pastor, L., and R.F. Stambaugh, "Liquidity risk and expected stock returns." *Journal of Political Economy*. Vol. 111, No. 3 (2003), pp. 642-685.
Piotroski, J. D. "Value investing: the use of historical statement information to separate winners from losers." *Journal of Accounting Research*. Vol. 38 (2000), pp. 1-41.
Roncalli, T. "Alternative risk premia: what do we know?" (2017).
Roncalli, T., and G. Weisang, "Risk parity portfolios with risk factors." *Quantitative Finance.* Vol. 16, No. 3 (2016), pp. 377-388.
Ross, S. "The arbitrage theory of capital asset pricing." *Journal of Economic Theory*. Vol. 13, No. 13 (1976), pp. 341-360.
Sharpe, W. F. "Capital asset prices: a theory of market equilibrium under conditions of risk." *The Journal of Finance*. Vol. 19, No. 3 (1964), pp. 425-442.
Suhonen, A., M. Lennkh, and F. Perez, "Quantifying backtest overfitting in alternative beta strategies." *The Journal of Portfolio Management*. Vol. 43, No 2 (2017), pp. 91-105.
Tancar, R., and J. Viebig, "Alternative beta applied–an introduction to hedge fund replication." *Financial Markets and Portfolio Management*. Vol. 22, No. 3 (2008), pp. 259-279.

Appendix

Clustering Algorithm

The agglomerative hierarchical clustering algorithm is a recursive method that begins with each object as a singleton cluster. At each step, it groups two objects into a new cluster, according to 1) *"the behavior of different objects within a cluster that must be as similar as possible";* and at the same time 2) *"the behavior of objects that do not belong to the same cluster must be as distinct as possible from the other groups"* (see Boillat et al. [2002]). The algorithm repeats again, iteratively, until finalizing with one cluster that contains all the objects. We must then decide the number of clusters we want to have by imposing a cutoff coefficient. As opposed to other clustering techniques, we do not need to assume *ex ante* the number of clusters. Furthermore, the dendrogram is a clear and useful visualization tool since it allows to recognize the groups that are "strongly" formed, and the ones that are "weakly" formed; and it can identify sub-groups within broader groups. In our tree, the *leafs* are the 173 indices, each one expressed as a vector of standardized daily returns for the period starting on 17 June 2010 and ending on 5 April 2017. (Note that 1.6% of the data are missing values. It is also worth mentioning that these missing values share a common characteristic. Dates are the same and they mostly concern indices which use Asia or Japan as their investment universe.)

Daily returns are indeed standardized to eliminate any possible differences coming from leverage effects. To measure the dissimilarity between clusters to form a new group we rely on the Ward method, which uses the "*increase in the total within-cluster sum of squares (the sum of the squares of the distances between all objects in the cluster and the centroid of the cluster) as a result of joining two clusters*", computed as follows (see, e.g., MATLAB Documentation web site for the *linkage* function):

$$d(r,s) = \sqrt{\frac{2n_r n_s}{n_r + n_s}} \|\overline{\boldsymbol{x}}_{\boldsymbol{r}} - \overline{\boldsymbol{x}}_{\boldsymbol{s}}\|_2 \qquad (2)$$

where $\|\cdot\|$ is the Euclidean distance; $\overline{\boldsymbol{x}}_{\boldsymbol{r}}$ and $\overline{\boldsymbol{x}}_{\boldsymbol{s}}$ are the centroids of clusters $\boldsymbol{r}$ and $\boldsymbol{s}$; $\boldsymbol{n_r}$ and $\boldsymbol{n_s}$ are the number of elements (objects) in clusters $\boldsymbol{r}$ and $\boldsymbol{s}$. Being closely related to the centroid method, we prefer the Ward method because the factor $\sqrt{\frac{2n_r n_s}{n_r+n_s}}$ encourages the formation of clusters with a similar number of objects, rather than, for instance, one or few clusters with a large number of objects and lots of singleton clusters. Another intuitive way to compute the Ward dissimilarity is:

$$c(r,s) = \left(\sum_{i=1}^{n_{r+s}} \|\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{\boldsymbol{r+s}}\|_2^2\right) - \left(\sum_{i=1}^{n_r} \|\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{\boldsymbol{r}}\|_2^2 + \sum_{i=1}^{n_{r+s}} \|\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{\boldsymbol{s}}\|_2^2\right) \quad (3)$$

$$= \frac{n_r n_s}{n_r + n_s} \|\overline{\boldsymbol{x}}_{\boldsymbol{r}} - \overline{\boldsymbol{x}}_{\boldsymbol{s}}\|_2^2$$

where the first argument out of the parenthesis is the sum of the squared distances between each of the objects in the "new cluster" (formed by merging the cluster r and s) and its centroid $\overline{\boldsymbol{x}}_{\boldsymbol{r+s}}$; and the second argument inside the parenthesis is the sum of the sum of the squared distances between the objects in each of the two "old clusters", r and s, and their respective centroid $\overline{\boldsymbol{x}}_{\boldsymbol{r}}$ and $\overline{\boldsymbol{x}}_{\boldsymbol{s}}$. By subtracting the first argument from the second one we thus calculate the increase in "dissimilarity" when joining the two "old clusters" into one "new cluster". Thus, $c(r,s)$ can be seen as a "cost function" of joining clusters r with s. At each step, the two clusters with the minimum "cost" are joined into a new cluster. (For a more detailed explanation, one can for instance refer to the lecture notes of Prof. Cosma Shalizi, Statistics 36-350: Data Mining, Fall 2009, from Carnegie Mellon University.)

The cutoff threshold determines the number of clusters. There is no rule to decide its value. Common practice is to set the threshold when there is a big jump in distance; or to set it according to the characteristics of the data. In the case of ARP indices, threshold could be set according to the number of categories that the industry uses.

Additional Results

Exhibit 9

Number of Indices by Asset Class and Factor					
	Equities	**Rates**	**Credit**	**Commodities**	**Currencies**
Carry	8	20	3	21	14
Value	13	3	1	2	5
Momentum	14	3	1	4	2
Trend	4	4	2	4	7
Volatility Carry	14	7	2	18	10
Mean Reversion	6				2
Size	11				
Quality	12				
Low Vol/Beta	12				
Profitability	4				
Merger Arbitrage	1				

Indices included in the sample used to perform our analyses, by asset class and factor. From a bigger sample, the indices with no available data from 17 June 2010 to 5 April 2017 have been excluded. In this table we include indices denominated in various currencies and of "excess return" and "total return" type.

Exhibit 10

Descriptive Statistics: Equity Indices

Factor	N	$\widehat{\mu}_{1/N}$	$\widehat{\sigma}_{1/N}$	$\widehat{skew}_{1/N}$	$\widehat{kurt}_{1/N}$	$\min\{\widehat{\mu}_i\}$ $i \in [1, N]$	$\max\{\widehat{\mu}_i\}$ $i \in [1, N]$	$\min\{\widehat{\sigma}_i\}$ $i \in [1, N]$	$\max\{\widehat{\sigma}_i\}$ $i \in [1, N]$
Carry	2	-0.60%	5.42%	-0.09	4.23	-1.95%	0.75%	5.48%	9.17%
Value	7	-1.32%	3.04%	0.51	5.87	-3.65%	0.17%	3.90%	6.22%
Momentum	7	2.10%	5.57%	-0.39	4.69	-6.50%	11.97%	4.76%	12.84%
Trend	4	-0.84%	4.77%	-0.60	6.24	-4.11%	0.76%	2.89%	8.67%
Volatility Carry	10	6.29%	5.33%	-2.85	19.87	1.66%	9.93%	2.05%	12.44%
Mean Reversion	3	3.52%	8.35%	-1.02	29.35	0.63%	5.44%	5.53%	14.75%
Size	7	-0.34%	3.14%	-0.24	4.36	-4.33%	2.01%	3.91%	7.32%
Quality	6	1.47%	3.07%	0.20	5.39	-1.96%	4.00%	3.52%	7.13%
Low Vol./Beta	7	-1.38%	2.80%	-0.40	6.17	-8.79%	2.49%	2.44%	8.63%
Profitability	1	-0.82%	3.29%	-0.06	4.50	-0.82%	-0.82%	3.29%	3.29%

Descriptive statistics for equity indices for each factor using the sample period from 17 June 2010 to 5 April 2017. For each factor, an equally weighted index is constructed using all the indices of that group to compute the results. Only USD-denominated indices of type "excess return" are included. N is the number of indices in the sample for each factor; $\widehat{\mu}_{1/N}$ the historical annualized mean return of an equally weighted index in which its components are the sample indices for each factor; $\widehat{\sigma}_{1/N}$ the historical annualized standard deviation of the equally weighted index; $\widehat{skew}_{1/N}$ the historical skewness of the equally weighted index; $\widehat{kurt}_{1/N}$ the historical kurtosis of the equally weighted index; $\min\{\widehat{\mu}_i\}, i \in [1, N]$ the historical mean return of the index with the lowest mean return from the ones used to construct the 1/N index; $\max\{\widehat{\mu}_i\}, i \in [1, N]$ the historical mean return of the index with the highest mean return from the ones used to construct the 1/N index; $\min\{\widehat{\sigma}_i\}, i \in [1, N]$ the historical standard deviation of the index with the lowest standard deviation from the ones used to construct the 1/N index; and $\max\{\widehat{\sigma}_i\}, i \in [1, N]$ the historical standard deviation of the index with the highest standard deviation from the ones used to construct the 1/N index.

Exhibit 11

Descriptive Statistics: Rates, Credit, Commodities and FX

Factor	N	$\widehat{\mu}_{1/N}$	$\widehat{\sigma}_{1/N}$	$\widehat{skew}_{1/N}$	$\widehat{kurt}_{1/N}$	$min\{\widehat{\mu}_i\}$ $i \in [1, N]$	$max\{\widehat{\mu}_i\}$ $i \in [1, N]$	$min\{\widehat{\sigma}_i\}$ $i \in [1, N]$	$max\{\widehat{\sigma}_i\}$ $i \in [1, N]$
Panel A: Rates									
Carry	13	1.24%	1.97%	0.06	5.98	0.09%	5.05%	0.60%	7.10%
Value	3	0.76%	2.82%	0.07	4.73	-0.39%	1.37%	2.52%	6.00%
Momentum	2	0.28%	3.22%	0.05	4.86	0.18%	0.38%	1.47%	5.91%
Trend	3	2.48%	2.90%	0.28	6.91	2.19%	2.73%	2.85%	4.13%
Volatility Carry	6	3.72%	5.35%	-2.49	17.00	0.55%	8.77%	0.50%	11.29%
Panel B: Credit									
Carry	2	3.60%	2.87%	-0.15	5.83	3.25%	3.94%	3.02%	3.02%
Trend	1	4.02%	7.34%	0.01	6.34	4.02%	4.02%	7.34%	7.34%
Volatility Carry	1	10.50%	4.33%	-1.64	19.45	10.50%	10.50%	4.33%	4.33%
Panel C: Commodities									
Carry	21	1.78%	2.61%	-0.26	4.69	-3.98%	4.55%	0.69%	14.21%
Value	2	0.81%	5.50%	-0.01	3.42	-2.19%	3.81%	3.68%	11.38%
Momentum	4	0.56%	11.75%	-0.18	4.45	-2.84%	3.65%	9.35%	19.40%
Trend	4	1.74%	6.42%	-0.48	9.21	0.45%	3.12%	1.82%	9.76%
Volatility Carry	18	6.90%	5.51%	-2.14	16.94	1.44%	15.68%	5.24%	13.55%
Panel D: FX									
Carry	13	0.27%	5.87%	-0.37	5.10	-1.78%	2.73%	3.54%	10.41%
Value	5	0.67%	3.87%	0.44	21.37	-2.39%	2.21%	3.95%	8.17%
Momentum	2	-0.33%	7.34%	-0.07	8.29	-0.72%	0.06%	7.16%	7.94%
Trend	7	1.01%	4.92%	-0.07	6.06	-1.12%	6.53%	2.59%	10.17%
Volatility Carry	10	1.64%	3.57%	-4.81	64.72	-0.42%	6.18%	1.21%	9.02%
Mean Reversion	2	2.86%	9.16%	-3.60	116.55	-0.09%	5.80%	4.28%	16.07%

Descriptive statistics for rates, credit, commodity and FX indices for each factor using the sample period from 17 June 2010 to 5 April 2017. For each factor, an equally weighted index is constructed using all the indices of that group to compute the results. Only USD-denominated indices of type "excess return" are included. N is the number of indices in the sample for each factor; $\widehat{\mu}_{1/N}$ the historical annualized mean return of an equally weighted index ; $\widehat{\sigma}_{1/N}$ the historical annualized standard deviation of the equally weighted index; $\widehat{skew}_{1/N}$ the historical skewness of the equally weighted index; $\widehat{kurt}_{1/N}$ the historical kurtosis of the equally weighted index; $min\{\widehat{\mu}_i\}, i \in [1, N]$ the historical mean return of the index with the lowest mean return from the ones used to construct the 1/N index; $max\{\widehat{\mu}_i\}, i \in [1, N]$ the historical mean return of the index with the highest mean return from the ones used to construct the 1/N index; $min\{\widehat{\sigma}_i\}, i \in [1, N]$ the historical standard deviation of the index with the lowest standard deviation from the ones used to construct the 1/N index; and $max\{\widehat{\sigma}_i\}, i \in [1, N]$ the historical standard deviation of the index with the highest standard deviation from the ones used to construct the 1/N index.

Exhibit 12

Correlations

Panel A: G10 FX Carry Indices					
	IB2	IB5	IB6	IB1	IB4
IB2		0.81	0.62	0.84	0.96
IB5	0.81		0.67	0.87	0.83
IB6	0.62	0.67		0.65	0.58
IB1	0.84	0.87	0.65		0.87
IB4	0.96	0.83	0.58	0.87	
Panel B: Commodity Trend Indices					
	IB2	IB5	IB4	IB6	Bench.
IB2		0.75	0.37	0.72	-0.15
IB5	0.75		0.61	0.82	-0.32
IB4	0.37	0.61		0.59	-0.18
IB6	0.72	0.82	0.59		-0.10
Bench.	-0.15	-0.32	-0.18	-0.10	
Panel C: Equity Size Indices					
	IB1	IB4	IB6	Bench	
IB1		0.53	0.67	0.24	
IB4	0.53		0.33	-0.25	
IB6	0.67	0.33		0.03	
Bench.	0.24	-0.25	0.03		
Panel D: Equity Trend Indices					
	IB3	IB5	IB4	IB6	Bench.
IB3		0.80	0.70	0.86	-0.04
IB5	0.80		0.74	0.80	-0.16
IB4	0.70	0.74		0.64	0.06
IB6	0.86	0.80	0.64		-0.03
Bench.	-0.04	-0.16	0.06	-0.03	
Panel E: Equity Value Indices					
	IB1	IB4	IB6	Bench.	
IB1		0.49	0.52	0.25	
IB4	0.49		0.37	0.14	
IB6	0.52	0.37		0.04	
Bench.	0.25	0.14	0.04		
Panel F: Equity Low Volatility Indices					
	IB1	IB4	IB6	Bench.	
IB1		0.02	0.34	-0.23	
IB4	0.02		0.17	-0.28	
IB6	0.34	0.17		-0.43	
Bench.	-0.23	-0.28	-0.43		

Cross-correlations between the indices of each group and correlations of the indices with the benchmark (when applicable) for the six premia selected to carry the correlations and drawdowns analyses. For each group and coefficient, we show the average, minimum and maximum values. The benchmark indices are the S&P GSCI for the commodity trend premium, the MSCI ACWI for equities premia with global universe and the S&P 500 for equities premia with US universe. The sample periods begin on January 2006 for FX carry, equity size, value and low volatility; and on January 2007 for commodity trend and equity trend premia, and end on April 2017 for the six groups. Only USD-denominated indices of type "excess return" are included.

Exhibit 13

Maximum Drawdowns

Panel A: G10 FX Carry Indices

Provider	Drawdown	Start Date	End Date	Duration
IB2	37.07%	24/07/07	02/02/09	559
IB5	29.29%	11/04/13	20/01/16	1014
IB6	23.45%	11/04/13	03/05/16	1118
IB1	15.74%	11/04/13	11/02/16	1036
IB4	32.45%	24/07/07	02/02/09	559

Panel B: Commodity Trend Indices

Provider	Drawdown	Start Date	End Date	Duration
IB2	15.52%	29/04/11	01/07/14	1159
IB5	14.00%	05/12/08	25/08/10	628
IB4	4.27%	20/01/16	10/03/17	415
IB6	22.04%	08/04/11	16/07/14	1195
Bench.	69.47%	03/07/08	20/01/16	2757

Panel C: Equity Size Indices

Provider	Drawdown	Start Date	End Date	Duration
IB1	12.24%	29/09/08	21/11/08	53
IB4	12.74%	29/09/08	21/11/08	53
IB6	23.44%	27/02/07	09/01/08	316
Bench.	56.78%	09/10/07	09/03/09	517

Panel D: Equity Trend Indices

Provider	Drawdown	Start Date	End Date	Duration
IB3	14.01%	03/07/14	04/11/16	855
IB5	14.69%	09/03/09	16/11/12	1348
IB4	10.87%	09/03/09	16/11/12	1348
IB6	34.84%	09/03/09	04/11/16	2797
Bench.	59.62%	31/10/07	09/03/09	495

Panel E: Equity Value Indices

Provider	Drawdown	Start Date	End Date	Duration
IB1	28.48%	18/05/07	06/03/09	658
IB4	21.55%	22/06/07	30/03/09	647
IB6	25.85%	03/03/10	06/07/16	2317
Bench.	56.78%	09/10/07	09/03/09	517

Panel F: Equity Low Volatility Indices

Provider	Drawdown	Start Date	End Date	Duration
IB1	6.21%	20/06/16	06/02/17	231
IB4	25.42%	27/10/08	24/02/14	1946
IB6	11.48%	20/11/08	23/10/09	337
Bench.	59.62%	31/10/07	09/03/09	495

Maximum drawdowns, with drawdown starting date, ending date, and duration, of the indices for the six risk premia selected to carry the correlations and drawdowns analyses; and their respective benchmarks. The benchmark indices are the S&P GSCI for the commodity trend premium, the MSCI ACWI for equities premia with global universe and the S&P 500 for equities premia with US universe. The sample periods begin on January 2006 for FX carry, equity size, value, and low volatility; and on January 2007 for commodity trend and equity trend premia, and end on April 2017 for the six groups. Only USD-denominated indices of type "excess return" are included.

Exhibit 14

Overfitting Evidence: Results by Provider and Currency Denominations

Provider	N	$\overline{\mu}_{backtest}$	$\overline{\mu}_{live}$	$\overline{\mu}_{diff}$	$\overline{\eta}_{diff}$	% $Indices$ $\overline{\mu}_{diff} < 0$	$\overline{\sigma}_{diff}$	% $Indices$ $\overline{\sigma}_{diff} < 0$	$SR\ haircut$	% $Indices$ $SR_{diff} < 0$	$\overline{m}$	$\overline{T} - \overline{m}$
IB1	14	7.31%	1.19%	-6.12%	-5.07%	92.86%	-0.50%	85.71%	-60.57%	92.86%	2835	1074
IB2	10	8.09%	2.88%	-5.21%	-4.31%	80.00%	-1.09%	70.00%	-52.84%	70.00%	2429	1161
IB3	1	10.57%	5.99%	-4.58%	-4.58%	100.00%	-2.77%	100.00%	-20.79%	100.00%	652	1440
IB4	26	6.46%	1.59%	-4.87%	-4.77%	92.31%	-1.33%	84.62%	-64.04%	88.46%	2835	819
IB5	9	10.19%	-0.41%	-10.60%	-8.39%	100.00%	-6.17%	88.89%	-78.31%	100.00%	2341	1120
IB6	26	8.07%	0.91%	-7.16%	-5.93%	96.15%	-0.58%	69.23%	-96.49%	96.15%	2641	1226
IB7	36	6.39%	-0.20%	-6.58%	-5.83%	97.22%	-0.67%	66.67%	-85.42%	91.67%	2966	757
IB8	69	4.27%	1.12%	-3.15%	-3.14%	79.71%	-1.56%	88.41%	-71.69%	72.46%	3103	692
IB9	33	2.06%	-2.86%	-4.93%	-2.26%	84.85%	-0.66%	69.70%	-97.29%	78.79%	2375	350
IB10	6	5.60%	1.86%	-3.74%	-2.69%	100.00%	-0.85%	100.00%	-52.78%	100.00%	4670	707
IB11	18	9.59%	1.57%	-8.03%	-4.04%	94.44%	-1.65%	77.78%	-76.16%	94.44%	2875	1057
IB12	7	5.05%	0.34%	-4.71%	-4.09%	100.00%	-2.08%	100.00%	-91.28%	85.71%	3508	542
USD	200	6.06%	0.29%	-5.76%	-4.52%	90.00%	-1.18%	78.00%	80.66%	87.50%	2860	828
EUR	52	5.15%	1.11%	-4.04%	-3.58%	88.46%	-1.76%	86.54%	70.71%	75.00%	2943	709

Overfitting evidence, classified by provider and by currency denomination below. The backtest period and live period differ among the indices. $\boldsymbol{N}$ is the number of indices in the sample; $\overline{\boldsymbol{\mu}}_{\boldsymbol{backtest}}$ the annualized mean backtested returns; $\overline{\boldsymbol{\mu}}_{\boldsymbol{live}}$ the annualized mean live returns; $\overline{\boldsymbol{\mu}}_{\boldsymbol{diff}}$ the average difference between annualized mean backtested returns and annualized mean live returns; $\overline{\boldsymbol{\eta}}_{\boldsymbol{diff}}$ the median difference between annualized mean backtested returns and annualized mean live returns; % $\boldsymbol{Indices}\ \overline{\boldsymbol{\mu}}_{\boldsymbol{diff}} < \boldsymbol{0}$ the percentage of indices with mean live returns lower than mean backtested returns; $\overline{\boldsymbol{\sigma}}_{\boldsymbol{diff}}$ the average difference between annualized backtested returns' volatilities and annualized live returns' volatilities; % $\boldsymbol{Indices}\ \overline{\boldsymbol{\sigma}}_{\boldsymbol{diff}} < \boldsymbol{0}$ the percentage of indices with live returns' volatilities lower than backtested returns' volatilities; $\boldsymbol{SR\ haircut}$ the median Sharpe ratio percentage decrease between backtested and live returns; % $\boldsymbol{Indices}\ \boldsymbol{SR}_{\boldsymbol{diff}} < \boldsymbol{0}$ the percentage of indices in which the Sharpe ratio of the live period is lower than the Sharpe ratio of the backtested period; $\overline{\boldsymbol{m}}$ the average duration of the backtested period (in number of days); and $\overline{\boldsymbol{T}} - \overline{\boldsymbol{m}}$ the average duration of the live period (in number of days), with $\overline{\boldsymbol{T}}$ being the average total observation period.

Exhibit 15

Overfitting Evidence: Results by Asset Class

Asset Class	N	$\overline{\mu}_{backtest}$	$\overline{\mu}_{live}$	$\overline{\mu}_{diff}$	$\overline{\eta}_{diff}$	% $Indices$ $\overline{\mu}_{diff} < 0$	$\overline{\sigma}_{diff}$	% $Indices$ $\overline{\sigma}_{diff} < 0$	$SR\ haircut$	% $Indices$ $SR_{diff} < 0$	$\overline{m}$	$\overline{T} - \overline{m}$
Equities	79	4.96%	-0.41%	-5.37%	-4.61%	82.28%	-2.10%	82.28%	-70.66%	77.22%	2901	631
Fixed Income	53	3.33%	0.74%	-2.58%	-2.50%	90.57%	-1.28%	90.57%	-78.76%	81.13%	2793	867
Commodities	52	7.02%	0.63%	-6.39%	-5.54%	96.15%	-0.59%	78.85%	-69.84%	96.15%	3111	843
FX	29	5.19%	1.63%	-3.56%	-2.98%	86.21%	-0.54%	68.97%	-86.03%	79.31%	2836	1111
Total	**213**	**5.09%**	**0.41%**	**-4.68%**	**-4.09%**	**88.26%**	**-1.31%**	**81.69%**	**-72.48%**	**83.10%**	**2916**	**807**

Overfitting evidence, classified by asset classes. The backtest period and live period differ among the indices. N is the number of indices in the sample; $\overline{\mu}_{backtest}$ the annualized mean backtested returns; $\overline{\mu}_{live}$ the annualized mean live returns; $\overline{\mu}_{diff}$ the average difference between annualized mean backtested returns and annualized mean live returns; $\overline{\eta}_{diff}$ the median difference between annualized mean backtested returns and annualized mean live returns; % $Indices\ \overline{\mu}_{diff} < 0$ the percentage of indices with mean live returns lower than mean backtested returns; $\overline{\sigma}_{diff}$ the average difference between annualized backtested returns' volatilities and annualized live returns' volatilities; % $Indices\ \overline{\sigma}_{diff} < 0$ the percentage of indices with live returns' volatilities lower than backtested returns' volatilities; $SR\ haircut$ the median Sharpe ratio percentage decrease between backtested and live returns; % $Indices\ SR_{diff} < 0$ the percentage of indices in which the Sharpe ratio of the live period is lower than the Sharpe ratio of the backtested period; $\overline{m}$ the average duration of the backtested period (in number of days); and $\overline{T} - \overline{m}$ the average duration of the live period (in number of days), with $\overline{T}$ being the average total observation period.

Exhibit 16: Dendrogram – All Asset Classes

Dendrogram (hierarchical tree) representing the 173 indices of the sample that are USD-denominated and of "excess return" type. By setting the cutoff at 90, 12 groups are formed. The Ward dissimilarity method is used, with the Euclidean distance as the distance measure and the daily returns as inputs for the sample from 17 June 2010 to 5 April 2017.

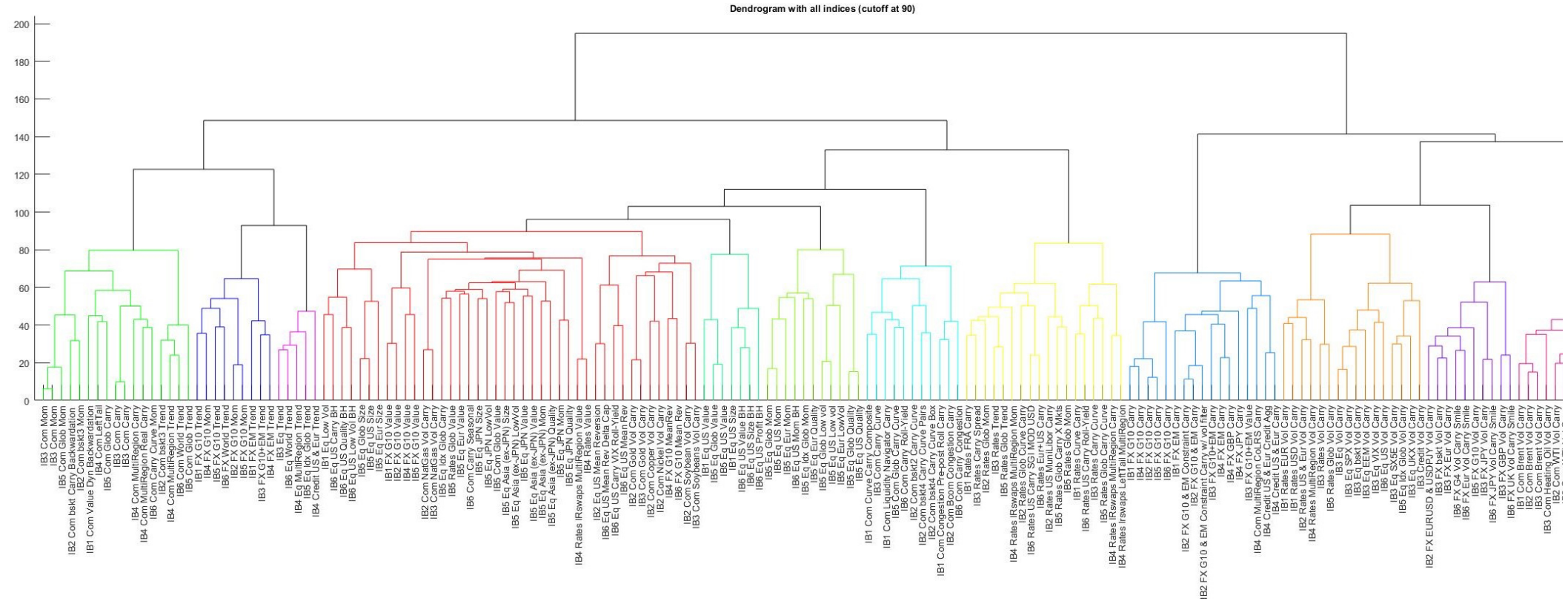


Exhibit 17: Dendrogram – Equity Indices

Dendrogram (hierarchical tree) representing the 84 equity indices of the sample, EUR- or USD-denominated and of "excess return" type. By setting the cutoff at 80, 10 groups are formed. The Wa dissimilarity method is used, with the Euclidean distance as the distance measure and the daily returns as inputs for the sample from 17 June 2010 to 5 April 2017.

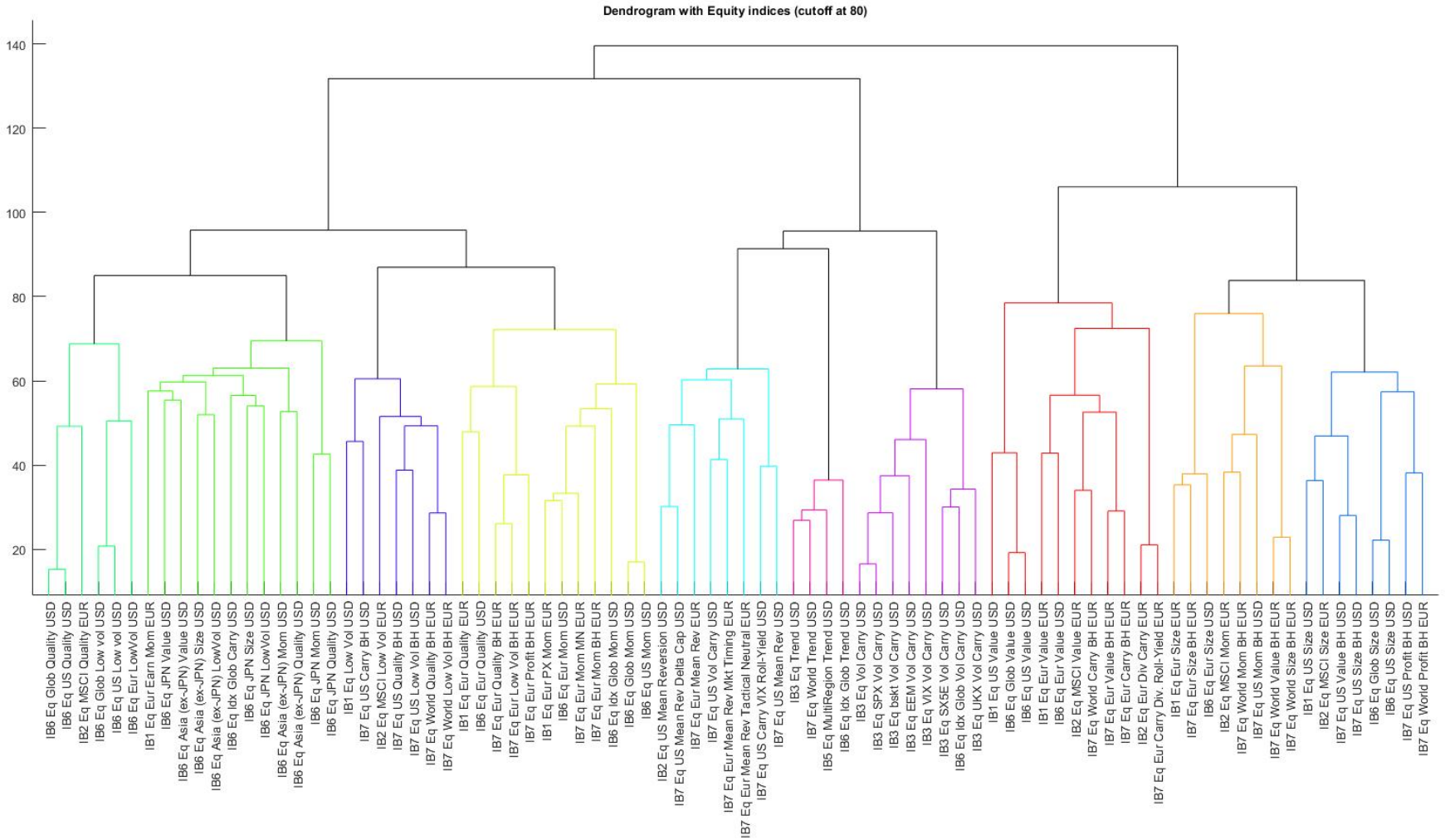
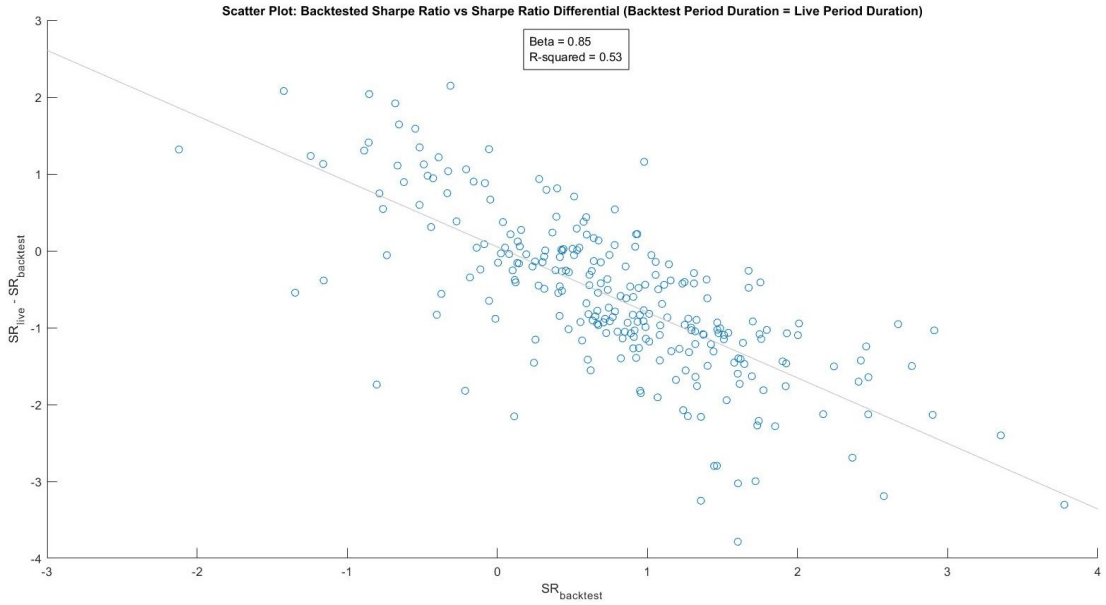


Exhibit 18: Backtested Sharpe Ratio vs. Live Sharpe Ratio

The scatter plot shows the relationship between 255 backtested Sharpe Ratios and the Sharpe Ratios differential between live and backtest periods. Statistics are calculated imposing the same duration for the sample periods of the backtests and the sample periods of the live data.



[1] For example, BarclayHedge gives the estimate of $2.7 trillion in 2015 and $3 trillion in 2016.

[2] "From Diversified Asset Classes to Factor-Driven Index Portfolios & the Re-Packaging of Active Investments". Citi Business Advisory Services

[3] Transparency is an argument quite often put forward in favour of traditional rule-based investing vs for example long-short strategies or hedge funds. Yet, careful investigation can reveal that alpha providers can be as transparent if not even more transparent than other managers. They will be (or should be) transparent to their investors but not to market participants.

[4] In fact, any strategy or asset that fulfils the above characteristics could be considered as an Alternative Risk Premium. For instance, Carhart et al. [2014] include Cat bonds or Real Estate in their "Exotic Betas" universe. In this paper, we focus on the most common premia offered by the "ARP industry".

[5] Price momentum strategies are quite often mentioned as examples for that third and last set of explanations. Yet, the source of momentum strategy profitability is still debatable. For some it comes from investors' behaviour and can be classified as market anomalies. For others, as for instance Chordia and Shivakumar [2002], momentum payoffs stem from time-varying expected returns and are therefore risk-based.

[6] Note that some styles are common for all asset classes, others for a few of them, and some are even asset-specific.

[7] It is certainly interesting to analyse if all three dimensions have an impact in terms of portfolio diversification and which one could have the greatest effect, if any. Somehow, results of the clusters analysis in the following section should help in answering the question.

[8] *Investment* refers to the research paper of Fama and French [2015]. *Skew* captures price differential between puts and calls. *Correlation* intends to extract the premium of dispersion trades.

[9] Complete results can be obtained upon request. The subsample was selected on the basis of the period of the data available. For each of the six categories we could use data covering the financial crisis and its aftermath. Note that the sample period used in this section is longer than the one used to compute the descriptive statistics shown in Exhibits 10 and 11 in the Appendix.

[10] Note that the low correlation with the benchmark argument naturally disappears once ARP are included into existing long-only portfolios. The inclusion of ARP indeed leads to the transformation of the risk-

return profile of the portfolios. Adding *equity size* for instance tilts an existing long-only equity portfolio towards small caps.

[11] Note that trend strategies are not to be confused with momentum strategies even if they are sometimes called "time-series momentum". In a classic momentum strategy, long and short positions are taken according to a relative performance criterion and that irrespective of the presence or not of trends among the different components the sample is made up of.

[12] One should note that the construction of the index provided by IB1, even though classified as a low volatility index by its provider, seems to be more similar to a low beta strategy since it applies a risk parity approach to compute the weight of its components. It is thus worthwhile to remark that some ARP providers do not always differentiate in names between "low volatility" indices and "low beta" indices, leading to misunderstandings. While the "low volatility" premium is constructed by taking long (short) positions in assets that have low (high) historical volatilities, a "low beta" index, as its name can tell us, will rank and pick assets according to their "market beta". Hence, any financial asset with relatively high volatility but low covariance to any market proxy, will be held long in the "low beta" premium index, but sold short in the "low volatility" premium index.

[13] EUR-denominated indices are included here to enrich the sample as 29 out of the 54 USD-denominated equity indices that make up the sample come from the same provider. However, the effect of currency denominations is very limited here when it comes to calculating the Euclidean distances on z-scores in the clustering algorithm. Indeed, the clustering technique looks for similarities in the way the indices covary together. On that, currency translation through forward rates has little impact.

[14] Biases and survivorship bias in particular have attracted the attention of researchers for quite some time. One can cite for example the famous paper of Brown et al. [1992] published in the beginning of the nineties that estimate the impact such biases can have in the case of the mutual funds industry. The analysis for hedge funds followed naturally. One can for instance think of the papers of Liang [2000] or Fung and Hsieh [2000]. The debate regarding the severity of the biases and on their impact on the performance analysis is still current as shown by the recent paper of Aiken et al. [2013].

[15] Nine indices were excluded from our initial sample either because the backtest period or the live period was too short to be relevant.

[16] In relation with the recent results obtained by McClean and Pontiff [2018], one could have expected the overfitting bias to differ among asset classes. Liquid and less complex instruments that typically characterize equity strategies coupled with the fact that long series of data are readily accessible, could

have translated into more severe overfitting bias and therefore higher haircuts than for FI or commodities strategies. The latter is yet not confirmed by referring to the results shown in Exhibit 15.