A Machine Learning Approach for Detecting Closet Index Trackers

Fabrizio Basso*

*Student ID: 10512476
*Emails: 10512476@mydbs.ie; fabrbasso@gmail.com

Applied Financial Analysis - CA2 Course Code: B9FT106

Abstract

Closet Index Tracking (or Closet Indexing and Index Hugging) relates to the practice of a fund manager claiming to actively manage an investment portfolio when in reality the fund closely tracks an index. This paper explores the problem of identifying a Closet Index Tracker Fund. This paper first provides a general overview of the regulatory background and the challenges it poses. Then it moves to analyse a set of funds investing in European Large capitalization Equities according to Morningstar's classification. After calculating a set of statistics on these funds against a set of indices, the paper outlines a procedure to spot potential passive tracker funds relying on clustering unsupervised machine learning algorithms.

Contents

1 Introduction: Regulatory Background

This section aims to provide a general overview on how the regulators have so far tackled the Index Hugging issue with a special attention to the Irish case.

1.1 Europe Goes on War to Closet Index Tracking

The European Securities and Markets Authority (ESMA)'s publication on February the 2nd, 2016, titled 'Supervisory work on potential closet index tracking' (link) [SA19], brought up the issue of the closet index tracking to the attention of the financial industry. By Closet Indexing, ESMA defines the practice whereby a UCITS while identifying its investment strategy as active in their documentation to the public, in fact, implements a passive investment strategy. At the same time, the 'phoney active fund' charges management fees in line with those of actively managed funds. ESMA's report concluded that between 5% to 15% of the funds "could be closet trackers". Spurred by ESMA's publication, national financial watchdogs and regulatory authorities started to test, probe and verify the investment strategies of the entities under their supervision. As a result of these inquiries, a number Consultation Papers, Guidance, Guidelines and Policy Statements have been issued Europe-wide. In Ireland, for instance, the Central Bank (CBI) issued an 'Industry Letter' on the 'Thematic Review of Closet Indexing' (link) [oI19], where it shortlisted a total of 182 funds, out of 2550 analysed, for further review for potential closet index tracking practice. In the UK, as a result of its studies, the FCA issued the Policy Statement 19/4 (link) [Aut19] significantly increasing the transparency whereby the management of existing funds is described to investors. The fallouts for a fund falling under the watchdog scrutiny for closet indexing can be on two counts:

- **Reputational**: In UK some Asset Management companies have been forced to re-issue their marketing material and notify the investors that the Fund was passively replicating an index; and
- **Monetary**: Again, in the UK, 64 funds had to compensate investors a total of £34m, while in Norway, one of the leading national asset management had to refund investors for NKr345m.

1.2 CBI's Analysis and Methodology

The CBI in its review of the fund industry on closet indexing analysed as many as 2'550 Irish authorized UCITS funds classified as active, for a total of 15'500 share classes. Each share class was analysed against a pool of 2'500 indices using a set of statics. As a result of this study, the Central Bank shortlisted 182 funds for further review or 7.1% of the sample. In its analysis, the CBI used the following metrics to identify Closet Indexing funds:

- **TEV**: The Tracking Error Volatility (TEV) measures as the standard deviation of the difference of the index and fund daily returns. The lower this metric is the closer the fund returns are to those of the index;
- **Beta**: Beta is calculated as the coefficient of a linear regression of a fund's returns on an index returns. The closer this metric is to one the more similar the fund is to the benchmark;
- **R2**: another byproduct of the linear regression performed to calculate the Beta statistics, the R^2 measures the proportion of the variation of the fund explained by the index.

In addition to the metrics mentioned above, in the CBI's Letter to the Industry, there are five further elements of great importance. Three of these elements are additional information:

- **Wide pool of Indices**: When evaluating if a fund is active or passive, the CBI did not limit the analysis to a fund's disclosed benchmark, but it considered a wider universe made of 2'500 indices;
- **Values for Money**: The CBI gave great emphasis to the fact that a management style of a fund must be aligned with its fees structure. If a fund has a passive management style, but its fees are in line with those of declared passive funds this is not a reason for concern for the CBI; and

- **Ongoing Concern**: The Central Bank stated very clearly that the Closet Indexing analysis is not a one-off exercise. It should be an ongoing control and examination about a fund's management style and what has been promised to the investors and the fees charged on them.

The remaining two relevant elements are the "untold" details. In presenting its analysis, the CBI omitted, on purpose, to disclose the following pivotal information:

- **Analysis Timeframe**: There is no reference on the time horizon used for the analysis; and
- **Metrics Thresholds**: The metrics used are known, but the significant levels after which a fund is flagged as a closet tracker are unknown.

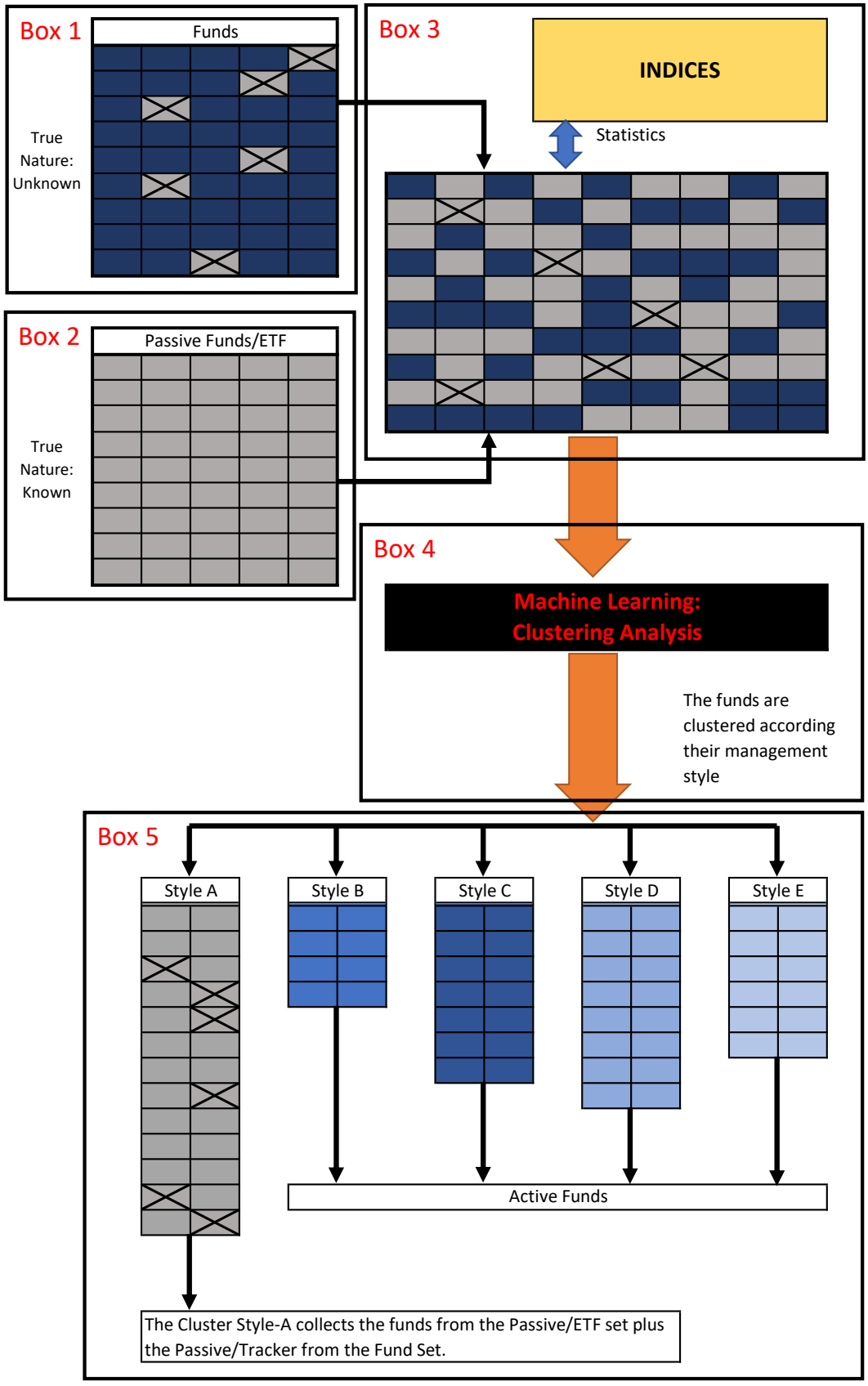
The lack of clarity around these elements created a grey area about which funds can or cannot be classified as a Closet Tracker. As a result, developing a proper methodology to review and monitor funds performances against the Closet index tracking practice is of paramount importance in the industry.

2 How to Track a Tracker: the Methodology

The diagram on the next page outlines the methodology used to detect closet tracker funds. This paper considers a 1-Year timeframe. However, this approach can be replicated over any desired time period. There are five critical steps in the procedure:

1. The starting point is the basket of **funds branding themselves as actively managed but whose real nature is unknown** (Box n.1, where each square represents a fund). Among these funds, some are passively tracking an index (represented by the stricken-out squares in grey). These are the fund that this process aims to identify.
2. The second set of funds is introduced (Box n.2), The fund from this lot are all **passively managed, and their nature is well-known**. For instance, they are ETFs aiming at tracking an underlying declared benchmark/index.
3. These two baskets of funds are merged (Box n.3). The resulting set is composed by three main groups of funds: (1) Active Funds -from box n.1- whose nature is **unknown**; (2) Passive Funds -from box n.1- whose nature is **unknown** and (3) Passive Funds -from box n.2- whose nature is **known**. At this stage, for each fund, the set of statistics used by the CBI to describe a fund's management style are computed against a wide selection of indices. Following the CBI's approach, this basket of indices also includes, but it is not limited to, the funds and the ETFs benchmarks.
4. The style descriptive statistics are then used to train an unsupervised Machine Learning algorithm performing a clustering analysis on the dataset (Box n.4). In this framework, unsupervised learning must be used as the funds from Box n.1 have no labels while those from Box n.2 share all the same label. As a result, the funds are grouped by management styles according to their style metrics.
5. As all the passive tracker funds from Box n.2 share the same management style, they are very likely to be grouped into the same cluster (Box n.5). **So the cluster containing these funds is the one identifying the "closet trackers"**. The funds from Box n.1 grouped in this cluster are those that require further analyzes to assess their real nature and to verify if their fees structure is consistent with their management style.

Finding a Index Tracker: the Methodology



It should be clear that the approach used in this paper differs from the one used from the CBI to the extent that no specific threshold is set to divide the active funds from the closet trackers. Rather than making assumptions about the limits used by the regulator, this process tries to identify the passive funds branded as active, relying on the similarities that they have with the declared passive funds.

3 The Dataset

The starting dataset used for the analysis is made up of three main groups of data. All the time series consist of daily observations starting as far as the 1st of January 2010, or a later date if they were not available at that time. The three baskets of data are the following:

1. **Funds**: This is a set of 40 Euro-denominated funds belonging to the Morningstar category investing in "European Large Cap Equity". Each fund is identified by its Bloomberg Code. These funds are either managed with reference to a Benchmark or are total return funds with no declared benchmark. In addition to the time series, the information set for each fund also covers the Bloomberg code of the relative Benchmark and the running management fees (see Fig. 1). At this stage nothing is known about the **actual** management style of these funds, a part being classified as active by Morningstar.

	Code	Benchmark Code	Fees
0	ACMESIE LX Equity	NDDLE15 Index	1.0900
1	RCMSTYL LX Equity	SX5T Index	0.7100
2	ALEPTEU LX Equity	STGPREZU Index	1.0000
3	AEEI2EC LX Equity	Total Return	2.5000
4	CAIXEIA LX Equity	MXEU Index	0.1500
5	AXWEOFC LX Equity	MXEU Index	1.0000

Figure 1: Funds Information Set.

2. **Passive Funds/ETFs**: It is a collection of 59 Euro denominated ETFs having as a benchmark an European index of large-capitalisation companies. Each ETF is identified by its Bloomberg Code. In addition to the time series, the information set for each fund also covers the Bloomberg code of the relative tracked index and the running management fees (see Fig. 2). During the analysis, the number of ETFs is reduced from 59 to 31. The reason behind this screening is that not all the ETFs experience the same level of liquidity: as an ETF reported closing price is the last traded contract, illiquid ETFs may display artificial divergences from their benchmark.

	Security	Description	Category	Subcategory	Benchmark	Fees
0	SX7EEX GY Equity	iShares EURO STOXX Banks 30-15 UCITS ETF DE (X...	Equities	Exchange Traded Products	SX7315T Index	0.51
1	FEZ US Equity	SPDR EURO STOXX 50 ETF (U.S.)	Equities	Exchange Traded Products	SX5U Index	0.29
2	SX5EEX GY Equity	iShares Core EURO STOXX 50 UCITS ETF DE (Xetra)	Equities	Exchange Traded Products	SX5T Index	0.10
3	MSE FP Equity	Lyxor EURO STOXX 50 DR UCITS ETF Class Dist (E...	Equities	Exchange Traded Products	SX5T Index	0.20
4	EUN2 GY Equity	iShares Core EURO STOXX 50 UCITS ETF EUR Dist ...	Equities	Exchange Traded Products	SX5T Index	0.10
5	BNKE FP Equity	Lyxor EURO STOXX Banks DR UCITS ETF Class Acc ...	Equities	Exchange Traded Products	SX7T Index	0.30

Figure 2: ETF Information Set.

3. **Indices**: It is a collection of 89 Euro denominated large-capitalisation equity indices of European based companies. These indexes are calculated by several sources (Eurostoxx, Euronext, Bloomberg, S&P, etc.) according to different aggregation methodologies (market capitalisation, active shares, volatility constrained, etc.) and coverage (in terms of size, sector, exclusion of specific companies or sectors, etc.). This set includes all the funds and ETFs' benchmarks, and it is the set of indices used to assess the funds' management style.

4 Funds and ETFs: Calculating the Statistics Set

As explained in paragraph 2, the Funds and ETFs datasets are merged, and their daily returns are compared against those from the Indices set. To characterise a fund management style, this paper employs the same metrics outlined in the CBI document. As a result, linear regressions are estimated on the returns of each fund against the indices returns taken one-by-one. The timeframe used for the analysis is one year. After each regression is fitted, the **Beta** and the R^2 are stored in a cross-sectional table, with the Funds/ETFs on the columns while the Indices are on the rows. As for the Beta metric, since the scope is to measure how much different it is from 1, its value is transformed as follow:

$$\hat{\beta^T} = \left|\hat{\beta} - 1\right| \tag{1}$$

The TEV is calculated as the standard deviation of the difference between a fund and an Index's returns. Again the results are stored in a cross-section table. As a result of this step, three cross-section matrices are produced describing the investment style of each fund against every index in the dataset. The following tables reproduce for each of the three features, the first five rows and ten columns:

1. **Beta:**

	ACMESIE LX Equity	RCMSTYL LX Equity	ALEPTEU LX Equity	AEEI2EC LX Equity	CAIXEIA LX Equity	AXWEOFC LX Equity	AXWECEI LX Equity	MEREMAI LX Equity	ETDD FP Equity	CSEMICE LX Equity
BE500 Index	0.0368678	0.424485	0.384412	0.0701813	0.0265991	0.0535709	0.0819003	0.0887743	0.0887147	0.0475178
SX5P Index	0.0606809	0.478713	0.452742	0.0548235	0.0202158	0.0414809	0.0675666	0.110376	0.0928114	0.0294493
SXXE Index	0.111478	0.485567	0.450077	0.0135995	0.125009	0.141706	0.00161842	0.18351	0.0137958	0.0304395
SXXP Index	0.0478293	0.430034	0.390228	0.0636115	0.0259269	0.0573885	0.0716535	0.0988298	0.0788491	0.0410786
EUETMP Index	0.394237	0.554507	0.494692	0.333047	0.392618	0.410132	0.332675	0.391667	0.368657	0.327677

Figure 3: Beta Cross-Sectional Matrix.

2. **R2:**

	ACMESIE LX Equity	RCMSTYL LX Equity	ALEPTEU LX Equity	AEEI2EC LX Equity	CAIXEIA LX Equity	AXWEOFC LX Equity	AXWECEI LX Equity	MEREMAI LX Equity	ETDD FP Equity	CSEMICE LX Equity
BE500 Index	0.839086	0.340804	0.299592	0.861985	0.906513	0.883322	0.867062	0.71594	0.90162	0.868564
SX5P Index	0.754435	0.264306	0.223817	0.7916	0.868185	0.856457	0.798044	0.645058	0.85871	0.792958
SXXE Index	0.864717	0.329724	0.289505	0.886736	0.886953	0.879667	0.899877	0.696032	0.946671	0.901015
SXXP Index	0.82767	0.337351	0.296672	0.859298	0.91615	0.884303	0.858573	0.706694	0.893531	0.865843
EUETMP Index	0.284996	0.175336	0.173324	0.287458	0.303048	0.294612	0.283236	0.273972	0.26033	0.307209

Figure 4: R2 Cross-Sectional Matrix.

3. **TEV:**

	ACMESIE LX Equity	RCMSTYL LX Equity	ALEPTEU LX Equity	AEEI2EC LX Equity	CAIXEIA LX Equity	AXWEOFC LX Equity	AXWECEI LX Equity	MEREMAI LX Equity	ETDD FP Equity	CSEMICE LX Equity
BE500 Index	0.310681	0.664833	0.746074	0.318426	0.230213	0.255452	0.316619	0.426195	0.271811	0.301046
SX5P Index	0.384782	0.708278	0.795613	0.388111	0.272758	0.281524	0.386175	0.477354	0.323115	0.375882
SXXE Index	0.297718	0.710287	0.78486	0.284878	0.27169	0.280712	0.269784	0.460206	0.194616	0.260656
SXXP Index	0.322227	0.668792	0.749532	0.320725	0.218079	0.254881	0.324956	0.434144	0.280621	0.303613
EUETMP Index	0.705334	0.75745	0.822314	0.749048	0.680833	0.680396	0.756441	0.724112	0.765798	0.721757

Figure 5: TEV Cross-Sectional Matrix.

The second step aims to find out for each fund the index that best describes its returns in each metric considered. For instance, when the fund ACMESIE LX is analysed according to the beta statistic, the index against which it scores the lowest value is selected and stored. The same procedure is followed for the TEV and the R^2. The only difference is that for the R^2 the index with the highest value is selected. For each fund, the respective benchmark is added to the list of chosen indices in this way. Eventually, the output is a cross-sectional matrix, as shown in Fig. 6. Here, each fund is matched to a list of indices (max four) according to the unique values in the list.

	Beta	R2	TEV	Benchmark
ACMESIE LX Equity	NDDLE15 Index	NDDLEMU Index	NDDLEMU Index	NDDLE15 Index
RCMSTYL LX Equity	SX5EDFT Index	N150 Index	SX5EDFT Index	SX5T Index
ALEPTEU LX Equity	SX5EDFT Index	N150 Index	SX5EDFT Index	STGPREZU Index
AEEI2EC LX Equity	NDDLEMU Index	S&P_Euro_75	S&P_Euro_75	Total Return
CAIXEIA LX Equity	NDDLE15 Index	MSDEE15N Index	MSDEE15N Index	MXEU Index
AXWEOFC LX Equity	NDDLE15 Index	MSDEE15N Index	MSDEE15N Index	MXEU Index
AXWECEI LX Equity	MSER Index	SXXT Index	SXXT Index	SXXT Index
MEREMAI LX Equity	NDDLE15 Index	S&P_EURO_PLUS	S&P_EURO_PLUS	M7EM Index
ETDD FP Equity	DJST Index	SX5HUN Index	SX5HUN Index	SX5T Index
CSEMICE LX Equity	FTEFC1 Index	NDDLEMU Index	NDDLEMU Index	MXEM Index

Figure 6: Indices shortlisted for Each Funds.

The table reports a "Total Return" entry under the benchmark column when a fund has not declared benchmark. The set of indices selected for each fund represents the best candidates to evaluate if the fund is a closet tracker or not.

As a final step in preparing the cross-sectional data for the analysis, a specific label is created for every match between a fund and the list of potential tracked indices. The metrics describing each couple are retrieved from the tables calculated in the preceding steps. All the duplicates are removed. Fig. 7 shows the first ten rows of the resulting table. In the case of the ETFs, only the metrics against the respective benchmarks have been considered. Overall. The chart captures 175 couples.

	beta	R2	TEV	Type	Fees	Benchmark Kind
Fund:ACMESIE LX Equity, Index:NDDLE15 Index	0.009256	0.836802	0.311759	Fund	1.09	0.0
Fund:ACMESIE LX Equity, Index:NDDLEMU Index	0.103376	0.871452	0.288825	Fund	1.09	0.0
Fund:RCMSTYL LX Equity, Index:SX5EDFT Index	0.208607	0.276730	0.623354	Fund	0.71	0.0
Fund:RCMSTYL LX Equity, Index:N150 Index	0.482851	0.397260	0.704746	Fund	0.71	0.0
Fund:RCMSTYL LX Equity, Index:SX5T Index	0.527797	0.287292	0.748876	Fund	0.71	0.0
Fund:ALEPTEU LX Equity, Index:SX5EDFT Index	0.153629	0.243193	0.721752	Fund	1.00	0.0
Fund:ALEPTEU LX Equity, Index:N150 Index	0.436496	0.362407	0.763137	Fund	1.00	0.0
Fund:ALEPTEU LX Equity, Index:STGPREZU Index	0.482569	0.284327	0.809872	Fund	1.00	0.0
Fund:AEEI2EC LX Equity, Index:NDDLEMU Index	0.006387	0.890447	0.280012	Fund	2.50	1.0
Fund:AEEI2EC LX Equity, Index:S&P_Euro_75	0.038571	0.893360	0.278072	Fund	2.50	1.0

Figure 7: Cross-Sectional data for each couple of fund and candidate tracked index.

5 Preliminary Data Analysis

Some essential features of the data can be inferred by a preliminary graphical exploration of the data. Fig. 7 below shows in box-plots the relative distribution of the statistics between Funds and ETF.

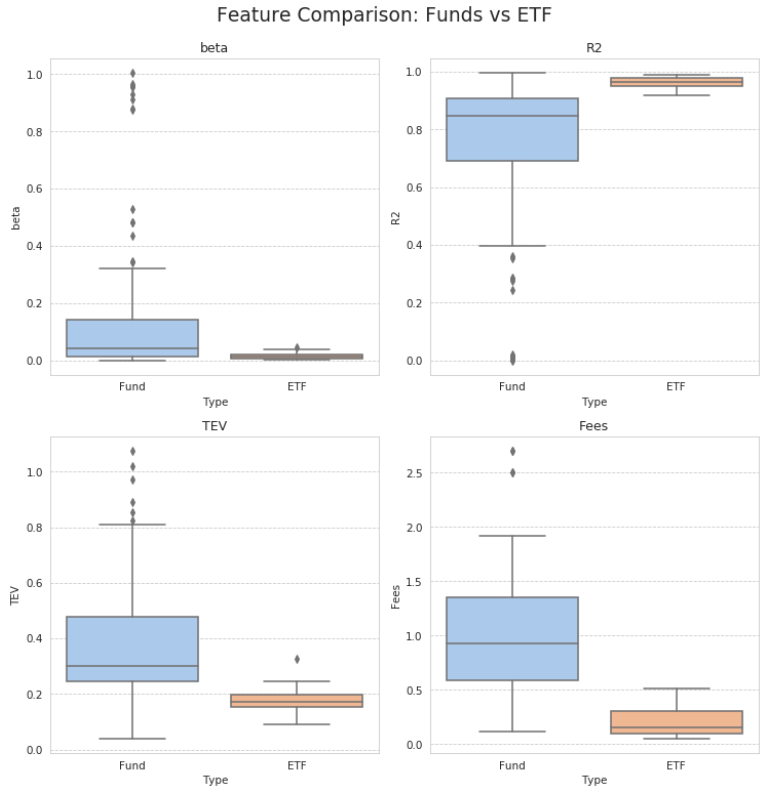


Figure 8: Relative distribution of style metrics divided by funds and ETFs.

1. The distribution of the selected metrics seems to be significantly different when collected from Funds and ETFs. For TEV and R2, the range of the ETFs' readings lies completely in the tail of the funds' distribution.

2. However, the two distributions overlap for all the three metrics describing the fund behaviour. This means that in the funds' pool some share a management style very similar to the passive funds/ETF;

3. The graph presenting the fees reveals another important feature. Some funds have management fees aligned with those of the ETFs. Indeed, if a fund behaves as a passive one but its fee profile is aligned as well, it would be considered still acceptable according to the "Value for Money" logic considered by the CBI.

Further information can be extrapolated from the features' scatter plots (Fig. 9). The critical upshot from these graphs is the relative spatial localisation of the ETF (orange dots) against the funds (blue dots). In all the plots, the ETFs are localised in a specific area meaning that they display consistently a similar set of characteristics that are well described by the selected features. Equally important, there are blue dots (funds) within or close the orange cluster. These funds (not necessarily the same in every graph) display a passive hallmark for the selected couple of features represented in each scatter plot. However, if a fund exhibits similar characteristics to ETFs across all the features couples, it is very likely to be clustered together with the passive funds by the machine learning algorithm.

Figure 9: Features Pairplots.

6 Data Pre-Processing

As many other machine learning algorithms, clustering analysis can be affected by the scale of the variables it processes. By using Fig. 7 to analyse this issue, in the specific case of the cross-sectional data used here, it can be noted that R2 usually scores values as bigger as twice of the size of the beta and the TEV. This difference can affect the analysis as algorithms relaying on Euclidean distance might lean to overweight the importance of R2 in comparison to the other features. Moreover, the scatter plots in Fig. 7 highlight the presence of some outliers that might affect the results. As a result of these considerations, all the features are re-scaled using the following scaling methods: (1) **Standardization**, and (2) **Robust Scaler**. Where the former will take all the features to the same scale and variance and the latter will provide an approach more resilient to the presence of outliers. The results of these transformations are shown in the following graphs.

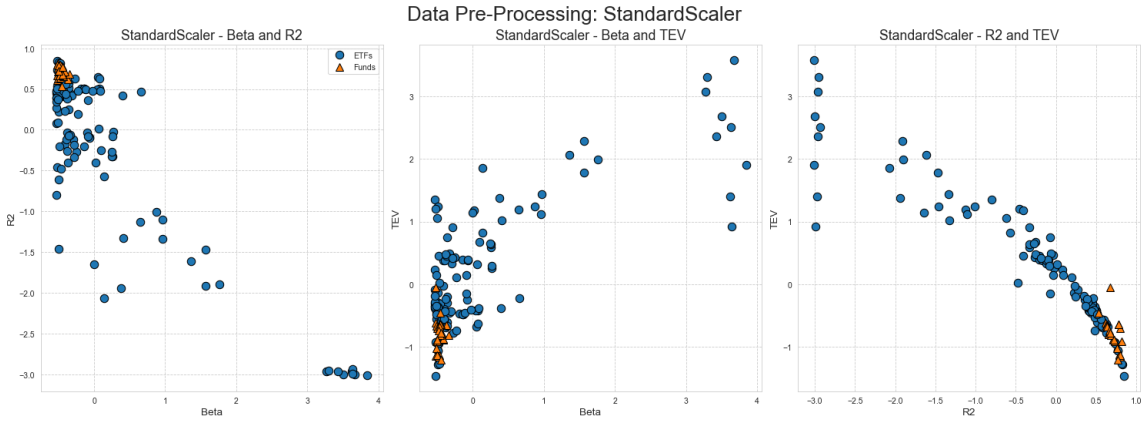


Figure 10: Data Preprocessing: Standard Scaler.

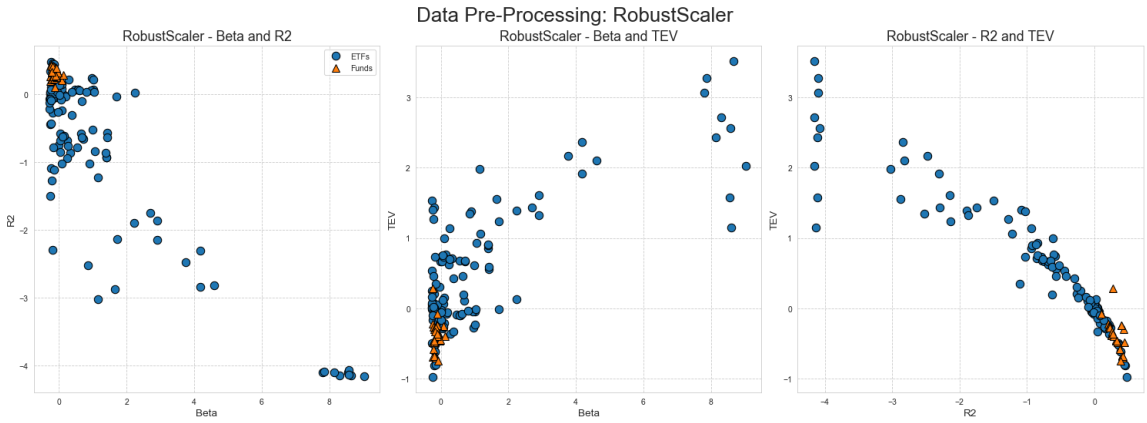


Figure 11: Data Preprocessing: Robust Scaler.

Although the two sets of graphs look similar, the relative scale of the features is sensibly different, and this can have a disparate impact on how the elements are clustered together. From the graphs above, it can be noted that most of the EFTs (orange triangles) are grouped in a specific area of the space. However, some funds (blue dots) are located in the same region as well: these funds display a behaviour comparable with the passive one.

7 Clustering Algorithms and Number of Clusters

In this paper two different clustering approaches have been considered:

1. **K-Means**; and
2. **Gaussian Mixture**.

The first model is a distance-based algorithm and, in this paper, it relies on the "Euclidean" distance to group together the elements. The Gaussian Mixture algorithm is a distribution-based model: it assumes that there are a certain number of Gaussian distributions, and each of these distributions represents a cluster. Since for this dataset there is no ground truth available, a specific study to understand the optimal number of clusters is necessary. Within this particular framework, each cluster represents a different style of management and the cluster grouping together the ETFs along with some other funds should be considered as representative of the "Passive & Closet Tracker" funds.

The following analyses are conducted to understand the right number of clusters:

1. **Elbow Method - For K-Means**;
2. **BIC and AIC analysis - Gaussian Mixture only**

7.1 Elbow Method - K-Means

The Elbow method relies on the measure of the dispersion of a cluster's elements from its centroid, measured as the sum of squared distances. Since this metric depends on the Euclidean measure for the distance, it is better suited to identify the ideal number of clusters for the K-Means model. The analysis covered the data in the three formats considered so far: raw, scaled with the standard-scaler and with the robust scaler. The results are reported in Fig. 12.

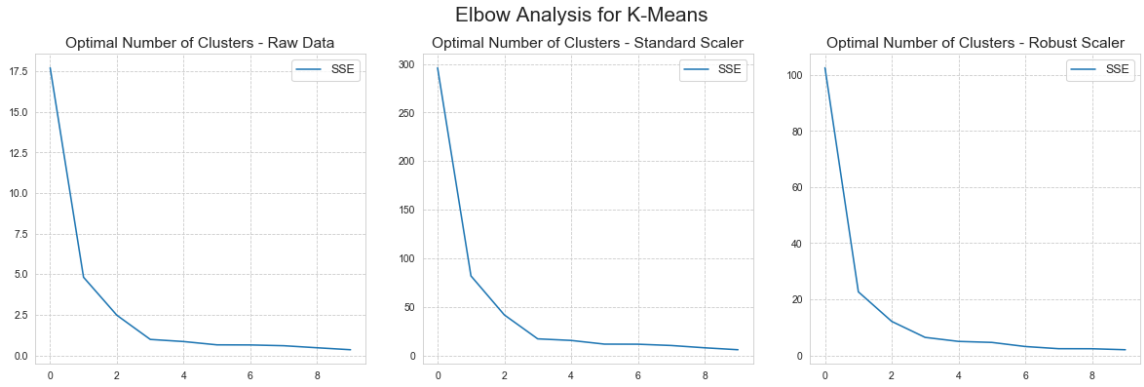


Figure 12: K-Means - Elbow Analysis

The three graphs seem to suggest that three clusters are the optimal levels, as the lines flatten out for higher numbers. As a result, the K-Means algorithm is fitted using three clusters as input.

7.2 BIC and AIC Method - Gaussian Mixture

In this section, BIC and AIC metrics are computed for the Gaussian Mixture model using different numbers of clusters as input. The two statistics can provide different optimal results and, given the relatively small size of the dataset, BIC is preferred as it applies higher penalties to additional clusters. Fig. 13 reports the results. According to the BIC metric, the optimal number of clusters for the Gaussian Mixture model occurs at four. This number differs from the one selected for K-Means. However, the goal of the analysis is not to define all the investment styles but to find the one that can be labelled as passive/closet tracker. As long as the model can outline the ETFs and Passive Funds cluster correctly, the number of the remaining clusters/styles is relatively irrelevant. As a result, the Gaussian Mixture model is fitted using four clusters as input.

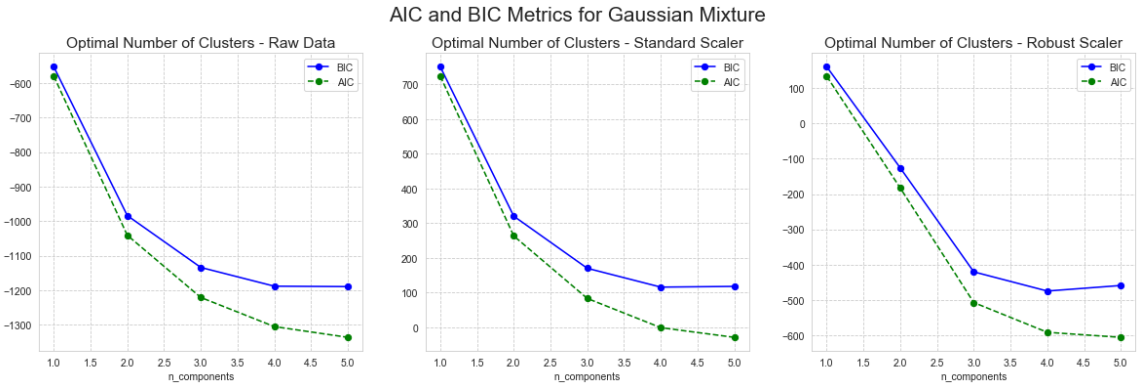


Figure 13: Gaussian Mixture - AIC and BIC metrics

8 Cluster Analysis

Based on the preceding analysis, a K-Means Model with three clusters and a Gaussian Mixture Model with four clusters are fitted on the data. The graphs represented in this section, which shows the results of the clustering process, should be compared with the Fig. 10 and Fig. 11 from Section 6 to locate the cluster that incorporates the ETFs straight away. The models are trained on the data scaled using the Standard-Scaler and the Robust-Scaler.

The results obtained on the Standard-Scaled data are presented in Fig. 15 and Fig. 14.

Figure 14: K-Means - Clusters on the Standard-Scaled Data

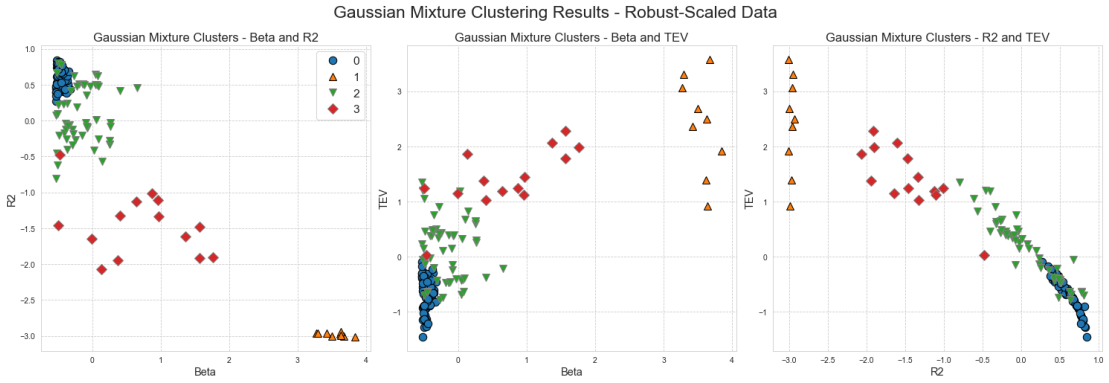


Figure 15: Gaussian Mixture - Clusters on the Standard-Scaled Data

By comparing the graphs above with those in Section 6, the cluster that includes the ETFs/-Passive Funds is the n.1 among those identified by the K-Means method and the n.0 among those spotted by the Gaussian Mixture Model. The same graphs are produced when the models are

fitted on the Robust-Scaled data and reported in Fig. 17 and Fig. 16.

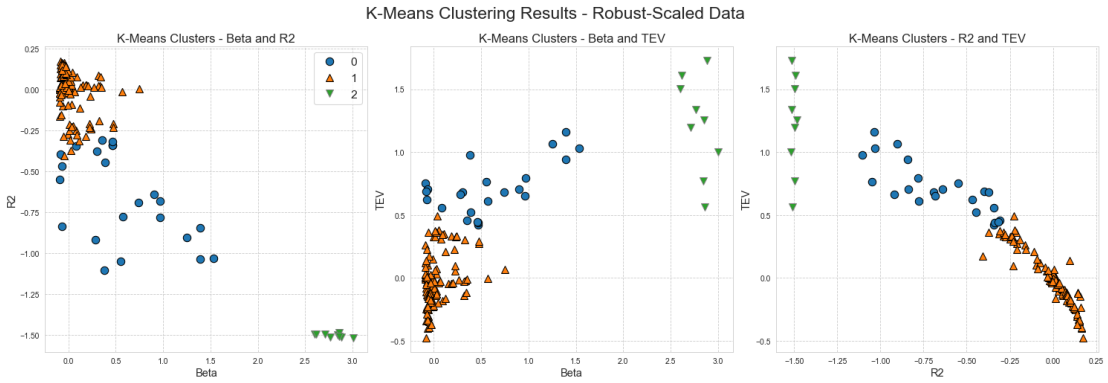


Figure 16: K-Means - Clusters on the Robust-Scaled Data

Figure 17: Gaussian Mixture - Clusters on the Robust-Scaled Data

The results from the Robust-Scaled data using the Gaussian Mixture model are only apparently different. The model uses different labels to identify the same cluster: the former cluster n.0 is now labelled as n.2, but the included instances seem to overlap. With the only irrelevant exception on the different labels, the clusters identified on the Robust-scaled data seem to not differ from those spotted on the Standard-Scaled Data. The same does not apply for the clusters established by the K-Means model. In this case, switching from the standard-scaled data to the robust-scaled means a significant shift in the boundaries. The cluster n.1 derived from the K-Means model covers the area of the graph where ETFs are. However, its boundaries fluctuate significantly according to the data pre-processing procedure applied. Based on that, it can be inferred that the four clusters selected with the Gaussian Mixtures are more stable and better defined than the three defined by the K-Means model. A further confirmation arrives from the comparison of the Silhouette diagrams of the two set of clusters, see Fig. 18 and Fig. 19. In the diagram, each knife-shaped form represents a cluster, with the height indicating the number of instances and the width the Silhouette coefficient. The dashed vertical line represents the Silhouette score for each number of cluster. When most of the instance or the whole cluster is below this level, it means that the cluster itself is not well defined. As a result, the three K-Means clusters seem relatively worse than those defined with the Gaussian Mixture model. The Silhouette analysis have been extended to cover the outputs from different numbers of clusters. From these graphs, it can be inferred that K-Means clusters start to be well defined when they are set at five. However, this is not consistent with the information from the Elbow analysis. On the contrary, the Gaussian Mixture model seems to deliver more robust and consistent results. The four clusters derived from the Gaussian Mixture are better defined than those in K-means and adding further clusters does not improve the clusters' shapes, consistently with the silhouette analysis. Based on this evidence, it is fair to state that the clusters produced by the Gaussian Mixture are better suited for identifying the potential closet tracker funds.

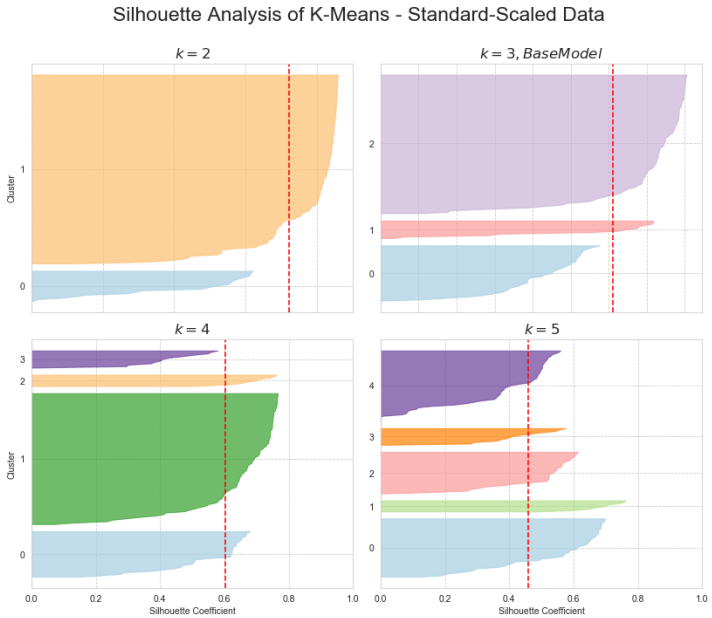


Figure 18: Silhouette Diagram - K-Means

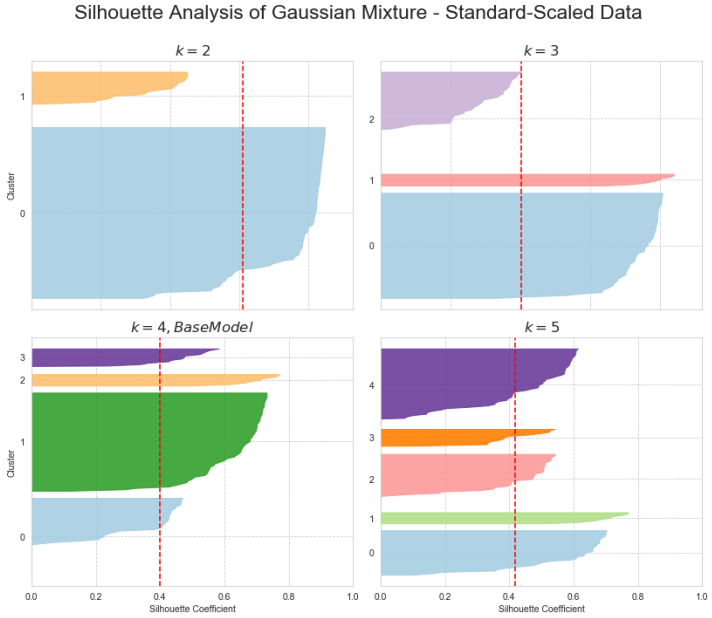


Figure 19: Silhouette Diagram - Gaussian Mixtures

9 Analysis of the Passive Cluster

Before starting to analyze the cluster, the results produced by the Gaussian Mixture model on the standard and robust scaled data are compared. The goal is to confirm that the funds identified for further analysis is the same in both clusters. A simple diagnostic on the two clusters produces the following results:

```
Verify that the Passive Fund cluster, built using Gaussian Mixture model on Standard-Scaled
and Robust-Scaled data, aggregates the same set of Funds:
 True

List of passive funds candidates:
 ['AEEI2EC LX Equity' 'CAIXEIA LX Equity' 'AXWEOFC LX Equity'
 'AXWECEI LX Equity' 'ETDD FP Equity' 'CSEMICE LX Equity'
 'CSIEEDE LX Equity' 'PAMEULF BB Equity' 'STHOPMI FP Equity'
 'FIDFEBI LX Equity' 'SGEURIE LX Equity' 'METVSRI FP Equity'
 'NATISRI LX Equity' 'INGEHDI LX Equity' 'PIPTFEI LX Equity'
 'SSEIIEU LX Equity' 'TEMGRIA LX Equity' 'UBSEITL LX Equity'
 'VANESII ID Equity']

Number of passive funds candidates:
 19

Numebr of the unique ETFs in the Cluster derived from Standard Scaled data:
 27

Numebr of the unique ETFs in the Cluster derived from Robust Scaled data:
 27
```

Figure 20: Diagnostic on 'Passive' Cluster

From the figures reported, it can be noted that the cluster embeds the vast majority of the ETFs used for the analysis, 27 out of 31. On the top of that, the cluster also aggregates 19 funds. Among these funds, the focus is on those that have a fee's profile inconsistent with their passive nature. This approach is consistent with CBI's "Value for Money" strategy in analysing a fund's behaviour. Coherently, the list of 19 funds is filtered to eliminate those with a fee profile consistent with their passive nature, using 60bps as a threshold. After this step, the number of potential closet trackers shrinks from 19 to 12. For these 12 funds, the procedure identifies 24 matches with indices where they display features consistent with a "closet tracker" behaviour. Moreover, all these funds require the investors to pay a level of fees not consistent with their investment approach. These are the funds that should fall under the regulatory scrutiny to verify if the investment strategy is properly disclosed, and the investors are correctly informed. The full list of matches between funds and indices are reported in Fig. 21

In the table, the funds are sorted from high to low according to their management fees. In the last part of the paper, the first three funds of the list are explored in greater details to verify that the insights generated by this procedure are reliable. For each fund, a visual analysis against the selected indices is performed along with a more detailed review of the regression used to produce the fund metrics.

	Cluster Class	Comparison	Fund or ETF	Total Return	Fees	Fund Code
0	0	Fund:AEEI2EC LX Equity, Index:NDDLEMU Index	Fund	1	2.5	AEEI2EC LX Equity
1	0	Fund:AEEI2EC LX Equity, Index:S&P_Euro_75	Fund	1	2.5	AEEI2EC LX Equity
2	0	Fund:FIDFEBI LX Equity, Index:NE700862 Index	Fund	0	1.92	FIDFEBI LX Equity
3	0	Fund:FIDFEBI LX Equity, Index:MXEM Index	Fund	0	1.92	FIDFEBI LX Equity
4	0	Fund:FIDFEBI LX Equity, Index:SXXT Index	Fund	0	1.92	FIDFEBI LX Equity
5	0	Fund:UBSEITL LX Equity, Index:S&P_EURO_PLUS	Fund	1	1.87	UBSEITL LX Equity
6	0	Fund:UBSEITL LX Equity, Index:NDDLE15 Index	Fund	1	1.87	UBSEITL LX Equity
7	0	Fund:METVSRI FP Equity, Index:SXXT Index	Fund	0	1.5	METVSRI FP Equity
8	0	Fund:METVSRI FP Equity, Index:LCXE Index	Fund	0	1.5	METVSRI FP Equity
9	0	Fund:METVSRI FP Equity, Index:SLVT Index	Fund	0	1.5	METVSRI FP Equity
10	0	Fund:STHOPMI FP Equity, Index:NDDLE15 Index	Fund	0	1.15	STHOPMI FP Equity
11	0	Fund:AXWEOFC LX Equity, Index:NDDLE15 Index	Fund	0	1	AXWEOFC LX Equity
12	0	Fund:TEMGRIA LX Equity, Index:NE700862 Index	Fund	0	0.93	TEMGRIA LX Equity
13	0	Fund:TEMGRIA LX Equity, Index:MXEM Index	Fund	0	0.93	TEMGRIA LX Equity
14	0	Fund:PAMEULF BB Equity, Index:S&P_EURO_PLUS	Fund	0	0.92	PAMEULF BB Equity
15	0	Fund:PAMEULF BB Equity, Index:E300 Index	Fund	0	0.92	PAMEULF BB Equity
16	0	Fund:NATISRI LX Equity, Index:NDDLE15 Index	Fund	0	0.9	NATISRI LX Equity
17	0	Fund:INGEHDI LX Equity, Index:FTEFC1 Index	Fund	0	0.81	INGEHDI LX Equity
18	0	Fund:INGEHDI LX Equity, Index:NDDLEMU Index	Fund	0	0.81	INGEHDI LX Equity
19	0	Fund:INGEHDI LX Equity, Index:MXEM Index	Fund	0	0.81	INGEHDI LX Equity
20	0	Fund:AXWECEI LX Equity, Index:SXXT Index	Fund	0	0.76	AXWECEI LX Equity
21	0	Fund:AXWECEI LX Equity, Index:MSER Index	Fund	0	0.76	AXWECEI LX Equity
22	0	Fund:CSEMICE LX Equity, Index:NDDLEMU Index	Fund	0	0.65	CSEMICE LX Equity
23	0	Fund:CSEMICE LX Equity, Index:MXEM Index	Fund	0	0.65	CSEMICE LX Equity
24	0	Fund:CSEMICE LX Equity, Index:FTEFC1 Index	Fund	0	0.65	CSEMICE LX Equity

Figure 21: Fund/Index matches with Passive characteristics

9.1 Potential Closet Tracker n.1: AEEI2EC LX:

This is a Luxembourg domiciled fund. According to the Fund's Key Investor Information Document (KIID): "*The Fund's investment objective is to provide capital growth and income over the long-term, by investing primarily in units or shares of UCITS or other UCIs, providing exposure to the European equity market. To achieve its objective, the Fund invests primarily in a diversified portfolio of equity mutual funds, managed by leading international asset managers, which mainly invest in European equities. The Fund is actively managed.*" ("underline" added by the Author, (link)). This is an actively managed fund of funds and, according to the data downloaded from Bloomberg, the ongoing management fees are 2.5%. However, the analysis suggests that this fund might be classified as a closet tracker against two indices: NDDLEMU and the S&P Euro 75. As initial pieces of evidence, a graphical comparison of the funds daily and cumulated returns is presented (Fig.23 and Fig.24).

These graphs seems to confirm that over the last year the fund's performances have been remarkably similar to those of two identified indices. In the scatter plot of the daily returns, most of the points lie in the proximity of the identity line. The regressions of the fund's daily returns on the two indices confirm that the Beta is quite close to one and the R^2 in the range of 0.9. The regressions also show that the constant is very close to zero, at -0.23 and -0.38 basis points.

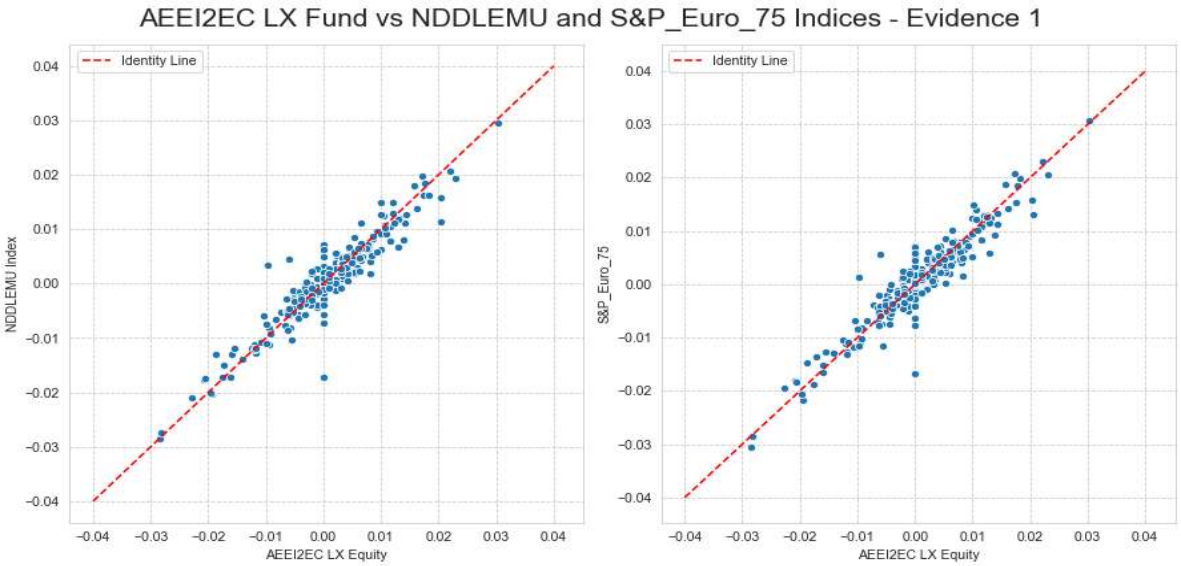


Fig. 22 - Daily Returns Vs Indices

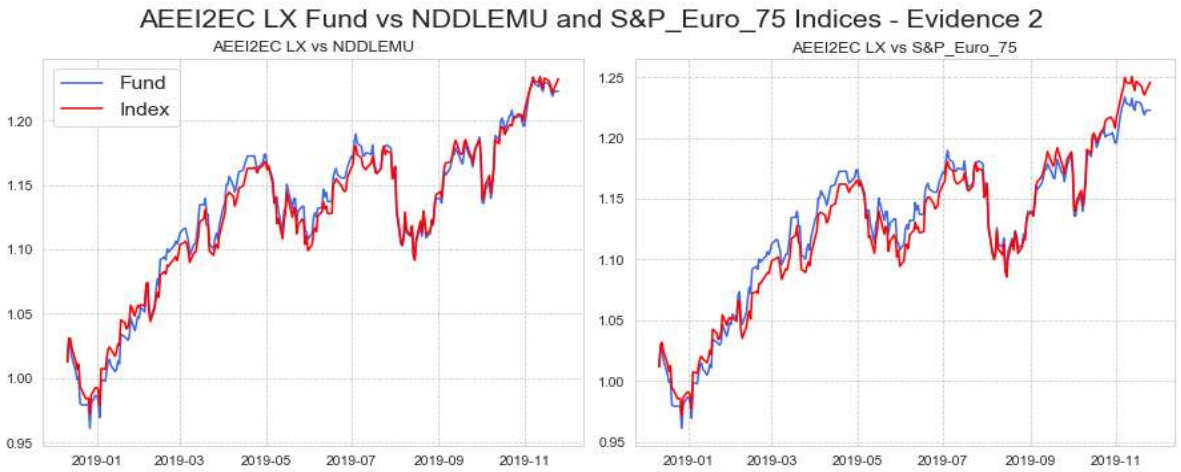


Fig. 24 - Cumulated Returns Vs Indices

Dep. Variable:	y		R-squared:		0.890	
Model:	OLS		Adj. R-squared:		0.890	
Method:	Least Squares		F-statistic:		2016.	
Date:	Sat, 07 Dec 2019		Prob (F-statistic):		4.39e-121	
Time:	14:56:50		Log-Likelihood:		1114.8	
No. Observations:	250		AIC:		-2226.	
Df Residuals:	248		BIC:		-2219.	
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.293e-05	0.000	-0.128	0.898	-0.000	0.000
x1	0.9936	0.022	44.897	0.000	0.950	1.037

Fig. 25 - OLS Regression AEEI2EC vs NDDLEMU

Dep. Variable:	y		R-squared:		0.893	
Model:	OLS		Adj. R-squared:		0.893	
Method:	Least Squares		F-statistic:		2078.	
Date:	Sat, 07 Dec 2019		Prob (F-statistic):		1.55e-122	
Time:	14:56:50		Log-Likelihood:		1118.2	
No. Observations:	250		AIC:		-2232.	
Df Residuals:	248		BIC:		-2225.	
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.843e-05	0.000	-0.218	0.828	-0.000	0.000
x1	0.9614	0.021	45.581	0.000	0.920	1.003

Fig. 26 - OLS Regression AEEI2EC vs S&P Euro 75 Index

9.2 Potential Closet Tracker n.2: FIDFEBI LX:

This is a Luxembourg domiciled fund. According to the KIID, the Fund "*Aims to provide long-term capital growth with the level of income expected to be low. At least 70% invested in the shares of blue chip companies in countries that are members of the Economic Monetary Union, and at least 70% denominated in Euro. Currently, there are nineteen member countries but if other countries join in the future, then investment in these countries may also be considered for inclusion in the fund.*" The Fund "*Has the freedom to invest outside the fund's principal geographies, market sectors, industries or asset classes."* It "*may invest in assets directly or achieve exposure indirectly through other eligible means including derivatives. Can use derivatives with the aim of risk or cost reduction or to generate additional capital or income, including for investment purposes, in line with the fund's risk profile. The fund has discretion in its choices of investments within its objectives and policies.*" (link). The running management fees are 1.92%, according to Bloomberg. The document does not specify if the fund is active or passive. However, the KIID's description is more consistent with an actively managed fund. The analysis suggest that this fund might be classified as a closet tracker against three indices: NE700862, MXEM and the SXXT. As initial evidence, a graphical comparison of the funds' daily and cumulated returns is analysed (Fig.27 and Fig.28).

These graphs seems to confirm that over the last year the fund's performances have been remarkably similar to at least two of the three selected indices. Indeed, the MXEM index seems to lag significantly behind when the cumulated performance is considered. In the scatter plot of the daily returns, most of the points lie in the proximity of the identity line. The regressions of the fund's daily returns on the two indices confirm that the Beta is quite close to one and the R^2 in the range of 0.9. However, some critical insight comes from the review of the fund against the MXEM Index. Despite having all the main statistic in line with those observed against the other two indices, the fund seems to outperform significantly against the MXEM. Additional information on this score is discussed in the analysis of the next fund. With regards to the other two Indices, the fund seems to track their performances quite closely.

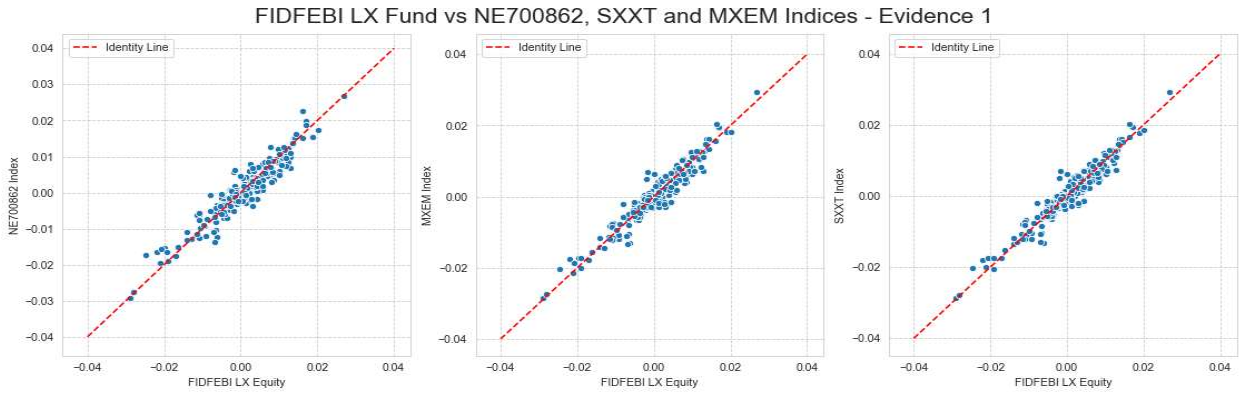


Fig. 27 - Daily Returns Vs Indices

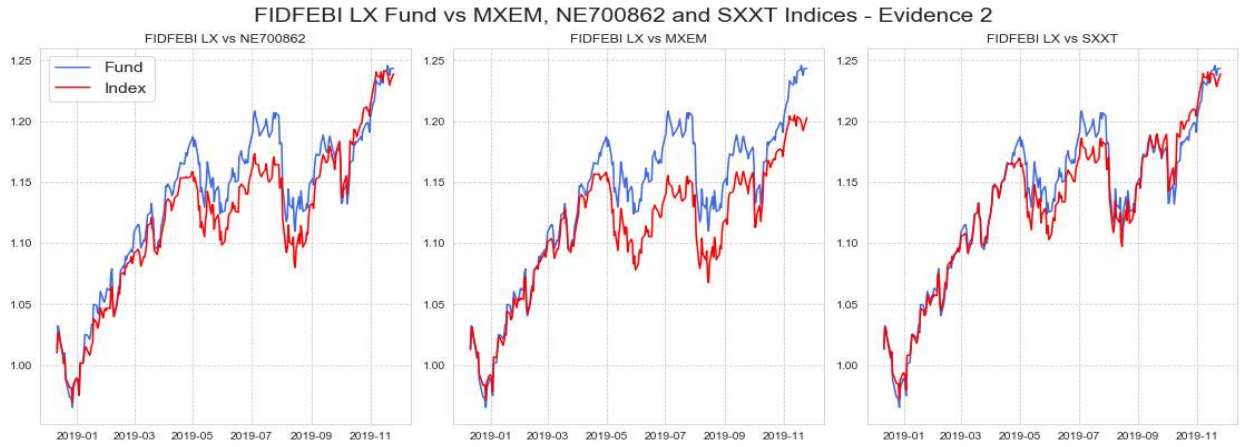


Fig. 28 - Cumulated Returns Vs Indices

Dep. Variable:	y	R-squared:	0.899
Model:	OLS	Adj. R-squared:	0.899
Method:	Least Squares	F-statistic:	2206.
Date:	Sat, 07 Dec 2019	Prob (F-statistic):	1.98e-125
Time:	14:07:08	Log-Likelihood:	1133.3
No. Observations:	250	AIC:	-2263.
Df Residuals:	248	BIC:	-2256.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.659e-05	0.000	0.100	0.921	-0.000	0.000
x1	1.0012	0.021	46.967	0.000	0.959	1.043

Fig. 29 - OLS Regression FIDFEBI vs NDDLEMU

Dep. Variable:	y	R-squared:	0.926
Model:	OLS	Adj. R-squared:	0.925
Method:	Least Squares	F-statistic:	3094.
Date:	Sat, 07 Dec 2019	Prob (F-statistic):	4.58e-142
Time:	14:07:08	Log-Likelihood:	1171.9
No. Observations:	250	AIC:	-2340.
Df Residuals:	248	BIC:	-2333.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0002	0.000	1.071	0.285	-0.000	0.000
x1	0.9748	0.018	55.621	0.000	0.940	1.009

Fig. 30 - OLS Regression FIDFEBI vs NDDLEMU

Dep. Variable:	y	R-squared:	0.928
Model:	OLS	Adj. R-squared:	0.928
Method:	Least Squares	F-statistic:	3194.
Date:	Sat, 07 Dec 2019	Prob (F-statistic):	1.16e-143
Time:	14:07:08	Log-Likelihood:	1175.6
No. Observations:	250	AIC:	-2347.
Df Residuals:	248	BIC:	-2340.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.311e-05	0.000	0.236	0.814	-0.000	0.000
x1	0.9800	0.017	56.517	0.000	0.946	1.014

Fig. 31 - OLS Regression FIDFEBI vs NDDLEMU

9.3 Potential Closet Tracker n.3: UBSEITL LX

This is a Luxembourg domiciled fund. According to the Fund's KIID: " *The investment fund invests primarily in the shares of Eurozone companies. The portfolio is focused primarily on securities from large companies, though it is strategically supplemented by securities from small and medium-sized companies. Working on the basis of well-grounded analyses by our local investment specialists, the fund manager combines carefully selected equities of various companies from various countries and sectors with the objective of exploiting interesting return opportunities while keeping the level of risk under control.*" (link). The running management fees are 1.87%, according to Bloomberg. The document does not specify if the fund is active or passive. However, the KIID's description is more consistent with an actively managed fund. The analysis suggest that this fund might be classified as a closet tracker against two indices: the S&P Euro Plus and the NDDLE15. As initial pieces of evidence, a graphical comparison of the funds daily and cumulated returns is analysed (Fig.32 and Fig.33).

The graphs related to the cumulated performances show again the attributes highlighted when analysing the FIDFEBI LX fund against the MXEM Index. Despite all the main behavioural metrics suggesting that the fund UBSEITL LX might passively track the NDDLE15 and the S&P Euro Plus indices, the performances between the fund and the indices over the last year are quite different. For instance, against the S&P Euro Plus the fund under-performs by about 7.5%, while versus the NDDLE15 the under-performance is in the range of 5%. Equally important, the two under-performances are gradually accumulated over the year, and not due to one-off effect. What seems to emerge here is a **shortcoming of the metrics used in the analysis**. This topic is going to be discussed in the next and final paragraph.

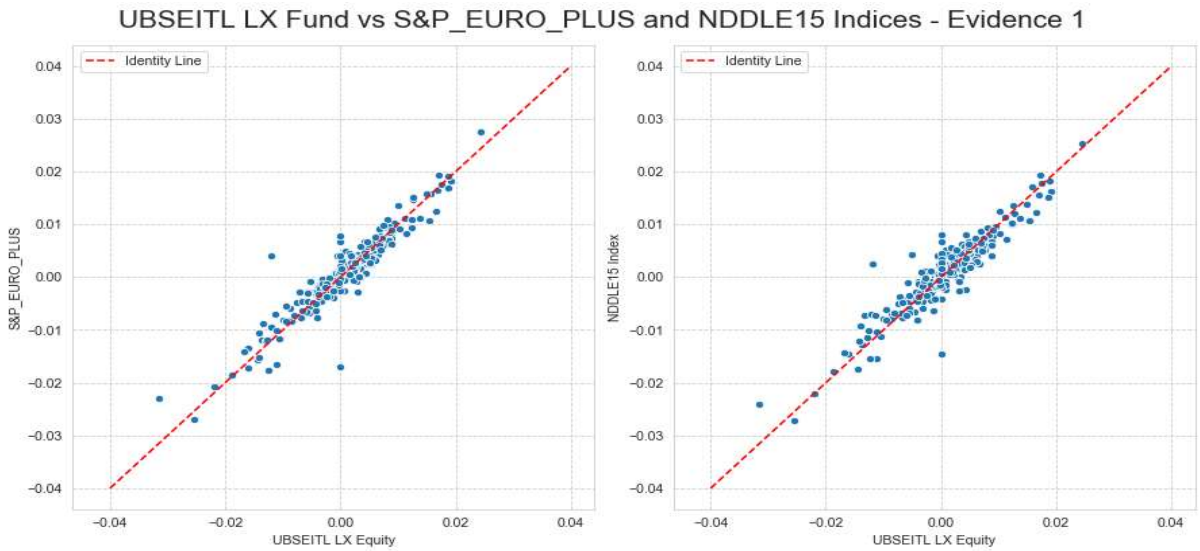


Fig. 32 - Daily Returns Vs Indices

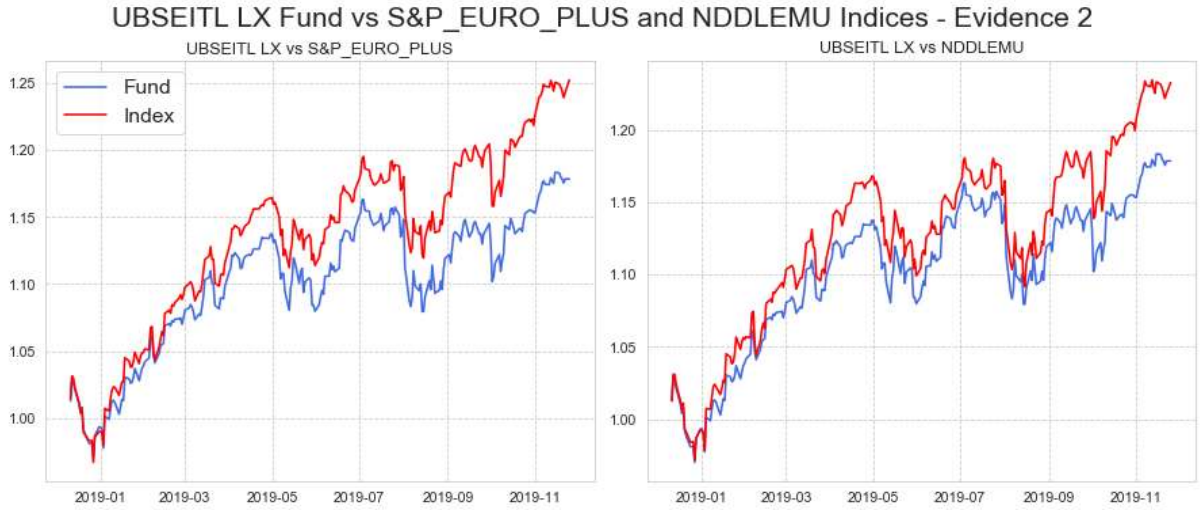


Fig. 34 - Cumulated Returns Vs Indices

Dep. Variable:	y	R-squared:	0.867
Model:	OLS	Adj. R-squared:	0.866
Method:	Least Squares	F-statistic:	1615.
Date:	Sat, 07 Dec 2019	Prob (F-statistic):	1.36e-110
Time:	13:05:41	Log-Likelihood:	1090.5
No. Observations:	250	AIC:	-2177.
Df Residuals:	248	BIC:	-2170.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0001	0.000	-0.749	0.454	-0.001	0.000
x1	1.0655	0.027	40.191	0.000	1.013	1.118

Fig. 35 - OLS Regression UBSEITL vs NDDLEMU

Dep. Variable:	y	R-squared:	0.890
Model:	OLS	Adj. R-squared:	0.890
Method:	Least Squares	F-statistic:	2016.
Date:	Sat, 07 Dec 2019	Prob (F-statistic):	4.39e-121
Time:	13:05:41	Log-Likelihood:	1114.8
No. Observations:	250	AIC:	-2226.
Df Residuals:	248	BIC:	-2219.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.293e-05	0.000	-0.128	0.898	-0.000	0.000
x1	0.9936	0.022	44.897	0.000	0.950	1.037

Fig. 36 - OLS Regression UBSEITL vs S&P Euro Plus Index

10 Conclusions

In this paper a procedure for identifying closet tracker funds has been developed and tested. This critical element of this procedure is an unsupervised machine learning clustering algorithm that tackles the goal of clustering a set of funds according to their management style. A control group of known passive funds (ETFs) has been added to the set of funds, and this trick should allow to single out the broader group of potential closet tracker once the funds are clustered. A Gaussian Mixture model delivers the most promising results in terms of funds clustering according to their management styles. Most of the funds from the control group are grouped in the same cluster as expected. However, by merely comparing the performances of three potential closet tracker against their candidate indices, a severe shortcoming in the metrics used to summarize funds' style has emerged. **The metrics used to outline the funds' investment approach seems to not take in due account the positive or negative alpha generated in the funds.** Significant over or under-performances can be accumulated over the period under review without being adequately captured by the statistics used in the analysis (Beta, TEV and R^2). This paper offers some evidence in favour of the opportunity to integrate them with additional metrics to better describe a fund's behaviour. Since these are the very metrics used by the Irish regulator to oversight investment funds, it should evaluate the possibility to add additional metrics relying on the experience from other countries or supranational entities such as the ESMA. For instance, the latter included in its study also the active equity component for each fund against its benchmark, while the Danish regulator used the asset turnover. It must be stressed that the procedure outlined in this paper and its underlying logic remain valid and are not affected by the set of metrics selected. On the contrary, the increase in the amount and quality of the data is very likely to increase the reliability of the output. To provide a measure of the information that are not captured by the current metrics, they have been applied on a simulated fund and its benchmark. This fund (Fig. 22) shows a beta of 0.995, a R2 of 0.95 and a TEV of 0.246 against the synthetic benchmark. All the metrics would qualify this fund as a closet tracker of the index. However, the fund manages to outperform the index by more than 10% contradicting the preceding statement. **When a fund has a sizeable part of its performances driven by a component that is unrelated to the market directions, these three metrics alone fail.**

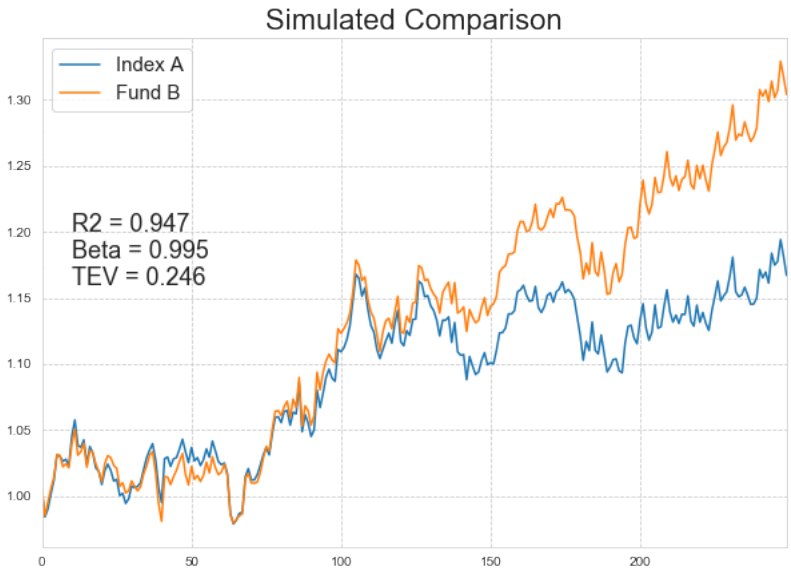


Figure 22: Fund/Index Simulated Case Study

References

[Aut19] Financial Conduct Authority. Asset management market study - further remedies - policy statement - ps19/4. *Policy Statement*, 2019.

[oI19] Central Bank of Ireland. Thematic Review of Closet Indexing. *Central Bank of Ireland*, 2019.

[SA19] European Securities and Markets Authority. Supervisory work on potential closet index tracking. *Statement*, 2019.