SHORT STUDY SERIES

Nº 47, julho 2025

FINANÇAS

FGV INVEST

SECTOR ANALYSIS ON THE S&P 500: A PCA APPROACH

ALEXIS PETRI M COSTA

FGV EESP
CENTRO DE TREINAMENTO EM PROCESSOS DE INVESTIMENTO

Sector Analysis on the S&P 500: A PCA Approach

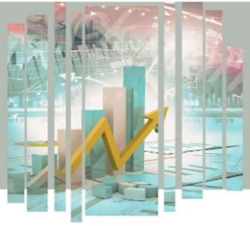
Alexis Petri M Costa[1]

ABSTRACT

This work proposes a simple exercise of reclassifying the S&P 500 stocks into new clusters based solely on price comovements using Principal Component Analysis and Hierarchical Clustering. It aims to discuss the rationale, challenges, and results of clustering stocks based on the similarity of their weight profile on the eigenvectors. As a result, the author identified significant misalignments between each sector's components and their newfound peers. This, in turn, raises questions about the efficacy of incorporating single sector ETFs into a portfolio, suggesting that some of these investment vehicles are neither diversified nor "pure play" instruments.

Key-words: sector-specific exchange-traded funds, Principal Component Analysis, Hierarchical Clustering, Global Industry Classification Standard.

Alexis Petri Magalhaes Costa has a M.Sc. from Fundação Getúlio Vargas – Escola de Economia de São Paulo (FGV-EESP) and a B.Eng. from Universidade de São Paulo – Escola Politécnica (USP-POLI).



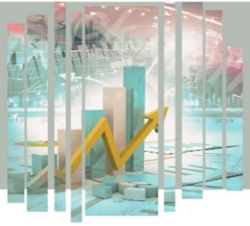
1. INTRODUCTION

The S&P 500 is the U.S.'s main stock market index, and arguably the most followed index in the world. The index is composed of 503 stocks, as of 2025, representing the largest 500 companies in the United States, by market capitalization, with a layer of discretion from the Index Committee and auxiliary eligibility criteria. It has been balanced at least quarterly since its creation, in 1957.

Since 1999, the S&P 500 stocks have been classified into sectors (and finer subclasses) through the GICS (Global Industry Classification Standard), in a partnership between MSCI and S&P Global. For over 25 years, they have disclosed the individual components of the traditional sectors (currently 11), resulting from the analysis of each company on a set of qualitative and quantitative criteria, but having revenues as the centerpiece.

These sectors, since inception, have been available for trading through ETFs by State Street Global Advisors (Select Sector SPDR ETF), which have had frequent recalibration in tandem with the revisions of the GICS. These ETFs complement the toolkit for portfolio construction, which is explicitly mentioned as part of the GICS's use cases. MSCI cites "contribution to portfolio performance" and "sector rotation strategies", both of which are readily implemented through ETFs. For completeness, however, MSCI does not limit itself to ETF-related use cases, adding "support for peer comparison analysis" and "company research within a specific industry" to the list.

It stands to reason that the choice of incorporating a sector-specific ETF into one's portfolio would follow a rationale that is supported by statistical evidence. MSCI explicitly states that the GICS does not follow a "purely statistical approach, which is solely financial market-based and backward looking, using past returns." (MSCI Inc., 2024). This work aims to highlight purely that statistical component, providing data to defend or reject the usage of GICS (and thus, sector-specific ETFs) for portfolio construction. Principal Component Analysis (PCA) reveals mixed evidence regarding both the distinctiveness of each sector and the internal consistency of their constituent stocks.



The exercise specified in this article is not presented as the definitive way to cluster stocks. Instead, the method herein is intended as a modular component within a larger exploratory process, and its conclusions are subject to refinement as new approaches are proposed. Further dialogue is welcome.

For a deeper dive on clustering techniques, one may refer to Salvati (2024), and Rotondi and Russo (2024), the latter using PCA for stock pairs trading strategies. Similar strategies may rely on different tools, such as cointegration-based or copula-based pairing (Rad, Low and Faff, 2016). The model described in this study can be redesigned to apply these tools, among others, and compare results.

2. MODEL OVERVIEW

The top 243 stocks were drawn from the S&P 500 to represent over 90% of the index by market capitalization, as of May 2025. Historical prices were collected from January 2000 (or later, depending on when each company went public) to May 2025. This time period was further split into subsets of roughly 8, 6 and 3.5 years (the time axis was split into three, four, and seven equal parts), resulting in 14 subsets. Each subset contained only the stocks that had a complete data series of prices for its time frame. This sectioning helps mitigate overfitting, taking into consideration PCA's sensitivity to noise. The S&P 500 ETF from State Street, SPY, was also included in the dataset, as a benchmark for validation purposes.

Principal Component Analysis was performed on the daily log-returns of all stocks (plus SPY) in each time-defined subset. For each of the 14 subsets, stocks were clustered based on Euclidean distances computed across their first eleven principal components, equal-weighted, using Ward's method. Subsequently, a symmetric relationship matrix was created (244x244), where each element represented a distance measure between two stocks. A consolidated relationship matrix was then created, where each element was the harmonic mean of the 14 distances available for each stock pair.

Deviating from Sarmento and Horta (2020), the first eleven principal components were considered (totaling roughly 97.5% of the explained variance)

instead of five (95%), as this iteration generated more stable clusters across the 14 subsets.

Finally, a new hierarchical clustering algorithm was applied to the consolidated relationship matrix, using the same method and parameters. Finally, the entire algorithm was rerun including the eleven Select Sector SPDR ETFs, with the goal of assessing where each ETF mapped onto the newly created groups.

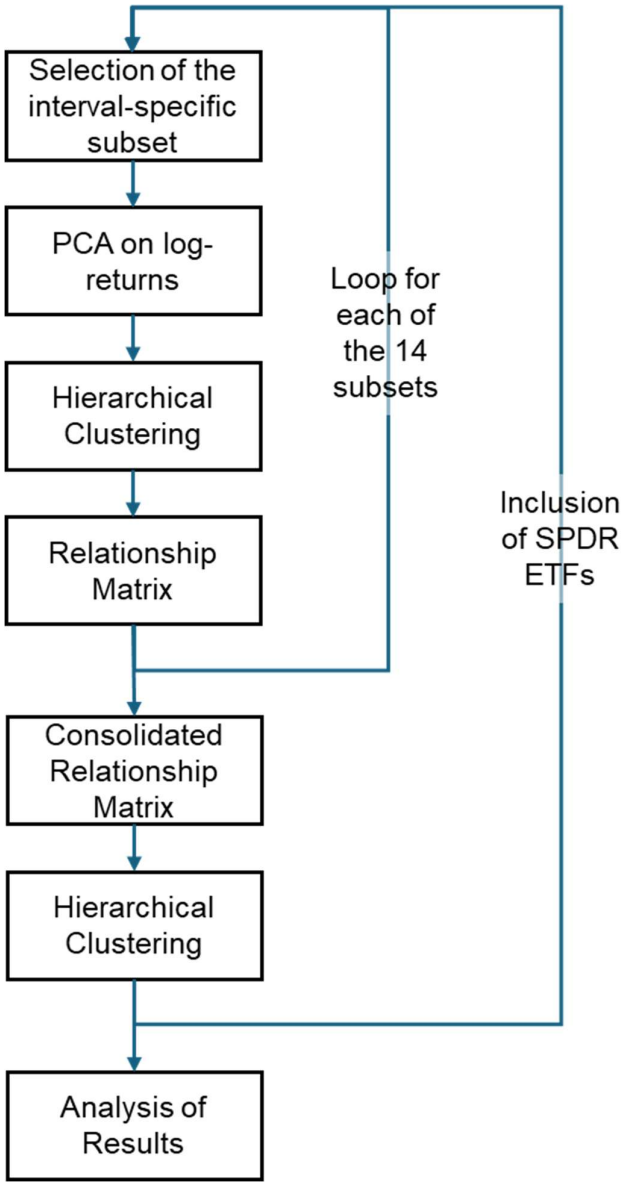


Fig 1. Summary of the clustering algorithm.

3. RESULTS

For simplicity, the number of clusters was chosen from the elbow of the cluster-count curve relative to the cut-off parameter, yielding 19 clusters. The largest group had 24 stocks, all others ranging from 9 stocks (four clusters) to 17 (one cluster). The inclusion of the SPDR ETFs did not significantly impact this arrangement – only a handful of tickers were reassigned to new clusters. These changes are most likely distortions caused by the fact that these ETFs are linear combinations of other stocks.

Some clusters had a high alignment to the GICS sectors, such as Energy, Health Care, REIT and Utilities. Others were more dispersed. The figure below summarizes the findings in a table of stock count at each intersection of sectors and clusters.

Stock Count	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Financials	15	12	5			4	1	2	2										
Tech			4	4	14		5	2		10									
Indust.			3	5	1	1		5	2		16	4	6						
Cons.Disc.						3	1	2	7	1			2	6					
Comm.							4			2		1			5				
Materials								1				3	1				2		
Staples								1	1					5	8				
Health Care									3			1				24			
Energy																	15		
REIT						1												9	
Utilities					1			1											9

Fig 2. Number of stocks shared by each sector and cluster.

Figure 3 provides a similar view but weighted by market capitalization. Each cell displays the percentage of the S&P 500 total market capitalization instead of the number of stocks, adding up to 90% as that had been the cut-off level for stock selection.

% Weight	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Financials	5	4	1			1	0	0	2										
Tech			1	0	11		1	1		15									
Indust.			0	0	0	1		1	0		3	1	1						
Cons.Disc.						0	2	0	1	4			0	2					
Comm.							8			0		0			2				
Materials								0				1	0				0		
Staples								0	0					3	2				
Health Care									1			0				7			
Energy																	3		
REIT						0												1	
Utilities					0			0											1

Fig 3. Percent market cap shared by each sector and cluster.

While the complete list of stocks is provided in the Annexes, some comments are worth highlighting.

a. Stocks within the Financials sector were distributed across three destinations in roughly equal proportions: most bank stocks flowed to cluster 1, insurance stocks to cluster 2, and the remaining third was absorbed into six newly formed groups, alongside companies classified by GICS as Industrials, Information Technology, or Consumer Discretionary.

b. Stocks within the IT sector were also partially scattered, with 72% of the selected stocks stabilizing in two main clusters. The first, cluster 5, was formed by companies in the semiconductors and hardware space. The second, cluster 10, was formed by large companies in the general IT space that cater to companies and individuals alike. These include AAPL and MSFT, ADBE and ADSK, ORCL and CSCO. This group also included AMZN, which distanced itself from GICS's Consumer Discretionary. EA and TTWO were, to some extent, exceptions – they are classified by the GICS as Communication Services.

c. Perhaps unsurprisingly to market participants, GOOG/GOOGL, NFLX and META, which are classified under the Communication Services sector, were grouped with some technology companies (cluster 7) that

operate mostly in the SaaS space or cybersecurity (CRM, NOW, FTNT, PAWN). Other five Communication Services companies were instead grouped with the largest cluster of Consumer Staples stocks. These were CHTR, CMSA, T, TMUS and VZ.

d. Cluster 3 contained stocks from the sectors of Financials, Technology and Industrials, and included names like SPGI, MSCI, ADP and FICO. The ETF for the S&P 500, SPY, was grouped in this cluster. SPY consistently displayed the highest PC1 weight, and the lowest weight (in absolute value) for all other components, adding further strength to the results.

e. Some groups were less intuitive, like cluster 6, of private equity (APO, BX, KKR) with auto manufacturers (GM, F). Worth noting that this grouping would not unwind with a different threshold value (e.g., the distances between the pairs APO-BX and GM-F are not significantly smaller than APO-GM and BX-F).

f. Cluster 17 incorporates all energy stocks, plus two stocks that are classified under the Materials sector. These are the miners NEM and FCX, raising the question of whether this association is the effect of randomness, or true similarities that drive stock comovements. It is undeniable, however, that *some* stock pairings will be spurious.

Finally, when incorporating the SPDR ETFs, the adherence (or lack thereof) to the GICS becomes evident. Figure 4 illustrates where (in which cluster) each ETF was grouped by the algorithm. The third column (%Sector in Cluster) shows how many stocks of a given sector fell into the same cluster as their respective ETF. Clusters that were heavily composed of stocks of the same sector attracted their respective ETF (100% of Energy stocks in cluster 17, 90% of REITs in cluster 18, 86% of Health Care stocks in cluster 16, 82% of Utility stocks in cluster 19). The table was mostly ordered by this column, while also trying to keep similarly themed clusters close together. With the sole purpose of facilitating the readability of the output, an attempt was made to describe the common themes of each cluster on the same table.

ETF	Definition	%Sector in Cluster	Cluster	Common theme(s)
XLE	Energy	100%	17	Energy
XLRE	REIT	90%	18	REITs
XLV	Health Care	86%	16	Health Care
XLU	Utilities	82%	19	Utilities
XLP	Staples	53%	15	Consumer Staples (Household products) and Broadband/Network Operators
-	-	-	14	Consumer Staples and Discretionary (mixed): Retailers
XLB	Materials	43%	12	Materials, Infrastructure and Environmental Services
XLI	Industrials	37%	11	Industrials: Defense, Engineering, Components and Rail
-	-	-	4	Industrials: Precision Tech and Electronics
XLF	Financials	37%	1	Financials: Banking
-	-	-	2	Financials: Insurance (P&C, Life and brokers) and Exchanges
XLK	Techonology	36%	5	Tech: Semiconductors and Hardware
-	-	-	10	Tech: Broad Consumer and Commercial Software/Hardware
-	-	-	3	Tech & Financials: Enterprise Software, Ratings
XLC	Communication	33%	7	Tech: Cyber, SaaS, Communications
XLY	Discretionary	32%	9	Cyclical Brands, Payments Infrastructure and Other Specialized Industrials
-	-	-	6	Private Equity and Selected Industrials (Automakers and Mobility-related)
-	-	-	8	Tech Disruptors and Others

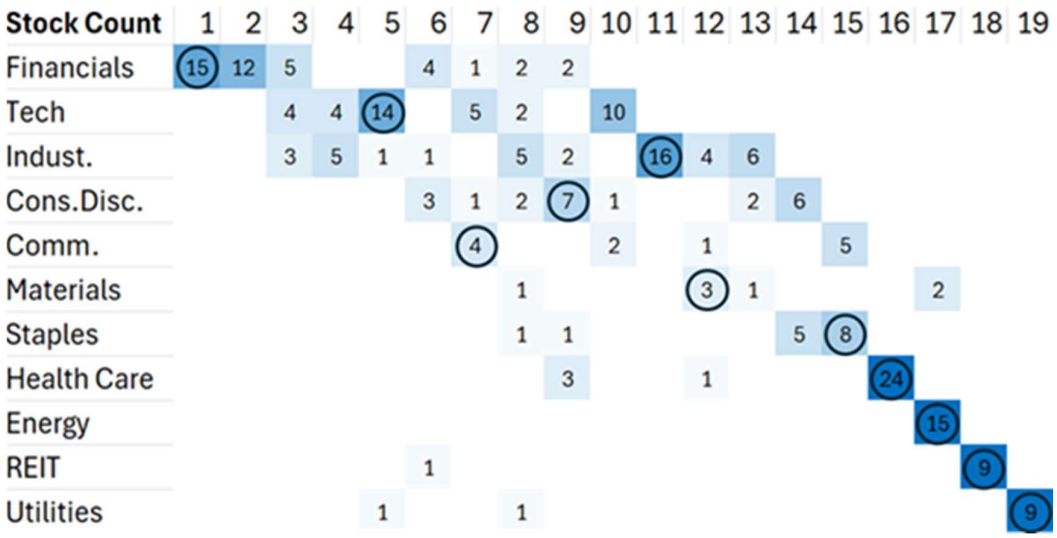


Fig 4. List of ETF per cluster, with Figure 2 revisited, overlayed with ETF assignments.

4. CONCLUSION AND POSSIBLE EXTENSIONS

Although it is not novel to advocate for parsimony in applying GICS classifications to portfolio construction, this work hopefully has provided an example of guidelines for a quantitative assessment. While some ETFs hold mostly stocks with distinct PCA profiles, others do not. Thus, adding these ETFs to one's portfolio may introduce unintended exposures or factor tilts. As a self-evident example, adding XLC (Communication Services) will cause some of XLP (Consumer Staples) elements to carry over, due to overlapping stock behavior.

Naturally, the algorithm showcased is only one of many ways this exercise can be done, nor is it without limitations. Adding more stocks, changing the time axis, parameters, or criteria, or further incorporating new tools could produce somewhat different results.

Hopefully this work served as a contribution to the discussion, and together with alternative models and extensions, paints a clearer picture of true distribution of outcomes from using GICS sector-specific ETFs.

5. REFERENCES

MSCI Inc. (2024), Global Industry Classification Standard (GICS) Methodology. https://www.msci.com/indexes/documents/methodology/0_MSCI_Global_Industry_Classification_Standard_GICS_Methodology_20240801.pdf

Rad, H., Low, R. K. Y., and Faff, R. (2016), The profitability of pairs trading strategies: distance, cointegration and copula methods. Quantitative Finance, 16(10), 1541-1558. https://doi.org/10.1080/14697688.2016.1164337

Rotondi, F., Russo, F., (2024), Machine Learning for Pairs Trading: a Clustering-based Approach. Working paper, Bocconi University. http://dx.doi.org/10.2139/ssrn.5080998

Salvati, F. (2024), Innovative Clustering Approach for S&P 500 Companies: Beyond Sector Boundaries and Cyclical – Non-Cyclical Portfolios. Master's thesis, Politecnico di Milano.

Sarmento, S. M., Horta, N. (2020). Enhancing a pairs trading strategy with the application of machine learning. Expert Systems with Applications, 158, 113490.

S&P Dow Jones Indices (2025), S&P U.S. Indices Methodology. https://www.spglobal.com/spdji/en/documents/methodologies/methodology-sp-us-indices.pdf

6. ANNEXES

Comprehensive list of stocks on each of the 19 clusters.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
AMP	AFL	ACN	AME	ADI	APO	CRM	ABNB	AXON	AAPL	CAT	APD	BA	AZO	CHTR	ABBV	BKR	AMT	AEP
AXP	AIG	ADP	APH	AMAT	BLK	FTNT	CARR	BKNG	ADBE	CMI	CPRT	FAST	COST	CL	ABT	COP	CCI	D
BAC	AJG	CTSH	GLW	AMD	BX	GOOG	CEG	CMG	ADSK	CSX	CTAS	FDX	HD	CMCSA	AMGN	CVX	DLR	DUK
BK	ALL	FI	JCI	ANET	CBRE	GOOGL	COIN	CVS	AMZN	DE	DHR	GWW	KR	KDP	BDX	EOG	EQIX	EXC
C	AON	FICO	LHX	AVGO	F	META	CRWD	HCA	CDNS	EMR	DIS	MAR	LOW	KMB	BMY	FANG	O	NEE
COF	BRK-B	FIS	PWR	DELL	GE	NFLX	CTVA	HLT	CSCO	ETN	ECL	RCL	ORLY	KO	BSX	FCX	PLD	PEG
GS	CB	MCO	ROP	HWM	GM	NOW	DASH	MA	EA	GD	LIN	SHW	PG	MDLZ	CI	HES	PSA	SO
JPM	CME	MSCI	TEL	INTC	GRMN	PANW	GEV	MCD	IBM	HON	RSG	TT	ROST	MO	COR	KMI	SPG	SRE
MET	MMC	MSI	URI	KLAC	KKR	PYPL	ICE	MNST	INTU	ITW	WM	UPS	TGT	PEP	ELV	MPC	WELL	XEL
MS	NDAQ	PAYX		LRCX		TSLA	KVUE	NKE	MSFT	LMT			TJX	PM	EW	NEM		
PNC	PGR	SPGI		MU		WDAY	OTIS	SBUX	ORCL	MMM			WMT	T	GILD	OKE		
SCHW	TRV	SPY		NVDA			PCAR	TDG	SNPS	NOC				TMUS	IDXX	OXY		
TFC		VRSK		NXPI			PLTR	TMO	TTWO	NSC				VZ	ISRG	PSX		
USB				QCOM			UBER	V		PH					JNJ	SLB		
WFC				TXN				YUM		RTX					LLY	VLO		
				VST						UNP					MCK	WMB		
															MDT	XOM		
															MRK			
															PFE			
															REGN			
															SYK			
															UNH			
															VRTX			
															ZTS			