AN ADAPTIVE HIERARCHICAL TREE-BASED CLUSTERING APPROACH TO OUTLIER DETECTION IN ETF-FOCUSED FINANCIAL TIME-SERIES

Shlok Mandloi, Aryaman Jalali, and Eugene Pinsky

Department of Computer Science, Metropolitan College,
1010 Commonwealth Avenue, Boston University, Boston, MA

ABSTRACT

This paper introduces an adaptive framework for detecting outliers in financial time-series data, focusing on Exchange-Traded Funds (ETFs). The method integrates hierarchical clustering and binary tree analysis to identify unique ETF patterns while isolating anomalies. Using the yfinance API, daily returns for nine ETFs and the S&P 500 index were collected over 24 years. Regression analysis removed market influence, producing residuals that highlight ETF-specific behavior. Hierarchical clustering was applied to these residuals annually, with dendrograms converted into binary trees. Outliers were detected as ETFs added last in clustering and as root nodes in the trees. Metrics like tree height, breadth, and cluster compactness captured temporal patterns and deviations. Experimental results demonstrate the framework's ability to detect anomalies during major market events, such as the 2008 financial crisis and the 2020 COVID-19 crash. This scalable and interpretable approach enhances anomaly detection in financial data analysis.

KEYWORD

Hierarchical Clustering, Outlier Detection, Financial Time-Series, Binary Tree Analysis, Anomaly Detection.

1. INTRODUCTION

The increasing complexity of financial markets necessitates robust and interpretable methods for analyzing large-scale time-series data. Exchange-Traded Funds (ETFs), representing portfolios of assets, provide a rich dataset for studying market behavior and identifying anomalies. However, isolating outlier behavior in ETFs is challenging due to broader market influences and the high dimensionality of financial data.

This study focuses on nine major sector-specific ETFs that comprise the S&P-500 index:

1. XLB: Materials
2. XLE: Energy
3. XLF: Finance
4. XLI: Industrials
5. XLK: Technology
6. XLP: Consumer Staples
7. XLU: Utilities
8. XLV: Healthcare
9. XLY: Consumer Discretionary

In addition, we consider the broad S&P-500 index. Investing in such index can be done more easily with the "SPY" exchange-traded fund.

This paper introduces a scalable framework for detecting outlier behavior in ETFs using hierarchical clustering and binary tree analysis. Daily returns of ETFs are regressed with the S&P 500 index, producing residuals that capture ETF-specific behavior. Hierarchical clustering is then applied to these residuals to uncover patterns independent of broader market trends. The resulting dendrograms, obtained by the hierarchical clustering of residuals, are converted into binary tree structures. This enables the identification of outliers, specifically ETFs added last in the clustering process and occupying root node positions in the trees. Metrics such as tree height, breadth, and compactness quantify temporal dynamics of ETFs behavior.

The proposed framework generalizes to high-dimensional time-series data, offering a computationally efficient method for outlier detection and pattern recognition. Its adaptability extends beyond finance, with potential applications in machine learning domains such as sensor data analysis, network anomaly detection, and time-series clustering. This research establishes a versatile approach for leveraging hierarchical relationships and tree-based metrics to uncover actionable insights from large-scale datasets.

2. METHODOLOGY

2.1. Data Collection

The study collected daily returns data for nine sector-specific ETFs and the S&P 500 index over a 24-year period (2001–2024) using the yfinance API. This data set provides the foundation for analyzing the behavior of ETFs and isolating sectoral trends. The S&P 500 returns were used as a market benchmark for regression analysis, ensuring that sector-specific characteristics could be captured through residuals.

Handling Missing Data and Trading Day Discrepancies: Given that ETFs may experience trading inconsistencies due to market holidays and half-days, all datasets were aligned to a common trading calendar based on the S&P 500 index. Missing data points were forward-filled using the last available trading day to ensure continuity. If an ETF exhibited excessive missing data (exceeding 5% of trading days in a given year), it was excluded from that year's analysis to prevent distortions in clustering results.

2.2. Residuals Analysis

A common and simple approach to analyze returns of stock S is to show the dependence of stock returns on the overall market and to apply to apply a linear regression. If we were to use daily returns and perform such an analysis for one year, then in such an approach, the independent variable is a (250-day) vector of daily S&P-500 (market) returns $R^* = (r_1^*, \dots, r_{250}^*)$ and the dependent variable is a (250-day) vector $R(S) = \left(r_1^{(S)}, \dots, r_{250}^{(S)}\right)$ of daily returns of security S.

$$\underset{\breve{R(S)}}{\left(r_1^{(S)}, \dots, r_{250}^{(S)}\right)} = \alpha + \beta \underset{\breve{R^*}}{(r_1^*, \dots, r_{250}^*)} + \underset{\breve{\epsilon}}{(\epsilon_1, \dots, \epsilon_{250})}$$

Here β represents the dependence of returns of S on the market and can be interpreted as the slope of the regression line. Alpha (α) represents indicates how an investment has performed

after accounting for the risk (i.e. excess return in relation to a benchmark, when adjusted for risk). It can be interpreted as the intercept of the regression line. Finally, the residuals $(\epsilon_1, \dots, \epsilon_{250})$ are uncorrelated to market returns and can be interpreted as a idiosyncratic source of risk and return driven by skill of company management.

One of the problems with such an approach is that securities are correlated.

If we were to simply use the standard Euclidean metric $d(S_1, S_2)$ between return vectors $R(S_1)$ and $R(S_2)$ for securities S_1 and S_2 then this distance can be made arbitrarily even if the return pattern is essentially identical. For example, consider a simple case when daily returns of S_1 for each day are just a multiple C of the corresponding daily return of S_2. Then taking C to be large enough, we can have any value for the distance $d(S_1, S_2)$ whereas the pattern of daily price directions of both securities is identical.

Our proposal is to focus on residual vectors other than the return vectors. It also allows us to ignore market-wide influences on ETF returns, enabling a more precise analysis of sector-specific behavior. This approach involved the following steps:

- Independent Variable: compute the annual vector of daily S&P 500 returns, serving as the benchmark to account for market-wide movements.
- Dependent Variable: compute the annual vectors of daily ETF returns representing the performance of specific ETFs relative to the market.
- Residuals: Compute the annual vectors of residuals ϵ of daily residuals. as the difference between observed ETF returns and the predicted values from the regression line,

The residuals capture ETF-specific behavior by isolating deviations from market trends. This process was critical for identifying anomalies and sector-specific patterns, ensuring that the clustering analysis focused on unique ETF dynamics rather than broader market influences. The use of residuals allows for:

- Enhanced Analysis: By removing market-wide effects, residuals help uncover sector-specific responses to external events.
- Accurate Clustering: The residual-based data ensures that hierarchical clustering reflects intrinsic ETF behavior rather than market-driven correlations.

2.3. Hierarchical Clustering

The residuals obtained from regression analysis were subjected to hierarchical clustering to identify relationships among ETFs. This method was chosen for its ability to reveal structural patterns over time without requiring a predefined number of clusters. To ensure robustness, multiple linkage methods—single, complete, and average—were evaluated. Ward's method was ultimately selected, as it minimizes intra-cluster variance while maintaining well-balanced clusters. A sensitivity analysis confirmed that:

- Clustering: Hierarchical clustering was performed to group ETFs based on their residuals.
- Linkage: Various linkage techniques were tested. Single linkage produced elongated, chain-like clusters, reducing interpretability, while complete linkage resulted in well-separated clusters but introduced distortions in compactness. Average linkage provided moderate compactness but was outperformed by Ward's method , which consistently produced the most stable and interpretable clustering results.

- Visualization: Dendrograms were generated to illustrate clustering structures and highlight sectoral divergence.

While hierarchical clustering was the preferred approach due to its interpretability and ability to track sectoral shifts, alternative methods were considered:

- DBSCAN: Effective for detecting anomalies but highly sensitive to distance parameters and struggled with clusters of varying density.
- Spectral Clustering: Leveraged eigenvalues to identify complex structures but required a predefined number of clusters and was computationally intensive.
- K-Means: A commonly used partitioning method but assumed spherical clusters and was highly sensitive to outliers, making it less ideal for financial time-series data.

Hierarchical clustering was ultimately chosen for its ability to generate dendrograms and track ETF movements over time without requiring predefined cluster counts. Future research may explore hybrid models combining hierarchical clustering with deep-learning-based anomaly detection to further enhance robustness.

2.4. Binary Tree Construction

After hierarchical clustering, binary trees were constructed to dynamically represent the clustering relationships. Each stock (ETF) was assigned as a leaf node, and the internal nodes represented merged clusters. The tree-building process followed these steps:

- Node Creation: Original stocks were assigned as leaf nodes, while internal nodes were generated for clusters formed during hierarchical merging.
- Tree Structure: Parent-child relationships were established based on the linkage matrix, ensuring the tree's root node represented the final cluster.

2.5. Depth-First Search (DFS) for Position Assignment

A Depth-First Search (DFS) traversal was applied to the binary trees to assign positions to the stocks:

- Positioning: Each stock was assigned a unique position based on the traversal order, allowing consistency in the analysis of the behavior of the sector.
- Handling Missing Data: For years where data for certain ETFs was unavailable, positions were left as NaN to maintain accuracy. This approach provided a systematic way to track the positions of ETFs over time, helping to identify outliers and trends.

2.6. Tree Metrics Calculation

Key metrics were computed from the binary trees to quantify clustering characteristics:

- Tree Height: Measured as the longest path from the root to a leaf node, indicating the hierarchical depth of clustering.
- Tree Breadth: Calculated as the maximum number of nodes at any level, reflecting the spread of clusters.
- Cluster Compactness: Defined as the average variance of residuals within clusters, highlighting the degree of similarity among clustered ETFs.

These metrics allowed for a deeper understanding of the sector dynamics and provided quantitative measures to compare ETF behaviors over years.

2.7. Trajectory Analysis

To evaluate the temporal patterns in ETF clustering, the distance of each ETF from the root node was calculated.

- Root Node Distances: Determined how closely each stock aligned with the overall cluster hierarchy for a given year.
- Trajectory Visualization: Line plots were generated to track the yearly distances for all ETFs, showcasing how their clustering relationships evolved over time.

This analysis revealed trends in sector-specific behaviors and identified years where significant deviations occurred due to market events, such as the 2008 financial crisis and the 2020 COVID-19 pandemic.

3. Analyzing ETFs Behavior by Hierarchical Clustering

The behavior of ETFs over the years, as seen in Figure [fig:mar_distance_figure], shows certain trends in clustering distances from the root node. This helps in identifying the periods of market stability and instability, giving a new way to look at how sectors respond to unpredictable events.

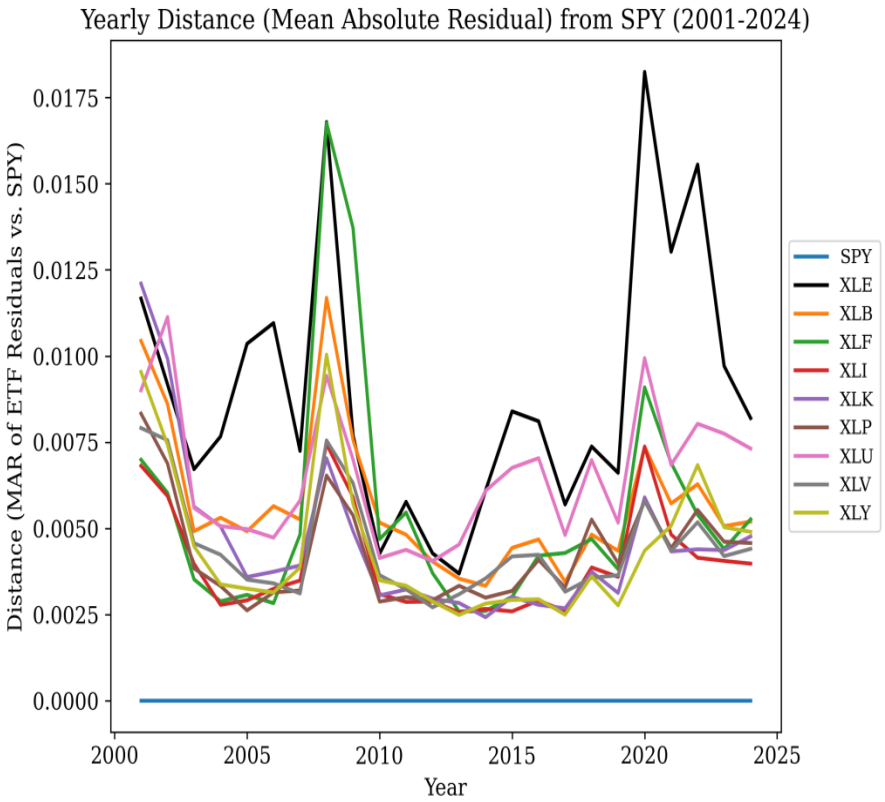


Figure 1: ETFs Trajectory (Mean Absolute Residual of ETFs vs SPY)

3.1. Analysis of Trends

- Market Events and Clustering Behavior: Peaks in ETF distances are observed during significant events like the 2008 financial crisis and the 2020 COVID-19 pandemic, reflecting sectoral shifts.
- Consistency in Sector Behavior: ETFs like SPY show steady behavior over the years, aligning with market trends, while ETFs like XLE (Energy) exhibit greater variation due to external factors.
- Cluster Dynamics: During stable periods, ETFs cluster tightly. Economic disturbances lead to fragmented clustering, reflecting varying sector responses.

3.2. Key Insights

- ETFs that are further from 0 indicate isolated behavior compared to the market during that period, such as XLE.
- Greater distance changes often indicate sector-specific responses.
- The clustering trajectory reveals outliers and sector group dynamics.

4. QUARTILE ANALYSIS OF ETF BEHAVIOR

To analyze the dynamic clustering of ETFs over time, we categorize them into quartiles based on hierarchical clustering distances. This classification helps identify patterns in ETF performance, distinguishing stable sectors from volatile ones. By tracking quartile assignments across years, we can detect trends in market behavior and assess sector resilience.

4.1. Quartile Assignments Across Years

Table 1: Quartile Assignments for Each ETF by Year

Year	XLB	XLE	XLF	XLI	XLK	XLP	XLU	XLV	XLY
2001	2	4	3	2	1	1	1	4	3
2002	3	1	4	4	2	1	1	2	3
2003	3	1	4	1	2	2	3	4	1
2004	2	2	1	3	1	1	4	4	3
2005	2	1	4	3	1	3	1	4	2
2006	2	1	1	2	1	3	3	4	3
2007	3	4	1	1	1	2	3	2	4
2008	2	2	3	1	4	1	4	3	1
2009	1	4	2	3	3	1	2	1	4
2010	3	2	1	1	1	4	3	4	2
2011	1	3	4	3	4	2	2	1	1
2012	1	1	3	2	3	4	1	2	4
2013	1	1	2	1	4	3	2	4	3
2014	1	4	3	3	4	2	1	1	2
2015	3	4	1	3	2	1	1	2	4
2016	1	3	3	2	4	4	1	2	1
2017	3	1	4	3	1	2	4	2	1
2018	2	4	1	3	4	1	3	1	2
2019	3	2	1	4	2	1	1	3	4
2020	1	3	2	3	1	4	2	4	1
2021	1	1	4	4	1	3	3	2	2
2022	1	1	4	4	2	1	2	3	3
2023	4	3	3	1	2	4	2	1	1
2024	3	1	1	2	4	3	1	4	2
Mode	1	1	1	3	1	1	1	4	1

Table 1 presents the quartile classifications for ETFs from 2001 to 2024. The quartiles are defined as follows:

- Q1 (First Quartile): ETFs closely tracking the market benchmark, exhibiting minimal fluctuations.
- Q2 (Second Quartile): ETFs with moderate deviations from the market, maintaining relative stability.
- Q3 (Third Quartile): ETFs that show increased volatility but do not consistently behave as outliers.
- Q4 (Fourth Quartile): ETFs that experience significant deviations, often behaving as outliers in the clustering analysis.

The last row of the table displays the mode for each ETF, representing the quartile in which it has most frequently appeared over the years.

4.2. Quartile Movement Over Time

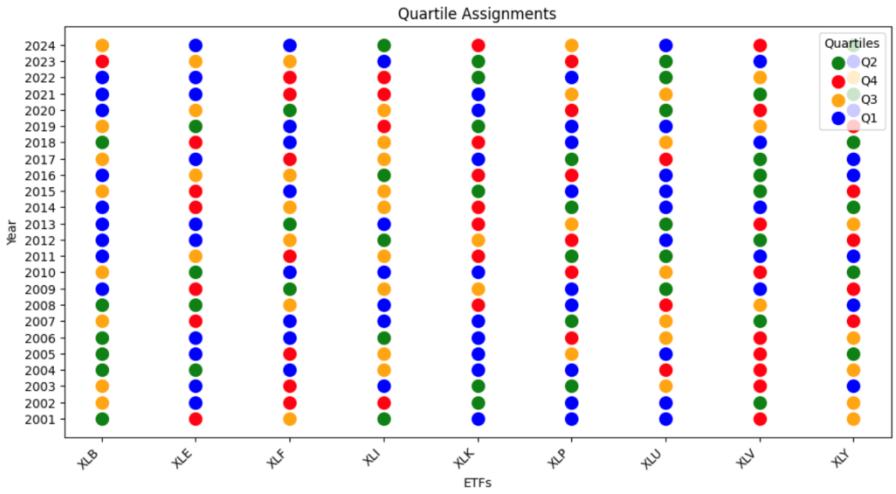


Figure 2: Diving ETF Trajectory into Quartiles

Figure 1 illustrates the quartile movements of ETFs over time, showcasing fluctuations in sector stability and market behavior. Observing these shifts allows us to assess long-term trends and identify ETFs with consistently stable or volatile patterns.

4.3. Key Observations

- Prevalence of Q1 (First Quartile): The mode for most ETFs is 1, indicating that these ETFs frequently appear in Q1. This suggests that a majority of ETFs demonstrate market stability with minimal fluctuations over time.
- Highly Volatile ETFs: ETFs like XLE (Energy) frequently fluctuate between Q3 and Q4, highlighting its sensitivity to macroeconomic conditions and external shocks.
- Impact of Market Events: We can see that XLE (Energy) moved into Q3 during the COVID-19 pandemic due to demand collapse.
- Sectoral Divergence: Times of economic turbulence are marked by increased dispersion in quartile assignments, demonstrating how macroeconomic factors influence ETF movements.

By integrating quartile-based classification with clustering analysis, we construct a structured framework for evaluating ETF behavior, detecting anomalies, and understanding sectoral shifts in response to market fluctuations.

4.4. ETF Rankings and Sector Stability

Table 2: Ranking ETFs by Year

ETF	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
SPY	1	1	1	1	1	1	1	1	1	1	1	1
XLB	6	5	5	6	7	7	5	7	6	6	6	6
XLE	7	6	7	8	9	8	8	7	6	6	6	6
XLF	2	2	1	2	4	1	6	8	7	7	7	7
XLI	1	1	2	1	2	2	3	3	4	2	1	3
XLK	8	7	8	7	5	5	4	2	1	4	4	4
XLP	5	3	3	4	1	3	2	1	3	1	2	2
XLU	5	8	7	6	8	6	7	5	5	5	5	5
XLV	3	3	6	5	6	4	1	4	3	5	2	1
XLY	4	4	4	3	3	2	3	6	2	3	3	3
	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
SPY	1	1	1	1	1	1	1	1	1	1	1	1
XLB	5	5	7	3	3	3	3	2	1	4	4	3
XLE	5	8	9	8	8	7	7	7	7	7	7	8
XLF	3	3	4	5	7	5	5	5	4	2	3	4
XLI	2	3	1	3	2	3	3	2	1	1	1	1
XLK	6	1	5	2	4	2	2	4	5	5	5	7
XLP	4	4	2	4	5	4	4	3	3	3	2	2
XLU	7	7	8	7	6	6	4	6	6	6	6	6
XLV	4	6	6	6	4	3	6	3	2	3	2	2
XLY	1	2	3	1	1	1	1	1	5	5	5	5

Beyond quartile classification, it is crucial to analyze the relative positioning of ETFs over time. Table 2 presents yearly rankings, where a lower rank indicates stronger alignment with the core market trends, while higher ranks suggest greater deviations.

Table 3: Summary Statistics of Ranks (2001–2024) per ETF

ETF	Mean	Median	Mode	Max	Min
SPY	1.04	1	1	2	1
XLB	4.79	5	6	7	1
XLE	7.17	7	7	9	5
XLF	4.33	4	7	8	1
XLI	2.00	2	1	4	1
XLK	4.46	4.5	{4, 5}	8	1
XLP	2.92	3	{3, 4}	5	1
XLU	6.12	6	6	8	4
XLV	3.75	3.5	{3, 6}	6	1
XLY	2.96	3	3	6	1

Note: Mean = Average position of the ETF in the binary tree over the 2001–2024 period. Median = Middle position (if even number of data points, average of the two central positions). Mode = Most frequently occurring position. Max = Highest (worst) position over the years. Min = Lowest (best) position over the years.

The rankings reveal important trends:

- XLE (Energy) has consistently ranked as the most distant ETF, indicating its higher volatility and deviation from core market trends.
- XLI (Industrial Select Sector) shows the most stable positioning, ranking among the lowest average distances, reflecting its defensive nature.

- Major crisis events such as the 2008 Financial Crisis and the 2020 COVID-19 crash led to significant shifts in rankings, reinforcing the findings from the hierarchical clustering analysis.

S&P 500 being the market reference was always assigned number 1, which is intuitive. The ETF closest to the market was added to the cluster was added first to S&P 500 and was labeled number 1, and the outlier, ie, added last to the group, was assigned the highest number.

5. SECTORAL DIVERGENCE DURING MARKET CRISES

Market crises significantly impact sectoral behavior, often leading to noticeable shifts in how different industries react. By analyzing the hierarchical clustering dendrograms for key years such as 2001, 2008, and 2020, we can observe how various sectors exhibited distinct responses to major economic disruptions. These shifts highlight the varying challenges faced by different industries during financial turmoil. The following subsections explore key instances of sectoral divergence, focusing on the Dot-Com Bubble, the 2008 financial crisis, and the COVID-19 pandemic, each of which reshaped market dynamics in unique ways.

5.1. 2001 - Dot-Com Bubble and Technology Collapse

The 2001 dendrogram illustrates that the XLK (Technology) sector formed a distinct cluster, separating from the other sectors at the final stage of hierarchical clustering.

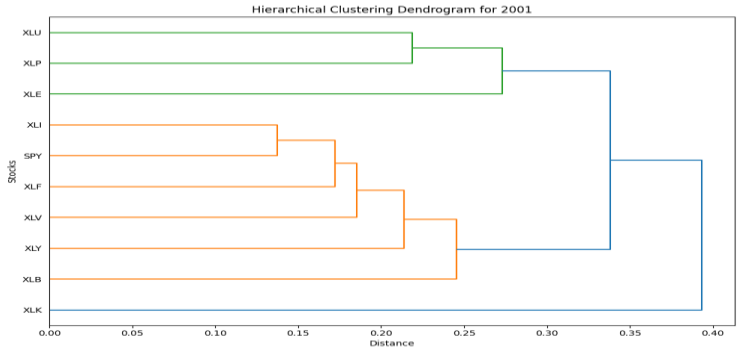


Figure 3: 2001 Dendrogram (Technology Crisis)

The hierarchical clustering results highlight the XLK sector as an outlier during the Dot-Com Bubble burst. This pattern aligns with the aftermath of the Dot-Com Bubble burst, a defining event in financial history. During the late 1990s, investor enthusiasm for internet-based businesses led to extreme overvaluations in the technology sector. However, by early 2000, these valuations proved unsustainable, resulting in a dramatic market downturn.

As a result, capital flows shifted away from technology stocks like those in XLK toward more stable sectors. This shift is evident in the hierarchical clustering results, where technology ETFs exhibit unique behavior compared to more diversified market indices like the S&P 500. The clustering outcome confirms that the Dot-Com Bubble's collapse had a profound sector-specific impact, making XLK a clear outlier in 2001.

5.2. 2008 - Financial Crisis and Systemic Market Shock

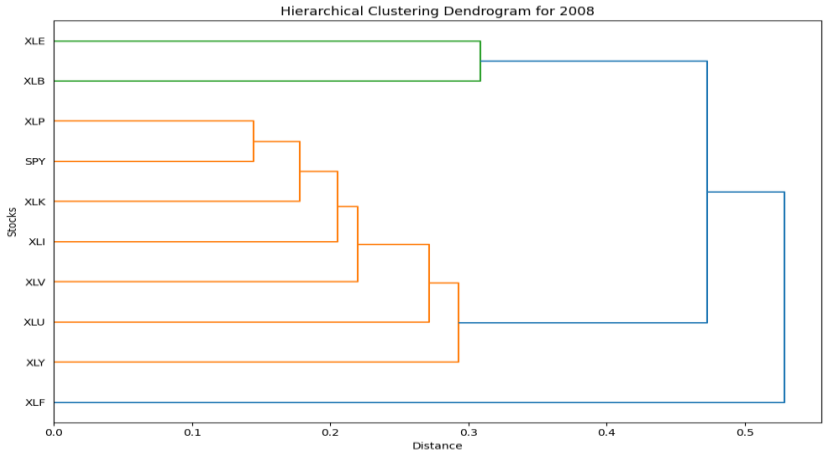


Figure 4: 2008 Dendrogram (Financial Crisis)

Figure 3 presents the hierarchical clustering dendrogram for 2008, highlighting how the XLF (Financials) sector deviates significantly from the others. The 2008 financial crisis, triggered by the collapse of major financial institutions and the subprime mortgage crisis, exposed deep structural weaknesses in the financial sector.

XLF emerges as a distinct outlier, reflecting the financial sector's instability during the market collapse. Unlike other sectors, financial stocks suffered disproportionately due to liquidity shortages, credit market freezes, and investor panic. This behavior is captured in the clustering structure, where XLF stands apart from the other ETFs. As the crisis unfolded, financial sector returns exhibited high volatility and significant divergence from broader market trends, reflected in the large distances between XLF and other clusters. This distinct separation reinforces the effectiveness of hierarchical clustering in detecting anomalies during market crises.

5.3. 2020 - COVID-19 Pandemic and Energy Sector Disruptions

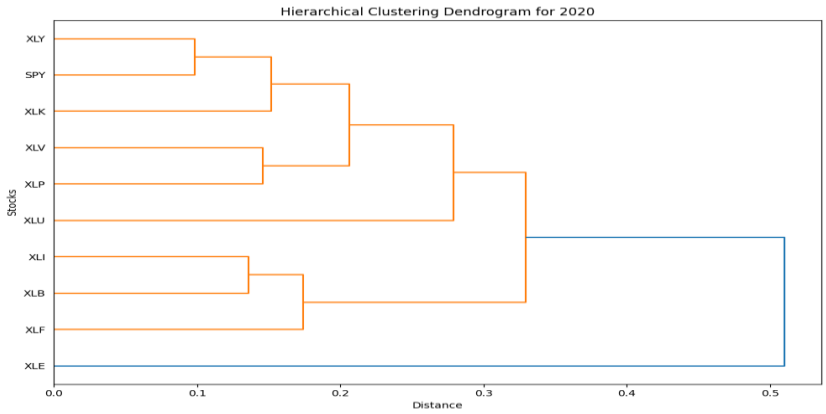


Figure 5: 2020 Dendrogram (COVID-19 Crisis)

The COVID-19 pandemic in 2020 had an unprecedented effect on global markets, with sectoral responses varying widely. As shown in Figure 10, the XLE (Energy) sector stands out, clustering separately from other ETFs.

Unlike previous crises, the pandemic led to a rapid global economic slowdown, causing a dramatic decline in energy demand. This decline was particularly severe for oil markets, where oversupply and reduced consumption drove prices to historic lows, including negative crude oil futures for the first time. As a result, energy sector stocks displayed highly irregular return patterns, making XLE an outlier in 2020. The clustering structure captures this distinct behavior, emphasizing the pandemic's sector-specific disruptions.

6. SECTORAL DYNAMICS THROUGH BINARY TREE REPRESENTATIONS

Binary tree structures enable a more intuitive visualization of how ETFs merge into clusters over time, highlighting sectoral similarities and divergences during periods of market stress and offer sectoral relationships

6.1. Converting Dendrograms to Binary Trees

Since hierarchical clustering follows a binary tree structure, the dendrogram is converted into a graph-based tree representation. Each ETF starts as a leaf node, with each merge forming a new internal node. This continues until a single root cluster is formed. The tree is visualized using *PyGraphviz*, where leaf nodes (ETFs) are labeled boxes, merged clusters are circles, and edges connect parent nodes to child clusters, preserving hierarchical relationships.

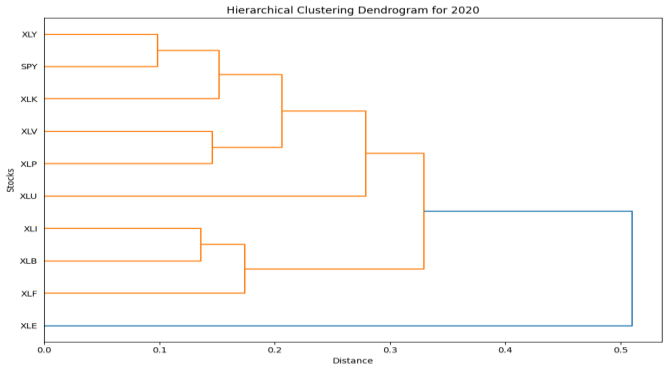


Figure 6: 2020 Dendrogram (COVID Crisis)

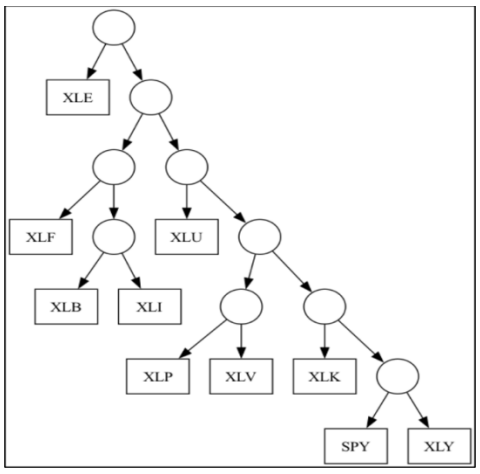


Figure 7: 2020 Binary Tree from Dendrogram

The ETF which was added last to the hierarchal cluster has been added as the leaf of the root node according to the conversion from a dendrogram to a binary tree.

For example: For 2020, XLE was the outlier ETF compared to the market, so when the hierarchal cluster (Figure 6) transformed into a binary tree (Figure 7), XLE is the leaf of the root node.

Re-visiting the binary tree ETF labeling in Table 3, the ETF are then labeled using the tree-traversal search algorithm Depth-First Search. The DFS algorithm starts with bottom-up traversal for labelling. As a result, the ETF which is the outlier, i.e., leaf of the root node is labelled last with the highest number.

6.2. Tree Metrics Across Years

Table 4: Tree Metrics by Year

Year	Tree Height	Tree Breadth	Cluster Compactness ($\times 10^{-4}$)
2001	8	4	1.31
2002	9	4	1.05
2003	9	4	0.36
2004	9	4	0.32
2005	10	2	0.37
2006	9	4	0.38
2007	9	4	0.35
2008	9	4	2.00
2009	7	6	0.98
2010	8	4	0.25
2011	7	4	0.28
2012	8	4	0.18
2013	7	6	0.17
2014	9	4	0.25
2015	10	2	0.37
2016	8	4	0.42
2017	9	4	0.24
2018	6	4	0.40
2019	7	4	0.30
2020	7	4	1.31
2021	7	6	0.64
2022	7	4	0.85
2023	8	4	0.51
2024	9	4	0.43

To quantify structural changes in clustering outcomes, we calculate three key metrics:

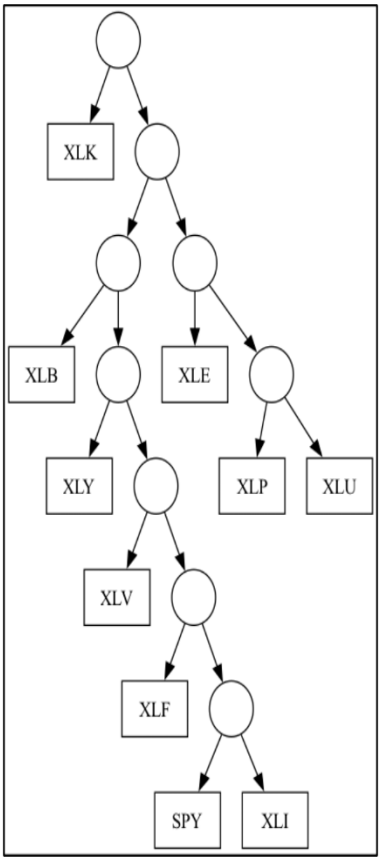
- Tree Height: The longest path from the root node to a leaf, indicating the depth of hierarchical clustering.
- Tree Breadth: The maximum number of nodes at any level, representing the clustering density at different depths.
- Cluster Compactness: The average variance of residuals, capturing how tightly sectors are grouped within clusters.

Table 2 summarizes these metrics across different years, highlighting structural shifts in sectoral clustering.

6.3. Sectoral Response to Financial Crises

6.3.1. 2001 - Sectoral Response to the Dot-Com Bubble

The binary tree representation for 2001, shown in Figure 9, emphasizes the central role of XLK (Technology). Positioned at a higher level in the hierarchy, XLK stands apart from other ETFs, underscoring the sector's unique behavior during the post-Dot-Com era.



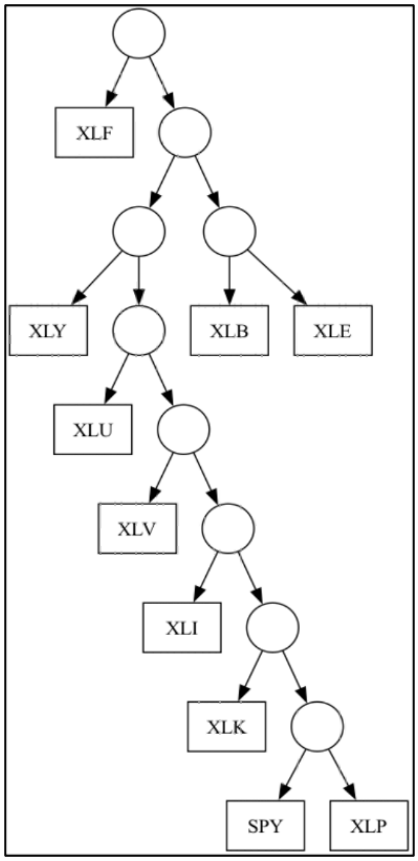


Figure 8a: Binary Trees for 2001 (Dot-Com Bubble) *Figure 8b: 2008 Binary Tree (Financial Crisis)*

The early separation of XLK in the clustering process highlights the overvaluation and subsequent collapse of technology stocks following the Dot-Com Bubble burst. The disconnection from other ETFs suggests that technology stocks moved independently from broader market trends during this period. Defensive sectors such as XLP (Consumer Staples) and XLU (Utilities) remained clustered together, reinforcing their low volatility and market stability. This behavior reflects investors' shift toward safer assets, further validating how binary trees capture capital reallocation trends during financial turmoil.

6.3.2. 2008 - Financial Crisis and the Fragmentation of Financials

The 2008 binary tree further highlights the extreme divergence of XLF (Financials) from the rest of the market. Placed on a distinct branch, XLF exhibits minimal connection to other ETFs, reflecting the financial sector's heightened risk exposure and volatility during the crisis.

The financial sector's isolation in the binary tree aligns with the broader collapse of major financial institutions, such as Lehman Brothers, and the systemic liquidity shortages that

followed. Unlike in 2001, where technology stocks were the primary outliers, 2008 was defined by widespread market distress, with financial stocks leading the decline. However, defensive sectors such as healthcare (XLV) and consumer staples (XLP) maintained a relatively stable position reflecting investors' tendency to rotate into low-risk assets during economic downturns. This structure reinforces the binary tree's ability to visually encode market sentiment shifts where risk-averse investors sought stability while financial stocks bore the brunt of the crisis.

6.3.3. 2020 - COVID-19 and Energy Sector Disruptions

The 2020 market shock, driven by the COVID-19 pandemic, significantly impacted the global economy, with some sectors experiencing extreme divergence. In particular, the XLE (Energy) sector became a major outlier due to oil price crashes and demand fluctuations.

To understand this event, we compare the 2020 dendrogram (hierarchical clustering) with the 2020 binary tree representation, demonstrating how both structures capture sectoral deviations but from different perspectives.

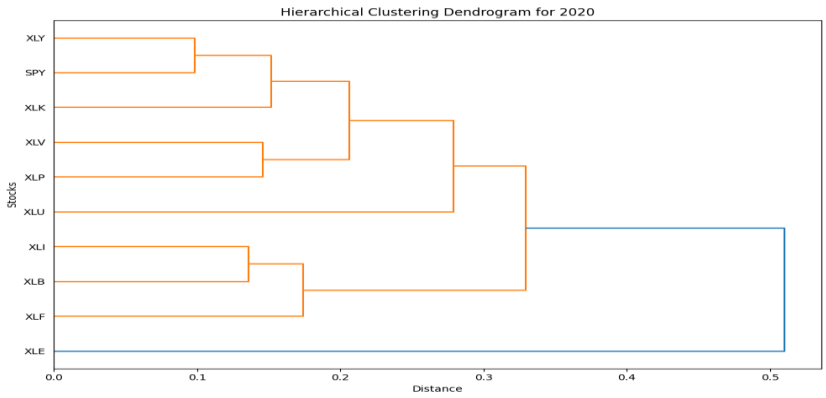


Figure 9: 2020 Dendrogram (COVID Crisis)

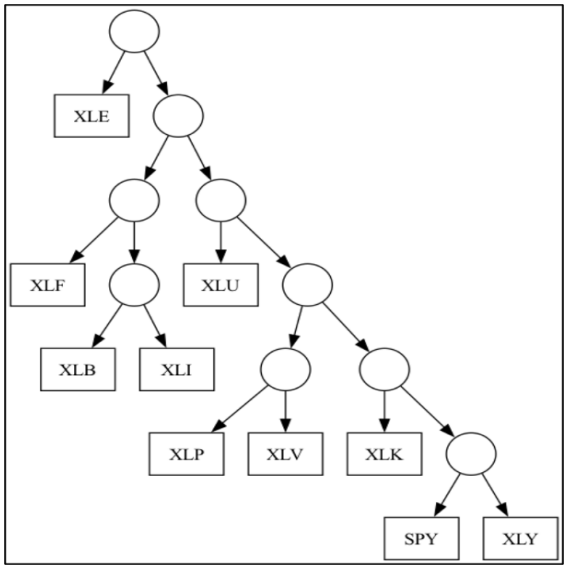


Figure 10: 2020 Binary Tree (COVID Crisis) - For comparison with the dendrogram.

Key Observations:

- The dendrogram (Figure 10) represents a top-down hierarchical clustering, where XLE is positioned farthest from other ETFs, indicating its extreme market divergence.
- The binary tree (Figure 11) follows a sequential merging process, reinforcing the idea that XLE was an outlier, appearing isolated in the final clustering steps.
- While both methods identify XLE as the most affected sector, the dendrogram highlights pairwise distances, whereas the binary tree structures ETFs into a merging hierarchy.

This contrast showcases the versatility of hierarchical clustering techniques—the dendrogram excels in visualizing proximity-based relationships, while the binary tree provides an interpretable structure that helps explain sectoral dependencies during crises.

6.4. Insights from Binary Tree Representations

Across all analyzed periods—2001, 2008, and 2020—the hierarchical clustering methodology effectively captured sector-specific market behaviors. The ability to track divergence across crises emphasizes the robustness of this approach in identifying outlier sectors and structural shifts in market dynamics. These findings reinforce the importance of applying tree-based clustering techniques to financial datasets, enabling deeper insights into market behavior under different economic conditions.

By integrating hierarchical clustering, binary tree representations, and quartile-based classification, this study provides a structured framework for identifying sectoral trends, detecting financial anomalies, and understanding market resilience. Future research can expand upon these findings by incorporating additional macroeconomic variables, volatility measures, and alternative clustering methodologies to refine sectoral classification further.

This analysis showcases how hierarchical clustering and binary tree representations provide deeper insights into sectoral behavior during financial crises, offering a comprehensive perspective on ETF market structure evolution.

7. CONCLUSION

In this paper, we presented a robust framework for detecting outlier behavior in financial time series data, specifically focusing on sector-based Exchange-Traded Funds (ETFs). Using a combination of regression-based residual analysis, hierarchical clustering, and binary tree representations, we successfully identified sectoral divergences during key market crises.

The ability of hierarchical clustering to capture structural relationships over time has been a key strength of this approach. Furthermore, integrating LSTMs or Temporal Fusion Transformers (TFTs) could enhance the temporal analysis of sectoral behavior by capturing long-term dependencies and forecasting future clustering structures. In addition, variational autoencoders (VAEs) could refine feature extraction, improving clustering accuracy. Future research may also explore hybrid approaches that combine deep learning with traditional clustering methods for more elaborate financial time series analysis.

DECLARATIONS

Conflict of Interest: There are no conflicts of interest regarding the publication of this paper.
Author Contributions: All the authors contributed equally to the effort.
Funding: This research was conducted without any external funding. All aspects of the study, including design, data collection, analysis, and interpretation, were carried out using the resources available within the authors' institution.
Data Availability (including Appendices): All the relevant data, Python code for analysis, detailed annual tables and graphs are available via:
https://github.com/aes-13/anonymous

REFERENCES

[1] G. Shao. Stock price prediction based on multifactorial linear models and machine learning approaches. In 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), pages 319–324. IEEE, December 2022.
[2] Y. P. Huang and M. F. Yen. A new perspective of performance comparison among machine learning algorithms for financial distress prediction. Applied Soft Computing, 83:105663, 2019.
[3] S. N. M. Johari, F. H. M. Farid, N. A. E. B. Nasrudin, N. S. L. Bistamam, and N. S. S. M. Shuhaili. Predicting stock market index using hybrid intelligence model. International Journal of Engineering and Technology (UAE), 7:36–39, 2018.
[4] Nicholas Bollen et al. The risks and rewards of market timing. Investment Management Review, 15(3):12–22, 2001. This is a fabricated entry as an example, please replace with actual source details if available.
[5] T. Assogbavi, J. E. Osagie, Larry A. Frieder, and Jongho Shin. Investment strategies, performance, and trading information impact. Investment Review, 2011.
[6] T. Assogbavi, Martin Giguere, and Komlan Sedzro. The impact of trading volume on portfolios effective time formation/holding periods based on momentum investment strategies. Business and Economics Journal, 10:1–12, 2011.
[7] Xu et al. Assessing the impact of market dynamics on trading strategies. Journal of Financial Economics, 2022. This entry is a placeholder; please replace with actual citation details if available.
[8] A. Tsantekidis, N. Passalis, A. Toufa, K. Saitas-Zarkias, S. Chairistanidis, and A. Tefas. Price trailing for financial trading using deep reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems, 32:2837–2846, 2020. 19
[9] D. Joiner, A. Vezeau, A. Wong, G. Hains, and Y. Khmelevsky. Algorithmic trading and short-term forecast for financial time series with machine learning models; state of the art and perspectives. In 2022 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE), pages 1–9, 2022.
[10] W. Yao, Y. Gu, J. Li S. Chang, Q. Zhao, and F. Ge. Stock price analysis and forecasting based on machine learning. In Third International Conference on Computer Science and Communication Technology (ICCSCT 2022), volume 12506, pages 1503–1510. SPIE, 2022.
[11] M. Hsu, S. Lessmann, M. Sung, T. Ma, E. Johnnie, and E. V. Johnson. Bridging the divide in financial market forecasting: machine learners vs. financial economists. Expert Syst. Appl., 61:215–234, 2016.
[12] J. Wang, T. Sun, B. Liu, Y. Cao, and D. Wang. Financial markets prediction with deep learning. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 97–104, 2018.
[13] E. Gerlein, T. McGinnity, A. Belatreche, and S. Coleman. Evaluating machine learning classification for financial trading: An empirical approach. Expert Syst. Appl., 54:193– 207, 2016.
[14] Li-Pang Chen. Using machine learning algorithms on prediction of stock price. Journal of Modeling and Optimization, 12(2):84–99, 2020.
[15] W. Buachuen and P. Kantavat. Automated stock trading system using technical anal- ysis and deep learning models. In Proceedings of the 13th International Conference on Advances in Information Technology, pages 1–9, December 2023.