FIN-525 Financial Big Data
Trading Strategies Given Market States

MFE

KHALFALLAH Selim SCIPER : 283616 **JELASSI Yosr** SCIPER : 286961

Contents

1 Abstract

In the ever-evolving landscape of financial markets, the integration of time period clustering into trading strategies emerges as a promising avenue for adapting to dynamic market conditions. This project addresses the challenge of navigating financial big data to uncover temporal structures within market behaviors. Utilizing clustering algorithms, we identify distinct market regimes with the goal of tailoring strategies to each temporal phase. This research contributes to the practical understanding and implementation of adaptive trading strategies in financial markets, emphasizing the need for resilience in the face of changing market dynamics.

2 Introduction

The financial markets present a complex and dynamic environment, necessitating adaptive strategies to navigate evolving conditions. This project delves into the integration of time period clustering into trading strategies. Our focus is on leveraging financial big data to uncover temporal structures within market behaviors. The subsequent sections detail our approach, from the data preprocessing to trading strategy evaluation, through time period clustering methods. The ultimate goal is to provide actionable insights for practitioners seeking to navigate the complexities of financial markets.

3 Data

3.1 SP100 data acquisition and preprocessing

At the start of our project, we planned on using the dataset of the SP100 index, provided as part of the course materials, specifically for educational purposes. This dataset focused on publicly traded companies whose stocks were constituents of the SP100 index between the years 2004 and 2008.
It comprised both trade files, which record the transactions made during specific time intervals, and BBO (Best Bid and Offer) files, providing information about the best available prices for securities at a given point in time. However, as our analysis progressed, it became evident that the SP100 dataset was not sufficiently comprehensive for our clustering and trading strategy development goals, encompassing only 87 assets. Recognizing this limitation, we opted to transition to a more extensive dataset derived from the SP500 index to achieve more robust and representative results.

Nevertheless, in the initial stages of our project, we utilized the SP100 dataset for training and testing our code, leveraging its lighter data load to expedite the development and refinement of our methodologies.

3.2 SP500 data acquisition and preprocessing

For the purpose of the project and to obtain significant results, we worked with SP500 from 2010, a more extensive dataset, featuring 401 assets. Similar to the SP100 dataset, this dataset included trade files, and Best Bid and Offer (BBO) files.
However, a challenge arose due to the sheer volume of data, resulting in a significantly smaller time period (T) for analysis, set at 20 days. Expanding the time window to cover a more extensive date range would have resulted in a dataset of considerable size, exceeding 100 gigabytes.

To address this scalability issue, the professor provided a preprocessed dataset aggregated at 60-second intervals with BBO files only, reducing the computational load while still offering a representative snapshot of market dynamics. This strategic adjustment facilitated the application of our clustering and trading strategy methodologies to a more comprehensive dataset, ensuring meaningful insights without compromising computational efficiency or dataset size.

4 Strategy

In our financial big data analysis methodology, we focus on constructing a detailed understanding of market dynamics by examining asset returns within distinct time intervals of the trading day. Specifically, we analyze returns between 10:00 (t_1) and 15:30 (t_2) to form clusters that encapsulate the market state during the majority of the trading day. The purpose of these clusters is to categorize days into similar market states based on the asset returns observed in the t_1 - t_2 window, using this period exclusively for cluster formation.

Once clusters are established, we proceed under the assumption that the market state remains consistent into the final segment of the trading day, between 15:30 (t_2) and 16:00 (t_3). This assumption allows us to apply a targeted trading strategy in the t_2 - t_3 window, leveraging historical data from days classified within the same cluster to predict market behavior. The selection of assets for this strategy hinges on ranking them by their t-statistic from the t_2 - t_3 returns, thereby focusing on assets that exhibit statistically significant performance. This ranking process treats the initial days within our rolling calibration window as a training set, where these days' data inform our understanding and selection of assets based on their performance in the t_2 - t_3 period.

The actual application of our trading strategy occurs on the last day of the rolling calibration window, which we consider as the testing sample. On this day, we execute trades on the top 5 ranked assets for long positions and the bottom 5 for short positions, based on their t-statistic ranking from the t_2 - t_3 returns. This approach not only allows us to adapt our strategy daily by shifting the calibration window forward but also extends our testing set continuously, providing a dynamic and iterative mechanism for refining our trading decisions. Through this rolling calibration, each iteration treats the latest day as a new opportunity to test our strategy against the market, ensuring our approach remains responsive to evolving market conditions and historical trends.

5 Methods

5.1 Selection of assets and time period

In the initial phase of the project, we encountered a significant data quality challenge: a substantial number of assets in our dataset exhibited numerous missing files, often providing data for only a limited number of days. This issue could potentially skew our analysis and hinder the reliability of our results.

To address this concern, we implemented a rigorous data preprocessing strategy. Our first step involved a comprehensive examination of the dataset to identify assets with adequate data coverage. Assets with a limited number of data points, spanning only a few days, were flagged as potential candidates for exclusion from our analysis. As a result of this selection process, we were

able to retain a curated subset of assets that demonstrated a high number of days with available data.

Furthermore, we only selected the days that were in common for all the retained assets, we ensured a consistent time frame for our analysis. This process resulted in a dataset consisting of 351 assets and 250 common days, forming the foundation for our subsequent analysis.

5.2 Choice of t_1, t_2, and t_3

In our approach to financial big data analysis, the computation of returns is pivotal for understanding asset price movements over specific intervals within the trading day. This analysis commences at 10:00 (t_1), a time selected for its balance between market stability post-opening fluctuations and the ability to capture early day trends. This ensures avoidance of the initial market volatility at opening, while still early enough to grasp emerging patterns of the day.

The analysis extends to 15:30 (t_2) to include a substantial portion of the trading day, providing a comprehensive view of market activity. This broad interval is chosen to offer an accurate assessment of the day's market state, avoiding the atypical activities typical of the market's opening moments.

The final analysis segment from 15:30 (t_2) to 16:00 (t_3) is deliberately short, under the assumption that the market state remains relatively stable in this brief period. This allows for the application of a targeted trading strategy based on the market state identified earlier in the day, aiming to leverage this stability before the market closes. The choice of this interval is strategic, allowing timely execution of trading actions, leveraging insights from the broader t_1 to t_2 analysis.

5.3 Opening Price and Closing Price

In the initial phase of our project, we planned to utilize the closing price column from our BBO files to calculate the cumulative returns of our trading strategy. Surprisingly, the returns calculated on the final day of the evaluation period were significantly higher than expected, with the strategy's returns reaching $R_s = 10^{124}$ and the benchmark's returns at $R_b = 10^{121}$. These results were evidently incorrect, suggesting a miscalculation likely due to the use of the closing price (`X.Close`) values. A further analysis revealed that these values exhibited several orders of magnitude variation from the start to the end of the day, leading to an unrealistic exponential growth in calculated returns. This discrepancy indicated a potential flaw in the data processing or the formula used for computing returns.

Given the incoherence of these results, calculating any strategy evaluation metrics became impractical. Consequently, we decided to shift our focus to using the opening price (`X.Open`)for our computations. Although the opening price might not typically be the first choice for a trading strategy due to its volatility and potential for overnight news effects, we found the data to be more coherent and suitable for testing our strategy and performing portfolio evaluation. This adjustment allowed us to proceed with a more realistic and practical analysis of our trading approach.
To illustrate, Figure 1 shows as an example IBM's opening and closing prices for 2010-10-27. The plot suggests a relatively stable opening price throughout the day, while the closing price demonstrates significant volatility. Notably, the closing price exhibits extreme fluctuations, ranging from near 1 to almost 110 within the same trading day. This extreme range implies that if returns

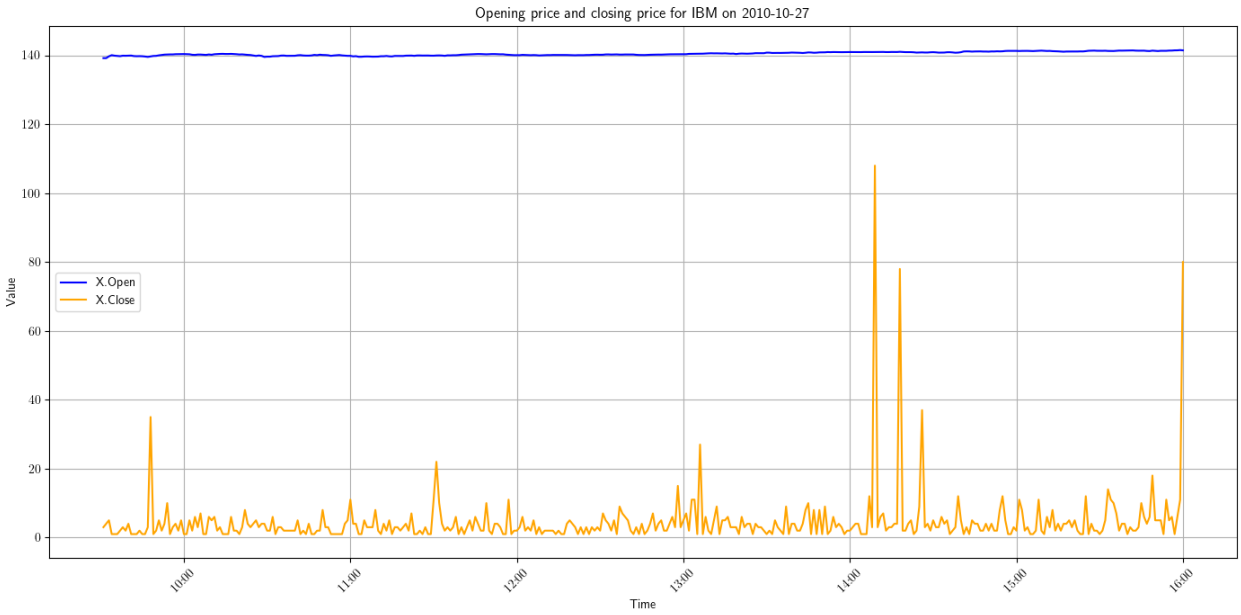

Figure 1: Opening price vs. closing price for IBM on October 27th, 2010

are computed based on closing prices, one might observe exceptionally high returns, which could potentially translate into explosive cumulative returns over time. Such pronounced volatility and the associated outsized returns could skew our analysis of a trading strategy's performance, especially when only a limited number of data points (such as 3 specific times of day) are taken into consideration.

5.4 Computation of the Returns Matrix

The calculation of returns is foundational to our analysis, providing insights into the dynamics of asset prices over specific time intervals throughout the trading day.

Our methodology for calculating returns is designed to address the complexities of financial data analysis. The initial step involves identifying potential data challenges, such as the presence of insufficient data points in the opening price column. To mitigate these issues, we have developed a versatile function that accepts a boolean parameter (`log`), enabling the computation of either simple or logarithmic returns. This function is specifically tailored to handle scenarios where the dataset may be incomplete or improperly structured, ensuring our analysis remains robust and reliable.

We employ a strategy for constructing the returns matrix, particularly focusing on the critical timestamps t_1 (10:00) and t_2 (15:30). Given the possibility of encountering NaN values at these precise moments—which could significantly impair our analysis or lead to data omission-we implement a tailored approach to data imputation. For t_1, a forward-looking method is adopted, whereby we select the first available valid opening price immediately following any NaN occurrences. Conversely, for t_2 and t_3, a backward-looking strategy is applied, opting for the nearest preceding valid value to mitigate the impact of missing data. This decision is intending to minimize the temporal

gap between t_2 and t_3 to ensure continuity and accuracy in our returns matrix.

6 Clustering

For the creation and evaluation of asset clusters within our dataset, we adopt the Louvain clustering algorithm. This method is celebrated for its adeptness at uncovering community structures within complex networks, making it fitting for the analysis of financial data. The efficiency and scalability of the Louvain method enable us to navigate the complex relationships among assets effectively, thereby facilitating a nuanced understanding of financial time series data through the lens of dynamic network adaptability.

The clustering analysis employs returns matrices generated from both regular and logarithmic returns. This dual approach allows for a comparative evaluation, shedding light on the impact of different returns calculation methods on the clustering process. By exploring these variations, we aim to ascertain the most coherent and stable clustering configuration, which is crucial for the integrity of our financial analysis.

We anticipate identifying approximately five to six distinct clusters, each representing unique market dynamics and correlations. To ensure the precision of our clustering efforts, we plan to conduct a series of experiments focusing on optimizing key parameters. These include the choice between simple returns (Experiment A) and log returns (Experiment B), and the determination of the most effective time window for analysis. We know that, in theory, the time window (T) and the number of assets used (N) are supposed to be around $T = N/3$ or $T = N/4$. Here, T represents the time window in terms of the number of days, and N represents the total number of assets. For that, in our experiments, options for the time window (T) range from considering the entirety of the dataset to analyzing segmented portions, such as one-third ($T = N/3$) or one-quarter ($T = N/4$) of the data.

7 Optimization

To enhance the robustness and efficiency of our trading strategy, we have applied several optimization techniques aimed at enhancing computational speed and scalability. A critical component of this optimization strategy was the one-time calculation of asset returns between specific time points—namely t_1, t_2, and t_3—for each asset. By doing so and storing these results immediately, we eliminated the necessity of repeatedly accessing larger data files.

Moreover, we streamlined our process by computing and storing the returns matrices for the $t_1 - t_2$ and $t_2 - t_3$ intervals only once, thus removing the need for redundant calculations. This approach highlights the effectiveness and scalability of our methodology across large datasets, allowing the development of a trading strategy that is both rapid and scalable.

In addition, our strategy's data management efficiency was enhanced by adopting the Parquet format for data storage. Known for its superior performance and compatibility with substantial datasets, the Parquet format has substantially reduced the time needed for subsequent analyses. This is achieved by minimizing redundant computations and enabling quicker data retrieval.

8 Strategy Evaluation

To evaluate our strategy effectively, we implemented a benchmark model that closely mirrors the primary approach of our strategy. This model utilizes t-statistics to select the top 5 assets for a long position and the bottom 5 assets for a short position. It then calculates the mean differences to derive daily returns. However, this benchmark model does not incorporate the clusters identified through our clustering process, offering a comparative baseline that highlights the added value of our clustering-based strategy.

Our evaluation methodology encompasses several metrics that are critical indicators of the overall profitability and efficiency of this trading strategy. We begin by computing the daily returns, which, for a given day, are determined by the difference between the mean expected returns for the top 5 long assets and the mean expected returns for the bottom 5 short assets. This calculation framework enables us to assess the performance of our clustering-based strategy in comparison to the benchmark strategy across various dimensions, including daily returns, volatility, cumulative returns, and the Sharpe ratio.

9 Results

9.1 Clustering

Tables 1 summarizes the results of our clustering experiments.

Experiment	Returns Type	Time Window	Mean Clusters	Std. Deviation
A.1	Regular	$T = N$	4.46	0.59
A.2	Regular	$T = N/3$	4.95	0.53
A.3	Regular	$T = N/4$	5.99	0.64
B.1	Log	$T = N$	4.58	0.68
B.2	Log	$T = N/3$	4.98	0.51
B.3	Log	$T = N/4$	6.02	0.53

Table 1: Summary of Clustering Experiments

We observed that the results of our clustering experiments were relatively stable across different settings, with the usage of logarithmic returns slightly increasing the mean number of clusters. A notable observation was that the smaller the time window, the higher the number of clusters, indicating a correlation between time window size and clustering granularity.

In the course of our analysis, we found ourselves deliberating between Experiment A.3 and B.3. Both experiments showed promise in aligning with our expectations of obtaining five to six clusters, which was our targeted range for a coherent group segmentation. Experiment A.3, with a time window of $T = N/4$, yielded an average of 5.99 clusters with a standard deviation of 0.64, while Experiment B.3, under the same time window but using logarithmic returns, produced a slightly higher mean of 6.02 clusters with a slightly lower standard deviation of 0.53. Despite the close resemblance in outcomes, we ultimately leaned towards Experiment A.3. Our preference was guided by the results fitting marginally better with our initial expectations in terms of achieving a desirable cluster count that falls within our specified range.

As a result, we have opted to proceed with regular returns and a time window $T = N/4$. This

decision is expected to enhance the predictive prowess of our model, providing a solid foundation for our subsequent financial analyses.

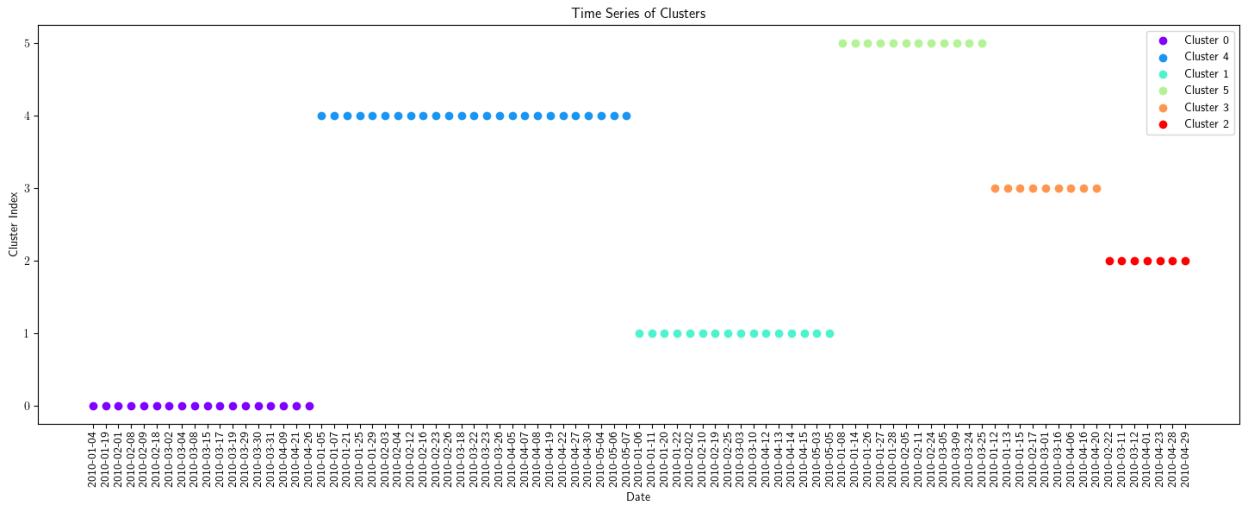

Figure 2: Time series of clusters for experiment A.3

In analyzing the results of the clustering, we also noticed that consecutive days tend to belong to the same cluster. This observation provides a reassuring insight into the logic behind our methodology, as the consistency in cluster assignments for consecutive days aligns with our expectations and reinforces the validity of our approach.

9.2 Strategy Evaluation

In our evaluation of the trading strategies, we focused on critical performance metrics including mean daily returns, volatility, cumulative returns, and the Sharpe ratio. These metrics were instrumental in assessing the profitability, risk, and risk-adjusted performance of our clustering-based strategy relative to a benchmark strategy.

9.2.1 Daily Returns and Volatility

The clustering-based strategy demonstrated a mean daily return of 2.17%, which was slightly higher than the 1.99% observed for the benchmark strategy. This indicates a marginal improvement in daily profitability through the application of the clustering approach. Regarding risk, as measured by volatility, the clustering-based strategy exhibited a volatility of 2.39%, compared to the benchmark's 2.14%. This suggests a marginally higher level of risk associated with the clustering-based strategy.

9.2.2 Statistical Significance

A t-test was performed to determine the statistical significance of the difference in mean daily returns between the two strategies. The results yielded a T-Statistic of 0.708 and a P-Value of 0.240, indicating that the difference in mean daily returns between the clustering-based strategy and the

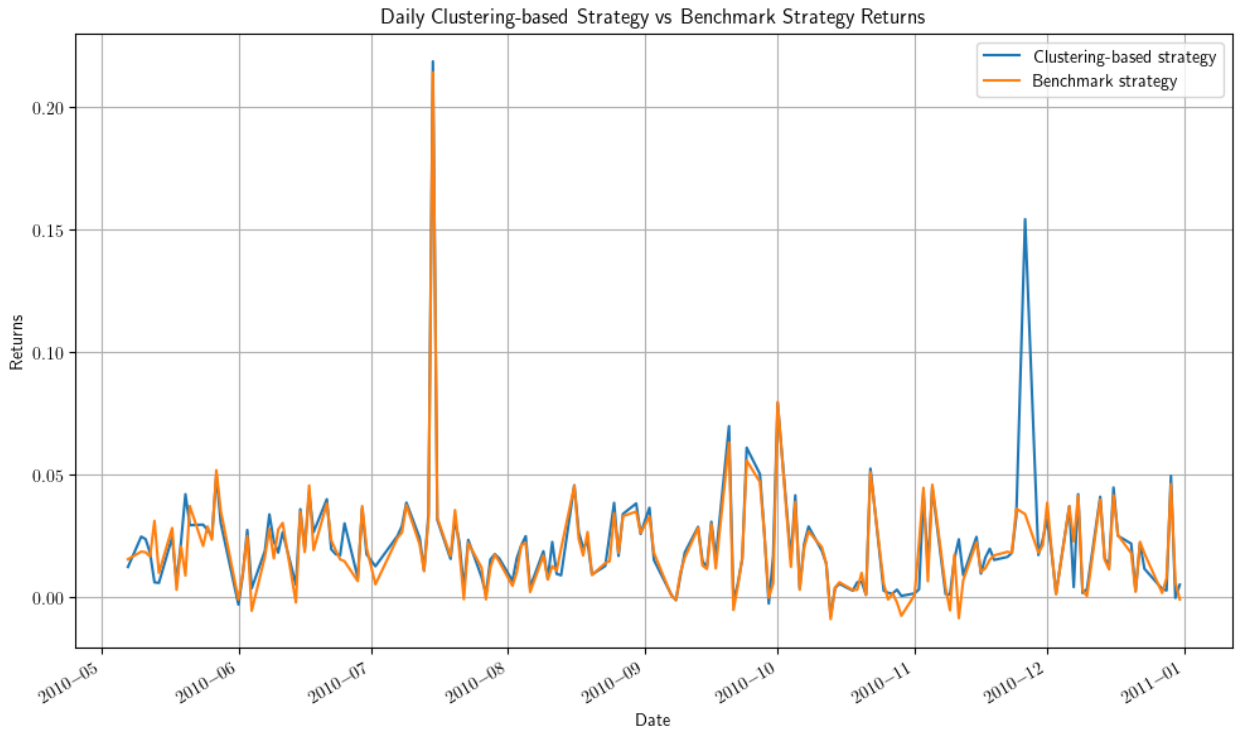

Figure 3: Daily returns of clustering-based strategy vs. benchmark strategy

benchmark is not statistically significant at the conventional 0.05 level.

While the mean daily return for the clustering-based strategy is higher, the lack of statistical significance suggests that this difference could be due to chance. It is important to note that the evaluation of a trading strategy typically does not rely solely on daily returns. The broader context of a strategy's performance, including its cumulative returns and risk-adjusted returns, often provides a more comprehensive assessment of its efficacy.

9.2.3 Cumulative Returns

The evaluation period concluded with the clustering-based strategy achieving a final cumulative return of 3121.07% over 164 days, substantially surpassing the benchmark strategy's return of 2340.34%. This significant outperformance in cumulative returns underscores the potential advantage of employing a clustering-based approach, despite the similar mean daily returns and slightly higher volatility.

9.2.4 Sharpe Ratio

The Sharpe Ratio, which measures risk-adjusted returns, was calculated to be 0.902 for the clustering-based portfolio and 0.924 for the benchmark. The proximity of these ratios indicates that both strategies deliver comparable returns per unit of risk. However, the clustering-based strategy's marginally lower Sharpe Ratio reflects its higher volatility.

9.3 Time Efficiency

It is noteworthy that our algorithm offers high results in terms of its efficiency, as the entirety of the iteration loop workflow, encompassing clustering, t-statistics calculations, ranking, trading and result computations requires only 35 seconds to run entirely. Comparatively, the benchmark strategy, excluding the cluster-related steps, requires 22.5 seconds. This thin difference highlights the scalability of our methodology, demonstrating the ability to handle large datasets efficiently.

10 Discussion

Our analysis has showcased remarkable scalability and efficiency, particularly when applied to large-scale financial datasets. The optimization techniques we employed have significantly enhanced computational speed, enabling our strategy to effectively handle extensive data volumes.

However, our initial decision to use closing prices led to an unexpected issue—an explosion of values towards the end of the trading day. This resulted in disproportionately large returns that compromised the integrity of our analysis, preventing a realistic and reliable evaluation of our trading strategy against the benchmark. In response, we shifted from closing prices to opening prices. While this change might seem counter-intuitive for a typical trading strategy, it allowed us to test our strategy on more consistent data. It remains unclear whether the issue with closing prices was due to incorrect data or extreme volatility. Further research is needed, possibly involving a different dataset, to explore this anomaly in depth.

Notably, the analysis revealed that, while the clustering-based strategy offers a marginal improvement in mean daily returns, this difference was not statistically significant. However, the strategy's substantial cumulative returns point to its potential for superior long-term growth. One might wonder why the cumulative returns were so high despite the modest daily returns. This phenomenon could be attributed to the compounding effect of small gains, which, over time, can lead to significant growth, especially in volatile markets where the strategy might capitalize on specific trends more effectively than the benchmark.

Another aspect of our findings was the optimal performance achieved with a time window of $T = N/4$, for both simple and logarithmic returns. This particular range's efficacy suggests that it might capture a sweet spot for balancing the trade-off between responsiveness to market conditions and the stability of the clustering outcomes. Further research is required to delve into why this time window is optimal and how it aligns with the temporal dynamics of market data, potentially offering insights into improving clustering algorithms for financial analysis.

Our methodology assumes market state consistency between t_1 - t_2 and t_2 - t_3, a presumption that does not always hold true, given the volatility observed in closing prices. This discrepancy highlights a potential misalignment between our theoretical models and the realities of market behavior, suggesting a need for our strategies to adapt more dynamically to market volatility.

Lastly, our strategy's current configuration does not facilitate actionable decisions between t_1 and t_2, limiting trading actions to the interval between t_2 and t_3. This limitation highlights a gap in our ability to leverage intra-day data, pointing to potential areas for enhancing our strategy's responsiveness and effectiveness throughout the trading window.

11 Conclusion

In essence, this project has been an enriching experience, primarily due to our exploration of clustering techniques in the context of financial big data. Departing from our usual scope, we immersed ourselves in the world of financial markets, where working with extensive time series datasets and implementing clustering algorithms played a central role. This endeavor deepened our comprehension of the complexities inherent in financial markets and revealed the potential for innovative clustering strategies to navigate these complexities. Despite encountering various challenges, we are gratified by the robustness and performance of our clustering methodology.

12 Contributions

This project was made possible through the dedicated efforts of our team members. Yosr took charge of the technical aspects, handling data pre-processing, calculating returns, implementing clustering algorithms, evaluating the strategy's returns, managing the rolling calibration, benchmark strategy and the evaluation metrics against the benchmark strategy. Selim focused on evaluating the strategy's returns. Both team members collaborated on writing the report.

We extend our gratitude to Professor Damien Challet and Teaching Assistant Federico Baldi Lanfranchi for their invaluable guidance and support throughout this project.