Predicting end-of-day Market State Using Intraday Data Clustering

Marc Pitteloud, Marine de Rocquigny du Fayel, Pierre-Hadrien Levieil Group E
FIN-525, EPFL, Switzerland

***Abstract*—In this study, we propose a data-driven approach for predicting end-of-day market states using intraday trading data. We employ a clustering technique based on the Louvain algorithm to group similar trading days based on their intraday price movements and trading volumes. This clustering process helps us to develop various portfolio strategies tailored to distinct market conditions. We evaluate the performance of multiple investment strategies, including Naïve, Tangent, Top-Bottom, Momentum, and Risk Parity portfolios, with and without the use of clustering. Our results show that clustering improves the performance of most strategies, yielding higher cumulative returns compared to the standard approaches. Using this demonstrate the value of utilizing cluster-based predictions in market forecasting.**

I. Introduction & Motivation

Financial markets are complex and dynamic, with asset prices and trading volumes fluctuating throughout the day. Predicting these fluctuations is a challenging task, particularly when trying to anticipate end-of-day market conditions. This paper presents a method for predicting the market state at the end of each trading day using intraday data from the first part of the trading session. Specifically, we employ clustering techniques to group trading days with similar intraday patterns and leverage these clusters to inform predictions for end-of-day trades.

The primary objective of this project, undertaken as part of the Financial Big Data (FIN-525) course at EPFL, is to explore how clustering intraday data can enhance the prediction of end-of-day market behavior. We utilize a dataset of intraday trading data for 49 stocks over a period of ten months in 2012. Using the Louvain algorithm, we identify distinct clusters of trading days that share similar intraday behaviors, and subsequently, use these clusters to inform investment strategies. By comparing portfolio strategies based on clustered and non-clustered data, we aim to demonstrate the practical benefits of clustering in real-world market predictions.

II. Data

Our data consists of ETFs (Exchange Traded Funds) data collected over the 2008-2012 period for X stocks, provided to us by the FIN-525 course teaching team.

The data provided by the professor was accessed via the shared drive. We began by exploring the dataset to select the most suitable time period for our analysis. Our goal was to choose a period with relatively stable market conditions, so we excluded 2007-2008 due to the financial crisis. Next, we analyzed the amount of missing data by examining the number of days missing per stock for each month from 2009 to 2012. Based on this analysis, we determined that the optimal period was from March 1, 2012, to December 31, 2012, with exactly 56 stocks at our disposition.

After selecting this period, we investigated the number of missing minutes per stock (see Figure 1).

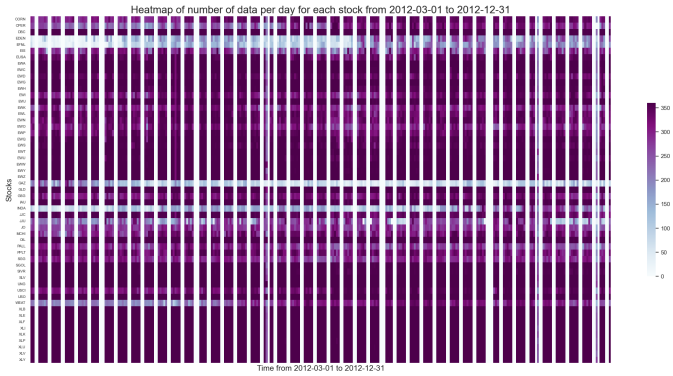


Figure 1. Amount of missing minutes per stock.

Four stocks exhibited significant gaps in the data, and two additional stocks had more than 25% of the total minutes missing. Consequently, we decided to remove these six stocks from our dataset.

For the remaining stocks, we addressed missing minutes by filling them using a weighted average computed over a rolling 5-minute window. Specifically, the weighted average was calculated as follows:

$$\text{WA} = \frac{\sum_{i=1}^{5} (\text{Ask Volume}_{t+i} + \text{Bid Volume}_{t+i}) \cdot \text{Price}_{t+i}}{\sum_{i=1}^{5} (\text{Ask Volume}_{t+i} + \text{Bid Volume}_{t+i})}$$

This approach ensured the integrity of our dataset while mitigating the impact of missing data. We therefore, focus on a total of 49 stocks over 210 days.

III. Methods

A. Clustering

The foundation of our project lies in the clustering algorithm employed to group trading days based on their intraday behavior. Specifically, we clustered trading days using vectors of stock log returns calculated for each minute prior to 15:30, our designated end-of-day threshold. To achieve this, we processed our cleaned dataset by filtering out trades occurring after 15:30 and computed the minute-by-minute log returns for each stock. The resulting data was organized into a matrix where each row represented a trading day and each column corresponded to the log returns for all stocks at each minute of the day. The matrix had dimensions of Number of Days $\times$ (360 $\times$ Number of Stocks), where 360 is the number of minutes between 9:30 and 15:30.

From this matrix, we calculated the correlation matrix C_{corr}, along with its eigenvalues (λ) and eigenvectors ($\mathbf{v}$). We then constructed the matrix C_0, which is the sum of two components: C_r and C_m. These components are defined as follows:

$$C_r = \sum_{i \text{ s.t. } \lambda_i \leq \lambda_+} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$$

where:

- λ_i are the eigenvalues of the correlation matrix C_{corr},
- $\mathbf{v}_i$ are the corresponding eigenvectors,
- λ_+ is a hyperparameter threshold determining which eigenvalues are included in the sum.

$$C_m = \lambda_M \mathbf{v}_M \mathbf{v}_M^\top$$

where:

- $\lambda_M = \max_i \lambda_i$ is the largest eigenvalue,
- $\mathbf{v}_M$ is the eigenvector corresponding to λ_M.

The input to the Louvain clustering algorithm was computed as:

$$|C_{\text{corr}} - C_0|$$

where C_{corr} is the original correlation matrix, and C_0 is the reconstructed matrix derived from the dominant and thresholded eigenvectors.

Finally, we applied the Louvain clustering algorithm [1], implemented using the Community Louvain Python library [2], to this input. The algorithm identified five distinct clusters of trading days, providing meaningful groupings for subsequent analysis.

In order to incorporate new days into our existing clusters, we opted not to rerun the Louvain clustering algorithm for each additional day. Instead, we used correlation analysis to determine the most suitable cluster for a new day. Specifically, we computed the correlation between the new day's state vector and the centroids of each cluster, assigning the day to the cluster with the highest correlation. This approach was chosen to ensure the stability of the existing clusters, particularly to prevent changes in their composition or in the total number of clusters, which is determined dynamically by the Louvain algorithm and not as a hyperparameter.

B. Investment Strategies

Once the market clusters were identified, we developed various investment strategies based on historical patterns within each cluster. The goal was to explore different portfolio construction methods to determine which approach yields the most favorable risk-adjusted returns. The strategies implemented include the tangency portfolio, a naïve portfolio, top-bottom portfolio, momentum portfolio, and risk parity portfolio.

Each of these strategies comes with its own advantages and drawbacks, yet they all share a common objective—to outperform the market by leveraging insights derived from similar historical market conditions.

For all the strategies, the first step involved filtering out past days that did not belong to the same cluster as the current day under analysis. Once the relevant historical data was identified, various portfolio allocation techniques were applied to optimize investment decisions based on the patterns observed within the cluster.

1) Tangency Portfolio: Our first approach was to implement a basic tangency portfolio based on historical values within each market cluster. The tangency portfolio, also known as the mean-variance efficient portfolio, aims to maximize the Sharpe ratio by optimizing the trade-off between risk and return. It assumes that investors seek to achieve the highest expected return per unit of risk.

The portfolio weights are computed using the following formula:

$$w^* = \frac{\Sigma^{-1}\mu}{\mathbf{1}^T \Sigma^{-1} \mu}$$

where:

- w^* represents the optimal weight vector of the assets,
- μ is the vector of expected returns,
- Σ is the covariance matrix of asset returns,
- $\mathbf{1}$ is a vector of ones ensuring weights sum to one.

By applying this methodology, we aimed to construct an optimal portfolio within each identified market state.

2) Naïve Portfolio: In this approach, we used a straightforward investment strategy that takes historical returns as the primary decision-making factor. For each stock, if the average return in past days belonging to the same cluster was positive, we took a long position; if it was negative, we took a short position. This simplistic strategy assumes that past performance is indicative of future trends.

The investment weights were calculated as follows:

$$w_i = \begin{cases} \frac{r_i}{\max(r_{i>0})} & \text{if } r_i > 0 \\ \frac{r_i}{|\min(r_{i<0})|} & \text{if } r_i < 0 \end{cases}$$

where:

- w_i is the weight allocated to stock i,
- r_i is the mean log return of stock i,
- $\max(r_{i>0})$ represents the highest positive mean log return in the cluster,
- $\min(r_{i<0})$ represents the lowest negative mean log return in the cluster.

Pros::

- **Simplicity:** Easy to implement and interpret.
- **Trend following:** Benefits from persistent stock performance.

Cons::

- **No diversification:** Relies solely on past performance without considering risk.
- **Overfitting risk:** Might lead to poor performance in changing market conditions.

3) Top-Bottom Strategy: This strategy builds upon the naïve approach by focusing only on the historically best and worst-performing stocks within each cluster. Instead of taking positions in all available stocks, we selected a subset based on their past performance. Specifically, we longed the top-performing stocks and shorted the bottom-performing stocks based on their historical mean returns within the cluster.

Unlike other strategies where portfolio weights sum to one, in this approach, the total sum of weights is zero, meaning the portfolio is constructed to be market-neutral, with equal exposure to long and short positions. This approach aims to benefit from relative performance rather than overall market movements.

The portfolio weights were assigned as follows:

$$w_i = \begin{cases} \frac{1}{|T|} & \text{if } r_i \geq Q_{1-\alpha} \\ -\frac{1}{|B|} & \text{if } r_i \leq Q_\alpha \\ 0 & \text{otherwise} \end{cases}$$

where:

- w_i is the weight of stock i,
- r_i represents the mean historical return of stock i within the cluster,
- $Q_{1-\alpha}$ and Q_α are the upper and lower quantiles based on the specified threshold α,
- T and B represent the sets of top and bottom-performing stocks, respectively.

Since we assign equal weights to the selected stocks, the long and short positions balance each other, resulting in a total portfolio weight of zero.

Pros::

- **Focus on strong trends:** The strategy concentrates on assets with clear performance patterns.
- **Market neutrality:** The zero-sum weight approach reduces exposure to overall market movements, focusing purely on relative performance.
- **Potential for higher returns:** By filtering out underperforming mid-range stocks, capital is allocated more efficiently.

Cons::

- **Limited diversification:** The strategy may lack exposure to a broad range of assets, increasing concentration risk.
- **Potential overfitting:** Historical top/bottom performers may not maintain their trends in future periods.
- **High turnover:** Frequent portfolio rebalancing may lead to increased transaction costs.

4) Momentum Portfolio: In this modified momentum strategy, we allocate weights based on the past performance of stocks within the same cluster. Unlike the previous long-only approach, this strategy takes both long and short positions, aiming to benefit from positive trends while hedging against underperformers.

The portfolio weights are computed as follows:

$$w_i = \begin{cases} \frac{r_i}{\sum_{j \in P} r_j} & \text{if } r_i > 0 \\ -\frac{r_i}{\sum_{j \in N} |r_j|} & \text{if } r_i < 0 \\ 0 & \text{otherwise} \end{cases}$$

where:

- w_i is the weight of stock i,
- r_i represents the mean historical return of stock i,
- P denotes the set of stocks with positive historical returns,
- N denotes the set of stocks with negative historical returns.

Finally, the portfolio is normalized to ensure that the sum of absolute weights equals one:

$$\sum |w_i| = 1$$

Pros::

- **Balanced exposure:** Captures both upside and downside trends.
- **Improved risk management:** Offsets losses by shorting underperforming assets.
- **Market neutrality:** Potential to generate returns independent of market direction.

Cons::

- **Higher complexity:** More complex than a long-only approach.
- **Transaction costs:** Increased due to shorting.
- **Volatility risks:** Can be sensitive to sudden market reversals.

5) Risk Parity: Building on the work of [**?**], the risk parity investment strategy aims to allocate portfolio weights such that each asset contributes an equal amount of risk to the overall portfolio. This method is widely used because it

promotes diversification and avoids concentration in a few assets, making it a robust approach for risk management.

The portfolio weights in the risk parity strategy are computed based on the inverse of the asset risk (standard deviation), ensuring that assets with higher volatility receive lower allocations. The mathematical formulation is as follows:

$$w_i = \frac{\frac{1}{\sigma_i}}{\sum_{j=1}^{N} \frac{1}{\sigma_j}}$$

where:

- w_i is the weight of stock i,
- σ_i represents the standard deviation of returns for stock i,
- N is the total number of assets in the portfolio.

The objective is to achieve a portfolio where the risk contribution RC_i from each asset i is equal:

$$RC_i = w_i \cdot \sigma_i = \frac{1}{N} \sum_{j=1}^{N} w_j \cdot \sigma_j$$

This strategy ensures that no single asset dominates the portfolio risk, leading to a more balanced and resilient investment.

Pros::

- **Diversification:** Ensures balanced risk exposure across all assets.
- **Stability:** Tends to perform well in volatile markets.
- **No strong return assumptions:** Does not rely on predicting asset returns.

Cons::

- **Computationally intensive:** Requires frequent rebalancing based on changing volatilities.
- **Underperformance in trending markets:** May lag behind more aggressive strategies in strong bull markets.
- **Sensitivity to estimation errors:** Misestimation of volatility can lead to suboptimal allocations.

IV. Results

In Figure 2, we present the results of our different investment strategies. The comparison includes the performance of all five strategies applied in their standard form (without using clustering) and their clustered variants, where only the historical data from days belonging to the same cluster as the day under analysis is used. As shown, clustering significantly enhances the cumulative returns for most strategies, with the exception of the risk parity portfolio, which exhibits almost identical returns for both methods. Importantly, all strategies in their improved version achieve positive returns, indicating their overall effectiveness.

A. Comparison with Market Performance

It is worth noting that our clustered strategies have outperformed the S&P 500 index, which yielded approximately 16% returns in 2012. In contrast, our best-performing strategies, such as the clustered top-bottom and naive portfolios, delivered significantly higher cumulative returns over the same period. This highlights the effectiveness of leveraging historical similarities to enhance investment decision-making.

B. Effectiveness of Clustering

A key takeaway from the results is that clustering generally improves the performance of our strategies compared to their standard counterparts. This indicates that filtering historical data based on similar market conditions provides valuable information that can be exploited to enhance returns. By isolating periods of comparable market behavior, our approach effectively reduces the influence of unrelated market noise, allowing for more targeted and informed investment decisions.

C. Risk Parity Portfolio Insights

Interestingly, the risk parity portfolio shows similar performance in both the standard and clustered versions. Estimating the variance-covariance matrix accurately is inherently challenging, as it requires a large amount of data to produce stable estimates. However, the fact that clustering does not degrade the results suggests that filtering by clusters helps reduce noise while compensating for the loss of data by focusing only on relevant periods. This could imply that our clustering approach provides sufficient diversity within clusters to maintain robust risk parity allocations.

Overall, our findings suggest that clustering enhances performance for most strategies by isolating relevant periods and reducing noise. The approach proves to be an effective way to refine decision-making and potentially increase profitability.

Additionally, Figure 3 highlights the relative performance of the strategies. The Top-Bottom strategy stands out as the most profitable by a wide margin, followed by the Naive strategy and then the Momentum strategy. We theorize that these returns could be greatly improved by using more days and most importantly more stocks, as we might be limited in performance by our amount of data.

Figure 2:

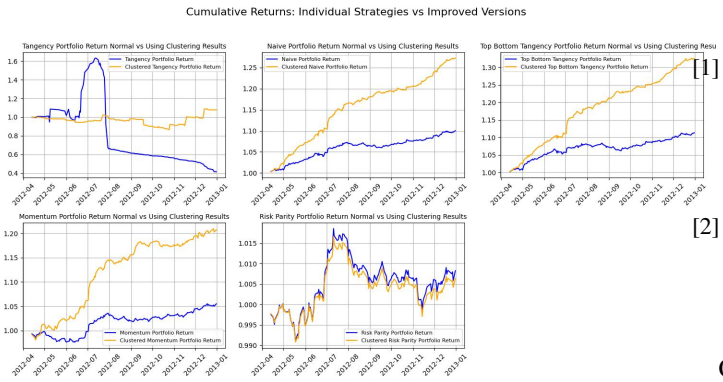
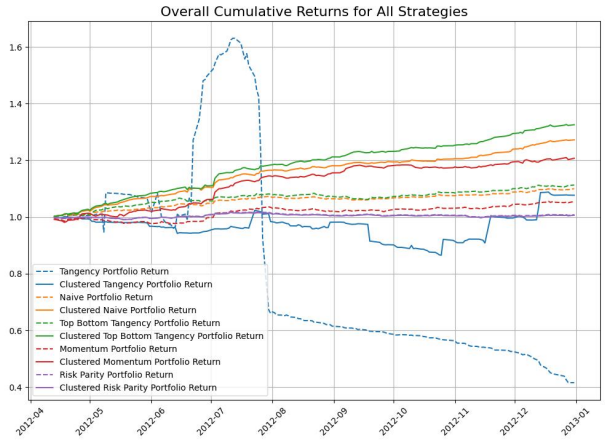


Figure 3:



V. Conclusion & Discussion

In this study, we successfully applied clustering techniques to intraday trading data in order to predict end-of-day market states and inform investment strategies. Our results show that clustering enhances the performance of most strategies, with the Top Bottom strategy emerging as the most profitable. This suggests that leveraging clusters of similar market conditions can yield better risk-adjusted returns compared to traditional strategies that do not account for these patterns.

While the clustering approach improved portfolio performance, there are several avenues for future research. First, the performance of all strategies could benefit from expanding the dataset to include more stocks and a larger time period. Additionally, exploring more advanced clustering techniques or incorporating machine learning models for predicting returns could further improve the accuracy and profitability of the strategies. Overall, the results demonstrate the potential of using intraday data clustering as a tool for enhancing market predictions and optimizing investment decisions.

References

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, Oct. 2008. [Online]. Available: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008

[2] "Scikit network louvaij python documentation." [Online]. Available: https://scikit-network.readthedocs.io/en/latest/tutorials/clustering/louvain.html

VI. Code

Our code and documentation is available at: (link)