Applied Statistical and Machine Learning Methods in the U.S Equities Markets

Evran Ozkul, Quang Duong and Sahand Ahmad

EECS 545 final report - Fall 2007

1 Introduction

Principles of machine learning provide several powerful methods for analyzing and learning from data sets. The stock market is a natural environment to apply these learning methods considering its stochastic nature, high dimensionality, and numerous sectors. We are motivated to develop an effective and profitable trading strategy based on integrating machine learning methods with time series models.

We propose a new method of combining clustering and cointegration methods to identify and refine groups of stocks between which exist some relationships that can be exploited for higher trading profits. Our approach is to employ clustering methods to identify groups/baskets of stocks inherently correlated based on historical data. We then apply the cointegration method to discover and consequently these cointegrated stocks.

The methodology for cointegration was introduced in 1987 in Engle and Granger's seminal paper "Cointegration and Error Correction: Representation, Estimation, and Testing" [3]. They concluded that if non-stationary time series vectors are each stationary after taking the first difference, and if a linear combination of these untransformed vectors results in stationarity as well, then these vectors are claimed to be cointegrated. That is we seek to find linear combinations of non-stationary variables, that will result in a stationary white noise like process, with constant mean and variance. In this analysis, we parameterize our linear combination through ordinary least squares. Engle and Granger's original analysis involved economic time series, where they found that consumption and income were cointegrated, as well as long and short term interest rates.

Recently, cointegration has become a popular tool in explaining relationships amongst financial variables, in addition to economic applications. Numerous cointegration applications have been investigated within financial time series, many focused on similar datasets involving the world's equity and credit markets. Many of these papers have used cointegration to explain relationships between securities within different global markets [10, 8], however few has investigated cointegration as a feasible trading methodology. Furthermore, none to our knowledge, has investigated integrating machine learning methods to improve the effectiveness of cointegration based trading.

We seek to answer the following question: Will the integration of clustering in the cointegration method generate any extra profits compared with cointegration trading without clustering?

Section 2 starts the discussion with an overview of the model and data, and a detailed description of our trading program's main components: clustering and cointegration. Our empirical study in section 3 explains the set up and objective of our experiments, our evaluation method, and then our analysis of the results. Section 4 reviews our research questions, summarize our results, discuss some highlighted achievements and limitations, and suggests some further research.

2 Methodology

2.1 Basic definitions

First of all, we would like to define some terminology employed throughout this paper. Stock returns are the percentage change in value of the underlying stock price. Volatility is the standard deviation of returns. Volume measure denotes the number of shares traded in a given day. PE ratio is evaluated as ratio of stock price to corporate earnings. Market capitalization total is the equity value of a firm or more precisely the total number of shares outstanding multiplied by its stock price. Short selling refers to the method used to profit from securities' decline in value. A short seller borrows a security from a broker and sells it in the marketplace, only to repurchase the share later at a lower price in order to return the share to the broker, and close out the short position. In contrast, long simply means to purchase a stock [1].

2.2 Data and model

Our data source is the Bloomberg Database, where we obtain the closing price times series for the 500 stocks constituting the S&P 500. This data set includes trading information of 500 stocks over a period of 757 days, reflecting 3 years of time series data for the S&P 500. The data includes a series of daily closing prices and trading volumes $\{(p_t^i, v_t^i)|t = [1, 757], i = [1, 500]\}]$ where p_t^i denotes the closing price for stock i on day t and v_t^i represents stock i's trading volume on that day. Another set of data is $\{(s^i, cap^i, PE^i)|i = [1, 500]\}$, in which s^i is the industry sector that stock i belongs, cap^i is the market capitalization of the company of stock i, and PE^i denotes its PE ratio. As our trading strategy relies on the movement of stock prices, not their absolute values, we incorporate the log return measure, denoted as $r_t^i = \ln\frac{p_{t+1}^i}{p_t^i}$, in our clustering and cointegration methods.

2.3 Learning and trading methods

Our trading program consists of two main modules: 1. the clustering module takes in historical data of stocks for the last quarter $T = 65$ days and then separates them into basket. 2. the cointegration module uses the remaining data to learn the correlations between stocks within each basket, and execute trading to exploit those whose prices have a tendency to move together. We address cointegration as the backbone of our trading policy and then describe clustering as an enhancement device to cointegration.

2.3.1 Cointegration

Engle and Granger proposed parameterizing the cointegrating vector β_1 by ordinary least squares regression [3]:

$$\hat{Y_{1t}} = \beta_0 + \beta_1 Y_{2t} + \epsilon t$$

where $\hat{Y_{1t}}$ are the predicted logged prices and Y_{2t} are the observed prices. We use $\bar{Y_{1t}}$ and $\bar{Y_{2t}}$, two historical time series vectors of logged prices for two different stocks, in place of $\hat{Y_{1t}}$ and Y_{2t} in our cointegration method. All possible pairs within every cluster, will be parameterized and tested. Therefore, smaller clusters will result in low computational complexity, as the number of combinations is equal to :

$$\sum_{g=1}^{K} \frac{n_g \times (n_g + 1)}{2}$$

where n_g is the number of elements per cluster, and K is the number of clusters.

In the linear model β_0 will be set to zero so that we will attempt to find offsetting pairs with mean zero. The next step in testing for cointegration is to determine whether the linear combinations are indeed stationary, and are not serially correlated. In time series analysis, a simple t-test based on an auto-regressive process would be constructed of the form

$$\hat{X}_t = \beta_0 + \beta_1 X_{t-1} + \epsilon t$$

where in this case $\hat{X}_t$ is defined as$\hat{X}_t = Y_{1t} - \beta_1 Y_{2t}$, and we are testing $\hat{X}_t$, the linear combination of the cointegrating vectors, for stationarity, also known as a unit root test. In the econometric literature, one of the most popular root tests is the Dickey-Fuller test statistic (DF) based on the null hypothesis H_0: $\beta_1 = 1$ [6].

The graph below shows two corrrelated stock prices, which move together and tend to converge after any divergence between them occurs.

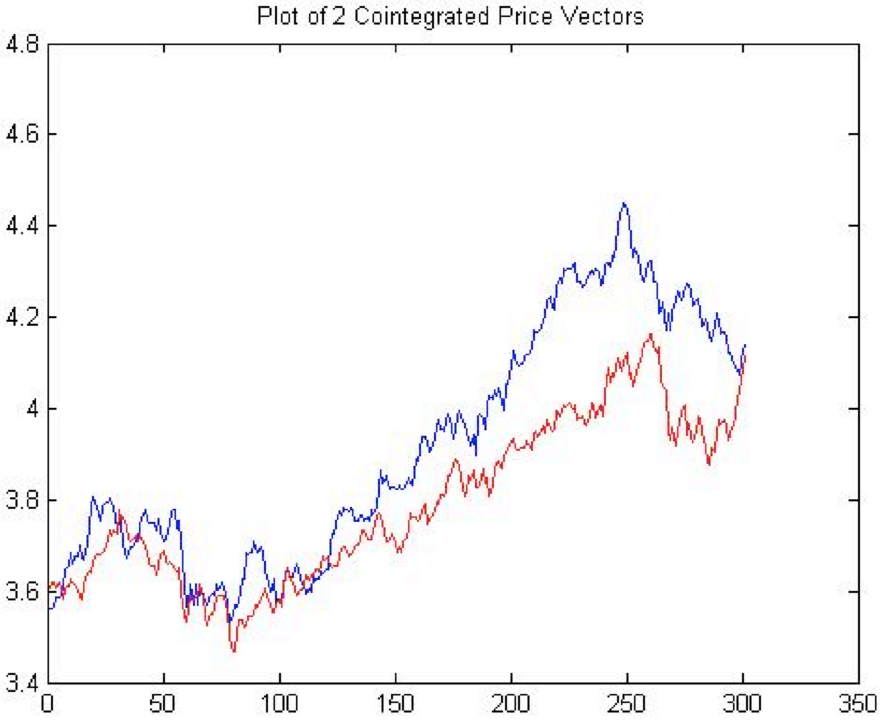


Figure 1. Two correlated stocks' behavior

However, the literature indicates that the Dickey-Fuller t-statistic does not follow a standard t-distribution as the sampling distribution of this test statistic is skewed to the left with a long, left tail [6]. Therefore, a Dickey Fuller Table must be referenced instead of a normal t-table. For large data sets with observations greater than 500, a Dickey-Fuller Statistics of less than -3.96 is necessary to obtain a p-value of 0.01 [6].

If the parameterization above results in a significant Dickey Fuller statistic, the pair of stocks are said to be cointegrated. The next step would be to implement a "pairs trading strategy." As the name implies, pairs trading involves a pair of securities, typically long one stock and short another. In this case, our cointegrating vector indicates the weights of the long-short relationship. Pairs trading is based on the predicted value $\hat{Y_{1t}}$ relative to the true Y_{1t} value. If $\hat{Y_{1t}} - Y_{1t}$ is greater than some positive threshold, this will be interpreted as a signal to buy Y_{1t} and short sell $\beta_1 Y_{2t}$. If the difference is negative, that is Y_{1t} is overvalued relative to the predicted value, this will be interpreted as a signal to short sell Y_{1t} and buy $\beta_1 Y_{2t}$. This is commonly referred to as a mean reverting pairs trading strategy, that is the difference between Y_{1t} and $\hat{Y_{1t}}$ will revert to its long term mean. As we have tested the series for cointegration, we have determined the long term mean of this linear combination to be zero.

2.3.2 Clustering

As hidden relationships between stocks are extremely complex, employing only cointegration in investigating these interactions may oversimplify their complexity, and lead to incorrect conclusions that could possibly cause decline in profits. Observing that cointegration relies mainly on stock-return data motivates us to employ clustering to take advantage of other information about stocks valuable to identifying their underlying correlations. Separating stocks in different baskets and limiting cointegration among same-basket stocks may be argued to eliminate some profitable trading options and cause loss of profitability. However, Clustering is more than just to get rid of some trading options arbitrarily, but to use additional information to selectively filter out unreliable stock correlations. This argument will be further examined and justified in our empirical study.

Each stock is a data point in our clustering state space, whose dimensions are defined as features believed to have the capability of differentiating them in a meaningful way. The features incorporated as our clustering state space's dimensions are: trading volume, returns, volatility, industry, PE ratio, and market capitalization. We choose these features because first of all, we have access to their data. More importantly, they possess some properties that could possibly differentiate stocks with respect to their prices' dynamics. For instance, stocks in the same industry tend to have some mutual influences. Big companies with high market capitalization are more likely to react to market's changes in some similar ways. Profitable and popular stocks with high returns and high trading volume may be an expression of some hidden correlations between their businesses. Although time series data of stock prices is our main input, time series clustering is not considered in our clustering method due to its complexity and moreover, its very limited as Keogh et al. argues [9]. Therefore, we compress each stock's returns and volume data series into average returns and volume over time $T = 65$ days.

Besides those features, we introduce one novel addendum that provides some insights about these complex stock interactions, using factor analysis: eigenvectors of the covariance matrix of stock returns [11]. Given an $n \times T$ matrix of stock returns, we compute its covariance matrix, which is then used to calculate the covariance matrix's eigenvectors. The objective is to explain the covariability among a number of observable variables, in this case stock returns over a period of T days, in terms of smaller number of unobservables called factors, corresponding to the covariance matrix's eigenvectors [11]. Therefore, we choose the most representative dimension or eigenvector $\{eg^i | i \in [1, 500]\}$ that has the highest eigenvalue to incorporate as another dimension in our clustering method, besides the other existing features.

The clustering method of choice in several financial applications is hierarchical clustering [2, 5, 4, 7], as it does not require specifying in advance the number of clusters, and its graphical representation, dendrogram, helps manual stock data investigation that relies more on analysts' experience and intuition. Therefore, we decided to adopt hierarchical clustering for our model. Besides, given that for each clustering method, there are many parameters whose value range requires a lot of exploration, focusing on hierarchical clustering allows us to have more time fine-tuning our model .

Since we are exploring the effects of some features and have no prior knowledge our their interdependence, we calculate the distance between two points in our clustering data space simply by the Euclidean distance, with an exception of the industry sector dimension:

$$d(i,j) = \sqrt{\sum_k (f_k^i - f_k^j)^2 + (\frac{s^i \neq s^j}{\alpha})^2}$$

where f_k is one of the features: average returns, average trading volume, volatility, PE, and market capitalization. $s^i \neq s^j$ returns 0 if i and j are form the same industry and 1 otherwise. α is some scaling factor to make $(s^j \neq s^i)$ comparable with the rest of the features. After some experiments with different scaling factor, we set $\alpha = n$. Moreover, features are scaled to have unit variance, which allows the model to keep balance between different features' variability.

The major difference between our approach and existing clustering financial applications is that our state space has more dimensions that just volatility [2], returns and volatility [4], or industry sector's general health [5]. On the one hand this helps to inject more information in our clustering method, but on the other hand leads to the problem of selecting which features to include. Using all dimensions or features does not guarantee optimal performance in terms of trading profits due to the curse of dimensionality in our clustering method. Furthermore, the features that we choose to consider have not been thoroughly proved to truly reflect any underlying relations between stocks, which means that including one more dimension does not necessarily reveal more hidden stocks' interactions. Our approach is then to explore some subsets of the features empirically and comparing the profitability of the testing scenarios to determine which feature combinations are worth considering.

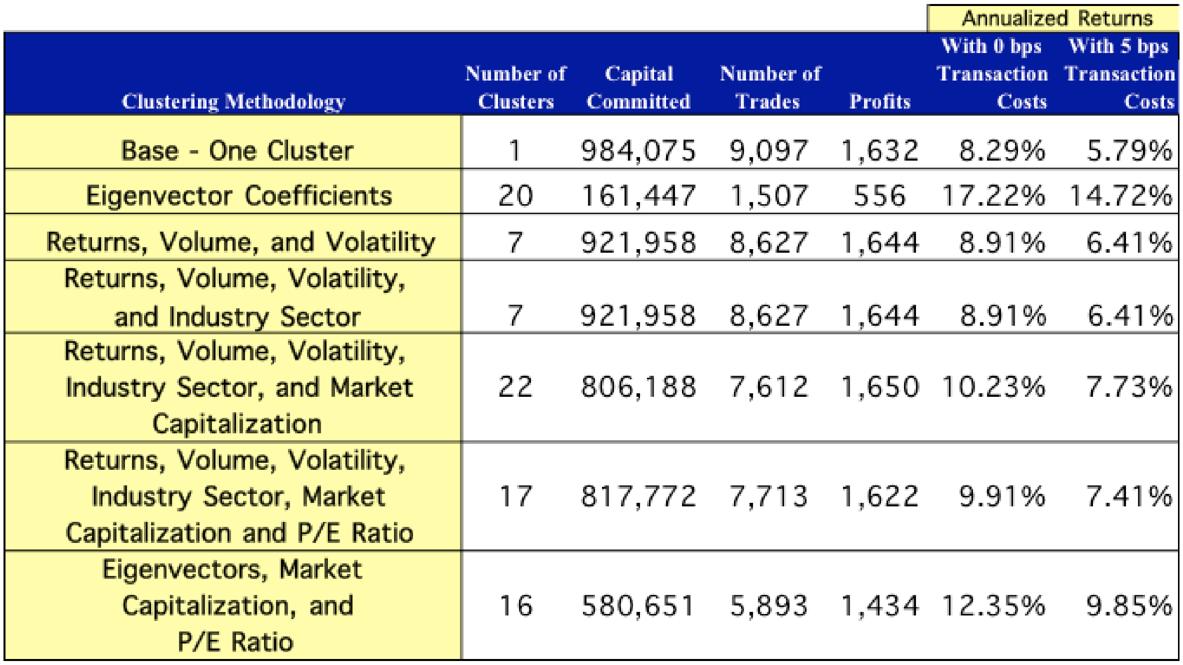
The number of clusters plays an important role in the performance of our trading program. The more clusters, the more unstable and unreliable relations between stocks are eliminated. At the same time, too many clusters repudiate meaningful interactions between stocks that can be exploited in cointegration. Therefore, we decide to limit the number of clusters as $\frac{N}{10}$, because profit tends to decline as the number of clusters exceeds this number. Another component of our clustering module is the linkage measure used as an instrument to merge and create clusters. Our preliminary study of three linkage computation methods: centroid, min and max, shows no significant differences between them in terms of profitability; thus, we choose centroid for the clustering's linkage computation.

3 Empirical study and analysis

A priori, we do not know which variables will add value to a clustering methodology, nor do we know how many dimensions to include. We have restricted the number of clusters per method to be less than $\frac{N}{10}$, as a very large number of clusters would dramatically restrict trading opportunities. We have developed 6 base case scenarios with varying dimensions, within each base case, 9 different tree cutting thresholds are implemented in order to better understand analyze the the clustering sensitivities.

Interestingly, the clustering method bases solely on the coefficients of the first eigenvector enjoyed the greatest performance, yielding an annualized return of 17.22% greatly outperforming the base case scenario of one group including all stocks, which yields 8.29%. Moreover, in order to effectively penalize the high frequency of trades, an estimate of 5 basis points (or 0.05%) is assessed as a transaction cost. This cost is an estimate of the costs to trade through a brokerage house. This is only an estimate as transaction costs are dependent on volume of trades as well as a function of liquidity of each asset.

Different scenarios are run based on various combinations of the aforementioned variables, in order to evaluate which variable or variable combination results in the highest return performance, as shown in Table 1.



					Annualized Returns	
Clustering Methodology	Number of Clusters	Capital Committed	Number of Trades	Profits	With 0 bps Transaction Costs	With 5 bps Transaction Costs
Base - One Cluster	1	984,075	9,097	1,632	8.29%	5.79%
Eigenvector Coefficients	20	161,447	1,507	556	17.22%	14.72%
Returns, Volume, and Volatility	7	921,958	8,627	1,644	8.91%	6.41%
Returns, Volume, Volatility, and Industry Sector	7	921,958	8,627	1,644	8.91%	6.41%
Returns, Volume, Volatility, Industry Sector, and Market Capitalization	22	806,188	7,612	1,650	10.23%	7.73%
Returns, Volume, Volatility, Industry Sector, Market Capitalization and P/E Ratio	17	817,772	7,713	1,622	9.91%	7.41%
Eigenvectors, Market Capitalization, and P/E Ratio	16	580,651	5,893	1,434	12.35%	9.85%

Table 1. Performance measures of hierarchical clustering with different parameters

A base case is established to be one large group, so that there is no clustering influences affecting the cointegration based trading methodology. According to the scenarios that were run, it appears that the clustering method based on the coefficients of the first eigenvector are the most significant improvement over the base case. Market Capitalization also appears to significantly increase the results of the cointegration trading methodology.

Worthy of note is that the other methodologies employed in clustering also enjoyed many instances of outperformance relative to the base case, implying that the information extracted through clustering has allowed for the cointegration based trading to run more effectively. Since the job of evaluating different clustering versions in cointegration can not be done by reasoning about the characteristics of the features, our empirical study provides us with a quantitative method of measuring their power. It appears that clustering is an effective method to improve cointegration and furthermore, some versions of clusterings with a particular set of parameters perform better than the others. Although reasoning and comparing these features is not interpretable and justifiable, our inituiton about the relative performance of the 1st eigenvector's coefficients is that similar coefficient contribute similarly to the covariance structure of the data set. Therefore, there is potentially powerful information in these coefficients and can create natural groups within the data set.

Additionally, cointegration was evaluated under more restrictive scenarios, namely an ADF statistic of -4.25 and -4.50. Under these more restrictive ADF statistics, there was strictly monotonic increases in performance accross all scenarios, further justifying the merit of cointegration based trading methodologies. Please refer to the Appendix for complete summary statistics.

4 Conclusions

In this study, we propose a method for combining clustering and cointegration in stock trading, and empirically analyze its power and show that our strategy indeed provide a return significantly superior to the return we would achieve with our baseline strategy employing cointegration without clustering. Furthermore, we explore one novel feature that has significant positive effects on our trading strategy: the greatest-coefficient eigenvector of the covariance matrix of stock returns.

As the space for exploration in this problem domain is huge, our approach contains some limitations. First of all, the learning part of our trading program uses data from a certain time period (the past three years), and thus may discover relations that may not be persistent for a different or longer set of time. However, since we focus more on short-term trading, this problem can be ignored as we assume that in the reasonably near future, all interactions between stocks remain relatively unchanged. Besides, as our access to data is limited, we did not incorporate more information in our clustering method, such as earning reports, companies' investment, debt and so on. Moreover, we only use daily closing stock prices, which limits our results as real world trading occurs in the scale of minutes or even second. Data availability again restricts what can be done. Another limitation is that the exploration of different feature subsets is not sufficiently comprehensive, which suggests there may be other feature combinations and parameter values yielding higher profits that are not examined in the scope of this project.

Future extensions of this research therefore can address those current limitations.

In particular, more information about stocks can be incorporated and examined in the clustering module to explore more underlying stock correlations and result in greater profitability. A more systematic approach to explore different feature combinations can certainly be a potential separate project.

In summary, this project helps us gain hands-on experience on the applications of clustering and the myriad tasks of tuning, selecting and specifying the dimensions, parameters, computation methods, and so on. We also learn how to evaluate and test various versions of clustering empirically using our cointegration trading simulation. Additionally, we can conclude that in general, cointegration is an effective trading methodology, and that integrating additional information from other variables can significantly increase the performance of the method. However, just as with any statistical model the power of the variables can change substantially over time.

References

[1] Z. Bodie, A. Kane, and A.J. Marcus. *Investments Irwin/McGraw-Hill series in finance, insurance, and real estate*. Irwin/McGraw-Hill, 1999.

[2] P.L. Cooley, R.L. Roenfeldt, and N.K. Modani. Interdependence of Market Risk Measures. *The Journal of Business*, 50(3):356–363, 1977.

[3] R.F. Engle and CWJ Granger. Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, 55(2):251–276, 1987.

[4] M. Gavrilov, D. Anguelov, P. Indyk, and R. Motwani. Mining the stock market: Which measure is best. *Proc. of the 6th ACM SIGKDD*, 2000.

[5] M.C. Gupta and R.J. Huefner. A Cluster Analysis Study of Financial Ratios and Industry Characteristics. *Journal of Accounting Research*, 10(1):77–95, 1972.

[6] J.D. Hamilton. *Time series analysis*. Princeton University Press Princeton, NJ, 1994.

[7] T. Hastie, R. Tibshirani, J.H. Friedman, et al. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[8] K. Kassimatis and S.I. Spyrou. Stock and credit market expansion and economic development in emerging markets: further evidence utilizing cointegration analysis. *Applied Economics*, 33(8):1057–1064, 2001.

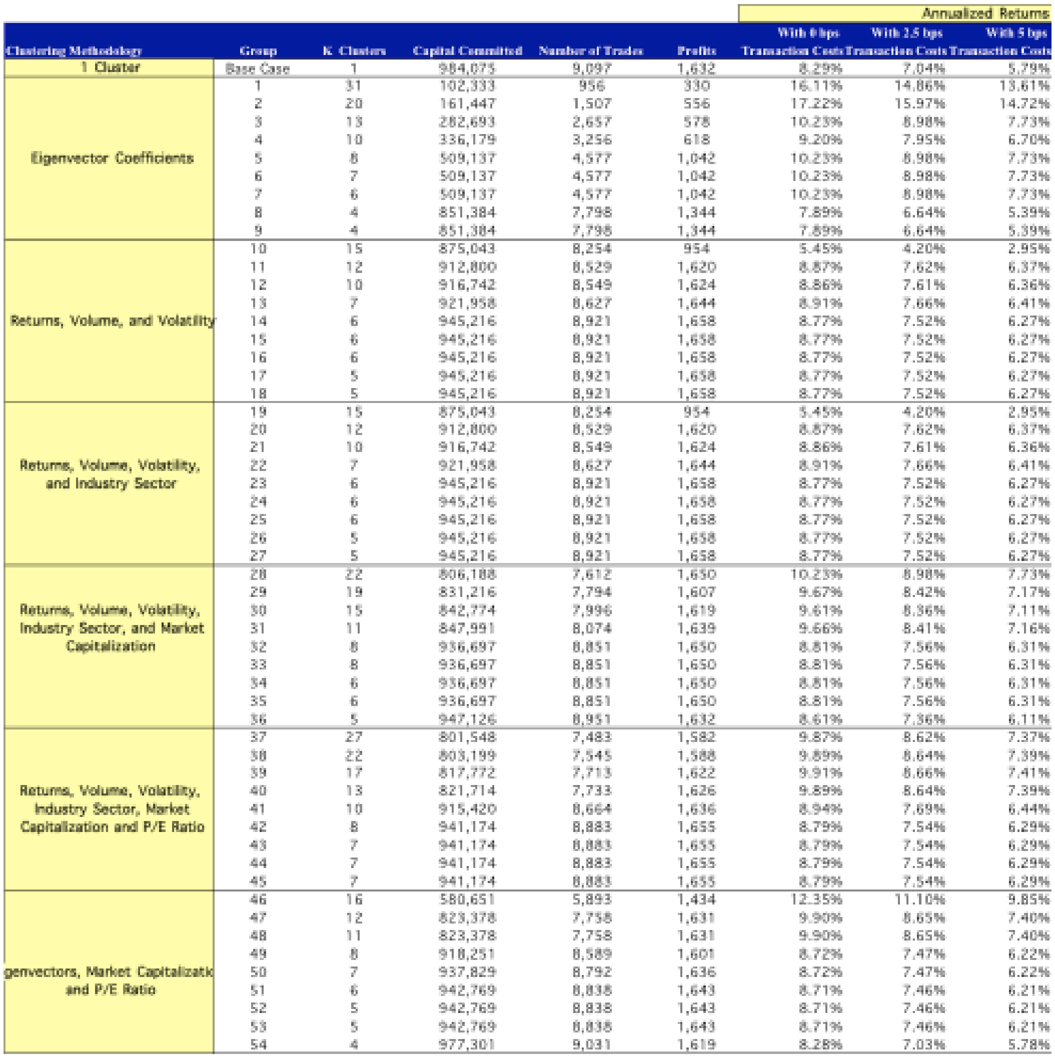
[9] E. Keogh and J. Lin. Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowledge and Information Systems*, 8(2):154–177, 2005.

[10] E. Porras. A test of cointegration between security markets of latin american nations (...). 2004.

[11] A.C. Rencher. *Methods of multivariate analysis*. Wiley New York, 1995.

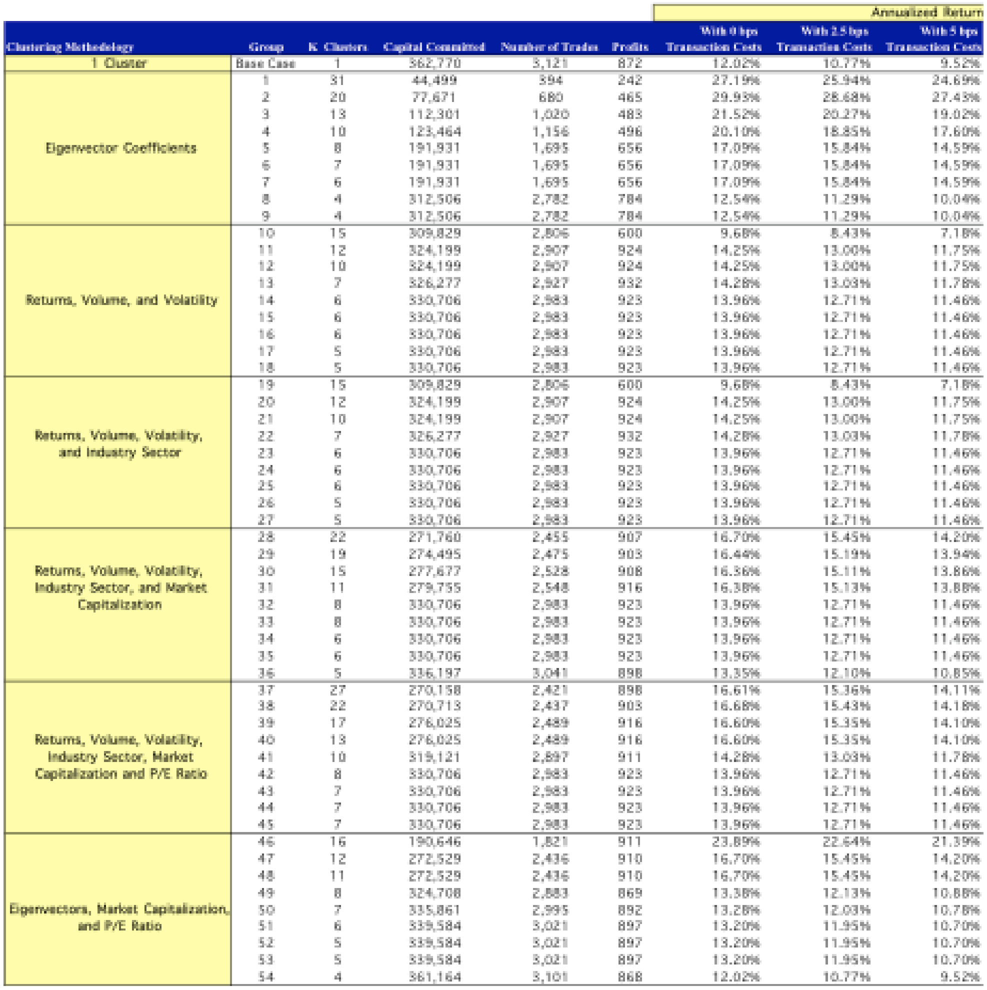
Appendix

Figure 2A. Summary statistics for and ADF threshold = -3.96 (.01 p-value)



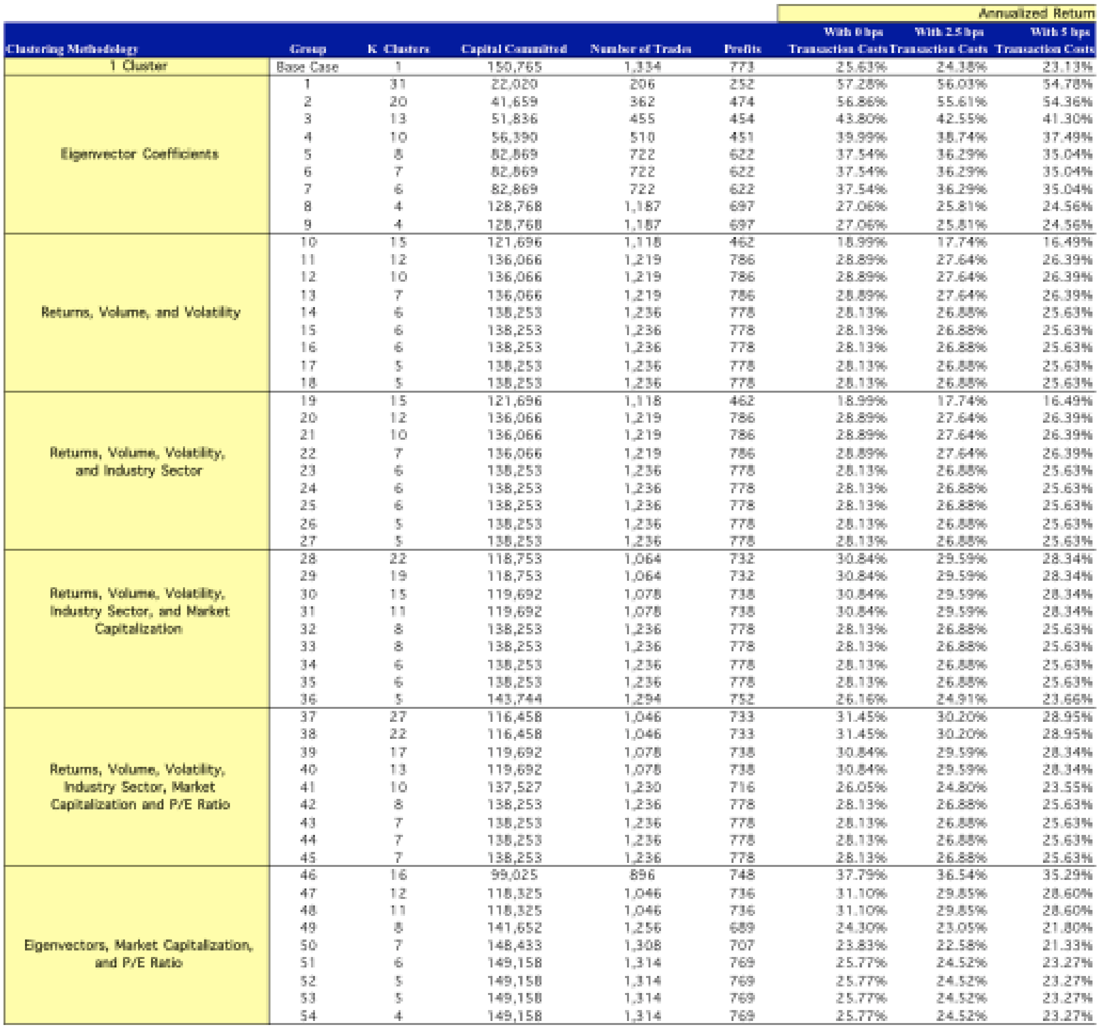
						Annualized Returns		
						With 0 bps	With 2.5 bps	With 5 bps
Clustering Methodology	Group	K Clusters	Capital Commited	Number of Trades	Profits	Transaction Costs	Transaction Costs	Transaction Costs
1 Cluster	Base Case	1	984,075	9,097	1,632	8.29%	7.04%	5.79%
Eigenvector Coefficients	1	31	102,333	956	330	16.11%	14.86%	13.61%
	2	20	161,447	1,507	556	17.22%	15.97%	14.72%
	3	13	282,693	2,657	578	10.23%	8.98%	7.73%
	4	10	336,179	3,256	618	9.20%	7.95%	6.70%
	5	8	509,137	4,577	1,042	10.23%	8.98%	7.73%
	6	7	509,137	4,577	1,042	10.23%	8.98%	7.73%
	7	6	509,137	4,577	1,042	10.23%	8.98%	7.73%
	8	4	851,384	7,798	1,344	7.89%	6.64%	5.39%
	9	4	851,384	7,798	1,344	7.89%	6.64%	5.39%
Returns, Volume, and Volatility	10	15	875,043	8,254	954	5.45%	4.20%	2.95%
	11	12	912,800	8,529	1,620	8.87%	7.62%	6.37%
	12	10	916,742	8,549	1,624	8.86%	7.61%	6.36%
	13	7	921,958	8,627	1,644	8.91%	7.66%	6.41%
	14	6	945,216	8,921	1,658	8.77%	7.52%	6.27%
	15	6	945,216	8,921	1,658	8.77%	7.52%	6.27%
	16	6	945,216	8,921	1,658	8.77%	7.52%	6.27%
	17	5	945,216	8,921	1,658	8.77%	7.52%	6.27%
	18	5	945,216	8,921	1,658	8.77%	7.52%	6.27%
Returns, Volume, Volatility, and Industry Sector	19	15	875,043	8,254	954	5.45%	4.20%	2.95%
	20	12	912,800	8,529	1,620	8.87%	7.62%	6.37%
	21	10	916,742	8,549	1,624	8.86%	7.61%	6.36%
	22	7	921,958	8,627	1,644	8.91%	7.66%	6.41%
	23	6	945,216	8,921	1,658	8.77%	7.52%	6.27%
	24	6	945,216	8,921	1,658	8.77%	7.52%	6.27%
	25	6	945,216	8,921	1,658	8.77%	7.52%	6.27%
	26	5	945,216	8,921	1,658	8.77%	7.52%	6.27%
	27	5	945,216	8,921	1,658	8.77%	7.52%	6.27%
Returns, Volume, Volatility, Industry Sector, and Market Capitalization	28	22	806,188	7,612	1,650	10.23%	8.98%	7.73%
	29	19	831,216	7,794	1,607	9.67%	8.42%	7.17%
	30	15	842,774	7,996	1,619	9.61%	8.36%	7.11%
	31	11	847,991	8,074	1,639	9.66%	8.41%	7.16%
	32	8	936,697	8,851	1,650	8.81%	7.56%	6.31%
	33	8	936,697	8,851	1,650	8.81%	7.56%	6.31%
	34	6	936,697	8,851	1,650	8.81%	7.56%	6.31%
	35	6	936,697	8,851	1,650	8.81%	7.56%	6.31%
	36	5	947,126	8,951	1,632	8.61%	7.36%	6.11%
Returns, Volume, Volatility, Industry Sector, Market Capitalization and P/E Ratio	37	27	801,548	7,483	1,582	9.87%	8.62%	7.37%
	38	22	803,199	7,545	1,588	9.89%	8.64%	7.39%
	39	17	817,772	7,713	1,622	9.91%	8.66%	7.41%
	40	13	821,714	7,733	1,626	9.89%	8.64%	7.39%
	41	10	915,420	8,664	1,636	8.94%	7.69%	6.44%
	42	8	941,174	8,883	1,655	8.79%	7.54%	6.29%
	43	7	941,174	8,883	1,655	8.79%	7.54%	6.29%
	44	7	941,174	8,883	1,655	8.79%	7.54%	6.29%
	45	7	941,174	8,883	1,655	8.79%	7.54%	6.29%
genvectors, Market Capitalizatic and P/E Ratio	46	16	580,651	5,893	1,434	12.35%	11.10%	9.85%
	47	12	823,378	7,758	1,631	9.90%	8.65%	7.40%
	48	11	823,378	7,758	1,631	9.90%	8.65%	7.40%
	49	8	918,251	8,589	1,601	8.72%	7.47%	6.22%
	50	7	937,829	8,792	1,636	8.72%	7.47%	6.22%
	51	6	942,769	8,838	1,643	8.71%	7.46%	6.21%
	52	5	942,769	8,838	1,643	8.71%	7.46%	6.21%
	53	5	942,769	8,838	1,643	8.71%	7.46%	6.21%
	54	4	977,301	9,031	1,619	8.28%	7.03%	5.78%

Figure 2B. Summary statistics for and ADF threshold = -4.25 (smaller than .01 p-value)



						Annualized Return		
Clustering Methodology	Group	K Clusters	Capital Committed	Number of Trades	Profits	With 0 bps Transaction Costs	With 2.5 bps Transaction Costs	With 5 bps Transaction Costs
1 Cluster	Base Case	1	362,770	3,121	872	12.02%	10.77%	9.52%
Eigenvector Coefficients	1	31	44,499	394	242	27.19%	25.94%	24.69%
	2	20	77,671	680	465	29.93%	28.68%	27.43%
	3	13	112,301	1,020	483	21.52%	20.27%	19.02%
	4	10	123,464	1,156	496	20.10%	18.85%	17.60%
	5	8	191,931	1,695	656	17.09%	15.84%	14.59%
	6	7	191,931	1,695	656	17.09%	15.84%	14.59%
	7	6	191,931	1,695	656	17.09%	15.84%	14.59%
	8	4	312,506	2,782	784	12.54%	11.29%	10.04%
	9	4	312,506	2,782	784	12.54%	11.29%	10.04%
Returns, Volume, and Volatility	10	15	309,829	2,806	600	9.68%	8.43%	7.18%
	11	12	324,199	2,907	924	14.25%	13.00%	11.75%
	12	10	324,199	2,907	924	14.25%	13.00%	11.75%
	13	7	326,277	2,927	932	14.28%	13.03%	11.78%
	14	6	330,706	2,983	923	13.96%	12.71%	11.46%
	15	6	330,706	2,983	923	13.96%	12.71%	11.46%
	16	6	330,706	2,983	923	13.96%	12.71%	11.46%
	17	5	330,706	2,983	923	13.96%	12.71%	11.46%
	18	5	330,706	2,983	923	13.96%	12.71%	11.46%
Returns, Volume, Volatility, and Industry Sector	19	15	309,829	2,806	600	9.68%	8.43%	7.18%
	20	12	324,199	2,907	924	14.25%	13.00%	11.75%
	21	10	324,199	2,907	924	14.25%	13.00%	11.75%
	22	7	326,277	2,927	932	14.28%	13.03%	11.78%
	23	6	330,706	2,983	923	13.96%	12.71%	11.46%
	24	6	330,706	2,983	923	13.96%	12.71%	11.46%
	25	6	330,706	2,983	923	13.96%	12.71%	11.46%
	26	5	330,706	2,983	923	13.96%	12.71%	11.46%
	27	5	330,706	2,983	923	13.96%	12.71%	11.46%
Returns, Volume, Volatility, Industry Sector, and Market Capitalization	28	22	271,760	2,455	907	16.70%	15.45%	14.20%
	29	19	274,495	2,475	903	16.44%	15.19%	13.94%
	30	15	277,677	2,528	908	16.36%	15.11%	13.86%
	31	11	279,755	2,548	916	16.38%	15.13%	13.88%
	32	8	330,706	2,983	923	13.96%	12.71%	11.46%
	33	8	330,706	2,983	923	13.96%	12.71%	11.46%
	34	6	330,706	2,983	923	13.96%	12.71%	11.46%
	35	6	330,706	2,983	923	13.96%	12.71%	11.46%
	36	5	336,197	3,041	898	13.35%	12.10%	10.85%
Returns, Volume, Volatility, Industry Sector, Market Capitalization and P/E Ratio	37	27	270,158	2,421	898	16.61%	15.36%	14.11%
	38	22	270,713	2,437	903	16.68%	15.43%	14.18%
	39	17	276,025	2,489	916	16.60%	15.35%	14.10%
	40	13	276,025	2,489	916	16.60%	15.35%	14.10%
	41	10	319,121	2,897	911	14.28%	13.03%	11.78%
	42	8	330,706	2,983	923	13.96%	12.71%	11.46%
	43	7	330,706	2,983	923	13.96%	12.71%	11.46%
	44	7	330,706	2,983	923	13.96%	12.71%	11.46%
	45	7	330,706	2,983	923	13.96%	12.71%	11.46%
Eigenvectors, Market Capitalization, and P/E Ratio	46	16	190,646	1,821	911	23.89%	22.64%	21.39%
	47	12	272,529	2,436	910	16.70%	15.45%	14.20%
	48	11	272,529	2,436	910	16.70%	15.45%	14.20%
	49	8	324,708	2,883	869	13.38%	12.13%	10.88%
	50	7	335,861	2,995	892	13.28%	12.03%	10.78%
	51	6	339,584	3,021	897	13.20%	11.95%	10.70%
	52	5	339,584	3,021	897	13.20%	11.95%	10.70%
	53	5	339,584	3,021	897	13.20%	11.95%	10.70%
	54	4	361,164	3,101	868	12.02%	10.77%	9.52%

Figure 2C. Summary statistics for and ADF threshold = -4.5 (smaller than .01 p-value)



						Annualized Return		
Clustering Methodology	Group	K Clusters	Capital Committed	Number of Trades	Profits	With 0 bps Transaction Costs	With 2.5 bps Transaction Costs	With 5 bps Transaction Costs
1 Cluster	Base Case	1	150,765	1,334	773	25.63%	24.38%	23.13%
Eigenvector Coefficients	1	31	22,020	206	252	57.28%	56.03%	54.78%
	2	20	41,659	362	474	56.86%	55.61%	54.36%
	3	13	51,836	455	454	43.80%	42.55%	41.30%
	4	10	56,390	510	451	39.99%	38.74%	37.49%
	5	8	82,869	722	622	37.54%	36.29%	35.04%
	6	7	82,869	722	622	37.54%	36.29%	35.04%
	7	6	82,869	722	622	37.54%	36.29%	35.04%
	8	4	128,768	1,187	697	27.06%	25.81%	24.56%
	9	4	128,768	1,187	697	27.06%	25.81%	24.56%
Returns, Volume, and Volatility	10	15	121,696	1,118	462	18.99%	17.74%	16.49%
	11	12	136,066	1,219	786	28.89%	27.64%	26.39%
	12	10	136,066	1,219	786	28.89%	27.64%	26.39%
	13	7	136,066	1,219	786	28.89%	27.64%	26.39%
	14	6	138,253	1,236	778	28.13%	26.88%	25.63%
	15	6	138,253	1,236	778	28.13%	26.88%	25.63%
	16	6	138,253	1,236	778	28.13%	26.88%	25.63%
	17	5	138,253	1,236	778	28.13%	26.88%	25.63%
	18	5	138,253	1,236	778	28.13%	26.88%	25.63%
Returns, Volume, Volatility, and Industry Sector	19	15	121,696	1,118	462	18.99%	17.74%	16.49%
	20	12	136,066	1,219	786	28.89%	27.64%	26.39%
	21	10	136,066	1,219	786	28.89%	27.64%	26.39%
	22	7	136,066	1,219	786	28.89%	27.64%	26.39%
	23	6	138,253	1,236	778	28.13%	26.88%	25.63%
	24	6	138,253	1,236	778	28.13%	26.88%	25.63%
	25	6	138,253	1,236	778	28.13%	26.88%	25.63%
	26	5	138,253	1,236	778	28.13%	26.88%	25.63%
	27	5	138,253	1,236	778	28.13%	26.88%	25.63%
Returns, Volume, Volatility, Industry Sector, and Market Capitalization	28	22	118,753	1,064	732	30.84%	29.59%	28.34%
	29	19	118,753	1,064	732	30.84%	29.59%	28.34%
	30	15	119,692	1,078	738	30.84%	29.59%	28.34%
	31	11	119,692	1,078	738	30.84%	29.59%	28.34%
	32	8	138,253	1,236	778	28.13%	26.88%	25.63%
	33	8	138,253	1,236	778	28.13%	26.88%	25.63%
	34	6	138,253	1,236	778	28.13%	26.88%	25.63%
	35	6	138,253	1,236	778	28.13%	26.88%	25.63%
	36	5	143,744	1,294	752	26.16%	24.91%	23.66%
Returns, Volume, Volatility, Industry Sector, Market Capitalization and P/E Ratio	37	27	116,458	1,046	733	31.45%	30.20%	28.95%
	38	22	116,458	1,046	733	31.45%	30.20%	28.95%
	39	17	119,692	1,078	738	30.84%	29.59%	28.34%
	40	13	119,692	1,078	738	30.84%	29.59%	28.34%
	41	10	137,527	1,230	716	26.05%	24.80%	23.55%
	42	8	138,253	1,236	778	28.13%	26.88%	25.63%
	43	7	138,253	1,236	778	28.13%	26.88%	25.63%
	44	7	138,253	1,236	778	28.13%	26.88%	25.63%
	45	7	138,253	1,236	778	28.13%	26.88%	25.63%
Eigenvectors, Market Capitalization, and P/E Ratio	46	16	99,025	896	748	37.79%	36.54%	35.29%
	47	12	118,325	1,046	736	31.10%	29.85%	28.60%
	48	11	118,325	1,046	736	31.10%	29.85%	28.60%
	49	8	141,652	1,256	689	24.30%	23.05%	21.80%
	50	7	148,433	1,308	707	23.83%	22.58%	21.33%
	51	6	149,158	1,314	769	25.77%	24.52%	23.27%
	52	5	149,158	1,314	769	25.77%	24.52%	23.27%
	53	5	149,158	1,314	769	25.77%	24.52%	23.27%
	54	4	149,158	1,314	769	25.77%	24.52%	23.27%

Contribution credits:

1. Proposal: Ozkul, Duong, and Ahmad (writing).
2. Progress report: Duong and Ozkul (writing)
   Clustering module implementation: Duong
   Cointegration implementation: Ozkul
3. Final paper: Duong and Ozkul (writing)
   Hierarchical clustering implementation: Duong
   Spectral clustering implementation: Ahmad (not included, due to lack of write-up and explanations)
   Cointegration implementation: Ozkul
4. Poster presentation: Ozkul and Duong