Investor Regime Analysis using Self-Organising Maps and Hierarchical Clustering

---

Quant Awards 2019

Abstract

This study investigates how modern machine learning techniques can be used to cluster equity flows to define investor regimes and inform the creation of a hedge fund style investment model. Equity flow data is analysed using the artificial neural network technology and visualisation tool known as self-organising maps. This analysis informs the number of investor regimes to look for, the optimal number was found to be four. The Hierarchical clustering algorithm and dynamic-time warping distance measure are implemented to determine four investor regimes. These regimes are characterised by stability and average weekly returns. The results of which informed the creation of a portfolio model. The performance of the investment model is evaluated by comparing it to a risk-free rate. The portfolio is compared quantitatively to one created by repeating the methodology, relying solely on return data to inform the regimes in this iteration. The model created using equity flows outperformed the one made by clustering returns. performance was measured using the Sharpe ratio. The model created using equity flows was calculated to have a Sharpe ratio of 0.24 while the model created using the analysis of return data is shown to have a Sharpe ratio of -0.21. By examining the probability matrices of both models, we see that the regimes created by clustering equity flow data are shown to be more stable than the regimes created using return data.

Introduction

In recent decades, advanced statistical techniques and machine learning technologies have revolutionized how we process and analyse data (Domingos 2012). The term Machine Learning (ML) was coined by IBM in the 1970's, it refers to the field of study involving machines and computer programmes capable of performing useful tasks or gaining insight from data without being explicitly programmed to do so (Burkov 2019). The development of easy to use programming languages such as Python and R has contributed to the wide usage of ML algorithms. R is the primary tool used in this study to carry our data cleaning, analysis and the implementation of the machine learning algorithms. This study looks at how modern ML clustering and visualisation techniques can be used to cluster equity flows to define investor regimes. Equity flow data is clustered using hierarchical clustering and dynamic time warping. Intuitive explanations and definitions of these terms will be provided.

The word 'learning' can be deceiving as machines cannot learn the same ways in which humans do. The purpose of this catchy name was to encourage further research into this area, push the best talent to work for IBM and impress clients (Domingos 2012). These modern techniques have the potential to revolutionise the quantitative investing industry. This study discusses and tests how ML assists in portfolio management in the modern day and the potential for financial applications in the future.

Portfolio managers have many tools and techniques at their disposal (Becker and Reinganum 2018, Kahn 2018). One of many effective strategies is the consideration of what "investor regime" the market resides in to assist in making investment decisions. An "investor regime" refers to a state the market is in, where equity is moving and what regions are exhibiting high, low or neutral returns (Ang, 2004).

This study relies on the established principle used by investors that if the market resides in a certain regime in each week then it is most likely to be in the same regime in the following week. These regimes inform the creation of a portfolio informed by investor regimes, the details of which are included. The same strategy is carried out again but this time clustering the regional return data to define regimes. We examine how profitable the resulting investment model is by comparing it to a risk-free rate. One can compare the use of equity flow data and return data when defining regimes and investigate the validity of using this modern technology to inform international portfolio management.

Self-organising maps and hierarchical clustering are used in tandem with the dynamic time warping distance measure are used on return and equity flow data to determine investor regimes and inform the creation of a profitable investment model. The results and methodology of this experiment are given in detail. We begin with a brief and intuitive explanation to the terms; Machine learning, supervised learning, unsupervised learning, self-organising maps, hierarchical clustering and dynamic time warping. Following on from that this study will describe how these techniques can be used to inform international portfolio diversification decisions.

Literature Review

Portfolio Management and Investor Regimes

Modern portfolio theory states that diversification of security returns with lower correlation should yield more favourable results for an investor (Levy & Sarnat, 1970). There exists a variety of different approaches that an investor can take to diversify a portfolio, including diversification by sector and by country or region. In a 1970 paper by Levy and Sarnat discusses the high degree of correlation between security returns in a single economy and presents the benefits of diversifying assets internationally in comparison to holding assets across different industries domestically (Levy & Sarnat, 1970). For many years international diversification has been an established portfolio management strategy and has grown more popular in recent decades (Hitt, 2006).

Equity Flows

Equity flows is the change in asset allocation across sector, countries or regions. Change of flows occur when investor move, buy or sell assets. It is established that they can be used to assist in forecasting future equity returns (Froot 2001). A 2001 study by Froot, O Connell and Seasholes showed their persistent nature, meaning that equity flows in general are less volatile and more reliable than returns (Froot 2001). This study uses machine learning techniques to analyse equity flow data to determine investor regimes and creating a portfolio that tracks investor behaviour in those regimes. We examine the hypothesis that equity flow data is more persistent/stable than that of equity return data by forming market regimes with each and creating transition/probability matrices of how the regimes change or stay in each regime from week to week.

Types of Learning

There exists a variety of different learning methods in which ML can be achieved. The three main types of learning will be outlined very briefly here. First, we consider the area of **Supervised learning**, where the dataset used in the algorithm must be labelled data (Burkov 2019). Each element of the data set can contain various facets of information. A single datapoint might refer to a person and each person might have several features such as gender, height, weight ect (Burkov 2019). In the context of this study, each data point is a week and the features of that data point are the equity flow values in that week. When implementing **unsupervised** learning, the dataset is a collection of unlabelled examples (Burkov 2019). **Reinforcement learning/ competitive learning** lies between supervised learning and unsupervised learning. It operates through continuing interactions between a learning system and the environment (Haykin, 2009), competitive learning is the learning type of main relevance in this study.

Self-Organising Maps

The first ML analysis tool used in this study is that of the **self-organising map**. A self-organising map (SOM), is a type of artificial neural network (Kohonen, 1990). They can reduce the dimensionality of data thus making them useful for visualization (Kohonen, 1990). Prof Teuvo Kohonen developed this data analysis technique in the 1980's (Kohonen, 1990). An established use for this technology is in the area of exploratory analysis, in examining the structure and finding patterns in large datasets (Kaski, 1997). An effective exploratory analysis tool is essential for analysts working on large and complicated data sets. Analysis of these data sets can sometimes be difficult and time-consuming (Kaski, 1997).

Like most ANN's, SOM's operate in two modes; training and mapping (Kohonen, 1990). The training part of the operation involves building the map using input examples. The mapping part of the operation automatically classifies a new input vector (Kohonen, 1990. This form of training is known as **competitive learning**. When a training example is fed to the network, its Euclidean distance to all weight vectors is computed. The Euclidean distance between two time series, V and W;

$$V = v_1, v_2, \ldots \ldots v_n \qquad (1)$$
$$W = w_1, w_2, \ldots \ldots w_n \qquad (2)$$

by the following formula

$$E(v, w) = \sum_{i=1}^{n} (v_i - w_i)^2 \qquad (3)$$

The map space is pre-defined before the training process. The space consists of nodes arranged in a rectangular or hexagonal grid; the dimensions of this grid are pre-set. SOM's can be helpful in the area of data visualisation. Data science is more than just building machine learning models; it's also about explaining the models and using them to drive data-driven decisions. Displaying data in an informative and visually appealing way can play a very important role of presenting data in a powerful and credible way.

Data

Equity flow data

The primary data set used in this study is that of the equity flow data. This was provided by State Street Global markets. The data are derived from data held by State Street Bank & Trust (SSB). SSB the largest mutual fund custodian in the US and hold roughly 40% of the industry's funds under custody (Froot, 2001). An approximate estimate to the quantity of assets under custody by SSB is $6 trillion. The period selected for analysis was from 7th of January 2012 to 18th of August 2018. The dataset was originally in daily format but was transformed to weekly in the statistical software; R. A breakdown of the regions can be found in Appendix I.

MSCI and LIBOR

The **MSCI Index** is a measurement of stock market performance in a particular area, it is the industry's accepted gauge of global stock market activity. The weekly MSCI data was downloaded from Bloomberg.com. The MSCI indices were used to calculate the total regional weekly returns.

The risk-free rate used in this study is the LIBOR dollar rate. It is the average interest rate at which leading banks borrow funds from other banks in the London market. It is a widely used global "benchmark" or reference rate for short term investments.

Methodology

The first task carried out during this study was to choose an appropriate number of investor regimes to look for. The exploratory analysis tool used was the SOM. Once the optimal number of regimes were determined, the weekly equity flow data was clustered using the hierarchical clustering algorithm and the dynamic time warping distance measure. Four investor regimes were determined and were characterised by average weekly returns and their stability. A portfolio model was constructed based on the analysis of these four regimes.

Number of regime selection by SOM

This helpful technique allows one to get an overall sense of a large dataset and quickly observe what the results look like when different numbers of regimes are chosen. By examining the resulting maps.

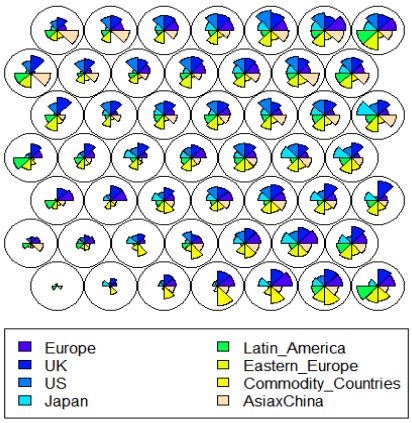


Figure 1: Example of 7x7 SOM, produced in R using equity flow data

In figure 1, This technique allows an analyst to get an overview of the dataset in question. Shown below is an example of a self-organising map produced using the total flow dataset, its smaller size makes it easier to read. Inside each node/circle there is a coloured wedge representing the magnitude of equity flow in a certain region. This magnitude can be compared by their relative size. For example, the top right corner of the SOM is populated by nodes representing weeks of high equity flow into all regions whereas in the bottom left corner, the size of the wedges are small which represent weeks of low or negative equity flow across the regions.

Determining Investor Regimes using Hierarchical Clustering and Dynamic-Time Warping

There exists a multitude of different **ML algorithms**, all possess individual purposes and advantages. The ML algorithm in focus here is that of **hierarchical clustering** (Steinbach, 2000). Hierarchical clustering either falls into the top-down or bottom-up category. The method used for these clusters was a 'bottom-up' method. Bottom-up algorithms treat each data point as a single cluster in the beginning and then merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all data points (Steinbach, 2000). Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC.

Dynamic time warping (DTW) was the selected distance measure for this study. Once a clustering algorithm is chosen the next task is to select an appropriate distance measure. A **distance measure** defines what is the measure of similarity or dissimilarity between two data points. An example which highlights the importance of a distance measure is if you wished to write a program which calculated the time taken to get to a destination in a car. The most standard measure of similarity is that of the Euclidean distance. This measure, in terms of the car example, would calculate the 'birds' eye' view distance between two points on a map. In real life this is not a good measure of the distance a car must travel, as it must stay on roads and in some cases roads way have one-way systems.

Each distance measure has different specifications of what defines a cluster, so a certain clustering distance measures might be preferred depending on what types of clusters one wishes to obtain. Time-Series data can pose some challenges, partly due to factors like large size and dimensionality. A first important issue is to decide whether clustering must be governed by a "shape-based" or "structure-based" dissimilarity concept.

DTW is a 'shape-based' clustering algorithm (Berndt, Clifford 1994). It clusters together time series that have similar shapes. Consider time series S and T;

$$S = s_1, s_2, \ldots \ldots s_n \qquad (4)$$
$$T = t_1, t_2, \ldots \ldots t_n \qquad (5)$$

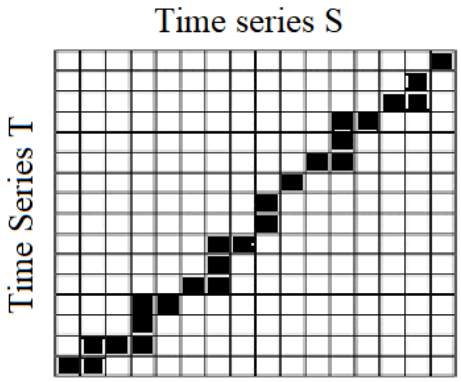
The DTW measure computes the difference between a point on S to every other point on T. It then iteratively does this to every point on S and creates a matrix out of these values. The **warping path** is the path taken to get from the lower left matrix entry up to the upper right matrix entry (Berndt, Clifford 1994). The DTW algorithms clusters together time series that are

computed to have the smallest warping path, w, between them. This can be expressed as;

$$DTW(S,T) = min_W\left[\sum_{k=1}^{p} \partial(w_k)\right] \quad (6)$$

An example of a warping path between two time series is given below in figure 2.

Figure 2: Small example SOM, produced in R using equity flow data.



An advantage of dynamic time warping is that having dates out of sync across time series will not affect the results of clusters obtained. This can prove particularly useful when analysing international time-series data where time zones can cause differences in date/times of close of market (Berndt, Clifford 1994). Dynamic time warping simply considers the overall shape of the time series when measuring similarity. This can save time in data cleaning, getting date and times to match up exactly across many time-series can be time consuming and thus costly for companies (Berndt, Clifford 1994).

Exploratory Analysis using SOMs

The first step of the analysis process involved the equity flow dataset. Exploratory analysis was primarily carried out using SOMs, as discussed previously, this has been established as a useful tool for this stage of analysis. This technique allows an analyst to get an overview of the dataset in question. Shown below is an example of a self-organising map produced using the total flow dataset, it's smaller size makes it easier to read.

This type of investment model relies on the principle that if the market is in a certain regime one week then it will most likely be in the same regime in the next week. This theory is examined with the use of probability matrices for each set of clusters found, we can see what the probability is for the market to change regime or to stay in the same regime.

The Sharpe ratio allows investors to compare the return of an investment to its risk. In general, the higher the Sharpe ratio, the more attractive the portfolio is to an investor. The recognized Sharpe ratio is

$$S = \frac{R_p - R_f}{\sigma_p} \quad (7)$$

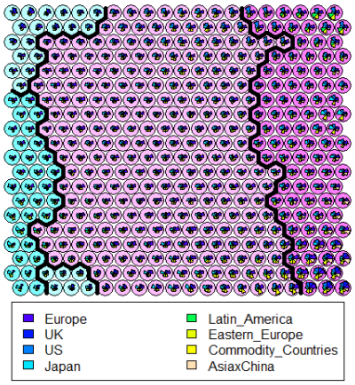
R_p is the return of the portfolio, R_f is the risk-free rate and σ_p is the standard deviation of the portfolio's excess annualised return.

Results and Discussion

Exploratory Analysis using SOMs

Shown overhead in figure 3, is a SOMs produced in Rstudio, using the kohonen package using the equity flow data. This analysis techniques allows one to get a general sense of the overall structure of the dataset. The SOM grid consists of many circular nodes, one can set the desired number of nodes depending on the size and nature of the dataset. The below grids consist of 400 nodes (20x20). Inside each node, there are eight wedges of varying size, each wedge representing the magnitude of equity flow. There is approximately 350 data points of average weekly equity flow from 2012 to the present day. The above SOM was trained with a dataset of similar magnitude to the number of nodes it possesses. For example, in the bottom right of both figures one can observe that each of these nodes correspond to weeks during this period where the equity flows across all regions are of a large positive magnitude. Each node roughly corresponds to a week during this period. A disadvantage to a SOM of this size is that is can cause the nodes to be difficult to read.

Figure 3: Full size SOM produced in R using equity flow data from 2012-2018.



Four distinct regions can be observed in figure 3. This was not the case in SOM produced where the number of clusters were set to be larger than four. This result informed the selection of four investor regimes moving forward in this research.

Characterising Regimes by Returns and Stability

Stability of Market Regimes

Probability matrices are also known as transition matrices, they display the probability of transitioning from one state to another. The following probability matrices illustrate the differences between using equity flow data to and return data to define market regimes. The two probability matrices shown in this section demonstrate the stability of the market regimes found created by first the equity flow data and secondly by utilising the return data.

Table 1: Probability matrix of market regimes created by equity data

	1	2	3	4
1	0.779	0.110	0.000	0.110
2	0.190	0.660	0.050	0.100
3	0.059	0.235	0.706	0.000
4	0.321	0.196	0.000	0.482

Table 2: Probability matrix of market regimes created by regional returns.

	1	2	3	4
1	0.152	0.261	0.326	0.261
2	0.126	0.336	0.263	0.274
3	0.139	0.278	0.306	0.278
4	0.115	0.229	0.364	0.292

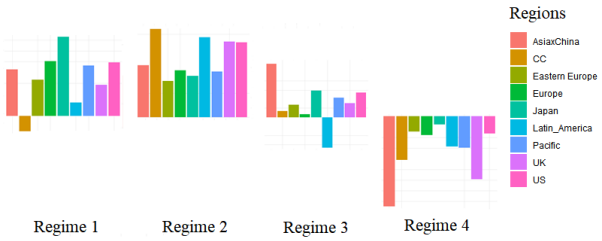
Each cell of the matrices represents the probability of the market changing to a certain regime or staying in the regime it is already in. The largest numbers in table 3 are those along the diagonal from the top left corner to the bottom right. This indicates that the regimes are relatively stable, if the market is in a certain regime then it is most likely to stay in that regime in the following week. This may be a contributing factor to the investment model created by equity flows outperforming the model created by return data.

This is not the case in table 4, where the numbers along the diagonal are not significantly larger than those in all other positions in the matrix. This shows the unstable nature of the market regimes created by the analysis of return data. This supports the hypothesis that equity data is more persistent and stable than equity return data.

Characterising Regimes by Average Weekly Returns

We examine the regimes obtained by clustering the total weekly equity flows from the period of January 2012 to July 2018. The average returns in each of the four regimes are found and inform the three investment models.

Figure 4: Return analysis of market regimes created by clustering regional equity flow



In the above plots, one can see that regime 1 and 2 are characterized by positive returns, regime 3 is generally low positive returns and returns are on average negative during regime 4 across all regions. The portfolio takes a long position in all regions except the commodity countries (neutral) and Latin America (short) during regime 1. In regime 2 the model takes a long position in all regions. In regime 3, the model

takes a long position in Asia, Japan, Pacific, UK and US. It takes a neutral position in all other regions except in Latin America where it takes a short position. For regime 4, the model takes a short position in all regions. Larger versions of this plot and the return analysis comparison can be found in appendix IV.

Figure 5: Cumulative returns of hedge fund model informed by analysis of regional returns

Figure 5 above shows how the portfolio outperforming the risk-free rate. This proves the concept of the potential of this technology to create inform the creation of profitable portfolios. The details of this model can be found below in table 1.

Table 1: Hedge fund model created from equity flow data performance figures compared to the risk-free rate.

	Hedge Fund Model	**Risk Free Rate**
Total Returns	7.21%	4.46%
Annual Cumulative Returns	1.054%	0.660%
Volitility	1.64%	-
Sharpe ratio	0.24	0

Conclusions

The self-organising map technology was a useful and effective tool in the exploratory analysis of the equity flow data. By repeatedly using this technology and changing the number of clusters one can quickly and easily determine what is appropriate number of clusters to use when implementing the more sophisticated clustering algorithm of hierarchical clustering with dynamic time warping distance measure. Many SOMs were produced in order to determine the optimal number of regimes to look for. An example of one of these maps produced using the equity flow data is given in figure 3 where four distinct regions can be observed. Thus for the next portion of the study, four regimes were determined and analysed to inform the portfolio.

The results of this study showed that the portfolio model results in higher returns when the equity flow data is used to create the market regimes. The model created by clustering equity flow data was calculated to have a Sharpe ratio of of 0.24 while the model created using the analysis of return data is shown to have a Sharpe ratio of -0.21. The full results of the analysis of creating a portfolio by relying only on return data t create the regimes can be found in appendix IV.

Furthermore, the probability matrix of market regimes produced from equity flow data, shows that these clusters are more stable than clusters produced from clustering return data. These results show some of the benefits of using equity flow data to inform portfolio management decisions and strengthens the hypothesis that equity flows are more persistent and stable to returns.

The potential to use modern machine learning techniques to create profitably investment models has been shown during this study. Furthermore, this study shows the advantages of determining investor regimes using equity flow data in comparison to using return data. This is shown in the fact that the resulting portfolio is more profitable and less risky in addition to the regimes being more reliable. Self-organising maps were helpful in the exploratory analysis of large financial datasets and assisted in the selection of an appropriate number of market regimes to define going forward in the research. Hierarchical clustering used in conjunction with dynamic time warping were successfully implemented to inform a regime portfolio.

References

Levy, H., & Sarnat, M. (1970). Diversification, portfolio analysis and the uneasy case for conglomerate mergers. *The journal of finance*, *25*(4), 795-802.

Hitt, M. A., Tihanyi, L., Miller, T., & Connelly, B. (2006). International diversification: Antecedents, outcomes, and moderators. Journal of Management, 32(6), 831-867.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464-1480.

Kaski, S. (1997). Data exploration using self-organizing maps. In *Acta polytechnica scandinavica: Mathematics, computing and management in engineering series no. 82*.

Ang, A., & Bekaert, G. (2004). How regimes affect asset allocation. *Financial Analysts Journal*, *60*(2), 86-99.

Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Andriy Burkov.

Berndt, D. J., & Clifford, J. (1994, July). Using dynamic time warping to find patterns in time series. In *KDD workshop* (Vol. 10, No. 16, pp. 359-370).

Becker, Y. L. and M. R. Reinganum (2018). "The Current State of Quantitative Equity Investing", CFA Institute Research Foundation.

Kahn, R. N. (2018). "The Future of Investment Management", CFA Institute Research Foundation.

Steinbach, M., Karypis, G., & Kumar, V. (2000, August). A comparison of document clustering techniques. In *KDD workshop on text mining* (Vol. 400, No. 1, pp. 525-526).

Froot, K. A., O'connell, P. G., & Seasholes, M. S. (2001). The portfolio flows of international investors. *Journal of financial Economics*, *59*(2), 151-193.

Appendix I – Region breakdown by countries

Countries were split up into the following regions during analysis. The data in question was supplied to the research team with these pre-set groups

Table 2: Region breakdown by country

Europe	EE	CC	Asia+	UK	US	Japan
France	Czech Republic	Australia	Hong Kong	UK	US	Japan
Austria	Hungary	Canada	Malaysia			
Belgium	Israel	Norway	Indonesia			
Denmark	Russia	New Zealand	Singapore			
Finland	Turkey		Thailand			
Greece	Czech Republic		Taiwan			
Ireland			South Korea			
Italy			Egypt			
Netherland						
Portugal						
Sweden						
Spain						
Germany						
Switzerland						

Appendix II – MSCI and Equity flow and Return Data Details by Region

Table 3: Return data breakdown by region

Countries	**Volitility**	**Average Total return**	**Anualised cumulative returns**
Europe	20.50%	16.76%	2.36%
Eastern Europe	19.16 %	11.75%	1.68%
Asia+	18.6%	12.98%	1.85%
Commodity Countries	23.24%	5.27%	0.77%
Latin America	21.51%	9.46%	1.37%
US	16.74%	22.18%	3.06%
UK	18.44%	12.82%	1.83%
Japan	15.67%	23.62%	3.24%

Appendix III – Results of determining regimes based on return data instead of equity flow data

Market regimes created by clustering returns

Here we examine the results from obtained by clustering the weekly returns from the period of January 2012 to July 2018. The average returns in each of the four regimes are found and inform the three investment models.

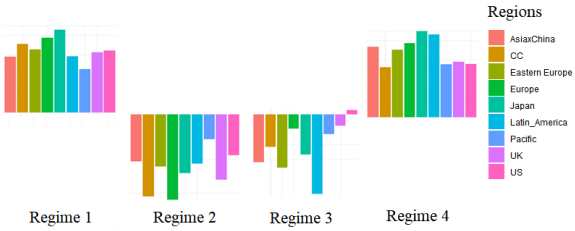


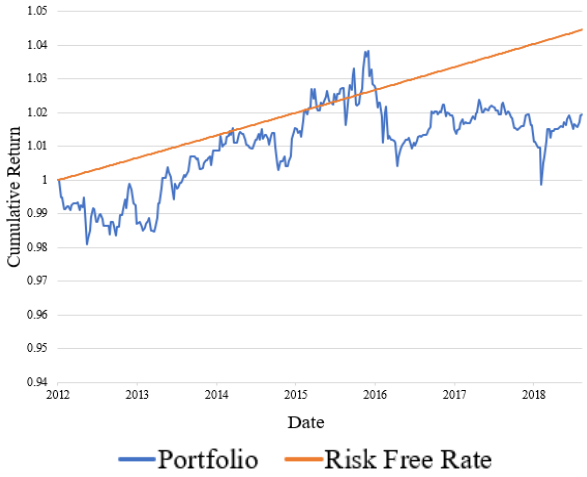
Figure 6: Return analysis of market regimes created by clustering regional returns.

In the above plots, one can see that regime 1 and 2 are characterized by positive returns, regime 3 is generally low positive returns and returns are on average negative during regime 4 across all regions. The model takes a long position in all regions during regime 1. In Regime 2 the model takes a short position in all regions. In Regime 3, the model takes a short position in all countries except Europe, pacific, UK and US which it takes a neutral position in. The model takes a long position in all regions during regime 4.

Table 4: Hedge fund model created from return data performance figures compared to the risk free rate.

	Portfolio	**Risk Free Rate**
Total Returns	1.0196	1.0446
Annual Cumulative Returns	0.29%	0.660%
Volility	1.79%	-
Sharpe ratio	-0.21	0

Figure 7: Cumulative returns of hedge fund model informed by analysis of equity flow data



Appendix IV – Enlarged plots of average equity returns

Figure 7:Average weekly regional returns of market regimes created by clustering equity flows

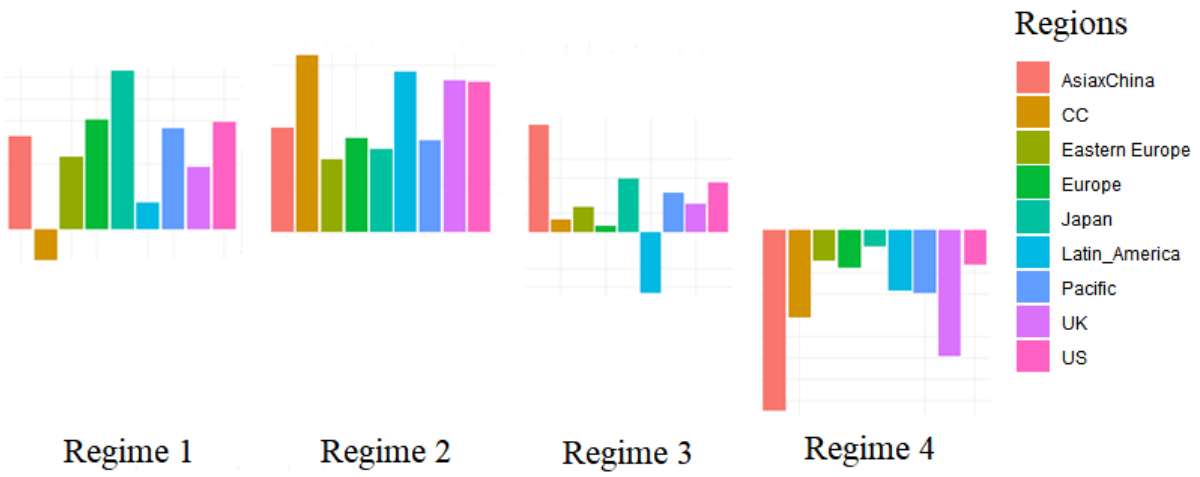


Figure 8:Average weekly regional returns of market regimes created by clustering return data