OFR STAFF DISCUSSION PAPER
OFFICE OF FINANCIAL RESEARCH

No. 2015-01 | January 23, 2015

Clustering Techniques and Their Effect on Portfolio Formation and Risk Analysis

Victoria Lemieux
University of British Columbia
Vancouver, Canada
vlemieux@mail.ubc.ca

Payam S. Rahmdel
University of British Columbia
Vancouver, Canada
rahmdel@magic.ubc.ca

Rick Walker
Middlesex University London
London, UK
r.walker@mdx.ac.uk

B.L. William Wong
Middlesex University London
London, UK
w.wong@mdx.ac.uk

Mark D. Flood
Office of Financial Research
mark.flood@treasury.gov

Clustering Techniques And their Effect on Portfolio Formation and Risk Analysis

Victoria Lemieux
University of British Columbia
Vancouver, Canada
vlemieux@mail.ubc.ca

Payam S.Rahmdel
University of British Columbia
Vancouver, Canada
rahmdel@magic.ubc.ca

Rick Walker
Middlesex University London
London, UK
r.walker@mdx.ac.uk

B.L. William Wong
Middlesex University London
London, UK
w.wong@mdx.ac.uk

Mark Flood
Office of Financial Research
Washington DC, USA
mark.flood@treasury.gov

ABSTRACT

This paper explores the application of three different portfolio formation rules using standard clustering techniques—K-means, K-mediods, and hierarchical—to a large financial data set (16 years of daily CRSP stock data) to determine how the choice of clustering technique may affect analysts' perceptions of the riskiness of different portfolios in the context of a prototype visual analytics system designed for financial stability monitoring. We use a two-phased experimental approach with visualizations to explore the effects of the different clustering techniques. The choice of clustering technique matters. There is significant variation among techniques, resulting in different "pictures" of the riskiness of the same underlying data when plotted to the visual analytics tool. This sensitivity to clustering methodolgy has the potential to mislead analysts about the riskiness of portfolios. We conclude that further research into the implications of portfolio formation rules is needed, and that visual analytics tools should not limit analysts to a single clustering technique, but instead should provide the facility to explore the data using different techniques.

General Terms

Algorithms, Design, Economics, Experimentation

Keywords

Clustering Techniques, Financial Stability Monitoring, Visual Analytics

1. INTRODUCTION

Monitoring threats to financial stability is not a single-input operation. It typically requires integration and analysis of numerous datasets, which may have many data points and many attribute dimensions. Techniques from data science may help alleviate some of the resulting information-processing burdens for macroprudential supervisors. Cluster analysis groups similar data objects into clusters where the classes or clusters have not been defined in advance [1]. A cluster of data objects is thus a form of data compression [2]. In finance, clustering can segment stock data into portfolios for additional analysis. In this paper we explore the application of different clustering techniques to a large financial data set (16 years of daily CRSP stock data) to determine how the choice of clustering technique might affect an analyst's perception of the riskiness of different portfolios in the context of a prototype visual analytics system designed for financial stability monitoring.

The results in this paper are preliminary. Our prototype visual analytics tool draws upon only two of many available approaches to the financial stability risk monitoring. A comparative analysis of different financial stability risk analysis methods is available at Bisias et al. [3]. Also, our clusterings of stocks do not use traditional portfolio allocation rules, which incorporate a wide range of practical goals and constraints, such as retirement planning, formal investment mandates, tax laws, etc.[4]. For present purposes, "portfolio" simply means a group of stocks that exhibit similar characteristics across a range of dimensions (e.g. ask price, bid price, returns, etc.). We apply standard clustering algorithms to daily equities data to determine these portfolios.

The point is to demonstrate the sensitivity of visual renderings of overall portfolio risk to plausible variation in the portfolio-selection rule. That is, the different clustering techniques place individual data objects (e.g., stocks) into different clusters (i.e. create different aggregations, or portfolios) from the same dataset. We hypothesize that this may affect analysts' perception of levels of risk for particular portfolios (and the data objects that comprise them) in the context of a visual analytics system—the RiskMapper tool—designed to support financial stability monitoring. This potential for misperception occurs because the clusters may place in dif-

SIGMOD '14, June 2014, Snowbird Utah, USA

ferent parts of the "magic quadrant" visualization in the tool.

Section 2 of the paper provides a review of background literature for the study. Section 3 discusses the two-phase methodology that we followed in our experiments as well as description of the dataset. Experimental results have been highlighted in Section 4 followed by our conclusion in Section 5

2. BACKGROUND LITERATURE

From the perspective of machine learning, data clustering is an unsupervised learning algorithm with the principal task of partitioning a set of unlabeled data objects into homogenous groups with a similar pattern. Cluster analysis is a fundamental operation in many data analysis and information retrieval procedures [5], [6], [7], [8], [9], [10]. In finance, clustering algorithms have been widely used in applications such as market segmentation [11], credit scoring [12], [13] and bankruptcy prediction [14], [15], [16], [17]. The pattern recognition community has proposed hundreds of clustering algorithms.[18], [19], [20], [21], [22].

Clustering techniques are heuristic approaches, and no two algorithms will generate the same results [23], [24]. In addition, no single method can identify every kind of cluster shape and structure. A number of possible solutions exist for the problem of cluster variability. Given a dataset, there are two main approaches:, 1) applying a set of different clustering algorithms; or 2) applying one clustering technique but with different parameter adjustments at each round. The results are typically stored in multiple cluster sets or a *cluster ensemble* [23]. It may be difficult to sort or filter multiple, theoretically plausible cluster datasets based on a priori quantitative measures or domain knowledge [24]. One solution to this problem is to identify the so-called "stable" clusters that consistently contain the same records across the results of different methods. Examples of such attempts to evaluate cluster ensembles include [25], [26], [27], [28].

3. EXPERIMENT METHODOLOGY

To determine the effects of different clustering techniques on how an analyst might perceive the riskiness of different portfolios in the context of our prototype visual analytics system, we designed a two-phased experiment. In phase one, we applied three different clustering techniques to samples of our data set to determine the degree of variability within and between portfolios formed from samples of the data. In the second phase, we imported a sample portfolio created in the first phase into our prototype visual analytics tool—the RiskMapper—to determine the effect of any variability on visualization of the riskiness of the portfolios.

3.1 Dataset

The CRSP database is a well-known comprehensive database for historical security prices and returns information. It contains various financial datasets such as US Stock, US Treasury, Historical Indexes, etc. [29]. In our experiments, we use the CRSP US Stock database, comprising daily market and corporate action data for securities with primary listings on the New York Stock Exchange (NYSE and NYSE Amex), NASDAQ Stock Market, and Archipelago Exchange. Our sample covers sixteen years' worth of daily data, from January 1998 to December 2013, for the full CRSP equities universe. This period encompasses the period where financial crisis of 2007-2009 occurred. This dataset contains more than 29 million records, i.e. data objects, of daily stock transactions.

3.2 Phase One Treatment

In Phase One, we generated portfolios with three different clustering algorithms and visualized the correlation between the clustering methods using Kosara et al.'s parallel sets [30] and traditional parallel coordinate plots. Because this is a data-driven approach, one might naively assume that different clustering methods would generate similar results, with slight variations from one to another. A macroprudential analyst is likely to focus on the response of the final risk calculations to exogenous economic factors, rather than on the impact of a technical choice of a portfolio-selection rule.

To determine these effects, we tested two standard partition-based clustering techniques, i.e. K-means and K-medoids (partition around medoids (PAM)) [31], and one hierarchical clustering technique against the dataset. In the first set of experiments, we extracted six subsets by random sampling of the original data, where each sample contains 10,000 records. For the implementation of the clustering algorithms, we appied R's built-in functions *kmeans*, *pam* and *hcluster*, respectively, using the default parameter settings for each method (e.g., squared Euclidean distance as the default distance measure in K-means), and setting the number of clusters at 7 as suggested by R's optimization function. The clusterings produced a four-dimensional vector in a cluster ensemble matrix, where each row corresponds to a particular stock and columns represent the stock's ID and cluster into which particular stocks had been placed using each of the three clustering techniques. This matrix was the input to the visualizations.

Another way of assessing variations between different clustering techniques is to treat the time dimension as an *independent* variable and observe the results in a single trading day. Instead of random sampling of the data, in the second set of Phase One experiments, we examined the cluster variations by choosing the data of a single day (we selected 4 Jan 2007 at random) as the input data and applied the same clustering algorithms and parameter settings to the data. The results appear in Section 4.1.

3.3 Phase Two Treatment

Phase Two entailed importing the sample portfolios created by the three clusterings from Phase One into the RiskMapper tool to see how the system projected the riskiness of each portfolio. For Phase Two, we used the Phase-One samples imported that treated the time dimension was treated as an independent variable, since time is a necessary input for computation of the risk measures used in the RiskMapper tool.

3.3.1 Description of RiskMapper tool

The RiskMapper is a prototype visual analytics tool [32] designed to aid macroprudential supervisors in monitoring financial stability. Visual analytics is defined as the "science of analytical reasoning facilitated by interactive visual

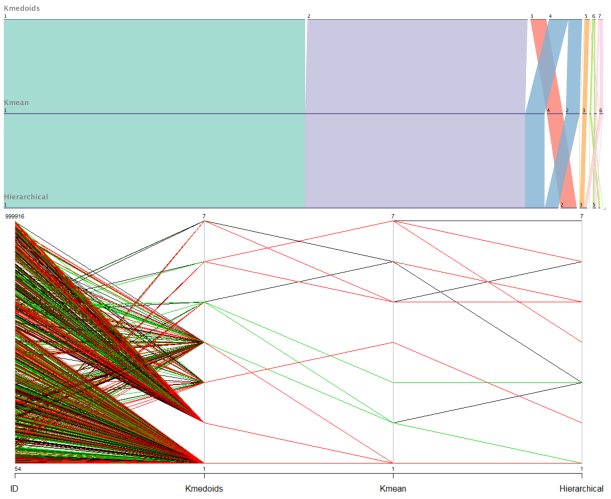

Figure 1: Results of clustering methods on the first subset of the CRSP Stock Daily using Parallel Sets (top) and parallel coordinate plot (bottom).

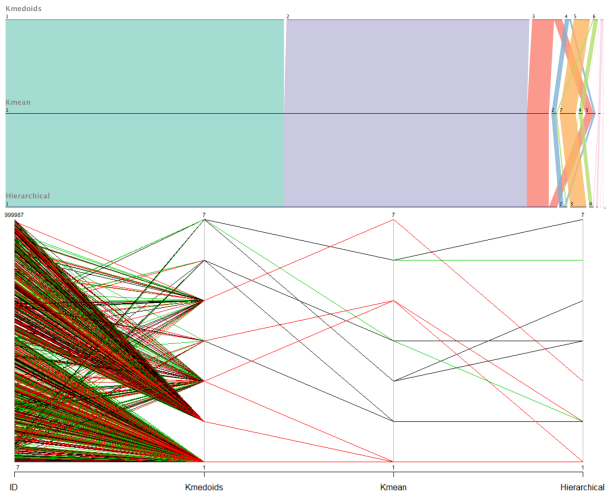

Figure 2: Results of clustering methods on the second subset of the CRSP Stock Daily using Parallel Sets (top) and parallel coordinate plot (bottom).

interfaces," [33] and is a relatively new approach that may be esepcially valuable in exploring and understanding the vast amounts of heterogeneous and uncertain data when the objectives and outcomes of analysis are exploratory. In the RiskMapper, portfolios of financial contracts (e.g., stocks) are plotted on a "magic quadrant" visualization. The x and y axes of the quadrant map invariant relationships that are functionally related across different systemic financial processes. The approach originates from the field of cognitive systems engineering and representation design, with applications in the analysis and design of complex systems such as nuclear power plants [34]. The implementation of the RiskMapper in this study represents "riskiness" in terms of liquidity (x axis) and market attractiveness (y axis). As the measure for market attractiveness we used an implementation of the absorption ratio of Kritzman et al. [35]. The liquidity measure is an implementation of Kyle and Obizhaeva's price-impact measure, [36], [37], [38]. Against these axes, we can then portray the performance of the stocks in relation to these functional invariants. In our study, we computed market capitalisation of the stocks - one possible performance indicator of a company's net worth - and portrayed it in relation to the axes. Such a portrayal would present dangerous situations such as highly capitalised stocks operating under high risk conditions, enabling regulators to monitor the performance of these stock portfolios and perhaps providing them with early warning signals. The RiskMapper [32] is still under development and the efficacy of the measures on the x and y axes are likely to change in subsequent versions of this tool.

4. EXPERIMENT RESULTS

4.1 Phase One Results

The results of the Phase One experiments appear in Figure 1, Figure 2 and Figure 3. The figures show the correlation between the results of distinct clustering techniques using parallel sets visualization [30] and parallel coordinate plots. The advantage of parallel sets is that they show the density of each cluster by the thickness of the bars; the parallel coordinate plot better illustrate the paths in which clusters vary. Figure 1 and Figure 2 show the result of two selected samples of the CRSP data, each containing 10,000 records. There are differences across the clustering techniques within a given sample, as well as variation across samples. For instance, in Figure 1, cluster number 4 in the K-mediods approach (coloured in blue) belongs to two distinct clusters in the K-means, but is back to one cluster in the hierarchical approach. Similarly, in Figure 2, clusters number 1 (coloured in green), 2 (coloured in purple), and a portion of 3 (coloured in orange) in the K-medoids merge into one cluster in both the K-means and the hierarchical approach. The paths that each of the clusters follow from one to another are more apparent in the parallel coordinate plot at the bottom of each figure. Figure 3 shows the results of clustering on a single trading day that illustrates a similar behaviour in cluster variations. This considers only the 699 stocks that appeared in the CRSP data as trading on 4 Jan 2007. In this example, variations between the methods are even higher than the previous two samples. Although this single day of trading is non-representative, selection of a single day of trading was sufficient for our initial prototype. The objective of the experiment was to to draw attention to a key factor in analysts' perception of systemic risk, which is grouping the financial transactions using different, but similar, mathematical techniques. Nevertheless, future experiments will be carried out with dates other than the first or last days of a trading year.

The results clearly show inconsistency between the outputs of different clustering methods. However, one concern is how these variations may affect an analyst's perception of systemic risk. Therefore, the cluster ensemble needs to be plotted on the RiskMapper tool.

4.2 Phase Two Results

The results of Phase Two appear in Figures 4, 5, and 6. The x axis of the RiskMapper refers to liquidity measure, represented by *phigh*, which shows the probability that the

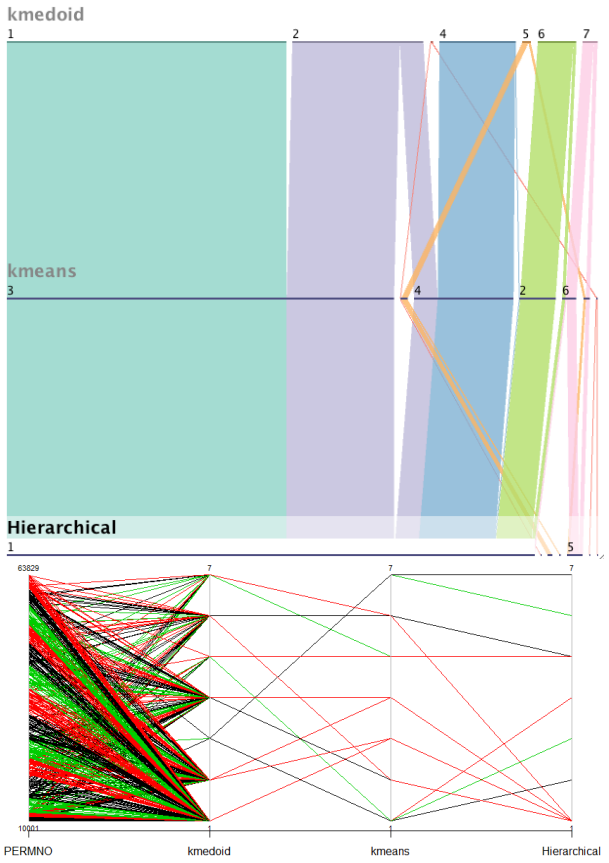

Figure 3: Results of clustering methods of the CRSP data of 04/Jan/2007 using Parallel Sets (top) and parallel coordinate plot (bottom)

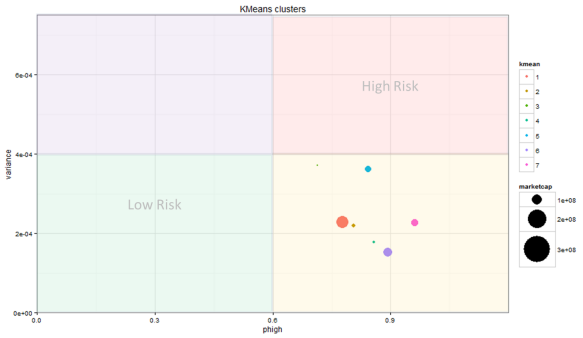

Figure 4: Results of the K-means clustering on the RiskMapper.

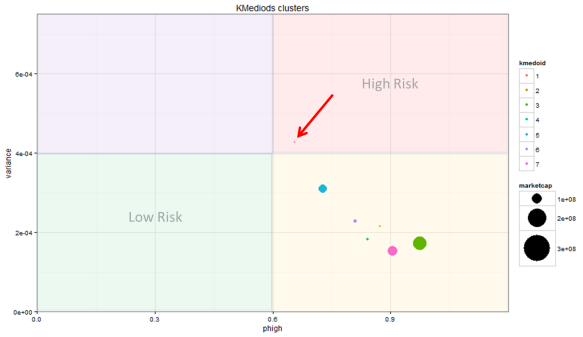

Figure 5: Results of the K-medoids clustering on the RiskMapper. Note the location of the small dot pointed by the arrow in the High Risk quarter.

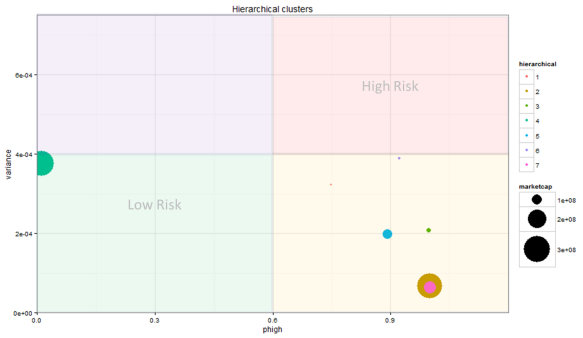

Figure 6: Results of the Hierarchical clustering on the RiskMapper.

equity (or cluster) is in a state of high liquidity as defined by the price-impact measure, [36, 37]. The y axis corresponds to the market attractiveness represented by the absorption rate over a lagging twenty-day period. Figure 4 illustrates the result of mapping the portfolios formed via the K-means method. Here, each dot represents one cluster. The size of each dot indicates market capitalization of the stocks in the portfolio, not the size of the given portfolio. Sizes can be quite large in absolute terms even if the dot in the visualization is small because of the portfolio's relative proportion of total market capitalization. The result of K-medoids, depicted in Figure 5, shows an interesting behaviour. Cluster number 1 (indicated by the arrow) appears in the "high risk" quadrant of the RiskMapper. This cluster is, in fact, the largest cluster among the other six clusters, (see the parallel sets plot in Figure 3), even though the dot is small (because of the proportion of total market capitalization represented by the portfolio). The result of hierarchical clustering also varies significantly . There we see that cluster number 4 stands apart from the other clusters.

5. CONCLUSION

The results demonstrate the variability in portfolio formation that can occur using different clustering techniques and their possible effects on risk perception when imported into a visual analytics tool. These results are very preliminary,

but suggest several avenues for future research. One limitation of this exploratory study is that we have only used samples of the CRSP dataset. A future direction will be to test for the same effects using the entire CRSP dataset, different time slices of the dataset and different dimensions e.g. risk weighting of the stocks. Another limitation is that we have explored only three out of many possible clustering techniques. In future, we would also like to explore the effects of other techniques, and for all of these techniques, the effects that occur when the number of clusters is adjusted up or down. Furthermore, while we have shown that there is significant variability between clustering techniques, resulting in differences in placement of the portfolios on the RiskMapper plots, we do not know the effect of this variability on analysts' perception of the dots. That is, will analysts perceive the portfolios as "risky" or not. Thus, a future direction of study is to conduct controlled experiments to assess the effects of dot placement on visual perception of the relative riskiness of portfolios.

In spite of the limitations, these initial exploratory results highlight that it is important for analysts to be aware of the effects of different clustering techniques on portfolio formation and downstream risk assessment. The results also underscore the value of building in interactions in visual analytics tools, and possibly other types of analytics tools as well. Such interactivity would enable analysts to experiment and change between different clustering/portfolio formation techniques to facilitate exploration and iterative representation of the data space. In this way, analysts avoid the potential for model risk that arises from use of a single clustering technique for aggregation or summarization of large, high dimensional datasets. Exploring ways to implement this functionality into the RiskMapper tool, and testing the results, will also be a future direction of our research.

6. ACKNOWLEDGEMENT

The work of Payam S. Rahmdel was supported in part by the MITACS Elevate Fellowship. The work of William Wong was supported in part by the Peter Wall Institute for Advanced Studies at the University of British Columbia.

7. REFERENCES

[1] S. J. Taylor, *Modelling Financial Time Series*. World Scientific Publishing, 2007.

[2] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006.

[3] D. Bisias, M. Flood, A. W. Lo, and S. Valavanis, "A survey of systemic risk analytics," *Annual Review of Financial Economics*, vol. 4, no. 1, pp. 255–296, 2012.

[4] J. Maginn, D. Tuttle, J. Pinto, and D. McLeavey, *Managing Investment Portfolios: A Dynamic Process*. CFA Institute, 2007.

[5] D. Judd, P. McKinley, and A. Jain, "Large-scale parallel data clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 871–876, August 1998.

[6] R. Xu and I. Wunsch, D., "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, pp. 645–678, May 2005.

[7] D. Fasulo, "An analysis of recent work on clustering algorithms," *Department of Computer Science & Engineering*, 1999.

[8] S. Bhatia and J. Deogun, "Conceptual clustering in information retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 28, pp. 427–436, June 1998.

[9] C. Carpineto and G. Romano, "A lattice conceptual clustering system and its application to browsing retrieval," *Machine Learning*, vol. 24, no. 2, pp. 95–122, 1996.

[10] E. J. and G. Frederix, "Finding regions of interest for content extraction," *Proc. SPIE 3656, Storage and Retrieval for Image and Video Databases VII*, vol. 3656, pp. 501–510, 1998.

[11] A. Musetti, "Clustering methods for financial time series," *Swiss Federal Institute of Technology*, 2012.

[12] N.-C. Hsieh, "Hybrid mining approach in the design of credit scoring models," *Expert Systems with Applications*, vol. 28, no. 4, pp. 655 – 665, 2005.

[13] D. Zhang, S. Leung, and Z. Ye, "A decision tree scoring model based on genetic algorithm and k-means algorithm," in *Third International Conference on Convergence and Hybrid Information Technology*, vol. 1, pp. 1043–1047, November 2008.

[14] J. D. Andres, P. Lorca, F. J. de Cos Juez, and F. Sanchez-Lasheras, "Bankruptcy forecasting: A hybrid approach using fuzzy c-means clustering and multivariate adaptive regression splines (MARS)," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1866 – 1875, 2011.

[15] M. A. Boyacioglu, Y. Kara, and O. K. Baykan, "Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in turkey," *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 3355 – 3366, 2009.

[16] P. R. Kumar and V. Ravi, "Bankruptcy prediction in banks and firms via statistical and intelligent techniques - a review," *European Journal of Operational Research*, vol. 180, no. 1, pp. 1 – 28, 2007.

[17] P. Alam, D. Booth, K. Lee, and T. Thordarson, "The use of fuzzy clustering algorithm and self-organizing neural networks for identifying potentially failing banks: an experimental study," *Expert Systems with Applications*, vol. 18, no. 3, pp. 185 – 199, 2000.

[18] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, pp. 264–323, September 1999.

[19] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification, 2nd Edition*. Wiley-Interscience, 2000.

[20] L. Kaufman and P. Rosseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc., 1990.

[21] M. L. D. S. Brian S. Everitt, Sabine Landau, *Cluster Analysis*. John Wiley and Sons, 1993.

[22] S. Theodoridis and K. Koutroumbas, eds., *Pattern Recognition*. Academic Press, 1999.

[23] A. Fred and A. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27,

pp. 835–850, Jun 2005.

[24] J. Moguerza, A. Munoz, and M. Martin-Merino, "Detecting the number of clusters using a support vector machine approach," in *Artificial Neural Networks - ICANN 2002* (J. Dorronsoro, ed.), vol. 2415 of *Lecture Notes in Computer Science*, pp. 763–768, Springer Berlin Heidelberg, 2002.

[25] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[26] H. C. Romesburg, *Cluster Analysis for Researchers,*. Lulu Press, 2004.

[27] A. Fred, "Finding consistent clusters in data partitions," in *Multiple Classifier Systems* (J. Kittler and F. Roli, eds.), vol. 2096 of *Lecture Notes in Computer Science*, pp. 309–318, Springer Berlin Heidelberg, 2001.

[28] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. 0, pp. 53–65, 1987.

[29] "http://www.crsp.com/."

[30] R. Kosara, F. Bendix, and H. Hauser, "Parallel sets: interactive exploration and visual analysis of categorical data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, pp. 558–568, July 2006.

[31] L. Kaufman and P. Rousseeuw, *Clustering by Means of Medoids*. Faculty of Mathematics and Informatics, 1987.

[32] B. L. W. Wong and V. L. Lemieux, "Briefing note for the design concept: contracts and the "Risk Map"," *CIFER Research Working paper, $www.cif erresearch.org/products_service/products$*, pp. 1 – 10, 2013.

[33] J. J. Thomas and K. A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.

[34] A. M. Rasmussen, J.; Pejtersen and L. P. Goodstein, *Cognitive systems engineering*. Wiley and Sons, 1994.

[35] M. Kritzman, Y. Li, S. Page, and R. Rigobon, "Principal components as a measure of systemic risk," *MIT Sloan Research Paper*, 2010.

[36] A. S. Kyle and A. A. Obizhaeva, "Market microstructure invariants: Empirical evidence from portfolio transitions," *Working paper, College Park, University of Maryland*, 2011.

[37] A. S. Kyle and A. A. Obizhaeva, "Market microstructure invariants: Theory and implications of calibration," *Working Paper, College Park, University of Maryland*, 2011.

[38] Office of Financial Research, "Annual report 2013," *http://www.treasury.gov/press-center/press-releases/Pages/jl2248.aspx*, 2013.