Stock Selection Based on Cluster Analysis

Newton Da Costa, Jr
Department of Economics, Federal University of Santa Catarina

Jefferson Cunha
Department of Economics, Federal University of Santa Catarina

Sergio Da Silva
Department of Economics, Federal University of Santa Catarina

Abstract

We put forward a technique based on cluster analysis to group stocks in spot markets according to a risk-return criterion. We show how an informed investor will make money using the cluster analysis to select stocks of major companies from North and South America.

Keywords: Stocks, Cluster Analysis

JEL classification: M21, G11, G15

1. Introduction

Cluster analysis (Johnson and Wichern 1992) sorts variable sets according to their degree of correlation. One can preset the number of clusters and then let analysis group the variables or, alternatively, one can let analysis itself identify natural clusters. In either case, results are displayed in tree diagrams. Here "fruits" are the variables and "branches" show the distance (degree of association) between them. An advantage of using cluster analysis over other regression techniques is its classifying nature.

We employ the cluster analysis to show that stocks from selected companies of the Americas can be categorized according to their degree of integration. Stock returns are likely to be similar in a region thanks to geography and macroeconomic features, just to name a few. So identification of stock clusters allows one to track those with similar returns but different risks. Much literature focuses on integration of international stock exchange indices (e.g. Da Costa et al 2005) but there is scarce work on integration of individual stocks on an international basis.

Once stocks are grouped by cluster analysis, an informed investor can use the output in his interests. He will, for instance, look for same-return stocks and then choose to minimize risks. Or else he will pick a cluster of same-risk stocks and high return.

The rest of the paper is organized as follows. Section 2 presents data. Section 3 analyzes them. And Section 4 concludes.

2. Data

Our variables are as follows. Return, risk, earnings-price ratio, book value-price ratio, sales-price ratio, sales-number of stocks ratio, and dividend yield. To group the stocks, we take daily data from two time windows, namely 2 January 1997–31 December 1999 and 2 January 2000–31 December 2001. The return and risk variables are calculated between the first and last day of either window. The values for the other variables are those at the last day of either window. We aim to evaluate whether a particular cluster sequence in the first window is worth keeping in the second time window.

From a large set of 1,959 stocks of companies from North and South America listed in the Economatica database, we select those with daily average trade volume greater than one hundred thousand dollars. Doing so we leave out low-liquidity stocks with risk premiums that are artificially elevated by low trade volumes. This leaves us with a shorter set of 816 company stocks from Argentina, Brazil, Chile, Colombia, Mexico, Peru, Venezuela, and the United States. These still made up 99.96 percent of the overall volume traded in such countries' stock exchanges over the period. And these also represent 85.8 percent of the total value of the companies listed in the stock exchanges.

To reckon return and risk we take closing values of the stocks in US dollar terms. These values are corrected for dividends, splits, and other events. We decide to leave out missing-value stocks from the sample. So the sample ends up with 476 stocks.

A stock's total return R is reckoned according to

$$R = \left(\frac{P_t - P_{t-1}}{P_t} \right) \cdot 100$$

where P_t is the stock's closing value at time period t.

Risk (standard deviation σ) is given by

$$\sigma = \sqrt{\sum_{i=1}^{n} P_i \left[r_i - E(r) \right]^2}$$

where r_i is a feasible return, $E(r)$ is expected return, and P_i is the probability related to $r_i - E(r)$.

To prevent units from interfering with their relative weight prior to the groupings, we take standardized values SV_i for all variables X_i, i.e.

$$SV_i = \frac{X_i - \overline{X}_i}{\sigma}$$

where the overbar stands for average values.

The more similar two stocks A and B are, the shorter the Euclidian distance d_{AB} between them, i.e.

$$d_{AB} = \sqrt{\sum_{i=1}^{p} \frac{\left[(x_i(A) - x_i(B)\right]^2}{p}}$$

where $x_i(A)$ gives the location of stock A compared to plane i's origin, and p is space size, i.e. the number of variables.

3. Analysis

We take the hierarchical clustering algorithm of Aldenderfer and Blashfield (1984), and carry out the cluster analysis with Statistica 5.0. The stocks' correlation coefficients are used as inputs in a likelihood matrix. Similar stocks are then merged. And the resulting clusters are displayed in a dendogram, which is a diagram representing the hierarchical organization of the stocks' relationships.

We employ Sokal and Michener's (1958) technique to add a stock in an already existing cluster. The technique considers average returns across time to compensate for random movements. After doing the hierarchical clustering, we employ Ward methodology to sort the clusters. Figure 1 shows the resulting dendogram.

At the arbitrary cut level of 20, we get 10 clusters. Every stock is assumed to have the same weight in a cluster. So returns are given by the mean. (Risk and the other variables are also given by means.) Risk-return features are displayed in Figure 2.

The preferred clusters in Figure 2 are those with negatively sloped straight lines, because these represent higher returns with lower risks. Cluster 2 is made up of only one stock and looks best at first sight. But it also presents large risk-return. Then we consider it as an outlier and leave it out from analysis. As result, cluster 6 becomes the best, followed by clusters 1, 8, and 10 respectively. This sequence gives clusters with similar returns and increasing risks.

Once the best cluster is chosen, the investor can perfect his choice by selecting a stock within the cluster considering the country where the company is based in. And here he might wish to take country risk into account. Figure 3 shows the distribution of clusters according to place of origin. (To improve resolution, four stocks with very high returns are left out.)

Table 1 summarizes the clusters' features for the first time window. (Though cluster 6 is made up of only one stock we decide not to leave it out because of its good fundamental-analysis variables.)

We then move on to consider the second time window. Figure 4 displays risk and return for the same ten clusters selected for the first time window. As can be seen, preferred clusters continue to be 6, 1, 8, and 10 respectively. What is more, this cluster sequence presents now even higher returns with lower risks. So the investor will profit (or lose less) by keeping the sequence selected for the first time window.

Figure 5 shows risk and return of the cluster sequence for the second time window. The sequence is located at the bottom left. This means greater volatility and lower returns throughout and happens because the stock exchanges went lower in the second time window. Figure 6 compares risk and return of the two time windows. As one moves from clusters 6 to 1, 8, 10, 4, 7, 9, 5, 3, and 2 respectively, the relationship between risk and

return gets weaker. Finally Table 2 shows variable values in the second time window for the cluster sequence selected in the analysis of the first time window.

4. Conclusion

This paper shows how an investor can profit (or lose less) using cluster analysis to select stocks. To illustrate our case we take major company stocks of North and South America at two distinct time windows.

The investor choosing stocks according to the clusters sorted in the first time window (2 January 1997–31 December 1999) is found to profit (or to minimize losses) in the second time window (2 January 2000–31 December 2001).

Table 1. Clusters' features, 2 January 1997–31 December 1999

Cluster	Number of Stocks	Return (%)	Risk (%)	Earnings-Price Ratio	Book-to-Price Ratio	Sales-Price Ratio	Sales-Number of Stocks Ratio	Dividend Yield (%)
1	94	44.204 (78.084)	32.8 (10.4)	0.058 (0.045)	0.462 (0.314)	1.258 (0.879)	37.326 (20.996)	2.478 (1.931)
2	1	15171.324 (0.000)	80.8 (0.0)	0.000 (0.000)	0.011 (0.000)	0.005 (0.000)	1.118 (0.000)	0.000 (0.000)
3	46	751.744 (1211.769)	73.3 (29.2)	0.011 (0.033)	0.183 (0.196)	0.245 (0.266)	5.510 (5.404)	0.207 (0.529)
4	20	43.921 (99.121)	55.4 (15.6)	0.139 (0.171)	1.223 (0.806)	1.112 (0.940)	7.019 (14.636)	8.451 (5.563)
5	5	–47.605 (27.483)	66.6 (22.3)	–0.445 (0.628)	6.869 (5.199)	3.384 (2.470)	0.038 (0.056)	2.578 (3.567)
6	1	1.300 (0.000)	28.4 (0.0)	0.085 (0.000)	0.590 (0.000)	1.368 (0.000)	77.613 (0.000)	1.057 (0.000)
7	77	106.480 (731.848)	59.3 (17.0)	–0.029 (0.414)	1.050 (0.660)	1.672 (1.466)	8.320 (17.351)	1.669 (2.012)
8	23	133.144 (284.003)	37.9 (15.9)	0.064 (0.044)	0.512 (0.330)	1.772 (1.322)	69.128 (61.060)	1.580 (1.556)
9	13	27.980 (120.861)	63.2 (44.4)	–0.126 (0.309)	0.956 (0.931)	3.673 (3.382)	24.096 (22.760)	2.838 (4.366)
10	196	200.383 (1125.174)	41.5 (13.8)	0.032 (0.083)	0.384 (0.351)	0.609 (0.857)	12.149 (10.411)	1.483 (1.836)

Note
Standard deviations of the averages are in brackets

Table 2. Clusters' features, 2 January 2000–31 December 2001

Cluster	Number of Stocks	Return (%)	Risk (%)	Earnings-Price Ratio	Book-to-Price Ratio	Sales-Price Ratio	Sales-Number of Stocks Ratio	Dividend Yield (%)
1	94	15.680 (45.307)	41.1 (13.5)	0.036 (0.074)	0.472 (0.349)	1.587 (1.636)	47.640 (34.120)	2.279 (1.680)
2	1	–91.800 (0.000)	93.8 (0.0)	–0.009 (0.000)	0.195 (0.000)	0.071 (0.000)	1.259 (0.000)	0.000 (0.000)
3	46	–35.875 (40.223)	83.4 (23.1)	–0.170 (0.899)	0.396 (0.313)	0.509 (0.499)	6.663 (5.776)	0.857 (2.027)
4	20	–8.531 (33.577)	49.2 (13.3)	0.115 (0.126)	1.546 (1.020)	1.835 (1.472)	14.240 (36.739)	10.105 (8.341)
5	5	–3.507 (48.338)	64.1 (20.9)	–2.809 (6.118)	10.517 (13.053)	9.994 (17.780)	0.031 (0.042)	0.000 (0.000)
6	1	13.495 (0.000)	31.8 (0.0)	0.164 (0.000)	0.885 (0.000)	2.421 (0.000)	151.229 (0.000)	2.400 (0.000)
7	77	–10.579 (52.838)	57.4 (20.1)	0.037 (0.309)	1.515 (2.119)	2.637 (2.352)	10.459 (23.924)	4.315 (11.977)
8	23	13.150 (57.677)	47.1 (19.5)	0.005 (0.135)	0.519 (0.359)	1.913 (1.478)	72.205 (62.717)	1.322 (1.444)
9	13	–6.112 (36.053)	69.4 (50.4)	–0.187 (0.542)	1.031 (0.901)	3.484 (3.774)	24.769 (27.927)	13.479 (41.754)
10	196	4.538 (53.312)	47.9 (17.7)	–0.008 (0.447)	0.519 (0.662)	0.855 (1.023)	15.449 (17.615)	2.111 (4.128)

Note
Standard deviations of the averages are in brackets

Figure 1. Dendogram using Ward methodology and taking Euclidian distances, 2 January 1997–31 December 1999

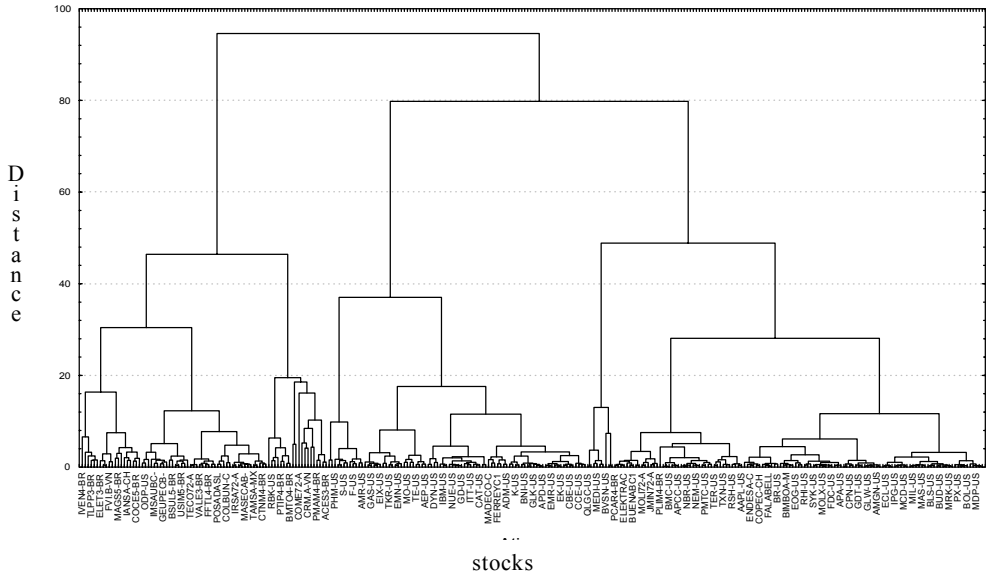


Figure 2. Clusters' average return and risk, 2 January 1997–31 December 1999

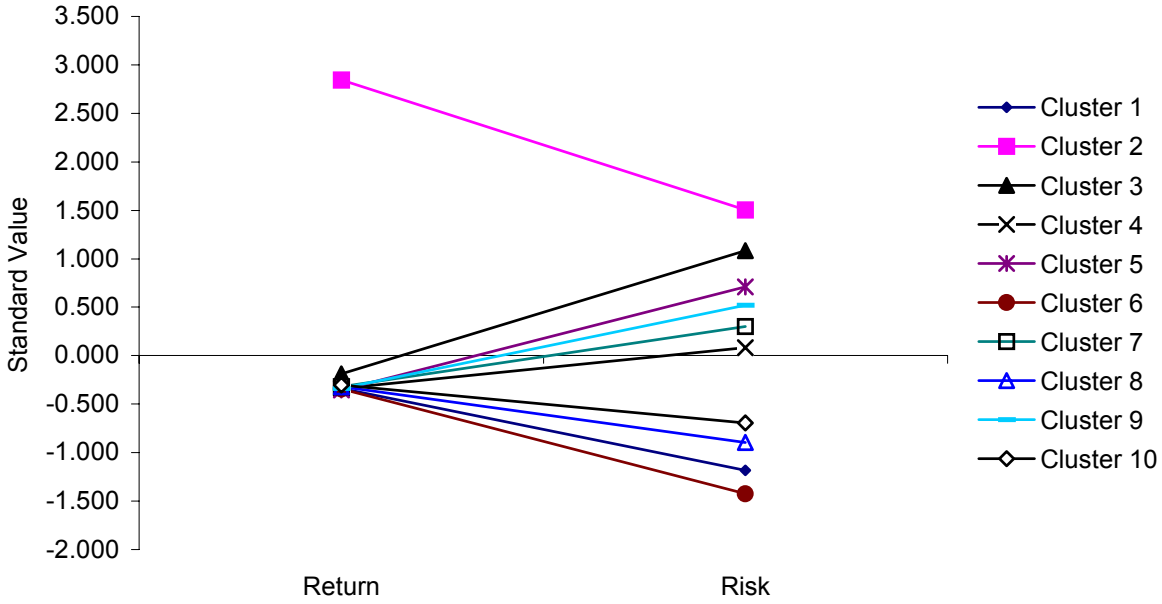


Figure 3. Return and risk of stocks in a cluster,
2 January 1997–31 December 1999

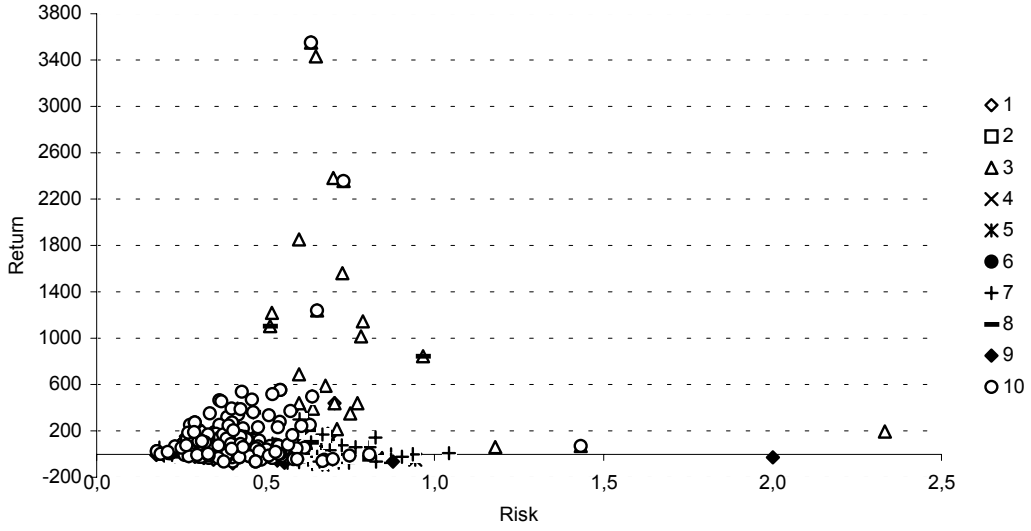


Figure 4. Clusters' average return and risk,
2 January 2000–31 December 2001

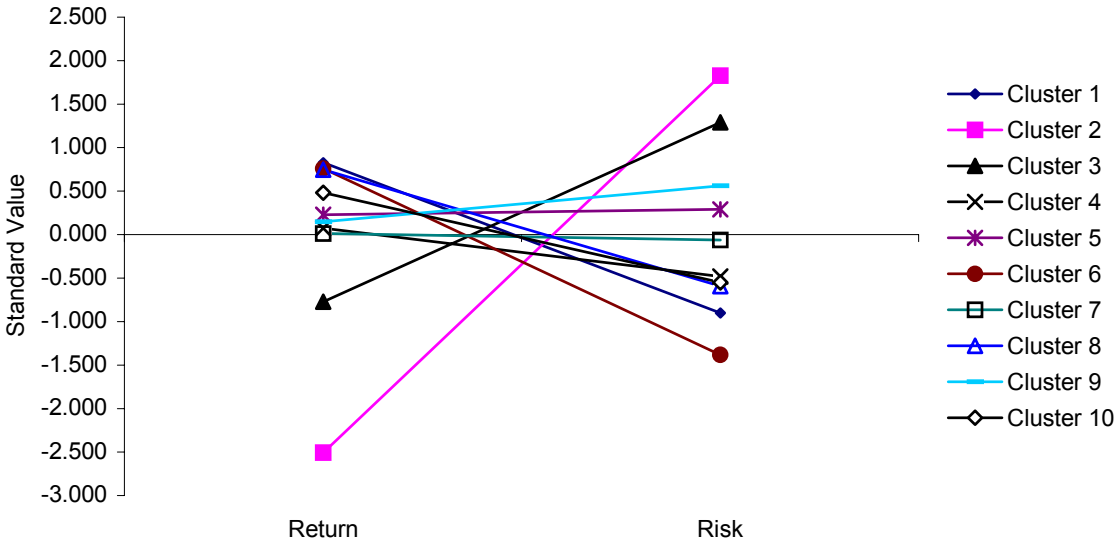


Figure 5. Return and risk of stocks in a cluster,
2 January 2000–31 December 2001

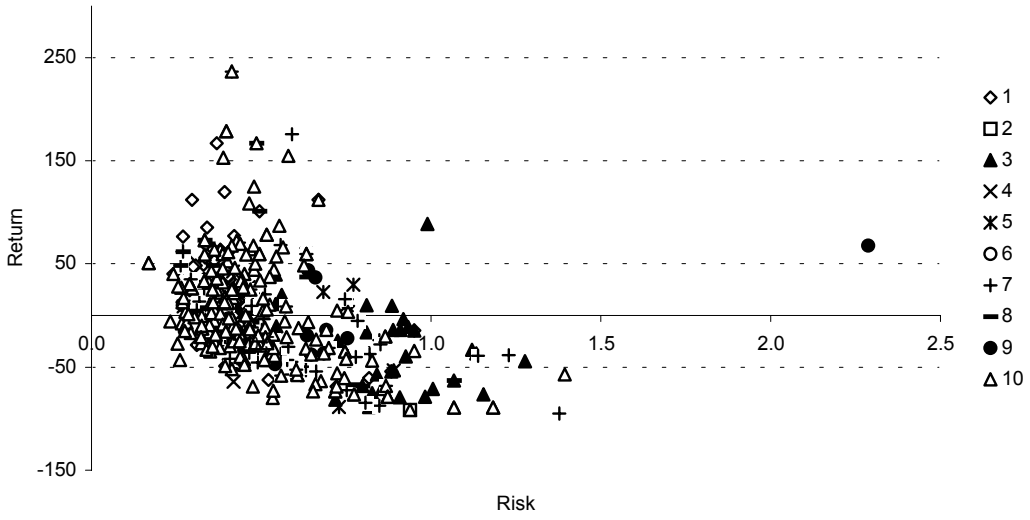
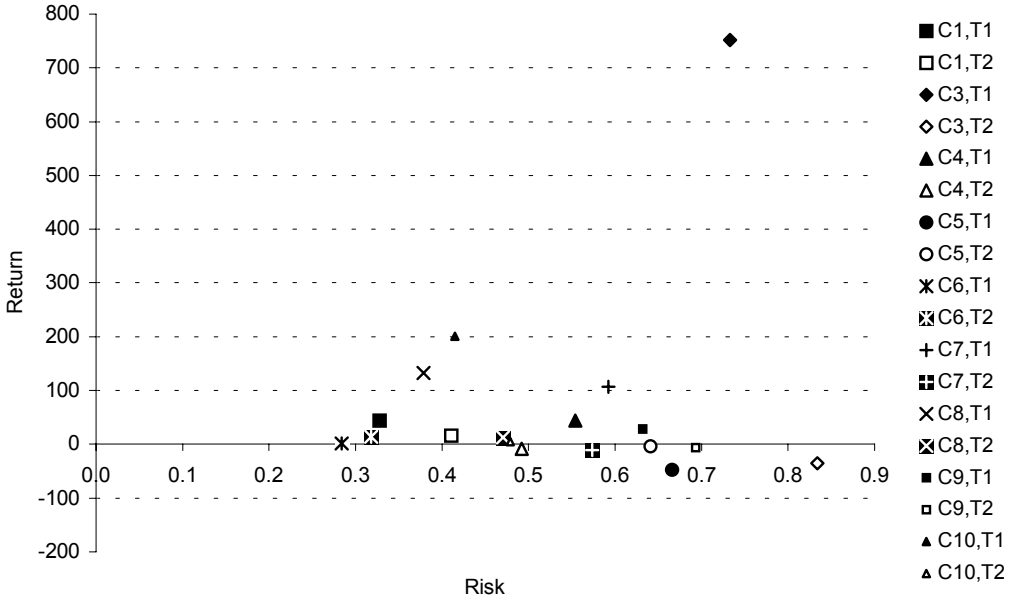


Figure 6. Clusters' average return and risk of a cluster C(n) in time window T(n),
2 January 1997–31 December 1999 and 2 January 2000–31 December 2001



References

Aldenderfer, M.S., and R.K. Blashfield (1984) *Cluster Analysis*. Sage University Paper: Quantitative Applications in the Social Sciences, n. 44.

Da Costa, Jr, N., S. Nunes, P. Ceretta, and S. Da Silva (2005) "Stockmarket comovements revisited", *Economics Bulletin* **7**, 1-9.

Johnson, R. A., and D.W. Wichern (1992) *Applied Multivariate Statistical Analysis*, Prentice-Hall International: New Jersey.

Sokal, R., and C.D. Michener (1958) "A statistical method for evaluating systematic relationships" *University of Kansas Scientific Bulletin* **38**, 1409-1438.