DOI:10.26524/jms.2017.11

Journal of Management and Science ISSN: 2249-1260 | e-ISSN: 2250-1819

Application of data mining to grading indian industries on the basis of financial ratios

G. MANIMANNAN[1], R. LAKSHMI PRIYA[2] AND V. ANISH PREETHI[3]

[1]ASSISTANT PROFESSOR AND HEAD, DEPARTMENT OF STATISTICS, DRBCCC HINDU COLLEGE, CHENNAIL
[2]ASSISTANT PROFESSOR, DEPARTMENT OF STATISTICS, DR. AMBEDKAR GOVERNMENT ARTS COLLEGE, VYASARPADI, CHENNAI
[3]ASSISTANT PROFESSOR, DEPARTMENT OF STATISTICS, DRBCCC HINDU COLLEGE, CHENNAI

Abstract: An attempt is made to introduce a new method of rating the top ranking industriess on the basis of certain financial ratios. It is well known that the financial ratios are being used as a yardstick by researchers for many purposes. About 500 industries from public and private sectors were considered for each year from 2007 to 2012, which were ranked according to their net sales. Twenty financial ratios were carefully chosen out of numerous ratios that could give different notion of the objectives and have significant meaning in the literature. The unique feature of this study is the application of factor, k-mean clustering and discriminant analyses as data mining tools to exploit the hidden structure present in the data for each of the study periods. Initially, factor analysis is used to uncover the patterns underlying financial ratios. The scores from extracted factors were used to find initial groups by k-mean clustering algorithm. A few outlier industries, which could not be classified to any of the larger groups, were discarded as some of the ratios possessed higher values. The clusters thus obtained formed the basis for the further analyses as they inherited the structural patterns found by the factor analysis. The cluster analysis was followed by iterative discriminant procedure with original ratios until cent percent classification was achieved. Finally, the groups were identified as industries belonging to Grade A, Grade B and Grade C in that order, which exhibit the behavior of High performance, Moderate performance and Low performance. From the present study it was observed that a little over 90% of the total variations of the data were explained by the first five factors for each year. These five factors revealed the underlying structural patterns among the twenty ratios that were initially considered in the analysis. Also only three clusters could be meaningfully formed for each of the periods. It is also interesting to note that the clusters could be arranged by magnitude of their mean vectors on selected ratios, thus permitting the groups to be identified on the basis of their performance.

Key Words: *Data mining, Financial Ratios, Factor Analysis, k-means Clustering and Multivariate Discriminant Analysis.*

1. INTRODUCTION

Among various techniques used in financial statement analysis, the ratio analysis is the most powerful tool for financial analysis. As ratios are simple to calculate and easy to understand, they have been extensively used by researchers for many purposes in recent years. They include numerous statistical models that have been developed for

prediction of corporate failure (Beaver, 1966; Taffer, 1982); bond rating (Pinches, 1973; Copland and Ingram 1984); firm‘s performance (Bayldon, Woods and Zafiris, 1984) and corporate health (Prasant, Mishra and Satpathy, 1996). However, these models typically link a set of ‒independent‖ variables to a ‒dependent‖ variable that can take two or more discrete values. All these models use prior group information in classification pertaining to a known number of groups. Usually two groups are considered in finance problem where the operational objective is to assign the firm or company to one of the groups after data analysis (for example firms classified as sick Vs. non-sick, etc.,). In this paper, an attempt is made to analyse the performance of industries based on the financial ratios, where no assumptions are made with regard to the number of group or any other structural patterns in advance. The Objective of the present study, therefore is to uncover the inherent groups or classes that would reflect the performance of top ranking industries in india, using the concepts of data mining.

2. METHODOLOGY

This section brings out the discussion of the database, the ratios selected and the Data Mining Techniques.

2.1. Database

The financial data published by *Capital Market* was considered as the database. It is to be noted that the publisher excluded banking and state corporations from the data, as their comparison is meaningless. However, only top 500 industries are carefully thought about for the analysis for each year from 2007 to 2012 based on their net sales. Among the listed industries, number of industries varied over the study period (*Table 1*) owing to removal of those industries for which the required data are not available.

2.2. The Ratios

The number of ratios that can be calculated from a typical set of financial statements is much large to in incorporate in this study. Moreover, due to constraints discussed in the above section, only twenty financial ratios are carefully chosen that gave meaningful interpretation. The different ratios computed are given in Appendix.

2.3. Data Mining Techniques

Although data mining is a new term, the technology is not. Data Mining or Knowledge Discovery in Databases (KDD) is the process of discovering previously unknown and

potentially useful information from the data in databases. In the present context data mining exhibits the patterns by applying few techniques namely, factor analysis, **k**-means clustering and discriminant rule. Mining enables industries to determine relationship among ‒internal‖ factors such as price, product positioning or staff skills and ‒external‖ factors such as economic indicators and competition. It also enables the company owners to determine the impacts of sales, customer‘s satisfaction and corporate profits to place their company performance in perspective. As such KDD is an iterative process, which mainly consist of the following steps;

Step 1: Data Cleaning and Integration
Step 2: Data selection and transformation
Step 3: Data Mining
Step 4: Knowledge representation

Of these above iterative process Steps 3 and 4 are most important. If clever techniques are applied in Step 3, it provides potentially useful information that explains the hidden structure. This structure discovers knowledge that is represented visually to the user, which is the final phase of data mining.

Table 1

Number of industries in the analysis before and after Data Pruning

Year	Number of Industries	
	Before	After
2007	500	435
2008	500	462
2009	500	470
2010	500	347
2011	500	363
2012	500	381

2.3.1. Factor Analysis

In the present study, factor analysis is initiated to uncover the patterns underlying financial ratio variables (*Appendix*). Factor analysis reduces the variable space to a smaller number of patterns that retain most of the information contained in the original data matrix. In factor extraction method the number of factors is decided based on the proportion of sample variance explained. Orthogonal rotations such as Varimax and Quartimax rotations are used to measure the similarity of a variable with a factor by its factor loading. In factor analysis, the interest is centered on the parameter in the factor model that estimated values of the common factor, called *factor scores.* These scores are subjected to further analysis to mine the data.

2.3.2. k-Means Clustering Algorithm

A nonhierarchical clustering algorithm suggested by MacQueen (1967) also known as ***unsupervised classification*** is the next technique in data mining. This process divides

the data set into mutually exclusive group such that the members of each groups are as close as possible to one another and different groups are as far as possible from another. Generally this technique uses Euclidean distances measures computed by variables. Since the group labels are unknown for the data set, **k**-means clustering is one such technique in applied statistics that discovers acceptable classes. Thus forming the nuclei of clusters or groups as seed points exhibited in factor analysis. The number of cluster **k** is determined as part of the clustering procedure.

2.3.3. Discriminant Analysis

Many researchers have used aprior group information for classification and model buildings using discriminant Analysis (DA) to achieve their objectives. In the present study, iterative discriminant analysis is used to exhibit groups graphically and judge the nature of overall performance of the industries. This process re-allocated the industries that were assigned a group label by **k**-means clusters as a seed point. Re-allocation is subjected until cent percent classification is attained, by considering the classification of group obtained in iteration **t** as the input into the next iteration **t+1**. It is to be noted that the concept of performing repetitive DA is new in accessing the performance of the top rated industries in terms of net sales.

3. ALGORITHM

A brief algorithm to grade the industries during each of the study period based on their overall performance is described below:

Step 1: Factor analysis is initiated to find the structural pattern underlying the data set and scores were extracted.

Step 2: **k** –means analysis partitioned the data set into **k**-clusters using factor scores as input matrix.

Step 3: Repeat Steps 1 and 2 until meaningful groups are obtained, by removing outliers in each cycle.

Step 4: Discriminant analysis is then performed with the original ratios by considering the groups formed by the **k**-means algorithm.

Step 5: Repeat step 4 until cent percent classification is achieved from iteration **t** to the next iteration **(t+1)** for some **t**.

4. RESULTS AND DISCUSSION

As mentioned in Section 2.3.1 Varimax and Quartimax criterion for orthogonal rotation have been used for the pruned data. Even though the results obtained by both the criterions were very similar, the varimax rotation provided relatively better clustering of financial ratios. Consequently, only the results of varimax rotation are reported here. We have decided to retain 90 percent of total variation in the data, and thus accounted consistently five factors for each year with eigen values little less than or equal to unity. *Table 2* shows variance accounted for each factors

Table 2

Percentage of Variance explained by factors (Year-wise)

	Variance explained					
Factors	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
1	50.2	47.0	55.1	56.3	58.6	59.1
2	26.4	26.3	23.4	22.6	21.4	17.2
3	6.8	8.2	6.0	5.3	5.2	5.7
4	4.9	5.2	4.5	4.1	4.2	5.6
5	3.3	3.4	3.1	3.3	2.9	3.2
Total	**91.6**	**90.1**	**92.1**	**91.6**	**92.3**	**90.8**

From the above table we observe that the total variances explained by the extracted factors are over 90 percent, which are relatively higher. Also, for each factor the variability is more or less the same for the study period, though the number of industries in each year, after data cleaning and selection, kept varying owing to various reasons. The financial ratios loaded in the factors are presented in *Table 3*. Only those ratios with higher loadings are indicated with asterisk (*) symbol. From the *Table 3* it is clear that the clustering of financial ratios is stable during the study period. We observed slight changes in factor loadings during the periods considered. The differences in factor loadings may be due to statistical variations in the original data.

After performing factor analysis, the next stage is to assign initial group labels to each company. Step 2 of the algorithm is explored with factor score extracted by Step 1, by conventional **k**-means clustering analysis. Formations of clusters are explored by considering 2-clusters, 3-clusters, 4-cluster and so on. Isolated groups with few industries are discarded from the analysis as outliers. A few financial ratios for these outlier industries are comparatively high or low to those excelled in the analysis. Out of all the possible trials, 3-cluster exhibited meaningful interpretation than two, four and higher clusters. Having decided to consider only 3 clusters, it is possible to rate a company as Grade **A**, Grade **B** or Grade **C** depending on whether the company belonged to Cluster 1, Cluster 2 or Cluster 3 respectively. Cluster 1 (Grade **A**) is a group of industries that have high values for the financial ratios, indicating that these industries are performing well. The industries with lower values for the financial ratios are grouped into Cluster 3 (Grade **C**). This suggested that Cluster 3 is a group of industries with low-profile. Cluster 2 (Grade **B**) are those industries which perform moderately well as compared to the Cluster 1 and Cluster 3.

In spite of incorporating the results for each year, only the summary statistics are reported in *Table 4*. The first column in *Table 4* provides the groupings done by cluster analysis. The second column gives the groupings after the application of discriminant analysis until 100 percent classification is achieved. Column three indicates the number of cycles required for convergence.

Table 4

Number of industries in the clusters

Years	Initial Cluster			Converged Discriminant			Number of Cycles
	1	**2**	**3**	**1**	**2**	**3**	
2007	148	241	46	87	326	22	12
2008	73	332	57	166	262	34	22
2009	37	117	316	73	100	297	11
2010	37	252	58	26	266	55	10
2011	36	130	197	34	112	217	11
2012	51	50	280	21	46	214	12

1 – Grade **A** 2 – Grade **B**

3 – Grade **C**

Table 4 indicates that majority of industries are in the moderate performance category except for the year 2007, 2008 to 2010. The possible reasons that kept most of the industries in lower profile in the years 2011 to 2012 may be due to the political uncertainty in New Delhi and 2G spectrum scandal. The reasons for more industries grouping into low-profile category in the year 2009 due to general election for Indian parliament. *Figure 1* through *6* show the groupings of industries into 3 clusters for each year of the study span. It is interesting to note that the mean vectors of these clusters can be arranged in the increasing order of magnitude as show in *Table 5*. We rated the members in the first cluster as Grade **A**, and the second as Grade **B** and the third as Grade **C.** Industries belonging to Grade **A** category are the ones that performs better than those of Grade **B** and Grade **C**. Similarly the industries belonging to Grade **B** category are superior to those of Grade **C**, indicating the members in the category Grade **C** are at a low profile in terms of the ratios considered in the present analysis.

5. CONCLUSION

The purpose of this paper was to identify the meaningful groups of industries that are rated as best with respect to their performance in terms of net sales using data mining techniques. An attempt is made to analysis the financial data relating to public and private sector industries over a period of six years from 2007 to 2012. Each year‘s data is analysed separately. Initially, factor analysis is used to identify the underlying structure in the 20 financial ratios. The factor scores are used to partition the industries into different clusters by using **k**-means clustering algorithm.

After obtaining the initial clusters using the factor scores, original ratios are considered for further analysis. Discriminant analysis is then iteratively performed on the initial groups, by re-allocating members from iteration **t** to the clusters obtained in iteration **(t+1)**, until the process converges, that is, a member belonging to a cluster is assigned to itself.

The industries could be grouped only to 3 clusters for each year. The members of Cluster 1 are found to have high values for the financial ratios and hence they performed well.

Thus, the members of Cluster 1 are labeled as Grade **A** industries. Similarly, the Cluster 2 included industries, which performed moderately well and the Cluster 3 with low-profile industries.

The present analysis has shown that only 3 groups could be meaningfully formed for each year. This indicates that only 3 types of industries existed over a period of six years. Further, the industries find themselves classified into *High* (Grade **A**), *Medium* (Grade **B**) and *Low* (Grade **C**) categories depending on the financial ratios. A generalization of the results is under investigation to obtain a unified class of 3 groups of industries for any given year.

REFERENCES

1. Altman E I (1968), Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, The Journal of Finance, vol.23, pp. 589-609
2. Anderson T W (1984), An Introduction to Multivariate Statistical Analyis, 2/e, John Wiley and Sons, Inc., New York
3. Breaver W H (1966), Financial Ratios as Predictors of Failure, Journal of Accounting Research, pp. 179-192
4. Capital Market, Indian Corporate Database, India(2007 to 2012)
5. David Wishart (1999), Clustering Methods for Large Data Problems, Bulletin of International Statistical Institute, Book1, pp. 437-440
6. David Wishart (2001), K-Mean Clustering with Outlier Detection Mixed Variables and Misssing Values, www. clustan.com
7. David Wishart (1998), Efficiency hierarchical cluster analysis for data mining and knowledge discovery, Computing Science and Statistics, pp. 257-263
8. Everitt (1980), Cluster Analysis, Halsted Press, Division of John Wiley and Sons, New York
9. Pieter Adriaans and Dolf Zantinge (1996), Data Mining, Addison-Wesly Longman Limited, England
10. Prasanna Chandra (1997), Financial Management Theory and Practice, 4/ed, Tata McGraw-Hill Publication Company Limited, New Delhi
11. Richard A Johnson and Dean W Wichern (1992), Applied Multivariate Statistical Analysis, 3/ed, Prentice-Hall of India Private Limited, New Delhi

Appendix

1.	Gross Profit / Net Sales	PBDT/NS
2.	Net Profit / Net Sales	PAT/NS
3.	Earning Before Interest and Tax /Total Assets	EBIT/NS
4.	Net Profit/Total Assets	PAT/A
5.	Net Profit before tax/Net Sales	PBT/NS
6.	Net Profit/Net Worth	PAT/NW
7.	Operating Profit/Net Sales	PBDIT/NS
8.	Operating Profit /Gross Sales	PBDIT/GS
9.	Gross Profit/Gross Sales	PBDT/GS
10.	Operating Profit/Total Assets	PBDIT/A
11.	Net Sales / Total Assets	NS/A
12.	Gross Profit / Total Assets	PBDT/A
13.	Cost of Sales/Net Sales	COGS/NS
14.	Cash Flow/Net Sales	CF/NS
15.	Cash Flow/Net Worth	CF/NW
16.	Net Worth/Net Sales	NW/NS
17.	Retained Earning/Total Assets	RP/A
18.	(Net profit/Net Worth) * (1- Payout)	SGR
19.	Earning Before Interest and Tax/Interest	TIER
20.	Pay out ratio	PAY_OUT

Table 3
Financial Ratios in Rotated Factors (Year -wise)

Ratios	2007					2008					2009					2010					2011					2012				
	Factors					Factors					Factors					Factors					Factors					Factors				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
PBDIT/NS	*					*					*					*						*				*				
PBDT/NS	*					*					*					*						*				*				
COGS/NS	*					*					*					*						*				*				
PBDIT/GS	*					*					*					*						*				*				
PBDT/GS	*					*					*					*						*				*				
CF/NS	*					*					*					*						*				*				
PAT/NS	*					*					*					*						*				*				
PBT/NS	*					*					*					*						*				*				
NW/NS	*					*					*					*						*				*				
NS/A	*					*					*					*						*				*				
PBDIT/A		*					*					*						*				*						*			
PBIT/A		*					*					*						*				*						*			
PBDT/A		*					*					*					*					*						*			
PAT/A		*					*					*					*					*						*			
SGR			*						*				*					*				*							*		
CF/NW			*					*					*					*				*							*		
PAT/NW			*					*					*					*				*							*		
RP/A			*				*						*					*				*						*			
TIER				*						*				*					*					*						*	
PAY_OUT					*				*						*					*				*						*

***** Indicates financial ratios highly loaded in respective factors.

Table 5 Centroids of Final Groups

Ratios	2008			2009			2010			2011			2012		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
PAY_OUT	.8698	.7544	.6457	.3154	.2837	.1418	.3106	.2734	.0852	.3027	.1915	.1608	.2955	.2307	.0492
PBDIT/GS	.1877	.1428	.0935	.2220	.1654	.1338	.2180	.1572	.1491	.2030	.1965	.1263	.3264	.1536	.0637
PBDT/GS	.1385	.0987	.0712	.1848	.1400	.0773	.2004	.0958	.0727	.1770	.1241	.0683	.2905	.0956	-.0294
PAT/NS	.0861	.0783	.0592	.1499	.0839	.0500	.1360	.0615	.0093	.1250	.0674	.0329	.1725	.0520	-.0885
PBDIT/NS	.2098	.1578	.1067	.2433	.1814	.1459	.2314	.1652	.1638	.2231	.2121	.1382	.3382	.1673	.0713
NW/NS	.5815	.5673	.2949	.7209	.5094	.4292	1.099	.5459	.5347	1.158	.5783	.5387	1.379	.5670	.3715
CF/NS	.0996	.0860	.0677	.1711	.0839	.0651	.1415	.0760	.0520	.1428	.1027	.0528	.1988	.0801	-.0379
COGS/NS	.9192	.8909	.8452	.9161	.8468	.7977	.9276	.8937	.7872	.9257	.8638	.8087	1.034	.8963	.6985
PBT/NS	.1154	.0834	.0662	.1585	.1303	.0540	.1736	.0745	.0178	.1507	.0835	.0407	.2369	.0601	-.0871
PBDT/NS	.1547	.1090	.0808	.2022	.1532	.0839	.2127	.1062	.0724	.1912	.1362	.0743	.3014	.1037	-.0341
PBDIT/A	.1881	.1748	.1210	.2199	.1867	.1166	.1984	.1429	.1282	.1723	.1719	.1065	.1997	.1384	.0687
PBDT/A	.1408	.1323	.0831	.1859	.1539	.0661	.1851	.0932	.0565	.1569	.1113	.0616	.1851	.0929	-.0065
NS/A	2.095	1.062	.9498	1.553	1.164	.9651	1.133	1.093	1.022	1.078	.9500	.8850	1.327	.9435	.7580
PAT/A	.0932	.0761	.0582	.1167	.1023	.0387	.1118	.0538	.0062	.1025	.0576	.0298	.1120	.0497	-.0478
PBIT/A	.1567	.1522	.1018	.1944	.1562	.0932	.1687	.1172	.0834	.1461	.1351	.0817	.1674	.1043	.0306
RP/A	.0825	.0526	.0455	.1003	.0708	.0278	.0859	.0388	-.0005	.0817	.0443	.0187	.0822	.0359	-.0505
PAT/NW	.3823	.1779	.1454	.2468	.2371	.1116	.1921	.1331	.0015	.1844	.1378	.0689	.1802	.1157	-.2239
CF/NW	.4450	.2044	.1653	.2704	.2283	.1531	.2012	.1624	.1119	.2116	.1896	.1097	.1838	.1742	-.0720
SGR	.0548	.0438	.0323	.2042	.1663	.0794	.1466	.0961	-.013	.1443	.1071	.0417	.1306	.0839	-.2309
TIER	5.847	5.154	3.597	8.209	7.510	2.408	12.67	2.753	1.405	13.14	2.451	2.250	23.07	3.328	.3136

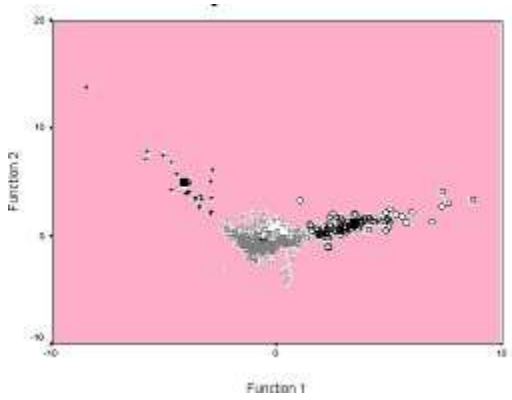
A- High Performance B- Moderate Performance

C- Low Performance

Clustered Groups

Figure 1. Year 2007

Figure 2. Year 2008



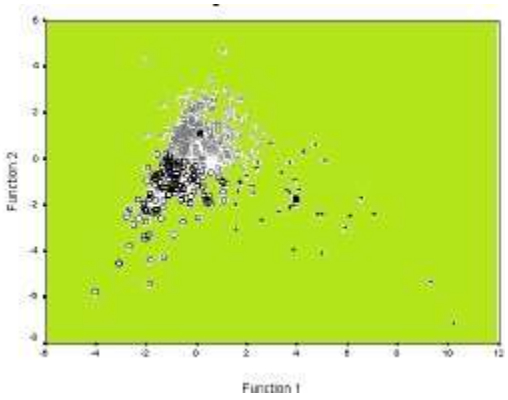
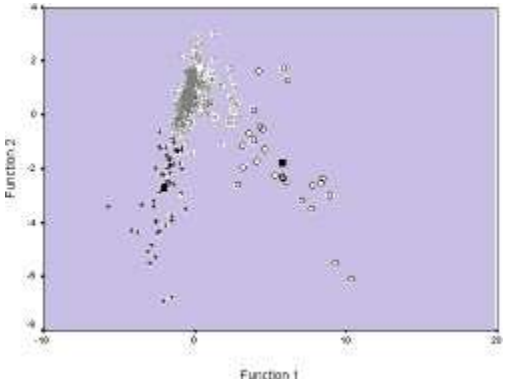


Figure 3. Year 2009

Figure 6. Year 2010



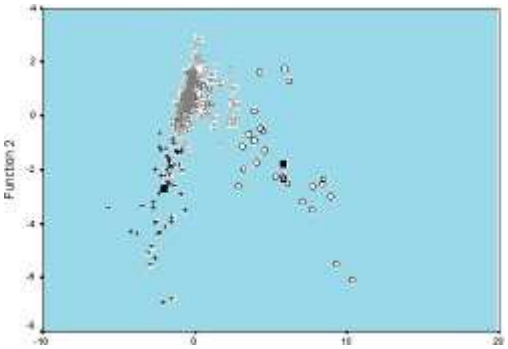
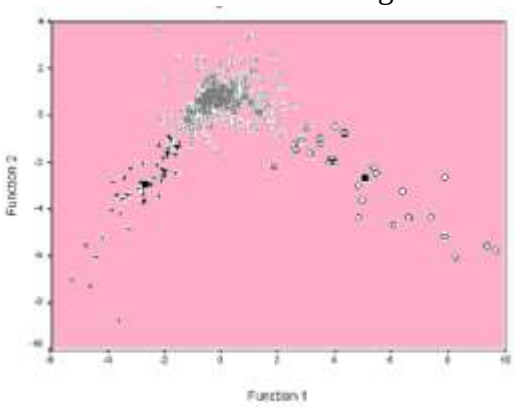
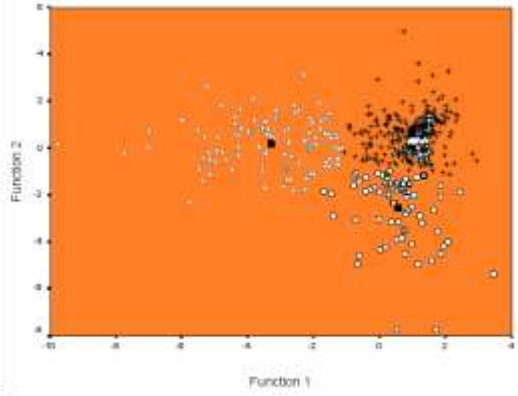


Figure 5. Year 2011

Figure 6. Year 2012





O Cluster 1 (Grade **A**) □ Cluster 2 (Grade **B**) ■ + Cluster 3 (Grade **C**) Cluster Center