

## Paper ST-03

**Find Potential Fraud Leads Using Data Mining Techniques**

Qiling Shi, Carrie McBrayer

NCI Information Systems, Inc., Nashville, Tennessee

**ABSTRACT**

The purpose of this study is to use data mining techniques such as principal component analysis and clustering technique to find the potential fraud leads using SAS®. Here fraud leads are the providers with suspicious or extremely aberrant billing activities that should be investigated in more details. We will identify the results from a supposed fraud matrix which also give rise to the concern. This example of fraud matrix ranks eleven providers by seven fraud indices, where these seven indices are given equal importance. A fraud index can be the rank of number of services billed during the holidays or provided to a deceased client. We will combine the multiple fraud indices into one comprehensive index and rank the providers by this comprehensive index. The top providers should be the fraud leads which are on the top of the alert list. Also by clustering providers into different categories using these seven fraud indices will give people a better picture of fraud maps. SAS procedures such as PROC PRINCOMP, PROC SORT, and PROC UNIVARIATE, PROC TABULATE, PROC PRINT, ODS HTML, PROC CLUSTER, PROC TREE, PROC SQL and DATA STEPS are employed to do the data analysis.

**INTRODUCTION**

In today's challenging economy, organizations need to be prepared to fight fraudulent activities. Being prepared for fraud is essential to the health of any organization. The various high-profile schemes that have recently come to light across the country serve as a reminder that all CPAs should remain constantly vigilant, looking out for the red flags that could indicate fraud.<sup>1</sup> Things become more complicated when you have so many red flags since more investigations need to be invested to see the details. The investigation of a suspected fraud can include many pitfalls. The potential for serious problems is due primarily to the adversarial nature of the fraud examination process. It is equally important for reporting purposes that information that has no bearing on the case be excluded. Irrelevant information confuses the reader, makes the report unnecessarily complex, and raises questions about matters of no consequence. By the same token, one must be alert to assertions that an incomplete report has been rendered.

Therefore it is very likely that you have multiple fraud indices according to one provider and you could not choose only one of them in the report. The fraud leads should be significant in multiple indices. It is very hard to notice the significance at the first glance since providers' ranks in these indices always vary. Hence we used the data mining techniques to quickly find those fraud leads. By this way, we would be able to reduce investigating time and improve the efficiency of fraud fighting significantly.

**METHODOLOGY**

Our analysis is based on the example of a fraud matrix ranking eleven providers by seven fraud indices. The seven fraud indices are: "Fraud\_index1" - "Fraud\_index7". Each index is equally important. In order to combine the multiple fraud indices into one comprehensive index and rank the providers by this comprehensive index, we use the PRINCOMP procedure in SAS to transform the observed ranks.

The PRINCOMP procedure performs principal component analysis. Principal Components Analysis (PCA) is a technique used to reduce multidimensional data sets to lower dimensions for analysis. Its goal is to extract the important information from the table, to represent it as a set of new orthogonal variables called principal components, and to display the pattern of similarity of the observations and of the variables as points in maps. The number of principal components is less than or equal to the number of original variables.

Principal components have a variety of useful properties:

- The eigenvectors are orthogonal, so the principal components represent jointly perpendicular directions through the space of the original variables.
- The principal component scores are jointly uncorrelated. Note that this property is quite distinct from the previous one.
- The first principal component has the largest variance of any unit-length linear combination of the observed variables. The  $j$ th principal component has the largest variance of any unit-length linear combination orthogonal to the first  $j-1$  principal components. The last principal component has the smallest variance of any linear combination of the original variables.
- The scores on the first  $j$  principal components have the highest possible generalized variance of any set of unit-length linear combinations of the original variables.
- The first  $j$  principal components provide a least-squares solution to the model

$$Y = XB + E$$

We suppose these seven fraud indices would not have very huge variations in the ranks, so we can take the first principal “Prin1” as our comprehensive index. And rank the providers by this comprehensive index “Prin1” which has the highest eigenvalue of the dataset. It represents the most significant relationship between the data dimensions. The top providers should be the fraud leads which are on the top of the alert list.

We cluster providers into different categories using these seven fraud indices using PROC CLUSTER. PROC CLUSTER computes squared Euclidean distances between these seven fraud indices. We choose the clustering method to be average linkage. In average linkage, the distance between two clusters is the average distance between pairs of observations. So the final dataset will give people a better picture of fraud maps on fraud ranks and fraud categories.

## PROGRAM CODE

```
/*-----*/
/* An example of fraud matrix */
/* (rankings of 11 providers by 7 fraud indices) */
/* Suppose the 7 fraud indices are equally important */
/* This paper is to give an example on how to combine */
/* multiple fraud indices into one comprehensive index */
/*-----*/

title1 'An Example of Fraud Matrix';
data fraud_matrix;
    infile 'C:\Documents and Settings\shiq\Desktop\New Folder\MWSUG\matrix.txt';
    input ID $5. Fraud_index1 Fraud_index2 Fraud_index3 Fraud_index4 Fraud_index5
    Fraud_index6 Fraud_index7;
run;

/* Use the PRINCOMP procedure to transform the observed ranks.*/
/* use the first principal component which has the highest */
/* eigenvalue of the dataset. */
/* It represents the most significant relationship between the*/
/* data dimensions. */

proc princomp data=fraud_matrix n=1 out=fraud_rank
    standard;
    var Fraud_index1--Fraud_index7;
run;

proc sort data=fraud_rank;
    by Prin1;
run;

proc univariate data=fraud_rank normal plot;
    histogram Prin1 /normal cfill=ligr cframe=liy barwidth=8 cv=black;
```

```

        inset mean std max min sum n nmiss median skewness;
run;

data fraud_rank;
    set fraud_rank;
    rank = _n_;
run;

ods html;
ods listing close;
ods graphics on;

proc tabulate data=fraud_rank;
    var Prin1;
    table Prin1*min Prin1*q1  Prin1*median  Prin1*q3 Prin1*p95 Prin1*max;
run;

proc print data=fraud_rank;
    title2 "Providers' Fraud Ranks Ordered by the First Principal Component";
run;
ods graphics off;
ods html close;
ods listing;

*cluster method;
*Method = average;

title "Cluster Analysis - Avarage Linkage";
proc cluster method=average data=fraud_matrix outtree=clust1;
    var Fraud_index1--Fraud_index7;
    id ID;
run;

proc tree horizontal nclusters=5 data=clust1 out=clust2;
    id ID;
run;

proc sql;
    create table merge as
    select a.*, b.cluster
    from fraud_rank a, clust2 b
    where a.ID=b.ID;
quit;

ods html;
ods listing close;
ods graphics on;

proc print data=merge;
    title2 "Providers' Fraud Ranks and Fraud categories";
run;
ods graphics off;
ods html close;
ods listing;

```

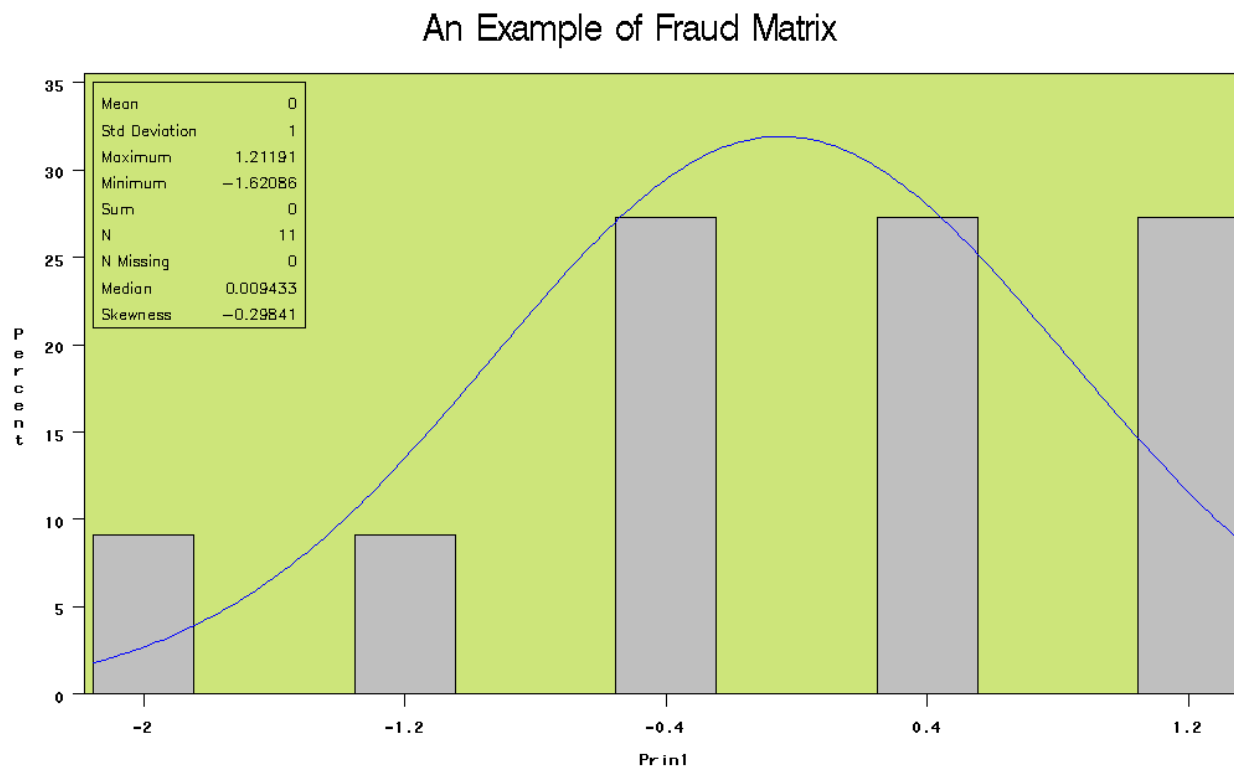
## RESULTS

The demonstrated data set “FRAUD\_MATRIX” has 11 observations and 8 variables. One variable is called “ID” which presents the provider’s identifications. The other seven variables are fraud indices: “Fraud\_index1” - “Fraud\_index7”. After PCA, we can get our comprehensive index “Prin1”.

Table 1: Basic Statistics of the first principal component “Prin1”.

Prin1	Prin1	Prin1	Prin1	Prin1	Prin1
Min	Q1	Median	Q3	P95	Max
-1.62	-0.77	0.01	0.94	1.21	1.21

Figure 1: Distribution of the first principal component “Prin1”.



From the above graph, we can see that “Prin1” can be binned into 5 categories which mean we may classify our providers into 5 fraud categories. This can be verified from the clustering later on.

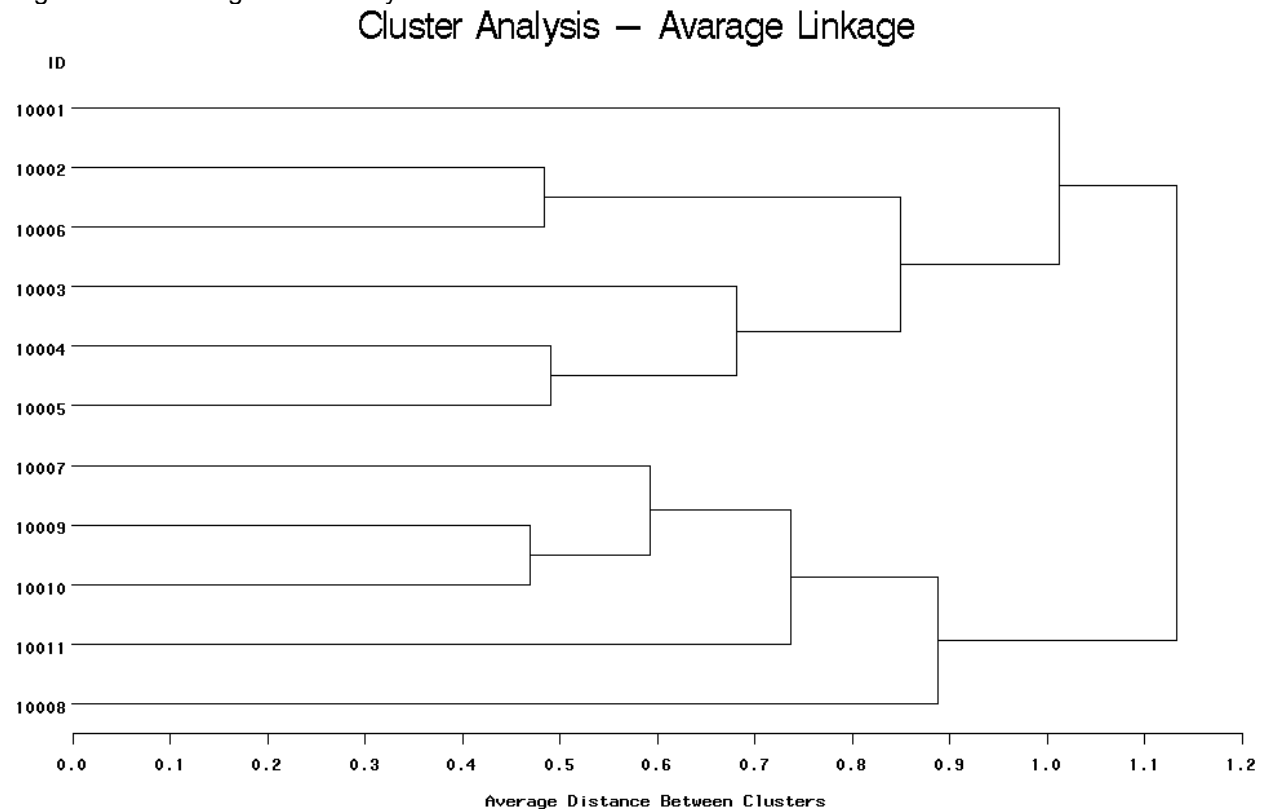
The following is the result of ranking providers by the comprehensive index “Prin1”.

Table 2: Fraud ranking of providers by the first principal component “Prin1”.

An Example of Fraud Matrix Providers' Fraud Ranks Ordered by the First Principal Component										
Obs	ID	Fraud_index1	Fraud_index2	Fraud_index3	Fraud_index4	Fraud_index5	Fraud_index6	Fraud_index7	Prin1	rank
1	10002	2	2	4	3	1	1	1	-1.62086	1
2	10006	6	1	2	2	3	4	2	-1.29859	2
3	10004	4	5	9	4	2	5	3	-0.77299	3
4	10003	3	4	5	1	5	11	8	-0.60850	4
5	10005	5	6	7	5	4	10	4	-0.36193	5
6	10001	1	8	1	9	8	9	6	0.00943	6
7	10007	7	10	6	11	6	6	5	0.61442	7
8	10008	8	3	11	6	11	2	10	0.74834	8
9	10009	9	7	8	8	7	8	11	0.93768	9
10	10011	11	11	3	7	10	3	9	1.14109	10
11	10010	10	9	10	10	9	7	7	1.21191	11

After we do the cluster analysis by average linkage method, we can see that the providers can be clustered into five different categories.

Figure 2: Clustering Providers by Fraud indices.



The following table is our final result which gives the information about the providers' fraud ranks and five fraud categories.

Table 3: Providers' Fraud Ranks and Fraud Categories.

<i>Cluster Analysis - Avarage Linkage Providers' Fraud Ranks and Fraud categories</i>										
ID	Fraud_index1	Fraud_index2	Fraud_index3	Fraud_index4	Fraud_index5	Fraud_index6	Fraud_index7	Prin1	rank	CLUSTER
10002	2	2	4	3	1	1	1	-1.62086	1	2
10006	6	1	2	2	3	4	2	-1.29859	2	2
10004	4	5	9	4	2	5	3	-0.77299	3	3
10003	3	4	5	1	5	11	8	-0.60850	4	3
10005	5	6	7	5	4	10	4	-0.36193	5	3
10001	1	8	1	9	8	9	6	0.00943	6	5
10007	7	10	6	11	6	6	5	0.61442	7	1
10008	8	3	11	6	11	2	10	0.74834	8	4
10009	9	7	8	8	7	8	11	0.93768	9	1
10011	11	11	3	7	10	3	9	1.14109	10	1
10010	10	9	10	10	9	7	7	1.21191	11	1

## CONCLUSION

From Table 3 we know that the top two providers (ID 10002 and 10006) are the most likely potential fraud leads and they belong to the same fraud category. Provider ID 10004, 10003 and 10005 which rank as 3rd, 4th and 5th places respectively should be classified as the second category of potential fraud leads. Provider ID 10001 and 10008 should be paid special attention to since they varied much in the seven fraud indices and do not belong to any categories. From the results of this study, we can see that data mining techniques such as principal component analysis and clustering can help people to find the potential fraud leads effectively.

## REFERENCES

1. The Practical and Ethical Considerations in Fraud Examination, CPA Journal, Joseph T Wells, June 2009.
2. "Detecting Medicaid Data Anomalies Using Data Mining Techniques", Southeast SAS Users Group Conference, 2010.

## DISCLAIMER

All opinions and suggestions stated in this paper do not necessarily reflect the opinions and suggestions of NCI Information Systems, Inc.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the primary author at:

Qiling Shi, Mathematics PhD, Certified Fraud Examiner  
 NCI Information Systems, Inc.  
 2636 Elm Hill Pike, Suite 115  
 Nashville, TN, 37214  
 Email: shiq@nciinc.com Email: shiqiling@gmail.com  
 Website: <http://www.linkedin.com/in/qilingshi>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.