LLOYDS BANKING GROUP

Segmentation Analysis Using Correspondence Analysis

Martin Harrison
Lloyds Banking Group
August 2011

Summary

LLOYDS BANKING GROUP

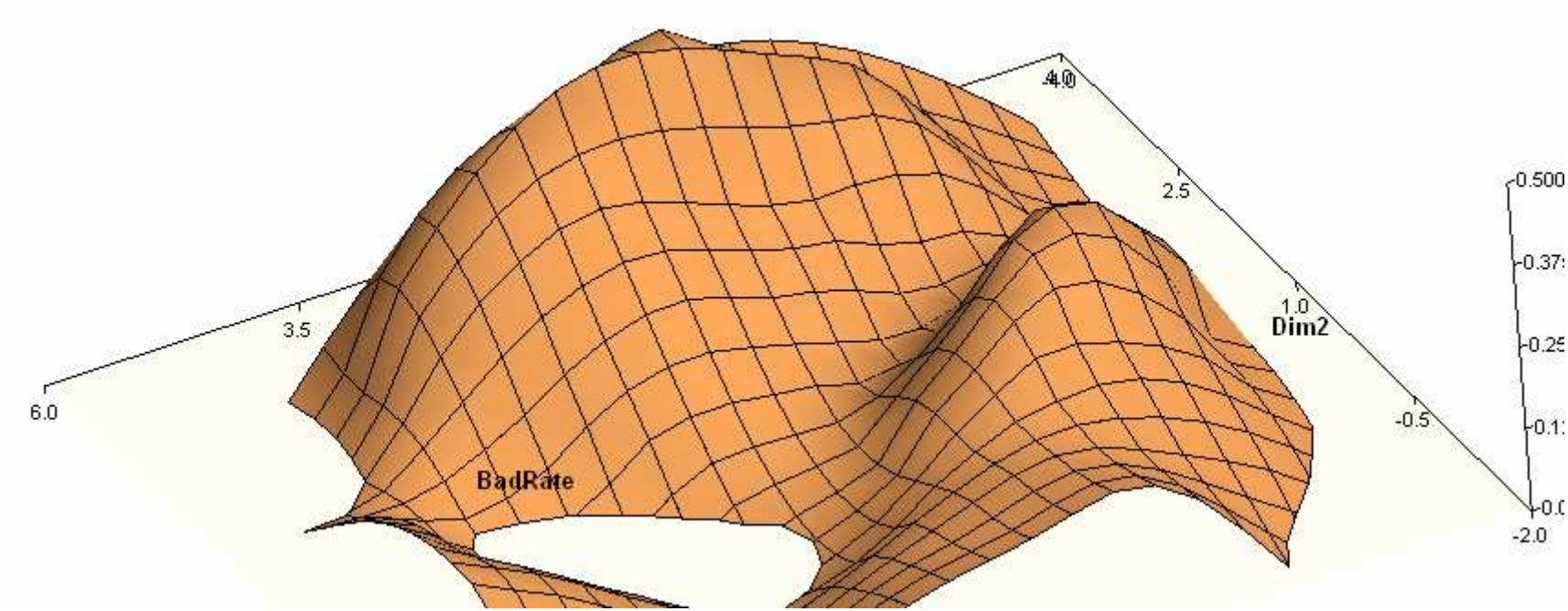
- Scorecard segmentation splits are used for several different reasons:
  - Business judgment.
  - Data availability.
  - Different profile of bads.
- Identification of Scorecard/Model segmentation splits is not undertaken with any standard industry technique.
- In many occasions segmentation is decided by expert judgement or some analysis on Information Value. This can cause many issues:
  - Lack of expert knowledge.
  - Time consuming trial and error.
  - Lack of mathematical credence to choosing those segmentors in the first place.
  - A desire to split due to severity of ‘bads’ as apposed to different types of bads.
- Correspondence Analysis gives us an opportunity to solve all of these issues by:
  - Allowing a non-expert to test several possible segmentors at once.
  - Giving a mathematical justification for the variables to be used to segment cases.

CORRESPONDENCE ANALYSIS CAN BE USED AS A SCORECARD SEGMENTATION AID.

Contents

LLOYDS BANKING GROUP

1. What is Correspondence Analysis?
2. Simple Correspondence Analysis.
3. Multiple Correspondence Analysis.
4. Use of Correspondence Analysis in Segmentation.
5. Examples.
6. Further Work.



Correspondence Analysis

LLOYDS BANKING GROUP

- Correspondence analysis is a multivariate statistical technique which (for those who have used it) is similar in concept to Principle Components Analysis but applies to Categorical Data.
- It is able to profile cases without a need for a defined target.
- The interpretation of the data can be explained in a very simple way.
- It provides a means of summarising a set of data and displaying it in a 2-d or 3-d graphical output. Which can aid you in developing a case for segmentation with stakeholders.
- It can be run using a simple procedure in SAS (Also equivalent procedures in R and Matlab).

Correspondence Analysis Example

LLOYDS BANKING GROUP

- We want to determine if there is any relationship between the job type/level and the level of smoking.

- For example in the summary data below 25 of the Senior Employees are non-smokers but only 4 of the senior employees are heavy smokers. We can then use correspondence analysis to determine what the relationship is.

Job Type	Smoking			
	None	Light	Medium	Heavy
Sr. Managers	4	2	3	2
Jr. Managers	4	3	7	4
Sr. Employees	25	10	12	4
Jr. Employees	18	24	33	13
Secretaries	10	6	7	2
Total	**61**	**45**	**62**	**25**

Example taken from http://www.statsoft.com/textbook/correspondence-analysis/

Simple Correspondence Analysis

- First of all calculate the matrix **D** (the Matrix of Inertias) where:

$$d_{ij} = \frac{1}{\sqrt{Total\ Sample\ Size}} \frac{(observed\ \ frequency - expected\ \ frequency)}{\sqrt{expected\ \ frequency}}$$

- This is essentially the formula for calculating the chi-square statistic.

- If a value of d_{ij} differs greatly from zero it is said that it this variable differs greatly from the norm.

- We also get

$$\sum_i \sum_j d_{ij} = \frac{\chi^2}{n}$$

i.e. The total chi-square information (inertia of the matrix).

- This goes by several names in correspondence analysis. The term inertia is taken from mechanics and means the difference from the overall norm.

Simple Correspondence Analysis

- We now begin the process of performing the correspondence analysis. First of all we create the create the matrix which is equivalent to the sum of squares. This is **D'D.**

- Eigenvectors of a matrix are those that, after being multiplied by the matrix, remain proportional to the original vector. For each eigenvector, the corresponding eigenvalue is the factor by which the eigenvector changes when multiplied by the matrix.

- An interesting property of **D'D** is that it's eigenvalues have a similar property to the of the elements of **D.** In that:

$$\sum_r \lambda_r = \frac{\chi^2}{n} \qquad \lambda_1, \lambda_2, \ldots, \lambda_{\min(r-1, c-1)}$$

- Because we have eigenvalues we can also calculate the corresponding eigenvectors. The number of possible eigenvectors we get is the minimum of the number of rows and number of columns.

Example

LLOYDS BANKING GROUP

- If we go back to our smoking example. We can see that the matrix **D'D** will have 3 eigenvalues (4 columns – 1). We also know that the total of the eigenvalues add up to the total chi-square information in the data. So if we rank order the eigenvalues

Axis (Factor)	Eigenvalue	% of Inertia	Cumulative %
1	0.07476	87.6	87.6
2	0.01002	11.8	99.5
3	0.00041	0.5	100

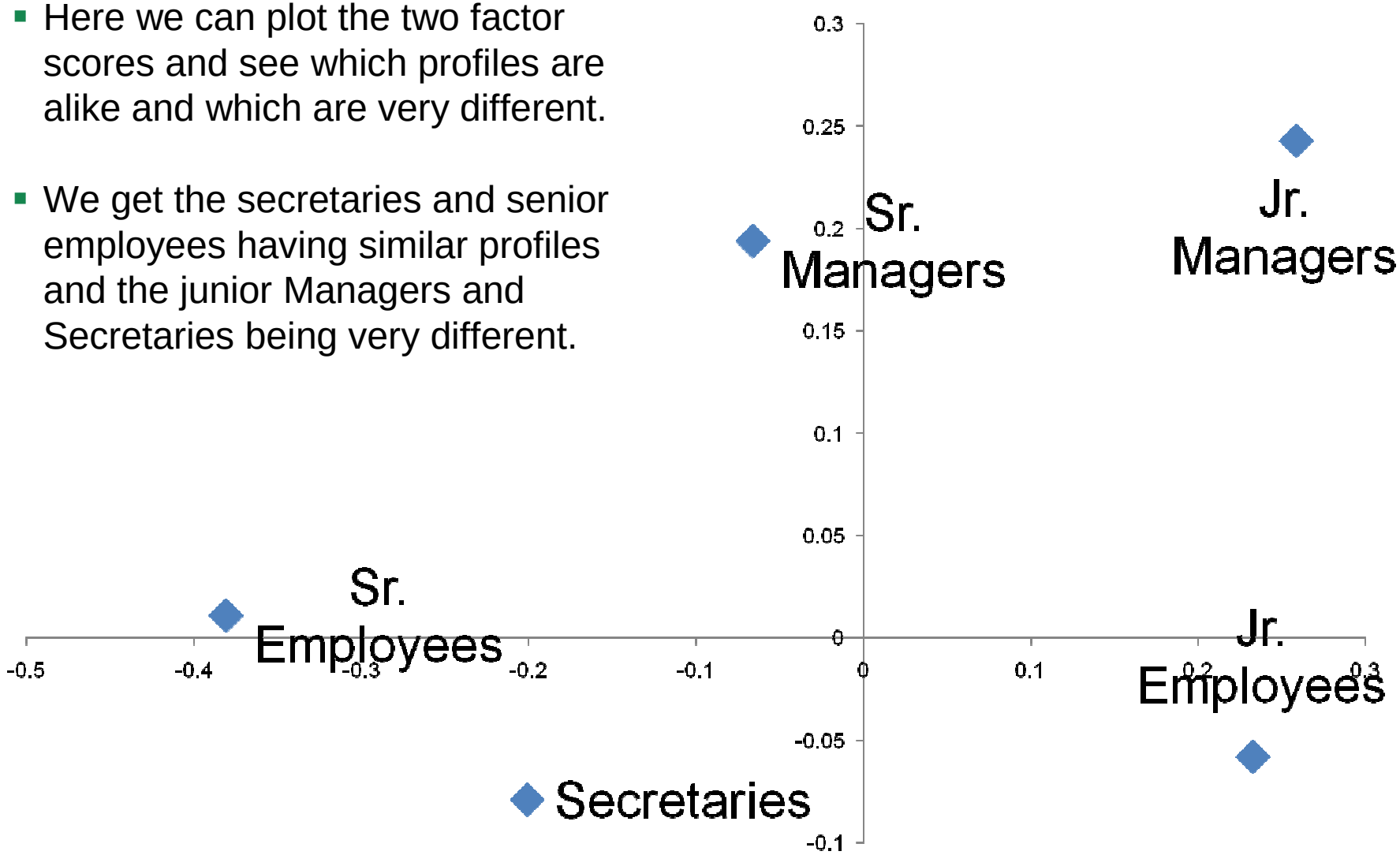
- We see that the first two eigenvalues accounts for most (99.5%) of the inertia and when we include the second factor we account for nearly all the variability. For this reason we only need look at the first two factors. The eigenvectors (or as they are referred to in CA the factor scores) are as follows:

Job Type	Factor Score 1	Factor Score 2
Sr. Managers	-0.066	0.194
Jr. Managers	0.259	0.243
Sr. Employees	-0.381	0.011
Jr. Employees	0.233	-0.058
Secretaries	-0.201	-0.079

Example taken from http://www.statsoft.com/textbook/correspondence-analysis/

Example

- Here we can plot the two factor scores and see which profiles are alike and which are very different.
- We get the secretaries and senior employees having similar profiles and the junior Managers and Secretaries being very different.



Multiple Correspondence Analysis

LLOYDS BANKING GROUP

- While simple correspondence analysis looks at the profiles of different variables.

- With multiple correspondence analysis we are able to look at the profiles of different cases. If we take the example from earlier. We can break down the table into :

Case Number	Senior Manager	Junior Manager	Senior Employee	Junior Employee	Secretary	None	Light	Medium	Heavy
1	1	0	0	0	0	1	0	0	0
2	1	0	0	0	0	1	0	0	0
3	1	0	0	0	0	1	0	0	0
4	1	0	0	0	0	1	0	0	0
5	1	0	0	0	0	0	1	0	0
...	.	.	.	.	.	.	.	.	.
...	.	.	.	.	.	.	.	.	.
...	.	.	.	.	.	.	.	.	.
191	0	0	0	0	1	0	0	1	0
192	0	0	0	0	1	0	0	0	1
193	0	0	0	0	1	0	0	0	1

- Here we see the example that case number 5 is a Senior Manager and a Light Smoker.

- This will allow us to create profiles to fit each different case based upon the variables we put in.

Burt Table

LLOYDS BANKING GROUP

- In order to perform the correspondence analysis we to summarise this table. Multiplying it by it's transpose.This is referred to as the ***Burt Table***.

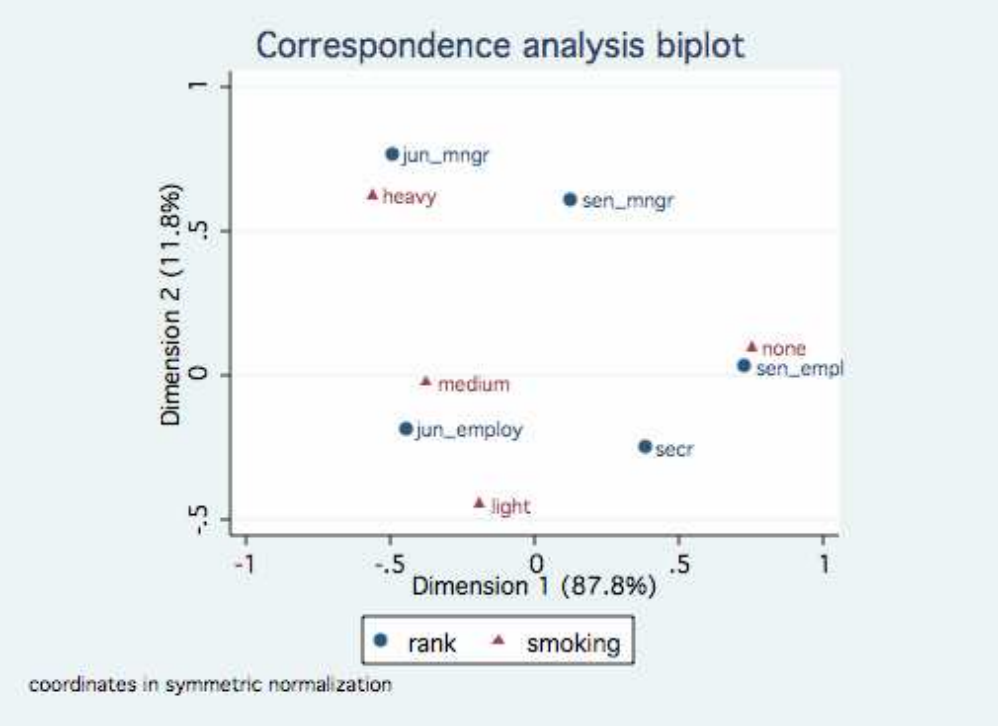
- So for the previous example we get:

	Employee					Smoking			
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)
(1) Senior Managers	11	0	0	0	0	4	2	3	2
(2) Junior Managers	0	18	0	0	0	4	3	7	4
(3) Senior Employees	0	0	51	0	0	25	10	12	4
(4) Junior Employees	0	0	0	88	0	18	24	33	13
(5) Secretaries	0	0	0	0	25	10	6	7	2
(1) Smoking:None	4	4	25	18	10	61	0	0	0
(2) Smoking:Light	2	3	10	24	6	0	45	0	0
(3) Smoking:Medium	3	7	12	33	7	0	0	62	0
(4) Smoking:Heavy	2	4	4	13	2	0	0	0	25

- We get a similar table to the earlier example but including all cross-tabulations. If we perform the same algebraic manipulations as with the simple correspondence analysis we can analyse the interactions between multiple variables.

Graphical Interpretation

- From this we can now visualise which of the characteristics are corresponded to each other.
- Those which appear to be close to each other are highly corresponded and those which are far apart are antitheses.
- For example Senior Employees tend to be non-smokers and Senior Managers tend not to be light smokers.



Applications of Correspondence Analysis

- The main current use for this comes from marketing analytics. Profiling customers in order to provide the best targeted products.

- In credit risk we can use the different profiles in many ways:
  - Segmentation within scorecard building.
  - Find out which profiles are generally bad customers.
  - Find out which of the generated profiles contain dormant credit card customers.
  - We can look at the migration between these profiles as a method for explaining instability within scorecards.

- The primary focus of this paper is to look at the use of correspondence analysis in Segmentation.

Correspondence Analysis in Segmentation

LLOYDS BANKING GROUP

- The main aspect we are going to look at is the possibility of finding new segmentors when we come to build new scorecards.
  - Logistic Regression finds variables which can lead you toward a defined target.
  - Correspondence Analysis does not require a target, however it is able to find predictors for different types of targets.

- We now provide an algorithm for how to use correspondence analysis in a scorecard build for segmentation. It breaks down as follows:

1. Select a list of predictive variables.
2. Variable Classing.
3. Coding Correspondence Analysis.
4. Profile the cases.
5. Cluster the profiled cases.
6. Analyse the clusters.

1. Select a list of predictive variables.

LLOYDS BANKING GROUP

- The first step is to select a list of possible variables. This can be done in several ways.

1. Expert Knowledge:
   - If you have the knowledge available from previous scorecard/model builds you may be able to add in several variables you have seen in the past that can be good ways to segment a population.

2. Previous Scorecards and sub-population splits:
   - This will give you a list of variables you know have been good at identifying your previous targets.

3. A ‘first-run’ build of a scorecard/model
   - If you are building a brand new scorecard/model and do not have any expert knowledge or know how previous scorecards have been build. The suggestion is to build a ‘first-run’ model. This will give a similar result as method 2, where we have a list of variables which we know are good at predicting the target. What we wish to know is do these variables lean towards the same type of target or do some lean towards different ones.

2. Variable Classing

- To use correspondence analysis we need to ‘bin’ or ‘bucket’ each of the variables into categories. How you do this will effect how the correspondence analysis. Again there are several methods we can take in binning the variables for example:

1. Leaving the categorical variables as they are. For example ‘flags’ can remain as they are.

2. Use the binning from a previous scorecard.

3. Create a binary split in the variable based upon weight of evidence toward the desired target.
   - What we mean by this is creating two characteristics:
     - Variable1_Good
     - Variable1_Bad

     where variable1_Good leans towards having a higher proportion of good customers and Variable1_Bad has a higher proportion of bad customers.
   - What we are able to with this information is see if any of the ‘bad’ characteristics correspond to any of the good characteristics.
   - If they do this is definite evidence for a segmentation.

3. Coding the Correspondence Analysis

- We can now look at each of these variables in terms of the correspondence analysis in SAS.

```
proc corresp mca data=Table outc=Scores;
Tables (List of variables);
run;
```

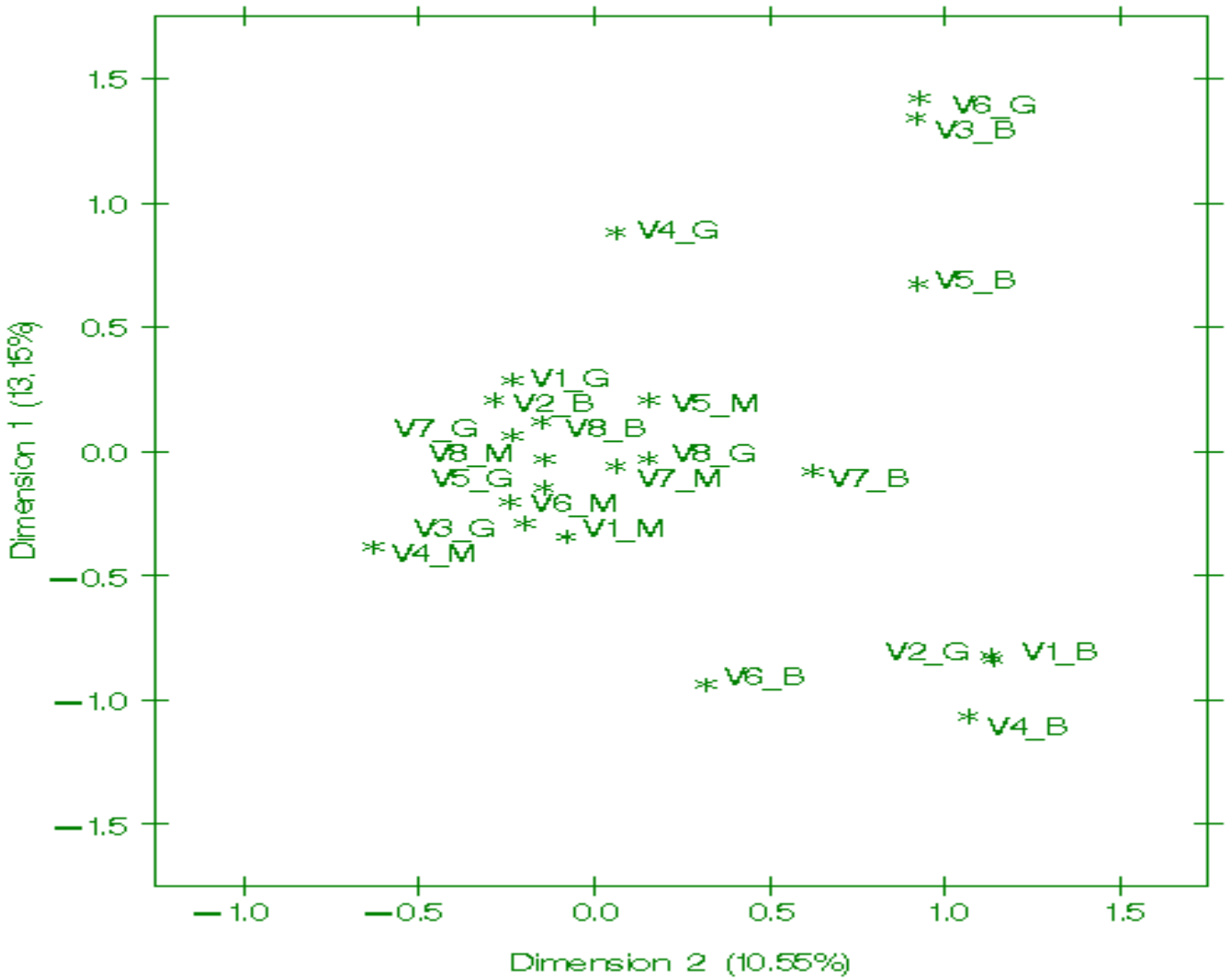
- The list of variables here are the classed categorical variables.

- This will produce the following results:
  - The Inertia output for all possible eigenvalues.
  - The factor scores for each characteristic.
  - Summary Statistics.

- The output dataset contains a list of characteristics and the factor scores which we can use later.

- Using the %Plotit Macro we can display the relationships between the characteristics. The ones which are close together are highly corresponded. Those which are far apart are antitheses.

Example of output from %Plotit Macro



4. Profiling Cases

- We can now take the information we have on the different characteristics and apply that to each of the cases.
- We do this using the factor scores for each of the characteristics and the initial multiple correspondence table.
- If we go back to our original example we have the first case being:

Case Number	Senior Manager	Junior Manager	Senior Emp	Junior Emp	Sec	None	Light	Medium	Heavy
1	1	0	0	0	0	1	0	0	0

- We can calculate the factor score for case 1 as
  - Score 1 = FS1(Senior Manager) + FS1(None) = Dim1
  - Score 2 = FS1(Senior Manager) + FS1(None) = Dim2
- We can then plot each of the cases on the same two dimensional graph.
- We can then use cluster analysis to profile cases into particular groups.

4. Profiling Cases

LLOYDS BANKING GROUP

- We can profile cases using Proc Score.

```
proc score data=Initial Table score=Coor
out=Scored;
 var (List of Characteristics);
run;
```

- This uses the output of the scores from Proc Corresp. You will need to transpose the dataset used to be run in Proc Score.

- The input table for this must be the original table containing the 1's and 0's for each characteristic.

- This will then produce two new variables in a new dataset. SAS automatically call's these dim1 and dim2. Where each score is the sum of the factor scores of the category they fall into.

5. Cluster the Different Profiles

LLOYDS BANKING GROUP

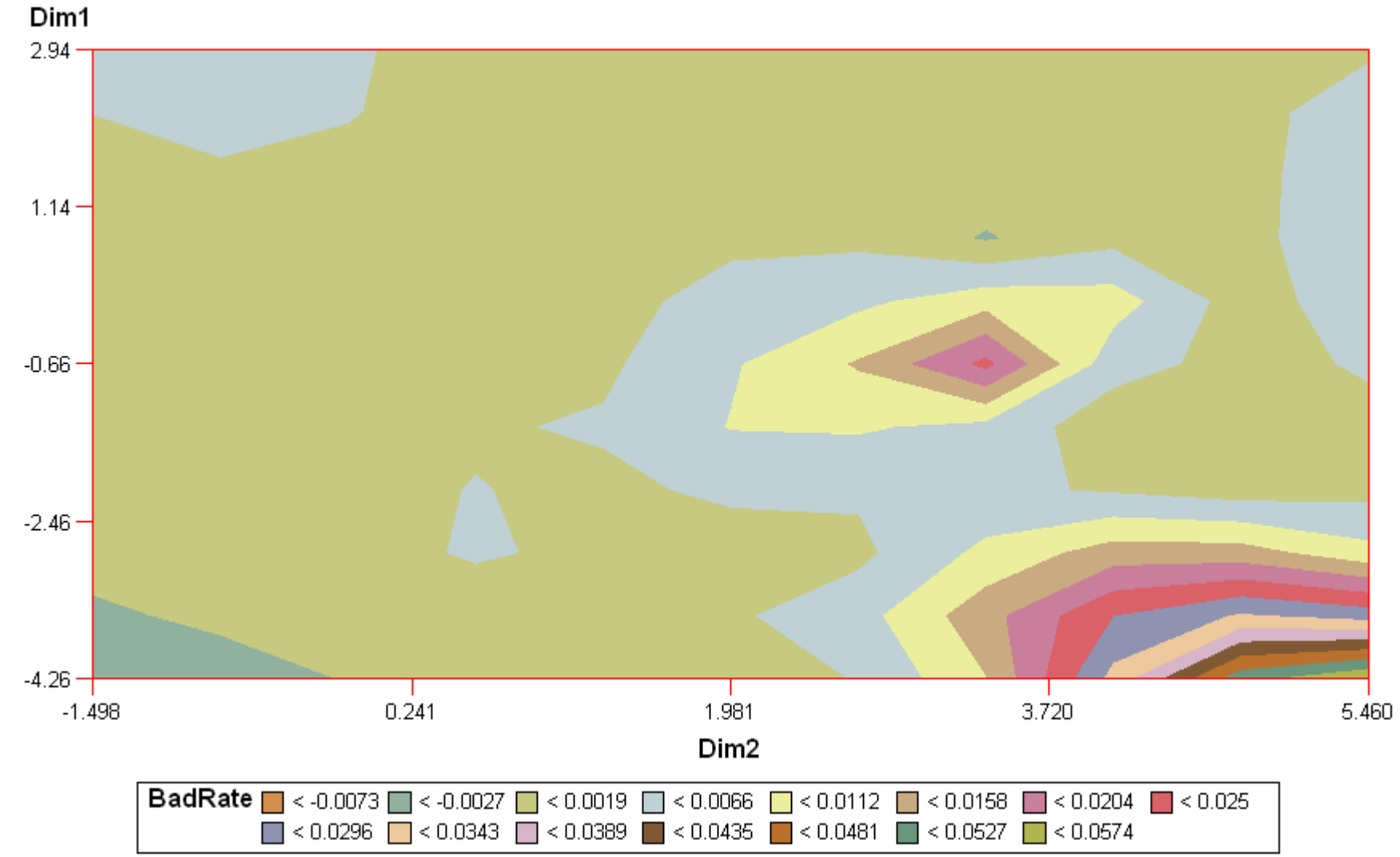
- As we stated earlier the correspondence analysis in SAS automatically produces two variables or scores for each case called Dim1 and Dim2. Using Proc Fastclus we are able to group these cases together.

```
proc fastclus data=Scored maxc=50 out=clus;
var dim1 dim2;
run;
```

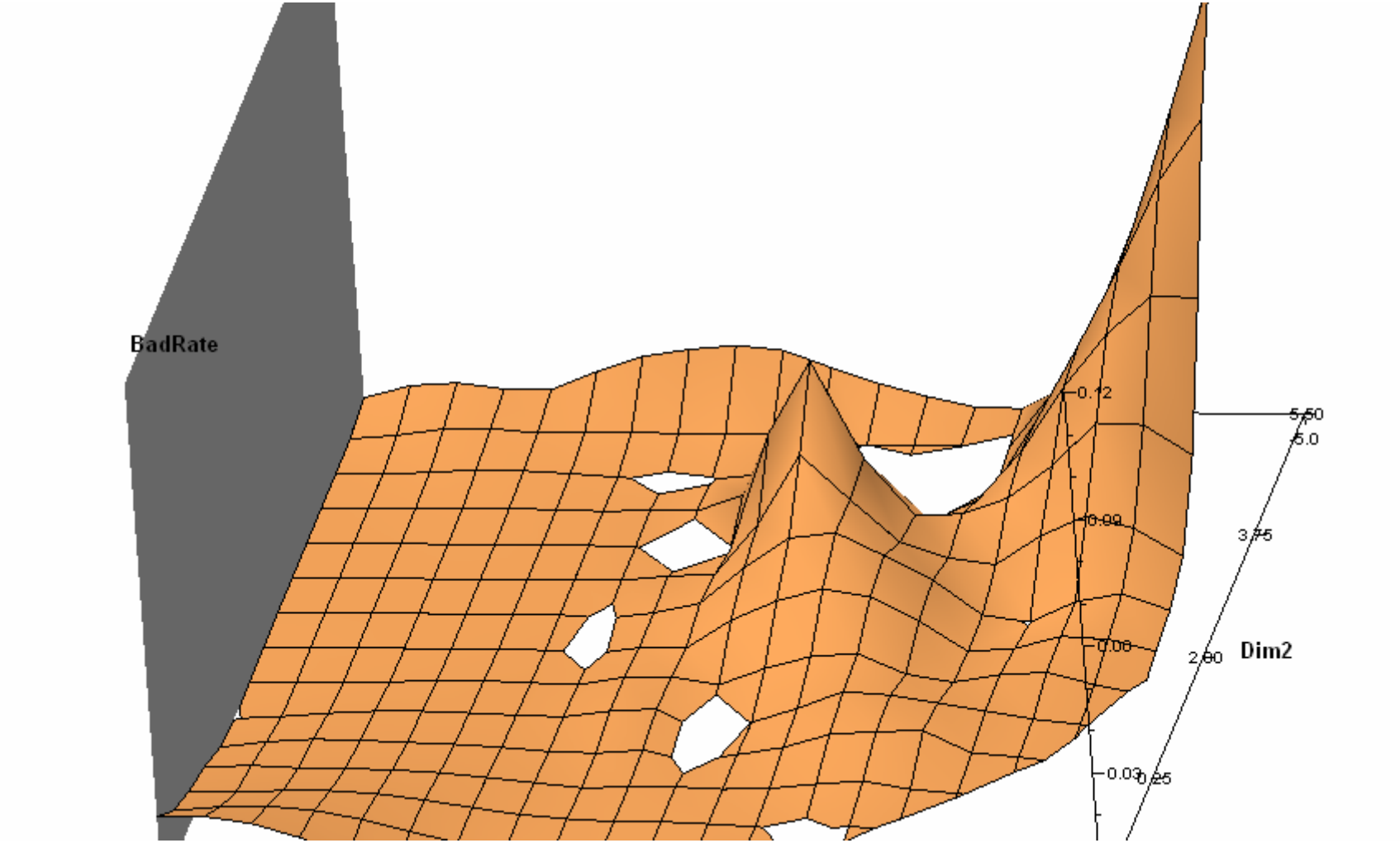
- This will create 50 clusters (profiles) of customers. You should choose clusters so that the profiles will be different enough to allow for good visualisation while still containing a sufficient volume.
- The output dataset will contain the cluster number the case falls into.
- Using this we can run a simple Proc Freq to determine which clusters have a higher volume of our targets.
- A much nicer way to look at this is to display out output graphically. We essentially have 3-dimensions to display. We take the cluster mean to be our factor scores for those in that profile and plot this against the bad rate for this cluster.

6. Analyse the Clusters

LLOYDS BANKING GROUP



6. Analyse the Clusters



6. Analyse the Clusters.

LLOYDS BANKING GROUP

- Using the graphical displays we can try and locate what is driving these clusters.
- The simplest way of doing this is finding out which of characteristics in the target clusters differ in proportion from the norm.
- For example if we have:

Variable 1	% in 'Target' Cluster	% in Population
Variable1_Good	10%	50%
Variable1_Bad	90%	50%

- We can say that Variable 1 is driving the cluster because it has a higher proportion of the bads in the target area. If we also have:

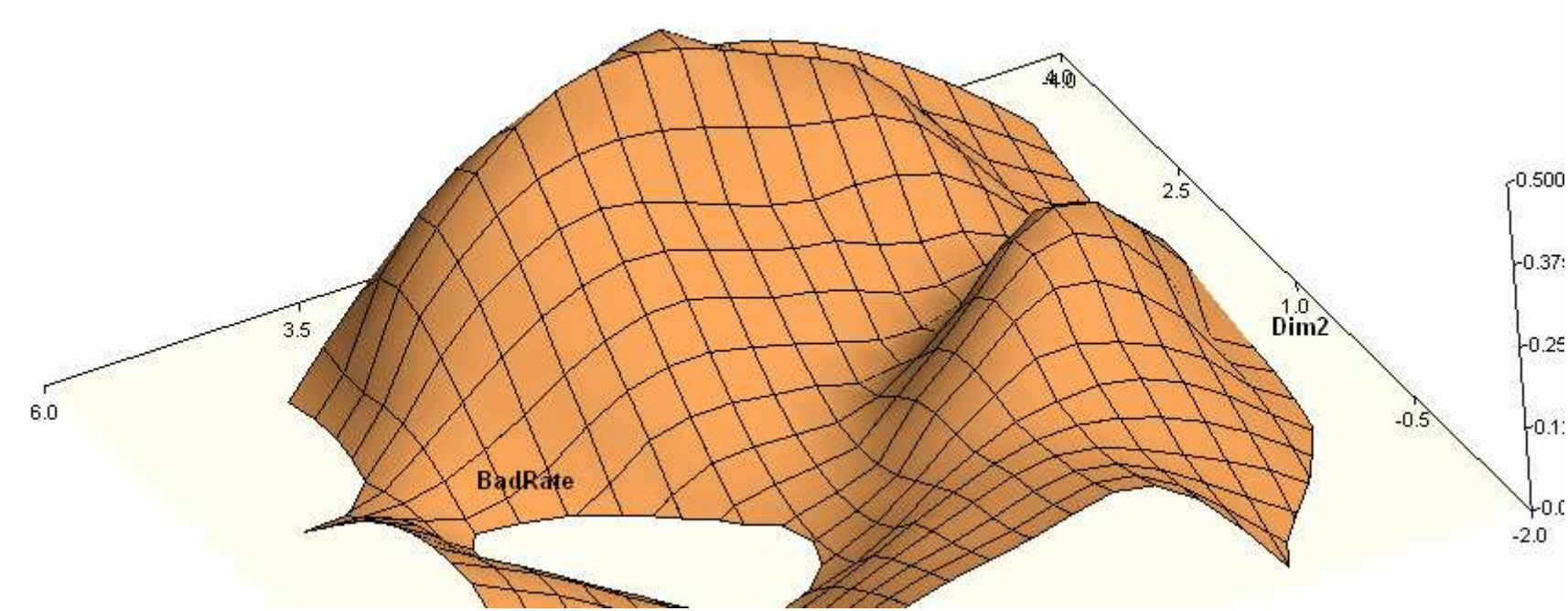
Variable 2	% in 'Target' Cluster	% in Population
Variable2_Good	90%	50%
Variable2_Bad	10%	50%

- We can now say that we may want to segment on variable 1 because the target cluster is shows a difference in what is driving it.

Further Examples

LLOYDS BANKING GROUP

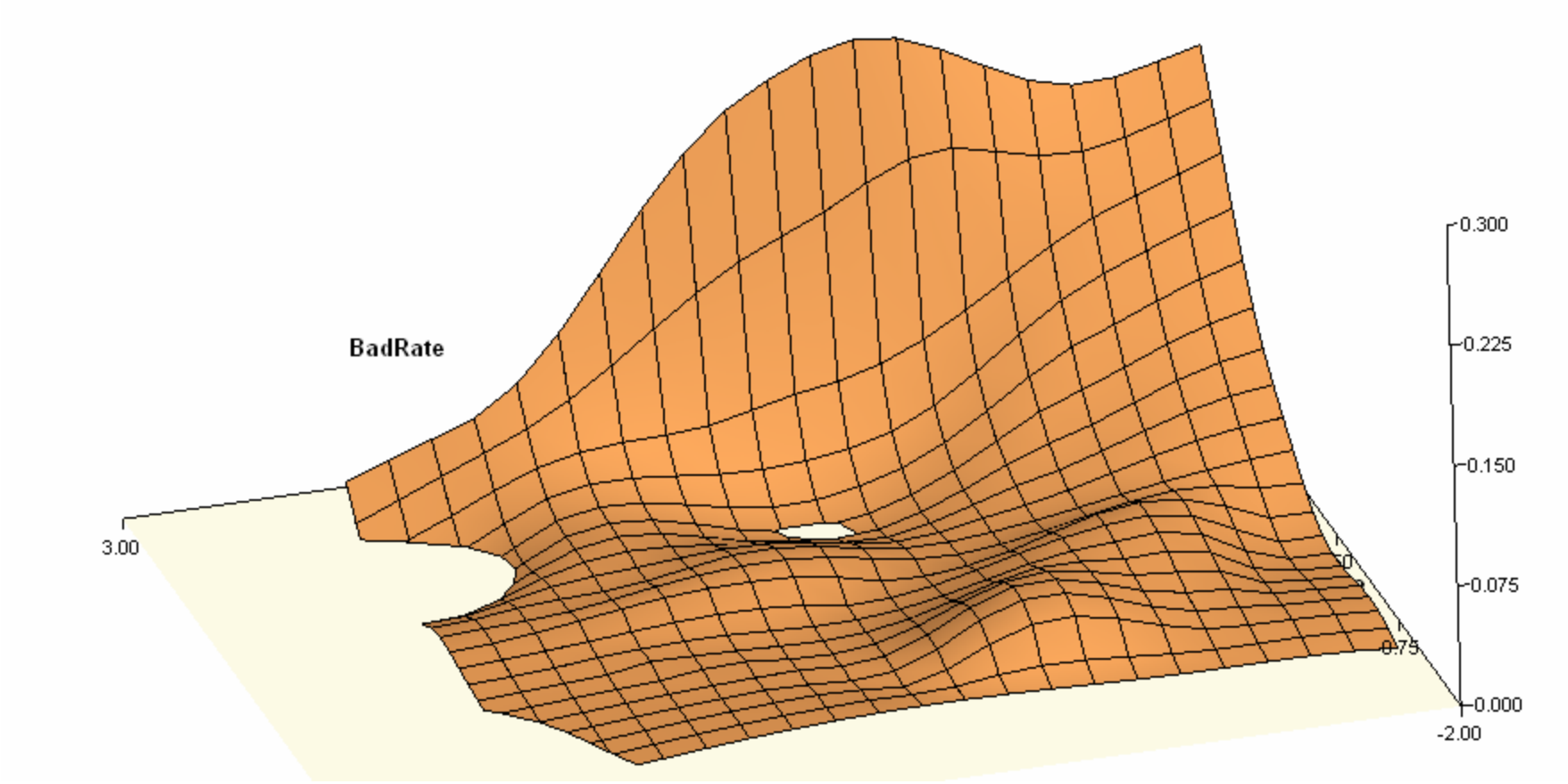
- Here we have an example of some credit risk data.
- Again we see a couple of different ‘hills’. This indicates there may be a split to be had in the data.
- In fact it was concluded that there should be a ‘time on books’ split for this particular example.



Further Examples

LLOYDS BANKING GROUP

- Analysis of some fraud data. Graph shows that the majority of the bads are in one peak.
- This suggests that there is no need for a segmentation on the model.

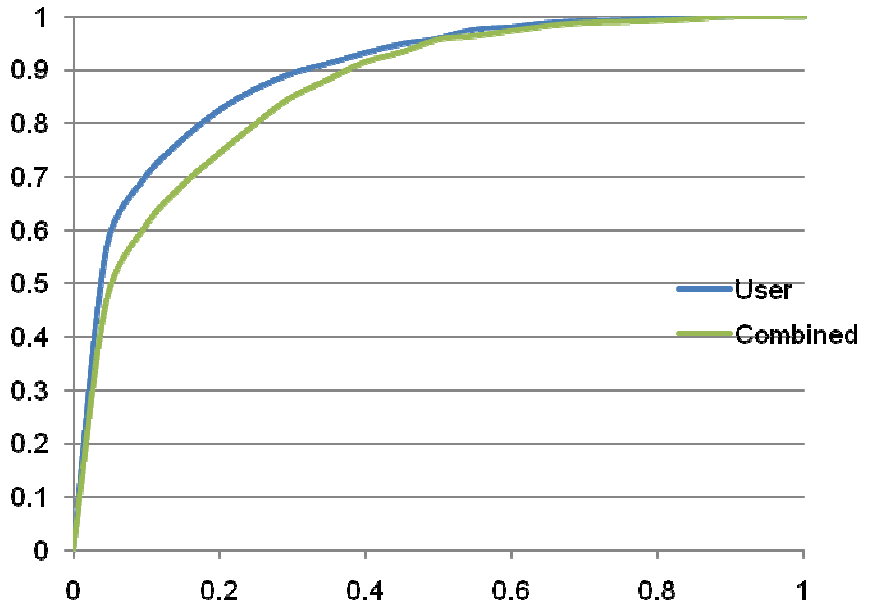


Example of use

- We found such an example within one of our model builds. In fact our example graphs show the two hills that we wish to segment on. If we look at the correspondence graph we can see we need to split on:
  - V6_G
  - V3_B

- We then ran this through our usual scorecard building procedure to see if we could get a Gini uplift from using this segmentation.

- We found the following:
  - Parent Model Gini – 72.57%
  - Composite Gini from the Segmentation – 78.26%



Further Work – Model Monitoring

LLOYDS BANKING GROUP

- One possible application of this is to look into how correspondence analysis can be used to monitor our existing scorecards/other models.
- We could monitor the models based upon different clusters. If the model is working we should see the main level of targets drop off.
- However if we see new ‘hills’ arising in the model we can see what is driving that change.
- We can also look at stability of the profiles and see how people are moving between them.
- Would mean that we might be able to create an additional segment to an existing model rather than a complete reweight/rebuild.

Pros and Cons

LLOYDS BANKING GROUP

Cons

- There is still a requirement to have some expert judgement. Knowing whether the split makes business sense as well as mathematical sense.
- Using too many variables can make the output overly complex and tough to analyse. (Hence the suggestion).
- Requires some idea of variables to put into the model.
- Governance of this model would need addressing.

Pros

- Powerful visualisation tools for segmentation.
- Runs quickly using SAS.
- Provides Mathematical credence to the models.
- Provides insight into new segmentations that have not been given a business case before.