

# ADVANCED PRINCIPAL COMPONENTS ANALYSIS: A REAL-LIFE APPLICATION USING *STATISTICA 6* STATSOFT, INC.

**DATE:** October 30<sup>th</sup> 2001

**AUTHOR:** Walter Kuypers, IBM Amsterdam

## ◆ Introduction

In my final year of chemistry studies, I was given the opportunity to go abroad for a placement and final project at the De Montfort University (DMU) of Leicester in 1999.

DMU has always valued the the importance of research, therefore it has many centers of research, including Lubricant Research Centre (LRC).

The LRC carries out certain investigative queries from industry. My combined placement and final project was based on one of industry's queries funded by the British government and an industrial partner British Petroleum (BP).

## ◆ Lubricants

In industry many long lubricants tests are done to investigate lubricants on their performance of dispersability and suspendability. Important aspects in these investigations are the amount of degradation and oxidation of lubricants. The basic function of lubricant is the reduction of friction, transfer of heat and the suspension of contaminants. The ability of a lubricant to remain effective in the presence of these contaminants is very important. The amount of contaminants (which are by-products of the combustion process) and the interaction of oils towards these contaminants gives an indication of the performance of oils.

## ◆ The Project

The project was based on trying to produce a more cost effective method of testing Lubricants to the CEC 1431 60 hour car-engine test standards and its further developments. The aim was to determine whether a 6 hour engine test with an oil sump reduction of 75-80% could lead to the same or even more specific information about the characteristics of oils than the non-reduced sump 60 hours test method.

The data derived from the analysis of oil samples taken from the engine at different hours and rotation speeds (RPM) was a result of a diversity (in total 27) of chemical

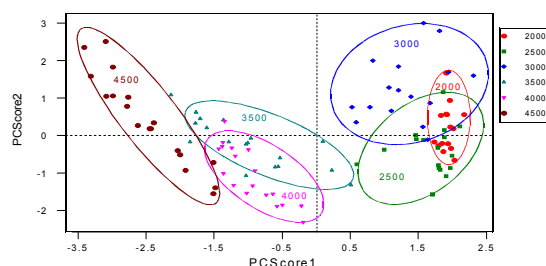
measurements on these oil samples. Some examples are particle size analysis, Infrared, viscosity, and light absorbance. When the 27 variables are taken into account as well as the 60,000 rows which result from replicate measurements, it is possible to say that a univariate analysis is inappropriate here. The solution to the analysis of complex databases like these is Multivariate Data Analyses (MDA). In this project Principal Component Analysis (PCA) was used to detect hidden phenomena mainly through reducing the variables to find the most important ones, finding patterns in the data and classifying any combination of objects or variables.

By using PCA the total amount of variables needed for oil characterization was reduced by 30% leading to a cost reduction of 45,000 British pounds. In addition, it was statistically proven that a 6 hour test provided the same results as a 60 hour test, which led to 90% reduction in manpower.

## ◆ Statistical software

Minitab Inc.'s SCAN release 1.0 (1995) and StatSoft's *STATISTICA* release 6.0 were used to carry out the Principal Components Analysis.

The picture below shows an example of a score (objects) plot output chart from SCAN.



## ◆ *STATISTICA 6.0*

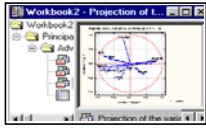
Due to my great interest and curiosity in statistical software I was happy to be able to work with *STATISTICA 6.0*. When I compare SCAN with *STATISTICA* I can see a lot of advantages such as:

- many algorithms to choose from (SCAN only 2),
- 3 dimensional charts (SCAN only has 2D charts),
- the ease of using external databases (Web, DB2, Oracle). The function of importing spreadsheets from other sources makes this tool very user friendly (no manual conversion needed),

A disadvantage of using *STATISTICA* compared to SCAN is the wide range of possibilities. The structure of the package is complex and low in transparency for a starting user.

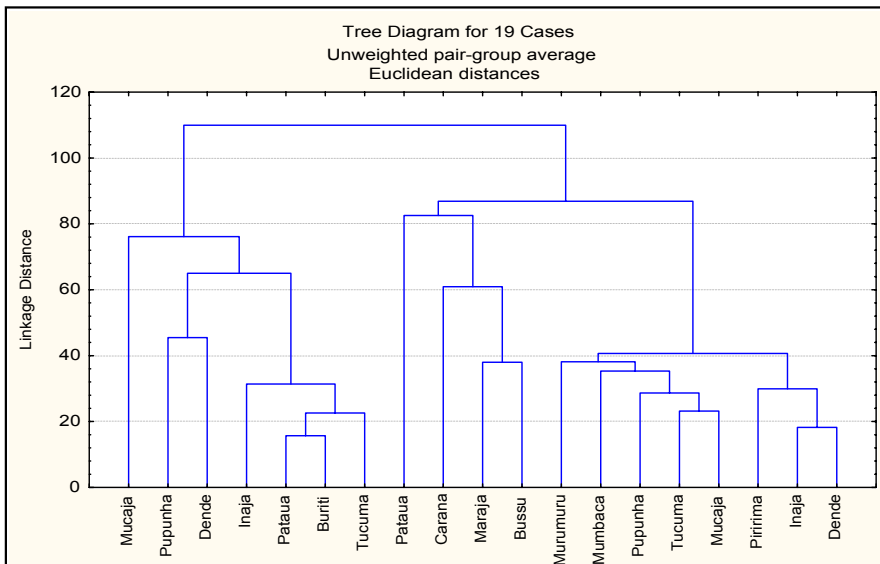
# STATISTICA 6

## EXAMPLE WITH PALM OILS ANALYZED ON DIFFERENT CHEMICAL CHARACTERISTICS



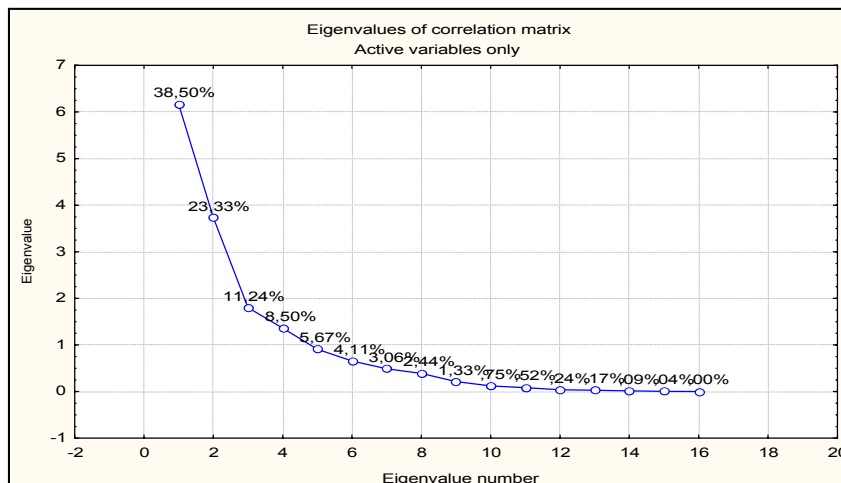
The following data are the outcome of an analysis of different Palm oils (different brands and different types) based on several chemical characteristics such as the different acids it contains. The data set contains 19 different Palm oils which were analyzed on the variables:

Yield	RI	Acidity	Saponifi
Iodineln	Isaponi	Caprilic	Capric
Lauric	Miristic	Palmitic	Palmitol
Stearic	Oleic	Lenolic	Class



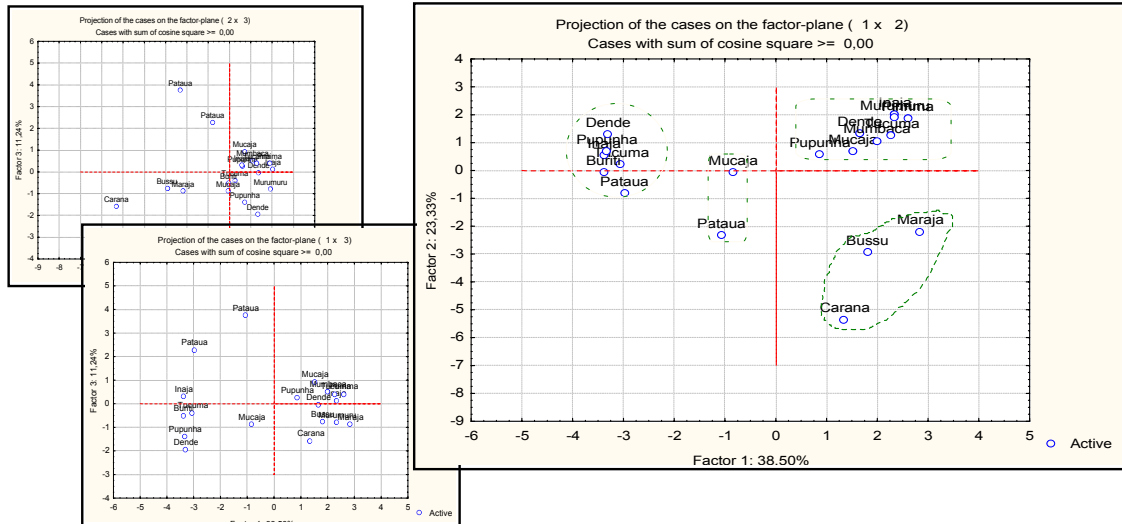
The Dendrogram shows the way the different Palm oils cluster. The dendrogram clearly shows two clusters between the first 7 oils and the further 12. In the second step we can try to see which variables contribute to the spread in the data.

First the amount of variation in the data that can be explained by different spread factors (Principal Components) has to be investigated. The figure below shows how many PC's explain how much of the variation in the data.

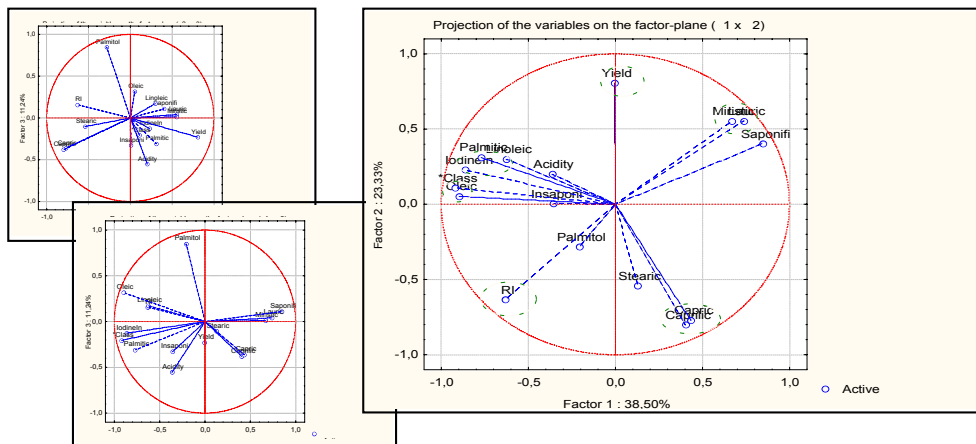


From the Eigenvalue plot it is possible to say that with 3 principal components around 72% of the variation in the data can be explained. So instead of investigating all 16 variables in a univariate way, we have now reduced the amount of variables to 3 main factors (principal components).

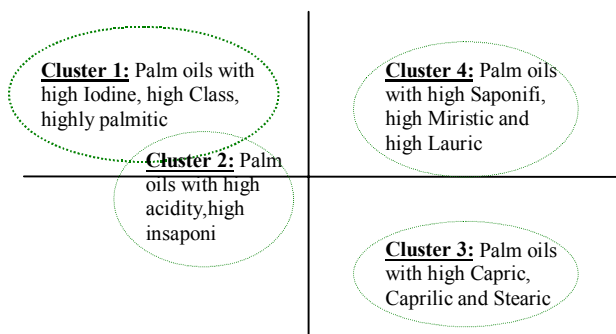
The question is now, how it is possible to visualize the differences between Palm oils? To see if Palm oils are different from each other the case score plot is needed. This shows how the different objects (cases) score on the variables.



In the factor score plot it shows that there are clear differences between Palm oils. Now the question rises: Which variables contribute to the spread along Factor 1 and Factor 2 as these 2 factors already explain more than 60% of the variation? This can be seen in the variable score plot.



From the cases score plot (factor 1 vs factor 2) roughly 4 clusters of Palm oils can be defined, which can then be described as shown below.



The variable score plot shows variables that cluster within close range with each other when a high percentage of variation in the data explained. Looking at the variable score plot on the previous page, the following variables cluster:

**Cluster 1:** Capric & Caprilic

**Cluster 2:** Miristic & Lauric

**Cluster 3:** IodineLn & Palmitic & Linoleic

This last cluster doesn't seem to be a good cluster when the 3<sup>rd</sup> factor is compared with the 1<sup>st</sup> factor. This would have been more visible if a 3D variable score plot option was available.

A second possibility for variable reduction determination is the use of the correlation matrix of the variables. The higher the correlation (either negative or positive) between two variables, the higher the equality in information. This can result in rejecting the most expensive analysis. The *STATISTICA* correlation matrix is shown below:

Variable	Correlations (Palms)									
	RI	Saponifi	IodineLn	Caprilic	Capric	Lauric	Miristic	Palmitic	Oleic	*Class
RI	1,00	-0,74	0,40	0,20	0,14	-0,76	-0,73	0,22	0,52	0,50
Saponifi	-0,74	1,00	-0,72	0,01	0,06	0,77	0,72	-0,53	-0,69	-0,79
IodineLn	0,40	-0,72	1,00	-0,46	-0,49	-0,40	-0,39	0,63	0,76	0,96
Caprilic	0,20	0,01	-0,46	1,00	0,99	-0,18	-0,24	-0,46	-0,53	-0,35
Capric	0,14	0,06	-0,49	0,99	1,00	-0,15	-0,22	-0,47	-0,54	-0,38
Lauric	-0,76	0,77	-0,40	-0,18	-0,15	1,00	0,83	-0,48	-0,64	-0,56
Miristic	-0,73	0,72	-0,39	-0,24	-0,22	0,83	1,00	-0,40	-0,58	-0,52
Palmitic	0,22	-0,53	0,63	-0,46	-0,47	-0,48	-0,40	1,00	0,55	0,71
Oleic	0,52	-0,69	0,76	-0,53	-0,54	-0,64	-0,58	0,55	1,00	0,77
*Class	0,50	-0,79	0,96	-0,35	-0,38	-0,56	-0,52	0,71	0,77	1,00

If the first assumption is tested the following correlation percentages can be found (minimum of 70% correlation was chosen):

**Cluster 1:** Capric & Caprilic - Correlation 99%

**Cluster 2:** Miristic & Lauric - Correlation 83%

**Cluster 3:** IodineLn & Palmitic & Linoleic - all score below 70% correlation

### **Final Conclusion:**

Multivariate Data analysis by using *STATISTICA* 6 Principal Component Analysis & Classification analysis can be used to determine hidden phenomena in the data by a using case score plots and variable score plots. Besides detecting hidden phenomena and classifying it, it is also of great use in determining possibilities of variable reduction, leading to huge cost savings for companies who "live" from data.

With some basic guidance and education *STATISTICA* should be an ideal package for universities to use when analyzing data coming from experiments, projects, or research. Also ideal is the option of screen catching, which is helpful in making presentations or creating pictures in reports. As the described example above is only a small part of the wide range this package can be used for, *STATISTICA* should be interesting for mathematical sciences as social sciences as well.

One of the things lacking in *STATISTICA*'s visualization tools is the option for a 3D variable and case factor scatterplot. Also an option of rotation within the 3 dimensional scatterplot would be a powerful option for students to better understand what clustering means.

A very interesting package resembling a good book. The longer you read, the more you want to know...