

modeFRONTIER: as a statistical tool

modeFRONTIER: como herramienta estadística

En este artículo se presentan las herramientas de análisis estadístico y de análisis multivariado disponibles dentro de la versión 4 de modeFRONTIER. El objetivo es el de demostrar como tales herramientas puedan utilizarse para capturar y analizar las más importantes informaciones contenidas en bases de datos de todo tipo. Sólo por medio de estas tecnologías es posible descubrir y reconocer las relaciones escondidas o no inmediatamente visibles entre las variables y los parámetros de un problema.

La importancia y la utilidad de estas herramientas se presentan aquí utilizando un simple ejemplo de tipo médico-estadístico. A partir de una base de datos con informaciones médicas de distintos pacientes, como por ejemplo sexo, edad, peso, tensión sistólica y nivel de colesterol, es posible determinar cuáles son los factores que estadísticamente condicionan este nivel y además reconocer posibles riesgos para otros pacientes.

Nowadays, the fact that the use of statistical software can improve processes or drive and speed up the development of new products is well-known. The aim of this article is to show that modeFRONTIER can be considered as a complete and comprehensive software for data analysis able to perform a statistical evaluation of a database.

The Design Space in modeFRONTIER version 4 can be considered as a stand-alone environment, where the user is able to perform extensive and complete statistical analyses of data derived from different contexts, making it a compelling tool for the decision-making process.

The variety of tools, reports and charts that can be used to explore data and perform complex statistical or engineering analyses are:

- Visualize data using several charts such as scatter plots or bubble charts;
- Monitor trends and time series by means of the history plots with its moving average and Bollinger bounds;
- Visualize data distributions by using histograms, probability and cumulative distribution plots;
- Find out important linear relations between variables using the correlation chart and the scatter matrix that summarize all these effects in a single chart;
- Find series outliers by means of useful charts such as Box-Whiskers or Quantile-Quantile (Q-Q) plot;
- Verify whether or not a series of data corresponds to a given distribution using the distribution fitting, the histogram plot and the Q-Q plot;
- Check the effects of the parameters on the outcomes using useful statistical tools such as the DOE main effect or the interaction effects;
- Perform several statistical tests (e.g. t-Student analysis, ANOVA).

In the post-processing panel, a complete new environment

named Multivariate Analysis (MVA) includes tools to:

- Organize designs into groups according to a given rule and look for clusters of data (hierarchical and partitive clustering);
- Build Self Organizing Maps (SOMs) in order to have an easy-to-read bi-dimensional representation of complex multi-dimensional data.

Moreover the user can also take advantage of the Response Surface Methodology (RSM), which allows the construction of meta-models of data and eventually to perform virtual optimizations.

It is very difficult, perhaps nearly impossible, to effectively analyze and summarize a huge amount of data without the help of a good statistical analysis tool. The following example demonstrates that modeFRONTIER can be considered as a complete tool for making statistical analysis of complex multi-dimensional data.

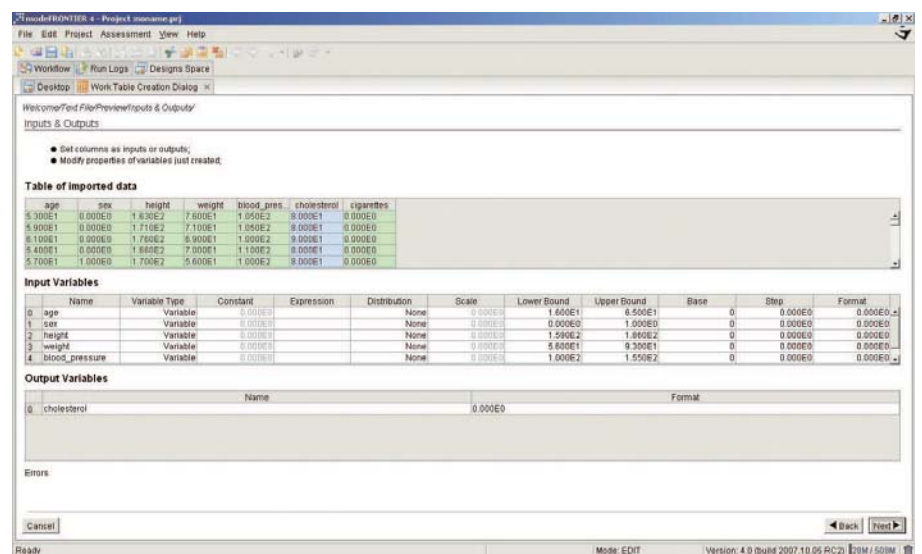


Figure 1: One step of the Data Wizard, the tool for importing data into modeFRONTIER

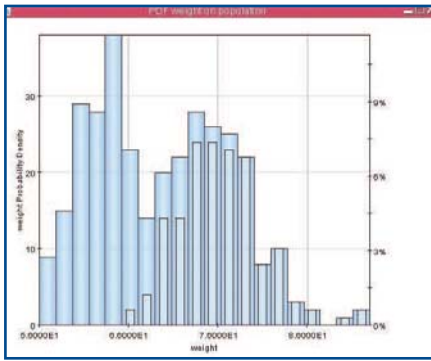


Figure 2: This histogram chart shows that the population weight has a distribution with two distinct peaks. This is due to the overlapping of two distinct Gaussian distributions for men and women.

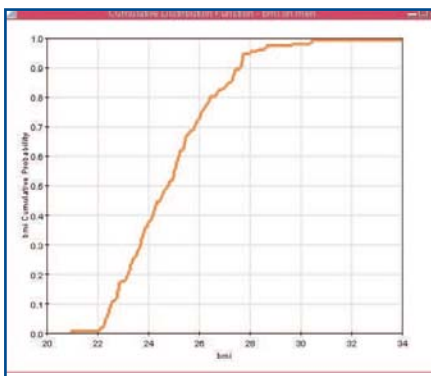


Figure 3: modeFRONTIER chart showing the ECDF of the population BMI.

Evaluating risks of LDL cholesterol

Suppose that we have collected the sex, age, height, weight, number of cigarettes smoked per day and the systolic blood pressure of a certain group of people and, finally, their level of LDL cholesterol. This data could be the result of a medical investigation which, could also consider many other aspects of patient health. All the selected quantities aim at monitoring some of the main risk factors of cardiovascular diseases. The data contained has been generated artificially and for demonstration purposes only, taking into account gerent heath information.

Load data and manage work tables

We can suppose that this kind of data is usually well-organized in a file, where columns collect the age (in years), sex (0 for men, 1 for women), height (in cm), the weight (in kg), blood pressure (in mm hg), LDL cholesterol level (in mg/dl) and the num-

ber of cigarettes smoked per day. By means of the Data Import Wizard (Figure 1), it is very easy to load the data into modeFRONTIER. During the import phase, the user can remove rows and columns containing useless data, specify the role of each column, insert objectives and constraints if any and set up the visualization format for numbers.

In this example, the variable cholesterol is set as output while all the others are set as inputs. Moreover, thanks to the work table capabilities it is also possible to insert additional columns containing derived data. In this example, we introduce the ratio between the weight (in kg) and the squared height (in m) as the body mass index (BMI) that is often used to identify if a person is of normal weight or not. With the Find tool, it is easy to select designs which satisfy certain conditions (e.g. age less than a given value and weight greater than a specified limit) and thus subdivide designs into categories or create new tables of data. For example, the BMI values usually considered normal lie between 18.5 and 25.0, hence one can easily determine if a patient's BMI exceeds or falls below this range.

Histograms

Once the data has been loaded, it is straightforward to build the histogram charts by defining the number of classes. The probability density functions which better fit the data are highlighted, and the user can visualize them superimposed on top of the histogram. It can be seen in Figure 2, that the population weight has a distribution with two distinct peaks. If one marks the designs corresponding to the men, the designs are consequently highlighted in the histogram. Obviously, the same operation can be performed by dividing data according to other criteria and consequently constructing subgroups of data which can be analyzed separately.

Cumulative Distribution

The cumulative distribution plot indicates the probability that a given event arises in the population. It reports the experimental cumulative distribution (or ECDF) together with the most probable theoretical CDF, if any.

Figure 3 shows that there is roughly a 50% probability to find a man who is overweight, corresponding to a BMI of 25. Only a small portion (less than the 5%) of the male population is obese, having a BMI greater than 30.

Box-Whiskers

The Box-Whiskers plot summarizes certain information about the data, such as the average and its confidence interval, the quartiles and the outliers. The confidence limit is an estimate of the average with lower and upper limits, increasing the uncertainty in our estimate of the true average. The narrower the interval, the more precise the estimate. Confidence limits are expressed in terms of a confidence coefficient, with 95% being the most commonly used. Outliers the designs that fall outside an interval centered in the average and with semi-amplitude of 1.5 times the standard deviation. In Figure 4, it is seen that, if we consider the blood pressure, there are four outliers which can be easily selected and eventually categorized.

It is possible to build a Category Box-Whiskers which plots the data taking into account a subdivision of the designs into risk categories. In Figure 5, the population age is considered. It can be seen that, statistically, the most risky age for cardiovascular diseases ranges between 29.5 and 46 years (the first and the third quartiles of the medium and the high risk distributions have been considered to define this range). However, if we also examine the low risk series, the age range should be enlarged up to 58. Moreover, it can be seen that the densest half is located in the highest

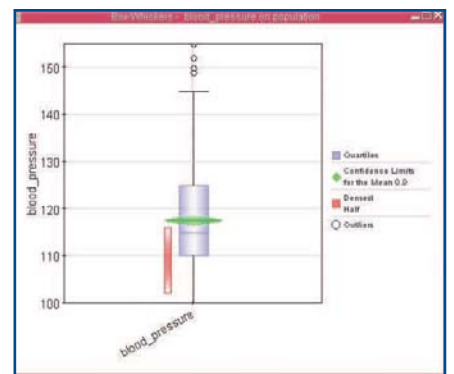


Figure 4: modeFRONTIER showing a box-whiskers chart of the blood pressure.



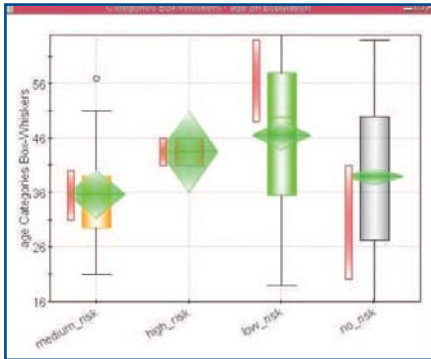


Figure 5: Box-whiskers showing four different categories of risk.

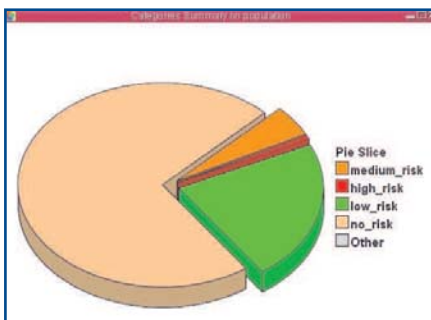


Figure 6: A pie plot providing a global view on how the population is organized. In this case a large majority of the subjects do not present any risk (according to our own subdivision), while roughly the 5% has a medium/high risk.

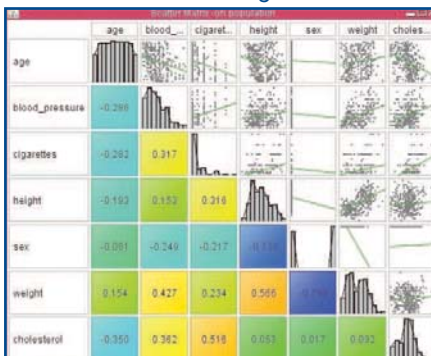


Figure 7: A scatter matrix summarizes linear relations between variables

part. This means that the risk, even low, is statistically higher for increasing ages; this statement is corroborated by the fact that the densest half of the no risk distribution is located in the lower part.

Relations between variables

The Correlation matrix and the Scatter matrix (Figure 7) are useful tools to check linear relationships between variables. The correlation coefficient is a measure of the closeness of the linear relationship between two variables. The correlation coefficient is a dimensionless number from -1.00 (representing a perfect negative cor-

relation) to +1.00 (represent a perfect positive correlation). Positive values of the correlation indicate a tendency of the two variables to increase together. When the coefficient is negative, large values of the first variable are associated with small values of the second one.

Risk level	bmi		cigarettes		cholesterol		blood pressure
High	>30	and	>0	and	>130	and	>130
Medium	>22	and	>0	and	>100	and	>130
Low	>25			and	>100	or	>130
None	Otherwise						

The correlation ranking reports the most relevant connections, but only relatively high values of correlation coefficients should be considered as reliable. In the Scatter matrix, the scatter plots together with the regression lines help to understand how the designs are distributed in the space. All the graphs can be enlarged and explored by just clicking on their left-high corner.

In Figure 7, it can be seen that the most important negative correlations involve the sex and the weight (-0.799) and the sex and the height (-0.728). The two most important positive relations involve the height and the weight (0.566) and the cholesterol and the number of cigarettes (0.516).

DOE Main Effects and Student Analysis

In constructing a DOE main effect graph, it is first necessary to identify effect and factors. The factor domains are split into two equal intervals containing the lowest and the highest values, identified with a - and a + respectively. In this way, two separate distributions are created and then plotted with respect to the chosen effect, using a layout very similar to the Box-Whiskers one. The resulting graph identifies the factors that influence the effect showing any direct or inverse relation between factors and the effect. Figure 8 shows that the consumption of tobacco and the blood pressure have a direct effect on the cholesterol level, while the age and the height have an inverse effect.

By performing a t-Student analyses of data on the output variable, it is possi-

ble to understand which are the most important factors for the cholesterol levels. In Figure 9, it is quite clear that the high consumption of tobacco and high blood pressure are the two main factors which contribute to high levels of cholesterol.

Finally, it is interesting to point out that it is possible to generate a statistical report for every variable; this report represents a descriptive statistics tool that contains all the most important univariate statistics and graphs that completely characterize the data series. The report can be saved in different formats, to allow the user to collect results and reuse them in a subsequent context.

Multivariate Analysis (MVA)

In this example, the number of variables does not allow a compact visualization of designs. Actually, if the dimension is higher than 4 or 5 it becomes prohibitive, and somewhat useless, to plot all the information contained in the database using classical 2-dimensional charts. Obviously, this represents an important limit to the users understanding of the data. For this reason, it is often extremely difficult group of similar designs, identify outliers, and understand the design space. One strategy to solve this problem is to use a Self Organizing Map

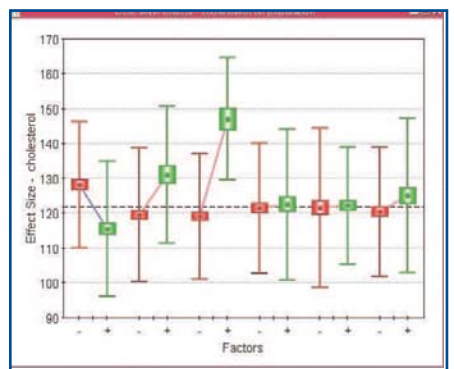


Figure 8: DOE Main Effects of the cholesterol level. This chart reveals that the consumption of tobacco and the blood pressure have a direct effect on the cholesterol level.

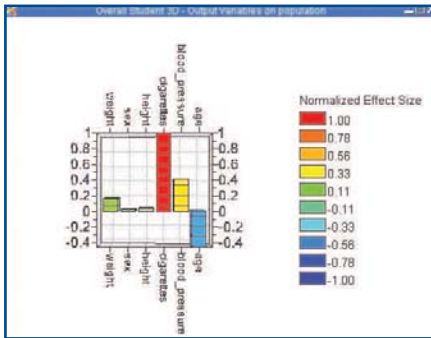


Figure 9: *t*-Student analyses on the output variables identifying the most important factors.

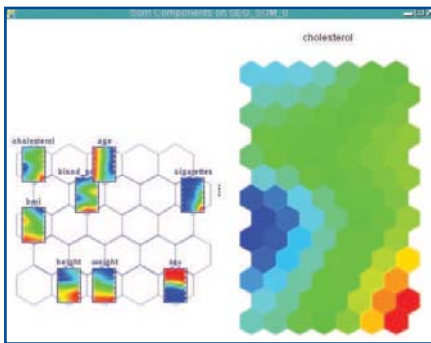


Figure 10: SOM components plot. This tool allows a global view of the database.

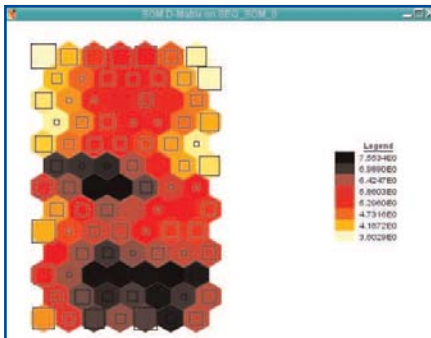


Figure 11: D-matrix of the SOM expressing the average distance between neurons.

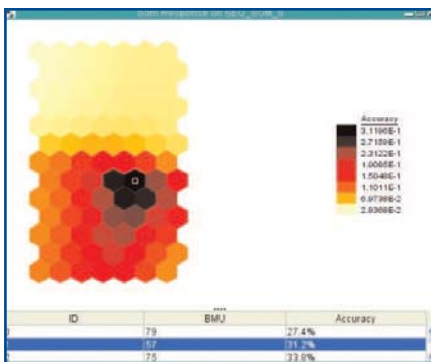


Figure 12: SOM response, this chart reports the best matching unit for each point. In this example, it allows the estimation of the cholesterol level of new patients.

(SOM) which is based on an unsupervised and competitive learning of a neural network. A SOM is able to map

the designs, belonging to a multi-dimensional space, onto a lower dimensional space, preserving the original data topology and density.

SOMs are part of the new multivariate analysis environment. Following a step by step a user-friendly wizard, the creation of a SOM is really easy. In this case, all variables were considered to build the SOM except for the cholesterol and the BMI (this was derived).

When a self organizing map is created, a new table is added to the modeFRONTIER project. Different graphical ways are available to visualize results: the first one is the SOM components plot (Figure 10), where all the database components are displayed. This allows a global view of the database and detects any relation between variables.

In this example, it is important to note that the cholesterol variable has a relatively smooth colored map. This variable was neglected during the map creation, so this indicates that its behavior is related in some sense with the other components. The relationship can be observed by seeing that the maximum values of the cholesterol are located in the lower-right corner of the map, so are the cigarettes, the weight, the blood pressure and the height components. The age and the sex have similar maps, simply rotated, with well separated red and blue zones: this can be seen as a demonstration that the examined population represents females and males of all ages.

Other charts are available along with the SOM components. For example, the D-matrix expresses the average distance between a neuron and its neighborhoods making it possible to detect if there are clusters of data and to judge if they are well separated or not. In Figure 11, it seems that there is no significant clustered distribution of data, and the designs are uniformly spread on the map (the dimension of the square is proportional to the number of designs pertaining to a given neuron) especially in the brightest zones of the map, where the distances between designs are minimal.

Now suppose that we have done another medical investigation on a second population, relatively homogeneous to the first one, collecting the same data of the first investigation, except for the cholesterol level. This time, the goal is to understand how this second population is distributed and therefore, to predict the cholesterol level for all these patients. In this way, it is possible to identify situations, such as high values of cholesterol, without having any experimental evidence. To do this, the user can load a new table of data in modeFRONTIER and plot a SOM response (Figure 12). This plot shows, for each new patient in the new table, the Best Matching Unit (BMU) of the SOM and affinity (a can be read as an accuracy value) between the value and the corresponding BMU. The BMU represents a kind of reference situation for the design under consideration, and here, the unknown cholesterol for the patient can be taken from the corresponding BMU.

Conclusions

In this article, the statistical and multivariate analysis tools available in modeFRONTIER have been briefly presented. The aim was to demonstrate how these tools can be used to capture the most important information contained in a database and discover hidden or not immediately evident relations between variables. The importance and usefulness of these tools were shown via an example where the Self Organizing Map was used for two different purposes; firstly as an effective representation of multidimensional data and, secondly, as a prediction tool.

Article written by Silvia Poles (ESTECO) and Massimiliano Margonari (EnginSoft), and edited by Gino Duffett (AperioTec).

APERIO
TECNOLOGIA EN INGENIERIA

Olivella 8
08870 Sitges (Barcelona)
tel 938 945 092
fax 938 113 957

email info@aperiotec.es
web www.aperiotec.es