Development of the KY-methods for use on toxicity prediction

○ Kohtaro Yuta　In Silico Data, Ltd. (Japan)　(http://www.insilicodata.com)

◆ **Research object : Developing new powerful data analysis methods that are specialized in toxicity evaluation**

◆ **Main difficulty to execute toxicity evaluation**

1. Structural diversity of compounds is quite large
2. Number of samples used in the analysis is very large
3. Quite complex of the toxicity expression mechanism
4. High classification and prediction value is required

Data analysis techniques that is normally used are shortage in toxicity evaluation

Data analysis technique that is normally used is a shortage in toxicity evaluation

Suitable for toxicity prediction

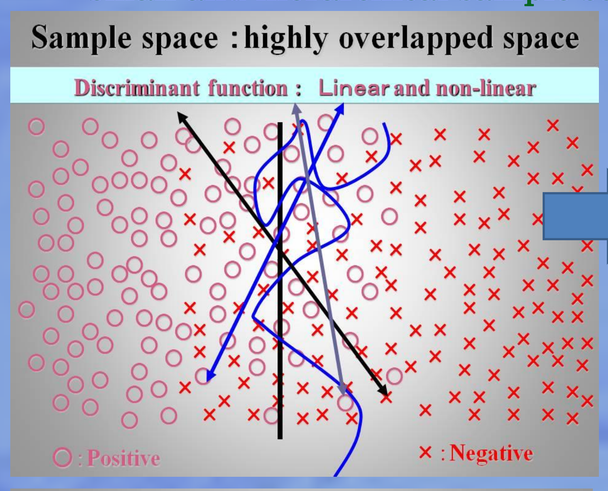
New data analysis methods **(KY-methods)** have been developed in order to solve toxicity evaluation problems

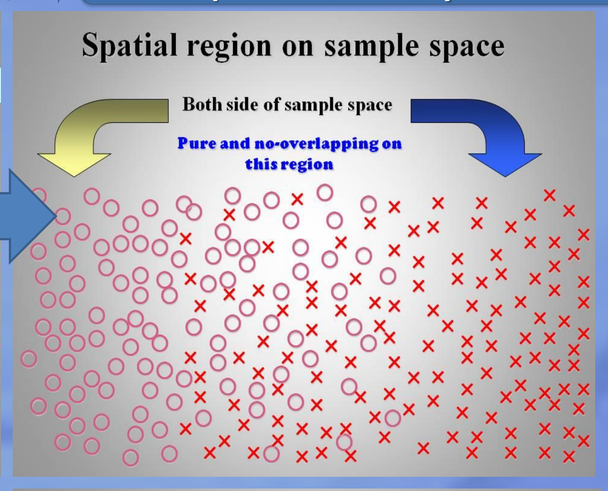
Outstanding features of the KY(K step Yard sampling)-methods

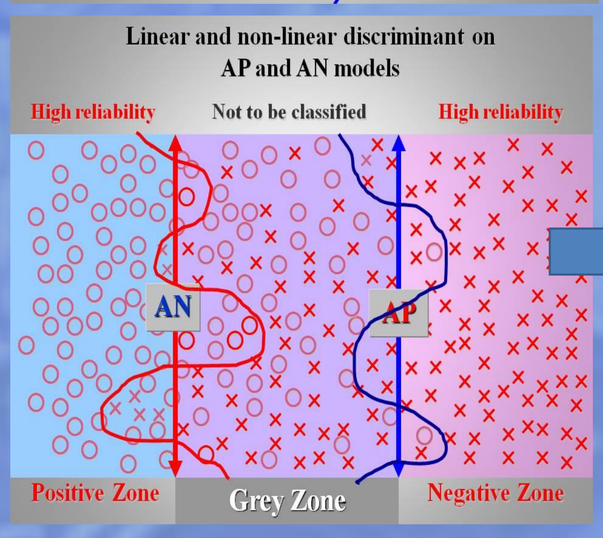
◆ Binary classification ◆

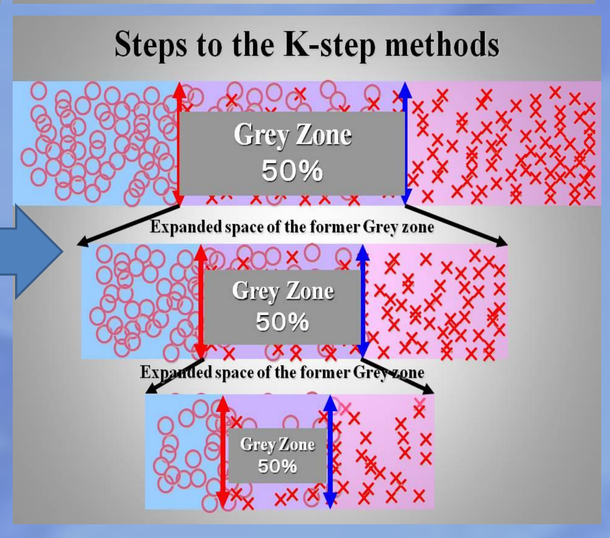
1. Constantly achieve perfect (100%) classification under any conditions
   - Highly overlapped class sample data set
   - Quite large number of sample data set (tens and several thousands of)
2. Starting sample set was divided into
   - small and clean sample set
   - small and hierarchical sample set

⇨ Repeat these operation, until all samples are correctly classified









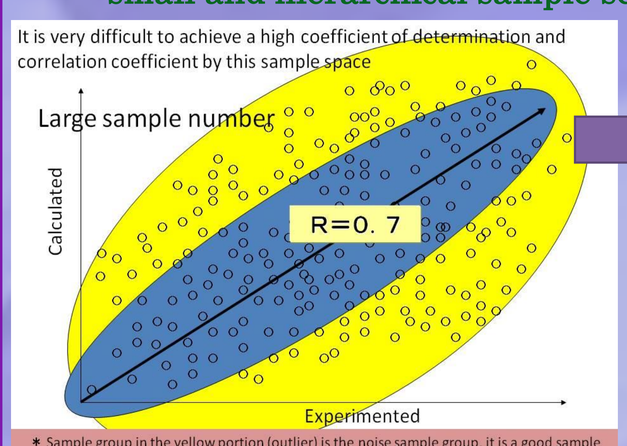
◆ **Variation of the "KY-methods on binary classifier"**

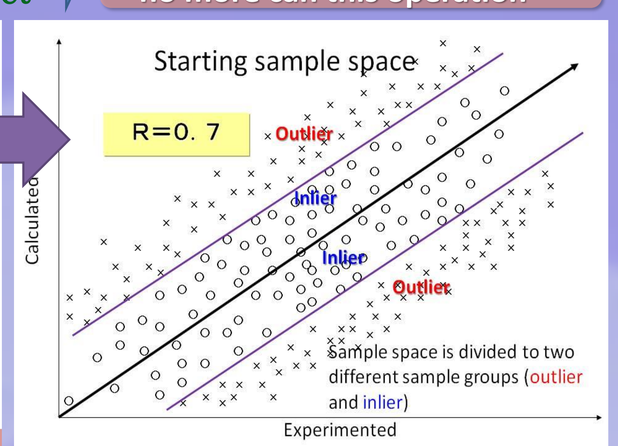
Binary classification ; 3 approaches
1. Two model KY– discriminant method
2. One model KY– discriminant method
3. Model free KY– discriminant method

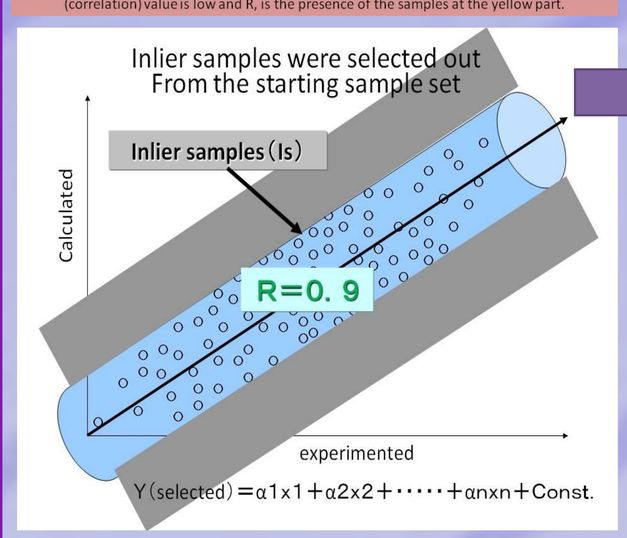
◆ Fitting : Regression analysis ◆

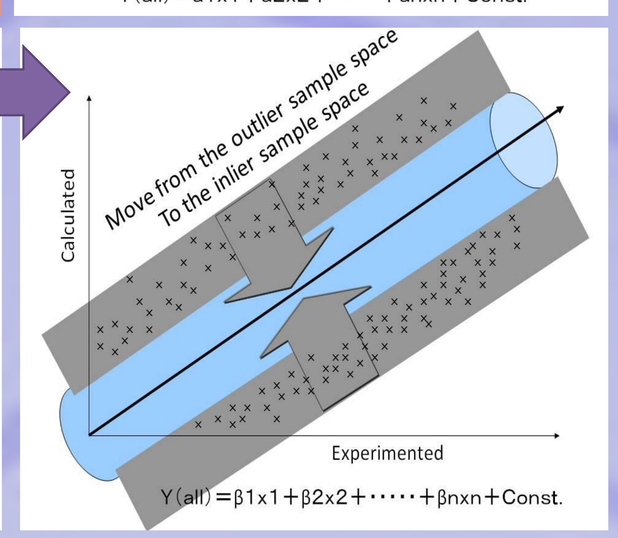
1. Constantly achieve high coefficient of correlation and high decision coefficient under any conditions
   - Widely distributed sample data space
   - Quite large number of sample data set (tens and several thousands of)
2. Starting sample set was divided into
   - 'inlier' and 'outlier' sample set
   - small and hierarchical sample set

⇨ Repeat these calculation, until no more can this operation









◆ **Variation of the "KY-methods on regression methods"**

Fitting (multi regression); 3 approaches
1. KY–fitting with discriminant method
2. Three zone KY–fitting method
3. Model free KY–fitting method

List of application examples of the KY method and conclusions

Example 1 (Binary data) : Ames test sample data set

About 7000 Ames test sample dataset ⇒ **Perfect classification**

＊ Usual multi-variate and pattern recognition methods can't

Example 2 (Binary data) : Skin sensitization sample data set

About 600 sample dataset ⇒ **Perfect classification**

＊ Usual multi-variate and pattern recognition methods can't

＊ Poster P05-21 Sato et.al, Euro Tox 2013

Example 3 (Continuous data): Fish toxicity sample data set

About 800 Fish toxicity sample dataset ⇒ High decision coefficient

＊ Usual multi-variate and pattern recognition methods can't

◆ **Results and conclusions:**

Very high correlation coefficient and perfect classification (100%) were realized by newly developed KY-methods.

On the toxicity research field, it is almost impossible and hard to achieve excellent and stable evaluation results by using the conventional data analysis methods.

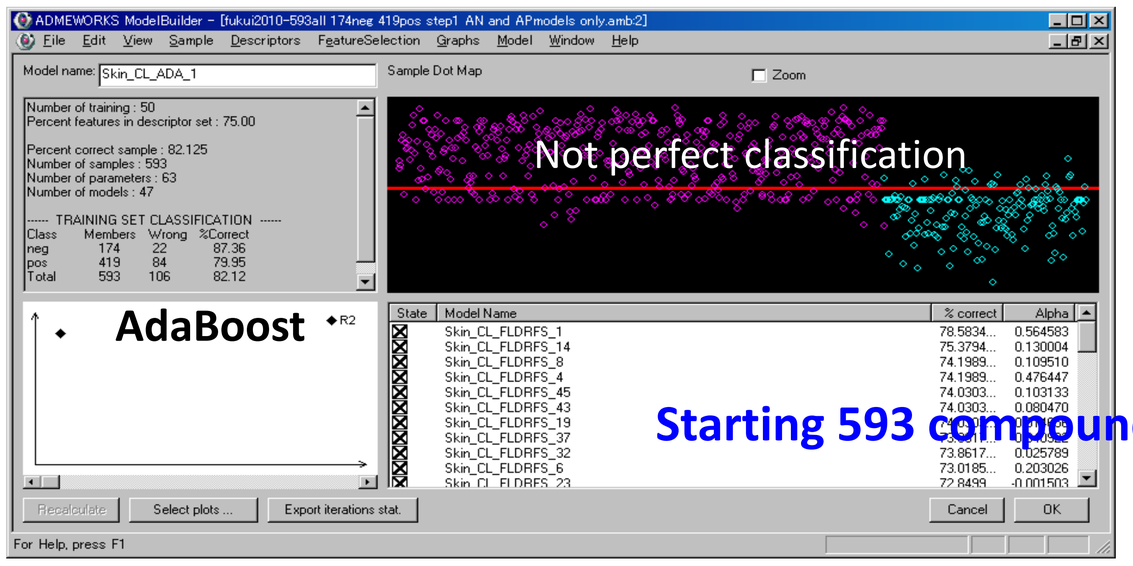
* All six KY-methods are patented and pending applications in JP, USA and EU and some were Korea and China.

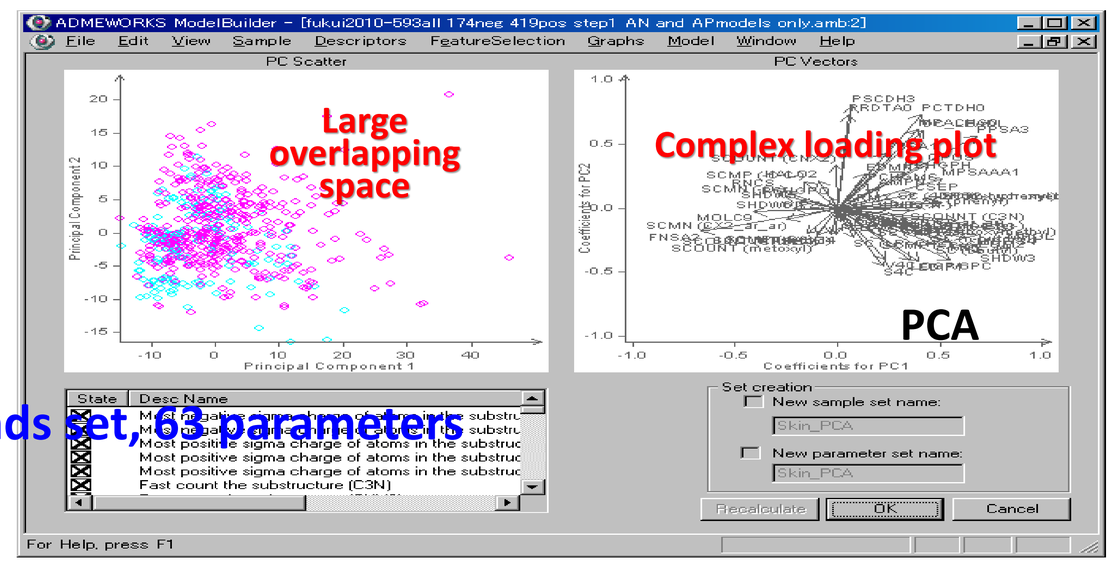
In Silico Data
Miracles by the KY-methods

Big toxicity data analysis by the combination of the PCA method and the KY-methods

○ Kohtaro Yuta　　In Silico Data, Ltd.　(http://www.insilicodata.com)

◆ Apply multi-step and re-sampling technologies by the **KY-methods** to the **PCA** for handling big toxicity data◆





When a large number of sample data were applied to,
1. binary classification method → perfect classification is not feasible.
2. PCA method → clear sample plot and simple loading plot are difficult to generate.

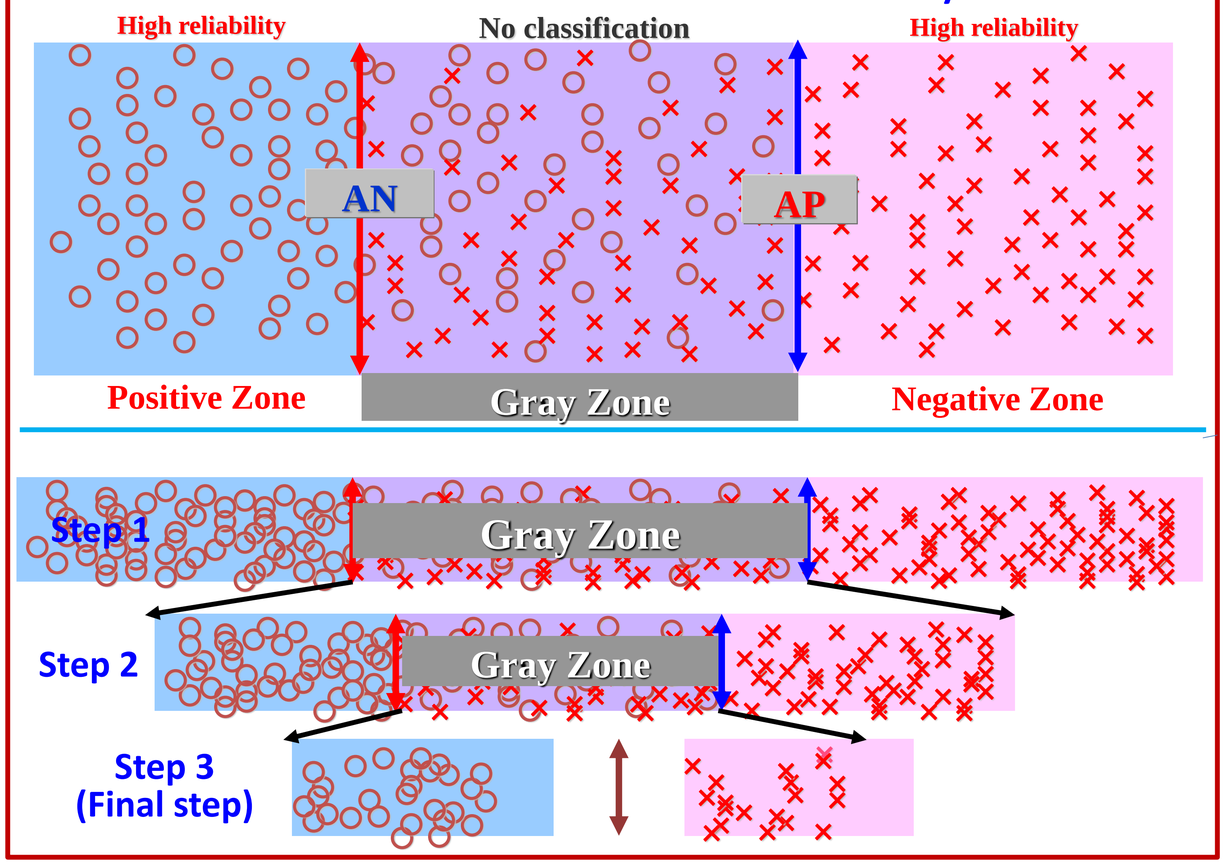
When used in combination of the KY-method and PCA, this was generate great result.
1. KY-method → perfect classification is achieved.
2. PCA method → clear sample plot and simple loading plot are generated.

☐What is the KY-methods

◆ Features of the KY (K-step Yard sampling) methods

1. Always achieve perfect (100%) classification under any conditions
   - Highly overlapped class sample data set
   - Quite large number of sample set (tens and several thousands of)
2. Starting sample set was divided into
   - small and clean sample set
   - small and hierarchical sample set
3. Applicable not only the discriminant but multi-regression analysis

" Two model KY-method for Discriminant analysis "



◆ Used samples : Skin sensitization data
Total ; 593, Positive 419, Negative; 174

◆ Classification result by **the KY-method** (100% correct)

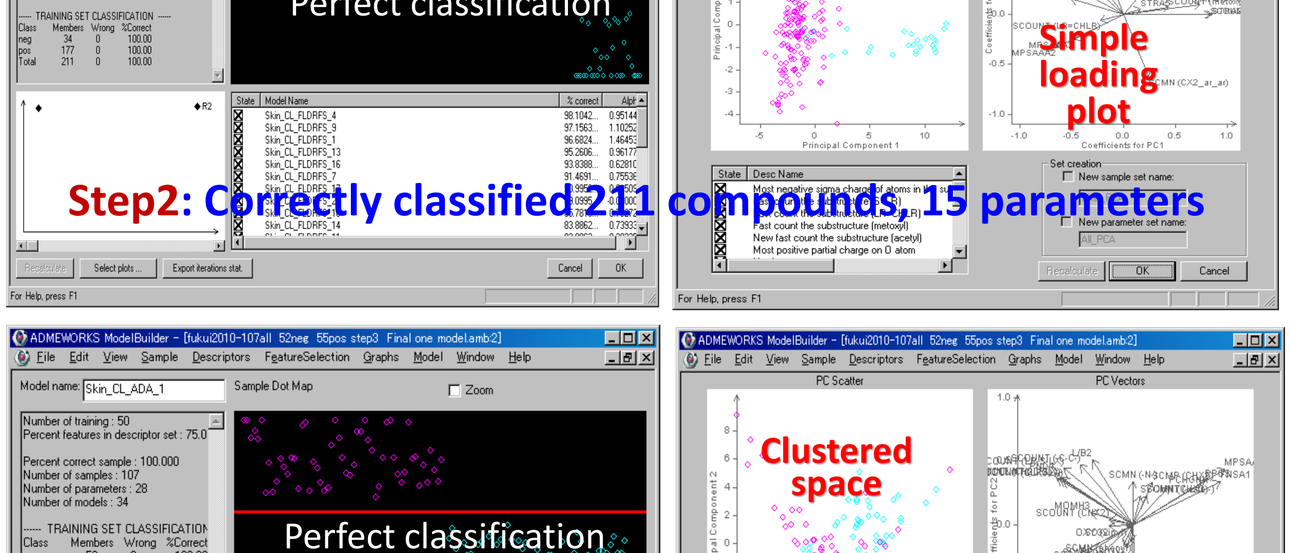
Step1; Positive 187　Negative 88　Grey zone 318
Step2; Positive 177　Negative 34　Grey zone 107
Step3; Positive 55　Negative 52　Grey zone 0

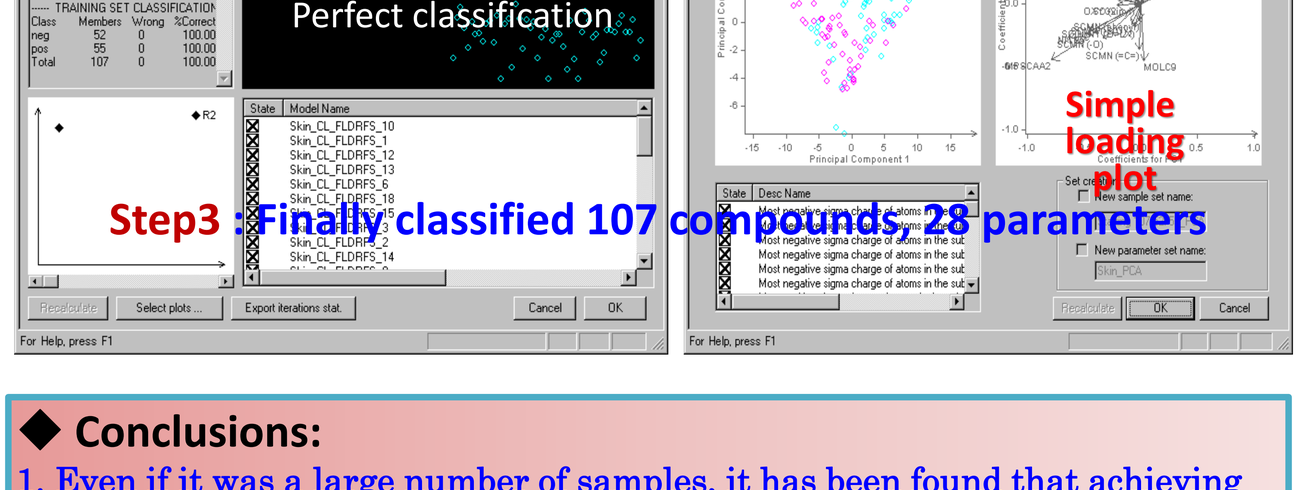
Perfect classification(100%)

◆ Classification Results by various methods (63params)

Methods	Total	Positive	Negative
N.N.	86.0%	86.2%	85.6%
SVM	91.7%	98.3%	75.9%
LDA	87.0%	95.2%	67.2%
KNN (K=5)	77.7%	86.9%	55.8%
AdaBoost	82.1%	80.0%	87.4%

◆ Classification results of compounds which are included in the "High reliability" space of the "KY-methods"





◆ Conclusions:

1. Even if it was a large number of samples, it has been found that achieving an excellent data analysis results by using a combination of the KY-method and PCA.
2. Perfect (100%) classification was achieved by the KY-method
3. As a result of applying PCA for resampled sample set by the KY-method, it was possible to obtain a clean clustered sample space and much more simple and clear loading plot.

All research works and screen displays were executed and generated by the ADMEWORKS : ModelBuilder program developed by FJQS (Fujitsu Kyushu Systems Ltd.)