



Make Sense of Your Data™

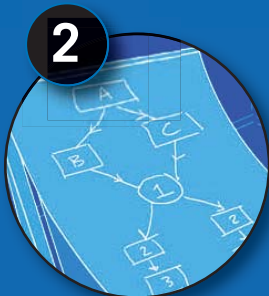
Just Three Simple Steps:

1



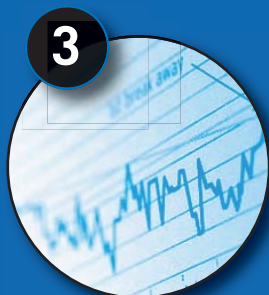
Provide Your Data:
spreadsheet or file

2



Specify Your Objective:
overview, relationship,
classification

3



Name Your Project:
result is an active report

CASE STUDY: Sweet Success

Sounds obvious — consumers expect white sugar to be white, but this isn't always the case if the sugar making process is not kept in control. Ash, the inorganic matter in sugar, is measured as the residue left after burning a sugar sample of specified size. Too much ash can grey the sugar. To keep sugar white, the amount of ash in the sugar must be kept low. To ensure this, a white sugar production plant used a traditional, but costly, wet chemistry test to determine two important qualities of the final product, ash content and color.

MULTIVARIATE ANALYSIS

Multivariate Analysis (MVA) is a data-analysis technique based on using all the measurements or variables in the dataset together. This is in contrast to classical analysis which looks at only one or two variables at a time. The basis of MVA is that the information from the dataset lies in how the measurements relate to each other. MVA was used to examine a new product test method, rapid fluorescence measurement, and to compare its effectiveness to the traditional wet chemistry test method.

OBJECTIVE, ANALYSIS, AND DATA

The objective of this study was to determine whether spectral analysis could be used in place of a wet chemistry test technique to determine the quality of white sugar. In this case study, the data-analytic approach was Partial Least Squares, or PLS. With this you look at the quantitative relationships in the data—in this case between the rapid fluorescence measurements and the color and amount of ash in the samples.

The spectral analysis was made on 106 samples of sugar with known color and levels of ash. A prediction set with 6 samples of sugar with known color and levels of ash was used to validate the model.

Fluorescence studies were made on 106 samples using the excitation wavelength 240 nm and emission wavelengths 275-560 nm followed by the determination of ash content and color.

WHAT IS THE NOISE LEVEL?

Noise level in the dataset is a measure of how well the data are representative of typical production. In this case, we have a very controlled process with accurate data collection. With these conditions, if the noise level is more than 20%, the prediction error for future sugar samples (deviation between predicted and actual Y values) will be substantial. If the noise level is too high, the dataset should not be used. The noise level in these data was 7% for color and 4% for ash content, indicating a very good sample.

THE SCORE PLOT

In this case study, we started with a score plot. The X-part summary comprises 6 scores. These scores are new variables which summarize the data. They are combinations of rapid fluorescence measurements and are similar to the Dow Jones or NASDAQ indices which summarize a table of time points (rows) by stock prices (columns). The scores are weighted averages of the original variables, hence providing a good summary of all the rapid fluorescence measurements, with the most informative scores being the two first. The weights (also called loadings) provide information about the relative importance of the individual rapid fluorescence measurement and about their correlations.

When you plot the two first scores against each other you get a picture, or map, of the dataset where the observations (the sugar samples) are seen as points in the plot. This allows you to see which observations are similar (near each other) and which are dissimilar (far away from each other). You can also drill down to see why observations are disparate and display a contribution plot showing which rapid fluorescence measurements caused the differences.

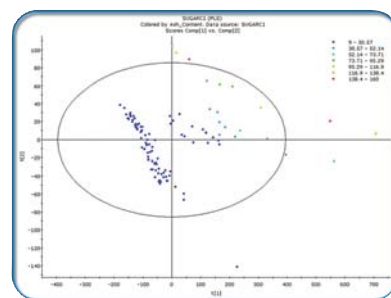


Figure 1—Summary of X-part of Sugar Samples (Score Plot)

In this case we see a nice grouping of data points on the left side of the plot and a number of outliers—data points spread out in a wide and random pattern. The groups correspond to normal operation (to the left) and abnormal to the right (process start up). A feature of EZinfo allows us to remove the outliers and recreate the scores plot with the data points to the left that formed the tight group (see Figure 2).

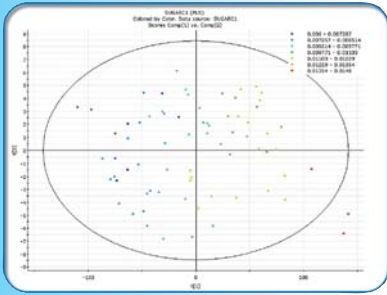


Figure 2—Summary of X-part of Sugar Samples (Score Plot) After Removing Outliers

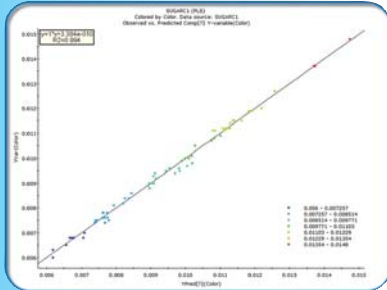


Figure 3—Observed vs. Predicted of Training Set, Y = color

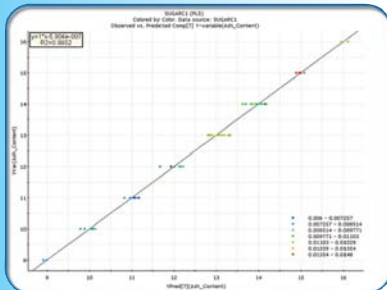


Figure 4—Observed vs. Predicted of Training Set, Y = ash content

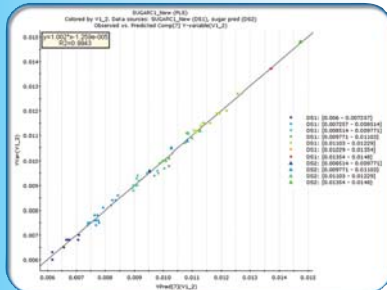


Figure 5—Observed vs. Predicted of Training Set and Prediction Set

The scores $t[1]$ and $t[2]$ are the two most important indices in relating rapid fluorescence measurement to the color and ash content of the sugar. The plot of $t[1]$ vs. $t[2]$ gives a picture of the X-part of the data. The plot (colored by ash content) shows the possible presence of atypical sugar samples, as well as groupings, similarities, trends, and other patterns in the data. Atypical sugar samples lie outside the ellipse.

The plots of the X-scores versus the Y-scores. i.e., $(t[1], u[1])$, $(t[2], u[2])$, etc., show the relationships between X and Y in each layer, each component. The first, $(t[1], u[1])$ is usually the most important (see Figure 2).

OBSERVED VS PREDICTED Y-VALUES

The Observed vs. Predicted plot displays the predicted (horizontal axis) and actual values (vertical axis) for the specified response. It is a measure of the goodness of the model.

Figures 3 and 4, the Observed vs. Predicted plots for color and ash content respectively, indicate that the data used for this case is “good” by the tight grouping of points along the 45 degree line.

To further validate the ability effectiveness of this test method, a separate prediction set of data is imported into the model and the color and ash content are predicted and displayed in an Observed vs. Predicted plot (Figure 5). Once again, a strong correlation between the observed and predicted values is illustrated, confirming that the rapid fluorescence measurement data can be used to accurately predict color and ash content.

RESULTS

EZInfo’s Observed vs. Predicted plot showed that a new fluorescence test (rapid fluorescence measurement) provided similar results to a more expensive wet chemistry test.

EZinfo automatically generated a model, based on the original data set. or “training set”, that accurately determined the correlation between the rapid fluorescence measurement data and the quality criteria, color and ash content. The

accuracy was confirmed by predicting the results for additional samples of known color and ash content by means of the model and comparing the predicted quality values with the actual (Table 1).

Color	Predicted Color	Ash Content	Predicted Ash Content
0.0148	0.0147375	15	14.8823
0.0096	0.00952743	13	13.2682
0.0105	0.0102733	15	15.0776
0.0112	0.011275	14	13.8989
0.0111	0.0108507	13	13.1919
0.0108	0.0108223	13	13.3147

Table 1—Prediction Dataset for Both Y’s of Prediction Samples

Calculating the predictive standard deviation for the six prediction set samples shows that color can be determined with a precision of around 6% and ash content 20%. This precision was sufficiently good to allow the rapid fluorescence measurements to be used to determine the color and ash content within the desired specifications.

This allowed the manufacturer to confidently change their quality control method to a lower cost, less complicated and more accurate process. Additionally, the fluorescence test required only a few minutes, while the wet chemistry test took several hours. This resulted in the ability to immediately correct the process if unsatisfactory quality was determined by the spectroscopy, which gave substantially increased quality and reduced costs.



About Umetrics . . .

Umetrics develops software for design of experiments and multivariate data analysis, for the individual user as well as for on-line continuous and batch processes. We provide training at more than 25 worldwide locations and on-site consulting services. We are committed to supporting our clients in their mission to control data flow by conveying our advanced expertise in multivariate technology.

Umetrics is now owned by MKS Instruments Inc., with the acquisition finalized in January, 2006. Our general manager is Nouna Kettaneh-Wold. Umetrics has offices located in Sweden, United Kingdom and USA, and employs just over 50 people.



www.ezinfo.net

www.umetrics.com

Headquarters:

Umetrics Inc.
17 Kiel Ave.
Kinnelon NJ 07405
USA
Phone: +1 973 492 8355
Fax: +1 973 492 8359
Email: info.us@umetrics.com

Offices:

Umetrics AB
Stortorget 21
SE-211 34 Malmö
Sweden
Phone: +46 (0)40 6642580
Fax: +46 (0)40 6642585
Email: info.se@umetrics.com

Umetrics AB
Box 7960
SE-90719 Umeå
Sweden
Phone: +46 (0)90 184800
Fax: +46 (0)90 184899
Email: info.se@umetrics.com

Umetrics UK Ltd.
Woodside House,
Windfield, Windsor
SL4 2DX, UK
Phone: +44 (0)1344 885605
Fax: +44 (0)1344 885410
Email: info.uk@umetrics.com

Umetrics USA
70 Rio Robles
San Jose, CA 95134
USA
Phone: +1 408 750 0300
Fax: +1 408 750 2990
Email: info.us@umetrics.com