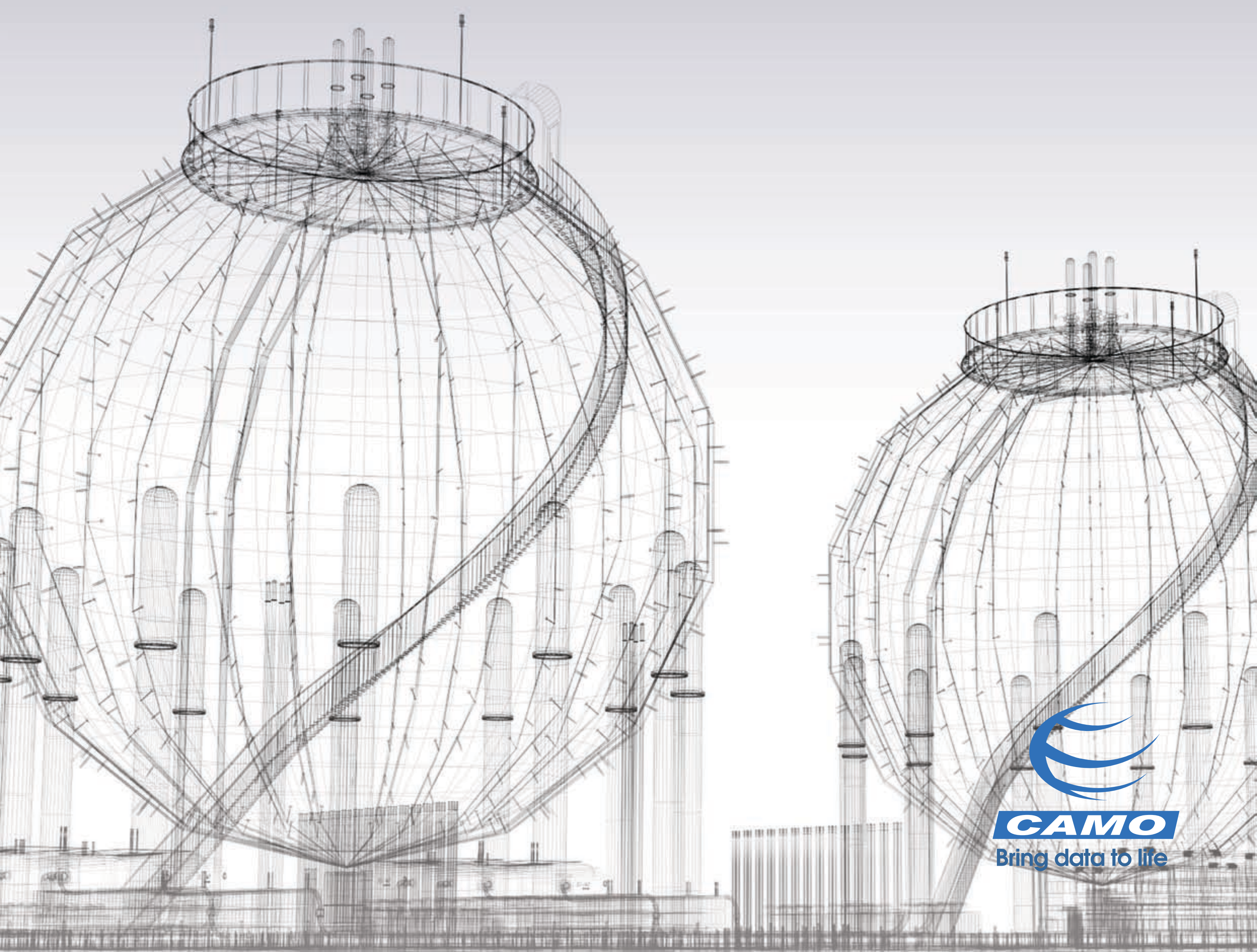
WHITE PAPER

IMPROVE THE RELIABILITY OF YOUR PROCESS MODELS FOR BETTER PROCESS UNDERSTANDING & CONTROL: **A HOLISTIC APPROACH FOR ENGINEERS**

While the First Principle Models typically used in engineering have many strengths, techniques such as Multivariate Data Analysis and Design of Experiments are increasingly being used in combination with traditional approaches to give engineers more robust tools for process modelling, product development and quality control.

By Dr. Frank Westad, Chief Scientific Officer, CAMO Software



CONTENTS

INTRODUCTION

Chemical engineering is a discipline that to a certain degree relies on deterministic or first principle models. These models may describe flow rates, mass balance, energy and other principal properties of a system. But to quote the famous chemist/statistician George Box: “All models are wrong but some are useful”.

Many of the models or formulae found in the literature have certain parameters (or constants) in them, some derived from theory and confirmed by experiments, others estimated from empirical data. One example of the former is the Venturi equation based on Bernoulli’s equation and conservation of energy. As a young student in chemistry where chemical engineering was one subject in the mandatory curriculum, these equations were at that time considered the “truth”. However as one got into the details of a particular subject it was clear that the basic equation did not tell the whole story: there were always assumptions behind those (simple) equations.

In retrospect one realizes that the constants in e.g. fluid mechanics had arisen more from an empirical origin than constants given by laws of nature. Other equations may be valid in a system given a specific operational window, but not in general.

MULTIVARIATE DATA ANALYSIS IN PROCESS MODELLING

The field of multivariate data analysis (MVA) relies more on empirical measurements from all instruments and sensors available in a system.

Typically, in the actual application one tries in the best possible way to find the relationship between raw materials, all steps in the process and the end quality as illustrated in Figure 1. Each row is an observation and each column is a variable. Depending on the column size of each block they may be weighted as a function of the number of variables to have the same possible impact in the model. Alternatively they may be weighted according to the assumed underlying dimensionality [1].

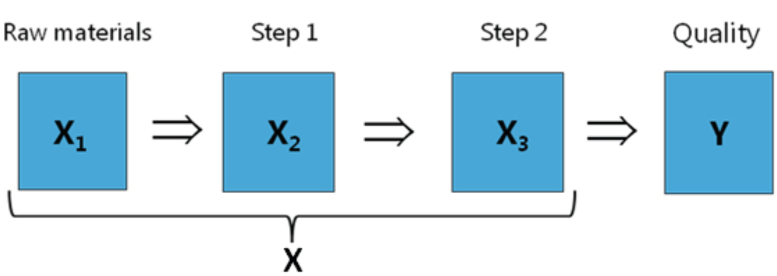
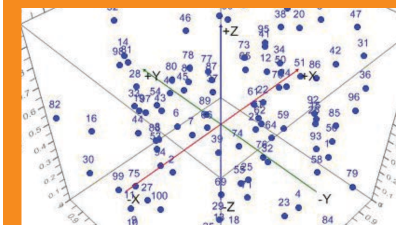


Figure 1. A general illustration of a multiblock model

The two main approaches may be described as top-down (deduction) or bottom-up (induction) and depending on the scientific community one may favor one or the other. Modelling a system from theory is rewarding if the “map matches the terrain” and hypotheses from first principles are confirmed. However, there may be disturbances or unknown systematic variations that cannot be explained by the first principle models. This can be changes in raw materials, external influences on the system etc.

One example is the assumption that the sum of the octane number for individual hydrocarbons in a fuel blend gives the total octane number. When applying e.g. LC-MS to measure the individual m/z fragments and use those intensities in the model for octane, it has been shown that there are deviations from the linear assumption.

WHAT IS MULTIVARIATE DATA ANALYSIS?



Multivariate Analysis is the investigation of many variables, simultaneously, in order to understand the relationships that may exist between them. This can be as simple as analyzing two variables right up to millions.

While traditional (univariate) statistical approaches such as mean, median, standard deviation etc serve their purposes for investigating and understanding simple systems, when the relationships between variables become more complex, a single variable cannot adequately describe the system. This is the case for most processes, and especially in chemical processes.

To find out more, download our introductory guide to Multivariate Analysis including illustrations of how it gives you a better understanding of a process compared to traditional process control tools.

What is Multivariate Analysis ?

There is, however, no reason why the two modelling philosophies cannot be combined to have "the best of both worlds". A procedure could be the following:

- Collect data for the input (raw material and all unit operations measurements) and output (quality, yield etc.) of the process (Figure 1). In the multivariate statistical terminology we may name these X and Y, respectively (if you have a background from control theory you may use other terminology).
- Model the output using the first-principle models which may include all or some of the individual variables above to give the theoretical output. We may call this $\hat{Y}_{Theory}$
- Compute the deviation from the observed output and the theoretical output, also known as the residual or error, ΔY
- Use multivariate regression in a multi-block context to relate the predictor variables X to the residual, $\Delta Y = f(X)$ Assumed non-linear relationships between individual variables X_k and y_j
- The combined prediction is then a function of the two: $\hat{Y}_{Combined} = \hat{Y}_{Theory} + \hat{\Delta Y}_{Empirical}$

Another approach is to apply an empirical model inside a state-space model to estimate the parameters that are inputs in the first principle model [2]. This ensures a more dynamic use of the empirical model.

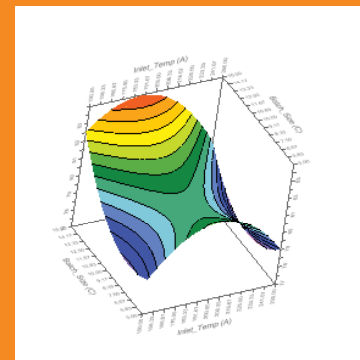
It should be mentioned that it is important to have an estimate of the uncertainty of the reference methods for the variables in Y. As the residuals from the theoretical is expected to be random noise if the model is correct one should be careful not to overfit in the second stage of the model.

Thus, the "goodness" of this model in terms of the average error must be compared to the reference method's uncertainty. The uncertainty may be due to the measurement principle itself but also because of the challenge of finding a representative sample. One example is in oil and gas production where the fiscal data as output from the separators is regarded as the truth, but it is not trivial to relate this back to per second predictions of the composition of multiphase flow from a flow meter mounted at the pipeline.

DESIGN OF EXPERIMENTS AS A CONFIRMATORY TOOL

Design of Experiments (DoE) has been recognized as a valuable tool to make planned experiments and make conclusions about critical process variables with as little effort as possible. It is only when a structure data table such as a factorial design is the basis for experiments that one can estimate one effect independently of the other. In addition, interaction and higher order effects may be estimated given the actual confounding pattern of the design. DoE also plays a role to confirm any hypothesis that may emerge from analyzing historical data from a complex process.

WHAT IS DESIGN OF EXPERIMENTS?



Design of Experiments (also known as experimental design) is a systematic approach to defining the most effective experimental plan for analyzing a system that uses the least amount of experimental effort to gain the most experimental information.

It allows the user to gather empirical knowledge i.e. knowledge based on the analysis of experimental data and not on theoretical models.

Adopting a Design of Experiments strategy can lead to cheaper and better product development by making 'optimized' decisions and reducing costly 'trial and error' approaches.

Read more about Design of Experiments here

The multivariate latent variable regression methods can handle redundancy between variables i.e. if several variables tell the same story about the underlying data structure one does not have to take some of them out due to the well-known collinearity problem in classical multiple linear regression. However, when several process variables show high correlation to the e.g. the product quality it is not evident what is the cause and effect in the process.

An analogy would be: The number of tourists and mosquitoes in Northern Norway show a high correlation but the causal variable is the temperature. To break this dependency one may use DoE to find the causal relationship. In this context it is important to distinguish between controlled and uncontrolled (observed) variables. It may also be that some of the variables are part of a mixture, thus the results are best interpreted in a mixture design response surface plot as the variables are not independent (they always sum to a certain percentage of the mixture).

METAMODELLING

However, it is not always practical to change process settings from a structured designed table in a real-scale process as the end product quality may be out-of-spec for some of the runs. Unless one can conduct experiments in a laboratory or pilot-plant scale this may be counter-productive and not go well with the plant manager. If one cannot do real-life experiments one alternative is to perform simulations of the process as described by first principles. However, the high-dimensional models in simulation software meet challenges such as increasing the speed of numerical solvers, developing tools for automated model simplification and global sensitivity analysis.

One emerging cross-disciplinary field of science which addresses these challenges is so-called 'metamodelling' [3]. Metamodels are statistical models mapping parametric variation to variation in the state variables throughout the entire relevant parameter space [4]. By this approach one can model the system "once and for all". Metamodelling will also give information of the sensitivity of the state space parameters, X, and which ones that are the most important in influencing the output Y.

A procedure for a metamodel of a first principle model process is given below:

- Perform initial simulations for a number of combinations of input variables X and Y by applying all background knowledge about the system to find the initial best parameter settings.
- Establish the relationship directly between X and Y with a multivariate latent regression method, e.g. PLS regression [5].
- From interpretation of this model space one may find blank spots that are not spanned by the existing experiments. Use a DoE strategy to find new combinations of the input variables (parameters in principle models in the simulation software).
- Run the simulation with these new set of experiments.
- Find the relationship between input and output by use of the proper data analytical methods. If the DoE is an orthogonal design the effect of the parameters can be estimated independently of each other with Analysis of Variance (ANOVA).
- Given the precision of the empirical model, new prediction of the output given new parameter settings can be performed with the empirical model, provided the new settings lie inside the model space. Confirm by running the simulation software for chosen settings

SUMMARY

Although the concepts of modelling are different depending on the tradition in various fields of science there should be no reason why theory and empirical results cannot co-exist. If the hypotheses from theory are confirmed by empirical data this is a validation of the system per se. However, seeing something new in the empirical data opens up for innovation and may lead to new theory [6].

For a confidential discussion on how Multivariate Data Analysis and Designed Experiments can help improve your process and equipment performance, please **contact our consulting team**

ABOUT CAMO SOFTWARE

Founded in 1984, CAMO Software is a recognized leader in Multivariate Data Analysis and Design of Experiments software. Our flagship product, The Unscrambler® X, is known for its ease of use, world-leading analytical tools and data visualization. More than 25,000 people in 3,000 organizations use our solutions to analyze large or complex data sets, improve process or equipment monitoring and build better predictive models. We help our clients make better decisions through deeper insights from their data, reduce R&D costs, improve production processes and product quality.

REFERENCES

[1] Lopes J.A., Menezes J.C., Westerhuis J.A., Smilde A.K. Biotechnol Bioeng. 2002 Nov 20;80(4):419-27.
[2] Psichogis, D.C., Ungar, L. https://www.cis.upenn.edu/~ungar/papers/OLD/psichogios_AICHE.pdf
[3] Kleijnen JPC: Design and Analysis of Simulation Experiments. New York, USA: Springer;, 1 2007
[4] Tøndel et al. BMC Systems Biology 2011, 5:90 http://www.biomedcentral.com/1752-0509/5/90
[5] Martens, H., Martens, M. Multivariate analysis of quality - an introduction, Wiley, Chichester (2000)
[6] http://www.experiment-resources.com/empirical-research.html

CAMO SOFTWARE
PRODUCTS & SERVICES

Our powerful yet easy to use and affordable solutions are used by engineers around the world in a wide range of industries

The Unscrambler® X
Leading multivariate analysis software used by thousands of data analysts around the world every day. Includes powerful regression, classification and exploratory data analysis tools.
TRIAL VERSION | READ MORE

Unscrambler® X Prediction Engine & Classification Engine
Software integrated directly into analytical or scientific instruments for real-time predictions and classifications directly from the instruments using multivariate models from The Unscrambler® X.
TRIAL VERSION | READ MORE

Unscrambler® X Process Pulse
Real-time process monitoring software that lets you predict, identify and correct deviations in a process before they become problems. Affordable, easy to set up and use.
TRIAL VERSION | READ MORE

Consultancy and Data Analysis Services
Do you have a lot of data and information but don't have resources in house or time to analyze it? Our consultants offer world-leading data analysis combined with hands-on industry expertise.
READ MORE | CONTACT US

Our partners
CAMO Software works with a wide range of instrument vendors and data formats used in the petrochemical refining industry. For more information please contact your regional CAMO Software office or visit
www.camo.com/partners

Training
Our experienced, professional trainers can help your team use multivariate analysis to get more value from your data. Classroom, online or tailored in-house training courses from beginner to expert levels available.
READ MORE | CONTACT US

Find out more ›

For more information please contact your regional CAMO office or email **sales@camo.com**

www.camo.com

NORWAY
Nedre Vollgate 8,
N-0158
Oslo
Tel: (+47) 223 963 00
Fax: (+47) 223 963 22

USA
One Woodbridge Center
Suite 319, Woodbridge
NJ 07095
Tel: (+1) 732 726 9200
Fax: (+1) 973 556 1229

INDIA
14 & 15, Krishna Reddy
Colony, Domlur Layout
Bangalore - 560 071
Tel: (+91) 80 4125 4242
Fax: (+91) 80 4125 4181

JAPAN
Shibuya 3-chome Square Bldg 2F
3-5-16 Shibuya Shibuya-ku
Tokyo, 150-0002
Tel: (+81) 3 6868 7669
Fax: (+81) 3 6730 9539

AUSTRALIA
PO Box 97
St Peters
NSW, 2044
Tel: (+61) 4 0888 2007

CAMO

Bring data to life