Pure & Appl. Chem., Vol. 61, No. 6, pp. 1033–1040, 1989.
Printed in Great Britain.

Data treatment—from the laboratory to industry

Sándor Kemény, Gábor Chikány, Éva Thury, Péter Láng, Dorottya Miklós

Department of Chemical Engineering, Technical University of Budapest
H-1521 Budapest, Hungary

Abstract - An useful engineering strategy is outlined connecting phase equilibrium experiments and modelling with pilot plant laboratory distillation and column calculations. The information content of experimental data on different properties are available and the role of planned experiments is discussed, together with an example of error propagation analysis. A novel method for checking thermodynamic consistency of data banks is proposed.

INTRODUCTION

As neither the available models nor the experimental data are complete, it is important to include experimental verification steps into the strategy of modelling a distillation technology. Figure 1 shows a scheme for a typical approach to a low or moderate pressure distillation separation problem.

WHAT TO MEASURE

Assuming that the models applied are almost perfect, the question arises, which measurable property is the most informative (has the largest sensitivity) on model parameters, considering also the precision and costs of the measurements. In other words, what type of experiments are to be carried out with a certain precision in order to obtain maximum information on (less uncertain values of) model parameters.
To get an idea, the effect of a parameter of a simplified model (one-parameter UNIQUAC) was investigated for measurable properties by computer simulation on the example of a 2-methyl-propanol-1 - water system.
The original value of the parameter, calculated from the UNIQUAC parameters taken from Gmehling's book (ref. 1), using the Antoine vapor pressure equation to get the heat of vaporization of pure components, was -6802 J/mole. Upon changing its value by 2% (that is to -6666 J/mole) the following average deviations were found for VLE, LLE and γ^{∞} calculations:

	VLE		LLE		$\Delta\gamma_1^{\infty}$	$\Delta\gamma_2^{\infty}$
	ΔT	Δy	$\Delta x'$	$\Delta x''$	($\gamma_1^{\infty}\sim 40$)	($\gamma_2^{\infty}\sim 3.5$)
absolute	1.6	0.01	0.055	0.008	8.55	0.33
relative(%)	1.5	2.6	10.7	25.6	~20	~10

If the relative precision of the measurements is assumed equal, the order of informativity is:

$$x'', \gamma_1^{\infty} > x', \gamma_2 > y+T$$

This means that the composition of the water-rich phase in LLE and γ^{∞} of the alcohol is the most informative, followed by the composition of the organic phase of LLE together with the infinite dilution activity coefficient of water. The less sensitive experiment is the VLE in this example. These conclusions are far from being generally valid, but similar calculations may be performed for any mixture.

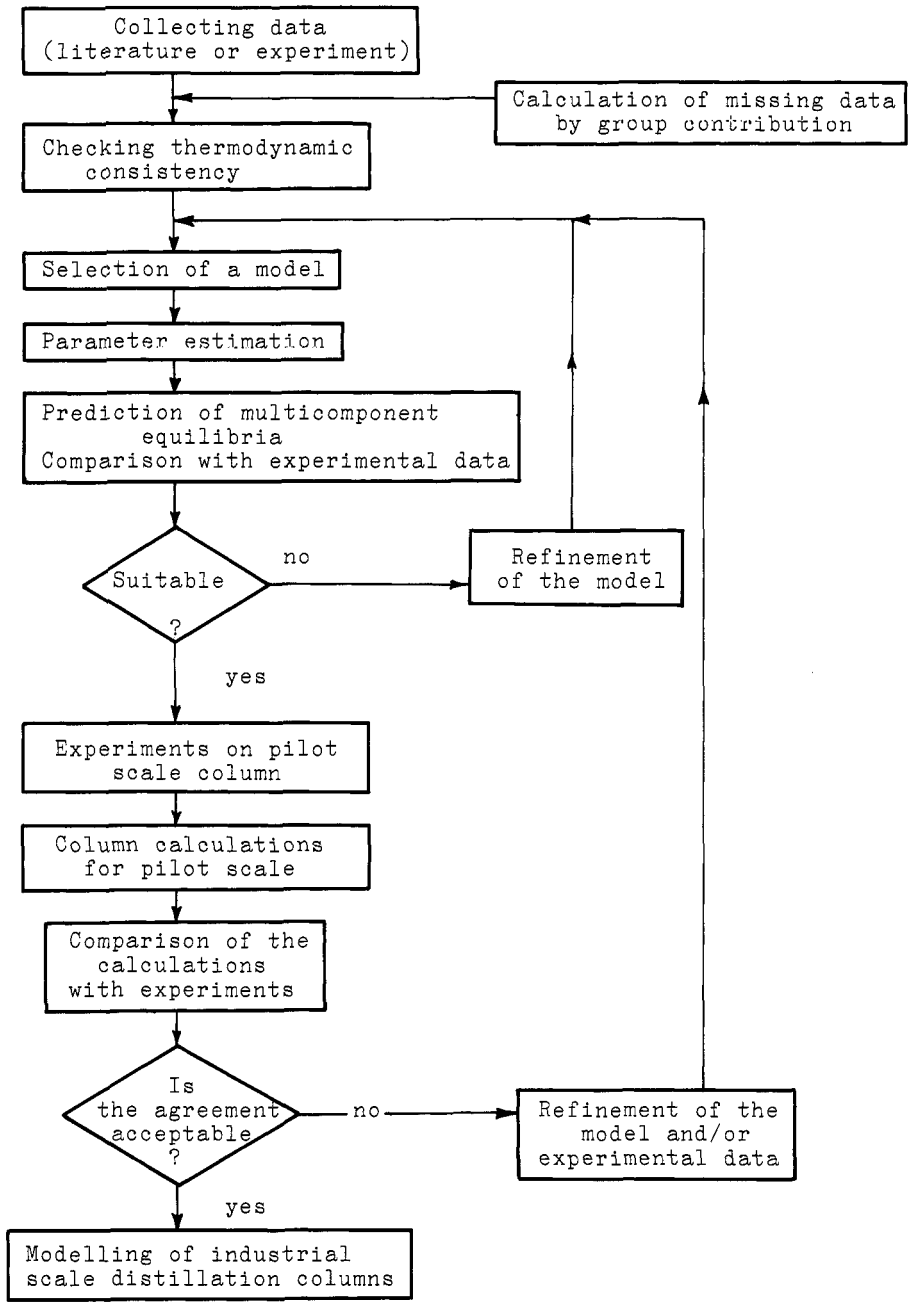

Fig. 1. Strategy of the design of a distillation column

NEED FOR THE DESIGN OF EXPERIMENTS

To the best of our knowledge, Sutton and MacGregor (ref. 2) proposed first that methods of optimization of experiments should also be applied to VLE measurements. In order to maximize the information stored in the parameters, they proposed to select those concentration values for measurements which would minimize the volume of the joint confidence region of model parameters. This is equivalent to the maximization of the determinant of variance - covariance matrix of parameters.
Howat and Swift (ref. 3) show a rather convincing illustrative example of VLE measurements for 2-methyl-butene-1 - isoprene system, applying a two-parameter model. The relative volatility in this system is about 1.1.
Plotting the determinant of the inverse covariance matrix of parameters as

a function of the two concentration values where the measurements would take place, it was found that the maximum information value is obtained at 0.3 and 0.75 mole fraction value for 2MB1. The optimum is shallow. Therefore, simulated experiments proceeded in the ranges 0.2 - 0.4 and 0.6 - 0.8, nine points altogether:

x: 0.2, 0.25, 0.3, 0.35; 0.60, 0.65, 0.70, 0.755, 0.80.

It is shown that this design contains the same piece of information as a 19-point equidistant experiment.

CONTROLLING AND IMPROVING PRECISION OF AN EXPERIMENT

VLE measurement in a dynamic still (ref. 4) is taken as an example. The main sources of systematic errors are:

- improper construction of the equipment (ref. 5), causing partial condensation of vapors or overheating of the liquid;
- bias of the temperature measurement caused by heat conduction and radiation; this may be avoided by calibrating the thermometer in the still itself under the condition of real VLE measurements;
- errors of concentration determination (e.g. partial evaporation of certain constituents of samples);
- impurity of materials

All these effects are to be excluded or diminished by careful construction and operation of the equipment. The last item is discussed in the following simulation study on the example of a water-contaminated acetone - methanol - propanol system.
First, the water content of pure compounds (acetone and methanol) was prescribed, then pseudo-binary mixtures were composed (acetone - methanol + water etc.) at 9 points each. Taking UNIQUAC parameters from the literature (ref. 1), the vapor-liquid equilibria for the ternary systems were predicted. Then by treating the generated ternaries as pseudo-binaries, UNIQUAC parameters were estimated. Using these pseudo-binary UNIQUAC parameters, isobaric ternary (acetone - methanol - propanol) equilibrium data were predicted at 20 points, the predicted T and y values were compared with those found using the original UNIQUAC parameters. The results are summarized in Table 1.

TABLE 1. Errors of phase equilibrium calculation caused by impurities.

mole fraction of water in acetone	mole fraction of water in methanol	SRQT	SRQY1	SRQY2
0.0005	0.001	2.33×10^{-5}	1.04×10^{-8}	1.07×10^{-8}
0.005	0.01	8.60×10^{-3}	3.80×10^{-6}	4.04×10^{-6}

$SRQX = \sum_{i}^{20}(x_i^o - x_i^w)^2/20$ x_i^o and x_i^w are obtained with and without taking into account water content, respectively.

It can be readily seen that at the really achievable minimum water content (ref. 6) the bias is well within the usual measurement errors, but on increasing the concentration of the contaminant the distortion becomes significant.

An investigation of random errors by error propagation analysis may help the experimentalist to identify the main error sources and gives an idea on variances of the parameter estimation. The following example taken from an old work (ref. 7) illustrates the method; the values of precision shown are out of date, however.
The error variances are estimated from repeated experiments or from the readability of instruments (readability limit = $\pm 3\sigma$). The error of pressure measurement is the sum of those of the barometer, the two levels of the manostat and of the oscillation of the outside pressure. The readability of the barometer scale is about 0.005 mmHg, that is $\sigma_b \simeq 0.1$ mmHg. During the reading of the water level of manostat 2 mm is distinguishable, therefore $\sigma_h \simeq 0.66$ mm, which corresponds to 0.05 mmHg, but as the difference of the two levels

is read, it is multiplied by $\sqrt{2}$: $\sigma_{man} \simeq 0.07$ mmHg. Oscillation of the outside pressure during the equilibrium measurement is taken as $\sigma_{out} \simeq 0.2$ mmHg. The resulting variance is

$$\sigma_P^2 = \sigma_b^2 + \sigma_{man}^2 + \sigma_{out}^2 = 10^{-2} + 5\times10^{-3} + 4\times10^{-2} = 5.5\times10^{-2} \text{ (mmHg)}^2$$

$$\sigma_P \simeq 0.235 \text{ mmHg}$$

The variance of the measured temperature is estimated from repeated experiments as $\sigma_T = 0.05$ K.

For the experimental determination of the composition, refractivity index vs mole fraction calibration curve was determined first. The variance of a point in the n(x,T) space is expressed by the error propagation law:

$$\sigma_{n(x)}^2 = \sigma_{n,o}^2 + \left(\frac{\partial n}{\partial T}\right)^2 \sigma_T^2 + \left(\frac{\partial n}{\partial x}\right)^2 \sigma_{x,o}^2$$

where the first term stands for the refractometer reading, the second one is the contribution of the temperature measurement, while the third term comes from the error committed when composing the mixture from pure materials by weight; this last contribution was proved to be negligibly small in the specific example.

The readability of the refractometer scale is about 5×10^{-4}, therefore $\sigma_{n,o} \simeq 8\times10^{-5}$, while on the thermometer scale 1K can be distinguished, $\sigma_T \simeq 0.133$ K is taken.

The average slope of the n vs T curve is about 5.5×10^{-4}. Thus,

$$\sigma_{n(x)}^2 = 6.4\times10^{-9} + 8.5\times10^{-9} = 1.5\times10^{-8}$$

As each calibration measurement was repeated three times, this variance is divided by three: $\sigma_{n(x)}^2 = 5\times10^{-9}$.

When this calibration curve is used to determine the composition from measured refractivity index values, the variance is approximated by

$$\sigma_x^2 = \left(\frac{\partial x}{\partial n}\right)^2 (\sigma_{n,o}^2 + \sigma_{n(x)}^2) = 6^2 \times 2\times5\times10^{-9} \simeq 4\times10^{-7} ,$$

where $\sigma_{n(x)}^2$ is the variance of the calibration curve which is roughly taken equal to $\sigma_{n,o}^2$, thus $\sigma_x = \sigma_y \simeq 6.3\times10^{-4}$.

To check if all important error sources were considered, VLE experiments were made at liquid compositions which were close to each other, see Fig. 2. To describe these points (the two clouds separately), any smooth curve is satisfactory, and the Wilson equation was chosen. The resulting variances were estimated by s_r^2 calculated from the residuals.

Using error propagation results, the variance of a vapor composition point with x and T independent variables subject to error is given below and compared with s_r^2:

$x_{benzene}$	σ_y^2	+ $\left(\frac{\partial y}{\partial x}\right)^2 \sigma_x^2$	+ $\left(\frac{\partial y}{\partial T}\right)^2 \sigma_T^2$	= $\sigma_{y(x,T)}^2$	s_r^2
0.13	4×10^{-7}	4.55×10^{-7}	5.6×10^{-10}	8.5×10^{-7}	2.5×10^{-6}
0.8	4×10^{-7}	4.19×10^{-7}	4.2×10^{-10}	4.4×10^{-7}	8.5×10^{-7}

It is seen that s_r^2 and $\sigma_{y(x,T)}^2$ are of the same order of magnitude in both regions investigated, that is, the results of error propagation calculations

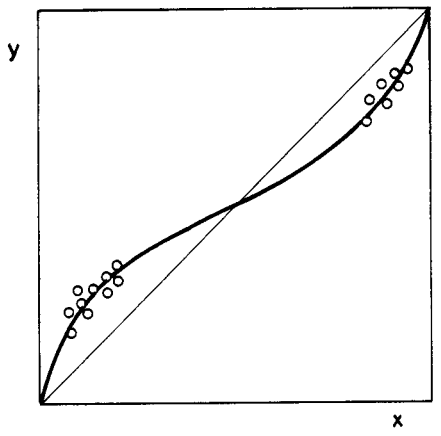

Fig. 2. Cloud-repetition

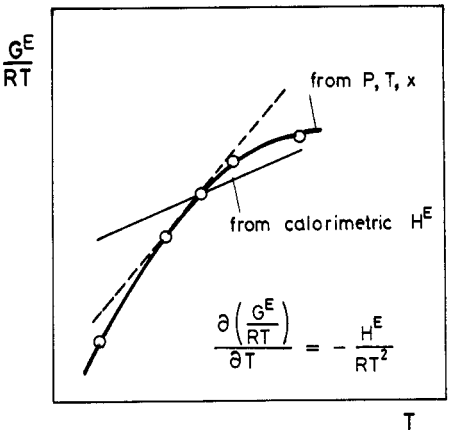

Fig. 3. Gibbs-Helmholtz analysis

are justified, the variances obtained are appropriate for using them in a parameter estimation procedure. It is also seen that the dominant error is that of the composition measurement, which is to be refined. Similar results were obtained for $\sigma^2_{P(x,T)}$.

CONSISTENCY

It means accordance with the rules of thermodynamics and/or meeting obvious requirements.

Internal consistency is checked within a data set. Typical examples are :
- checking of outliers;
- Gibbs-Duhem analysis of binary VLE data (x,y,T,P) (ref. 8, 9);
- pure component vapor pressure as the limiting value for VLE (ref. 10).

External consistency is investigated between different data sets. Subcases:
- results of different authors regarding the same property. If the conditions are identical, the data are simply plotted together, or deviation plots are prepared. If the conditions are different, the measured data are usually projected on a common base. Typical examples: second virial coefficient B vs T plots from pure fluid PVT data or B_{12} vs T from mixture PVTx data (ref. 11);
- series-consistency for properties of homologous compounds: plots in carbon number or normal boiling points etc. (ref. 12);
- with data on other properties. This item is discussed below.

In the era of huge data banks it became important, at least for the users (correlators), to check data coming from different sources for different properties but to be used together. A well known method for this task is the multiproperty analysis (ref. 13). The essence of the method is a study of residuals of different properties obtained as deviations from a function fitted at the same time to all properties. As an example we give the simultaneous treatment of density and enthalpy data:

$$w_\rho \sum_i \left[\rho_i - \rho_i(\hat{\underline{\theta}})\right]^2 + w_h \sum_j \left[h_j - h_j(\hat{\underline{\theta}})\right]^2 = \min$$

Here $\rho(\underline{\theta})$ and $h(\underline{\theta})$ are related through thermodynamics, $\underline{\theta}$ is the vector of parameters, the w weights (which may also be different for each i and j measurement point) are taken from statistical or heuristic considerations. The residuals $\rho_i - \rho_i(\hat{\underline{\theta}})$ and $h_j - h_j(\hat{\underline{\theta}})$ are plotted and examined. Algorithmized statistical tests for trend and shift are useful (ref. 14).
The crucial point of this kind of investigation is the inevitable use of an appropriate model (an equation of state in the example), which may seriously distort the residuals falsifying also the conclusions. The assignation of weights is also ambiguous.

A NOVEL APPROACH: DATA BANK CONSISTENCY

The method to be outlined is a generalization of the Gibbs-Helmholtz analysis, where G^E values calculated from experimental VLE data (x,T,P) are plotted

against the temperature, and the slope of the G^E vs T curve is compared with that calculated from calorimetric H^E data (ref. 8, 9) for a mixture of the same composition, as is illustrated in Fig. 3.

As H^E is usually not available at the same composition where G^E is measured, interpolation may be required. For different compositions (and also different temperatures) a whole set of curves and slopes is to be examined. The aim of the generalized method is to characterize the measure of inconsistency, preferably separately for different regions of independent variables T, P, x. The method is model-free, making use of strict thermodynamic relations only. Equations of the thermodynamics are mainly differential (or partial differential) equations, containing derivatives with respect to T, P and x, and in most cases these derivatives are not measured directly but calculated from experimental data.

The three ways for treating the problem will be illustrated by the example of a first order ordinary differential equation (ref. 15):

$$\frac{dy}{dx} + \Theta_1 y = 0 \ ;$$

a.) the exact solution of the differential equation is:

$$y = y_o \exp(-\Theta_1 x)$$

where y_o may be known or taken as a second Θ_2 parameter;

b.) taking numerical derivatives:

$$\left(\frac{dy}{dx}\right)_i \simeq \frac{\Delta y}{\Delta x} = \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}} \ ,$$

the solution is proceeded by iteration;

c.) formal integration of the differential equation and numerical integration of the data:

$$y_i - y_{i-1} + \Theta_1 \int_{x_{i-1}}^{x_i} y(x)\, dx = 0 \ .$$

Method a.) usually cannot be applied because the differential equations of the thermodynamics are much too complicated for solving them analytically. Using method b.), the random errors are exaggerated during differentiation, while in the integration applied in method c.) these random errors are filtered out. This is advantageous as systematic deviations are looked for, thus, the third method was chosen.

The idea was taken from chemical engineering, where properties as component mass flows are related through balance equations.

Some of the properties are measured. If the number of measured properties exceeds the system's number of degrees of freedom (the number of variables reduced by the number of relations between them), the redundancy allows the tracing of gross (that is substantial systematic) errors (ref. 16). The system of balance equations is as follows:

$$\underline{\underline{W}}\ \underline{X} - \underline{b} = \underline{0} \ ,$$

where $\underline{\underline{W}}$ is the matrix of coefficients of balance equations,

$\underline{X}$ contains the true values of properties (e.g. component mass flows),

$\underline{b}$ stands for the right hand side of equations.

For the really measured properties this equation is not fulfilled, the difference is the balance error:

$$\underline{\underline{W}}\ \underline{x} - \underline{b} = \underline{f}$$

where $\underline{x}$ are the measured values of properties,

$\underline{f}$ is the balance error.

The estimation criterion according to the maximum likelihood principle is given as

$$(\underline{X} - \underline{x})^T\ \underline{\underline{V}}^{-1}\ (\underline{X} - \underline{x}) = \min$$

with the condition

$$\underline{\underline{W}}\ \underline{X} - \underline{b} = \underline{0} \ ,$$

where $\underline{X}$ stands the estimated values of properties,

$\underline{\underline{V}}$ is the variance-covariance matrix of the measured properties.

The measure of deviations is defined as

$$q^2 = (\underline{\underline{W}}\,\underline{x} - \underline{b})^T (\underline{\underline{W}}\,\underline{\underline{V}}\,\underline{\underline{W}}^T)^{-1} (\underline{\underline{W}}\,\underline{x} - \underline{b})$$

Formulae for nonlinear "balance" equations are also available (ref. 16). The use of the method is illustrated on the example of an investigation of the consistency of pure component vapor pressure and heat of vaporization data.
The Clausius-Clapeyron equation for the saturation, neglecting the volume of the liquid and treating the vapor as an ideal gas, is written as

$$\frac{d \ln p^o}{dT} = \frac{\Delta H^{vap}}{R\,T^2}$$

Upon an integration from T_i to T_{i+1} and rearrangement, the following equation is obtained:

$$- \ln p^o(T_{i+1}) + \ln p^o(T_i) - \int_{T_i}^{T_{i+1}} \frac{\Delta H^{vap}}{RT^2}\, dT = 0 .$$

The balance variables (elements of the $\underline{x}$ vector) are: $\ln p^o(T_i)$, $i=1,\ldots,n$;

$$\int_{T_i}^{T_{i+1}} \frac{\Delta H^{vap}}{RT^2}\, dT, \qquad i = 1, \ldots, n-1 ;$$

there are n−1 balance equations altogether.
The elements of $\underline{b}$ are zero. The $\underline{\underline{W}}$ coefficient matrix assumes the following form:

$$\underline{\underline{W}} = \begin{pmatrix} -1 & +1 & 0 & 0\ldots0 & 0 & -1 & 0\ldots0 \\ 0 & -1 & +1 & 0\ldots0 & 0 & 0 & -1\ldots0 \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ 0 & 0 & 0 & 0\ldots0 & +1 & 0 & \ldots\ldots -1 \end{pmatrix}$$

The n number of points should not exceed the number of experimental points but it must be essentially greater than two, otherwise the integration between the two edges may cancel the systematic deviations. The results of a sample calculation are shown in Fig. 4, where an increasing systematic error was superposed to the true vapor pressure curve. The main features of the method, i.e., the sensitivity and the possibility of separate examination of regions, are well illustrated.

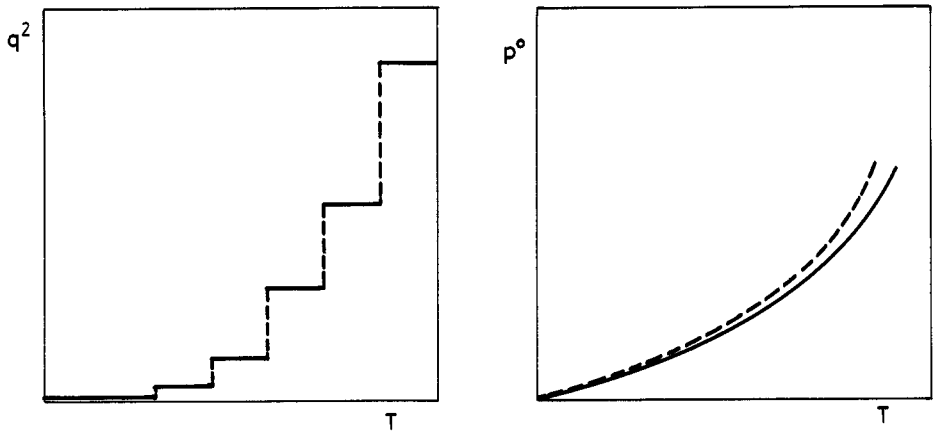

Fig. 4. Analysis of consistency for vapor pressure and heat of vaporization data.

USE OF MODELS AND PARAMETER ESTIMATION

In principle, it would be possible to calculate the properties of a multicomponent mixture using molecular data only. For this purpose models are used, parameters of which are not adjusted to the measured data. As the other limiting case, it is also possible to measure the required data directly; thus, no model would be required. E.g., if the boiling point of a multicomponent mixture of a certain composition is required, it can be measured.
From engineering point of view, none of these limiting cases is practicable. We do not have good enough models for ab initio calculations, on the one hand, while the phase equilibrium data in a whole concentration, temperature and pressure range are required for engineering calculations on the other. In the intermediate region models are used with parameters adjusted to experimental data:

ab initio calculations (no adjustable parameter)	← models with adjustable parameters →	measuring the required data directly (no model)

If the required properties in a region of interest are mapped by experiments, we are close to the right edge, and what we really need is an interpolation, the representation of the map in a mathematical form. Going from the right to the left, less (or no) multicomponent experimental information is available, and models of greater predictive ability are required.
This predictive ability is connected (more or less closely) with the theoretical background of the models.
To understand the basis of the predictive ability of a model is not always easy. E.g., the simplifying assumptions of group contribution models (similar groups coming from different molecules are not distinguished) are usually emphasised without stressing the major gain of these models: the assumption of the homogeneity of molecules is dropped, that is, the smearing approximation is not applied, and the molecules are treated as heterogeneous.
Physically sound models have parameters of physical meaning, the maximum information is to be extracted from the experimental data with respect to the parameters. A general method for this purpose is the properly applied maximum likelihood estimation procedure, which is proved to give the parameters with the smallest achievable uncertainty (ref. 17, 18).

REFERENCES

1. J. Gmehling, U. Onken, W. Arlt, P. Grenzheuser, U. Weidlich and B. Kolbe, Vapor-Liquid Equilibrium Data Collection, DECHEMA, Frankfurt/M (from 1977).
2. T.L. Sutton and J.F. MacGregor, Canad. J. Chem. Eng. 55, 609-613 (1977).
3. C.S. Howat and G.W. Swift, Fluid Phase Equil. 14, 289-301 (1983).
4. J. Manczinger and K. Tettamanti, Per. Polytechn. Chem. Eng. (Budapest) 10 183 (1966).
5. E. Hála, J. Pick, V. Fried and O. Vilim, Vapour-Liquid-Equilibrium, Pergamon, Oxford (1968).
6. J.A. Riddick and W.B. Bunger, Organic Solvents, Wiley, New York (1970).
7. S. Kemény, Ph.D. Thesis Budapest (1976).
8. J.D. Olson, Fluid Phase Equil. 14, 383-392 (1983).
9. B. Janaszewski, P. Oracz, M. Goral and S. Warycha, Fluid Phase Equil. 9, 295-310 (1982).
10. H.C. Van Ness, S.M. Byer and R.E. Gibbs, AIChE J. 19 238-244 (1973).
11. K.E. Starling, J.L. Savidge and K.H. Kumar, Fluid Phase Equil. 27 203-219 (1986).
12. J.D. Chase: Chem. Eng. Progr. April, 63-67 (1984).
13. K.W. Cox, J.L. Bono, Y.C. Kwok and K.E. Starling, Ind. Eng. Chem. Fundam. 10, 245 (1971).
14. K. Kollár-Hunek, S. Kemény, K. Héberger, P. Angyal and É. Thury, Fluid Phase Equil. 27, 405-425 (1986).
15. Y. Bard, Nonlinear Parameter Estimation, Academic Press, New York (1974).
16. G.A. Almásy and T. Sztanó, Problems of Control and Information Theory 4, 57 (1975).
17. T.A. Duever, S.E. Keeler, P.M. Reilly, J.H. Vera and P.A. Williams, Chem. Eng. Sci. 42, 403-412 (1987).
18. S. Kemény, J. Manczinger, S. Skjold-Jorgensen and K. Tóth, AIChE J. 28, 20-30 (1982).