A Convenient Hybrid Method for Obtaining Liquid–Liquid Equilibrium Data in Ternary Systems

Radu C. Racoviţă,[1] Adrian Victor Crișciu,[1] Sergiu Sima,[1] and Catinca Secuianu[1,2,*]

[1]*Department of Inorganic Chemistry, Physical Chemistry and Electrochemistry, Faculty of Applied Chemistry and Materials Science, University Politehnica of Bucharest, 1-7 Gh. Polizu Street, S1, 011061 Bucharest, Romania, catinca.secuianu@upb.ro;*

[2]*Department of Chemical Engineering, Imperial College London, South Kensington Campus, SW7 2AZ London, United Kingdom, c.secuianu@imperial.ac.uk*

Supplementary Information

1. Primary data processing

As mentioned above, the data processing method first proposed by Upchurch[1] is an integral part of the experimental protocol. As such, the mathematical translation of this method should not be regarded as actual "data modelling", but as a necessary step in the progression from raw experimental data (titration volumes and refractive indices) towards processed data (binodal curve and tie-line compositions), complete with consistency checks (via mass balance) and associated confidence intervals. In comparison with the original, purely graphical, primary processing method, this is the main advantage of the mathematical reformulation proposed in this paper: it allows a systematic analysis of uncertainties and provides statistical tools for the interpretation of consistency criteria.

1.1. Regression model

As stated above, the goal in this phase of data processing (called henceforth primary data processing) is not to provide thermodynamic model for the phase behavior of the system under study, but to allow tie lines to be calculated from raw (volumetric and refractometric) data and to provide a basis for estimating uncertainties and for the correct interpretation of mass balance checks. As such, the regression model used in this phase is based on strictly empirical polynomial functions for both the binodal and the refractometric curve, whose order was established by leave–one–out–cross–validation, as described below. In order to construct a proper regression procedure, it should be recognized that, along the binodal curve, both mole fractions are observed quantities,

[*]Corresponding author: catinca.secuianu@upb.ro (C. Secuianu)

so both are affected by comparable experimental errors. If normal distribution of errors is considered and correlation is neglected both within and across experimental observations ((x_1, x_2, n_D) triplets along the saturation curve), the total negative log-likelihood of the data set (including cloud-point and refractometric data) is:

$$\mathcal{L} = \sum_{i=1}^{n}\left(x_{2i} - bin(x_{1i}^{*}; a_B)\right)^2 + (x_{1i} - x_{1i}^{*})^2 + \left(n_{D,i} - ref(x_{1i}^{*} - a_R)\right)^2 \quad (1)$$

where

$$bin(x; a_B) = \sum_{i=0}^{p_B} a_{Bi} x_i$$

and

$$ref(x; a_R) = \sum_{i=0}^{p_R} a_{Ri} x^i$$

are the polynomial models for the binodal and refractometric curve, p_B and p_R are, respectively, their polynomial orders, a_B and a_R are the corresponding coefficient sets and x^{*}_{1i}, $i = \overline{1,n}$ are the "true" (unobservable) mole fractions for component 1.

It should be noted that, if the normality and independence conditions are rigorously met, minimizing the objective function $\mathcal{L}$ given by (1) provides maximum likelihood estimators for parameters a_B, a_R and x^{*}_{1i}. Moreover, even if some of these conditions are violated, the above function still provides, upon minimization, good parameter estimates. However, in this case, any statistical inference should be based on non-parametric methods, as the maximum likelihood theory is no longer applicable. Such methods, namely classic and moving block bootstrap, will be used below to obtain confidence intervals for mass balance consistency checks.

1.2. Regression algorithm

The parameters of the primary data processing model are a_B (polynomial coefficients for the binodal curve), a_R (polynomial coefficients for the refractometric curve) and x^{*}_{1i} ("true", unobservable, mole fractions for component 1). Their best estimates are the solutions of the gradient system:

$$\nabla\mathcal{L} = 0$$

which decomposes into three blocks:

A block of polynomial regression equations for the binodal. Simple algebra shows that the derivatives of L with respect to a_{Bk} can be written as:

$$\frac{\partial \mathcal{L}}{\partial a_{Bk}} = \sum_{j=0}^{p_B} a_{Bj} \sum_{i=1}^{n} x_{1i}^{*j} x_{1i}^{*k} - \sum_{i=1}^{n} x_{2i} x_{1i}^{*k}$$

which leads to

$$\forall k = \overline{0, p_B} \bullet \frac{\partial \mathcal{L}}{\partial a_{Bk}} = 0 \equiv \hat{X}_B^{*T} \hat{X}_B^{*} a_B = \hat{X}_B^{*T} x_2 \qquad (2)$$

where $\hat{X}_B^{*}$ is the polynomial design matrix for the binodal:

$$\hat{X}_B^{*} = \begin{pmatrix} 1 & x_{11}^{*} & \cdots & x_{11}^{*p_B} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1n}^{*} & \cdots & x_{1n}^{*p_B} \end{pmatrix}$$

These are the classical orthogonal equations for polynomial regression with predictor x_1^*.

A block of polynomial regression equations for the refractometric curve. In much the same manner, the subsystem

$$\forall k = \overline{0, p_R} \bullet \frac{\partial \mathcal{L}}{\partial a_{Rk}} = 0$$

reduces to

$$\hat{X}_R^{*T} \hat{X}_R^{*} a_R = \hat{X}_R^{*T} x_2 \qquad (3)$$

where $\hat{X}_R^{*}$ is the polynomial design matrix for the refractometric curve:

$$\hat{X}_R^{*} = \begin{pmatrix} 1 & x_{11}^{*} & \cdots & x_{11}^{*p_R} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1n}^{*} & \cdots & x_{1n}^{*p_R} \end{pmatrix}$$

A block of independent polynomial equations for x_1^*. Finally, the derivatives with respect to x_{1k}^* can be written as:

$$-\frac{\partial \mathcal{L}}{\partial x_{1k}^*} = bin'(x_{1k}^*; a_B)[x_{2k} - bin(x_{1k}^*; a_B)] + (x_{1k} - x_{1k}^*) + ref'(x_{1k}^*; a_R)\left[n_{D,k} - ref(x_{1k}^*; a_R)\right] \quad (4)$$

This particular structure of the gradient system leads to the following regression algorithm:

1. Perform all required pre-processing on the raw data (i.e. transform volumetric into compositional data and transform rectangular into Gibbs– Roseboom triangular coordinates).

2. Initialize the "true" mole fractions for component 1 with the experimental ones:

$x_1^* \leftarrow x_1$

3. Solve subsystem (2) to obtain current estimates for the binodal coefficients a_B.

4. Solve subsystem (3) to obtain current estimates for the refractometric coefficients a_R.

5. Solve equations (4) to obtain new estimates for the "true" mole fractions x_1^*. This step can be performed either using a polynomial equation solver, or *via* a Newton–type method, using the current estimates as initial guesses.

6. Repeat steps 3 to 5 above until some convergence criterion is met. In this work, the convergence criterion was chosen to ensure that the relative change in the parameter with the highest absolute value is lower than $\sqrt{\varepsilon_M}$, where ε_M is the machine-precision.

If the experimental data are not too noisy, this algorithm will converge in 3 to 5 iterations.

The polynomial orders p_B and p_R were chosen such as to strike the best compromise between bias and variance, measured, respectively, by the coefficient of determination R^2 and the leave–one–out cross–validation error LOOCV. These are defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left[z_i^{exp} - p\left(x_i^{exp}\right)\right]^2}{\sum_{i=1}^{N}\left[z_i^{exp} - \bar{n}\right]^2}$$

where the superscript *exp* designates an experimentally observed quantity, z is the model response (either mole fraction for component 2 or refractive index), x is the mole fraction of the reference component along the binodal line, $\bar{z}$ is the average response along the binodal line, p is the fitting polynomial and n is the number of experimental points and

$$LOOCV = \frac{1}{n}\sum_{i=1}^{N}\left[z_i^{exp} - p_{-i}\left(x_i^{exp}\right)\right]^2$$

where p_{-i} is the regression polynomial estimated from the original set without the i^{th} observation. All the above calculations were performed during a preliminary run, using independent classical regression models (without considering the errors in the independent variable x_1) in order to speed up the process.

As an example, **Figure 3** illustrates this procedure as applied to select an optimal degree polynomial for the refractometric curve along the trichloroethylene + water + ethanol binodal line. The figure suggests that a cubic polynomial is the optimal model for the underlying data. A lower-degree polynomial is too stiff to represent the data with any accuracy, while a higher-degree polynomial offers an insignificant gain in representation accuracy while significantly increasing the cross-validation error. This is due to the added flexibility, leading to overfitting.

Example results for the trichloroethylene + water + ethanol system are shown in **Figures 4** and **6**. As it can be seen, both curves provide an excellent fit to the data.

1.3. Residual analysis

A necessary step to perform after parameter estimation is residual analysis. This provides information about the applicability of the maximum likelihood theory thus guiding the ensuing uncertainty analysis. The objectives of this process are first to detect any correlation or autocorrelation among residuals both within and across observations, and, if necessary, to suggest a particular error distribution for uncertainty estimation. It is motivated by the observation that, if the regression model is correct, the residuals are actually samples of the underlying error distributions. For the purposes of this work, residuals are defined as follows:

• For the mole fraction of component 1 (the reference component):

$$\varepsilon_{*,i} = x_{1i} - x_{1i}^{*}$$

• For the mole fraction of component 2:

$$\varepsilon_{bin,i} = x_{2i} - bin(x_{1i}^{*}; a_B)$$

• For the refractive index:

$$\varepsilon_{ref,i} = n_i - ref(x_{1i}^*; a_R)$$

The analysis aimed at detecting any relationship among residuals, both across and within observation. Therefore, the main tool for this purpose was chosen to be Kendall's rank correlation coefficient, as it is nonparametric so does not impose any particular form on the correlation function. In order to interpret its values, a hypothesis test was conducted, based on a null hypothesis of no correlation. This means that the null distribution of the correlation coefficient can be estimated from permutation re-samplings. In this work, the null distribution was estimated on the basis of 5000 random permutations, so the confidence intervals could be computed directly from the quantiles of these samples. The results are reported in **Table S1** for the correlation between residuals and predictor x_1^* and in table S2 for the autocorrelation among residuals (only lag-1 autocorrelation was considered).

Table S1. Correlation coefficients for residuals and associated confidence intervals.

Correlation type	τ	Confidence interval (95%)
ε_* *vs.* x^*_1	0.268	[-0.346, 0.333]
ε_{bin} *vs.* x^*_1	-0.007	[-0.333, 0.320]
ε_{ref} *vs.* x^*_1	-0.150	[-0.333, 0.333]
ε_* *vs.* ε_{bin}	-0.007	[-0.333, 0.333]

Table S2. Autocorrelation coefficients for residuals and associated confidence intervals.

Correlation type	τ	Confidence interval (95%)
ε_*	0.382	[-0.346, 0.333]
ε_{bin}	0.500	[-0.333, 0.320]
ε_{ref}	0.000	[-0.333, 0.333]

These results show that the both the correlation of residuals along the binodal line within observations and the correlation of residuals along the refractometric curve across observations are statistically insignificant. However, the same results show a statistically significant correlation of residuals across observations along the binodal curve. Although this result is to be expected, due to the specific sequence of experimental steps described in the main paper, it shows that the

estimates obtained from the above–mentioned algorithm are *not* maximum likelihood estimates, so any inference on them should be nonparametric.

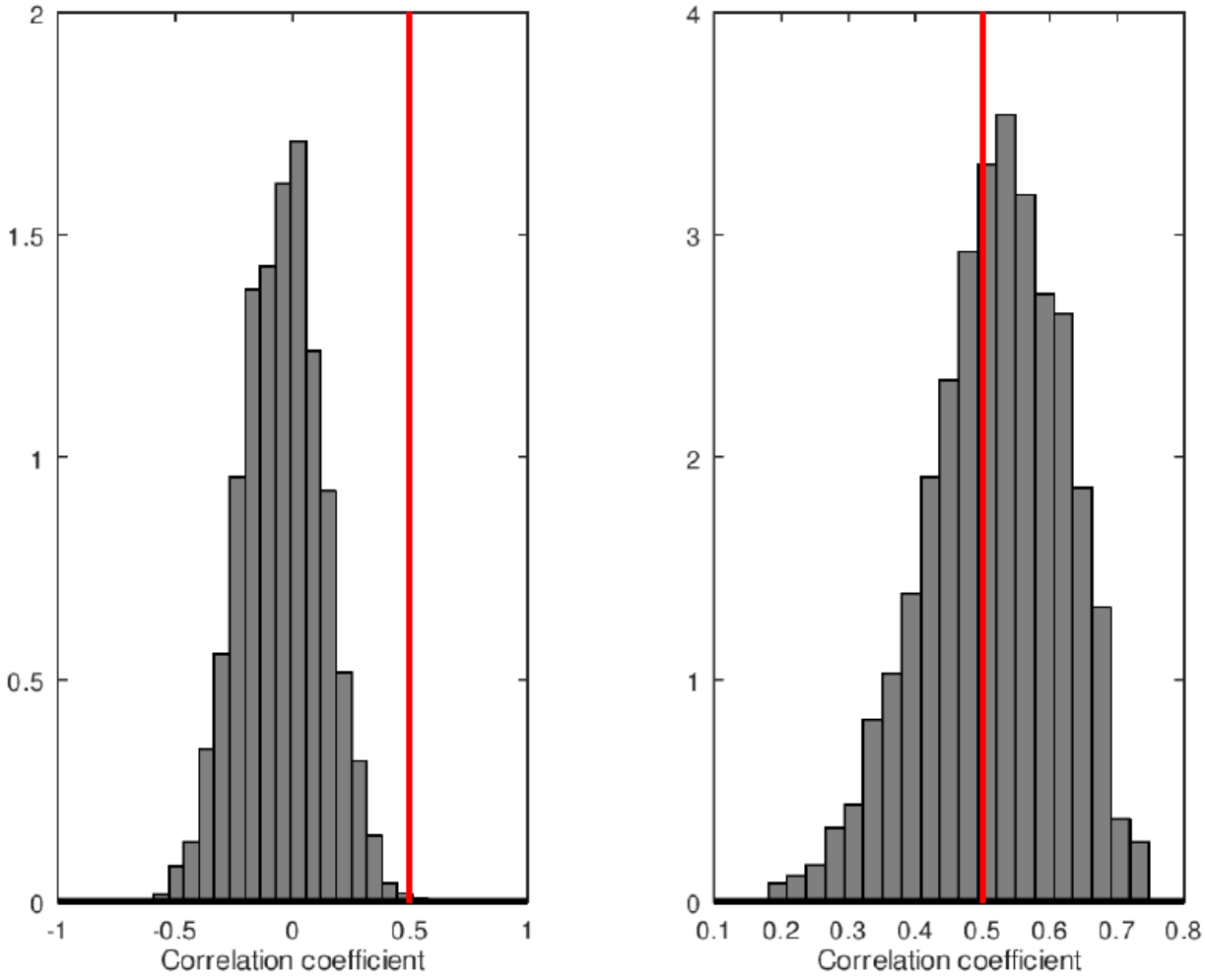

Figure S1. Mobile block bootstrap calibration to preserve autocorrelation in compositional residuals. The thick solid red line corresponds to the observed value. Block sizes are, respectively 1 (left panel) and 6 (right panel).

In this work, statistical inference is performed *via* bootstrap. Due to the specific autocorrelation structure among residuals, resampling from compositional residuals was performed *via* moving block bootstrap,[2] and resampling from refractive index residuals via classical bootstrap.[3] The block length for moving block bootstrap was chosen to preserve autocorrelation (as it is a feature of the experimental design), i.e. the bootstrap distribution of τ to be centered around the observed value. **Figure S1** illustrates the concept.

1.4. Tie line calculation

Tie line calculation is a very straightforward process. The necessary data are the feed composition and the pair of refractive indices for the coexistent phases. In fact, refractive indices provide enough information for tie line calculation, but the feed (global composition) provides a redundancy necessary for consistency tests.

In order to compute tie lines, each refractive index is used to reverse interpolate the corresponding mole fraction for the reference component:

$$x_1^{*L_1} : \; n_D^{L_1} = bin\left(x_1^{*L_1}; a_R\right)$$

and

$$x_1^{*L_2} : n_D^{L_2} \; = bin\left(x_1^{*L_2}; a_R\right)$$

It should be noted that, due to the construction of the regression model, these compositions are "true" mole fraction of the reference component. The second mole fraction is then computed directly from the binodal polynomial model:

$$x_2^{L_1} = bin\left(x_1^{*L_1}; a_B\right), \qquad x_2^{L_2} = bin\left(x_1^{*L_2}; a_B\right)$$

The distance from feed to tie line is then computed using standard analytical geometry formulae. As an example, **Figure 7** shows the results of these calculations. As it can be seen, tie lines are almost exactly aligned with the corresponding feed compositions, which proves the validity of the above regression procedure.

1.5. Consistency analysis

As shown in **Figure 7**, tie lines and feed compositions can be almost perfectly aligned. However, as with any data, the distance from feed to tie line is a stochastic quantity, subject to random fluctuations due to experimental errors. As such, its value is meaningless by itself, any interpretation requiring an associated sampling distribution or, at the very least, an estimation of a confidence interval. In this work, confidence intervals for these distances were determined as inter-quantile intervals from bootstrap approximations to the sampling distributions, based on the following procedure:

1. Perform bootstrap resampling on each set of residuals. Residuals along the refractometric curve were resampled by classic bootstrap,[3] as they are non-correlated. Residuals along the binodal were resampled by moving block bootstrap,[2] as they exhibit statistically significant autocorrelation.

2. From each bootstrap sample, form a synthetic data set by adding the sample to the corresponding calculated value:

$$x_1 = x_1^* + \varepsilon_*$$

$$x_2 = bin\ (x_1^*; a_B) + \varepsilon_{bin}$$

$$n = ref\ (x_1^*; a_R) + \varepsilon_{ref}$$

3. Perform the above-mentioned calculations (i.e. regression and tie line calculation) on each synthetic data set and collect the results.

4. From the results for feed-to-tie-line distance, build sampling distributions and estimate confidence intervals as inter-quantile ranges. Based on these calculations, a tie line is deemed consistent with respect to mass balance if the confidence interval for the corresponding feed-to-tie-line distance covers 0, distances being taken as positive or negative according to the position of the feed on one side or the other of the tie-line. **Table S3** reports the tie lines along with their feed-to-tie-line distances and **Table S4** reports the corresponding 95% confidence intervals for these distances. In these tables δ stands for the feed-to-tie-line distance, and the data (Gibbs-transformed molar fractions) are for the trichloroethylene + water + ethanol system.

Table S3. Tie lines and feed-to-tie-line distances for trichloroethylene + water + ethanol at 294.15±0.01 K and 1008±1 mbar.

No.	x_1^{L1}	x_2^{L1}	x_1^{L2}	x_2^{L2}	δ
1	0.7013	0.3499	0.3139	0.3275	0.0000
2	0.7345	0.3296	0.2817	0.3065	0.0041
3	0.7584	0.3126	0.2554	0.2875	0.0065
4	0.7751	0.2995	0.2345	0.2711	0.0066
5	0.7966	0.2812	0.1950	0.2374	0.0033
6	0.8268	0.2526	0.1386	0.1830	0.0037

As can be seed, at a confidence level of 95% and allowing for the effect of experimental errors, all tie lines are consistent with respect to mass balance. The confidence intervals are tight enough to support this hypothesis with a high degree of accuracy.

Table S4. Feed-to-tie-line distances 95% confidence intervals for trichloroethylene + water + ethanol at 294.15±0.01 K and 1008±1 mbar.

No.	Confidence interval for δ (95%)
1	[-0.0001, 0.0058]
2	[-0.0003, 0.0100]
3	[-0.0012, 0.0127]
4	[-0.0011, 0.0130]
5	[-0.0002, 0.0105]
6	[-0.0002, 0.0108]

1.6. Further results

Applying the above-discussed procedure for the other three systems (namely isooctane + ethyl acetate + acetonitrile, water + acetonitrile + chloroform and cyclohexane + ethyl acetate + acetonitrile) yields the results reported in **Tables S5**, **S6** and **S7**.

Table S5. Tie lines given in terms of Gibbs-transformed molar fractions and feed-to-tie-line distances for isooctane + ethyl acetate + acetonitrile at 294.15±0.01 K and 1008±1 mbar.

No.	x_1^{L1}	x_2^{L1}	x_1^{L2}	x_2^{L2}	Confidence interval for δ (95%)
1	0.27255	0.18767	0.60165	0.17693	[-0.01390, 0.00150]
2	0.23315	0.17277	0.67085	0.15684	[-0.01218, 0.00245]
3	0.17385	0.13345	0.73310	0.13025	[-0.01365, 0.00181]
4	0.18020	0.14567	0.70510	0.14359	[-0.01086, 0.00070]

Table S6. Tie lines given in terms of Gibbs-transformed molar fractions and feed-to-tie-line distances for water + acetonitrile + chloroform at 294.15±0.01 K and 1008±1 mbar.

No.	x_1^{L1}	x_2^{L1}	x_1^{L2}	x_2^{L2}	Confidence interval for δ (95%)
1	0.58260	0.57625	0.80965	0.31515	[-0.09547, 0.11567]
2	0.47500	0.58058	0.83825	0.26630	[-0.06599, 0.07298]
3	0.31970	0.44548	0.88135	0.19061	[-0.02330, 0.00314]
4	0.24230	0.33238	0.90490	0.14948	[-0.06695, 0.00799]

Table S7. Tie lines given in terms of Gibbs-transformed molar fractions and feed-to-tie-line distances for cyclohexane + ethyl acetate + acetonitrile at 294.15±0.01 K and 1008±1 mbar.

No.	x_1^{L1}	x_2^{L1}	x_1^{L2}	x_2^{L2}	Confidence interval for δ (95%)
1	0.33690	0.16905	0.73005	0.13276	[-0.01890, 0.00507]
2	0.32440	0.16524	0.75635	0.12150	[-0.01526, 0.00325]
3	0.29820	0.15588	0.76975	0.11544	[-0.02647, 0.00893]
4	0.22915	0.12324	0.82015	0.09085	[-0.03253, 0.00119]

As can be seen from these tables, all reported tie-lines are mass-balance consistent. However, for some tie-lines, the 95% confidence interval for the feed-to-tie-line distance is quite large. This can be attributed to the propagation of experimental errors from the cloud-point titration stage, which is a visual procedure, so prone to more experimental noise than more automated techniques. Moreover, this proves the point made above, that interpreting mass-balance consistency results (such as feed-to-tie-line distance) is meaningless without considering the uncertainty limits due to experimental errors.

2. The invariance of interval coverage under affine transformations

All calculations described above are performed in the *Gibbs–Rooseboom* coordinate system. In this section, a proof of invariance is given for the interval coverage property under the affine transformation linking *Gibbs–Rooseboom* and rectangular coordinates.

Let $\left(\overline{x_1, x_2}\right)$ denote the line segment delimited by $x_1, x_2 \in \mathbb{R}^2$. Also, let $y = Ax + T$ be the image of $x \in \mathbb{R}^2$ under the affine transformation determined by the deformation matrix A and the translation vector T. From standard analytical geometry

$$x \in \left(\overline{x_1, x_2}\right) \Leftrightarrow \exists \alpha \in [0,1] \bullet x = \alpha x_1 + (1-\alpha) x_2$$

With these notations, the image of any interior point can be successively rewritten as follows:

$$\begin{aligned} y &= Ax + T \\ &= A\left[\alpha x_1 + (1-\alpha) x_2\right] + T \\ &= \alpha (Ax_1) + \alpha T + (1-\alpha)(Ax_2) + (1-\alpha) T \\ &= \alpha (Ax_1 + T) + (1-\alpha)(Ax_2 + T) \\ &= \alpha y_1 + (1-\alpha) y_2 \end{aligned}$$

This proves that, under any affine transformation of the Euclidean plane:

$$x \in \left(\overline{x_1, x_2}\right) \Leftrightarrow y \in \left(\overline{y_1, y_2}\right)$$

where $y = Ax + T$ is the image of point *x* under the affine transformation.

This proves that the coverage property of any line segment is maintained under any affine transformation, so the above assessment of tie-line consistency with respect to mass balance is independent of the coordinate system.

3. Comparison of experimental data herein with available literature data

As mentioned before, for the water + ethanol + trichloroethylene and the isooctane + ethyl acetate + acetonitrile system, some previous liquid–liquid equilibrium data at ambient pressure exist in the literature, albeit at slightly different temperatures. **Figures S2-S5** represent graphical comparisons of our experimental results with past work, as indicated in the figure captions.

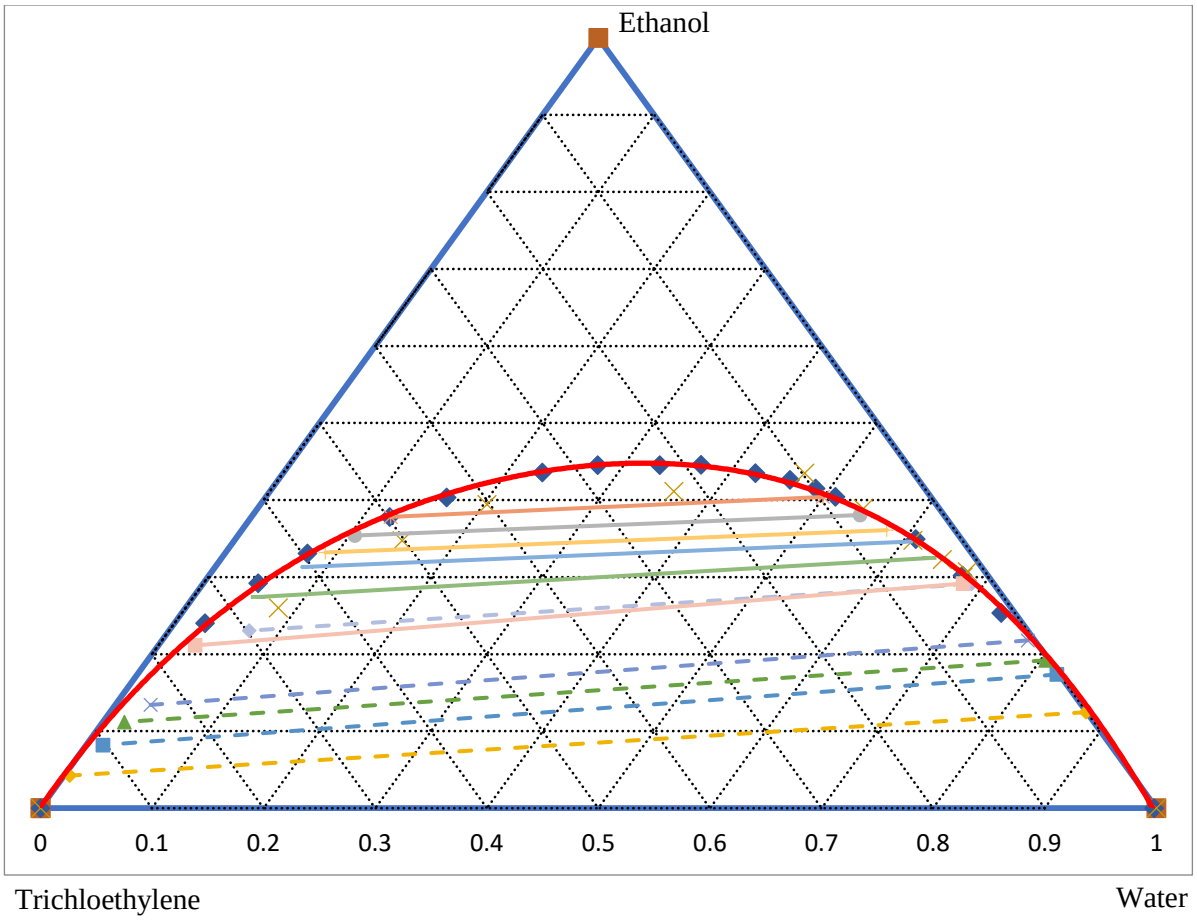

Figure S2. Experimental binodal curve (blue diamonds) and tie-lines (full color lines) for trichloroethylene + water + ethanol at 294.15 K compared with binodal curve data (mustard X's) and tie-lines (dashed color lines) for the same system at 294.65 K published by Hayden *et al.*[4]

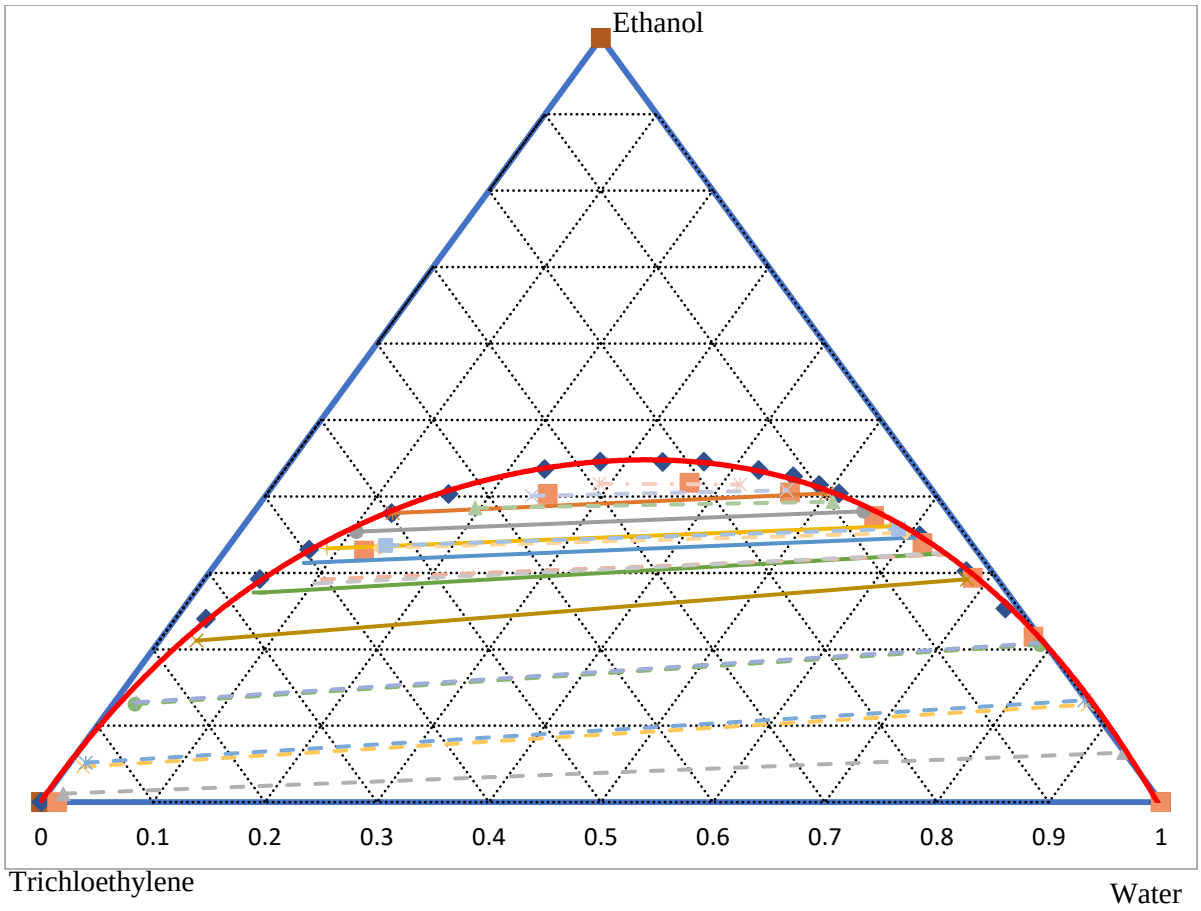

Figure S3. Experimental binodal curve (blue diamonds) and tie-lines (full color lines) for trichloroethylene + water + ethanol at 294.15 K compared with binodal curve data (brown full squares) and tie-lines (dashed color lines) for the same system at 293.15 K published by Reinders and de Minjer[5]

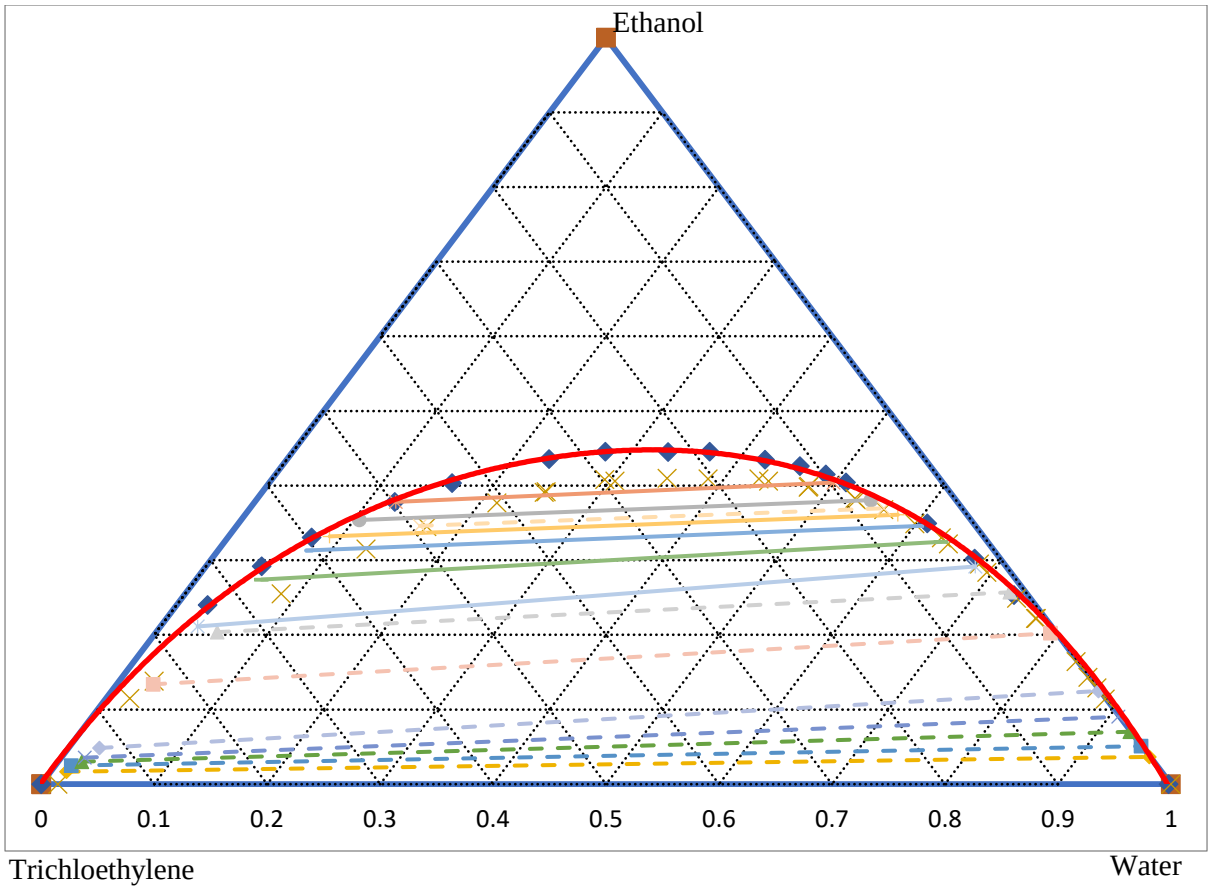

Figure S4. Experimental binodal curve (blue diamonds) and tie-lines (full color lines) for trichloroethylene + water + ethanol at 294.15 K compared with binodal curve data (mustard X's) and tie-lines (dashed color lines) for the same system at 298.15 K published by Colburn and Phillips[6]

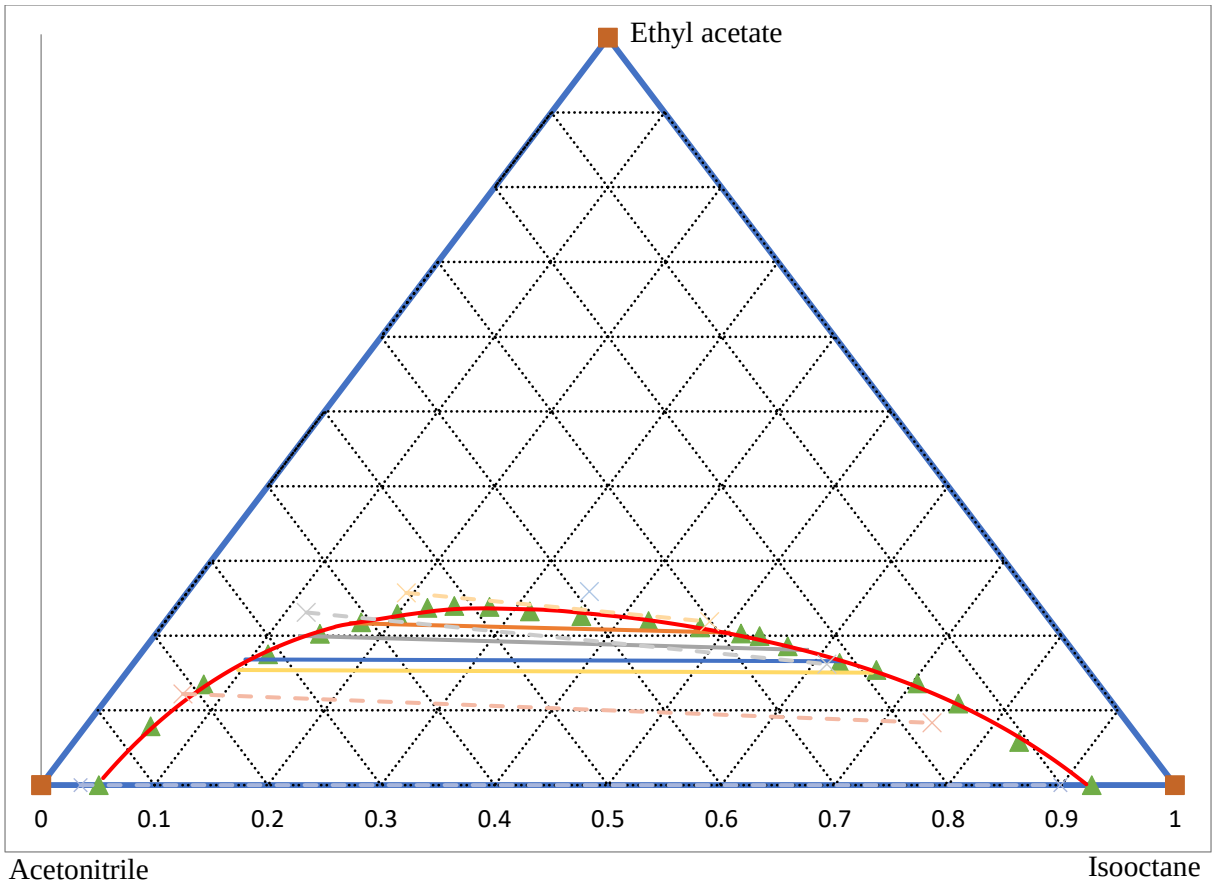

Figure S5. Experimental binodal curve (green triangles) and tie-lines (full color lines) for acetonitrile + isooctane + ethyl acetate at 294.15 K compared with tie-line data (colored X's) and tie-lines (dashed color lines) for the same system at 298.15 K published by Ooms *et al*.[7]

4. Conclusions

A mathematical model was developed based on Upchurch's[1] purely graphic procedure for primary data processing. This model provides:

- A greater degree of accuracy for the required calculations.
- A way of performing statistical inference on all computed quantities, as exemplified above for feed-to-tie-line distances for trichloroethylene +water + ethanol. It should be stressed that such calculations, although very important in assessing the significance of the results, are impossible in Upchurch's original graphic method.

Moreover, it was proven that mass-balance consistency assessment for tie-lines is invariant under affine transformations of the Euclidean plane, so it needs not be repeated after reverting to

Cartesian coordinates and is meaningless unless uncertainties due to experimental errors (noise) are taken into account.

References

(1) Upchurch, J. C.; Winkle, M. V. Liquid-liquid equilibria heptadecanol-water-acetic acid and heptadecanol-water-ethanol. *Ind. Eng. Chem*. **1952**, *44*, 618−621.

(2) Lahiri, S. N. Resampling methods for dependent data. *Springer series in statistics,* **2003**.

(3) Chernick, M. R. Bootstrap methods: A guide for practitioners and researchers, 2nd ed. *John Wiley & Sons,* **2008**.

(4) Hayden, N. J.; Diebold, J.; Noyes G. Phase Behavior of Chlorinated Solvent + Water + Alcohol Mixtures with Application to Alcohol Flushing. *J. Chem. Eng. Data* **1999**, *44,* 1085–1090.

(5) Reinders, W.; de Minjer, C. H. Vapor-Liquid Equilibria in Ternary Systems IV. The System Water-Ethanol-Trichloroethene. *Recl. Trav. Chim. Pays Bas* **1947**, *66,* 552-563.

(6) Colburn, A. P.; Phillips, J. C. Experimental Study of Azeotropic Distillation – Use of Tricholoethylene in Dehydration of Ethanol. *Trans. Am. Inst. Chem. Eng.* **1944**, *40,* 333-359.

(7) Ooms, T.; Steven Vreysen, S.; Baelen, G. V.; Gerbaud, V.; Rodriguez-Donis, I. Separation of ethyl acetate–isooctane mixture by heteroazeotropic batch distillation. *Chem. Eng. Res. Des*. **2014**, *92*, 995–1004.