A Dynamic Library for Physical and Thermodynamic Properties Correlations

Mordechai Shacham*

Department of Chemical Engineering, Ben Gurion University of the Negev, Beer-Sheva 84105, Israel

Neima Brauner

School of Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel

A dynamic physical properties (DPP) library of a new type of architechture is discussed. The library consists of a property evaluation program and a database which contains original experimental data and regression models, as well as statistical and graphical information for assessing their quality. The property evaluation program identifies "smooth" regions, where optimal regression models are fitted to the data, and discontinuities (transient regions), where statistically valid models cannot be obtained. The property evaluation program and the databases can be continuously updated to include the current state of the art regression models, experimental data, and regression techniques. The advantages of the DPP library structure over the traditional correlation libraries are demonstrated using an example of correlating heat capacity for a solid substance, where the experimental data includes smooth and transient regions.

1. Introduction

Correlation equations of physical and thermodynamic properties are being used extensively in process calculations and computations. Those correlations provide the property values as a function of process conditions, such as temperature, pressure, and composition. The correlations are developed usually by fitting regression models to experimental data.

Correlations are presented in forms which are intended to make their use most effective in process calculations. In the era of the slide-rule calculations, correlations were presented in the form of graphs and charts. Nowadays, they are usually presented as model equations, where the parameters are obtained by regression of experimental data. Such correlations are collected to form libraries of physical and thermodynamic properties. The libraries contain correlations for many different properties for a large number of chemical compounds. One of the first well-known libraries was published in the book of Reid et al.[1] Libraries that include more property correlations for many more compounds are available now (see, for example, Daubert and Danner[2]).

The existing correlation libraries have several serious limitations. They do not include all of the available information of the experimental data but rather the part that can be expressed in an acceptable accuracy by the correlation equation used in the particular library. Thus, when different regions in the data require different model equations for their representation, parts of the range where experimental data are available may be excluded from the correlation. Because of the time it takes to develop a large library, the standards that were set at the initial stages of the project (for model equations and regression algorithms that are being used and for the error estimation techniques that are employed) can be very far from the current state of the art in later stages of the same project.

Development of new, more accurate and stable regression models for various properties is an ongoing continuous process. Such a development is made possible by the availability of high-speed high-precision computers, where the regression model complexity becomes a secondary concern to the accuracy of the calculated properties, and also by the availability of new stepwise regression algorithms and regression diagnostic techniques. Improved measuring devices and measurement techniques enable one to obtain more accurate experimental data. Regression of more accurate data provides model parameters that represent the particular property with higher precision.

Unfortunately, the current correlation libraries cannot be regularly updated to include the most accurate and stable correlations for the various properties. The structure of those libraries is static, in the sense that there is one form of correlation (or a limited number of forms) for a particular property. The library cannot accept a new correlation of a different form, even if it is much more accurate. For example, if vapor pressure data are correlated with the Antoine equation, Wagner's[9] equation cannot be used for a particular compound, even when the latter is proved to be more accurate. The libraries do not contain the experimental data that were used for obtaining the regression parameters for the various correlations. Therefore, the user has to resort back to the original references of the experimental data (if those are still available) for assessing the accuracy of the library model in comparison with a new, potentially more accurate model.

In this paper, some of the shortcomings of the structure and philosophy of the existing libraries will be demonstrated, and a new concept of the "dynamic physical properties (DPP) library" will be introduced.

* To whom correspondence should be addressed. Phone: 972-7-6461481. Fax: 972-7-6472916. E-mail: shacham@bgumail.bgu.ac.il.

10.1021/ie990681a CCC: $19.00
Published on Web 04/27/2000

The DPP library contains a property evaluation program and databases of experimental data and regression models. The databases can be continuously updated and the property evaluation program can be revised, or even completely replaced, without altering the databases. The dynamic and flexible nature of the DPP library ensures that property values can be obtained in the full range where experimental data are available and that the most up-to-date correlation equations are used to calculate these values. In addition, the DPP library can serve as an archive of pertinent experimental data, making the data easily accessible even when the associated publications are not.

In section 2 of the paper some basic concepts related to optimal regression models are briefly reviewed. In section 3 the principles of the DPP library are described. In section 4 is an example that demonstrates the advantages of the DPP library over the current correlation libraries.

2. Basic Concepts—Definition of an Optimal Regression Model

The property evaluation program of the DPP library should provide optimal regression models for calculating the various properties. Following is a brief review of the basic concepts related to the definition of optimal regression models.

Most of the widely used property correlations include either linear or linearizable equations. A standard linear regression model can be written as

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + ... + \beta_n\mathbf{x}_n + \epsilon \qquad (1)$$

where $\mathbf{y}$ is an N vector of the dependent variable, $\mathbf{x}_j$ ($j = 1, 2, ..., n$) are N vectors of explanatory variables, β_0, β_1, ..., β_n are the model parameters to be estimated, and ϵ is an N vector of stochastic terms (measurement errors). It should be noted that an explanatory variable can represent an independent variable or a function of one or more independent variables.

A certain error (disturbance, imprecision, and noise) in the explanatory variables should also be considered. Thus, a vector of an explanatory variable can be represented by $\mathbf{x}_j = \tilde{\mathbf{x}}_j + \delta\mathbf{x}_j$, where $\tilde{\mathbf{x}}_j$ is an N vector of the expected value of $\mathbf{x}_j$, and $\delta\mathbf{x}_j$ is an N vector of stochastic terms due to noise. The errors in the dependent variable (ϵ) and the explanatory variables ($\delta\mathbf{x}_j$) cannot be measured but can be estimated. If estimates on the experimental errors are available, these can be used for $\delta\mathbf{x}_j$ and ϵ. Otherwise, it is usually assumed that the data are correct up to the last decimal digit reported. In such cases, the average rounding error can be used (approximately 3×10^{-t}, where t is the number of reported digits after the decimal point; see Stewart[3]). If functions of the independent variables are used or data transformation is carried out, the error propagation formula can be used to calculate the resultant $\delta\mathbf{x}_j$.

The vector of estimated parameters, $\hat{\boldsymbol{\beta}}^{\mathrm{T}}$, is usually calculated via the least-squares error approach, by solving the normal equation $\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$, where $\mathbf{X} = [1, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$ is an $N(n + 1)$ data matrix and $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{A}$ is the normal matrix. An alternative option is solving the overdetermined system of equations, $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}$ using QR decomposition (Press et al.[4]). The QR decomposition method requires more arithmetic operations than the solution of the normal equation but is known to be less sensitive to numerical error propagation (see Brauner and Shacham[5]).

From among the widely used statistical tests and criteria, the variance and confidence intervals on parameter estimates are used in this study for assessing the accuracy and stability of regression models. The sample variance $s^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)/(N - n - 1)$ is a measure of the quality of the fit. The confidence interval ($\Delta\beta_j$) on a parameter estimate is defined by

$$\beta_j = \hat{\beta}_j \pm \Delta\beta_j; \quad \Delta\beta_j = t(\nu,\alpha)\, s\sqrt{a_{jj}} \qquad (2)$$

where a_{jj} is the jth diagonal element of A^{-1}, $t(\nu,\alpha)$ is the statistical t distribution corresponding to ν degrees of freedom ($\nu = N - (n + 1)$) and a desired confidence level α, and s is the standard error of the estimate. Clearly, if $|\hat{\beta}_j| < \Delta\beta_j$, then the zero value is included inside the confidence interval and the parameter value is not significantly different from zero. A model which includes insignificant parameters is unstable. The most severe effects of instability are incorrect representation of the derivatives and absurd property values for even a small range of extrapolation (see, for example, Brauner and Shacham[10]). To eliminate the influence of the correlation between explanatory variables on the individual confidence intervals, confidence intervals on orthogonalized variables (no correlation between the variables) are used as indicators.

A statistically valid model is defined as a model where all of the parameter estimates are significantly different from zero (all $|\hat{\beta}_j| > \Delta\beta_j$ for orthogonalized variables). The optimal model is a statistically valid model which yields a minimal variance.

It should be noted that in some cases theory dictates a nonlinear regression model for concise representation of a particular property (instead of using series expansion, such as polynomial or quadratic representation). In such cases initial application of nonlinear regression to obtain the optimal parameters of the nonlinear model is recommended. Additional model terms can then be added, as necessary, considering the residual as the dependent variable.

3. Principles of a Dynamic Physical Properties (DPP) Library

The DPP library is envisioned as consisting of two parts: a property evaluation program (PEP) and a database containing experimental data and optimal regression models. The basic unit in the database is the "single compound−single property worksheet". The worksheet is divided into two sections: the data section and the correlation's section. In the data section, the raw experimental data, error estimates for both the independent and dependent variables, and references for the data and error estimates are stored. It should be emphasized that the original experimental data are kept without any smoothing or removal of outliers and other uncertain data. The error estimates are used by the PEP to determine the most accurate model for representing a particular property in view of the uncertainties in the data.

The correlation's section contains the optimal model for smooth regions and the pertinent experimental data for transient regions. For the optimal model of a particular region, the complete model definition including the parameter values, 95% confidence intervals, a plot of experimental data versus calculated values, and

a residual plot are included. The reference for the particular model (when it was developed and by whom) is also shown. For a transient region the pertinent experimental data and a plot of these data are displayed.

The user can directly access the worksheet of a particular property in order to assess the precision of the data or the available correlations. More frequently, the database will be accessed by the PEP in order to provide numerical point values, derivatives of a particular property, or an integral of the property over a range of the independent variable, according to requests from a calling program (i.e., a process simulator). The PEP may use the stored correlations for this purpose, or it may use the stored data to obtain new correlations. The algorithms for the operation of the PEP, for the case where new correlations are requested, follow.

Evaluation of Point Property and Derivatives.

1. Identify smooth and transient regions in the experimental data and set boundaries for the various regions. Store this mapping of data for future use.

2. Characterize the region where the property value or its derivative is requested. If this is a transient region, go to 4; otherwise, proceed to 3.

3. Fit an optimal (most accurate and stable) model to the data in the selected region. Store the optimal model for future use.

4. Calculate the requested value (property or its derivative) using the optimal model (in the case of a smooth region) or linear interpolation (in the case of a transient region).

5. If additional values in the same region are requested, go back to 4; if additional values in a different region are requested, go back to 2; otherwise, exit.

Evaluation of Integrals of the Property over a Space of the Independent Variables.

1. Identify smooth and transient regions in the experimental data and set boundaries for the various regions. Store this mapping of data for future use.

2. Characterize all of the regions that are included in the requested integral range. If no smooth regions are involved, go to 4; otherwise, proceed to 3.

3. Fit optimal (most accurate and stable) models to the data in the selected smooth regions. Store the optimal models for future use.

4. Calculate the integral by integrating the optimal model, using symbolic manipulation to obtain the analytical expression, if possible (in the case of a smooth region), or via interpolation (in the case of a transient region).

5. If additional integrals involving the same regions are requested, go back to 4; if additional integrals in different regions are requested, go back to 2; otherwise, exit.

If stored correlations are used, only the fourth and fifth steps of the two algorithms are executed.

It should be noted that during construction of the optimal model the PEP can exclude one or more insignificant independent variables from the model. If differentiation or integration with respect to such a variable is requested, a warning message is issued by the program.

Descriptions of the algorithms for the selection of the optimal regression model and the identification of smooth and transient regions follow.

Selecting the Optimal Regression Model Using the SROV Procedure.[6] For fitting the optimal model in the smooth regions, the use of the SROV procedure is proposed. This procedure is a stepwise regression program based on orthogonalized variables. It uses an initial pool of explanatory variables (the independent variables and/or their functions) to select the ones that should be included in the optimal model (optimal in the sense described in section 2) and to calculate the respective parameter values. The same procedure also yields various indicators that can identify the dominant cause preventing the addition of more variables to the model, thus limiting its precision.

The SROV procedure is described in detail by Shacham and Brauner.[6] In this procedure, the selection of a new variable to enter the model is based on three indicators: a correlation indicator (YX_j), a collinearity indicator (TNR_j), and an indicator which measures the signal-to-noise ratio in the correlation indicator (CNR_j). The SROV procedure consists of successive phases, where in the first phase an initial (nearly optimal) solution is found. In the subsequent phases the variables are rotated in an attempt to improve the model. Every phase of the procedure consists of successive stages, where at each stage one of the explanatory variables is selected to enter the regression model as an additional variable (basic variable). The remaining explanatory variables (nonbasic variables) and the independent variables are updated by subtracting the information which is collinear with the variables already included in the model. This updating generates nonbasic variables and a residual of the dependent variable, which are orthogonal to the basic variables set.

At each stage, the strength of the linear correlation between an explanatory variable $\mathbf{x}_j$ and a dependent variable $\mathbf{y}$ is measured by $YX_j = \mathbf{y}^T\mathbf{x}_j$, where $\mathbf{y}$ and $\mathbf{x}_j$ are centered and normalized to a unit length. The value of $|YX_j|$ is in the range [0, 1]. In the case of a perfect correlation between $\mathbf{y}$ and $\mathbf{x}_j$ ($\mathbf{y}$ is aligned in the $\mathbf{x}_j$ direction), $|YX_j| = 1$. In case $\mathbf{y}$ is unaffected by $\mathbf{x}_j$ (the two vectors are orthogonal), $YX_j = 0$. The inclusion of a variable $\mathbf{x}_p$, which has the highest level of correlation with $\mathbf{y}$ (YX_p value is the closest to 1) in the basic set, will affect the maximal reduction of the variance of the regression model. Therefore, the criterion $\mathbf{x}_p = \mathbf{x}_j\{\max |YX_j|\}$ is used to determine which of the nonbasic variables should preferably be included in the regression model at the next stage, provided that the $CNR_p > 1$ and $TNR_p > 1$ tests are both satisfied. The addition of new variables stops when for all of the nonbasic variables either $CNR_j > 1$ or $TNR_j > 1$, which indicates that the information included in these variables and/or in the residual of $\mathbf{y}$ is already at the experimental noise level.

Identification of Smooth Regions, Transient Regions, and Discontinuities. There are many instances where a particular property cannot be accurately represented by a single model equation over the whole range of interest. Typical examples include properties of solids where abrupt changes in properties may occur because of phase transitions between various possible crystal structures. The change may be gradual, in which case the transitional region can be represented by a statistically significant model (albeit different

Table 1. Worksheet Section Describing Part of the Solid Heat Capacity Data of HBr

name:	hydrogen bromide, hydrobromic acid; HBr
property:	solid heat capacity
reference:	Giaugue and Wiebe[7]
notation:	T = temperature (K)
	C_p = heat capacity (cal/mol·K)
	del T = estimated error in T
	del C_p = estimated error in C_p
data:	no. of points: 70

no.	T	C_p	del T (K)	del C_p (%)
1	15.72	1.831	0.05	0.30
2	17.81	2.16	0.05	0.30
3	19.57	2.615	0.05	0.30
4	22.32	3.01	0.05	0.30
5	25.49	3.459	0.05	0.30
6	30.16	3.955	0.05	0.30
7	34.58	4.415	0.05	0.30
8	39.15	4.827	0.05	0.30
9	43.75	5.16	0.05	0.30
10	48.32	5.453	0.05	0.30
11	52.93	5.832	0.05	0.30
12	57.8	6.171	0.05	0.30

models may be required for representing adjacent regions), or it may occur abruptly, introducing discontinuities and inconsistent changes in the property values. A region where the property changes are inconsistent with theory or with accepted models is herein denoted a "transient region", whereas regions that can be represented by statistically valid models (that are appropriate for that particular property) are denoted as "smooth regions".

The SROV procedure can be helpful in identifying the boundaries of the various regions, because inclusion of a data point belonging to a different region will usually seriously impair the quality of the fit between the optimal model and the data. Typically, this impairment will show up as a large increase in the variance and in the confidence intervals, a considerable increase in the number of terms included in the optimal model, and a nonrandom residual plot. To use the SROV procedure effectively for this purpose, a preliminary identification of the potential boundary points is required. This identification can be done by inspection of the plot of the experimental data or by use of the "difference from a moving average" (DFMA) technique (see below) to identify data points belonging to different regions.

In the DFMA procedure, the difference of each data point (dependent variable) from the moving average (the average of two adjacent points) is calculated. All data points for which this difference is larger (in absolute value) than the average value of the differences are suspected as being boundary points, belonging to transient regions or being outliers. Following this preliminary identification, the SROV procedure is used for final determination of the topology of the data set.

In the next section an example will be presented. This example demonstrates the use of the DFMA and SROV procedures for identification of the various regions and for fitting of the optimal regression models in smooth regions. The example also demonstrates the proposed library structure and some of its advantages. The calculations related to this example were carried out with the regression program of the POLYMATH 5.0 [copyrighted by M. Shacham, M. B. Cutlip, and M. Elly (http://www.polymath-software.com)] package. The SROV procedure was implemented with MATLAB 5.2 [trademark of MathWorks Inc. (http://www.mathworks.com)]. The database of the library was implemented with Excel [trademark of Microsoft Corp. (http://www.microsoft.com)].

4. Correlation of Heat Capacity Data of Solid HBr—An Example

Data of heat capacity versus temperature of solid HBr, which was published by Giauque and Wiebe,[7] are used in this example. The original data were sorted in an increasing order of the temperature values in order

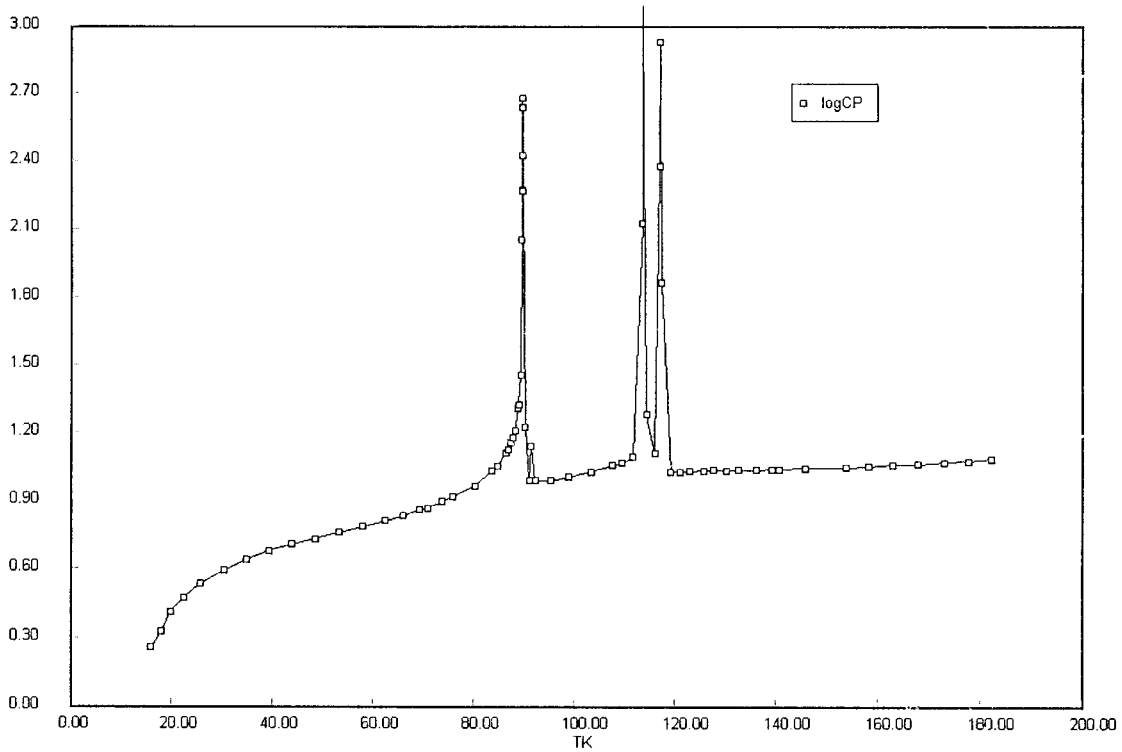


Figure 1. Heat capacity of solid HBr (C_p on a logarithmic scale).

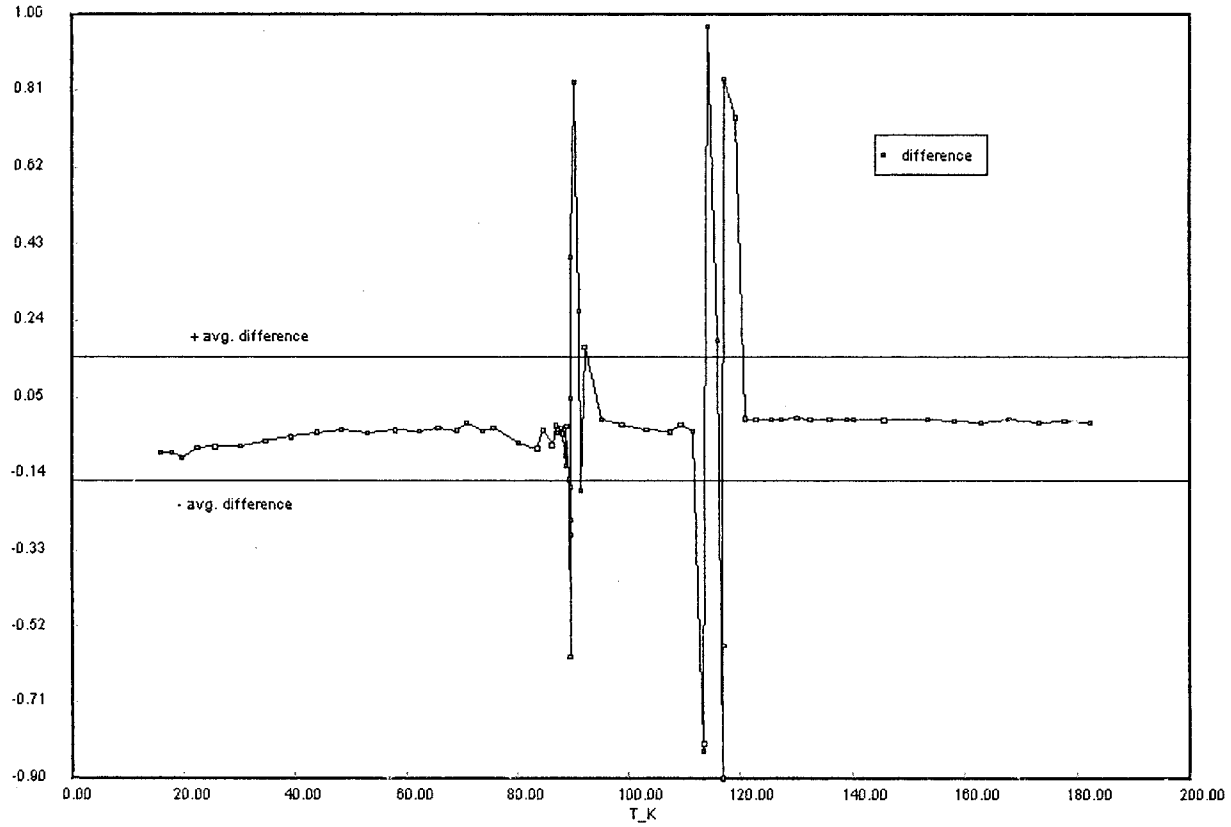


Figure 2. Difference between point values and moving average values for the heat capacity data.

to make the analysis easier. The section of the single component−single property worksheet, where part of the data are listed, is shown in Table 1. The precision of the temperature measurements (as reported by Giauque and Wiebe[7]) is $\delta T = 0.05$ K for all of the data points. The estimated error of the heat capacity data for most of the points is $\epsilon = 0.3\%$; for a few points (not shown in Table 1) there is an uncertainty regarding the precision of data.

In Figure 1 the heat capacity is plotted versus the temperature, where a logarithmic scale is used for heat capacity. Solid HBr exhibits two phase transitions, which are associated with discontinuities in the heat capacity versus temperature curve, as shown in Figure 1. This plot can be used for a preliminary identification of the boundaries of the various regions just by inspection. Alternatively, the DFMA procedure can be used for an automatic preliminary identification. In Figure 2 the calculated difference of each data point from the moving average is plotted versus the temperature. The two straight lines in this plot represent positive and negative values of the averaged absolute differences (=0.154). All points lying outside these two bounds are presumed as excluded from the current smooth region.

The descriptions of the five regions that were identified (by both inspection and the DFMA procedure) are shown in Table 2. Further validation of the region boundaries and the characterization of a particular region (smooth or transient) using the SROV procedure will be discussed in conjunction with the identification of the appropriate model for the various regions.

Discussion of the construction of the optimal regression models for the various regions follows.

Construction of the Optimal Regression Model for Region 5. Region 5 is the only region that has been included in the correlation library published by Daubert and Danner.[2] They have recommended the use of a fourth-order polynomial to model the data in this region. Shacham and Brauner[8] investigated collinearity considerations in polynomial regression using these data. We will be using the algorithm proposed by Brauner and Shacham[5] to construct the optimal polynomial model for the data. Using this algorithm, the temperature data are first transformed to yield a variable distribution in the range of [−1, 1] [via the z transformation, $z = (2x - x_{max} - x_{min})/(x_{max} - x_{min})$] and the heat capacity data are normalized by dividing all of the values by the maximal value in the range. Various powers of z (up to the 15th power) are used as an initial pool of explanatory variables that can potentially be included in the regression model. The SROV procedure is used to identify the explanatory variables that should eventually be included and to calculate the respective model parameters.

Table 2. Description of the Regions of the Heat Capacity Data

region no.	point range	temperature range (K)	type	est. error in T (δT; K)	est. error in C_p (ϵ; %)
1	1−29	15.72−89.23	smooth	0.05	0.30
2	30−38	89.39−91.23	transient	0.05	uncertain
3	39−45	92.01−111.47	smooth	0.05	0.30
4	46−52	113.31−117.09	transient	0.05	uncertain
5	53−70	118.99−182.09	smooth	0.05	0.30

In Figure 3 the section of the worksheet describing the optimal regression model is shown. The description includes the model equations, the parameter values (including 95% confidence intervals), and the variance. The model includes nonconsecutive polynomial terms z, z^2, and z^5 and a free parameter. The stability of the model is implied by its statistical validity (all of the confidence intervals are smaller than the respective parameter values). The model description includes also

Name:	Hydrogen Bromide, Hydrobromic Acid; HBr			
Property:	Solid Heat Capacity			
Region 5 (smooth)	Range:	118.99 K - 182.09 K		
by:	Shacham, M. and Brauner, N.	Date:	8.23.1999	
Notation:	T - Temperature (K)			
	Cp - Heat capacity (cal/mol-K)			
Model:	CPn=a0 + a1*z + a2*z^2 + a3*z^5			
	Cpn=Cp/12.32	z=(2*T-182.09-118.99)/(182.09-118.99)		

Parameters		
Parameter	Value	95% conf. Int.
a0	0.914182	0.002337
a1	0.05035	0.004052
a2	0.02428	0.004308
a3	0.012245	0.006204
Variance	7.63E-06	

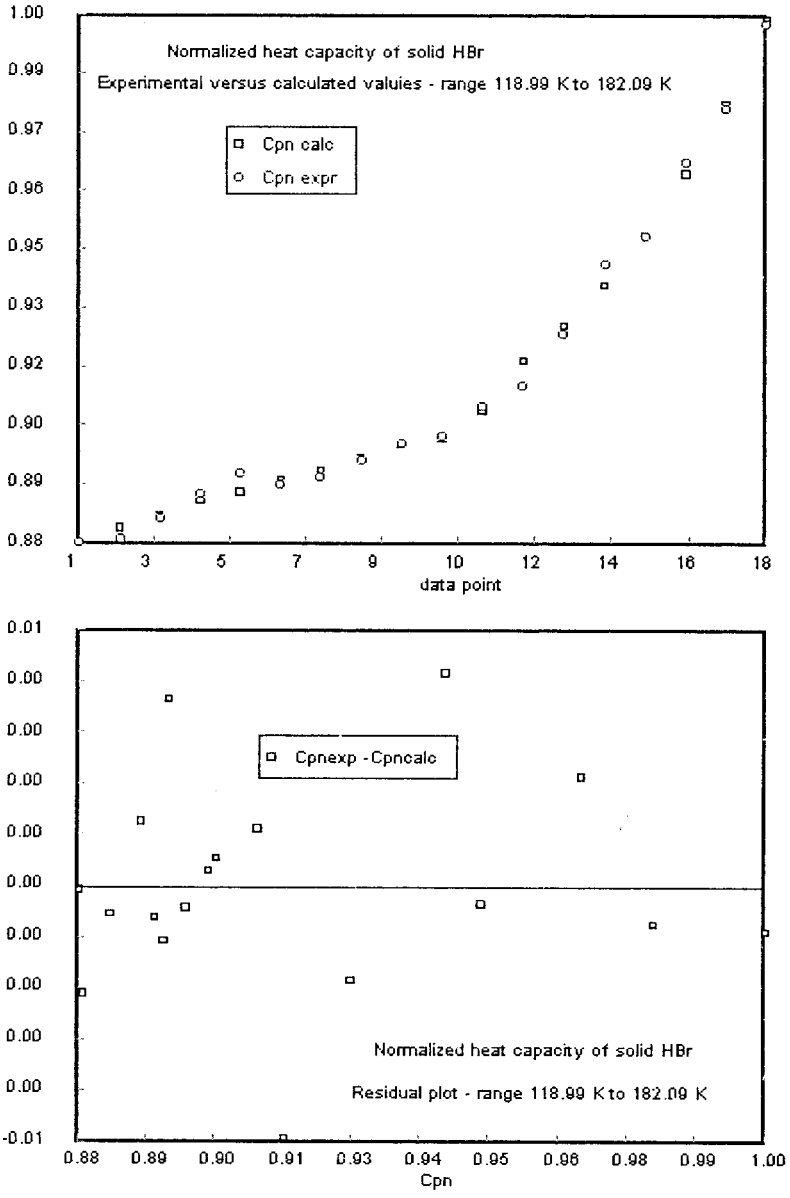


Figure 3. Section of the worksheet describing the optimal model for region 5.

a plot of the experimental data points along with the calculated data points and the residual plot. As shown, the residuals are randomly distributed, the maximal error is about 0.66%, and the average error is about 0.3%, the same as the estimated experimental error in the heat capacity data. Thus, the fit is satisfactory. Daubert and Danner[2] estimated the accuracy of their correlation (obtained for part of region 5) to be ±3%, which is significantly less accurate than the herein proposed model.

The form of the model description as shown in Figure 3 is appropriate for dual use. A computer program for carrying out additional calculations can import the definition of the model equations, the parameter values, and the bounds on the range of its applicability. In addition, a user can directly access the model definition for assessing the precision of the experimental data and the available correlations.

To verify that the bound of this region is indeed at $T = 118.99$ K (as found in the preliminary analysis and shown in Table 2), the SROV procedure was applied to the data of this region including an additional point (at $T = 117.09$ K). In this case, an optimal (statistically valid) model containing seven nonconsecutive polynomial terms—z^5, z^7, z^9, z^{11}, z^{12}, z^{13}, z^{14}, and z^{15}—has been found, with a variance of 4.372×10^{-5}. Thus, with the inclusion of the additional point, the number of terms in the model has increased from four to eight, and

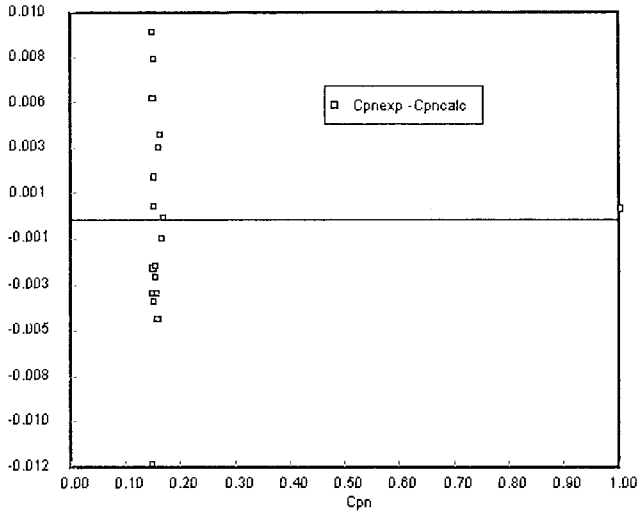


Figure 4. Residual plot of the optimal model in region 5 when the point at $T = 117.09$ K is included.

at the same time the variance has increased by almost an order of magnitude. The residual plot for this case, shown in Figure 4, demonstrates most dramatically that the point added does not belong to the same population as the rest of the data in this region.

Construction of the Optimal Regression Model for Region 1. An attempt to fit a polynomial model to the data in region 1 may yield poor results, as can be seen in Figure 5, where a fifth-order polynomial has been fitted to the data. The model is statistically invalid, because some of the confidence intervals are larger than the (absolute) parameter values and the variance is very large (see Table 3). Using a large pool of polynomial terms and the SROV procedure to select the terms that should be included in an optimal model yields more accurate results. Still, the model does not provide a random error distribution, implying that it can be further improved.

Table 3. Fifth-Order Polynomial Representation of Heat Capacity versus Temperature in Region 1 (Model: $C_p = a_1TK + a_2TK^2 + a_3TK^3 + a_4TK^4 + a_5TK^5$)[a]

variable	value	95% confidence interval
a_1	0.888 226 7	0.891 986 9
a_2	−0.076 540 7	0.077 662 8
a_3	0.002 593 7	0.002 316 8
a_4	-3.609×10^{-5}	2.862×10^{-5}
a_5	1.762×10^{-7}	1.253×10^{-7}

[a] Variance = 4.568 094 2.

The exponential increase of C_p values, starting at about $T = 85$ K, can be moderated by fitting a curve to $\log(C_p)$ instead of C_p itself. Employing the SROV procedure on normalized values of $\log(C_p)$ versus z yields the results shown in Figure 6. The optimal model includes 10 nonconsecutive polynomial terms—z, z^4, z^5, z^6, z^8, z^{10}, z^{12}, z^{13}, z^{14}, and z^{15}—and a free parameter. While this model seems to contain excessively large number of terms, it is statistically valid and stable, because all of the confidence intervals are smaller than the respective parameter values.

Figure 6 includes also a plot comparing the experimental data points for region 1 with the calculated values. The fit looks satisfactory. The residuals plot for the optimal model shows that the error is randomly distributed and the maximal error is about 1% in $\log(C_p)$ and about 2% in C_p values. Thus, the average error in this case is considerably higher than the estimated experimental error, implying that there may be a more appropriate model of higher accuracy.

The results for this example demonstrate that in some cases regression models with a large number of terms should be used in order to achieve a high level of accuracy. Considering that the level of complexity is not a major issue, when the model equations and the parameter values are directly imported by other programs, there is no need to refrain from the use of

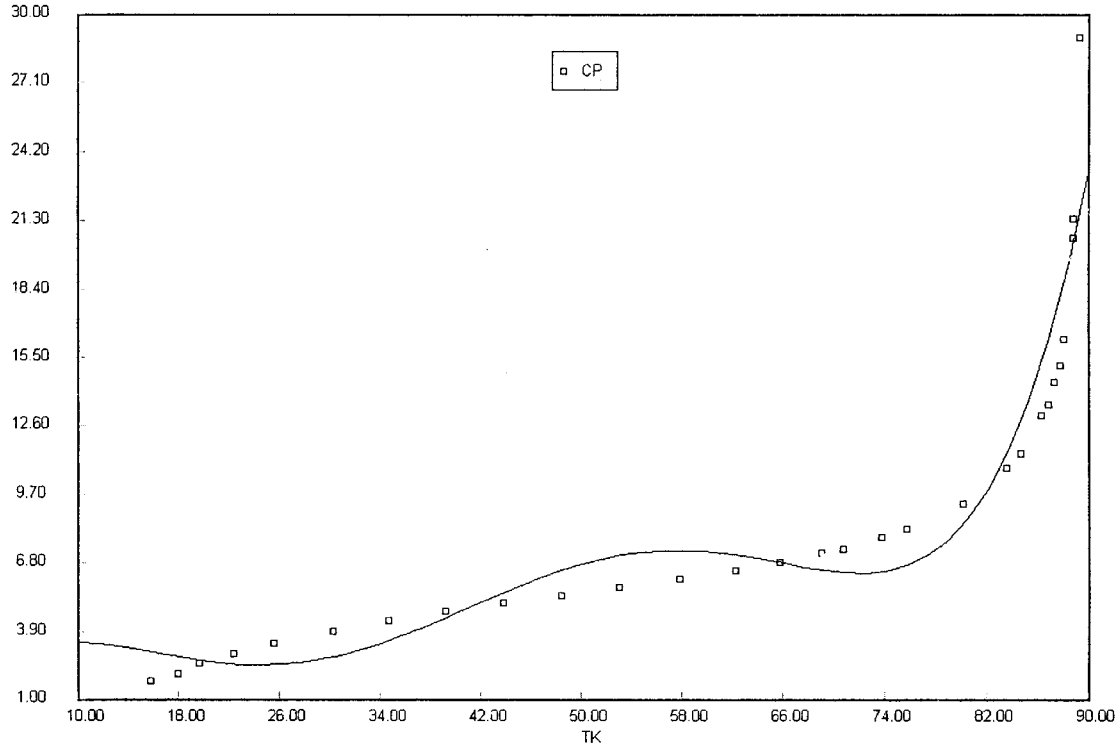
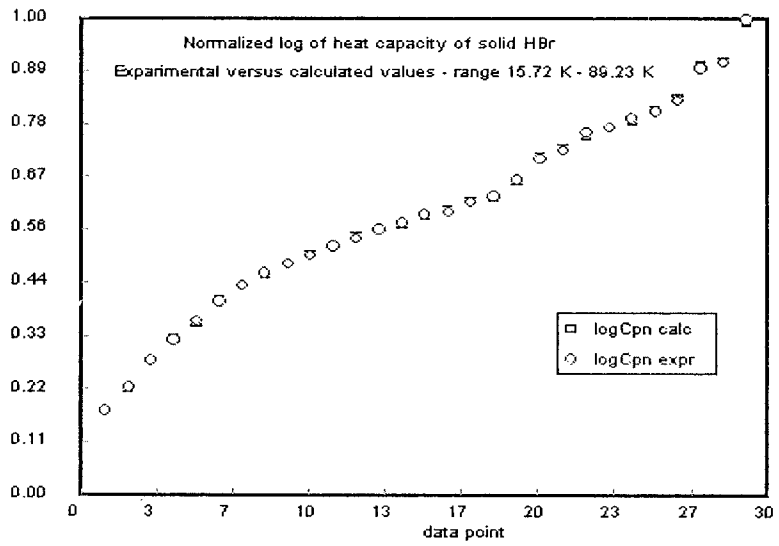


Figure 5. Experimental values versus calculated curve, region 1, fifth-order polynomial representation.

Name:	Hydrogen Bromide, Hydrobromic Acid; HBr				
Property:	Solid Heat Capacity				
Region 1 (smooth)	Range:	15.72 K - 89.23 K			
by:	Shacham, M. and Brauner, N.		Date:	8.23.1999	
Notation:	T - Temperature (K)				
	Cp - Heat capacity (cal/mol-K)				
Model:	logCPn = a0 + a1*z + a2*z^4 + a3*z^5 + a4*z^6 + a5*z^8 + a6*z^10 + a7*z^12 + a8*z^13 + a9*z^14 + a10*z^15				
	logCPn=log (Cp)/log(29.08)		z=(2*T-89.23-15.72)/(89.23-15.72)		

Parameters		
Parameter	Value	95% conf. Int.
a0	0.521899	0.004884
a1	0.145115	0.009377
a2	-1.40286	1.341462
a3	0.219044	0.052508
a4	14.67631	11.53583
a5	-58.5728	38.81345
a6	109.2208	62.90551
a7	-95.9653	49.10486
a8	-0.76847	0.325016
a9	32.10905	14.8174
a10	0.81156	0.284009
Variance	2.03E-05	



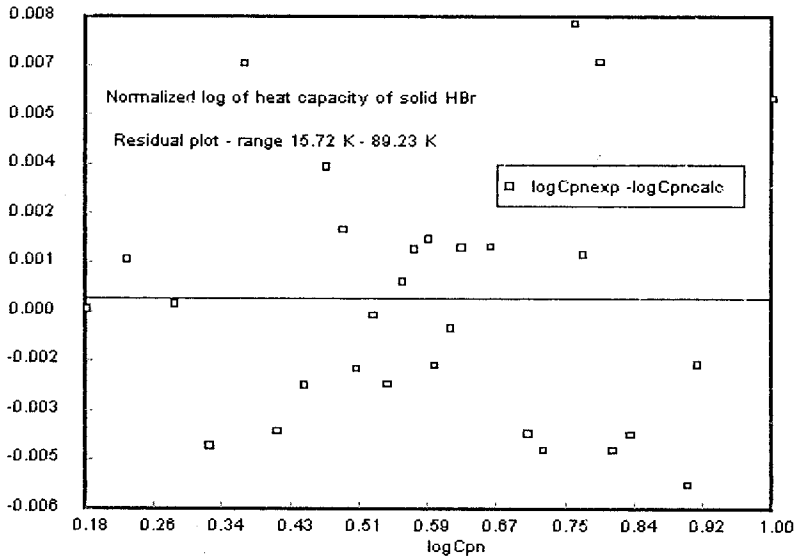


Figure 6. Section of the worksheet describing the optimal model for region 1.

complex models as long as they are statistically valid and of high accuracy.

Analysis of the Data in a Transient Region (Region 2). Figure 7 shows the section of the worksheet describing the pertinent information for a transient region (region 2). The description includes the experimental data, its range, and plot of the heat capacity versus temperature in this region. An attempt to fit a model to these data has indicated that there is no statistically significant polynomial model representation in this region, implying that a high level of uncertainty is associated with calculations involving such a region. One of the working assumptions that can be adopted is that the measured data points are more accurate than any (statistically insignificant) model fitted to the data. Based on this assumption, the use of linear interpolation for calculations involving such regions can be justified.

Regions 3 (smooth) and 4 (transient) are basically not different from the regions that were discussed so far, and therefore their analyses are not presented.

5. Conclusions

The proposed DPP correlations library can provide more information and of higher accuracy than the existing traditional libraries. The information in the DPP library is easily accessible and updateable. Higher accuracy is achievable by identifying and separately

Name:	Hydrogen Bromide, Hydrobromic Acid; HBr			
Property:	Solid Heat Capacity			
Region 2 (transient)	Range:	89.39 K - 91.23 K		
by:	Shacham, M. and Brauner, N.		Date:	8.23.1999
Notation:	T - Temperature (K)			
	Cp - Heat capacity (cal/mol-K)			
Model:	No statistically valid model found.			
	Use linear interpolation			
Data	No. of points: 9			
No.	**T**	**Cp**		
1	89.39	114		
2	89.45	190		
3	89.49	268		
4	89.5	483		
5	89.53	437		
6	89.59	187		
7	90.01	17.02		
8	90.91	9.87		
9	91.23	14.09		

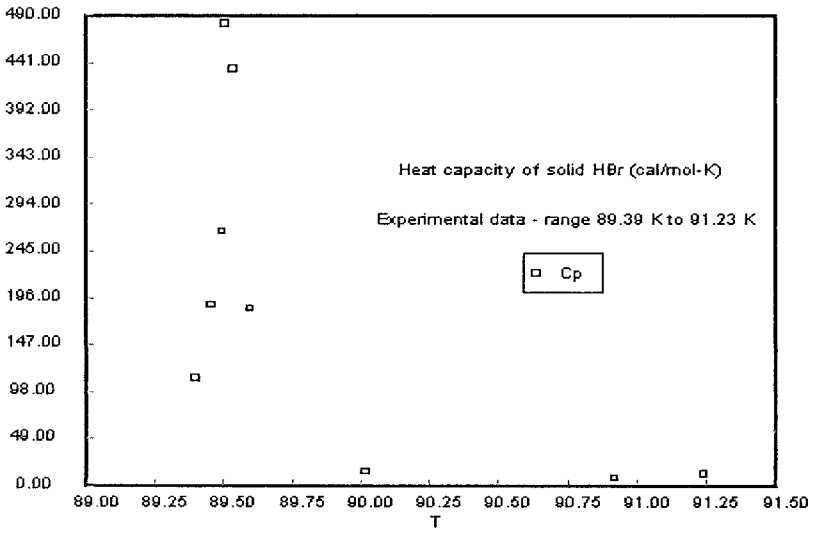


Figure 7. Section of the worksheet presenting the data of region 2.

treating smooth and transient regions and by using state of the art regression techniques for modeling the data in the smooth regions. Thus, the most accurate representation of the data is provided for the whole range where data are available.

The user is provided with the information needed to analyze the appropriateness of the suggested correlations for his purposes and may request the generation of new correlations with different model equations as necessary. The plots included in the library can show the user the true effects of extrapolation outside the smooth regions, so that he can select a more sensible calculation method if extrapolation leads to absurd results.

The whole set of experimental data and the model equations of the various regions (including the parameter values) are available for import by other programs. Because there is no need to manually copy equations and numbers, the complexity of the regression model becomes a minor concern. Therefore, it is no longer necessary to rely on simple, compact models, which may sacrifice the accuracy that is available in the experimental data. Complex, more precise models can be used as long as they are proven to be stable.

The library can be easily updated by copying the single component−single property worksheet and introducing the necessary changes: addition of new data, fitting of a different model, or employment of a more up to date regression package, to obtain a more precise correlation in a new copy of the worksheet.

The proposed library structure fits very well into the current trend of using "open system architecture" in software development in general and in process simulation in particular. Thus, it can be expected that this structure will replace the traditional structure in the development of correlation libraries.

Literature Cited

(1) Reid, R. C.; Prausnitz, J. M.; Sherwood, T. K. *The Properties of Gases and Liquids*, 3rd ed.; McGraw-Hill: New York, 1977.

(2) Daubert, T. E.; Danner, R. P. *Physical and Thermodynamic Properties of Pure Chemicals: Data Compilation*; Hemisphere Publishing Co.: New York, 1989.

(3) Stewart, G. W. Collinearity and Least Squares Regression. *Stat. Sci.* **1987**, *2*, 68−100.

(4) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flanerry, B. F. *Numerical Recipes in FORTRAN, The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: Cambridge, U.K., 1992.

(5) Brauner, N.; Shacham, M. Considering Error Propagation in Stepwise Polynomial Regression. *Ind. Eng. Chem. Res.* **1999**, *38* (11), 4477−4485.

(6) Shacham, M.; Brauner, N. Considering Precision of Experimental Data in Construction of Optimal Regression Models. *Chem. Eng. Process.* **1999**, *38*, 477−486.

(7) Giaugue, W. F.; Wiebe, R. The Heat Capacity of Hydrogen Bromide from 15 K to its Boiling Point and its Heat of Vaporization. *Am. Chem.* **1928**, *50*, 2193−2203.

(8) Shacham, M.; Brauner, N. Minimizing the Effects of Collinearity in Polynomial Regression. *Ind. Eng. Chem. Res.* **1997**, *36* (10), 4405−4412.

(9) Wagner, W. New Vapor Pressure Measurements for Argon and Nitrogen and a New Method for Establishing Rational Vapor Pressure Equations. *Cryogenics* **1973**, *13*, 470.

(10) Brauner, N.; Shacham, M. Role of Range and Precision of the Independent Variable in Regression of Data. *AIChE J.* **1998**, *44* (3), 603−611.

Received for review September 7, 1999
Revised manuscript received January 21, 2000
Accepted January 24, 2000

IE990681A