ELSEVIER

Fluid Phase Equilibria 185 (2001) 21–29

FLUID PHASE EQUILIBRIA

www.elsevier.com/locate/fluid

Quantitative review and delivery of reliable physical property data: development of DIPPR® Environ 2001™ database and estimation software

A.A. Kline [a,*], C.R. Whitten [a], M.S. Heward [b], M.R. Trumbell [a], P.M. Wells [a], T.N. Rogers [a], D.A. Zei [a], M.E. Mullins [a]

[a] *Department of Chemical Engineering, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA*
[b] *EPCON International, 16360 Park Ten Place, Houston, TX 77084, USA*

Abstract

The goal of American Institute of Chemical Engineers Design Institute for Physical Property Data (AIChE DIPPR®) Project 911 has been to develop a comprehensive database of physical properties for chemicals that are regulated by various agencies of the United States government, and are important to the chemical process industry. Project 911 collects and quantitatively reviews environmental, safety and health (ESH) data for over 1000 chemicals and 56 physical properties. Project 912 analyzes and uses published estimation methods and develops new algorithms to generate predicted values where experimental data do not exist. Physical properties within Project 911 include aqueous solubility, octanol–water partition coefficients, vapor pressure, aquatic toxicity, bioconcentration factor, flash point, and activity coefficients at infinite dilution. Data are reviewed qualitatively for purity of chemicals and type of experiment, reported precision of measured data, and agreement with other investigators. An extensive quantitative review of the Project 911 database uses statistical quality control (SQC) techniques, where individual data points are compared to the highest rated data value from the qualitative review. The SQC review also tests data values using thermodynamic relationships. Recommended data values and estimation techniques are delivered to the user by a new Visual Basic™ software product, Environ 2001™. Results to date show an error rate of 1.5% for nearly 130,000 data values in the Project 911 database.

Keywords: DIPPR® Environ 2001™; Estimation software; Statistical quality control

1. Introduction

In order to refine the criteria for the quality assessment of the American Institute of Chemical Engineers Design Institute for Physical Property Data (AIChE DIPPR®) Project 911 environmental, health and safety (ESH) data compilation, Michigan Technological University (MTU) personnel have identified

* Corresponding author. Tel.: +1-607-254-6330; fax: +1-607-255-0305.
E-mail address: aakline@mtu.edu (A.A. Kline).

0378-3812/01/$20.00
PII: S0378-3812(01)00453-8

quantitative data checks that have been computerized as a statistical quality control (SQC) system. The goal has been to implement and use an SQC system that satisfies the quality assurance/quality control (QA/QC) needs required by the project steering committee and by the rest of the technical community. The computerized checks do not supplant the qualitative QA/QC procedures already in place, but are additional review mechanisms. The SQC system provides a method of assessing the error rate over the course of the project, allows the project team to characterize and categorize the errors, and helps to indicate the steps necessary to correct problems in our QA/QC system. To date, 130,000 data values in the Project 911 database have been evaluated. Recommended data values from Project 911 are available through a Windows™ format software product, Environ 2001™, commercially available from EPCON International. Environ 2001™ also contains estimation techniques developed under DIPPR® Project 912 (ESH estimation), to provide data where experimental values are not available. The ultimate goal is to develop, thoroughly review, and deliver a high quality ESH database and additional estimation techniques that can be used to support engineering and regulatory calculations.

2. Defining the SQC system

Two types of quantitative checks are performed on the Project 911 data values, as outlined below. Additional information on the Project 911 data review process is found in Kline et al. [1].

The "Level 1" internal consistency check is defined in terms of a range of deviation from the highest qualitatively rated data value for that chemical and physical property. The percent deviation (tolerance value) criterion for raising a "flag" is set according to the property of interest. Tolerance values for physical properties within the Project 911 database are found in Table 1.

The "Level 2" data evaluation method involves the comparison of data values for a given property code using a comparison to another property, an algebraic calculation involving one or more other properties, or other review techniques, as described below. Data values that successfully complete a Level 1 or 2 review are designated with a Q1 or Q2 in the Project 911 database, respectively.

2.1. Other review methods

Empirical checks, or "family plots" for inconsistencies in the range of values for a property among a homologous series are being used. The "smoothness" of the property value trends for increasing molecular weight within family groups or sometimes between closely related families can provide an indication of possible "outliers". Some properties as a function of carbon number are also being examined. This concept could be expanded to properties that are a function of temperature being evaluated at a specified temperature and plotted versus molecular weight. The use of family plots for comparing data has been especially useful in the oxygen demand properties, organic-carbon/water partition coefficient (Koc) [2], and bioconcentration factor (BCF) review.

Scatter plots have been used to plot related data against each other, and then a tolerance value is applied. The closer the plotted data lies to a 45° line, the better the agreement of the data with expected behavior. This technique was used for the comparison of log Koc and log Kow data, using a tolerance value of 30%. The scatter plots were used for comparison of some dissimilar chemical types, as well as within the individual family groups.

Table 1
DIPPR® project 911 database SQC review results

Property	Q1 criteria[a]	Total points	AF[b]	AD[c]	DE[d]	% DE
Biochemical oxygen demand (BOD, 1a)	1a ≤ 1cn	986	0	4	4	0.41
Dichromate chemical oxygen demand (COD, 1b)	1b ≤ 1cc	622	14	0	1	0.16
Permanganate chemical oxygen demand (COD, 1bp)	1bp ≤ 1b	146	0	0	0	0
Theoretical oxygen demand, carbonaceous (ThOD, 1cc)	1cc ≤ 1cn	864	4	0	3	0.35
Theoretical oxygen demand, combined (ThOD, 1cn)	1cc ≤ 1cn	594	0	0	1	0.17
Octanol/water partitioning, Kow	>20%	5077	389	12	21	0.41
Organic carbon/water partitioning, Koc	log Koc vs. log Kow Scatter plot; >30%	1348	38	5	0	0
Bioconcentration factor (BCF)	> ±1 logarithmic unit	2554	241	5	0	0
Molecular weight	>5%	14165	57	0	95	0.67
Liquid density at 25°C	>10%	7747	59	12	624	8.05
Solubility of chemical in water	>30%	7470	1036	167	287	3.84
Melting point	>10%	11933	192	16	201	1.68
Normal boiling point (NBP)	>10%	14580	246	12	66	0.45
Vapor pressure at 25°C	>50%	4933	522	29	78	1.58
Molecular diffusivity in air	>30%	1638	3	1	17	1.04
Molecular diffusivity in water	>30%	1517	22	4	32	2.10
Surface tension at 25°C	>10%	2448	133	20	33	1.35
Ideal gas heat of formation	>20%	3622	324	20	102	2.80
Critical temperature	>10%	6237	36	2	10	0.16
Critical pressure	>10%	5042	128	16	40	0.79
Critical volume	>10%	2733	30	8	14	0.51
Heat of vaporization at 25°C	>20%	1908	62	6	49	2.57
Heat of vaporization at NBP	>20%	2286	38	5	15	0.70
Activity coefficient of chemical in water	>50%	1227	321	10	10	0.82
Activity coefficient of water in chemical	>50%	344	16	6	1	0.29
Henry's law constant	>50%	2945	535	50	30	1.02
Lower flammability limit	>25%	3605	175	12	61	1.69
Upper flammability limit	>50%	3246	93	9	31	0.96
Flash point	>10%	6665	493	23	33	0.50
Autoignition temperature	>15%	3563	110	12	24	0.67
Net heat of combustion	>10%	2389	98	29	39	1.60
Total		124434	5425	495	1992	1.55

[a] Q1 percent difference for comparison of data value to "highest rated data value" for a particular chemical and physical property combination (Level 1 SQC review).
[b] AF: anomaly flagged: value is taken correctly from the literature; but does not agree with other data for the same chemical.
[c] AD: anomaly dropped: value is AF, and rating has been lowered to make a new highest rated.
[d] DE: data entry: data entry error that has been corrected.

3. Classification of errors

When performing an analysis using the SQC system, the "actual error" rate and "flag" rate must be carefully separated. Each data value flagged by the SQC software is examined and analyzed to determine the classification or type of error. All data that successfully passed the SQC check will have a "Q1" or "Q2" entered into the Project 911 database.

For those data values that do not pass the SQC criteria, the original article is reviewed so that an error code can be attached to the data value. Those values receiving an error code of "AF" (anomaly flagged) or "AD" (anomaly, rating dropped) are reviewed by an MTU investigator to determine their accuracy, and a recommendation made as to their disposition. Data values that are determined to be data entry mistakes will be flagged as "DE" (data entry error) and corrected in the Project 911 database.

Values that have been flagged with an "AF" code are values that are outside the tolerance range when compared to the highest qualitatively rated data value for a particular chemical and physical property, but the value has been correctly transcribed from the original literature reference. An anomaly flag of "AD" is used when a number of literature sources report data values that are in agreement with each other, but are outside the tolerance range when compared to the highest qualitatively rated data value. The original literature references are reviewed, and a determination is made as to whether a transcription error has occurred with the highest rated value, or whether different experimental conditions or techniques were used for the various literature sources. A reliability assessment is also made about the authors of the literature source, based on the experience of MTU investigators. If an error in transcription has occurred, the highest rated value is corrected and labeled "DE". If it is found that the highest rated value uses a less reliable experimental technique, or there are questions about the quality of work based on the authors, the highest rated value is labeled "AD", the qualitative numeric rating is lowered, and the SQC system is run again to check the data against the new highest rated value. The "DE", "AF", and "AD" codes are not displayed within the Environ 2001^{TM} software product, but are documented by the SQC tracking system.

Data values that are flagged with an error code of "DE" are rechecked by the SQC system during the subsequent SQC reviews. At that time, it is anticipated that the error will not be repeated, and the values will receive a rating of "Q1" or "Q2". An anomaly tracking form is kept with the database reference article, so that MTU staff have a complete record of any changes made to data values from a particular reference. The output files from the SQC system, which are a compilation of all errors identified when a particular quantitative criteria check has been run on a set of data, are logged and dated in a notebook and maintained in a file according to the physical property on which the SQC analysis was completed. The SQC output files can be cross referenced to the individual data errors on the anomaly tracking forms attached to each Project 911 database reference paper.

4. Results from the SQC review system

SQC results through the 1999 calendar year are found as Table 1. For each physical property listed, the total number of data values reviewed, the Level 1 tolerance value, and the number and type of error are given. As shown in Table 1, over 120,000 data values have been reviewed using the Level 1 tolerance values and other methods. The overall error rate for the Project 911 database is about 1.5%.

An unusually high number of data entry errors were found for the liquid density values. This was mostly due to misinterpreting tabulated data in various references. For example, data was displayed in an article with a notation at the top of the column that the value was "$\times$1000" or "$\times$1E-03". It is not always clear from different references whether this refers to the values in the table having been already multiplied by the factor at the top of a column, or if the user needs to multiply by this factor when using the data. Cross comparison of data from various sources using the SQC process allowed us to see which liquid density values were entered incorrectly due to this, and corrections were made to the Project 911 database. This

same type of multiplier for an entire column of tabulated data is used occasionally for other physical properties, but in our experience is most prevalent in the liquid density literature values.

From Table 1, it can be seen that some properties have a high number of "AF" codes relative to the total number of available data points. This is especially true for bioconcentration factor, activity coefficient of chemical in water, solubility of chemical in water, Henry's law constant, and flash point. As explained earlier, these "AF" codes are not errors. Rather, they illustrate the wide disagreement of experimental data values in published literature. As additional data are entered into the Project 911 database and SQC reviews continue, the number of "AF" codes will be reduced as MTU personnel will be able to designate a "highest rated value" with greater confidence. The "highest rated value" from the Project 911 database is available to the user in the Environ 2001^{TM} software product.

Over 4600 data values for aquatic toxicity properties are also part of the Project 911 database. Aquatic toxicity values do not use the previously described SQC techniques, but undergo extensive review and intercomparison at the time of data entry. This is due to the different methodology used to enter comments and experimental conditions into the Project 911 database for the aquatic toxicity physical properties. Additional information on the aquatic toxicity data review is available from MTU [3].

5. Review of temperature dependent properties

Because there are usually only one or two sets of temperature dependent coefficients in the Project 911 database for a particular physical property and chemical combination, it is not possible to intercompare them to determine a "highest rated" set of coefficients. As currently configured, the Project 911 database does not include the individual experimental data values used to build a data curve to which the appropriate temperature dependent equation form is regressed. As a result, an alternative method to review temperature dependent properties has been used, as described below. Results from reviewing the temperature dependent coefficients are found as Table 2.

Table 2
SQC results for temperature-dependent properties within the project 911 database

Property	Total points	AR[a]	RC[b]	DL[c]
Liquid density	462	0	0	9
Vapor pressure	595	2	14	8
Vapor viscosity	419	0	0	0
Liquid viscosity	417	1	0	10
Surface tension	391	0	2	2
Liquid thermal conductivity	407	0	9	10
Vapor thermal conductivity	367	0	0	10
Liquid heat capacity	452	2	15	28
Ideal gas vapor heat capacity	491	0	4	11
Heat of vaporization	382	4	1	13
Total	4383	9	45	101

[a] AR: anomaly re-regressed: original Project 911 data file re-regressed with additional data, new set of coefficients put in database.

[b] RC: range change: value of T_{min} or T_{max} changed from original.

[c] DL: deleted: set of coefficients deleted for various reasons.

5.1. Review methodology

The method for checking the temperature dependent coefficients within the Project 911 database is as follows:

1. A download of all available temperature dependent coefficients was exported to an MS Excel™ spreadsheet.
2. The applicable temperature dependent equation forms were applied to each set of coefficients. Temperature values between the defined minimum and maximum temperature (T_{min} and T_{max}) were used to generate plots.
3. The plots were studied for abnormalities. Apparent errors were flagged and the data rechecked for data entry errors or for other extraneous causes of error.

Sets of coefficients that behaved as expected were given a "Q1" rating in the Project 911 database. Those sets of coefficients that did not behave as expected were further reviewed using additional criteria

5.2. Further review of temperature dependent coefficients

All questionable sets of coefficients that originated at MTU were checked against the original regression plots for transcription errors. Additional experimental data were added, if available. The data were re-regressed using the appropriate equation form, and the new set of coefficients checked for expected curve shape behavior. If the new set of coefficients were superior to the set currently in the database, it was labeled with an "AR" (anomaly re-regressed), and the new set of coefficients replaced the older coefficients in the Project 911 database. There were a total of nine sets of coefficients with "AR" ratings, as shown in Table 2.

Some temperature dependent curves showed unexpected maxima, minima, or inflection points. These curves were reviewed, and the T_{min} and T_{max} values adjusted to eliminate inflection points in the valid correlation temperature range. Curves were generated using the new temperature range, and checked for expected behavior. Sets of coefficients that had temperature ranges adjusted were labeled with a "RC" (range change) code in the database. There were a total of 45 sets of coefficients with a rating of "RC", as shown in Table 2.

The final type of rating used on the temperature dependent coefficients is "DL", which means a set of coefficients were deleted from the database. These deletions occurred for various reasons, including:

1. $T_{min} = T_{max}$ and only a single coefficient, whereas other regressions used three to five coefficients. These values were not intended to be used as a set of temperature dependent coefficients, but were meant to be specific data points with comments, such as "enthalpy of sublimation" or "solid thermal conductivity".
2. Suspect sets of coefficients that originated at MTU, but the original data files used to regress coefficients could not be located. The coefficients were removed from the Project 911 database, and will be replaced at a later date when the data files are rebuilt, regressed, and evaluated.
3. Sets of coefficients that originated from other sources, such as DIPPR® Project 801, but curve shapes did not show expected behavior. These sets of coefficients are undergoing further evaluation.

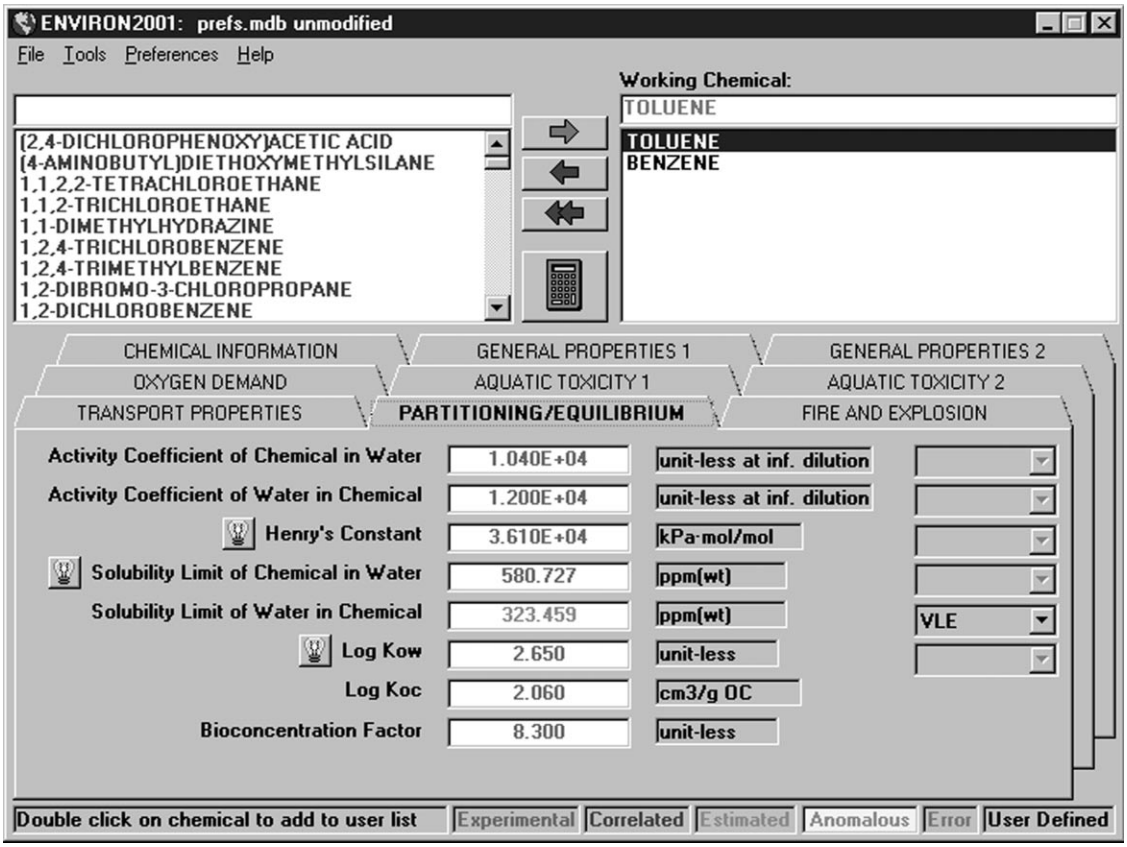

Fig. 1. Partitioning equilibrium for Toluene.

In all, 101 sets of temperature dependent coefficients were removed from the Project 911 database, as shown in Table 2, representing a 2.3% rate. As part of the continuing work under Project 911, additional sets of coefficients are being added as they are regressed from available experimental data.

6. Project 911 data delivery: Environ 2001™ software

After SQC review, recommended data values for the 56 physical properties within the Project 911 database and estimation methods from Project 912 are made available to the user in a Windows™ format software product. Environ 2001™ utilizes an MS Visual Basic™ interface and MS Access™ databases. An example screen is shown as Fig. 1, showing some of the available physical properties. The user is able to select amongst various data sources and estimation methods for each chemical and physical property combination (Fig. 2). Results of data searches can be exported to MS Excel™ spreadsheets, or printed as hard copy reports. Additional information about Environ 2001™ can be obtained from EPCON International [4].

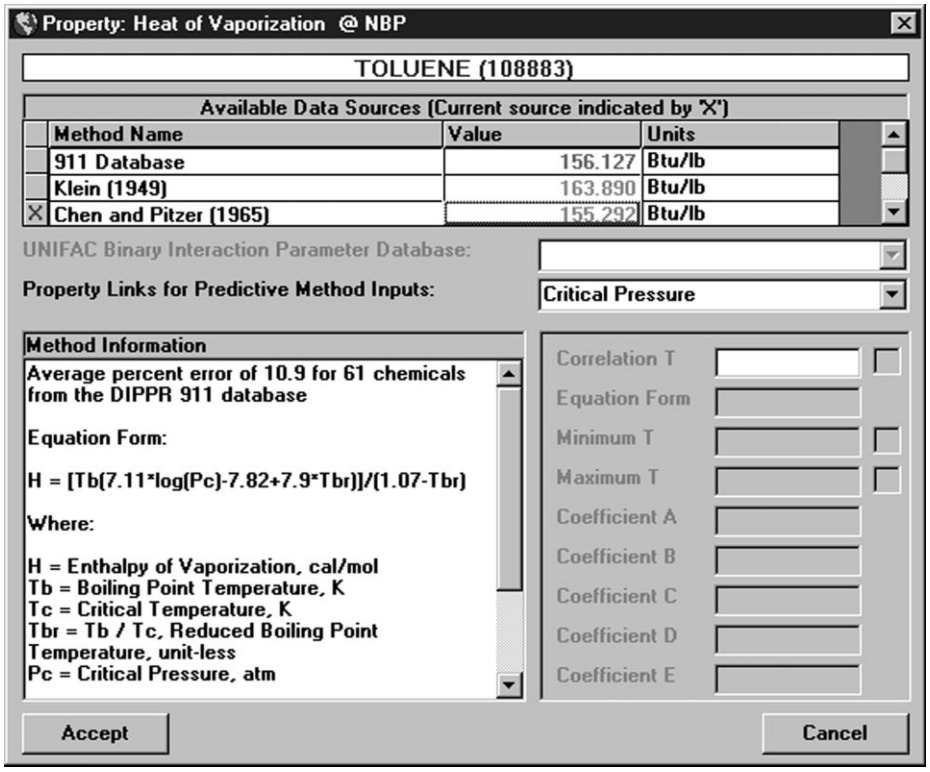

Fig. 2. Chen and Pitzer method information for heat of vaporization @ NBP.

7. Conclusions

To date, over 120,000 data values in the Project 911 database have undergone an SQC review. An additional 4383 sets of temperature dependent coefficients were reviewed using other techniques. The overall error rate for the Project 911 database is 1.5%. Liquid density had an unusually high error rate of 8%, due to misinterpreting the use of multipliers accompanying tabulated data values. Some physical properties have a large number of "AF" values relative to the total number of data points. These values are not errors in the database due to transcription, but instead reflect the wide range of data in the published literature for a particular property, such as activity coefficient of chemical in water, or flash point. The recommended data values from the DIPPR® Project 911 database and estimation techniques from Project 912 are accessible to the user through a new MS Windows™ based software product, Environ 2001™, available from EPCON International.

Acknowledgements

The authors wish to acknowledge continuing financial support from the Design Institute for Physical Property Data of the American Institute of Chemical Engineers (AIChE DIPPR®).

References

[1] A.A. Kline, C.R. Szydlik, T.N. Rogers, M.E. Mullins, Fluid Phase Equilib. 150/151 (1998) 421–428.
[2] J.R. Baker, J.R. Mihelcic, D.C. Luehrs, J.P. Hickey, Water Environ. Res. 69 (2) (1997) 136–145.
[3] A.A. Kline, D.A. Zei, C.R. Whitten, D.C. Luehrs, T.N. Rogers, E.V. Lutz, J.R. Mihelcic, S.D. Radecki, M.E. Mullins, J.C. Metsa, DIPPR® Project 911 Policy and Procedures Manual, Michigan Technological University, 1999.
[4] EPCON International, 16360 Park Ten Place, Houston, TX 77084, USA, www.epcon.com.