An Overview of Compiling, Critically Evaluating, and Delivering Reliable Physical Property Data from AIChE DIPPR® Projects 911 and 912[1]

A. A. Kline,[2,3] C. R. Szydlik,[2] T. N. Rogers,[2] and M. E. Mullins[2]

[1] Paper presented at the Thirteenth Symposium on Thermophysical Properties, June 22-27, 1997, Boulder, Colorado, U.S.A.

[2] Department of Chemical Engineering
Michigan Technological University, Houghton, MI 49931

[3] To whom correspondence should be addressed.

ABSTRACT

The goal of AIChE DIPPR® Projects 911 and 912 is to develop a comprehensive, consolidated database of physical properties for those chemical species which are regulated by various agencies of the U.S. government, and are important to the chemical process industry. Environmental, safety and health (ESH) properties are the prime focus of the data collection and validation efforts of the two projects. Project 911, a database compilation effort, collects data for 700 chemicals and 55 physical properties. These properties include aqueous solubility, viscosity, vapor pressure, flash point, octanol-water partition coefficient, and bioconcentration factor. Project 912 is a complementary effort which focuses on the review of existing physical property prediction techniques and developing new estimation algorithms where none exist. Limited mixture data (*e.g.*, infinite dilution vapor/liquid equilibrium measurements) are also being researched.

Work is continuing on the critical assessment of the quality of data within the Project 911 database. Available literature data are compiled and categorized according to quality. Recommended data values and correlation statistics are provided as part of the Project 911 software product. To automate the data evaluation effort, a computerized Statistical Process Control (SPC) data review system has been designed. The Project 911 database is being developed to support engineering and regulatory calculations and to work in tandem with the estimation protocols established by Project 912 to predict properties for chemicals not readily available through literature sources.

KEY WORDS: environmental; physical property data; statistical process control

1. INTRODUCTION

In response to a critical need to fill major gaps in the information required to make environmentally responsible engineering decisions and meet regulatory requirements, in 1991 AIChE/DIPPR® (American Institute of Chemical Engineers, Design Institute for Physical PRoperty Data) initiated DIPPR® Projects 911 and 912 - Environmental, Safety and Health Data Compilation and Estimation Manual at Michigan Technological University (MTU). The goal of AIChE/DIPPR® Projects 911 and 912 is to assemble a collection of carefully evaluated experimental data from the literature and estimation methods covering properties and parameters crucial to the fields of environmental protection, process safety, and health. Due to the lack of carefully evaluated experimental data for many of these parameters, good estimation methods are a necessity. The advent of molecular structure-based prediction methods, combined with the dramatic increase of computer availability and power, has made the development of sophisticated prediction/estimation software practical. The results of this project will greatly aid engineers in designing cleaner, safer manufacturing processes; and in evaluating the fate and risk associated with chemicals in the environment.

2. QUALITATIVE DATA SCRUTINY

Every effort is made to ensure that the thousands of pieces of data in the Project 911 database are recorded accurately. The quality assurance procedures for the data entry process includes two major steps; 1) screening of a journal article by an MTU investigator to determine data quality, 2) a screening of the data entry process, and item by item check for each chemical and physical property combination contained in a journal article. Literature data values are

rated on a scale of 0 to 2 for each of five categories, as shown in Table I. The sum of the individual scores forms a qualitative rating for comparing physical property data values from different literature sources. The highest rated data values and qualitative rating codes are made available to the user through a DOS compatible software program.[1] A Windows™ based product incorporating the Project 912 estimation techniques is under development.

3. QUANTITATIVE DATA ANALYSIS USING THE STATISTICAL PROCESS CONTROL (SPC) SYSTEM

In order to refine the criteria for the quality assessment of the DIPPR® Project 911 data compilation, the MTU investigators have identified quantitative data checks that are being computerized to initiate a Statistical Process Control (SPC) system. The goal is to implement an SPC system that satisfies the quality assurance/quality control needs required by the project steering committee and by the rest of the technical community. The new computerized checks *do not* supplant the qualitative QA/QC procedures already in place, but are extra precautions in addition to those measures. The SPC 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.

3.1. Defining the SPC System

Two types of checks are being performed on the Project 911 data values, as outlined below.

1. **Internal Check Within a Physical Property for Data Consistency:** This is an internal consistency check within a physical property for each individual chemical. This "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. Example tolerance values for some of the physical properties within the Project 911 database are found in Table II. Data values that pass the Level 1 SPC evaluation are denoted in the software database product with a "Q1" quality code.
2. **Intercomparison of Physical Properties and Additional SPC Criteria:** This "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 comparison to a specified range.

3.2. Physical Property Intercomparisons Within the SPC System

A list of Level 2 quantitative checks for some of the DIPPR® 911 properties is shown in Table III. Data values that pass the Level 2 data analysis are designated in the Project 911

software product with a "Q2" quality code. Further information on the types of Level 2 intercomparisons that are being implemented is given below.

Upper or lower limits: Most physical and chemical properties have a typical range of values; e.g.- flammability must have a range between 0 and 100%, and gas diffusivities usually fall within a decade.

Order of Magnitude: Related property values have a narrow range of relative values; e.g. liquid diffusivity should not exceed gas diffusivity.

Theoretical Data Relationships: The thermodynamic interrelationships of properties provide the best check for many colligative properties. One such example is the relationship between the temperature dependent vapor pressure equation, the heat of vaporization at the normal boiling point, and the normal boiling point (NBP). In such interrelationships one must first define the property with the highest quality code and/or the ones most easily and accurately determined by direct measurement on which to base the evaluation. In the previously mentioned case, the NBP may be most easily measured. The vapor pressure equation should then predict a pressure of one atmosphere at the NBP, and the slope of the equation at that point (on a ln P vs 1/T basis) should be proportional to the heat of vaporization. Similar interrelationships include the theoretical ties between activity coefficients and partitioning values (Henry's law constants, K_{ow}, etc.); heats of combustion and heats of formation; and many others. Some limiting values or ranges of values may also be predicted by ideal thermodynamic assumptions. For example, the

critical compressibility, Z_c, usually falls between 0.25 and 0.31. Since:

$$Z_c = \frac{P_c \underline{V}_c}{RT_c}$$

this relationship may be used to test the consistency of the critical properties, with P_c, T_c, and V_c, being the critical pressure, critical temperature, and critical molar volume, respectively.

Comparison to Predicted Values and "Rules of Thumb": With the development and selection of DIPPR® Project 912 prediction techniques, estimated values for most of the data values of concern to the project will be calculated. Consistent deviation from the values produced by carefully evaluated prediction methods may be used to flag property values well outside the range of anticipated error. The MTU project team has already employed this approach to great effect for the fire and explosion parameters. Systematic errors may also be spotted this way if inconsistencies with well identified data sets occur. "Rules of thumb" tend to be simplified versions of these estimation methods (e.g. - Trouton's rule), and are generally restricted to predicting limiting property values or data ranges.

Family Plots: Empirical checks for inconsistencies in the range of values for a property among a homologous series are being implemented. 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.

3.3. Classification of Errors

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

For those data values that do not pass the SPC 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 data files.

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 SPC 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 Project 911 software product, but they are documented by the SPC tracking system.

Data values that are flagged with an error code of "DE" are rechecked by the SPC system during the next SPC review of the Project 911 database. 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 SPC system, which are a compilation of all errors identified for the criteria checks run, are logged and dated in a notebook and maintained in a file according to the physical property on which the SPC analysis was completed. The SPC 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 SPC REVIEW SYSTEM

Results to date from the SPC review system are found in Table II. 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. The "Q1" quality code designates a successfully completed Level 1 review, and a "Q2" is used for a successfully completed Level 2 review.

The total number of data values in the Project 911 database as of May 1997 that will be subject to review by the SPC system is 60,000. From Table II, it can be seen that 27,131 data values (45% of the total) have successfully completed a Level 1 SPC analysis, as designated by the "Q1" quality code rating. The total number of data entry errors ("DE" code) is 261, which is less than one percent error due to data entry mistakes for the physical properties evaluated to date. The low occurrence of reduced ratings ("AD" code; 0.27% of total data values analyzed) suggests that the qualitative rating system that has been applied to all database references is an effective screening tool for determining the reliability of literature data. Less than two percent of the data values that have completed a Level 1 review were judged to be questionable when compared to data values from other data sources ("AF" code). The largest number of "AF" codes occurred for the Heat of Formation. Project 911 defines this property as the Heat of Formation for the chemical in the form of an ideal gas at standard state. In some cases where the Heat of Formation for an ideal gas is not available, the Heat of Formation for a liquid at standard state has been entered in the database, with appropriate comments to identify this fact to the database user. As a result, many of the "AF" flags for the Heat of Formation are due to the difference in phases. It should be noted that as additional data values are acquired and entered in the Project 911 database, the data values that are currently

flagged as anomalies will be rechecked by the SPC system, and they may eventually achieve a "Q1" rating.

One Level 2 SPC analysis has been completed, the comparison of melting point to normal boiling point. As shown in Table II, the percentage of errors again is very low. Most of the "AF2" codes (Anomaly Flagged, Level 2 review) were due to chemicals that decompose when they are heated. This makes accurate experimental measurements for physical property data difficult, and the resulting data available in the literature are not always in good agreement.

As more Level 2 evaluations are completed, it will be important to remember that the total "flag" count may differ considerably from the actual number of errors logged. This will be a result of not double counting errors (when more than one evaluation method is employed for a single physical property) and will also help prevent comparing "apples to oranges", such as comparing homologous series "outliers" with interrelated thermodynamic property inconsistencies.

Future work on the analysis of the results obtained from using the SPC system include developing a report format compatible with the production of Pareto charts to document the source of the errors and the error rate for the DIPPR® 911 project. This is a standard method of tabulating statistical errors.[2] Pareto charts will help to identify steps in the data entry and evaluation process that need improvement, if any. Physical properties where a large number of anomalies are identified will also require further review to determine the reasons behind inconsistencies in the published literature.

5. CONCLUSIONS

The goal of AIChE DIPPR® Projects 911 and 912 is to develop a comprehensive, consolidated database of physical properties for those chemical species which are regulated by various agencies of the U.S. government, and are important to the chemical process industry. Environmental, safety and health (ESH) properties are the prime focus of the data collection and validation efforts of the two projects.

Work is continuing on the critical assessment of the quality of data within the Project 911 database. Available literature data are compiled and categorized according to quality. To automate the data evaluation effort, a computerized Statistical Process Control (SPC) data review system has been designed. To date, 45% of the data values within the Project 911 database that will be subject to evaluation by the SPC system have completed a Level 1 analysis. The SPC review has shown a data entry error rate of 1%, and a data anomaly rate of 2%. Those data values that have been flagged as anomalies will be reviewed again by the SPC system when additional data values are available for comparison purposes. Recommended data values and qualitative and quantitative rating codes are provided as part of the Project 911 DOS software product.

ACKNOWLEDGEMENT

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. AIChE DIPPR® Project 911 Environmental, Safety, and Health Data Software, *AIChE 1997 Publications Catalog*, (AIChE, New York, NY, 1997), p. 7.

2. D.J. Wheeler and D.S. Chambers, *Understanding Statistical Process Control*, *Second Edition*, (SPC Press, Inc., Knoxville, TN, 1992).

Table I. Qualitative Evaluation Criteria

Criteria	Rating of 0	Rating of 1	Rating of 2
Experimental Conditions	Not stated	Stated without range	Stated with range
Purity of Chemicals	Not stated	Stated with no range or as received with range	Stated with range, purified, or calculated
Experimental Technique	Not stated or Not acceptable	Stated briefly	Described in detail
Reported Accuracy and/or Internal Precision - QA/QC Procedures	Not stated or >20%	Stated with 5 to 20%	Stated with 1 to 5%
Qualitative Agreement with Other Work	Does not agree	Agrees	Agrees well with other work

Table II. Results from the SPC Review System

Physical Property	Number of Values Evaluated	Number and Type of Error AF / DE / AD	Level 1 Tolerance Value	Quality Code
Molecular Weight	5669	5 / 16 / 1	± 5%	Q1
Melting Point	3891	88 / 20 / 5	± 10%	Q1
Normal Boiling Point	5628	20 / 34 / 10	± 10%	Q1
Diffusivity in Air	1282	8 / 1 / 0	± 30%	Q1
Diffusivity in Water	1141	63 / 4 / 1	± 30%	Q1
Heat of Formation	1510	173/ 50 / 21	± 20%	Q1
Critical Temperature	3469	22 / 5 / 0	± 10%	Q1
Flash Point	3256	73 / 121/ 25	± 10%	Q1
Heat of Combustion	1285	76 / 10 / 9	± 10%	Q1
TOTAL	27,131	508 / 261 / 72		
		AF2 / DE / AD		
Melting Point less than Normal Boiling Point	3635	56 / 7 / 1		Q2

AF: Anomaly Flagged, Level 1; value is correct as transcribed from literature source
AF2: Anomaly Flagged, Level 2; value is correct as transcribed from literature source
DE: Data Entry Error
AD: Anomaly, Rating Dropped; highest rated value is transcribed correctly from literature source, but does not agree with multiple other literature values.

Table III. Examples of Level 2 Quantitative Checks for the SPC System

Liquid Heat Capacity > Vapor Heat Capacity.
Liquid Thermal Conductivity > Vapor Thermal Conductivity
Henry's Law Constant $\leq$ Vapor Pressure / Aqueous Solubility
Normal Boiling Point is 0.5 to 0.9 times the Critical Temperature
Aqueous Solubility $\geq$ 1 / Activity Coefficient (within a given tolerance)
Bioconcentration Factor < log K_{ow}
Liquid Density (does not include polymers) falls within a range of 0.5 to 8 g/cm^3
Diffusivity in Air < diffusion rate of helium in air, and within the range of 0.1 to 1 cm^2/sec
Diffusivity in Water < diffusion rate of hydrogen in water, and within the range of 1 to 10 x 10^{-5} cm^2/sec
Molecular Weight comparison to values calculated from IUPAC atomic weights