Reliability of Uncertainty Assignments and Its Impact on Generating Recommended Data

Qian Dong

Thermodynamics Research Center (TRC)
National Institute of Standards and Technology (NIST)
Boulder, CO, U.S.A.

Types of Physicochemical Data

Measured (Observed) Values

Derived Values

Correlated (Estimated) Values

Recommended Values

Example

Simmrock et al. , **Critical Data of Pure Substances**, 1986

The IUPAC Project on Vapor Liquid Critical Properties, (1987 ~ 2002)

> "Use of this compilation except by those already knowledgeable in the subject cannot be recommended because the authors apparently did not understand the difference between observed and estimated values, and as a result they reported critical properties of about **200** more substances than the 400 or so that have been studied experimentally. They then included estimated values in tests of methods of estimation, thus invalidating the results of the tests".

Uncertainty and Data Quality

- "A measurement result is complete only when accompanied by a quantitative statement of its uncertainty".

- Attribute 1: "Uncertainties" and Attribute 2: "Data Integrity" from "Data Quality Assurance for Thermophysical Property Databases – Applications to the TRC SOURCE Data System", ***J. Chem. Inf. Comput. Sci.* 2002, *42*, 473-480**

- An essential piece of quality information expected by users of the databases.

- Required in order to decide if the result is adequate for its intended purpose and to ascertain if it is consistent with other similar results.

Functionality of Uncertainty

1. Used as a basis for selecting the “best” value from duplicate measurements

2. Used as a basis for calculating weighting factors for taking average or for fitting data to models

3. Used as a measure of tolerance for the property value or for propagation to derived properties

Terminology of Measurement Uncertainty

- Common Understanding – “Interval around the result of a measurement that contains the ‘true value’ with high probability”.

- Formal definition by ISO – “A parameter associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand”.

Assignment of Uncertainty in Databases

- Database Design:
  - Stored in the form of a numerical value
  - Interpreted as a bias for the associated property value
  - Graphical representation

- Assignment Criteria:
  - Sample purity
  - Special effects – contamination or decomposition
  - Measurement technique
  - Completeness and adequacy of the procedure description
  - Agreement among duplicate measurements
  - Estimates of uncertainty, precision, or reproducibility by authors
  - Experience/records of the investigator

Methods in Selecting Best Values

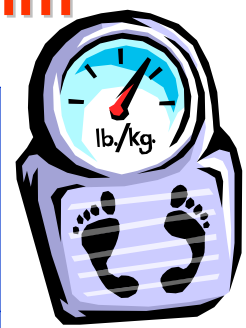
Multiple-point property values through

- Weighted-average-value method (WAV)
- Estimated uncertainty
- Publication year
- Sample purity
- Experimental objective

One- or two-point property value through

- Measurement-related factors (sample purity, methods, instruments control)
- Characters of the compound (stability, temperature)
- Similar structures
- Critically reviewed and newly developed models
- Reliable data sources

Codes of Weighting Factor Used in Algorithm



Publish Year	>1960 > 0.8	1900~1960 0.2 ~ 0.6	<1900 > 0.02	
Default Uncertainty	>1960 <2	1900~1960 3 ~ 5	<1900 10	
Object	A 1	B 0.8	C 0.6	0.4 D
Sample Purity	High % 1	Low % 0.5	UN > 0.02	

Computer Algorithm for Constants

Step 0
Preliminary Screening

first data set selected

Step 1
Weigh Factor Value
Weighted-Average-Value
Absolute Average Deviation

second data set selected

Step 2
Weighted-Average-Value
Based on Actual Deviation

third data set selected

Step 3
Repeat of Step 2
At a Stricter Level

Recommended Values

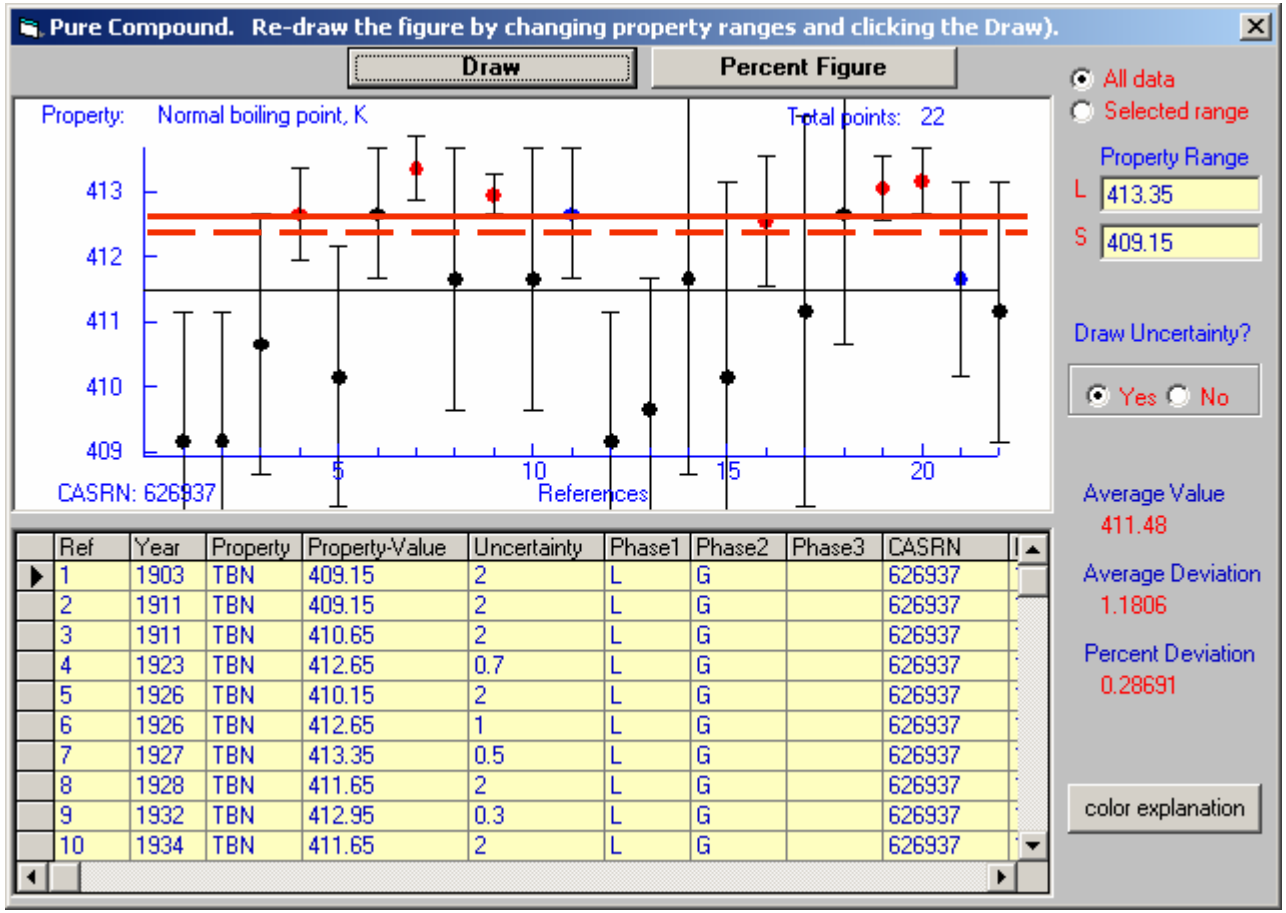
The Uncertainty Assignment Has a Large Subject Component

- Experienced database professionals who are competent to make fair judgment and understand the probable effect of various phenomena on the final estimate
- It is dependent on the information given by the investigator in the document or in previous document
- Different compilers will probably make different assignments based on the same information
- The same compiler will make different judgments toward the same information at different times

Difficulties in Assigning reliable uncertainty

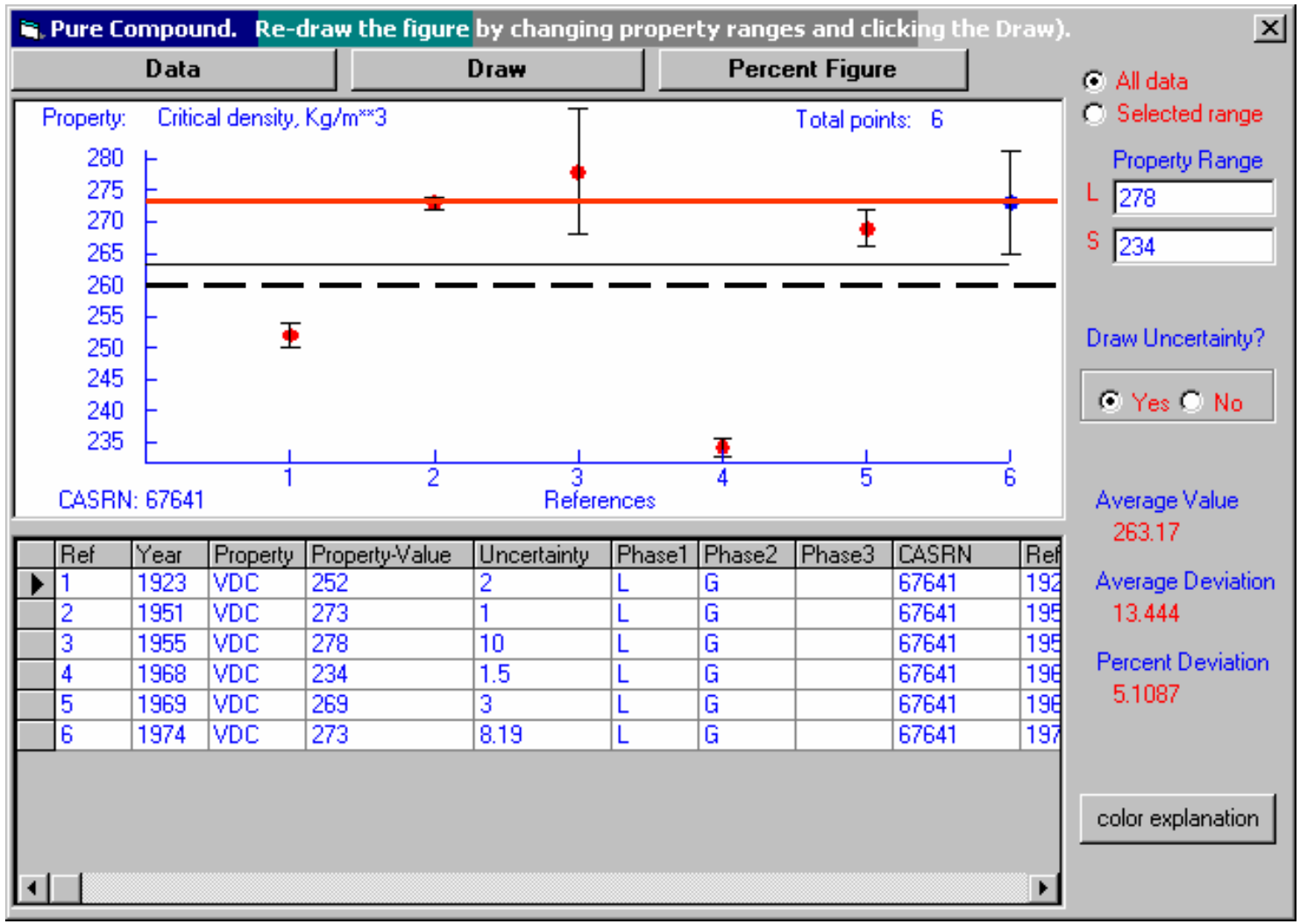
- Inconsistent assignments
- Missing values
- Lower limit of uncertainty by investigators
- Incomplete description in the reports
- Variations in the reported results (Plots)

Data Discreteness



Ref	Year	Property	Property-Value	Uncertainty	Phase1	Phase2	Phase3	CASRN
1	1903	TBN	409.15	2	L	G		626937
2	1911	TBN	409.15	2	L	G		626937
3	1911	TBN	410.65	2	L	G		626937
4	1923	TBN	412.65	0.7	L	G		626937
5	1926	TBN	410.15	2	L	G		626937
6	1926	TBN	412.65	1	L	G		626937
7	1927	TBN	413.35	0.5	L	G		626937
8	1928	TBN	411.65	2	L	G		626937
9	1932	TBN	412.95	0.3	L	G		626937
10	1934	TBN	411.65	2	L	G		626937

The Impact of Uncertainty on Recommended Data



Ref	Year	Property	Property-Value	Uncertainty	Phase1	Phase2	Phase3	CASRN	Ref
1	1923	VDC	252	2	L	G		67641	192
2	1951	VDC	273	1	L	G		67641	195
3	1955	VDC	278	10	L	G		67641	195
4	1968	VDC	234	1.5	L	G		67641	196
5	1969	VDC	269	3	L	G		67641	196
6	1974	VDC	273	8.19	L	G		67641	197

Key Issues

1. Database Design
2. Data Quality
3. Knowledge Supporting System
4. Integrated Process

Key Issue 1 – Database Design

Information Completeness –

A well-designed raw data repository with complete supporting information (metadata) and reliable uncertainty assessments

Required Information –

- Substance characteristics (stability)
- Sample Purity
- Measurement Techniques
- Estimation of uncertainty by authors

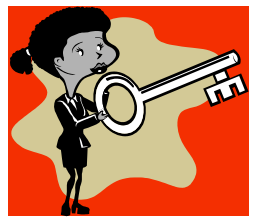
Key Issue 2 – Data Integrity

Data Integrity – the correctness of the information stored in the database, and selection has to be made based on the clean data

Current Status of Database Quality – a large-scale numeric database without critical evaluation may have an error rate of 2-5%.

Major Error Types –

- Typographical errors
- Unit-conversion errors
- Report interpretation errors
- Metadata compilation errors
- Original report errors

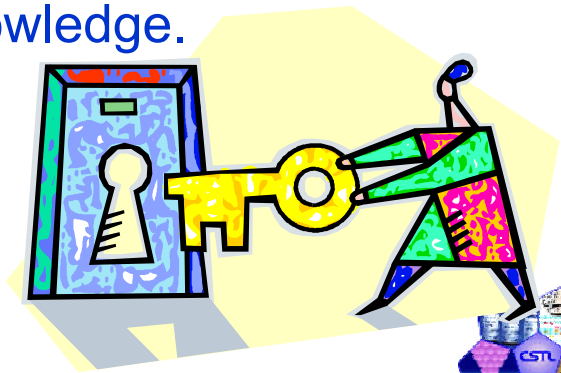


Systematic and iterative approach

Key Issue 3 – Knowledge Supporting System

- Scientific experiment is a complicated process
- Experimental data tend to have uncertainty or error
- Evaluation of scientific data is extremely difficult, no way to guarantee its absolute correctness
- The true value of physicochemical property needs repeat examinations

The need to do evaluation on measured data and the need to be able to explain to users the quality of data, i.e. enabling the users of this data to be aware of its limitation and/or problem, demanding supporting system based on the domain knowledge.



Key Issue 4 - Integrated Process