**MATCH**
*Communications in Mathematical and in Computer Chemistry*

MATCH Commun. Math. Comput. Chem. **73** (2015) 303-326

ISSN 0340 - 6253

# Applicability Domain of Nonlinear Property-Property Relationships – Example: Estimation of Vapour Pressure

**Joachim Altschuh[1,*], Dieter Lenoir[2], Florian Rehfeldt[1,3], Rainer Bruggemann[4]**

*[1] Helmholtz Zentrum München, German Research Centre for Environmental Health, Institute of Computational Biology, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany*
*[2] Helmholtz Zentrum München, German Research Centre for Environmental Health, Molecular Exposomics, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany*
*[3] Georg-August-University, Third Institute of Physics – Biophysics, Friedrich-Hund-Platz 1, 37077 Göttingen, Germany*
*[4] Leibniz-Institute of Fresh Water Ecology and Inland Fisheries, 12587 Berlin, Germany*

(Received September 3, 2014)

## Abstract

We discuss some property-property relationships (PPRs) for the estimation of vapour pressure with regard to their predictive power in terms of accuracy and applicability. Seven different PPRs mostly based on the Clausius-Clapeyron equation are analysed, including for example the method of Mackay or those of Fishtine/Vetere. A data set of 375 compounds was compiled, which contains all required quantities as measured ones. Several criteria are defined to determine the accuracy of the estimation methods. By applying partial order theory, it turns out that two equations by Fishtine/Vetere are optimal, i. e. yield the best estimation results, albeit they are not comparable. According to the main focus of this paper our data set is discussed in terms of chemical structures. Structural elements, such as carbonyl-function, amine-function, halogen substitution etc., were identified, which were suitable to characterize the diversity of a set of organic chemicals. Given the composition of the current data set, it was necessary to reduce the number of structural elements to eight. Finally, we introduce the ad-matrix, describing the quality of an estimation equation with respect to a certain structural element. The numerical differences of the $ad_{ij}$-values among different estimation equations are not large. Hence a fuzzy partial order approach was applied to get the best vapour pressure estimation equations with respect to a distinct structural element. Four estimation methods can be recommended, concerning their accuracy with respect to different chemical structures.

[*] Corresponding author: altschuh@helmholtz-muenchen.de

# 1 Introduction

In general, the validity of an estimation equation can be discussed in terms of accuracy and applicability. Both terms are closely interrelated. A higher required accuracy usually implies lower applicability, i. e. the validity of the estimation equation with respect to certain structural classes. The methods we apply here can be useful for any quantitative structure-activity relationship or property-property relationship. However, the technique of analysis for linear models has made large steps toward a conceptually closed theory (see for example [1-3]). It is interesting to note that already in 1994 the need of an analysis of analogies in chemical structures and estimation methods has been stressed [4].

The vapour pressure is an important thermodynamical property in its own right. Moreover, it is also a key property in the estimation of fate and distribution of chemicals in the environment. In consequence, many publications originate from fate modelling of environmental chemicals within the context of risk assessment (see e.g. [5-7]) or from investigations on aerosols [8,9].

Experimental determination of the vapour pressure is often time-consuming and expensive; measurements are especially complicated for compounds with low and very low vapour pressure [10]. Consequently, estimation methods to predict vapour pressures are of increasing importance. The number of vapour estimation methods is large [11-13]. In the following we cite Barley et al. [8]: “The number of vapour pressure equations in the literature that could be combined with estimated $T_b$ -(boiling temperatures)- values is large, although several equations are variations on each other.” In the same reference we find another sentence, which is characterizing the problem with the validation of estimation equations of the vapour pressure: “The selection of a vapour pressure estimation method for use (in the modelling of aerosol formation) is always going to be a compromise between accuracy, complexity and coverage of all the required functional groups.” Here, in this paper, we are confronted with this problem, too: Sophisticated estimation methods may use parameters whose values are specific for chemical groups (see for instance [14-16]). The advantage is the chance for a high accuracy for specific chemical classes but there is also a chance for a high input error as the group parameters have to be estimated, too (compare the famous Mc O’Neill parabolas, see for example [17]). Here, we do not discuss group-based methods at all; instead our aim is to verify the applicability domain of property-property estimation methods which do not contain group-specific parameters and which are considered as being applicable to a wide range of chemical classes.

Our paper is organized as follows.

(i) Methods: Here we discuss some of the vapour pressure estimation methods and present seven different estimation equations. After introducing the data set, we define the criteria, which will be used for analyzing the accuracy and applicability. As more than one criterion will be necessary to evaluate the accuracy, partial order theory will be applied, which will be briefly introduced. The discussion of accuracy is, however not the main point of this paper. Our focus lies on the applicability part and consequently our data set is discussed in terms of chemical structures. A scheme of structural elements introduced previously [18] is used to describe and quantify the chemical characteristics of the data set of environmental chemicals.

(ii) Results: With respect to the accuracy two variants of the Fishtine/Vetere equations yield the best estimations of vapour pressure. When the applicability domain has to be quantified, it is necessary to obtain statistically robust results. Therefore, the decisive quantity $ad_{ij}$ (ad: applicability domain) is introduced. This matrix ("ad-matrix") describes the relative number of chemicals being no outliers with respect to the $j^{th}$ structural element, applying the $i^{th}$ estimation equation. The numerical differences of the $ad_{ij}$-values among different estimation methods are not large. Hence, a fuzzy partial order approach was applied to get optimal vapour pressure estimation equations concerning the $j^{th}$ structural element.

(iii) Conclusion: We present a methodology, which follows the traditional concept of estimation equations, where the chemical structures are not explicitly built-in. When this restriction is accepted, then our approach can be summarized as:

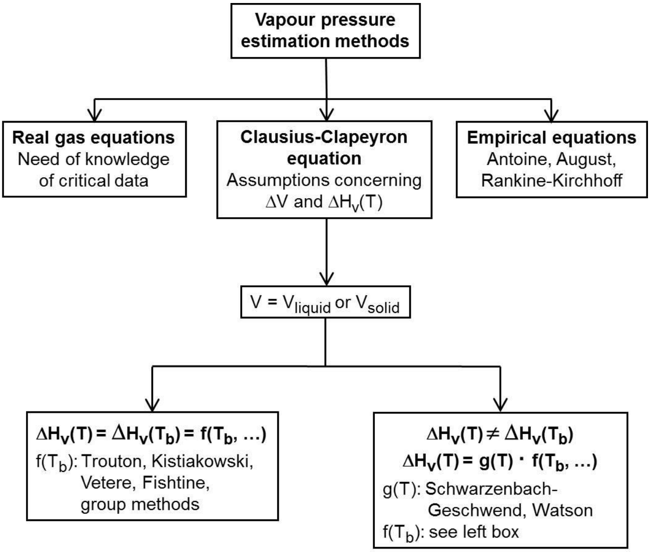
- Find criteria for accuracy and apply partial order to obtain the relative best (i. e. optimal) estimation equations.
- Find structural elements and test the estimation equations concerning the number of outliers.
- From the relative number of outliers the central $ad_{ij}$-values are derived. These in turn relate the quality of the estimation method with the structural element of interest.

# 2 Methods

## 2.1 Property-property relationships for vapour pressure estimation

Estimation methods for physicochemical and biological properties play an important role in different fields (see for example [11,19,20]). Basically, these methods fall into two categories: (i) Quantitative structure-property relationships (QSPRs) make use of molecular structure descriptors; (ii) quantitative property-property relationships (PPRs) use other experimental data (e. g. boiling points) [12]. In the case of vapour pressure most of the PPRs are developed from the Clausius-Clapeyron equation and have therefore a sound thermodynamical basis. In order to break this equation down to easily accessible chemical properties, various approaches have been developed to account for the temperature dependence of the enthalpy of vaporization as well as dealing with the enthalpy of vaporization at boiling temperature. These procedures finally result in a number of semi-empirical estimation methods (see [11], for example).

For the present validation study we have compiled seven PPRs which are mostly based on the Clausius-Clapeyron equation. Some others are empirical equations such as the August or the well-known Antoine equation. It may be useful to give first an overview by a flow chart (Figure 1).



Vapour pressure estimation methods
Real gas equations
Need of knowledge of critical data
Clausius-Clapeyron equation
Assumptions concerning ΔV and $\Delta H_V(T)$
Empirical equations
Antoine, August, Rankine-Kirchhoff
$V = V_{liquid}$ or $V_{solid}$
$\Delta H_V(T) = \Delta H_V(T_b) = f(T_b, \ldots)$
$f(T_b)$: Trouton, Kistiakowski, Vetere, Fishtine, group methods
$\Delta H_V(T) \neq \Delta H_V(T_b)$
$\Delta H_V(T) = g(T) \cdot f(T_b, \ldots)$
$g(T)$: Schwarzenbach-Geschwend, Watson
$f(T_b)$: see left box

**Figure 1:** Schema of vapour pressure estimation methods.

The equations for the estimation of the vapour pressure $p_v$ are given in detail below. The boiling point $T_b$, the melting point $T_m$, and the molecular mass M are required as experimental data. In some cases there is a need of extrapolation from the solid to the subcooled liquid state. This is done by eq. (1) [21]:

$$f(T_m) = \begin{cases} 0 \quad \text{for liquids at 298 K} \\ 2.9532 \cdot (1 - T_m/T) \quad \text{for solids at 298 K} \end{cases} \tag{1}$$

T is the absolute temperature.

**AU1 (August)** [22]

$$\log p_v = -5.87 \cdot T_b/T + 11.03 \tag{2}$$

**AU2 (August)** [23]

Parameters of eq. (2) as fitting quantities at $p_v = p_0$ and $T = T_0$

$$\log p_v = - \beta/T \cdot \log (p_o/p_s) + \log p_o + \beta/T_o \cdot \log (p_o/p_s) + 2.125 + f(T_m) \tag{3a}$$

$$\text{with } \beta = (T_o \cdot T_b)/(T_o - T_b),\ \log p_o = 6.68,\ T_o = 1750 \text{ K} \tag{3b}$$

**AN (Antoine)** [19]

$$\log p_v = 5.0057 + A \cdot B \cdot C + f(T_m) \tag{4a}$$

$$A = (2.09/T_b) + (0.4747/T_b) \cdot \ln T_b \tag{4b}$$

$$B = (0.81\ T_b + 18)^2 \tag{4c}$$

$$C = [1/(0.81\ T_b + 18)] - [1/(T + 18 - 0.19\ T_b)] \tag{4d}$$

**WA (Watson)** [19]

$$\log p_v = 0.43429 \ln p_v + 5.0057 \tag{5a}$$

$$\ln p_v = K_F \cdot T_b \cdot (8.75 + R \cdot \ln T_b)\ A / (1.9274 \cdot T_b) \tag{5b}$$

$$A = 1 - B - C \tag{5c}$$

$$B = (3 - 2\ y)^m / y \tag{5d}$$

$$C = 2\ m\ (3 - 2\ y)^{(m - 1)} \cdot \ln y \tag{5e}$$

$$K_F = 0.99 \tag{5f}$$

with $y = T/T_b$, R is the gas constant in cal/(mol · K). The parameter m is calculated as follows:

$$T_m < 298 \text{ K:} \quad m = 0.19 \tag{5g}$$

$$T_m > 298 \text{ K:} \quad m = \begin{cases} 0.36 & \text{for } T/\ T_b > 0.6 \\ 0.8 & \text{for } 0.6 > T/\ T_b > 0.5 \\ 1.19 & \text{for } T/\ T_b < 0.5 \end{cases} \tag{5h}$$

**MC (Mackay)** [24]

$$\log p_v = 0.43429 \ln p_v + f(T_m) + 5.0057 \tag{6a}$$

$$\ln p_v = -(4.4 + \ln T_b) \cdot [(1 + K) \cdot (T_b/T - 1) - K \cdot \ln T_b/T] \tag{6b}$$

$$K = 0.803 \tag{6c}$$

**FV1 (Fishtine/Vetere)** [20]

$$\log p_v = 0.43429 \ln p_v + f(T_m) + 5.0057 \qquad (7a)$$

$$\ln p_v = (\Delta H_V\ (T_b)/R)\ [(1 + K) \cdot (1/T_b - 1/T) + K/T_b \cdot \ln (T_b/T)] \qquad (7b)$$

$$\Delta H_V(T_b) = \Delta S_V \cdot T_b \qquad (7c)$$

$$\Delta S_V = b \cdot (a_1 + a_2 \cdot \log T_b + (1/M) \cdot \sum a_{i+2} \cdot T_b^i)\ \ (i = 1, 2, 3) \qquad (7d)$$

with the following values for $a_i$ and b

| b | $a_1$ | $a_2$ | $a_3$ | $a_4 \times 10^3$ | $a_5 \times 10^6$ |
|---|---|---|---|---|---|
| 1.03 | 10.604 | 3.664 | 0.09354 | 1.035 | -1.345 |

**FV2 (Fishtine/Vetere)** [20]

$$\log p_v = 0.43429 \ln p_v + f(T_m) + 5.0057 \qquad (8a)$$

$$\ln p_v = (\Delta H_V(T_b)/R)\ [(1 + K) \cdot (1/T_b - 1/T) + K/T_b \cdot \ln (T_b/T)] \qquad (8b)$$

$$\Delta H_V(T_b) = \Delta S_V \cdot T_b \qquad (8c)$$

$$\Delta S_V = 13.91 + 3.27 \cdot \log M + 1.55 \cdot A/M \qquad (8d)$$

Preliminary studies have shown that is reasonable to modify the term A, which is originally given by eq. (8e) [20], as follows:

$$A = \begin{cases} [T_b - (263 \cdot M)^{0.581}]^{1.037} \ \ \text{for } T_b - (263 \cdot M)^{0.581} < 0 & (8e) \\ 0 \ \ \text{otherwise} & (8f) \end{cases}$$

## 2.2 Vapour pressure data set

Measured vapour pressure data are required for the development and validation of estimation methods. Many data are available from the literature and are used here. Beyond this, six additional vapour pressures have been determined experimentally. These additional compounds belong to the class of aliphatic alcohols and phenols, respectively, which has shown striking number of outliers in former studies [25]. The gas saturation method [10,26] was applied and phenanthrene was used for calibration of the method. The obtained value is in fairly good agreement with the values obtained by other methods [27]. The measured vapour pressure of the six compounds is contained in the Supplementary Material. In general, the values were measured in the temperature range from 40–160 °C in five steps.

For a validation study of the PPRs given in section 2.1, not only vapour pressure data but also the corresponding boiling points (and melting points) are required. In order to maintain the same information content for all seven PPRs, we restricted ourselves to those compounds where all needed properties are available. This resulted in a set of 375 organic chemicals from various chemical classes. A full list of the compounds and their $p_V$, $T_b$, and $T_m$ data is given in the Supplementary Material.

## 2.3 Criteria

To test the validity of the estimation equations, we introduce different criteria following closely an earlier publication [18]. As a basic criterion we introduce the mean square error (MSE). This quantity combines variance and the bias, see [28].

$$K_1 = \text{MSE} \tag{9}$$

However, the MSE is a highly aggregated criterion. Therefore we introduce additional criteria, which are more specifically related to the different types of errors.

We consider specifically the number of deterministic "outliers", $N_Q$, as a basic quantity. $N_Q$ is defined as the number of compounds for which the quotient between the experimental ($P_{exp}$) and the estimated value ($P_{est}$) of the considered substance property P

$$F_Q(P) = P_{exp}/P_{est} \tag{10}$$

deviates by a certain factor $F_Q \neq 1$. In this paper we consider those compounds as outliers for which $F_Q \leq 0.1$ or $F_Q \geq 10$, respectively, holds. (Note: P is used as symbol for any chemical property, whereas PPR is used, when a specific estimation method concerning a single property is meant.)

The number of outliers has to be related to the total number of compounds N, hence the second criterion is given by:

$$K_2 = N_Q/N \tag{11}$$

Further criteria are developed on the basis of the linear regression equation relating measured and estimated values.

$$P_{exp} = a \cdot P_{est} + b \tag{12}$$

In general, both, $P_{est}$ and $P_{exp}$ are to be considered as stochastic quantities; therefore the regression analysis is performed following the geometric mean technique [29]. In order to get the criteria oriented in the same way (the better, the lower the value of the specific criterion) we set:

$$K_3 = \text{abs}(1 - a) \tag{13}$$

and

$$K_4 = \mathrm{abs}(b) \tag{14}$$

Criterion $K_1$ is a screening criterion. It will turn out that the criterion $K_2$ is related to the applicability and $K_3$ and $K_4$ are bias-related criteria. The set $\{K_2, K_3, K_4\}$ will be used simultaneously, i. e. we do not want to favour one of these three criteria over the others or to combine them numerically, for example by weighted sums.

We speak of $K_i(PPR_j)$ as the $i^{th}$ criterion applied on the $j^{th}$ PPR estimation equation.

## 2.4 Partial order concept

### 2.4.1 Motivation

We introduced four criteria in order to examine the validity of the estimation equations. Experiences show [18,25] that the criteria $K_2$, $K_3$, $K_4$ are in general not highly correlated. Therefore, it is meaningful to study them simultaneously. To perform the analysis in this way, the concept of partial order has proven to be suitable [30]. In the following the concept of partial order is briefly introduced.

### 2.4.2 Basic definition

***Standard partial order***

We call a certain estimation equation (briefly written as $PPR_1$) worse than another one ($PPR_2$) if the following holds:

$$PPR_1 \geq PPR_2 : \Leftrightarrow [K_2(PPR_1), K_3(PPR_1), K_4(PPR_1)] \geq$$
$$[K_2(PPR_2), K_3(PPR_2), K_4(PPR_2)] \tag{15}$$

The ≥-relation between vectorial quantities is not defined a priori. Nevertheless, there are several possibilities for a meaningful definition, for example within the concept of majorization [31-33]. Here one of the simplest is selected, given by eq. (16):

$$[K_2(PPR_1), K_3(PPR_1), K_4(PPR_1)] \geq [K_2(PPR_2), K_3(PPR_2), K_4(PPR_2)] : \Leftrightarrow$$
$$K_2(PPR_1) \geq K_2(PPR_2) \text{ and } K_3(PPR_1) \geq K_3(PPR_2) \text{ and } K_4(PPR_1) \geq K_4(PPR_2) \tag{16}$$

Consider two tuples [1, 2, 3] and [3, 4, 3]. Then eq. (16) is fulfilled and hence – according to eq. (15) – we can write: [3, 4, 3] ≥ [1, 2, 3]. Consider another pair of tuples: [1, 2, 5] and [3, 4, 1]. Then eq. (16) is not fulfilled and hence an order relation such as ≥ cannot be valid between these two tuples. If eqs. (15) and (16) are not fulfilled, $PPR_1$ and $PPR_2$ are “incomparable”. I. e. the definitions from eqs. (15) and (16) constitute a partial order. According to eqs.

(15) and (16) the partial order is defined as a pair (X, IB) where X is a set of objects (here the set of PPRs) and IB is the set of criteria, here $\{K_2, K_3, K_4\}$. The set (X, IB) is called a partially ordered set (poset).

***Fuzzy partial order***

According to definitions in eqs. (15) and (16) an order relation is only possible if the $\leq$-relation of the single components of the criteria vector is valid for all three criteria simultaneously. This requirement does not take care for slight numerical differences, which may be thought of as non-relevant. Therefore also fuzzy-concepts can be applied, which are explained in depth elsewhere [30,34]. Consider the two tuples [3, 4, 3] and [2.9, 5, 3]. Then the incomparability is caused by the slight numerical difference with respect to the first component of the tuple. The only problem is: Which difference is considered as relevant? Therefore a "tolerance parameter" $\alpha_{cut}$ is introduced. Low values of $\alpha$ mean that large differences in the numerical values are considered as irrelevant, whereas for values of $\alpha_{cut} \rightarrow 1$ only very small numerical differences will be thought of as being irrelevant and hence the values are equivalent. Research is still going on, how an optimal value of $\alpha_{cut}$ can be defined. In this paper that poset will be selected that is informative enough.

***Further definitions concerning partial order***

Maximal elements: Elements x for which no other element $y \in X$ can be found with $y \geq x$

Minimal elements: Elements x for which no other element $y \in X$ can be found with $y \leq x$

Cover relation: let $x \leq z$ and $z \leq y$ then it follows by the transitivity axiom that $x \leq y$. Pairs of objects x, y which are comparable and for which **no** object z can be found for which $x < z$ and $z < y$ are following a cover relation, denoted by "≤:". In case of $x \leq y$, x is "covered by" y, or y "is covering" x.

Hasse diagram: Objects are related to each other by cover relations. By cover relations a visualization of the partial order set is possible. Often this type of diagrammatical representation is called "Hasse diagram" [35]. In general, objects x, y for which $x > y$ is valid are drawn with x in a vertical higher position than y.

## 2.5 Structural elements

In order to better analyse the applicability domain of a given PPR we have introduced structural elements $S_i$, which characterize the chemical features of an organic compound. The 15 structural elements have been chosen to fit our needs, i. e. the description of environmentally relevant organic chemicals and their physicochemical properties [18]. However, additional or other types of structural elements could be applied depending on the composition of a specific data set or a different property/activity, for example. Here the use of dictionaries of chemical structures as developed in chemoinformatics could be helpful [36].

The 15 structural elements used here are shown in Table 1 together with some examples of organic compounds, which bear the given structural element. Basically, the concept of introducing s structural elements allows to discuss chemical classes with different degree of refinement, namely s not-necessarily disjoint sets of compounds to $2^s - 1$ necessarily disjoint classes.

**Table 1.** Definition of structural elements $S_i$

| No. | Structural element $S_i$ | Examples<br>(compounds that bear the corresponding structural element) |
|---|---|---|
| 1 | Aromatic | Benzene, naphthalene, biphenyl |
| 2 | Nonaromatic cyclic | Cyclohexane |
| 3 | Nonaromatic C=C | Butadiene |
| 4 | Nonaromatic C≡C | Acetylene |
| 5 | Halogen substitution | Bromobenzene, chloroform |
| 6 | N-O function | Nitrobenzene |
| 7 | C=N or C≡N | Benzonitrile |
| 8 | $NR_3$ | Aniline, trimethylamine |
| 9 | C=O function | Acetone, acetic acid |
| 10 | OH function | Ethanol, phenol |
| 11 | Phosphororganic | Triethylphosphate |
| 12 | Sulfurorganic | Thiophene |
| 13 | Topological genus<br>(more than one ring) | Biphenyl, camphene |
| 14 | Heterocyclic | Atrazine |
| 15 | R-O-R function | Diethylether, methyl benzoate |

The compounds may then be easily characterized by tuples of digits indicating the presence ("1") or non-presence ("0") of the corresponding structural element. Chlorobenzene, for example, is a halogenated aromatic chemical; thus $S_1$ and $S_5$ are present and the tuple reads [1,0,0,0,1,0,0,0,0,0,0,0,0,0,0]. Obviously, the presence of a certain structural elements does not imply the absence of other structural elements. Further examples have been given elsewhere [18]. Using these tuples, distinct subsets of larger data set can easily be created and characterized.

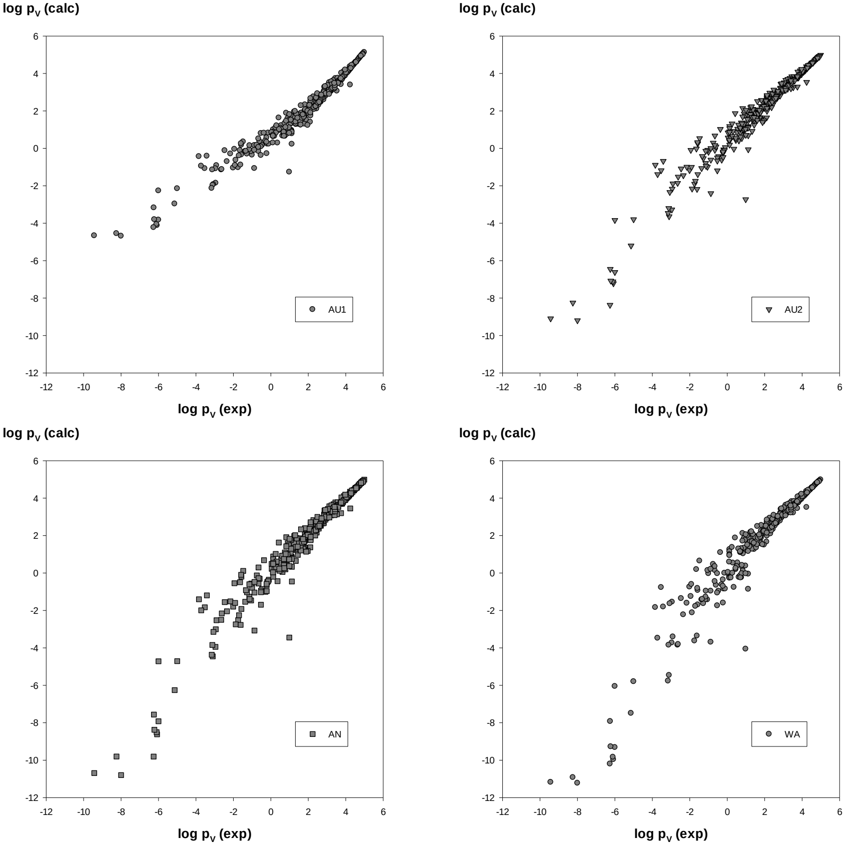
Our idea can now be explained more precisely: We want to validate nonlinear estimation methods with classes as large as possible, taking into account that probably some structure dependencies must be acknowledged.

# 3 Results and Discussion

## 3.1 Accuracy

### 3.1.1 Measured vs. calculated data

Figure 2 shows the estimation results from the 7 PPRs by comparing experimental and calculated vapour pressures of the whole data set of 375 compounds. By simple optical inspection one can see that the estimation error for MC, FV1, FV2, and AU2 is relatively small and constant over the entire vapour pressure range. These four PPRs seem to yield the best overall results, whereas the results obtained from AU1 seem to be the worst. The PPRs AN and WA tend to underestimate the vapour pressure in particular at low values, whereas AU1 overestimates at low values. There are also a couple of chemicals with poor estimation results independent of the PPR used.



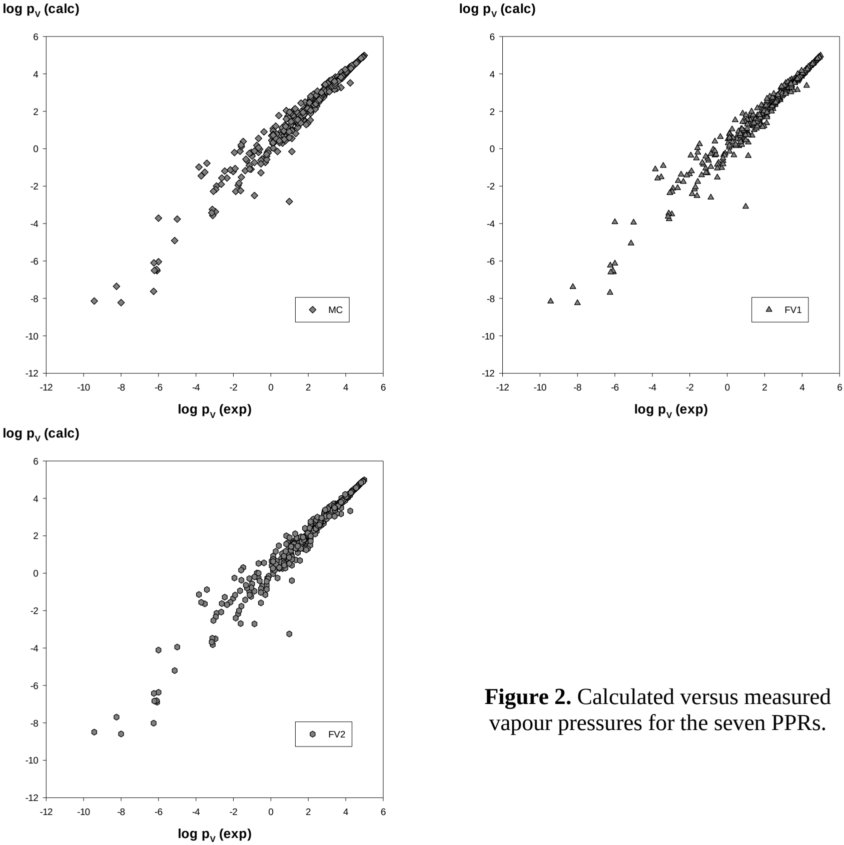


**Figure 2.** Calculated versus measured vapour pressures for the seven PPRs.

A decision about the "best" PPR from these figures only is not reasonable. To discuss these differences further, the criteria defined in section 2.3 have been calculated and will be analysed using the partial order concept described in section 2.4.

## 3.1.2 Accuracy with respect to the criteria

### *Results with respect to any single criterion $K_1$ – $K_4$*

The four criteria describing the validity of the seven PPRs are summarized in Table 2. The values are rounded to three decimals.

With respect to $K_1$ we find the following order among the estimation equations:

$K_1$: FV1 < FV2 < MC < AU2 < AN < WA < AU1 #1

**Table 2:** The four criteria of the seven PPRs (for definition see section 2.3).

| PPR | $K_1$ | $K_2$ | $K_3$ | $K_4$ |
|---|---|---|---|---|
| AU1 | 0.668 | 0.128 | 0.297 | 0.923 |
| AU2 | 0.363 | 0.083 | 0.023 | 0.266 |
| AN | 0.388 | 0.080 | 0.074 | 0.154 |
| MC | 0.361 | 0.061 | 0.046 | 0.263 |
| FV1 | 0.279 | 0.056 | 0.033 | 0.099 |
| FV2 | 0.289 | 0.053 | 0.015 | 0.105 |
| WA | 0.652 | 0.120 | 0.128 | 0.201 |

This sequence altogether confirms the optical inspection of the measured vs. calculated data from Figure 2, and provides a deeper insight into the comparison of the PPRs. Based on the MSE-criterion $K_1$ (see section 2.3) FV1 is the best and AU1 the worst estimation equation. We further observe that FV1 < FV2 and AU2 < AN. However, the reason for a high position of a PPR in the sequence #1 is not obvious. Is MC for example worse than FV2 because of the bias or because of its scatter of values? The analysis of the criteria $K_2$-$K_4$ is helpful in this respect.

With respect to $K_2$ we find the following sequence:

$K_2$: FV2 < FV1 < MC < AN < AU2 < WA < AU1 #2

This sequence differs from that obtained from $K_1$. The order of the pairs FV1 and FV2 as well as of AU2 and AN are reversed comparing the results of $K_1$ and $K_2$, respectively.

As can be seen in the sequences #3 and #4, which result from the criteria $K_3$ and $K_4$, respectively, there are many inversions of the orders, indicating that the medium to poor validation results as found by sequence #1 have different reasons.

$K_3$: FV2 < AU2 < FV1 < MC < AN < WA < AU1 #3

$K_4$: FV1 < FV2 < AN < WA < MC < AU2 < AU1 #4

For all criteria the PPR AU1 is always the worst, whereas FV2 is the best estimation equation in two of the three sequences #2-#4, and second best in the other two. In general, however, it is difficult to decide on an order among the PPRs considering all three criteria $K_2$, $K_3$, and $K_4$ simultaneously.

***Results from applying partial order theory***

Based on the three criteria $K_2$, $K_3$, and $K_4$ applied on seven PPRs and the set of 375 compounds an evaluation matrix is derived, whose rows are related to the seven PPRs and whose columns are numerical values referring to $K_2$, $K_3$, and $K_4$.

The correlation matrix (see Table 3) shows that there is no perfect correlation and hence partial order concepts come into play.

**Table 3:** Pearson correlation index of the four criteria.

| | $K_3$ | $K_4$ |
|---|---|---|
| $K_2$ | 0.857 | 0.704 |
| $K_3$ | | 0.912 |

Application of partial order, eqs. (15) and (16), leads to a Hasse diagram shown in Figure 3.

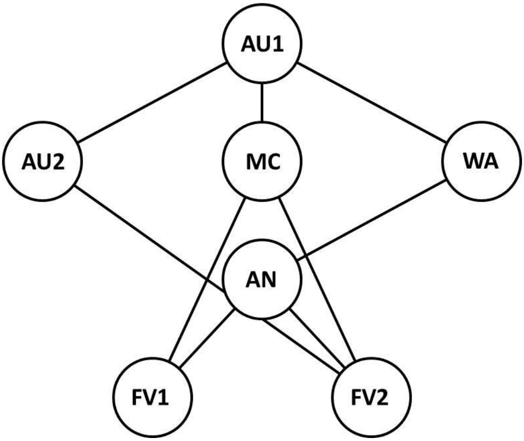


**Figure 3:** Hasse diagram based on the criteria $K_2$, $K_3$, and $K_4$.

A number of conclusions can be drawn by analyzing this Hasse diagram:

- FV1 and FV2 are minimal elements. Inspection of the three criteria (Table 2) shows that FV1 is better than FV2 in only $K_4$, whereas FV2 is (slightly) better than FV1 in $K_2$ and $K_3$.
- With respect to all three criteria AU1 is the worst estimation equation.
- The PPRs AU2, MC, and WA are found in the middle range of the Hasse diagram, i. e. there are some PPRs which are better and some others which are worse in all three criteria.
- According to the three criteria $K_2$, $K_3$, and $K_4$ possible sequences can be found: for example FV1 < AN < WA < AU1 (the longest sequence) or FV2 < AU2 < AU1 (a shorter one). In the terminology of partial order theory these subsets, where each object (each

PPR) is comparable to each other, are called chains. Identification of chains provides important information: The values of all criteria are simultaneously (weakly) increasing, when starting at the bottom and progressing up to the top along a chain.

- AU2 is the estimation equation with the highest degree of incomparabilities. For example FV1 can be compared with almost all other estimation equations, FV2 being the only exception. AU2 can only be compared with two other estimation methods, namely AU1 and FV2.
- WA cannot be compared with MC. Although both coincide with respect to the modelling of $\Delta H_V(T_b)$ (both: Kistiakowski-approach), they differ in the description of the temperature dependency: WA follows the Watson approach, whereas MC is a linear approximation of $\Delta H_V(T)$, introducing the parameter K.
- AU2 could be drawn at the same horizontal level as AN without violating the laws of partial order. Nevertheless the simple method AU2 cannot be compared with AN. Although the Antoine equation has three adjustable parameters, but AU2 only two, the two PPRs are not comparable: There is only a slight numerical difference in $K_2$, but with respect to $K_3$ and $K_4$ it is found: $K_3$: AN > AU2 and $K_4$: AN < AU2. Bias and scatter lead to the incomparability between AN and AU2.

## 3.2 Applicability Domain

### 3.2.1 Outliers with respect to the structural elements

Figure 4 shows the distribution of the 15 structural elements in the data set of 375 compounds. Aromatic ($S_1$), halogen substitution ($S_5$), and hydroxy-groups ($S_{10}$) are well, whereas phosphororganic chemicals ($S_{11}$) and triple bonds ($S_4$) are only poorly represented in the data set. Due to the fact that some $S_i$ are not well represented in the data set and since we want to get statistically robust results, we retain only those $S_i$ whose representation exceeds the median of the $N_i$-distribution, i. e. 36. Thus eight structural elements remain.

For all seven PPRs the number of outliers (as defined in eq. (10)) are shown in Table 4 for each of the eight structural elements. We call the number of outliers referring to the j$^{th}$ structural element of the i$^{th}$ PPR $N_Q^{(i)}(S_j)$.

In Table 5 some striking observations, which can be deduced from Table 4, are summarized.

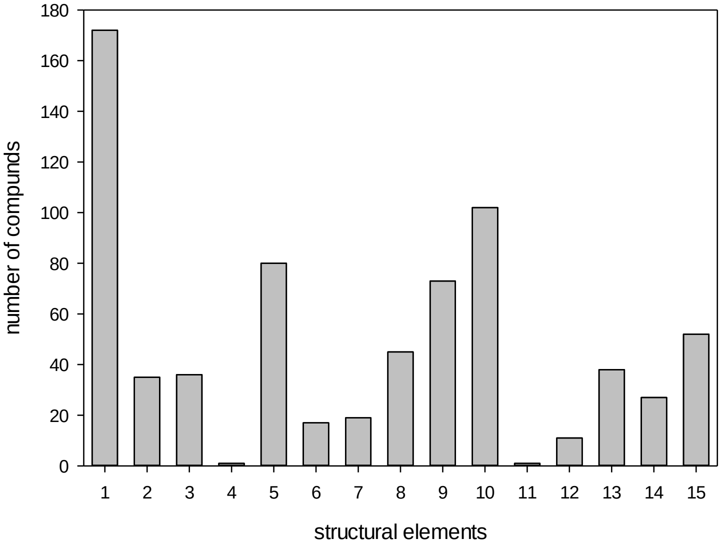


**Figure 4:** Distribution of the structural elements within the data set.

**Table 4:** Absolute number of outliers $N_Q^{(i)}(S_j)$ of the seven PPRs with respect to the remaining eight structural elements $S_j$.

| $S_j$ | $N_j$ | $N_Q^{(i)}(S_j)$ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | AU1 | AU2 | AN | MC | FV1 | FV2 | WA |
| $S_1$ | 172 | 30 | 11 | 17 | 8 | 8 | 8 | 27 |
| $S_3$ | 36 | 2 | 0 | 0 | 0 | 0 | 0 | 1 |
| $S_5$ | 80 | 6 | 3 | 1 | 2 | 2 | 2 | 3 |
| $S_8$ | 45 | 5 | 6 | 6 | 6 | 6 | 7 | 7 |
| $S_9$ | 73 | 6 | 3 | 2 | 3 | 3 | 3 | 6 |
| $S_{10}$ | 102 | 22 | 20 | 12 | 15 | 13 | 12 | 15 |
| $S_{13}$ | 38 | 19 | 6 | 14 | 4 | 5 | 4 | 17 |
| $S_{15}$ | 52 | 4 | 3 | 0 | 1 | 1 | 1 | 6 |

**Table 5**: Best and worst PPR with respect to the structural elements.

| $S_i$ | structural element | best PPR | worst PPR |
|---|---|---|---|
| $S_1$ | aromatic | MC, FV1, FV2 | AU1 |
| $S_3$ | C=C | AU2, AN, MC, FV1, FV2 | AU1 |
| $S_5$ | halogens | AN | AU1 |
| $S_8$ | $NR_3$ | AU1 | FV2, WA |
| $S_9$ | C=O | AN | AU1, WA |
| $S_{10}$ | OH | AN, FV2 | AU1 |
| $S_{13}$ | topological genus | MC, FV2 | AU1 |
| $S_{15}$ | R-O-R | AN | WA |

The number of outliers of the $i^{th}$ PPR with respect to the $j^{th}$ structural element $S_j$ as shown in Table 5 is misleading because the number of compounds bearing the structural element $S_j$ is strongly varying. Therefore, it is convenient to introduce the quantity $ad_{ij}$ in order to get results independent of the realisation for each structural element.

$$ad_{ij} = 1 - N_Q^{(i)}(S_j)/N_j \tag{17}$$

Obviously, the less the relative number of outliers, the better the corresponding $i^{th}$ PPR. The ad-matrix is given in Table 6.

**Table 6:** The $ad_{ij}$-values of the seven PPRs and the eight $S_i$.

| $S_i$ | AU1 | AU2 | AN | MC | FV1 | FV2 | WA |
|---|---|---|---|---|---|---|---|
| $S_1$ | 0.826 | 0.936 | 0.901 | 0.953 | 0.953 | 0.953 | 0.843 |
| $S_3$ | 0.944 | 1 | 1 | 1 | 1 | 1 | 0.972 |
| $S_5$ | 0.925 | 0.963 | 0.988 | 0.975 | 0.975 | 0.975 | 0.963 |
| $S_8$ | 0.889 | 0.867 | 0.867 | 0.867 | 0.867 | 0.844 | 0.844 |
| $S_9$ | 0.918 | 0.959 | 0.973 | 0.959 | 0.959 | 0.959 | 0.918 |
| $S_{10}$ | 0.784 | 0.804 | 0.882 | 0.853 | 0.873 | 0.882 | 0.853 |
| $S_{13}$ | 0.5 | 0.842 | 0.632 | 0.895 | 0.868 | 0.895 | 0.553 |
| $S_{15}$ | 0.923 | 0.942 | 1 | 0.981 | 0.981 | 0.981 | 0.885 |

$S_3$ differentiates the seven PPRs in only three values. Interestingly, $S_{13}$ differentiates strongly, namely by six different values, i. e. almost each PPR has its own value.

The well-known amoeba diagrams can help visualizing the differences between two PPRs with respect to their applicability domain. In Figure 5 such an amoeba diagram is shown, which compares the two estimation equations FV1 and FV2. As can be seen, it cannot be stated that FV1 has less outliers than FV2 nor that FV1 has generally more outliers than FV2. Additionally, FV1 and FV2 behave differently with respect to the structural elements. Whereas for the topological genius $ad_{FV2,topol} > ad_{FV1,topol}$, holds, the situation is reversed with respect to the structural element $NR_3$, i. e. $ad_{FV2,NR3} < ad_{FV1,NR3}$ holds in this case. In other words, FV1 and FV2 are not only incomparable with respect to the accuracy criteria $K_2$-$K_4$, but also with respect to their applicability domain.

The seven estimation equations will lead to 21 pairwise amoeba diagrams or to one amoeba diagram with seven lines, which are close to each other. Such a figure with 7 lines or a set of 21 pairwise amoeba diagrams is certainly not suitable for a general analysis. Instead once again concepts of partial order are applied.

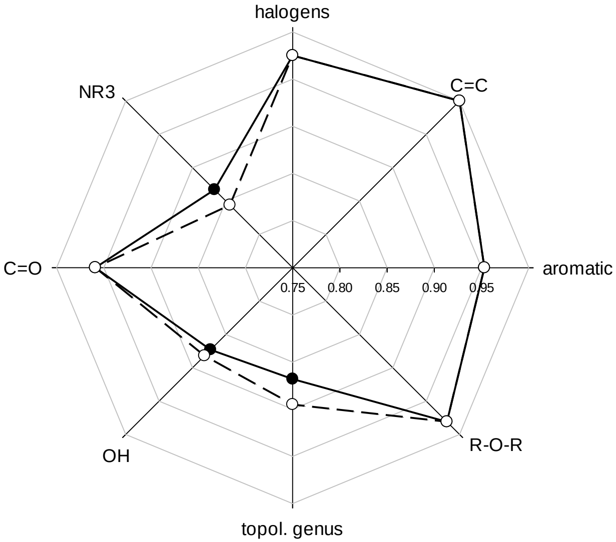


**Figure 5:** Amoeba diagram for the comparison of FV1 (full line) and FV2 (dashed line) due to the $ad_{ij}$-values from Table 6.

### 3.2.2 Fuzzy partial order

The ad-matrix is the basis for the partial order analysis. The seven PPRs are now characterized by the $ad_{ij}$-values. The larger the value the better the results obtained by the corresponding PPR. Examination of Table 6 shows that the numerical differences can be very small. Therefore the concept of fuzzy posets is applied (cf. sect. 2.4.2). As we are interested in verifying strong differences in ad-values, the tolerance parameter $\alpha_{cut}$ was selected in that way that at least six equivalence classes are found, which is more than 75% of all PPRs. Figure 6 shows a Hasse diagram, whose construction is based on concepts of fuzzy theory.

The Hasse diagram from Figure 6 shows two striking aspects. A chain can be found with WA as worst, FV2 as middle ranged and MC and FV1 as relatively best PPRs. There are two maximal elements (MC $\cong$ FV1 and AN) and one isolated element (AU1). In Table 7 we consider all possibly incomparable PPRs due to the Hasse diagram in Figure 6, such as AU1 and AN, or AU2 and FV1 and identify the structural elements by which a PPR is favoured over the other one.

As can be expected from Table 6, AU1 with its bad value with respect to $S_{13}$ and its pretty good value with respect to $S_8$ is isolated. With respect to WA, only $S_8$ and $S_{13}$ lead to the observed incomparability. AN is better than FV1, FV2, and MC with respect to the structural elements $S_5$, $S_9$, and $S_{15}$. In some cases $S_9$ and $S_{10}$ favour AN. For three pairs of PPRs, namely (AN, FV1), (AU2, AN), and (AN, MC), the structural elements $S_3$ and $S_8$ do not differentiate among the PPRs. Taken $S_3$ alone, five PPRs are not differentiated among each other.

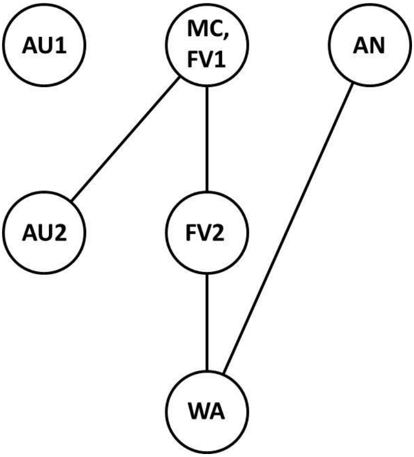


**Figure 6:** Hasse diagram of seven PPRs based on ad-values and generated by applying fuzzy concepts with $\alpha_{cut} = 0.95$. "Good" PPRs are here at the top of the diagram in contrast to Figure 3, where deviations from best values are the basis.

**Table 7:** Comparison of PPRs with respect to certain structural elements (those structural elements are shown, for which the PPRs show better or equal results than others)

| $PPR_1$ | $PPR_2$ | $S_i$: $PPR_1 > PPR_2$ | $S_i$: $PPR_1 = PPR_2$ | $S_i$: $PPR_1 < PPR_2$ |
|---|---|---|---|---|
| AU1 | AU2, AN, FV1, FV2, MC | 8 | - | 1, 3, 5, 9, 10, 13, 15 |
| AU1 | WA | 8, 15 | - | 1, 3, 5, 9, 10, 13 |
| AU2 | AN | 1, 13 | 3, 8 | 5, 9, 10, 15 |
| AU2 | FV2 | 8 | 3, 9 | 1, 5, 10, 13, 15 |
| AU2 | WA | 1, 3, 8, 9, 13, 15 | 5 | 10 |
| AN | FV1 | 5, 9, 10, 15 | 3, 8 | 1, 13 |
| AN | FV2 | 5, 8, 9, 15 | 3, 10 | 1, 13 |
| AN | MC | 5, 9, 10, 15 | 3, 8 | 1, 13 |

A closer examination of Table 7 shows:

- $S_1$ and $S_{13}$ are causing a preference of {FV1, FV2, MC} compared to the remaining PPRs.
- WA > AU1 and WA > AU2: in both cases $S_{10}$ is causing this preference. WA < AU1 and WA < AU2: in both cases $S_8$ and $S_{15}$ are causing this preference.

- AN is incomparable with five PPRs, namely AU1, AU2, FV2, and MC, and FV1: $ad_{AN,3}$ = 1, therefore AN is better than WA and AU1.
- AU2 (as one of the minimal elements) can only be compared with MC and FV1, i. e. AU2 is in all eight structural elements worse than MC and FV1. However, AU2 cannot be compared with AU1, FV2, WA, and AN. This incomparability is caused by a pretty complex pattern of structural elements. Details can be deduced from Table 6.

# 4 Discussion

## 4.1 Our methods compared to others

Validation concepts have been described before (see for instance [2,37,38]). However, they often concentrate on the accuracy aspects of the validation, such as division of the data set into training and test sets and cross-validation procedures. These methods are well established and do not need any reinvestigation. Compared to these aspects, besides of the Williams plot [37] and the use of leverage values (known in regression analysis as ‘influence statistics’ (hat matrix) [39]), methods to evaluate the applicability domain of estimation methods are not in broad use.

We have chosen the vapour pressure as example for a chemical property to be estimated. Our chosen set of seven estimation methods is sufficiently large enough to introduce our methodological approach. The advantage of this selection is given by its simplicity and the availability of a data set with adequate chemical diversity. A complete analysis of a broader set of estimation methods for vapour pressures must be subject to additional studies.

## 4.2 Summary of results

Seven PPRs are identified to estimate the vapour pressure from other macroscopic properties. By simultaneous evaluation of three accuracy criteria, we obtained a Hasse diagram (see Figure 3) where the Fishtine/Vetere equations FV1 and FV2 turned out to be the relative best ones, whereas the simple method AU1 is the worst one. These results do not take into account that different PPRs may turn out to be better applicable, when specific substance sets (of specific chemical classes) are selected. Here we finally investigated eight substance subsets with respect to one criterion (number of outliers, $K_2$). Each of these sets is characterized by the presence of one of the eight most relevant structural elements for all substances. Checking the ad-matrix for each PPR and substance subset (see Figure 6) the relative best PPRs are MC and FV1 and AN. Even for the overall worst PPR AU1 the analysis by fuzzy partial order set theory has shown that there is one structural element ($S_8$), for which AU1 is better than all other PPRs. Although the PPR AN is a maximal element, i. e. yielding good results with respect to the ad-matrix, it is incomparable with FV1 and MC. The structural elements $S_5$, $S_9$, $S_{10}$, and $S_{15}$ favour AN over FV1 and MC, whereas $S_1$ and $S_{13}$ favour FV1 and MC over AN.

The concept of validation of estimation methods by checking the accuracy first and then identifying those substance subsets with low number of outliers might be summarized by a general evaluation procedure, which is shown (and proposed for general application in the field of QSAR) in the next section.

## 4.3 Summary of the evaluation procedure

In Figure 7 the procedure to check the accuracy and applicability of PPRs is shown in general terms. Following the flow chart shown in Figure 7 the analysis includes (i) selecting suitable estimation methods, (ii) applying suitable accuracy criteria to find the relatively best PPRs, (iii) describing the chemical diversity by structural elements, and (iv) applying (fuzzy) partial order to identify the applicability domain.

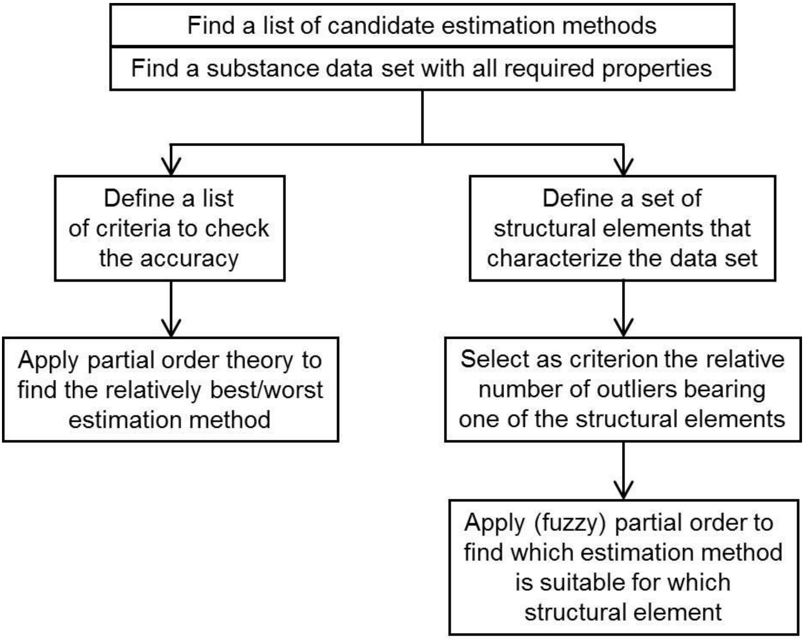

Find a list of candidate estimation methods
Find a substance data set with all required properties
Define a list of criteria to check the accuracy
Define a set of structural elements that characterize the data set
Apply partial order theory to find the relatively best/worst estimation method
Select as criterion the relative number of outliers bearing one of the structural elements
Apply (fuzzy) partial order to find which estimation method is suitable for which structural element

**Figure 7:** Procedure to evaluate property-property estimation methods.

## 4.4 Conclusion

We started with seven vapour pressure estimation equations and 15 structural elements. According to the accuracy criteria defined here, the Fishtine/Vetere equations FV1 and FV2 were optimal (i. e. yielded the best overall estimation results), whereas the empirical equation AU1 is the least element in partial order and is not recommendable. However, the inclusion of structural elements can change this result. In order to obtain statistical robust results the set of structural elements was reduced to eight. Here the representation by our basis set of 375 substances was sufficiently satisfying. Application of the ad-matrix relates the outliers to the eight remaining structural elements. Because the numerical differences were too small, we applied fuzzy partial order. Now the optimal estimation equations are AN, MC1, FV1, and AU1. The inclusion of structural elements, i. e. the applicability aspect, justified the simple August-equation (AU1) as an equation suitable for specific substance classes.

# References

[1] P. Gramatica, Principles of QSAR models validation: internal and external, *QSAR Comb. Sci.* **26** (2007) 694–701.

[2] A. Tropsha, Best practices for QSAR model development, validation and exploration, *Mol. Inf.* **29** (2010) 476–488.

[3] R. Khosrokhavar, J. B. Ghasemi, F. Shiri, 2D Quantitative structure–property relationship study of mycotoxins by multiple linear regression and support vector machine, *Int. J. Mol. Sci.* **11** (2010) 3052–3068.

[4] D. H. Rouvray, Similarity studies 1. The necessity for analogies in the development of science, *J. Chem. Inf. Comp. Sci.* **34** (1994) 446–452.

[5] S. Galassi, A. Provini, E. Halfon, Risk assessment for pesticides and their metabolites in water, *Int. J. Environ. Anal. Chem*. **65** (1996) 331–344.

[6] L. Carlsen, A combined QSAR and partial order ranking approach to risk assessment, *SAR QSAR Environ. Res.* **17** (2006) 133–146.

[7] L. Carlsen, R. Bruggemann, Y. Sailaukhanuly, Application of selected partial order tools to analyze fate and toxicity indicators of environmentally hazardous chemicals, *Ecol. Indicators* **29** (2013) 191–202.

[8] M. H. Barley, The critical assessment of vapour pressure estimation methods for use in modelling the formation of atmospheric organic aerosol, *Atmos. Chem. Phys.* **10** (2010) 749–767.

[9] S. Compernolle, K. Ceulemans, J. F. Müller, Technical note: Vapor pressure estimation methods applied to secondary organic aerosol constituents from a-pinene oxidation: an intercomparison study, *Atmos. Chem. Phys.* **10** (2010) 6271–6282.

[10] K. Nass, D. Lenoir, A. Kettrup, Calculation of the thermodynamic properties of polycyclic aromatic hydrocarbons by an incremental procedure, *Angew. Chem. Int. Ed. Engl.* **34** (1995) 1735–1736.

[11] E. J. Baum, *Chemical Property Estimation – Theory and Application*, Lewis Pub., Boca Raton, 1998.

[12] J. C. Dearden, Quantitative structure–property relationships for prediction of boiling point, vapor pressure, and melting point, *Environ. Toxicol. Chem.* **22** (2003) 1696–1709.

[13] A. R. Katrizky, M. Kuanar, S. Slavov, C. D. Hall, Quantitative correlation of physical and chemical properties with chemical strcture: Utility for prediction, *Chem. Rev.* **110** (2010) 5714–5789.

[14] E. Voutsas, M. Lampadariou, K. Magoulas, D. Tassios, Prediction of vapor pressures of pure compounds from knowledge of the normal boiling point temperature, *Fluid Phase Equil.* **198** (2002) 81–93.

[15] E. Panteli, E. Voutsas, K. Magoulas, D. Tassios, Prediction vapor pressures and enthalpies of vaporization of organic compounds from the normal boiling point temperature, *Fluid Phase Equil.* **248** (2006) 70–77.

[16] Y. Nannoolal, J. Rarey, D. Ramjugernath, Estimation of pure component properties: Part 3. Estimation of the vapor pressure of non-electrolyte organic compounds via group contributions and group interactions, *Fluid Phase Equil.* **269** (2008) 117–133.

[17] R. Brüggemann, U. Drescher–Kaden, *Einführung in die modellgestützte Bewertung von Umweltchemikalien – Datenabschätzung, Ausbreitung, Verhalten, Wirkung und Bewertung*, Springer, Berlin, 2003.

[18] R. Brüggemann, J. Altschuh, A validation study for the estimation of aqueous solubility from n-octanol/water partition coefficients, *Sci. Total Environ.* **109-110** (1991) 41–57.

[19] W. J. Lyman, W. F. Reehl, D. H. Rosenblatt, *Handbook of Chemical Property Estimation Methods, American Chemical Society*, McGraw Hill, New York, 1990.

[20] R. C. Reid, J. M. Prausnitz, T. K. Sherwood, *The Properties of Gases and Liquids*, McGraw Hill, New York, 1977.

[21] S. H. Yalkowsky, Estimation of entropies of fusion of organic compounds, *Ind. Eng. Chem. Fundam.* **18** (1979) 108–111.

[22] R. E. Rathbun, D. Y. Tai, Vapor pressures and gas-film coefficients for ketones, *Chemosphere* **16** (1987) 69–78.

[23] H. Fromherz, *Physikalisch–chemisches Rechnen in Wissenschaft und Technik*, Verlag Chemie, Weinheim, 1973.

[24] D. Mackay, A. Bobra, D. W. Chan, W. Y. Shiu, Vapor pressure correlations for low–volatility environmental chemicals, *Environ. Sci. Technol.* **16** (1982) 645–649.

[25] J. Altschuh, R. Brüggemann, W. Karcher, Attempts to classify QSARs with respect to their validity: vapour pressure estimation as an example, *Sci. Total Environ.* **134** (1993) 1409–1419.

[26] K. Nass, *Die Messung von Dampfdrucken mit der Gassättigungsmethode*, Dissertation TU München, 1994.

[27] D. Lenoir, F. Rehfeldt, unpublished results.

[28] M. Precht, *Biostatistik*, Oldenbourg Verlag, München, 1987.

[29] E. Halfon, Regression method in ecotoxicology: A better formulation using the geometric mean functional regression, *Environ. Sci. Technol.* **19** (1985) 747–749.

[30] R. Brüggemann, G. P. Patil, *Ranking and Prioritization for Multi–Indicator Systems – Introduction to Partial Order Applications*, Springer, New York, 2011.

[31] A. W. Marshall, I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Acad. Press, San Diego, 1979.

[32] E. Ruch, R. Schranner, T. H. Seligman, Generalization of a theorem by Hardy, Littlewood and Polya, *J. Math. Anal. Appl.* **76** (1980) 222–229.

[33] E. Ruch, I. Gutman, The branching extent of graphs, *J. Comb. Inf. Sys. Sci.* **4** (1979) 285–295.

[34] R. Bruggemann, A. Kerber, G. Restrepo, Ranking objects using fuzzy orders, with an application to refrigerants, *MATCH Commun. Math. Comput. Chem.* **66** (2011) 581–603.

[35] E. Halfon, M. G. Reggiani, On ranking chemicals for environmental hazard, *Environ. Sci. Technol.* **20** (1986) 1173–1179.

[36] A. R. Leach, V. J. Gillet, *An Introduction to Chemoinformatics*, Kluwer, Dordrecht, 2003.

[37] P. Gramatica, A. Di Guardo, Screening of pesticides for environmental partitioning tendency, *Chemosphere* **47** (2002) 947–956.

[38] E. Papa, J. C. Dearden, P. Gramatica, Linear QSAR regression models for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors, *Chemosphere* **67** (2007) 351–358.

[39] B. S. Everitt, *The Cambridge Dictionary of Statistics*, Cambridge Univ. Press, Cambridge, 1998.

# 6 Supplementary Material

A list with the experimental vapour pressures as well as the boiling and melting temperatures of the 375 compounds is available from the corresponding author.