A simplified model for the characterisation of toxic releases: A regression approach

April 2003

Contribution to Work-package 8 of the OMNIITOX Project

Reinout Heijungs

Centre of Environmental Science, Leiden University

Contents

Summary

The USES-LCA model with its data set on 181 substances was used to learn to understand the possible role of regression models for a simplified model on the basis of a (future) base model. Prior to carrying out the regression analyses, however, a number of theoretical analyses of the mathematical structure of the USES-LCA model was undertaken. These give clues towards the model specification in the regression model: which variables to include, and which transformations (logarithms, squares, etc.) to perform.

After a combination of theoretical analysis, statistical analysis and trial-and-error, we were able to deduce regression models that account for a substantial amount of variance (often 70-90%) of the logarithm of the characterisation factors for aquatic and terrestrial ecotoxicity and human toxicity for different emission compartments (air, water and soil) on the basis of 2 or 3 variables. These variables are: a measure of toxicity (the MTC for ecotoxicity or the ADI and/or the TCL for human toxicity), a measure of persistence (the DT50, most often for the emission compartment), and sometimes the Henry coefficient. A closer analysis for one of the toxicity potentials reveals an error by a factor of less than 6 for almost all substances.

The analysis also makes clear that there is a strong and opposite relation between goodness-of-fit and data availability. Good predictions can be made on the basis of more variables, and thus for fewer substances. We conclude that it might be useful to develop not just one simplified baseline model, but rather to develop a hierarchy of estimation methods, for instance a low-quality one that is based on molecular weight alone, a somewhat better one that needs one additional parameter, a still better one that needs three parameters, etc. Predictions from each of these simplified models can be accompanied by an estimated uncertainty of the prediction.

1. Introduction

Characterisation factors for toxicity-related impact categories cover three dimensions: fate, exposure and effect. It is current practice to use multi-media fate models for the fate dimension, bioconcentration factors and intake models for the exposure, and reference doses, like no-effect concentrations and acceptable daily intakes for the effect part. All these models require two types of parameters: environment-specific parameters, like the wind speed and the fraction of organic carbon in soil, and substance-specific parameters, like the chemical's half life in water and the vapour pressure. Where the gathering of environment-specific parameters is a one-time job, substance-specific parameters must be collected for every substance under consideration. Especially for the fate and exposure part of the characterisation models, this substance-specific data demand is quite large. And as there are hundreds of thousands of substances that may be emitted by industry, and as there are important data gaps for a large part of those substances, it emerges that finding characterisation factors by running the current characterisation models is problematic for a large number of substances. It is therefore natural to look for shortcuts in finding characterisation factors.

Certain characterisation models start from the idea of poor data availability, and develop approaches that are based on completely different principles than the “correct” models. For instance, there are methods that come up with three indicators, one for persistence, one for bioaccumulation, and one for toxicity. In addition, these three indicators may be combined with a simple rule. A disadvantage of these methods is that both indictors and combination rules are based on “heuristic” considerations. More specifically, they are not derived from the “correct” model, but represent an independent view on which aspects are critical for toxicity assessment. A perhaps even more important disadvantage of such methods is that they provide metrics that are incommensurable with those provided by a more detailed model.

Another option is to establish statistical relationships between the outcome of the “correct” model and a subset of its underlying parameters. For instance, if a simple formula can be found that links the vapour pressure of a chemical to its characterisation factor for a number of substances, one might apply this formula to a number of different substances. The framework of regression analysis provides a basis for identifying such statistical relationships (see Meent et al., 2002). The present document focuses on an investigation of the feasibility of using a regression model to estimate characterisation factors in a metric that is commensurable with the more detailed ones. In a sense, such a regression model is comparable to the approaches

taken by analysts of (quasi) structural activity relationships (SAR/QSAR; see e.g. Lyman et al., 1990), which is also employed in many detailed fate and exposure models (such as USES; Linders & Jager, 1998).

2. The basics of regression analysis

Regression analysis (see, e.g., Dobson, 1983, Draper & Smith, 1998, Greene, 1997) is a widely-used statistical tool in ecology, econometrics and the behavioural sciences to identify relationships between one or more input variables (the "independent" variables) and one or more output variables (the "dependent" variables). Depending on the number of independent and dependent variables, one distinguishes simple regression, multiple regression and so on. Moreover, there are types of regression analysis with special features, known under names like logistic regression, constrained regression and weighted regression. In this text , the emphasis is on the classical multiple regression analysis.

It is assumed that there is one dependent variable, *y* (here: the characterisation factor). For a number of *N* cases (here: substances) a value of *y* is given. This set of values can be denoted as

$$y_1, y_2, \ldots, y_N \tag{1}$$

A conventional way of writing is with vector notation:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{pmatrix} \tag{2}$$

It is also assumed that there are *k* independent variables (here: substance-specific data, like the molecular weight and the half life in air) for the same *N* cases. The value for case (= substance) *i* on variable *j* is denoted by x_{ij}. A tableau of these values can be written as

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} \end{pmatrix} \tag{3}$$

One of the purposes of regression analysis is to submit the hypothesis of a linear relationship between **X** and **y** to a statistical test, and to estimate the optimal coefficients in this relationship. Let the relationship be specified as

$$\hat{y}_i = a + x_{i1}b_1 + x_{i2}b_2 + \cdots + x_{ik}b_k \tag{4}$$

or in matrix notation as

$$\hat{\mathbf{y}} = a + \mathbf{X}\mathbf{b} \tag{5}$$

where a is the intercept, and **b** is a vector of coefficients. One writes $\hat{\mathbf{y}}$ instead of $\mathbf{y}$ to make clear that a perfect fit is not achievable, and that the lefthand side will be an approximation to the measured data. The coefficients a and **b** are still to be determined here, and one needs to establish a criterion when the fit is said to be best and the coefficients are said to be optimal. This criterion is a formal translation of the requirement that $\hat{\mathbf{y}}$ should be as close as possible to $\mathbf{y}$. The conventional operationalisation of this requirement is that the mean square deviation of measured[1] and estimated dependent variable is minimal. Thus,

$$\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{6}$$

should be minimised.

From this objective function, values for the coefficients a and **b** can be estimated. Moreover, under certain assumptions concerning the distribution of the error term of the fit, the standard errors of the estimates a and **b** can be computed. A formal test of significance of each coefficient can be performed with the t-statistic: it shows if a certain coefficient differs "significantly" from zero, hence if the data provides evidence that the variable is a predictor variable. The overall goodness-of-fit is expressed in the coefficient of determination, R^2. It is given by

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{7}$$

where $\bar{y}$ denotes the mean value of the elements of **y**. This coefficient measures the fraction of variance in the dependent variable that is accounted for by the statistical model. A value of 1 indicates a perfect fit, a value of 0 complete misfit, and anything in between a partial fit. For instance, $R^2 = 0.6$ means that 60% of the variance in **y** (here: the characterisation factor) can be explained[2] by the model. Whether or not this is good enough, is open to discussion. Econometricians and psychologists are often glad to find such values, while the natural sciences will often not be satisfied with such a result. It also depends on the purpose of the analysis. There is a statistical test for the significance of the coefficient of determination. It

[1] It is usual to speak about measured and estimated values, even though in the present case the former set does not represent truly measured values, but model outcomes.

[2] Explained here refers to a non-causal and purely statistical relationship: it means that the regression model accounts for the given percentage of variance.

expresses the probability that the real R^2 is zero, based on an *f*-statistic. Another interesting result of regression analysis is the standard error of the estimate

$$se = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{N-k}} \quad (8)$$

It can be used to construct confidence intervals around predictions made by the regression model. If $\hat{y}_i$ is such a prediction, a 95%-confidence interval is given by

$$\left[\hat{y}_i - t_2(df = N-k, p = 0.05)se, \hat{y}_i + t_2(df = N-k, p = 0.05)se\right] \quad (9)$$

where t_2 indicates the critical value of the *t*-distribution at the specified number of degrees of freedom and the chosen two-sided significance level of 95%.

3. The role of characterisation models

One might be tempted to submit the characterisation factors directly to a regression analysis. However, this is not a very efficient approach. The reason is, that one of the problems of regression is to choose the right model. The issue of model specification suffers from a number of problems.

A first question is: which variables should be part of the model equation? Although a *t*-test of a regression gives a clue to the question if a variable matters at all, things are in reality not so easy. The assumptions of applying the *t*-test for a coefficient are not always completely met. These assumptions include the requirement that the residual follows a normal distribution with a constant standard deviation, and that the independent variables are non-correlated. Especially when some of the independent variables are strongly correlated, the issue of multicollinearity arises. This may produce a high R^2 with low *t*-values for the individual coefficients. In the present case, it is likely that some of the variables will be highly correlated. The half life in sea water, for instance, will in many cases be the same as that for fresh water. And terrestrial no-effect levels are often derived from aquatic no-effect levels. Use of these types of "prior" information seems to be essential in preventing multicollinearity, and resolving the model specification.

A next question concerns the functional relationship employed. Although regression analysis is ordinarily linear regression, this does not mean that the independent variables should directly come from the parameters that are present. They may rather be replaced by their square, square root, logarithm, or any other transformed form. In principle, this leads to a very large number

of regression models to be tested: one in which the vapour pressure is the independent variable, one in which it is the square of the vapour pressure, one in which it is the logarithm of the vapour pressure, etc. Fortunately, we have prior knowledge of the underlying model. On the grounds of the structure of the model equations, we can deduce the best way in which certain variables should occur in the regression equation. For some other variables, this is less easy from theoretical considerations, but we can employ numerical simulations to study the dependence of a dependent variable on one of the independent variables. We then pick one case, i.e. one characterisation factor for one substance, keep all input parameters constant but vary one parameter, say the air-water partition coefficient over a very large range of values. The result can be plotted, and the resulting graph often suggests the mathematical form of the dependence: linear, logarithmic, hyperbolic, etc.

4. Preparatory work: removing the effect aspect

Let us first examine the structure of the characterisation model. It consists of a large amount of model equations, but the essential features can be grasped by decomposing it into two or three elements: fate, effect and, for human endpoints, intake. For the ecosystem endpoints we then have the form

$$Q_{fi} = E_f F_{fi} \tag{10}$$

where F_{fi} represents the fate factor that accounts for the transport of a chemical from the initial compartment i to the final compartment f, and E_f represents the effect factor that accounts for the sensitivity of a chosen endpoint in final compartment f. The characterisation factor is then given by Q_{fi}. For human endpoints the form is

$$Q_i = EI_f F_{fi} \tag{11}$$

where I_f accounts for the transfer of the chemical by intake route f.

Both forms point to a multiplicative model, and are clearly not in agreement with the linear additive form that is assumed by the regression analysis. However, a logarithmic transformation leads to the additive form:

$$\ln Q_{fi} = \ln E_f + \ln F_{fi} \tag{12}$$

and

$$\ln Q_i = \ln E + \ln I_f + \ln F_{fi} \tag{13}$$

An additional advantage of the logarithmic transformation is that the differences in scale of the characterisation factors are reduced. After all, one should realise that characterisation factors are normally specified per kg of chemical, regardless of whether it is a relatively harmless substance that is regularly emitted in quite large quantities (like dichloromethane) or an extremely poisonous substance that is severely regulated (like dioxin). Characterisation factors can easily span 10 orders of magnitude or more, and such large differences in scale can create a biased picture in the least squares fit of regression analysis. A logarithmic transformation reduces the scale differences to a much better range, with a much lower number of possibly influential data points.

A next step is, in the case of ecosystem toxicity, to investigate the influence of the fate and effect part on the characterisation factors. The effect factor is in general determined by some no-effect level (*NEL*, e.g. an LC_{50}) a to which a safety factor (*SF*) is applied:

$$E_f = \frac{1}{PNEC_f} = \frac{1}{SF \times NEL_f} \tag{14}$$

Both no-effect level and safety factor are not influenced by the fate parameters, like vapour pressure and octanol-water partition coefficient, but they are specified as measured data items (for the no-effect level) and fixed data items (for the safety factor). The *PNEC* itself is therefore of direct relevance in the regression equation. And as we chose to use the logarithm of the effect factor, and $\ln 1/x = -\ln x$, we obtain

$$\ln Q_{fi} = -\ln PNEC_f + \ln F_{fi} \tag{15}$$

for the regression equation.

5. First analysis with real data

The USES-LCA model (Huijbregts, 1999) with its data set on 181 substances was used to learn to understand the possible role of regression models for a simplified model on the basis of a detailed model.

Choosing as an example the fresh water compartment as the emission compartment and as the target ecosystem compartment (hence the FAETP for emissions to fresh water: $FAETP_{water}$), the pattern looks quite promising; see Figure 1. A regression analysis yields an R^2 of 0.88, with only one independent variable. This means that for the chosen case only 12% of the variance in the logarithm of the characterisation factor can be explained by the fate modelling. Indeed, the

variation in the concentration in fresh water is much smaller (three orders of magnitude) than the variation in the characterisation factor (ten orders of magnitude).

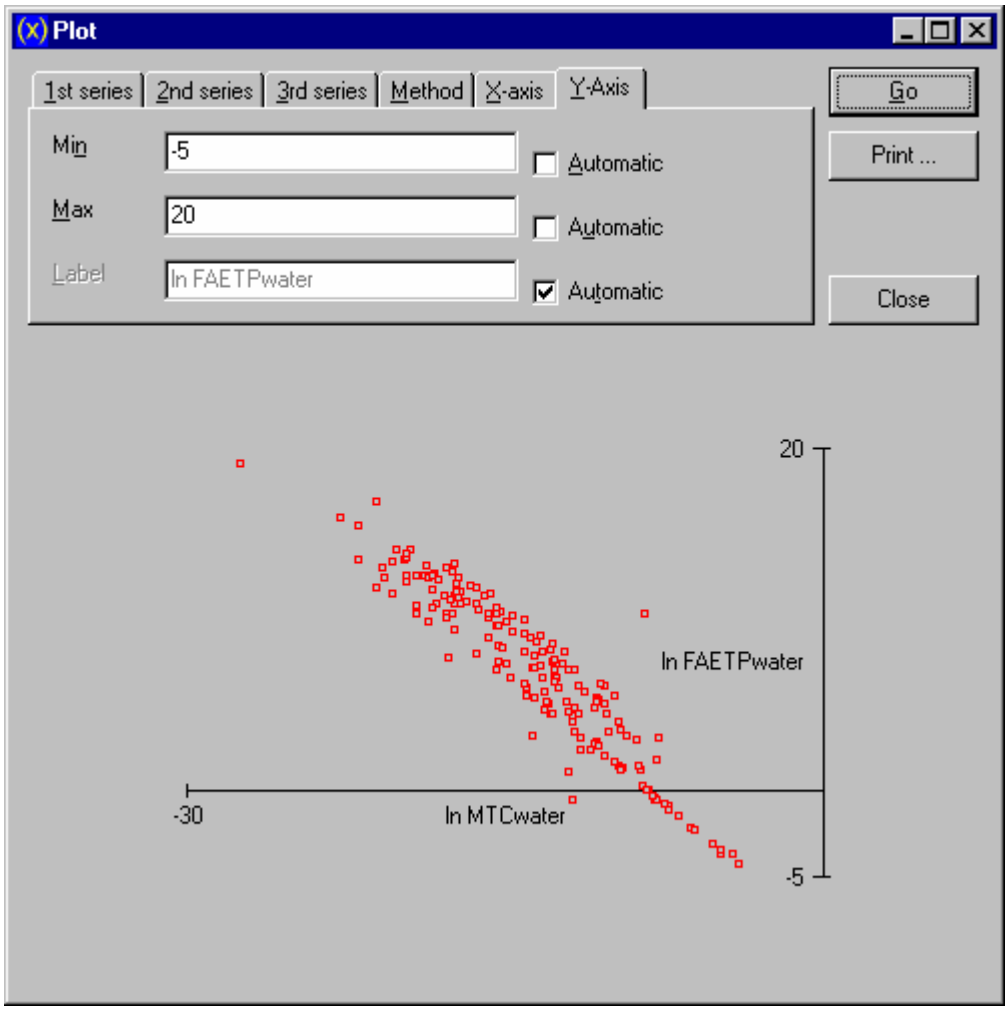

Figure 1. Relation between the logarithm of the effect parameter (MTCwater) and the logarithm of the characterisation factor (FAETPwater) for effects of emissions to fresh water on freshwater aquatic ecosystems.

Notice that, in the above, the emission and the effect compartment were the same. The fate aspect included degradation in the fresh water compartment and transport to other compartments, but not transport from other compartments. If we want to understand the multi-media behaviour of chemicals more completely, we should address relationships between, say, the air as an emission compartment and the fresh water as a target compartment. Figure 2 shows how the logarithms of these variables depend. A regression analysis gives an R^2 of 0.66. This means that the fate aspect is in this case responsible for 34% of the variance of the logarithm of the characterisation factor.

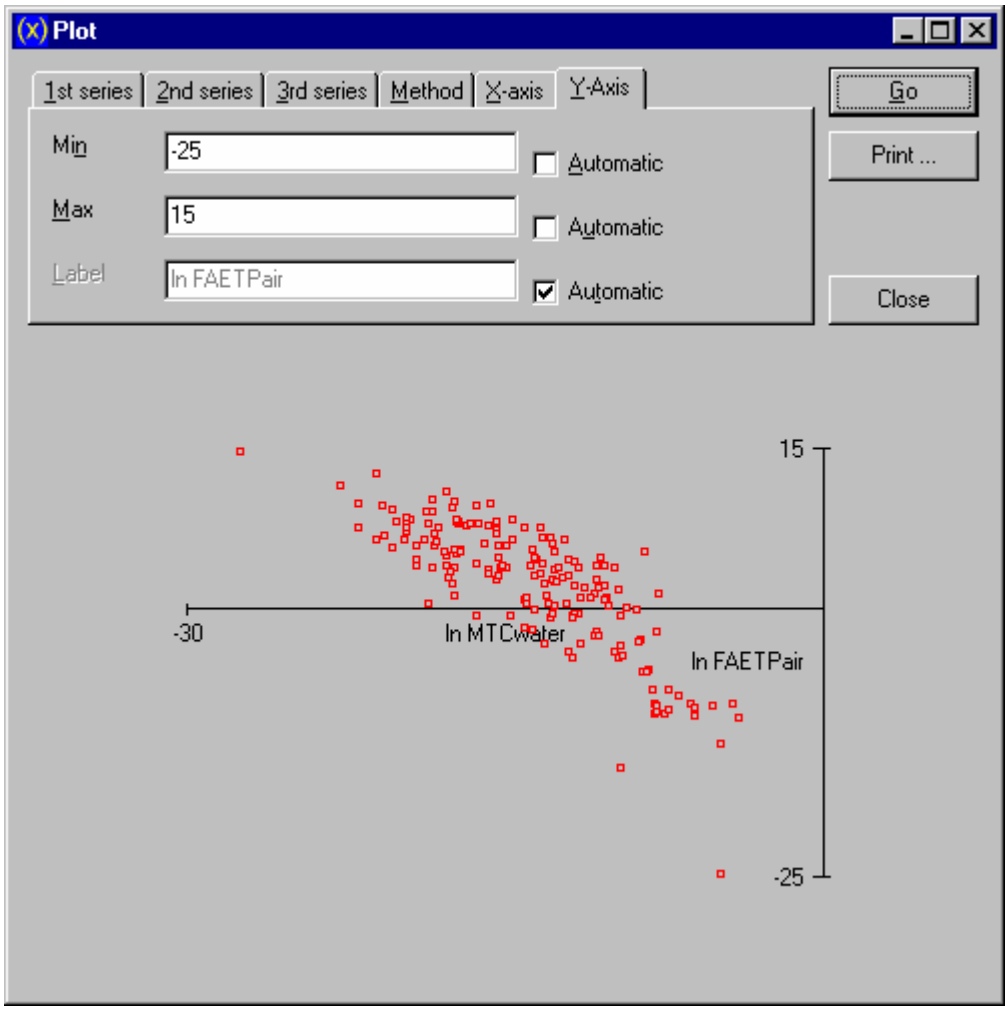

Figure 2. Relation between the logarithm of the effect parameter (MTCwater) and the logarithm of the characterisation factor (FAETPair) for effects of emissions to air on freshwater aquatic ecosystems.

It appears that a separate study of the dependence between fate parameters and the concentrations in various compartments is an important step to make.

6. Theoretical analysis of the fate aspect

The fate aspect of a characterisation model can be defined as the operation of a fate factor that transforms an emission size (E_i) into a concentration (PEC_f):

$$PEC_f = F_{fi} \times E_i \qquad (16)$$

For the emission size, it is usual to always use a unit amount of 1 kg (or 1 kg per time unit). The fate factor itself is determined by a large number of parameters. The general structure of

the fate model can be derived as a system of differential equations on the basis of the conservation of mass (Heijungs, 1995):

$$\frac{\mathrm{d}\mathbf{c}(t)}{\mathrm{d}t} = \mathbf{A}\mathbf{c}(t) + \mathbf{b}(t) \tag{17}$$

where $\mathbf{c}(t)$ is the vector of concentrations in the various compartments at time t, $\mathbf{b}(t)$ is the vector of emission sizes to these compartments, and $\mathbf{A}$ is a square matrix that contains all the multi-media transport and degradation coefficient that govern the fate of the chemical. As we are interested in the steady-state situation, the lefthand side of the differential equations will vanish, and the time parameter at the righthand side can be dropped. Thus, one obtains

$$\mathbf{0} = \mathbf{A}\mathbf{c} + \mathbf{b} \tag{18}$$

When $\mathbf{A}$ and $\mathbf{b}$ are specified, one can compute the concentration vector by matrix inversion:

$$\mathbf{c} = -\mathbf{A}^{-1}\mathbf{b} \tag{19}$$

When only one element of the emission vector $\mathbf{b}$ is non-zero, namely the one for the initial compartment i, when this emission size is 1 by definition, and when one is interested in the concentration in one specific compartment, the final compartment f, this equations reduces to

$$PEC_f = -\left(\mathbf{A}^{-1}\right)_{if} \tag{20}$$

where we have replaced $(\mathbf{c})_f$ by the more familiar PEC_f. Although only one element of a large matrix is involved in connecting the concentration in the final compartment to the initial release compartment, this element is part of an inverted matrix. It is important to realise that one element of an inverted matrix depends on all elements of the original matrix. Therefore, the structure of the original matrix is relevant for an analysis of the causal relationships in the characterisation model.

The matrix with fate coefficients $\mathbf{A}$ is built up from a number of elements[3]:

$$\left(\mathbf{A}\right)_{ij} = \begin{cases} ADV_{ij} + DIFF_{ij} & (i \neq j) \\ -\sum_k ADV_{ik} + DIFF_{ik} - kdeg_i & (i = j) \end{cases} \tag{21}$$

The essential elements here are the following:

- ADV_{ij} represents the advective flow from compartment i to compartment j;
- $DIFF_{ij}$ represents the diffusive flow from compartment i to compartment j;
- $kdeg_i$ represents the degradation in compartment i.

All these terms are in their turn composed of several other parameters. For instance,

[3] We have left out some volume-related coefficients in these equations, because these are the same across the set of substances, and would complicate the discussion. A model-specific change of such a parameter induces changes in regression constants, but not in the regression models.

$$kdeg = \frac{\ln 2}{DT_{50}} \tag{22}$$

and similar formulas can be specified for the advective and diffusive flows, often involving partition coefficients..

Let us start with the ignoring the transport from other compartments, and restrict the discussion to degradation within one single medium. The matrix **A** then has a diagonal structure:

$$\mathbf{A} = \begin{pmatrix} -kdeg_1 & 0 & \cdots \\ 0 & -kdeg_2 & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix} \tag{23}$$

and its inverse is simply

$$\mathbf{A}^{-1} = \begin{pmatrix} -1/kdeg_1 & 0 & \cdots \\ 0 & -1/kdeg_2 & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix} \tag{24}$$

This leads to

$$\ln PEC = -\ln kdeg = \ln DT_{50} - \ln \ln 2 \tag{25}$$

as a relationship within one single compartment, neglecting multi-media transport. The constant term ln ln 2 has no meaning, because it is based on the simplification to leave out the volume-related terms.

7. Second analysis with real data

So far the theoretical simplification. Figure 3 shows the situation of the fresh water compartment as emission and target compartment, with the logarithm of the (bio)degradation on the horizontal and the logarithm of the concentration on the vertical axis. In this figure the non-degradable substances (metals) have been left out.

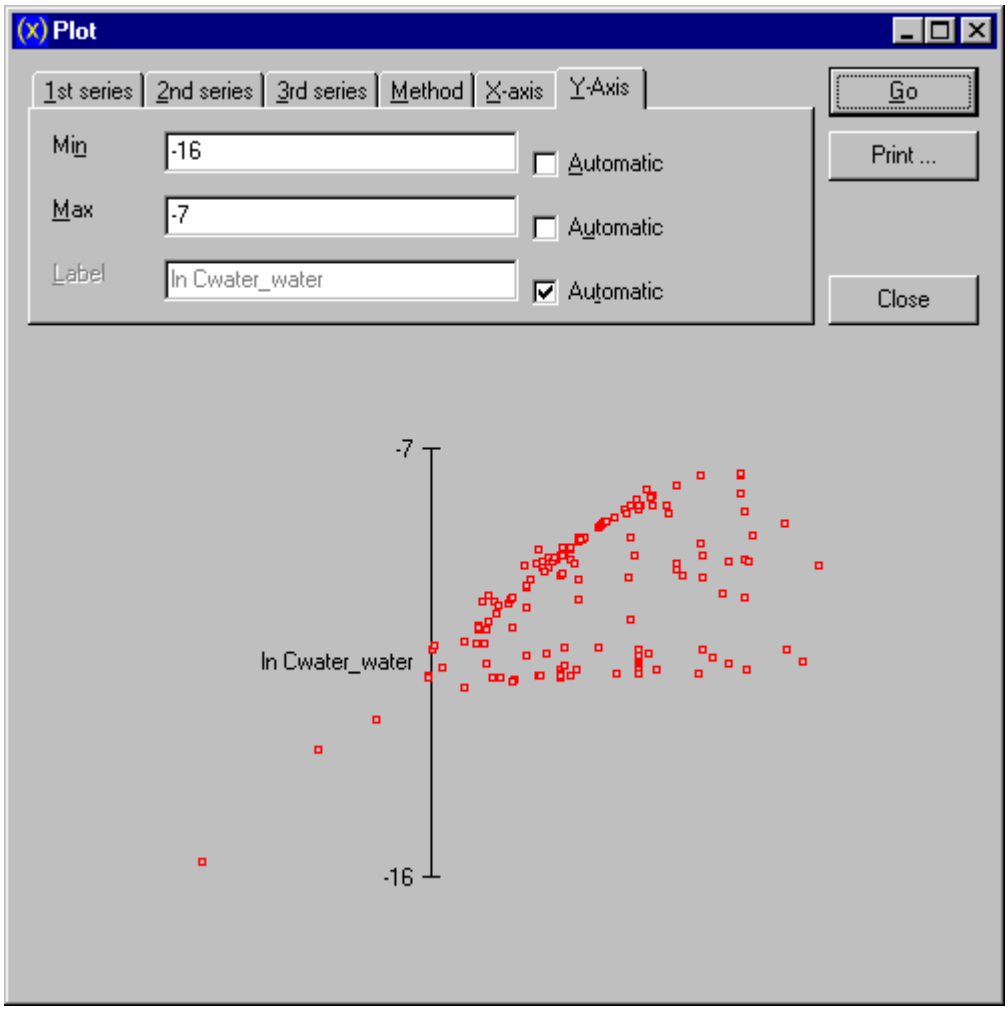

Figure 3. Relation between the logarithm of the half life in water (DT50water) and the logarithm of the concentration (Cwater_water) for emissions to water.

We see a clear straight line that appears to mark a sort of forbidden area at the left upper side of the graph. The area at the right lower side of the line contains points corresponding to substances that have migrated to other compartments before they would have been annihilated by the decay process. For similar cases, for instance for the atmospheric compartment, we see a similar straight line with a forbidden area.

A further analysis of the fate aspect then includes the multi-media transport aspect. The analytical methods employed above can no longer be used, because the structure of the **A** matrix and hence its inverse matrix becomes too complex. We have to rely on a combination of intuition, visual inspection and statistical analysis. It is clear that some sort of partition coefficient should be accountable for the multi-media transport. When we extract the transport aspect of the previous aquatic example, a regression on the logarithm of the Henry's law

constant and the logarithm of the octanol-water partition coefficient yields an R^2 of 0.42. This regression, however, is restricted to only 45 substances for which all these data items were available.

8. Some regression results

Table 1 gives an overview of some selected regression results, where metals have been left out. The total analysis included 168 substances.

Table 1. Regression results for 168 out of 181 substances (excluding metals).

Dependent variable	Independent variables	*N*	R^2
ln FAETPair	ln MTCwater, ln HENRY, ln DT50water	45	0.912
ln FAETPair	ln MTCwater, ln DT50water	157	0.725
ln FAETPwater	ln MTCwater, ln HENRY, ln DT50water	45	0.906
ln FAETPwater	ln MTCwater, ln DT50water	157	0.913
ln FAETPsoil	ln MTCwater, ln HENRY, ln DT50water	45	0.804
ln FAETPsoil	ln MTCwater, ln DT50water	157	0.695
ln TETPair	ln MTCsoil, ln HENRY, ln DT50soil	17	0.918
ln TETPair	ln MTCsoil, ln DT50soil	56	0.35
ln TETPwater	ln MTCsoil, ln HENRY, ln DT50soil	17	0.837
ln TETPwater	ln MTCsoil, ln DT50soil	56	0.219
ln TETPsoil	ln MTCsoil, ln HENRY, ln DT50soil	17	0.926
ln TETPsoil	ln MTCsoil, ln DT50soil	56	0.85
ln HTPair	ln ADI, ln TCL, ln DT50air	45	0.903
ln HTPair	ln ADI, ln DT50air	139	0.624
ln HTPwater	ln ADI, ln TCL, ln DT50water	46	0.859
ln HTPwater	ln ADI, ln DT50water	144	0.794
ln HTPsoil	ln ADI, ln TCL, ln DT50soil	46	0.852
ln HTPsoil	ln ADI, ln DT50soil	144	0.765

We see that the regression results are not too bad: in many cases we find an R^2 of 0.8 or higher for a large part of the substances on the basis of only two or three parameters.

Let us consider one of the regression models in more detail. The FAETP for emissions to water can be estimated with an R^2 of 0.913 on the basis of only the MTC for aquatic ecosystems and the half-life in water. A plot of the estimated values versus the actual values is shown in Figure 4.

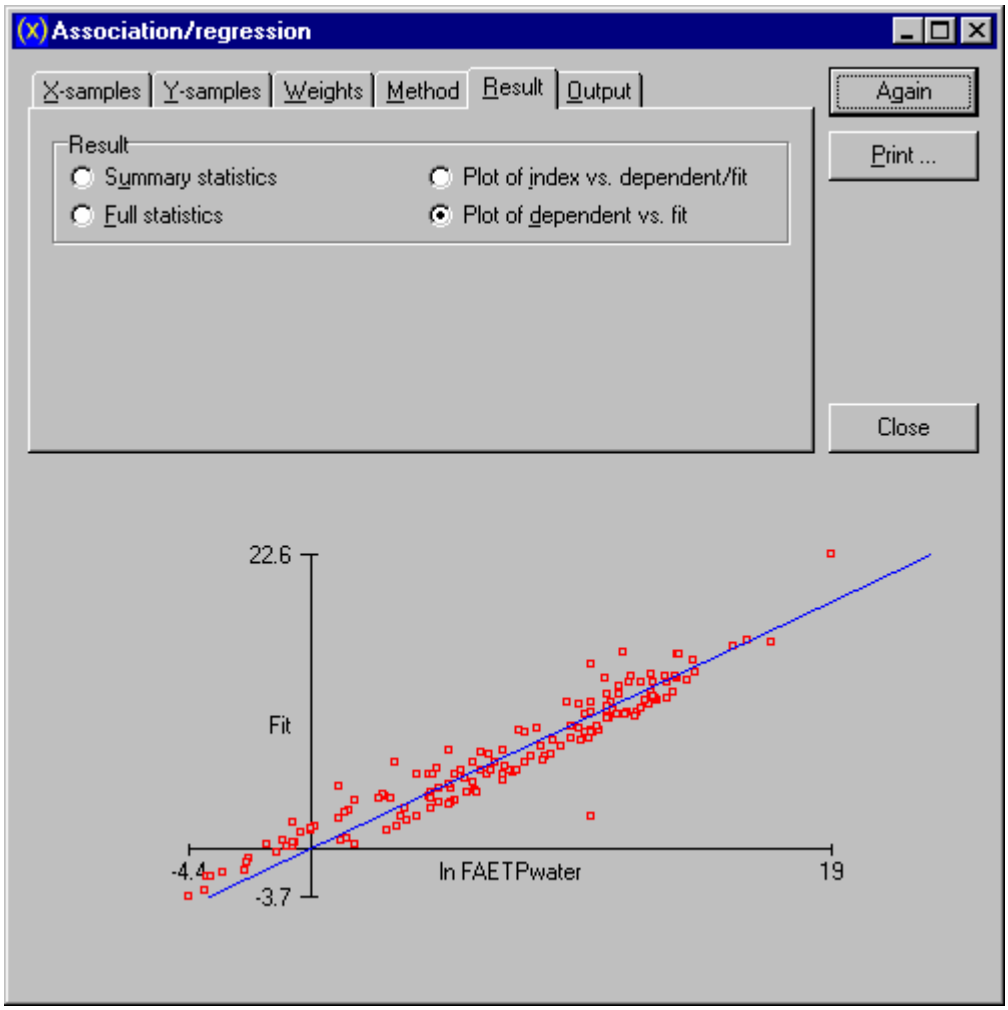

Figure 4. The data points for the logarithm of the characterisation factor (FAETPwater) for effects of emissions to fresh water on freshwater aquatic ecosystems against the regression prediction on the basis of two parameters (MTCwater and DT50water) for 157 out of 181 substances (excluding metals and other substances for which either of the two independent or the dependent variable is not available), and the regression line that indicates perfect predictions.

The regression line is given by

$$\ln FAETP_{water} = -8.8 - 1.06 \ln MTC_{water} + 0.358 \ln DT50_{water} \quad (26)$$

All three coefficients are highly significant[4]. Most data points correspond quite well to the fitted line. Remarkable is the data point that lies midway the positive horizontal axis but much too low. This is the data point for total-PAH. Also remarkable, but less so, is the data point in the right upper corner, for 2,3,7,8-TCDD. Figure 5 shows the correspondence between fit and actual value on a non-logarithmic scale.

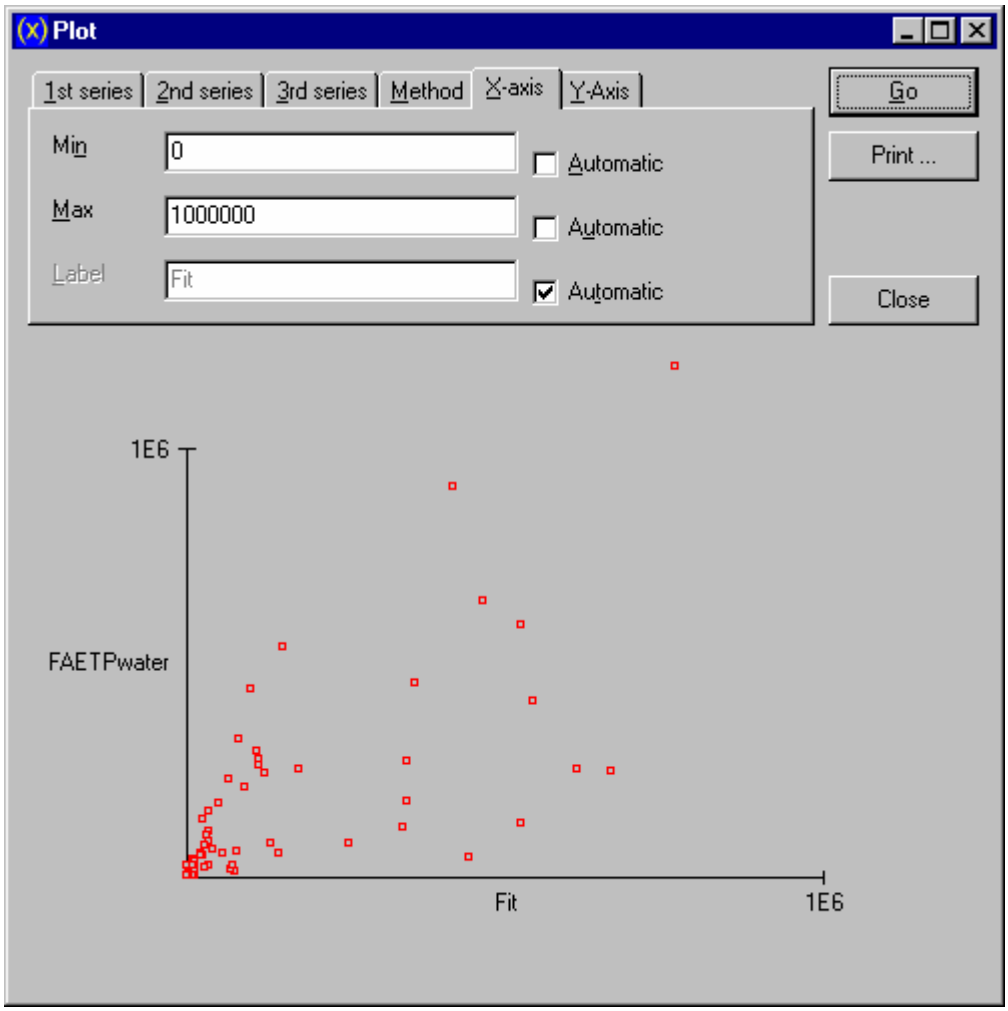

Figure 5. Same as Figure 4, but now for the backtransformation from the logarithm to the normal FAETPwater. Data points with very high values for FAETP or fit have fall outside the plotted region.

Division of the actual value by the estimated value gives ratios of which the largest part lies between 0.1 and 6. The poorest fit are obtained for DDT (0.0193) and total-PAH (ratio of

[4] The statistical details are: constant: se = 0.437, t = –20.1; ln MTC: se = 0.0266, t = –39.9; ln DT50water: se = 0.061; t = 5.87. The standard error of the estimate is 1.48.

2320). This means that for almost all 157 substances considered, the estimated FAETP is wrong by a factor of less than 6.

In the above analysis, metals were omitted, because no half-life value is available. For metals, a strategy of including the logarithm of the solid-water partition coefficient turns out to be needed, although it adds little in some of the regressions; see Table 2.

Table 2. Regression results for the metals.

Dependent variable	Independent variables	*N*	R^2
ln FAETPair	ln MTCwater, ln Kp	17	0.805
ln FAETPwater	ln MTCwater, ln Kp	17	0.817
ln FAETPsoil	ln MTCwater, ln Kp	17	0.804
ln TETPair	ln MTCsoil, ln Kp	10	0.999
ln TETPwater	ln MTCsoil, ln Kp	9	0.787
ln TETPsoil	ln MTCsoil, ln Kp	9	0.999
ln HTPair	ln ADI, ln TCL, ln Kp	14	0.883
ln HTPwater	ln ADI, ln TCL, ln Kp	14	0.654
ln HTPsoil	ln ADI, ln TCL, ln Kp	14	0.788

It is questionable to what extent a regression model for metals (or, for that sake, inorganics) is needed. The number of metals to be considered is quite limited, and these substances require only few data items. It is likely that the main need for regression model lies in the large class of organic compounds.

Prediction of characterisation factors proceeds as follows in the case of FAETP for emissions to water with two explanatory variables that was considered above. Assume that insufficient data is available to use the base model for the substance captafol. The regression equation (26) can be used to find an estimate of the ln FAETPwater, based on the ln MTCwater (–17.4) and the ln DT50water (6.35). This estimate is 11.9, with a standard error of 1.48. The 95%-confidence interval is given by [9.0, 14.8]. This agrees well with the actual value of 13.2.

9. Discussion

The material presented should not be seen as a definite study or even a definite working plan. Only a limited number of regression analyses and analytical studies were performed. The purpose was merely to investigate the possibility and usefulness of estimating characterisation factors by means of regression models. In that respect, the message of this study is reasonably positive. Only two or three parameters suffice to estimate characterisation factors, often within one order of magnitude. Equations, such as (26), can be constructed that may serve to estimate characterisation factors, given a rather limited number of data items.

On the other hand, one should understand the philosophical problems involved in the set-up of the present research. We have used a model to generate outcomes, and now use a regression model to estimate the outcomes. This, of course, is no validation whatsoever. It is therefore problematic to speak about the degree to which an estimate is "correct". It is correct with respect to the model that is used, i.c. USES-LCA.

For this feasibility study, the USES-LCA model with its data set for 181 substances was used (Huijbregts, 1999). When a different model is used, or when a different data set is used, the results of the study will change. There are two ways in which changes may occur:

- the regression coefficients and goodness-of-fit of a particular regression model may change;
- the choice of the "best" regression model may change (other choice of parameters, e.g. DT50 soil, Kow etc.).

When a different model and/or data set is selected, part of the work reported should be carried out again. However, the theoretical considerations reported will remain to be valid, and the findings on the most critical variables will probably also not change.

Related to this point is the dynamic aspect that is involved in the use of an estimation method while at the same time providing a detailed method that may be used if more data is available. On the basis of a model and data available now, regression equations may be provided to estimate characterisation factors for substances that are not covered by the current databases. When the required data items become available in due time, there is no need to use the estimated (Simple Base Model) characterisation factor anymore, but the "correct" characterisation factor can be calculated with the detailed Base Model. This will lead to an updated, improved characterisation factor for that substance. But it will also be possible to redo the regression analysis, and to update the regression coefficients (or even to update the optimal model specification) on the basis of the newly acquired knowledge, which is a

relatively easy job to do. In principle, this may affect, however, all hitherto calculated estimates of characterisation factors. Therefore, a Bayesian perspective on the incorporation of new information and the updating of beliefs – and thus characterisation factors – should be developed within the OMNIITOX IS to deal with this.

In Table 1, there are two alternative regression models for each characterisation factor. For instance, the FAETPair can be estimated on the basis of either two or three independent variables. The three-variable estimate is much better than the two-variable estimate (R^2 = 0.912 versus 0.725), but it is based on much fewer substances (N = 45 versus 157). It is therefore also much less applicable to new chemicals. It appears that there may be a need for more than just a detailed Base model and an estimation model (Simple Base Model), but that several estimation models are needed. for instance, one estimation model should be applicable for almost every chemical, so on the basis of data that are always available. The price of such a model is that its estimates are quite inaccurate. For chemicals for which some more data are available, a second estimation model should be available. It is less widely applicable, but it produces better results. In this way a hierarchy of models could be developed, from extremely simple to detailed via moderately advanced. Figure 6 shows the regression results for an estimation on the basis of only the molecular weight, a parameter that is almost always available. The R^2 is almost 0.3, suggesting a not too good fit, but at least an estimate that is far better than tossing a coin or no estimate at all. It is an interesting point for discussion how the trade-off between data availability and predictive accuracy should be made, especially in the context of the possibility of a hierarchical structure of several regression models. If the latter is chosen for, the need for a Selection Method should be re-considered of course.

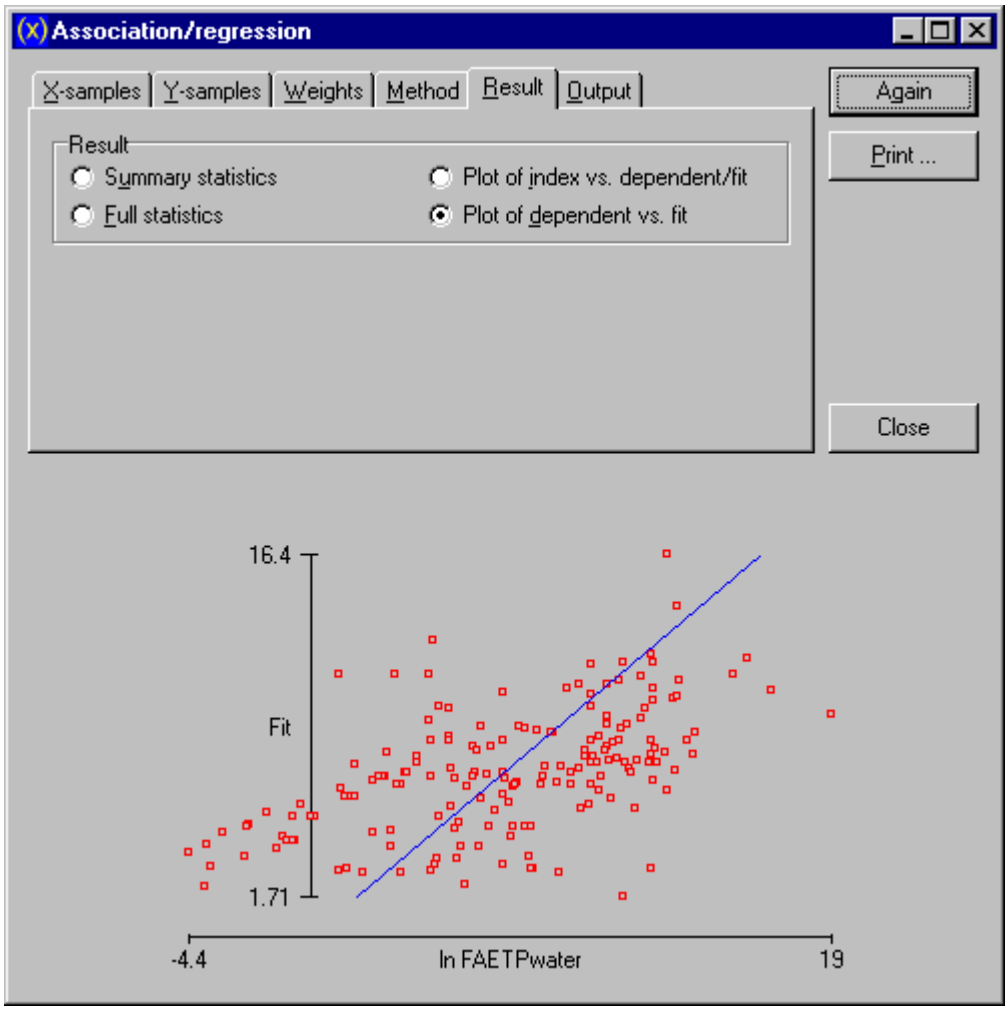

Figure 6. The data points for the logarithm of the characterisation factor (FAETPwater) for effects of emissions to fresh water on freshwater aquatic ecosystems against the regression prediction on the basis of only the molecular weight for 176 out of 181 substances (excluding substances for which either the independent or the dependent variable is not available), and the regression line that indicates perfect predictions.

One development that should be considered is the possibility to construct within the OMNIITOX IS a continuum of simplified base models, that is integrated with the base model itself. When a user wants to calculate a characterisation factor, he uses the OMNIITOX IS, and is prompted to enter all substance-related parameters that he knows. When this is the molecular weight, the IS comes up with a prediction with a large uncertainty range. When the user also enters the melting point, a different and more accurate prediction is returned. When the user enters all parameters that are needed for the base model, the simplification is automatically skipped. Because there are hundreds of parameters, and a user may have any number and combination of these, there may literally be millions of regression models. A good information

system may calculate the coefficients, the prediction and the uncertainty of prediction for one such model in a second or less, it may be a good strategy to use this type of online simplification at request.

How much time is needed to develop an estimation model (Simple Base Model) on the basis of the final OMNIITOX data set and Base Model, depends on the number of toxicity potentials and emission compartments to be covered (e.g. HTPair, HTPwater, HTPseawater, ..., TETPindustrialsoil), the number of levels within a possible hierarchical structure of regression models (coverage of substances, availability of data and accuracy of estimations), and the types of substances to be captured (organics, inorganics, metals, ...).

Besides estimates of the regression coefficients, the regression framework offers standard errors for the coefficients and a standard error of the estimate as well. These can be used to indicate confidence intervals for any estimate made by the regression equation. If, in addition, standard errors for the results of the detailed Base model are available, a weighted regression can be performed with the inverse of the squares of these standard errors as weights. It may be expected that certain data points (characterisation factors) are much more uncertain than others, so regression results may well change upon the introduction of error-weighted regression. When no such information is available, an alternative strategy would be to use as a default that the standard error of a characterisation factor is proportional to its value. This would at least reduce the influence of extremely high values, such as for 2,3,7,8-TCDD.

Acknowledgements

The author gratefully acknowledges the financial support of the European commission through the "Sustainable and Competitive Growth"-research program given to the OMNIITOX-project (Operational Models aNd Information tools for Industrial applications of eco/TOXicological impact assessments, EC Project number: G1RD – CT – 2001 – 00501) which consists of 11 partners: Antonio Puig, S.A., Spain; ESA and IEI, Chalmers University of Technology, Sweden; European Chemicals Bureau, Joint Research Centre, Italy; Ecole Polytechnique Federale de Lausanne, Switzerland; CML, Leiden University, Netherlands; The Procter & Gamble Company, Belgium; Randa Group S.A., Spain; Stora Enso Oyj, Sweden; IPL, Technical University of Denmark, Denmark; IER, University of Stuttgart, Germany; Volvo Technological Development, Sweden.
Discussions and comments by Arjan de Koning and Jeroen Guinée are appreciated and acknowledged.

References

Dobson A, 1983. Introduction to statistical modelling. Chapman and Hall, London.

Draper NR, Smith H, 1998. Applies regression analysis. Third edition. John Wiley & Sons, Inc.

Greene WH, 1997. Econometric analysis. Third edition. Prentice-Hall International, Inc., 1997.

Heijungs R, 1995. Harmonization of methods for impact assessment. Environmental Science & Pollution Research 2:4 (1995), 217-224.

Huijbregts MAJ,1999. Priority assessment of toxic substances in the frame of LCA. Development and application of the multi-media fate, exposure and effect model USES-LCA. Interfaculty Department of Environmental Science, University of Amsterdam, Amsterdam.

Linders JBHJ, Jager DT, 1998. Uniform system for the evaluation of substances (USES). Version 2.0. RIVM, Bilthoven.

Lyman WJ, Reehl WF, Rosenblatt DH, 1990. Handbook of chemical property estimation methods. American Chemical Society, Washington.

Meent D van de, Struijs J, Sijm D, 2002. Simple recipe for calculating multi-substance toxic pressure as indicator of the ecological impact of toxic substances. In: Abstract proceedings of SETAC Europe 12th Annual Meeting, 12-16 May 2002, Vienna, 69.