



## Using multiple regression in estimating (semi) VOC emissions and concentrations at the European scale

Patrik Fauser<sup>1</sup>, Marianne Thomsen<sup>1</sup>, Alberto Pistocchi<sup>2</sup>, Hans Sanderson<sup>1</sup>

<sup>1</sup> National Environmental Research Institute, Aarhus University, Department of Policy Analysis, Roskilde, Denmark

<sup>2</sup> European Commission, DG Joint Research Centre, Ispra (VA) Italy

### ABSTRACT

This paper proposes a simple method for estimating emissions and predicted environmental concentrations (PECs) in water and air for organic chemicals that are used in household products and industrial processes. The method has been tested on existing data for 63 organic high-production volume chemicals available in the European Chemicals Bureau risk assessment reports (RARs). The method suggests a simple linear relationship between Henry's Law constant, octanol-water coefficient, use and production volumes, and emissions and PECs on a regional scale in the European Union. Emissions and PECs are a result of a complex interaction between chemical properties, production and use patterns and geographical characteristics. A linear relationship cannot capture these complexities; however, it may be applied at a cost-efficient screening level for suggesting critical chemicals that are candidates for an in-depth risk assessment. Uncertainty measures are not available for the RAR data; however, uncertainties for the applied regression models are given in the paper. Evaluation of the methods reveals that between 79% and 93% of all emission and PEC estimates are within one order of magnitude of the reported RAR values. Bearing in mind that the domain of the method comprises organic industrial high-production volume chemicals, four chemicals, prioritized in the Water Framework Directive and the Stockholm Convention on Persistent Organic Pollutants, were used to test the method for estimated emissions and PECs, with corresponding uncertainty intervals, in air and water at regional EU level.

### Keywords:

VOC

Emissions

PEC

Multiple regression

ECB reports

### Article History:

Received: 30 March 2010

Revised: 21 June 2010

Accepted: 25 June 2010

### Corresponding Author:

Patrik Fauser

Tel: +45-46301236

Fax: +45-46301114

E-mail: paf@dmu.dk

© Author(s) 2010. This work is distributed under the Creative Commons Attribution 3.0 License.

doi: 10.5094/APR.2010.017

### 1. Introduction

Chemicals in the environment arising from human activities pose issues concerning human health and environmental risks (EC, 2005; Karjalainen, 2005). Emission quantification at different life-cycle stages of chemicals and chemical containing products is the natural starting point of exposure assessment and is the key to any modeling effort aimed at deriving predicted environmental concentrations (PECs) in air, water and soil. Emissions are often related to production and use of chemicals and the complex nature of chemical emission patterns makes quantification of emissions and consequently environmental concentrations somewhat uncertain. In many cases the predominant uncertainties in an exposure assessment are indeed related to the uncertainties of emission inventories.

There are many ways to perform an emission inventory. The emission inventory guidebook prepared by the United Nations Economic Commission for Europe/European Monitoring and Evaluation Programme (UNECE/EMEP) Task Force on Emissions Inventories and Projections (UNECE, 2005) to support reporting under the UNECE Convention on Long-Range Transboundary Air Pollution (LRTAP) (UNECE-CLRTAP, 2010) and the EU directive on national emission ceilings 2001/81/EC provides a comprehensive state-of-the-art methodological guide for estimating atmospheric emissions (EC, 2001).

Comprehensive emission and concentration estimates have been made, for single priority chemicals in the context of risk

assessments, by the EU member states and coordinated by the formerly known European Chemicals Bureau (ECB); this work is accessible in the form of publicly available risk assessment reports (RARs) (EC Chemicals, 2010). Priority chemicals or groups of chemicals require immediate attention because of their potential effects on man or the environment (Lerche et al., 2002; Thomsen et al., 2008). They have been found on the basis of information submitted by manufacturers and importers, the European Commission, in consultation with Member States. Four priority lists have been defined, comprising in all 141 chemicals (EC Priority Lists, 2010). To date full risk assessments presented in draft or finalized reports are available only for 78 chemicals of the priority lists (EC Chemicals, 2010). The ECB RARs include data, modeling results and expert judgments, based on the information in IUCLID (EC-IUCLID, 2010). IUCLID is a tool for data collection and evaluation within the EU-Risk Assessment Programme and comprises the largest set of uniformly reported data for organic compounds and metals that are directly applicable for the EU. Volatile Organic Compounds (VOCs) and semi-VOCs are an important sub-category due to their abundant use in chemical industry and consumer products and capability of long-range transport in the environment.

In accordance with Directive 67/548/EEC (EC, 1967) and Regulation (EEC) 93/793 (EC, 1993), exposure related information must be provided for notified new chemicals and for priority existing chemicals, and particularly information on proposed use. When neither measured nor estimated exposure data are provided by the responsible industry (i.e. the notifier of a new chemical,

which can be the manufacturer or importer of a priority existing chemical, respectively), the information on proposed use will be useful to competent authorities for developing emission scenarios. They are in most cases based on more in-depth studies of the environmental emission of chemicals used in the different industrial categories, as defined in the European Commission Technical Guidance Document on risk assessment (TGD) (EC, 2003a). So far documents concerning emission scenarios have been developed for 10 out of 16 industrial categories (EC, 2003b). The emission of a chemical at different stages of its life cycle should thus be estimated by order of preference from:

- (1) Specific information for the given chemical (e.g. from producers, product registers or open literature),
- (2) Specific information from the emission scenarios document, available for several industrial categories,
- (3) Emission factors as included in emission tables in the TGD where the emission is given as a fraction of the produced or used amount.

Although information from industry is preferable, it is often the case that no such information is publicly available. For this reason, the emission estimates in the existing RARs are obtained as a result of various possible combinations of expert judgment and empirical assumptions, based on generic scenarios as defined in the TGD (EC, 2003b). For each chemical, environmental concentrations are then calculated from emissions using mathematical models such as EUSES (EC, 2004), and compared with monitored values. Emission estimates are thus a key parameter for estimating fate and exposure in complex models that involves a wide range of fundamentally different input parameters with varying uncertainties, making a risk assessment a time consuming task requiring an extensive effort in data collection and evaluation (Penman et al., 2001).

In December 2006 the Council of Ministers adopted a new EU regulatory framework for Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) (EU, 2007). This has warranted new methodologies, tools and technical guidance for the practical implementation of REACH with an important purpose to facilitate and speed up the risk assessment process. As data on emissions and concentrations of chemicals in specific compartments are not estimated routinely and, in general, there is no obvious functional relationship between chemical specific parameters and emissions and concentrations that can be derived from the methodology described in the TGD (EC, 2003b), a challenge can be faced when predicting emissions and concentrations of other chemicals using actually available information, in a proactive and cost-effective way.

In this work, we demonstrate how data presented in the RARs, can be used to estimate spatial emission patterns and concentrations of chemicals in a quick and reliable way for screening level applications. The method thus bypasses some complex steps that are inherent in emission and exposure modeling. We perform a critical test of the applicability of the estimation model and the model domain is evaluated and tested towards existing data. We focus on VOCs and semi-VOCs, which cover a large group of chemicals predominantly found in industrial processes and in many consumer products. Of the 78 risk assessed chemicals, those containing metals, e.g. cadmium and zinc, have been excluded in the analysis, mostly due to lack of applicability of vapor pressure. This leaves a selected subset of data for 63 chemicals that have been subjected to regression analysis by purpose of unveiling data and describing simple relationships between production and use quantities, physico-chemical properties and emissions and concentrations at regional level. In the last part of the paper we illustrate how the method can be used to predict emissions and concentrations in air and in surface water for currently non-assessed chemicals that are prioritized in the Water Framework

Directive (EC, 2008) and the listed substances in the Stockholm Convention on Persistent Organic Pollutants (SC-POPs, 1997).

## 2. Methods and Applications

A method aimed at estimating emissions and PECs for VOCs and semi-VOCs has been designed and tested. It comprises two types of tools, namely correlation or pattern recognition followed by regression models, including Partial Least Square Regression (PLS-R) and Multiple Linear Regression (MLR).

The basic information that is present in finalized and draft RARs includes:

- (1) Produced and used amount of specific chemical in the EU,
- (2) Names (locations) of producers and importers (no link is available between company name and produced/processed amount of chemicals),
- (3) Products and uses that the chemicals are associated with; these are categorized as “use in closed systems”, “use resulting in inclusion into or onto matrix”, “non-dispersive use” and “wide-dispersive use”,
- (4) Emissions to wastewater, air, soil, surface water, sea/estuaries and landfills on a local (non-dispersive and wide-dispersive), regional and continental scale,
- (5) Predicted Environmental Concentrations (PECs) for the same compartments and spatial scales as above,
- (6) Physico-chemical properties,
- (7) Ecotoxicological and human toxicological parameters,

Items (1), (4), (5), and (6) are relevant for predicting emissions and PECs in air and water at screening level assessments.

In this work we focus on 63 chemicals; their emissions and PECs on a regional (EU) scale. The chemicals are used as training set for calibration and cross-validation of regression models based on PLS-R and MLR. The analysis is made in two steps:

Step 1: Predictor parameters (PP), i.e. parameters that are entered in the models, are produced amounts in the EU, used amounts in the EU together with the physico-chemical parameters  $\log H$  and  $\log K_{ow}$ , compiled and explained in Table 1. The target parameters (TP) that are used to calibrate the models, describe regional conditions, represented by regional emissions to air and water (summed emissions to surface water and waste water). The regional scale is relevant for chemicals that enter the environment through diffuse use. It takes into account the further distribution and fate of the chemical upon release. Furthermore PEC regional is assumed to be a steady-state background concentration (EC, 2003a).

Step 2: PPs are produced and used amounts,  $\log H$ ,  $\log K_{ow}$ , 1<sup>st</sup> order degradation rate in surface water ( $k_{sw}$ ) and regional emissions to air and water. TPs comprise regional PECs to air and water. Data on emissions and PECs at regional conditioned scenarios are compiled in Table 2.

Prior to selecting PPs for optimal modeling of TPs, the PPs and TPs were log-transformed to approximate normal distribution of data. Normal distribution can be assumed when skewness  $< \pm 2 \times$  Standard Error of skewness and kurtosis  $< \pm 2 \times$  Standard Error of Kurtosis (SYSTAT, 1997). The normal distribution criteria are met for all selected parameters in Table 1. Parameters  $\log H$  and regional emission to air showed slightly skewed distributions by having longer right tail and left tail, respectively, than those of a normal distribution (Mateu, 1997). The latter are nonetheless included in the analysis. Furthermore, parameters have been auto scaled, i.e. the mean subtracted and divided by standard deviation, to obtain equal variances and mean zero, and approximate homoscedastic noise between variables (Mateu, 1997; Larsen, 2006).

**Table 1.** Production and use amounts, Henry's Law constants (*H*), octanol–water coefficients (*K<sub>ow</sub>*) and 1<sup>st</sup> order degradation rates in surface water (*k<sub>sw</sub>*) from risk assessment reports (training chemicals). These parameters are used as predictor parameters in the PLS and MLR/LR models

CAS no.	Substance Name	Production in EU (t/y)	Use in EU (t/y)	log <i>H</i>	log <i>K<sub>ow</sub></i>	<i>k<sub>sw</sub></i> (1/d)
100-42-5	Styrene	3 743 000	90 000	2.36	3.02	0.047
101-77-9	4,4'-methylenedianiline	432 000	436 000	-6.35	1.59	0.00036
103-11-7	2-ethylhexyl acrylate	70 000	14 494	2.36	3.88	0.047
106-46-7	1,4-dichlorobenzene	25 500	1 892 000	2.39	3.38	0.046
106-99-0	buta-1,3-diene	1 892 000	100 000	3.86	1.99	n.a.
107-02-8	Acrylaldehyde	100 000	1 250 000	0.785	-0.89	0.047
107-13-1	Acrylonitrile	1 250 000	5 641	0.982	0.25	0.00462
107-64-2	Dimethyldioctadecylammonium chloride	5 651	17 250	-10	3.8	0.0047
108-88-3	Toluene	16 750 000	16 450 000	2.72	2.65	0.0237
108-95-2	Phenol	1 829 100	187 700	-1.65	1.47	0.05
109-66-0	Pentane	55 000	50 000	5.02	3.45	0.047
110-65-6	but-2-yne-1,4-diol	185 000	900 000	-4.69	0.73	0.047
110-82-7	Cyclohexane	880 000	9 000	4.17	3.44	0.046
110-85-0	Piperazine	10 000	5 000	-1.59	-1.24	0.00462
111-77-3	2-(2-methoxyethoxy)ethanol	20 000	108 000	-2.53	-0.682	0.047
112-34-5	2-(2-butoxyethoxy)ethanol	46 600	46 630	-2.35	0.56	0.047
1163-19-5	bis(pentabromophenyl)ether	0	15 000	1.64	6.27	n.a.
117-81-7	bis-(2-ethylhexyl)phthalate	595 000	476 000	6.64	7.5	0.0139
120-82-1	1,2,4-trichlorobenzene	7 000	1 400	2.37	4.05	0.0047
123-91-1	1,4-dioxane	2 500	2 000	-0.453	-0.29	n.a.
127-18-4	Tetrachloroethylene	164 000	10 120	3.32	2.53	n.a.
141-97-9	Ethyl Acetoacetate	10 000	8 210	-0.982	0.25	0.047
1570-64-5	4-chloro- <i>o</i> -cresol	15 000	49 975	0.217	3.09	0.03
1634-04-4	tert-butyl methyl ether	3 030 000	2 313 000	1.75	1.06	0.00462
25154-52-3	Nonylphenol	73 500	78 500	1.04	4.48	0.00462
26447-40-5	Methylenediphenyl Diisocyanate	790 000	689 000	1.79	4.5	0.116
26-761-40-0	di-"isodecyl" phthalate	280 000	200 000	2.05	8.8	0.014
28553-12-0	di-"isononyl" phthalate	185 200	107 200	1.62	8.8	0.014
32534-81-9	Diphenyl ether, pentabromo derivative	0	1 642 500	1.04	6.57	n.a.
32536-52-0	Diphenyl ether, octabromo derivative	4 000	450	1.02	6.29	n.a.
60-00-4	Edetic acid (EDTA)	53 900	31 114	-10	-5.01	0.035
62-53-3	Aniline	530 000	560 550	-0.823	0.9	0.046
67-66-3	Trichloromethane (chloroform)	302 800	271 000	2.45	1.97	0.0312
67774-74-7	Benzene, C10-13 alkyl derivatives	450 000	280 000	1.97	8.31	0.047
71-23-8	Propan-1-ol	5 000	30 100	-0.931	0.34	0.047
71-43-2	Benzene	7 247 000	10 000	2.63	2.13	0.047
75-05-8	Acetonitrile	10 000	138 000	0.463	-0.34	0.00462
75-56-9	Methyloxirane	1 445 000	1 495 000	0.938	0.055	0.0046
75-91-2	tert-butyl hydroperoxide	750 000	14 200	0.386	0.7	0.00462
7664-39-3	Hydrogen fluoride	245 000	245 000	0.315	-1.4	n.a.
7722-84-1	Hydrogen peroxide	750 000	670 000	-1.99	-1.5	0.139
77-78-1	Dimethyl sulphate	25 000	20 000	-0.533	0.16	0.047
79-01-6	Trichloroethylene	138 000	100 100	3.01	2.29	1.39x10 <sup>-6</sup>
79-06-1	Acrylamide	100 000	838 300	-4.52	-1	0.047
79-10-7	Acrylic acid	810 000	21 583.2	-1.56	0.46	0.047
79-11-8	Chloroacetic acid	145 000	120 000	-3.70	0.2	0.0462
79-20-9	Methyl acetate	30 000	162 351.5	0.808	0.18	0.047
79-41-4	Methacrylic acid	40 000	10 900	-1.06	0.93	0.047
80-05-7	4,4'-isopropylidenediphenol	700 000	690 000	-5.39	3.4	0.047
80-62-6	Methyl methacrylate	470 000	388 690	1.41	1.38	0.047
81-14-1	4'-tert-butyl-2',6'-dimethyl-3',5'-dinitroacetophenone	0	35	-1.59	4.3	n.a.
81-15-2	5-tert-butyl-2,4,6-trinitro- <i>m</i> -xylene	0	67	-1.22	4.9	n.a.
84-74-2	Dibutyl phthalate	26 000	1 100	-0.568	4.57	0.047
85535-84-8	Chloro alkanes, C10-13	15 000	13 208	1.26	6	1.66x10 <sup>-10</sup>
85535-85-9	Chloro alkanes, C14-17	102 500	65 300	0.691	7	n.a.
85-68-7	Benzyl butyl phthalate	45 000	36 000	-0.903	4.84	0.0462
88-12-0	1-vinyl-2-pyrrolidone	30 000	0	-2.87	0.4	0.0462
90-04-0	<i>o</i> -anisidine	1 000	1 000	-1.54	1.18	0.0125
91-20-3	Naphthalene	200 000	40 950	1.56	3.55	0.00462
95-76-1	3,4-dichloroaniline	12 000	4 100 000	-1.30	2.7	0.039
98-01-1	2-furaldehyde	7 000	42 350	-0.751	0.41	0.047
98-82-8	Cumene	4 100 000	3 742 000	3.00	3.55	0.00462
994-05-8	2-methoxy-2-methylbutane	250 000	287 000	1.92	1.55	6.93x10 <sup>-7</sup>

n.a.: not available

**Table 2.** Emissions to water and to air and predicted environmental concentrations (PECs) in water and air from risk assessment reports (training chemicals). Emissions are used as target parameters in Step 1 and as predictor parameters in Step 2. PECs are used as target parameters in Step 2

CAS no.	Substance Name	Emission reg Water (t/y)	Emission reg Air (t/y)	PEC reg Water ( $\mu\text{g/L}$ )	PEC reg Air ( $\mu\text{g/m}^3$ )
100-42-5	Styrene	282	2 615	0.052	0.034
101-77-9	4,4'-Methylenedianiline	84.9	n.a.	0.01	$4.6 \times 10^{-15}$
103-11-7	2-Ethylhexyl acrylate	38.8	10.6	0.0058	0.00079
106-46-7	1,4-Dichlorobenzene	48.65	782.5	0.04	0.074
106-99-0	Buta-1,3-Diene	120	1 435	0.073	0.0257
107-02-8	Acrylaldehyde	5.1	672	0.02	0.03
107-13-1	Acrylonitrile	4.3	191	2.81	0.261
107-64-2	Dimethyldioctadecylammonium chloride	9.6	n.a.	5.1	n.a.
108-88-3	Toluene	9 200	39 240	6.26	6.92
108-95-2	Phenol	675	9 683	2.41	0.026
109-66-0	Pentane	6.81	3 745	0.00037	0.32
110-65-6	But-2-Yne-1,4-Diol	76.94	0.0087	0.28	$2.6 \times 10^{-9}$
110-82-7	Cyclohexane	737.8	6 895	0.05	0.35
110-85-0	Piperazine	2.4	2.7	0.59	$9.5 \times 10^{-6}$
111-77-3	2-(2-Methoxyethoxy)ethanol	986.1	88.2	10	0.002
112-34-5	2-(2-butoxyethoxy)ethanol	2280	585	10	0.013
1163-19-5	bis(pentabromophenyl)ether	126.2	2.9	0.094	0.0054
117-81-7	bis-(2-ethylhexyl)phthalate	598.7	54.6	2.2	0.0075
120-82-1	1,2,4-trichlorobenzene	11.94	0.317	0.00952	$5.46 \times 10^{-4}$
123-91-1	1,4-dioxane	130.2	289.5	1.3	0.02
127-18-4	Tetrachloroethylene	16.06	12 090	0.011	0.88
141-97-9	Ethyl Acetoacetate	3.6	5	0.04	$3.7 \times 10^{-8}$
1570-64-5	4-chloro-o-cresol	39.75	4.65	n.a.	n.a.
1634-04-4	tert-butyl methyl ether	711	14 000	1.5	0.75
25154-52-3	Nonylphenol	95.7	31.8	0.6	0.00314
26447-40-5	Methylenediphenyl diisocyanate	n.a.	0.7138	0.00138	0.000206
26-761-40-0	di-"isodecyl" phthalate	393.5	22.4	1.8	0.0007
28553-12-0	di-"isononyl" phthalate	164.5	13.5	0.7	0.0004
32534-81-9	Diphenyl ether, pentabromo derivative	0.5746	4.3	0.0015	0.00035
32536-52-0	Diphenyl ether, octabromo derivative	0.692	0.775	0.0036	0.00014
60-00-4	Edetic acid (EDTA)	2 895	n.a.	95	n.a.
62-53-3	Aniline	12	21	0.13	0.00022
67-66-3	Trichloromethane (chloroform)	124.1	992.8	0.828	0.145
67774-74-7	Benzene, C10-13 alkyl derivatives	87.9	0	0.07	n.a.
71-23-8	Propan-1-ol	1 068	2 108	8.59	0.0945
71-43-2	Benzene	2 585	18 290	0.275	1.54
75-05-8	Acetonitrile	422.1	5 246	2.41	0.4
75-56-9	Methyloxirane	17	75	0.067	0.0054
75-91-2	tert-butyl hydroperoxide	19.34	49.59	0.261	0.00336
7664-39-3	Hydrogen Fluoride	n.a.	n.a.	n.a.	n.a.
7722-84-1	Hydrogen Peroxide	1 752	216.3	3	0.00223
77-78-1	Dimethyl Sulphate	n.a.	n.a.	n.a.	n.a.
79-01-6	Trichloroethylene	429.6	5 220	0.35	0.47
79-06-1	Acrylamide	9.12	0.066	0.05	$3.56 \times 10^{-8}$
79-10-7	Acrylic Acid	36.3	54	0.4	0.002
79-11-8	Chloroacetic acid	317.8	14.4	0.068	$2.38 \times 10^{-4}$
79-20-9	Methyl acetate	195	1 328	0.85	0.13
79-41-4	Methacrylic acid	13	4	0.14	0.0001
80-05-7	4,4'-isopropylidenediphenol	5.371	2.135	0.12	$2.08 \times 10^{-7}$
80-62-6	Methyl methacrylate	52.1	2 380	0.14	0.05
81-14-1	4'-tert-butyl-2',6'-dimethyl-3',5'-dinitroacetophenone	1.05	n.a.	0.11	0.00001
81-15-2	5-tert-butyl-2,4,6-trinitro-m-xylene	2.01	n.a.	0.18	$3.9 \times 10^{-5}$
84-74-2	Dibutyl phthalate	94.5	111	0.4	0.006
85535-84-8	Chloro alkanes, C10-13	204.5	0.03942	0.33	0.012
85535-85-9	Chloro alkanes, C14-17	85.06	17.14	0.745	0.00612
85-68-7	Benzyl butyl phthalate	n.a.	n.a.	n.a.	n.a.
88-12-0	1-vinyl-2-pyrrolidone	7.365	15.84	0.0388	$3.52 \times 10^{-5}$
90-04-0	o-anisidine	n.a.	n.a.	n.a.	n.a.
91-20-3	Naphthalene	14.79	4 394	0.03	0.14
95-76-1	3,4-dichloroaniline	6.8	0.0037	0.08	$2.1 \times 10^{-6}$
98-01-1	2-furaldehyde	n.a.	n.a.	n.a.	n.a.
98-82-8	Cumene	615	1 242	0.0003	0.07
994-05-8	2-methoxy-2-methylbutane	1 246	7 490	0.52	0.34

n.a.: not available

Regression models can be made for all forms of emissions to environmental compartments. In this study, we find the coefficients  $\alpha_i$  ( $i$  represents the number of predictor parameter) to the multiple regression models, as defined in Equations (1) and (2):

Step 1 (Estimating emissions):

$$TP = \alpha_0 + \alpha_1 \text{Production} + \alpha_2 \text{Use} + \alpha_3 \log H + \alpha_4 \log K_{ow} \quad (1)$$

Step 2 (Estimating PEC):

$$TP = \alpha_0 + \alpha_1 \text{Production} + \alpha_2 \text{Use} + \alpha_3 \log H + \alpha_4 \log K_{ow} + \alpha_5 k_{sw} + \alpha_6 \text{Emission to water} + \alpha_7 \text{Emission to air} \quad (2)$$

In the next section, the results are discussed. PCR and PLS were compared to reveal the inherent amount of correlation and co-linearity between PP and TP. Whereas PCR represents the inherent correlation, i.e. without fitting patterns in X, including all the PPs, to correlate optimal with TP data, PLS is an iterative process where the maximum amount of variation in X fitting optimal to the pattern in TP is found. PCR consists of two steps, where the first step is a PCA carried out on X after which the principal components (PCs) are used as predictors in an MLR. PLS-R is a bilinear modeling approach, where the PPs are projected onto a small number of underlying latent variables in an iterative process. In PLS-R, the TP data are used actively in determining the latent variables ensuring the highest possible relevance for prediction of TP in first PC. The number of PCs increases until no further increase in the explained TP-variance is achievable; i.e. maximum explained variance is obtained and further inclusion of PCs increases the noise in the model.

An important assumption for the MLR method is that the PPs are linearly independent. Optimal PPs with highest explanatory capacity were selected based on the results for PCR and PLS in parallel to stepwise linear regression (SYSTAT, 1997; CAMO ASA, 2005). The best simple MLR results are presented together with a visualization of latent variables in loading plots from the PLS-R models in the section below.

The stepwise approach we propose for deriving data and using these as PPs for predicting TPs for not assessed chemicals, then consists of the following tasks:

Step 1: Estimating emissions to air and water

(a) Retrieve and log-transform gross production and use data in the EU and physico-chemical properties;  $\log H$  and  $\log K_{ow}$  for a VOC of interest.

(b) Enter a regression model [Equation (1)] made from the training set to estimate the emissions as TP.

Step 2: Estimating PECs to air and water

(c) Retrieve and log-transform data on  $k_{sw}$ . Retrieve log-transformed data on emissions, derived in Step 1, as PP.

(d) Enter a regression model [Equation (2)] made from the training set to estimate the PECs as TP.

In order to accomplish the above stated steps, one should use the best data available in the specific situation to which the estimate is referred. Here we focus on data available for estimates at the level of continental Europe. It is essential that production and use for a chemical are not both zero. In such a case, the equations will still give emissions and PECs different from zero, which is not meaningful.

Total annual use in the EU can be computed from the mass balance:

$$\text{Quantity used} = \text{Quantity produced} + \text{Quantity imported} - \text{Quantity exported} \quad (3)$$

Production, import and export data can be found from e.g. EUROSTAT (EC-EUROSTAT, 2010): "PRODCOM annual sold and annual total", either on a national basis or on summed EU scale.

Physico-chemical parameters can be found from a variety of databases. In this analysis, mainly the IUCLID (EC-IUCLID, 2010) and HSDB (TOXNET-HSDB, 2010) databases were used. In this way, a compilation of approximate information could be retrieved for a large number of existing VOCs and semi-VOCs.

### 3. Results and Discussions

#### 3.1. Building regression model for emissions and PECs with production, use data and physico-chemical parameters

Multivariate data analysis is performed by purpose of exploring the existence of any inherent general patterns in a unique data set extracted from the 63 risk assessment reports that can be used for future screening risk assessment of new compounds. MLR/LR models for the seven target parameters (TPs) have been derived and the coefficients,  $\alpha_i$  to Equations (1) and (2), are shown in Table 3. The predictor parameters (PPs) were selected by a stepwise regression procedure in SYSTAT. In parallel, PCR and PLS models were used for selection of PPs in MLR/LR models based on the weighted criteria: (1) maximum orthogonality and (2) highest explanatory capacity as discussed below in relation to Figures 1 and 2. Only the optimum MLR/LR are shown in Table 3, i.e. models with highest  $R^2$ ,  $n$  and  $F$ -ratio and lowest  $p$ -value and RMSEP.

The usual limit used in the interpretation of a  $p$ -value is 0.05 (or 5% significance level). As observed from Table 3, the  $p$ -value is below 0.05 in all models included.

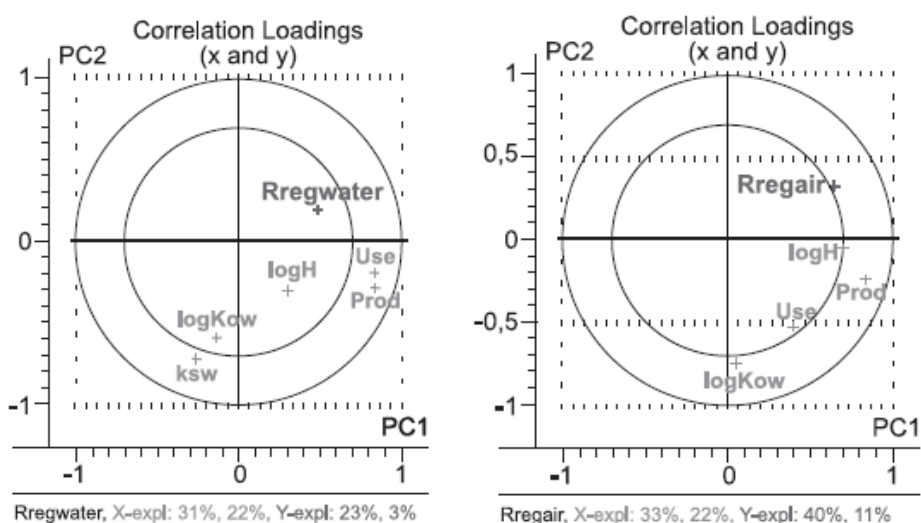
The plots in Figures 1 and 2 show the explanatory capacity, correlation patterns and correlation loading weights (CLW) between PPs and TPs for the first two principal components of the six partial least squares regression (PLS-R) models. The basic principle is that all plots include two circles representing 50% and 100% correlation; i.e. the variables between the two circles are the important ones. Furthermore, X-variables having zero loading in one principal component (PC) and a positive or negative loading in another are orthogonal; i.e. describes patterns that are independent or orthogonal to each other. The loadings of PPs with respect to TPs shows the importance of each PP in the principal components, i.e. PC1 and PC2, with respect to the X-variance in each PC used for explaining Y, i.e. the individual TPs. The used X-variance in PC1 and PC2 for explaining the TP-variance by the first two principal components are given in percent for each correlation loading plots in Figures 1 and 2.

#### 3.2. Estimating emissions

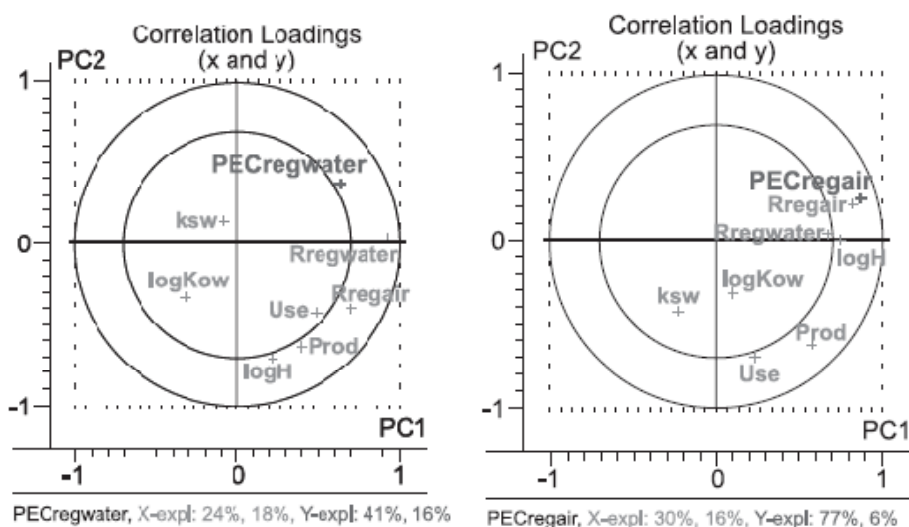
The right plot in Figure 1, modeling the emissions to air (*Rregair*), shows that  $\log H$  and production are the most important X-variable in PC1 using 33% X-variance for describing 40% of the variance in *Rregair*.  $\log K_{ow}$  is the most important X-variable in PC2 which uses 22% X-variance for describing only 11% of the variance in *Rregair*.  $\log K_{ow}$  has zero loading in PC1 and is orthogonal to  $\log H$ . The PLS-R model indicates that an increase in production volume is correlated to an increase in  $\log H$ ; at least for the dataset given in Table 1, excluding the outlier chemicals. The left plot, modeling the emissions to water (*Rregwater*), shows that the important X-variables with respect to explanation of *Rregwater* are use and production in PC1 using 31% X-variance for explaining 23% variance in *Rregwater*. PC2 does not contribute to any further reduction on the residual variance of the model.

**Table 3.** MLR/LR models of target parameters, where  $\alpha_0$  to  $\alpha_7$  are the regression coefficients in Equations (1) and (2),  $n$  is the number of cases,  $R^2$  and  $Q^2$  are the correlation coefficients based on calibration and leave-one-out cross-validation, respectively. The  $F$ -ratio is the regression sum of squares divided by unexplained residual variance,  $p$ -value is the significance level for the modeled variation to be real, RMSEP is the root mean square of calibration, and RMSEC the root mean square error of predictions, expressed in the same units as the target parameters. Bold figures indicate good combination of high  $R^2$ , RMSEP,  $n$ , high  $F$ -ratio and low  $p$ -value. Cells in grey comprise Step 1 (emission modeling) and all cells comprise Step 2 (PEC modeling)

	Constant	log production	log Use	log H	log $K_{ow}$	log $k_{sw}$	log Rregwater	log Rregair	Model performance parameters						
Regression coefficients	$\alpha_0$	$\alpha_1$	$\alpha_2$	$\alpha_3$	$\alpha_4$	$\alpha_5$	$\alpha_6$	$\alpha_7$	$n$	$R^2$	$Q^2$	F-ratio	$p$ -value	RMSEP	RMSEC
log Rregwater	-0.0559	0.392	-	-	-	-	-	-	52	0.12	0.063	7.27	0.0095	0.98	0.95
log Rregair	2.48	-	-	0.502	-0.362	-	-	-	51	0.45	0.37	19.4	<1.0x10 <sup>-5</sup>	1.34	1.26
log PECregwater	-1.77	-	-	-0.103	-	-	0.650	-	<b>46</b>	<b>0.50</b>	0.42	<b>22.8</b>	<1.0x10 <sup>-5</sup>	<b>0.68</b>	0.63
log PECregair	-3.52	-	-0.168	-	-	-	-	1.00	<b>48</b>	<b>0.71</b>	0.66	<b>54.0</b>	<1.0x10 <sup>-5</sup>	<b>1.21</b>	1.10



**Figure 1.** Loading plots for PLS models using production, use, log  $K_{ow}$  and log H for 63 chemicals as predictor parameters to predict the target parameters; emissions to water and air, respectively, at the EU scale. The outer ellipse indicates 100% and the inner ellipse indicates 50% of explained variance, respectively. The PLS models are used for selecting potential predictor parameters and corresponding coefficients in Equation (1).



**Figure 2.** Loading plots for PLS models using production, use, log  $K_{ow}$ , log H,  $k_{sw}$ , Rregwater, and Rregair for 63 chemicals as predictor parameters to predict the target parameters; PEC in water and air, respectively, at the EU scale. The outer ellipse indicates 100% and the inner ellipse indicates 50% of explained variance, respectively. The PLS models are used for selecting potential predictor parameters and corresponding coefficients in Equation (2).

MLR models comprising the explanatory variables  $\log K_{ow}$  and  $\log H$ , Use or Production were tested and the root mean square error of predictions (RMSEP) by these models was compared to the RMSEP by use of the PLS–R model as visualized in Table 3. The best fit and F–statistics is thus obtained by selecting  $\log H$  and  $\log K_{ow}$  as PPs in a MLR model for predicting *Rregair* and production as PP in a simple LR for predicting *Rregwater*. The model performance parameters in Table 3 are poor for the *Rregwater* model and reasonable for the *Rregair* model. This indicates that predicting emissions is more complex than can be deduced from simple relationships between emissions and the chosen parameters. Emission modeling requires an in–depth assessment of emission activities, use amounts under specified conditions and corresponding emission factors. However, the models will be used in this paper, and the results will be assessed with associated model errors.

### 3.3. Estimating PECs

From the left plot in Figure 2 it is seen that the most important X–variables with respect to PEC in water (*PECregwater*) are *Rregwater*, but also *Rregair* and  $\log H$  have high importance. Production, use and  $\log K_{ow}$  have lesser importance and  $k_{sw}$  are of minor importance. From the right plot it is seen that the most important X–variables are *Rregair* and  $\log H$ . *Rregwater* is just below the 50% correlation circle in the direction of PC1. Production volume is contributing equally to PC1 and PC2 indicating a splitting of chemicals into two trend patterns: the direction of PC1 explaining patterns of direct correlation between an increase in production volume and *PECregair*, and another trend in the direction of PC2 where production volume is inversely related to *PECregair*. The high negative loading of use in the PC2 indicates an inverse correlation to *PECregair*.

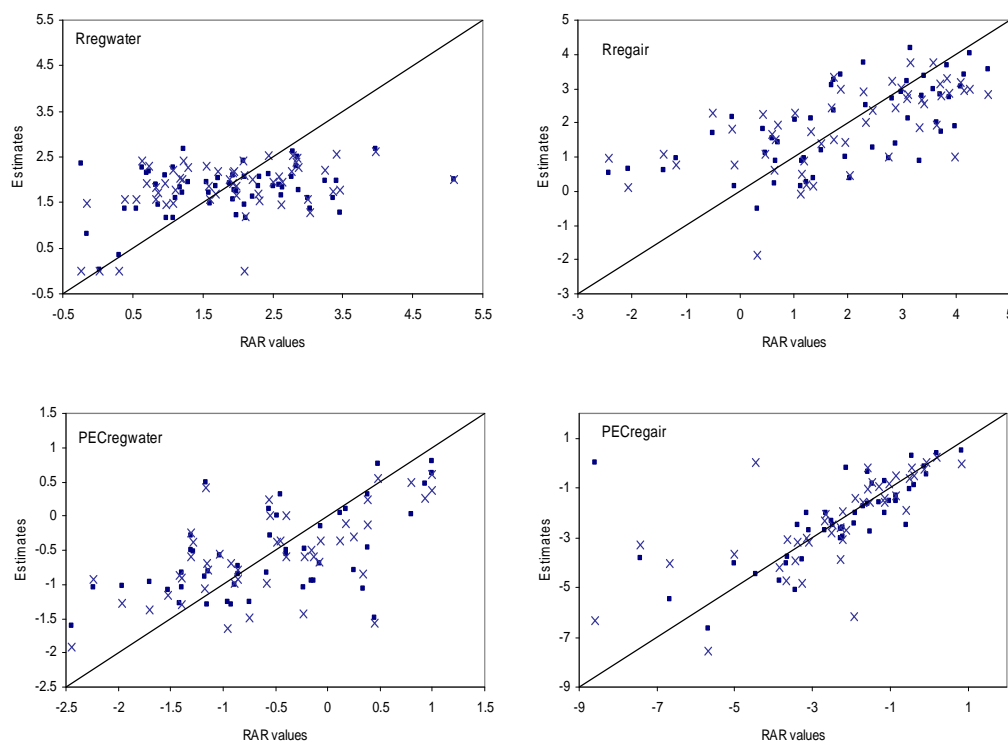
For *PECregwater*, the parameters *Rregwater* together with  $\log H$  give the best model performance. This is supported by  $\log H$

having highest loading in PC2 and *Rregwater* having highest loading in PC1 and zero loading in PC2. For *PECregair*, the parameters *Rregair* and use give best possible model performance. This is supported by *Rregair* having highest loading in PC1 and Use having highest loading in PC2. However, the two parameters are not totally orthogonal as visualized in the loading plot in Figure 2.

Prediction of PEC from the chosen parameters and emissions yield regression coefficients for air and surface water of  $R^2 = 0.71$  and 0.50, respectively. Both regression models require an estimate of emissions to the respective medium. When emission estimates are available, estimates of PEC can be obtained at the certainty level reflected by the average prediction error, RMSEP, given in Table 3.

### 3.4. Evaluation of MLR/LR models

The training set is used for cross-validating the emissions and PECs for each chemical by removing it from the training set and making a PLS and MLR/LR model with the remaining chemicals. Figure 3 shows the cross–validation in a scatter diagram between reported emissions (RAR) and predicted emissions from PLS (boxes) and MLR/LR (crosses) models. Between 79% and 93% of all model predictions are within one order of magnitude of the RAR values. When using this model performance criteria the PLS estimates for *Rregair* and *PECregwater* are more accurate than the MLR/LR estimates and for *Rregwater* and *PECregair* the MLR/LR estimates are more accurate than the PLS estimates. For estimating emissions the data points deviating more than one order of magnitude from the RAR values are overestimating emissions, i.e. conservative estimates, predominantly at low emissions. For estimating PECs the data points deviating more than one order of magnitude from the RAR values are underestimating PECs, predominantly at low PECs.



**Figure 3.** Scatter plots of RAR values vs. leave–one–out PLS cross–validated estimates (squares) and MLR/LR estimates (crosses) for each of the four target parameters; *Rregwater*, *Rregair*, *PECregwater* and *PECregair*. Model MLR/LR performance parameters are shown in Table 3.

The coefficients of variation (= standard error/mean value) for the MLR/LR models in log-units are 0.19, 0.39, 0.67, and 0.39 for *Rregwater*, *Rregair*, *PECregwater*, and *PECregair*, respectively. The standard error is an estimate of the deviation of the MLR/LR value for every RAR value. The most reliable model in terms of the model performance parameters in Table 3 and in Figure 3 is for *PECregair* and the least reliable model is for *Rregwater*.

For comparison to the MLR/LR model performance parameters in Table 3, RMSEP, which may be interpreted as the average prediction error, and expressed in the same units as the original response values, were 0.7 and 0.9 for *PECwater* and *PECair*, respectively.

### 3.5. Estimating emissions and PECs to air and surface water for non-assessed chemicals

One way to use multiple regression analysis, as shown above, is to predict emissions and PECs in air and surface water of chemicals for which a RAR, or generally speaking comprehensive information on emissions and environmental concentrations, is not available. Chemicals listed in the Water Framework Directive as Priority Substances and Certain Other Pollutants (According to Annex II of the Directive 2008/105/EC) are candidates for testing the water model and chemicals listed in the Stockholm Convention on Persistent Organic Pollutants are candidates for testing the air model. A number of the listed chemicals are already included as RAR chemicals and thus used in the model training set. Excluding metal complexes and pesticides, which are not comprised in the model domain, production, use, log *H* and log *K<sub>ow</sub>* can be found in EUROSTAT, IUCLID and HSDB for the following test chemicals: 1,2-dichloroethane (CAS no. 107-06-2), dichloromethane (CAS no. 75-09-2), tetramethylbutyl phenol (CAS no. 140-66-9), and carbon tetrachloride (CAS no. 56-23-5).

Emissions and PECs are estimated from regression analysis of the training set, Equations (1) and (2). All data for the above four test chemicals are shown in Table 4.

For all predicted TPs in Table 4, intervals are stated. These are found from the standard errors of the MLR/LR models, which are in log-units. The intervals of absolute values in Table 4, do therefore not exhibit normal distributions, but state the boundaries of the mean ± standard error.

## 4. Conclusions

In this paper, a new method was tested for estimating emissions and PECs of certain high-production volume semi-VOCs and VOCs in air and water, on a regional EU scale. The model domain comprises data from the European Chemicals Bureau risk assessment reports (RARs) of 63 VOCs and semi-VOCs that are used in a variety of industrial and domestic activities and products.

Metal complexes and pesticides have been excluded due to deviating use patterns especially for the pesticides and missing volatilities for the metals. Estimation methods for these chemicals must be built from a training set of similar chemicals.

The method uses simple linear relationships between chemical properties, i.e. Henry's Law constant, octanol-water coefficient, and production and use volumes. Being aware of the complex relationship between physico-chemical properties, production and use patterns of chemicals and their emissions and PECs in the environment, the methodology does not attempt to describe emission or dispersion processes. The method explores the existence of any inherent general patterns in the unique data set using multivariate data analysis. In this way a complex and time consuming analysis is bypassed, but at the same time it is important to bear in mind the domain of the method and that critical errors may occur when using chemicals that have deviating physico-chemical properties, production and use patterns and amounts. In addition to using a proper training set of chemicals, the method must be designed and tested for a representative region (e.g. EU, US, Asia). Other domains may not reveal linearity and it may be necessary to include other parameters to explain emissions and PECs.

Uncertainty measures are not available for the RAR data; however, cross validation of the applied regression models reveals that between 79% and 93% of all emission and PEC estimates are within one order of magnitude of the reported RAR values. The coefficients of variation (= standard error/mean value) in log-units are 0.19, 0.39, 0.67, and 0.39 for emission to water, emission to air, *PECwater*, and *PECair*, respectively. The standard error is an estimate of the deviation of the multivariate linear regression value for every RAR value. The most reliable model in terms of the model performance parameters is for *PECair* and the least reliable model is for emission to water.

The method is simple to use as required data on production and use amounts in industry and in downstream products can be retrieved e.g. from EUROSTAT and Henry's Law constant and octanol-water coefficient can be retrieved e.g. from the IUCLID database. This information is readily available for many chemicals in a transparent and uniformly comparable form. For other parts of the world similar data have to be retrieved for relevant regions. However, it is not within the scope of this work to assess whether such data are available outside EU.

The method has been tested with four chemicals, prioritized in the Water Framework Directive and the Stockholm Convention on Persistent Organic Pollutants. Emissions and PECs in air and water have been estimated, with corresponding uncertainty intervals, at regional EU level.

**Table 4.** Test chemicals. MLR/LR model coefficients in Table 3 are used to predict emissions to water and air (Equation 1) and PECs to water and air (Equation 2). Production and use are from EUROSTAT and physico-chemical parameters are from IUCLID and HSDB. Intervals are calculated from standard error of log-normal distributions

CAS no.	Substance Name	Production in EU (t/y)	use in EU (t/y)	log <i>H</i>	log <i>K<sub>ow</sub></i>	Emissions regional Water (t/y)	Emissions regional Air (t/y)	PEC regional Water (µg/L)	PEC regional Air (µg/m <sup>3</sup> )
107-06-2	1,2-dichloroethane	1 452 000	1 345 000	1.99	1.48	230 (83-640)	880 (63-12 000)	0.36 (0.18-0.71)	0.025 (0.006-0.10)
75-09-2	Dichloromethane	300 000	243 000	2.46	1.25	120 (49-300)	1 800 (97-33 000)	0.21 (0.07-0.60)	0.068 (0.024-0.19)
140-66-9	Tetramethylbutyl phenol	22 630	22 860	1.04	4.0	45 (22-92)	36 (8.9-150)	0.16 (0.05-0.55)	2.0x10 <sup>-3</sup> (1.8x10 <sup>-4</sup> -2.2x10 <sup>-2</sup> )
56-23-5	Carbon tetrachloride	22 330	20 790	3.37	2.73	45 (22-92)	1 500 (87-26 000)	0.09 (0.02-0.45)	0.085 (0.033-0.22)

In conclusion the method can be used as a screening level tool to estimate emissions and PECs in air and water for chemicals not subjected to RARs, for the goal of chemical fate and transport modeling evaluation. The procedure can also be seen as supportive to the development of a RAR, as it allows simplification and acceleration of the process of emission and concentration estimates, and quick prioritization of critical chemicals and environmental compartments, which can be a time consuming task.

As a screening level procedure, it provides only a first approximation estimate, although it is observed that often emission inventories themselves have intrinsically high uncertainties (Breivik et al., 2006); therefore, the proposed procedure appears promising when only limited data are available and a quick response is required. It should also be noticed that when pursuing assessments at a local scale, the spatial distribution of point emissions, may have a very strong impact on predictions. In general, however, the method is expected to provide more and more reasonable estimates as moving to regional and continental scales.

### Acknowledgements

This research was financially supported in part by the European Union under European Commission FP6 Contract No. 003956 (NoMiracle Project). The Authors wish to thank all colleagues that helped with discussions during the development of the work.

### References

- Breivik, K., Vestreng, V., Rozovskaya, O., Pacyna, J.M., 2006. Atmospheric emissions of some POPs in Europe: a discussion of existing inventories and data needs. *Environmental Science and Policy* 9, 663-674.
- CAMO ASA, 2005. The Unscrambler 9.02; Oslo, Norway.
- EC Chemicals, 2010. European Commission, DG JRC, European Chemicals Bureau (ECB) (coordination), Risk assessment Reports, various years. <http://ecb.jrc.ec.europa.eu/existing-chemicals/>.
- EC, EUROSTAT., 2010. European Commission, Statistics Database. [http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search\\_database](http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database).
- EC, IUCLID., 2010. European Commission, DG JRC, European Chemicals Bureau (ECB) (coordination), The International Uniform Chemical Information Database (IUCLID). <http://iuclid.echa.europa.eu/>.
- EC, Priority Lists., 2010. European Commission, DG JRC, European Chemicals Bureau (ECB) (coordination), Priority Lists. <http://ecb.jrc.ec.europa.eu/esis/index.php?PGM=ora>.
- EC, 2008. European Commission, Priority Substances and Certain Other Pollutants (According to Annex II of the Directive 2008/105/EC), [http://ec.europa.eu/environment/water/water-framework/priority\\_substances.htm](http://ec.europa.eu/environment/water/water-framework/priority_substances.htm).
- EC, 2005. Environment and Health, EEA Report No 10/2005, Copenhagen.
- EC, 2004. European Union System for the Evaluation of Substances 2.0 (EUSES 2.0). Prepared for the European Chemicals Bureau by the National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands (RIVM Report no. 01900005) (<http://ecb.jrc.ec.europa.eu/euses/>).
- EC, 2003a. Technical Guidance Document in support of Commission Directive 93/67/EEC on Risk Assessment for new notified substances, Commission Regulation (EC) No 1488/94 on Risk Assessment for existing substances and Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market. (<http://ecb.jrc.it/tgd/>).
- EC, 2003b. Technical Guidance Document in support of Commission Directive 93/67/EEC on Risk Assessment for new notified substances, Commission Regulation (EC) No 1488/94 on Risk Assessment for existing substances and Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market. Part IV. ([http://ecb.jrc.it/Documents/TECHNICAL\\_GUIDANCE\\_DOCUMENT/EDITION\\_2/tgdpart4\\_2ed.pdf](http://ecb.jrc.it/Documents/TECHNICAL_GUIDANCE_DOCUMENT/EDITION_2/tgdpart4_2ed.pdf)).
- EC, 2001. Directive 2001/81/EC of the European Parliament and of the Council of 23 October 2001 on national emission ceilings for certain atmospheric pollutants. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:309:0022:0030:EN:PDF>
- EC, 1993. Council Regulation 793/93/EEC of March 1993 on the evaluation and control of risks of existing substances. Official Journal of the European Communities, L48/1.
- EC, 1967. COUNCIL DIRECTIVE 67/548/EEC of 27 June 1967 on the approximation of laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances. <http://ecb.jrc.ec.europa.eu/legislation/1967L0548EC.pdf>.
- EU, 2007. The Regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH). [http://ec.europa.eu/enterprise/sectors/chemicals/reach/index\\_en.htm](http://ec.europa.eu/enterprise/sectors/chemicals/reach/index_en.htm).
- Karjalainen, T., 2005. Commission research in Action: tackling the hormone disrupting chemicals issue, EUR report 21941.
- Larsen, P.V., 2006. Regression and analysis of variance. <http://statmaster.sdu.dk/courses/st111/>.
- Lerche, D., Sorensen P.B., Larsen, H.S., Carlsen, L., Nielsen, O.J., 2002. Comparison of the combined monitoring-based and modelling-based priority setting scheme with partial order theory and random linear extensions for ranking of chemical substances. *Chemosphere* 49, 637-649.
- Mateu, J., 1997. Methods of assessing and achieving normality applied to environmental data. *Environmental Management* 21, 767-777.
- Penman, J., Kruger, D., Galbally, I., Hiraiishi, T., Nyenzi, B., Emmanuel, S., Buendia, L., Hoppaus, R., Martinsen, T., Meijer, J., Miwa, K., Tanabe, K. (eds), 2001. Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories. IPCC National Greenhouse Gas Inventories Programme.
- SC POPs, 1997. Stockholm Convention on Persistent Organic Pollutants (POPs), <http://chm.pops.int/Portals/0/Repository/conf/UNEP-POPS-CONF-4-AppendixII.5206ab9e-ca67-42a7-afee-9d90720553c8.pdf#AnnexC>
- SYSTAT, 1997. Statistics. SPSS Inc., Chicago IL, USA.
- Thomsen, M., Knudsen, L.E., Vorkamp, K., Frederiksen, M., Bach, H., Bonefeld-Jorgensen, E.C., Rastogi, S., Fauser, P., Krongaard, T., Sorensen, P.B., 2008. Conceptual framework for a Danish human biomonitoring program. *Environmental Health: A Global Access Science Source* 7 (SUPPL. 1), art. no. S3.
- TOXNET, HSDB. Hazardous Substances Data Bank (HSDB) - Comprehensive, peer-reviewed toxicology data for about 5,000 chemicals. United States National Library of Medicine, TOXNET Toxicology Data Network, 2010. <http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB>.
- UNECE, 2005. UNECE/EMEP Task Force on Emissions Inventories and Projections, EMEP/CORINAIR Emission Inventory Guidebook - 3<sup>rd</sup> Edition, EEA Technical report No. 30, Copenhagen.
- UNECE-CLRTAP, 2010. Convention on Long-range Transboundary Air Pollution, <http://www.unece.org/env/Irtap/fulltext/1979.CLRTAP.e.pdf>.