ELSEVIER

Journal of Hazardous Materials A91 (2002) 63–80

Journal of Hazardous Materials

www.elsevier.com/locate/jhazmat

Selection of a representative set of chemical accidents from a complex data matrix for the development of environment–accident index

Åsa Scott [a,*], Mats Tysklind [b], Ingrid Fängmark [a]

[a] *Swedish Defence Research Agency, Division of NBC Defence, SE-901 82 Umeå, Sweden*
[b] *Department of Chemistry, Environmental Chemistry, Umeå University, S-901 87 Umeå, Sweden*

Received 14 June 2001; received in revised form 24 June 2001; accepted 27 November 2001

Abstract

Chemical accidents often lead to negative consequences for the environment. Preparedness and proper actions are, therefore, essential components in order to minimise environmental effects. To assist and facilitate this work, a proposed planning tool, the environment–accident index (EAI), was formulated by Scott [J. Hazard. Mater. 61 (1998) 305]. As a result of a first validation of the index, based on 21 chemical accidents, the database was complemented with 42 additional accidents covering a broader spectrum of chemicals. The additional accidents were collected by means of an inquiry and their environmental consequences are, so far, unknown. The collected data had an overrepresentation of accidents involving petroleum products (69%). Because of the overrepresentation of this group of chemicals in the material, the data was skewed with respect to chemical properties. Since the model should be valid for a variety of chemical accidents, a method was needed which enabled a proper and unbiased selection of a representative subset of accidents to be used in development and validation of the model. For this purpose, the possibility to use multivariate data analysis in combination with statistical design was investigated. The result showed the feasibility of this method in the selection of a representative subset from a complex and skewed large dataset. Within the new dataset, 53% were accidents involving petroleum products and 47% involved other chemicals. The selected accidents will be used in further work to evaluate the environmental consequences, for model development and model validation.

Keywords: Environment–accident index; Chemical accidents; Selection; Multivariate design; PCA

* Corresponding author. Tel.: +46-90-10-68-29; fax: +46-90-10-68-00.
E-mail address: asa.scott@foi.se (Å. Scott).

0304-3894/02/$ – see front matter
PII: S0304-3894(01)00387-9

1. Introduction

1.1. Background

Chemical accidents often lead to negative consequences for the environment. Preparedness and proper actions are, therefore, essential components in order to minimise environmental effects. People dealing with the accidents will face many questions. How toxic is the chemical? How will the chemical behave in the environment? Which are the spreading properties of the environment and which are the resulting consequences for the environment? Instead of trying to collect such data at the time of the accident this work should already be done at the planning stage. To assist and facilitate this work a proposed planning tool, the environment–accident index (EAI), was formulated by Scott [1]. The EAI was formulated as a simple equation based on chemical inherent properties and properties of the surroundings at the accident sites, such as soil and groundwater conditions. The magnitude of the index could then be used to judge the consequences for the environment in order to take proper actions at an accident site.

A first validation of the EAI, based on 21 chemical accidents, demonstrated that the index has the capacity to become a useful tool for the ranking and classification of what kind of further risk assessment to be performed [1]. However, to achieve this the variables describing the surroundings need to be more detailed and spreading related descriptors, such as volatility and density, should be added. A disadvantage with the dataset used for the validation is that it only consisted of 21 accidents, with a focus on petroleum products. Therefore, the database was complemented with 42 additional accidents covering a broader spectrum of chemicals.

In the development of the EAI it is necessary to make an unbiased judgement of relevant variables to include in the formula and to estimate their relative importance. To avoid focusing on petroleum products and thus increase the chemical diversity, there is a need for a method to select a representative subset of accidents. To address the above the following strategy is proposed:

I. To collect a larger database of accidents, together with relevant numerical descriptors to be used in the development of the index.
II. To condense the large number of dependent descriptors in the database into a few orthogonal independent descriptors.
III. To use the new latent descriptors in a statistical experimental and multivariate design to select a minimum number of representative accidents.
IV. To evaluate the consequences for the environment of the representative set of accidents, e.g. by an expert panel.
V. Model development.
VI. Model validation.

In the first paper [1], accidents involving both organic and inorganic chemicals was included in the validation. This paper deals with the chemical accidents involving organic chemicals only and with parts I–III in the proposed strategy.

The objectives of this paper were to select a representative set of chemical accidents for the following development of the EAI and to investigate if it is possible to use multivariate statistical design for the selection.

2. Materials and methods

2.1. Chemical accidents

The material used in the project was data on chemical accidents that occurred between 1986 and 1999 containing two parts, one part with accidents from the first validation and other with additional accidents. The 42 additional accidents were mostly collected by means of an inquiry containing questions about chemical accidents. All information concerning the accident material is available from the author [2]. The inquiry was sent to the public environmental health offices and the rescue services of all 289 municipalities in Sweden. The information asked for was a short description of the place of the accident, cause, chemical involved, amount of the stored or transported chemical and eventual damage to the environment. Following the inquiry, a deeper investigation was performed to gain a more detailed knowledge about both the chemical and the place of the accident. In the new database, the data had an overrepresentation of accidents involving petroleum products (69%). The tendency was confirmed by statistics from the Swedish Rescue Services Agency [3–6], which showed that 60–77% of chemical accidents occurred between 1996 and 1999, involved petroleum products. The dataset was thus skewed with respect to chemical properties. Because the EAI also shall be valid for other organic chemicals, there was a need for a method to select a representative dataset for the calibration of the index.

Another problem in the selection procedure was that variables describing inherent properties of the chemicals (see below) tended to be more heavily weighted than the other variables. These variables would, therefore, easily receive too much importance in the final selection. It was of utmost importance that descriptors for the surrounding were given equal attention. To address the above problems statistical multivariate design was used in the selection procedure.

2.1.1. Descriptor variables

The collected material (dataset) consisted of 58 chemical accidents (objects) described by 10 parameters (variables) (Table 1). Additional information showing the places where the accidents occurred is given in Table 2. Three accidents (numbers 6, 7 and 52) involved chemicals with a very high viscosity. These accidents were excluded in the calculations because of the extreme chemical properties that also would create an unbalanced design. A balanced design will be achieved by a selection of test objects projected far from each other in the score plot (Section 2.2.1) but not on the extreme periphery of the plot. This left 55 chemical accidents to be modelled.

The descriptor variables were: kinematic viscosity, υ (mm^2/s); water solubility, S_w (wt.%); amount of the stored or transported chemical, m (metric tonnes); acute toxicity for water living organisms, Tox (mg/l); and properties of the surrounding environment. Toxicity and amount are important variables to judge effects on the environment. However, it is the

Table 1
Dataset with chemical accidents and descriptors[a]

No.	Chemical	CAS-no.	P_v (kPa)	v (mm^2/s)	D (kg/m^3)	S_w (wt.%)	m (metric tonnes)	Tox (mg/l)	m/Tox (l)	DNW (m)	DGS (m)	SGS	K' (m/day)	n (%)
1	Kerosene/jet fuel	8008-20-6	0.1	1.5	808	0.1	10	3.1	3.20E+09	6	2	1	1	55
2	Kerosene/jet fuel	8008-20-6	0.1	1.5	808	0.1	22.6	3.1	7.30E+09	3	0.6	0.5	1	55
3	Kerosene/jet fuel	8008-20-6	0.1	1.5	808	0.1	24.2	3.1	7.80E+09	15	1.5	1	3	35
4	Kerosene/jet fuel	8008-20-6	0.1	1.5	880	0.1	38.8	3.1	1.30E+10	5	0.01	1	1	15
5	Petroleum liq.	115-86-6	0.1	9.5	880	0.1	88	100	8.80E+08	5	0.5	1	1	15
6	Petroleum liq.	8012-95-1	0.1	95	919	0.1	0.047	100	4.70E+05	17.5	2	0.1	1	15
7	Heating oil no. 5	64741-45-3	1	41	965	0.1	483	55	8.80E+09	10	10	1	3	33
8	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	328	2.6	1.30E+11	1	1.5	1	3	35
9	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	24.6	2.6	9.50E+09	100	5	1	3	35
10	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	42.6	2.6	1.60E+10	9	3	1	1	55
11	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	9.84	2.6	3.80E+09	50	2	0.5	2	35
12	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	8.2	2.6	3.20E+09	3	1.5	1	1	15
13	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	4.2	2.6	1.60E+09	15	4	1	1	15
14	Diesel fuel/heating oil no. 1	6833430-5	0.5	2.75	820	0.1	1.3	2.6	5.00E+08	3	6	0.5	1	43
15	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	24.6	2.6	9.50E+09	5	0.2	1	1	15
16	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	37	2.6	1.40E+10	7	0.35	1	1	15
17	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	49	2.6	1.90E+10	15	2	1	2	35
18	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	240	2.6	9.20E+10	220	4.4	1	2	35
19	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	0.5	2.6	1.90E+08	600	0.01	1	3	33
20	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	0.122	2.6	4.90E+07	1600	3	1	1	15
21	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	13.1	2.6	5.00E+09	5	8	1	1	55
22	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	4.9	2.6	1.90E+09	1	3	1	1	43
23	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	0.326	2.6	1.30E+08	2000	6	0.1	1	43
24	Diesel fuel/heating oil no. 1	68336-30-5	0.5	2.75	820	0.1	24.6	2.6	9.50E+09	350	3	1	1	43
25	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	4075	2.6	1.60E+12	13	7.2	1	3	33
26	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	0.8	2.6	3.10E+08	40	2	1	2	35
27	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	0.122	2.6	4.70E+07	0.01	1	1	1	15
28	Diesel fuel/heating oil no. 1	68334-30-5	0.5	2.75	820	0.1	8	2.6	3.10E+09	17	13	1	u	u
29	Gasoline	86290-81-5	70	1	750	0.01	16.5	4.9	3.40E+09	3	3	1	1	43
30	Gasoline	86290-81-5	70	1	750	0.01	28.5	4.9	5.80E+09	25	0.01	1	2	35
31	Gasoline	86290-81-5	70	1	750	0.01	38.2	4.9	7.80E+09	100	0.01	1	3	33
32	Gasoline	86290-81-5	70	1	750	0.01	35.3	4.9	7.20E+09	2	16	1	1	43
33	Gasoline	86290-81-5	70	1	750	0.01	0.75	4.9	1.50E+08	8	2	0.5	1	15
34	Gasoline	86290-81-5	70	1	750	0.01	26	4.9	5.30E+09	0.01	12	1	1	15

Table 1 (*Continued*)

35	Gasoline	86290-81-5	70	1	750	0.01	35.5	4.9	7.20E+09	2	1.87	0.5	1	43
36	Gasoline	86290-81-5	70	1	750	0.01	12	4.9	2.40E+09	9	8	0.1	1	15
37	Gasoline	86290-81-5	70	1	750	0.01	22.5	4.9	4.60E+09	200	6	1	1	15
38	Gasoline	86290-81-5	70	1	750	0.01	35.3	4.9	7.20E+09	100	9.4	1	2	35
39	Gasoline	86290-81-5	70	1	750	0.01	18	4.9	3.70E+09	7	2.6	1	1	15
40	Gasoline	86290-81-5	70	1	750	0.01	33.6	4.9	6.90E+09	7	5	1	1	15
41	Methanol	67-56-1	12.8	0.8	790	90	35	13200	2.70E+06	125	33	1	1	55
42	Isopropanol	67-63-0	4.2	3	790	90	0.18	2285	7.90E+04	300	2	1	2	35
43	Benzene	71-43-2	10.1	1	880	1.8	1232	5.3	2.30E+11	3	2	1	3	35
44	Chlorobenzene	108-90-7	1.2	1	1110	0.05	1000	4.7	2.10E+11	30	2.3	1	1	15
45	Phenol	108-95-2	0.05	3.2	1070	8	90	3.3	2.70E+10	4	4.1	0.5	u	u
46	Methyl methacrylate	80-62-6	3.7	2.25	940	1.5	47	159	3.00E+08	6	4	1	1	55
47	*n*-Butylacetate	123-86-4	1.2	1	880	0.007	4.8	18	2.70E+08	25	23	0.5	3	33
48	Vinylacetate	108-05-4	12	1	930	0.02	29.8	18	1.70E+09	7	7.6	0.5	2	35
49	Styrene	100-42-5	0.6	0.9	910	0.01	60	9.1	6.60E+09	1	3.1	1	3	33
50	4-Chloro-*m*-cresol	59-50-7	u	2.25	1370	0.001	0.002	7.6	2.60E+05	200	1.5	0.5	1	15
51	Trichloroethylene	1979-01-06	7.7	0.9	1460	0.1	2	16	1.30E+08	30	7	1	3	35
52	DEHP	117-81-7	vl	82.2	990	0.1	3.2	500	6.40E+06	30	1	1	3	33
53	Benzenesulphonic acid	98-11-3	u	2.25	1295	90	1.04	5.5	1.90E+08	15	u	0.5	u	u
54	Butyladehyde	123-72-8	12	1	810	7	4.1	19	2.20E+08	35	3.2	0.5	1	55
55	Formic acid	64-18-6	4.3	1.5	1220	90	6.1	70	8.70E+07	30	3.2	0.5	1	15
56	Glyphosate	38641-94-0	vl	2.25	1170	90	0.016	86	1.80E+05	2	16.5	1	1	43
57	Fluaziname	79622-59-6	vl	2.25	1300	0.01	0.002	0.05	4.20E+07	5	3	1	2	35
58	Metamitrone	41394-05-2	vl	2.25	u	0.2	0.021	101.7	2.10E+05	100	4	1	1	55

[a] Tox: acute toxicity to fish, *Daphnia* or algae (LC_{50} or EC_{50}); K': classed hydraulic conductivity (1–3).
vl: very low; u: unknown.

Table 2
Score values for PCA of the dataset[a]

Obs. no.	Obs. Name	PC1	PC2	PC3
1	**Arboga**	**0.56**	**0.24**	**−0.12**
2	Haninge	0.52	−0.23	−0.40
3	*Solna*	*0.01*	*1.28*	*1.21*
4	Gallivare	−0.32	2.22	−0.70
5	Eskilstuna	1.89	3.18	−2.10
8	Ystad	−0.41	1.83	0.25
9	**Nykoping**	**0.24**	**1.19**	**1.41**
10	Mariestad	0.50	0.51	−0.28
11	Torsby	0.50	0.17	0.36
12	Hallsberg	−0.02	1.10	−1.44
13	Ange	0.22	0.75	−1.10
14	Solna	0.83	−0.66	−1.01
15	**Karlshamn**	**−0.24**	**1.71**	**−1.16**
16	*Sodertalje*	*−0.22*	*1.59*	*−1.13*
17	Boden	0.09	1.21	0.18
18	Vaidotai	0.14	1.08	0.77
19	Orebro	0.31	2.46	2.48
20	Nykoping	0.85	0.42	0.23
21	*Helsingborg*	*0.65*	*0.17*	*−0.49*
22	Robertsfors	0.41	0.53	−1.09
23	Torsby	1.51	−1.61	0.57
24	Norrtalje	0.61	0.55	0.38
25	Vastervik	−0.43	1.61	0.66
26	Eksjo	0.58	0.83	0.55
27	Hofors	0.04	0.96	−2.77
28	Falkoping	0.45	0.47	−0.55
29	Nykoping	−1.91	−0.88	−0.24
30	*Kil*	*−2.44*	*1.04*	*1.40*
31	**Kungsbaca**	**−2.53**	**8.38**	**2.50**
32	*Hidinge*	*−1.90*	*−1.24*	*−0.52*
33	Falkoping	−1.78	−1.56	−0.73
34	Lycksele	−2.62	−0.71	−2.54
35	**Bracke**	**−1.88**	**−1.55**	**−0.50**
36	Grangesberg	−1.84	−2.38	−1.05
37	Norrtalie	−2.03	−0.75	0.09
38	Ragunda	−1.91	−0.76	1.14
39	Oskarshamn	−2.28	−0.48	−0.70
40	Harryda	−2.30	−0.60	−0.78
41	**Astorp**	**1.30**	**−2.86**	**1.15**
42	*Karlstad*	*2.15*	*−0.49*	*1.47*
43	Oxelosund	−1.03	0.33	1.10
44	*Kotka*	*−0.44*	*0.23*	*−0.52*
45	**Sundsvall**	**2.33**	**0.15**	**−1.77**
46	Hango	1.07	−0.62	−0.04
47	**Molndal**	**−0.28**	**−1.87**	**1.19**
48	*Angelholm*	*−0.72*	*−1.51*	*0.30*
49	**Malmo**	**−0.86**	**0.58**	**0.64**
50	**Visby**	**2.05**	**−0.17**	**−0.19**
51	Vetlanda	0.79	0.87	1.85

Table 2 (*Continued*)

Obs. no.	Obs. Name	PC1	PC2	PC3
53	Vetlanda	3.91	−1.61	0.91
54	Malmo	0.53	−2.27	0.62
55	*Malmo*	*1.91*	*−1.74*	*−0.44*
56	Eslov	4.32	−0.91	−0.69
57	Varberg	1.39	0.35	0.14
58	Vallinge	3.02	0.68	0.70

[a] Accidents in the training (bold) and validation sets (italic). The accidents are named (Obs. name) after the place where they occurred.

toxicity in relation to the amount that is crucial for the effects on the environment at an accident site. A less toxic chemical (high value = high concentration of the chemical is needed to cause effect) can in large amounts cause severe damage to the environment as well as a small amount of a highly toxic chemical (low value = low concentration of the chemical is enough to cause effect). Therefore, the ratio of amount/toxicity (*m*/Tox) was calculated and used in the selection procedure. The water solubility of a chemical is important both for the spreading properties and for how toxic the chemical will be to the environment and the viscosity contributes to describe the (horizontal) spreading of the chemical. With respect to the first validation two new descriptor variables for chemical inherent properties, were added to the dataset: density, D (kg/m^3); vapour pressure, P_v (kPa). These new variables are important to describe the vertical transportation of the chemical in water and soil and evaporative losses to the air. Properties related to the chemicals were gathered from various literature sources and databases [7–18].

The properties of the surroundings were described by the distance to nearest well, lake or watercourse, DNW (m); the depth to groundwater surface, DGS (m); the slope of the groundwater surface and the flow direction, SGS (leaning towards a well lake or watercourse (=1), horizontal surface (=0.5) and no well lake or watercourse in the flow direction (=0.1)); and the permeability of the soil. In the first validation, the permeability of the soil was expressed as type of soil, i.e. sand, gravel or clay. However, more quantitative variables were needed to better describe the behaviour of the chemicals in the different types of soils. Two new variables were, therefore, introduced, replacing the type of soil variable, viz. the hydraulic conductivity for each chemical and soil, K' (m/day) and the porosity of the soil, n (%). The calculated values of hydraulic conductivity were assumed to be under saturated soil conditions and were calculated as shown by Engström and Gustavsson [19] in formula (1).

$$K' = \frac{\upsilon_w}{\upsilon'} K_w \tag{1}$$

The variable υ_w is the kinematic viscosity for water, υ' is the kinematic viscosity for the chemical and K_w the hydraulic conductivity in the specific soil for water. The new variables give a better description on how the chemicals will behave in the soils in case of a spill than just using the type of soil. The dataset can be considered as consisting of two groups of variables, υ, S_w, D and P_v describing the inherent properties of a chemical and DNW, DGS, SGS, K' and n being a second group related to the properties of the surroundings at the accident site. The amount of a chemical involved in an accident and thus also the ratio *m*/Tox, are properties which do not belong to any of the two groups.

2.1.2. Explanation to how the variables are used

Petroleum products are not pure compounds but mixtures of different hydrocarbons and the values of toxicity, density, vapour pressure and viscosity presented in the literature, are given as intervals. It was, therefore, decided to use the arithmetic mean of the lowest and highest value in the calculations.

The type of soil at an accident site is not always clear and easy to judge. At industrial sites, the soil might consist of filling which can be mixtures of many different materials. However, a filling is often highly permeable and was, therefore, regarded as sand or gravel in the calculations of the hydraulic conductivity.

The amount of the chemical used was the maximal transported or stored amount of the pure compound (or mixture). The reason for this is that a worst-case scenario is desired when calculating the EAI to avoid underestimation of the situation. Therefore, mixtures with water were recalculated as pure compounds.

The water solubility (S_w) was used as follows: if the solubility is given as <1 wt.%, the value 0.1 wt.% was used. If the solubility is given as $\ll 1$, the value 0.01 wt.% was used. If the solubility is given as complete mixable or >90 wt.%, the value 90 wt.% was used.

For the hydraulic conductivity (K') which covered several magnitudes of range, it was considered precise enough to make a division into three classes: ($=1$) for values <1 m/day, ($=2$) for values 1–10 m/day and ($=3$) for values >10 m/day.

For the variable DNW, the ditches connected to watercourses were also accounted for due to the risk of further spreading of the chemical from ditches to larger watercourses or lakes. For accidents where the chemical was spilled directly into water, DNW was set to 0.01 m for modelling–technical reasons.

For the variable DGS, the accidents where there were no data available on the depth to the groundwater surface, the mean value of the other accidents, with known DGS and the same type of soil, was used as estimation. For accidents where the chemical was spilled directly into water, DGS was set to 0.01 m for modelling–technical reasons.

For some chemicals the vapour pressure is very low and the losses to air were considered negligible. Hence, the data were replaced with vl = very low in the data table (Table 1). For other accidents information on vapour pressure for the chemical and hydraulic conductivity and porosity for the surrounding could not be found and were, therefore, labelled u = unknown in the table.

2.1.3. Pre-treatment of dataset

The variables P_v, D, m/Tox, S_w, DNW and DGS were log transformed and the variable viscosity (v) was employed a square root transform to approach normality. Before calculation the data was mean centred and scaled to unit variance in order to allow each variable equal opportunity to influence the model.

2.2. Methods

Principal component analysis (PCA) was selected as means to get an overview of the dataset (Table 1) when taking all variables into account, a strategy described by Eriksson [20]. This overview formed a basis for the selection of a subset of representative accidents for training and validation of the EAI.

2.2.1. Data analytical method: PCA

PCA is an analytical projection method designed to extract the systematic variation in large data tables and get an overview of patterns and trends in the data. PCA can handle dependent descriptor variables, such as the group of chemical property variables. Another advantage is that calculations are made without incorporating any assumptions concerning physical laws or the mathematical model. The reason for choosing PCA and not, for example, factor analysis (FA) lies in the function of the underlying model. The objective of this paper is to make a selection from the original dataset to get a set of objects where all different types of chemical accidents are represented. Therefore, a maximal spreading or variability among the objects is desired, which is a specific feature of PCA on the contrary to FA which explains the structure or the correlation in the data. A comparison of these two methods has been done by Jackson [21].

By analysing the multivariate descriptor matrix with PCA the original large number of descriptors are contracted to a few information-rich principal components (PCs). We use these PCs as design variables because they are independent and can thus be used as variables in statistical designs.

In PCA, displaying the dominant trends in plots facilitates the interpretation of the variation in the data. These plots are used to study relationships between objects (score plot) and between variables (loading plot). Since, directions in the score and loading plots are the same, these plots together can be used to study which variables have large influence on the objects and vice versa. Outliers can be detected using the statistic Hotelling's T^2, which is a multivariate generalisation of Student's *t*-test, Eriksson et al. [22]. This statistic can detect observations that are extreme or that do not fit the PCA model well. In the variables score plots, the Hotelling's T^2 defines the normal area corresponding to, for example, 95 or 99% tolerance level and is given by a tolerance ellipse. In this work, the 95% level was used, meaning that N (=55 accidents) $\times$ 0.05 = 2.75 observations (accidents) were expected to be outside the ellipse. For a thorough presentation of PCA [22].

The number of significant PCs are usually determined using cross-validation but in this study, the number of significant components in the model was evaluated using the eigenvalue criteria. An eigenvalue above 1 was considered significant, as suggested by Jackson [21].

2.2.2. Statistical experimental design

The selection of a subset of chemical accidents had a need to be done in such a way that it would cover a sufficient range of variation in all the important variables for the index. By using statistical multivariate design, such as factorial or fractional factorial designs, schemes are generated that introduce systematic variation of all variables simultaneously. Varying all the variables at the same time in such systematic way will insure that the whole experimental area is investigated which is well described by Box et al. [23].

With the present type of data the "experiments" were chemical accidents, which contained natural and uncontrolled data and they could not be performed at the desired combination of variables and levels. In addition, the more or less correlated variables in Table 1 needed to be condensed into a few dominant and orthogonal descriptors, before a statistical design could be applied. By subjecting the data table to PCA, new descriptors were derived. These new variables, the PCs, which summarise the information present in the original variables, are also commonly referred to as principal properties (PPs), as described by Skagerberg

[24]. Their applicability as design variables comes from the fact that PPs are few, orthogonal and contain most of the information in the data from which they were derived. This method, statistical multivariate design, hence allows a large number of variables to be a part of the design.

Chemical accidents well separated from each other in the score plot of the PCA were obvious candidates for selection because they represent a systematic spread in all properties.

Mathematically, the selection was based on a two level full factorial design in the derived PPs, i.e. the score values. The design was complemented with three center points to provide information about the curvature and to give a rough estimation about quadratic and interaction terms. The selections were made according to a 2×2^3 (=16 accidents) design. This means that from each design level two representative accidents were selected. Then, for future modelling and validation (see strategy in Section 1.1) the 16 selected accidents were divided in two subsets, a training set for the building of a model and a validation set for the validation of the model each containing eight accidents and complemented with two and one center points, respectively.

2.2.3. Evaluation of the selected accidents

Because of constrains in the experimental space there was a need to evaluate that the selection fulfils the desired criteria. Therefore, PCA was also used to evaluate how well the selected accidents span the property space of the dataset (chemical inherent properties and properties of the surroundings). By studying the separate PCA score plots calculated for the two groups, respectively, it would be possible to evaluate if the selected accidents efficiently span the domain of the variables within each group.

A separate evaluation of the diversity and coverage of the selected accidents with respect to the range of each individual variable, according to the original dataset, was also made.

3. Results and discussion

A three-component PCA model was calculated for the whole dataset (Fig. 1a–d). The model explained 52% (PC1 (22.3%) + PC2 (16.7%) + PC3 (12.5%)) of the variation in the dataset (R2(X)) but the predictive power was low according to cross validation. This is to be expected with this type of data, containing "natural" non-designed data, such as soil and groundwater conditions and can be accepted since prediction, in this case, was not the objective with the PCA modelling. Instead, the number of significant components was evaluated using the eigenvalue criteria.

Looking at the score plot of the two first components (Fig. 1a) the two major chemical groups of the dataset, diesel fuel (accidents 8–28) and gasoline (accidents 29–40), can be discerned as clusters of data points. From the corresponding loading plot (Fig. 1b), it is obvious that the chemical inherent properties vapour pressure (P_v) and *m*/Tox (negative influence), water solubility (S_w) and density (D) (positive influence) influence PC1. Three variables: viscosity (υ), the slope of the groundwater surface (SGS) (positive influence) and depth to groundwater surface (DGS) (negative influence) explain PC2. The third component is controlled by the properties of the surroundings (Fig. 1d) and hence, the accidents are not equally clustered along PC3 in the score plot (Fig. 1c). PC3 is mainly explained by variables

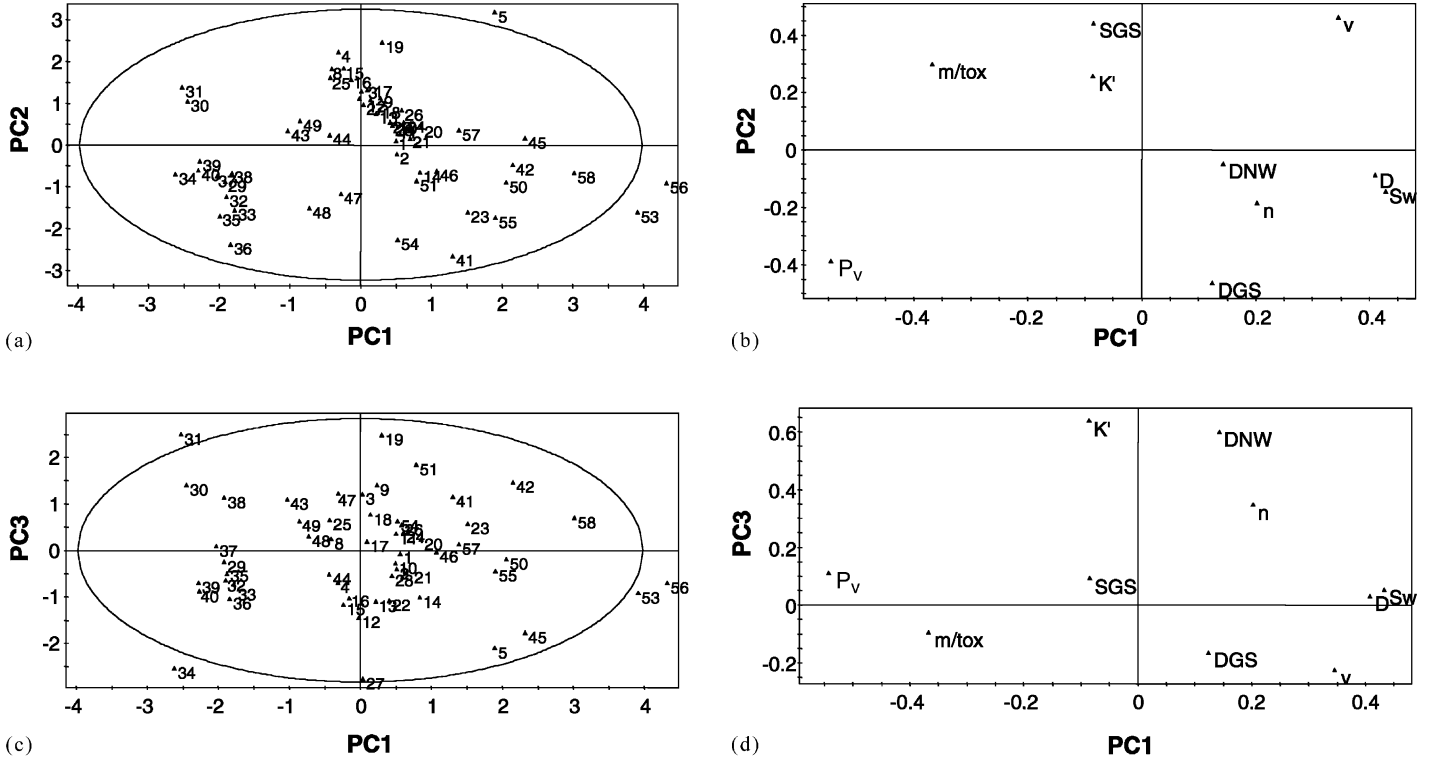

Fig. 1. PCA overview of the dataset. PC2/1 and PC3/1: variable scores (a and c) and variable loadings (b and d). Hotelling's T^2 (0.05) is given by the tolerance ellipse. For identification of objects (chemical accidents) and variables, see Tables 1 and 3.

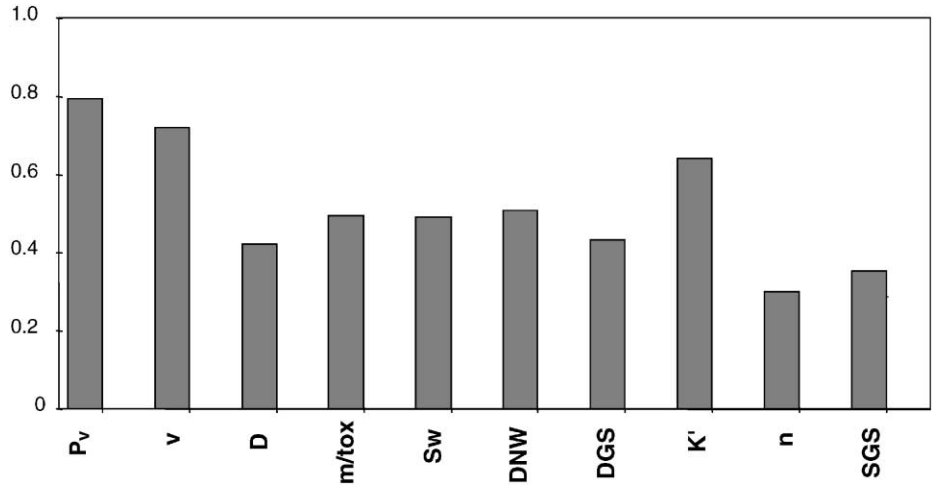

Fig. 2. Model overview. Shows the extent to which each variable contributes to the model after three calculated components. From 0 = no contribution to 1 = total contribution.

related to the properties of the surroundings: the hydraulic conductivity (K'), the distance to nearest well, lake or watercourse (DNW) and the porosity (n) which have a positive influence in the third component. Hence, PC1 can be considered a chemical property vector and PC2 and 3 mainly as surrounding property vectors. Looking at the model overview (Fig. 2) after three calculated components the properties of the surroundings and the chemical inherent properties are equally important for the model even though the overall importance is slightly higher for the chemical inherent properties.

When the calculated PCs or PPs (PC1–3 in Table 2) were used as design variables in a 2^3 full factorial design, 5–10 chemical accidents were found at each design level (Table 3). Of those, one representative accident from each design level plus two center

Table 3
2^3 design with the candidates on each design level and the selected chemical accidents[a]

Design levels			Candidates	Selected
t[1]	t[2]	t[3]		
−	−	−	29, 32, 33, 34, 35, 36, 39, 40	32, 35[a]
+	−	−	2, 14, 46, 50, 53, 55, 56	50[a], 55
−	+	−	4, 12, 15, 16, 44	15[a], 16
+	+	−	1, 5, 13, 21, 22, 10, 27, 28, 45	21, 45[a]
−	−	+	37, 38, 47, 48	47[a], 48
+	−	+	23, 41, 42, 51, 54, 58	41[a], 42
−	+	+	8, 25, 30, 31, 43, 49	30, 31[a]
+	+	+	3, 9, 11, 17, 18, 19, 20, 24, 26, 57	3, 9[a]
0	0	0	1, 44, 49	1[a], 44, 49[a]

[a] Accidents that form the training set.

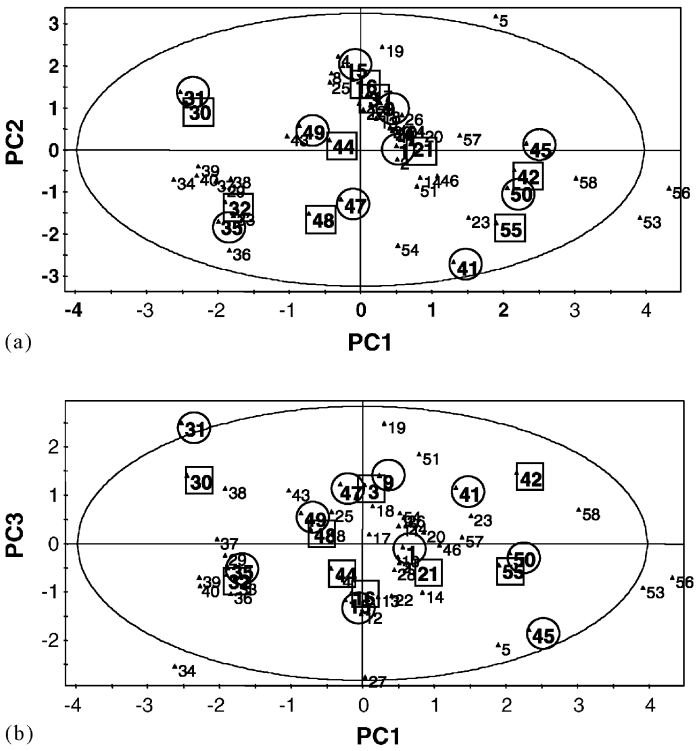

Fig. 3. Selected training and validation sets. Accidents in the training set are marked with a circle and the accidents in the validation set are marked with a square.

points was selected to be included in the training set. Then the procedure was repeated, although with only one center point, to form the validation set. In the selection, the most extreme accidents were avoided (with the exception of accident number 31, which is a good example of a transport accident and therefore included). Other knowledge about the accidents (such as the magnitude of available information, investigations made after the accident, etc.) was also considered. The selected accidents are also shown in the score plots of Fig. 3a and b. Fig. 3a and b demonstrate that the selected accidents cover the investigated PP space.

3.1. Analysis of selected accidents

Because of constrains in the experimental space it was also important to evaluate the selection to see if the selected subset was the best possible and fulfils the desired criteria for further modelling.

Further, in order to get a picture of how well the selected accidents represented the two groups of variables in the dataset (chemical properties and properties of the surroundings) separate PCAs were calculated for each group, respectively.

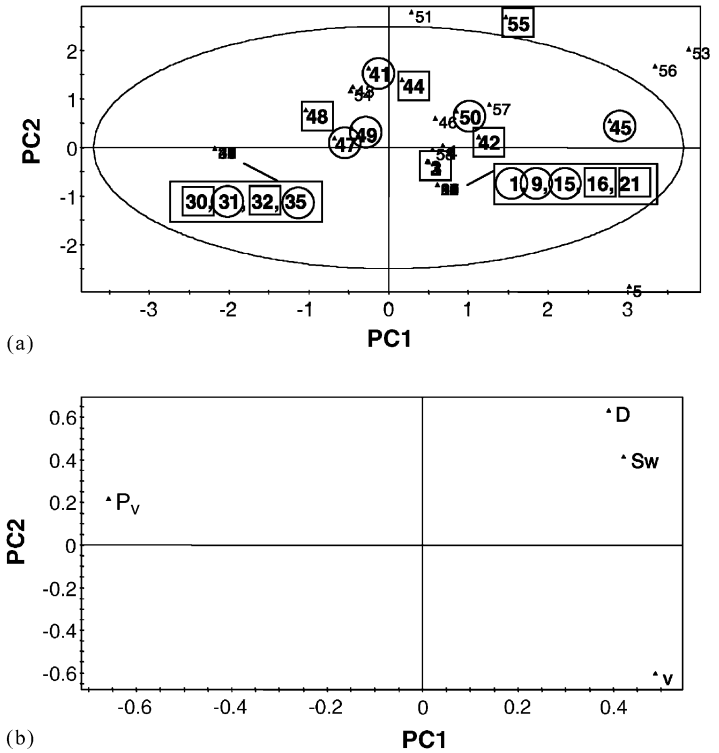

Fig. 4. Coverage of the range of variation in chemical properties for the selected accidents, PC2/PC1: (a) variable scores; (b) variable loadings. Hotelling's T^2 (0.05) is given by the tolerance ellipse.

The resulting two-component score plot of the chemical inherent properties is shown in Fig. 4a. The two groups of diesel fuel and gasoline can clearly be discerned. The loading plot (Fig. 1b) demonstrate that PC1, which has more influence on the model than PC2 (49.7% compared to 25.1%), is explained by all variables. The most important of those is vapour pressure, which has a large negative influence. Density, water solubility and viscosity are separated along PC2. The positions of the chemical property variables within the loading plot will give different spreading scenarios in each of the four quadrants. Depending on in which of the quadrants the chemical accidents are positioned they will represent different spreading scenarios regarding the chemical inherent properties. The selected accidents (Fig. 4a) are not covering all quadrants and hence, not all types of spreading scenarios. However, since the original dataset, based on real cases, does not either have a balanced coverage of all four quadrants, the selected accidents can be regarded as the best possible, covering the real situation.

In Fig. 5a, showing the two-component PCA score plots of the properties of the surroundings, the selected accidents are randomly distributed without strong groupings. The loading plot (Fig. 5b) shows that DGS has the main influence on PC1, followed by SGS and K' and that PC2 is explained by DNW. Only the porosity of the soil (n) is less well

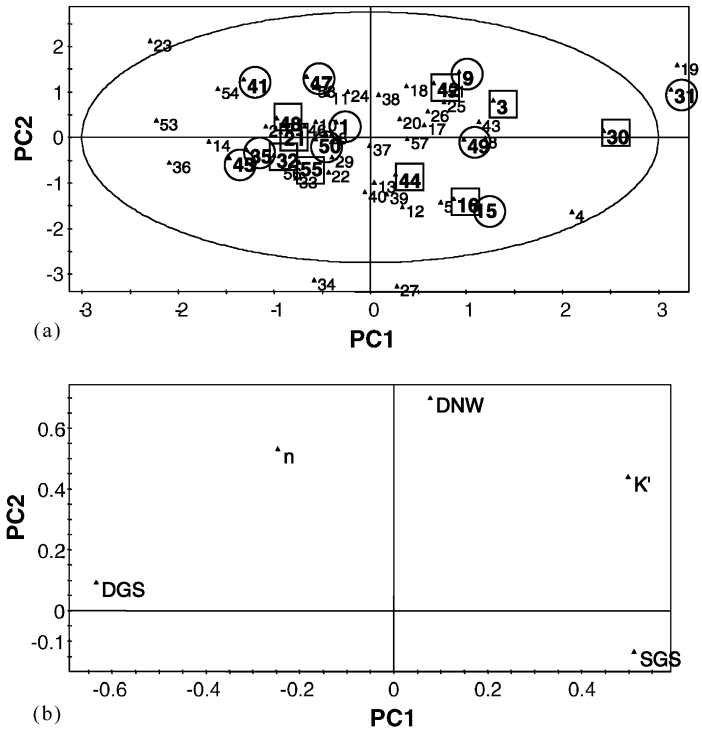

Fig. 5. Coverage of the range of variation in properties of the surroundings for the selected accidents, PC2/PC1: (a) variable scores; (b) variable loadings. Hotelling's T^2 (0.05) is given by the tolerance ellipse.

explained. The four quadrants will represent different spreading scenarios, based on the positions of the variables, with respect to the properties of the surroundings. Of the selected accidents number 31 is a weak outlier, influenced by K', but overall the score plot in Fig. 5a shows that the selected accidents are well representing the different spreading scenarios and will be good representatives of the original dataset with respect to the properties of the surroundings.

The analysis of the separate score plots shows that the selected accidents are representative for the original dataset with respect to both groups of variables. However, a closer look at each variable according to the selected accidents is necessary to see if the selected accidents are well distributed within the variables and not only covering a single measure. This can be seen in the column plots of each variable (Fig. 6). In Fig. 6, all variables seems to be quite well represented, within their respective ranges, of the selected training and validation sets, except for variables SGS and DNW. This means that these types of surrounding circumstances could be underestimated. SGS has one class (0.1; no well, lake or watercourse) that is not represented. However, this situation, pointing at a low-risk scenario for spreading of the chemical to the water environment, is not very common (two of 55 accidents) and will, therefore, probably not cause any underestimation of the situation when using the EAI in the future. The variable DNW also has one scenario missing and that is release of chemical

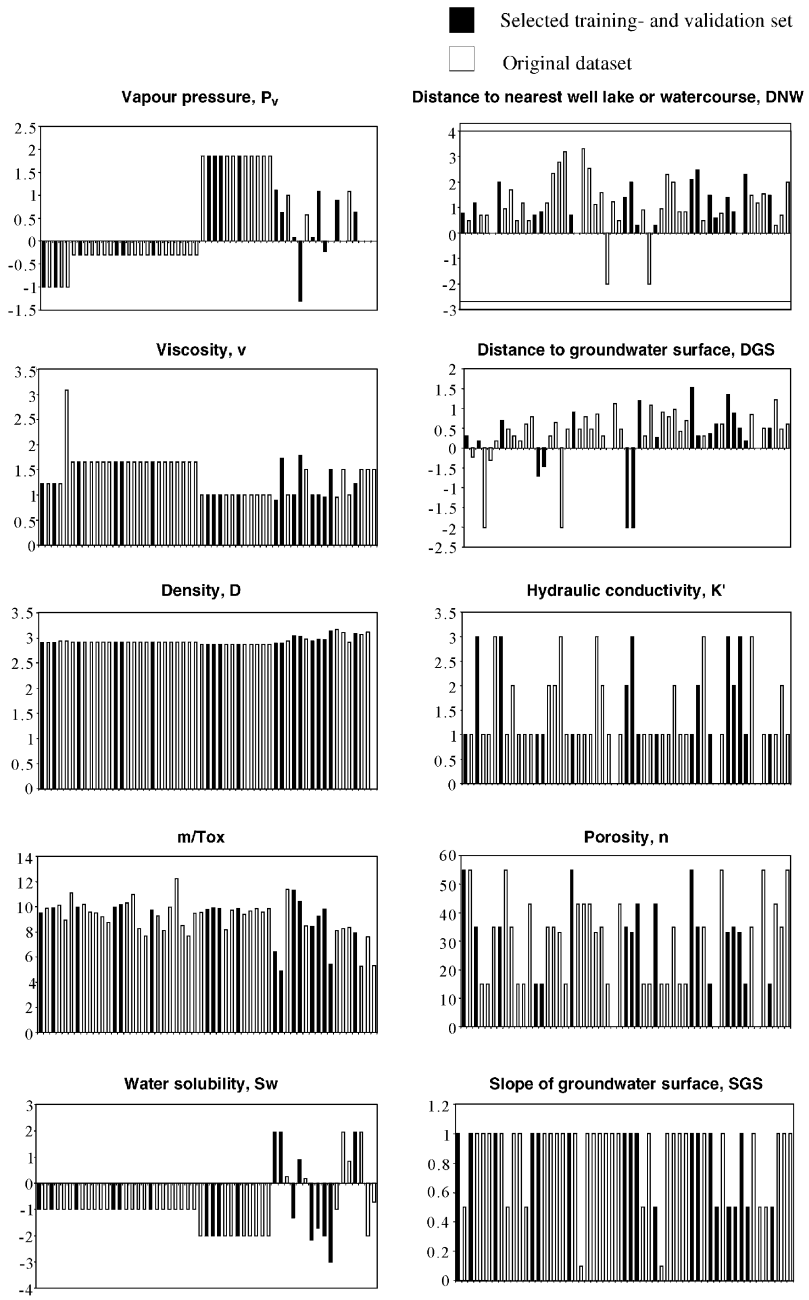

Fig. 6. Column plots showing, for each variable, the pattern of each accident in the original dataset in light grey colour. The selected training and validation sets are dark grey coloured. Transformed variables have been used due to large range of variation in the data.

directly into water, DNW = 0.01 m. However, in the original EAI [1], DNW between 0 and 10 m was regarded to belong to the same risk category and was, therefore, given the same point during the calculations. Therefore, this type of scenario can be covered by other selected accidents that occurred within 10 m from a well, lake or watercourse, such as accident numbers 32, 15, 45 and 49.

4. Conclusions

The objectives of this paper were to select a representative set of chemical accidents for the further development of EAI and to study if it is possible to use PCA in combination with statistical experimental design to do the selection. The assembled dataset consisted of both laboratory produced data and natural data (such as the ones concerning the soil and the groundwater) and was highly skewed because it covered a large range of compound specific properties and properties of the surrounding. Conclusions that can be made from this work are:

- Although the material is complex and the selection was made according to a less stringent significance criterion, eigenvalue 1, PCA and multivariate design can be used to make an objective selection of representative subsets of data to be used for modelling and evaluation of the EAI.
- Evaluation of the selection has demonstrated that both the descriptors for chemical inherent properties and properties of the surroundings were well represented in the selected sub set of chemical accidents. The only exceptions were the variable SGS and DNW but the different scenarios were either low-risk or covered by other accidents and will, therefore, probably not cause any underestimation of the situation when using the EAI in the future.
- The selected subset of chemical accidents can thus be used in the work of developing criteria for how to judge the environmental consequences of a chemical accident.

Acknowledgements

This work was financially supported by the Center for Environmental Research in Umeå (CMF) and the Swedish Research Agency (FOI), which are gratefully acknowledged. Special thanks to all friends and colleagues for fruitful discussions and advice on the subject.

References

[1] Å. Scott, J. Hazard. Mater. 61 (1998) 305–312.
[2] Å. Scott, Material from chemical accidents occurred (1986–1999), Swedish Defence Research Agency, Umeå, Sweden.
[3] Swedish Rescue Services Agency, Räddningsinsatser 1996, Karlstad, 1997.
[4] Swedish Rescue Services Agency, Räddningstjänst i siffror, Karlstad, 1998.
[5] Swedish Rescue Services Agency, Räddningstjänst i siffror, Karlstad, 1999.
[6] Swedish Rescue Services Agency, Räddningstjänst i siffror, Karlstad, 2000.

[7] BUA report, Serie, German Chemical Society, Wissenschaftliche Verlagsgesellschaft, Stuttgart, 1995.
[8] CONCAWE, Gas oils (diesel fuels/heating oils), Product dossier no. 95/107, Brussels, 1996.
[9] CONCAWE, Gasolines, Product dossier no. 92/137, Brussels, 1992.
[10] CONCAWE, Kerosenes/jet fuels, Product dossier no. 94/106, Brussels, 1995.
[11] CONCAWE, Heavy fuel oils, Product dossier no. 98/109, Brussels, 1998.
[12] G. Hommel, Handbuch der Gerfärlichen Güter, Springer, Berlin, 1987.
[13] HSDB, Hazardous substances databank, The National Library of Medicine, USA.
[14] ChemFinder, Internet provided database on chemicals, www.ChemFinder.com, Mars, 1999.
[15] Association of Swedish Chemical Industries, Miljöskyddsblad.
[16] E. Nikkunen, R. Leionen, A. Kultamaa, Environmental properties of chemicals, Research report 91/1990, Ministry of the Environment, Environmental Protection Department, Helsinki, 1991, p. 1084, ISBN 951-47-3539-0.
[17] OECD, High volume production chemicals program, Draft-risk assessment reports on human & environmental effects, 1990.
[18] Svenska brandförsvarsföreningen, Farligt godskort.
[19] P. Engström, K. Gustavsson, Rörlighet hos förorenande vätskor, särskilt petroleumprodukter i mark-och grundvatten: en kunskapsöversikt, Uppsala Universitet, 1988, p. 102.
[20] L. Eriksson, A strategy for ranking environmentally occurring chemicals, Ph.D. thesis, Research Group for Chemometrics, Department of Organic Chemistry, Umeå University, 1991.
[21] J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991, ISBN 0-471-62267-2.
[22] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Introduction to Multi- and Megavariate Data Analysis Using Projection Methods (PCA & PLS), Umetrics AB, 1999, p. 490.
[23] G.E.P. Box, W.G. Hunter, J.S. Hunter, Statistics for Experimenters, Wiley, New York, 1978.
[24] B. Skagerberg, Principal properties in design and structural description in QSAR, Ph.D. thesis, Umeå University, 1989.