NILU
OPPDRAGSRAPPORT NR: 32/1982
REFERANSE: 20382
DATO: JUNE 1982

PRINCIPAL COMPONENT ANALYSES OF PAH DATA FROM SUNDSVALL

BY

RONALD C. HENRY*

NORWEGIAN INSTITUTE FOR AIR RESEARCH
P.O.BOX 130, N-2001 LILLESTRØM
NORWAY

*On leave from Environmental Research & Technology, Inc.
2625 Townsgate Rd.
Westlake Village,
California 91361, USA

ISBN 82-7247-327-5

TABLE OF CONTENTS

PRINCIPAL COMPONENT ANALYSES OF PAH DATA FROM SUNDSVALL

1 INTRODUCTION

This report covers the analysis of intercorrelation between components of the PAH at four sites in Sundsvall, Sweden. Differences between sites in average concentrations or frequency distribution are not discussed. Principal component analysis (PCA) is used to find statistically independent linear combination of the PAH's, of which only three or four can explain 75% or more of all the variability in the data. The form of these independent components reveals the underlying structure of the data as a whole. The principal components are interpreted as the balance of effects of meteorology and source composition. Target transformation factor analysis is applied to test the consistency of the observed PAH correlations with assumed source PAH compositions.

2 DESCRIPTION OF THE DATA

The measurements consist of the sum of particulate and vapor phase concentration of PAH's and fluorides. Sampling techniques and analysis are described elsewhere (1). Of more than 30 PAH's analysed, 21 were consistently found in measurable quantities at the four sites. The PCA analysis was limited to these compounds and fluoride. Table 1 lists them, along with their identifying numeric codes, compound 1050 suffered from some analytical problems. Typical minimum detectable limits of the compounds are also given in Table 1. All the data are for 24 hour sampling periods, in some cases concentrations from two 12 hour samples were summed to give a 24 hour composite sample. Samples were collected about once a week between June 1980 and July 1981.

The collection sites are shown in Figure 1. There are between 37 and 44 complete days of data from each site.

Table 1: Variables in PCA.

Code number	Name	Typical minimum detectable limit (ng/m^3)
1000	Fluorides	≃ 10
1010	Naphtalene	0.1
1040	Biphenyl	0.1
1050	Acenaphthene	0.1
1060	Fluorene	0.1
1070	Dibenzothiophen	0.1
1080	Phenanthrene	0.1
1090	Anthracene	0.1
1120	1-methylphenanthrene	0.1
1130	Fluoranthene	0.1
1140	Pyrene	0.1
1150	Benzo(a)fluorene	0.1
1160	Benzo(b)fluorene	0.1
1170	Benzo(a)anthracene	0.1
1180	Chrysene/Thriphenylene	0.1
1190	Benzo(b,j.k)fluoranthene	0.2
1210	Benzo(e)pyrene	0.2
1220	Benzo(a)pyrene	0.2
1240	O-phenylenepyrene	0.2
1260	Benzo(g h i)perylene	0.2
1280	Coronene	0.2

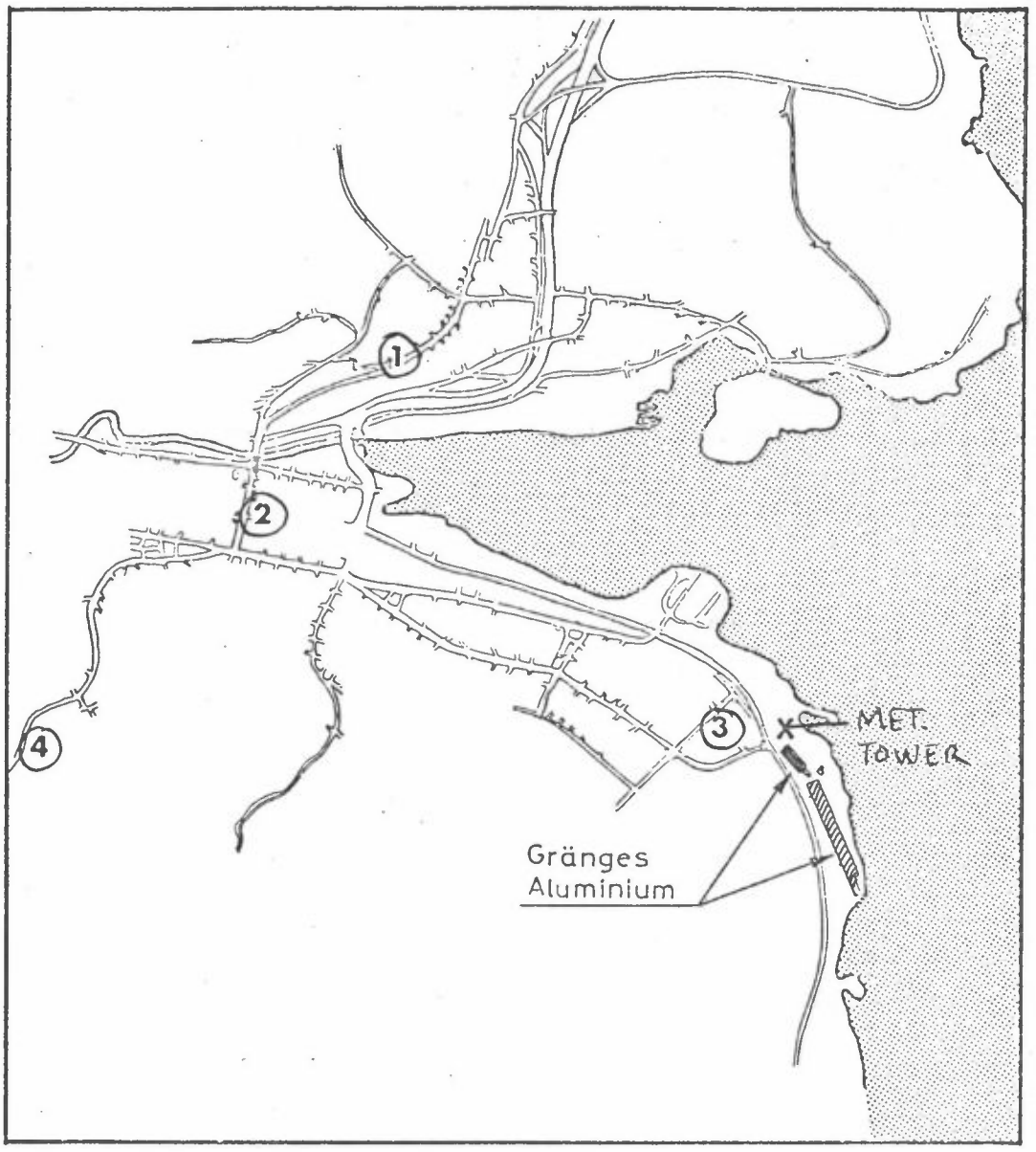

Figure 1: PAH sampling sites in the Sundsvall area.
1 = Haga, 2 = Købmannsgatan (KGT)
3 = Kubikenborg (KU), 4 = Sidsjön (SID)

3 PRINCIPAL COMPONENT ANALYSIS (PCA) METHODOLOGY

By means of eigenvectors of the correlation matrix, the normalized variables (PAH's) are expressed as

$$\tilde{V}_{ik} = \frac{V_{ik} - \bar{V}_i}{\sigma_i} = \sum_{j=1}^{N} a_{ij} P_{jk} \qquad (1)$$

where:

V_{ik} = kth value of the ith variable
$\tilde{V}_{ik}$ = normalized variable (average equals zero and standard deviation of one)
$\bar{V}_i$ = average of the ith variable
σ_i = standard deviation of the ith variable
a_{ij} = loading of the ith variable on the jth component
P_{jk} = kth value of the jth component

Further discussion and description of PCA of air quality variables is found in Henry and Hidy (2) . Table 2, reproduced from this paper, gives the basic means for interpretation of the component loadings, the a_{ij} in (1) above. The following analysis is based on interpretation of the principal component loadings.

3.1 Results

The results will be discussed in terms of what they reveal of the relative variation of the PAH's, connection to source composition, and consistency between sites. First, broad, common feature of analysis shared by all sites are discussed, followed more detailed description.

3.1.1 Overview of results

The analysis presented here is based on PCA of the 22 variable of Table 1, normalized to total PAH for that day.

Table 2

INTERPRETATION OF COMPONENT LOADINGS LISTED IN TABLES 4 & 5

Mathematical formula $Z_{ik} = \sum_j A_{ij} P_{jk}$, for K = 1,...,M data points.

ith Normalized observable — Component loading — jth Principal component

Value of A_{ij} (x1000)	Meaning
Near zero (-200 to +200)	Almost no correlation or dependence of variable i on component j.
Large positive (+900 to +1000)	Strong positive correlation. If the component is positive, the variable is above its average value; if the component is negative, the variable is below its average; if the component is zero the variable is near its average value.
Large negative (-900 to -1000)	Strong negative correlation. If the component is positive, the variable is below its average; if the component is negative the variable is above its average value.
Moderate ($\pm$200 to $\pm$900)	Same interpretation as above but with a proportionately less strong correlation.

PCA was also carried out on 29 variables, not normalized to total PAH and for the same 22 variables also not normalized. The normalization helps discount the overwhelming influence of a few samples with very high absolute loadings. The use of 29 variables (many of them equal to zero) resulted in a correlation matrix so illconditioned that its eigenvectors could not be determined.

As shown in Table 3, only 5 to 8 independent combinations of the 22 variables are needed to explain all but 10% of the variation in these 22 variables. In fact only 3 principal components explain 68-78% of the total variability of the data. Thus, PCA shows that there are, effectively three major sources of PAH variation and two to five minor sources. Some of these minor sources probably reflect minor problems in the data quality, so that there are 3 or 4 major components that are likely related to sources and meteorology.

The single combination of variables able to account for the maximum of variance is defined as the first principal component. At all sites this component has a similar structure of component loadings (see Table 4). The light PAH's 1010 and 1040 have opposite sign to the vast majority of the other PAH's. This means that, much of the time, when 1010 and 1040 are high (above average) the others are low (below average) and vise versa. Compound 1280 is an exception, it is not strongly associated with the first component. This one principal component explains 35 to 48% of the total variability, and is thus a major feature of the data. Without further analysis, including meteorological variables, the physical interpretation of this, or other components is not certain, yet it is clear that fluoride (1000) and PAH's 1130 to 1240 are negatively correlated with 1010 and 1040. This may be a sampling artifact with low values of 1010 and 1040 due to lower collection efficiencies during high temperature periods when, relatively more of the higher molecular weight PAH's would be found. It could also be caused by actual physical or chemical degradation of light PAH's or negatively correlated source emissions. The medium molecular weight PAH's 1050 to 1120 are not so heavily involved in this

Table 3: Overview of PCA results.

Site	Number of significant components	Percentage of total variance of the date explained by these	Number of days of data
KGT	5	91.1	31
KU	7	89.1	42
HAGA	8	92.3	37
SID	8	90.7	38

Table 4: Loadings of first principal component (x10000) at each site (see Table 2 for guide to interpretation).*

*variables with the most consistent behaviour are boxed.

Variable	Site			
	KU	KGT	HAGA	SID
1000	6814	6743	4720	5791
1010	-6724	-6685	-6981	-5429
1040	-7587	-4551	-6200	-6079
1050	-3846	-4354	920	-1494
1060	1868	6331	-438	-1949
1070	3579	7305	3569	2578
1080	5104	9088	3863	1929
1090	5961	7588	4154	1028
1120	348	5184	2464	465
1130	5315	9537	7774	5031
1140	8823	9596	7290	7191
1150	6928	8465	5343	6245
1160	9024	9024	6544	4612
1170	8717	7216	7928	8851
1180	9136	8508	8386	8855
1190	6876	7583	7184	8177
1210	8556	7316	7623	8292
1220	8214	3531	7142	7698
1240	6112	5998	6615	8067
1260	6671	2290	7317	7949
1280	1025	626	-1714	-959
% of total variability explained	45.2	48.4	35.3	35.0

component, as is the case with 1280. These compounds tend to figure strongly in other components.

In particular, 1280 is thought to be a "tracer" for vehicular sources of PAH's. At the KGT site, located close to heavy trafic, there is a strong principal component highly correlated with 1260 and 1280 (see Table 5).

Table 5: Vehicular source component from PCA on KGT data.

Variable	Component loading (x 10000)
1000	-4046
1010	4075
1040	5945
1050	1339
1060	- 372
1070	-4823
1080	-2185
1090	-1324
1120	-1055
1130	-2110
1140	1216
1150	- 626
1160	-1048
1170	1481
1180	- 181
1190	4339
1210	5932
1220	6940
1240	5983
1260	9121
1280	8278

It explains 19% of the variability at KGT. Similar, but less strong indication of vehicular source influence are found at all the sites. Other major principal components vary in structure from site to site. Also at each site there is a component explaining about half of the variability of compound 1050 which is believed to be related to some analytical difficulties.

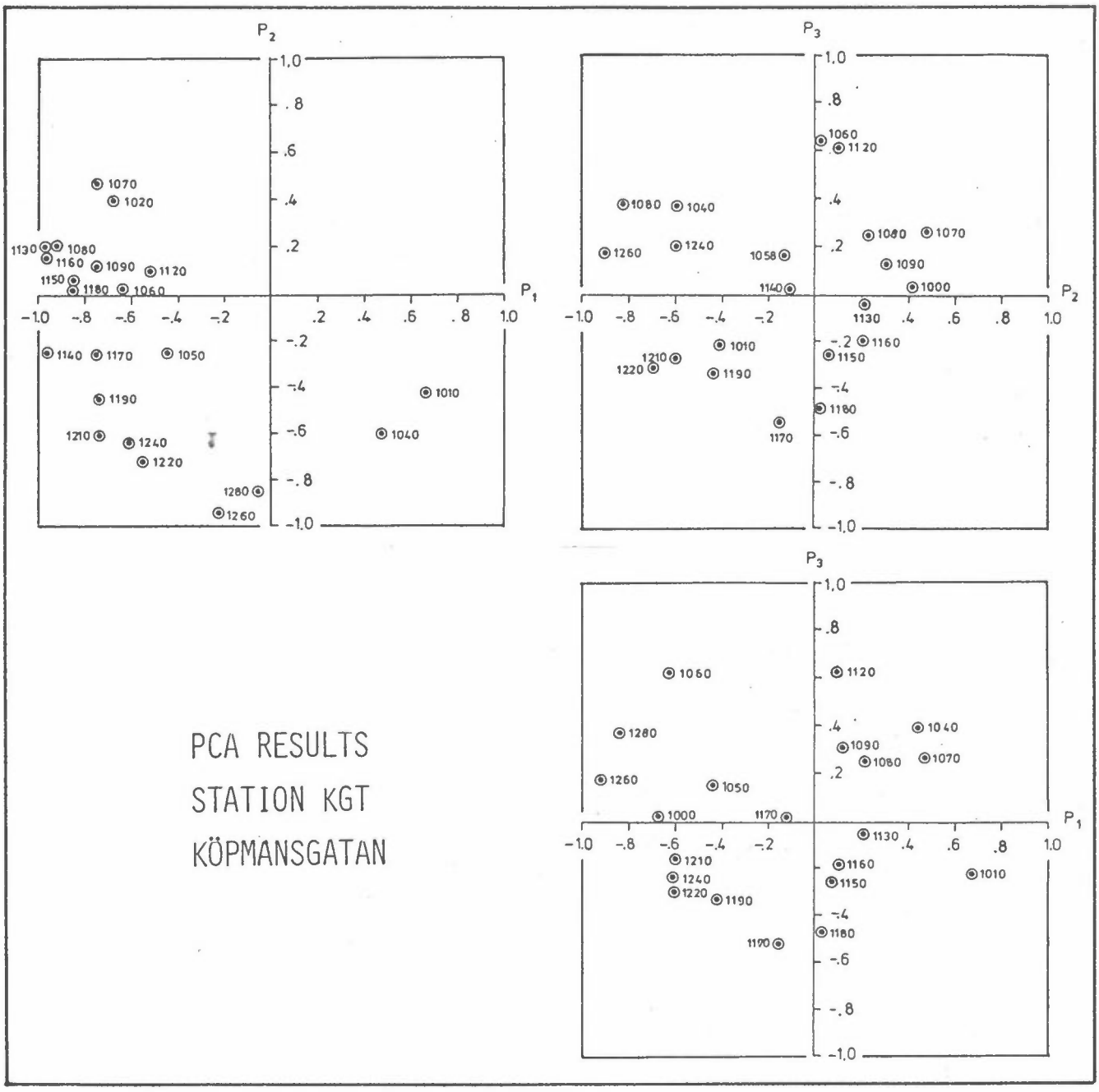

Figure 2: Principal component analysis results for Köpmansgatan.

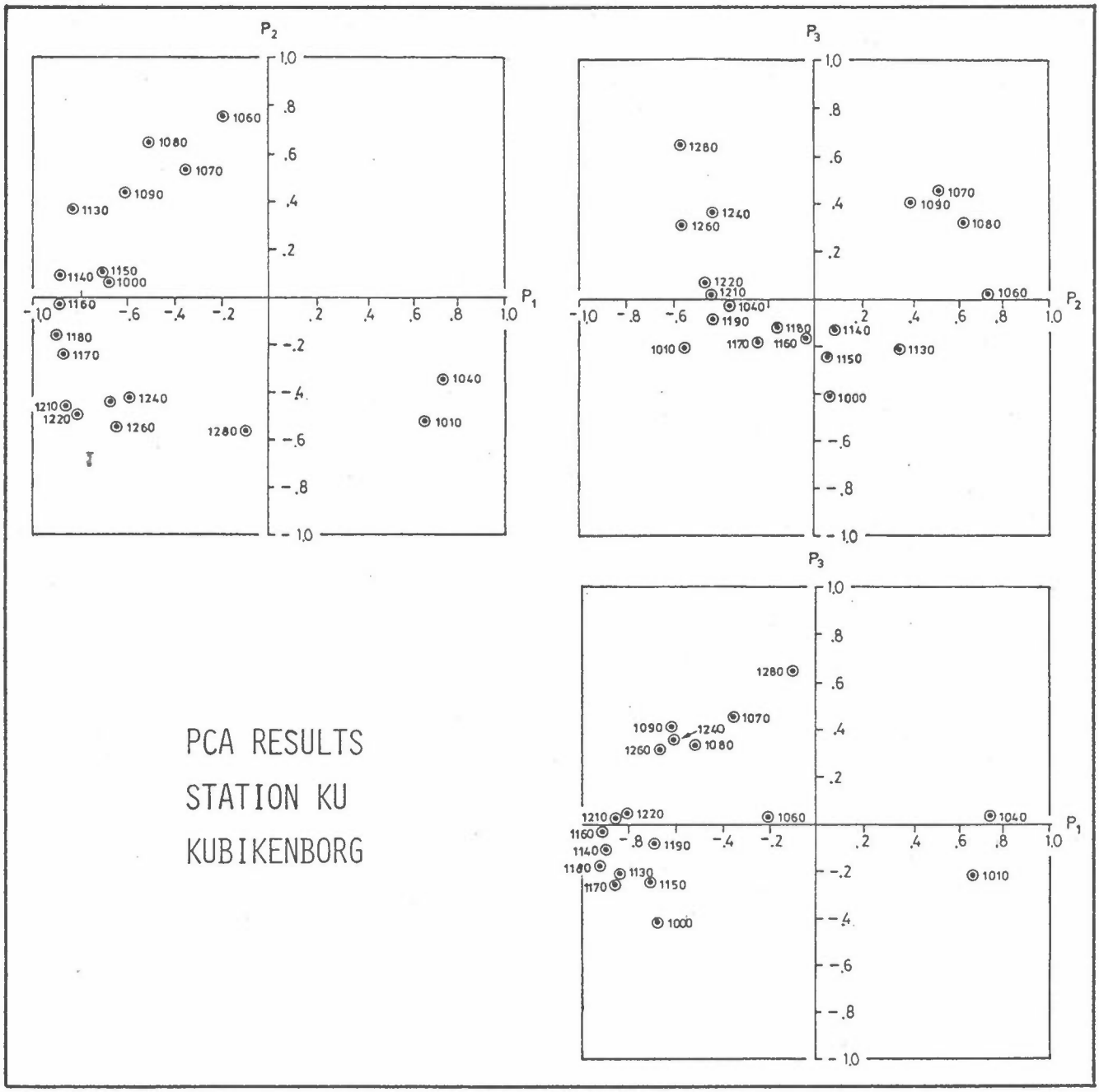

Figure 3: Principal component analysis results for Kubikenborg.

In general, the structure of the data as revealed by PCA is about as consistent from site to site as can be expected for sites with only about 40 days data out of more than a one year period.

3.1.2 Specific results

The detailed structure of the first three major components for KU and KGT (sites with the highest PAH concentrations) are given in Figures 2 and 3. In these figures the loadings of components 1, 2 and 3 are plotted against each other. The object of these plots is to graphically display the groupings of the variables. Note that the statistical uncertainty in each point is about ± 0.1. If we think of variable 1000 (Fluoride) as a tracer for aluminium plant emissions and 1280 (Coronene) as a tracer for vehicular source, then the graphs of P_1 vs. P_2 show the PAH's spread out between these two poles, the lighter ones tending nearer fluoride and the heavier ones nearer coronene. This is in accordance with our knowledge of the composition of the source emissions. Traffic tending to dominate in the heavier PAH's while the aluminium plant tends to have more lighter PAH's in its emissions. However, the plots of P_1 vs. P_3 and P_2 vs P_3 show that all is not so simple. The grouping of heavier PAH's 1240, 1260, 1280 is seen to hold in general, probably reflecting a common major source (traffic) but the middle weight PAH's are widely spread out. Compound 1070, 1080 and 1090 tend to group together perhaps reflecting a common source ancestry, such as long range transport or some local source such as incineration or oil burning.

3.2 Effects of outliers and choice of variables on PCA results

PCA relies on analysis of correlation which are very sensitive to data outliers. It was decided to run the PCA with and without days with questionable and/or very high (or very low) PAH loadings. To this end, data from KU and KGT sites were screened and three days with extreme values removed from each data set and the PCA repeated. No significant change (> ± .1) in component loadings of the major principal components were found. Also, the PCA for the KU site was repeated without variables 1050 and 1120, which were zero fairly often at this site. Again no significant changes in

loadings were observed due to this change in choice of variables. It is concluded that the results presented above are stable and not the result of a few unusual values.

4 TARGET TRANSFORM FACTOR ANALYSIS

The principal component model of the PAH data can be used to provide an improved estimate of the source PAH composition. The methodology for this is based on so called target transform factor analysis (TTFA), see Malinowski and Howery (35) for the background of this technique. The basic idea is to transform the chemical species balance (CSB) model into a form which corresponds to the PCA model. Then the PCA model is transformed to match a "target" assumed source composition vector as closely as possible, in a least squares sense. Thus, the information from the inter-correlations of the PAH's is transferred to the CSB model using the PCA model. This procedure is rather technical and the details are beyond the scope of this report.

The results are given in Tables 6,7 and 8 for the Aluminium plant, vehicular sources and a long range transport component. Each of these three tables gives the assumed or target vector, the closest approximation based on the PCA and the ratio of the two.

Table 6: Target transformation results for the aluminium plant source.

PAH NUMBER	ASSUMED SOURCE * COMPOSITION	BEST-FIT SOURCE COMPOSITION	BEST-FIT / ASSUMED
1010	6.80	-35.55	-5.23
1040	4.30	0.24	0.06
1050	6.00	-4.84	-0.81
1060	12.60	9.68	0.77
1070	3.00	3.42	1.14
1080	16.90	24.00	1.42
1090	4.50	3.32	0.74
1120	3.90	3.56	0.91
1130	5.80	9.04	1.56
1140	6.80	5.88	0.86
1150	1.00	1.46	1.46
1160	0.80	0.90	1.12
1170	0.76	-0.01	-0.01
1180	1.00	0.74	0.74
1190	1.70	2.06	1.21
1210	0.80	1.36	1.70
1220	1.20	0.82	0.68
1240	1.00	1.44	1.44
1260	2.60	2.83	1.09
1280	2.40	2.06	0.86

* As a percentage of total PAH

Table 7: Target transformation results for vehicular sources.

PAH NUMBER	ASSUMED SOURCE COMPOSITION *	BEST-FIT SOURCE COMPOSITION	BEST-FIT / ASSUMED
1010	16.50	-71.38	-4.33
1040	8.10	-3.80	-0.47
1050	0.00	-7.31	
1060	20.30	11.21	0.55
1070	4.30	6.75	1.57
1080	25.40	37.30	1.47
1090	8.00	4.39	0.55
1120	0.00	4.37	
1130	3.90	15.73	4.03
1140	2.60	6.76	2.60
1150	1.70	2.17	1.28
1160	1.70	1.39	0.82
1170	1.00	0.00	0.00
1180	2.60	1.74	0.67
1190	0.00	1.50	
1210	1.00	0.77	0.77
1220	0.30	0.22	0.74
1240	0.20	1.07	5.33
1260	0.90	0.79	0.88
1280	0.00	0.97	

* As a percentage of total PAH

Table 8: Target transformation results for long range transport.

PAH NUMBER	ASSUMED SOURCE COMPOSITION *	BEST-FIT SOURCE COMPOSITION	BEST-FIT / ASSUMED
1010	3.90	-31.50	-8.08
1040	0.50	-2.02	-4.05
1050	10.90	5.41	0.50
1060	6.50	5.72	0.88
1070	4.00	2.53	0.63
1080	27.00	28.40	1.05
1090	3.80	2.01	0.53
1120	0.90	0.85	0.94
1130	11.70	16.25	1.39
1140	8.20	9.91	1.21
1150	1.00	2.45	2.45
1160	1.20	1.53	1.27
1170	2.40	1.45	0.61
1180	3.90	5.42	1.39
1190	4.50	4.15	0.92
1210	2.40	2.24	0.93
1220	1.70	1.20	0.71
1240	1.20	1.13	0.95
1260	1.50	2.12	1.41
1280	0.30	0.74	2.47

* As a percentage of total PAH

Note that the lighter PAH's 1010, 1040 and 1050 are often negative. This physically impossible result implies that the observation of these PAH's may be seriously flamed since the data cannot be made compatible with a physically reasonable model. This is caused by the negative correlation of these PAH's with most of the PAH's, or noted in the PCA discussion. It is likely that meteorological conditions are affecting the collection efficiency for the lighter PAH's, thus confusing the PCA results.

Except for these light PAH's, there is good ability of the PCA model to be transformed in a manner consistent with the assumed source composition. Almost all PAH's from the TTFA are within a factor of two of their target. The major exception is compound 1170, which is predicted by TTFA to be much smaller in vehicular and LRT (long range transport) sources than assumed. Also, for LRT, the PAH's 1130 and 1240 are given by TTFA to be a factor 4 and 5 greater than was assumed. No specific reason for these discrepancies are known at this time. However, in general the TTFA gives results within the large uncertainty bounds for the assumed source copositions.

5 CONCLUSIONS

PCA analysis of the correlation of the PAH's shows that:

a) There are three major and two to five minor independent sources of variability of the 22 variables in the analysis

b) There are only slight difficulties in the analysis caused by data quality or analytical methods, except for the light PAH's where changes in sampling efficiency may be a problem.

c) The major principal components at each site are similar and indicates a good correlation between fluoride and many PAH's, especially compounds 1130 to 1240. This indicates the possible ubiquitous influence of the aluminium plant in the Sundsvall area.

d) Using coronene as a tracer of vehicular PAH impact, this source was clearly a major factor in PAH variability at all sites, especially the KGT site, near strong traffic sources.

e) At least one relatively strong source of PAH's 1070, 1080 and 1090, seems to be present. Its nature is not known, but could be long range transport or due to local combustion processes.

f) Target transformation factor analysis (TTFA) showed that the observed correlations of the PAH's are on the whole consistent with current limited knowledge about source compositions, except for light PAH's, which could not be explained in a physically meaningful way.

g) TTFA indicates that compound 1170 is a good tracer for aluminium plant PAH emissions.

h) Coronene (1280) is seen to be a good tracer for vehicular PAH's, except that the effect of aluminium production must also be taken into account.

6 REFERENCES

(1) Thrane, K.E.
Mikalsen, A.
High volume sampling of airborne polycyclic aromatic hydrocarbons using glass fibre filters and polyurethane foam.
Atmos. Environ. 15, 909-918 (1981).

(2) Henry, R.C.
Hidy, G.M.
Multivariate analysis of particulate sulfate and other air quality variables by principal components. Part 1. Annual data from Los Angeles and New York. *Atmos. Environ.* 13, 1581-1596 (1979).

(3) Malinowski, E.R.
Howery, D.G.
Factor Analysis in Chemistry.
John Wiley & Sons, New York.
Pp. 50-53 (1980).

NORSK INSTITUTT FOR LUFTFORSKNING
(NORGES TEKNISK-NATURVITENSKAPELIGE FORSKNINGSRÅD)
POSTBOKS 130, 2001 LILLESTRØM
ELVEGT. 52.
TLF. (02) 71 41 70

RAPPORTTYPE	RAPPORT NR.	ISBN--82-7247- 327-5
Oppdragsrapport	OR 32/82	
DATO	ANSV.SIGN.	ANT. SIDER
JUNE 1982	B. Ottar	18
TITTEL Principal component analyses of PAH data from Sundsvall		PROSJEKTLEDER B. Sivertsen NILU PROSJEKT NR. 20382
FORFATTER(E) Ronald C. Henry		TILGJENGELIGHET** A OPPDRAGSGIVERS REF.
OPPDRAGSGIVER Nordisk Ministerråd (NMR)		
3 STIKKORD (á maks. 20 anslag) PAH	Statistisk analyse	Aluminiumverk

REFERAT (maks. 300 anslag, 5-10 linjer)

TITLE

ABSTRACT (max. 300 characters, 5-10 lines.
Concentrations of fluoride and polycyclic aromatic hydrocarbons (PAH) have been determined in ambient air in Sundsvall, Sweden. Principal component analysis has been used to find statistically independent linear combustion of th PAH's. The results show that some of the PAH's are associated with the major sources in the area, such as the aluminium industry and traffic.

**Kategorier: Åpen - kan bestilles fra NILU A
Må bestilles gjennom oppdragsgiver B
Kan ikke utleveres C