Stockholms universitet

An analysis of trends and source composition of Arctic lower tropospheric aerosols using Positive Matrix Factorization

Elin Magnusson

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2016:6
Matematisk statistik
December 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2016:6**
http://www.math.su.se

An analysis of trends and source composition of Arctic lower tropospheric aerosols using Positive Matrix Factorization

Elin Magnusson*

December 2016

Abstract

In this thesis we analyse over thirty years of continuous observations of 21 constituents in the aerosol and two gases of the lower Arctic troposphere at Alert, Canada on northern Ellesmere Island. Aerosols are defined as minute solid and/or liquid particles suspended in air. The analysis is an extension and expansion of the analysis in the report "*Arctic lower tropospheric aerosol trends and composition at Alert, Canada: 1980-1995*" (Sirois and Barrie, 1999). The analysis is made in two parts, the first part is a time series analysis of the observed atmospheric concentrations between 1980 and 2013. The second is a multivariate analysis called Positive Matrix Factorization (PMF) to reduce the dimensionality of data by factorizing constituents/gases into sources. Relevant background information and theory are explained including a detailed description of PMF. The PMF analysis is made using the licensed program PMF2 while all other model and data handling is done using R. For the time series analysis, the complete data set as well as a subset of data for the winter (peak) season were analysed separately. Long term and seasonal trends have been described using cubic smoothing splines. Since Alert is situated above the polar circle, during a large part of the year it is in complete darkness until the polar sunrise occurs: this affects the compositions of the aerosol. Therefore, a new approach for the PMF analysis was made in contrast to Sirois and Barrie (1999). Two different factorizations are made: one for the dark part of the pollution period and one for the light part of the pollution period. The two factorizations are made for 19 aerosol constituents and the gases: ozone and gas phase mercury, with data between 1995 and 2007. Some of the key results are: (i) a drastic drop of aerosol sulphate, acidity, ammonium and metals related to oil combustion in the mid 1990's, most likely due to the collapse of the economy of the former Soviet Union, and (ii) the spring time correlation shown between O3 , Hg and Br related to photochemistry involving sea salt in snow after polar sunrise.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: elin.maria.magnusson@gmail.com. Supervisor: Gudrun Brattström.

Acknowledgements

I would like to thank my supervisor Leonard Barrie for the opportunity to write this article and for a great collaboration, a big thank you to Sangeeta Sharma as well. I would also like to express my gratitude to my supervisor Gudrun Brattström for valuable feedback and help when ever in need.

Contents

Chapter 1

Introduction

In this analysis we focus on measured concentrations of aerosols and the two gases Ozone and Mercury, observed in the Arctic part of Canada. Aerosols are very small solid or liquid particles dispersed in air. The science of airborne pollution is a relatively new research field and it was not until the mid seventies that observations of aerosols were first made for a complete year in the Alaskan Arctic. The information gained from monitoring the aerosols was of interest and since 1980 aerosols have been measured at the high Arctic station Alert, Canada. The special weather conditions are well suited to monitor aerosols. Since Alert is above the polar circle the sun never rises during the winter months and this makes for very stable weather conditions. Thus, during the dark part of the year the residence times of aerosols are very long.

This thesis is a part of an article which will be published in collaboration with Leonard Barrie and Sangeeta Sharma. In the thesis we will look closer at the statistical part of the analysis while the article will be more concentrated on the interpretation and the chemical analysis made. The analysis is a continuation of L.A. Barrie's earlier research and primarily the article "*Arctic lower tropospheric aerosol trends and composition at Alert, Canada: 1980-1995*" (Sirois and Barrie, 1999). I strongly recommend to give this article a quick read before continuing with this thesis, since it is heavily referenced.

As in the previous article, the statistical analysis in this thesis is an analysis in two parts, a time series analysis followed by Positive Matrix Factorization (PMF), a non-negative dimension reduction technique. There is high interest in how concentrations of aerosol constituents have changed since 1980 and looking at the seasonality of the constituents, which the time series analysis provides.

The second part of the analysis is done in order to lower the dimension of data and looking at sources of the constituent concentrations. In this context, a constituent is either a chemical element or compound which is factored into sources such as Sea-salt, Soil and Smelter. Although the general goal of the two different analysis techniques are the same as in the report from 1999, there are some changes in how we conduct them. In our analysis we use smoothing splines to explain the long term trend and seasonality, in the last report this was done with polynomial, sine and/or cosine curves. For the PMF analysis we have divided the data into two data sets with one factorization made during the light period and one during the dark.

1.1 Data

Weekly samples of atmospheric aerosols have been collected at the high Arctic station Alert, Canada, from 1980 until present day. A filter is placed each week on a plateau 210 m above sea level and 6 km from the main base. The filters are removed and then cut into eight pieces and analysed in different ways. Concentrations of aerosols are analysed and then divided by the amount of air that has gone through the filter while in place, resulting in measuring the weekly concentrations in units of nanograms per cubic meter. Thus the concentration is a mean over the specific week. In order to measure constituents the filter has to be in place for a week, otherwise too many chemical constituents will not be detectable above the filter background especially in the cleaner summer season. Three different methods have been used to analyse the constituent loadings in the filter; Liquid Ion Chromatography(IC), Instrumental Neuron Activation Analysis (INAA) and Inductively Coupled Plasma emission or mass spectroscopy (ICP).

Every fourth week a blank filter is handled exactly as the rest of the filters but after it has been put in place it is immediately removed. The field blanks together with the analytical detection limit are used in order to determine the Operational Detection Limit (ODL) for the mass concentrations in air. Thus, the filed blanks include a measurement of contributions to blanks from handling and analysis as well as that from the original filter content.

The lifetime of aerosols varies greatly throughout the year which clearly can be seen in the filter loadings. The aerosol residence time in the lower atmosphere of the Arctic lies about 3 to 7 weeks during October to May, while during the warmer half of the year it is somewhere around 3 to 7 days. This together with seasonality in the north to south transport means there is strong seasonality in Arctic aerosol components.

In the previous report (Sirois and Barrie, 1999) 18 constituents were analysed. For our analysis five more constituents were available: the gases Ozone (O_3) and Gaseous Elemental Mercury (GEM) along with the aerosols Equivalent Black Carbon (EBC), Nickel (Ni) and Iron (Fe). In Table 1.1 all constituents are listed with percentage of missing data, percentage of Below Detection Limit (BDL), time range of observation as well as maximum and minimum. The IC analysis is made throughout the time period while INAA stopped in January 2007 and ICP analysis in June 2006.

Contrary to the weekly filters the concentrations for O_3, EBC and GEM are instrumentally collected. These were not reported as weekly mean concentrations but rather hourly means for O_3 and EBC and six hourly means for GEM. The black carbon is measured by drawing air through a instrument with a filter attached which in turn is illuminated. The amount of light that goes trough the filter is then calculated as a function of time. This instrument also uses a blank filter as reference point. The GEM is collected in three basic steps i) pre-concentration of GEM onto a trap; ii) removal of the Hg from the trap by thermal desorption and iii) detection and quantification of the Hg. The details of the method can be found in Steffen et al. (2008) among others. In order to analyse the data set as a whole, the mean of these concentrations have been taken to match the time points for the weekly measurements of the aerosols. The information of the data set in Table 1.1 shows the averaged data, more information of the original data sets can be seen in section 1.1.2.

Constituent	Observation Period		% NA	Min	Max	Operational Detection Limit ng m^{-3}		% BDL	Analytic method
	First	Last				Min	Max		
SO_4^{-2} (Sulfate)	July 1980	Oct. 2013	0	10.68	3495.27	1.42	27.26	0.36	IC
H^+	July 1980	Oct. 2013	3.38	0.02	48.39	0.01	0.6	39.79	IC
Br^-	July 1980	Oct. 2013	1.19	0.01	88.44	0.11	1.24	11.88	INAA (IC[2])
NH_4^+ (Ammonium)	July 1980	Oct. 2013	0	4.57	465.81	6.51	68.56	9.86	IC
NO_3^- (Nitrate)	July 1980	Oct. 2013	0.18	4.73	294.86	1.99	17.37	0.59	IC
Na^+	July 1980	Oct. 2013	0.06	1.01	1807.9	1.85	9.06	4.93	IC
Cl^-	July 1980	Oct. 2013	0.06	1.23	3406.21	2.83	20.46	14.25	IC
K^+	July 1980	Oct. 2013	1.8	0.08	137.92	0.2	8.22	9.5	IC
Pb	July 1980	May 2006	0.77	0.01	9.34	0.01	1.23	18.17	ICP
V	July 1980	Dec. 2006	1.96	0.01	3.01	0.01	0.15	2.49	INAA
Mg	July 1980	May 1997	0.6	0.6	662.51	0.9	7.38	1.25	ICP
Zn	July 1980	May 2006	1.85	0.05	59.53	0.07	4.7	7.72	ICP
Cu	July 1980	Dec. 2006	1.2	0.02	71.86	0.03	0.67	8.73	ICP (INAA[2])
Ca^{2+}	July 1980	Dec. 2006	0.6	0.44	2603.53	0.65	44.58	1.43	ICP (INAA[2])
Mn	July 1980	Dec. 2006	0.3	0.01	35.88	0.01	0.31	1.96	INAA (ICP[2])
I	July 1980	Dec. 2006	7.15	0.01	4.46	0.02	0.29	6.59	INAA
Al	July 1980	Dec. 2006	0.53	0.38	2219.22	0.57	39.85	1.43	INAA
MSA (Methanesulfonic acid)	July 1980	Oct. 2013	3.92	0.01	55.43	0.05	2.49	2.97	IC
Ni	July 1980	May 2006	1.69	0.01	0.55	0.01	0.55	35.13	ICP
Fe	July 1980	May 2006	0.77	1.3	1106.02	0.64	14.08		ICP
O_3[1]	Dec. 1991	Dec. 2006	6.83	1.27	48				Instrumental
EBC[1]	May 1989	Dec. 2012	5.95	3.75	343.52	5	5		Instrumental
GEM[1]	Jan 1995	Oct. 2013	12.8	0.19	2.45				Instrumental

[1] Modified data, see section 1.1.2
[2] Method used to impute missing values, see section 1.1.3

Table 1.1

1.1.1 Missing data and BDL

Almost all constituents have missing values during the observed period of time but for most constituents the total percentage is very low, as seen in Table 1.1. Some of the constituents have been analysed by more than one method and for these concentrations there has been a possibility to use data from an alternative method in order to eliminate missing values. As the missing values of the different techniques often are in different weeks this is achievable. The constituents for which this was possible were Ca, Cu, Br and Mn. For the first three there was a small but significant difference between the INAA and the ICP analysis, with the consultation of Leonard Barrie the choice was made to fill gaps in ICP Cu and Ca with values from INAA Cu and Ca and the other way around for Mn. The IC analysis improved greatly after 1986, therefore the concentrations from the INAA Br analysis was used overall and was imputed with values from IC Br.

For some constituents, the below detection limit varies depending on improvement of chemical analysis technique, clean filter composition and/or handling. The observations are calculated from nanograms of constituent found in the filter divided by the air volume that has gone through the filter between time of placement to extraction of filter. If only a small amount of air has gone through the filter and the amount of a constituent measured is small, the concentration

cannot be measured.

The BDL values have been treated as in Barrie and Sirois, 1999, by taking two thirds of the BDL for that period of time, considering the comparability to the last report. The decision of including BDL's is motivated by the fact that they represent information; namely that the concentration in this week was lower than the BDL. In Section 2.6.2 the theory of this is further discussed. As mentioned, because of the generally higher concentrations of aerosols during the winter season the number of BDL decrease immensely during the colder half of the year.

1.1.2 O_3, GEM, EBC

In Table 1.2 information about the unedited data sets can be seen. The concentrations for O3, GEM and EBC measured at time resolutions of less than a week have been averaged for a week to match the time resolution of the weekly filter data. In order to do the PMF analysis this is a necessity. For the time series analysis the use of the reduced data sets is motivated by comparison to other constituents and the inconvenience in using such a large number of observations.

Constituent	First observation	No. of observations	No. of NA	Min	Max
EBC	29 May 1989	206 740	31160	0.01	1794.27
O_3	31 Dec. 1991	192 869	16 656	1	57
GEM	1 Jan. 1995	29 231	4048	0.002 (0)[1]	4.56

[1] Smallest value is 0 but has been removed since values must be positive

Table 1.2

As can be see in Figure 1.1 there are values measured of O_3, GEM and EBC during the time the filter that measure the aerosols is switched. Usually, there is a short period of time between the weekly filter is being removed and the next one put in. There are 26 occurrences where the gap between one filter has been taken out and before another one has been placed that was more than 24 hours. Disregarding these 26 occurrences the mean of the time between one filter being removed and the next put in is 11.75 minutes. Of these 26 occurrences 10 are from after the time O_3, EBC and GEM observations began. The mean of the three concentrations was taken from end point to end point. The loss of accuracy should not be affected if some observation lies in between the time the filters are being switched, except for these 10 time points.

Using the mean of data between one filter being removed until the next is removed is made for gain of data points for each mean and speed in computation since we do not have to look through the large number of observations of O_3, GEM and EBC to find and remove observations when there is no filter in place. The data for O_3, EBC and GEM has been removed that lies in gaps larger than 24 hours between filters. So each observation used in the analysis is a mean from one filter is removed until the next is removed except for 10 times where the mean is only during that filter is in place. An illustration of this can be seen in Figure 1.1

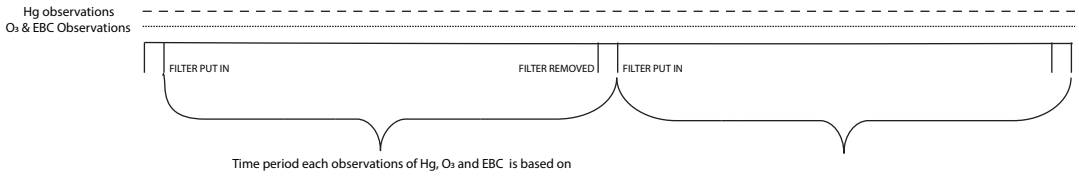


Figure 1.1: Illustration of observations

If all observations during a particular week are missing, the corresponding mean is treated as a missing observation. In some weeks there are very few values which makes the averaged observation unreliable or unrealistic. The decision was made to treat all weekly means as missing if more than 51 observations were missing for EBC and O_3, and 9 for GEM. The numbers 51 and 9 are around 30 % of the mean number of observations in each week. In Table 1.3 information about the new reduced data vectors are shown. A noticeable thing is the much reduced maximum value for EBC which is due to the high fluctuation from hour to hour. The maximum seen in Table 1.2 is followed by an hour of a below detection limit value. As a comparison O_3 is much more consistent from hour to hour so only a small drop can be seen in the overall maximum.

Constituent	No.of observations	No. of missing obs. (all missing)	No. of missing obs. (30% or more)	Min	Max
EBC	1211	63	77	3.75	343.52
O_3	1127	12	72	1.27	48
GEM	972	42	124	0.19	2.45

Table 1.3

1.1.3 Missing values for Br, Mn, Cu, Ca

As mentioned in Section 1.1.1, there are some components that have been analysed in several ways. We use either the most trusted way or the vector of concentrations with least missing values as the main vector, while the alternative method is used to fill in missing observations. For Br and Mn the INAA was used as the main analysis technique and IC to fill in gaps for Br while ICP was used to fill in gaps for Mn. For Cu and Ca the ICP analysis was used as the prime vector and the INAA analysis observations were used to fill in missing values. All tests in this section are used on the natural logarithm of data since the data vectors are approximately lognormally distributed for all constituents in this section.

Br

For Br the IC analysis was highly improved after 1986 as can be seen in Figure 1.2, therefore we will look separately at the data before 1987 and one from 1987 onward. In Figure 1.2 and 1.3 we can see that the observations seem better correlated after 1986. Although IC Br is measured the longest and has the least missing values, the INAA Br is considered a more precise technique.

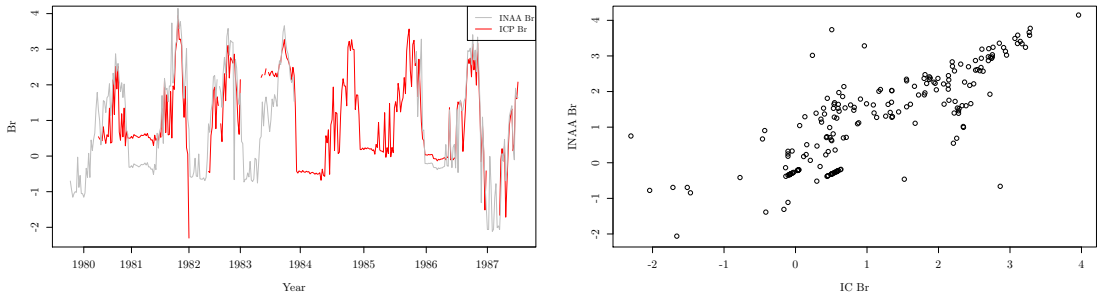


Figure 1.2: Br 1980-1987

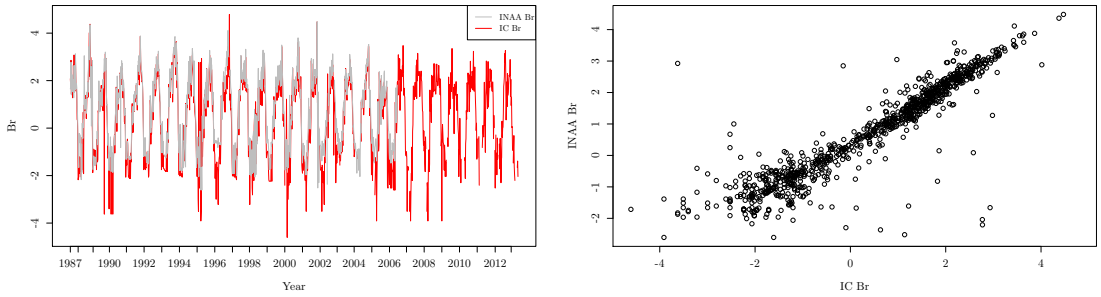


Figure 1.3: Br 1987 and forward

The two different correlation tests between IC Br and INAA Br, one for 1980-1986 and the other one for 1987-2013, can be seen in Table 1.4 and 1.6. Much of the badly correlated observations, especially in the data 1980-1986, can be attributed to the different techniques having different BDL, which clearly shows in Figure 1.2 and 1.3. The detection limit was decreased drastically for IC but also somewhat for INAA in 1986. It would have been preferable to not use IC Br before 1987 but a large part of the missing data for INAA Br lies between 1984-1986 which is seen in 1.4, so even if the imputed values before 1987 are less accurate they are still a good estimate.

	t	cor	df	p.value
Pearson's test	18.97	0.80	201	0.00

Table 1.4: Pearson's χ^2 test for correlation between IC and INAA Br (1980-1986)

	conf.low	conf.high
Confidence intervall	-0.01	0.42

Table 1.5: 95% confidence interval diff. of mean between IC and INAA Br (1980-1986)

While the correlation test is a good measurement of how similar the two time series are it does not take into account that they are on different levels. A paired T-test was performed to look at the difference in mean of the two data vectors in order to compare the levels between the two different analysis techniques. In Table 1.5 and 1.7 the 95% confidence interval of the difference in mean between the techniques from a paired T-test are shown. From this we can read that even though the data between 1980-1986 has worse correlation it seems to be better leveled than the data between 1987-2007. The IC Br after 1986 have a lower ODL than INAA Br but the difference between the means are still good enough not to make adjustments in the level of data.

	t	cor	df	p.value
Pearson's test	71.71	0.92	934	0.00

Table 1.6: Pearson's χ^2 test for correlation between IC and INAA Br (1987-2007)

	conf.low	conf.high
Confidence intervall	-0.38	-0.11

Table 1.7: 95% confidence interval diff. of mean between IC and INAA Br (1987-2007)

Mn

For Mn as for Br the main technique used was INAA while ICP was the analysis method for patching the missing values. In Figure 1.4 and Table 1.8 the correlation between the two analyses are very high. From Table 1.9 there is no significant difference in mean either, so the two different techniques seem to be comparable.

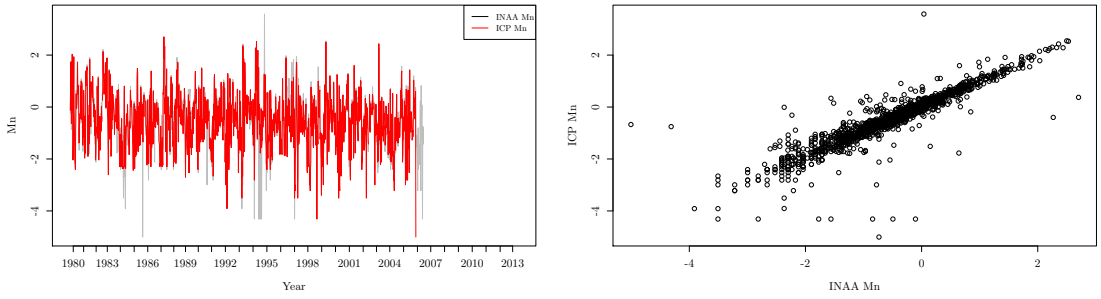


Figure 1.4: Mn

	t	cor	df	p.value
Pearson's test	75.27	0.90	1273	0.00

Table 1.8: Pearson's χ^2 test for correlation between ICP and INAA Mn

	conf.low	conf.high
Confidence intervall	-0.14	0.02

Table 1.9: 95% confidence interval diff. of mean between ICP and INAA Mn

Cu

For Cu the observed vector with least missing observations was ICP Cu so that was chosen as the main technique and INAA Cu as the spare values. For ICP Cu there are only 10 missing observations during the observed period, but while ICP Cu is stopped in June 2006 INAA Cu continues until December 2006 so most imputed values occur in the end of 2006. From the T-test in Table 1.11 the level of data is somewhat different. This is most likely a result of the different ODL limits but the confidence interval is very close to zero.

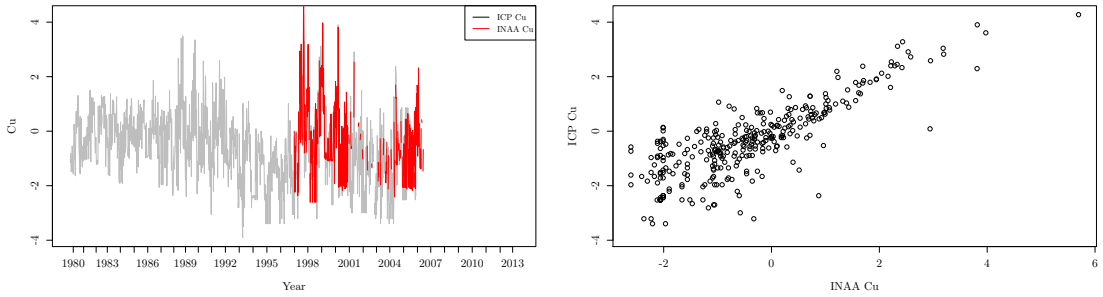


Figure 1.5: Cu

	t	cor	df	p.value
Pearson's test	25.41	0.82	323	0.00

Table 1.10: Pearson's χ^2 test for correlation between ICP and INAA Cu

	conf.low	conf.high
Confidence intervall	0.01	0.32

Table 1.11: 95% confidence interval diff. of mean between ICP and INAA Cu

Ca

The same choice as for Cu was made for Ca, to use ICP as the primary method, since those observations are less patchy than INAA. Although ICP has less missing observations, here the INAA Ca observations continue further than ICP Ca. What we can see from Figure 1.6 is that when INAA was first used the BDL was far better for ICP, while at the time ICP was stopped the BDL seems to be at the same level for both techniques. From the scatter plot these BDL values is the cluster shown to the left.

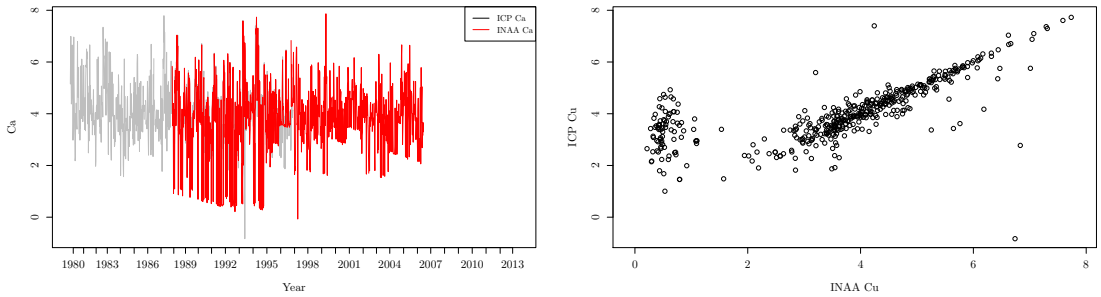


Figure 1.6: Ca

Since the cluster of BDL would have a large influence, for the imputation these were removed from INAA Ca and treated as NAs. There are only 5 missing values during the time ICP Ca was observed, so almost all values used from INAA Ca lie after the time the detection limit was improved. For Ca the mean of the data is a bit higher for ICP Ca than for INAA Ca when the lowest BDL values have been removed from the analysis.

	t	cor	df	p.value
Pearson's test	36.90	0.89	362	0.00

Table 1.12: Pearson's χ^2 test for correlation between ICP and INAA Ca

	conf.low	conf.high
Confidence intervall	-0.22	-0.02

Table 1.13: 95% confidence interval diff. of mean between ICP and INAA Ca

Chapter 2

Method & Theory

In this report we will mainly look at two different statistical analysis methods: first a time series analysis is performed on the constituents and then Positive Matrix Factorization is used to look at source components in the aerosol. The analysis and data handling is done using R except for the PMF analysis which is done using the licensed program PMF2.

2.1 Cubic smoothing spline

In contrast to most other splines, cubic smoothing splines appear as an optimization problem. For example, the commonly used regression splines use placed knots where piecewise polynomials join smoothly. For smoothing splines we want to find a function $f(x)$ so that the penalized sum of squares is minimized (Hastie and Tibshirani, 1990)

$$PRSS = \sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda \int (f''(t))^2 dt, \tag{2.1.0.1}$$

where the only criterion of f is that it is twice differentiable with continuous derivatives and where λ is a fixed constant. It is evident that the first term of (2.1.0.1) is the least square fit of data while the second term is a roughness penalty term which trades off smoothness to fidelity to data via λ. A unique minimizer to this problem can be shown to be a natural cubic spline with $n-2$ internal knots (Green and Silverman, 1994).

For timepoints $a < t_1, ...t_n < b$ we let $h_i = t_{i+1} - t_i$ and define the $n \times (n-2)$ matrix $\mathbf{N}_{ij}$ for $i = 1, ..., n$ and $j = 2, ..., n-1$ with entries

$$n_{j-1,j} = h_{j-1}^{-1}, \quad n_{jj} = -h_{j-1}^{-1} - h_j^{-1}, \quad n_{j+1,j} = h_j^{-1} \quad \text{and} \quad n_{ij} = 0 \quad \text{for} \quad |i-j| \geq 2$$

and the $(n-2) \times (n-2)$ matrix $\mathbf{\Omega}$ with entries

$$\omega_{i,i} = \frac{1}{3}(h_{i-1} + h_i) \quad \text{for} \quad i = 2, ..., n-1,$$

$$\omega_{i,i+1} = \omega_{i+1,i} = \frac{1}{6}h_i \quad \text{for} \quad i = 2, ..., n-2$$

and

$$\omega_{ij} = 0 \quad \text{for} \quad |i-j| \geq 2.$$

By standard numerical linear algebra $\mathbf{\Omega}$ is strictly positive definite (Green and Silverman, 1994) and we can define the matrix $\mathbf{K} = \mathbf{N}\mathbf{\Omega}^{-1}\mathbf{N}^T$.

It can be shown that f and γ specify a natural cubic spline if and only if (Green and Silverman, 1994)

$$\mathbf{N}^T f = \mathbf{\Omega}\gamma. \tag{2.1.0.2}$$

If (2.1.0.2) is satisfied then the roughness penalty term can be written as

$$\int_a^b f''(t)^2 dt = \gamma^T \mathbf{\Omega}\gamma = f^T \mathbf{K} f \tag{2.1.0.3}$$

and PRSS sum (2.1.0.1) can be rewritten as

$$PRSS = (y-f)^T(y-f) + \lambda f^T \mathbf{K} f = f^T(I + \lambda\mathbf{K})f - 2y^T f + y^T y. \tag{2.1.0.4}$$

Since $\lambda\mathbf{K}$ is non-negative definite $(\mathbf{I} + \lambda\mathbf{K})$ is strictly positive definite and it follows that for a fixed λ and by setting $f = (\mathbf{I} + \lambda\mathbf{K})^{-1}y$ this uniquely minimizes (2.1.0.4).

The parameter λ is controlling the smoothness of the spline, another way of controlling the smoothness is by setting the Equivalent Degrees of Freedom (*edf*) which is defined as the trace of the hat (or smoother) matrix $\mathbf{S}_\lambda = (\mathbf{I} + \lambda\mathbf{K})^{-1}$.

2.2 ARMA(p, q)-process

The autoregressive moving-average (ARMA) process $\{Y_t\}$ can be expressed as

$$Y_t - \phi_1 Y_{t-1} - ... - \phi_p Y_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q}, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

for Z_t white noise distributed as $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ if the Z_t's are uncorrelated and $\mu_Z(t) = 0,\ E[Z_t^2] = \sigma^2$.

The fit of the ARMA-process is calculated by maximizing the likelihood for a Gaussian time series $\{Y_t\}$. With the help of the innovations algorithm we can recursively find one step ahead predictions $\hat{Y}_{n+1}$

$$\hat{Y}(t)_{n+1} = \begin{cases} \sum_{j=1}^{n} \theta_{nj}(Y(t)_{n+1-j} - \hat{Y}(t)_{n+1-j}) & \text{for} \quad 1 \leq n < max(p,q) \\ \phi_1 Y_n + ... + \phi_p Y_{n+1-p} + \sum_{j=1}^{p} \theta_{nj}(Y(t)_{n+1-j} - \hat{Y}(t)_{n+1-j}) & \text{for} \quad n \geq max(p,q) \end{cases}$$

with B as the backward shift operator $B^j Y_t = Y_{t-j}$ we define $W(t)$ as

$$W(t) = \begin{cases} \sigma^{-1} Y_t & \text{for} \quad t = 1, ..., max(p,q) \\ \\ \sigma^{-1} \phi(B) Y_t & \text{for} \quad t > max(p,q) \end{cases}$$

and $E(Y_{n+1} - \hat{Y}_{n+1})^2 = \sigma^2 E(W_{n+1} - \hat{W}_{n+1})^2 = \sigma^2 r_n$. Where θ_{nj} and r_n is determined by the innovations algorithm for $n \geq m$, $m = max(p,q)$.

With this the Gaussian Likelihood can be expressed as

.

$$L(\phi, \theta, \sigma^2) = \frac{1}{\sqrt{(2\pi)^n r_0, ..., r_{n-1}}} \exp(-\frac{1}{2\sigma^2}) \sum_{j=1}^{n} \frac{(Y_j - \hat{Y}_j)^2}{r_{j-1}},$$

with maximum likelihood estimators

$$\hat{\sigma}^2 = \frac{1}{n} S(\hat{\phi}, \hat{\theta})$$

$$S(\hat{\phi}, \hat{\theta}) = \sum_{j=1}^{n} \frac{(Y_j - \hat{Y}_j)^2}{r_{j-1}}$$

And for values of $\hat{\phi}, \hat{\theta}$ that minimize

$$l(\hat{\phi}, \hat{\theta}) = ln(\frac{1}{n} S(\hat{\phi}, \hat{\theta})) + \frac{1}{n} \sum_{j=1}^{n} ln(r_{j-1})$$

The maximum likelihood estimation must be solved numerically.

2.3 GARCH(m, n)-process

While the ARMA-process is used to calculate the conditional mean the GARCH(m, n)-process (Generalized Autoregressive Conditional Heteroskedasticity) is used to calculate a conditional variance that is allowed to change over time, in comparison with classical time series analysis where a constant variance is assumed.

Let ε_t denote a real-valued discrete-time stochastic process, and ψ_t the information set (σ-field) of all information through time t. The GARCH(m, n)-process is then defined as below (Bollerslev, 1986).

$$\varepsilon | \psi_{t-1} \sim N(0, h_t) \tag{2.3.0.1}$$

$$h_t = \omega + \sum_{i=1}^{n} \alpha_i \varepsilon_{t-1}^2 + \sum_{i=1}^{m} \beta_i h_{t-1} \tag{2.3.0.2}$$

For $m \geq 0$, $n > 0$, $\omega > 0$, $\alpha > 0$ and $\beta \geq 0$. The GARCH- and ARMA-process are compatible with each other (Brockwell and Davis, 2002) so a time series can be described by an ARMA-GARCH model.

2.4 Definitions

- AICc

For determining the order of the ARMA-process the AICc criterion is used as recommended by Brockwell and Davis (2002), which is defined as

$$AICc = L(\phi_p, \theta, \frac{1}{n} S(\hat{\phi}, \hat{\theta})) + \frac{2(p+q+1)n}{n-p-q-2}$$

- Autocovariance

$$\gamma_Y(t,h) = \mathrm{Cov}(Y_{t+h}, Y_t)$$

- Weak stationarity of order 2

A time series $\{Y_t\}$ is weakly stationary if

- $\mu_Y(t)$ is independent of t.
- $\gamma_Y(t,h)$ is independent of t for every h, so we write $\gamma_Y(h)$ instead.

- Autocorrelation

$$\rho_Y(h) = \frac{\gamma_Y(h)}{\gamma_Y(0)}$$

- IID noise

$\{Y_t\}$ is called iid noise, $\{Y_t\} \sim \mathrm{IID}(0, \sigma^2)$, if the Y_t's are independent and $\mu_Y(t) = 0,\ E[Y_t^2] = \sigma^2$.

- Ljung-Box test

The test statistic Q_{LB} follows a χ_h^2 distribution under the null hypothesis H_0: Residuals are independent normally distributed noise

$$Q_{LB} = n(n+2) \sum_{j=1}^{h} \hat{\rho}(j)^2 / (n-j)$$

2.5 Change point model

A change point model can be used when an ordered data vector has different distributions within the vector and thereby divide the data at the change point or points. The change points can be optimized in several ways. In this thesis we look at optimizing by the mean. The change point τ can be estimated by the least square estimation of $\hat{\tau}$, with the length of the series Y_t as T with μ_1 and μ_2 the respective mean before and after the change point, defined as

$$\hat{\tau} = argmin_\tau \left(min\{\sum_{t=1}^{\tau}(Y_t - \mu_1)^2 + \sum_{t=\tau+1}^{T}(Y_t - \mu_2)^2\} \right) \tag{2.5.0.1}$$

We can write the sum of squares of residuals as

$$S_\tau^2 = \sum_{t=1}^{\tau}(Y_t - \hat{\mu_1})^2 + \sum_{t=\tau+1}^{T}(Y_t - \hat{\mu_2})^2 \tag{2.5.0.2}$$

Where $\hat{\mu_1}$ and $\hat{\mu_2}$ are the least square mean estimators of the mean for the first 1-τ and τ+1-T respectively, with this it follows with μ being the overall mean that

$$\sum_{t=1}^{T}(Y_t - \mu)^2 = S_\tau^2 + V_\tau^2 \tag{2.5.0.3}$$

For V_τ as

$$V_\tau = (\frac{\tau(T-\tau)}{T})^{1/2}(\mu_1 - \mu_2) \tag{2.5.0.4}$$

we see that

$$\hat{\tau} = argmin_\tau(S_\tau^2) = argmax_\tau(V_\tau^2) = argmax_\tau|V_\tau| \quad (Bai,\ 1994)$$

2.6 Positive matrix factorization

2.6.1 Introduction to receptor modelling

The multivariate factor analysis technique Positive Matrix Factorization (PMF) is used to lower the dimension of data. In environmental data the method is used to find the number and composition of sources. In this context a source can be the two measured factors Na and Cl make up the interpretable source *salt*. In air pollution research this is commonly referred to as receptor modelling, based on air monitoring data and determining the sources of air pollution, i.e. a large data set of variables is collected , e.g. chemical elements or compounds, that together can be factorized into sources such as sea-salt, smelter, soil etc . One of the key features in aerosol constituents is the lack of negative data since it is measured in concentrations. The natural physical constraints was the motivation for the development of PMF; all other then-current methods had their limitations when used on physical data. Any factor analysis receptor model has to adhere to some basic physical constraints (Henry, 1987; Hopke, 2003);

- The model has to reproduce the original data.
- All sources have a positive percentage of all elements, so the predicted source compositions must be non-negative.
- A source cannot give off negative mass so the predicted source contributions to the aerosol must all be non-negative;
- The sum of the predicted elemental mass contributions for each source must be less than or equal to total measured mass for each element; the whole is greater than or equal to the sum of its parts.

The basis of Positive Matrix Factorization is the general factor model as is Principal Component Analysis (PCA) and other dimension reduction techniques: for the observed $n \times m$-matrix $\mathbf{X}$ with n observations and m variables we have

$$\boldsymbol{X} = \boldsymbol{GF} + \boldsymbol{E} \tag{2.6.1.1}$$

for $p \leq m$, $\mathbf{G}$ is the factor scores matrix with dimension $n \times p$, $\mathbf{F}$ a $p \times m$ loadings matrix and $\mathbf{E}$ the error matrix with same dimension as $\mathbf{X}$. In a receptor modelling setting p is the number of sources.

The commonly used dimension reduction technique in environmental sciences and otherwise is PCA, which traditionally is based on the eigenvectors of the covariance matrix but later also the Singular Value Decomposition (SVD). A problem with using the covariance matrix is that $\mathbf{X}$ has to be centered, this results in a loss of information about the initial scale of variables. For the SVD the matrix $\mathbf{X}$ is expressed as

$$\boldsymbol{X} = \boldsymbol{USV^T} + \boldsymbol{E} \tag{2.6.1.2}$$

Where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices, while $\mathbf{S}$ is a diagonal matrix. It can be shown that (2.6.1.2) can be written on the form of (2.6.1.1) (Comero et al., 2009). Just as in the case of PCA there is a least square property so the solution truncated to p components of (2.6.1.2) minimizes the Frobenius norm of $\mathbf{E}$. For any prechosen factorization with rank p of $\mathbf{X}$ can be defined as

$$\{\mathbf{G},\mathbf{F}\} = \underset{G,F}{argmin}||\mathbf{X}-\mathbf{GF}||_F = \underset{G,F}{argmin}\sum_{ij}(x_{ij}-\hat{x}_{ij})^2 \tag{2.6.1.3}$$

where $\mathbf{G}$ and $\mathbf{F}$ are required to be of previously selected rank p. The solution to (2.6.1.3) has minimum variance if and only if the precision of all x_{ij} is the same (Paatero and Tapper, 1993). Although (2.6.1.3) can be solved there is not a unique solution, for any non-singular square matrix $\mathbf{T}$ there is a rotation of the solution that doesn't effect the residual matrix $\mathbf{E}$

$$\boldsymbol{X} = \boldsymbol{GF} + \boldsymbol{E} = \boldsymbol{GTT^{-1}F} + \boldsymbol{E} = \boldsymbol{\bar{G}\bar{F}} + \boldsymbol{E} \tag{2.6.1.4}$$

The singular value decomposition is not invariant to scaling, which means that different decompositions arise if different measurements are used from one column (or row) to another. To correct this problem different scaling techniques for $\mathbf{X}$ have been used for PCA, the different scalings were studied in relation to a weighted least square (Paatero and Tapper, 1993, 1994).

$$\{\mathbf{G},\mathbf{F}\} = \underset{G,F}{argmin}\sum_{ij} w_{ij}(x_{ij}-\hat{x}_{ij})^2 \tag{2.6.1.5}$$

What Paatero and Tapper (1993) showed was that optimization, i.e the smallest variance of (2.6.1.5) is obtained if each data point is scaled individually based on their uncertainties $\sigma(x_{ij})$. If the rank of the standard deviation matrix is one and the SVD of $\mathbf{X}$ is rotatable so that all matrices are positive, only then the solution by SVD is optimal. If not, the effects of the point by point scaling is shown to be a scaled matrix that cannot be recreated by any factor analysis based on SVD. As mentioned one of the basic constraints by Henry (1987) for a good receptor model was that the model has to reconstruct the data. These were the main motivations to a new factor analysis model.

2.6.2 PMF

What Paatero and Tapper (1994) proposed in order to address the problems with then-current factor models was to minimize (2.6.2.1), with respect to the constraints that both $\mathbf{G}$ and $\mathbf{F}$ should be non-negative.

$$Q(E) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{e_{ij}^2}{\mathrm{s}_{ij}^2} \tag{2.6.2.1}$$

where e_{ij} is defined as

$$e_{ij} = x_{ij} - \hat{x}_{ij} = x_{ij} - \sum_{k=1}^{p} f_{ik}g_{kj} \tag{2.6.2.2}$$

For an estimated or known standard deviation matrix $\mathbf{S}$ with entries s_{ij}. As can be seen in (2.6.2.1) the problem with non-optimal scaling is addressed. In the initial paper on PMF the minimization of $Q(E)$ was solved via an iterative altering least squares algorithm which kept one of the matrices $\mathbf{G}$ or $\mathbf{F}$ constant in each iterative step while the other one was optimized to minimize Q, until convergence occurred. This way of minimizing $Q(E)$ can be very slow if the factors are far from being orthogonal. In order to speed up the optimization a faster algorithm where both matrices are optimized in the same iterative step was created.

An optimization scheme that is able to vary $\mathbf{G}$ and $\mathbf{F}$ in each iterative step was described by Paatero (1997) and implemented in the program PMF2. The function (2.6.2.1) took on a more complicated form (2.6.2.3). In this new enhanced object function the non-negativity constraint is implemented as penalty functions. Beginning with positive pseudorandom numbers as initial matrices, we can see that when g_{ik} and f_{kj} goes to zero the penalty functions will be large so the algorithm will continue to search for a minimum.

$$\begin{aligned}\bar{Q}(E,G,F) =& Q(E) + P(G) + P(F) + R(G) + R(F) = \\ & \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{e_{ij}^2}{\mathrm{s}_{ij}^2} - \alpha\sum_{i=1}^{m}\sum_{k=1}^{p}\log g_{ik} - \beta\sum_{k=1}^{p}\sum_{j=1}^{n}\log f_{kj} \\ & + \gamma\sum_{i=1}^{m}\sum_{k=1}^{p} g_{ik}^2 + \delta\sum_{k=1}^{p}\sum_{j=1}^{n} f_{kj}^2 \end{aligned} \qquad (2.6.2.3)$$

While $P(G)$ and $P(F)$ prevent $\mathbf{G}$ and $\mathbf{F}$ from becoming negative, $R(G)$ and $R(F)$ are called *regularization functions* which remove the rotational indeterminacy and control the scaling of the left and right factors. The R functions are always needed in the algorithm to prevent singularities in the model from rotation and scale changes. The coefficients α, β, γ and δ each control the strength of their function and are given smaller and smaller values so that the final values are negligible but larger than zero. To solve each iterative step the algorithm uses first Gauss-Newton or Newton-Raphson and between each step a rotational substep is performed. In each substep a rotation matrix $\mathbf{T}$ and its inverse $\mathbf{T}^{-1}$ is determined so that $\bar{Q}$ is minimized with respect to the new factor matrices $\bar{\mathbf{G}}$ and $\bar{\mathbf{F}}$ ($\mathbf{GT}$ and $\mathbf{T^{-1}F}$). The main part of $\bar{Q}$ is not changed by the rotations so the substeps only minimize the sum of P and R in (2.6.2.3). The algorithm would work without the substeps but a factor of circa 2 of gain in speed can be attained.

As we could see with PCA, PMF also has rotational ambiguity as (2.6.1.4) (Paatero and Tapper, 1994), this leads to a non-unique solution of the algorithm. An illustration of the ambiguity in Figure 2.1 is shown with simulated data of iron and silicon, where each pair of lines corresponds to a different source profile. Since the two factors are not correlated to each other two factors are needed to reproduce the source profiles. The true profiles lie somewhere in between the x and y axes and the pair of solid lines containing all points. Although there is a rotational ambiguity in PMF as well as in any other general factor analysis model the non-negativity constraint reduces the number of rotations and in some cases even produces a unique solution to (2.6.2.1) if all rotations produce some negative elements in matrices $\bar{\mathbf{G}}$ and $\bar{\mathbf{F}}$.

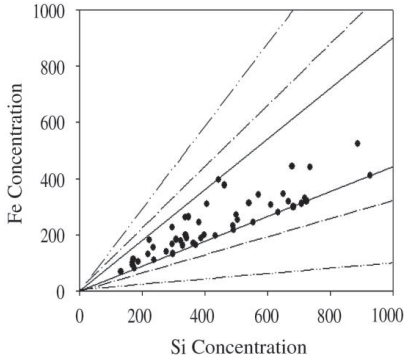


Figure 2.1: Simulated data showing multiple possible source profiles that could be used to fit the data. (Paatero et al., 2002)

Often the true standard deviation matrix $\mathbf{S}$ is unknown. A study (Kim and Hopke, 2007) shows that PMF analysis using estimated and sample-species specific standard deviations give very similar solutions. Typically in receptor modelling there are three different kinds of data points; known, BDL and missing values. Missing data have historically been treated in three different ways, either remove any row in $\mathbf{X}$ if it contains a missing value or remove the complete column. This approach is typically made if a large percentage of a row or column is missing, this is a very crude way of treating data unless it is a very large part that is missing. The third approach is to impute values, usually the arithmetic or geometric mean of the columns of $\mathbf{X}$, and for these values raise the standard deviation by a constant so the imputed values get less weight in the minimization of Q. One study from Huang et al. (1999) where either samples with one or more missing value were removed vs. an imputation of the mean showed that using imputed values gave more realistic factor solutions and less frequent occurrence of physically unrealistic factors.

The BDL values also have three similar approaches. The BDL data actually give information about the value being very small so removing rows or columns with BDL values results in a big loss of information. The most common approach is to take a small constant such as $\frac{1}{2}$ times the detection limit and use that value, this seem to originate from Polissar et al. (1998) who suggested the following use of data in PMF.

- $x_{ij} = v_{ij}$ For known values
- $x_{ij} = \frac{d_{ij}}{2}$ For BDL values
- $x_{ij} = \tilde{v}_j$ For missing values

And their respective standard deviations as

- $\sigma_{ij} = u_{ij} + \frac{d_{ij}}{3}$ For known values
- $\sigma_{ij} = \frac{\bar{d}_j}{2} + \frac{d_{ij}}{3}$ For BDL values
- $\sigma_{ij} = 4 \cdot \tilde{v}_j$ For missing values

For v_{ij} as known values, u_{ij} the analytical uncertainty, and d_{ij} the analytical detection limit. Where the analytical uncertainty can include sources of uncertainty such as bias in sampling, storage, measurement conditions and a random effect error. We let $\bar{d_j}$ denote the arithmetic mean of the BDL values of variable j and $\tilde{v}_j$ the geometric mean of variable j. There have been more suggestions similar to Polissar et al. (1998) how to estimate the error matrix **S**. Most suggest a combination of error in data and measurement error such as the detection limit, minimum or otherwise.

2.7 PMF2

The program PMF2 implements the PMF algorithm to minimize (2.6.2.3), for more details on the algorithm see Paatero (1997), with several options implemented such as error model, rotations, robustness and more.

2.7.1 Error models

In PMF2 there are four different error models to choose from. The standard deviation matrix **S** is either set beforehand or the matrix is calculated in each iterative step. Below we see the two models that are of interest in this thesis, the other two models include an error model set beforehand and one specific for Poissonian data. The error models which are iteratively calculated all use the fitted matrix $\mathbf{Y} = \mathbf{GF}$ in the standard deviation matrix **S**.

• EM=-10, in this model the data is assumed to be lognormally distributed and in each iterative step **S** is recalculated.

$$s_{ij} = \sqrt{t_{ij}^2 + 0.5 \cdot v_{ij}^2 \cdot |y_{ij}|(|x_{ij}| + |y_{ij}|)}$$

Where t_{ij} is assumed to be a measurement error, v_{ij} the logarithm of the geometric standard deviation and y_{ij} the fitted value in the matrix $\mathbf{Y} = \mathbf{GF}$ calculated in each step. t_{ij} is recommended to be given a small value such as the below detection limit which can be seen as a measurement error.

• EM=-14 is a general error structure where no assumption on distribution is made and recommended for environmental work.

$$s_{ij} = t_{ij} + u_{ij}\sqrt{max(|x_{ij}|, |y_{ij}|)} + v_{ij} \cdot max(|x_{ij}|, |y_{ij}|)$$

Usually, u_{ij} is set to zero unless data is Possonian and then often v_{ij} is put to zero instead. This model and EM=-10 are recommended for receptor modelling but using EM=-10 is recommended to be used with caution since all data have to be strictly lognormally distributed. If v_{ij} is not large in comparison to size of data, EM=-14 and EM=-10 are basically the same model, so in most cases it is "safer" to use EM=-14 since no assumptions have to be made.

If each standard deviation is known, which rarely is the case, all error models can be used, setting all elements u_{ij} in **U** and v_{ij} in **V** to zero which results in $s_{ij} = t_{ij}$.

2.7.2 Robust mode

In PMF2 there is an option to run the iterations in robust mode, so outliers don't have as large a pull on Q. In environmental data outliers can arise for many different reasons, they can come from e.g. sample contamination, fault in measurement or just a couple of weeks with extremely high concentrations. It can be hard to determine the cause of an outlier; regardless of the origin they can have a large influence on the solution. The Robust mode in PMF2 downweight the outliers based on the Huber influence function, that modifies Q. Where in the robust mode the least square formulation becomes

$$Q = \sum_{i=1}^{m}\sum_{j=1}^{n}(\frac{e_{ij}}{h_{ij}s_{ij}})^2 \tag{2.7.2.1}$$

Where

$$h_{ij}^2 = \begin{cases} 1 & \text{if} \quad |\frac{e_{ij}}{s_{ij}}| \leq \alpha \\ \frac{|e_{ij}/s_{ij}|}{\alpha} & \text{if} \quad |\frac{e_{ij}}{s_{ij}}| > \alpha \end{cases}$$

So that each data point only can have a certain influence to the fit, it is recommended to use the robust mode for environmental work since outliers, whether caused by the measurement process or not, often are a part of data. Per default the outlier distance is α=4.

2.7.3 Explained variance

In order to assess how many sources p are to be used in the model several aspects need to be considered, one of them being the explained variance (EV) or rather the non-explained variance NEV. The explained variance matrix, defined by Paatero (2004b), is either a $n \times (p+1)$ or $(p+1) \times m$ matrix depending on if it describes the scores matrix (EVG) or the loadings matrix (EVF). The quantity of EV is dimensionless and ranges from 0.0 to 1.0, the value sums up how well a factor element explains one row or column in the observed matrix $\mathbf{X}$.

The explained variation matrix for $\mathbf{G}$ is a $n \times (p+1)$ matrix where the first p columns corresponds to the variance explained in each row by each factor. The *p+1*th column correspond to the residuals, so the matrix has been written as if the residuals were an extra factor. This results in the *p+1*th column being a measurement of the variance explained by the residuals i.e. the unexplained variance of the model in each row. For the explained variance matrix of $\mathbf{G}$ the elements are calculated as

$$EV(\mathbf{G})_{ik} = \frac{\sum_{j=1}^{m}|g_{ik}f_{kj}|/s_{ij}}{\sum_{j=1}^{m}(\sum_{h=1}^{p}(|g_{ih}f_{hj}| + |e_{ij}|)/s_{ij})} \qquad \text{for } k = 1, ..., p \tag{2.7.3.1}$$

and

$$EV(\mathbf{G})_{ik} = \frac{\sum_{j=1}^{m}|e_{ij}|/s_{ij}}{\sum_{j=1}^{m}(\sum_{h=1}^{p}(|g_{ih}f_{hj}| + |e_{ij}|)/s_{ij})} \qquad \text{for } k = p+1 \tag{2.7.3.2}$$

For $i = 1, ..., n$ and where Equation (2.7.3.1) explains how much each element g_{ik} explains the ith row of $\mathbf{X}$ while (2.7.3.2) explains how much e_{ij} explain the ith row. Similar equations also hold for calculations of EV($\mathbf{F}$) where columns instead of rows of $\mathbf{X}$ are explained.

2.7.4 Normalization of factor matrices

There are several ways of normalizing the output matrices $\mathbf{G}$ and $\mathbf{F}$ from PMF2, the normalization is made to properly compare the solutions. Seven different options (Paatero, 2004b) can be chosen such as normalizing by the maximum or mean value of each column of $\mathbf{G}$ or $\mathbf{F}$. In this analysis the option used is normalizing so that the mean value of each column in $\mathbf{G}$ equals 1 as to compare with results from Sirois and Barrie (1999), where the same normalization method was used.

Immediately after output the matrices $\mathbf{G}$ and $\mathbf{F}$ are normalized by dividing and multiplying the columns and the rows by the mean of respective column of $\mathbf{G}$ so that each normalized column of $\mathbf{G}$ has mean 1. Since the PMF is invariant to scaling this does not influence the solution in any way.

The columns of $\mathbf{G}$ and rows in $\mathbf{F}$ are divided and multiplied respectively by

$$\overline{\mathbf{G}^i} \qquad \text{for } i = 1, ..., p$$

Chapter 3

Analysis & Results Time series

3.1 All year time series analysis

In order to describe the temporal variation of the weekly concentrations, time series analysis models are fitted to each constituent. The time series analysis is solely made for a descriptive rather than a forecasting purpose. All components were analysed for the all year data, with some constituents additionally analysed only for the colder part of the year to look closer at the peak season. The data set that was analysed is the one described in Table 1.1. The model for the analysis can generally be expressed as.

$$C_t = m_t + s_t + Y_t \tag{3.1.0.1}$$

Where C_t are the weekly concentrations, m_t a long term trend, s_t a seasonal component and Y_t a noise component. Many constituents in the data set are lognormally distributed or the variance grows increasingly with the level of data, in these cases a temporary transformation of the concentrations ($\log(C_t)$ or $\sqrt(C_t)$) has been made in order to have fluctuations not dependent on the level of the series. As recommended by Brockwell and Davis (2002), first the long term trend was identified and removed, if no long term trend was found the mean of the series was removed in order to center the series. Then a seasonal trend was removed to obtain (weakly) stationary residuals, before an autoregressive moving-average process was fit to Y_t. Where Y_t is defined as

$$Y_t = \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \qquad \text{for } p, q = 1, 2 \text{ and } \varepsilon_t \sim N(0, \sigma) \tag{3.1.0.2}$$

The median monthly wind speed at Alert varies between 2 and 9 km/h, which corresponds to around 340-1600 km in a week on a spatial scale so the pollution is expected to be somewhat homogeneous over larger areas. We choose to limit the values of p and q to be at most two, which is both physically realistic and sufficient since it seems a larger order only overfit the models.

Each constituent was analysed in the same manner by first assessing if any transformation of the concentrations was needed in order to obtain normally distributed residuals ε_t. Transformations

were made for all constituents except for Ozone and Mercury in the all-year data. The lack of BDL and the fact that the two are gases, not aerosols, might be the reason for the more normally distributed observations. All other constituents are approximately lognormally distributed so a temporary transformation to $\log(C_t)$ was made in the fitting of trends and for describing the time series Y_t.

The long term trends are both a descriptive tool to see how the overall aerosol concentrations have developed over the years, but also a step towards estimating the underlying process of the noise Y_t. Since the data exists for such a long time span, linear and lower degree polynomial trends are for most constituents neither sufficient or physically realistic estimators of the trends. The decision to use splines was made in order to describe the long term trends accurately. Several methods can be used. A disadvantage of using splines is that they are hard to use for prediction but since the trends are fitted purely for a descriptive purpose this is not a problem. There are several types of splines, in this paper we have used smoothing splines described in Section 2.1. An advantage of using cubic smoothing splines is that it presents itself as an overall optimization problem in comparison to e.g. regression splines where we have to choose where in the time series the knots are set which easily becomes very arbitrary.

As mentioned in Section 2.1 the way of controlling the degree of smoothing is either to set λ or the equivalent degrees of freedom edf=tr($\mathbf{S}_\lambda$). The higher the trace of the smoothing matrix is the more the penalty term in Equation (2.1.0.1) becomes superfluous i.e. the solution becomes an interpolating twice differentiable function at the extreme and at the other end as the edf are dropping the solution will go to the least squares line.

The long term trends were chosen by adjusting the equivalent degrees of freedom, first by looking at plots so that all larger trends in the time series are covered, but the line doesn't include very local trends. After choosing a long term trend, that trend together with the seasonal trend is removed and we fit an ARMA-process to the remaining Y_t. We then go back and readjust the degrees of freedom a bit lower/higher and redo this process until the long term trend is optimized in terms of looking at the acf, QQ-plot, time sorted plot and Ljung-Box test of the residuals. In Figure 3.1 an example of three lines; one overfitted, one underfitted and also the chosen smooth line is displayed. We can see that the line with $edf = 5$ goes too smoothly through the large trends in data, while the line with $edf = 15$ show some more local trends with a local minimum in 1992.

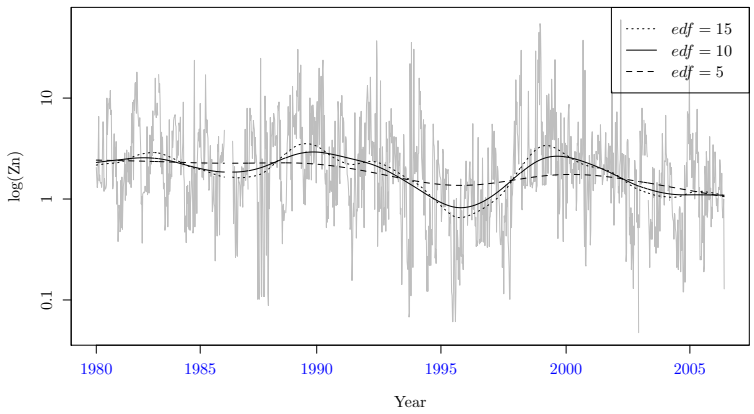


Figure 3.1: Long term trends

After the long term trends in Figure 3.2 were removed from each respective constituent time series we consider a seasonal trend. From the nature of aerosols and the weather conditions in the Arctic part of Canada we know that there exists strong seasonality in the aerosol constituents. Each weekly observation has exact times of the insertion and removal of the filter so we can use this to sort the observations. Because of this, it was possible to estimate smoothing splines as seasonality trends since the now detrended observations are centered and can be seen as independent observations throughout the year. The technique used for the smoothing splines is similar to the one used for the long term trends, where we set the degrees of freedom to adjust the amount of smoothing. Another way of removing the seasonal trend would have been to remove the monthly means but for many of the constituents the trend within one month is too drastic to accurately remove the seasonality that way.

When looking at the time sorted data there is seasonality present for all observed variables. Initially different degrees of freedom were considered for different constituents but the conclusion was that for all constituents the degrees of freedom of the spline was around edf=10 so this was set for all seasonal trends. That is where the most temporal variation was captured without overfitting i.e. not letting individual observations influence the curve too much.

When both detrended and deseasonalized we wish to have weakly stationary time series of the constituents. As already stated, in this section the stochastic process has natural limitations due to how long the air travels and the stability of the aerosols, so a maximum for p and q is set for 2 it means there is no higher process than an ARMA(2,2).

Initially we considered diagnostics such as acf, pacf, QQ-plots and Ljung-Box test of the residuals of the deseasonalized time series and then fit the best model based on the AICc criterion. If there is autocorrelation in the residuals or they seem to not be Independent and Identically Distributed (IID) normal observations, we go back and look at what can be changed in the model and whether the BDL and missing values seem to affect the residuals.

Another common problem that arises when estimating trends of time series is the existence of missing values but overall we do not consider this a large problem. In general the missing values are spread out in such a way that there are not more than a couple of weeks missing in a row,

which should not effect the estimated long term trend. There are some exceptions though, as can be seen in Figure 3.2, the aerosol constituents H^+, MSA, I and Ozone all have one gap each which reaches more than a year. There is little to be done in retrieving the trend for the missing years, so when fitting the trends the missing values have been removed in order to fit the spline. The trends overall should not be impacted by the missing values but just before and after the large gaps in H^+, MSA, I and Ozone the trend might not be accurate.

3.2 All year results

The results from the time series analysis for the whole year data shows that an adequate model can be fitted for most constituent time series. In Figure 3.2 the long term trend of the 23 different concentrations is displayed and in Figure 3.3 the seasonal trends can be seen. The largest problem with data is not the missing values but the BDL observations which censor the data, i.e. are only known to be below a certain threshold. It is easy to spot by eye in Figure 3.2 and 3.3 the concentrations where the BDL's might impact the model heavily. The most evident is the observations of H^+ and also Ni where the BDL values lie in layers in the lower part of plots in Figure 3.3b and 3.3t. In these cases we cannot expect completely independent normally distributed residuals from the fitted model.

Overall the trends have been fitted to be 3-5 year trends for the observed period of time, there are trends that can be found on a smaller time scale. Generally if the degrees of freedom are increased, so that trends on a 1-3 year scale is fitted, the residuals of the underlying process Y_t has more autocorrelation between them and looks less normally distributed. This is why some trends don't show the same fluctuations as Sirois and Barrie (1999) since the trends are fitted on a longer period of time.

3.2.1 Long term trends

A long term trend was found for all constituents except for Na. Both simple linear regression and splines was looked at for Na and there was no indication of any long term trend. For the long term trends, variations in BDLs for a significant fraction of data is a disruption. A clear example can be seen in the Pb plot in Figure 3.2i where the operational detection limit was massively lowered in the beginning of the 1990s, after this we can see a clear decrease in the trend. By looking at the decrease we cannot say if there would be a decrease or not if the ODL would have been the same for the whole time period, although the trend from 1993 and onward should be correct. One other way the BDL values can influence the trend is by putting the trend on a higher/lower level than it should be but this problem does not effect the model as much as the difference level of the ODL.

SO_4^{-2} and H^+ have similar trends where both decrease and then come into a stable state before decreasing again in the last couple of years, the difference being that H^+ has a longer stable state in the middle and SO_4^{-2} is more stable in the beginning. The lowered ODL for H^+ in the end of the 1980s until the beginning of the 1990s might affect the trend for H^+. As mentioned earlier, because of the high number of BDL's the model for H^+ will only be roughly approximated.

Br^- barely has a long term trend. There is a slight decrease from 2010 and forward and a bump in Br^- in the first couple of years but as in the case with H^+ the trend is highly likely be due to the increase of the ODL. As mentioned, we can see that the same problem presents itself for

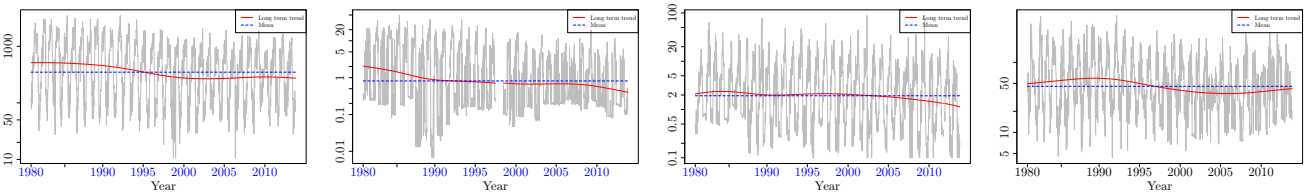
Pb, Ni, Fe and also in smaller scale for K^+, caution should be taken when looking at the trends where the ODL has large changes.

Zn and Cu have similar long term trends with much volatility and minima around 1995 followed by a peak in 2010 before fading the last years measured. As seen in Sirois and Barrie (1999) and as we will later show with the PMF analysis the two constituents Zn and Cu have a close connection to each other so a similar trend is not surprising.

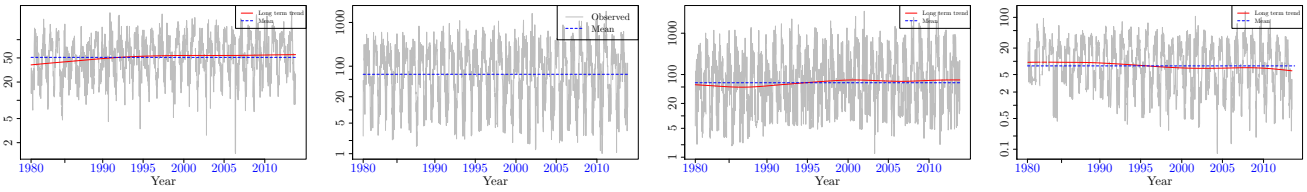
NH_4^+ has a clear maximum around 1990 and a maximum 15 year later before increasing some in the last years. V and Mn show similar trends where there is a decrease before being constant and then both show a decrease. Although, V starts to decrease in the middle of the 90s while Mn begin its decrease a decade later in the middle of the 00s.

The long term trends for Al, Fe, Ca, Mg, Pb, Zn, Cu and Ni very much correspond to the trends found in Gong and Barrie (2005), which displays trends for mentioned metals. The large difference is that a trend has been used for Na but as can be seen in Figure 3 (Gong and Barrie, 2005) the trend is very flat.

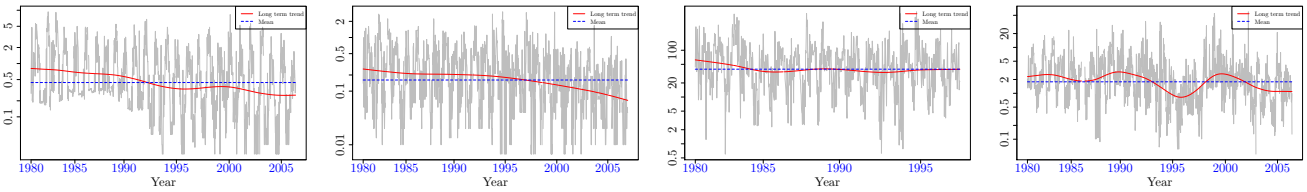
From being quite consistent for a decade Mercury shows a clear decreasing trend from 2005 and forward. This is supported by Cole and Steffen (2010) who analysed long term trends of GEM observed at Alert between 1995-2007. It is also mentioned that a small decline was noticed but no trend could be established in an analysis made on data between 1995-2005 (Temme et al., 2007). The overall decrease seem to be due to decline in anthropogenic emissions (Zhang et al., 2016).



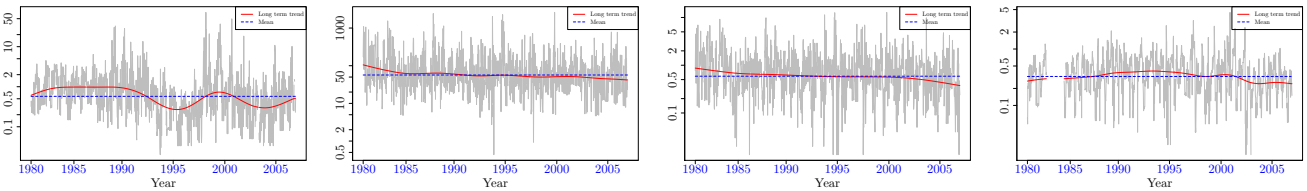
(a) SO_4^{2-} (b) H^+ (c) Br^- (d) NH_4^+



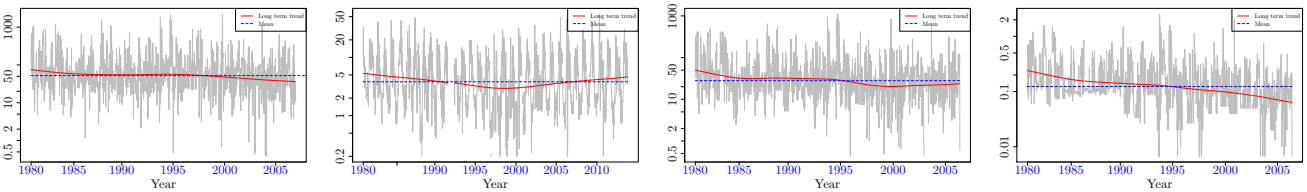
(e) NO_3^- (f) Na^+ (g) Cl^- (h) K^+



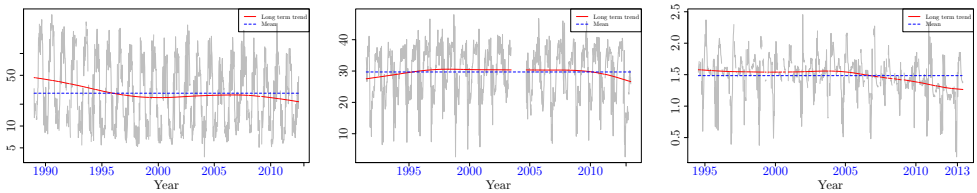
(i) Pb (j) V (k) Mg (l) Zn



(m) Cu (n) Ca^{2+} (o) Mn (p) I



(q) Al (r) MSA (s) Fe (t) Ni



(u) EBC (v) O_3 (w) Hg

Figure 3.2: Long term trends

Long term trends of all concentrations, note that all plots are on log scale except for O_3 and Hg

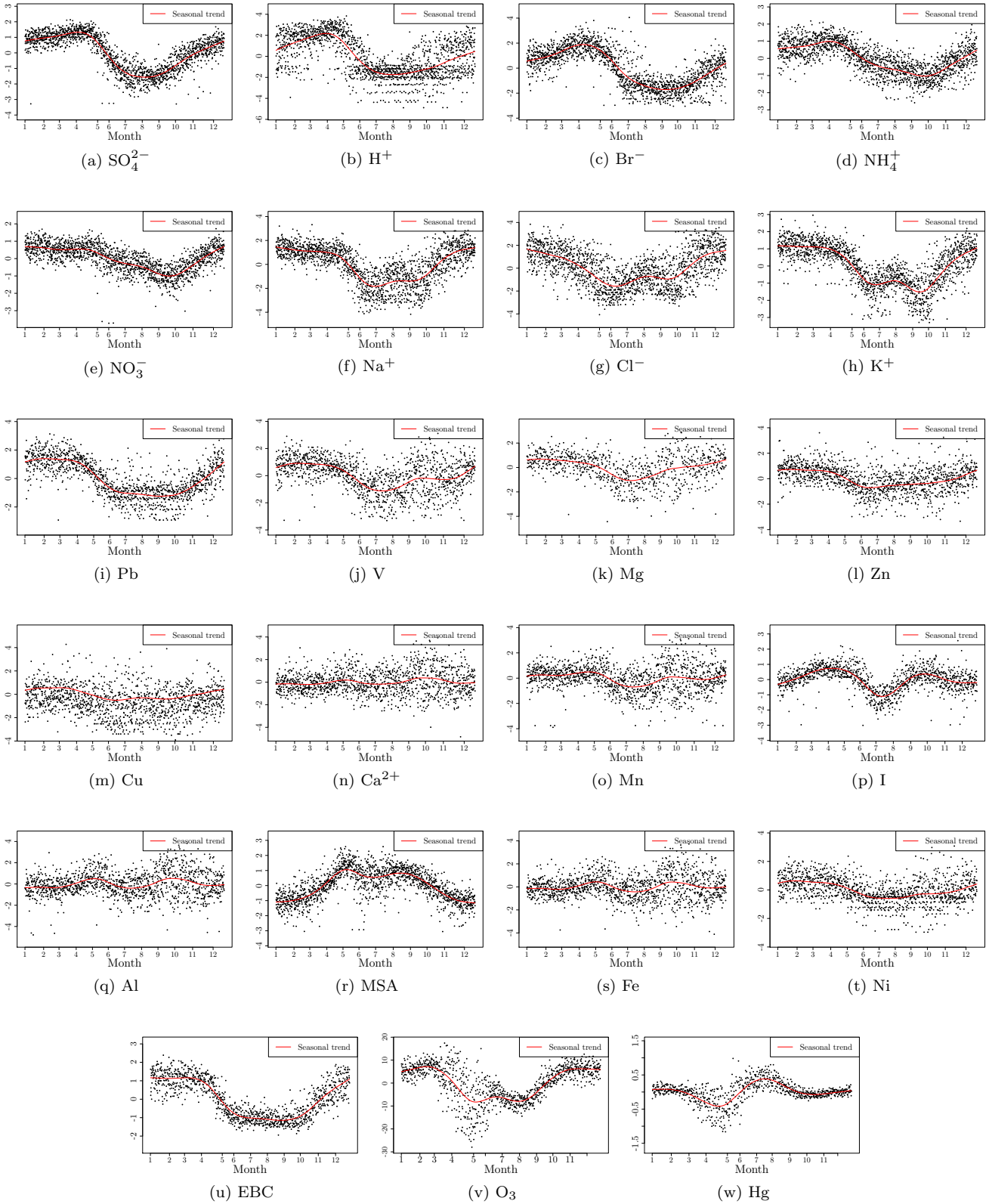
3.2.2 Seasonal trends

Contrary to the long term trends the seasonal trends are smoother. Many of the aerosol constituents have very strong seasonal patterns associated with the lifetime of aerosols in the Arctic air mass due to the variations in transport and removal processes.

The seasonal trends estimated correspond well, at the very least visually, to the seasonal trends for the constituents presented in Sirois and Barrie (1999) and Gong and Barrie (2005). So we can conclude that the seasonal trends haven't changed much in the last 16 years since the last report was published. The seasonality of all aerosol constituents analysed in this report can be seen in the last report together with Ni and Fe analysed in Gong and Barrie (2005), the exception is EBC and the gases Ozone and Mercury.

The same data set of Mercury in this analysis is also analysed in Steffen et al. (2014). From their analysis we see the same pattern (Fig.2 Steffen et al. (2014)) as in our analysis with a decrease during springtime where the variance is larger followed by a peak in June with a consistent concentration during the cold half of the year. It is also shown that the minimum of Mercury has changed in period of time between 1995-2001 and 2002-2007, most likely related to increase in temperature.

In Ozone there are two minimas in the seasonal trend where the first occurs during spring and the variance is largely increased which is associated with the polar sunrise and ozone depletion. It has been found that the depletion of O_3 is related to the depletion of Mercury which also decreases markedly in the spring. A negative correlation between O_3 and Bromide during spring has been found (Barrie et al., 1988), we will examine this relationship further in the PMF analysis.



(a) SO_4^{2-} (b) H^+ (c) Br^- (d) NH_4^+

(e) NO_3^- (f) Na^+ (g) Cl^- (h) K^+

(i) Pb (j) V (k) Mg (l) Zn

(m) Cu (n) Ca^{2+} (o) Mn (p) I

(q) Al (r) MSA (s) Fe (t) Ni

(u) EBC (v) O_3 (w) Hg

Figure 3.3: Seasonal trends
Seasonal variation of all concentrations, after long term trend has been removed.

3.2.3 ARMA-process

The numbers of p and q vary between constituents and out of a first glance a direct pattern cannot be detected. When we look closer we can see that the concentrations where the ODL varies throughout time such as Br, Pb, K^+ and Ni have a tendency to have MA(2)-terms. This might be due to the need to smooth the errors over time. For H^+ a AR(2)-process has been fitted but due to the high percentage of BDL values in data, 39.79 % , there is no way of fitting a model that fulfill the model criteria of weakly stationary Y_t and iid normal residuals.

In Figure 3.3 the variance during the seasons vary some. This does not show in the residuals of the models except for the gases Ozone and Mercury where we can see a clear seasonality in the variance. In Figure 3.4 concentrations of Ozone and Mercury are plotted with removed long term trend and seasonality with a clear systematic increase of variance each year.

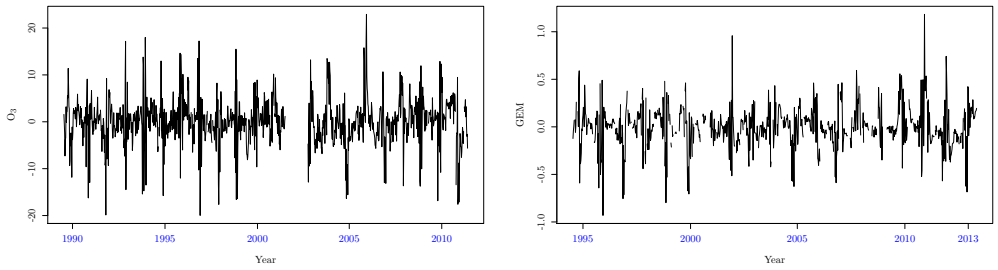


Figure 3.4: Ozone and Mercury without long term and seasonal trends

To properly model the concentrations in addition to an ARMA-process we use the GARCH-process to model the conditional variance. For both gases we use a GARCH(1,1)-process in combination with an ARMA(1,1)-process for Ozone and ARMA(2,1)-process for Mercury. In Table 3.1 we see that fitted parameters to the GARCH-process and in Figure 3.5 the conditional variances are plotted which have a clear seasonal pattern.

Gas	ω	α	β
O_3	2.949	0.374	0.521
GEM	0.002	0.443	0.556

Table 3.1

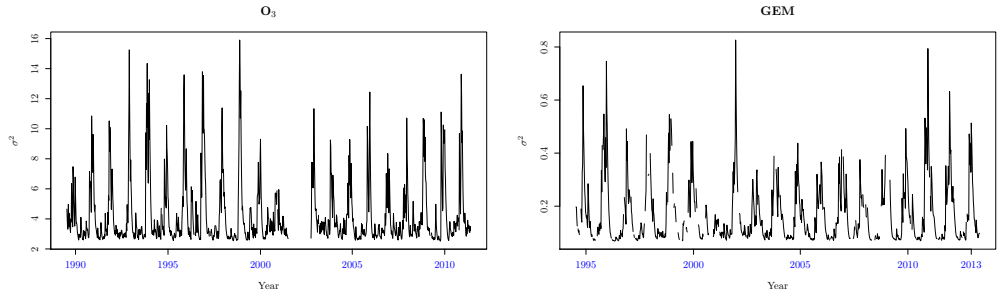


Figure 3.5: Conditional σ^2 for Ozone and Mercury

We will see in the winter time series analysis that the number of p and q are reduced for all constituents analysed in both time periods. We can suspect the number of BDL values will influence this but also that the processes of Y_t during the year are profoundly different so a higher number of p and q are needed to explain the series.

Constituent	Long term trend		Seasonal Cycle	ARMA	ARMA coefficients			
	R^2	*df*	R^2	R^2	φ_1	φ_2	θ_1	θ_2
SO_4^{-2}	0.05	5	0.77	0.03	0.06	0.47	0.27	−0.34
H^+	0.04	5	0.52	0.07	0.79	−0.5		
Br	0.02	6	0.7	0.09	0.87		−0.48	−0.05
NH_4^+	0.05	5	0.57	0.09	0.1	0.5	0.31	−0.33
No_3^-	0.02	4	0.55	0.06	1.1	−0.24	−0.72	
Na	0		0.63	0.04	−0.22	0.4	0.53	−0.21
Cl^-	0.01	6	0.46	0.05	0.48		−0.18	
K^+	0.01	5	0.64	0.02	−0.02	0.45	0.34	−0.34
Pb	0.08	7	0.58	0.06	0.75		−0.35	−0.12
V	0.08	5	0.3	0.08	0.37			
Mg	0.03	6	0.24	0.09	0.35			
Zn	0.15	10	0.22	0.09	0.78		−0.45	−0.11
Cu	0.2	8	0.1	0.29	0.83		−0.58	
Ca	0.04	7	0.04	0.1	0.46	−0.15		
Mn	0.03	5	0.12	0.1	0.33			
I	0.06	9	0.45	0.08	0.67	−0.35		
Al	0.02	5	0.09	0.08	0.14	0.32	0.14	−0.26
MSA	0.03	5	0.54	0.07	−0.36	0.32	0.73	
Fe	0.05	6	0.07	0.09	0.5	−0.22		
Ni	0.11	5	0.21	0.12	0.67		−0.28	−0.09
EBC	0.04	5	0.77	0.03	0.87		−0.48	−0.24
O_3	0.02	5	0.59	0.05	0.68		−0.35	
GEM	0.09	6	0.47	0.05	0.81	0	−0.45	

Table 3.2

3.3 Winter time series analysis

By the all-year time series analysis the large part BDL values, strong seasonality and varying variance motivates the decision to reduce the data set and only look at the winter part of the year. For the PMF analysis we will use data between November to May to reduce the number of BDL values in data. A natural choice would be to use the same time period for the winter time series but the seasonality will influence the time series if not a shorter period is used.

A choice was made to analyse the same constituents as in the previous report together with the new chemical compounds/elements, which all have peaks during the winter/spring. So, as in the last report (Sirois and Barrie, 1999) we will use the data between 1st of January and 1st of May for SO_4^{-2}, NO_3^-, H^+ and NH_4^+ while January 1st to April 1st for Zn, Pb, Fe, Ni, K^+, xV, xMn, EBC, GEM and O_3, where xV and xMn are the V and Mn associated with non-soil, calculated as

- xV=**V**-0.0013**Al**
- xMn=**Mn**-0.0065**Al**.

As we will see in the PMF analysis there is a large part Mn and V that is associated with Soil which is heavily loaded with Al. Since there is an interest in looking at the trends of Mn and V which is not associated with Soil we analyse xMn and xV. The same measurement of xMn and xV was used in Sirois and Barrie (1999) and the relationship between Mn and Al is considered to be constant.

The two are analysed in the last report as well. The added constituents in the new analysis have been added to the January to April analysis. We chose to do this for Fe and Ni because of the metal characteristics which follows Zn and Pb while Ozone, Mercury and to some extent EBC have a change in variance during April and May.

The overall process of the winter time analysis is the same as for the all-year described in Section 3.1. In the last report no transformation was needed for any of the concentrations, in this extended data sets log-transformations were needed for all metals (Zn, Pb, Fe, Ni) and including the alkali metal K^+. For xV and xMn the transformation of $\sqrt{(C_t)}$ was made in order to obtain normally distributed residuals. During the analysis we momentarily transform the concentrations as $\sqrt{xMn}$ and $\sqrt{xV}$. In Figure 3.6k and 3.6l the data and trend has been transformed back to normal scale.

The long term trends were analysed as for the all-year data by smoothing splines with various degree of freedom. When analyzing the winter data the mean and variance for SO_4^{-2}, H^+ and NH_4^+ seem to have a large change in the mid 1990s. A decision was made to adapt a change point model for these three constituents.

As we can see in Figure 3.6a-3.6c the dominant part of the values before the change point lies above the overall mean and respectively most values lie below after the change point. A least square method (see Section 2.5) was used to estimate the change point in the time series, where we treat data as a linear process and the mean was used to asses the change point. After the change point has been set we divide the vector in two parts and continue with the analysis as before but with two separate vectors.

In the all year data it was known that a seasonal trend existed. For the winter data this is not the case so each constituent had to be assessed individually. After the long term trend is removed we look at the autocorrelation function of $C_t - m_t$ to see if there is seasonality in the

data. The data should naturally be autocorrelated at around lag 17 for the Jan-May data and 13 for Jan-April, depending on how many weekly observations are fitted in the time span, if there exists a seasonal trend.

We fit the ARMA-process as described in the all year analysis, with looking at the AICc value, acf and pacf plots. As mentioned the natural settings of the transportations of the aerosols are restricted and although we could limit the number of p and q in the ARMA(p,q)-process, this is not necessary. All time series is fitted either an AR(1)-, Ma(1)- or ARMA(1,1)-process, which all are well-fitted with white noise residuals.

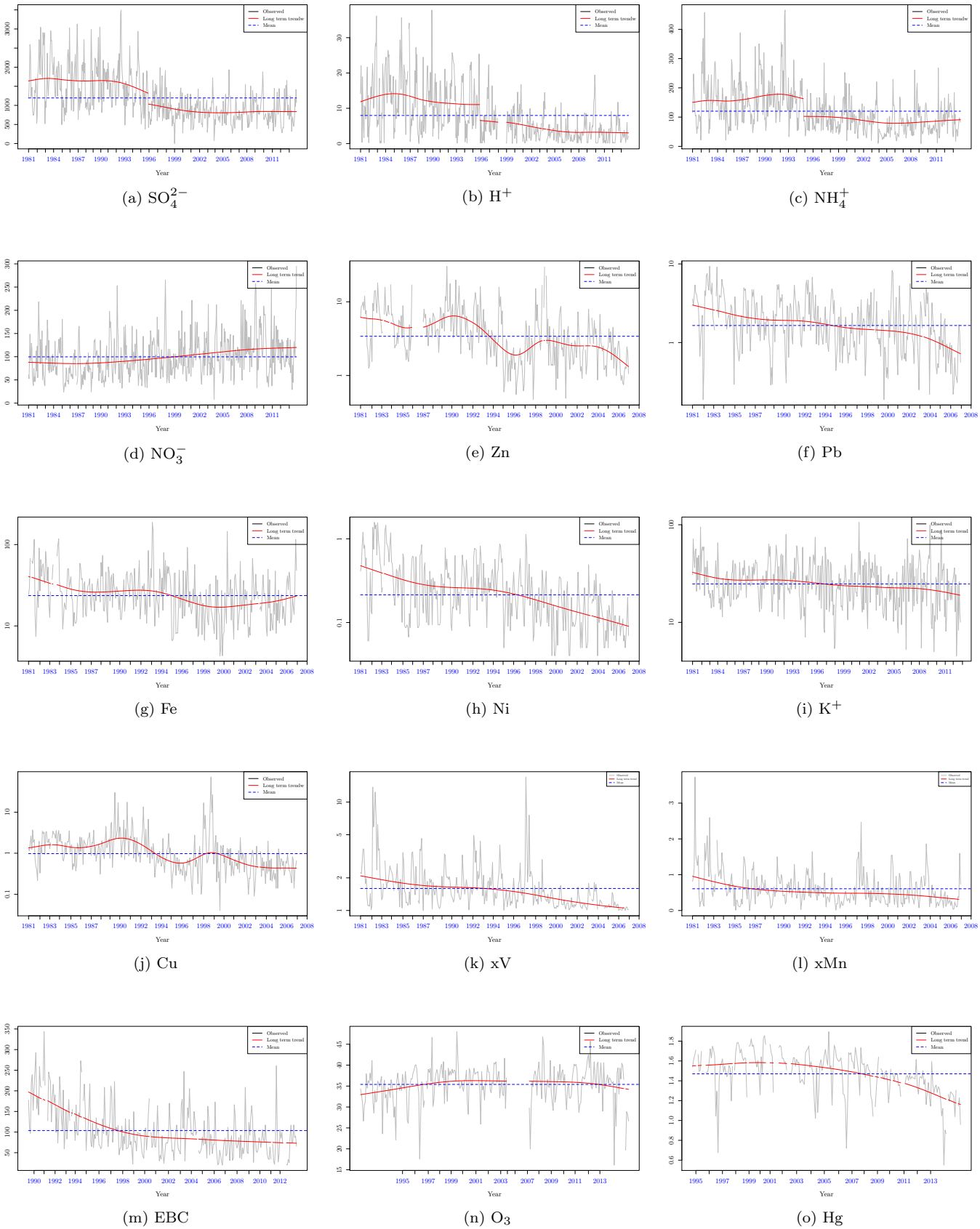
3.4 Winter time series results

For the three constituents for which we used a change point model to optimize $\hat{\tau}$ the result is very similar: for SO_4^{-2} and H^+ the change point was estimated to the last observation of 1995 and for NH_4^+ the last observation of 1994. That the change point lies in the last observation in a year makes sense both from a statistical and an environmental point of view, since there is a clear jump between the first of May until the first of January.

3.4.1 Long term trends

The results for the long term trends are displayed in Figure 3.6 as well as in Table 3.3. Most trends look more or less the same as in Figure 3.2 but more refined, the large difference is the change point models which have less smooth trends divided into two parts. The problem with the BDL values seen in the all-year analysis is almost non-existent in the winter peak data, especially for H^+ and most metals.

The large decrease of Sulfate in the beginning of the 1990's has been attributed to the fall of the Soviet Union (Sharma et al., 2004). This is most likely the cause for the decrease of H^+ and NH_4^+ as well (Bodhaine and Dutton, 1993). In Figure 3.2m black carbon has a steep decreasing trend from the beginning of 1990 until 1998 when it levels out. This decreasing trend has also been recognized as a consequence of the collapse of the Soviet Union economy (Sharma et al., 2004).



(a) SO_4^{2-} (b) H^+ (c) NH_4^+

(d) NO_3^- (e) Zn (f) Pb

(g) Fe (h) Ni (i) K^+

(j) Cu (k) xV (l) xMn

(m) EBC (n) O_3 (o) Hg

Figure 3.6: Long term trends-Winter
a)-d) based on Jan-May data, normal scale
e)-j) based on Jan-April, log scale
k)-o) based on Jan-April, normal scale

3.4.2 Seasonality and AR/MA-process

From the analysis we can conclude that the constituents with a strong seasonal trend were SO_4^{-2}, NO_3^-, H^+ and NH_4^+, xV, xMn, GEM and O_3. The metals and EBC have no significant month to month trend from January to April, we can also hint this in Figure 3.3.

In Table 3.3 all detrended time series Y_t are sufficiently fitted with an AR(1)-process, except for black carbon where a MA(1)-process was a better fit and Mercury which is fitted with an ARMA(1,1)-process. Since we are only using the 1st of January to 1st of April data for Ozone and Mercury the variance is consistent, as we can see in Figure 3.3v and 3.3w it is in the shift between March and April that the variance starts to blow up. Judging from diagnostics all residuals can be considered white noise. We can conclude that using the all-year data will in some cases lead to misconceptions on long term trends and underlying processes. When looking at the all year data, the large part of low and BDL values will get to much influence especially when transforming data to a logarithmic scale. When only looking at the peak season, clearer trends and time series processes can be seen.

Constituent	Long term trend		Seasonal Cycle		ARMA	ARMA coefficients	
	R^2	*df*	R^2	*df*	R^2	φ_1	θ_1
SO_4^{-2} (1980-1995)	0.04	4	0.17	3	0.17	0.47	
SO_4^{-2} (1996-2013)	0.04	4	0.16	3	0.05	0.28	
NO_3^-	0.09	4	0.04	5	0.15	0.42	
H^+ (1980-1995)	0.03	4	0.15	3	0.15	0.44	
H^+ (1996-2013)	0.11	4	0.23	3	0.19	0.39	
NH_4^+ (1980-1994)	0.03	4	0.04	3	0.24	0.5	
NH_4^+ (1995-2013)	0.04	4	0.09	3	0.15	0.33	
Zn	0.35	9	0		0.12	0.51	
Pb	0.18	5	0		0.18	0.46	
Fe	0.14	6	0		0.06	0.27	
Ni	0.29	4	0		0.16	0.47	
K^+	0.06	5	0		0.08	0.29	
Cu	0.33	9	0		0.06	0.3	
Vx	0.31	4	0.02	5	0.19	0.54	
Mnx	0.13	4	0.01	5	0.22	0.51	
EBC	0.36	4	0		0.05		0.27
Ozone	0.06	4	0.11	5	0.03	0.21	
Mercury	0.32	4	0.24	6	0.05	0.71	−0.46

Table 3.3

Chapter 4

PMF analysis

4.1 Analysis

We now proceed in the analysis and will look at the PMF analysis of the data set. All concentrations will be analysed except for Fe. It is only used in the time series analysis since Fe is measured by ICP mass spectroscopy and is not total Fe but rather the acid soluble fraction of total aerosol iron. Our main goal with the analysis is reducing the dimension of $\mathbf{X}$ as $\mathbf{X}$=$\mathbf{GF}$, where $\mathbf{X}$ is a $n \times m$ matrix with n observed weeks and m aerosol constituents. Which in turn is made to obtain interpretable sources of the measured factors.

4.1.1 Reproduction

The first step in the analysis is to recreate the PMF analysis made in Sirois and Barrie (1999). The main problem is the lack of information on the analysis made, such as handling of missing values and choice of error model. Recalling from Section 2.6.2, the main objective of PMF is to minimize Q (2.6.2.1) where s_{ij} has to be estimated. Even though the PMF analysis is quite robust the choice of $\mathbf{S}$ is vital in reconstructing the initial analysis. Unfortunately the way of reconstructing the analysis without all information is trial and error which is a very time consuming process. Throughout the PMF analysis the robust mode is used with default outlier distance $\alpha = 4$.

Handling of missing BDL observations

In accordance with Sirois and Barrie (1999) we set, for d_{ij} as the detection limit,

- $x_{ij} = \frac{2}{3} \cdot d_{ij}$ For BDL values

We wish to have errors so that small changes of the constant $\frac{2}{3}$ will not change the outcome of the analysis, therefore a sensitivity analysis by using $\frac{1}{2}$ instead (the constant recommended by Polissar et al. (1998)) is made to make sure no major changes are made to the outcome. As for the errors we trust that using a fixed analysis error will compensate for the higher errors of the BDL values. We come back to this in the next section.

Almost all concentrations have some missing values and since all rows in the matrix **X** must be complete we quickly realize that excluding all rows with one or more missing values will reduce the data immensely. The option is to impute missing values which is also a method validated by previous literature (Section 2.6.2). By looking at the number of missing values in each row we choose a combination of the two. Since if a large amount of constituents is missing in a row that observation has no, little, or a misleading contribution to the analysis.

We chose to impute values as suggested by Polissar et al. (1998). Namely, to substitute missing values as the geometric mean, we use the seasonal geometric mean which are more accurate than the overall. For each constituent the geometric mean aggregated by month is calculated and imputed in the corresponding missing values. For the errors of the missing values we also follow the same guidance and put the standard deviations four times higher for the imputed values. If there is 6 or more missing values in one row we choose to discard that row. There are very few rows with more than 5 missing values, e.g. in the reproduction only 3 out of 442, but if there are, it is often one type of analysis method missing (INAA, ICP or IC) which might influence the analysis.

Error model and normalization

In the previous analysis Sirois and Barrie (1999) the observations are assumed to be lognormally distributed, an explicit error model is not mentioned but a reasonable assumption is that error model -10 was used since it corresponds to lognormally distributed factors.

$$s_{ij} = \sqrt{t_{ij}^2 + 0.5 \cdot v_{ij}^2 \cdot |y_{ij}|(|x_{ij}| + |y_{ij}|)}$$

What can be altered is **T**, **V** and the handling of missing values, since there are only guidelines on the three and no clear restrictions, each s_{ij} can be altered infinitely. When estimating the errors, **T** is often set as a small value such as the BDL, min or the least significant digit in the measurement (Kim and Hopke; Lee et al., 1999; Polissar et al., 1998). In this error model it functions as an imposed error: both as a measurement of analytical uncertainties and to control the errors of small values. When **T** and **V** are both used, the uncertainty matrix **S** is a trade-off between a fixed analysis error and the uncertainty in concentrations, so small values will have a higher error in relation to size. Since we know the BDL values are small but not how small they have a higher uncertainty than known values, which naturally will be implemented when using **T**. Using **T** also helps the algorithm that minimize **Q** to converge since there is a risk of iterating very small errors for values close to zero.

For this error model **V** is theoretically the logarithm of the geometric standard deviation although this is usually set as a constant. After exploring using different values of v_{ij} for different constituents we conclude that for the consistency of the analysis a constant is preferable and instead using different values of t_{ij} for different chemical compounds. If both **V** and **T** change with regards to constituent and/or season the analysis will not be comparable if e.g. using different periods of time. We also conduct the analysis using different values of **T** such as the BDL and minimum for each constituent along with using different small constants between $\frac{1}{4}$ and 1 times the values.

After a great deal of trials and almost as many errors we select to use

- $t_j = \frac{1}{2} max(d_j) \quad$ for $j = 1, ..., m$

- $v = 0.4$

which almost reproduces the analysis of 1980 to 1995 data by Sirois and Barrie (1999), some tries have been closer to the values but in turn unstable with more local minimas of **Q**.

As in Sirois and Barrie (1999) the matrices **G** and **F** have been normalized by dividing and multiplying the columns and the rows by the mean of respective column of **G**, so each normalized column of **G** has mean 1. This will ensure that the results of the two analyses are comparable.

4.1.2 New analysis

For the extended analysis there are some limitations in how long the different concentrations are measured. First of all Mg only has observations until May 1997 so in a PMF analysis beyond 1997 Mg has to be excluded. For the new constituents Black carbon, Ozone and Mercury there is a lower limit where they can be included since measurements started in 1989, 1991 and 1995 respectively.

Due to the limitations, different time periods were analysed

- 1980-2007 without Mg
- 1995-2007 without Mg
- 1989-2007 without Mg with EBC
- 1995-2007 without Mg with EBC
- 1995-2007 without Mg with EBC, O_3 and Hg

What the analyses from the different periods of time and with change of constituents shows is that factors from the last report are more or less stable. The main change is the magnitude of the factors, as a result this is expected from what we have seen from the time series analysis. A good example of this is the factor named PHOTO-S in Sirois and Barrie (1999). When looking at 1995-2007 the factor loadings are almost half the loading between 1980-1995, which are likely driven by changing anthropogenic emissions of sulphur dioxide gas to the atmosphere.

So even though the results are interesting, and are validated by the time series analysis, they are also very much what was expected from the analysis of Sirois and Barrie (1999). Thus a different analysis was undertaken; comparing PMF analysis for the dark time of year before polar sunrise to that for the light time of year after polar sunrise.

After complete darkness in January and February the sun rises in March at Alert and it quickly gets lighter. Thus photochemistry of the atmosphere is switched on. In Figure 3.3 it can be seen that many concentrations either drop or increase drastically when this occurs. There have been studies of the change showing that the polar sunrise is an important factor of changes in the composition of the lower Arctic atmosphere (Barrie et al., 1988). As this new analysis shows interesting results and a clear change in factors and factor loadings, this will be our main focus for the PMF analysis.

Error model

When reproducing the results we used the error model -10 which is used for lognormally distributed observations. When introducing Ozone and Mercury in the model this error model no longer holds since they're clearly not lognormally distributed. Instead we use EM -14 which should more or less be the same model although our **V** is quite high. When looking at the value

of Q when switching between the two error models Q is higher when using EM -10 than for EM -14. An outcome of this is that the relative errors are lower for EM -14.

When doing sensitivity analysis, using EM-14, results in different solutions, so the error model used doesn't give a robust result. It seems the trade-off between the fixed error in **T** and the observation based error in **V** is too heavily weighted towards the fixed errors. Two things can be done: either raise the values of **V** or lower the values of **T**. Different approaches was made and in the end we choose the error that produced the model which resembled the results gained from using EM -10 together with same solution from sensitivity analysis of **T**. This was to use the minimum BDL instead of using the maximum BDL values seen in Table 1.1 as for the reproduction.

In the Light/Dark analysis we use,

- $t_j = \frac{1}{2} min(d_j) \quad$ for $j = 1, ..., m$
- $v = 0.4$

and $u_{ij} = 0$ for errors

$$s_{ij} = t_{ij} + u_{ij}\sqrt{max(|x_{ij}|, |y_{ij}|)} + v_{ij} \cdot max(|x_{ij}|, |y_{ij}|)$$

The handling of missing values and BDL observations is the same as in the reproduction analysis. There are two exceptions in the errors, since ozone and mercury are measured instrumentally they do not have a detection limit so instead of using the minimum detection limit we use the overall minimum instead. We could scale the errors and get the same solution. If we use e.g.

- $t_j = \frac{1}{5} min(d_j) \quad$ for $j = 1, ..., m$
- $v = 0.2$

We will get the same solutions which talks to the robustness of the model.

4.1.3 Selection of factors

We base our choice of factors on the number chosen in the last analysis, which was ten factors, as well as the interpretability and how physically realistic they are. We also look at diagnostics such as the value of Q. We can also look a the scaled residuals which should be ranging between -2 and 2 without visual patterns. There are exceptions to this though, if concentrations are almost completely explained by one factor this will show in much smaller residuals.

Ultimately the choice of model comes down to interpretability, so a choice of different factors which all had pros and cons statistically was presented. These were looked over by Leonard Barrie to see which solution was the most physically realistic. The changes in terms of concentration from the last report is that Mg is removed and Ni, EBC, Ozone and Mercury have been added to the analysis. Since the choice in 1999 was to use 10 factors and Mg was not its own factor the same or a higher number of factors is expected to explain all concentrations. For the dark period the choice of factors was between 10-13 and for the light period 11-14, since most factors are more active when the polar sunrise occurs and processes such as ozone and mercury depletion and aerosol Br production by photochemistry set in.

The choice was made to use 12 factors for the dark period and 13 for the light. For the light period there where two similar solutions with minimas of Q close to each other, while for the

dark period there was a solution with smallest minima of Q. So for the light period the most interpretable solution was chosen.

4.2 Results

The results from our PMF analysis will be displayed as in Sirois and Barrie (1999). For the dark and light period respectively we look at the factor loadings matrix **F** (corresponding to Table 5 in Sirois and Barrie (1999)), explained variance matrix (corr. to Table 4) and the factor scores plotted (corr. to Figure 5 and 6). The factor loadings are assumed to be normally distributed.

As mentioned the factor loadings and scores are normalized so we can compare the loadings between the different analyses. The factor names are chosen by L.A. Barrie, which are based on the interpretation of the loadings, the same technique used in Sirois and Barrie (1999). For a geoscientist the names of the loadings are interpretable. For example modified means that there are other constituents loaded on the factor than the original primary, *anthro* is anthropogenic i.e. man-made and *photo* stands for photochemical.

The factor ACID PHOTO-S, equivalent to PHOTO-S in the last report, is a factor mainly associated with SO_4^{-2} and H^+ but also NH_4^+, related to gas phase photochemical oxidation of sulphur dioxide to sulphuric acid and absorption of ammonia gas to form NH_4^+. Recall that the three concentrations which we used a change point model for are the main three components in this factor. When looking at the factor loadings we can clearly see the results from the time series analysis also here in the PMF analysis. When comparing the dark period to the light the loadings of SO_4^{-2}, H^+ and NH_4^+ are about the double for the light period compared to the last report. The loadings of SO_4^{-2} and H^+ are about the double once more from the light period, while NH_4^+ only has decreased with about a fifth.

Of all factors BROMIDE is one of the strongest; both when using different errors, number of factors and handling of BDL values there is always a factor driven by Br. From the tables we see that during the dark period there is also some SO_4^{-2} and Na loaded onto BROMIDE in a higher degree than in the light period. In the earlier analysis BROMIDE has nothing else loaded on it, this can both depend on the solutions chosen but also that especially Na is more correlated with Br in recent years.

	BROMIDE		ACID PHOTO-S		SEA-SALT		LEAD ANTHRO		MOD. SEA-SALT		OIL COMB.		SMELTER		ZINC		IODIDE		SOIL		BLACK CARBON		MOD. MSA	
SO_4^{-2}	63.58	±25.27	115.05	±22.65	6.19	±14.97	104.6	±22.81	172.24	±22.55	62.62	±21.09	0.7	±5.78	0.01	±0.10	8.14	±6.67	0.87	±7.68	18.51	±20.10	28.81	±15.71
H^+	0.01	±0.02	2.79	±0.10	0	±0.01	0	±0.01	0.01	±0.01	0	±0.02	0.01	±0.00	0.01	±0.00	0	±0.00	0	±0.01	0.01	±0.02	0.01	±0.01
Br^-	4.38	±0.24	0.07	±0.09	0.03	±0.12	0.35	±0.13	0.01	±0.08	0.21	±0.13	0	±0.05	0.06	±0.02	0.14	±0.05	0	±0.05	0	±0.11	0.05	±0.08
NH_4^+	0.33	±2.36	10.27	±2.43	0.16	±1.45	21.1	±2.96	17.33	±2.35	4.57	±1.71	0.01	±0.39	1.24	±0.47	0.96	±0.84	0.01	±0.38	0.04	±1.57	0.34	±1.73
NO_3^-	2.74	±2.79	0.04	±1.46	9.15	±2.91	12.01	±2.93	29.41	±3.56	1.38	±2.17	0.08	±0.95	1.28	±0.50	3.22	±1.67	3.45	±1.73	16.83	±3.14	0.09	±2.16
Na^+	18.73	±7.66	1	±2.93	167.61	±13.43	10.48	±5.33	50.02	±8.15	1.35	±3.26	0.56	±3.43	1.43	±0.61	1.97	±2.18	0.46	±3.53	3.9	±5.60	1.18	±4.11
Cl^-	0.22	±2.10	0.06	±1.67	370.92	±16.86	0.66	±3.02	4.7	±6.31	0.24	±1.96	0.67	±3.07	3.73	±1.23	0.31	±1.61	3.46	±3.88	5.36	±4.01	0.53	±3.51
K^+	3.09	±0.84	0.94	±0.49	8.05	±0.88	5.05	±0.78	2.5	±0.69	0.56	±0.54	0.01	±0.24	0.14	±0.06	0	±0.13	0	±0.13	0.31	±0.58	0.04	±0.40
Pb	0.15	±0.02	0	±0.00	0	±0.01	0.94	±0.05	0	±0.00	0	±0.00	0.06	±0.01	0.03	±0.01	0	±0.00	0.02	±0.01	0	±0.01	0	±0.00
V	0	±0.00	0	±0.00	0	±0.00	0.04	±0.01	0	±0.00	0.22	±0.01	0	±0.00	0.01	±0.00	0	±0.00	0	±0.00	0	±0.00	0	±0.00
Zn	0.13	±0.05	0.02	±0.05	0	±0.04	0.43	±0.10	0	±0.04	0.13	±0.07	0.56	±0.08	2.24	±0.19	0	±0.03	0.01	±0.06	0.01	±0.05	0.01	±0.05
Cu	0.02	±0.01	0	±0.01	0.01	±0.01	0.07	±0.02	0.01	±0.02	0	±0.01	0.82	±0.05	0.01	±0.01	0	±0.01	0	±0.01	0	±0.01	0	±0.01
Ca^{2+}	2.22	±1.16	0.01	±0.54	7.24	±1.59	0.01	±0.29	0.02	±0.77	1.43	±0.69	0.01	±0.26	1.61	±0.59	1.13	±0.73	51.43	±2.80	0.04	±0.92	0.01	±0.47
Mn	0	±0.00	0	±0.01	0	±0.01	0.26	±0.03	0	±0.00	0.08	±0.03	0.01	±0.00	0.03	±0.01	0	±0.00	0.38	±0.03	0	±0.01	0	±0.01
I	0	±0.00	0	±0.01	0	±0.01	0.03	±0.01	0	±0.01	0.02	±0.01	0.01	±0.00	0.01	±0.00	0.26	±0.02	0	±0.01	0	±0.01	0	±0.01
Al	0.01	±0.24	0.07	±0.45	0.01	±0.25	0	±0.13	0	±0.18	4.82	±0.91	0.66	±0.26	0.01	±0.19	0.17	±0.34	56	±2.30	0	±0.21	0	±0.18
MSA	0	±0.01	0.1	±0.03	0	±0.02	0	±0.01	0	±0.02	0	±0.01	0	±0.01	0.01	±0.00	0	±0.01	0	±0.01	0	±0.01	1.29	±0.06
Ni	0	±0.00	0	±0.00	0	±0.00	0.04	±0.01	0.01	±0.00	0.04	±0.01	0	±0.00	0.02	±0.00	0	±0.00	0.02	±0.01	0.01	±0.00	0	±0.00
EBC	2.78	±1.91	8.8	±2.24	0.98	±1.30	0.55	±2.55	0.23	±2.13	3.42	±2.27	0.12	±0.58	3.65	±1.01	0.18	±0.89	0.1	±1.53	50.49	±3.53	2.58	±2.17
Ozone	0.04	±1.22	3.48	±1.30	0.09	±1.14	0.03	±0.92	14.3	±1.94	0.02	±0.59	0.35	±0.58	1.41	±0.38	2.27	±1.00	2.23	±0.84	2.68	±1.57	8.94	±1.81
Mercury	0.01	±0.09	0.18	±0.07	0.01	±0.06	0	±0.07	0.61	±0.10	0	±0.04	0.01	±0.03	0.09	±0.03	0.08	±0.05	0.12	±0.05	0.11	±0.08	0.36	±0.09

Estimated factor loadings and standard deviations

Table 4.1: Factor loadings of dark period

SEA-SALT and MODIFIED SEA-SALT, equivalent to MIXED PHOTO-S/SEA-SALT in Sirois and Barrie (1999), are factors in both the dark and light period. From the time series analysis we could see that Na^+ has been stable throughout the measured period while Cl^- has had minor fluctuations. The main difference from the previous analysis is when we divide SEA-SALT into a dark and a light period the loadings of Na and Cl^- are much higher during the dark. The ratio between them also seem to be slightly changed with a little bit more Na loaded onto the factor during the dark period. For the MODIFIED SEA-SALT factor we can see another example of the drop in SO_4^{-2}, where the loadings in Table 4.1 and 4.2 are less but the explained variance in Table 4.3 and 4.4 is about the same as in corresponding tables in Sirois and Barrie (1999).

	ACID PHOTO-S		BROMIDE		MOD. SEA-SALT		SEA-SALT		NITRATE		ZINC		OIL COMB.		SMELTER		IODIDE-SOIL		SOIL		MOD. BC		MSA		PHOTO O3-HG	
SO_4^{-2}	179.08	±40.76	91.86	±37.25	242.71	±43.31	65.36	±27.98	25.69	±18.48	72.36	±41.22	55.69	±37.73	5.11	±6.76	0.12	±5.35	0.57	±17.82	60.44	±43.36	59.72	±23.42	0.06	±2.51
H^+	4.68	±0.20	0	±0.03	0	±0.04	0	±0.03	0	±0.02	0	±0.03	0	±0.02	0	±0.02	0	±0.01	0	±0.01	0	±0.04	0	±0.02	0.1	±0.02
Br^-	1.15	±0.36	9.81	±0.73	0	±0.12	0.02	±0.32	0.21	±0.34	0.02	±0.39	0.11	±0.26	0.01	±0.08	0.36	±0.23	0	±0.03	0.01	±0.34	0	±0.07	0.83	±0.24
NH_4^+	15.83	±4.52	0.16	±3.69	32.58	±5.12	0.09	±2.50	13.4	±3.88	9.81	±4.64	9.24	±5.51	0.06	±0.87	0.03	±1.40	0.73	±3.00	0.07	±3.13	1.8	±2.32	5.52	±2.66
NO_3^-	0.22	±2.22	0.37	±4.02	15.73	±4.84	11.8	±3.79	50.72	±4.73	0.4	±3.29	0.11	±4.06	0.87	±0.63	0.06	±1.93	6.92	±3.05	10.7	±4.33	3.28	±1.95	0.02	±0.84
Na^+	0.45	±2.84	9.45	±4.56	43.9	±5.65	97.08	±8.28	0.03	±1.32	0.21	±3.93	0.49	±5.27	1.8	±1.18	3.26	±1.90	0.07	±2.67	11.87	±5.43	0.16	±1.86	0.01	±0.64
Cl^-	0.02	±0.69	0.13	±1.87	0.06	±2.05	149.55	±7.70	0.35	±1.81	0.1	±1.30	0.04	±1.76	1.38	±0.61	0.2	±1.33	0.79	±1.29	0.06	±1.94	0.05	±0.75	3.94	±1.54
K^+	1.26	±0.58	0.62	±0.62	4.95	±0.78	5.58	±0.77	0.26	±0.27	1.93	±0.81	0.94	±0.82	0	±0.09	0.02	±0.25	0.02	±0.40	2.39	±0.81	0.01	±0.25	0.05	±0.12
Pb	0.1	±0.03	0	±0.02	0.1	±0.03	0.02	±0.02	0	±0.01	0.39	±0.05	0.08	±0.03	0.06	±0.02	0	±0.01	0	±0.01	0.11	±0.04	0	±0.01	0	±0.01
V	0	±0.01	0.02	±0.01	0	±0.01	0	±0.01	0	±0.01	0	±0.01	0.18	±0.02	0	±0.00	0.03	±0.01	0	±0.00	0.01	±0.01	0	±0.00	0.01	±0.01
Zn	0	±0.05	0.05	±0.08	0	±0.08	0.01	±0.05	0.23	±0.07	1.05	±0.13	0	±0.09	1.03	±0.14	0.01	±0.04	0.06	±0.06	0.03	±0.07	0.04	±0.04	0.01	±0.03
Cu	0	±0.01	0.01	±0.01	0.02	±0.02	0.02	±0.01	0	±0.02	0.01	±0.02	0.03	±0.03	1.34	±0.09	0	±0.01	0	±0.02	0	±0.02	0	±0.01	0.01	±0.01
Ca^{2+}	0.03	±1.09	0.06	±1.14	0.03	±1.55	6.34	±1.58	5.84	±2.28	0.04	±1.34	2.77	±2.38	0.06	±0.29	0.06	±1.74	63.72	±4.37	1.26	±1.86	0.04	±1.17	0.08	±1.46
Mn	0	±0.02	0	±0.03	0.14	±0.05	0.01	±0.02	0.01	±0.04	0.15	±0.04	0	±0.05	0	±0.01	0.2	±0.04	0.49	±0.06	0	±0.03	0.02	±0.02	0.02	±0.03
I	0.1	±0.02	0	±0.02	0	±0.01	0.09	±0.02	0.12	±0.03	0	±0.02	0	±0.02	0.03	±0.01	0.35	±0.03	0	±0.01	0.03	±0.03	0	±0.01	0	±0.02
Al	0.07	±1.29	0.06	±1.76	0.12	±2.23	0.05	±1.14	0.03	±1.07	1.22	±1.81	0.03	±1.38	0.04	±0.35	24.51	±2.95	76.98	±4.96	0.03	±1.43	0.67	±1.71	0.05	±1.42
MSA	0	±0.07	0	±0.09	0	±0.07	0.01	±0.08	0.01	±0.12	0	±0.04	0	±0.10	0.01	±0.03	0.01	±0.12	0.01	±0.15	0	±0.09	8.02	±0.36	0.01	±0.06
Ni	0.01	±0.01	0	±0.00	0	±0.01	0.01	±0.00	0	±0.01	0.02	±0.01	0.08	±0.01	0.01	±0.00	0	±0.00	0.01	±0.01	0	±0.01	0	±0.00	0	±0.00
EBC	3.47	±1.72	3.14	±1.90	0.1	±2.06	0.19	±1.75	0.58	±1.49	2.96	±2.76	0.08	±2.29	0.37	±0.39	0.15	±1.02	0.07	±1.18	37.81	±3.19	0.03	±0.90	4.78	±1.55
Ozone	0.08	±0.87	0.01	±0.60	2.07	±1.31	0.01	±0.53	0.01	±0.51	0.06	±1.14	0.05	±1.57	0.03	±0.35	0.04	±0.65	1.68	±1.03	5.49	±1.44	0.06	±0.95	18.62	±1.53
Mercury	0.05	±0.06	0	±0.06	0.01	±0.08	0	±0.04	0.07	±0.08	0	±0.06	0	±0.08	0.01	±0.02	0.01	±0.05	0.09	±0.06	0.18	±0.08	0.04	±0.06	0.75	±0.09

Estimated factor loadings and standard deviations

Table 4.2: Factor loadings of light period

By introducing Ni into the analysis we see that a new factor has been formed together with V, called OIL COMBUSTION, this factor is present both in the dark and light period. Whereas V was mostly associated with LEAD ANTHRO (ANTHRO-S in the previous analysis) it is mainly loaded on OIL COMBUSTION in this analysis. With the new factor, V now has less unexplained variance in both Table 4.3 (6.1%) and 4.4 (4%) compared to 11% in Sirois and Barrie (1999). In Lee et al. (1999) a PMF analysis of pollution in Hong Kong was made with similar chemical elements/compounds where Ni and V were included and together formed their own component linked to Oil burning.

When looking at Table 4.1 and 4.2 we also notice that OIL COMBUSTION during the light period is only loaded with V, Ni and some Pb. Compared to the dark where there are small portions of mostly soil associated metals such as Al, Mn and Ca^{2+} but also some Br^-, SO_4^{2+} and NH_4^+ loaded on the factor.

Another factor that is present during both the dark and light as well as in the last analysis is MSA, originating from the atmospheric oxidation of the marine biogenic gas dimethyl sulphide. However, the loadings are much lower in the dark than the light due to less ocean productivity in winter than in spring. During the dark period O_3 and Hg are loaded on this factor (Modified MSA) and the amount loaded is also much less than in the light period. This can also be seen in the seasonality plot in Figure 3.3r where MSA is low during the dark part of the year and as soon as the sun rises it increases with a peak in the beginning of June.

SOIL is another important source of aerosol constituents, in all analyses it is associated with Al, Ca^+ and Mn. In the previous report V was also loaded on the factor but as we could see it is now mostly explained by the OIL COMBUSTION. Although V and Al are connected to each other in both analyses; during the light period V and Al are both loaded on IODINE-SOIL, while during the dark period they're both loaded on LEAD ANTHRO.

Another constituent connected with Aluminum is IODINE. In the previous analysis there is a

small loading of Al on IODINE. During the light period it is associated with Al as well as some other constituents so it is named IODINE-SOIL while during the dark period IODINE is mostly loaded with I, except for some very small loadings of Br^- and EBC. If we look at Table 4.3 and 4.4 we can see that during the dark winter I is most noticeably explained by one factor while during the spring light period it is more spread out.

	BROMIDE	ACID PHOTO-S	SEA-SALT	LEAD ANTHRO	MODIFIED SEA-SALT	OIL COMBUSTION	SMELTER	ZINC	IODIDE	SOIL	BLACK CARBON	MODIFIED MSA	Unexplained variance
SO_4^{2-}	9.1%	14.8%	1.2%	13.1%	30.7%	8.9%	0.2%	0%	1.7%	0.3%	3%	5.5%	11.5%
H^+	1.1%	84.4%	0.7%	0.2%	1.1%	0.6%	1%	2.7%	0.8%	1.6%	1.1%	1.3%	3.4%
Br^-	70.6%	1.8%	0.7%	7.8%	0.2%	5.1%	0.1%	2.5%	4.8%	0.1%	0.1%	1.9%	4.3%
NH_4^+	0.5%	13.6%	0.3%	24.9%	32%	6.7%	0%	2.4%	2%	0%	0.1%	0.7%	16.8%
NO_3^-	2.8%	0%	8.9%	12%	32.5%	1.4%	0.1%	2%	3.9%	4.3%	17.2%	0.1%	14.7%
Na^+	7.7%	0.8%	46.9%	5.6%	22.4%	0.7%	0.3%	1.5%	1.2%	0.3%	2%	0.7%	9.8%
Cl^-	0.1%	0.1%	81.5%	0.5%	2.5%	0.2%	0.3%	2.8%	0.3%	1.8%	3.3%	0.4%	6.4%
K^+	12.5%	5%	31%	20.2%	13.8%	2.6%	0.1%	1.4%	0%	0%	1.7%	0.3%	11.4%
Pb	17.3%	0%	0%	59.5%	0%	0%	6%	3.9%	0%	4.2%	0%	0%	9%
V	0.2%	0.1%	0%	17.5%	0.1%	67.4%	0.1%	4.4%	0%	0%	0.1%	4%	6.1%
Zn	7.2%	1%	0.3%	16.6%	0.1%	6%	17.5%	45.9%	0.2%	0.6%	0.5%	0.6%	3.7%
Cu	4.3%	0.3%	3.5%	13.6%	3.2%	0.4%	69.1%	2.1%	0.4%	0.9%	0.5%	0.4%	1.3%
Ca^{2+}	5.2%	0.1%	12.9%	0%	0%	3%	0%	3%	2.7%	58.4%	0.1%	0%	14.6%
Mn	0%	0%	0%	30.3%	0%	8.5%	1.9%	3.5%	0%	42.6%	0%	0.7%	12.4%
I	0.1%	0.5%	0.5%	11.2%	0.4%	5.8%	3.1%	2.5%	71.2%	0.7%	1.2%	1.6%	1.3%
Al	0%	0.4%	0%	0%	0%	10.2%	2.4%	0%	0.6%	75.8%	0%	0%	10.4%
MSA	0%	7.8%	0.1%	0%	0.1%	0.1%	0.1%	2.4%	0%	0%	0%	85%	4.4%
Ni	2.9%	2.5%	0.1%	21.5%	8.2%	19.4%	0.3%	8.2%	0.7%	12.6%	4.7%	0%	19%
EBC	4.4%	12.5%	2%	0.7%	0.5%	4.6%	0.2%	4.9%	0.3%	0.2%	59.5%	4.7%	5.5%
Ozone	0.1%	8.7%	0.2%	0.1%	34.2%	0%	0.8%	3.5%	5.6%	5.8%	6.3%	21.6%	13%
Mercury	0.8%	9.8%	0.4%	0.1%	33.5%	0.1%	0.8%	4.4%	4.7%	6.5%	6.1%	20.4%	12.5%

Total Variance of each Aerosol constituent explained by the factors

Table 4.3: Explained variance- Dark period

The nineteen aerosols constituents and the two gases are well explained by the model with all having less than 20% unexplained variance. Overall it seems that the model for the light period has less unexplained variance for most concentrations, a small part is explained by the extra factor. Probably the largest part is due to the fact that when the sun rises most aerosols and especially O_3 and Hg have a much clearer trend than during the dark period. An example of this is O_3 and Hg which during the dark period is spread out among several factors but during the light is collected into one. This reflects the strong photochemistry and simultaneous depletion of ozone and mercury in the light rather than the dark period. Even if we increase the number of factors during the dark period they do not collect themselves as a physically realistic factor. The component which has the most unexplained variance is Ni. It might be associated with other metals not included in the analysis or be related to analytical uncertainty. Nevertheless, as we could see it helps with the explanation of V through OIL COMBUSTION.

	ACID PHOTO-S	BROMIDE	MOD. SEA-SALT	SEA-SALT	NITRATE	ZINC	OIL COMBUSTION	SMELTER	IODIDE-SOIL	SOIL	MODIFIED BLACK CARBON	MSA	PHOTO-O3-HG	Unexplained variance
SO_4^{2-}	16.8%	9.6%	23.3%	6.6%	4.1%	6.7%	6.4%	0.7%	0%	0.1%	6.3%	8.6%	0%	10.6%
H^+	82.3%	0.2%	0.3%	0.2%	0.4%	0.1%	0.1%	0.1%	0.7%	0.1%	0.1%	0.6%	10.6%	4.1%
Br^-	10%	67.2%	0%	0.2%	2.6%	0.2%	1.3%	0.1%	4.3%	0%	0.1%	0%	12.1%	2%
NH_4^+	14.5%	0.2%	28.8%	0.1%	15.9%	8.8%	9%	0.1%	0%	0.9%	0.1%	2.2%	7.2%	12.3%
NO_3^-	0.3%	0.4%	15.5%	8.9%	45%	0.4%	0.1%	1.1%	0.1%	6.9%	10.8%	3.7%	0%	6.9%
Na^+	0.4%	6.8%	30.5%	36.4%	0%	0.2%	0.4%	1.4%	3.8%	0.1%	8.3%	0.2%	0%	11.4%
Cl^-	0.1%	0.2%	0.1%	74.6%	0.8%	0.2%	0.1%	2.3%	0.5%	2%	0.1%	0.1%	11.8%	7.1%
K^+	7.3%	3.6%	26.3%	22.2%	2.3%	9.3%	5.9%	0%	0.2%	0.2%	12.5%	0.1%	0.7%	9.4%
Pb	10.9%	0.1%	12.8%	2.4%	0.1%	33.7%	10.5%	5.1%	0.1%	0.1%	12.8%	0.2%	0.9%	10.5%
V	0.4%	9.2%	0.1%	0.1%	0.1%	0.9%	55.8%	1.4%	12.9%	0%	6.3%	0.1%	5.3%	7.3%
Zn	0.2%	2.5%	0.2%	0.4%	15.2%	39.4%	0.2%	19.5%	1%	3.9%	1.6%	3.4%	0.5%	12.1%
Cu	1.5%	1.8%	5.5%	3.8%	1.2%	1.4%	7.7%	67.2%	1.3%	1.2%	0.8%	0.8%	2.5%	3.3%
Ca^{2+}	0.1%	0.1%	0.1%	10.7%	11.6%	0.1%	4.4%	0.1%	0.1%	59.9%	2.7%	0.1%	0.2%	9.9%
Mn	0.2%	0.5%	14.3%	0.8%	1.2%	15.3%	0.2%	0%	17.6%	33.9%	0.6%	1.8%	2.6%	11%
I	15.5%	0.1%	0%	11.5%	18%	0.1%	0.2%	3.9%	37.2%	0.1%	4.6%	0.1%	0.3%	8.4%
Al	0.2%	0.1%	0.2%	0.1%	0.1%	2.3%	0%	0.1%	26.7%	61.8%	0.1%	1.2%	0.1%	7%
MSA	0.1%	0.1%	0.1%	0.2%	0.2%	0%	0.1%	0.4%	0.3%	0.1%	0.1%	96.6%	0.3%	1.3%
Ni	7%	0.6%	0.7%	7.9%	4.1%	10.3%	39.7%	2.9%	2.9%	7.9%	0.4%	3.8%	0%	11.9%
EBC	6.6%	6.9%	0.2%	0.4%	1.7%	5.3%	0.2%	1%	0.5%	0.2%	58.9%	0.1%	13.8%	4.3%
Ozone	0.3%	0.1%	8.5%	0.1%	0%	0.2%	0.2%	0.1%	0.2%	7%	18.8%	0.2%	59%	5.2%
Mercury	3.8%	0.2%	0.6%	0.1%	5.4%	0.1%	0.2%	0.6%	0.9%	7.1%	13.7%	3%	55.2%	9%

Total Variance of each Aerosol constituent explained by the factors

Table 4.4: Explained variance- Light period

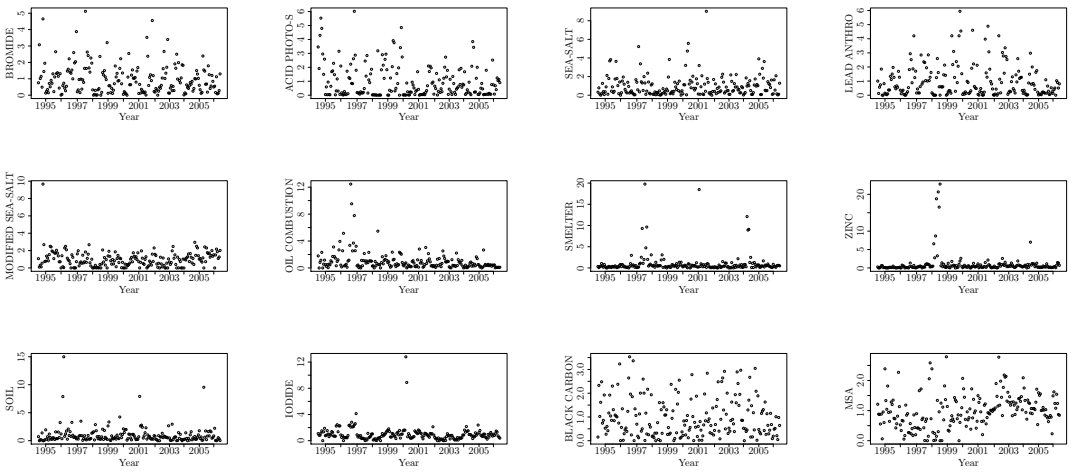


Figure 4.1: Factor Scores, Dark

The metal Zn is highly explained by the two factors ZINC and SMELTER. In Figure 4.1 and 4.2 we can see the factor scores plotted. During the dark, the two factors both share the same peaks while during the light, ZINC doesn't show the same peak as SMELTER does.

When comparing the factor score plots for the dark/light analysis with the factor scores plotted in Figure 7 and 8 in Sirois and Barrie (1999), the scores overall have a lot less seasonality due to the division of dark and light. Long term trends are present in many of the factors. Since a

number of factors are dominated by a single constituent such as MSA, ZINC and BROMIDE we could also examine the time series of the individual constituent concentrations.

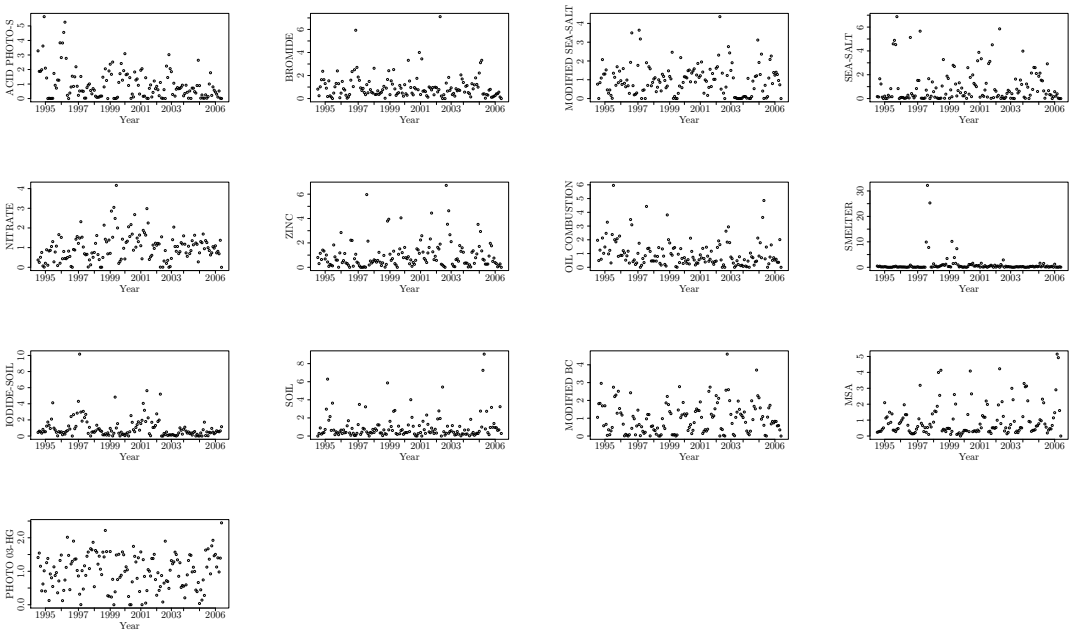


Figure 4.2: Factor Scores, Light

There is a lot to be learned from the factor scores plots, one of the interesting things is the relationship between the factor O3-HG and Br. From the time series analysis as well as the PMF analysis we could see that Ozone and Hg deplete as soon as the sun starts to rise. In Figure 4.3 the factor scores of PHOTO O3-HG are plotted against both the actual values of Bromide and the factor scores of BROMIDE. In a dimension reduction model allowing negative values, the concentrations of O_3, Hg and Br would most likely have formed a single component. In a PMF model this cannot happen since all loadings are positive.

From the plots we can both see the negative correlation of O3-HG and Br but also that Br almost solely is explained by BROMIDE. By Table 4.2 the loading of Br on BROMIDE is 9.81, we can see this in the plot since the factor scores are a bit less than a tenth of the actual values.

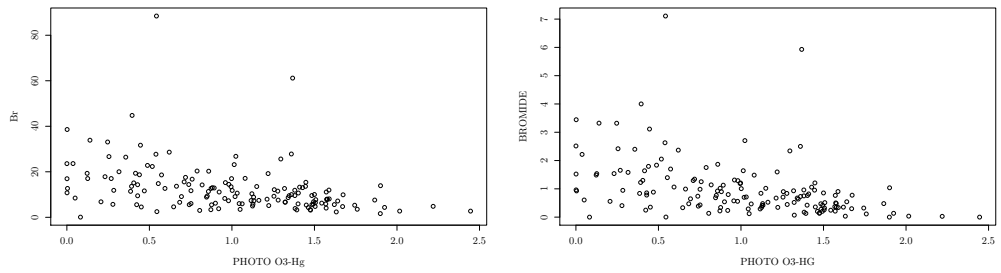


Figure 4.3: Scatter plot showing relationship between Br concentrations, BROMIDE Factor scores and the Photo O3-HG factor score.

In Table 4.5 and 4.6 two correlation tests are made which show a negative correlation between PHOTO O_3-HG and Br. This again confirms the conclusions in Barrie et al. (1988) regarding the negative correlation between ozone and bromide. The correlation is slightly stronger between the two factors than between the PHOTO O_3-HG and the actual values but the difference is so small no conclusions can be drawn by the difference.

	t	cor	df	p.value
Pearson's test	-5.04	-0.38	149	0.00

Table 4.5: Pearson's χ^2 test for correlation between PHOTO O_3-Hg and Br

	t	cor	df	p.value
Pearson's test	-5.69	-0.42	149	0.00

Table 4.6: Pearson's χ^2 test for correlation between PHOTO O_3-Hg and BROMIDE

Chapter 5

Discussion

The primary aim of this analysis was to continue the work done in Sirois and Barrie (1999), where two separate but connected analyses were made. Namely, a time series analysis and a Positive Matrix Factorization (PMF). As we explored data and looked at results from the reproduction of the 1980-1995 analysis in Sirois and Barrie (1999), we decided to change the analyses methods. For the time series analysis the main change was to use smoothing splines instead of polynomial, sine and/or cosine curves. For the PMF analysis we have made two different factorizations, one for data during winter time when the sun never rises and one period with sunlight. The decision of changing the way we estimate the trend curves was based on explaining the curves as physically realistic as possible, without overfitting the curves. The change of the PMF analysis was motivated by looking at the results we gained, which was what we could expect from the last report and the time series analysis. Therefore, a division of the data between the dark winter and light spring time was made which gave interesting results.

The time series analysis shows long term trends for all aerosol constituents and gases except for Na where no significant trend was found. The long term trends of the aerosol constituents and the gases analysed indicate changes in the arctic atmospheric chemical environment. We could see that some concentrations have dropped drastically, some have increased and some have fluctuated throughout the time period. One of the interesting things we found in the analysis was the change point models associated with the fall of the Soviet Union.

In order to correctly interpret the trends we need to have more information about the origin of the aerosols and independently gathered information on arctic haze. The time series analysis made in this thesis can be continued by looking at transport indexes and analyzing the cause of the trends which we only have discussed briefly.

The main motivation for using PMF instead of other dimension reduction techniques, e.g. PCA, is the non-negativity restraint and with that the ability to model physically realistic sources. Using positive matrix factorization has both its advantages and disadvantages. Compared to other dimension reduction techniques it is a very flexible modelling tool which allows us to analyse data with missing and/or BDL values very easily, compared to other analysis techniques.

In the PMF analysis we have both reproduced the analysis from Sirois and Barrie (1999) and factorized data for varying period of time, constituents, different errors etc. Overall, PMF is robust but there are numerous variables and constants that can be changed. The errors are the main change in different models, which can be altered infinitely if the true errors are unknown.

During the work on this data set we learned that an important part is to find a middle way of modelling the errors sufficiently uniformly so that factorizations of different time periods are comparable.

The main disadvantage of using PMF and the PMF2 program is the sometimes lack of theoretical background. Paatero writes

> "*it is essential to stress that the technical details of the [PMF] algorithm should in fact not concern the user of PMF. In the case of PMF it is possible to separate the question 'What is computed? i.e. which mathematical problem is solved by the program?' from the question of how it is computed....The second aspect should only be important for the application if it is suspected that the program does not in fact do what it is supposed to do*" (Paatero, 1997).

It often seems that by Paatero, PMF and especially PMF2 is constructed so that the user should not have to concern themselves with too many details of the computation, e.g. the specification of the algorithm used is sometimes scarce and the different error models are never derived properly.

Positive matrix factorization is a great tool to understand and explain sources without dealing with issues like scaling, since PMF is invariant if different measurements are used from one column (or row) to another. The two most common implementations of the method are the licensed program PMF2 used in this thesis and EPA's (US Environmental Protection Agency) free program PMF 5.0. While PMF2 offers more options with the choice of iterating errors, PMF 5.0 is a far more user friendly since PMF2 is implemented in MS-DOS using an .ini file to control options. Since the algorithm and the iteration process is fairly simple it could easily be implemented as a R or a MATLAB program.

The work on this data set has been interesting and the amount of information that can be gained is rich. The long time span of the data set is quite unique and the conclusions we have come to are only a small part of what can be explored. In the scope of this analysis we have looked at the data set both as a whole and at each separate constituent but there is still much that can be analysed. Unfortunately there are some obstacles with the data which complicates the analysis, such as the changing Over Detection Limit (ODL) and missing values. These problems might have a better solution than the uniform one that has been used in this analysis.

Despite the disadvantages of PMF, there is a huge amount of information we can extract by looking at sources in a receptor model setting using PMF. We have only looked at a small part of what could be analysed, as in the previous report we could look at time series for each factor score plotted in Figure 4.1 and 4.2 and use different combinations of concentrations for different purposes. We could look closer at correlations between different components as with PHOTO O_3-Hg and Br, analyse trends on a smaller scale and look at different dimension reductions.

When using positive matrix factorization we could see that it is essential to know both statistical theory and a considerable knowledge in the data that is to be analysed. Positive matrix factorization is an example of a method which is a compromise between statistical theory and interpretability. During the work on this thesis I have worked closely with Leonard Barrie and Sangeeta Sharma who have been interpreting data from a geoscientific view point. An important part of the work together has been to translate the statistical analysis to physically realistic and understandable results from a geoscience point of view.

Bibliography

J. Bai. Least squares estimation of a shift in linear processes. *Journal of Time Series Analysis*, 15:453–472, 1994.

L. Barrie, J. Bottenheim, R. Schnell, P. Crutzen, and R. Rasmussen. Ozone destruction and photochemical reactions at polar sunrise in the lower arctic atmosphere. *Nature*, 334(6178): 138–141, 1988. ISSN 00280836.

B. A. Bodhaine and E. G. Dutton. A long-term decrease in arctic haze at barrow, alaska. *Geophysical Research Letters*, 20(10):947, 1993. ISSN 00948276.

T. Bollerslev. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31(3):307 – 327, 1986. ISSN 03044076.

P. J. Brockwell and R. A. Davis. *Introduction to time series and forecasting*. Springer, New York, 2. ed. edition, 2002. ISBN 0-387-95351-5.

A. S. Cole and A. Steffen. Trends in long-term gaseous mercury observations in the Arctic and effects of temperature and other atmospheric conditions. *Atmospheric Chemistry & Physics*, 10(10):4661, 2010. ISSN 16807316.

S. Comero, L. Capitani, and B. Gawlik. *Positive Matrix Factorisation (PMF) - An Introduction to the Chemometric Evaluation of Environmental Monitoring Data Using PMF.* , JRC Scientific and Technical Reports, 2009. EUR 23946 EN.

S. Gong and L. Barrie. Trends of heavy metal components in the Arctic aerosols and their relationship to the emissions in the northern hemisphere. *Science of the Total Environment*, 342(Sources, Occurrence, Trends and Pathways of Contaminants in the Arctic Bidleman S.I.): 175 – 183, 2005. ISSN 0048-9697.

P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models : a roughness penalty approach*. Chapman & Hall, London, 1994. ISBN 0-412-30040-0.

T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. Chapman and Hall, London, 1. ed. edition, 1990. ISBN 0-412-34390-8.

R. C. Henry. Current factor analysis models are ill-posed. *Atmospheric Environ.*, 37:23–35, 1987.

P. K. Hopke. *A guide to positive matrix factorization*. Department of Chemistry, Clarkson University, Potsdam, NY 13699-5810, 2003.

S. Huang, K. A. Rahn, and R. Arimoto. Testing and optimizing two factor-analysis techniques on aerosol at Narragansett, Rhode Island. *Atmospheric Environment*, 33:2169 – 2185, 1999. ISSN 1352-2310.

E. Kim and P. K. Hopke. Source apportionment of fine particles in washington, dc, utilizing temperature-resolved carbon fractions. *Journal of the Air & Waste Management Association (Air & Waste Management Association)*, 54(7):773 – 785. ISSN 10962247.

E. Kim and P. K. Hopke. Comparison between sample-species specific uncertainties and estimated uncertainties for the source apportionment of the speciation trends network data. *Atmospheric Environment*, 41:567–575, 2007.

E. Lee, C. K. Chan, and P. Paatero. Application of positive matrix factorization in source apportionment of particulate pollutants in hong kong. *Atmospheric Environment*, 33:3201 – 3212, 1999. ISSN 1352-2310.

P. Paatero. Least square formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems.*, 21:1815–1820, 1997.

P. Paatero. *User's guide for positive matrix factorization programs PMF2 and PMF3, Part2: references*. University of Helsinki, Helsinki, Finland, 2004b.

P. Paatero and U. Tapper. Analysis of different modes of factor analysis as least squares fit problems. *Chemometrics and Intelligent Laboratory System*, 18:183–194, 1993.

P. Paatero and U. Tapper. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Envirometrics*, 5:111–126, 1994.

P. Paatero, P. K. Hopke, X.-H. Song, and Z. Ramadan. Understanding and controlling rotations in factor analytic models. *Chemometrics and Intelligent Laboratory Systems*, 60:253–264, 2002.

A. V. Polissar, P. K. Hopke, P. Paatero, W. C. Malm, and J. F. Sisler. Atmospheric aerosol over Alaska : 2. elemental composition and sources. *Journal of Geophysical Research. Atmospheres*, 103(D15):19045, 1998. ISSN 2169897X.

S. Sharma, D. Lavou, L. Barrie, S. Gong, and H. Chachier. Long-term trends of the black carbon concentrations in the canadian arctic. *Journal of Geophysical Research D: Atmospheres*, 109 (15):D15203 1–10, 2004. ISSN 01480227.

A. Sirois and L. A. Barrie. Arctic lower tropospheric aerosol trends and composition at Alert, Canada: 1980-1995. *Journal of Geophysical Research. Atmospheres*, 104(D9):11599, 1999. ISSN 2169897X.

A. Steffen, T. Douglas, M. Amyot, P. Ariya, K. Aspmo, et al. A synthesis of atmospheric mercury depletion event chemistry in the atmosphere and snow. *Atmospheric Chemistry and Physics*, 8:1445, 2008.

A. Steffen, J. Bottenheim, A. Cole, W. Leaitch, R. Ebinghaus, and G. Lawson. Atmospheric mercury speciation and mercury in snow over time at Alert, Canada. *Atmospheric Chemistry and Physics*, 14(5):2219–2231, 2014. ISSN 16807324.

C. Temme, P. Blanchard, A. Steffen, C. Banic, S. Beauchamp, L. Poissant, R. Tordon, and B. Wiens. Trend, seasonal and multivariate analysis study of total gaseous mercury data from the Canadian atmospheric mercury measurement network (camnet). *Atmospheric Environment*, 41:5423 – 5441, 2007. ISSN 1352-2310.

Y. Zhang, D. Jacob, H. Horowitz, L. Chen, H. Amos, E. Sunderland, D. Krabbenhoft, F. Slemr, and V. St. Louis. Observed decrease in atmospheric mercury explained by global decline in anthropogenic emissions. *Proceedings of the National Academy of Sciences of the United States of America*, 113(3):526–531, 2016. ISSN 10916490.