REPORT SERIES IN AEROSOL SCIENCE

N:o 145 (2013)

DATA CYCLE IN ATMOSPHERIC PHYSICS: FROM DETECTED MILLIVOLTS TO UNDERSTANDING THE ATMOSPHERE

HEIKKI JUNNINEN

Division of Atmospheric Sciences
Department of Physics
Faculty of Science
University of Helsinki
Helsinki, Finland

Academic dissertation

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in auditorium E204 Gustaf Hällströmin katu 2, on 14.1.14 at 12 o'clock

Helsinki 2013

Author: Heikki Junninen
Department of Physics
P.O.Box 64
FI-00014 University of Helsinki
Heikki.Junninen@helsinki.fi

Supervisors: Professor Markku Kulmala
Department of Physics
University of Helsinki

Professor Tuukka Petäjä
Department of Physics
University of Helsinki

Dr. Bo Larsen
Joint Research Center
European Commission

Professor Douglas Worsnop
Department of Physics
University of Helsinki

Reviewers: Docent Teemu Hölttä
Department of Forest Sciences
University of Helsinki

Professor Markus Olin
VTT Technical Research Center of Finland

Opponent: Senior Researcher Üllar Rannik
Institute of Physics
University of Tartu
Estonia

Print: ISBN 978-952-5822-80-9
ISSN 0784-3496
Helsinki 2013, Unigrafia Oy

PDF: ISBN 978-952-5822-81-6
http://ethesis.helsinki.fi
Helsinki 2013, Helsingin yliopiston verkkojulkaisut

Acknowledgements

The research presented in this work was carried out at the Department of Environmental Sciences in University of Kuopio, at the Joint Research Center of European Commission, Ispra, Italy, and at the Department of Physics of University of Helsinki. I would like to thank department heads of all three institutes for providing the facilities for carrying out the research Prof. Juhani Ruuskanen, Dr. Giovanni De Santi, and Prof. Juhani Keinonen.

I wish to thank my main supervisor Prof. Markku Kulmala for providing the tools and resources for my work and for a motivating work environment, also for the push needed to finalize this thesis. I thank Prof Tuukka Petäjä for the help with the thesis and for the years of collaboration.

I'm greatly thankful for Prof. Taisto Raunemaa for introducing me to aerosol sciences and to an artist like attitude for making science. Rest In Peace.

I thank Dr Bo Larsen for making my time in Ispra scientifically meaningful and Dr Covadonga Astorga Llorens for guidance and friendship.

I thank Prof Mikko Kolehmainen for introducing me to Matlab, to a new world of machine learning and for the innovative ideas to analyze environmental data sets.

Dr Douglas Worsnop is acknowledged for showing how a detailed data analysis can be done in even more detail, also for great and innovative discussions all around the globe.

I thank my friends and colleagues Dr Mikeal Ehn for critical ideas and Dr Mikko Sipilä for crazy ideas.

All colleagues at the department are greatly acknowledged for the pleasant working environment and for making us the "state-of-the-art"!

Finally, and most importantly I thank my family – Mom, Dad, brother and sister for love and support. I thank Diana Veersalu for filling my heart with love and my daughters, Helena and Elinor for making purpose for my life.

I acknowledge my limited knowledge by words of one of the first aerosol scientist, John Aitken (1839-1919): "Much, very much, still remains to be done. Like a traveller who has landed in an unknown country, I am conscious my faltering steps have extended but little beyond the starting point. All around extends the unknown, and the distance is closed in by many an Alpine peak, whose slopes will require more vigorous steps than mine to surmount. " (Aitken 1880)

Data cycle in atmospheric physics: From detected millivolts to understanding the atmosphere

Heikki Junninen

University of Helsinki, 2013

Abstract

In this thesis the concept of data cycle is introduced. The concept itself is general and only gets the real content when the field of application is defined. If applied in the field of atmospheric physics the data cycle includes measurements, data acquisition, processing, analysis and interpretation.

The atmosphere is a complex system in which everything is in a constantly moving equilibrium. The scientific community agrees unanimously that it is human activity, which is accelerating the climate change. Nevertheless a complete understanding of the process is still lacking. The biggest uncertainty in our understanding is connected to the role of nano- to micro-scale atmospheric aerosol particles, which are emitted to the atmosphere directly or formed from precursor gases. The latter process has only been discovered recently in the long history of science and links nature's own processes to human activities. The incomplete understanding of atmospheric aerosol formation and the intricacy of the process has motivated scientists to develop novel ways to acquire data, new methods to explore already acquired data, and unprecedented ways to extract information from the examined complex systems – in other words to compete a full data cycle.

Until recently it has been impossible to directly measure the chemical composition of precursor gases and clusters that participate in atmospheric particle formation. However, with the arrival of the so-called atmospheric pressure interface time-of-flight mass spectrometer we are now able to detect atmospheric ions that are taking part in particle formation. The amount of data generated from on-line analysis of atmospheric particle formation with this instrument is vast and requires efficient processing. For this purpose dedicated software was developed and tested in this thesis.

When combining processed data from multiple instruments, the information content is increasing which requires special tools to extract useful information. Source apportionment and data mining techniques were explored as well as utilized to investigate the origin of atmospheric aerosol in urban environments (two case studies: Krakow and Helsinki) and to uncover indirect variables influencing the atmospheric formation of new particles.

Keywords: Atmospheric aerosols, data mining, mass spectrometry, aerosol measurements, factor analysis

Table of Contents

List of publications

This thesis consists of an introductory review and 6 research articles. In the introductory part these papers are cited according to their roman numerals.

I. Junninen, H., M. Ehn, T. Petäjä, L. Luosujärvi, T. Kotiaho, R. Kostiainen, U. Rohner, M. Gonin, K. Fuhrer, M. Kulmala and D. R. Worsnop (2010) "A high-resolution mass spectrometer to measure atmospheric ion composition." *Atmospheric Measurement Techniques* 3 (4) 1039-1053 Doi 10.5194/Amt-3-1039-2010

II. Junninen, H., H. Niska, K. Tuppurainen, J. Ruuskanen and M. Kolehmainen (2004) "Methods for imputation of missing values in air quality data sets." *Atmospheric Environment* 38 (18) 2895-2907 Doi 10.1016/J.Atmosenv.2004.02.026

III. Junninen, H., A. Lauri, P. Keronen, P. Aalto, V. Hiltunen, P. Hari and M. Kulmala (2009) "Smart-SMEAR: on-line data exploration and visualization tool for SMEAR stations." *Boreal Environment Research* 14 (4) 447-457

IV. Junninen, H., J. Monster, M. Rey, J. Cancelinha, K. Douglas, M. Duane, V. Forcina, A. Muller, F. Lagler, L. Marelli, A. Borowiak, J. Niedzialek, B. Paradiz, D. Mira-Salama, J. Jimenez, U. Hansen, C. Astorga, K. Stanczyk, M. Viana, X. Querol, R. M. Duvall, G. A. Norris, S. Tsakovski, P. Wahlin, J. Horak and B. R. Larsen (2009) "Quantifying the Impact of Residential Heating on the Urban Air Quality in a Typical European Coal Combustion Region." *Environmental Science & Technology* 43 (20) 7964-7970 Doi 10.1021/Es8032082

V. Järvi, L., H. Junninen, A. Karppinen, R. Hillamo, A. Virkkula, T. Mäkelä, T. Pakkanen and M. Kulmala (2008) "Temporal variations in black carbon concentrations with different time scales in Helsinki during 1996-2005." *Atmospheric Chemistry and Physics* 8 (4) 1017-1027

VI. Hyvönen, S., H. Junninen, L. Laakso, M. Dal Maso, T. Grönholm, B. Bonn, P. Keronen, P. Aalto, V. Hiltunen, T. Pohja, S. Launiainen, P. Hari, H. Mannila and M. Kulmala (2005) "A look at aerosol formation using data mining techniques." *Atmospheric Chemistry and Physics* 5 3345-3356

Author's contributions

I am solely responsible for the summary of this thesis. In **Paper I** I participated in the experimental planning and did considerable work in the laboratory and ambient measurements as well as in programming the data analysis software. The methods part of the article is mainly written by me. In **Paper II** in addition to the planning the experiment, I am mainly responsible in experimental planning and wrote most part of the original software and for writing a big part of the paper. In **Paper III** I built the web interface to database and extended database with air mass back trajectory, emission register (EPER) and new particle formation classification. I am mainly responsible for the writing of the paper. In **Paper IV** I designed, coded, and ran the constrained matrix factorization (CMF) modeling. I participated in interpretation of the results as well as in writing and revision of the manuscript. In **Paper V** I guided the design and interpretation of the multilinear regression (MLR) modeling experiment and the interpretation. I participated in writing by commenting the paper. In **Paper VI** I took part in designing the experiment as well as in collecting and pretreating the data. I wrote some parts of the paper and helped interpreting the results.

1. Introduction

The atmosphere of the Earth is a complex system that is a result of the equilibrium between solid earth crust, oceans, biological activities (considering here the humans as a part of biosphere) and space. The gravitational field of the Earth holds the atmospheric gases from escaping to space while the life on the planet, seismic activity, and oceans modify the composition. The vast majority (99%) of the atmosphere consists of N_2, O_2. However, the remaining 1% is responsible for the majority of reactions taking place and for controlling the climate. An extreme example on how nano- to micro-scale substance has a huge impact on the whole planet is the atmospheric particle formation (Mäkelä, et al. (1997), Kulmala, et al. (2013)) followed by the cloud formation (Spracklen, et al. 2008). Clouds in general play a crucial role in total energy balance on the Earth (Carslaw, et al. 2010). The pathway to the new particle and cloud formation is initiated by the formation of only 1-10 million molecules of sulfuric acid per cm^3 (Eisele and Tanner 1993, Nieminen, et al. 2009, Petäjä, et al. 2011, Jokinen, et al. 2012). This number is very small, especially when compared to the total number of molecules in air, which is about 10^{19} molecules/cm^3. Even though sulfuric acid was recognized to be important for fog and cloud formation already 133 years ago by Aitken (Aitken 1880), recent studies have concluded that we have not yet achieved a complete understanding of the process (Sipilä, et al. 2010). H_2SO_4 is a fairly simple gas to measure and is one of the rare cases with a detection limit at the level of sub ppt (Eisele and Tanner 1993, Berresheim, et al. 2000, Petäjä, et al. 2009, Jokinen, et al. 2012). Many other important gases have much higher detection limits. Even a simple molecule like NH_3 is very difficult to measure at the low concentration levels relevant in the atmosphere (von Bobrutzki, et al. 2010). The constant need for more sensitive instrumentation goes in parallel with more complex instrumentation and more data rich applications. This combined with a new trend in atmospheric physics (and chemistry), namely the heavy utilization of advanced mass spectrometers, leads us to a situation where the amount of data generated by a single instrument significantly exceeds the level at which data can be simply plotted or visualized on a computer screen. New instrumentation requires substantial data preprocessing before interpretation can be started.

Modern mass spectrometers are not the only data rich applications in atmospheric physics. Long time operated measurement stations with multiple instruments, such as e.g. the SMEAR stations (Station for Measuring Forest Ecosystem–Atmosphere Relations; (Hari

and Kulmala 2005, Järvi, et al. 2009)) can be considered as a battery of instruments where all the measurements are related to each other and the measured variable is a combination of all or some separate instruments. To keep a measurement station operational over a long period of time is costly and includes a considerable workload. As an example, the most comprehensive of the SMEAR stations, SMEAR II has been working over 17 years now. The data produced is valuable and unique and should be treated accordingly. It is essential to have a good quality control and to run an effective database to guarantee security of and accessibility to the collected data. Todays' SMEAR stations in Finland comprise SMEAR I, Värriö since 1991; SMEAR II, Hyytiälä, since 1994; SMEAR III, Helsinki, since 2004; and SMEAR IV, Kuopio, since 2009. New stations are planned to be installed in Järvselja in Estonia and in China (site still to be decided).

As already mentioned, the atmosphere is a complex system where many components are linked to each other. By the use of comprehensive measurement stations, where a wide variety of parameters are measured at the same time and at the same location over a long period of time, we are able to perform studies that cannot be made otherwise. New particle formation in the atmosphere has been extensively studied during the past 15 years, at ground sites in boreal forests (Mäkelä, et al. 1997, Mäkelä, et al. 2000, Clement, et al. 2001, Boy and Kulmala 2002, Dal Maso, et al. 2005, Junninen, et al. 2008, Hussein, et al. 2009, Paasonen, et al. 2009, Yli-Juuti, et al. 2009, Ehn, et al. 2010, Manninen, et al. 2010, Vakkari, et al. 2011, Manninen, et al. 2013), in coastal regions (O'dowd, et al. 2002, O'dowd, et al. 2002, Vana, et al. 2008, Ehn, et al. 2010, Pikridas, et al. 2010, Pikridas, et al. 2012), at high altitude sites (Kivekas, et al. 2009, Bianchi, et al. 2013), at the savanna (Laakso, et al. 2008, Vakkari, et al. 2011, Hirsikko, et al. 2013, Laakso, et al. 2013) and above ground level by use of aircrafts (Crumeyrolle, et al. 2010, Schobesberger, et al. 2013). Most attention has been lend to observations and deterministic modeling in scenarios where known/hypothetical mechanisms have been tested against measured data (Korhonen, et al. 1997, Kerminen, et al. 2004, Grini, et al. 2005, Boy, et al. 2006, Sihto, et al. 2009, Monahan, et al. 2010, Leppa, et al. 2011, Paasonen, et al. 2012, Zhou, et al. 2013). However, comprehensive datasets can also lay the ground for a different approach, in which the data is explored without an initial hypothesis. Such an approach may be called the Let the Data Talk -approach, or data mining. By this, a full advantage is taken of the complete data set in order to avoid overlooking any variable, just because it "does not fit the theory".

As stated in the title of this thesis the term "data cycle" comprises the steps taken from the

very beginning of data being generated by the instrument(s) over the quality control and preprocessing of the data into a consolidated product to the comprehensive analysis, which aims at atmospherically relevant interpretations leading to an improved understanding of the atmosphere. The thesis is set up so that the whole data cycle is covered. In detail, the objectives are following:

1. Data collecting and preprocessing (Paper I-II). I start with the problem of having complex instrumental data that needs to be converted into a data product that can be used for decision-making. In other words, to get an insight into the chemical composition of atmospheric ions and ion clusters.
2. Data storing and visualization (Paper III). I am taking advantage of the existing SMEAR database, extend it and explore the ways to make it accessible to a broad, multi-disciplinary audience.
3. Data analysis (Paper IV-VI). I evaluate numerous multivariate methods for the extraction of information from data deriving from comprehensive atmospheric measurements. Still, the final goal is not to identify the most appropiate methods in itself, but to understand the underpinning environmental and atmospheric processes for the data, such as the origin of air pollutants (Paper IV and V) and new particle formation in the atmosphere (Paper VI).

2. *Preprocessing of measurement data*

All measured data need some sort of processing before it can be used for interpretation. Sensors hardly ever measure directly the physical quantity of interest. Commonly sensors are sensitive only to some physical property and as a response give out a measurable voltage or current. This electrical output from the sensor needs to be converted to a physically meaningful value. For the conversion one needs to know the response relationship of the sensor output and the physical quantity, the so called calibration curve. By exposing the sensor to a known magnitude of the physical quantity (calibration standard) and measuring the response from the sensor one can obtain a calibration curve. Single sensor devices have only one calibration curve, but more complex instruments might have multiple calibration curves all of which are to be taken in to account before the measured raw signal can be converted into the data representing a physical quantity.

Preprocessing here means the preparation of the measured raw data for any further data analysis. The same processing could also be called post-processing if it is thought as a process needed to perform after the measurements are conducted.

2.1. Processing time-of-flight mass spectrometry data

2.1.1. Instrumentation

One instrument that has a complex preprocessing scheme is atmospheric pressure interface time-of-flight mass spectrometer, APiTOF (Junninen, et al. 2010) (PAPER I). This instrument is equipped with an interface to sample ions in gas samples at atmospheric pressure and from ground electrical potential. The mass spectrometer analyzes the mobility of ions in vacuum, hence measuring the mass of the ions. The instrument is depicted in Figure 1; it consists of an inlet that delivers the gas sample through an orifice (300μm) to four consecutive, differentially pumped chambers. The first three are used for focusing the ions and the last one for measuring the time of flight of the ions. The time of flight is measured by giving an energy pulse that sends the ions to travel in the orthogonal direction of the initial travel path and by detecting the arriving ions at a multi channel plate (MCP detector) (Gonin, et al. 1998). This step is called extraction. The time of flight of the ions is determined as the time difference from the pulse to their detection at the MCP.

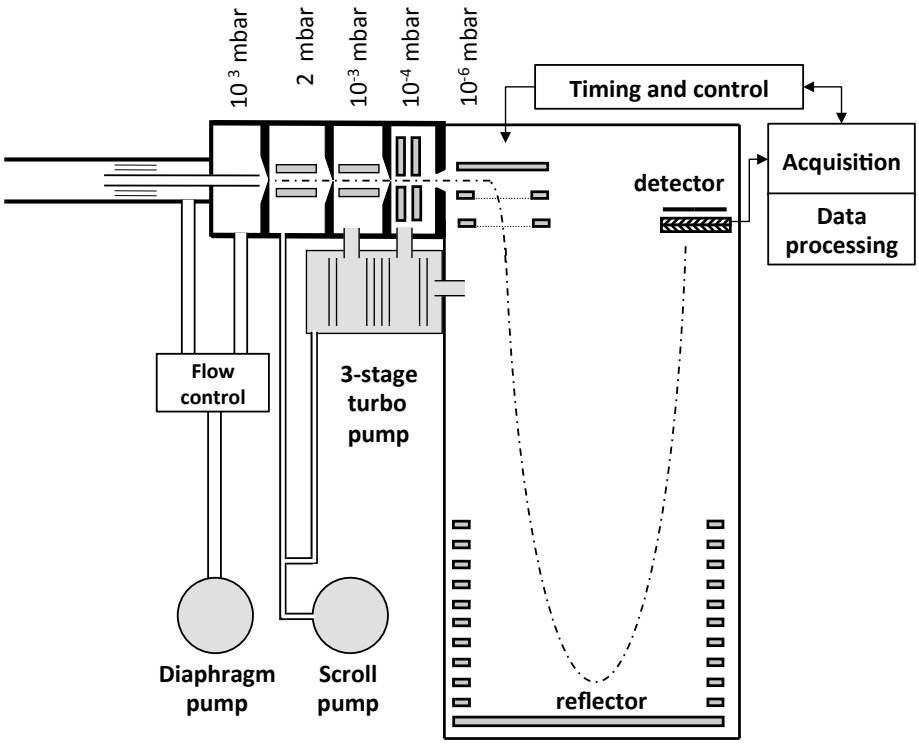

Figure 1. Schematics of APiTOF. The inlet for the gas sample is the opening at the left. Sample flows are regulated by a flow control unit. It is followed by 4 differentially pumped chambers, the first 3 of which are for ion focusing and the last for measuring the time of flight of ions.

The detector of the mass spectrometer is a multi channel plate, which is a kind of electron multiplier where the impact of a charged ions (or particles) generates a pulse of electrons. Every time those electrons collide to the channel walls even more electrons are released. The channels are organized so that the amount of collisions is maximized and high voltage across the plates drives the electrons towards the anode, where the total current is measured. This is the primary measurement of the APiTOF. However, this primary data is not recorded per se, but is processed further on hardware. There are two ways the primary signal can be processed. (1) The simplest is just to record an event of ion collision at its time of arrival. For this the current has to exceed a user defined threshold separating the signal from the electronic noise. (2) The second method is to record the magnitude of the event in addition to the time of arrival. If two ions hit the MCP at the same time twice the current will be generated. For the acquisition of the data from the first method a time-to-digital converter (TDC) can be used and for the second method an analog-to-digital converter (ADC) can be used. With the latter technique the operator first has to make a calibration, by which the signal from a single ion has to be measured. In case of the TDC this calibration is

not needed, since only the time of arrival is stored (magnitude information is lost). In both cases a correct threshold has to be applied.

APiTOF is pulsed at a frequency of 8-12kHz, corresponding in average to 10 000 pulses per second. During every pulse one spectrum is generated, which is called an extraction. For ambient applications not all generated spectra are saved. Instead 1-10 seconds of data is accumulated and then saved to the hard disk. In case of a 10 second accumulation, co-addition of 100 000 spectra will be saved. A spectrum consists of data points that are acquired by acquisition card at fixed sampling rate. The rate has to be set by the operator. Typically the sampling rate (actually sampling interval) is set to 400 pico seconds. In practice this means that if the APiTOF is being pulsed at a frequency of 10kHz, it takes 0.1ms to record one spectrum, if acquisition is done every 400^{th} ps the number of data points per spectrum is 250 000. In other words, a typical ambient air measurement generates 250000 data points every 10 seconds. Thus it goes without saying that such raw data requires preprocessing before attempts of interpretation can be commenced. Algorithms for preprocessing are described and discussed in the following (PAPER I).

2.1.2. Preprocessing software

In this thesis a complete set of preprocessing tools for analysis of time of flight mass spectrometry was developed; the package is called tofTools. The development was done in Matlab (MathWorks 2012). The main reasons for choosing Matlab were the compromise between easy coding and computational efficiency for large matrices, platform independence and easy graphical user interface adaptation. The software, tofTools consists of a set of command line functions that can be called from a script file for automation and batch processing, and a graphical user interface (GUI) for easy and quick visual data analysis. With this tool it is possible to generate a script for batch processing from the GUI. Figure 2 summarizes the entire process and in the following chapters all individual preprocessing steps are explained in detail.

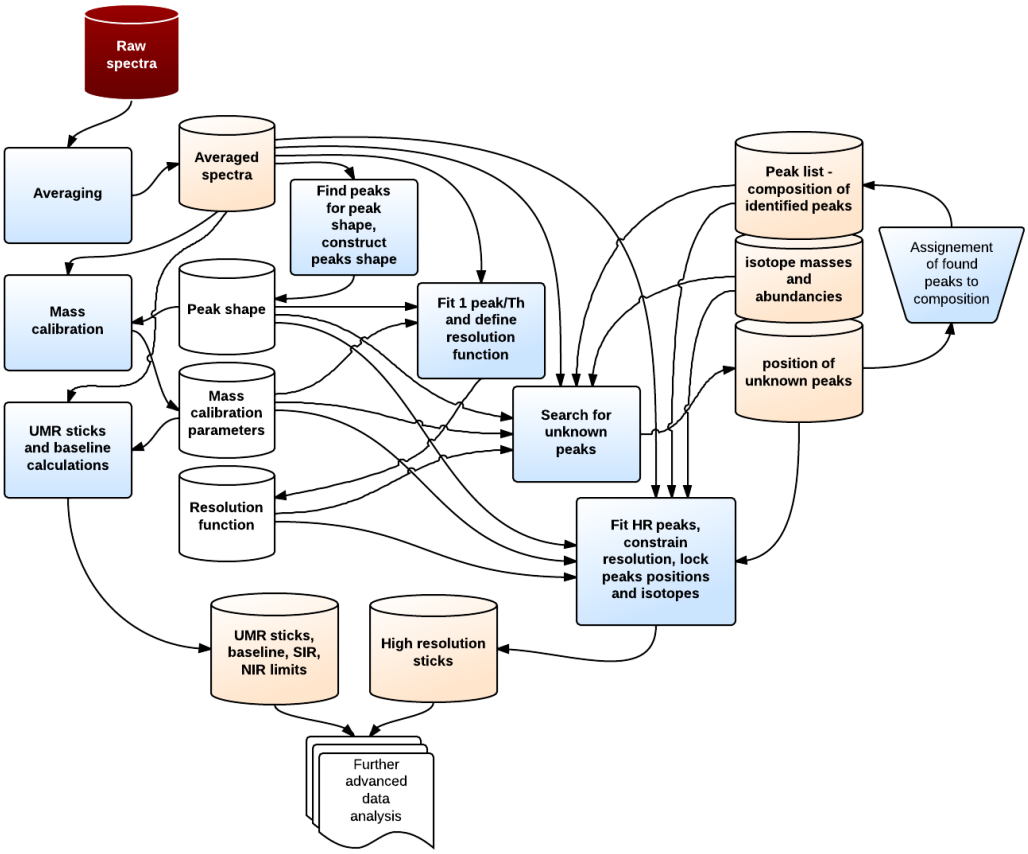

Figure 2. Process schematics of preprocessing of time-of-flight mass spectrometer data by tofTools. UMR – unit mass resolution, SIR – signal integration region, NIR – noise integration region, HR – high resolution

2.1.3. Averaging

Often the sampling is set up so that the acquisition time resolution is higher than actually needed for the working time resolution. This gives the benefit to filter out bad data (bad in the sense of failed, contaminated or interfered measurements) and choose a different working time resolution depending on the purpose. For example if data is acquired at a 5Hz resolution one can calculate fluxes, where very high time resolution is needed, but at the same time the data can be averaged to 10min time series for more traditional concentration data analysis.

The backbone for house keeping of the data preprocessing is an HDF-file (hierarchical data format, (HDF 2000-2010)), the files is appended every time a new preprocessing step is done. Finally, when all preprocessing is done this file contains the final result of the analysis. The HDF data format was chosen because it allows data compression that is transparent for the user and it supports large matrices and partial data access without

reading the whole dataset. This lowers considerably the memory requirements for the software.

Although the step is called averaging the actual mathematical operation that is performed, is summing. After this step the sum of all signals per time-of-flight bin (the earlier example had 250 000 bins, one bin 0.4ns wide) (unit is mV/extraction) is saved together with the number of pulses (extractions) made during the averaging period, the starting time of the averaging period and the single ion signal (mV/ion) for each time step. This information is then used to calculate a signal in units of ions/s.

2.1.4. Instrumental parameters

The mass spectrometer in constant ambient conditions is a fairly stable instrument. However, the performance of the instrument can vary a great deal when ambient conditions are changed (temperature and pressure) or ion-guiding voltages are altered. Some instrumental parameters can be pre-defined for fixed set of voltage settings (so called tuning settings) and measured data corrected in retrospect, such parameters are: peak-shape, transmission and resolution functions. These settings are not sensitive to the above-mentioned change in ambient conditions. However, a radical change in the ambient pressure will change the transmission. This has to be kept in mind when sampling from a high altitude mountain site or conducting air born measurements. Sampling at sea level compared to the sampling at a high mountain site, for example at 3500m, will change the pressure from 1atm to 0.6atm, which has a considerable influence on the transmission. The mass calibration is an instrumental parameter that is sensitive to ambient temperature - not as much to temperature of the sampling gas as to the temperature of the instrument itself.

2.1.4.1. Peak-shape

Often it is assumed in mass spectroscopy that the acquired signals are perfectly symmetric Gaussian peaks. In reality this is rarely the case. For a practical solution many mathematical approximations are available (Di Marco and Bombi 2001), but none of them can cover all the peak shapes. Previously a novel way has been presented of measuring the peak shape from real data and applying this numerical peak-shape-model in the pre-processing phase (DeCarlo, et al. 2006). In PAPER I this idea is developed further by applying a constraint to the peak-shape-model, by which the peak-shape can only be

monotonically increasing and/or decreasing (left side of the peak and right side, respectively).

The peak-shape is obtained by searching for single peaks in a one unit mass window. From each peak we subtract the position of the peak. Now all found peaks are centered at 0. Then all peaks are normalized by the width of the peak (sigma, assuming a Gaussian peak shape), which is determined by measuring the actual width at the half of the maximum signal intensity (*fwhm*) and converting it to sigma (σ) by Eq 1

$$\sigma = \frac{fwhm}{2\sqrt{2\log 2}} \qquad \text{Eq 1}$$

Where σ denotes the Gaussian width and *fwhm* denotes the full width at half maximum.

Additionally, also the signal intensity of all peaks is normalized to unity. In Figure 3 the individual normalized peaks are shown in panel A, and the actual peak-shape function in panel B.

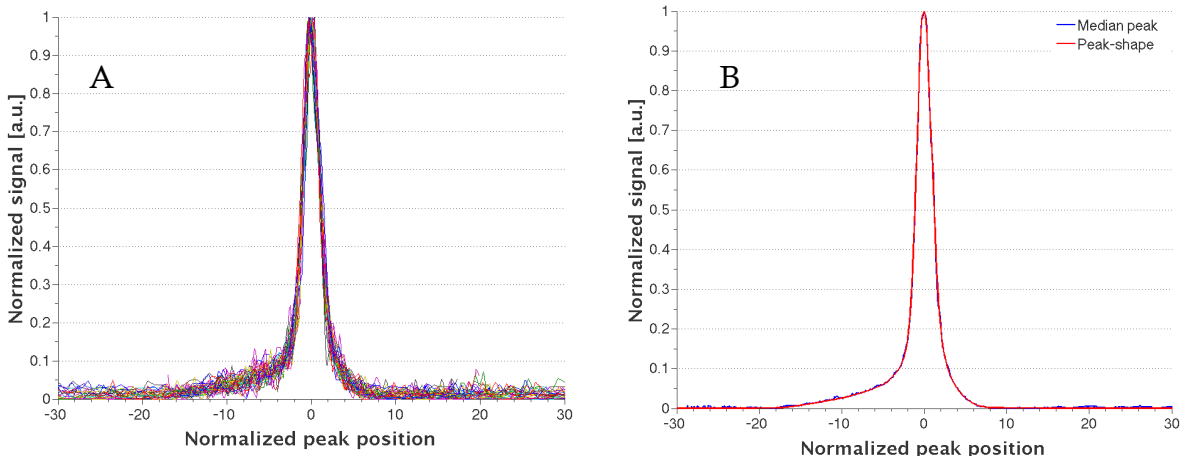

Figure 3. Measuring the Peak-shape. A) 40 found peaks overlaid; each individual peak centrum is removed and is normalize to unit intensity and width. B) Calculated median peak and fitted peak-shape.

The individual normalized peaks are averaged (median) and the monotonic peak-shape-function is fitted. The obtained peak-shape model can take any shape of the measured peaks and improves drastically the peak fitting. The only assumption taken here is that the peaks used for defining the peaks-shape function contain only one chemical compound.

2.1.4.2. Mass calibration

A time-of-flight mass spectrometer measures the time it takes for the ions to travel a fixed distance. A mass calibration is performed in order to relate the measured time to the ion mass (or mass/charge actually). Mass/charge is a quantity used in mass spectrometry and has the SI unit of kg/C, where kg is mass and C is Coulomb. However the use of this unit is not very practical; instead Thomson (Th) is used here. 1Th = 1 Da/e = 1u/e=1.0364×10^{-8} kg/C, where Da and u are Dalton and unified atomic mass units, respectively. 1Da is defined as 1/12 of mass of ^{12}C isotope mass and is 1.660×10^{-27}kg.

The ratio of mass/charge is physically well defined, but more accurate results are obtained when empirical calibration functions are used. Since the calibration functions are empirical there is no clear rule which function to use. The four functions for the mass calibration implemented in tofTools are listed in Table 1.

Table 1. Mass calibration methods implemented in tofTools.

Equations	Name in tofTools	Number of parameters
$M = {}^{m}/_{Q} = \left(\frac{t-b}{a}\right)^{2}$	2 parameter model	2
$M = {}^{m}/_{Q} = \left(\frac{t-b}{a}\right)^{p}$	3 parameter model	3
$M = {}^{m}/_{Q} = p_4 t^{p_1} + p_5 t^{p_2} + p_3$	2 parabola model	5
$M = {}^{m}/_{Q} = p_1 t^2 + p_2 t + p_3$	Quadratic model	3

Where m/Q – mass/charge, $a,b,c,p_1,p_2,p_3,p_4,p_5$ – empirical coefficients, t – flight time of an ion.

2.1.4.3. Resolution function

The resolution function expresses the resolving power of the mass spectrometer over a mass range. The resolving power is defined as:

$$R_m = \frac{m}{\Delta m} \qquad \text{Eq 2}$$

where R_m – resolving power for mass m, Δm – peak width at half maximum of the peak.

For TOFs of the Tofwerk brand – used for the present thesis - the resolving power is about constant over the mass range, except for low masses bellow 200Th (Tofwerk: the instrument manual). The mass dependency of the resolution can be corrected by Eq 3

$$R_m = \frac{m}{am + b} \qquad \text{Eq 3}$$

Where a and b are parameters to be measured by finding a linear dependency between the peak width (FWHM) and the position (mass). In an ideal case each of the peaks in the spectrum is a single component peak (consists of only one molecule) finding the parameters is straightforward. However, often it is not known if the peak has one or more components and the challenge is then to know which peaks are single peaks and which are not. One way to find the single component peaks is to assume that at least some of the peaks are single peaks and by fitting one Gaussian to all of the peaks and plotting the FWHM against mass (Figure 4). By this representation the multicomponent peaks appear wider than the single component peaks at the same mass and single peaks form a lower edge of the data cloud in FWHM-mass space. By finding the lower edge we find the single component peaks and can find the parameters in Eq 3.

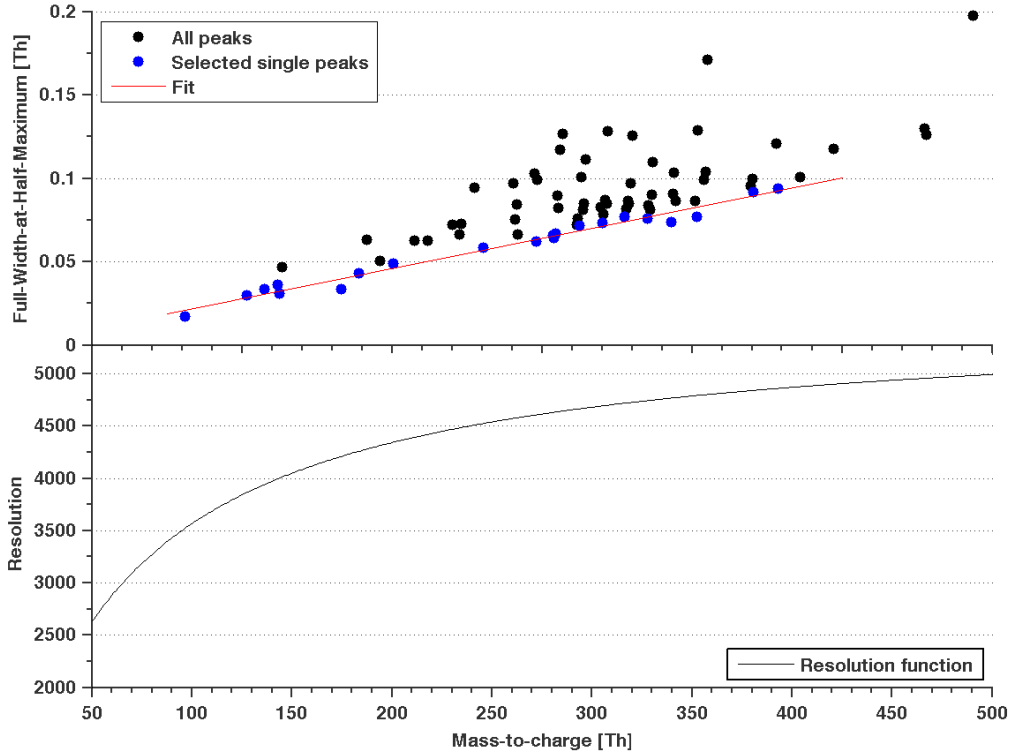

Figure 4. Resolution function. The upper panel shows the peak widths plotted against the mass. The black dots are multicomponent peaks and blue dots are peaks assumed to be single component peaks. The red line represents a linear fit to the data. In the lower panel the same fit is plotted using Eq 3 as the resolution function.

2.1.4.4. Transmission function

In APiTOF not all the ions have the same transmission through the APi-part of the instrument. The ion guides (ion lens and quadrupoles) have a specific mass window that depends on the operating voltages. Overall, the system is not simple to model theoretically and one way to estimate the ambient ion concentrations is to calibrate the instrument for mass dependent transmission. This has to be done for each voltage setting and instrument separately.

In order to define the transmission function the API-TOF is connected in parallel with an electrometer. Both instruments are sampling the same mobility classified sample from a Herrmann –type high resolution differential mobility analyzer (HDMA, (Herrmann, et al. 2000)). The HDMA uses high flow rates (sheath flow rates of up to 2000 l min^{-1} and sample flows of 15 l min^{-1}) and can classify ions with a mobility diameter ranging from 0.8 to 10 nm (Asmi, et al. 2009, Ehn, et al. 2011, Kangasluoma, et al. 2013). The calibration ions are

produced by a tube furnace or an electrospray (see details in (Kangasluoma, et al. 2013)). An example of a transmission function is depicted in Figure 5.

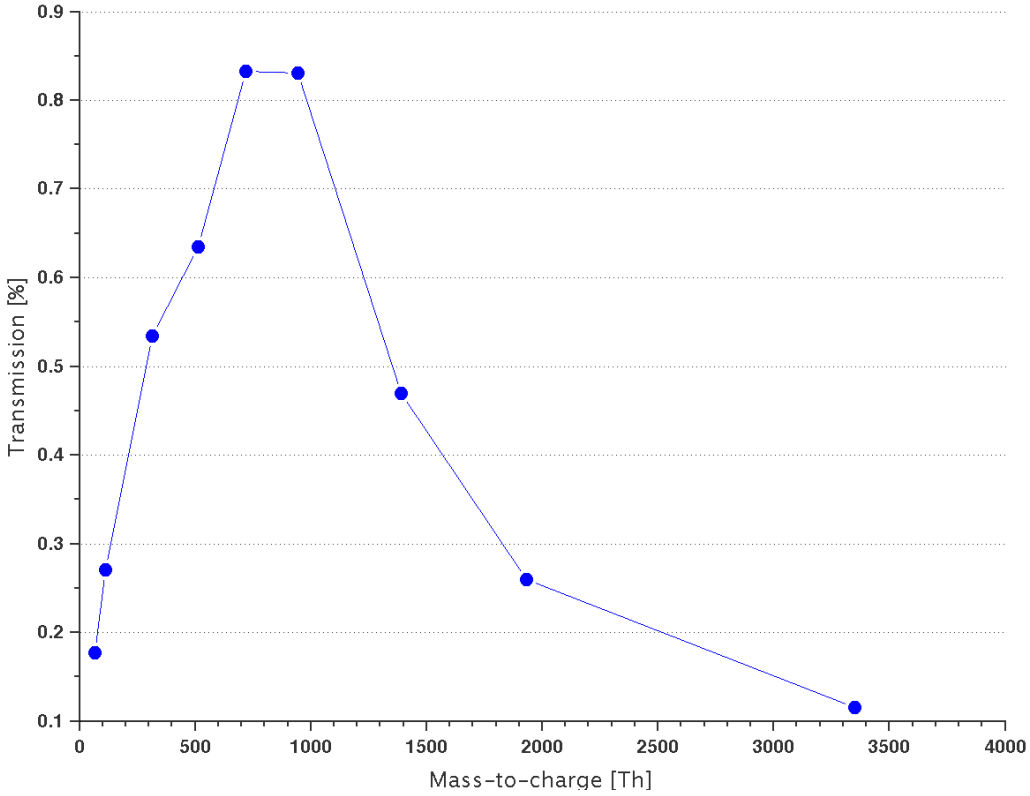

Figure 5. Transmission function of a APiTOF

2.1.5. Data reduction into a stick diagram

2.1.5.1. Unit mass resolution sticks

The raw data that the APiTOF produces can have 250000 data points per averaged time period. The traditional way to reduce the amount of data in mass spectrometry is to combine meaningful parts of the spectrum together. Unit Mass Resolution sticks (UMR-sticks) in tofTools will sum up the entire signal in 1Th window. This 1Th window is divided into the Signal-Integration-Region (SIR) and the Noise-Integration-Region (NIR). Finally the noise corrected data is calculated by subtracting the integrated signal in NIR from the integrated signal in SIR (Figure 6). Traditionally SIR and NIR are defined so that the centrum of the SIR is at the integer mass.

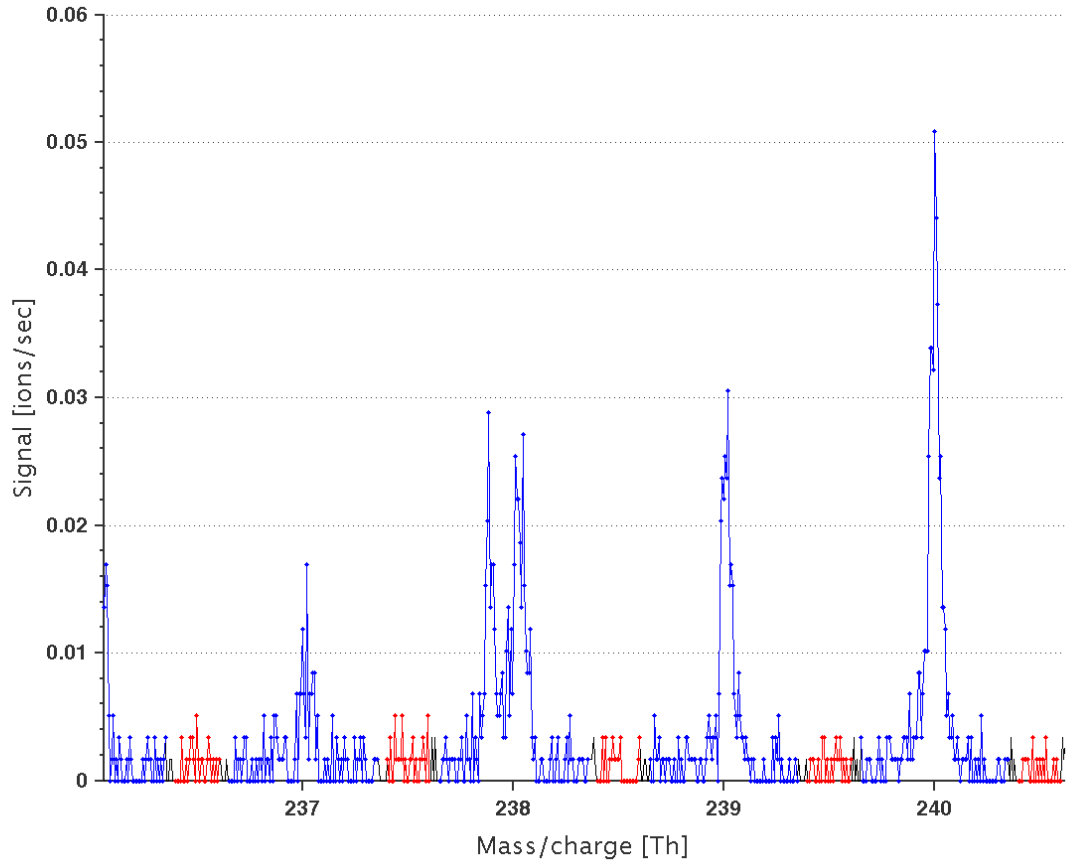

Figure 6. The Signal-Integration-Region (blue) is set to be 0.8Th wide and the Noise-Integration-Region (red) is set 0.2Th wide.

In Figure 6, peaks at 238Th show one potential problem of this kind of UMR sticking. If there are double peaks at a certain mass the chemical information is all summed up and information is blurred.

The second problem with the UMR sticks is the mass defect. Here we define the mass defect as a difference between the nominal mass (number of nucleons) and the exact mass (measured and referenced to the mass of ^{12}C, (NIST isotope database, (Coursey, et al. 2005)). The mass defect is the fraction of a mass of the atoms that is used as a binding energy. E.g. for the carbon atom the exact mass is 12.000Da and when summing up all the nucleons i.e. neutrons, protons and electrons ($6\times1.008\ 664+6\times1.007\ 276+6\times5.485\ 799\times10^{-4}$=12.0989, Mohr et al 2011) the mass difference will be 0.0989Da. It is valid for all of the elements that the mass of the nucleus is less than the combined mass of the individual nucleons, which may become a problem for UMR sticks for large molecules where the mass defect can be so large that the exact mass will be shifted from SIR to NIR. This will result in negative values since the signal will be "measured" from the wrong part of the spectrum. The shift due to

the mass defect can be rectified by applying mass defect correction to SIR and NIR limits, but this only helps when all peaks in a spectrum consists of similar substances e.g. all peaks are organic chemical compounds containing only C, H and O. However, if other elements are also present UMR sticks cannot be used reliably. For example two molecules with the nominal mass of 509Da can have an exact mass of 509.6025Da or 508.6727Da depending on the elemental composition ($C_{36}H_{77}$ and I_3O_8, respectively). If those two molecules are present in the sample spectrum UMR sticks cannot be used. The solution here is to use high resolution sticks.

2.1.5.2. High resolution sticks

The purpose of the high resolution (HR) sticks is the same as the UMR sticks, namely to present a spectrum in a more practical, reduced format. But instead of crudely integrating over a defined range in the spectrum, the HR sticks take a full advantage of pre-defined instrument parameters (mass calibration, peak shape, resolution function), the elemental composition of identified compounds and the isotopic abundances of elements.

HR fitting is most effective if a majority of the peaks are identified and the elemental composition has been assigned. The software (tofTools) calculates automatically all significant isotopes for each peak and uses them in fitting. When using the isotopes in HR fitting the magnitude of the isotopic peaks are fixed to the main peak according to natural abundances. If the peak-shape is defined, also this is used for all the peaks. Otherwise the peaks are assigned the shape of Gaussian function. The width of the peaks is calculated from a predefined resolution function with the only parameter fitted being the area of the peaks. This is obtained by the minimizing norm, Eq 4

$$Min\|\beta\hat{y} - y\| = Min\sum_i (\beta\hat{y} - y)_i \qquad \text{Eq 4}$$

where y is measured spectrum, $\hat{y}$ is fitted spectrum and $\boldsymbol{\beta}$ is the area of the overlapping peaks.

The procedure for fitting multiple overlapping constrained peaks is illustrated in Figure 7.

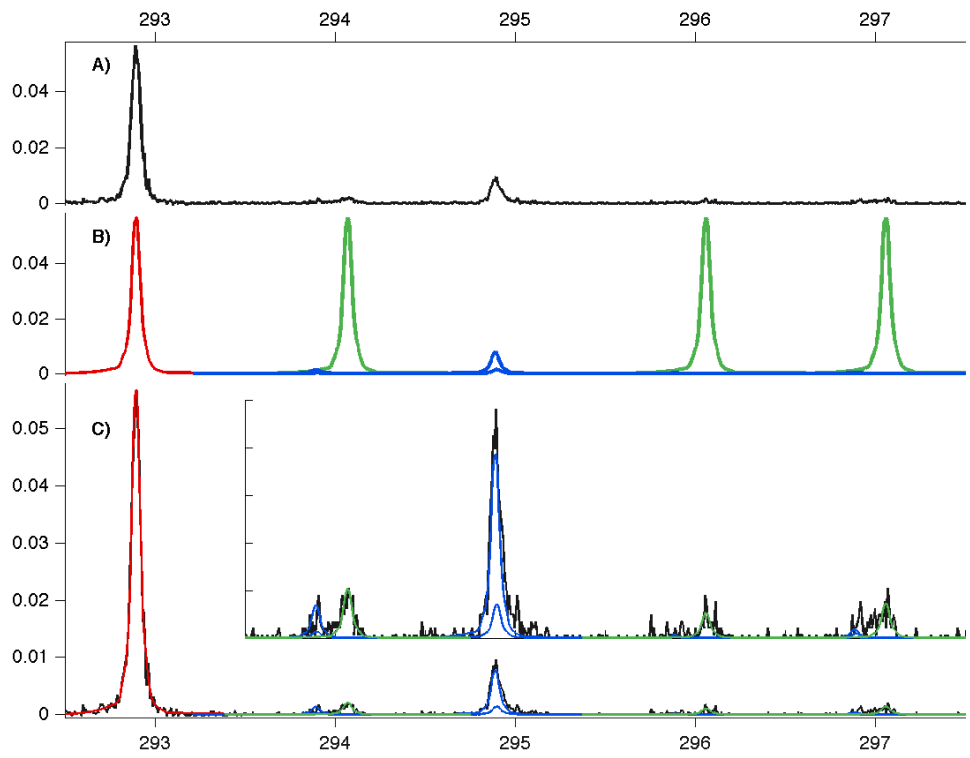

Figure 7. Fitting procedure step by step. A) Averaged spectrum. B) Peak kernels to be fitted to the data. The red peaks have a known composition and the isotopes (blue) associated. The ratio red/blue remains constant throughout the fitting. Green peaks are peaks found by the algorithm. All the peaks to be fitted have a predefined fixed peak shape, a predefined width from the resolution function and a fixed position based on a known composition or as result of the peak search algorithm. C) Peaks fitted to the data.

2.1.6. Finalizing data

When the spectrum is compressed into a stick data matrix the preprocessing of the raw mass spectrometer data is almost finished. Now the unit of the data is ions/s. To convert this to ions/cm^3 we have to apply the transmission function, defined earlier and account for the sample flow to the instrument (in case of APiTOFs with a 300um pinhole it is 0.75l/min). In case of artificial ionization in front of the instrument we have to account for the charging efficiency that is calibrated separately in laboratory (Petäjä, et al. 2009, Jokinen, et al. 2012).

The preprocessing steps of the APiTOF is one of the most complicated ones used in this work, but all instruments have similar steps when converting from raw signal to the final data product.

3. *Missing data imputation*

Generally all the observation datasets have some values that are missing and this issue has to be dealt with somehow. The need for imputing the missing data with some values or not comes from the requirements of the methods that will be used to analyze the data. If the methods are tolerant to missing values no imputation is needed. However, some methods like time series analysis, require continues data, and thus missing values has to be imputed.

A commonly used method is to substitute the missing values with a mean of the variable. This preserves the mean of the variable, but distorts all other statistical measures, like correlations and covariance and considerably affects the inherent structure of data. PAPER II studies the multiple imputation methods and proves the substitution with the mean to be the worst imputation scheme investigated.

The missing data imputation methods can be divided into two groups: univariate and multivariate methods. The univariate methods use only the very same variable that is being fixed, while the multivariate methods utilize all the measured information. In addition we developed (PAPER II) a hybrid method by which time wise small gaps were imputed first with the linear interpolation and afterwards with the multivariate methods. This method improved considerable the quality of imputed missing data. Even better results were obtained when the hybrid method was used with the multiple imputations scheme. Here not only one multivariate method was used but an average of all of them. This improved the robustness of the imputation, but computational cost increased as all the multivariate methods had to be calculated first.

3.1.1. Computational methods

In PAPER II complete data sets (in sense of not having missing data) from Helsinki and Belfast were used for testing and missing data was generated artificially. Different missing data patterns were used to mimic the real world appearance of missing data.

3.1.1.1. Univariate methods; Linear, Spline, Nearest neighbor

Univariate nearest neighbor is a straightforward and robust method for the missing data imputation. With this method no new data values are introduced as only the values present in the data are used. In some applications this is an important feature, however not in atmospheric sciences. In the nearest neighbor method, endpoints of a gap are used to fill in the missing values. In the linear interpolation method, the missing values are estimated

from a linear function between the endpoints of the gap. And finally, in the spline method cubic polynomials are constructed to series of measured data points. A comparison of the performance of the univariate methods as a function of gap length is shown in Figure 9. The critical gap lengths for meteorological and trace gas concentrations when using the linear interpolation in the Helsinki (white bar) and Belfast (black bar) data sets. RH relative humidity, T temperature, WD wind direction, and WS is wind speed (PAPER II).The Y-axis in the figure is an index of agreement (Willmott 1982) that is defined in Eq 5

$$d = 1 - \left[\frac{\sum_{i=1}^{N} (P_i - O_i)^k}{\sum_{i=1}^{N} (|P_i - \bar{O}| + |O_i - \bar{O}|)^k} \right], \qquad \text{Eq 5}$$

where k is either 1 or 2. In this work k set to 2, N is the number of imputations, O_i the observed data point, P_i the imputed data point, $\overline{O}$ is the average of observed data. Two methods, the nearest neighbor and the linear interpolations were performing equally well, but the cubic spline lost the performance very quickly when the gap length was increased.

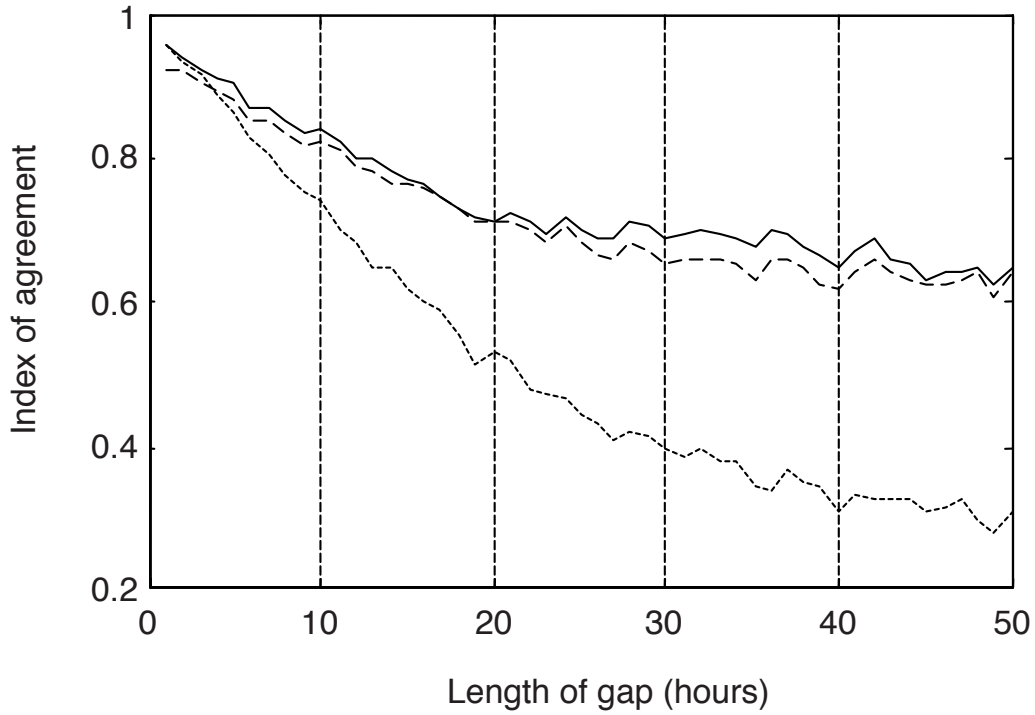

Figure 8. Performance of univariate interpolation methods as a function of gap length. Solid line—linear interpolation, dashed line—nearest neighbour interpolation and dotted line—cubic spline interpolation. (from PAPER II)

3.1.1.2. Multivariate nearest neighbor

The multivariate nearest neighbor method handles a row of N-variables as co-ordinate in an N-dimensional space and calculates Euclidian distances of the row in question to all other rows. Then the closest row is used to fill values for missing data. The distances are weighted down proportionally with the number of missing data in a row, to compensate the lesser reliability of the incomplete rows (Dixon 1979).

3.1.1.3. Self organizing maps, SOM

Self-organizing map (SOM) neural networks (Kohonen 1997) have been widely employed including applications to the atmospheric sciences (Kolehmainen, et al. 2001). The SOM belongs to a group of unsupervised neural networks, which do not need to have the "right answer" during the adaptation process. It learns the structure of the data and adapts itself to that. Later the map can be used to "predict" data. The basic idea of the SOM is to construct a mapping from a high dimensional input space to a low dimensional output space consisting typically of a two-dimensional array of map units.

The map units (neurons) have the same number of weights as there are input dimensions. Each weight is randomly initialized and gets iteratively changed when teaching data is presented to the map. At the end the map units represent the portion of the data that are the most similar to that map unit (best matching unit, BMU). During the adaptation process the missing data is ignored and for missing data imputation the corresponding value from BMU is used (see details in PAPER II).

3.1.1.4. Multilayer perceptron, MLP

The multilayer preceptor is a supervised learning artificial neural network. This is one of the most commonly used and among the most powerful neural network schemes with application in many fields, including atmospheric sciences (Gardner and Dorling 1999). The MLP utilizes feed-forward architecture and neurons are updated by back propagating the error. The network contains multiple layers with n-neurons in each layer, and all the neurons are connected to all the neurons and the weights on each connection are updated iteratively. The number of layers and neurons are design parameters and are normally defined experimentally. In PAPER II we used a 2-layer design with a separate network for each missing data pattern.

3.1.1.5. Regression based imputation

Regression-based imputation methods construct a linear regression model between the missing data and the available data. A separate model is made for each of the variables. Here, the method is based on iterated analysis of linear regression by using the expectation maximization (EM) algorithm (REGEM) (Schneider 2001).

3.1.1.6. Hybrid model in PAPER II

Finally a hybrid method was developed where small gaps were filled first by the linear interpolation method and then the multivariate method was applied. The idea was that the linear interpolation is very reliable for small gaps and if further used in the training set of multivariate methods it improves considerably the overall performance. The critical gap length that was imputed using the linear interpolations depends on the variable and was estimated separately for each of the variables. Figure 9 shows the critical gap length for two data sets (see details on how the critical gap was estimated in paper II). We see big differences between variables. The differences originate from the different atmospheric time scales of the variables.

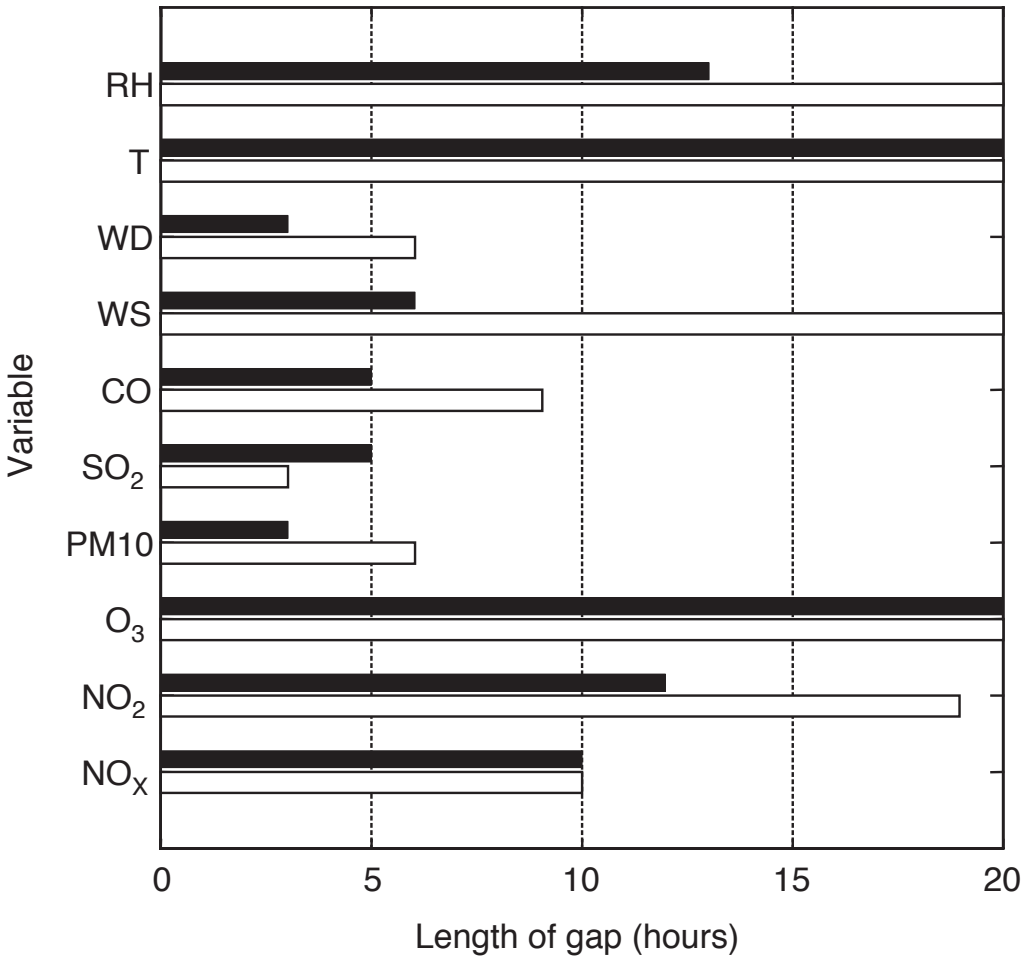

Figure 9. The critical gap lengths for meteorological and trace gas concentrations when using the linear interpolation in the Helsinki (white bar) and Belfast (black bar) data sets. RH relative humidity, T temperature, WD wind direction, and WS is wind speed (PAPER II).

3.1.2. Performance

Missing data imputation is a complicated task and it is preferable to avoid missing data in the first place. However, if missing data has to be imputed a general recommendation is to use the hybrid methods. First the small gaps are filled by linear interpolation and then multivariate methods are used. One of the most reliable and fastest methods is Multivariate Nearest Neighbor. Neural networks perform slightly better, but the usage is much more complicated. Table 2 summaries the pros and cons of different methods tested in this study.

Table 2. Summary for missing data imputation methods (more details see PAPER II).

Method	Performance		Speed
	Short gaps	Long gaps	
Linear interpolation	+ + + + +	+	+ + + + +
Linear regressions (REGEM)	+ + +	+ +	+ + +
Multivariate nearest neighbour (NN)	+ + +	+ + +	+ + + +
Self-organizing map (SOM)	+ + +	+ + +	+ + +
Multi-layer perceptron (MLP)	+ + +	+ + +	+ +
Hybrid methods	+ + + + +	+ + + +	+ + +
Multiple imputations	+ + + + +	+ + + + +	+

+ Poor, + + + + + Best.

4. Saving and presenting measurement data

4.1. Database

After the measured raw data have been converted to physically meaningful quantities, corrected for flawed data values and missing data have been imputed, the data-set is ready for further usage. However, first it has to find its way away from the researcher work computer and has to be saved to a reliable location together with all relevant meta-data and connectors to other data collected from the same measurement site and time. The best media for doing this is a relational database (Codd 1970). In this way other researches get access to the data.

A good example of usage of a relational database as measurement data storage and sharing media is the database for the SMEAR stations ((Hari and Kulmala 2005, Järvi, et al. 2009), PAPER III). In this example the database is build in MySQL (structured query language) language and holds about 170 variables that can be divided into 7 logical blocks: 1) Gases:

NO, NO_x ($NO+NO_2$) SO_2, O_3, H_2O, CO_2, CO for all 6 sampling heights; 2) Meteorology: Temperature, Pressure, RH (relative humidity), Wind speed and direction, precipitation, visibility; 3) Radiation: Global radiation, diffuse global radiation, direct global radiation (global radiation – diffuse global radiation), reflected global radiation, net radiation, PAR (photosynthetically active radiation), reflected PAR, UVA (ultraviolet A) radiation, UVB (ultraviolet B) radiation; 4) Aerosols: number concentration and size-distributions measured by differential mobility particle sizer (3-1000 nm, DMPS), black carbon (7 wavelength aethalometer), optical properties (3 wavelength nephalometer), mass concentration bellow 1um, 2.5um and 10um diameter particles (PM1, PM2.5 and PM10, impactor); 5) New particle formation classification according to (Dal Maso, et al. 2005); 6) Back-trajectories calculated using the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model (Draxier and Hess 1998) for 3 arriving heights and 4 days back. Together with trajectory coordinates also the meteorology along trajectory has been saved to database; 7) Emission point sources from the European Pollutant Emission Register (http://eper.eea.europa.eu/eper/; 10000 facilities).

Preprocessed measurement data is saved every two hours. However, it is not quality controlled and every few month manual quality control is performed for the data.

4.2. Web interface

The database has a user interface that allow quick and convenient data browsing and exploring, called Smart-SMEAR (www.atm.helsinki.fi/smartSMEAR) and is built using php-scripting language, Javascript and graphics package JPGraph - PHP Graph Creating Library (http://www.aditus.nu/jpgraph/).

The value of the tool is its simplicity. The idea is to give a first quick look and overview of the atmospheric composition. Figure 10 presents the main view of the smartSMEAR web site (PAPER III).

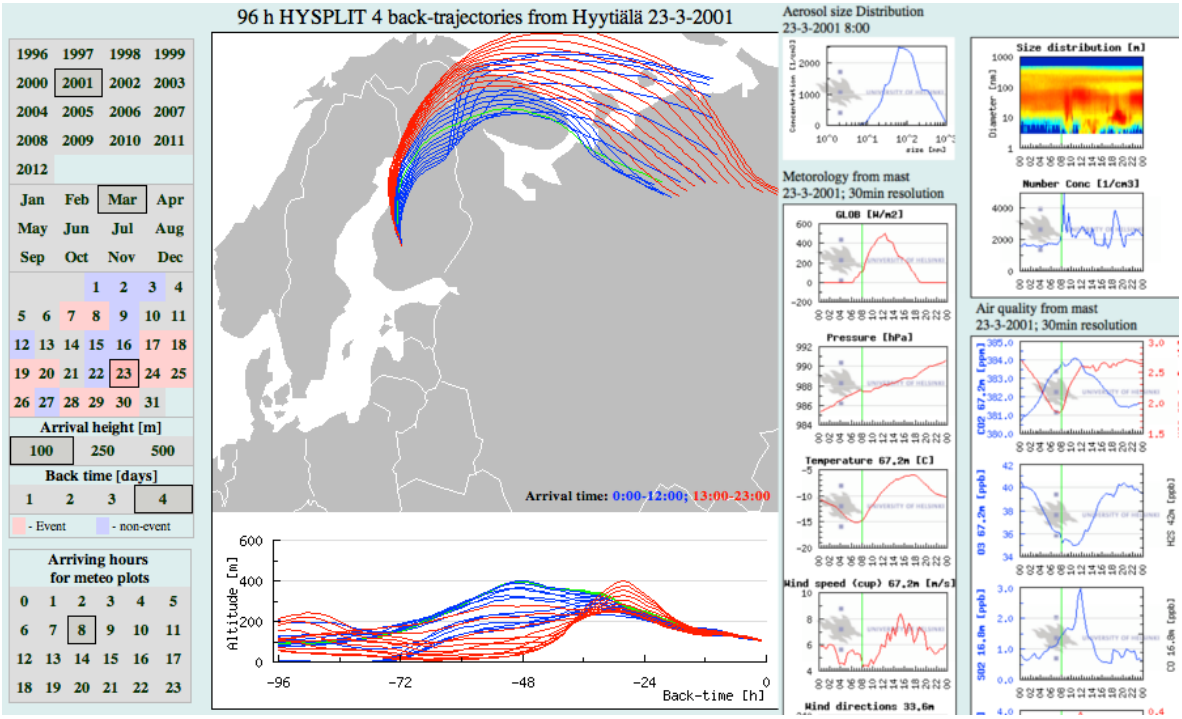

Figure 10. Illustration of the main view of the smarSMEAR. The left hand side is the selection panel, the middle shows back-wards trajectories, and the right hand side depictures aerosol, gas and meteorological data. In the actual web page also meteorology during the trajectory and some more gas concentration plots are presented.

Smart-SMEAR has been used for teaching in university classes and the results have been extremely encouraging for education of student groups with members from a variety of different disciplines. In the courses students can explore complicated set of variables to conclude whether some specific phenomena was occurring. Smart-SMEAR allowed them to skip data processing and plotting and jump directly to visual presentation of the data. After a quick exploration students could concentrate on more details by downloading the data through the interface and conduct further analysis. Smart-SMEAR was given an honorable mention in the University of Helsinki educational technology competition in 2007.

5. *Advanced data analysis by multivariate methods*

The selection of methods for extracting information from data is not trivial and is influenced by many factors. The most important factors in selecting one analysis method over others are: 1) the type of result desired; 2) the skill level of a researcher; and 3) the quality of the data. Data analysis can be divided into two main categories by the type of relationship being studied: dependent or interdependent. The dependent relationship means that the data consists of variables that are dependent on other variables and the model tries to find the statistical dependency and predict the dependent variable(s). In case of interdependent relationship we try to understand the structure of the data by either searching underlying latent variables or communalities, or classifying samples into groups. The schematic in Figure 11 illustrates a path for selecting the appropriate multivariate method for the analysis (adopted from Hair et al 2005).

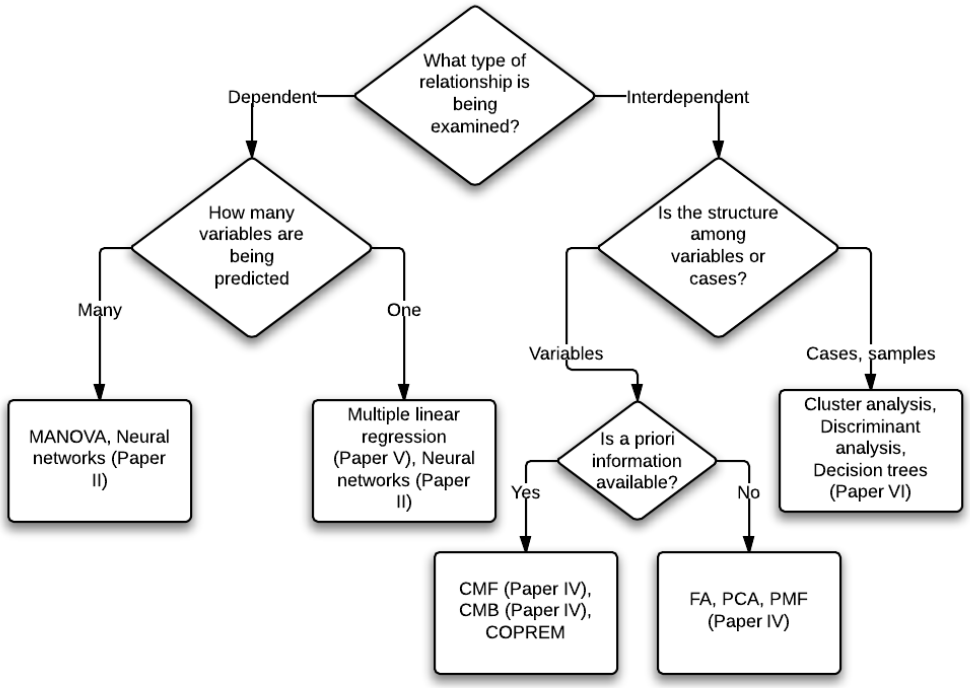

Figure 11. A scheme for selecting the appropriate multivariate method for the analysis. MANOVA – multivariate analysis of variance, CMF – constrained matrix factorization, CMB – chemical mass balance, FA – factor analysis, PCA – principal component analysis, PMF- positive matrix factorization, COPREM – Constrained physical receptor model (Wåhlin 2003).

5.1. Source apportionment

When analyzing data from any single atmospheric measurement station, often the question of air mass origin is raised. Whether the interest is to analyze details of clean or polluted air masses one needs to know the sources and routes affecting the composition of the air masses, in other words, to apportion the sources. Source apportionment is a common term that includes a wide variety of techniques in environmental data analysis where pollution sources are "back calculated" from accumulated data (signals from all sources are mixed) obtained at the receptor site.

When constructing a model for the source apportionment variables are not separated to dependent and interdependent variables, but instead hidden underlying variables are searched. Following the schematic in Figure 11 we are selecting a branch under interdependent relationship and structure between variables, which includes the methods CMB, CMF, and COPREM when a priori information is available and factor analysis, PC, and PMF when a priori information is not available. In this work I used CMB, CMF for source apportionment of PM10 sources from Krakow and Zakopane measurements. These two methods were compared to other multivariate receptor models: EPA-Unmix (Edge analysis), EPA positive matrix factorization, PCA coupled with multi-linear regression analysis (SI of paper IV).

5.1.1. Chemical mass balance, CMF

The Chemical Mass Balance model (Friedlan.Sk 1973) solves a set of linear equations where each receptor chemical concentration is a linear sum of products of source fingerprint profile abundances and source contributions. The model takes the source chemical fingerprint and receptor chemical time series as an input and solves the relative contribution. The model has been used widely in atmospheric studies and is one of the recommended methods for source apportionment by US EPA.

The model has several assumptions and requirements that make the applicability to real world use difficult. The CMB requires that we have accurate chemical fingerprint of all significant sources that the receptor site is exposed to. In addition the CMB assumes that the source fingerprint profiles do not change over time and that the tracers used in the profile do not react with each other and do not change phase by e.g. re-partitioning processes. However, these requirements can be relaxed slightly by requirement for uncertainty estimates for the ambient concentrations and source profiles. In other words, if the

chemical/physical transformation is known, this information can be incorporated to the CMB model through uncertainty estimates (Latella, et al. 2005, Demir and Saral 2011). On the other hand when all the requirements are fulfilled, CMB gives the most accurate source apportionment of all methods.

5.1.2. Positive Matrix Factorization, PMF

As a receptor model, the PMF solves a least squares problem (Paatero and Tapper 1994, Paatero 1997) using atmospheric measurements and the associated uncertainties. The sample matrix is defined as a product of two unknown factor matrices (Eq 6):

$$X=GF+E \qquad \text{Eq 6}$$

Where *X* is the known *n* x *m* matrix of measurements of *n* samples and *m* measurement parameters. The *G* is *n* x *p* matrix of source contributions of *p* sources (factors) and *F* is *p* x *m* matrix of source compositions. E is *n* x *m* matrix of residuals.

Both *G* and *F* are unknown and to be determined by the model including the positivity constraint. The positivity constraint is well justified, if data contains physical parameters where negative values are not defined. This is a way to constrain a purely statistical model by physical a priori information.

The unique model feature in the PMF is the utilization of error estimates Eq 7:

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{e_{ij}}{s_{ij}}\right)^2 \qquad \text{Eq 7}$$

where *Q* is the sum of squares to be minimized, *e* is prediction residual for each data point and *s* is error estimate for each data point.

As seen in Eq 7, each residual of data point is scaled by the corresponding error estimates. If errors are estimated correctly, this scaling greatly enhances the performance of the model. Ideally the error estimates are not just the measurement errors, but also the chemical/physical transformation during the travel from source to receptor site is taken into account. Latella and coworkers have scaled the error estimates by photochemical reactivity when modeling VOC's source apportionment in Bresso, Italy (Latella et al 2005). Since they also had source fingerprint measurements they could conclude that the error scaling by

photo-reactivity improved the PMF model results. Later, also the CMB has been modified to capture the chemical uncertainty (Demir and Saral 2011).

Both the accurate error estimation and the positivity constraints are means to improve the statistical model with physical information about the system in question. Pure statistical factor analysis has a rotational ambiguity and constraining the rotational space by physical knowledge reduces this ambiguity. Other ways the PMF can take advantage of a priori information is to use Paatero's multilinear engine (ME2) algorithm (Hopke, et al. 1999, Paatero 1999, Xie, et al. 1999) or Wåhlin's COPREM algorithm (Wåhlin 2003). Using these algorithms it is possible to build a factorization model where the measured source profiles are utilized in optimization step, kind of hybrid models between CMB and PMF. Here some of the factors are locked completely or partly within confidence limits. This further limits the rotational ambiguity and factorization result is physically more sensible (mathematically all the rotations are equal). The first mentioning of similar hybrid method is the transformation factor analysis (TTFA) (Hopke 1988). Since the level of constraints can be relatively high we called the model Constrained Matrix Factorization, CMF (PAPER IV) (Junninen, et al. 2009).

In paper IV we used CMF and CMB to apportion the PM10 sources in Krakow during wintertime. Table 3 summarizes the modeling results. In general the two models compared well and source contribution estimates (SCE) were within 95% confidence intervals. Both models computed the highest primary contributions to the PM10 mass to be from residential heating, especially in Zakopane. The second highest was industrial emissions and the traffic and dust resuspension were estimated to be only a minor contributions.

Table 3. Source apportionments (+-95% confidence interval) for Krakow and Zakopane sites using two receptor models, CMB and CMF. Unit is ug/m^3. (from paper IV)

		Krakow		Zakopane	
		CMB	**CMF**	**CMB**	**CMF**
home heating	residential coal combustion in small stoves and boilers	38 ± 11[a]	11 ± 5	43 ± 40	16 ± 16
	residential heating (wood, coke, oil)		13 ± 6	46 ± 20	58 ± 31
industrial power generation (coal)	LE-boilers (coal)	16 ± 3	17 ± 3	5.4 ± 3.5	5.5 ± 4.4
	HE-coal combustion	3.5 ± 1.2	13 ± 5	not significant	1.1 ± 0.9
secondary aerosol (inorganics)	sulfates, nitrates and chlorides	16 ± 2	16 ± 2	7.7 ± 3.7	9.4 ± 4.4
traffic and resuspension	vehicles	5.8 ± 2.0	3.7 ± 1.5	not significant	0.5 ± 0.4
	resuspension (incl. road salt)	2.1 ± 0.3	2.0 ± 0.3	not significant	1.2 ± 0.4
	mass coverage	82%	84%	91%	82%
	R^2	0.94	0.96	0.89	0.89

[a] In a large number of CMB runs, profiles for residential heating (coal, wood, coke, oil) resulted collinear and were

5.2. Feature selection by Multiple Linear Regression, MLR

Statistical models can be used to study complicated dataset, e.g. to search for the most correlating set of variables. However, the most correlating does not necessarily mean the true physical causality. The Multiple Linear Regression model, MLR, is a good tool when one variable is predicted by other variables, all the variables are normally distributed and multilinear dependency is expected (Hair, et al. 2005). MLR is defined by Eq 8:

$$\hat{y} = \sum_{i=1}^{n} \beta_i x_i + \beta_0 \qquad \text{Eq 8}$$

where $\hat{y}$ is the predicted variable, β– regression coefficients, β_0– intercept and n – number of variables in the model.

An experimental setup where multiple models are run and the best combinations of independent variables are searched for is called "feature selection" (Hand et al 2001). The MLR is fairly fast so that brute force method can be used by which all combinations of all variables are searched. However, if the number of independent variables is too high or the model is too slow some simplifications may be applied. Two main feature selection options exist, namely step-forward and step-backward selections (Hand et al 2001). The step-forward selection algorithm chooses the variable with the best correlation as a starting point and adds the second variable that produces the best MLR model and so on. The step-backward selection algorithm starts with a model that include all independent variables end gradually leaves out the worst independent variable. Both of the methods evaluate only a subset of all possible variation combinations.

In paper V the MLR was used with a brute force feature selection to find the variables that are the best to explain the black carbon concentration measured in urban SMEAR station (Järvi, et al. 2008). Three variables explained the BC concentration: traffic intensity, wind speed, and mixing height. Addition of an extra variable did not improve the quality of the model. Separate models were run for weekdays and weekends and the effect off traffic was prominent during weekdays.

5.3. Data mining by Discriminant Analysis and Clustering

Data mining is not a statistical method, but instead a selection of methods (Hand et al. 2001, Hastie 2001). The methods are not predefined and complete freedom is left to the researcher. The purpose of data mining is to extract information from data and present it in a more understandable and simpler form.

Data mining tools are used in paper VI to find the best set of variables measured at the SMEAR II station to explain days with events of new particle formation (Dal Maso et al 2005) with altogether 80 variables considered. The dependent variable was a categorical variable so the performance of all models was evaluated by using a misclassification rate, Eq 9:

$$\varepsilon = \frac{N_{missed} + N_{false}}{N_{total}} \qquad \text{Eq 9}$$

where N_{missed} is the number of event days classified as nonevents, N_{false} is the number of nonevent days classified as event days, and N_{total} is the total number of days classified.

This study used a selection of 7 methods to search the best two and the best three variables to separate non-event days from event days. 3 variable models were not significantly better than 2 variable models and the best 2 variable models to separate the nucleation events were RH and log(CS). The performance of all methods with the best variable is listed in the Table 4,

Table 4. Average and 95% confidence limits of 1000 runs. LDA – linear discriminant analysis, SVM – support vector machines, 10-NN –nearest neighbor, LDAQ - LDA with quadratic boundaries.

Method	error rate (%)	false events (%)	missed events (%)
LDA	11.9±0.2	11.7±0.2	12.2±0.3
logistic regression	12.3±0.2	11.3±0.2	13.3±0.3
linear regression	12.2±0.2	14.8±0.2	9.2±0.2
SVM (linear kernel)	11.9±0.2	11.7±0.2	12.0±0.2
10-NN	13.8±0.2	14.6±0.3	12.8±0.3
LDAQ	12.7±0.2	10.6±0.3	15.0±0.3
decision trees	14.2±0.2	6.5±0.2	23.1±0.4

6. *Review of papers*

Paper I describes the first adaptation of a newly developed mass spectrometer to atmospheric studies. First the laboratory tests for transmission, sensitivity and response are described. Next, the data analysis procedures are explained in details. Finally the instrument was deployed to the field stations (SMEAR III) and instrument characterization and the developed data processing tools were put in use for deconvoluting the first measurements of naturally charged ion spectra in Helsinki.

Paper II tackles the problem of missing data. In the paper multiple methods to impute missing data are evaluated and recommendation for the best approach is given. The performance of the investigated methods was found to vary from variable to variable. This information was utilized for the construction of a hybrid method that combines the simple univariate methods and the complex multivariate methods for the best performance.

Paper III describes a tool to explore and save data from a comprehensive measurement station. The tool is called smartSmear and is adapted to SMEAR II data. The paper describes in detail the structure of the database and the web based user interface. It also gives an example in form of a case study to demonstrate the power of the tool.

Paper IV describes an extensive particulate matter source apportionment campaign conducted in Krakow. Particulate matter in the atmosphere was sampled and chemically characterized at receptor sites as well as from potential emission. Several source apportionment models were tested as described in the supplementary material of the paper, but eventually two methods were identified to be the most reliable, namely the Chemical Mass Balance model (CMB) and Constrained Matrix Factorization (CMF). Wintertime residential heating was found to be the most dominant source of PM mass in Krakow city and surrounding areas. The traffic accounted only for 5% of the total mass. At the time of publishing the paper was chosen as news story of the month by the journal "Environmental Science and Technology".

Paper V discusses the temporal variations and sources of black carbon (BC) in the city of Helsinki. Black carbon concentrations in the city were found to correlate very well with traffic intensity, in both the weekly cycle and in the diurnal cycle. However, not all the variance was explained by traffic intensity and the feature selection routine based on the Multilinear Regression model was conducted to find the most important meteorological

parameters influencing the BC concentrations. As the result of the study BC was explained the best by traffic intensity, wind speed and planetary boundary layer height.

Paper VI compares a wide variety of the data exploratory methods to find the variables that could explain the new particle formation observed in Hyytiälä, SMEAR II station. The data used in the study was a result of 8 years of measurements and nearly 80 variables were included. As a summary from multiple methods the new particle formation was best explained by the surface area of pre-excising aerosols (condensation sink) and by the relative humidity.

7. Conclusions

A comprehensively the complete data cycle of atmospheric science was carried out for this thesis, from data collection, processing, analysis and final interpretation. The work concentrated more on data analysis and data treatment methods than on the final interpretation of results, but the motivation of the work is tied tightly to atmospheric physics. We applied a new instrument for the first time, conducting field and laboratory campaigns using an APiTOF mass spectrometer. Data analysis was performed by software (tofTools) developed in this thesis. The software treats all steps required for processing time of flight mass spectrometer data (PAPER I).

During the field campaign at the SMEAR II station we measured and identified the composition of low molecular weight (LMW) ionic species sampled directly from the ambient atmosphere. Negative ions comprised mainly inorganic acids and LMW organic acids. Daytime were dominated by sulfuric acid and its clusters while after sun-set the nitric acid monomer (NO_3^-) and dimer ($HNO_3NO_3^-$) became the most dominate anions. The positive ion spectra showed very little diurnal variation and were dominated by quinoline and pyridine cations (PAPER I, Ehn et al. 2010).

The instrument usage and development is ongoing and the software has been adapted to other mass spectrometry instrument applications (chemical ionization, ion mobility), all of which are built on top of the original APiTOF. Similarly to the instrument itself, the software (tofTools) developed in this thesis has been critical to data analysis in a number of scientific works (Ehn, et al. 2010, Junninen, et al. 2010, Ehn, et al. 2011, Kirkby, et al. 2011, Laitinen, et al. 2011, Lehtipalo, et al. 2011, Manninen, et al. 2011, Ehn, et al. 2012,

Jokinen, et al. 2012, Kulmala, et al. 2012, Kangasluoma, et al. 2013, Keskinen, et al. 2013, Kulmala, et al. 2013, Mohr, et al. 2013).

A database without an appropriate user interface is good only for data backup purposes. With an easily accessible user interface the database becomes a tool for scientific data analysis. The web based tool SmartSMEAR is an easy to use interface designed for the SMEAR database (PAPER III). This together with its search engine Smart-Search, has proven to be an efficient tool for studying atmospheric chemistry and atmospheric aerosol dynamics. It has been used in multiple courses and the data analysis for the paper of Mazon et al 2009 is in great extent done using SmartSMEAR.

Source apportionment is a powerful tool for the study of sources and processes that influence the atmospheric composition at a given receptor site. However, the usability of traditional factor analysis methods is limited due to the rotational ambiguity that makes interpretations very difficult. The Positive Matrix Factorization (PMF) is a significant step towards a physically constrained factorization model. By adding more constraints based on physical knowledge about the studied system of sources and receptors we developed the Constrained Matrix Factorization method, CMF. By complete or partially locking specific known factors, an increased degree of confidence was obtained that the calculated source contributions are indeed correct. Using data from chemical characterization of winter particulate matter samples collected from Krakow and Zakopane, we showed that the majority of the aerosol mass originates from residential heating (especially in Zakopane, a mountain village) and industrial power generation (in Krakow city these sources were equal) whereas traffic was only a minor source of the PM mass (<5%).

Contrary to these PM10 mass apportionment results, particle-bound black carbon in Helsinki clearly originated from traffic emissions. Specific meteorological factors enhanced atmospheric concentrations, namely a low wind speed and a shallow boundary layer.

New particle formation has been extensively studied for 15 years (Kulmala, et al. 2000, Kulmala 2003, Kulmala, et al. 2004, Kulmala, et al. 2007, Kulmala, et al. 2013). In order to extract more insight via full statistical analysis, we applied data mining tools to an eight years dataset with 80 variables in search of the best classifying variables. Multiple methods produced comparable results. We concluded that atmospheric nucleation requires both low air humidity and low condensation sinks (surface area of the pre-existing particles) (Hyvönen, et al. 2005, Hamed, et al. 2007). This is in good agreement with current theoretical knowledge. However, the result was somewhat surprising in the sense that the

data mining tools identified inhibition variables as the most important, not the production-variables (like solar radiation or atmospheric SO2 concentrations (Petäjä, et al. 2009)). Obviously both types of variable can be important, but according to these results clearly the strongest correlation of for lack of new particle formation is the presence of strong inhibition mechanisms, not the lack of production components, but the presence of strong inhibition mechanisms. Only when inhibition is lowered do we observe nucleation of new particles.

References

Aitken, J. (1880) "On dust, Fog and Clouds." *Proceedings of Royal Society of Edinburgh* XI

Asmi, E., M. Sipilä, H. E. Manninen, J. Vanhanen, K. Lehtipalo, S. Gagne, K. Neitola, A. Mirme, S. Mirme, E. Tamm, J. Uin, K. Komsaare, M. Attoui and M. Kulmala (2009) "Results of the first air ion spectrometer calibration and intercomparison workshop." *Atmospheric Chemistry and Physics* 9 (1) 141-154

Berresheim, H., T. Elste, C. Plass-Dulmer, F. L. Eisele and D. J. Tanner (2000) "Chemical ionization mass spectrometer for long-term measurements of atmospheric OH and H2SO4." *International Journal of Mass Spectrometry* 202 (1-3) 91-109

Bianchi, F., H. Junninen, J. Trostl, J. Duplissy, L. Rondo, M. Simon, A. Kurten, A. Adamov, J. Curtius, J. Dommen, E. Weingartner, D. R. Worsnop, M. Kulmala and U. Baltensperger (2013) "Particle nucleation events at the high alpine station Jungfraujoch." *Nucleation and Atmospheric Aerosols* 1527 222-225

Boy, M. and M. Kulmala (2002) "Nucleation events in the continental boundary layer: Influence of physical and meteorological parameters." *Atmospheric Chemistry and Physics* 2 1-16

Boy, M., O. Hellmuth, H. Korhonen, E. D. Nilsson, D. ReVelle, A. Turnipseed, F. Arnold and M. Kulmala (2006) "MALTE - model to predict new aerosol formation in the lower troposphere." *Atmospheric Chemistry and Physics* 6 4499-4517

Carslaw, K. S., O. Boucher, D. V. Spracklen, G. W. Mann, J. G. L. Rae, S. Woodward and M. Kulmala (2010) "A review of natural aerosol interactions and feedbacks within the Earth system." *Atmospheric Chemistry and Physics* 10 (4) 1701-1737

Clement, C. F., L. Pirjola, M. dal Maso, J. M. Mäkelä and M. Kulmala (2001) "Analysis of particle formation bursts observed in Finland." *Journal of Aerosol Science* 32 (2) 217-236

Codd (1970) "A relational model of data for large shared data banks." *Commun. ACM* 13 (6) 377-387

Coursey, J. S., D. J. Schwab and R. A. Dragoset Atomic Weights and Isotopic Compositions. National Institute of Standards and Technology, NIST (http://physics.nist.gov/Comp) 2005

Crumeyrolle, S., H. E. Manninen, K. Sellegri, G. Roberts, L. Gomes, M. Kulmala, R. Weigel, P. Laj and A. Schwarzenboeck (2010) "New particle formation events measured on board the ATR-42 aircraft during the EUCAARI campaign." *Atmospheric Chemistry and Physics* 10 (14) 6721-6735

Dal Maso, M., M. Kulmala, I. Riipinen, R. Wagner, T. Hussein, P. P. Aalto and K. E. J. Lehtinen (2005) "Formation and growth of fresh atmospheric aerosols: eight years of

aerosol size distribution data from SMEAR II, Hyytiala, Finland." *Boreal Environment Research* 10 (5) 323-336

DeCarlo, P. F., J. R. Kimmel, A. Trimborn, M. J. Northway, J. T. Jayne, A. C. Aiken, M. Gonin, K. Fuhrer, T. Horvath, K. S. Docherty, D. R. Worsnop and J. L. Jimenez (2006) "Field-deployable, high-resolution, time-of-flight aerosol mass spectrometer." *Analytical Chemistry* 78 (24) 8281-8289

Demir, S. and A. Saral (2011) "A New Modification to the Chemical Mass Balance Receptor Model for Volatile Organic Compound Source Apportionment." *Clean-Soil Air Water* 39 (10) 891-899

Di Marco, V. B. and G. G. Bombi (2001) "Mathematical functions for the representation of chromatographic peaks." *Journal of Chromatography A* 931 (1-2) 1-30

Dixon, J. K. (1979) "Pattern-Recognition with Partly Missing Data." *Ieee Transactions on Systems Man and Cybernetics* 9 (10) 617-621

Draxier, R. R. and G. D. Hess (1998) "An overview of the HYSPLIT_4 modelling system for trajectories, dispersion and deposition." *Australian Meteorological Magazine* 47 (4) 295-308

Ehn, M., H. Junninen, T. Petäjä, T. Kurten, V. M. Kerminen, S. Schobesberger, H. E. Manninen, I. K. Ortega, H. Vehkamäki, M. Kulmala and D. R. Worsnop (2010) "Composition and temporal behavior of ambient ions in the boreal forest." *Atmospheric Chemistry and Physics* 10 (17) 8513-8530

Ehn, M., H. Vuollekoski, T. Petäjä, V. M. Kerminen, M. Vana, P. Aalto, G. de Leeuw, D. Ceburnis, R. Dupuy, C. D. O'Dowd and M. Kulmala (2010) "Growth rates during coastal and marine new particle formation in western Ireland." *Journal of Geophysical Research-Atmospheres* 115

Ehn, M., H. Junninen, S. Schobesberger, H. E. Manninen, A. Franchin, M. Sipilä, T. Petäjä, V. M. Kerminen, H. Tammet, A. Mirme, S. Mirme, U. Hõrrak, M. Kulmala and D. R. Worsnop (2011) "An Instrumental Comparison of Mobility and Mass Measurements of Atmospheric Small Ions." *Aerosol Science and Technology* 45 (4) 522-532

Ehn, M., E. Kleist, H. Junninen, T. Petäjä, G. Lönn, S. Schobesberger, M. Dal Maso, A. Trimborn, M. Kulmala, D. R. Worsnop, A. Wahner, J. Wildt and T. F. Mentel (2012) "Gas phase formation of extremely oxidized pinene reaction products in chamber and ambient air." *Atmospheric Chemistry and Physics* 12 (11) 5113-5127

Eisele, F. L. and D. J. Tanner (1993) "Measurement of the Gas-Phase Concentration of H2so4 and Methane Sulfonic-Acid and Estimates of H2so4 Production and Loss in the Atmosphere." *Journal of Geophysical Research-Atmospheres* 98 (D5) 9001-9010

Friedlan.Sk (1973) "Chemical Element Balances and Identification of Air-Pollution Sources." *Environmental Science & Technology* 7 (3) 235-240

Gardner, M. W. and S. R. Dorling (1999) "Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London." *Atmospheric Environment* 33 (5) 709-719

Gonin, M., Y. H. Chen, T. Horvath, M. Theiss and H. Wollnik (1998) "Ion-detectors for TOF mass analyzers." *Nuclear Instruments & Methods in Physics Research Section B-Beam Interactions with Materials and Atoms* 136 1244-1247

Grini, A., H. Korhonen, K. E. J. Lehtinen, I. S. A. Isaksen and M. Kulmala (2005) "A combined photochemistry/aerosol dynamics model: model development and a study of new particle formation." *Boreal Environment Research* 10 (6) 525-541

Hair, J., W. C. Black, B. J. Babin, E. A. Rolph and R. L. Tatham. 2005. "Multivariate Data Analysis. 5 ed: Pearson Prentce Hall.

Hamed, A., J. Joutsensaari, S. Mikkonen, L. Sogacheva, M. Dal Maso, M. Kulmala, F. Cavalli, S. Fuzzi, M. C. Facchini, S. Decesari, M. Mircea, K. E. J. Lehtinen and A. Laaksonen (2007) "Nucleation and growth of new particles in Po Valley, Italy." *Atmospheric Chemistry and Physics* 7 355-376

Hari, P. and M. Kulmala (2005) "Station for measuring ecosystem-atmosphere relations (SMEAR II)." *Boreal Environment Research* 10 (5) 315-322

Hierarchial data format version 5 The HDF Group 2000-2010

Herrmann, W., T. Eichler, N. Bernardo and J. F. de la Mora (2000). "Turbulent Transition Arises at Reynolds Number 35000 in a Short Vienna Type DMA with a Large Laminization Inlet." *Journal*(Issue).

Hirsikko, A., V. Vakkari, P. Tiitta, J. Hatakka, V. M. Kerminen, A. M. Sundstrom, J. P. Beukes, H. E. Manninen, M. Kulmala and L. Laakso (2013) "Multiple daytime nucleation events in semi-clean savannah and industrial environments in South Africa: analysis based on observations." *Atmospheric Chemistry and Physics* 13 (11) 5523-5532

Hopke, P. K. (1988) "Target Transformation Factor-Analysis as an Aerosol Mass Apportionment Method - a Review and Sensitivity Study." *Atmospheric Environment* 22 (9) 1777-1792

Hopke, P. K., Y. L. Xie and P. Paatero (1999) "Mixed multiway analysis of airborne particle composition data." *Journal of Chemometrics* 13 (3-4) 343-352

Hussein, T., H. Junninen, P. Tunved, A. Kristensson, M. Dal Maso, I. Riipinen, P. P. Aalto, H. C. Hansson, E. Swietlicki and M. Kulmala (2009) "Time span and spatial scale of regional new particle formation events over Finland and Southern Sweden." *Atmospheric Chemistry and Physics* 9 (14) 4699-4716

Hyvönen, S., H. Junninen, L. Laakso, M. Dal Maso, T. Grönholm, B. Bonn, P. Keronen, P. Aalto, V. Hiltunen, T. Pohja, S. Launiainen, P. Hari, H. Mannila and M. Kulmala (2005) "A look at aerosol formation using data mining techniques." *Atmospheric Chemistry and Physics* 5 3345-3356

Järvi, L., H. Junninen, A. Karppinen, R. Hillamo, A. Virkkula, T. Mäkelä, T. Pakkanen and M. Kulmala (2008) "Temporal variations in black carbon concentrations with different time scales in Helsinki during 1996-2005." *Atmospheric Chemistry and Physics* 8 (4) 1017-1027

Järvi, L., H. Hannuniemi, T. Hussein, H. Junninen, P. P. Aalto, R. Hillamo, T. Mäkelä, P. Keronen, E. Siivola, T. Vesala and M. Kulmala (2009) "The urban measurement station SMEAR III: Continuous monitoring of air pollution and surface-atmosphere interactions in Helsinki, Finland." *Boreal Environment Research* 14 86-109

Jokinen, T., M. Sipilä, H. Junninen, M. Ehn, G. Lönn, J. Hakala, T. Petäjä, R. L. Mauldin, M. Kulmala and D. R. Worsnop (2012) "Atmospheric sulphuric acid and neutral cluster measurements using CI-APi-TOF." *Atmospheric Chemistry and Physics* 12 (9) 4117-4125

Junninen, H., M. Hulkkonen, I. Riipinen, T. Nieminen, A. Hirsikko, T. Suni, M. Boy, S. H. Lee, M. Vana, H. Tammet, V. M. Kerminen and M. Kulmala (2008) "Observations on nocturnal growth of atmospheric clusters." *Tellus Series B-Chemical and Physical Meteorology* 60 (3) 365-371

Junninen, H., J. Monster, M. Rey, J. Cancelinha, K. Douglas, M. Duane, V. Forcina, A. Muller, F. Lagler, L. Marelli, A. Borowiak, J. Niedzialek, B. Paradiz, D. Mira-Salama, J. Jimenez, U. Hansen, C. Astorga, K. Stanczyk, M. Viana, X. Querol, R. M. Duvall, G. A. Norris, S. Tsakovski, P. Wåhlin, J. Horak and B. R. Larsen (2009) "Quantifying the Impact of Residential Heating on the Urban Air Quality in a Typical European Coal Combustion Region." *Environmental Science & Technology* 43 (20) 7964-7970

Junninen, H., M. Ehn, T. Petäjä, L. Luosujärvi, T. Kotiaho, R. Kostiainen, U. Rohner, M. Gonin, K. Fuhrer, M. Kulmala and D. R. Worsnop (2010) "A high-resolution mass spectrometer to measure atmospheric ion composition." *Atmospheric Measurement Techniques* 3 (4) 1039-1053

Kangasluoma, J., H. Junninen, K. Lehtipalo, J. Mikkila, J. Vanhanen, M. Attoui, M. Sipilä, D. Worsnop, M. Kulmala and T. Petäjä (2013) "Remarks on Ion Generation for CPC Detection Efficiency Studies in Sub-3-nm Size Range." *Aerosol Science and Technology* 47 (5) 556-563

Kerminen, V. M., T. Anttila, K. E. J. Lehtinen and M. Kulmala (2004) "Parameterization for atmospheric new-particle formation: Application to a system involving sulfuric acid and condensable water-soluble organic vapors." *Aerosol Science and Technology* 38 (10) 1001-1008

Keskinen, H., A. Virtanen, J. Joutsensaari, G. Tsagkogeorgas, J. Duplissy, S. Schobesberger, M. Gysel, F. Riccobono, J. G. Slowik, F. Bianchi, T. Yli-Juuti, K. Lehtipalo, L. Rondo, M. Breitenlechner, A. Kupc, J. Almeida, A. Amorim, E. M. Dunne, A. J. Downard, S. Ehrhart, A. Franchin, M. K. Kajos, J. Kirkby, A. Kurten, T. Nieminen, V. Makhmutov, S. Mathot, P. Miettinen, A. Onnela, T. Petäjä, A. Praplan, F. D. Santos, S. Schallhart, M. Sipilä, Y. Stozhkov, A. Tome, P. Vaattovaara, D. Wimmer, A. Prevot, J. Dommen, N. M. Donahue, R. C. Flagan, E. Weingartner, Y. Viisanen, I. Riipinen, A. Hansel, J. Curtius, M. Kulmala, D. R.

Worsnop, U. Baltensperger, H. Wex, F. Stratmann and A. Laaksonen (2013) "Evolution of particle composition in CLOUD nucleation experiments." *Atmospheric Chemistry and Physics* 13 (11) 5587-5600

Kirkby, J., J. Curtius, J. Almeida, E. Dunne, J. Duplissy, S. Ehrhart, A. Franchin, S. Gagne, L. Ickes, A. Kurten, A. Kupc, A. Metzger, F. Riccobono, L. Rondo, S. Schobesberger, G. Tsagkogeorgas, D. Wimmer, A. Amorim, F. Bianchi, M. Breitenlechner, A. David, J. Dommen, A. Downard, M. Ehn, R. C. Flagan, S. Haider, A. Hansel, D. Hauser, W. Jud, H. Junninen, F. Kreissl, A. Kvashin, A. Laaksonen, K. Lehtipalo, J. Lima, E. R. Lovejoy, V. Makhmutov, S. Mathot, J. Mikkila, P. Minginette, S. Mogo, T. Nieminen, A. Onnela, P. Pereira, T. Petäjä, R. Schnitzhofer, J. H. Seinfeld, M. Sipilä, Y. Stozhkov, F. Stratmann, A. Tome, J. Vanhanen, Y. Viisanen, A. Vrtala, P. E. Wagner, H. Walther, E. Weingartner, H. Wex, P. M. Winkler, K. S. Carslaw, D. R. Worsnop, U. Baltensperger and M. Kulmala (2011) "Role of sulphuric acid, ammonia and galactic cosmic rays in atmospheric aerosol nucleation." *Nature* 476 (7361) 429-U77

Kivekas, N., J. Sun, M. Zhan, V. M. Kerminen, A. Hyvarinen, M. Komppula, Y. Viisanen, N. Hong, Y. Zhang, M. Kulmala, X. C. Zhang, Deli-Geer and H. Lihavainen (2009) "Long term particle size distribution measurements at Mount Waliguan, a high-altitude site in inland China." *Atmospheric Chemistry and Physics* 9 (15) 5461-5474

Kohonen, T. (1997) "The self-organising map, a possible model of brain maps." *Perception* 26 1-1

Kolehmainen, M., J. Ruuskanen, E. Rissanen and O. Raatikainen (2001) "Monitoring odorous sulfur emissions using self-organizing maps for handling ion mobility spectrometry data." *Journal of the Air & Waste Management Association* 51 (7) 966-971

Korhonen, P., M. Kulmala and Y. Viisanen (1997) "A theoretical study of binary homogenous nucleation of water-ammonium chloride particles in the atmosphere." *Journal of Aerosol Science* 28 (6) 901-917

Kulmala, M., U. Pirjola and J. M. Mäkelä (2000) "Stable sulphate clusters as a source of new atmospheric particles." *Nature* 404 (6773) 66-69

Kulmala, M. (2003) "How particles nucleate and grow." *Science* 302 (5647) 1000-1001

Kulmala, M., L. Laakso, K. E. J. Lehtinen, I. Riipinen, M. Dal Maso, T. Anttila, V. M. Kerminen, U. Hõrrak, M. Vana and H. Tammet (2004) "Initial steps of aerosol growth." *Atmospheric Chemistry and Physics* 4 2553-2560

Kulmala, M., I. Riipinen, M. Sipilä, H. E. Manninen, T. Petäjä, H. Junninen, M. Dal Maso, G. Mordas, A. Mirme, M. Vana, A. Hirsikko, L. Laakso, R. M. Harrison, I. Hanson, C. Leung, K. E. J. Lehtinen and V. M. Kerminen (2007) "Toward direct measurement of atmospheric nucleation." *Science* 318 (5847) 89-92

Kulmala, M., T. Petäjä, T. Nieminen, M. Sipilä, H. E. Manninen, K. Lehtipalo, M. Dal Maso, P. P. Aalto, H. Junninen, P. Paasonen, I. Riipinen, K. E. J. Lehtinen, A. Laaksonen and V. M. Kerminen (2012) "Measurement of the nucleation of atmospheric aerosol particles." *Nature Protocols* 7 (9) 1651-1667

Kulmala, M., J. Kontkanen, H. Junninen, K. Lehtipalo, H. E. Manninen, T. Nieminen, T. Petäjä, M. Sipilä, S. Schobesberger, P. Rantala, A. Franchin, T. Jokinen, E. Järvinen, M. Aijala, J. Kangasluoma, J. Hakala, P. P. Aalto, P. Paasonen, J. Mikkila, J. Vanhanen, J. Aalto, H. Hakola, U. Makkonen, T. Ruuskanen, R. L. Mauldin, J. Duplissy, H. Vehkamäki, J. Back, A. Kortelainen, I. Riipinen, T. Kurten, M. V. Johnston, J. N. Smith, M. Ehn, T. F. Mentel, K. E. J. Lehtinen, A. Laaksonen, V. M. Kerminen and D. R. Worsnop (2013) "Direct Observations of Atmospheric Aerosol Nucleation." *Science* 339 (6122) 943-946

Laakso, L., H. Laakso, P. P. Aalto, P. Keronen, T. Petäjä, T. Nieminen, T. Pohja, E. Siivola, M. Kulmala, N. Kgabi, M. Molefe, D. Mabaso, D. Phalatse, K. Pienaar and V. M. Kerminen (2008) "Basic characteristics of atmospheric particles, trace gases and meteorology in a relatively clean Southern African Savannah environment." *Atmospheric Chemistry and Physics* 8 (16) 4823-4839

Laakso, L., J. Merikanto, V. Vakkari, H. Laakso, M. Kulmala, M. Molefe, N. Kgabi, D. Mabaso, K. S. Carslaw, D. V. Spracklen, L. A. Lee, C. L. Reddington and V. M. Kerminen (2013) "Boundary layer nucleation as a source of new CCN in savannah environment." *Atmospheric Chemistry and Physics* 13 (4) 1957-1972

Laitinen, T., M. Ehn, H. Junninen, J. Ruiz-Jimenez, J. Parshintsev, K. Hartonen, M. L. Riekkola, D. R. Worsnop and M. Kulmala (2011) "Characterization of organic compounds in 10-to 50-nm aerosol particles in boreal forest with laser desorption-ionization aerosol mass spectrometer and comparison with other techniques." *Atmospheric Environment* 45 (22) 3711-3719

Latella, A., G. Stani, L. Cobelli, M. Duane, H. Junninen, C. Astorga and B. R. Larsen (2005) "Semicontinuous GC analysis and receptor modelling for source apportionment of ozone precursor hydrocarbons in Bresso, Milan, 2003." *Journal of Chromatography A* 1071 (1-2) 29-39

Lehtipalo, K., M. Sipilä, H. Junninen, M. Ehn, T. Berndt, M. K. Kajos, D. R. Worsnop, T. Petäjä and M. Kulmala (2011) "Observations of Nano-CN in the Nocturnal Boreal Forest." *Aerosol Science and Technology* 45 (4) 499-509

Leppa, J., T. Anttila, V. M. Kerminen, M. Kulmala and K. E. J. Lehtinen (2011) "Atmospheric new particle formation: real and apparent growth of neutral and charged particles." *Atmospheric Chemistry and Physics* 11 (10) 4939-4955

Mäkelä, J. M., P. Aalto, V. Jokinen, T. Pohja, A. Nissinen, S. Palmroth, T. Markkanen, K. Seitsonen, H. Lihavainen and M. Kulmala (1997) "Observations of ultrafine aerosol particle formation and growth in boreal forest." *Geophysical Research Letters* 24 (10) 1219-1222

Mäkelä, J. M., M. Dal Maso, A. Laaksonen, L. Pirjola, P. Keronen and M. Kulmala (2000) "Characteristics of the three years continuous data on new particle formation events observed at a boreal forest." *Nucleation and Atmospheric Aerosols 2000* 534 896-899

Manninen, H. E., T. Nieminen, E. Asmi, S. Gagne, S. Häkkinen, K. Lehtipalo, P. Aalto, M. Vana, A. Mirme, S. Mirme, U. Hõrrak, C. Plass-Dulmer, G. Stange, G. Kiss, A. Hoffer, N. Toeroe, M. Moerman, B. Henzing, G. de Leeuw, M. Brinkenberg, G. N. Kouvarakis, A. Bougiatioti, N. Mihalopoulos, C. O'Dowd, D. Ceburnis, A. Arneth,

B. Svenningsson, E. Swietlicki, L. Tarozzi, S. Decesari, M. C. Facchini, W. Birmili, A. Sonntag, A. Wiedensohler, J. Boulon, K. Sellegri, P. Laj, M. Gysel, N. Bukowiecki, E. Weingartner, G. Wehrle, A. Laaksonen, A. Hamed, J. Joutsensaari, T. Petäjä, V. M. Kerminen and M. Kulmala (2010) "EUCAARI ion spectrometer measurements at 12 European sites - analysis of new particle formation events." *Atmospheric Chemistry and Physics* 10 (16) 7907-7927

Manninen, H. E., A. Franchin, S. Schobesberger, A. Hirsikko, J. Hakala, A. Skromulis, J. Kangasluoma, M. Ehn, H. Junninen, A. Mirme, S. Mirme, M. Sipilä, T. Petäjä, D. R. Worsnop and M. Kulmala (2011) "Characterisation of corona-generated ions used in a Neutral cluster and Air Ion Spectrometer (NAIS)." *Atmospheric Measurement Techniques* 4 (12) 2767-2776

Manninen, H. E., S. Mirme, M. Ehn, K. Leino, S. Schobesberger, H. Junninen, E. Järvinen, J. Kangasluoma, T. Nieminen, R. Tillmann, F. Angelini, G. P. Gobbi, A. Mirme, S. Decesari, A. Wahner, T. Petäjä, D. R. Worsnop, F. Rohrer, T. F. Mentel and M. Kulmala (2013) "Does the Onset of New Particle Formation Occur in the Planetary Boundary Layer?" *Nucleation and Atmospheric Aerosols* 1527 567-570

MATLAB version 7.14.0 Natick, Massachusetts, USA: The MathWorks Inc. 2012

Mohr, C., F. D. Lopez-Hilfiker, P. Zotter, A. S. H. Prevot, L. Xu, N. L. Ng, S. C. Herndon, L. R. Williams, J. P. Franklin, M. S. Zahniser, D. R. Worsnop, W. B. Knighton, A. C. Aiken, K. J. Gorkowski, M. K. Dubey, J. D. Allan and J. A. Thornton (2013) "Contribution of Nitrated Phenols to Wood Burning Brown Carbon Light Absorption in Detling, United Kingdom during Winter Time." *Environmental Science & Technology* 47 (12) 6316-6324

Monahan, C., H. Vuollekoski, M. Kulmala and C. O'Dowd (2010) "Simulating Marine New Particle Formation and Growth Using the M7 Modal Aerosol Dynamics Modal." *Advances in Meteorology*

Nieminen, T., H. E. Manninen, S. L. Sihto, T. Yli-Juuti, R. L. Mauldin, T. Petäjä, I. Riipinen, V. M. Kerminen and M. Kulmala (2009) "Connection of Sulfuric Acid to Atmospheric Nucleation in Boreal Forest." *Environmental Science & Technology* 43 (13) 4715-4721

O'dowd, C. D., K. Hämeri, J. Mäkelä, M. Väkevä, P. Aalto, G. de Leeuw, G. J. Kunz, E. Becker, H. C. Hansson, A. G. Allen, R. M. Harrison, H. Berresheim, M. Geever, S. G. Jennings and M. Kulmala (2002) "Coastal new particle formation: Environmental conditions and aerosol physicochemical characteristics during nucleation bursts." *Journal of Geophysical Research-Atmospheres* 107 (D19)

O'dowd, C. D., K. Hämeri, J. M. Mäkelä, L. Pirjola, M. Kulmala, S. G. Jennings, H. Berresheim, H. C. Hansson, G. de Leeuw, G. J. Kunz, A. G. Allen, C. N. Hewitt, A. Jackson, Y. Viisanen and T. Hoffmann (2002) "A dedicated study of New Particle Formation and Fate in the Coastal Environment (PARFORCE): Overview of objectives and achievements." *Journal of Geophysical Research-Atmospheres* 107 (D19)

Paasonen, P., S. L. Sihto, T. Nieminen, H. Vuollekoski, I. Riipinen, C. Plass-Dulmer, H. Berresheim, W. Birmili and M. Kulmala (2009) "Connection between new particle

formation and sulphuric acid at Hohenpeissenberg (Germany) including the influence of organic compounds." *Boreal Environment Research* 14 (4) 616-629

Paasonen, P., T. Olenius, O. Kupiainen, T. Kurten, T. Petäjä, W. Birmili, A. Hamed, M. Hu, L. G. Huey, C. Plass-Duelmer, J. N. Smith, A. Wiedensohler, V. Loukonen, M. J. McGrath, I. K. Ortega, A. Laaksonen, H. Vehkamäki, V. M. Kerminen and M. Kulmala (2012) "On the formation of sulphuric acid - amine clusters in varying atmospheric conditions and its influence on atmospheric new particle formation." *Atmospheric Chemistry and Physics* 12 (19) 9113-9133

Paatero, P. and U. Tapper (1994) "Positive Matrix Factorization - a Nonnegative Factor Model with Optimal Utilization of Error-Estimates of Data Values." *Environmetrics* 5 (2) 111-126

Paatero, P. (1997) "Least squares formulation of robust non-negative factor analysis." *Chemometrics and Intelligent Laboratory Systems* 37 (1) 23-35

Paatero, P. (1999) "The multilinear engine - A table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model." *Journal of Computational and Graphical Statistics* 8 (4) 854-888

Petäjä, T., R. L. Mauldin, E. Kosciuch, J. McGrath, T. Nieminen, P. Paasonen, M. Boy, A. Adamov, T. Kotiaho and M. Kulmala (2009) "Sulfuric acid and OH concentrations in a boreal forest site." *Atmospheric Chemistry and Physics* 9 (19) 7435-7448

Petäjä, T., M. Sipilä, P. Paasonen, T. Nieminen, T. Kurten, I. K. Ortega, F. Stratmann, H. Vehkamäki, T. Berndt and M. Kulmala (2011) "Experimental Observation of Strongly Bound Dimers of Sulfuric Acid: Application to Nucleation in the Atmosphere." *Physical Review Letters* 106 (22)

Pikridas, M., A. Bougiatioti, L. Hildebrandt, G. J. Engelhart, E. Kostenidou, C. Mohr, A. S. H. Prevot, G. Kouvarakis, P. Zarmpas, J. F. Burkhart, B. H. Lee, M. Psichoudaki, N. Mihalopoulos, C. Pilinis, A. Stohl, U. Baltensperger, M. Kulmala and S. N. Pandis (2010) "The Finokalia Aerosol Measurement Experiment-2008 (FAME-08): an overview." *Atmospheric Chemistry and Physics* 10 (14) 6793-6806

Pikridas, M., I. Riipinen, L. Hildebrandt, E. Kostenidou, H. Manninen, N. Mihalopoulos, N. Kalivitis, J. F. Burkhart, A. Stohl, M. Kulmala and S. N. Pandis (2012) "New particle formation at a remote site in the eastern Mediterranean." *Journal of Geophysical Research-Atmospheres* 117

Schneider, T. (2001) "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values." *Journal of Climate* 14 (5) 853-871

Schobesberger, S., R. Vaananen, K. Leino, A. Virkkula, J. Backman, T. Pohja, E. Siivola, A. Franchin, J. Mikkila, M. Paramonov, P. P. Aalto, R. Krejci, T. Petäjä and M. Kulmala (2013) "Airborne measurements over the boreal forest of southern Finland during new particle formation events in 2009 and 2010." *Boreal Environment Research* 18 (2) 145-163

Sihto, S. L., H. Vuollekoski, J. Leppa, I. Riipinen, V. M. Kerminen, H. Korhonen, K. E. J. Lehtinen, M. Boy and M. Kulmala (2009) "Aerosol dynamics simulations on the connection of sulphuric acid and new particle formation." *Atmospheric Chemistry and Physics* 9 (9) 2933-2947

Sipilä, M., T. Berndt, T. Petäjä, D. Brus, J. Vanhanen, F. Stratmann, J. Patokoski, R. L. Mauldin, A. P. Hyvarinen, H. Lihavainen and M. Kulmala (2010) "The Role of Sulfuric Acid in Atmospheric Nucleation." *Science* 327 (5970) 1243-1246

Spracklen, D. V., B. Bonn and K. S. Carslaw (2008) "Boreal forests, aerosols and the impacts on clouds and climate." *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences* 366 (1885) 4613-4626

Vakkari, V., H. Laakso, M. Kulmala, A. Laaksonen, D. Mabaso, M. Molefe, N. Kgabi and L. Laakso (2011) "New particle formation events in semi-clean South African savannah." *Atmospheric Chemistry and Physics* 11 (7) 3333-3346

Vana, M., M. Ehn, T. Petäjä, H. Vuollekoski, P. Aalto, G. de Leeuw, D. Ceburnis, C. D. O'Dowd and M. Kulmala (2008) "Characteristic features of air ions at Mace Head on the west coast of Ireland." *Atmospheric Research* 90 (2-4) 278-286

von Bobrutzki, K., C. F. Braban, D. Famulari, S. K. Jones, T. Blackall, T. E. L. Smith, M. Blom, H. Coe, M. Gallagher, M. Ghalaieny, M. R. McGillen, C. J. Percival, J. D. Whitehead, R. Ellis, J. Murphy, A. Mohacsi, A. Pogany, H. Junninen, S. Rantanen, M. A. Sutton and E. Nemitz (2010) "Field inter-comparison of eleven atmospheric ammonia measurement techniques." *Atmospheric Measurement Techniques* 3 (1) 91-112

Wåhlin, P. (2003) "COPREM - A multivariate receptor model with a physical approach." *Atmospheric Environment* 37 (35) 4861-4867

Willmott, C. J. (1982) "Some Comments on the Evaluation of Model Performance." *Bulletin of the American Meteorological Society* 63 (11) 1309-1313

Xie, Y. L., P. K. Hopke, P. Paatero, L. A. Barrie and S. M. Li (1999) "Identification of source nature and seasonal variations of Arctic aerosol by the multilinear engine." *Atmospheric Environment* 33 (16) 2549-2562

Yli-Juuti, T., I. Riipinen, P. P. Aalto, T. Nieminen, W. Maenhaut, I. A. Janssens, M. Claeys, I. Salma, R. Ocskay, A. Hoffer, K. Imre and M. Kulmala (2009) "Characteristics of new particle formation events and cluster ions at K-puszta, Hungary." *Boreal Environment Research* 14 (4) 683-698

Zhou, L. X., M. Boy, T. Nieminen, D. Mogensen, S. Smolander and M. Kulmala (2013) "Modeling New Particle Formation with Detailed Chemistry and Aerosol Dynamics in a Boreal Forest Environment." *Nucleation and Atmospheric Aerosols* 1527 405-408