Data mining approaches to explaining aerosol formation

S. Hyvönen [a], H. Junninen[b], L. Laakso[b], M. Dal Maso[b], T. Grönholm[b], B. Bonn[b], P. Keronen[b], P. Aalto[b], V. Hiltunen[c], T. Pohja[c], S. Launiainen[b], P. Tunved[d], H.C. Hansson[d], P. Hari[e], H. Mannila[a] and M. Kulmala[b]

[a]Helsinki Institute for Information Technology, Department of Computer Science, University of Helsinki P.O.Box 68, FI–00014 University of Helsinki, Finland

[b]Department of Physics, University of Helsinki, P.O. Box 64, FI–00014 University of Helsinki, Finland

[c]Hyytiälä Forestry Field Station, University of Helsinki, Finland,

[d]Department of Applied Environmental Science, Stockholm University

[e]Department of Forest Ecology, Faculty of Agriculture and Forestry, University of Helsinki, Finland

Abstract: Atmospheric aerosol particle formation is frequently observed in various environments. Yet, despite numerous studies, processes behind these so called nucleation events remain unclear. In this work we describe the use of data mining techniques to detect factors influencing particle formation. These techniques are applied to a dataset of eight years of 80 variables collected at the boreal forest station (SMEAR II) in Southern Finland, including air pollutant, weather, gas and particle measurements. In a previous study classification methods have been used together with feature selection in order to understand what causes nucleation. Each day was classified as an event day, when a nucleation event occurs, or as a nonevent day, and looking at which features were selected gives us information on which factors are important for the aerosol formation process. This way it was possible to identify two key variables, relative humidity and preexisting aerosol particle surface (condensation sink), capable of explaining 88% of the nucleation events. Using these two variables a nucleation probability function could be derived. In this paper this nucleation probability function has been tested on data collected from other sites, Värriö in Northern Lapland and Aspvreten in Sweden. We show that in the extreme conditions in Värriö the nucleation parameter does not work, whereas in Aspvreten the two key variables can be used to identify nucleation events, though the nucleation parameter has to be adjusted slightly. The two key variables are related to mechanisms that prevent nucleation. One reason for the domination of preventive mechanisms could be the existence of more than one mechanism causing nucleation. Another intriguing phenomenon, possibly related to this, is the temporal variation of nucleation events. We have investigated temporal phenomena in nucleation by using classification methods in a sliding window.We discuss some aspects of this approach and present some results obtained.

Keywords: atmospheric aerosols; particle formation; data mining; classification

1 INTRODUCTION

Understanding atmospheric aerosol particle formation is an important issue in understanding the contribution of aerosols to climate change. Aerosol particle formation bursts, also called nucleation events, are frequently observed in various environments [Kulmala et al., 2004]; yet the driving mechanisms behind nucleation events are in many cases poorly understood [Hellmuth, 2005]. The physical and chemical complexity of the atmosphere makes it difficult to focus on the most relevant processes causing nucleation. This is where data mining techniques come to aid.

Previous studies have demonstrated that low atmospheric water content, low preexisting particle concentration and high solar radiation imply favorable

conditions for nucleation events [Boy and Kulmala, 2002]. Also the physical mechanisms, meteorological conditions [Nilsson et al., 2001] and chemical compounds related to particle formation [Weber et al., 1995; Korhonen et al., 1999; Birmili and Wiedensohler, 2000; O'Dowd et al., 2002; Bonn and Moortgat, 2003; Kulmala et al., 2004] have been studied.

Data mining techniques have been used by Hyvönen et al. [2005] to study atmospheric aerosol formation. Here we review some of those results, test them on data collected from other sites, and extend those results by applying classification methods in a sliding window to gain insight on temporal phenomena in nucleation.

2 DATA

Measurements used in this study where performed during the years 1996–2003 at the SMEAR II station, which is located in Hyytiälä, (61°51'N, 24°17'E, 180 m a.s.l.). Rannik [1998] describes the micrometeorology of the site. Measured variables include meteorological data: temperature, pressure, wind speed, wind direction, humidity and radiation (UV-A, UV-B, PAR, global, net, reflected global and reflected PAR); gas concentrations of NO, NO_x, SO_2, O_3, H_2O, CO and CO_2; and flux measurements of sensible heat, latent heat, momentum, CO_2, H_2O and O_3. More details on the data and measurements can be found in [Hyvönen et al., 2005; Vesala et al., 1998]. The continuous measurements of these variables have been averaged over 30 minute intervals. The meteorological and gas concentration measurements where performed at six different heights. Some of these variables correlate strongly. For this reason, we have removed a number of variables. These include latent heat flux, which correlates with water vapour flux; all radiations but one, as these correlate with each other. Note that it is impossible to remove all correlations due to the nature of atmospheric data; once the strong correlations are removed, one still must take care in interpreting the results.

In addition to this measurement data we include the condensation sink as one of our variables. The condensation sink is roughly related to the surface area of particles present in the air. For details on the definition and significance of the condensation sink see [Hyvönen et al., 2005; Pirjola and Kulmala, 1998].

The data set we analyzed consists of eight years of measurements averaged over 30 minute time intervals, with a large number of missing data. This data set is preprocessed as follows. First we calculate for each day the mean and standard deviation of each variable during daylight hours. This time window was chosen because in boreal regions such as Hyytiälä (situated 61 degrees North) the length of the day depends strongly on the time of the year. Next all variables with more than 800 days of missing values were excluded from the data set; after this we exclude any day with any missing variable. Missing data is mostly due to equipment failure or maintenance during which measurements are missing for several days or even weeks. These are not completely random, as maintenance breaks tend to occur when nucleation events are known to be rare. We do not use all different height measurements, but an average over all heights. Finally, the data is normalized to have zero mean and unit variance. This is done to make sure that all variables are treated equally; otherwise, variables with large numerical values would appear to be more important when days are compared.

To distinguish between days with new particle formation and days with no particle formation we used a database created by Dal Maso et al. [2005]. Days displaying a growing new mode in the nucleation size range prevailing over several hours are classified as event days. Days which are clear of all traces of particle formation are classified as non-event days. Days which can not unambiguously be classified as either event or non-event days are termed 'undefined' days, and removed from the data pool used in this study.

3 CLASSIFICATION METHODS

In studying the causes behind nucleation events one can consider the task as a classification problem: using the atmospheric data set we wish to classify each day as an event day or a non-event day. In fact, we are not only interested in finding a reasonable classifier, but finding out which variables are significant in separating event days from non-event days. The application of a wide range of classification methods applied to this data set has been reported by Hyvönen et al. [2005]; we mention here two of them, used later in this paper as well. For details, see e.g. [Hand et al., 2001; Hastie et al., 2001].

Linear discriminant analysis (LDA) aims to find a set of linear combinations of the original variables that best separates the classes. Such linear combinations are called linear discriminants. In a two-

class case only one linear discriminant is sought after. This gives the direction which separates the two classes best; it is thus the normal to the plane separating the classes.

Linear regression in turn predicts the output y via a linear model $y = \beta_0 + \sum_{j=1}^{n} \beta_j x_j$, where $(x_j)_{j=1}^{n}$ is our $n-$dimensional input data. This is usually used to predict quantitative outputs, but it can be used for classification tasks too. In the classification case we define y to be one for event days and zero for non-event days, and fit the regression model accordingly. Our input data consists of the measurement vectors for each day. Linear regression is computationally fast compared to e.g. LDA; which makes it attractive when a large number of classification tasks is to be performed.

Because of the correlations still present in the data, neither of these methods can be reliably used to estimate the significance of different variables in classification. But we are interested in particular in finding out which variables are significant in separating event days from non-event days. This can be done using foward stepwise selection of variables. In this approach we start with the variable which gives the best classification performance alone, and on each step we add the variable which results in the best classification performance together with the variable(s) already selected.

We have used cross-validation to estimate the performance of the classification methods used.

4 NUCLEATION PARAMETER

The main result obtained by Hyvönen et al. [2005] using a wide range of classification methods is that the most important variables in explaining nucleation events are the means of relative humidity (RH) and the logarithm of the condensation sink (CS). The data in terms of these two variables is presented in Figure 1a. This finding was supported by a number of different approaches. Different classification methods perform slightly differently, but these two key variables remain unchanged, while including further parameters does not improve the results notably. Linear discriminant analysis has the best performance. Using LDA together with the two key variables results in a 12% classification error. When the data is projected onto the first linear discriminant, points at one end of the line are mainly event, whereas points at the opposite end are mainly non-events. From this projected data it is possible to compute the probability of having an event day at

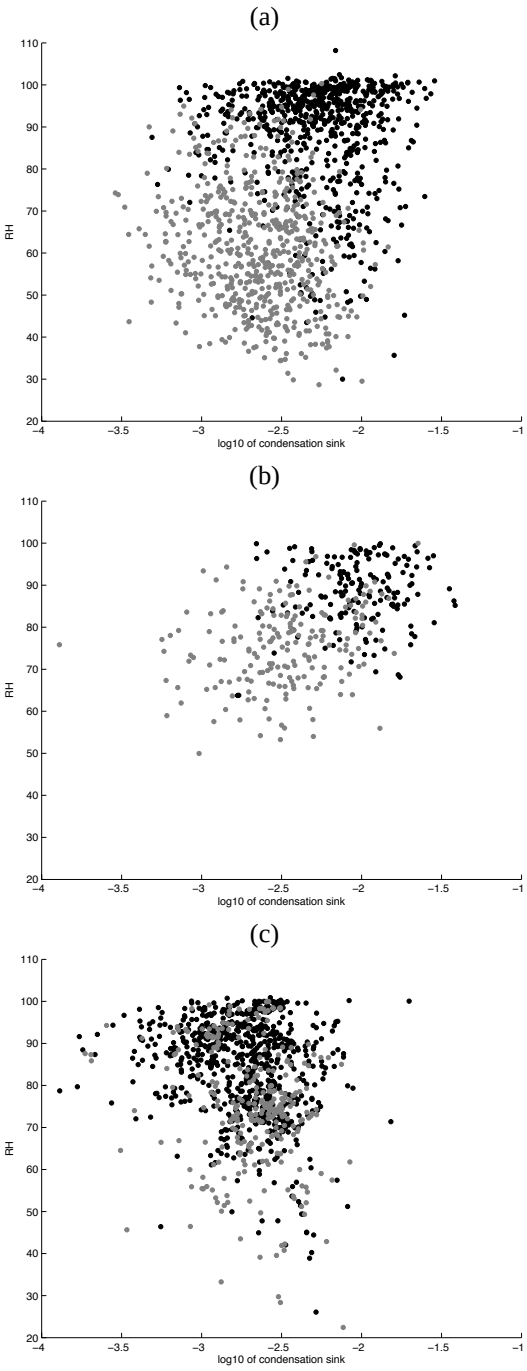

Figure 1: Event days (grey) and nonevent days (black) in the RH, $\log_{10}$CS-plane in (a) Hyytiälä [Hyvönen et al., 2005], (b) Aspvreten and (c) Värriö.

each point. This was done in [Hyvönen et al., 2005], where the following nucleation parameter describing the probability of nucleation was derived:

$$P_{nucl} = \frac{1}{1 + \exp(\beta_1 \log(CS) + \beta_2(RH))},$$

where $\beta_1 = 1.7$ and $\beta_2 = 0.13$. This nucleation parameter was derived using data from the Hyytiälä measuring station, so therefore it is certainly biased towards the conditions there. We have also tested it on data from other sites. The SMEAR I measuring station is situated in Värriö, in Eastern Lapland, way above the arctic circle (67°45'N, 29°37'E, 375 m a.s.l.). Stockholm University has a measuring station in Aspvreten, situated about 80 km southwest of Stockholm and 2km west of the Baltic coast (58° 48'N, 17°23'E, 20 m a.s.l.). Comparing Figures 1a–1c it is evident, that both of the more southern locations events and nonevents are nicely separated in the RH, $\log_{10}$CS-plane, whereas in the northernmost Värriö, where are very different from those in Hyytiälä, this clearly is not the case, so the nucleation parameter will not work there.

The nucleation parameter together with the proportions of events along the first linear discriminant in all three sites is given in Figure 2. We notice, that the Aspvreten data agrees with the nucleation parameter in shape, but there is a shift towards the right, so the nucleation parameter should be adjusted by a constant. This may indicate a latitudinal dependency that should be incorporated in the nucleation parameter.

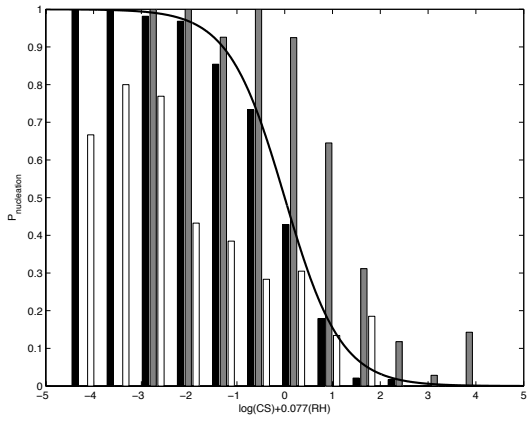

Figure 2: Nucleation parameter and proportion of events along the first linear discriminant. Black for Hyytiälä, grey for Aspvreten and white for Värriö.

5 TEMPORAL VARIATION IN NUCLEATION

Doing classification together with feature selection on the whole data set has given us the two variables, relative humidity and the condensation sink, capable of explaining 88% of the nucleation events. However, seasonal variation of all variables is strong, and nucleation also has a strong seasonal behaviour (see Figure 3). Hence, it is natural to ask whether explaining the events would be easier if we only wanted to do it in one season. Instead of fixing the definition of the seasons we approach this question by doing classification together with feature selection in a sliding window to see how the variables selected for classification vary seasonally. After a preliminary round involving a larger set of 20 variables we have picked only the five variables frequently selected by the classification methods. The results are presented for these variables only.

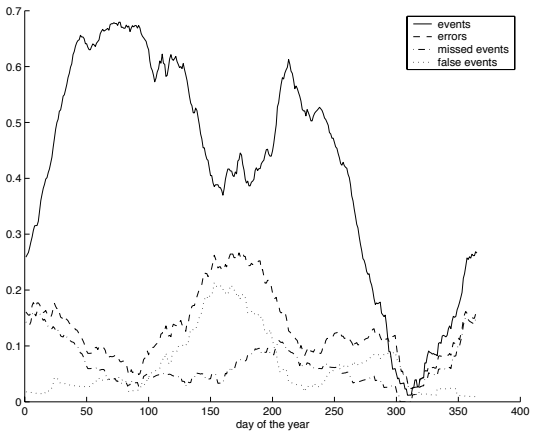

Figure 3: Proportion of events and errors in a sliding window of width 60 days.

We have used linear regression together with feature selection in a sliding window. Too short windows yield unstable results, because we do not have enough data in each window, so we have used a window width of 100 days. The regression weights in a sliding window are shown in Figure 4. A clear seasonal behaviour is present here. One can roughly distinguish between the four models presented in Table 1. LDA detects the same time intervals with models involving the same variables. Also using a shorter window gives similar results with slightly more noise. The erratic behavior of the models towards the end of the year is due to lack of events and low variation of parameters in the winter. This is not really a problem, because most models valid elsewhere work well also in the winter period.

variables	model1	model2	model3	model4
const	0.50	0.42	0.51	0.44
RH	0	0	0	-0.40
PAR	0.29	0	0	0
RPAR	0	0	0.33	0
logCS	-0.21	-0.31	-0.29	-0.10
sensheat	0	0.18	0	0

Table 1: Models picked from the results obtained by regression in a sliding window, see Figure 4.

Because we use sliding windows, the models overlap. To determine where the best switching points from one model to another are segmentation methods [Gionis and Mannila, 2003] can be used. Comparing the error rates of different numbers of time segments we concluded, that the best choice is using three time segments, see Table 2. So we use model 1 (Table 1) for days before April 6th, model 4 for days after September 7th, and model 2 in the time interval in between.

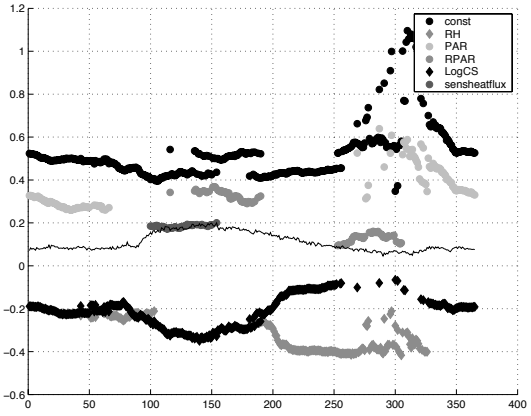

Figure 4: Regression weights in a sliding window of width 100 days. The starting point of the interval is indicated on the x-axis. The solid line shows the error in the sliding window.

Now if we use the three segment model instead of the one model valid all year, we get an approximately 10% error rate compared to the ca. 13% error rate obtained for linear regression.

Further insight into seasonal behaviour is gained if we look at how the error rate varies seasonally. Figure 5 shows the error rates and false negative error rates in a sliding window for the one model case in black and the three model case in grey. Results are similar for LDA.

segments	breaking points	models used	error
2	249	2,4	12.5
3	96,250	1,2,4	10.4
4	101,165,249	1,2,3,4	10.4
5	15,104,165,249	1,1,2,3,4	10.5

Table 2: Segmentation results, when the four models in Table 1 were used.

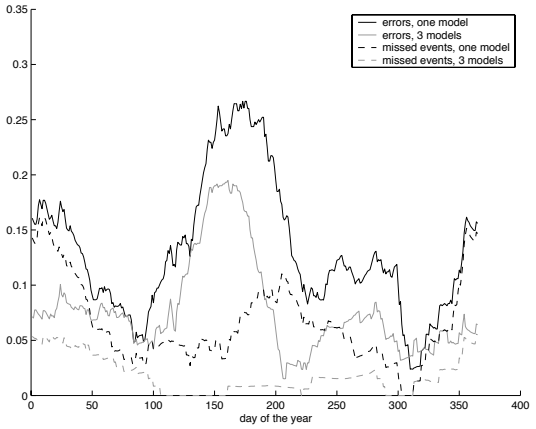

Figure 5: Proportion of events and errors in a sliding window of width 60 days for one model for the whole year in black and the three segment model in grey.

6 CONCLUSIONS

In our previous work [Hyvönen et al., 2005] we used a wide range of classification methods to gain insight on atmospheric aerosol formation. We were able to identify two key variables, relative humidity and the condensation sink, capable of explaining 88% of the nucleation events. Furthermore, a nucleation parameter describing the probability of nucleation was presented.

Here we have tested this parameter on data collected from other sites. We conclude that if the conditions are sufficiently similar, as they are in Aspvreten, then the two key variables still work well in predicting nucleation, though the nucleation parameter has to be adjusted slightly; however, if conditions are very different, as they are in Värriö situated far above the arctic circle, then the parameter no longer is valid.

Furthermore, we have investigated the temporal variation in the classification models. This has been done by doing regression in a sliding window. It is

evident, that there is seasonal variation in the optimal model: a natural division into a spring, summer and fall models appear. No winter model is needed, because there are practically no events and all models perform well. That different models work better in different seasons does not mean, that the mechanisms behind nucleation change; rather, it reflects on changes in event density and general meteorological conditions. In fact, the different models correlate, so all of them work fairly well on the whole time segment. It should be noted, that though the predictive power of the segmented model is higher than that of the one model case, the main goal here was not maximizing the classification perfomance, but gaining insight on the temporal behaviour of nucleation.

The condensation sink and relative humidity both relate to mechanisms that prevent nucleation from starting and particles from growing to detectable sizes. On the contrary, the solar radiation measured by PAR is known to be one of the key elements in the reaction chain; also, event days have high values of sensible heat flux, which describes heat energy transfer into the atmosphere. Even using this seasonally segmented model the error rate during summer is significantly higher than during the other seasons. The errors are almost exlusively false positives, since the summer events are in fact fairly rare. Explaining this remains a future challenge.

REFERENCES

Birmili, W. and A. Wiedensohler. New particle formation in the continental boundary layer: Meteorological and gas phase parameter influence. *Geophysical Research Letters*, 27:3325–3328, 2000.

Bonn, B. and G. Moortgat. Sesquiterpene ozonolysis: Origin of atmospheric new particle formation from biogenic hydrocarbons. *Geophys Res Letters*, 30:1585–1588, 2003.

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

Dal Maso, M., M. Kulmala, I. Riipinen, R. Wagner, T. Hussein, P. P. Aalto, and K. E. J. Lehtinen. Formation and growth of fresh atmospheric aerosols: Eight years of aerosol size distribution data from SMEAR II, Hyytiala, Finland. *Boreal Env. Res.*, pages 322–336, 2005.

Gionis, A. and H. Mannila. Finding recurrent sources in sequences. *7th International Conference on Research in Computational Molecular Biology (RECOMB)*, 2003.

Hand, D., H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

Hellmuth, O. Conceptual study on nucleation burst evolution in the convective boundary layer part I: Modelling approach. *Atmos. Chem. Phys. Discuss.*, 5:11413–11487, 2005.

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. A look at aerosol formation using data mining techniques. *Atmospheric Chemistry and Physics*, 5:3345–3356, 2005.

Korhonen, P., M. Kulmala, A. Laaksonen, Y. Viisanen, R. McGraw, and J. Seinfeld. Ternary nucleation of H_2SO_4, NH_3 and H_2O in the atmosphere. *J. Geophys. Res.*, 104:26349–26353, 1999.

Kulmala, M., V.-M. Kerminen, T. Anttila, A. Laaksonen, and C. O'Dowd. Organic aerosol formation via sulphate cluster activation. *J. Geophys. Res.*, 109:10.1029/2003JD003961, 2004.

Nilsson, E. D., J. Paatero, and M. Boy. Effects of air masses and synoptic weather on aerosol formation in the continental boundary layer. *Tellus*, 53B, 2001.

O'Dowd, C. D., P. Aalto, K. Hämeri, M. Kulmala, and T. Hoffmann. Atmospheric particles from organic vapours. *Nature*, 416:497–498, 2002.

Pirjola, L. and M. Kulmala. Modelling the formation of H2SO4-H2O particles in rural, urban and marine conditions. *Atmos. Res.*, 46, 1998.

Rannik, U. On the surface layer similarity at a complex forest site. *J. Geophys. Res.*, 103(D8):8685–8697, 1998.

Vesala, T., J. Haataja, P. Aalto, N. Altimir, G. Buzorius, and et al. Long-term field measurements of atmosphere-surface interactions in boreal forest ecology, micrometeorology, aerosol physics and atmospheric chemistry. *Trends in Heat, Mass and Momentum Transfer*, 4:17–35, 1998.

Weber, R. J., P. H. McMurry, F. L. Eisele, and D. J. Tanner. Measurements of expected nucleation precursor species and 3-500-nm diameter particles at Mauna Loa observatory, Hawaii. *J. Atmos. Sci.*, 52:2242–2257, 1995.