Atmos. Chem. Phys., 6, 5549–5557, 2006
www.atmos-chem-phys.net/6/5549/2006/
© Author(s) 2006. This work is licensed
under a Creative Commons License.

Atmospheric Chemistry and Physics

Using discriminant analysis as a nucleation event classification method

S. Mikkonen[1], **K. E. J. Lehtinen**[1,2], **A. Hamed**[1], **J. Joutsensaari**[3], **M. C. Facchini**[4], **and A. Laaksonen**[1]

[1]Department of Physics, University of Kuopio, P.O. Box 1627, 70211 Kuopio, Finland
[2]Finnish meteorological institute, Kuopio Unit, P.O. Box 1627, 70210 Kuopio, Finland
[3]Department of Environmental Sciences, University of Kuopio, P.O. Box 1627, 70211 Kuopio, Finland
[4]Istituto di Scienze dell'Atmosfera e del Clima – CNR, Italy Via Gobetti 101, 40 129 Bologna, Italy

Received: 24 August 2006 – Published in Atmos. Chem. Phys. Discuss.: 7 September 2006
Revised: 13 November 2006 – Accepted: 5 December 2006 – Published: 11 December 2006

Abstract. More than three years of measurements of aerosol size-distribution and different gas and meteorological parameters made in Po Valley, Italy were analysed for this study to examine which of the meteorological and trace gas variables effect on the emergence of nucleation events. As the analysis method, we used discriminant analysis with non-parametric Epanechnikov kernel, included in non-parametric density estimation method. The best classification result in our data was reached with the combination of relative humidity, ozone concentration and a third degree polynomial of radiation. RH appeared to have a preventing effect on the new particle formation whereas the effects of O_3 and radiation were more conductive. The concentration of SO_2 and NO_2 also appeared to have significant effect on the emergence of nucleation events but because of the great amount of missing observations, we had to exclude them from the final analysis.

1 Introduction

One of the central topics in atmospheric research is the effects of aerosols on climate change. Aerosol particles influence cloud formation and absorb or scatter solar radiation. It is well known that new particle formation can occur almost everywhere in the atmosphere (Kulmala et al., 2004) but despite of several years of investigation, many of the processes and factors behind the new particle formation in the atmosphere remain unclear.

The use of statistical methods has been almost non-existent in the investigation of nucleation events, although they are a powerful tool in the analysis of large measurement datasets. Most studies on ambient nucleation events have investigated only physical or chemical mechanisms of nucleation (e.g. O'Dowd et al., 2002; Boy and Kulmala, 2002). Hyvönen et al. (2005) introduced some statistical data mining methods to explain new particle formation in Hyytiälä station in Finland. However, their model is not as successful for our data, recorded in Po Valley, Italy (Hamed et al., 2006), which may be due to different environmental factors and/or differences in the amount of pollution between the measurement stations. The key variables of our model were relative humidity, radiation and ozone concentration, whilst in Hyvönen et al. (2005) only relative humidity and condensation sink were sufficient predictors for the best classification.

The aim of our study was to find a parameterization that is suitable for nucleation event classification for more than three years of measurements made in highly polluted Po Valley area in Italy. The measurements are discussed briefly in Sect. 2 but more details can be found in Hamed et al. (2006). We constructed a statistical discriminant analysis model with non-parametric kernel density estimate, described also in Sect. 2, and tested the model with several different combinations of trace gas and meteorological variables. The results of the model are introduced in Sect. 3 including the tests of the accuracy of the classification and the comparison of final parameterization with the parameterization from Hyvönen et al. (2005). The results are discussed in Sect. 4 and finally, in Sect. 5, general conclusions of the paper are drawn.

2 Methods

Our dataset consists of measurements made in 24 March 2002–30 April 2005 at San Pietro Capofiume (SPC) station in the Po Valley area, Italy. The event classification was made from the particle size-distribution measurements, which were carried out using a twin Differential Mobility Particle Sizer (DMPS) system. The DMPS system was operational on 814 days during the time period, which included 293 event days and 270 nonevent days, and 251 days that could not be

Correspondence to: S. Mikkonen
(santtu.mikkonen@uku.fi)

Published by Copernicus GmbH on behalf of the European Geosciences Union.

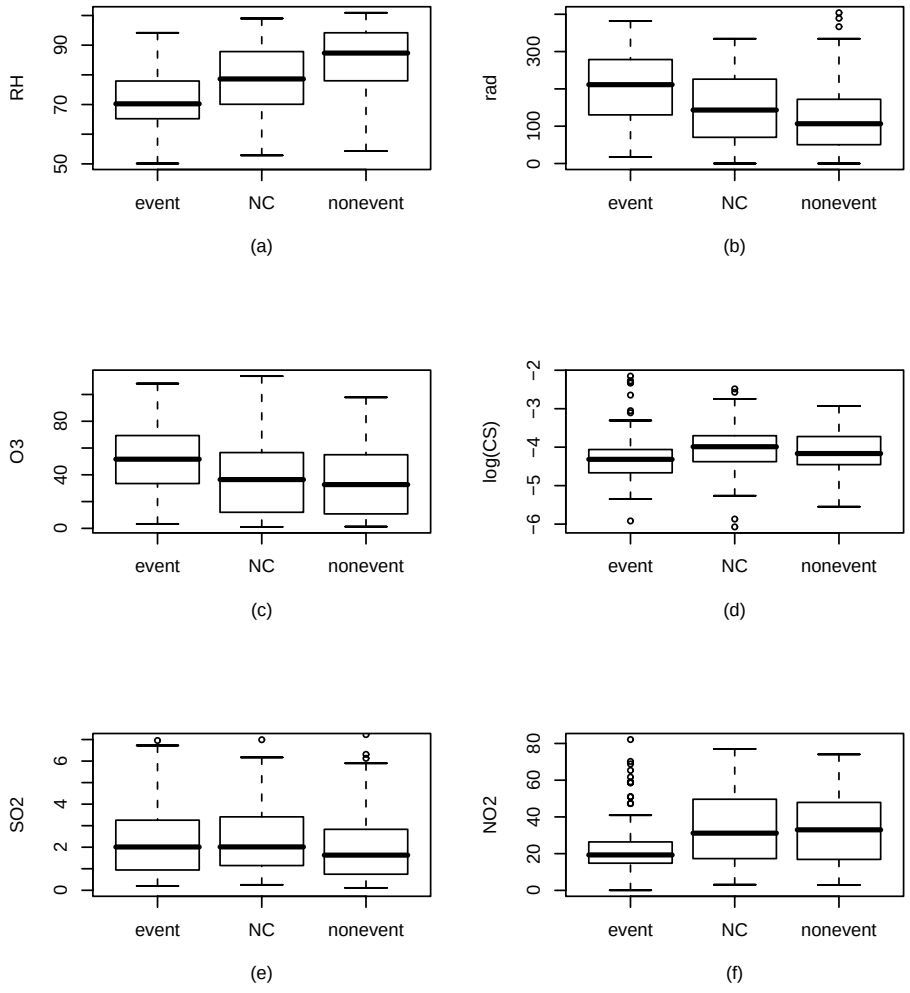

Fig. 1. Distributions of the predictor variables in event, nonevent and unclassified (NC) days. The length of the box represents the difference between the 25th and 75th percentiles, the horizontal line inside the box represents the median, the lengths of the dashed lines correspond to the largest and smallest values that are not outliers, and the outliers, labelled with o, are cases with the values more than 1.5 box-lengths from the 75th percentile or 25th percentile.

classified. A day is considered an event day if the formation of new aerosol particles starts in the nucleation mode size range and the mode is observed over a period of several hours showing signs of growth. If no new particle formation has been observed, the day is classified as a non-event day (NE). A large number of days did not fulfil the criteria to be classified either clear event or NE day and they are considered as unclassified days (NC). The classification method of nucleation events we used was visual analysis, based on the methods described by Mäkelä et al. (2000) and Dal Maso et al. (2005).

As predictor variables we used several different gas and meteorological parameters measured at SPC, including SO_2, NO, NO_2, NO_x, O_3, temperature, relative humidity (RH), wind direction and speed, global radiation, precipitation, and atmospheric pressure. During the measurement period, there were some missing data as well as some bad quality data. Therefore, the actual number of days used in the analysis was decreased. More details of the measurements and event classification can be found in Hamed et al. (2006). The daily averages used in this analysis are made from the whole 24 h of each day because we did not want to lose any information about conditions that might affect the occurrence of nucleation events by limiting the time span. Several different time windows were tested, including daylight hours (used in Hyvönen et al., 2005) and morning hours i.e. hours before and during the usual event start times, but 24 h averages gave the best classification results. It appears that 24 h averages are the best estimators for the conditions needed for nucleation taking place.

Figure 1 illustrates the distributions of the predictor values. Radiation (Wm^{-2}) (Fig. 1b) and ozone concentration ($\mu g\,m^{-3}$) (Fig. 1c) are clearly higher during event days than during nonevent days, whereas relative humidity (Fig. 1a) is

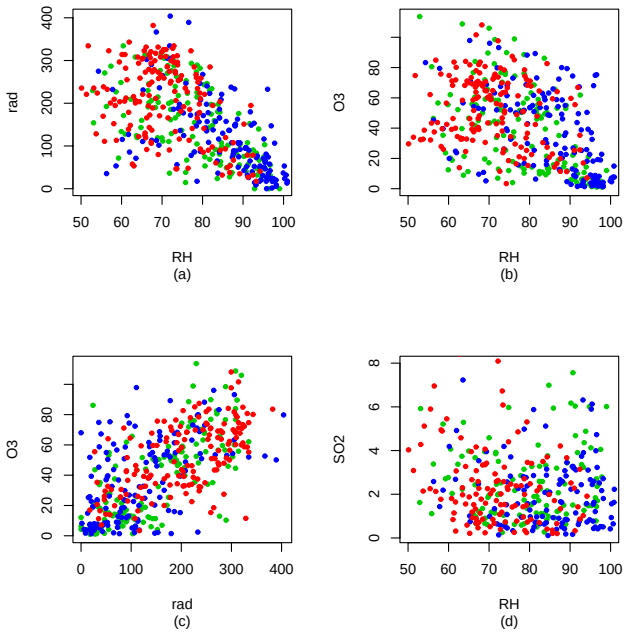

Fig. 2. Daily means of different predicting pairs of variables. Events are red, nonevents are blue and unclassified days are green.

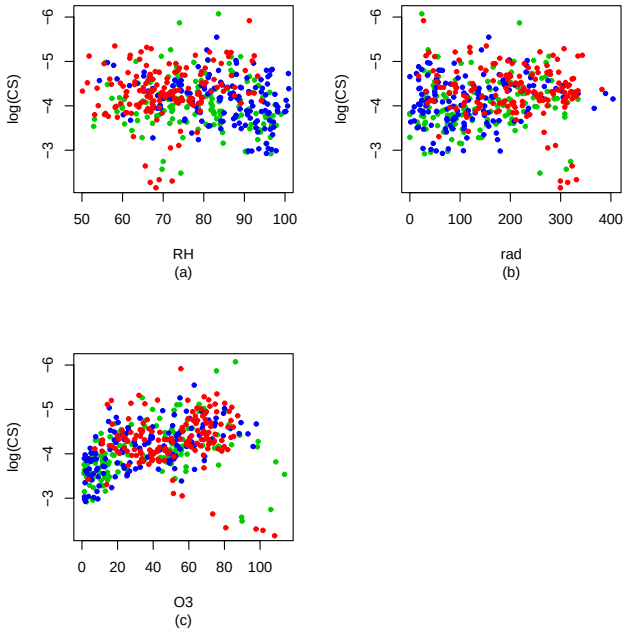

Fig. 3. Daily means logarithm of condensation sink paired with different predicting variables. Events are red, nonevents are blue and unclassified days are green.

lower during event days than during nonevent days. The difference between event and nonevent days in the distribution of the natural logarithm of condensation sink (Fig. 1d) is not so clear, which makes it an inadequate classification variable. The distribution of the concentration of SO_2 ($\mu g\,m^{-3}$) is also quite similar within different event classes (Fig. 1e) but the number of observations is low especially in event days, which may cause some bias to the distribution. The concentration of NO_2 ($\mu g\,m^{-3}$) is clearly lower during event days (Fig. 1f) but again it is questionable if the number of observations is sufficient.

Favourable conditions for nucleation events can be observed from Fig. 2. It is evident that high relative humidity is a preventing factor for the events, whereas radiation seems to have a clear positive effect on the emergence of nucleation events (Fig. 2a). Radiation and relative humidity also have a significant negative correlation with each other but the correlation is not strong enough to cause multicollinearity, so they can safely be used in the same model. High concentration of O_3 combined with low RH (Fig. 2b) and high radiation (Fig. 2c) seem also to have a positive effect on the emergence of the events. Ozone and radiation are clearly correlated (Fig. 2c), which is no surprise as $O_3$is formed photochemically, but as with the anticorrelation of Fig. 2a, simultaneous use of the two variables in the model is not prevented by multicollinearity. As stated earlier, the logarithm of the condensation sink seems to be an inadequate classification variable for our data. Figure 3 illustrates the classification ability of the (natural) logarithm of the condensation sink with different pairs. It can also be seen from Fig. 3c that the condensation sink and O_3 have a positive correlation, which can be expected because high levels of particulate matter and O_3 are commonly observed during pollution events.

2.1 Discriminant analysis

Several applicable methods have been introduced for classifying quantitative variables. In this paper, we used Discriminant Analysis (DA) with non-parametric Epanechnikov kernel (Epanechnikov, 1969) to find factors that classify the days as nucleation event days or nonevent days. The notation of this and the following section refers to SAS Institute Inc. (2004). We used two different methods to test the goodness of fit of the models: resubstitution, where the computed model is fitted to the same dataset from which it was estimated, and cross-validation, where the model is fitted to a different dataset than the one used in the estimation.

Discriminant analysis is a multivariate statistical analysis method, which is commonly used to build a predictive or descriptive model of group discrimination based on observed predictor variables and to classify observations into the groups. If the distribution within each group is multivariate normal, a parametric (linear or quadratic) method can be used to develop a discriminant function. Non-parametric discriminant methods are used when the normality assumption cannot be made. Non-parametric methods are based on group-specific probability densities and they are used to produce a classification criterion based on those probabilities. In our case, when analysing aerosol measurement data, the

Table 1. Resubstitution table for two different models.

Resubstitution			
predictors	Classification error	Missed events	False events
RH, O3, Radiation	3.36%	1.27%	5.67%
RH, log(CS)	22.82%	12.1%	34.7%

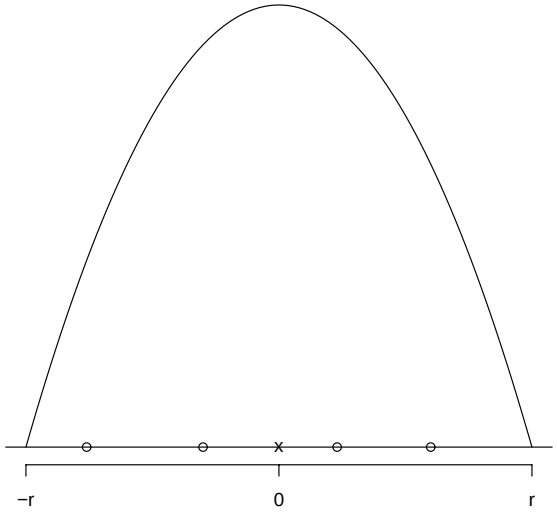

Fig. 4. Epanechnikov kernel in 1-dimensional case. The observation from the test set, marked with x, is classified in the group that has most observations from the training set, marked with o, within the radius r.

normality assumption is not realistic and we have to use a non-parametric kernel method. The non-parametric method is also more robust for multicollinearity, which might occur in this kind of analysis, where some variables measure partly the same effect.

2.1.1 Kernel estimation

The purpose of kernel estimation is to estimate the density function of observations without any distribution assumption. The proximity of observations is needed in kernel estimation. The squared distance between two observation vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, in group t is given by

$$d_t^2(\boldsymbol{x}, y) = (\boldsymbol{x} - y)' V_t^{-1} (\boldsymbol{x} - y),$$

where V_t is in our case the covariance matrix within the event classification group t.

The classification of an observation vector $\boldsymbol{x}$ is based on the estimated group-specific densities from the data. From these estimated densities, the posterior probabilities of group membership at $\boldsymbol{x}$ are evaluated. An observation $\boldsymbol{x}$ is classified into group u if setting $t{=}u$ produces the largest value of conditional probability $\mathrm{p}(t|\boldsymbol{x})$.

The kernel method uses a fixed radius, r, and a specified kernel, K_t, to estimate the group t density at each observation vector $\boldsymbol{x}$. Let $\boldsymbol{z}$ be a p-dimensional vector. Then the volume of a p-dimensional unit sphere bounded by $z'\boldsymbol{z}{=}1$ is

$$v_0 = \frac{\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}+1\right)}$$

where Γrepresents the gamma function.

Thus, in group t, the volume of a p-dimensional ellipsoid bounded by $\left\{\boldsymbol{z}|z' V_t^{-1} \boldsymbol{z}{=}r^2\right\}$ is

$$v_r\,(t) = r^p\,|V_t|^{\frac{1}{2}}\,v_0.$$

Several applicable functions for the kernel density have been defined. Out of these, the one that has been tested the most in numerous different statistical applications is the Epanechnikov kernel, given by

$$K_t\,(\boldsymbol{z}) = \begin{cases} c_1(t)\left(1-\frac{1}{r^2} z' V_t^{-1} z\right) & \text{if } z' V_t^{-1} \boldsymbol{z} \le r^2 \\ 0 & \text{elsewhere} \end{cases}$$

where

$$c_1\,(t) = \frac{1}{v_r\,(t)}\left(1+\frac{p}{2}\right).$$

The group t density at $\boldsymbol{x}$ is estimated by

$$f_t\,(x) = \frac{1}{n_t} \sum_y K_t\,(\boldsymbol{x} - y)$$

where n_t is the number of observations in group t, the summation is over all observations $\boldsymbol{y}$ in group t, and K_t is the specified kernel function. The posterior probability of membership in group t is then given by

$$p\,(t|x) = \frac{q_t\, f_t\,(x)}{f\,(x)}$$

where $f\,(x) = \sum_u q_u\, f_u\,(x)$ is the estimated unconditional density and q_t is the prior probability of group t. If the closed ellipsoid centred at $\boldsymbol{x}$ does not include any training set observations, $f(x)$ is zero and x cannot be classified in any of the groups, otherwise $\boldsymbol{x}$ is classified in the group that has the largest number of observations in the closed ellipsoid. The principle of the Epanechnikov kernel in 1-dimensional situation is illustrated in Fig. 4. The observation from the test set, marked with x, is classified in the group that has most observations from the training set, marked with o, within the radius r.

Table 2. Resubstitution table for the models for three-class data.

Cross-validation, 1000 simulations					
predictors		Mean	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean
RH, O3, Radiation	Total error	17.63%	0.0277	17.46	17.80
	missed events	15.24%	0.0436	14.97	15.52
	false events	20.29%	0.0531	19.96	20.62
RH, log(CS)	Total error	23.21%	0.0249	23.05	23.36
	missed events	13.67%	0.0395	13.42	13.91
	false events	33.83%	0.0495	33.52	34.14

Table 3. Total classification errors in cross-validation and misclassification rates for the models for the three-class data.

Resubstitution, three-class data					
predictors	Classification error	Missed events	nonevent to event	NC to event	NC to nonevent
RH, O_3, Radiation	13.47%	1.91%	4.96%	7.14%	2.86%
RH, log(CS)	43.15%	16.56%	31.21%	46.43%	29.29%

3 Results

In the first phase, we leave the unclassified days out of the analysis and try to separate the event and nonevent days. Hyvönen et al. (2005) presented a two-variable model for the data from Hyytiälä, Finland. In their model, Relative Humidity and the natural logarithm of the Condensation sink explained 88% of the nucleation events with a total classification error of 12%, but for our data, this model was too simplified. The model still explained almost 88% of the events in resubstitution but it also gave a large number of false events (i.e. predict a nonevent day to be an event day), which increased the total classification error to 22% (Table 1).

Since this was the best two-variable model found, we had to increase the number of predictors to get an applicable classification. With our model, we could explain almost 99% of the nucleation event days in resubstitution with the information from three different factors: relative humidity, which was the most significant variable, radiation, and ozone concentration. The effect of radiation was estimated with a third degree polynomial. The concentrations of SO_2 and NO_2 were also adequate predictors for the model but because of a great number of missing observations, they could not be used. The total classification error for the model was 3.36%.

Re-substitution is a good way to find the best model for the current data but if one needs to know how the model performs with different datasets, a cross-validation method should be used. For cross-validation, we constructed 1000 training sets and 1000 test sets from the original data with Bootstrap re-sampling method (Efron and Tibshirani, 1993). We computed both models for all training sets and tried to predict the event distributions of the test sets with the results. We computed the mean, standard deviation and 95% confidence interval from the total classification errors and misclassification rates from the 1000 repeated estimations. The two-variable model misclassifies on average 13.7% of the event days but it also predicts a great number of false events, which increases the total classification error to 23.2% (Table 2).

The model with RH, radiation and O_3 included missed a few more events than the two-variable model but it also produced fewer false events. The total mean classification error from 1000 repeats is 17.6%, which is clearly better than in the two-variable model, and we can say that the three-predictor model fits better to cross-validation data.

The flaw in this approach is that we did not take into account the days that were unclassified in the data. That kind of restriction would have lead to loss of almost one third of our data. Since we are planning to use the classification information in further analysis, we also needed to take into account the unclassified days. This is done by using the discriminant analysis to a three-class event-variable, where classes are event, nonevent and unclassified, instead of only two class variables, event and nonevent days.

As the unclassified class is not exactly independent from the event and nonevent classes, the discriminant analysis becomes slightly more unstable and the classification is not as good as it was with the restricted data. The total classification error for the three-predictor and three-class model was 13.5%, and the misclassification rate for event days was 1.9% (Table 3). The model had some difficulties in separating nonevent days from unclassified days. This happens probably

Table 4. Resubstitution table for model 1 for three-class data where all days are classified as event or non-event days.

Cross-validation, 1000 simulations, three-class data					
predictors		Mean	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean
RH, O3, Radiation	Total error	31.78%	0.0276	31.61	31.95
	missed events	24.56%	0.0473	24.26	24.85
	event to nonevent	6.61%	0.0269	6.44	6.78
	nonevent to event	7.17%	0.0293	6.99	7.35
RH, log(CS)	Total error	45.08%	0.0242	44.93	45.23
	missed events	19.83%	0.0506	19.51	20.14
	event to nonevent	11.88%	0.0391	11.64	12.12
	nonevent to event	29.16%	0.0501	28.85	29.48

Table 5. Daily means of different predicting pairs of variables. Events are red, nonevents are blue and unclassified days are green.

Resubstitution, NC days re-classified with the model					
predictors	Missed events	nonevent to event	NC to event	NC to nonevent	NC classification Failed
RH, O3, Radiation	1.27%	5.67%	27.86%	55.71%	16.43%
RH, log(CS)	12.1%	33.33%	54.29%	44.29%	1.43%

because a great number of unclassified days are most likely nonevent days.

The performance of the two-variable model with three classes is also worse than with two classes. The total classification error of the model is 43.2% even though it predicts correctly 83.3% of the event days. The two-variable model cannot distinguish unclassified days from event and non-event days in our data, as it classifies almost half of the unclassified days into event days. It also classifies 10.8% of event days into nonevent days and 31.2% of nonevent days into event days.

Cross-validation for the three-class data in Table 4 shows that adding the unclassified days to the analysis makes the analysis slightly more unstable and the classification errors increase. The three-predictor model misses again a few more events but most of those missed events are classified into unclassified group, while the two-variable model predicts most of the missed events into nonevents. In addition, the two-variable model produces a great number of false events.

It is also possible to force the model to re-classify all of the unclassified days into either event or nonevent days. To do this, we have to assume that every day is either an event day or nonevent day and the unclassified days in the data are only a result of insufficient classification method. As a result of this assumption we could make a resubstitution where event and nonevent days classified as they did when there were only two classes in the analysis and 27.86% of the unclassified days classified into event days (Table 5). The model fails to resubstitute 23 NC days; this is due to a tie for the largest group-classification probability. Resubstitution failures could be avoided by changing the width of the kernel, but for comparability we wanted to use the same kernel in every model. As already seen in Table 3, the two-variable model assumes that most of unclassified days (54.29%) are event days.

This resubstitution gives an estimate for the numbers of unclassified days, which should actually be classified into event and nonevent days. It is commonly assumed that most of the unclassified days are nonevent days, as the three-predictor model suggests. The two-variable model tends to overestimate the number of event days and classifies more than half of the unclassified days to event days.

4 Discussion

The best classification result in our data was attained with the combination of RH, O_3 and a third degree polynomial of radiation. Relative humidity was found to be a preventing factor for nucleation events, which has also been suggested in previous studies (e.g. Birmili et al., 2003; Boy and Kulmala, 2002). On the other hand, radiation and ozone concentration were found to have a positive effect on the new formation of particles. Radiation is known to be an essential factor in nucleation events (e.g. Birmili et al., 2003; Woo et al., 2001) and several previous studies support our finding that ozone is

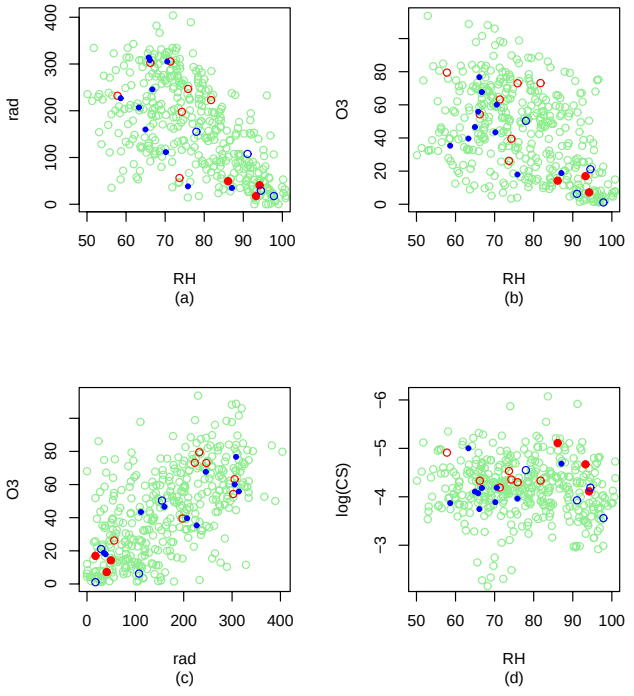

Fig. 5. The success of the Three-predictor model in comparison to the daily means of different predictor variables: green indicates correct prediction, solid red indicates missed event, red circle indicates nonevent day classified to event, solid blue indicates NC day classified to event, blue circle indicates NC day classified to nonevent.

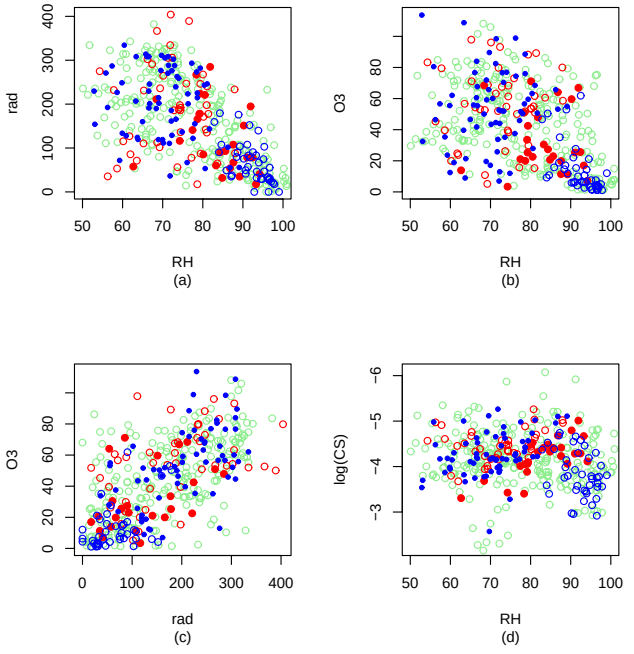

Fig. 6. The success of the two-variable model in comparison to the daily means of different predictor variables: green indicates correct prediction, solid red indicates missed event, red circle indicates nonevent day classified to event, solid blue indicates NC day classified to event, blue circle indicates NC day classified to nonevent.

a good indicator for new particle formation (e.g. Rodriguez et al., 2005).

Our parameterization differs greatly from the results of Hyvönen et al. (2005), who obtained the best results by using RH and condensation sink, both factors tending to prevent nucleation. The data used in their analysis had been collected from Hyytiälä, Finland, where the air is rather clean. The model with RH and condensation sink as its predictors was far too simple for our data. It appears that in highly polluted areas, like Po Valley, different predictors are needed to make an applicable classification. The concentrations of SO_2 and NO_2 also appeared to have a significant effect on the emergence of nucleation events, but because of the great number of missing observations, we had to exclude them from the final analysis. In the models where SO_2 or NO_2 was included, the significance of O_3 was reduced. Particularly NO_2 and O_3 measure the effect of pollution in nucleation and particle growth processes. It is also known that the concentration of SO_2 affects the nucleation of new particles as it is a precursor for sulphuric acid, whilst O_3 is assumed to have greater effect on the growth of new particles as it is an oxidising agent for VOC's, affecting thus the production of condensable organic species (Kulmala et al., 2003). When comparing the significance of these three variables, SO_2, NO_2 and O_3, it was seen that O_3 is clearly the most significant predictor of these three in our data, and SO_2, NO_2 have almost equal classification ability.

The three-predictor model gives an adequate resubstitution to the data; it misses only three events from the three-class data. Figure 5 shows that missed events, marked with solid red dots, are produced in a situation where relative humidity is high, while radiation and the ozone concentration are low, which is the exact opposite of the usual favourable conditions for an event day. (It appears that in two out of three of these events, the radiation levels peaked strongly just before event start. Remember, that we use 24 h averages in our modelling.) The distributions of predictor variables were illustrated in Figs. 1 and 2. Nonevent days classified as event days, marked with red circles, do not show as clear pattern even though relative humidity is lower than on average nonevent day in all cases and radiation is higher than on average nonevent days in six days out of seven. In addition, no clear pattern was detected with false events and false nonevents produced from the unclassified days. It is notable that all missed events were observed in December 2002 and they were all classified as weak events on the scale of Hamed et al. (2006).

For the two-variable model, the proportion of missed events increases when the relative humidity increases (Fig. 6a), whilst radiation and concentration of ozone decreases (Fig. 6c), just as in the three-predictor model, though the pattern is not as clear. The logarithm of condensation

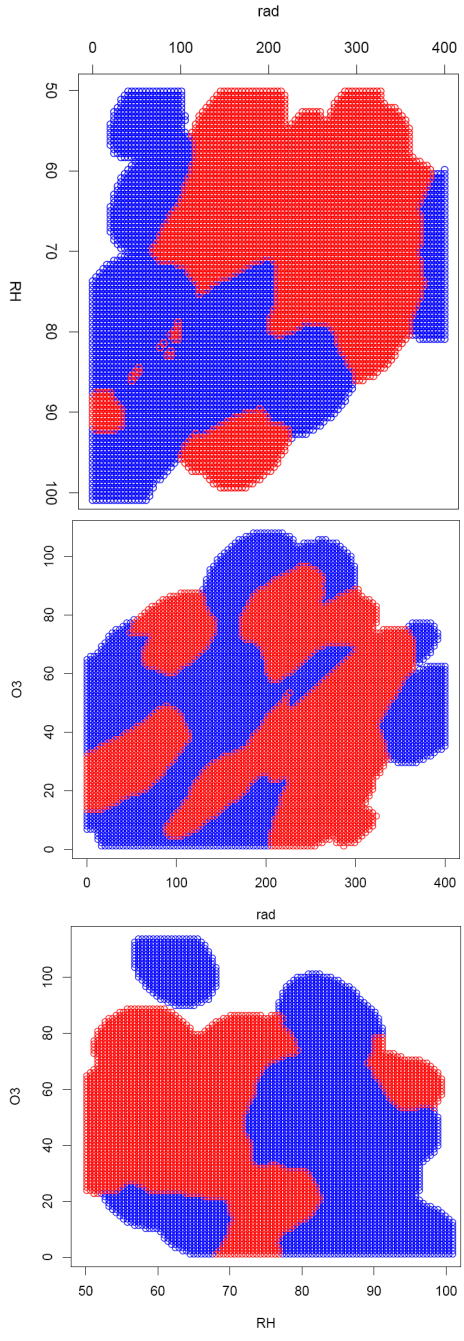

Fig. 7. Performing of the three-predictor model in simulated grids when one of the variables is set to constant. Red indicates predicted events and blue indicates predicted nonevents. In white areas, there are no training set observations nearby and prediction cannot be made.

sink, which was used as a predictor in two-variable model, seems not to have any effect on the occurrence of the prediction errors (Fig. 6d). As we saw from Fig. 1 and Fig. 3, the classification ability of the condensation sink in our data is questionable.

To demonstrate the performance of the discriminant function, we used two-dimensional grids, where the third predictor was set to constant, as test sets in the analysis. Figure 7a illustrates a situation where the ozone concentration is fixed to 60 $\mu g\,m^{-3}$, in Fig. 7b RH is set to 80%, and in Fig. 7c radiation is set to 150 $W\,m^{-2}$. It can be seen that in these cross-sections the areas where grid points are classified into event or nonevent are not continuous, however, in a three-dimensional grid they form a continuous volume when all variables are let to vary freely.

5 Conclusions

We analyzed a dataset collected from Po Valley, Italy during a period 24 March 2002–30 April 2005. Our findings show that in polluted areas, like Po Valley, more complicated processes control the emergence of the nucleation events than in clean areas. Our findings show, that high relative humidity has a preventing effect on the occurrence of new particle formation, while high radiation has a positive effect on the probability of nucleation events and helps the particles grow to detectable sizes. High ozone concentrations are detected on nucleation event days, and it is a good indicator for new particle formation, but it is not known for sure if it participates into the nucleation process. It is possible that O_3 oxidises VOC's, which produces condensable organic compounds (Hamed et al., 2006; Kulmala et al., 2004a) and thus participates into new particle formation.

As the analysis method of our study, we used discriminant analysis with Epanechnikov kernel, included in nonparametric density estimation method, to examine which of the meteorological and trace gas variables effect on the emergence of nucleation events. The best classification result in our data was reached with the combination of RH, O_3 and a third degree polynomial of radiation. The concentrations of SO_2 and NO_2 also appeared to have significant effects on the emergence of nucleation events but because of great amount of missing observations, we had to exclude them from the final analysis.

It is somewhat surprising that both radiation and ozone concentration belong to the set of the three variables that give the best statistical explanation of nucleation event occurrence, as ozone is generated in photochemical reactions. However, ozone concentrations obviously are dependent also on other factors than radiation, and may therefore yield extra information to the statistical model, which is likely related to concentrations of oxidized organics able to participate in fresh particle growth.

The model is easy to implement into an atmospheric model, which can be used to investigate the effect of nucleation events on the local aerosol budget, and therefore it is also important to have the unclassified days included in the analysis. However, even though this parameterization is the best possible in our data it may not be the best everywhere

(and might not be for our dataset if additional gas-phase parameters were available). It has already been shown that in clean boreal forest area only two parameters, RH and condensation sink, are needed for adequate classification so it is clear that additional work will be required before more general model can be presented.

Acknowledgements. This work was supported by Graduate school in Physics, Chemistry, Biology and Meteorology of Atmospheric composition and climate change.

Edited by: K. Hämeri

References

Birmili, W. , Berresheim, H., Plass-Dülmer, C., Elste, T., Gilge, S., Wiedensohler, A., and Uhrner, U.: The Hohenpeissenberg aerosol formation experiment (HAFEX): a long-term study including size-resolved aerosol, H2SO4, OH, and monoterpenes measurements, Atmos. Chem. Phys., 3, 361–376, 2003, http://www.atmos-chem-phys.net/3/361/2003/.

Boy, M. and Kulmala, M.: Nucleation events in the continental boundary layer: influence of physical and meteorological parameters, Atmos. Chem. Phys., 2, 1–16, 2002, http://www.atmos-chem-phys.net/2/1/2002/.

Dal Maso, M., Kulmala, M., Riipinen, I., Wagner, R., Hussein, T., Aalto, P. P., and Lehtinen, K. E. J.: Formation and growth of fresh atmospheric aerosols: eight years of aerosol size distribution data from SMEAR II, Hyytiälä, Finland, Boreal Environ. Res., 10, 323–336, 2005.

Hamed, A., Joutsensaari, J., Mikkonen, S., Sogacheva, L., Dal Maso, M., Kulmala, M., Cavalli, F., Fuzzi, S., Facchini, M. C., Decesari, S., Mircea, M., and Laaksonen, A.: Nucleation and growth of new particles in Po Valley, Italy, Atmos. Chem. Phys. Discuss., 6, 9603–9653, 2006, http://www.atmos-chem-phys-discuss.net/6/9603/2006/.

Hyvönen, S., Junninen, H., Laakso, L., Dal Maso, M., Grönholm, T, Bonn, B., Keronen, P., Aalto, P., Hiltunen, V., Pohja, T., Launiainen, S., Hari, P., Mannila, H., and Kulmala, M.: A look at aerosol formation using data mining techniques, Atmos. Chem. Phys., 5, 3345–3356, 2005, http://www.atmos-chem-phys.net/5/3345/2005/.

Efron, B. and Tibshirani, R.: An Introduction to the Bootstrap, London Chapman and Hall, 1993.

Epanechnikov, V. A.: Nonparametric estimation of a multidimensional probability density. Theory Probab. Appl., 14, 153–158, 1969.

Kulmala, M., Kerminen, V.-M., Anttila, T., Laaksonen, A., and O'Dowd, C.: Organic aerosol formation via sulphate cluster activation, J. Geophys. Res., 109, D04205, doi:10.1029/2003JD003961, 2004a.

Kulmala, M., Vehkamäki, H., Petäjä, T., Dal Maso, M., Lauri, A., Kerminen, V.-M., Birmili, W., and McMurry, P.: Formation and growth rates of ultrafine atmospheric particles: a review of observations, J. Aerosol Sci., 35, 143–176, doi:10.1016/j.jaerosci.2003.10.003, 2004b.

Morrison, D. F.: Multivariate Statistical Methods, Belmont CA: Thomson/Brooks/Cole, 2005.

O'Dowd, C. D., Aalto, P., Hämeri, K., Kulmala, M., and Hoffman, T.: Atmospheric particles from organic vapours, Nature, 416, 497–498, doi: 10.1038/416497a, 2002.

Mäkelä, J., Dal Maso, M., Pirjola, L., Keronen, P., Laakso, L., Kulmala, M., and Laaksonen, A.: Characteristics of the atmospheric particle formation events observed at a boreal forest site in southern Finland, Boreal Environ. Res., 5, 299–313, 2000.

Rodríguez, S., Van Dingenen R., Putaud, J.-P., Martins-Dos Santos, S., and Roselli D.: Nucleation and growth of new particles in the rural atmosphere of Northern Italy–relationship to air quality monitoring. Atmospheric environment, 39(36), 6734–6746, doi:10.1016/j.atmosenv.2005.07.036, 2005.

SAS Institute Inc.: SAS/STAT User's Guide, Version 9.1, SAS Publishing, 2004.

Woo, K. S., Chen, D. R., Pui, D. Y. H., and McMurry, P. H.: Measurement of Atlanta aerosol size distributions: Observations of ultrafine particle events, Aerosol Sci. Technol., 34, 75–87, doi:10.1080/02786820120056, 2001.