Atmos. Chem. Phys. Discuss., 5, S3904–S3906, 2005
www.atmos-chem-phys.org/acpd/5/S3904/
European Geosciences Union

Interactive comment on “A look at aerosol formation using data mining techniques” *by* S. Hyvönen et al.

S. Hyvönen et al.

Received and published: 15 November 2005

We thank the referees for their comments.

Author response to referee 1

1. Our statement in the end of section 2.2.1 that "the RH dependency of the condensation sink is stronger than the temperature effect" is misleading: neither dependency is particularly strong, as is evident for the former pair of variables by looking at Fig. 4. The two last sentences in section 2.2.1 could be replaced by "Thus, the condensation sink depends mainly on the ambient particle size distribution."

However, other correlations in the data do exist. We have removed very strongly correlating variables from the data, but the nature of atmospheric data is such that all

Full Screen / Esc

Print Version

Interactive Discussion

Discussion Paper

correlations cannot be removed without discarding most of the data. One could use e.g. PCA to remove multicollinearity (as mentioned in Sect. 3.2 in the context of clustering), but this sacrifices interpretability. Instead, we have taken the problem into account when interpreting our results. For example, we might have the problem that LDA will not always correctly assess the relative importance of the variables, which is related to the unreliability of forward selection of variables with LDA as presented in Table 6. However, LDA is still useful in determining the best pair of variables, which we are mainly interested in; the relative importance of each pair of variables is not of primary interest.

2. In our case standardization is necessary, for our data consists of variables of very different nature, measured in very different units. The standard deviations of the variables range from 0.06 to 230. Leaving them as is would mean that the variables with large std's would dominate the whole analysis. The only reasonable approach is to use standardization.

3. To make sure leaving out the undefined days does not bias our data significantly we checked the distributions of values of our key varaibles in both sets. They were not significantly different. The proportion of days with low RH and low CS, that is with high nucleation parameter values was slightly higher in the defined set, while the proportion of days with low nucleation parameter values was higher. This can be explained as a result of the event classification process: undefined days are often days with high particle concentration which makes event status classification more difficult, and also is linked to a high condensation sink. Figures illustrating this can be included in the final version if desired.

Author response to referee 2

The referee is right to point out that days in the "cloudy days" cluster are in fact "low radiation days" - the latter seemed overly formal to us, but since the former can be misleading it is resonable to rename the cluster.

Interactive comment on Atmos. Chem. Phys. Discuss., 5, 7577, 2005.

Full Screen / Esc

Print Version

Interactive Discussion

Discussion Paper