Atmos. Chem. Phys. Discuss., 5, S2314–S2315, 2005
www.atmos-chem-phys.org/acpd/5/S2314/
European Geosciences Union

Atmospheric Chemistry and Physics Discussions

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

Anonymous Referee #1

Received and published: 31 August 2005

General comments

This paper is in a good road to increase the usage of data mining techniques and other statistical methods in the analysis of aerosol datasets. Data Mining is an efficient way to find new and useful insights out of large masses of data and the methods of multivariate analysis, such as principle component analysis and discriminant analysis used in this paper, help to find significant interdependences where the single variable methods fail.

Specific comments

1. In Section 2.2.1. it is stated that there exists a strong dependency between Con-

densation sink and relative humidity, but they are still used simultaneously in linear discriminant analysis and logistic regression. Is there not a danger of multicollinearity among them? The same danger could arise when the concentration of H2O and relative humidity are used in the same model, like reported in Table 4.

2. Page 7583 lines 16-19: Standardization of variables does not change the statistical significance of variables if it is made properly. In some cases, it may even lose some information; a variable may possess enhanced variation because it nicely separates two or more clusters in the data (Milligan, 1995). Standardization would decrease the contribution of such a variable. Standardization of variables should be argued more precisely.

3. Just for curiosity; how would the 'undefined' days be distributed in Figure 4?

References

Milligan, G. W.: Issues in Applied Classification: Variable Standardization. Classification Society of North America Newsletter, Issue #38, February 1995.

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