EGU2020-2355, updated on 01 Aug 2021
https://doi.org/10.5194/egusphere-egu2020-2355
EGU General Assembly 2020

Data mining and machine learning to enhance new-particle formation identification and analysis

Martha A. Zaidan[1], Pak L. Fung[1], Darren Wraith[2], Tuomo Nieminen[1], Tareq Hussein[1,3], Veli-Matti Kerminen[1], Tuukka Petäjä[1], and Markku Kulmala[1]

[1]Helsinki University, Institute for Atmospheric and Earth System Research, Physics, Helsinki, Finland (martha.zaidan@helsinki.fi)

[2]School of Public Health and Social Work, Queensland University of Technology, Queensland 4000, Australia

[3]Department of Physics, The University of Jordan, Amman 11942, Jordan

Data Mining (DM) and Machine Learning (ML) have become very popular modern statistical learning tools in solving many complex scientific problems. In this work, we present two case studies that used DM and ML techniques to enhance new-particle formation (NPF) identification and analysis. Extensive measurements and large data sets related to NPF and other ambient variables have been collected in arctic and boreal regions. The focus area of our studies is the SMEAR II station located in Hyytiälä forest, Finland that is in the area of interest of the Pan-Eurasian Experiment (PEEX).

Atmospheric NPF is an important source of climatically relevant atmospheric aerosol particles. NPF is typically observed by monitoring the time-evolution of ambient aerosol particle size distributions. Due to the noisiness of the real-world ambient data, currently the most reliable way to classify measurement days into NPF event/non-event days is through a manual visualisation method. However, manual labour, with long multi-year time series, is extremely time-consuming and human subjectivity poses challenges for comparing the results of different data sets. In this case, ML classifier is used to classify event/non-event days of NPF using a manually generated database. The results demonstrate that ML-based approaches point towards the potential of these methods and suggest further exploration in this direction.

Furthermore, NPF is a very non-linear process that includes atmospheric chemistry of precursors and clustering physics as well as subsequent growth before NPF can be observed. Thanks to ongoing efforts, now there exists a tremendous amount of atmospheric data, obtained through continuous measurements directly from the atmosphere. This fact makes the analysis by human brains difficult, on the other hand, enables the usage of modern data science techniques. Here, we demonstrate the use of DM method, named mutual information (MI) to understand NPF events and a wide variety of simultaneously monitored ambient variables. The same results are obtained by the proposed MI method which operates without supervision and without the need of understanding the physics deeply.