Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2018-162-RC1, 2018

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Exploring nonlinear associations between atmospheric new-particle formation and ambient variables: an information theoretic approach" *by* Martha A. Zaidan et al.

Anonymous Referee #1

Received and published: 15 May 2018

In this study, an information theory approach has been used to analyze nonlinear correlations between different atmospheric variables (meteorological, aerosol, gas and radiation) and new particle formation (NPF) in Hyytiälä measurement station in Finland. The correlations have been evaluated by calculating the mutual information (MI) between different variables and event/non-event days using a nearest neighbor method developed earlier for other applications. The authors suggested the MI method could be widely used to evaluate correlations between different variables and NPF as well as other phenomena in the atmospheric science.

To my knowledge, this is the first study when an information theory approach (i.e. MI)

have been used in NPF analysis. Briefly, the results obtained show that NPF correlates with sulphuric acid and water concentrations, ultraviolet radiation, condensation sink and temperature. In the previous studies, those variables have also been associated to NPF. The MI seems to be a simple and effective method to analyses nonlinear correlations between ambient variables and NPF in large datasets without any supervision or prior knowledge of NPF mechanisms. Furthermore, it seems to be computationally relatively light and have only one free model parameter. Therefore, the method is a suitable tool to analyze large and complicated datasets in atmospheric science and other "Big data" applications. Especially, it can be used for screening suitable variables for more detailed analysis.

The manuscript (MS) is quite well organized and written. The content on the MS is in the scope of journal, although, e.g., Atmospheric Measurement Techniques could be a more relevant journal due to a technical aspect of the MS. Overall, the MS is suitable for publication in this journal because it introduces a new approach for atmospheric data analysis, especially for analysis of NPF involving complicated and nonlinear processes. However, some comments, suggestions and technical corrections should be considered and discussed before publication.

Specific comments:

Tittle

In title, "an information theoretic approach" is mentioned. I feel that "mutual information approach" or similar would be more descriptive.

Abstract (also results and discussion)

It is mentioned, "The applied mutual information method finds that the formation events correlate with sulfuric acid concentration. . ." Is there a specific level for MI (certain value) that indicates a correlation between different variables? Alternatively, are those variables the most correlating factors because they have the highest MI values? In

general, when you can say that some variables correlate or not when using the MI approach. Please discuss in MS text.

1. Introduction

In the introduction, the authors are citing only Hyvönen at al. (2005) for conducting similar data mining on parameters affecting NPF. It should be noted that Mikkonen et al. (2006, 2011) have used similar approachs, discriminant analysis and multivariate nonlinear mixed effects model, to analyze key factors contributing to the NPF and growth of formed particles, respectively. They showed that in more polluted environments, like San Pietro Capofiume, Melpitz and Hohenpeissenberg, the parameters found in Hyytiälä were not sufficient to predict NPF. Especially, when Hyvönen et al. (2005) did not found global radiation to be important variable for NPF, in Mikkonen et al. (2006, 2011) papers it was the most important variable. Other significant parameters related to NPF were RH, O3, SO2, NO2 and temperature, some of these found relevant also in this study. I think that those studies should also be considered in the introduction and later in discussion together with the study by Hyvönen et al. (2005).

Please, also summarize very briefly other studies in the field of aerosol/atmospheric science that have used information theory approaches or a MI method. Is there any specific area in which those methods have been used frequently (e.g. remote sensing)? After quick search, I found some previous studies: Preining (1971), Li et al. (2009, 2012), and Brunsell and Young (2008) but probably there are much more published studies.

2.2 Measured variables

Please, check the size range and measurement height of the particle number size distribution measurements. For instance, Nieminen et al. (2014) mentioned that size range was 3-500 nm until Dec 2004 and Dal Maso et al. (2005) that sampling height was 2 m above ground.

2.3 Derived variables

Condensation sink has been calculated from particle size distributions, which size ranges were not same for all measurement. Has this any effect on the results?

Proxy concentration of sulfuric acid has been calculated from other variables. Does this have any effect on the results (MI values, relative correlations)?

For simplicity, undefined days have been removed before MI calculations. What would be an effect on the results if the undefined days were included in MI calculations? I think that MI can easily be used to evaluate several discrete variables.

3.1 Data pre-processing

You have normalized continuous variables to have zero mean and unit variance so that large numerical values are not too significant in analysis. In general, is this always needed if you use a Euclidean distance in the nearest neighbor method when calculating the MI values as described in Fig. 4?

In the analysis, you have eliminated nighttime data points in atmospheric variables. I suppose that after that you have not exactly same number of data points at exactly same time when calculating distances between variables (see Fig. 4b). Please, clarify in the MS how you have considered this problem and what is time resolution for measured variables (hour average/1 min instantaneous).

3.2 Information Theory: A brief introduction

I feel that Shannon, the pioneer in information theory, should be mentioned in the MS (Shannon, 1948). I suppose that his pioneer work is not well known for typical readers of this journal.

3.2.2 Mutual Information

In the Fig. 3, MI and Pearson correlation coefficient are shown a standard test set. Is there any reference for that data set or is it publicly available (line 19 page 7)?

Furthermore, a comparison of MI results to the Pearson correlation is a bit questionable as assumptions of the Pearson correlation are not valid for these data and, e.g., the Spearman correlation should be used instead. It might not change the results significantly but the comparison would be more valid.

3.3 Mutual Information implementation: nearest neighbor method

Please insert suitable references for that chapter. I think that is not generally known in the field of atmospheric science. The used nearest neighbor method have been described, e.g., in papers by Kraskov et al. (2004) and Ross (2014) as mentioned in a previous chapter. Kraskov et al. (2004) described two different algorithms and the first one seems to be used here. The notations and equations seems to be exactly same than in Ross (2014). Please, indicate preferences in more detail in this chapter.

Have you calculated MI values only for event days or both event and non-event days? Do the calculated MI levels present the key factors contributing to the NPF or the factors that best separate event and non-event days form each other (or vice versa)? If you have calculated MI values only for event days, what are key factors contributing to the non-event days. Please clarify this in the MS because it is now unclear for me. Practically, is the discrete variable x a set of event days or a set of event and non-event days in the calculations? Furthermore, if you include undefined days in MI calculations, how does this affect the results? Have you already done any calculations with event, non-event and undefined days? Those results would be very useful when a capability of MI method in NPF analysis is evaluated and thus should be discussed in the MS.

Why have bivariate correlations only been inspected? During NPF, multiple different phenomena occur simultaneously and thus the analysis should be multivariate. Can multivariate analysis be conducted using the information theory approach?

Please indicate in MS text, which distance (Euclidean distance, I suppose) and k-value (3?) you have used in the calculations. Furthermore, indicate how you have practically calculated MI values (using Matlab/Python/etc. programs made by authors, commer-

cial programs, programs distributed by Ross (2014) or otherwise). Finally, the publisher encourages authors to deposit software, algorithms, and model code on suitable repositories/archives whenever possible (see the journal Data policy).

3.4 Mutual information: a simulation case study

Is this chapter needed? Is this simulation case relevant with the NPF analysis? In Fig. 3, you have already shown capability of the MI method to find non-linear correlations and this is, I think, only one example more and therefore you can remove it.

Have you used same program in this or Fig. 3 cases than in NFP calculations? For me, this and Fig. 3 cases look like continuous variables vs. continuous variables cases whereas NPF calculations are discrete variables vs. continuous variables cases.

I think that tests with simulated event/non-event data would be more relevant than solar spectrum data with very large temperature variation. Have you done any studies with simulated event/non-event data?

4 Results

A better title of this chapter is Results and Discussion because the chapter also includes discussion of the results (not only results, see classical IMRaD structure).

4.1 Correlation analysis between atmospheric variables and NPF

Is it possible find using the MI method whether the correlation is positive or negative (i.e., is lower or higher value more favorable) in relevant situations?

You mentioned that the temperature is associated with many atmospheric variables. I think that chemical reactions that produce condensable species depend on temperature so it could be also mentioned.

You stated that wind direction have little correlation with NPF and discussed that small correlation persists due to pollution from the westsouth - west (station building and city of Tampere). How about a local sawmill and a power plant in Korkeakoski, located ca.

6 km southeast of Hyytiälä (see e.g., Liao et al. 2011; Williams et al., 2011; Lopez-Hilfiker, 2014). Could the sawmill and the power plant have any influence on NPF and can you see this in MI values?

5 Conclusions

You stated: "The method also contains only one free parameter (the number of nearest neighbours k) and its value does not affect the results significantly". Have you tested several k-values? If this is a generally known fact, please add a suitable reference.

Can MI method use to analyze long-term changes in NPF (e.g. due to climate change)?

You discussed about automatic event classification algorithms. Please note a recent paper by Joutsensaari at al. (2018).

Figure 5.

Please indicate in a caption what does sigma and MI mean.

Figure 6.

"MI correlation level for a variety of atmospheric variables": Should NPF be mentioned in caption (MI correlation levels between NPF and a variety . . .). Also, indicate in caption that notations are shown in Table 1.

Figure 7.

Does blue color (MI=0) in large particles in Period 1 indicate that there is no data or no correlation? Please clarify this.

Figure 8.

Please indicate in caption what r_pb means.

Technical corrections:

Page 12, line 2: . . .2017) . As. . . => . . .2017). As. . .

Page 12, line 13: (e.g. . . .) => (i.e., . . .)

References:

Brunsell, N. A., and Young, C. B.: Land surface response to precipitation events using MODIS and NEXRAD data, Int. J. Remote Sens., 29, 1965-1982, 10.1080/01431160701373747, 2008.

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.

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, https://doi.org/10.5194/acp-5-3345-2005, 2005.

Joutsensaari, J., Ozon, M., Nieminen, T., Mikkonen, S., Lähivaara, T., Decesari, S., Facchini, M. C., Laaksonen, A., and Lehtinen, K. E. J.: Identification of new particle formation events with deep learning, Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2017-1189, in review, 2018.

Kraskov, A., Stögbauer, H., and Grassberger, P.: Estimating mutual information, Physical review E, 69, 066 138, 2004.

Li, Y., Xue, Y., Guang, J., Wang, Y., and Mei, L.: A retrieval algorithm for aerosol optical depth from MODIS multi-spatial scale data based on mutual information, 2009 IEEE International Geoscience and Remote Sensing Symposium, V-489-V-492, 2009.

Li, Y. J., Xue, Y., He, X. W., and Guang, J.: High-resolution aerosol remote sensing retrieval over urban areas by synergetic use of HJ-1 CCD and MODIS data, Atmos. Environ., 46, 173-180, 10.1016/j.atmosenv.2011.10.002, 2012.

Liao L., Dal Maso M., Taipale R., Rinne J., Ehn M., Junninen H., Äijalä M., Nieminen T., Alekseychik P., Hulkkonen M., Worsnop D.R., Kerminen V.-M. & Kulmala M. 2011. Monoterpene pollution episodes in a forest environment: indication of anthropogenic origin and association with aerosol particles. Boreal Env. Res. 16: 288–303.

Lopez-Hilfiker, F. D., Mohr, C., Ehn, M., Rubach, F., Kleist, E., Wildt, J., Mentel, Th. F., Lutz, A., Hallquist, M., Worsnop, D., and Thornton, J. A.: A novel method for online analysis of gas and particle composition: description and evaluation of a Filter Inlet for Gases and AEROsols (FIGAERO), Atmos. Meas. Tech., 7, 983-1001, https://doi.org/10.5194/amt-7-983-2014, 2014.

Mikkonen, S., Lehtinen, K. E. J., Hamed, A., Joutsensaari, J., Facchini, M. C., and Laaksonen, A.: Using discriminant analysis as a nucleation event classification method, Atmos. Chem. Phys., 6, 5549-5557, https://doi.org/10.5194/acp-6-5549-2006, 2006.

Mikkonen, S., Korhonen, H., Romakkaniemi, S., Smith, J. N., Joutsensaari, J., Lehtinen, K. E. J., Hamed, A., Breider, T. J., Birmili, W., Spindler, G., Plass-Duelmer, C., Facchini, M. C., and Laaksonen, A.: Meteorological and trace gas factors affecting the number concentration of atmospheric Aitken (Dp = 50 nm) particles in the continental boundary layer: parameterization using a multivariate mixed effects model, Geosci. Model Dev., 4, 1-13, https://doi.org/10.5194/gmd-4-1-2011, 2011.

Nieminen, T., Asmi, A., Dal Maso, M., Aalto, P. P., Keronen, P., Petaja, T., Kulmala, M., and Kerminen, V. M.: Trends in atmospheric new-particle formation: 16 years of observations in a boreal-forest environment, Boreal Environ. Res., 19, 191-214, 2014.

Preining, O.: Information theory applied to the acquisition of size distributions, J. Aerosol Sci., 3, 289-296, https://doi.org/10.1016/0021-8502(72)90050-X, 1972.

Ross, B. C.: Mutual information between discrete and continuous data sets, PloS one, 9, e87 357, 2014.

Shannon, C.E. (1948), "A Mathematical Theory of Communication", Bell System Technical Journal, 27, pp. 379–423 & 623–656, July & October, 1948

Williams, J., at al.: an overview of meteorological and chemical influences, Atmos. Chem. Phys., 11, 10599-10618, https://doi.org/10.5194/acp-11-10599-2011, 2011.

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-162, 2018.