Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2018-162-AC1, 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.

Martha A. Zaidan et al.

martha.zaidan@helsinki.fi

Received and published: 27 July 2018

Point-to-point response to referee 1:

We thank the reviewers for their encouraging and positive comments. The original comments (requiring a response) are shown in boldface. Our responses will be intercalated, and the final manuscript will be revised accordingly.

Specific comments:

Tittle
In title, "an information theoretic approach" is mentioned. I feel that "mutual

information approach" or similar would be more descriptive.
We will change the title to be: "Exploring non-linear associations between atmospheric new-particle formation and ambient variables: a mutual information approach ".

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.
In general, there is no specific level for MI or threshold that indicate a correlation between different variables which is also similar to Pearson correlation where this correlation value gives an only indication of the variables relationship. The value of MI depends on the distribution and the amount data. Unless mutual information gives very high value (very close to one) or a very low number (very close to zero), scientists need to make their own judgement about the variable correlation. In this case, similar variables are grouped based on their measurement types (traced gases, radiation, etc.), and their correlation level is ranked. The variables that have the highest mutual information level indicates that they are more favourable to NPF process compared to other variables. We will intercalate the following explanation in the abstract and result section.

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 approaches, discriminant analysis and multivariate non-linear 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 Hohen-

peissenberg, 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.

We will add more discussion about suggested publications into our introduction section.

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.

It is true that the measured particle size ranges were between 3-500nm until Dec 2004, and after that, it was extended to cover the size range from 3 nm to 1000nm. The sampling height is 35 m since it was moved to the tower in 2015. Previously, it was 2 m above the ground. Since we used the data until 2014, we correct this to be 2 m above the ground. We will revise this in the manuscript.

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)?

It is difficult to comment if different size ranges in CS calculation in the overall measurement is affected by the results. Nevertheless, we believe that this might impact only slightly the outcome because MI estimation is the average of MI for all data points. See equations (13) and (14). For sulfuric acid, we believe that other correlated variables used for calculating the proxy may have only a slight effect on the results. For example, the radiation is known to influence NPF, but in our calculation SO2 has the least correlation among traced gases to NPF. In this case, MI tries to compromise this and finding its mutual information for sulfuric acid.

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.

What we mean “for simplicity” is that we removed undefined days to prevent extra bias added to our data because the undefined data cannot be unambiguously classified as either an event or non-event day. Undefined days may belong to event or non-event days if further investigation is made. We expect that MI result will not be reliable if we include this group since our focus is only to find the relationship between NPF and atmospheric variables. We will clarify this in our manuscript.

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 al- ways needed if you use a Euclidean distance in the nearest neighbor method when calculating the MI values as described in Fig. 4?

Yes, you are right, we obtained the same results with and without normalization.

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).

We strived to find a common resolution for the calculation. Since we perform bivariate analysis, between NPF and an atmospheric variable, the time resolution varies for every variable. If a variable is measured every10 minutes, it means we used 10 minute time resolution. We will clarify this in the manuscript as suggested.

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.

We will mention Shannon's first work on this field in the manuscript.

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)?

Fig.3 uses the standard test set data, that is publicly available. The data is made available under the Creative Commons CC0 1.0 Universal Public Domain Dedication.

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.

The Spearman correlation gives similar result with Pearson correlation on this data, we can re-do this and include in the Fig.3.

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.

Since this is a continuous - discrete case, the reference should be Ross (2014), we will insert this in the manuscript.

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.

We calculated MI values for both event and non-event days. The MI attempts to find the best factors/variables that can differentiate between event and non-event days, so those are the atmospheric variable influencing NPF. As described earlier, undefined days are excluded to prevent extra bias added to our data because this group cannot be unambiguously classified as either an event or non-event day. We will clarify this in the manuscript.

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?

This method can perform only with bivariate case. To find interrelation correlations, we need to perform mutual information for every variable and make a plot matrix to analyse the impact of each variable.

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, commercial 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).
We mentioned in the manuscript that we used Euclidean distance with k=3. The software is the extension of Ross program, where we added extra features, such as the Numata scaling factor, see page 8. The software may be published later in Python and/or Matlab on in the first author's Github and whenever possible in the ACP.

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?
The nice thing about the second simulation study is that we know the underlying equation and shows the relationship between variables. This case study demonstrates how MI is able to estimate the relationship among input and output variables in the known model equations. The principle of continues-discrete MI method is also based on Kraskov (2004) and the simulation test was already done in Ross (2014). In other words, Fig. 5, is a validation study whereas Fig. 3 is just based on data. We do not

perform any study yet with simulated event / non-event data as we do not have the model equations to simulate the event and non-event day. It can be an extension to the present work.

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).
Thanks for the suggestion, we will change this to be: "Results and Discussion" in the manuscript.

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 favourable) in relevant situations?
Unfortunately, this method does not detect negative/positive correlation. As stated in the conclusion, this method will not replace completely the standard method, instead this method should be used in the first place before performing a deeper data analysis method, such as through histogram and scatter plots. MI acts as a detecting mechanism and Causality testing at a later stage can be used to understand the direction of flow of information from one variable to another.

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.
We will include this in our explanation.

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 ca6 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?
Unfortunately, we may not be able to see this from MI values. As mentioned earlier, the function of MI is for early correlation detection (which we may miss via Pearson correlation due to the non-linearity in variable relationship). Extra analysis and plotting are still required to understand a particular phenomenon.

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.
Yes, the result does not change significantly for our case. This fact was also mentioned in Ross (2014). We will add this reference there.

Can MI method use to analyze long-term changes in NPF (e.g. due to climate change)?
Probably yes, if we group NPF days into two categories based on their occurrence or frequency. Then, we compare between these groups and all atmospheric variables. Therefore, we may also learn what variables influence the increase of their occurrence, etc.

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

Figure 5. Please indicate in a caption what does sigma and MI mean. We will do.

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. Good point. We will do

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. Yes, it was due to no data available. Good point, we will clarify this.

Figure 8. Please indicate in caption what r_{pb} means. OK, we will do

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