Statistical Analyses of Environmental Field Measurements Project

Viet Le, Steven Job Thomas

December 21, 2019

1 Introduction

One of the major uncertainties in predicting the Earth's climate are the atmospheric aerosol particles. Being highly variable in space and time, aerosols are very difficult to quantify exactly. This leads to uncertainties in understanding the radiative balance in the atmosphere and predicting precipitation. These effects depend on the size distribution and also the chemical properties of aerosol particles.(Singla et al., 2018). Besides the mentioned effects, aerosols also effects human health adversely, Due to its complex interactions, research is being conducted across different spatial and temporal scales explain the climate effects of aerosols.(Größ et al., 2018)

While primary emissions are a huge source of aerosols particles, New Particle Formation (NPF) is one of the major sources of increase in the particle number in atmosphere and are seen to occur in diverse atmospheric conditions (Größ et al., 2018; Singla et al., 2018). NPF is also a major source of cloud condensation nuclei in the atmosphere.Several atmospheric parameters affect the formation of new particles in the atmosphere. Several studies have observed high relative humidity and high condensation sink to suppress NPF events while parameters like temperature have found to enhance the NPF events. Recently several studies have shown sulphuric acid to be one of the major precursors in NPF events. Some studies have suggested that RH prevents some volatile organic compounds (VOC) for ozonolysis reactions, thereby preventing the formation of cloud condensation nuclei.(Dada et al., 2017),

In this study, based on 2 years (2008-2009) of data from SMEAR II station at Hyytiälä, Finland, we use tools from statistics to relate atmospheric parameters, gaseous compounds like NO_x and O_3 with new particle formation events data and try to conclude the factors that favour NPF events. We also try to relate between the atmospheric parameters between different seasons and also events. Sulphur dioxide is also taken as a proxy to calculate sulphuric acid concentration taking into account its source and sink(Dada et al., 2017).

2 Data

The data used for the analysis in this study are from SMEAR II station of the University of Helsinki, Hyytiälä. The station has been providing the scientific community with several atmospheric measurements since 1995. Hyytiälä is a Boreal forest situated in southern Finland. It is considered to be a place with rural background as it is far away from major human activities and has very low levels of pollution. Measurements are made at several heights along the mast, but in this study we use all measurements taken from 16.8m. We use gas measurements of Ozone NO_x and SO_2. The parameters and gas concentrations were obtain from `https://avaa.tdata.fi/web/smart/smear/download`, for the years 2008 and 2009. Condensation sinks between the years 2004 and 2014 were given particle formation events between the same years. The events have been classified into three types based on the paper (Lyubovtseva et al., 2005). To process and analyse the data jupyter notebook was used.

2.1 Handling the Data

All parameters and gaseous concentrations data were downloaded and merged corresponding to the years the data were going to be analysed, from 2008 to 2009. The events data and condensation data were handled separately. All three event classes were combined into one and this was inserted into the events table.

In order to deal with outliers, several methods such as zscore, linear regression models, statistical models, etc. can be used. Here, the method of zscore was used to remove any outlier. Zscore method scales the data whose mean is considered to be 0 and the standard deviation is 1. Zscore assumes the data to be of normal distribution. Since our data isn't, we find the zscore for the difference of the preceding and succeeding values. Then the difference of the points that are too away from the mean of the differences are considered to be outliers and these are removed accordingly. The advantage of zscore is its quite simple and it takes into consideration both the mean and the amount of variability in the data set.

Once, the outliers were removed from the data set, we aggregated the all the atmospheric variables and condensation sinks into two forms. An hourly aggregate using average was taken to produce an hourly analyses of all the data at different months/seasons. Also, a daily aggregate using average was taken to understand the time series of these parameters. Since the values do not change instantaneously, the daily aggregate is quite good to understand these parameters.

2.2 Sulphuric Acid

Since several studies have understood sulphuric acid as one of the major precursors for NPF events, we calculate sulphuric acid concentration based on (Petäjä

et al., 2009). The major source of sulphuric acid in the atmosphere is the presence of sulphur dioxide which gets converted to sulphuric acid under radical reactions. The main sink for sulphuric acid are aerosol particles, for which we have the condensation sinks already calculated. The oxidation of sulphur dioxide takes place in the presence of hydroxyl radical OH, which is already obtained from ozone in the presence of UV radiation. UV-B wavelength is considered as it is one major wavelength that reaches the earth's surface. Hence using ultraviolet radiation (UV-B), SO_2 concentration and also condensation sink (CS) we calculate the proxy from the formula:

$$[H_2SO_4]_{proxy} = k \cdot \frac{[SO_2][UVB]}{CS}$$

where k is an empirically derived value which is 9.9 x $10^{-7} m^2 W^{-1} s^{-1}$

3 Results

All the variables taken into consideration were plotted initially for the whole of two years, to get a general idea of they have varied from 2008 to 2009.

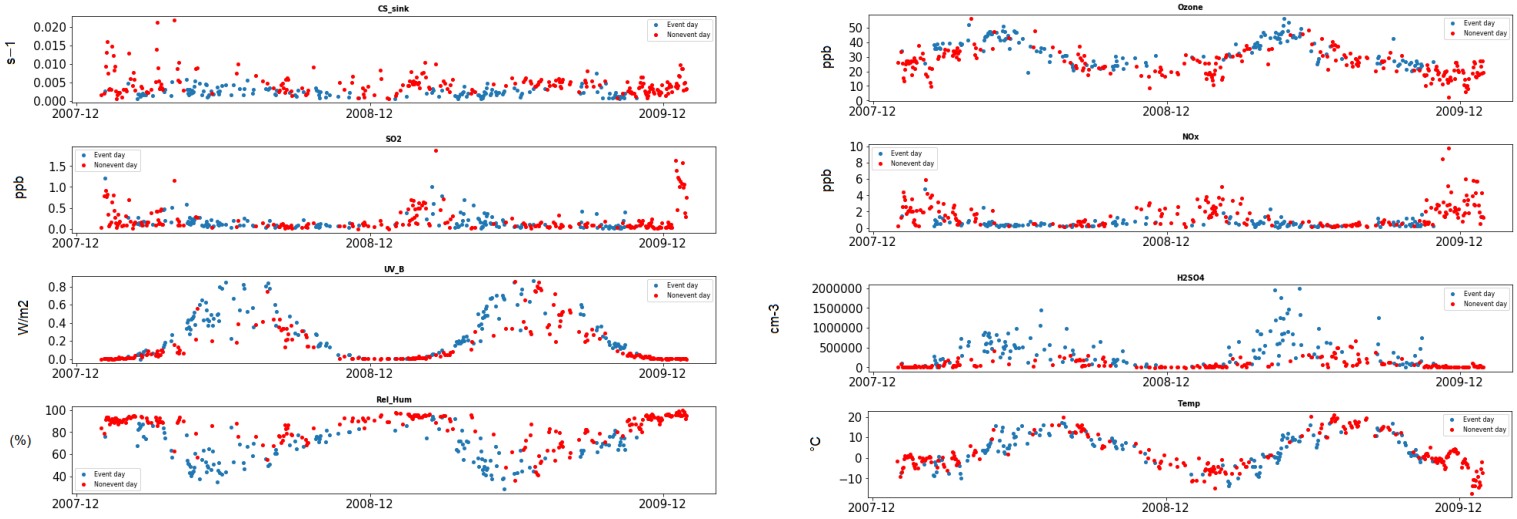

Figure 1: All the variables aggregated on a daily basis between the years 2008 and 2009. The blue circles correspond to event days while red circles correspond to non event days. Starting from right above: Ozone, NOx, Sulphuric Acid, Temperature, Relative Humidity, U-V(B) radiation, sulphur dioxide, Condensation Sink.

From the above plot, one can see that most of the event days can be accompanied with high UV-B radiation which will correspond to high temperature, high H_2SO_4 increased NPF events while high relative humidity, and high NOx suppress NPF events. From UV radiation plot, after summer, we can see that UV radiation does not play a major role that helps to differentiate between event

and non event days. This is in accordance with the paper by (Lyubovtseva et al., 2005), in which the UV (A) radiation is observed to have very little difference between event and non-event days during autumn. The temperature is not a good variable by itself to differentiate between NPF event and non event days. There are days when temperature has favoured NPF and this might be due to the reason at high temperatures, monoterpene emission tends to increase and favour NPF but it is not the case seen here. At very high temperatures there is also suppression of NPF, which is in accordance with the paper by (Dada et al., 2017). The plot of the condensation sinks show that NPF events tended to be favoured when condensation sink is generally.

Nuclear Particle Formation events were also plotted in a way where one can clearly see which month has the least and maximum number of events. From the the two years a total of 149 days had NPF while 229 days were considered to be days with no nucleation and 332 days being undefined. From figure 2, most maximum frequency of NPF events have taken place in the months of April and May during both the years, with 2009 showing more NPF during the same months. September and October is another month where one can see another set of NPF events, with 2009 having more event days. This frequency also seemed to coincide with that analysed for 20 years of data in the study by (Dada et al., 2017). The study also verified with a maximum number of events occurring in spring.

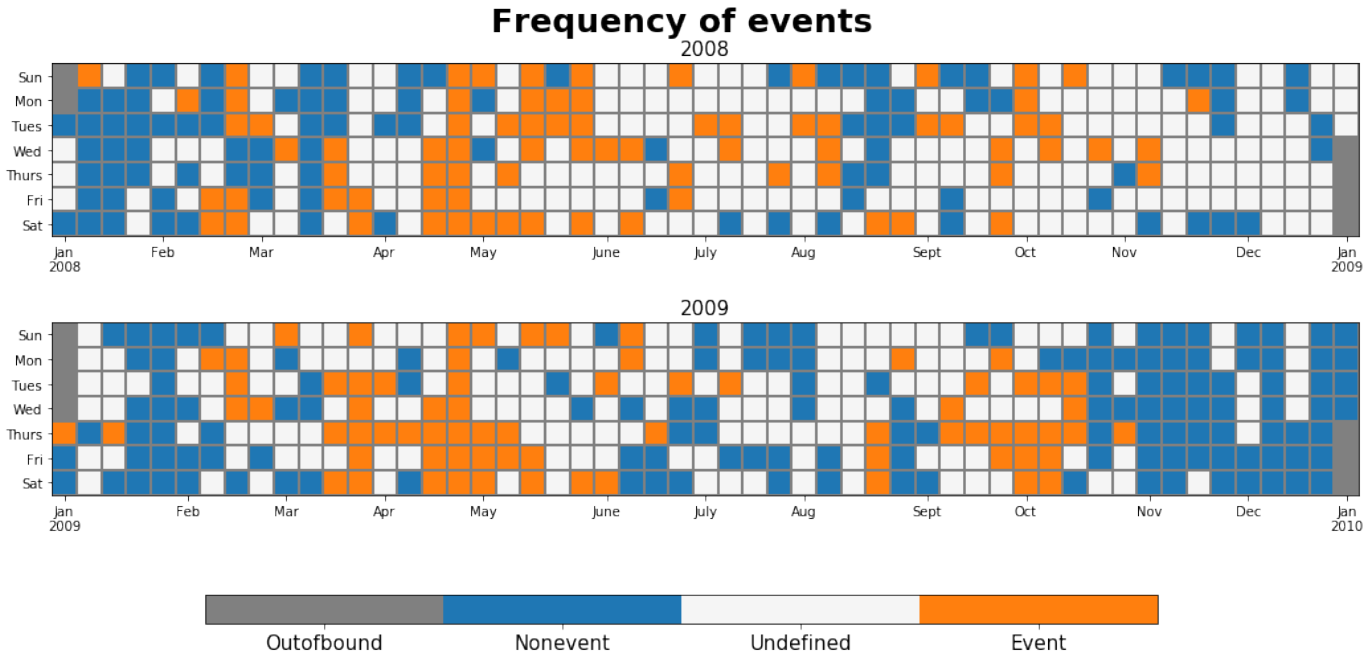

Figure 2: Plotting the frequency of events for both the years based on days.

3.1 Correlation Between the variables

A correlation was plotted among all the variables with respect to event and non-event days to see how the variables change with respect to the other. Relative Humidity as seen from Figure 1 are low on event days and this was also reflected in the correlation plot for event day. We find RH and temperature to be very highly negatively correlated on non-event days. Since we considered H_2SO_4 as a proxy for growth rate, we see that RH is negatively correlated with H_2SO_4 on even and non-event days, which explains that high RH suppresses NPF events.

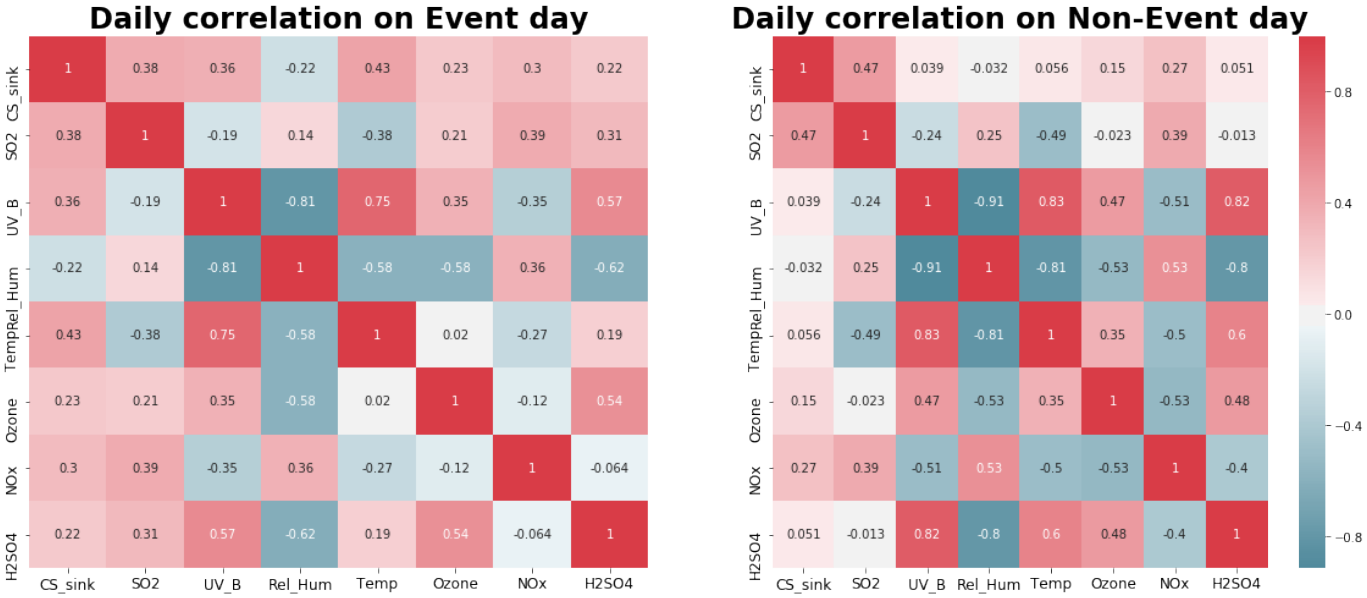

Figure 3: Correlation plot on event and non-event days explaining the relation between different variables.

3.2 During Different Seasons

Although four seasons were plotted, we will not be showing plots of how the variable change during all of them. We shall select the spring months, during which the NPF events seemed to maximum and also winter when there is least number of NPF events observed.

3.2.1 Spring

We have considered spring season in Hyytialä to during April and May. It is said that spring is said to last even during the first couple of weeks in June. In order to ease the analysis we have considered only April and May

Ultraviolet radiation (UV-B) rays becomes higher during the start of spring. There are clear diurnal patterns that are observed in temperature, relative humidity, ozone. This is accordance with the study done by (Lyubovtseva et al., 2005).

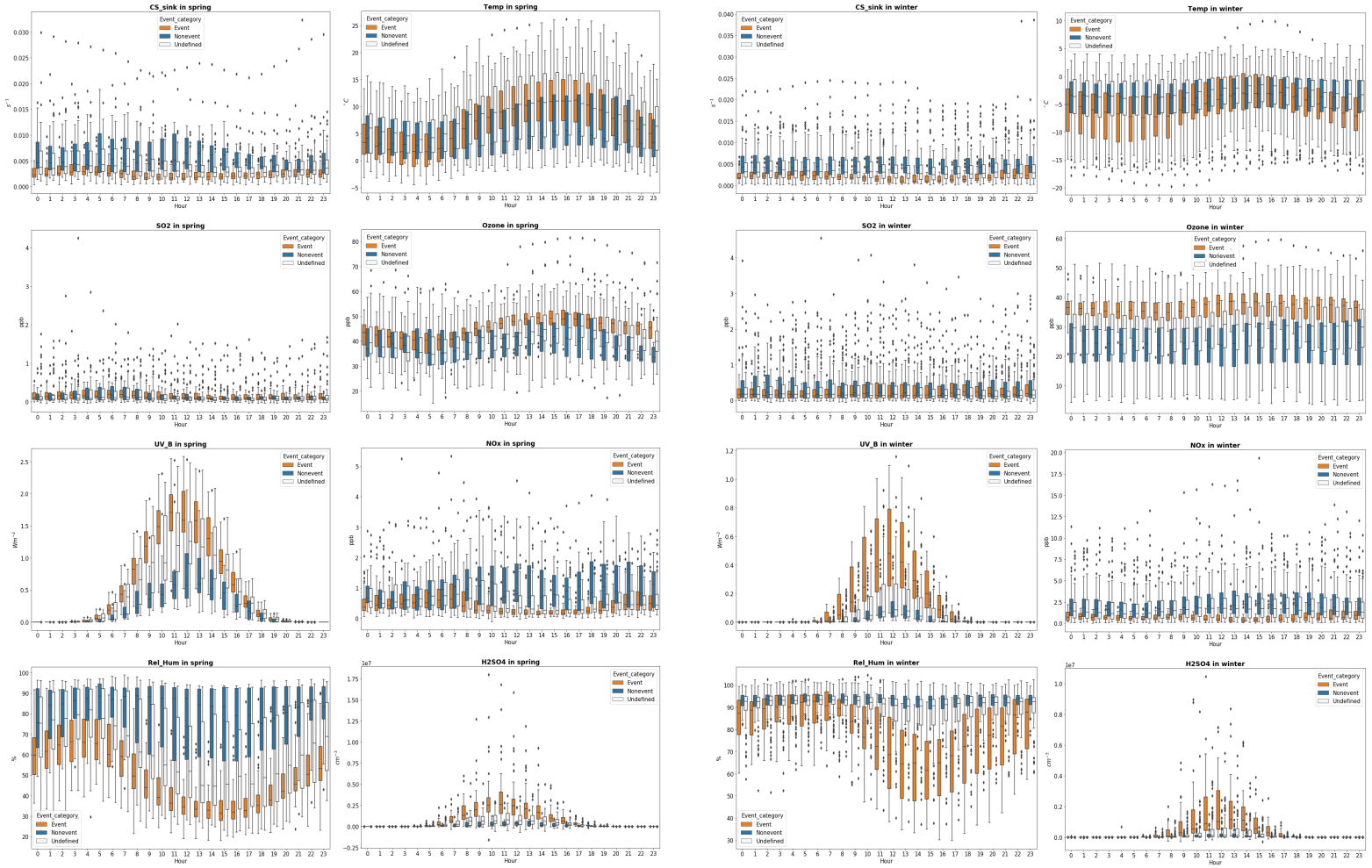

Figure 4: Diurnal variation of different variables plotted for Spring(left) and Winter(right)

The median concentration of NOx, seems to be very much lesser on event days than on non-event days. But this is not the same for SO_2, which is contradictory to the findings in (Lyubovtseva et al., 2005). But as our analysis comprised of only two years of data, it is insufficient to conclude this. The average temperature in spring during event days is much more than non-event days.

But during summer, the temperature during non-event days is much more than event days. During spring there is a clear evidence that, NPF events occur at high temperatures, with high ozone concentrations, high UV radiation and less relative humidity. High UV radiation increased sulphuric acid concentrations, which is clearly visible from Figure 1 and Figure 4(Spring). Nevertheless (Dada et al., 2017) has observed the same proxy to not be effective to explain NPF events. This can be noted from the proxy calculated when taking winter into consideration as well. A clear difference cannot be seen between the two. Rather it was noted by (Dada et al., 2017) that monoterpene oxidation products proved to be more effective to explain the growth rate and correlated well with NPF events.

3.2.2 Winter

We have considered the winter months to be from December to March. During December there is absence of nucleation events which can be confirmed from Figure 2. Temperature is quite low in December, and there is hardly any UV-B radiation throughout winter. But one can notice the nucleation events in January. The median temperature at which nucleation events have taken place is lesser than during non-event times. This is quite the opposite to that observed in spring, where nucleation events took place at higher temperatures. There is also greater concentration of SO_2, and NOx and very less concentration of sulphuric acid is observed due to absence of radiation. The higher concentration of SO_2 during winter is explained due to less mixing and low height of planetary boundary layer (Dada et al., 2017). Another interesting fact is that, the difference in the ozone levels between event and non-event days is much higher in winter, with ozone during event days being a lot higher. Whereas in spring this difference is quite less. During winter one can say that low temperatures and condensation sink, together with high SO_2 favoured events during the months from January to February.

3.3 NPF Events between 2008 & 2009

We wanted to see if the variables change considerably between the years 2008 and 2009 during spring and early summer, when the nucleation events were maximum. We performed a Kruskal-Wallis test to see if the variables change significantly. A non-parametric test was performed as the variables do not show a normal distribution.

All the variables except condensation sink proved to be not significantly different. The condensation sink in 2009 is lower during spring and early summer (Figure 5), and this also seemed to coincide with increased number of events during the same period, which can be confirmed from Figure 2.

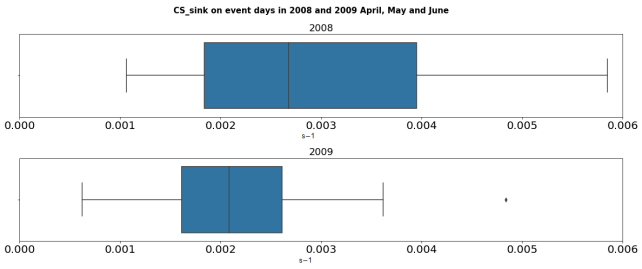

Figure 5: Comparison of condensation sink during months April, May and June between 2008 and 2009

3.4 K-Nearest Neighbour

We tried to form a model using K-Nearest Neighbour to predict all three event categories by using all obtained variables. One aspect about this modelling is that we ignored the temporal variation of the model and treated each observation as independent.

We randomly split the data into training and testing set. The model was put to train on the training data to obtain its own parameters, then it was used to predict on the test set. The prediction was then compared to true label of the test set via a Confusion Matrix.

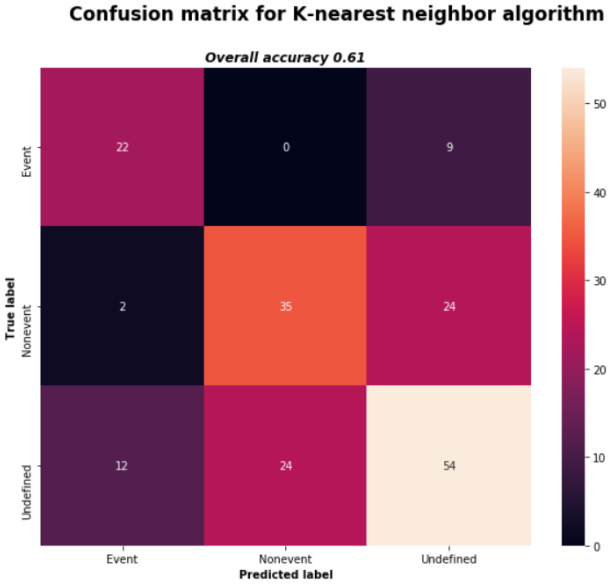

Figure 6: Confusion matrix on test set of K-Nearest Neighbour

The result shows that the accuracy is low at only 61 percent. However the algorithm has an extremely low false positive rate for misidentifying Event for Non-event and vice versa. This finding shows that there is some clear distinction between Event and Non-event which the model has efficiently identified.

In addition, k-fold cross validation with five folds were used to obtain more realistic accuracy for this model on new data. The result average accuracy across five folds is 0.52. Despite the low accuracy, we think that it is quite important to report this as it will improve analysis in the future.

4 Conclusion

We acquired and analysed ozone, temperature, condensation sink, UV irradiation, relative humidity, trace gases (sulphur dioxide, nitrogen oxides) and the proxy (sulphuric acid) for for the event and non-event days.

Maximum number of events took place in the year 2009. During spring and early summer the condensation sink was quite low in 2009 and hence could have attributed to the enhanced number of nucleation events in spring and early summer 2009. In spring, NPF events correlated well with high UV radiation, low relative humidity and condensation sink ,and high temperature and ozone. A good diurnal pattern was also observed during spring. As a diurnal pattern was observed between condensation sink and RH in a study by (Lyubovtseva et al., 2005), the same significant pattern could not be observed here. But this can also be attributed to the amount of data that was analysed here. The proxy H_2SO_4 calculated was not able to explain nucleation well either.

During winter, the event days took place at very low temperatures, higher relative humidity than observed during events in spring along with higher concentrations SO_2 and NOx. Sulphuric acid was also observed to be lesser than in spring. A clear diurnal pattern as observed in spring was not seen during winter. The NPF events that took place during winter is more complicated and more variables and data are needed to explain the nucleation during winter. The Sulphuric acid, a proxy that was calculated to observe nucleation events cannot clearly explain the same as it follows the trend of UV-radiation only.

References

Dada, L., Paasonen, P., Nieminen, T., Buenrostro Mazon, S., Kontkanen, J., Peräkylä, O., Lehtipalo, K., Hussein, T., Petäjä, T., and Kerminen, V.-M. e. a. (2017). Long-term analysis of clear-sky new particle formation events and nonevents in hyytiälä. *Atmospheric Chemistry and Physics*, 17(10):6227–6241.

Größ, J., Hamed, A., Sonntag, A., Spindler, G., Manninen, H. E., Nieminen, T., Kulmala, M., Hõrrak, U., Plass-Dülmer, C., and Wiedensohler, A. e. a. (2018). Atmospheric new particle formation at the research station melpitz, germany: connection with gaseous precursors and meteorological parameters. *Atmospheric Chemistry and Physics*, 18(3):1835–1861.

Lyubovtseva, Y. S., Dal Maso, M., Sogacheva, L., Bonn, B., Keronen, P., and Kulmala, M. (2005). Seasonal variations of trace gases, meteorological parameters, and formation of aerosols in boreal forests. *Boreal Environment Research*, 10(6):493–510.

Petäjä, T., Mauldin, III, R. L., Kosciuch, E., McGrath, J., Nieminen, T., Paasonen, P., Boy, M., Adamov, A., Kotiaho, T., and Kulmala, M. (2009). Sulfuric acid and oh concentrations in a boreal forest site. *Atmospheric Chemistry and Physics*, 9(19):7435–7448.

Singla, V., Mukherjee, S., Kristensson, A., Pandithurai, G., Dani, K. K., and Anil Kumar, V. (2018). New particle formation at a high altitude site in india:impact of fresh emissions and long range transport. *Atmospheric Chemistry and Physics Discussions*, pages 1–26.

Student Contribution

Viet Le

- Acquisition of Data and identifying outliers in the data.
- Plotted Frequency of events and plotted the variables during spring and summer.
- Plotted K-nearest neighbour algorithm.
- Also wrote the code for principal component analysis.
- Refined the programme.
- Wrote half of the report.

Steven Job Thomas

- Performed merging of data and aggregated them for further operations.
- Plotted the correlation plots for event and non-event days and also plotted the diurnal pattern for winter and autumn.
- Plotted the time series plot for all the variables.
- Performed Kruskal-Wallis test for the variables.
- Wrote the the half of the report.
- Refined the report.