A data-driven analysis of the impact of short-term exposure to air pollution on the respiratory rate of asthmatic adolescents

Pablo Andreu Sedeño

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2020

Abstract

Air pollution is considered to be an important health risk, which is increased for children, especially those with respiratory problems. Associations between the concentration of different pollutants and several indicators of respiratory problems in children, such as visits to clinics and hospital emergency rooms, have been shown in previous work, but none of them have demonstrated a causal effect. This dissertation establishes for the first time, a causal effect between the concentrations of three air pollutants: airborne particles with an aerodynamic diameter of less than 2.5 micrometres ($PM_{2.5}$), nitrogen dioxide (NO_2) and ozone (O_3), and the respiratory rate of asthmatic adolescents. To do so, a state-of-the-art causal discovery method has been used, evaluating the causal relationship for a period of time of up to eight hours between the exposure to the pollutant, and the response of the subjects' breathing rate. In fact, for more than 20% of the tested time intervals from 1 minute to 8 hours, the three pollutants were shown to directly affect the breathing rate of the subjects. Additionally, the exposure-response relationship has been studied for the three pollutants. For each of them, in the majority of the causal links found, an increase in the pollutant concentration to a higher value than the average of at least the past 200 minutes, results in an increase of the breathing rate of the subjects.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Pablo Andreu Sedeño*)

Acknowledgements

I would like to thank my project supervisor D.K. Arvind, as well as Zoë Petard, for their guidance and advice throughout this project.

I would also like to express my gratitude to my family and my girlfriend for all their love and support while I was studying abroad, which was essential during this tough year.

Finally, I hope that you are proud of me, my beloved grandmother, as you will always be in my heart.

Table of Contents

Chapter 1

Introduction

1.1 Motivation for the project

According to the World Health Organization (WHO), unhealthy environments were responsible for a 24% of the world's deaths in 2016 [28]. Air pollution is one of the main contributors, accounting for 4.2 and 3.8 million deaths worldwide in 2016 caused by outdoor and indoor pollution, respectively [21, 23]. The impact on low-middle income countries is much more pronounced, since the vast majority (98%) of their cities have air pollution levels which exceed the WHO guidelines for the pollutant concentrations, compared to a 56% of high-income ones [20]. As a result, around 91% of the deaths caused by outdoor air pollution in 2016 happened in those countries [21].

More than 9 out of 10 children live in highly polluted areas, which contributed to more than half a million deaths of children younger than 5 years old in 2016 [25]. They are usually very active, which results in a higher proportion of pollution intake with respect to their weight than adults [22]. Additionally, their body organs are not fully-developed, and the maturing process can be hampered by air pollutants [33].

People with asthma or chronic obstructive pulmonary diseases (COPD) also constitute a risk group. In fact, in low-middle income countries, one-quarter of the deaths from COPD are caused by indoor air pollution [23]. Numerous studies have shown a positive association between exposure to air pollutants and respiratory issues, including asthma exacerbations, supporting these claims [18, 37, 16, 33]. Nevertheless, there is also previous work studying children's respiratory health, in which a direct relationship between the intake of several pollutants and the prevalence of asthma or other respiratory problems has not been detected [12, 19].

Therefore, there is a necessity to establish the relationship between the exposure to

air pollution and the response of people with chronic respiratory conditions, especially of children or adolescents in low-middle income countries, who are the most affected population group. This would enable a better understanding of whether the pollutant concentration levels have a causal effect on the respiratory condition of the subjects, and if so, to predict what will be the impact of different pollution levels, in order to take preventative measures.

1.2 Project objectives

This project has two main objectives. The principal one is to determine whether there is a causal relationship, over different time lags, between the exposure of adolescent patients with asthma to three airborne pollutants: particulate matter with an aerodynamic diameter smaller than 2.5 μm ($PM_{2.5}$), nitrogen dioxide (NO_2) and ozone (O_3), and their respiratory rate. The secondary one is to establish how the respiratory rate changes with the concentrations of these pollutants: whether it increases when each of the pollutant levels increase.

This research is based on the data collected in Delhi, India by the Delhi Air Pollution: Health and Effects (DAPHNE) project [1], which gathers pollution data and its impact on asthmatic adolescents and on pregnant women. For this project, only the asthmatic adolescents data has been used, consisting of 127 subjects. Each one has between 1 and 3 trials registered in the DAPHNE dataset, which consist of time-series data collected from subjects for a period of up to 48 hours. The AIRSpeck-P [2] and RESpeck [3] wearable devices were used to monitor the personal exposure to particulate matter of different sizes (PM_1, $PM_{2.5}$ and PM_{10}) and the respiratory rate of the subjects, respectively. In addition to PM values, data for O_3 and NO_2 concentrations can be obtained from the AIRSpeck-S stationary monitors [2] in the subjects' schools and homes. These sensors have been developed at the University of Edinburgh.

$PM_{2.5}$, NO_2 and O_3 are three of the four pollutants for which the World Health Organization has defined guideline values [24] due to their health risks. The fourth one is sulphur dioxide (SO_2), which cannot be included because this data is not collected by the DAPHNE project. The effect on the subjects' breathing rate in the short term (up to a time lag of 8 hours from the exposure) has been evaluated for all three pollutants. Since NO_2 and O_3 data is not collected by the personal AIRSpeck device, the distance of the subjects to the static sensors at each timestamp has been taken into account so that only relevant data (< 1 km away) is included.

1.3 Novelty

This is the first research to study the short-term exposure-response relationship between $PM_{2.5}$, NO_2 and O_3, and the breathing rate at an individual level for asthmatic subjects. It has been evaluated for different time lags between the exposure and the response, from 1 minute up to 8 hours with one minute resolution for the first hour and every 10 minutes from then on. Unlike previous studies on the health effects of air pollution which use pollution data kilometres away from the subjects [36], personal exposure data has been used in the case of $PM_{2.5}$ and nearby data (< 1 km away from the subject) for NO_2 and O_3. Furthermore, a recently-published causal discovery method has been used to detect both linear and nonlinear relationships between the concentration of these pollutants and the respiratory rate.

1.4 Results achieved

A state-of-the-art causal discovery method has been implemented to establish causal relationship between concentration levels of $PM_{2.5}$, NO_2 and O_3, and the respiratory rate of asthmatic adolescents. The causal effects were present in 20% of the time intervals tested for up to 8 hours from exposure. Additionally, the exposure-response relationship has been studied for the three pollutants. For each of them, the respiratory rate increases in the majority of the causal links found, when the pollutant concentration is higher than the average for at least the previous 200 minutes.

1.5 Document structure

The dissertation document consists of 8 chapters. Chapter 2 contains the background of the project. The exploratory data analysis and the pre-processing of the pollutant and respiratory rate data is done in Chapters 3 and 4, respectively. In Chapter 5, several machine learning methods are used to demonstrate an association between the pollutants and the respiratory rate. The PCMCI causal discovery method is explained and used in Chapter 6 to determine the causal relationships. Chapter 7 analyses the exposure-response relationship between the pollutant concentration values and the breathing rate for those time lags in which a causal relationship has been found by the causal discovery method. Finally, Chapter 8 contains the conclusions of the project, summarising the achievements and suggesting future work to expand the research.

Chapter 2

Background

Air pollution is a matter of great concern, so there is a vast amount of research to study its impact on human health. Asthma exacerbations have been associated with the exposure to air pollution, which has been suggested to be caused by an inflammation of the respiratory airways produced by the pollutants, especially in countries with high concentration levels like India [10, 7]. However, the specific mechanisms are still unknown, as well as its short-term impact on the condition of asthmatic subjects, which was identified as a research gap [10], but has been filled in this project.

Asthmatic patients make use of rescue inhalers to help relieve the symptoms of asthma attacks [26]. Therefore, their usage is a good indicator of asthma exacerbations. A recent study analysed the short-term impact of particulate matter with a smaller diameter than 2.5 μm and the use of rescue inhalers in the United States [36]. The pollution data was assigned to the event from an air-quality control station from the same state. The result was an increase of less than 1% in the medication usage for a 12% increase in the $PM_{2.5}$ concentration. However, the fact that the station was within the same state does not guarantee to be near the monitored subject, with an average distance from the subject of 16 kilometres. Therefore, the pollutant concentration data recorded may not be similar to what the personal exposure of the subject actually was.

The same problem was identified recently in [34], where they demonstrated that the sparsely located air-quality control stations are not sufficient to identify all the pollutant concentration peaks. By installing community sensors in a region of California, United States, they were able to identify twice the number of air pollution episodes than with the government ones. This issue has been addressed in this project by directly using personal exposure data for $PM_{2.5}$ with a wearable device, and data from stations not more than 1km away from the subjects for NO_2 and O_3.

Asthma and other respiratory problems can lead to clinical visits or even hospital admissions. The association between air pollution and these events has been analysed by two different studies released in 2020. The first one evaluates the effect of different meteorological factors like temperature, humidity or air pressure, as well as particulate matter, NO_2, O_3 and SO_2 on clinical visits of children due to respiratory problems in Shanghai, China [14]. They performed a correlation analysis combined with a Poisson regression model to determine what the influence of an inter-quartile range increase of each pollutant concentration and meteorological factor was on the number of clinical visits of children because of respiratory conditions. The results indicated a positive association between both the meteorological factors and the pollutants with the number of children clinical visits. The meteorological factors had a stronger association, but high levels of NO_2 were also considered a risk factor [14]. Taking this into account, the meteorological factors measured by the AIRSpeck devices (temperature and humidity) have also been included in the analysis of this project. The study also claims to have used an extensive network of air-quality monitoring stations, but does not mention the average distance to the subjects.

The second study analyses the impact of NO_x, NO_2 and $PM_{2.5}$ on both visits to the hospital and hospital admissions due to bronchitis or asthma, not limiting it to children, in Silesian Voivodeship, Poland [15]. The method used is very similar, using once again a Poisson regression model to determine what is the effect of inter-quartile increases of the pollutant concentrations on the hospital visits and admissions. This method, although it can show the strength of the associations, it does not establish whether there is a causal relationship between them. In this project, a novel causal discovery method has been used to cover this gap. The method yielded a significant association between both NO_x and NO_2 and the hospital visits and admissions due to bronchitis and asthma, being the $PM_{2.5}$ association weaker, and only for long term exposures [15].

The exposure to several pollutants can also lead to an increase in the mortality risk. Beelen et al. studied the long-term effects of black smoke, NO_2, SO_2 and $PM_{2.5}$, which are traffic-related pollutants, on different types of mortality, including respiratory one [4]. The association with respiratory mortality was only significant for the black smoke and NO_2. These results agree with the previous ones in that NO_2 has a stronger association with respiratory problems than other pollutants such as $PM_{2.5}$, which will be assessed in this project with a stronger notion: causality.

The number of clinical visits, hospital admissions and on top of them, the mortality,

are due to strong alterations in the condition of the patients. In this project, a novel approach has been used, consisting of the analysis of the impact on the breathing rate of the subjects. This allows to capture more precisely how their condition changes with different exposures, without reaching situations where hospitalisation is needed, as there are different degrees of discomfort that the subject can suffer before that point.

A previous project [17] investigated the relationship between $PM_{2.5}$ and the respiratory rate of the DAPHNE asthmatic subjects using a selection of statistical methods and the causal discovery method adopted in this project. A causal relationship was established but due to lack of time it was not investigated in greater depth; the time lags were limited to the first 60 minutes at intervals of 1 minute for linear relationships between $PM_{2.5}$ exposure and the respiratory rate, and only for the 1, 5, 10, 15, 30, 45 and 60 minutes time lags for non-linear ones. This project has extended the previous study in several ways:

- The tested time lags have been extended for up to 20 hours with a 1 minute resolution for linear relationships, and up to 8 hours with a 1 minute resolution in the first hour and every 10 minutes from then on for non-linear ones.

- The causality relationship has been extended to include the effect of the NO_2 and O_3 pollutants. To do so, new data set was created with the AIRSpeck stationary sensor data to include concentrations of NO_2 and O_3 when the subjects are within 1 km from the sensors.

- The causality relationship between $PM_{2.5}$, measured using the stationary AIRSpeck for distances up to 1 km from subject, and their respiratory has been established, comparing it to the personal exposure $PM_{2.5}$ results.

- Detailed analysis has been conducted on the relative importance of personal $PM_{2.5}$, stationary $PM_{2.5}$, nitrogen dioxide and ozone concentrations on changes in respiratory rate

Chapter 3

Exploratory data analysis

The exploratory data analysis (EDA) is a common first step in data analysis tasks to understand the distribution and characteristics of the data, so that wrong interpretations of the results are avoided [11]. In this project, there are two different types of data: the one collected by the RESpeck wearable device, containing the respiratory rate data, and the one from the AIRSpeck devices, which consists of the meteorological data (temperature and humidity) and $PM_{2.5}$ personal exposure data from the wearable device, and the NO_2 and O_3 data gathered by the static one. There is data for 127 asthmatic subject. All of them are time series, which means that they consist of a series of data points collected through time [5].

The static data was only considered when the closest sensor (school or home sensors) was at most 1 km away from the subject at each timestamp. This resulted on an average distance of the subjects to the static sensors for which data was considered of approximately 156 metres. When the school sensor was the closest one, but had no data for that minute, the community sensor data was used instead (if available), since both school and community static sensors are located outdoor, while the home sensor is indoor. Several comparisons were made between the school and community sensors to see whether this decision was reasonable. They both yielded similar values, having exactly the same data for some trials because the community sensor was sometimes used as the school one. The results of these comparison are included in Appendix A.3.

Location, scale, shape and correlation measures have been taken from the available data. Location measures allow to know where the data is positioned (around what values), scale measures determine how dispersed the data is, shape measures evaluate how symmetric the distribution is and what its tails look like, and finally correlation ones uncover linear relationships between the different variables.

Each of these categories contain robust and non-robust measures. The values of non-robust measures like the mean are very affected when an outlier is included in the data, unlike robust measures which maintain similar values. This allows to detect whether there are extreme values that need to be eliminated later. The tables with the exact numeric results of the descriptive statistics have been included in Appendix A, and the results have been interpreted here.

3.1 RESpeck data

The data from the RESpeck device for each trial of each subject contains the following information:

Feature	Description
Timestamp	The date and time of the measurement, with 1 min. resolution.
Breathing rate	The average respiratory rate of the subject in the minute, measured in beats per minute (bpm).
Breathing rate std	Standard deviation of the respiratory rate per minute
Activity level	Intensity of the subject's activity (floating point number).
Activity type	The category of the activity the subject is carrying out.
Step count	The number of steps of the subject in the minute.

Table 3.1: Data collected by the RESpeck device.

The first thing noticed in the analysis is that 37% of the respiratory rate data is missing (143741 missing values out of 388050). This is partly because the data has been filtered so that if the activity type denotes that the subject was lying on the stomach, the recorded respiratory rate is set as missing data because it cannot be trusted. The same happens when the sensor is not being worn by the subject, which is detected when the activity level is below a threshold (0.013).

The location measures for the breathing rate are shown in Table A.1. The breathing rate data is located around 20 bpm. The mean (non-robust measure), the median and the mode (both robust measures) are very similar, and the maximum and minimum measurements are in reasonable ranges taking into account the first and third quartile values. This indicates that there may not be outliers in the respiratory rate data, or at least not with extremely high or low values. However, the data is being analysed to-

gether for all the trials and not individually, so there might still be some values outside the norm for some trials.

The relatively low standard deviation shown in Table A.2 (4.4) compared to the mean (21.4), indicates that the values tend to be concentrated around the mean. However, the significant difference between the IQR and the range ($\approx$ 30 bpm) may indicate elongated tails in the distribution. The usage of different sensors across trials contributes to a greater variability in the data, as well as different forced vital capacities of the subjects [9]. Figure 3.1 illustrates the differences in the distribution of 5 randomly chosen trials of different subjects. It is done with a violin plot, in which the white point in the centre indicates the mean and the wide black bar line indicates the inter-quartile range, with the top of the bar being the third quartile and the bottom the first one. Finally, it contains the kernel density plot showing the distribution of the data at each side of the central vertical axis (it is symmetric).

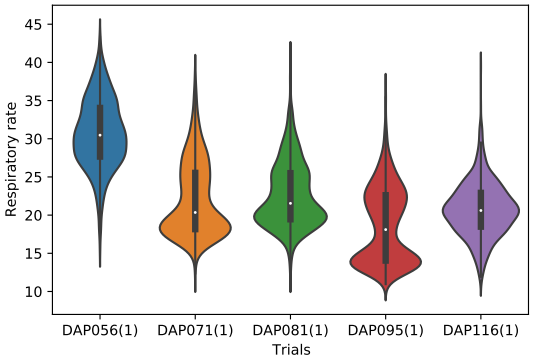

Figure 3.1: Respiratory rate violin plots for 5 random trials.

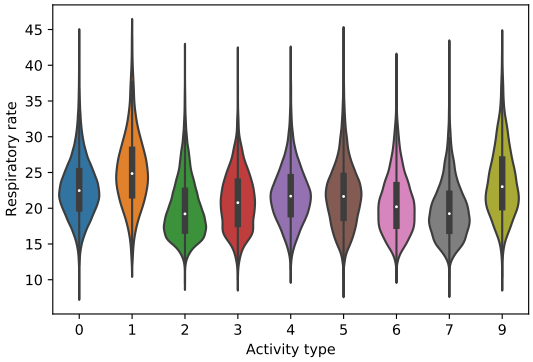

Figure 3.2: Respiratory rate distribution per activity type.

The shape measures of Table A.3 suggest a longer right tail in the probability distribution of the breathing rate (due to the positive skewness values), so that abnormal breathing rate values are usually greater than the mean, which is reasonable for asthmatic subjects whose respiratory rate can have unusual increases at certain points due to exacerbations. However the fact that the kurtosis value is lower than the one of a normal distribution (which is 3), indicates that there are no heavy tails (extremely low or high values) in the distribution.

Finally, the linear correlation between the respiratory rate and the activity level of the subjects have been computed with the Pearson's correlation coefficient and Kendall's tau. The former is not robust, since it divides the covariance of the two variables by the product of their standard deviations, and those measures are greatly

affected by outliers. The latter is robust since it only takes into account concordant and discordant pairs of data. A pair of data points (x_i, y_i) and (x_j, y_j) is concordant if when $x_i > x_j$, $y_i > y_j$ and when $x_i < x_j$, $y_i < y_j$. They are discordant when the opposite happens. Then, the number of discordant pairs is subtracted from the number of concordant ones and divided by the total number of pairs. Its value goes from -1 (negative linear relationship) to 1 (positive linear relationship) like Pearson's correlation coefficient.

The correlation measures have been computed for different time lags between the detected activity level and the respiratory rate. The results are shown in Table A.4, where a positive linear correlation is found for all lags, but it is not very strong (in the 0.2-0.3 range). The difference in the distribution of the breathing rate with different activity types is shown in Figure 3.2. Activity type 8 corresponds to subjects lying on the stomach, for which the data has been removed, so it is not included in the graph. Figure 3.3 shows the respiratory rate captured for the subject DAP001 in the second trial (DAP001(2)), which also shows the influence of the activity level during the night, as it drops down while sleeping. The impact of the activity type and level on the subjects' breathing rate has been removed in Section 4.4.

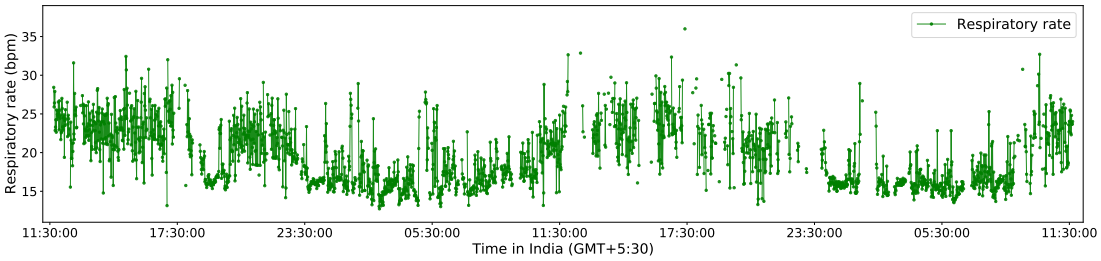

Figure 3.3: Respiratory rate data for trial DAP001(2).

3.2 AIRSpeck data

The AIRSpeck data can be divided in two: the one from the wearable sensors ($PM_{2.5}$, temperature and humidity), and the one from the static stations (which also includes NO_2 and O_3). The statistics of all of them will be analysed together in this section.

The static monitors register data every 5 minutes, as opposed to the 1 minute resolution of the wearable ones. Therefore, when resampling data to 1 minute, NO_2 and

O_3 have both a 63% of missing data, with respect to a 2% of personal exposure $PM_{2.5}$. The temperature does not have missing values and the humidity only a 0.2%.

The location measures for the AIRSpeck data are displayed in Table A.5. The mean $PM_{2.5}$ of the collected in Delhi is twice the median. There is also a significant difference between the mean and the median in the O_3 case (8750 vs 6859). This, together with the extreme maximum values found in the three pollutants (e.g. 65535 $\mu g/m^3$ for NO_2) and the temperature (with also an extreme minimum value), denotes the presence of outliers in the dataset, most likely due to measurement errors. All of them have either minimum or mode values of 0, which is due to the sensors having it as default value, so they have been filtered out in the pre-processing.

The mean and median values for the pollutants are much greater than the recommended 24-hour mean values set by the WHO in 25 $\mu g/m^3$ [24]. Their MAD and IQR (robust measures) shown in Table A.6 are also sizeable considering those guidelines. This might indicate that the personal exposure of the subjects to the pollutants in Delhi captured by the DAPHNE project is out of the reasonable range, but the data still needs to be calibrated, which have been done in Section 4.1, so conclusions cannot be drawn from the raw sensor data. Finally, the robust shape measures in Table A.7 indicate that there are no heavy tails in the distribution (extreme values) once the outliers are removed (the robust kurtosis is low, around 1.5), and that the data is mostly centred around the mean, with $PM_{2.5}$ having a longer tail of the distribution towards higher values than the mean (as it has a positive value for robust skewness) .

Since significant outliers have been detected in this EDA, only robust measures are reliable. Therefore, the correlation between the different AIRSpeck variables and the respiratory rate of the subjects have been computed with Kendall's tau. The results are displayed in Table A.8, showing very little correlation between them for the 5 time lags tested (all of them close to 0). However, Kendall's tau only tests for linear relationships, and the association between the variables may be non linear. In fact, the p-values for the statistic test are lower than 0.05 for all the features except NO_2, which means that the null hypothesis that says that there is no association between the variables can be rejected.

Chapter 4

Data pre-processing

The exploratory data analysis helped to determine the pre-processing steps that need to be carried out before any data analysis task is performed.

4.1 Calibration

The pollutant data has been gathered by different personal and static AIRSpeck sensors. The differences between those sensors is a potential issue, since a sensor can yield higher values than other one for the same air pollutant concentration. Therefore, in order to make their data comparable, they need to be calibrated. The calibration factors for $PM_{2.5}$ concentration data were already given by the Centre for Speckled Computing at the University of Edinburgh, since they were used for previous projects. However, static NO_2 and O_3 data had not been calibrated before.

The way the sensors are calibrated is by placing them near a reference Air Quality Monitoring Station for a period of at least 2 weeks, and comparing the collected data with the reference one to obtain the calibration factors that make the sensor data as similar as possible to the reference one. The calibration factors have been computed with a ridge regression model, explained in Chapter 5, which has been fitted with the data of the sensor to be calibrated, to predict the reference sensor data. In this way, the coefficients of the linear regression are obtained for calibrating the raw sensor data.

All the experiments were done for two different periods of time, since between October and November 2019 the sensors were put 1 km away from the reference sensor, and between January and April 2020 they were put right next to it. Three possibilities were considered for the set of features to train the machine learning model. The first one was to use the working and auxiliary electrodes for both NO_2 and O_3 (4 features).

The working electrode is the one responsible for the measurements, while the auxiliary one is meant to account for drift as a reference. The second one was to add the temperature and humidity values to the electrodes (6 features), and the last one to also include the $PM_{2.5}$ data (7 features).

Fitting the ridge regression model with all the available data could lead to overfitting, meaning that the model could adapt to the noise of the data, and have a bad performance on unseen data, although the performance on the training data is good. Since the objective is to have a good performance on new (unseen by the model) sensor data, two different calibrations have been performed for each pollutant. The first one fits the model with all the available data, and the second one performs cross-validation, splitting the data in 5, training the model 5 times (each one using a different split as test data and the rest as training data), and then averaging the performance of the 5 models on unseen data. Overfitting can be detected if the performance using all the data is much better than the one using cross-validation. Additionally, two different evaluation measures have been used: root mean squared error (RMSE) and mean absolute error (MAE). Only the latter is robust in the presence of outliers.

Figures 4.1 and 4.2 show for NO_2 and O_3 respectively, the MAE of the calibration with the three set of features, for the period of time between January and April 2020. The error when fitting the model with all the data is plotted in blue and, after cross-validation (CV), in orange. The trend is decreasing for both of them as the number of features used in the model increases.

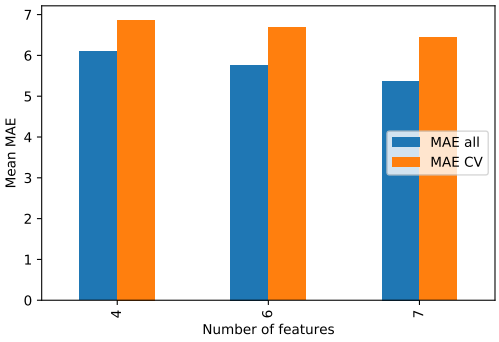

Figure 4.1: MAE of NO_2 calibration.

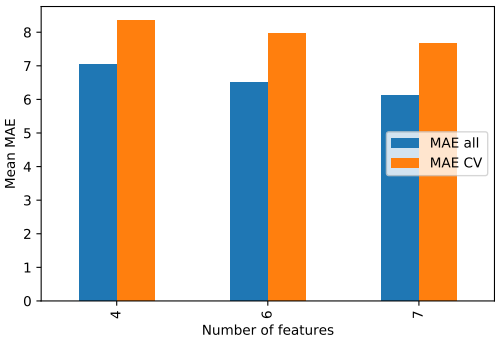

Figure 4.2: MAE of O_3 calibration.

Additionally, the relative increases of the cross-validation error (using unseen data to test the model) with respect to the fit-to-all error have been plotted for both pollutants in Figures 4.3 and 4.4. When looking at the MAE, the relative error increase goes up significantly with the number of features. The RMSE, however, has a peak when using only the electrode data (4 features). This is due to an error in the measurements of one

specific sensor, which makes RMSE go up as it is heavily influenced by outliers. This should not be taken into account, as before uploading it, the calibration of each sensor is evaluated manually to decide whether to include it or not (in case it has an error like this one). The significant difference in the cross-validation error with respect to the fit-to-all one when increasing the number of features is a sign of overfitting. Since the objective is to apply the calibration factors to new data, the model with only the electrode features (4 attributes) was selected for the calibration.

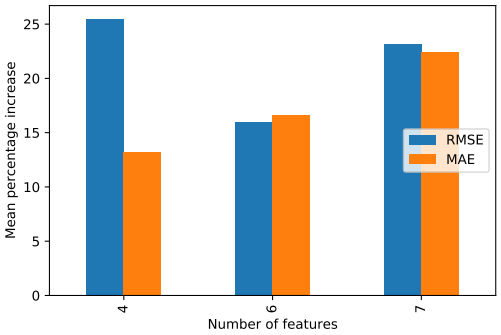

Figure 4.3: Mean percentage error increase from fit-to-all to CV for NO_2.

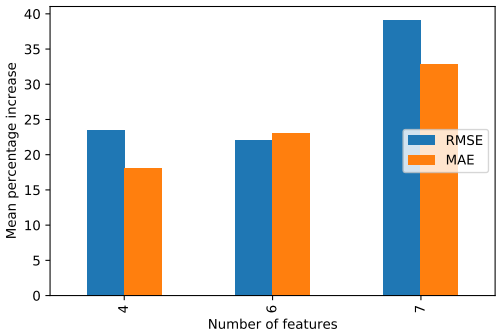

Figure 4.4: Mean percentage error increase from fit-to-all to CV for O_3.

Exactly the same trend happened in the period of time from October to November 2019, but the error values were higher, which is reasonable since the reference sensor was further away in this period. Therefore, the January-April period was selected for computing the calibration factors. The complete results, including the October-November period, the RMSE metrics and the sensor calibration leading to the RMSE increase in Figures 4.3 and 4.4 are included in Appendix B. Figure 4.5 shows the NO_2 data of a sensor before and after calibrating it with respect to the reference sensor.

4.2 Outlier removal

During the exploratory data analysis, outliers have been detected for most of the variables thanks to the use of robust and non-robust measures. These extreme values need to be removed before feeding the data to a statistical model, since they may affect its performance and could lead to wrong interpretations of the results. Two strategies have been tried out: Tukey's fences and winsorizing. The former considers as outliers the data points out of the $[Q_1 - k \cdot IQR, Q_3 + k \cdot IQR]$ interval, where k is usually 1.5, Q_1 and Q_3 are the first and third quartiles, and IQR is the inter-quartile range [38]. However, this technique, when applied individually to each subject, removed too

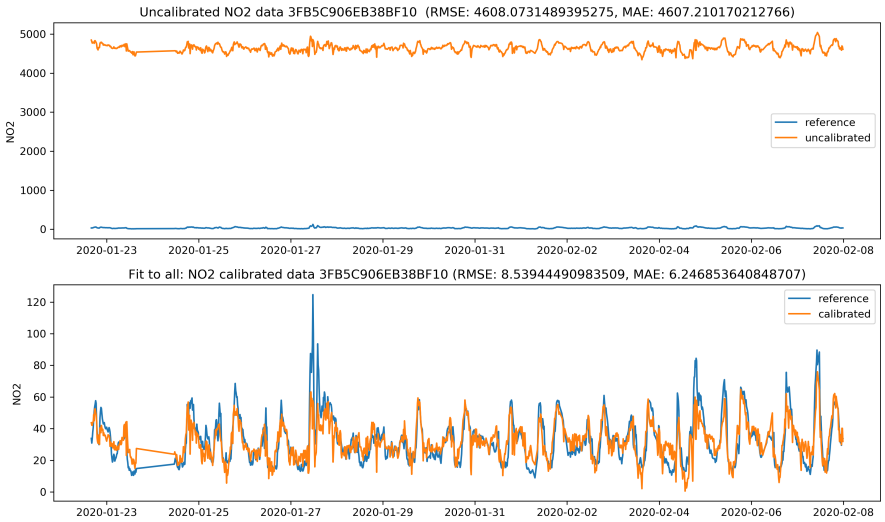

Figure 4.5: Sensor NO_2 data before and after calibrating to the reference one.

much data, including reasonable (although high) values that could give insights in the subsequent analysis. Therefore, winsorizing was used, which consists of removing the values above a certain quantile and below another one. A trade-off is required between the amount of non-outlier data that might be eliminated and the extreme values that need to be removed, so the 5^{th} and 95^{th} percentiles were selected as the boundaries beyond which data points are considered outliers, and therefore removed.

4.3 Missing data imputation

As reported in Chapter 3, the $PM_{2.5}$ data has a 2% of missing values, which is increased to 37% for the respiratory rate and 63% for NO_2 and O_3. Some statistical models, like the causal discovery method that has been used in this project, do not admit missing values, so they need to be imputed in the first place. An interpolation strategy has been selected to impute reasonable values, estimating the missing data points based on the present ones. However, with large gaps, the interpolation might turn unrealistic. Therefore, the maximum gap size over which to interpolate needs to be set.

Algorithm 1 was originally defined in [17], and has been reused in this project to determine the maximum gap for each variable, testing values from 5 to 60 minutes in 5 minutes steps. The set of trials used to test each variable have been selected as a trade-off between the amount of trials used (to be as high as possible) and the maximum percentage of missing data (to be as low as possible). Therefore, the thresholds

on the maximum percentage of missing data were 10%, 25%, 80% and 80% for $PM_{2.5}$, breathing rate, NO_2 and O_3, respectively. The maximum allowed mean absolute percentage error introduced due to the interpolation has been set to 15%, resulting in a maximum interpolation gap of 15 minutes for the respiratory rate, $PM_{2.5}$ and NO_2, and 25 minutes for O_3. After the standardising procedure explained in the next section, the remaining missing data has been imputed with the mean, to influence as little as possible the subsequent statistical models that make use of the data.

Algorithm 1 Maximum interpolation gap calculation [17]

Input: The maximum gap to test: max_gap
Let 'MAPE' be the Mean Absolute Percentage Error
$num\_trials =$ number of trials with enough data for the experiment
for all trials with enough data for the experiment **do**
  Randomly select $1000/num\_trials$ timestamps of the time series
  for all selected timestamps **do**
    Remove max_gap consecutive minutes beginning with the current timestamp
    Interpolate the gap
    Add the MAPE between the true and interpolated data to a counter
  end for
  Compute the MAPE average for the trial and add it to a $total\_MAPE$ counter
end for
return $total\_MAPE/num\_trials$

4.4 Normalisation

Normalisation (or standardisation) is a common procedure in data analysis to make data in different scales comparable. For example, when using two different attributes to fit a linear regression model, the contribution of each of them can only be compared with the resulting linear coefficients if they are on the same scale. In Section 3.1, the distribution of the respiratory rate of the subjects was shown to change with the type of activity they were carrying out, and it was positively correlated with the activity level. This influence can also be removed through standardisation, by subtracting the trend it produces throughout the day, so that the breathing rate no longer depends on it.

The main strategies for data normalisation are min-max and z-score normalisation [8]. The former consists of scaling the data into a specified interval with a minimum

and maximum value, which is usually [0,1]. However, this approach requires the different variables to have a specified minimum and maximum value, which is not the case. Therefore, the z-score strategy has been selected, which consists of scaling the data to a zero mean and a standard deviation of 1 by subtracting the mean to each data point and dividing the result by the standard deviation.

The mean of a time series can be thought of as the trend that the data follows through time. If the trend is subtracted, then the time series would be centered around zero. The Locally Weighted Scatterplot Smoothing method (LOWESS) was introduced in [6] and can estimate the trend value at a given point by first selecting the K nearest data points and then performing a linear regression weighted with the distance of each of the selected values to the target. The number of data points has been set to 30, so that the data 15 minutes before and after is used to estimate the trend. Figure 4.6 shows the trend computed using LOWESS for the DAP001(2) trial shown previously on Figure 3.3. The computed trend is subtracted from the data to remove the mean. Finally, the standard deviation of each time series is computed with a moving window of 30 minutes over the time series, and the result of subtracting the mean is divided by it. The standardised data for trial DAP001(2) is displayed in Figure 4.7.

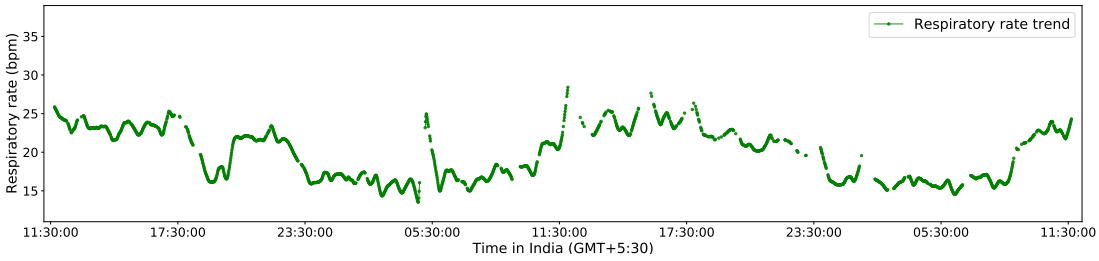

Figure 4.6: Respiratory rate trend for trial DAP001(2).

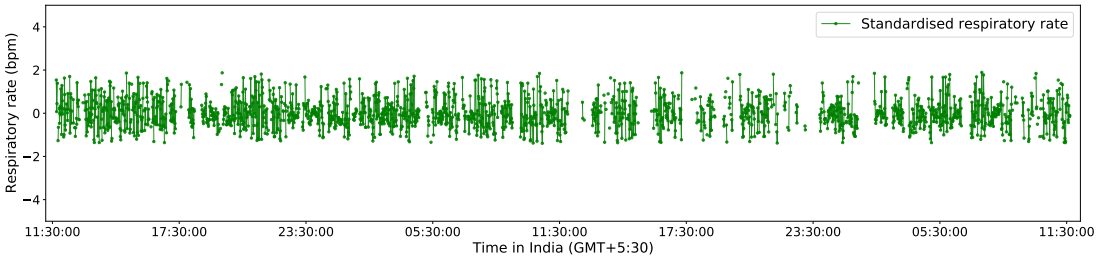

Figure 4.7: Respiratory rate standardised data for trial DAP001(2).

Chapter 5

Machine learning experiments

Before assessing the causality in the relationship between the pollutants and the respiratory rate, several preliminary experiments have been performed to demonstrate that an association between them exists. A variety of machine learning methods have been used for this task, since they are capable of capturing the relationship between different attributes and a target variable, and use it to predict the selected variable with any choice of values for the attributes. A regression task has been performed, in which the current breathing rate of a subject needs to be predicted using past respiratory rate and pollutant concentration values. The way to determine an association is to first compute the prediction error when only using past values of the breathing rate, and compare it to the error achieved when adding the data of a pollutant. If the error produced is lower in the latter case, it means that the pollutant concentration values encode information about how the breathing rate of the subject behaves. The evaluation measure used has been the mean absolute error (MAE), which is robust to outliers.

Both linear and non-linear machine learning methods have been selected to be able to capture any type of relationships between the variables. Two different linear methods have been tried out: ridge and Huber regression. Both aim to fit a line, plane or hyperplane to the available data (depending on the dimensionality of the data) and use it to predict the target variable at different attribute values, but they differ on the optimisation method used. Ridge regression minimises the mean squared error (MSE), while Huber regression does it with Huber loss. The latter uses the mean absolute error except for values close to zero (where MSE is used), so it is much more robust to outliers (since the error is not squared). Every machine learning method has several hyper-parameters that can be adjusted for a better prediction. In this case, the hyper-parameter adjusted for both was the L2 regularisation strength, which needs to be set

in a way that avoids overfitting but that makes a good fit of the data. 21 different regularisation values have been tested in a logarithmic scale from 10^{-5} to 10^{5} (with higher values there is less overfitting but a worse fit of the data).

Non-linear machine learning models are able to find patterns beyond linear relationships. Random forest has been the first non-linear model to be tried out, consisting on a group of decision trees, each one built with a random subset of the data, whose result is the average output of all the trees. The hyper-parameter tuned for this model was the maximum depth of the decision trees, which can control overfitting, testing from 2 to 2^{10} (using powers of two), as well as no restriction in the depth. K-nearest neighbours is the second non linear model, which yields the average target value of the K closest training data points to the evaluated one. In this case, the hyper-parameters were the number of neighbours (from 1 to 2^{10} in powers of 2), and whether the neighbour target values are all taken equally into account or if they are weighted taking into account the distance to the evaluated data point. Finally, support vector regression has been used, which is capable to uncover many types of non-linear relationships with the use of non-linear kernels to transform the data. The type of kernel used has been tuned to account for different types of relationships, including linear, polynomial of 2^{nd}, 3^{rd} and 4^{th} order, radial basis function and sigmoid transformations. These methods have been implemented with the scikit-learn Python package [27].

The way supervised machine learning is carried out with time series is to first apply a time lag to the attributes and use them to predict the current target variable value. In the case of predicting the breathing rate with its past values, if a time lag of 10 minutes is considered, several tuples will be made in which the second element is the breathing rate at a given timestamp (the target variable), and the first value is the respiratory rate 10 minutes before that timestamp (feature used for the prediction). All the time lags from 1 to 60 minutes have been tested for all the attributes in these experiments. For the hyper-parameter selection of the models, a cross-validation (CV) procedure has been followed, dividing the data into 5 folds as explained in Section 4.1 for the sensor calibrations. Since only pairs of data points are considered, covering all the gaps in the time series is not essential, so only the interpolation of missing data has been used without imputing the rest with the mean, as doing so could lead to artificial results.

The first step is to compute the prediction error of each model when using only the past breathing rate data. For each of the models, the cross-validation has been carried out for every time lag of the breathing rate in the first hour, and once the best parameters are selected, the models have been evaluated with a test set to assess their

performance on unseen data. The top 5 respiratory rate time lags with the lowest MAE were stored for each machine learning method for the subsequent experiments.

The process was then repeated, but using two attributes for the prediction: the past breathing rate and the past pollutant concentration. Algorithm 2 shows the machine learning pseudo-code used for each model and each of the 3 pollutants, in which the 60 time lags are tested for the pollutant, for each of the top 5 breathing rate time lags. Finally, the prediction error when only using the breathing rate at the 5 different time lags and when adding each of the pollutants, have been compared to determine whether they encode information about the respiratory rate behaviour.

Algorithm 2 Machine learning procedure with respiratory rate and pollutant data

```
Input: The pollutant to be tested and the machine learning model
for all trials do
    Get trial pre-processed data
    for all top 5 respiratory rate time lags do
        for all pollutant time lags from 1 to 60 minutes do
            Set the lagged breathing rate and pollutant data (using their respective lags)
            as input features and the original respiratory rate data as the expected output
            Split the data randomly into training set (80%) and test set (20%)
            for all hyper-parameters do
                Initialise the machine learning model with the hyper-parameters
                Split the training data in 5 folds
                Compute the CV error by averaging the errors produced when using each
                fold as validation set (and the rest as training set for the model)
                If the CV error is the lowest found so far, save the hyper-parameters used
            end for
            Initialise the model with the best set of hyper-parameters found
            Fit the model with the whole training data and compute the error (MAE)
            when predicting the test data
            Save the MAE for the current trial and combination of time lags
        end for
    end for
end for
```

The $PM_{2.5}$ personal exposure data was the first pollutant for which the association was tested. The five machine learning models yielded similar results. Figures 5.1 and

5.2 show the average MAE of all the trials, for the best performing linear and non-linear models (Huber regression and random forest), respectively. The full results have been included in Appendix C.1. The errors at the best respiratory rate time lag (1 minute) are very similar with and without the pollutant information, being almost the same for the linear model. The distance between them is increased with the subsequent time lags, but it is still small, showing that personal $PM_{2.5}$ for the tested trials encodes information about the breathing rate, although it does not add much value to only using the respiratory rate for the prediction.

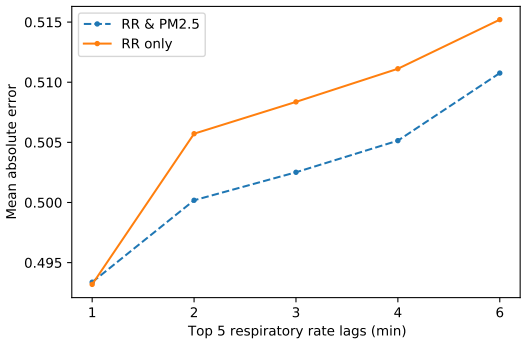

Figure 5.1: Huber regression average error comparison personal $PM_{2.5}$.

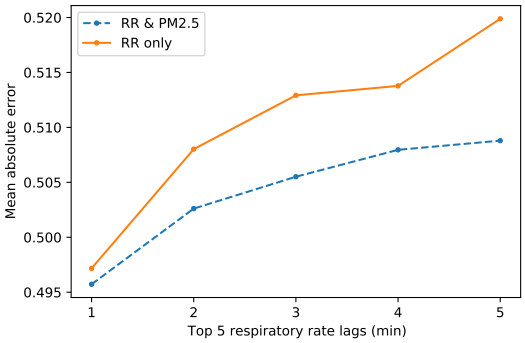

Figure 5.2: Random forest average error comparison personal $PM_{2.5}$.

Subsequently, the NO_2 and O_3 relationship with the breathing rate was evaluated. Huber and random forest regression were also the models that achieved the lowest average MAE for both pollutants. They obtained a significantly lower error when the pollutant information is included, of around 0.38, compared to the $PM_{2.5}$ personal exposure data, which achieved MAE values around 0.49 (very similar to only using the respiratory rate). Taking into account that the data has been standardised to approximately 0 mean and a standard deviation of 1, including the NO_2 or O_3 information results in a drop in the prediction error of approximately a 10% with respect to the standard deviation, and around a 22% reduction with respect to the MAE produced when only using past breathing rate values. This suggests a strong relationship between these gases and the respiratory rate of the tested subjects, as they contain information of its behaviour which is captured by the machine learning models for the prediction.

Nevertheless, the set of subjects used for personal $PM_{2.5}$ exposure and for NO_2 and O_3 is not the same. The reason for this is that only the static sensors for which a reasonable calibration was performed with respect to the reference sensor was included (they were revised individually). Therefore, to do a fair comparison, the personal $PM_{2.5}$ results have been plotted for the same subset of trials as the gases. The result for the best

two models are shown in Figures 5.3 and 5.4, and the rest are included in Appendix C.2. The average prediction errors are very similar for the three pollutants, and reduces significantly ($\approx 22\%$) with respect to only using the respiratory rate. Nevertheless, this does not mean that they have exactly the same effect on the breathing rate, but that they encode a similar amount of information to predict it.

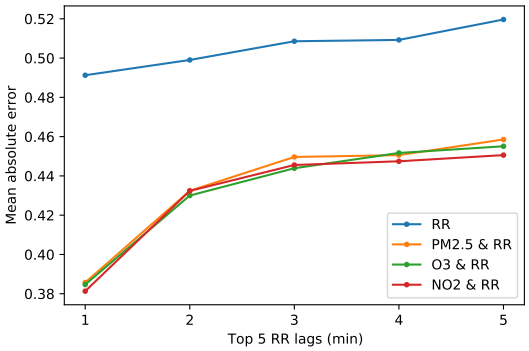

Figure 5.3: Huber regression average error comparison all pollutants.

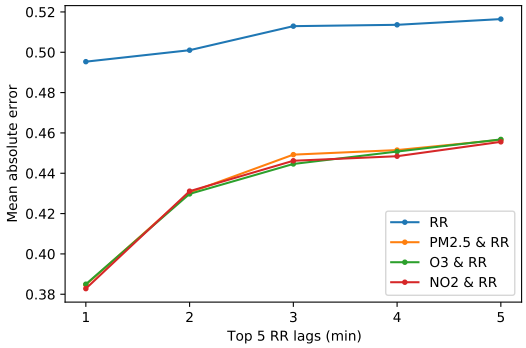

Figure 5.4: Random forest average error comparison all pollutants.

One further experiment was made to demonstrate the association, by evaluating the performance of the best two machine learning models (Huber and random forest regression) when only using the pollutant information, without the breathing rate, and comparing it to using only past breathing rate values. The results are shown in Figures 5.5 and 5.6, in which the three pollutants achieve a lower MAE than when using the respiratory rate. This is a sign of the strength of the association, since they contain more information about the breathing rate behaviour than the respiratory rate itself.

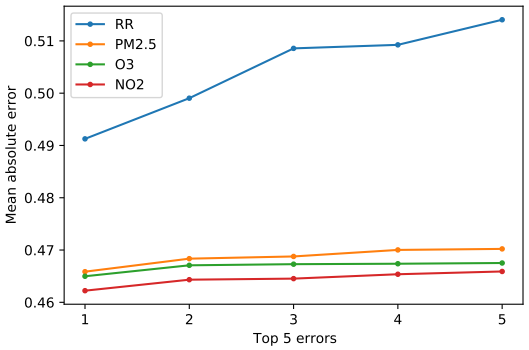

Figure 5.5: Huber regression average error comparison pollutants alone.

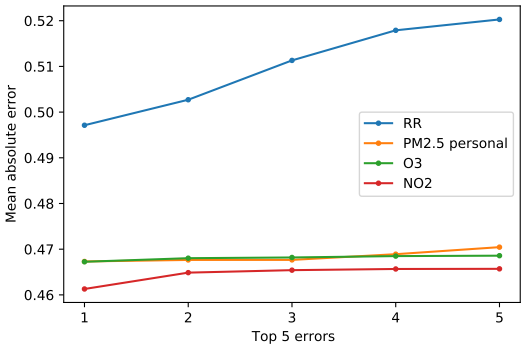

Figure 5.6: Random forest average error comparison pollutants alone.

In the last two experiments, the three pollutants had a very similar prediction performance, even though the NO_2 and O_3 data come from static sensors further away

from the subject than the $PM_{2.5}$ data. Therefore, a new experiment was devised to determine whether using personal exposure data had any actual benefit for the prediction, by fitting the models with both personal and static $PM_{2.5}$ data. The latter was obtained from the same static sensors as NO_2 and O_3, which also measure particulate matter. It was pre-processed the same way as the other pollutants, and its calibration was already given (as for personal $PM_{2.5}$) from previous projects. Once again, the results were very similar, with the static data actually giving a slightly better performance for both models, as seen in Figures 5.7 and 5.8. This has been explored further in Section 6.3 with a causal discovery method, to see whether the usage of wearable sensors really has an advantage, or if having nearby static ones as in [34] is sufficient.

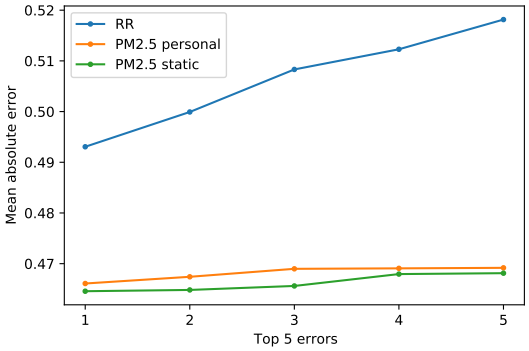

Figure 5.7: Huber regression average error personal vs static $PM_{2.5}$

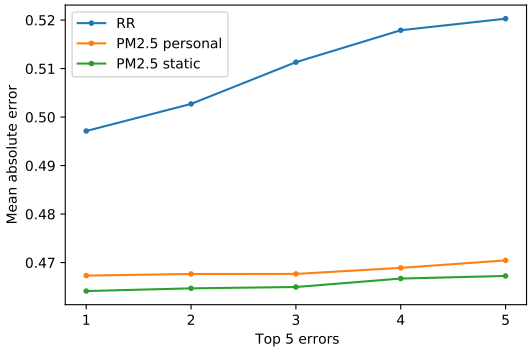

Figure 5.8: Random forest average error personal vs static $PM_{2.5}$.

When the different attributes are standardised to the same scale, the linear regression coefficients can tell which input feature affects the most the value of the predicted variable. This does not determine which variable contains more information about the target one (which was determined with the previous experiments), but rather which one makes the target value increase or decrease the most when its value is increased by one unit. Therefore, a final experiment was devised by fitting the Huber regression model with past breathing rate and the data from the three pollutants, using the best time lags found for each of them in the experiment shown in Figure 5.5. The result was that the increase of the past breathing rate (one minute ago) by one unit is the one that increases more the predicted value with a positive linear relationship, which is reasonable since the breathing rate one minute ago will be similar to the current one. It was followed by NO_2 and O_3 with a negative linear relationship, and $PM_{2.5}$ with a positive one. This, however, only takes into account a specific time lag in the first hour, so the exposure-response relationship has been further analysed in Chapter 7.

Chapter 6

Causal discovery method

In the previous chapter, an association has been demonstrated between the three pollutant concentrations and the respiratory rate. In this chapter, a method called PCMCI has been used to determine whether this relationship is causal, and at which time lags. PCMCI was introduced in [32] as a causal discovery method than can relate several large-scale time series and detect at which time lags there is a causal relationship with a previously selected target variable.

The method has several assumptions, detailed in [30]. The first one is the causal sufficiency, which assumes that all the elements that can be causes of at least two of the included variables take part in the analysis. To account for this, in addition to the pollutants and the past breathing rate, the temperature and the humidity have been included, as an association between meteorological factors and the condition of asthmatic subjects has already been demonstrated [14]. The causal Markov condition and faithfulness assumptions complement each other. The former means that if no causal link is found between a time lag of a variable and the target time series, then they are conditionally independent. The latter complements it stating that if two variables are conditionally independent, then the method will not find a causal link.

PCMCI also requires the different time series to be stationary, which means that at different points of time, the properties of the data will remain the same. This is true in this case, since the data has been standardised to be stationary in the mean and the standard deviation. Finally, it also assumes that the value of a variable in a certain point of time does not influence the value of another one at the same time.

The method can be divided in the condition-selection phase and the momentary conditional independence one (MCI). The former is a modified version of the PC algorithm introduced in [35], which aims to find a group of time-lagged variables that may

be a cause for the target variable (called the parent set). It focuses on maximising the true positives, so the result can include several false positives that need to be filtered out in the second stage. Algorithm 3 show how the condition-selection phase operates to obtain such set, testing for conditional independence between the variables.

Algorithm 3 Modified PC algorithm used in PCMCI [35]

Input: Time series data, target variable, significance level, maximum time lag, conditional independence test

Let the data be a collection of time series $\mathbf{X}_t = (X^1_t, X^2_t, \ldots, X^N_t)$ and the target variable X^j_t

Let τ_{max} be the maximum time lag tested

Let $X^j_{t-\tau}$ mean the variable X^j_t shifted with a τ time lag.

Initialise parent set of X^j_t ($P(X^j_t)$) with every variable at every time lag up to τ_{max}.

for $p = 0$ **to** $p =$size of $P(X^j_t)$ **do**

  for all variables $X^i_{t-\tau}$ in $P(X^j_t)$ **do**

    if $p == 0$ **then**

      Test $X^i_{t-\tau} \perp\!\!\!\perp X^j_t$ with the given conditional independence test

    else

      $S =$ first p variables in $P(X^j_t)$

      Test $X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid S$ with the given conditional independence test

    end if

    If the hypothesis cannot be rejected at the given significance level, tag $X^i_{t-\tau}$ to be deleted later from $P(X^j_t)$

  end for

  Delete tagged variables from $P(X^j_t)$

  Sort $P(X^j_t)$ in descending order with the test statistics of each variable

end for

return $P(X^j_t)$

The condition-selection stage can be seen as a dimensionality reduction step for the subsequent MCI test, obtaining a reduced set of variables to test for causality with the target one (which is a computationally expensive method). In the MCI phase, the independence between the target variable and each of the variables contained in the parent set returned by the PC algorithm is tested, conditioned not only on the rest of the parents of the target variable, but also on the parents of the evaluated variable from the provisional parent set. This is done to detect autocorrelation, and allows the

PCMCI method to be very robust against false positives [35]. Therefore, only those variables $X^i_{t-\tau}$ for which the null hypothesis $X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid P(X^j_t)$ and $P(X^i_{t-\tau})$ can be rejected with a given significance level, have a causal link with the target variable X^j_t. The existence of such links means that the values of the evaluated time series at the given τ time lag are a cause for the values of the target time series. A significance level of 5% has been chosen for the p-values, as it is standard in the literature [35].

Both parts of PCMCI require a conditional independence test to be defined. The Python package called Tigramite [29], which has been used for the method, supports several tests. The partial correlation test is the fastest, but it assumes a linear relationship between the variables. On the other hand, the test introduced in [31] consisting of a k-nearest neighbour algorithm for the conditional mutual information between variables (CMIknn), makes no assumption about the type of relationship, so it can uncover any non-linear dependency. The downside is that it is much more computationally expensive, and that it has less power detecting linear relationships than the partial correlation test [32]. The linear test has allowed to evaluate the linear relationship between the pollutants, humidity and temperature, and the breathing rate in time lags up to 20 hours with a 1 minute resolution. With the non-linear one, only time lags every 10 minutes up to 8 hours have been tested due to the vast amount of time it takes to run, although for the first hour every minute has been examined.

6.1 Personal $PM_{2.5}$ exposure causal relationship

The first objective with PCMCI was to reproduce the results obtained in [17] for the causal relationship between the personal $PM_{2.5}$ exposure, humidity and temperature, and the breathing rate of the subjects. Both linear and non-linear conditional independence tests were used, in which the tested time lags were the first hour with a 1 minute resolution for the linear approach, and the time lags of 1, 5, 10, 15, 30, 45 and 60 minutes for the non-linear one.

6.1.1 Linear PCMCI

The PCMCI results using the partial correlation test (which assumes linear relationships) were completely reproduced for the first hour, and they were further expanded up to a time lag of 20 hours, testing for every time lag with 1 minute resolution. The previous results for the first hour changed slightly when the maximum time lag was

expanded, since the possible variables in the parent set are increased. It is possible that some lagged $PM_{2.5}$ observations for which a link was found when only testing the first hour, is conditionally independent from the breathing rate of the subject given the information provided by a longer time lag, as explained in Algorithm 3.

In order to obtain meaningful results as least influenced as possible from the error introduced when imputing missing data, only trials for which less than 40% of their data was missing for the four time series involved in the analysis: breathing rate, personal $PM_{2.5}$, temperature and humidity, were selected. Additionally, those with less than 40 hours of data were discarded because PCMCI requires the time series data to be at least twice as long as the maximum time lag. The reason for this is that the last observation (at hour 20) also needs to be tested for a 20-hour time lag. As a result, 50 trials were selected for evaluation. An average of 58.82 causal links were found for each of them, with a minimum of 30 and a maximum of 80.

The average number of links per hour is rather constant, although there are two peaks, the first one between the 7^{th} and 10^{th} hours, and the second one at the end, in the 18^{th} hour. However, the distribution of links has a great variability depending on the subject, which is why the analysis has been made personalised. Two examples of this are displayed in Figures 6.1 and 6.2, where the former has a more pronounced short term effect, and the latter a long-term one (inside the first 20 hours). The distribution for all the trials has been included in Appendix D.1.1.

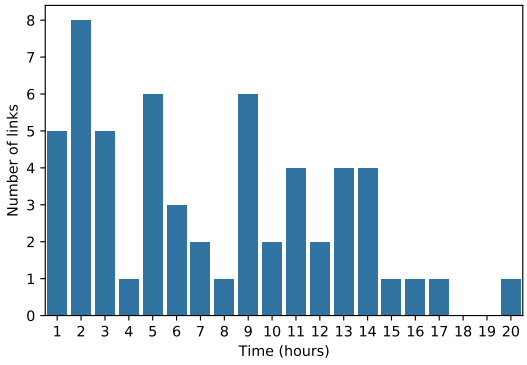

Figure 6.1: DAP095(1) causal links distribution for $PM_{2.5}$ in linear PCMCI.

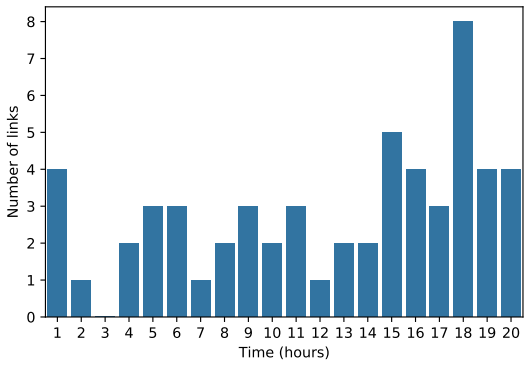

Figure 6.2: DAP067(2) causal links distribution for $PM_{2.5}$ in linear PCMCI.

6.1.2 Non-linear PCMCI

As mentioned before, the partial correlation test assumes that the relationship between the variables is linear, which could create spurious causal links if it is not the case.

Therefore, the CMIknn test is more reliable in the sense that it can discover any type of relationship. It is a much more computationally expensive method, and that is why in [17] only 7 time lags inside the first hours could be tested. However, those results could not be reproduced exactly, obtaining very similar values but not the same ones. This non-linear test requires a random permutation [31], so that different executions of the test yield slightly different results. To ensure that the outcomes of this research can be fully replicated, the same seed (0) for the random number generator has been set before each PCMCI execution.

The analysed time lags have been expanded up to 8 hours, but every minute could not be tested because it would take more time than what is available for this project. Therefore, every ten minutes from the start of the second hour to the end of the eight hour were tested, in addition to the 7 lags in the first hour. As 8 hours were tested instead of 20 for the linear one, more trials could be tested (55). An average of 7.76 causal links were found per trial, which is around three times the proportion of causal links with respect to the number of tested links obtained by the linear approach (15.84% and 4.9% respectively). The distribution of the total number of links found for all the subjects at the different time lags is shown in Figure 6.3, where there is a causality peak at the very short term ($<$ 30 minutes after the exposure) and a second one around the 7^{th} hour, which suggest a second reaction to the pollutant.

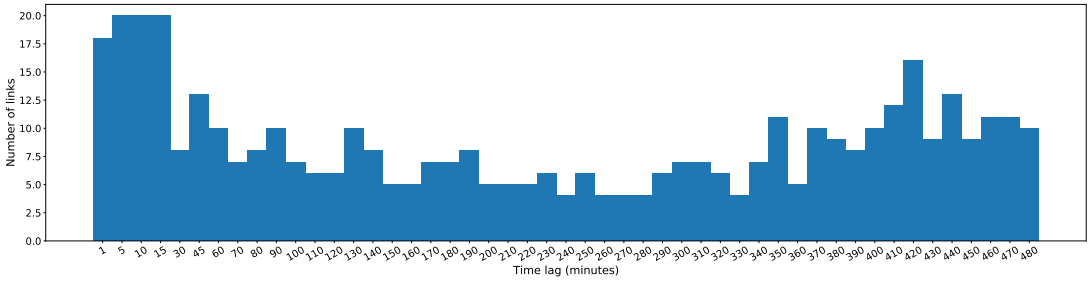

Figure 6.3: Non-linear PCMCI causal link distribution for $PM_{2.5}$

Causal links were found for every trial except for one (DAP030(1)), as shown in Figure 6.4, where the x-axis represents the 55 different trials. In order to check whether that subject's breathing rate is also influenced by the pollutant, the 480 time lags up to the 8^{th} hour (minute by minute) were evaluated for that trial. Five causal links were discovered, demonstrating that there is no trial for which the exposure to $PM_{2.5}$ influences its breathing rate at certain time lags. Additionally, to analyse the impact of

the pollutant in the shortest term, the causality of every time lag (minute by minute) in the first hour was evaluated. A quarter (25.39%) of the tested time lags for all the trials had a causal relationship with the breathing rate of the respective subjects. Only the trials DAP075(2) and DAP109(1) had no causal links in the first hour, but they had in the following ones. Overall, there was a peak in the number of links in the first 20 minutes and it starts decaying as the time passes, as it can be seen in Appendix D.1.2.

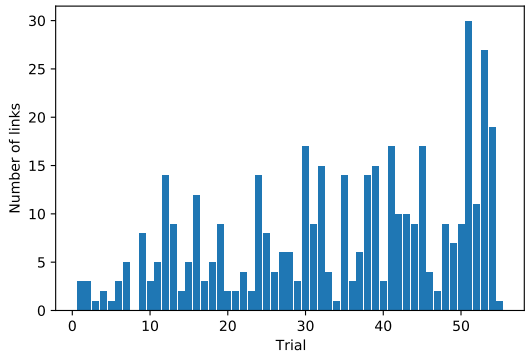

Figure 6.4: Number of links per trial for $PM_{2.5}$ with non-linear PCMCI.

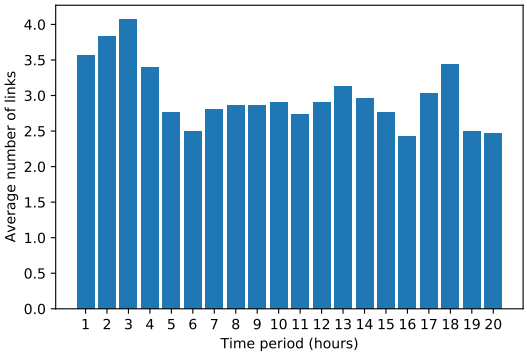

Figure 6.5: Average causal links distribution for $PM_{2.5}$ with linear PCMCI.

These results strongly suggest causality between $PM_{2.5}$ concentrations and the breathing rate of the subjects, having a high proportion of causal links considering that PCMCI has a very strong false positive control [35]. Appendix D.1.2 contains the complete results of this experiment for each trial and time lag.

6.2 Personal $PM_{2.5}$, NO_2 and O_3 causal relationship

In the previous section, the results achieved in [17] were reproduced and expanded, showing a strong causality between the exposure to $PM_{2.5}$ and the respiratory rate of the subjects. However, $PM_{2.5}$ is not the only pollutant that the subjects are exposed to. The causal sufficiency assumption of PCMCI states that all the possible causes for the breathing rate should be included in the analysis, to obtain results as accurate as possible. Therefore, the other two pollutants for which the DAPHNE dataset has information (NO_2 and O_3) have been added to the $PM_{2.5}$ personal exposure data.

6.2.1 Linear PCMCI

The first experiment was done with the partial correlation test, which allows to evaluate up to 20 hours, although making the linear relationship assumption. This can be

indicative of which relationship between the different pollutants and the respiratory rate has more linear components, which the non-linear test has less power to detect.

The selected set of trials is different from the one used in Section 6.1 because more variables are being considered and they have a higher percentage of missing data, as explained in Section 3.2. Additionally, some of the trials do not have calibrated values for NO_2 and O_3, since only those sensors with a good fit to the reference sensor when calibrating were selected. Therefore, only the trials with less than a 60% of missing data across all the variables (including the temperature and humidity) with at least 40 hours of data (as explained in the previous section) were considered (30 trials).

The distribution of the average number of causal links per trial between $PM_{2.5}$, NO_2 and O_3, and the breathing rate (all for the same set of 30 trials) are displayed in Figures 6.5, 6.6 and 6.7, respectively. The causality of the linear relationship between $PM_{2.5}$ and the respiratory rate seems to be stronger in the first hours after the exposure, in contrast to the one of the ozone, which is more pronounced in the last hours from the tested interval. The distribution of the NO_2 does not seem to follow a specific trend, having several spikes throughout the time interval. The distributions vary significantly with the evaluated subject, so the personalised analysis for each trial and pollutant has been included in Appendix D.2.1. The number of causal links found for each pollutant are very close to each other, with an average of 61.53 links per trial for NO_2, followed by O_3 with 60.33 and $PM_{2.5}$ with 59.93.

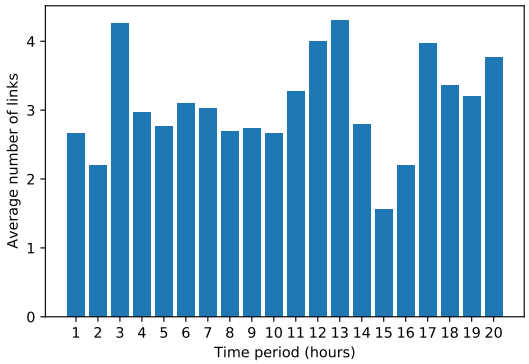

Figure 6.6: Average causal links distribution for NO_2 with linear PCMCI.

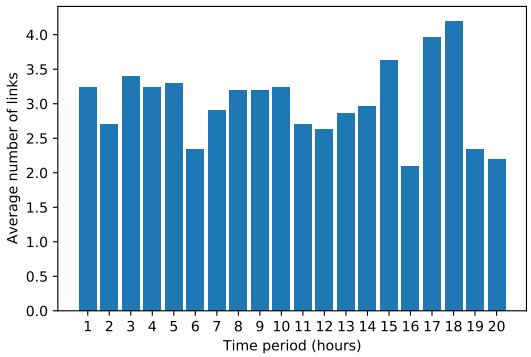

Figure 6.7: Average causal links distribution for O_3 with linear PCMCI.

These results provide insights into how the causality of the linear components in the relationship between the breathing rate and the different pollutants behaves over an extensive period of 20 hours. However, spurious results may occur if the relationship is not truly linear [32], which seems to be the case since the linear correlation between

them computed in Section 3.2 was very close to zero, especially for NO_2.

6.2.2 Non-linear PCMCI

Using the CMIknn test in the PCMCI method with the temperature, humidity, breathing rate and the three pollutants data is the most reliable experiment that can be done to determine the causal relationships between the different time series with the available data. The test is general enough to discover any type of relationship, and every possible cause for the breathing rate in the DAPHNE dataset has been included, since the time of the day and activity level influence were removed in Section 4.4, therefore satisfying the causal sufficiency assumption.

It is also the most computationally expensive experiment due to the generality of the test and the fact that two new time series have been added (NO_2 and O_3). Nevertheless, the same procedure has been followed, performing a detailed analysis of the first hour, and every ten minutes from then up to 8 hours. The execution of the algorithm had to be parallelised across several machines, using threads to run on every core of each of the machines (varying from 4 to 6 cores). The algorithm takes around 3 days to evaluate all those time lags for a single trial, being the only process running in the computer. There are 44 trials with less than 60% missing data and with enough data to test time lags up to 8 hours (compared to 30 when testing up to 20 hours), so it would have taken more than 4 months to run this experiment if it had not been parallelised.

Figures 6.8, 6.9, 6.10, 6.11, 6.12 and 6.13 show the distribution of the total number of causal links obtained between past respiratory rate, temperature, humidity, $PM_{2.5}$, NO_2 and O_3 values, and the respiratory rate for all the trials at the different time lags. The values shown for the first hour, where every time lag was tested, are the mean number of links every 10 minutes.

The causal effect of the past breathing rate values is the strongest in the first hour, but it decays swiftly as time passes. This is intuitive, since the breathing rate a minute ago will certainly influence the current one. The same happens with the meteorological factors, having a greater impact on the breathing rate in the first hour than the pollutants and decaying afterwards, but with a slighter effect than past respiratory rate values.

The causal effect of $PM_{2.5}$ on the breathing rate seems to be maintained through the different time lags, although a peak in the causality can be observed between the 7^{th} and 8^{th} hour, following a decrease produced after the 6^{th} hour. In contrast, the effect of the gases (NO_2 and O_3) is much more pronounced in the first hour, followed

by two peaks in the number of links for both of them, one in the mid term of the tested interval, and another one in the last hour.

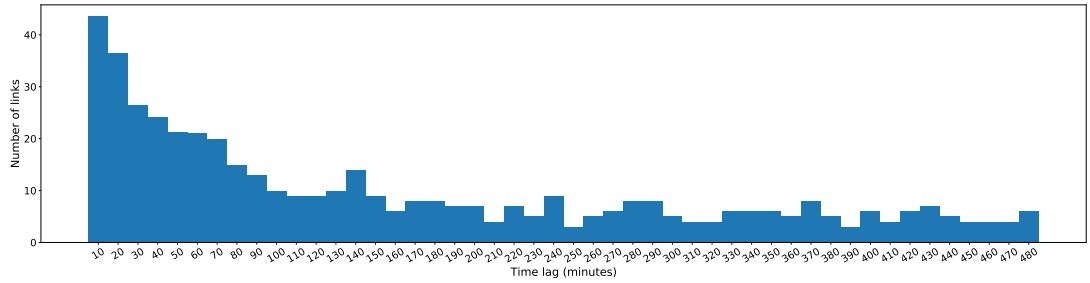

Figure 6.8: Non-linear PCMCI causal link distribution for the breathing rate.

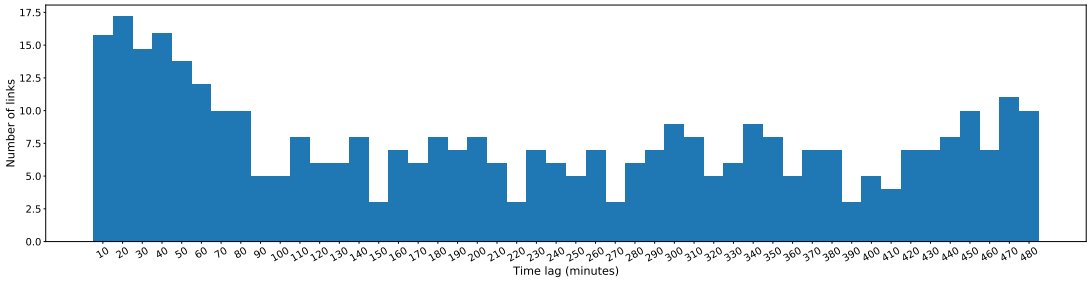

Figure 6.9: Non-linear PCMCI causal link distribution for the temperature.

Table 6.1 shows the proportion of causal links found for each of the lagged variables with respect to the number of tested links, for both the first hour with 1 minute resolution and the period from the 2^{nd} to the 8^{th} hour for which every time lag in ten minutes steps have been tested. After the past breathing rate values, the humidity and temperature have the strongest causal relationship with the breathing rate of the subjects, especially in the first hour after the exposure. This was also observed in [14], where the meteorological factors, including temperature and humidity, had a higher association in the short term with the clinical visits of children due to asthma or other allergies, although the impact on the breathing rate had never been examined.

The three pollutants obtained a sizeable proportion of causal links, meaning that more than 1 out of 5 of the evaluated lagged concentrations of these pollutants have a causal influence on the respiratory rate of the asthmatic subjects. The impact is increased in the first hour period from the exposure to around 1 out of 4 tested time lags. The proportion of causal links found for $PM_{2.5}$ has been reduced a 1.59% with

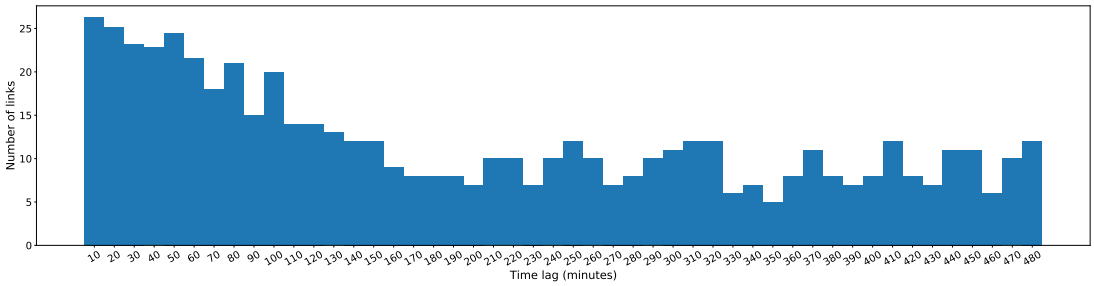

Figure 6.10: Non-linear PCMCI causal link distribution for the humidity.

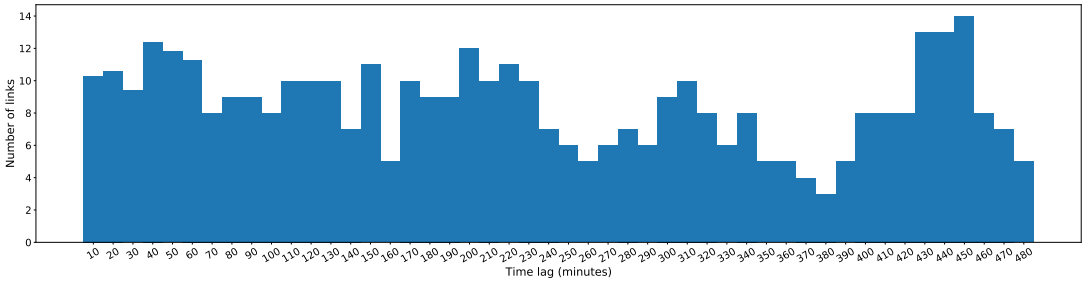

Figure 6.11: Non-linear PCMCI causal link distribution for $PM_{2.5}$.

respect to when it was the only pollutant included, because some of the links can be explained with the newly introduced NO_2 or O_3 information. Overall, the pollutant with the highest impact on the breathing rate was the NO_2, followed by $PM_{2.5}$ and O_3. The more aggressive effect of NO_2 in the short term compared to the other two has been detected in previous work when studying the effects of the pollutants on other indicators of respiratory problems (not on the breathing rate) [4, 14, 15].

These results demonstrate that the three pollutants (especially NO_2 in the 1^{st} hour) have a strong causal relationship with the variations in the breathing rate of the asthmatic subjects, while corroborating that the meteorological factors have an even higher impact, mainly on the very short term. The detailed results for each subject have been included in Appendix D.2.2.

The strength of the different causal links has also been assessed with the conditional mutual information value of the non-linear test, being stronger when the mutual information is higher. The resulting plots have also been included in Appendix D.2.2, which indicate that the strength of the links for the pollutants and meteorological factors does not depend on the time lag, but rather on the subject (they have a higher

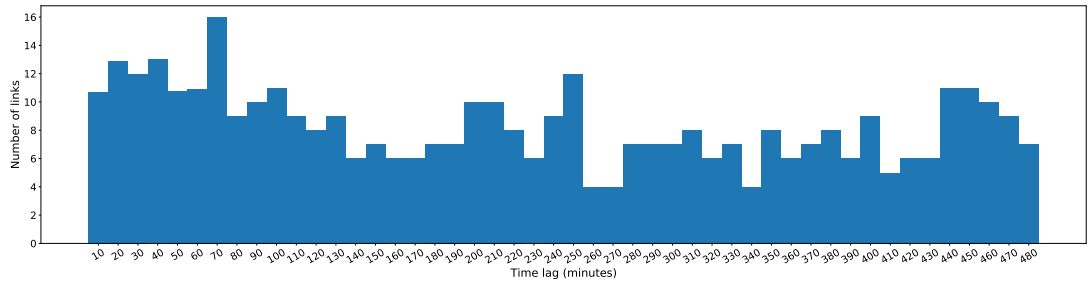

Figure 6.12: Non-linear PCMCI causal link distribution for NO_2.

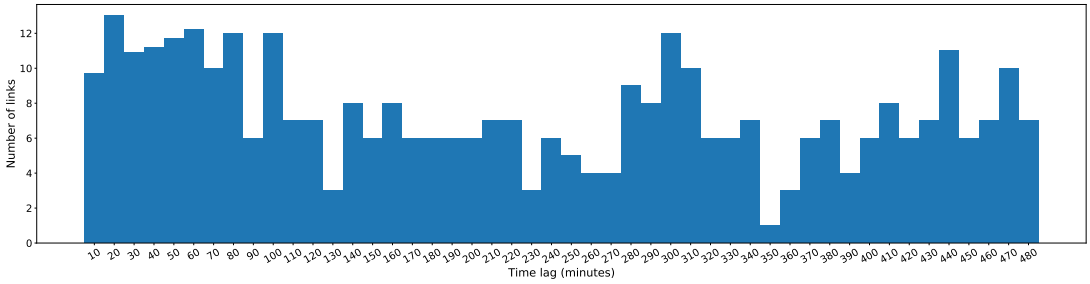

Figure 6.13: Non-linear PCMCI causal link distribution for O_3.

impact on some subjects, which is why a personalised analysis has been made). In contrast, the effect of past breathing rate values on the current one does depend on the time lag for all of them, with a decaying strength as the time lag increases.

6.3 Personal vs static $PM_{2.5}$

The machine learning experiments from Chapter 5 showed an almost identical performance when predicting future breathing rate values for $PM_{2.5}$ data collected from both static and wearable sensors. Therefore, a final experiment was designed with non-linear PCMCI to determine whether the use of static data could replace the one collected by sensors carried by the patients. If the PCMCI results of both of them are analogous, then there would be no need to use sensors to capture personal exposure data, as the use of several static sensors around the city would be sufficient.

The same experiment as in Section 6.2.2 was carried out, but substituting personal exposure $PM_{2.5}$ data with data from the static AIRSpeck sensors. Although the total proportion of causal links is similar to the one obtained with personal exposure $PM_{2.5}$

	Detailed 1st hour	**2nd to 8th hour**	**Total**
Breathing rate	65.45%	16.13%	45.14%
Temperature	33.86%	15.31%	26.22%
Humidity	54.39%	23.54%	41.69%
$PM_{2.5}$	24.92%	18.51%	22.28%
NO_2	26.63%	17.80%	22.99%
O_3	26.02%	15.48%	21.68%

Table 6.1: Proportion of causal links with the breathing rate found with PCMCI for the different lagged variables for the 1st hour, the 2nd to 8th hour period and for both.

data (21.83%), the distribution of the links is not the same. The causal links that coincide for both approaches only represent a 34.61% of the links found by the personal $PM_{2.5}$ data, and a 35.19% of the ones found by the static one. The plot showing the causal links differences has been included in Appendix D.2.2.7.

There can be two main reasons for this. The first one is that the static data has a greater percentage of missing data, mainly because the static sensors report data every 5 minutes and the personal ones do it every minute. Therefore, a much higher percentage of the data needs to be imputed (via interpolation or the mean as explained in Section 4.3), which makes it less reliable. The second one is that the static sensors may not gather what the subjects are really exposed to, due to being on average at 182.26 metres from them for the $PM_{2.5}$ data. The percentage of missed causal links (detected by personal but not static data) and false positives (only detected by static data) has been plotted for the trials that were close enough to the sensors ($<$ 30 metres on average) on Figures 6.14 and 6.15. There is a trend of increases on both percentages as the average distance from the sensors goes up, confirming this second point.

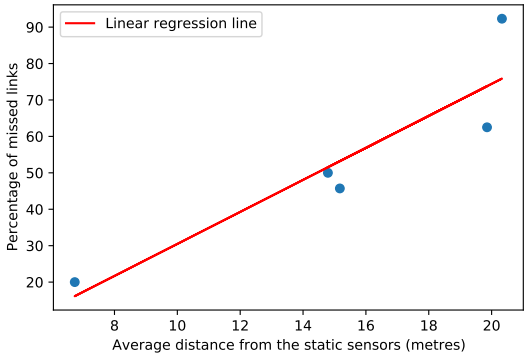

Figure 6.14: Missed links static $PM_{2.5}$

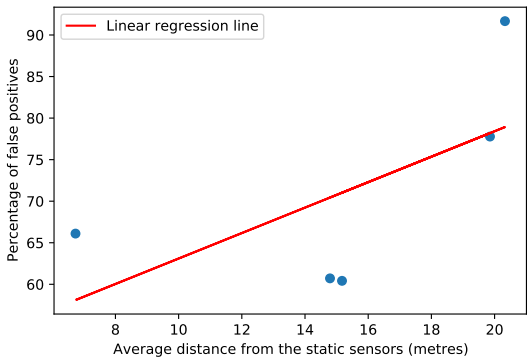

Figure 6.15: FP static $PM_{2.5}$

Chapter 7

Exposure-response relationship

PCMCI can determine whether there is a causal link between two time series with a certain time lag, but it does not yield the type of relationship between them. Therefore, a method is required to determine how the breathing rate behaves, in terms of increases or decreases, when the different pollutant concentrations increase at causal links between them and the respiratory rate.

The first attempt was to do it with the machine learning methods of Chapter 5. Since the PCMCI conditional independence test used is non-linear and uses a k nearest neighbour approach, the KNN machine learning model was the best suited method for the task. The initial experiment was done for $PM_{2.5}$ and each lag of the first hour, fitting the model with the same variables used in PCMCI and tuning the hyper-parameters. The breathing rate value was predicted after applying an inter-quartile range increase to the $PM_{2.5}$ concentration, and it was compared to the respiratory rate without the $PM_{2.5}$ increase. The result was that only a 5.13% of the causal links between $PM_{2.5}$ and the respiratory rate resulted in breathing rate increases when the pollutant concentration increased, as shown in the plot of Appendix E.1. Nevertheless, this result is not conclusive, since the error of the machine learning models shown in Chapter 5 is too high when compared to the scale of the data, which could easily affect the task of determining whether the breathing rate would go up.

A second (and much more reliable) experiment was developed. It is explained in Algorithm 4, where for each pollutant an increase is determined when there is a concentration value higher than the mean of the past values up to a maximum time window by more than a certain percentage threshold.

The percentage of the time lags where causal links were found, for which a pollutant concentration increase corresponds to a breathing rate increase (worsening the

patient's condition) for the three pollutants have been plotted in Figures 7.1, 7.2 and 7.3, where each line represents a different percentage threshold. The threshold does not seem to make a big difference in the percentage, as almost all of them follow the same trend. In contrast, for the three pollutants, the percentage increases as the considered window size goes up, always reaching more than a 50% for all the pollutants with window sizes of more than 200 minutes. These results suggest that in order to make the respiratory rate of the asthmatic subjects go up, the pollutant concentration needs to augment to higher levels than what the patient has been exposed to in the recent hours. The full results, detailed for each time lag and trial have been plotted in Appendix E.2.

Algorithm 4 Breathing rate variation with a pollutant concentration increase.

Input: Trial and pollutant to consider, time lags with causal links: *causal_time_lags*.
percentage_thresholds = All percentages from 5% to 300% in steps of 5.
time_windows = All window sizes from 10 min. to 4 hours in 10 min. steps.
for all *percentage_thresholds* and *time_windows* **do**
  for all *causal_time_lags* **do**
    for all pollutant values in the pollutant time series of the trial **do**
      if the pollutant value is higher that the average of the past time window values by more than the evaluated percentage threshold **then**
        Add to a counter the percentage of variation between the mean breathing rate in the time window and the one after the evaluated time lag.
      end if
    end for
    Compute the average variation dividing the counter by the number of added percentages and set it for the evaluated time lag, window and threshold.
  end for
end for

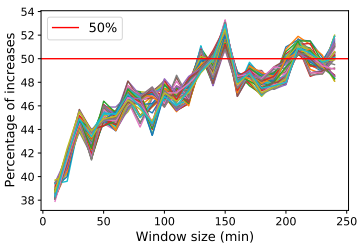

Figure 7.1: $PM_{2.5}$ increases

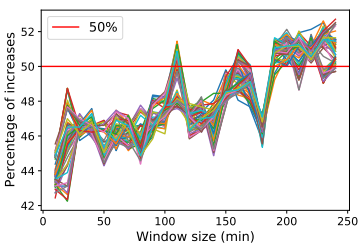

Figure 7.2: NO_2 increases

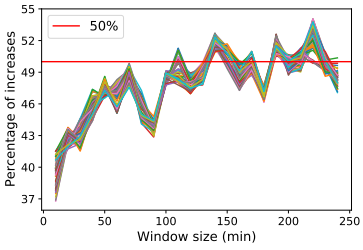

Figure 7.3: O_3 increases

Chapter 8

Conclusions

8.1 Discussion

For the first time, a causal relationship has been demonstrated between exposure to nitrogen dioxide, ozone and $PM_{2.5}$, and the breathing rate of asthmatic adolescents. The dataset gathered by the DAPHNE project has been used for this purpose, which includes the respiratory rate data of 127 asthmatic adolescents and the pollutant concentration data they were exposed to in Delhi, India. Wearable sensors were used for $PM_{2.5}$ and static sensors for NO_2 and O_3. The data has been pre-processed for the analysis as described in Chapter 4, including the calibration of NO_2 and O_3 data so that the values are comparable across sensors.

Firstly, the association between the pollutants and the breathing rate has been evidenced using a variety of linear and non-linear machine learning models to predict the current breathing rate of each subject based on past respiratory rate and pollutant concentration values. It was demonstrated with all the models, that the three pollutants encode information about the future breathing rate behaviour, even having a better prediction performance than past breathing rate values.

A recently-published causal discovery method (PCMCI) has also been used to assess the causality between them at different time intervals in the short term, for time lags ranging from 1 minute up to 20 hours for linear relationships, and up to 8 hours for non-linear ones. The previous work with this method in [17] assessing the causality between the respiratory rate and the past breathing rate, temperature, humidity and the $PM_{2.5}$ pollutant data for a limited set of time lags up to 60 minutes, has been extended in this work and corrected so that the non-linear PCMCI cases can be reproduced. The results showed a strong causal relationship between $PM_{2.5}$ and the breathing rate.

In order to satisfy the PCMCI causal sufficiency assumption, the experiments were repeated including the NO_2 and O_3 data in the analysis. A very strong relationship was found for the three of them, having a causal association in more than 20% of the evaluated time lags between the exposure to the pollutant and the breathing rate response of the subject, taking into account that the method has a strong false positive control. Therefore, the fact that the three pollutants directly cause changes in the breathing rate of the asthmatic adolescents subjects after a period of time from the exposure has been demonstrated. The pollutant with the highest proportion of causal links was NO_2 with a 22.99% of the total tested lags being causal, which is higher in the first hour with a 26.63%. It is also the pollutant with the stronger association found with other indicators of respiratory problems in previous work [4, 14, 15]. The other gas, O_3, has a very similar causal link distribution, being higher in the first hour and decaying afterwards, although with a lower proportion of links. $PM_{2.5}$ was the second pollutant with the highest percentage of causal links, having a lower proportion in the first hour than the two gases but maintaining the links in the subsequent hours, even having a peak between the 7^{th} and 8^{th} hours.

This analysis also corroborated that the humidity and temperature have a stronger causal association in the short term than the pollutants, as it was discovered in [14] for children clinical visits related to respiratory problems. It is much more pronounced in the first hour, decaying as the time lag between the exposure and the response goes up.

Furthermore, the importance of using personal exposure data for $PM_{2.5}$ instead of data from static sensors at a distance from the subject (as it is common in previous work like [36]) has been demonstrated. The machine learning experiments showed a very similar performance for the $PM_{2.5}$ data gathered from the personal sensors and the one from the static sensors when predicting the breathing rate. Therefore, the last PCMCI experiment was repeated but using the static sensor data for $PM_{2.5}$. The results were quite different, discovering with the static data only 34.61% of the links found with the wearable sensors, with a 64.81% of the causal links found by the static $PM_{2.5}$ not found by the personal exposure data (which could be seen as false positives). An important factor for this is that the static sensor has a larger amount of missing data, but also the distance of the subjects from the sensors was demonstrated to influence the amount of missed links and false positives for trials that were close on average to the static sensors (< 30 metres), showing the advantage of using wearable sensors.

The exposure-response relationship between the pollutants and the breathing rate has also been studied, using a sliding window strategy explained in Algorithm 4 to

determine whether the breathing rate goes up (which is an asthma symptom) when the different pollutant concentrations increase. The experiments concluded that for the majority of the time lags with causal links, the respiratory rate goes up when the pollutants concentrations increase to higher levels than the average of at least the past 200 minutes. From that time window on, more than 50% of the causal links resulted in breathing rate increases for the three pollutants. The use of the machine learning methods was considered not suitable for the task, due to the high prediction error they produced compared to the scale of the data.

Therefore, the principal objective of determining the causality between the pollutants and the breathing rate of the asthmatic subjects, as well as the secondary one of estimating the exposure-response relationship with the increases of the breathing rate, have been accomplished.

8.2 Future work

All of the mentioned experiments have been done personalised to each subject, so the obtained results will allow to estimate accurately the effect that the different pollution levels will have on each subject's respiratory condition. To do so, future work should focus on providing an equation form for each of the subjects that determines their breathing rate based on the lagged observations of pollutants, breathing rate and the meteorological factors for which a causal relationship has been found in this research.

Deep learning techniques which have been developed specifically to deal with time series could be used for this task. A long short-term memory, which is a recurrent neural network introduced in [13] that can take into account long term time lags, could be used to determine the coefficients of each of the variables of the equation. These variables need to be determined, most likely performing data transformations with the original lagged time series for which a causal link have been found in this project.

This is not a trivial task, since a vast amount of data is involved. The dataset for a trial with 150 causal links across the 6 variables included in PCMCI (the mean amount of causal links is approximately 183) and using a polynomial transformation of degree 3 for determining the equation variables, would consist of 585276 time series, more than 1685 million data points if the trial has 48 hours of data. This qualifies for a new research project, since different dimensionality reduction techniques should be tested, as well as different data transformations and machine/deep learning models to provide the most accurate analytical expression for the breathing rate of each subject.

Bibliography

[1] Daphne. https://gtr.ukri.org/projects?ref=NE%2FP016340%2F1. Online; accessed 22 July, 2020.

[2] DK Arvind, Janek Mann, Andrew Bates, and Konstantin Kotsev. The airspeck family of static and mobile wireless air quality monitors. In *2016 Euromicro Conference on Digital System Design (DSD)*, pages 207–214. IEEE, 2016.

[3] Andrew Bates, Martin J Ling, Janek Mann, and Damal K Arvind. Respiratory rate and flow waveform estimation from tri-axial accelerometer data. In *2010 International Conference on Body Sensor Networks*, pages 144–150. IEEE, 2010.

[4] Rob Beelen, Gerard Hoek, Piet A van Den Brandt, R Alexandra Goldbohm, Paul Fischer, Leo J Schouten, Michael Jerrett, Edward Hughes, Ben Armstrong, and Bert Brunekreef. Long-term effects of traffic-related air pollution on mortality in a dutch cohort (nlcs-air study). *Environmental health perspectives*, 116(2):196–202, 2008.

[5] Chris Chatfield. *Time-series forecasting*. CRC press, 2000.

[6] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

[7] Gennaro D'Amato. Effects of climatic changes and urban air pollution on the rising trends of respiratory allergy and asthma. *Multidisciplinary respiratory medicine*, 6(1):28, 2011.

[8] Salvador García, Julián Luengo, and Francisco Herrera. *Data preprocessing in data mining*. Springer, 2015.

[9] N Good, T Carpenter, GB Anderson, Ander Wilson, JL Peel, RC Browning, and J Volckens. Development and validation of models to predict personal ventila-

tion rate for air pollution research. *Journal of exposure science & environmental epidemiology*, 29(4):568, 2019.

[10] Michael Guarnieri and John R Balmes. Outdoor air pollution and asthma. *The Lancet*, 383(9928):1581–1592, 2014.

[11] Michael Gutmann and Arno Onken. Data mining and exploration lecture notes, University of Edinburgh, May 2020.

[12] T Hirsch, SK Weiland, E Von Mutius, AF Safeca, H Grafe, E Csaplovics, H Duhme, U Keil, and W Leupold. Inner city air pollution and respiratory health and atopy in children. *European respiratory journal*, 14(3):669–677, 1999.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] Yabin Hu, Zhiwei Xu, Fan Jiang, Shenghui Li, Shijian Liu, Meiqin Wu, Chonghuai Yan, Jianguo Tan, Guangjun Yu, Yi Hu, et al. Relative impact of meteorological factors and air pollutants on childhood allergic diseases in shanghai, china. *Science of The Total Environment*, 706:135975, 2020.

[15] Małgorzata Kowalska, Michał Skrzypek, Michał Kowalski, and Josef Cyrys. Effect of nox and no2 concentration increase in ambient air to daily bronchitis and asthma exacerbation, silesian voivodeship in poland. *International Journal of Environmental Research and Public Health*, 17(3):754, 2020.

[16] Ozlem Kar Kurt, Jingjing Zhang, and Kent E Pinkerton. Pulmonary health effects of air pollution. *Current opinion in pulmonary medicine*, 22(2):138, 2016.

[17] Sharan Maiya. Investigating the respiratory rate response to PM2.5 exposure in asthmatic adolescents. *Honours year dissertation, School of Informatics, University of Edinburgh*, June 2020.

[18] Rob McConnell, Kiros Berhane, Frank Gilliland, Stephanie J London, Hita Vora, Edward Avol, W James Gauderman, Helene G Margolis, Fred Lurmann, Duncan C Thomas, et al. Air pollution and bronchitic symptoms in southern california children with asthma. *Environmental health perspectives*, 107(9):757–760, 1999.

[19] Anna Mölter, Angela Simpson, Dietrich Berdel, Bert Brunekreef, Adnan Custovic, Josef Cyrys, Johan de Jongste, Frank De Vocht, Elaine Fuertes, Ulrike Gehring, et al. A multicentre study of air pollution exposure and childhood asthma prevalence: the escape project. *European Respiratory Journal*, 45(3):610–624, 2015.

[20] World Health Organization. Air pollution levels rising in many of the world's poorest cities. Online; accessed 22 July, 2020.

[21] World Health Organization. Ambient (outdoor) air pollution. Online; accessed 22 July, 2020.

[22] World Health Organization. Children and air pollution. Online; accessed 22 July, 2020.

[23] World Health Organization. Household air pollution and health. Online; accessed 22 July, 2020.

[24] World Health Organization et al. Who air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005: summary of risk assessment. Technical report, World Health Organization, 2006.

[25] World Health Organization et al. Air pollution and child health: prescribing clean air: summary. Technical report, World Health Organization, 2018.

[26] Alberto Papi, Giorgio W Canonica, Piero Maestrelli, Pierluigi Paggiaro, Dario Olivieri, Ernesto Pozzi, Nunzio Crimi, Antonio M Vignola, Paolo Morelli, Gabriele Nicolini, et al. Rescue use of beclomethasone and albuterol in a single inhaler for mild asthma. *New England Journal of Medicine*, 356(20):2040–2052, 2007.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[28] Annette Prüss-Üstün, Jennyfer Wolf, Carlos Corvalán, Robert Bos, and Maria Neira. *Preventing disease through healthy environments: a global assessment of the burden of disease from environmental risks*. World Health Organization, 2016.

[29] Jakob Runge. Tigramite. Online; accessed 4 August, 2020.

[30] Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, 2018.

[31] Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947, 2018.

[32] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019.

[33] Sundeep Salvi. Health effects of ambient air pollution in children. *Paediatric respiratory reviews*, 8(4):275–280, 2007.

[34] Edmund Seto, Graeme Carvlin, Elena Austin, Jeffry Shirai, Esther Bejarano, Humberto Lugo, Luis Olmedo, Astrid Calderas, Michael Jerrett, Galatea King, et al. Next-generation community air quality sensors for identifying air pollution episodes. *International journal of environmental research and public health*, 16(18):3268, 2019.

[35] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

[36] Austin M Williams, Daniel J Phaneuf, Meredith A Barrett, and Jason G Su. Short-term impact of pm2. 5 on contemporaneous asthma medication use: Behavior and the value of pollution reductions. *Proceedings of the National Academy of Sciences*, 116(12):5246–5253, 2019.

[37] Catherine Wyler, Charlotte Braun-Fahrländer, Nino Künzli, Christian Schindler, Ursula Ackermann-Liebrich, André P Perruchoud, Philippe Leuenberger, Brunello Wüthrich, The Swiss Study on Air Pollution, and Lung Diseases in Adults (SAPALDIA) Team. Exposure to motor vehicle traffic and allergic sensitization. *Epidemiology*, pages 450–456, 2000.

[38] Qi Zhou, Shaonan Li, Xiaopeng Li, Wei Wang, and Zhiguo Wang. Detection of outliers and establishment of targets in external quality assessment programs. *Clinica chimica acta*, 372(1-2):94–97, 2006.

Appendix A

Exploratory data analysis

A.1 RESpeck data

This section includes the numeric results of the descriptive statistics taken for the RESpeck device data.

Mean	Q1	Median	Q3	Minimum	Maximum	Mode
21.380	18.126	20.993	24.053	8.049	44.175	22.088

Table A.1: Respiratory rate location measures.

Standard deviation	Variance	MAD	IQR	Range
4.404	19.401	2.959	5.927	36.126

Table A.2: Respiratory rate scale measures.

Skewness	Galton's measure of skewness	Kurtosis	Robust kurtosis
0.634	0.033	0.544	1.216

Table A.3: Respiratory rate shape measures.

Measure \ Time lag	0 min	5 min	15 min	30 min	60 min
Pearson's correlation coefficient	0.335	0.245	0.224	0.208	0.197
Kendall's tau	0.298	0.284	0.271	0.258	0.238

Table A.4: Respiratory rate correlation measures.

A.2 AIRSpeck data

This section includes the numeric results of the descriptive statistics taken for the AIR-Speck devices data.

	Mean	Q1	Median	Q3	Minimum	Maximum	Mode
$PM_{2.5}$	81.261	18.231	41.044	97.73	0	19081.356	0
NO_2	4981.935	4632	4800	4944	0	32767	32767
O_3	8750.063	6677	6859.5	6969	0	65535	65535
Temp.	30.739	26.1	32.9	37.05	-46.800	128.8	0
Hum.	47.971	41.25	48.05	55.3	0	101.3	0

Table A.5: AIRSpeck data location measures.

	Standard deviation	Variance	MAD	IQR	Range
$PM_{2.5}$	178.470	31851.649	28.519	79.499	19081.356
NO_2	3734.743	1.395e+07	152	312	32767
O_3	11086.666	1.229e+08	134.833	292	65535
Temp.	10.700	114.488	5.200	10.950	175.600
Hum.	11.714	137.207	7	14.050	101.300

Table A.6: AIRSpeck data scale measures.

	Skewness	Galton's skewness	Kurtosis	Robust kurtosis
$PM_{2.5}$	37.771	0.426	2619.048	1.635
NO_2	6.142	-0.077	42.036	1.753
O_3	4.607	-0.250	20.254	1.683
Temp.	0.188	-0.242	15.171	0.918
Hum.	-0.605	0.032	3.087	1.219

Table A.7: AIRSpeck data shape measures.

Feature \ Time lag	0 min	5 min	15 min	30 min	60 min
$PM_{2.5}$	0.0063	0.01	0.0115	0.0106	0.008
NO_2	0.0003	0.0013	0.0018	0.0043	-0.0022
O_3	0.0203	0.0205	0.0209	0.0217	0.0154
Temp.	0.0689	0.0674	0.0653	0.0618	0.056
Hum.	-0.0056	-0.006	-0.0042	-0.0007	0.0039

Table A.8: AIRSpeck data Kendall's tau with breathing rate.

A.3 School and community sensors comparison

This sections shows some examples of the comparison made between the community and school sensor data (both outdoor sensors) to check whether they are similar.

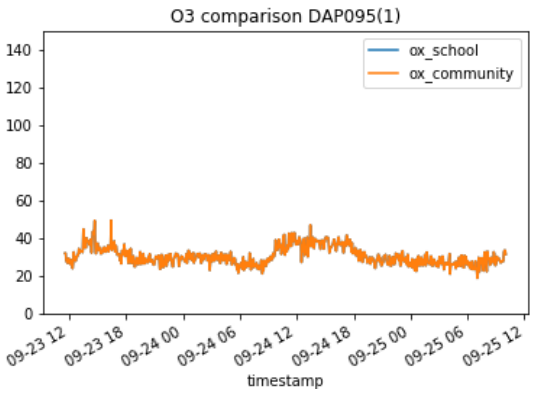

Figure A.1: Community sensor used as school sensor for O_3.

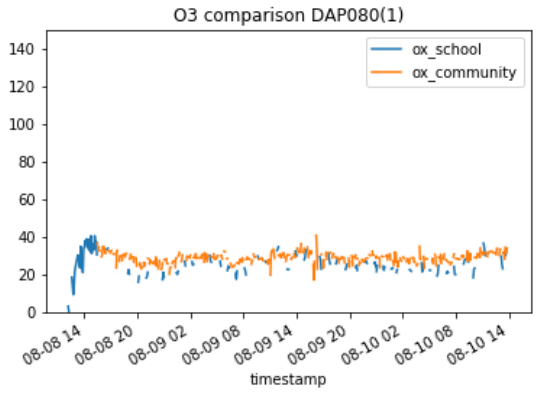

Figure A.2: First community and school sensor comparison for O_3.

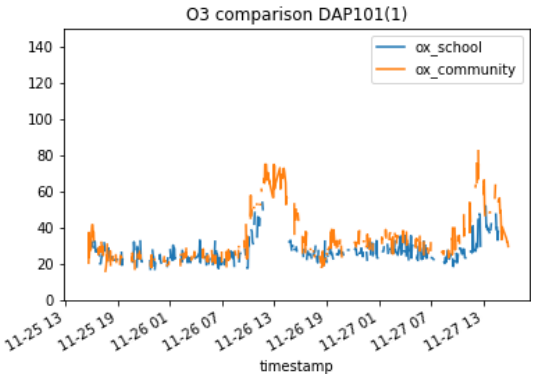

Figure A.3: Second community and school sensor comparison for O_3.

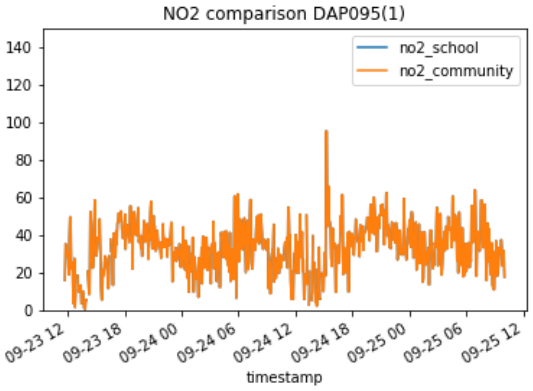

Figure A.4: Community sensor used as school sensor for NO_2.

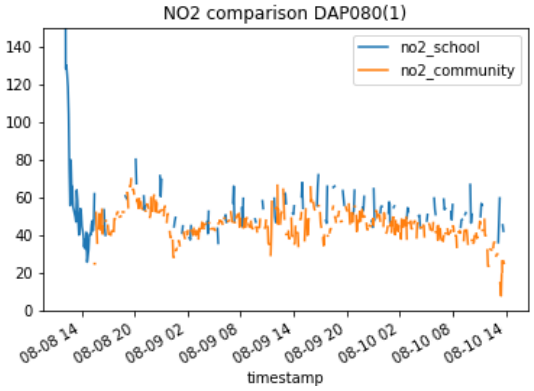

Figure A.5: First community and school sensor comparison for NO_2.

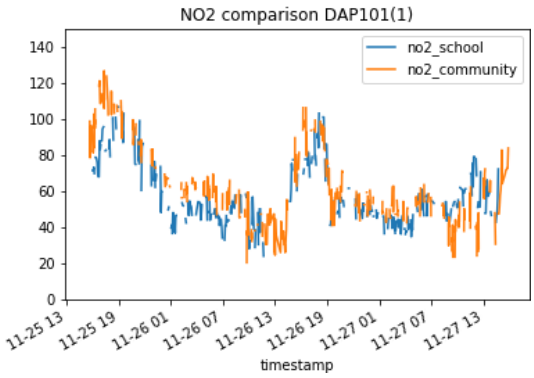

Figure A.6: Second community and school sensor comparison for NO_2.

Appendix B

Data pre-processing

B.1 October-November 2019 calibration

This section contains the graphs corresponding to the calibration of the October-November 2019 period of time, when the sensors were 1 km away from the reference one.

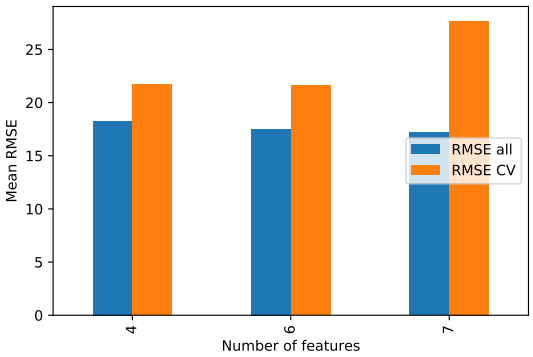

Figure B.1: Mean RMSE of fit-to-all and CV for O_3.

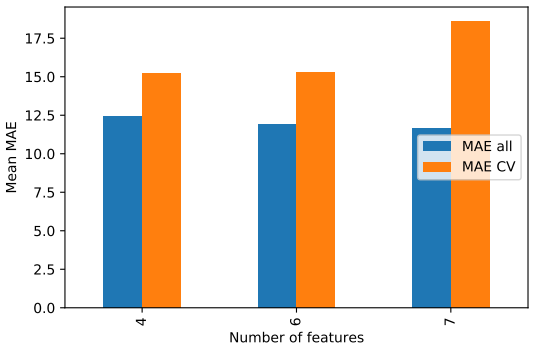

Figure B.2: Mean MAE of fit-to-all and CV for O_3.

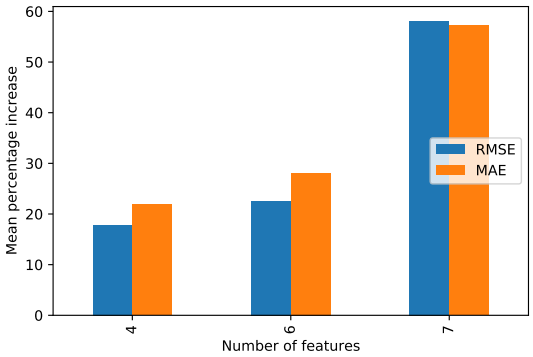

Figure B.3: Mean percentage error increase from fit-to-all to CV for O_3.

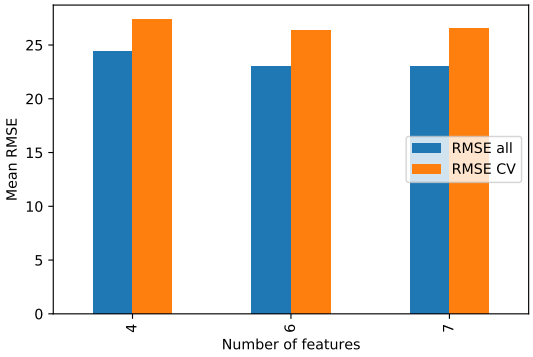

Figure B.4: Mean RMSE of fit-to-all and CV for NO_2.

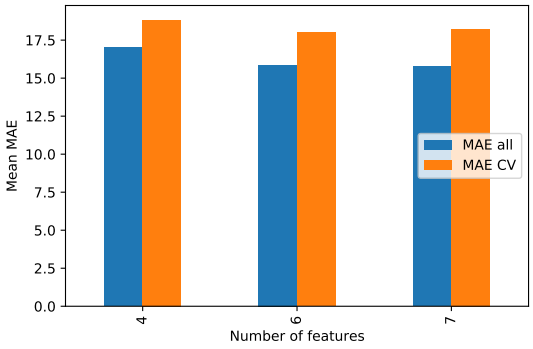

Figure B.5: Mean MAE of fit-to-all and CV for NO_2.

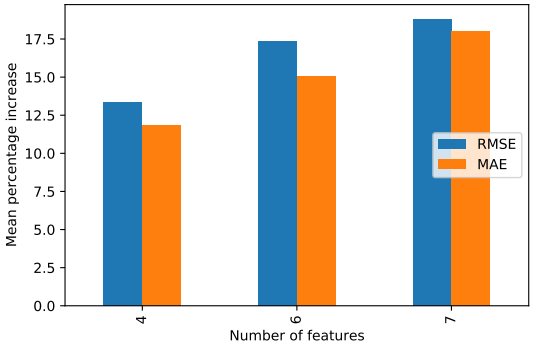

Figure B.6: Mean percentage error increase from fit-to-all to CV for NO_2.

B.2 January-April 2020 calibration

This section contains the graphs corresponding to the calibration of the January-April 2020 period of time, when the sensors were right next to the reference one.

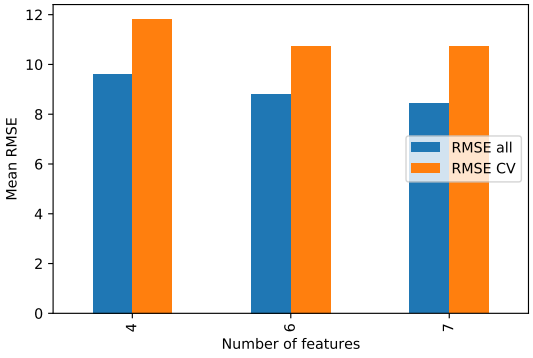

Figure B.7: Mean RMSE of fit-to-all and CV for O_3.

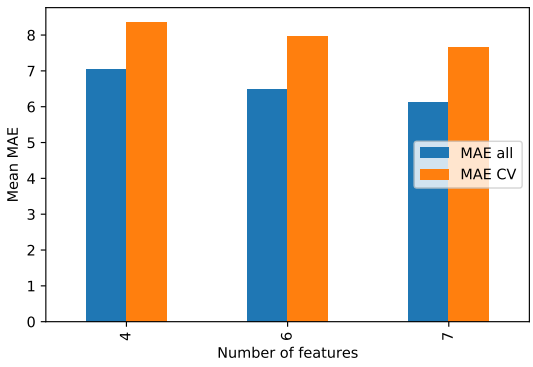

Figure B.8: Mean MAE of fit-to-all and CV for O_3.

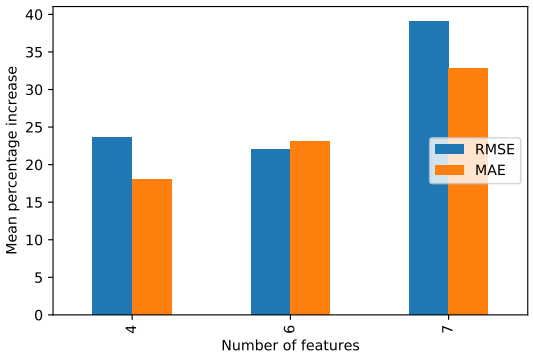

Figure B.9: Mean percentage error increase from fit-to-all to CV for O_3.

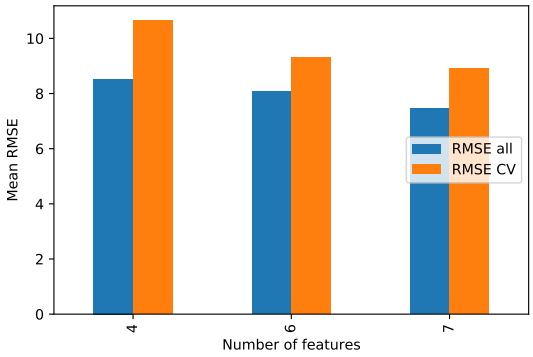

Figure B.10: Mean RMSE of fit-to-all and CV for NO_2.

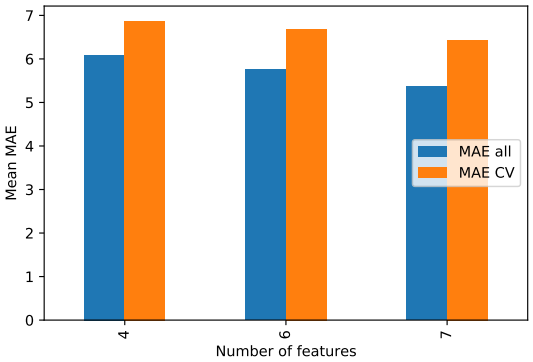

Figure B.11: Mean MAE of fit-to-all and CV for NO_2.

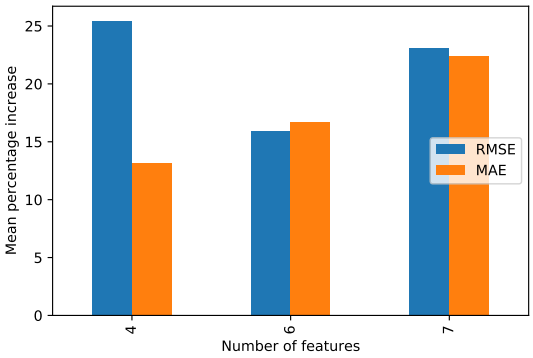

Figure B.12: Mean percentage error increase from fit-to-all to CV for NO_2.

B.3 2019-2020 calibration comparison

This section contains the comparison of the calibration between the October-November 2019 and January-April 2020 for the different sets of features.

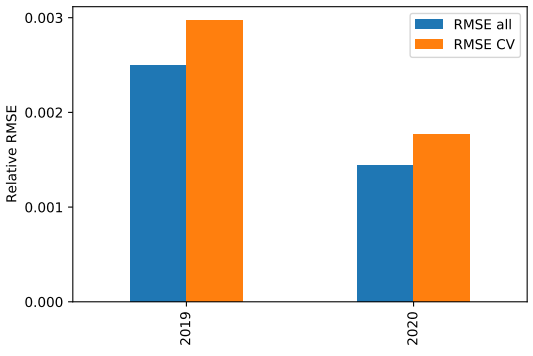

Figure B.13: Relative RMSE with 4 features for O_3.

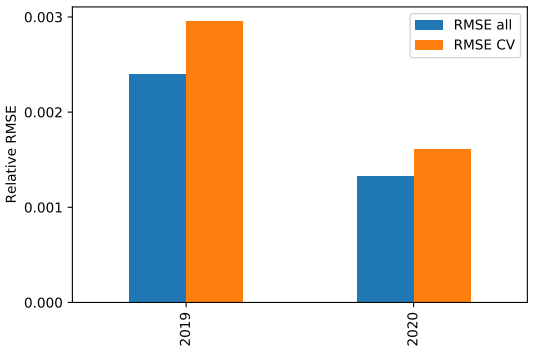

Figure B.14: Relative RMSE with 6 features for O_3.

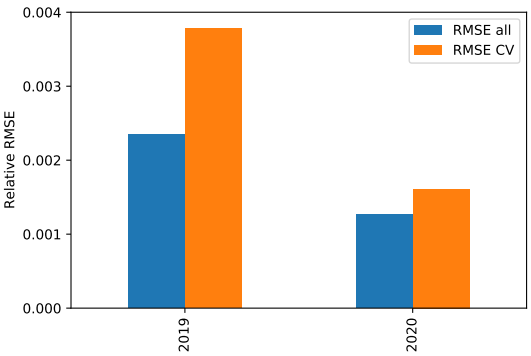

Figure B.15: Relative RMSE with 7 features for O_3.

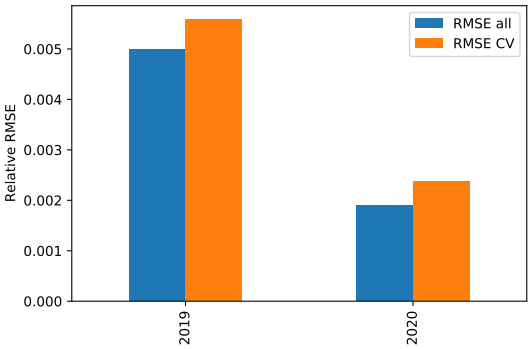

Figure B.16: Relative RMSE with 4 features for NO_2.

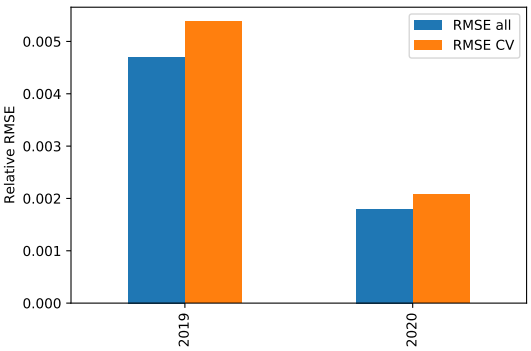

Figure B.17: Relative RMSE with 6 features for NO_2.

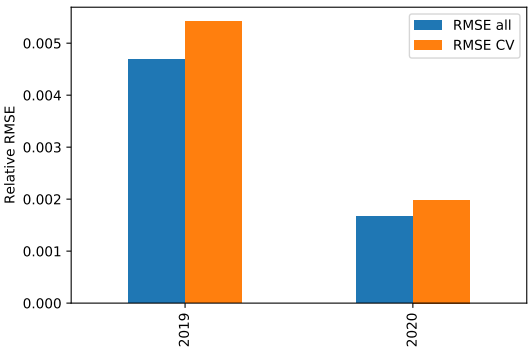

Figure B.18: Relative RMSE with 7 features for NO_2.

B.4 Sensor error

This section contains the plot of the sensor that resulted on an RMSE increase in Figures 4.3 and 4.4, whose data was not included in the calibration because of the error produced at a time point.

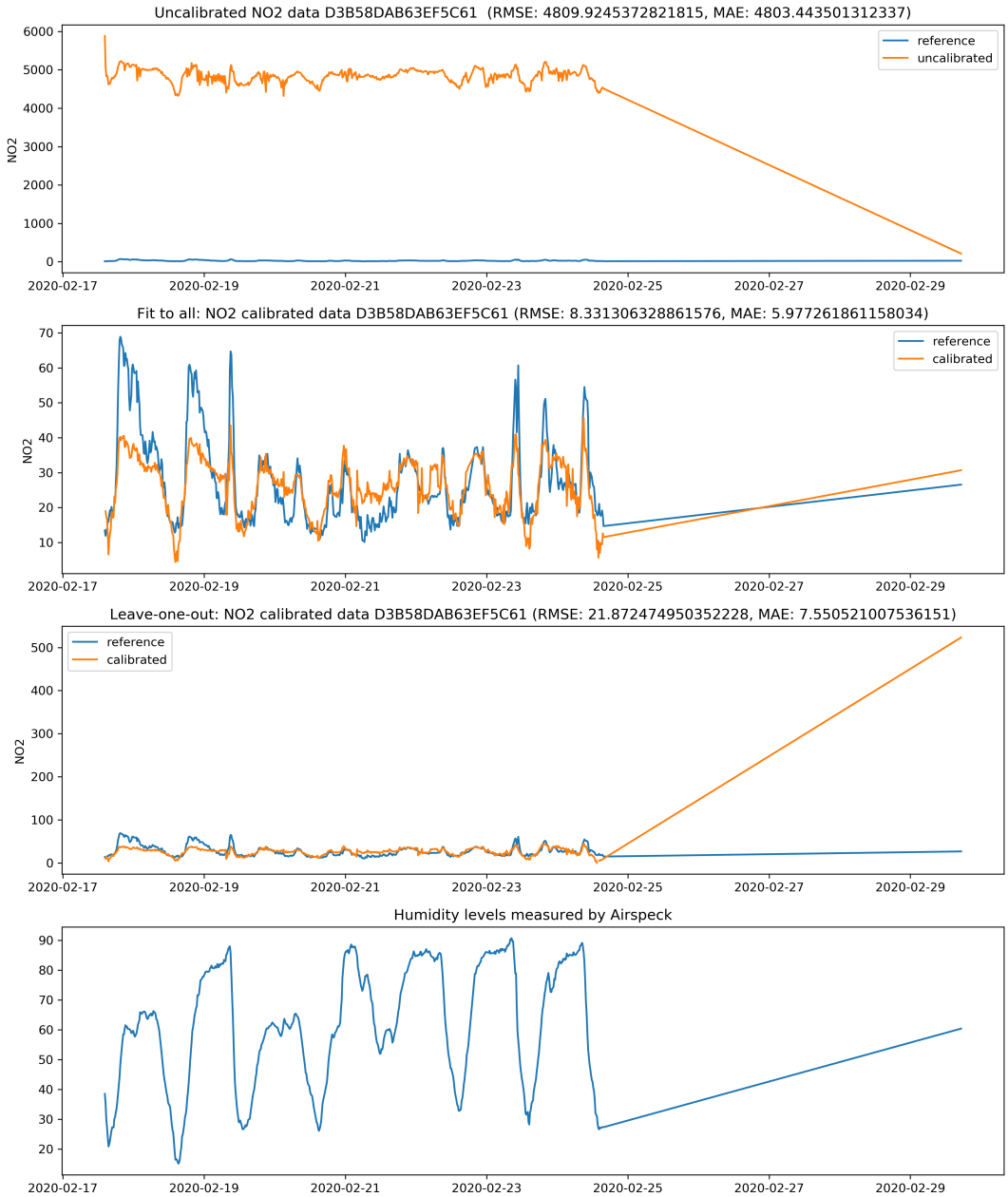

Figure B.19: Sensor NO_2 data before and after calibrating it, which failed at a time point.

Appendix C

Machine learning experiments

C.1 Personal $PM_{2.5}$ results

This section contains the machine learning results for all the trials for which personal $PM_{2.5}$ data is available.

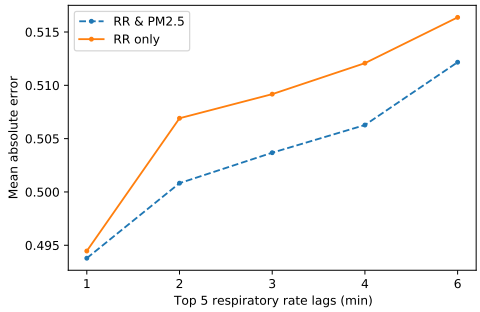

Figure C.1: Ridge regression average error comparison personal $PM_{2.5}$.

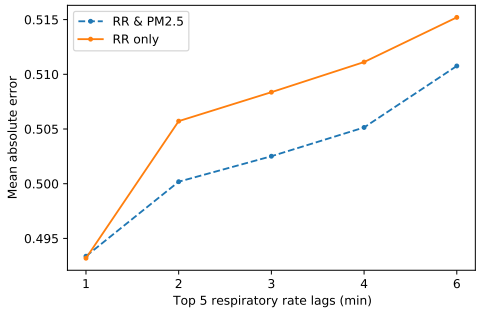

Figure C.2: Huber regression average error comparison personal $PM_{2.5}$.

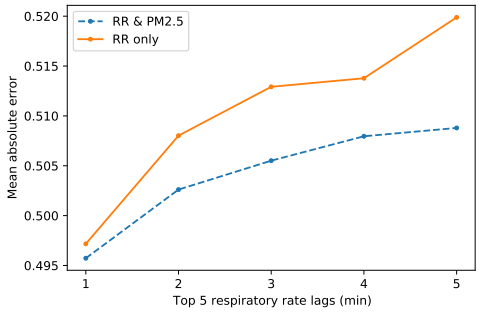

Figure C.3: Random forest average error comparison personal $PM_{2.5}$.

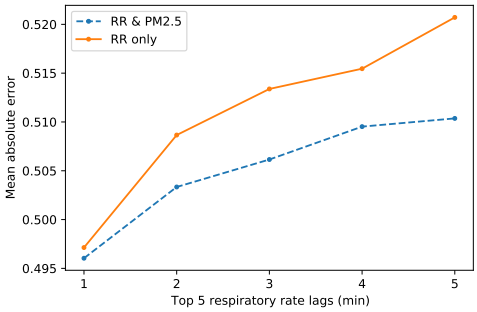

Figure C.4: KNN regression average error comparison personal $PM_{2.5}$.

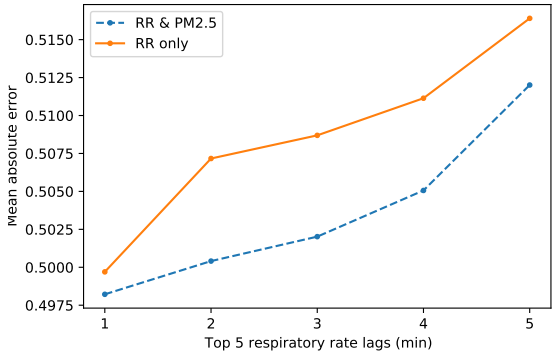

Figure C.5: SVR regression average error comparison personal $PM_{2.5}$.

C.2 Personal $PM_{2.5}$ results

This section contains the machine learning results for all the trials for which personal $PM_{2.5}$, NO_2 and O_3 data is available.

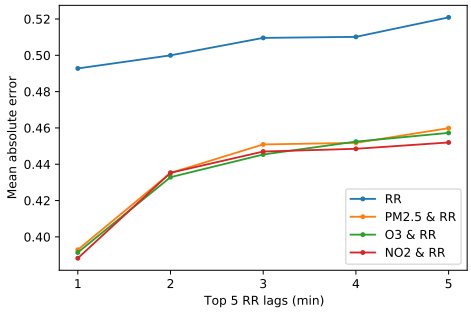

Figure C.6: Ridge regression average error comparison all pollutants.

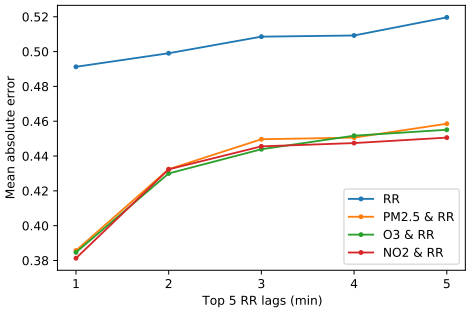

Figure C.7: Huber regression average error comparison all pollutants.

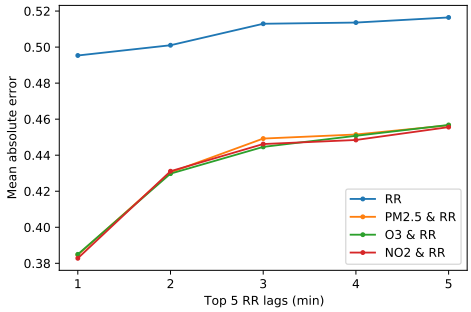

Figure C.8: Random forest average error comparison all pollutants.

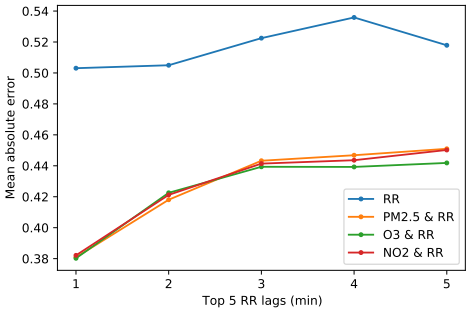

Figure C.9: KNN regression average error comparison all pollutants.

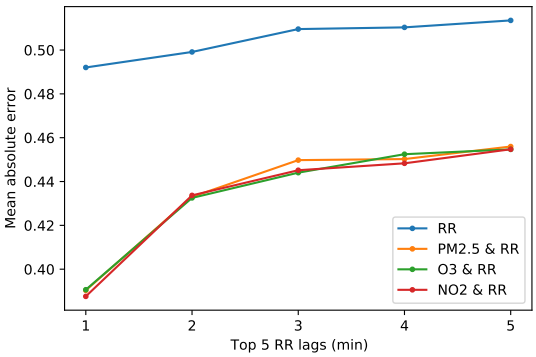

Figure C.10: SVR regression average error comparison all pollutants.

Appendix D

PCMCI experiments

D.1 Only for $PM_{2.5}$

D.1.1 Linear PCMCI

This section shows the link distribution for all the trials tested in linear PCMCI for personal $PM_{2.5}$ exposure data.

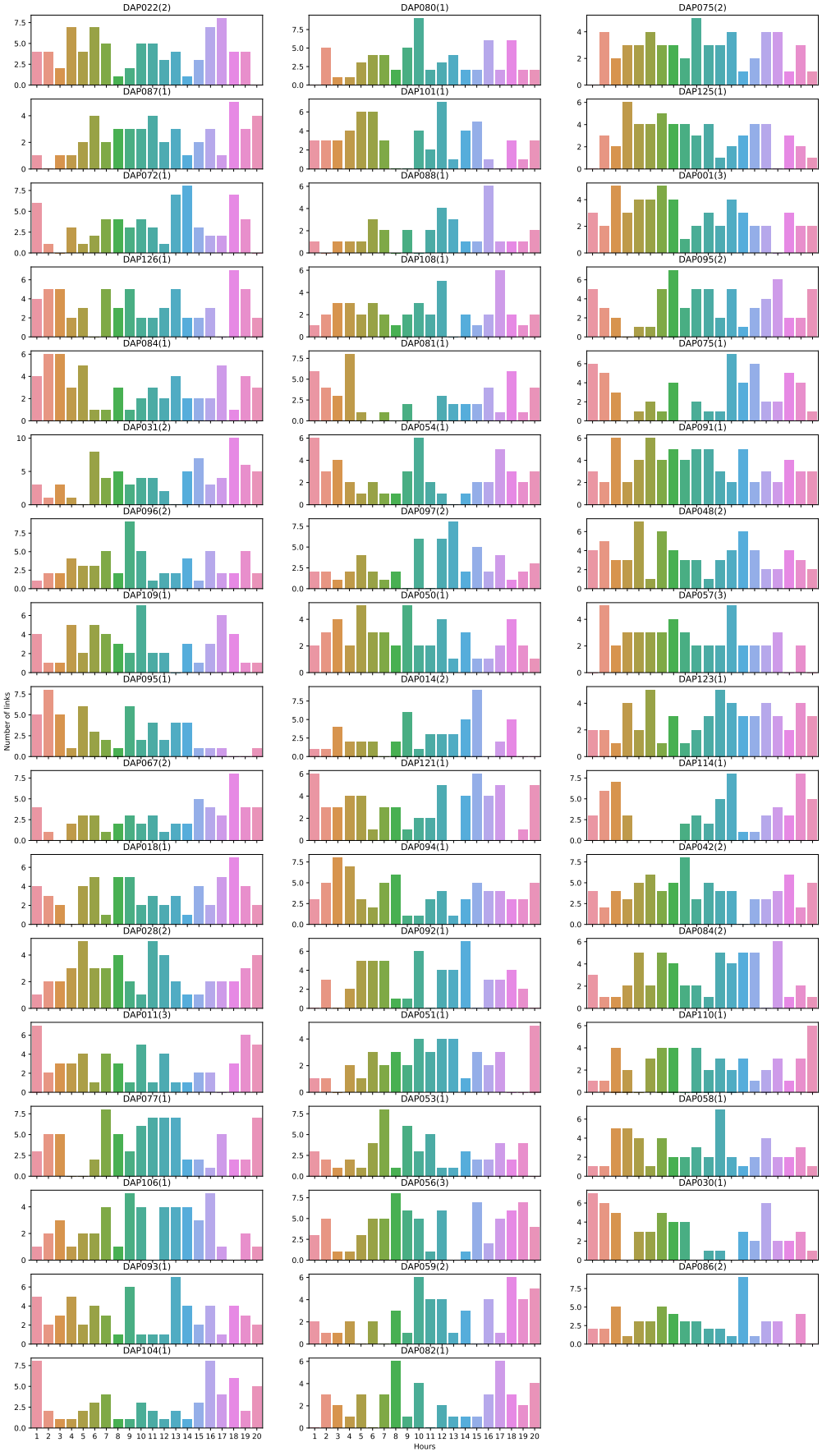

Figure D.1: Distribution of causal links for all the trials for linear PCMCI with personal $PM_{2.5}$ exposure data.

D.1.2 Non-linear PCMCI

This section contains the link distribution information for the trials tested in non-linear PCMCI for personal $PM_{2.5}$ exposure data that was not included in the main document, for both the 8 hours and the detailed first hour. It includes the color maps for both, which represent the p-values that PCMCI returns when testing causality with the breathing rate for each trial and each time lag. When the p-value is $<= 0.05$, then the result is statistically significant to determine that there is a causal link, which is plotted in green. Those combinations of trials and time lags who were close to be statistically significant ($0.05 <$ p-value $<= 0.1$) are plotted in yellow, and the rest in red.

D.1.2.1 8 hours

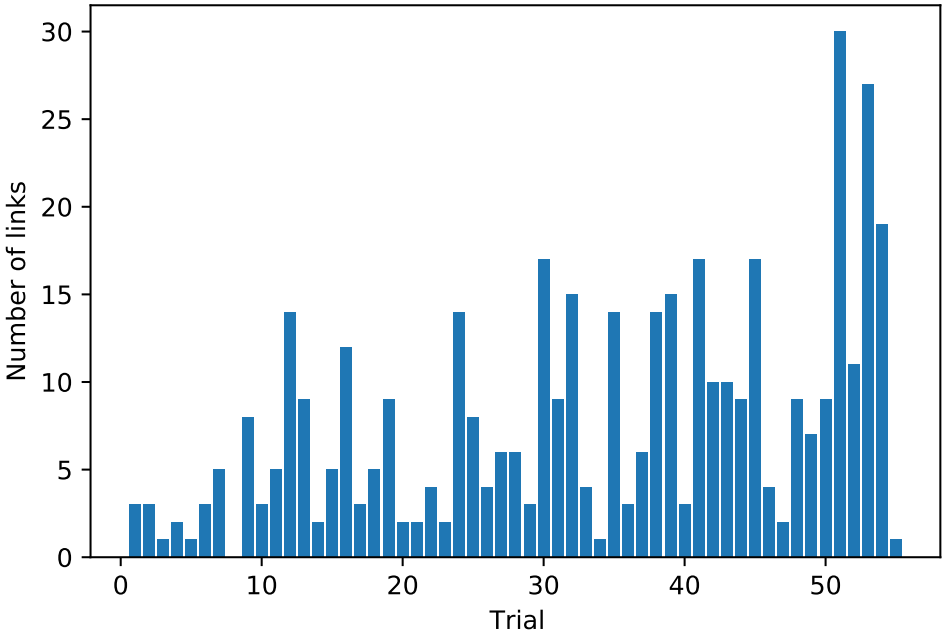

Figure D.2: Number of links per trial.

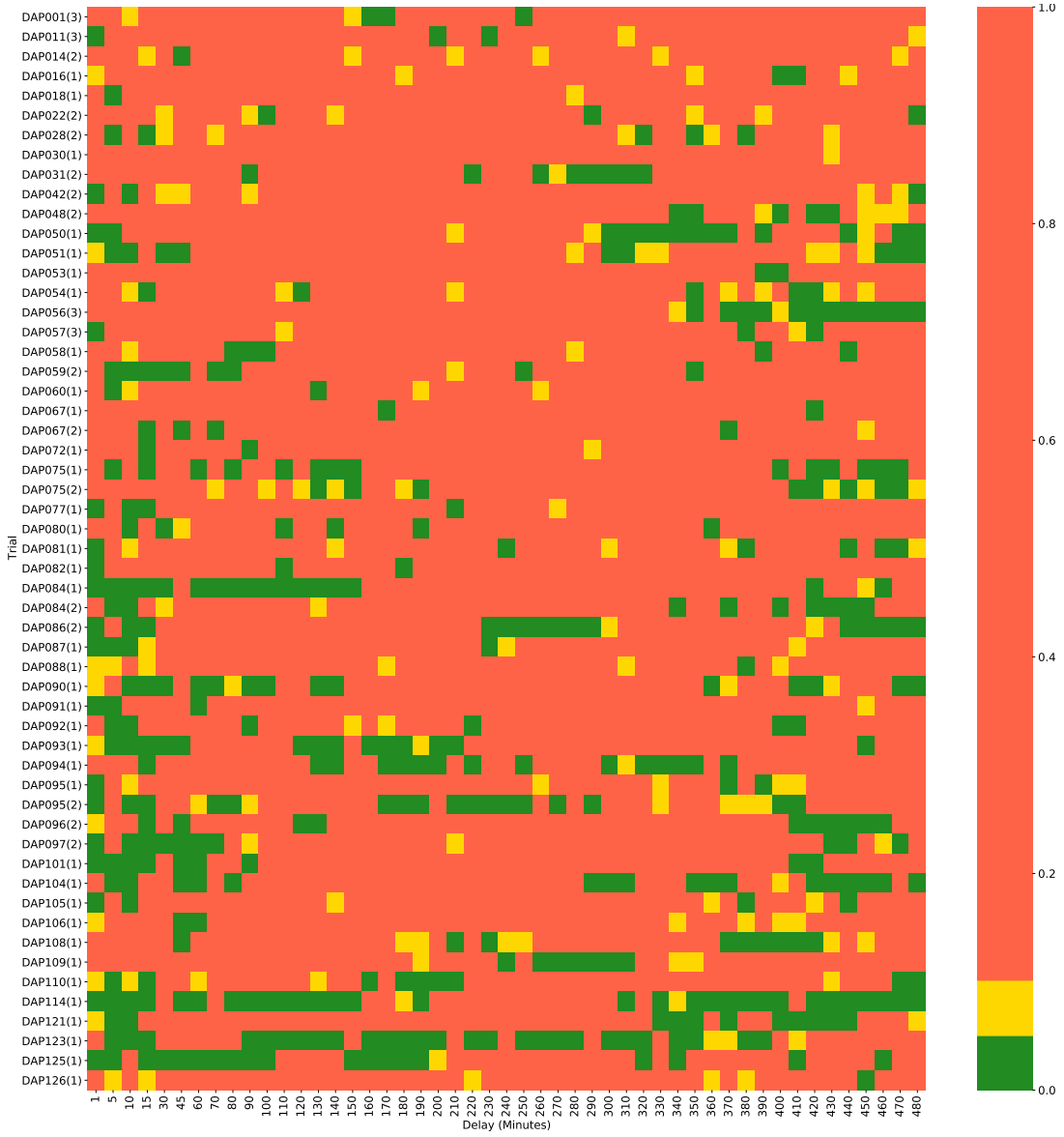

Figure D.3: Causal links p-values color map.

D.1.2.2 1 hour

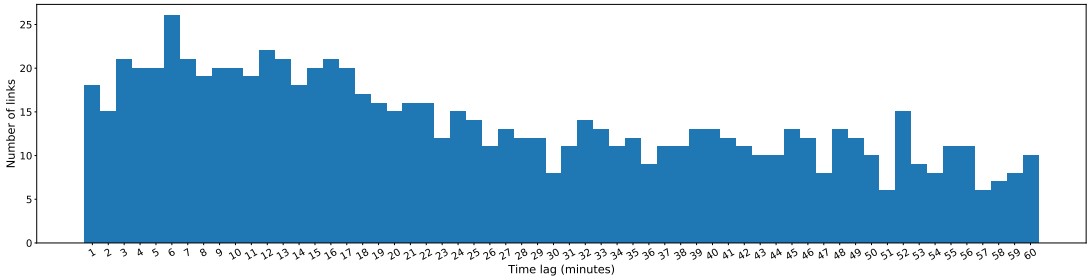

Figure D.4: Distribution of causal links over the first hour.

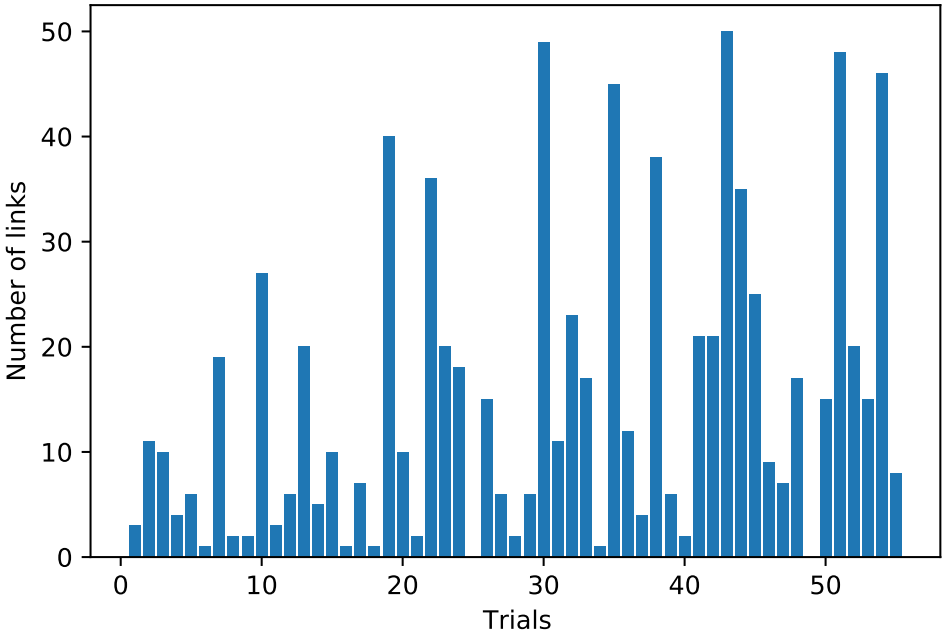

Figure D.5: Number of links per trial.

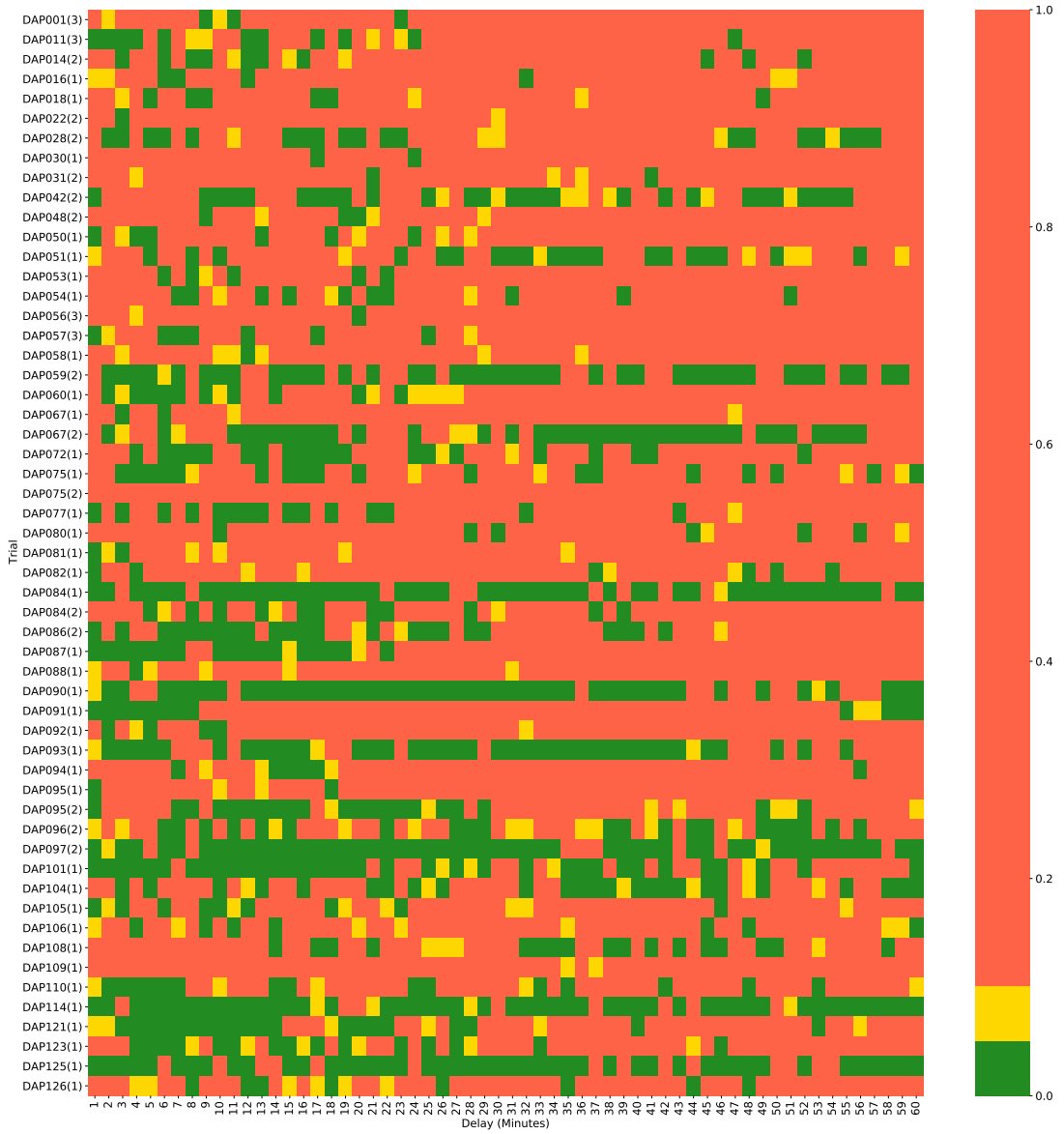

Figure D.6: Causal links p-values color map.

D.2 $PM_{2.5}$, NO_2 and O_3

D.2.1 Linear PCMCI

This section includes the linear PCMCI graphs that could not be included in the main document due to space restrictions.

D.2.1.1 $PM_{2.5}$

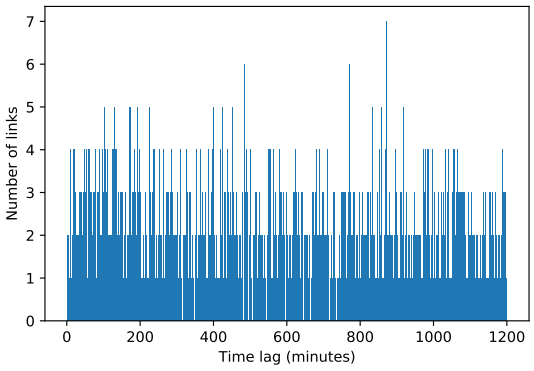

Figure D.7: Number of links per time lag for $PM_{2.5}$.

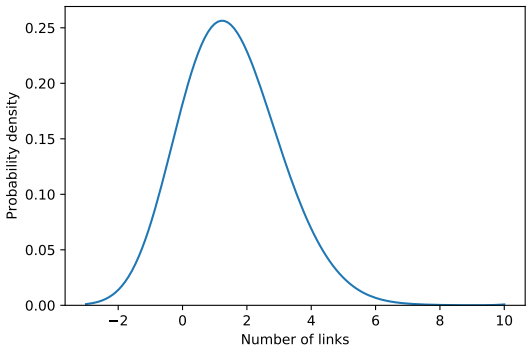

Figure D.8: Number of links distribution for $PM_{2.5}$.

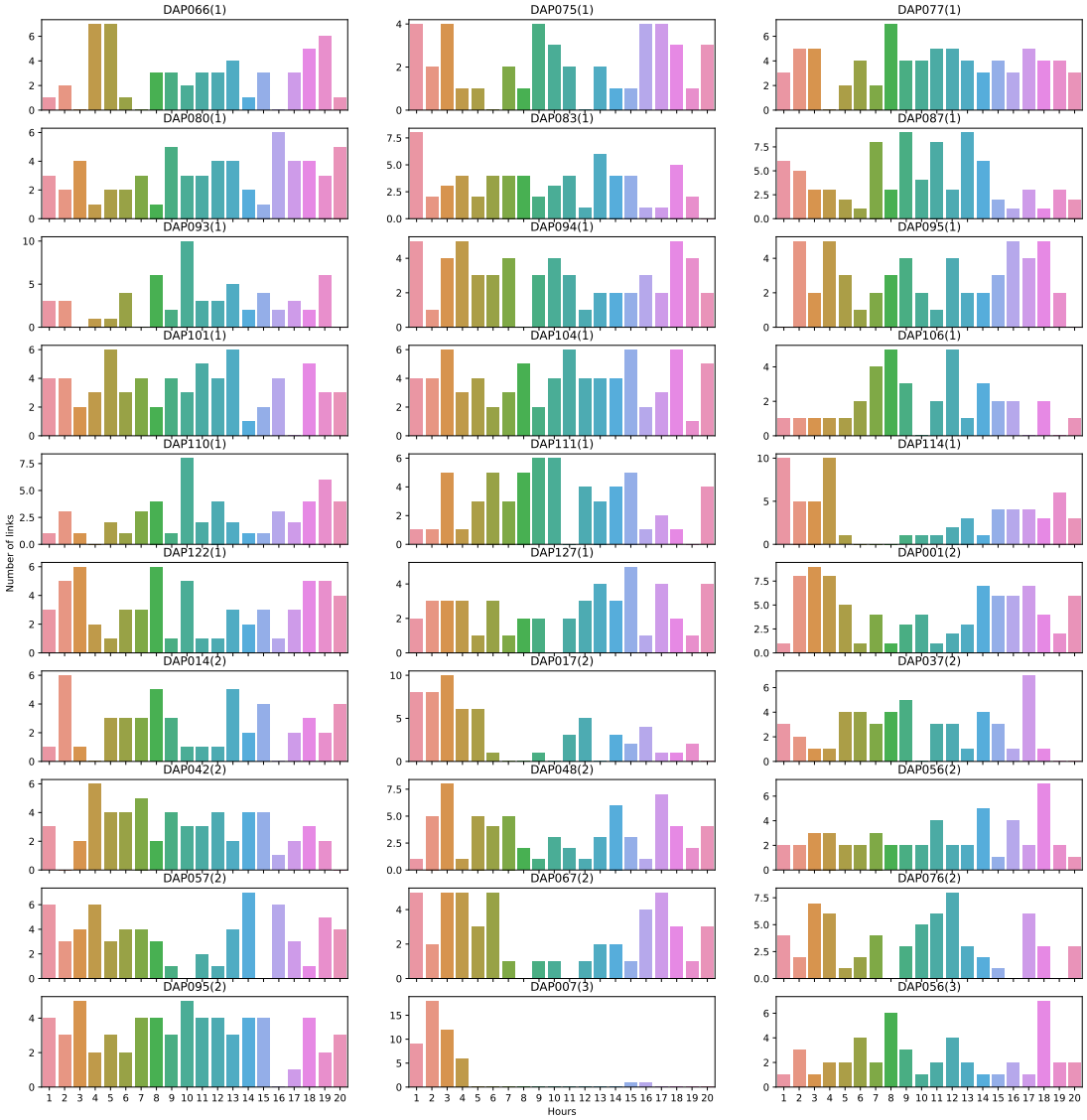

Figure D.9: Distribution of causal links for all the trials for linear PCMCI with personal $PM_{2.5}$ exposure data.

D.2.1.2 NO_2

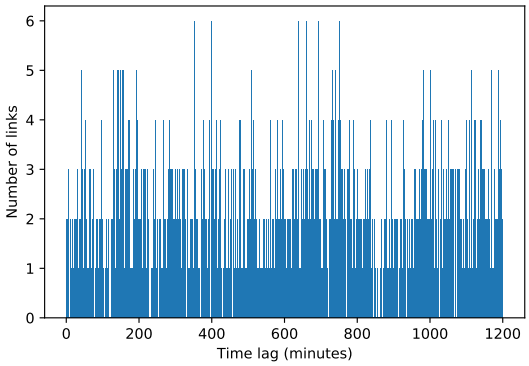

Figure D.10: Number of links per time lag for NO_2.

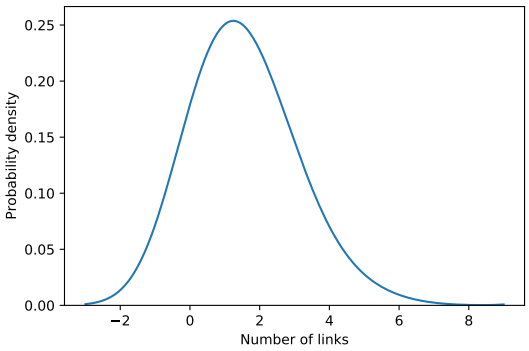

Figure D.11: Number of links distribution for NO_2.

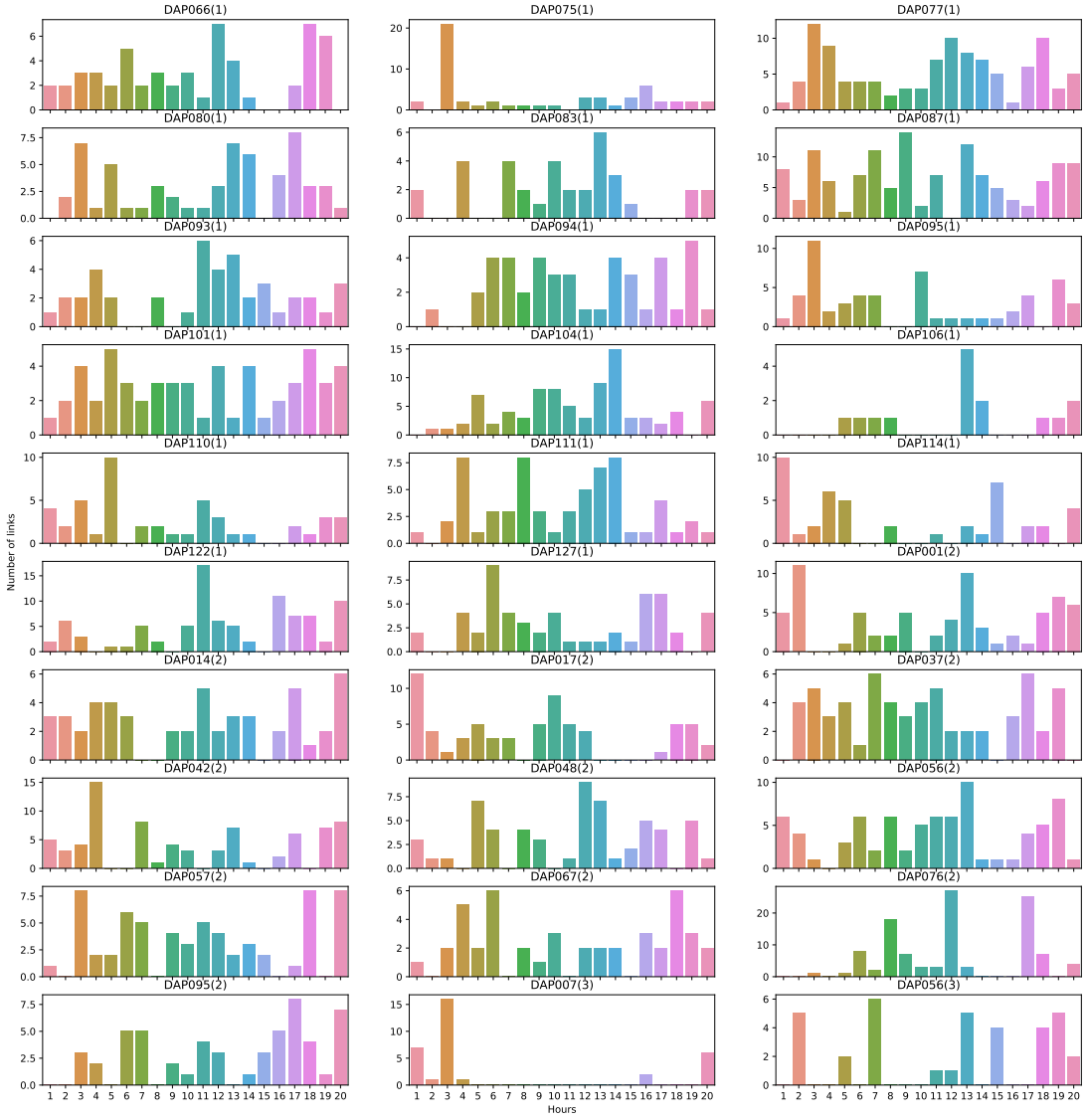

Figure D.12: Distribution of causal links for all the trials for linear PCMCI with NO_2 data.

D.2.1.3 O_3

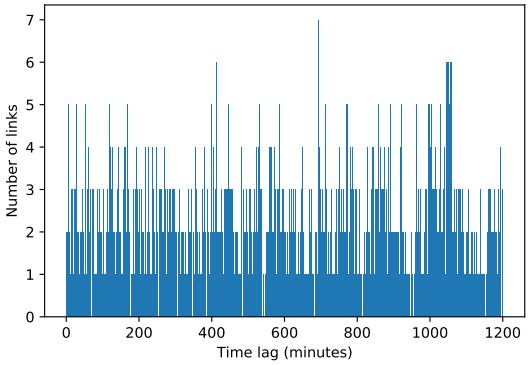

Figure D.13: Number of links per time lag for O_3.

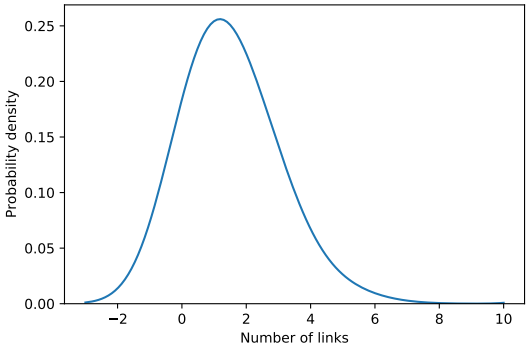

Figure D.14: Number of links distribution for O_3.

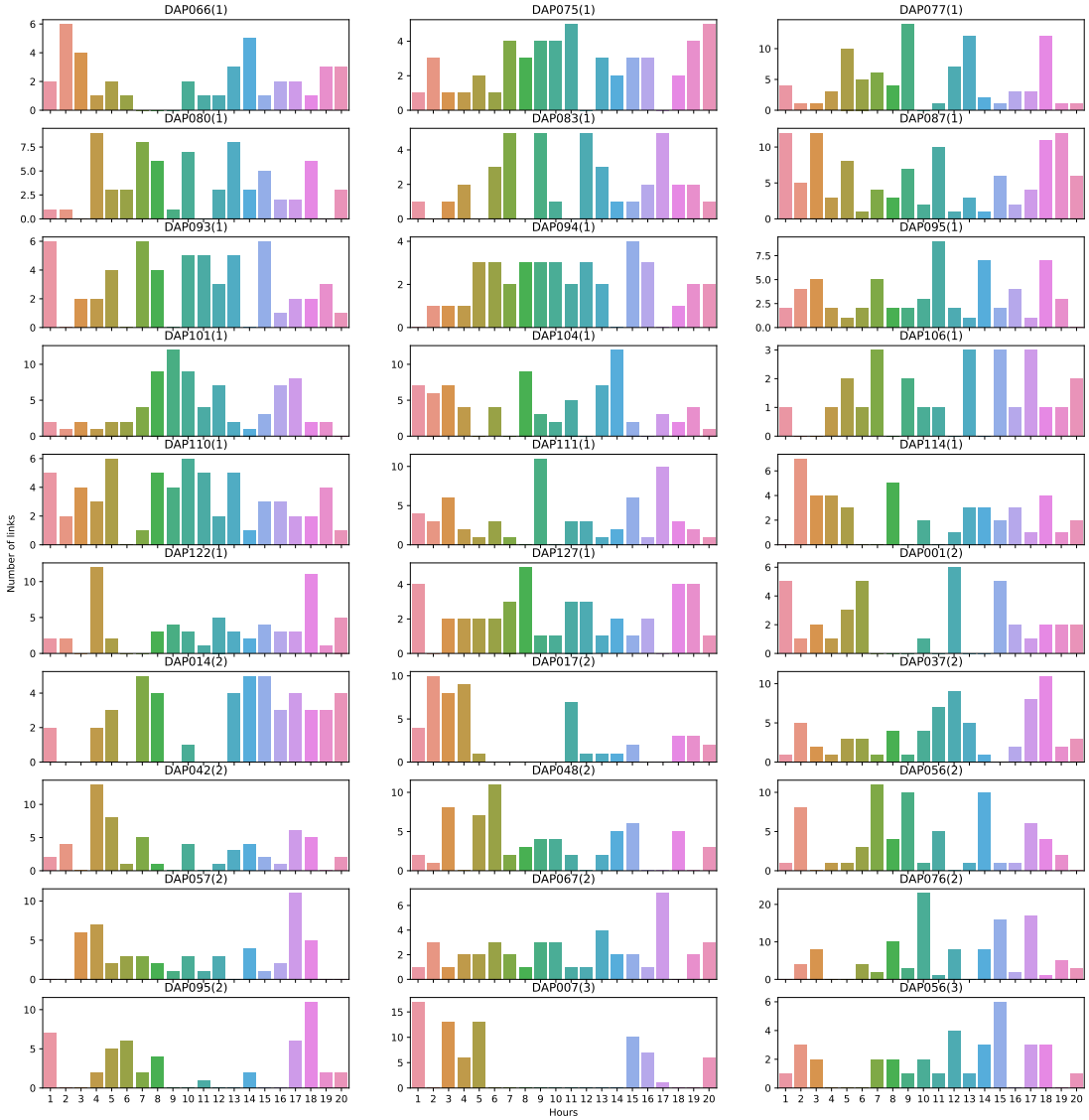

Figure D.15: Distribution of causal links for all the trials for linear PCMCI with O_3 data.

D.2.2 Non-linear PCMCI

This section contains the detailed results per trial of non-linear PCMCI when using past breathing rate values, temperature, humidity, $PM_{2.5}$, NO_2 and O_3 data, that could not be included in the main document due to space constraints. It includes the color maps that represent the p-values that PCMCI returns when testing causality with the breathing rate for each trial and each time lag. When the p-value is $<= 0.05$, then the result is statistically significant to determine that there is a causal link, which is plotted in green. Those combinations of trials and time lags who were close to be statistically

significant ($0.05 <$ p-value $<= 0.1$) are plotted in yellow, and the rest in red. It also includes the color map depicting the strength of the causal links found with a more intense tone when the link is stronger.

D.2.2.1 $PM_{2.5}$

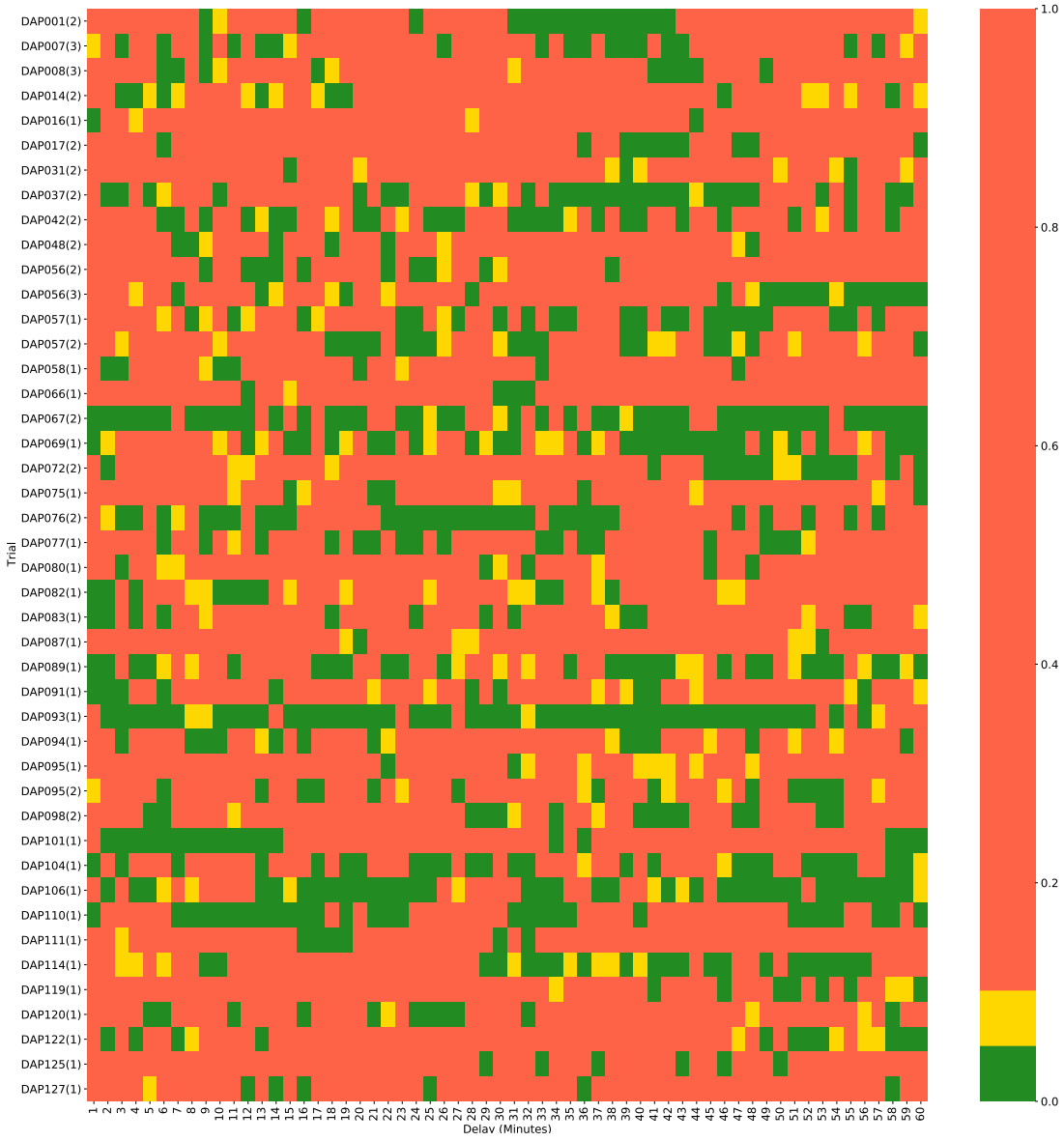

Figure D.16: Distribution of causal links for all the trials for non-linear PCMCI with $PM_{2.5}$ data for the 1^{st} hour

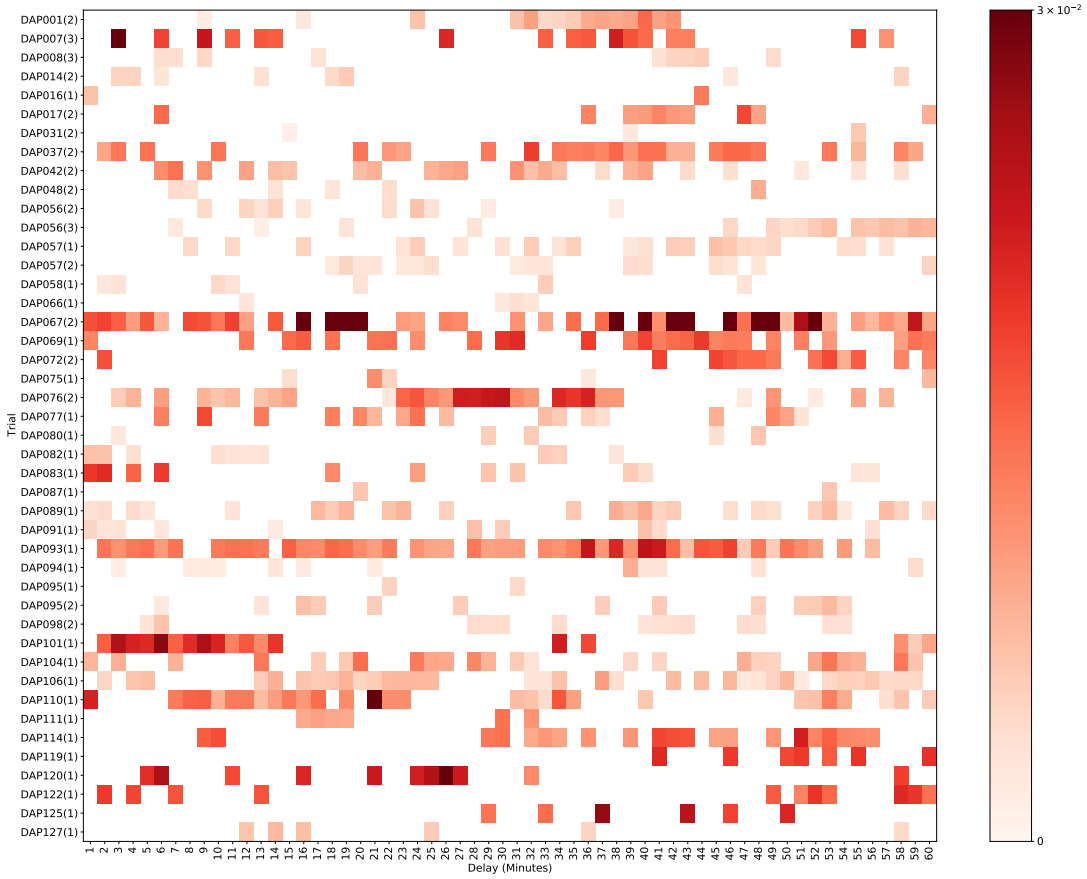

Figure D.17: Causal link intensity for non-linear PCMCI with $PM_{2.5}$ data for the 1^{st} hour

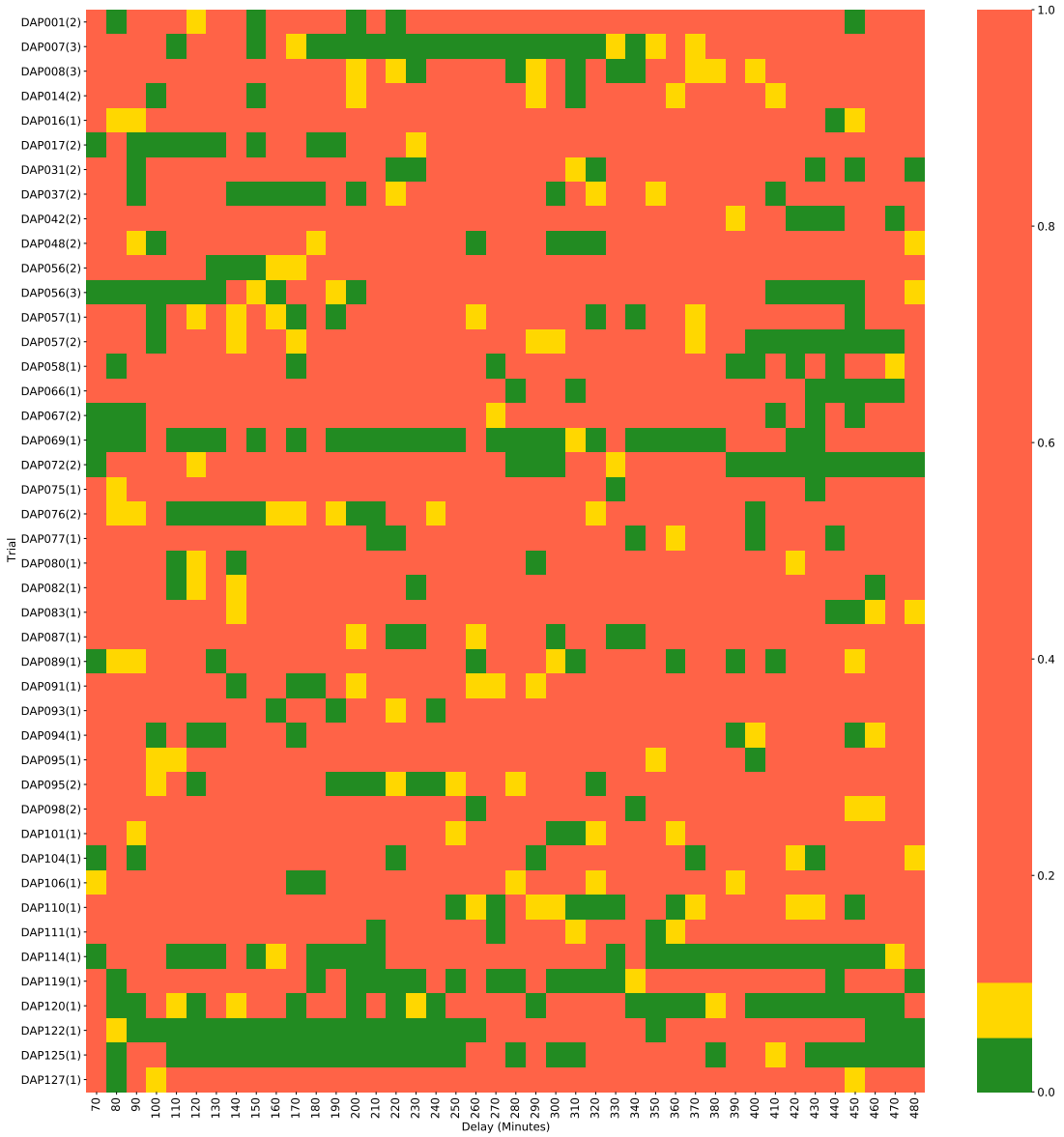

Figure D.18: Distribution of causal links for all the trials for non-linear PCMCI with $PM_{2.5}$ data from the 2^{nd} to 8^{th} hour

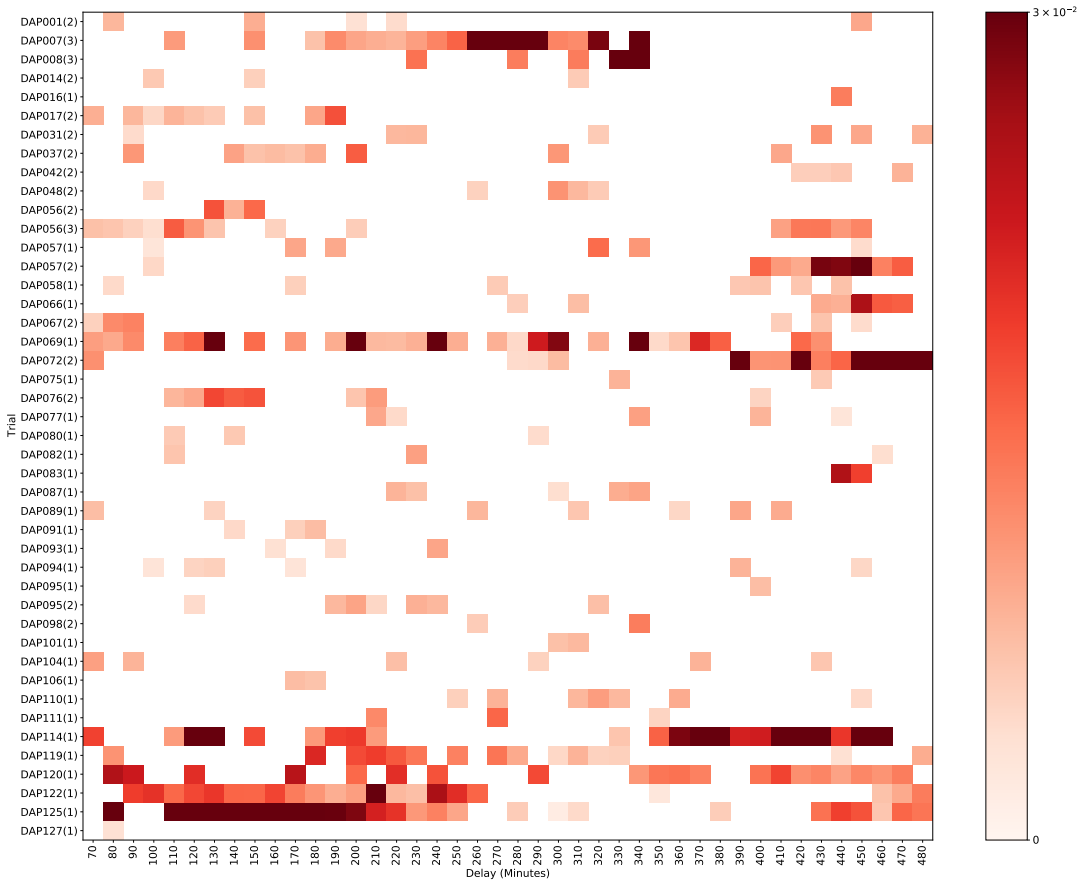

Figure D.19: Causal link intensity for non-linear PCMCI with $PM_{2.5}$ data from the 2^{nd} to 8^{th} hour

D.2.2.2 NO_2

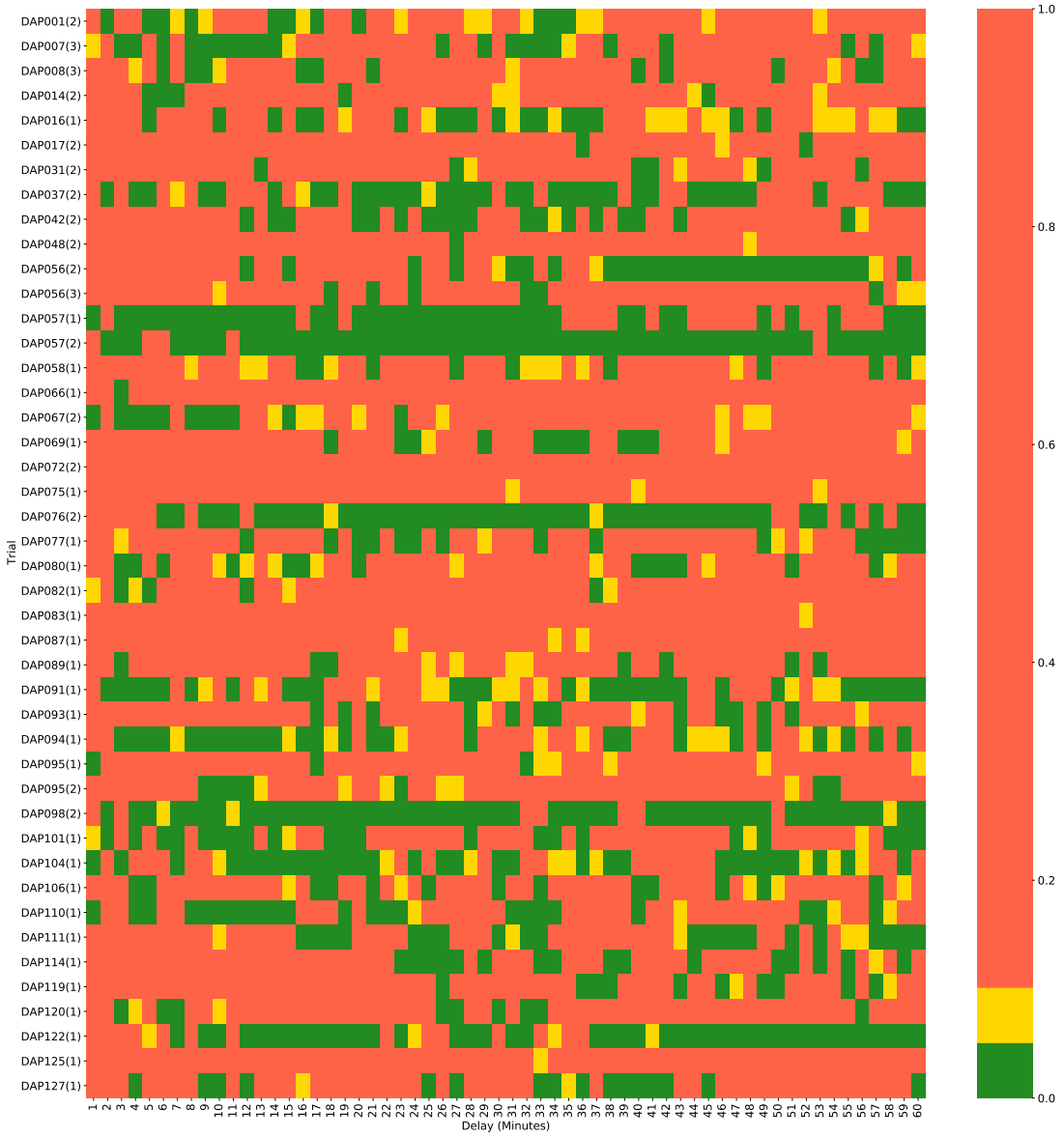

Figure D.20: Distribution of causal links for all the trials for non-linear PCMCI with NO_2 data for the 1^{st} hour

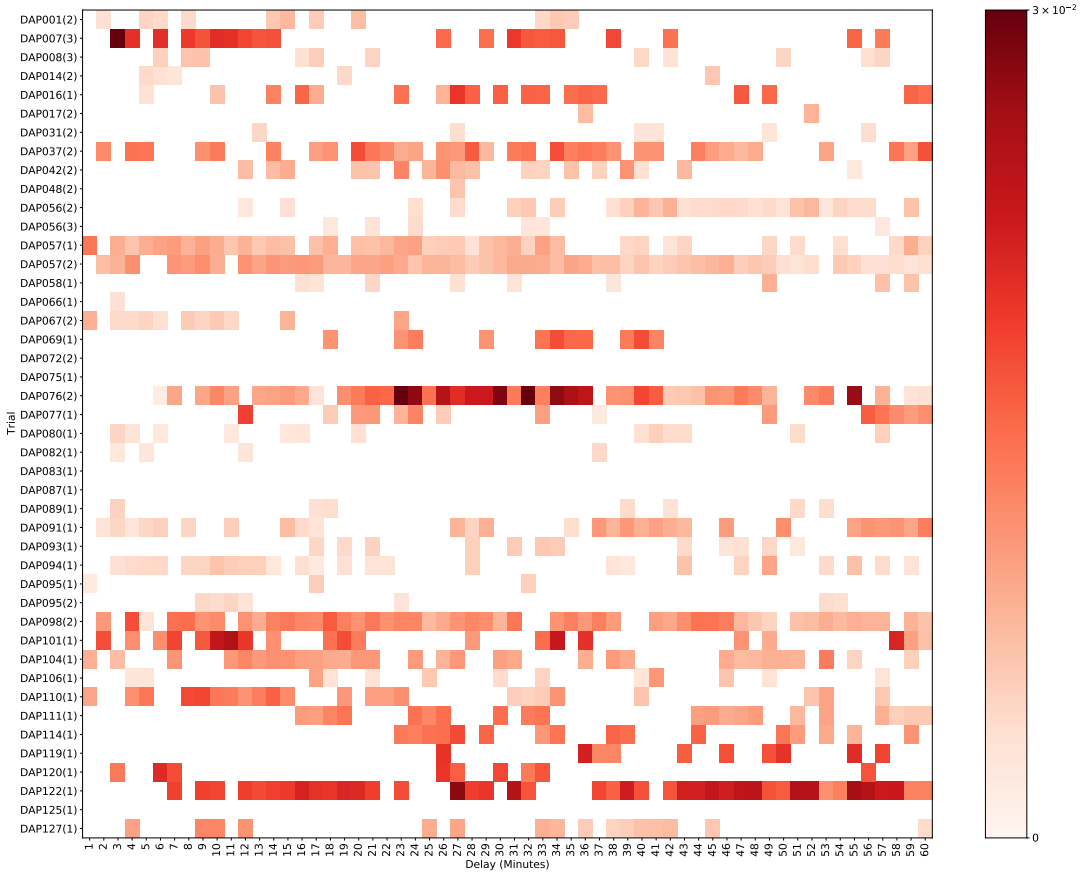

Figure D.21: Causal link intensity for non-linear PCMCI with NO_2 data for the 1st hour

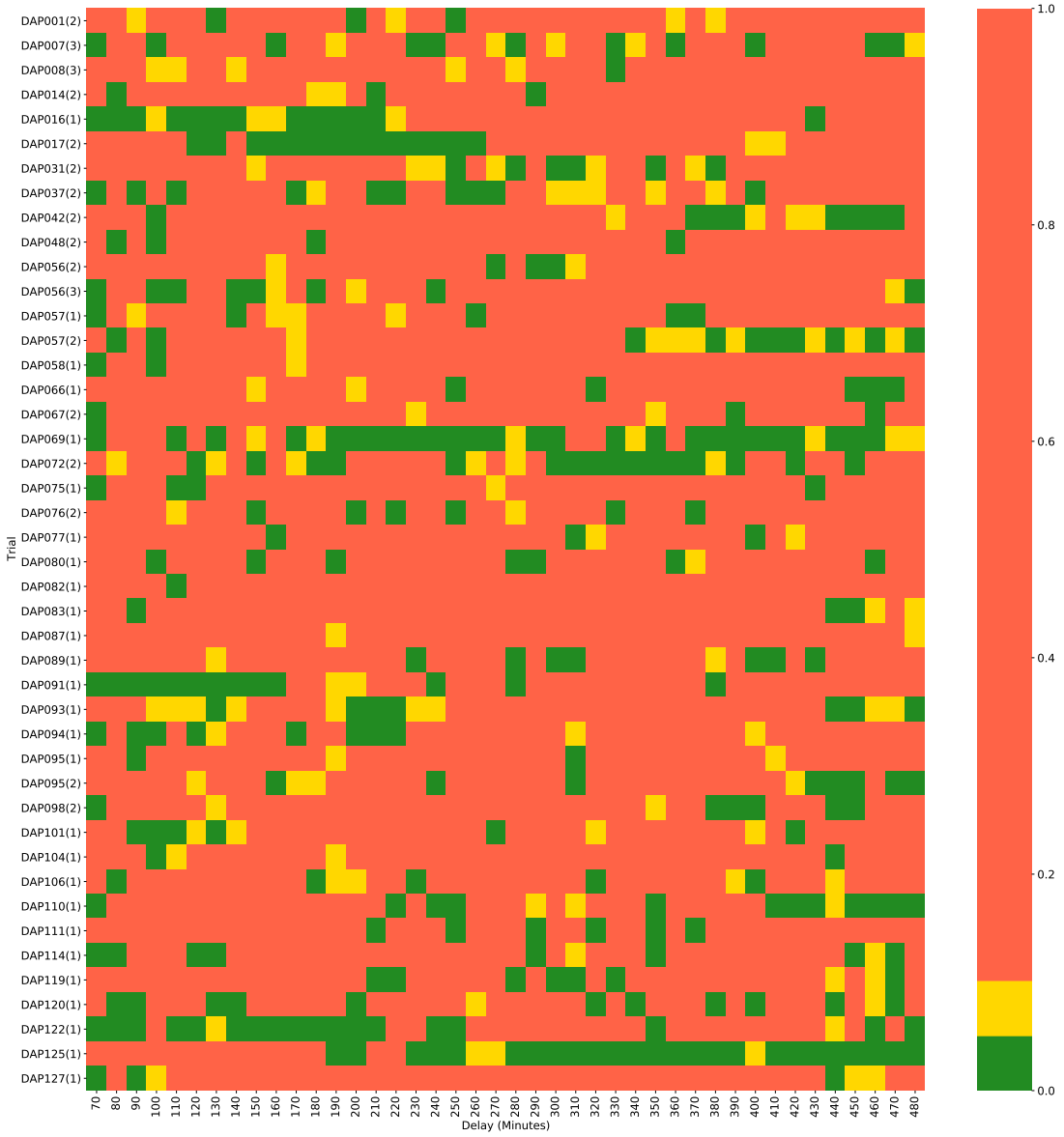

Figure D.22: Distribution of causal links for all the trials for non-linear PCMCI with NO_2 data from the 2^{nd} to 8^{th} hour

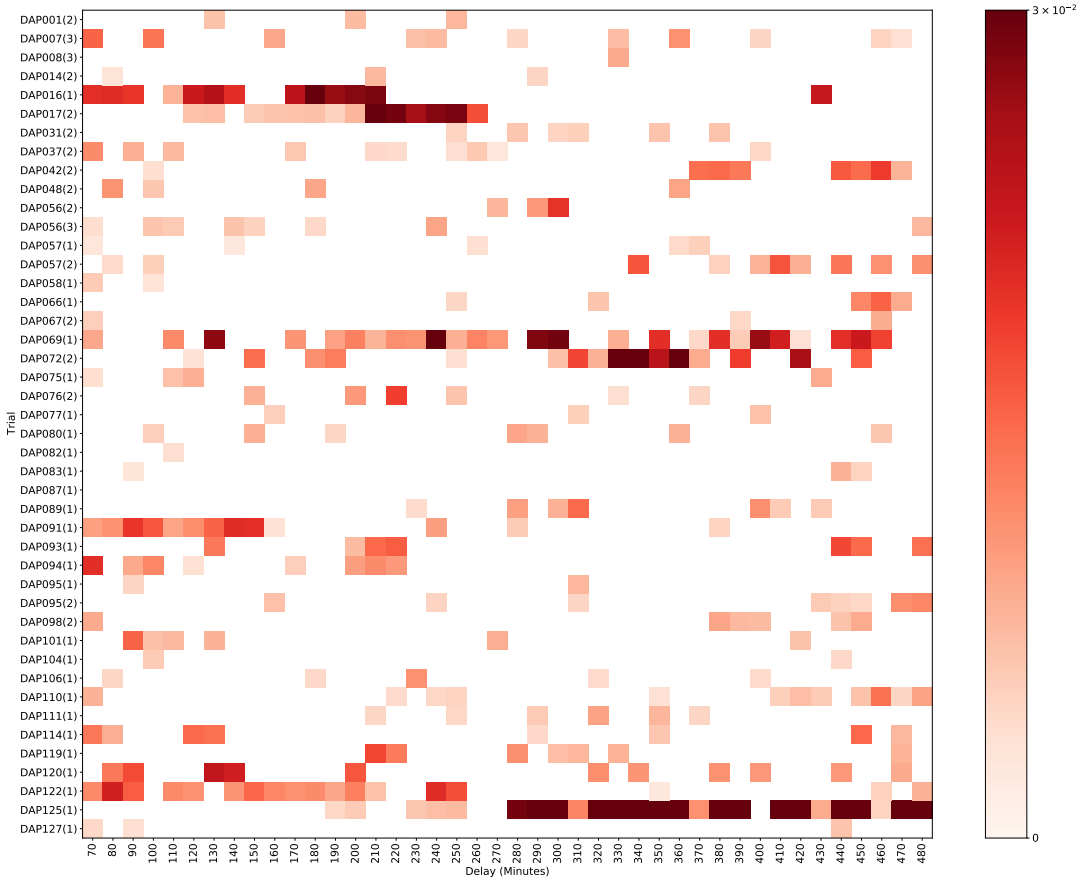

Figure D.23: Causal link intensity for non-linear PCMCI with NO_2 data from the 2^{nd} to 8^{th} hour

D.2.2.3 O_3

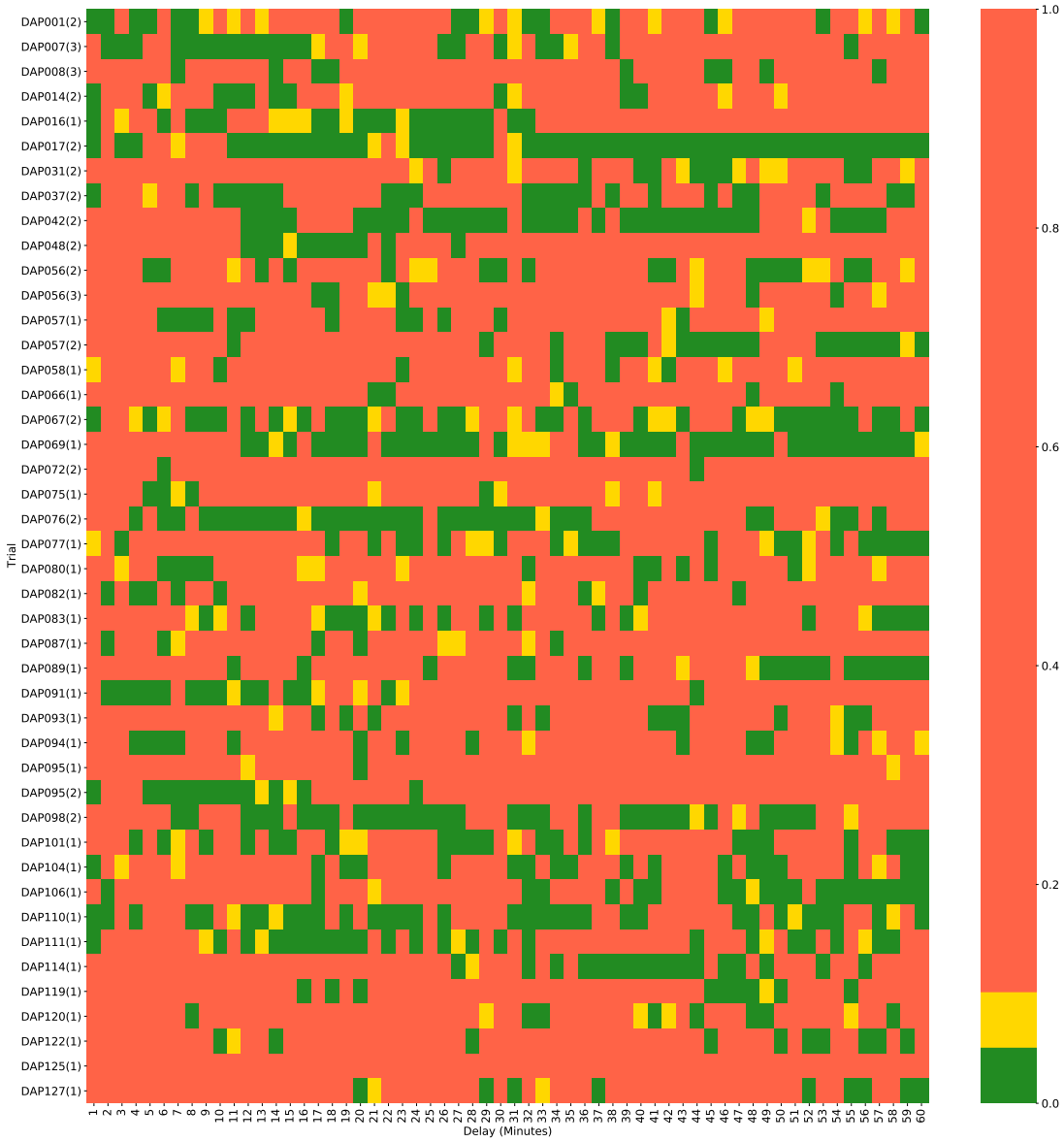

Figure D.24: Distribution of causal links for all the trials for non-linear PCMCI with O_3 data for the 1st hour

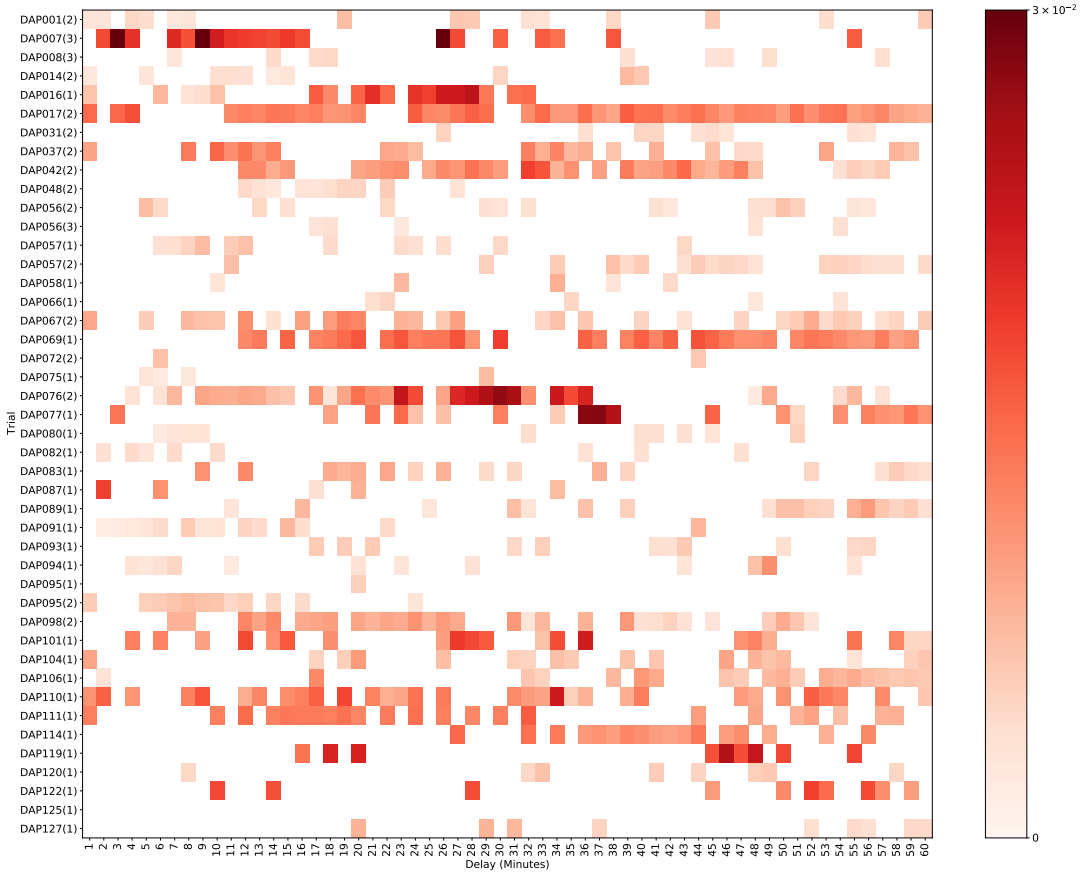

Figure D.25: Causal link intensity for non-linear PCMCI with O_3 data for the 1st hour

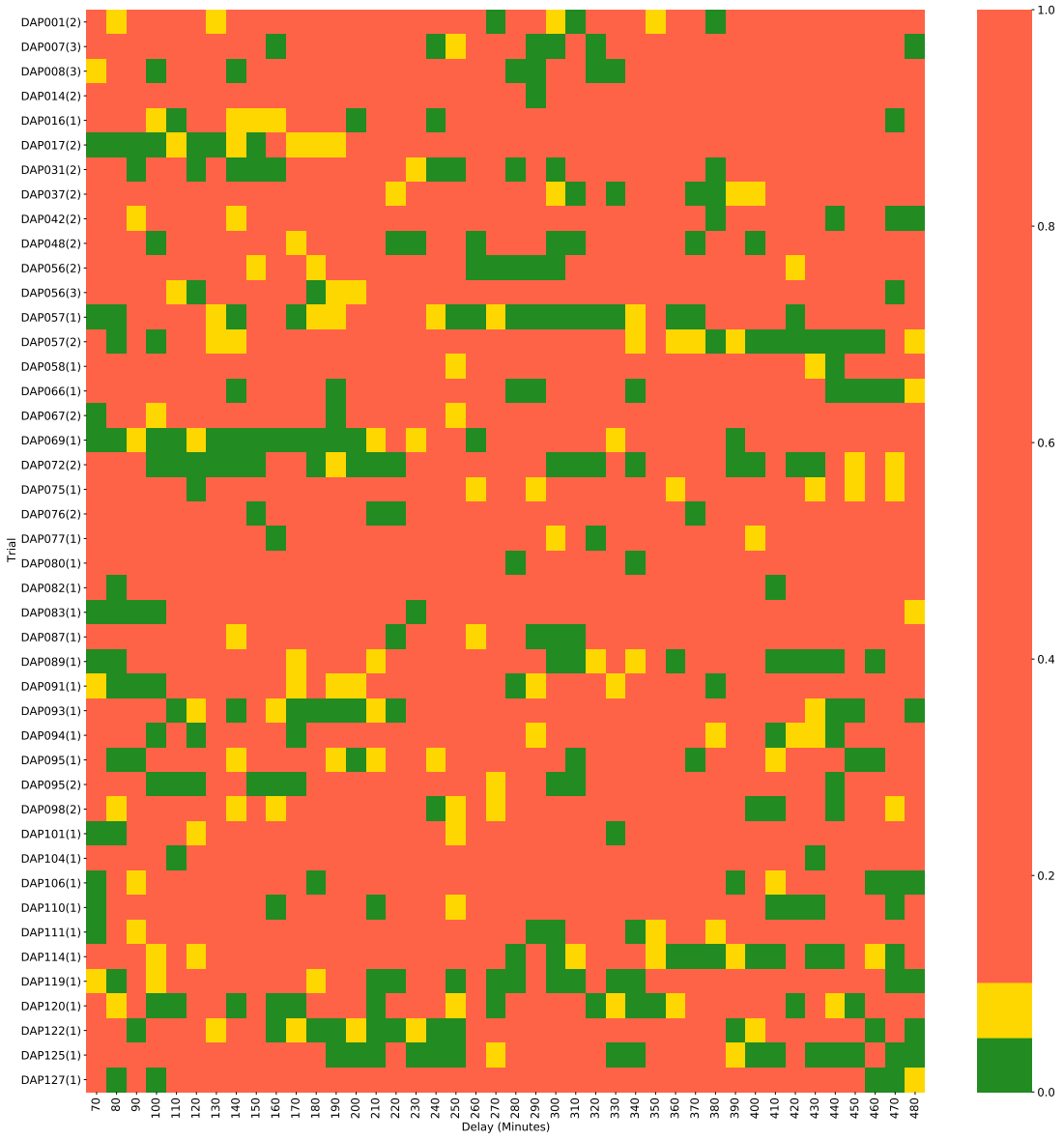

Figure D.26: Distribution of causal links for all the trials for non-linear PCMCI with O_3 data from the 2^{nd} to 8^{th} hour

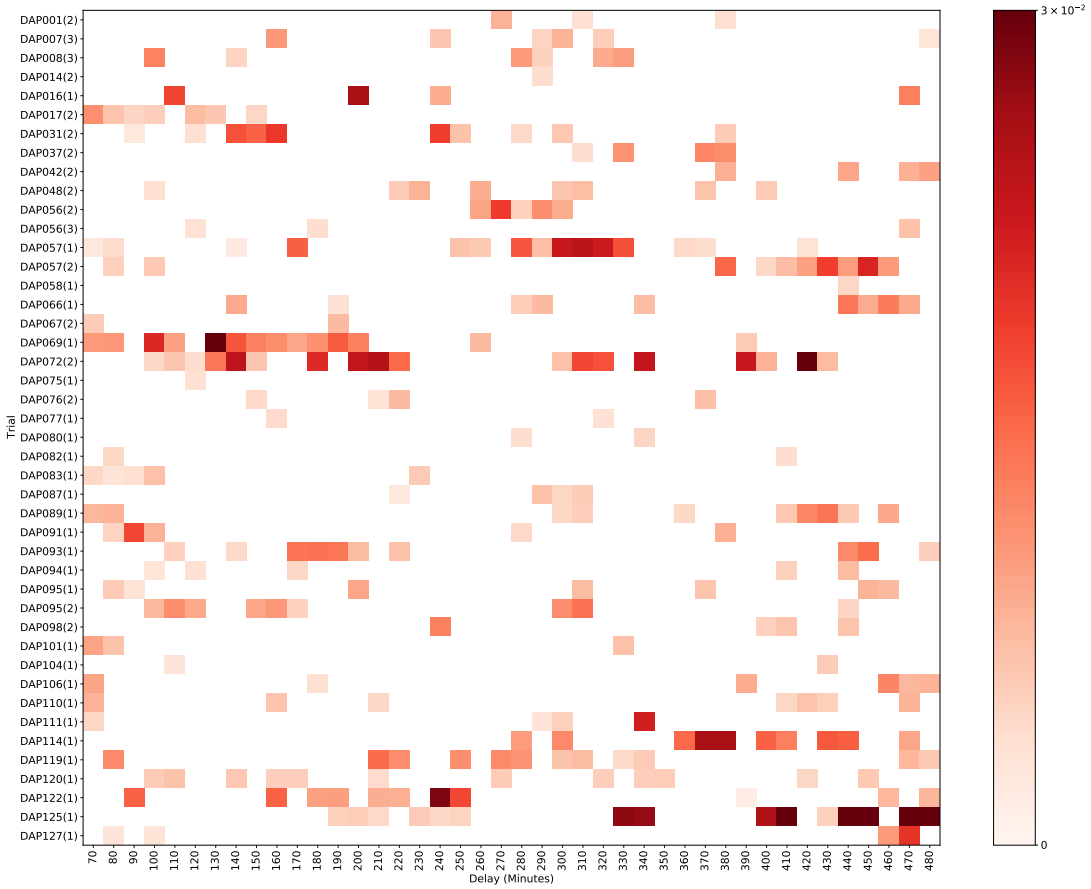

Figure D.27: Causal link intensity for non-linear PCMCI with O_3 data from the 2^{nd} to 8^{th} hour

D.2.2.4 Past breathing rate

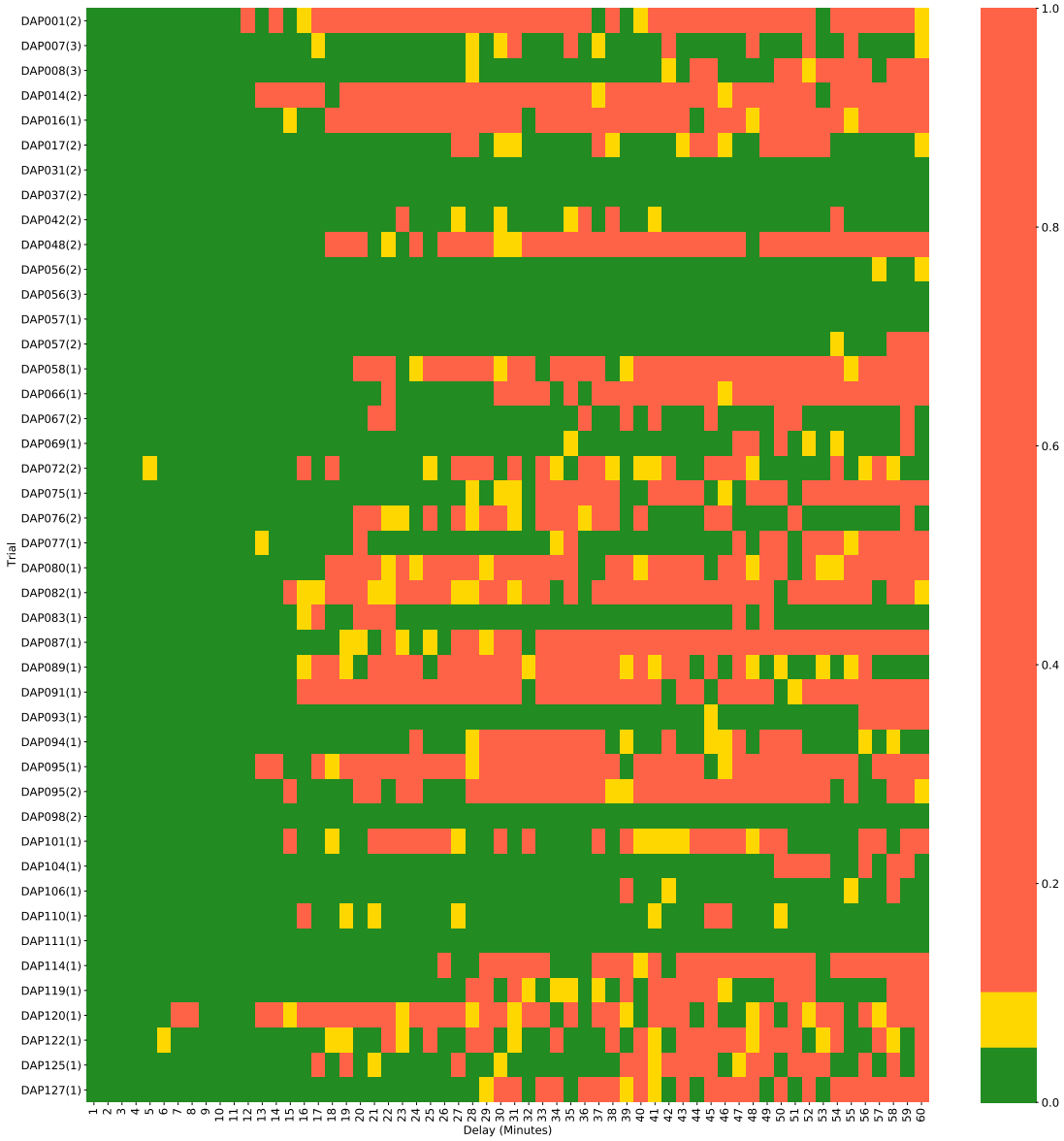

Figure D.28: Distribution of causal links for all the trials for non-linear PCMCI with past breathing rate data for the 1st hour

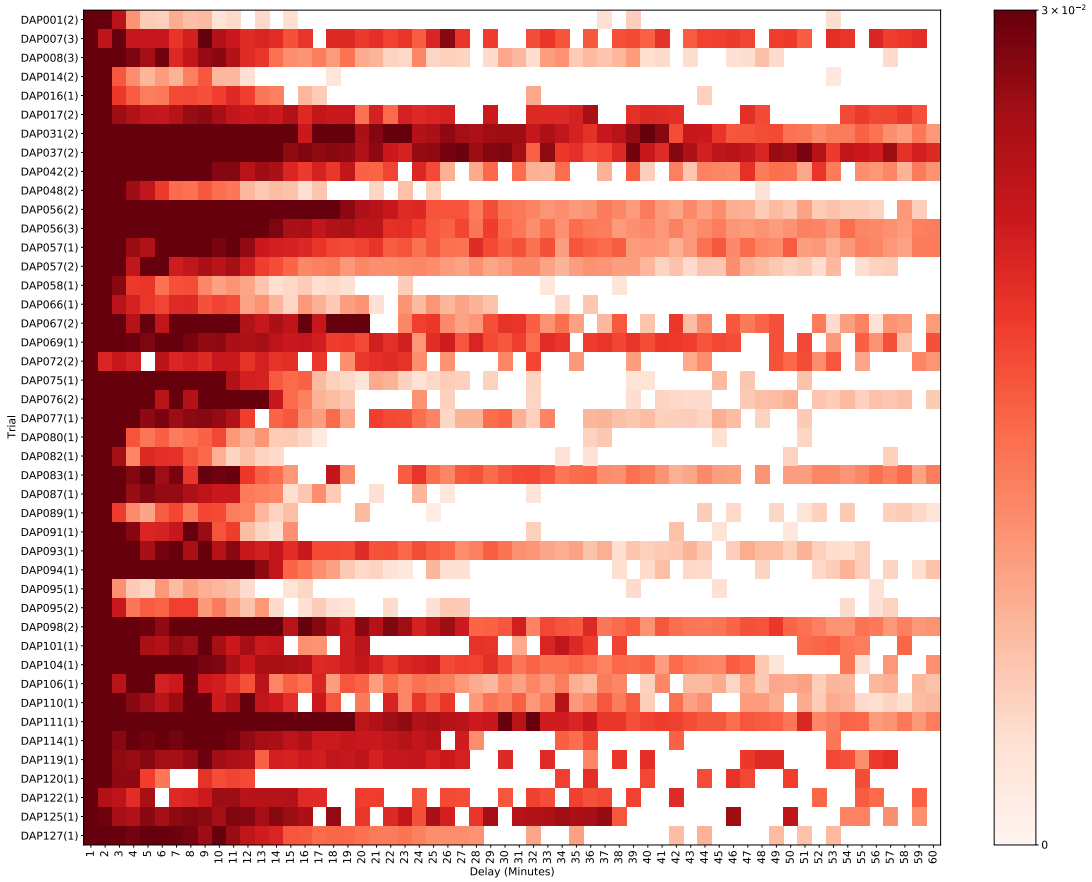

Figure D.29: Causal link intensity for non-linear PCMCI with past breathing rate data for the 1st hour

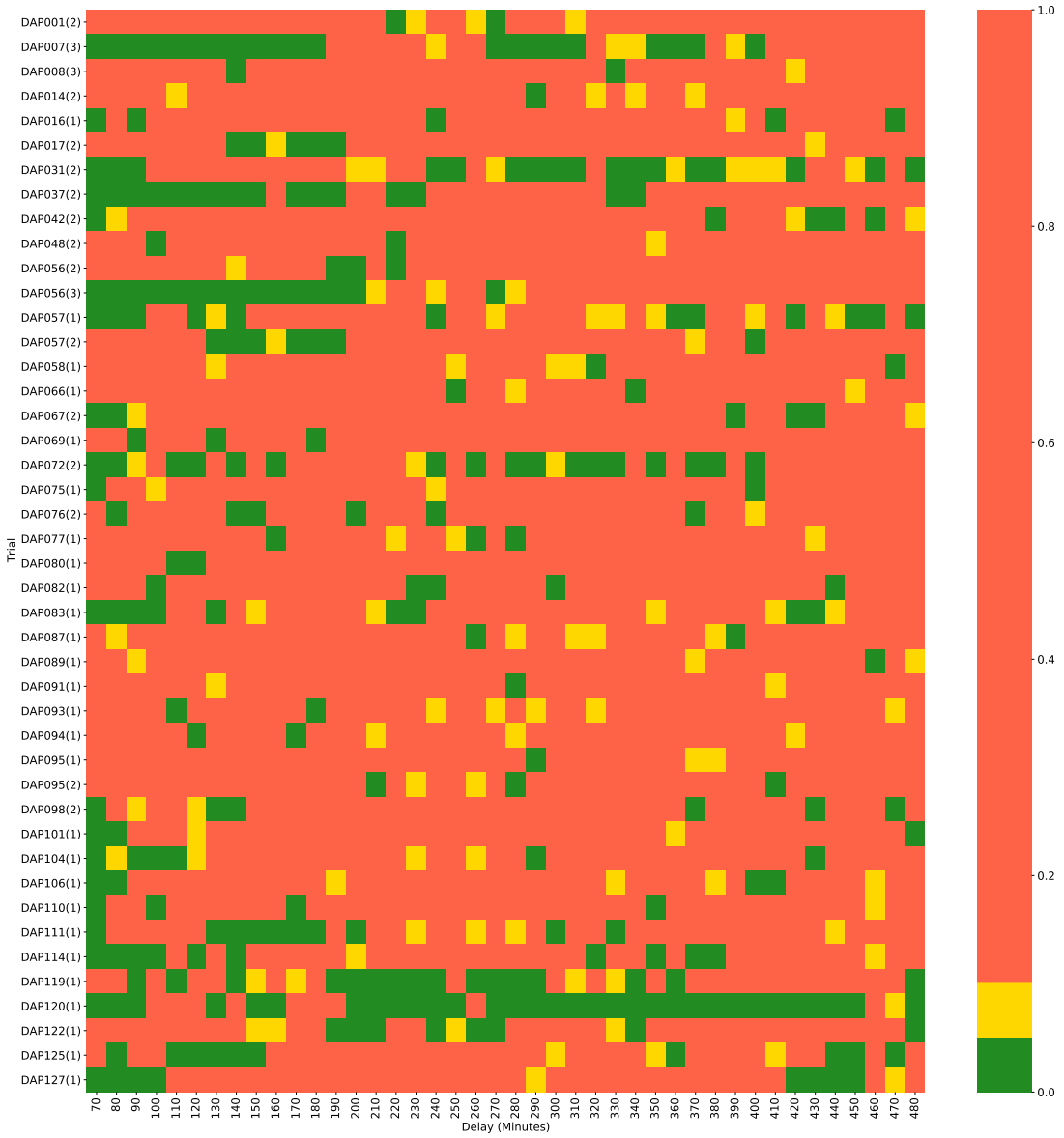

Figure D.30: Distribution of causal links for all the trials for non-linear PCMCI with past breathing rate data from the 2nd to 8th hour

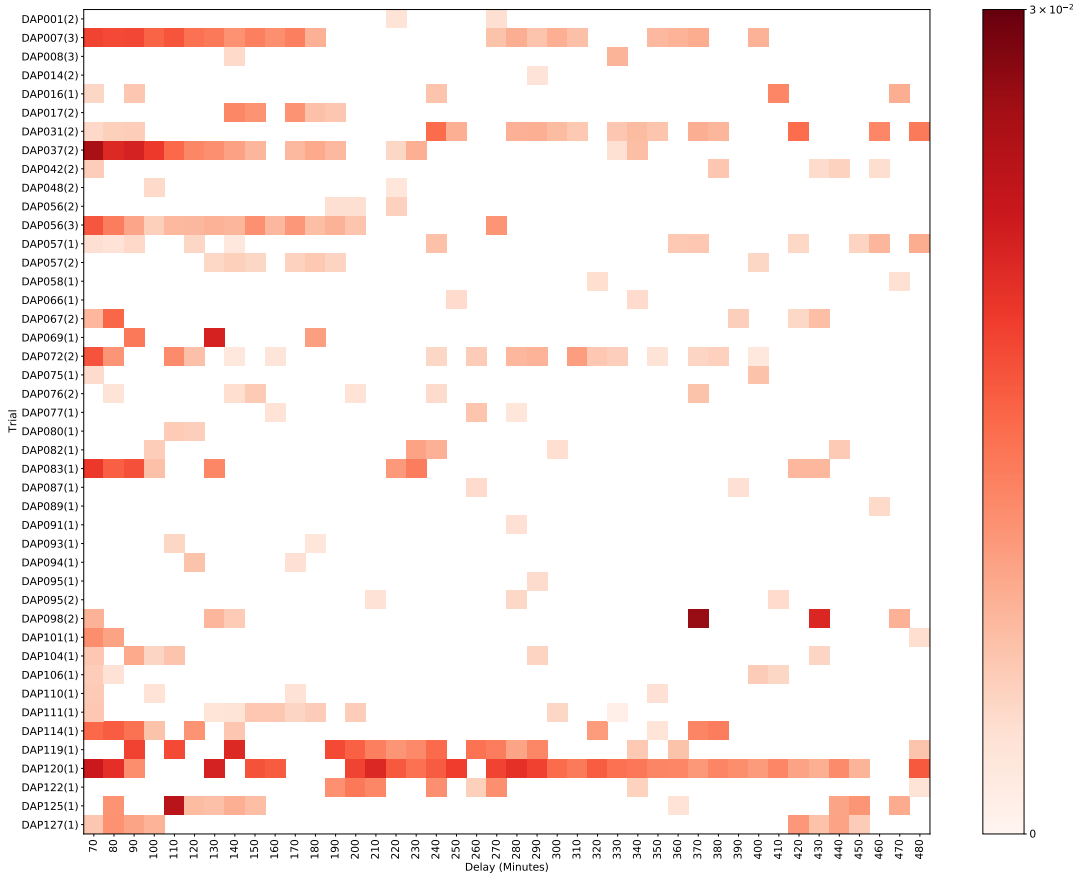

Figure D.31: Causal link intensity for non-linear PCMCI with past breathing rate data from the 2nd to 8th hour

D.2.2.5 Temperature

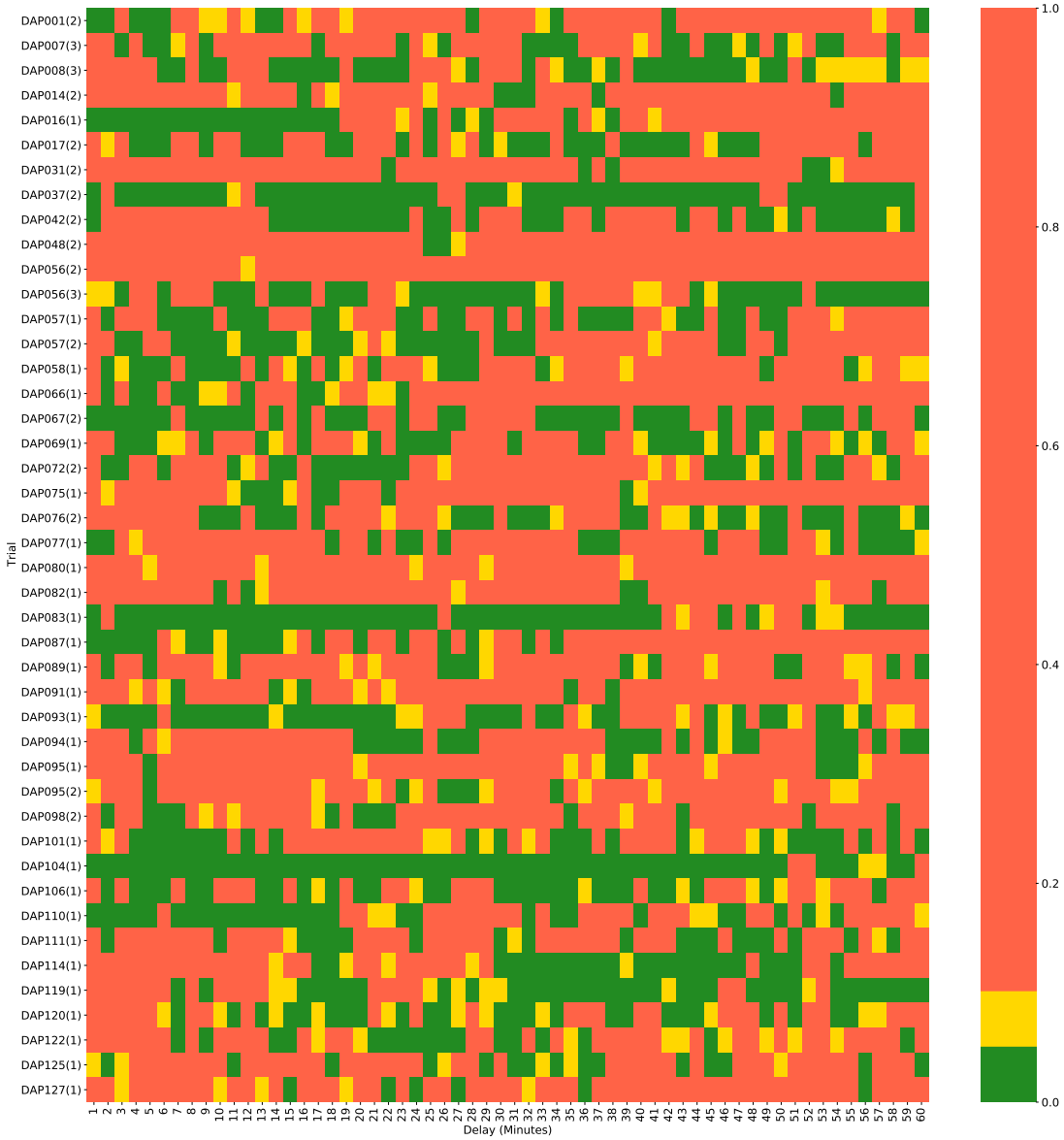

Figure D.32: Distribution of causal links for all the trials for non-linear PCMCI with temperature data for the 1st hour

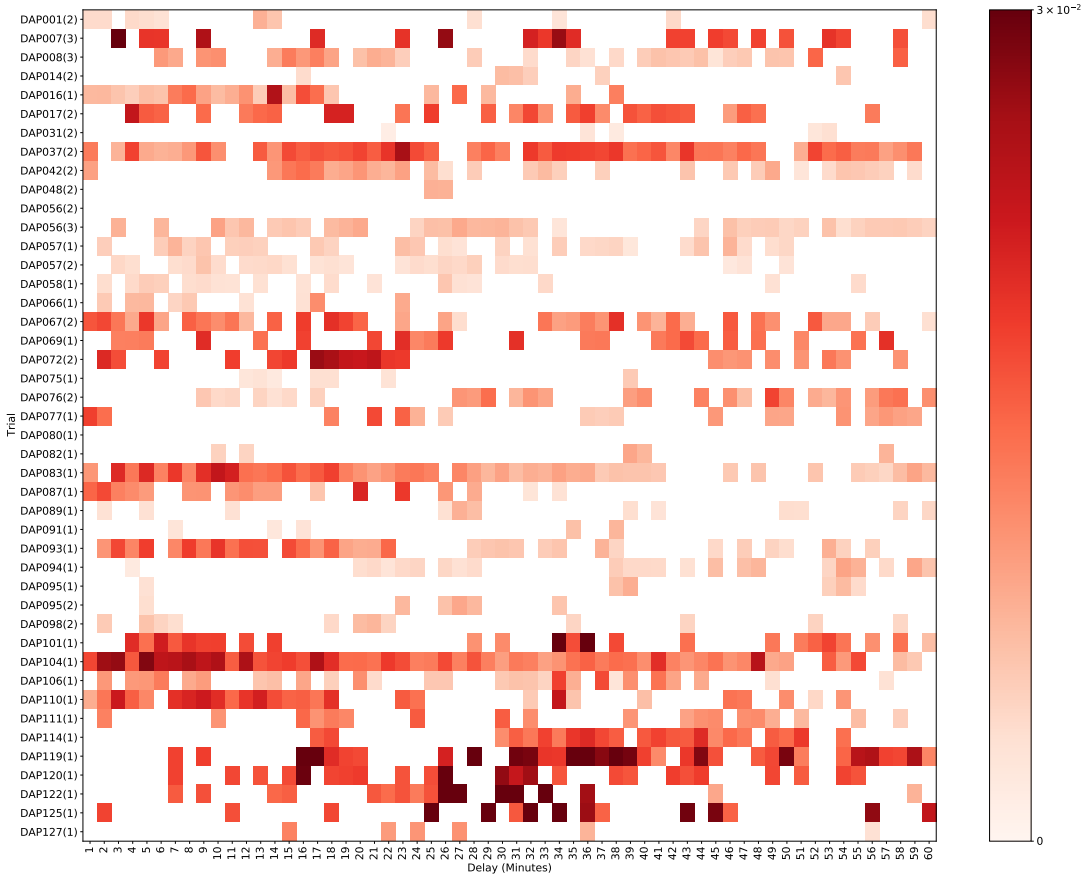

Figure D.33: Causal link intensity for non-linear PCMCI with temperature data for the 1st hour

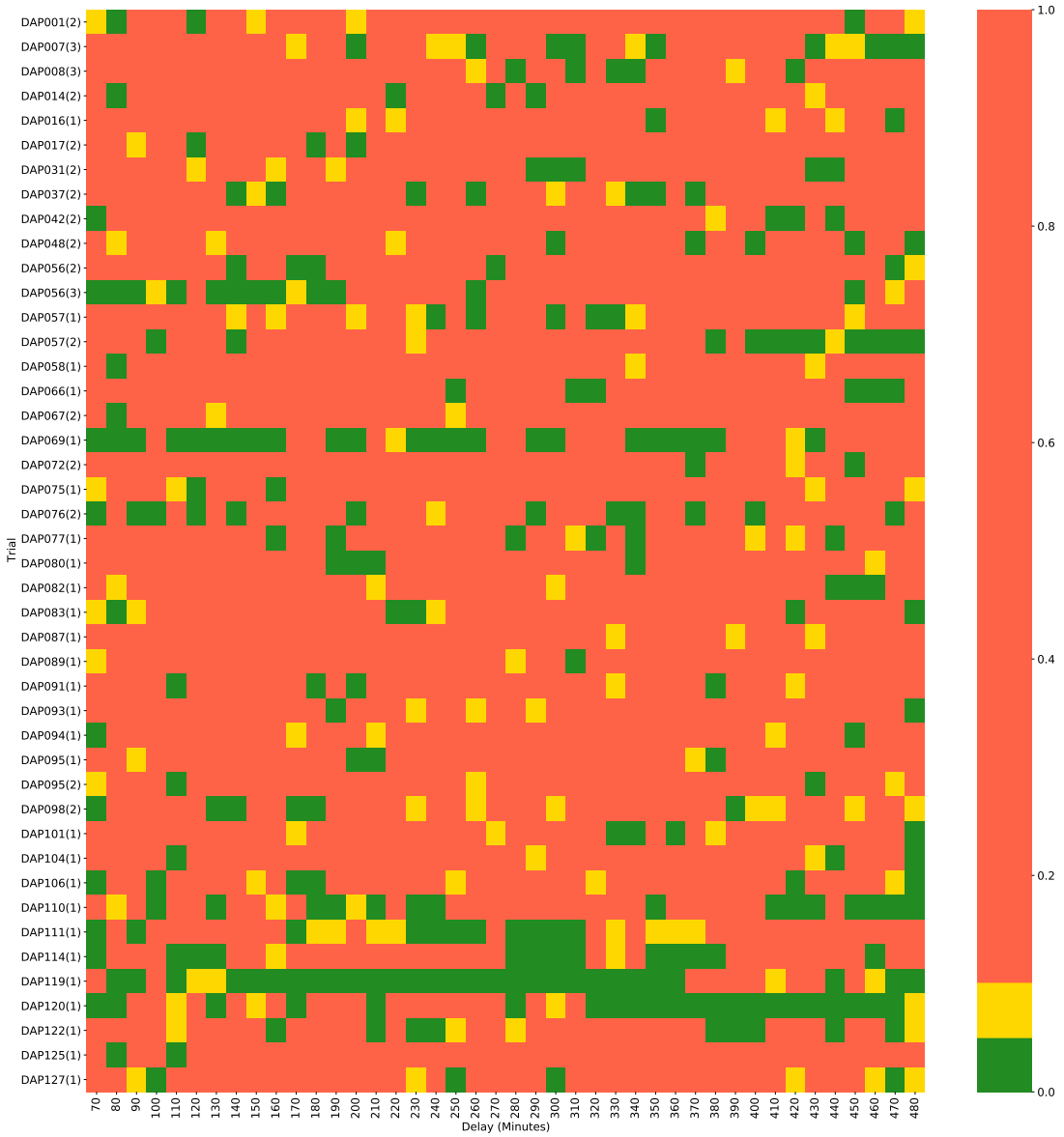

Figure D.34: Distribution of causal links for all the trials for non-linear PCMCI with temperature data from the 2nd to 8th hour

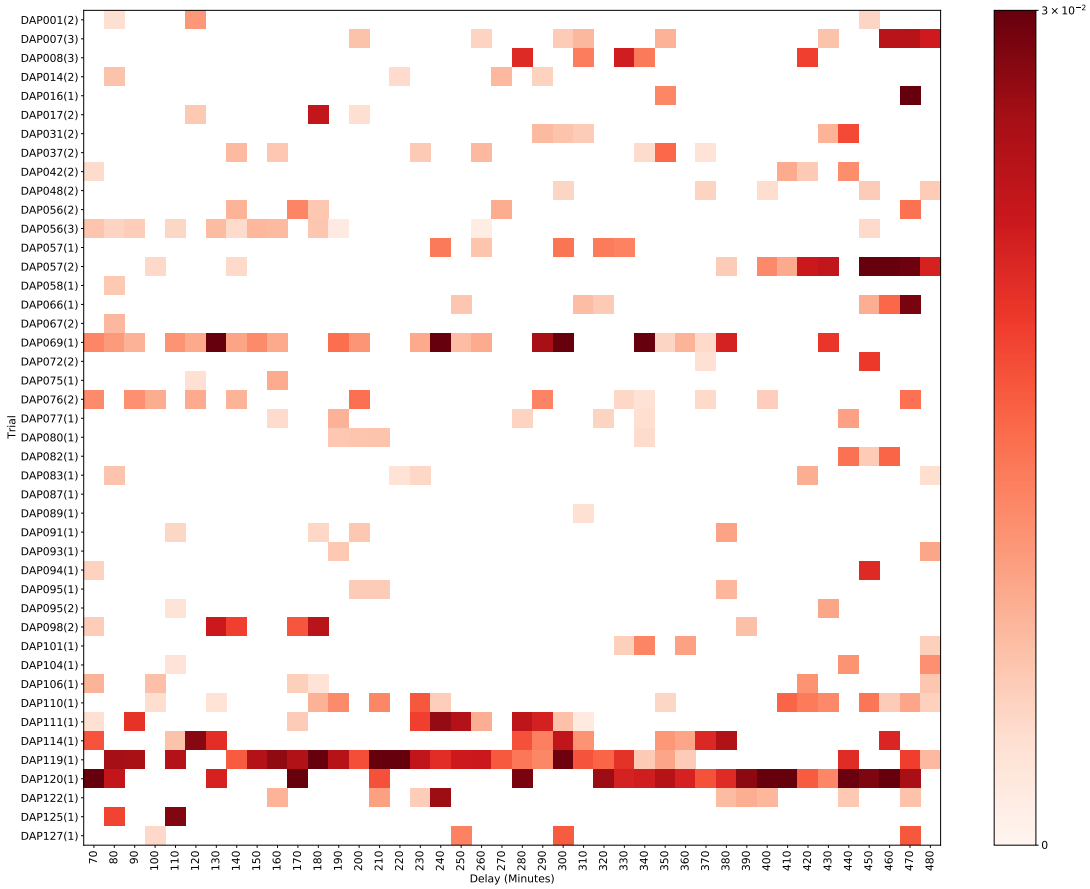

Figure D.35: Causal link intensity for non-linear PCMCI with temperature data from the 2^{nd} to 8^{th} hour

D.2.2.6 Humidity

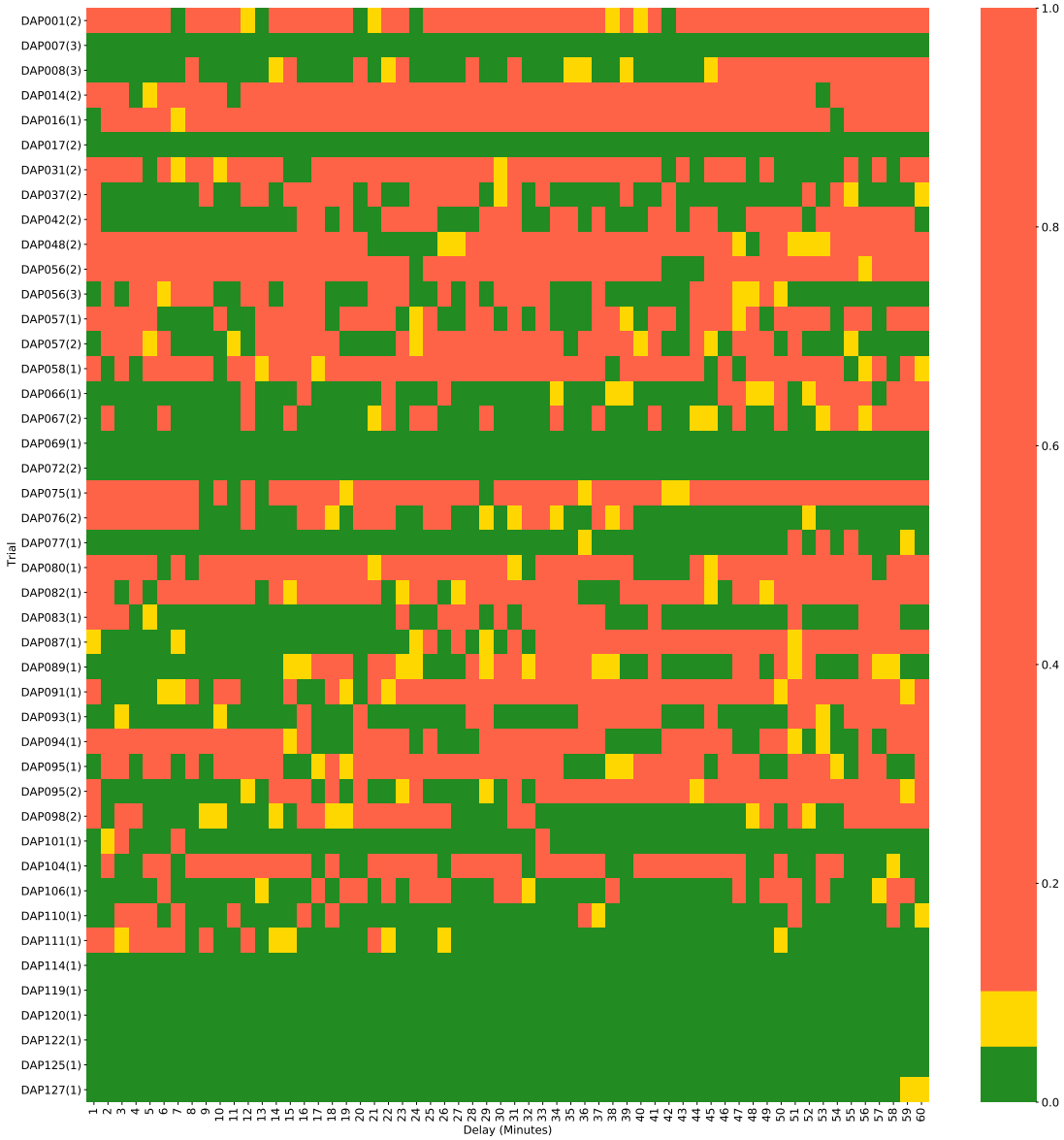

Figure D.36: Distribution of causal links for all the trials for non-linear PCMCI with humidity data for the 1st hour

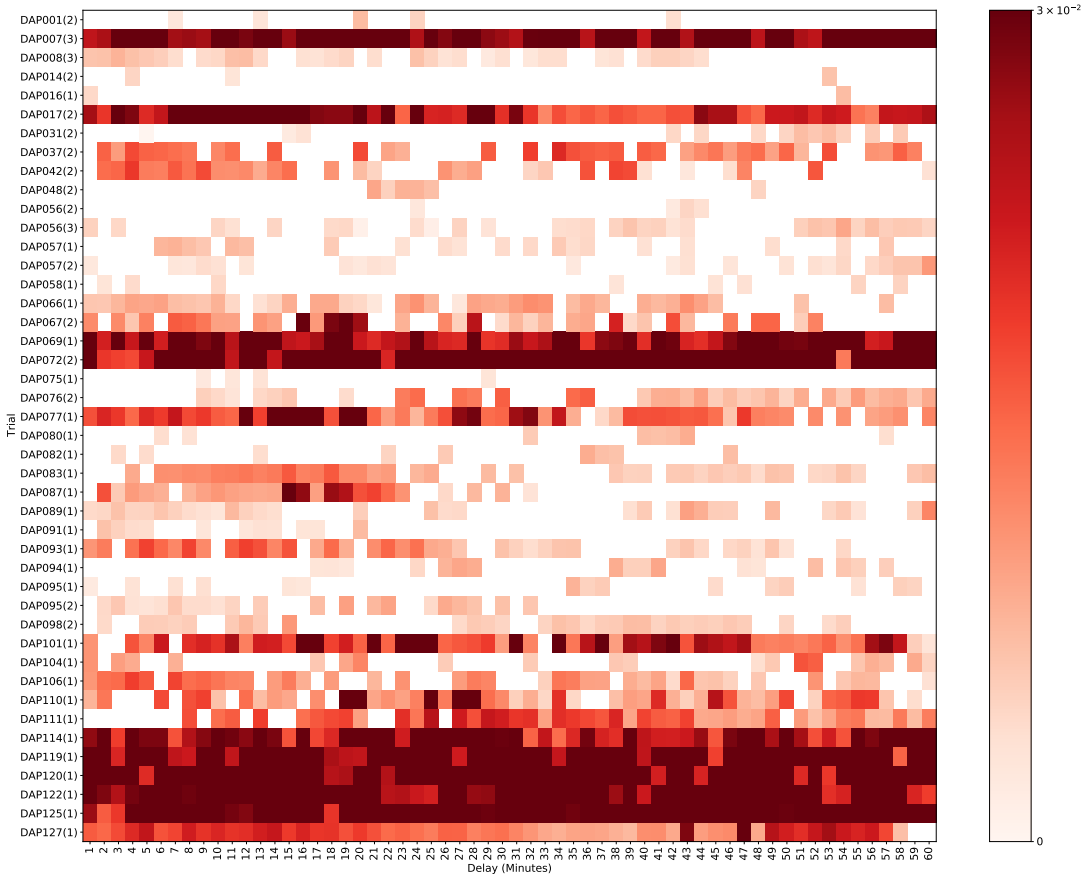

Figure D.37: Causal link intensity for non-linear PCMCI with humidity data for the 1st hour

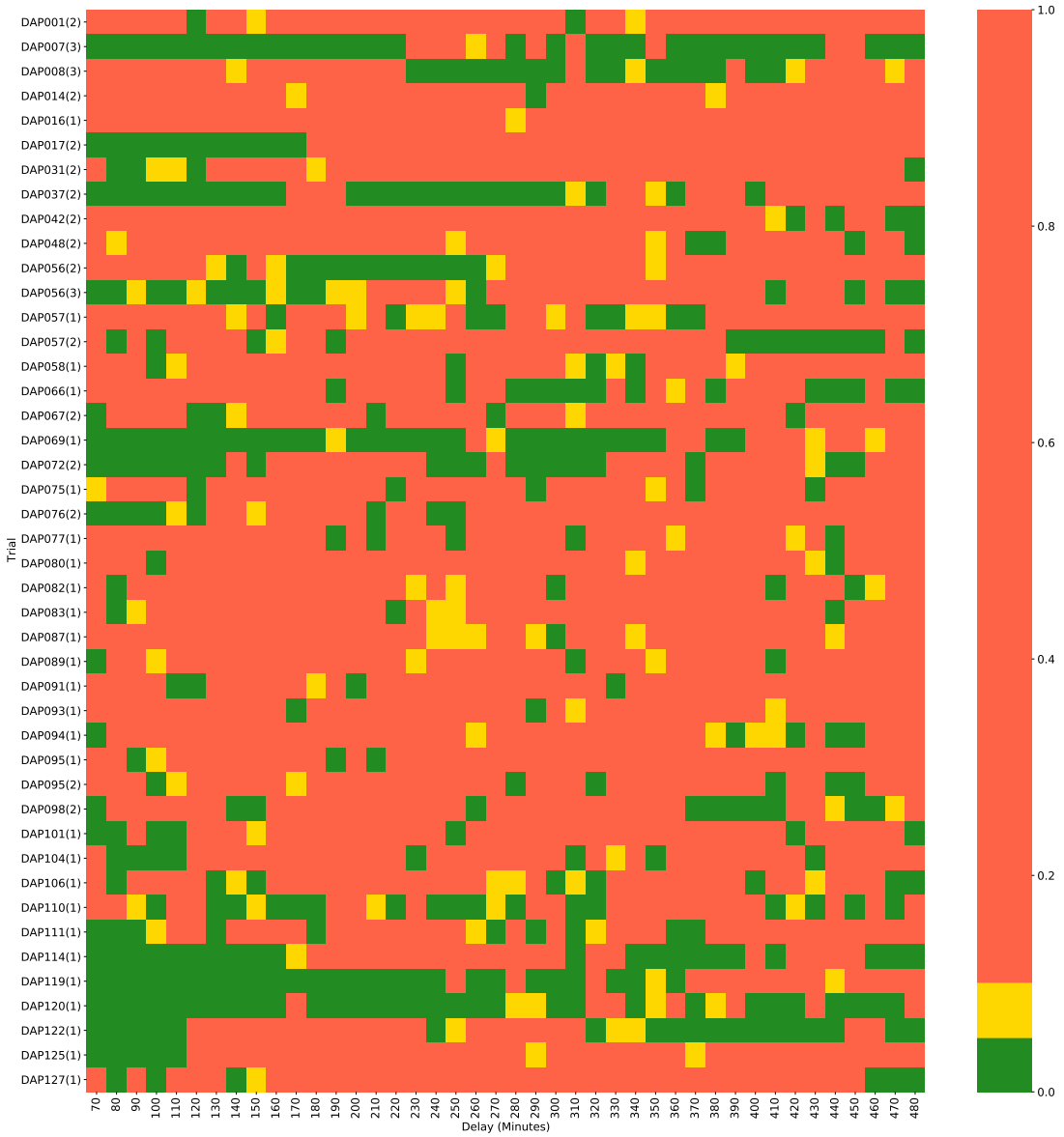

Figure D.38: Distribution of causal links for all the trials for non-linear PCMCI with humidity data from the 2nd to 8th hour

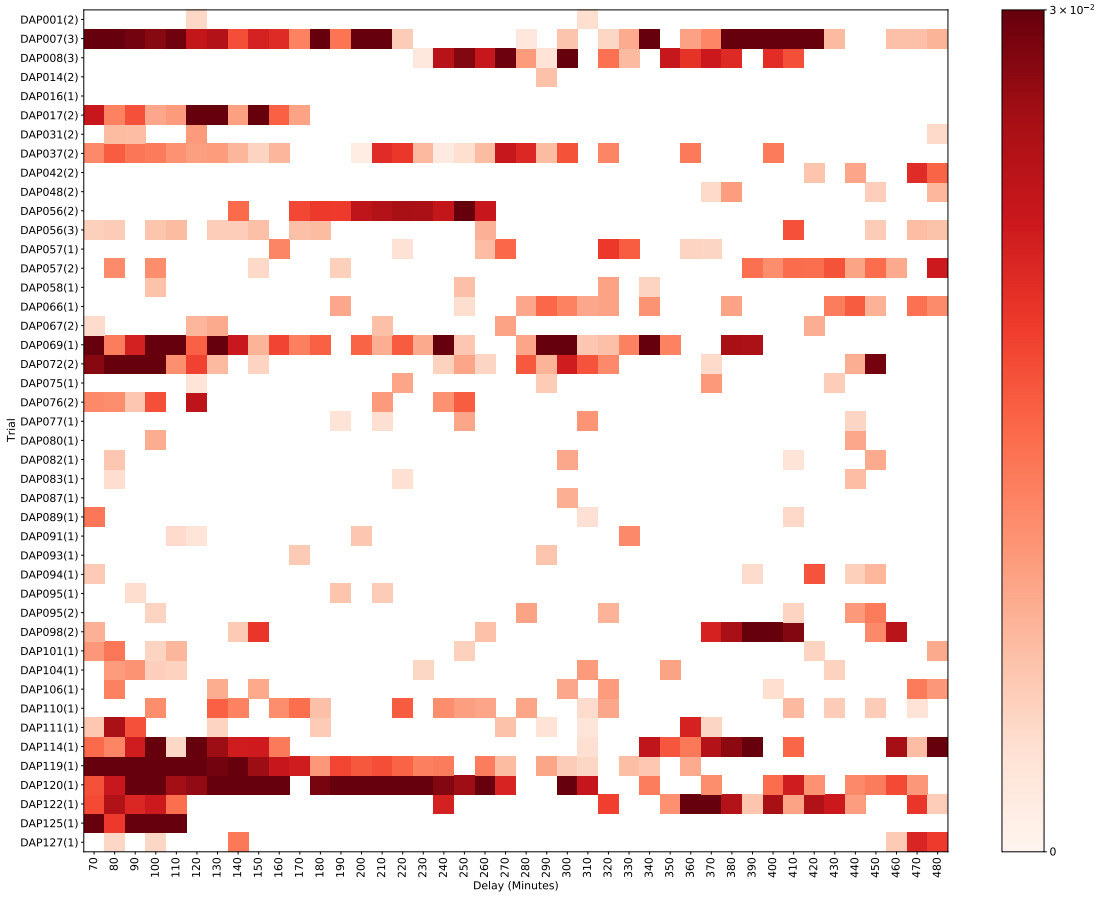

Figure D.39: Causal link intensity for non-linear PCMCI with humidity data from the 2nd to 8th hour

D.2.2.7 Personal vs static $PM_{2.5}$

The following color map shows the causal links that coincide for personal and static $PM_{2.5}$ in green, the ones only found by static $PM_{2.5}$ in blue, only for the personal data in yellow and the time lags for which no links were found in red.

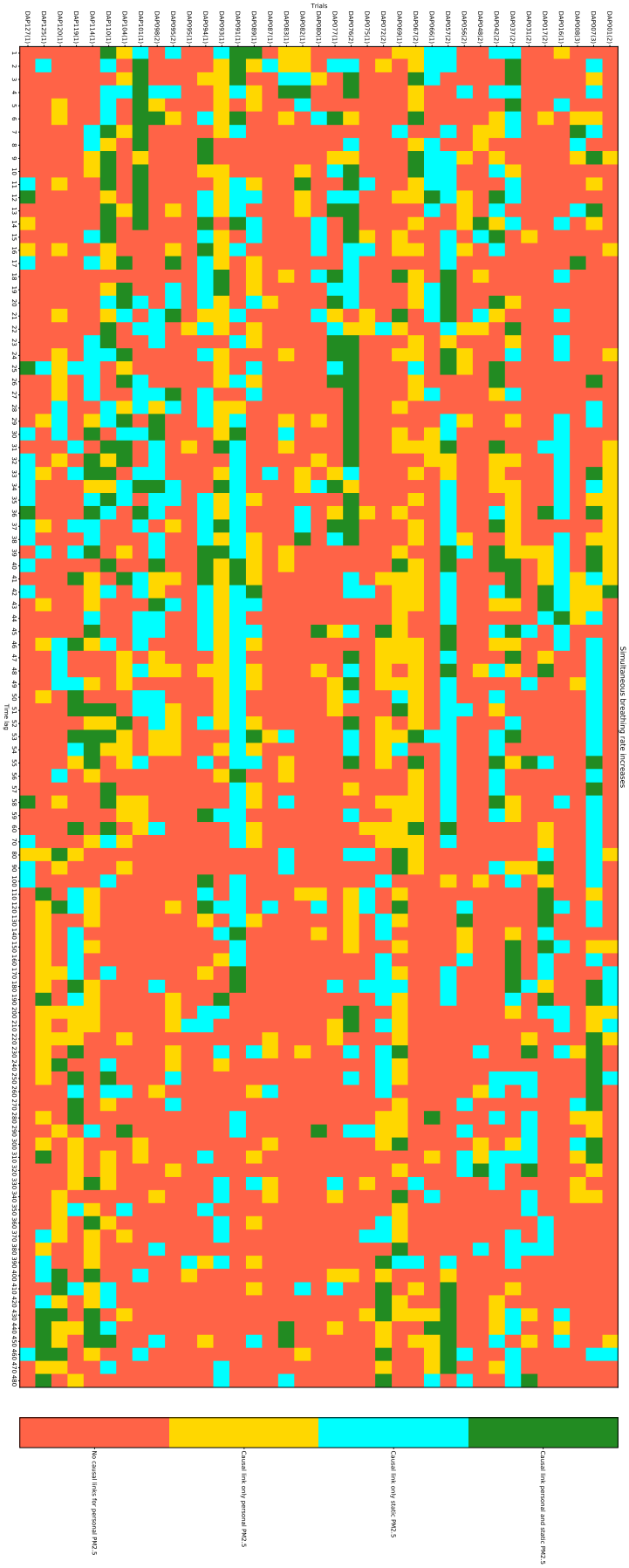

Figure D.40: Personal vs static $PM_{2.5}$ causal links

Appendix E

Exposure-response relationship

This chapter includes the plots from the experiments carried out to determine whether the breathing rate increases with the pollutant concentrations, that could not be included in the main document due to space constraints.

E.1 KNN machine learning model

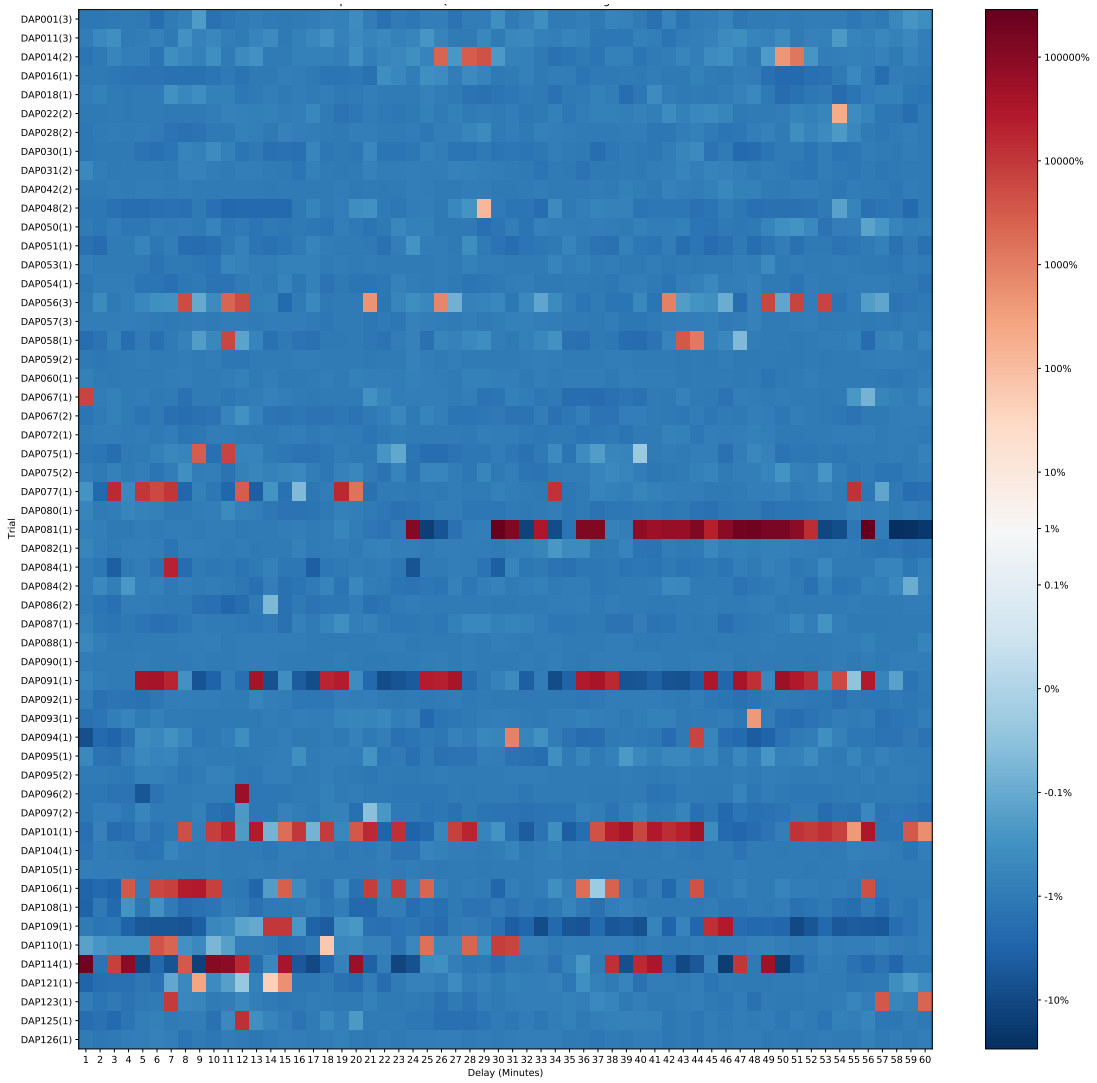

Figure E.1: Average changes in the breathing rate with an IQR increase of the $PM_{2.5}$ concentration, predicted by the KNN model.

E.2 Sliding window strategy

This section contains the results with the highest percentage of breathing rate increases with pollutant increases in the causal links. The white cells in the color maps indicate that there are no causal links there.

E.2.1 $PM_{2.5}$

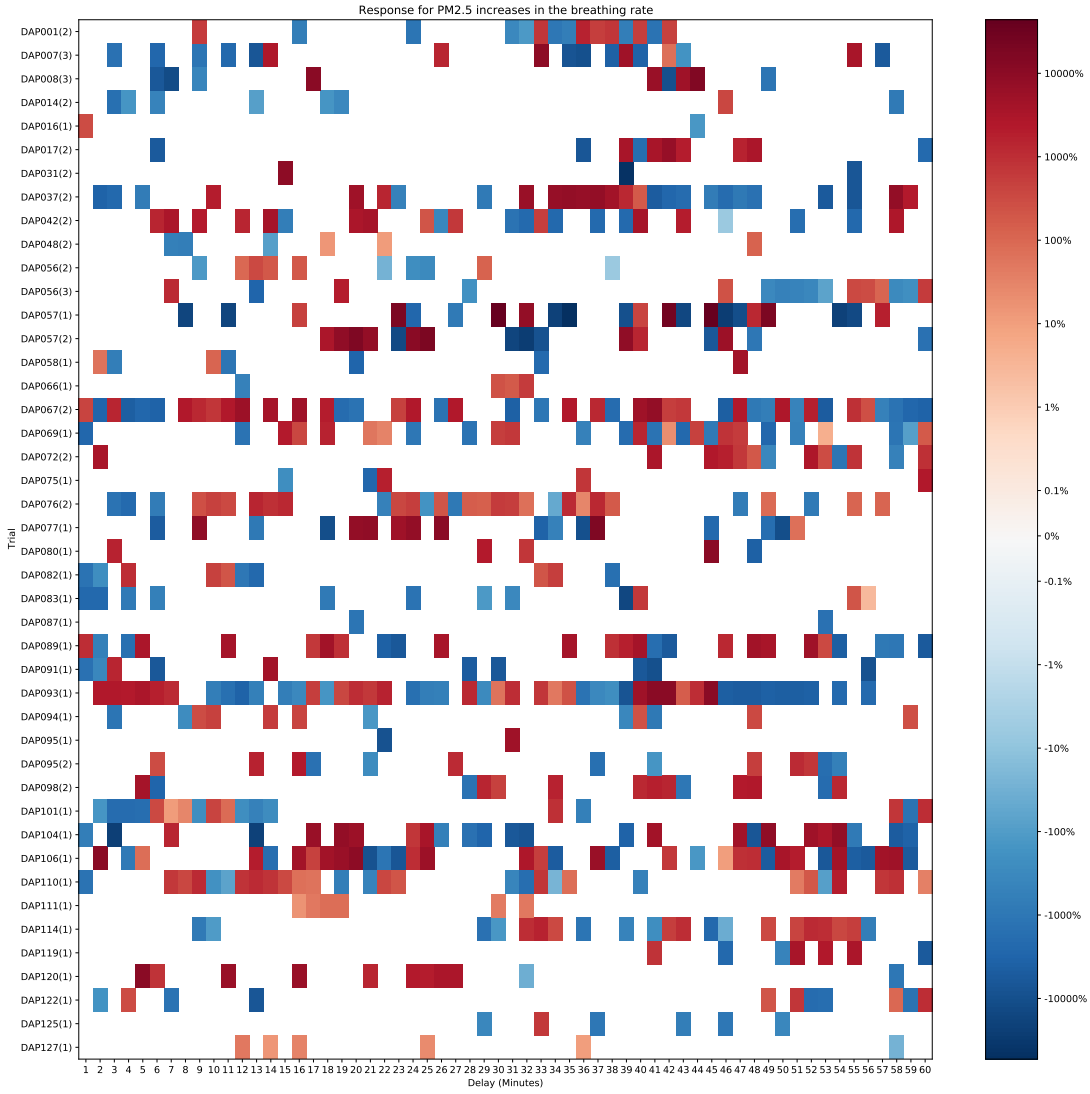

Figure E.2: Average changes in the breathing rate with an increase of the $PM_{2.5}$ concentration, with a window size of 150 minutes and a minimum increase threshold of 115% for the 1^{st} hour.

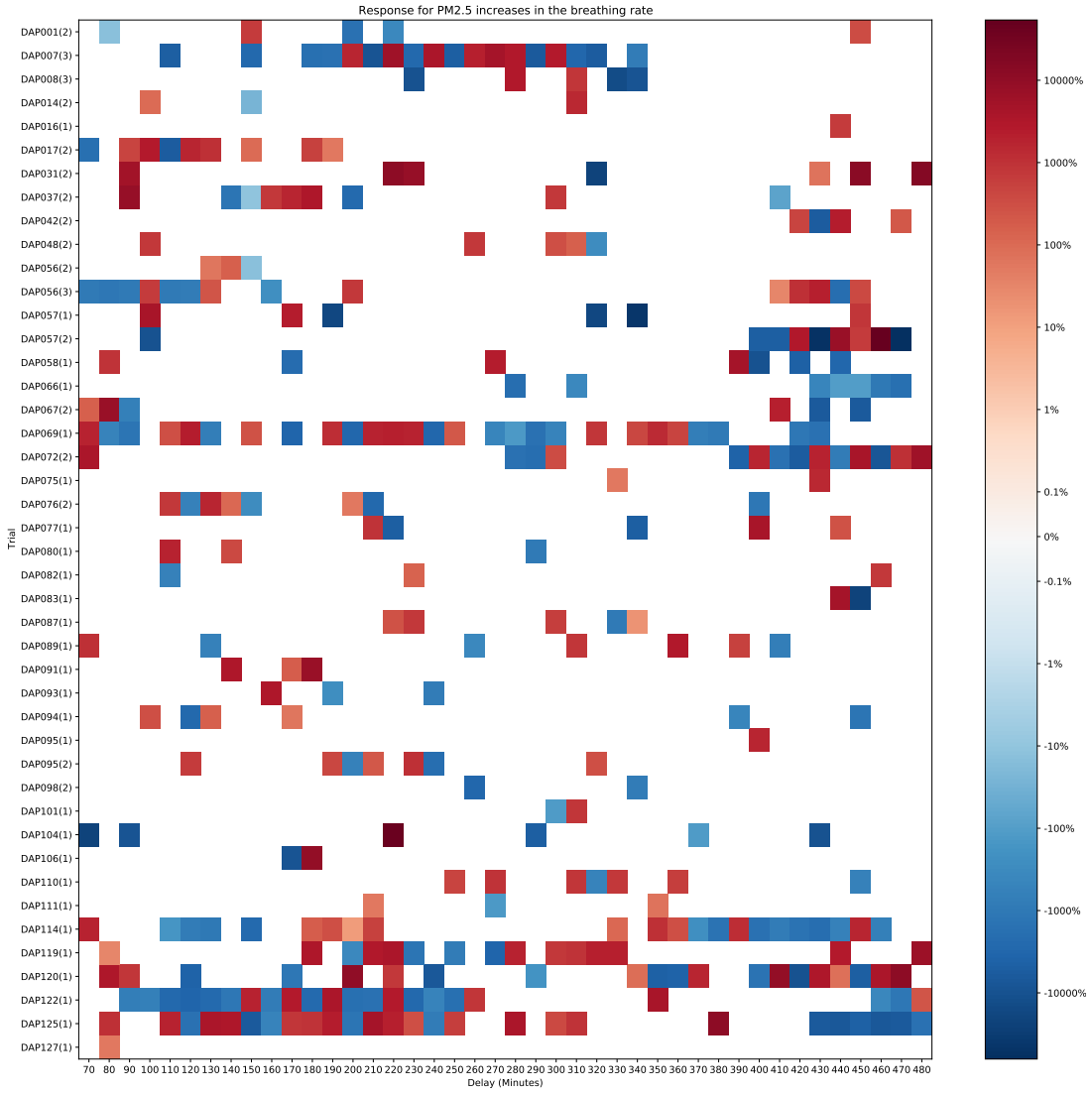

Figure E.3: Average changes in the breathing rate with an increase of the $PM_{2.5}$ concentration, with a window size of 150 minutes and a minimum increase threshold of 115% from the 2^{nd} to the 8^{th} hour.

E.2.2 NO_2

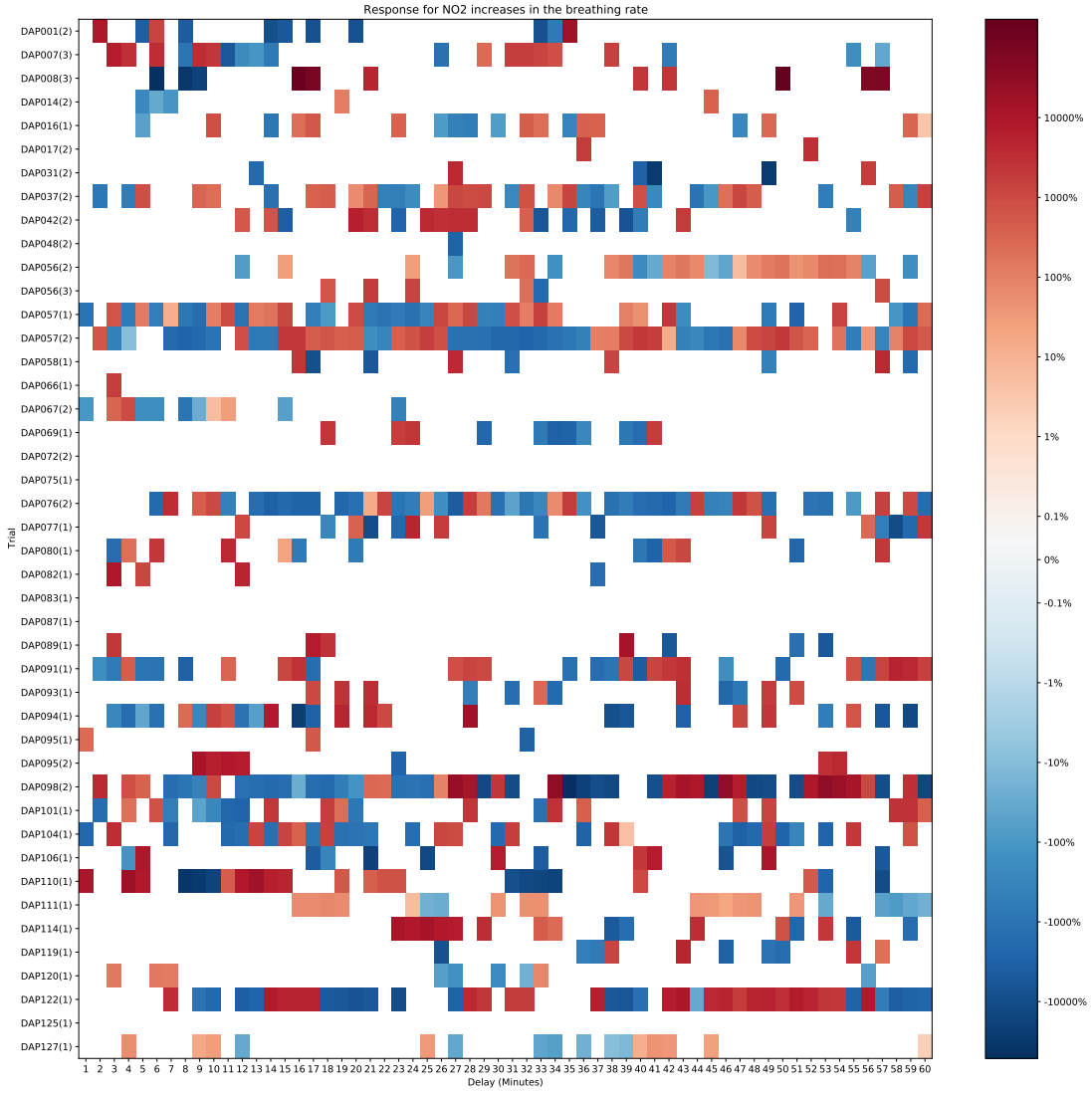

Figure E.4: Average changes in the breathing rate with an increase of the NO_2 concentration, with a window size of 240 minutes and a minimum increase threshold of 50% for the 1^{st} hour.

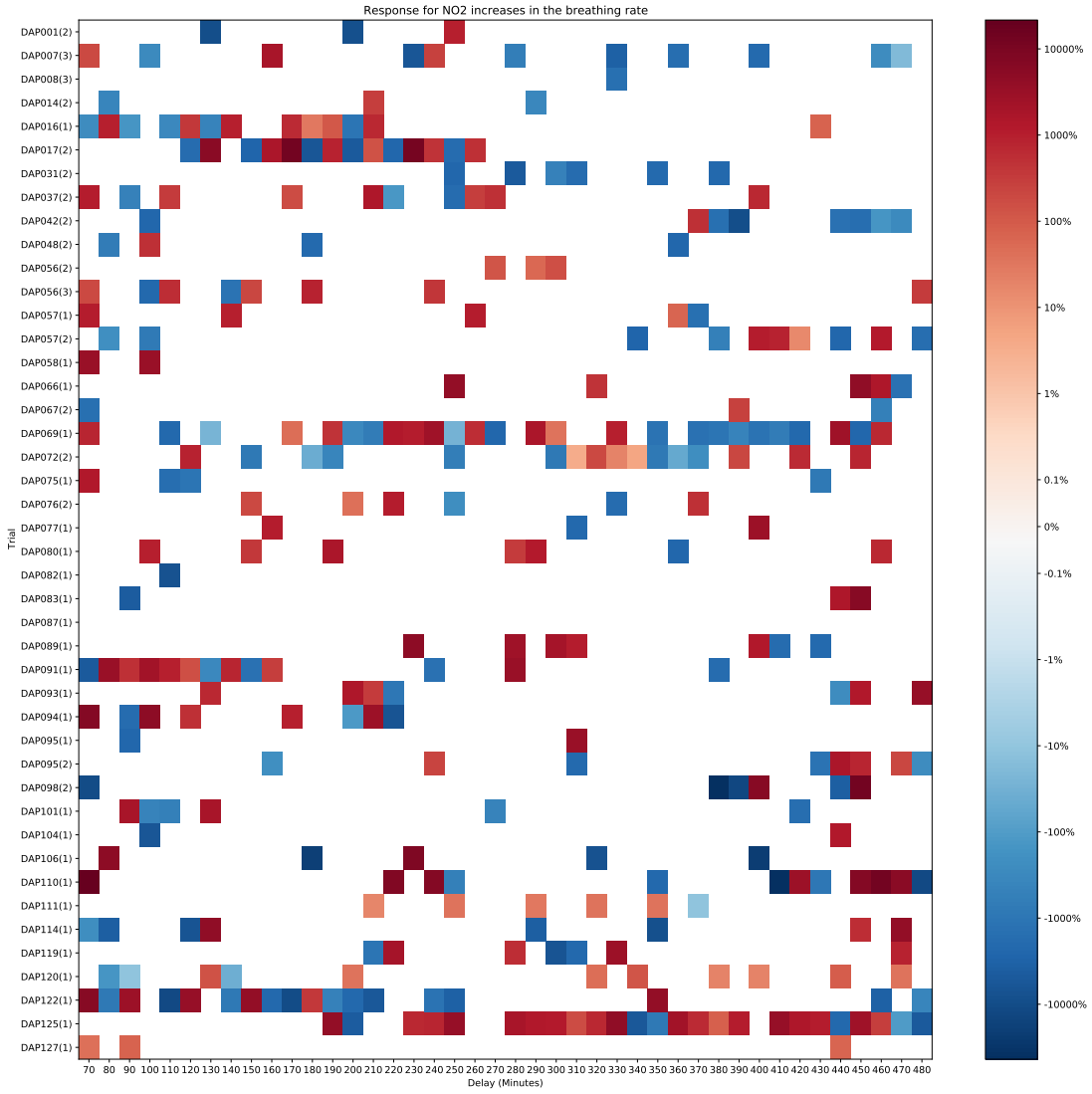

Figure E.5: Average changes in the breathing rate with an increase of the NO_2 concentration, with a window size of 240 minutes and a minimum increase threshold of 50% from the 2^{nd} to the 8^{th} hour.

E.2.3 O_3

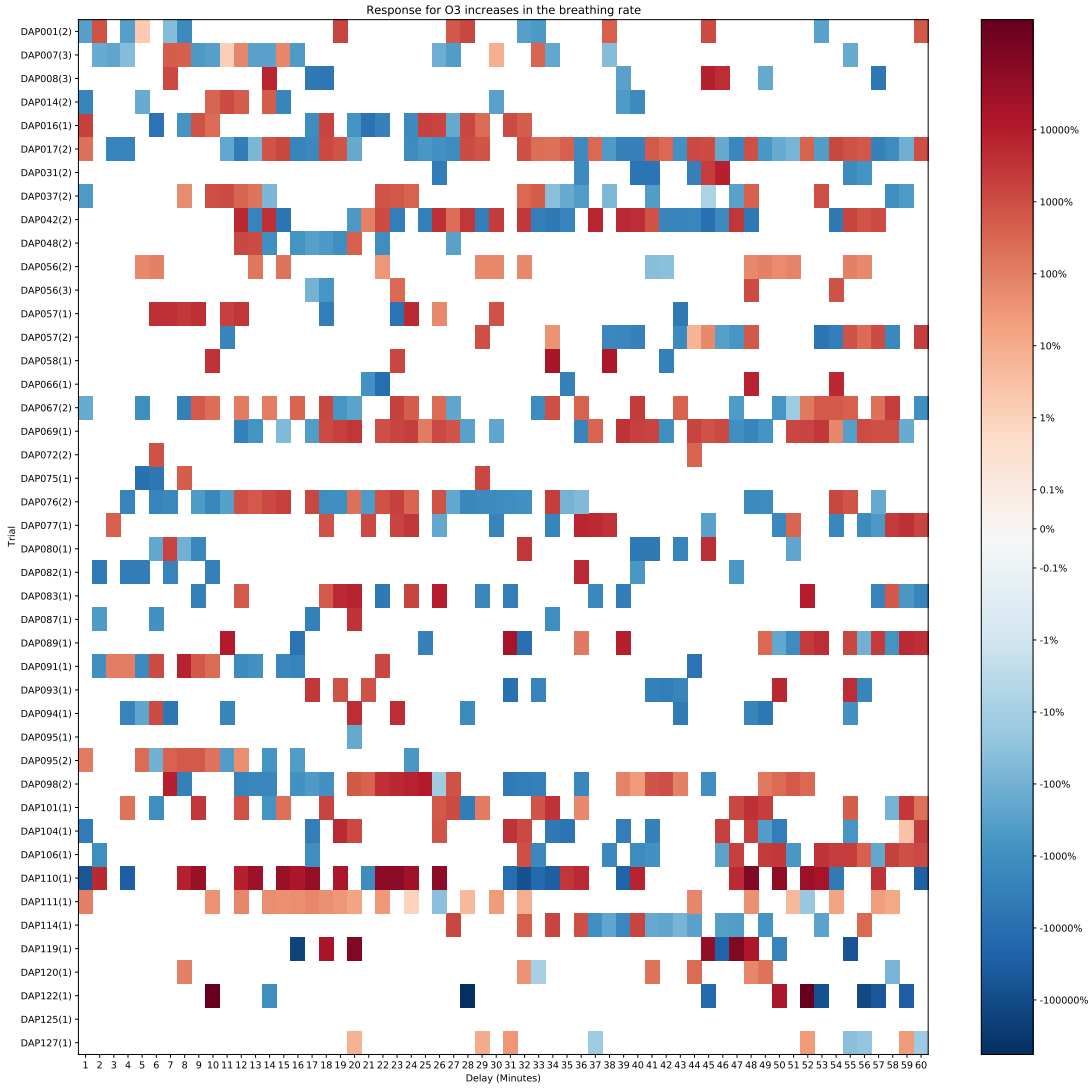

Figure E.6: Average changes in the breathing rate with an increase of the O_3 concentration, with a window size of 220 minutes and a minimum increase threshold of 110% for the 1st hour.

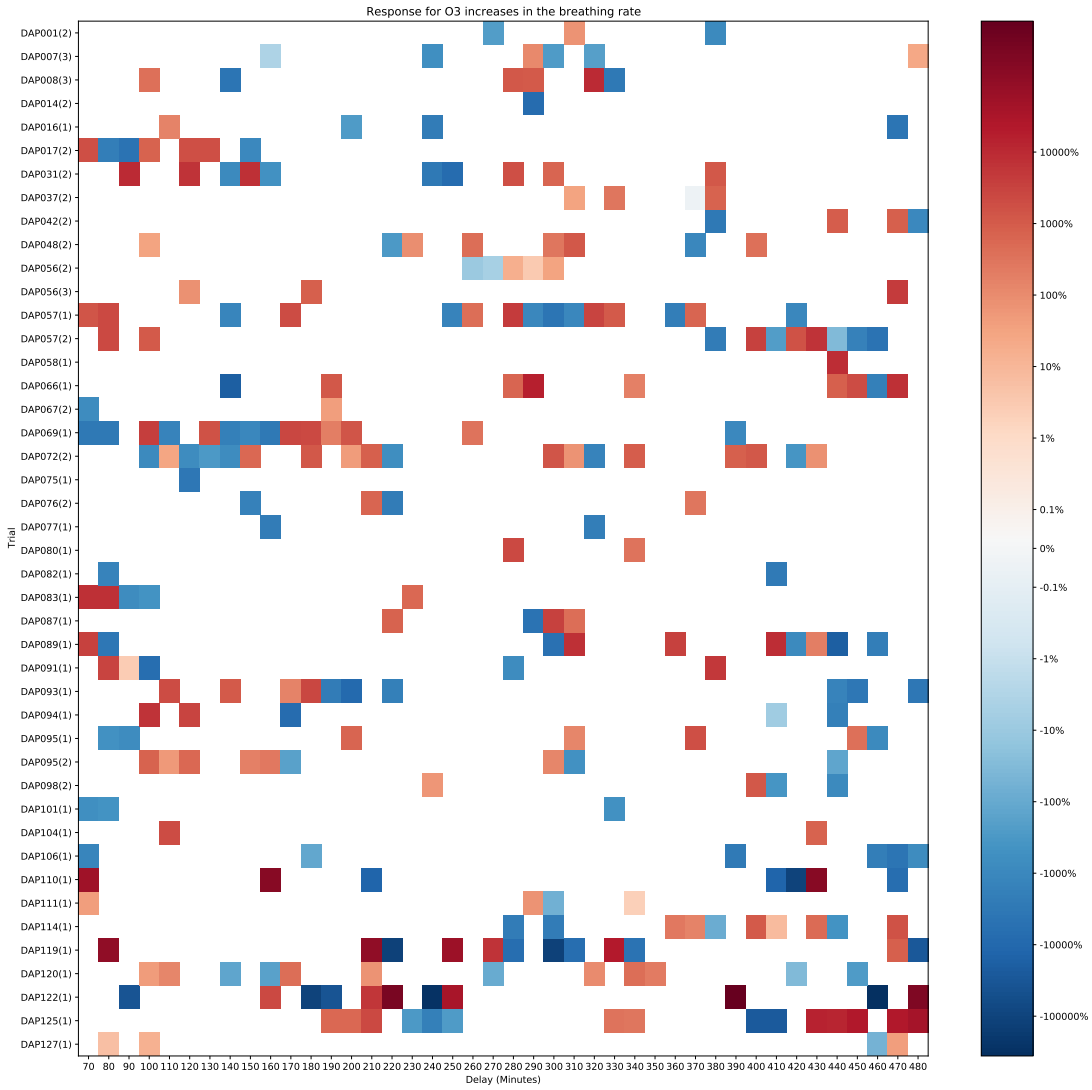

Figure E.7: Average changes in the breathing rate with an increase of the O_3 concentration, with a window size of 220 minutes and a minimum increase threshold of 110% from the 2^{nd} to the 8^{th} hour.

E.3 Breathing rate increases coincidences

The following color maps show the coincidences in the increase of the breathing rate when each of the pollutants increase. The white cells indicate that the breathing rate does not go up with any of the pollutants.

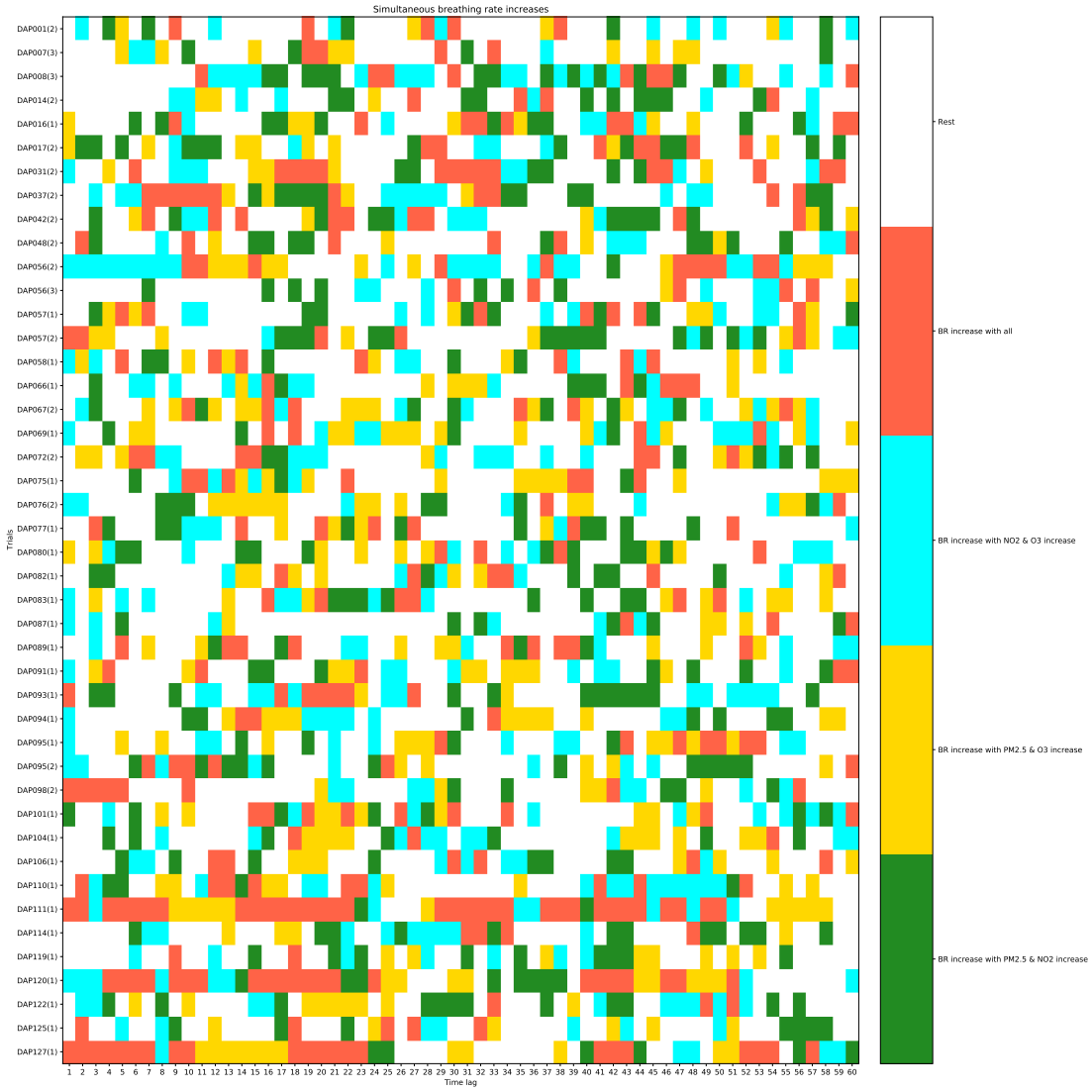

Figure E.8: Simultaneous breathing rate increases for increases of the three pollutants in the 1st hour.

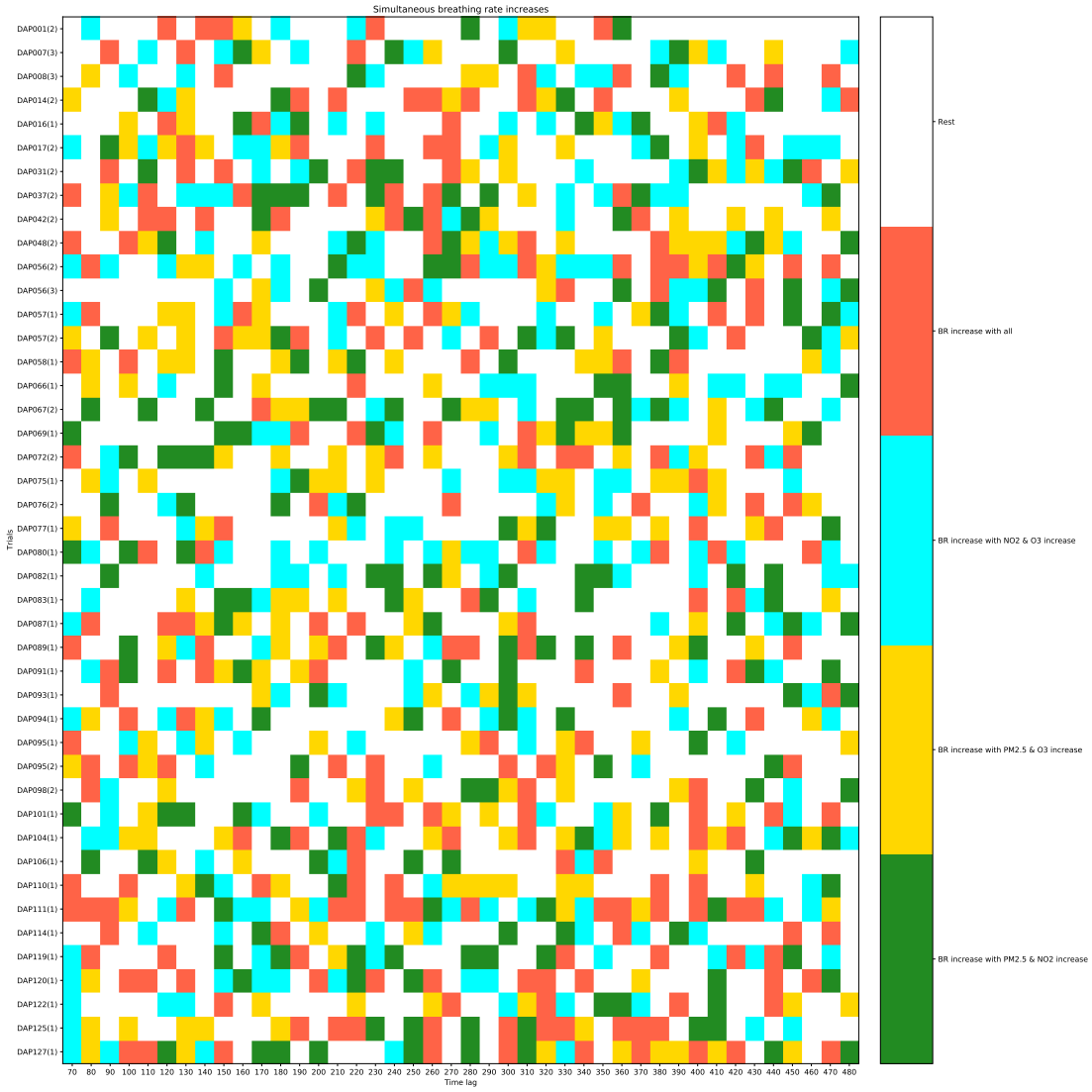

Figure E.9: Simultaneous breathing rate increases for increases of the three pollutants from the 2nd to the 8th hour.