The Impact of Factors - Size Fractions of Airborne Particles, Temperature, Relative Humidity and Activity - on Coughing Episodes in Asthmatic Adolescents

Huacheng Song

MInf Project (Part 2) Report
Master of Informatics
School of Informatics
University of Edinburgh

2023

Abstract

On World Asthma Day, May 7th, 2019, a warning was issued regarding the adverse impact of fine particulate matter in the air on individuals suffering from chronic respiratory diseases. These tiny particles have the potential to enter the lungs, exacerbating symptoms such as difficulty breathing, coughing, and chest tightness. While a multitude of studies have established a correlation between air pollution and respiratory issues, including coughing, a definitive causal relationship has not been established. This is due to the influence of various confounding factors that could also affect coughing. In this study, we utilized data from the DAPHNE and INHALE studies, focusing on asthma patients, to investigate the exposure-response relationship by analyzing PM values based on the diameter of airborne particles (0.38-17µm). The study employed state-of-the-art causal discovery methods while accounting for various complex factors, such as humidity, temperature, activity levels, and breathing rates of subjects, to build a sophisticated causal network. Our findings provide the first empirical evidence demonstrating that different size fractions of particles in the air can result in short- and long-term changes in coughing among patients, supporting a robust causal relationship.

Research Ethics Approval

The ethical approval for the DAPHNE study was granted by the Institute Ethics Committee of the All India Institute for Medical Science, Delhi (Reference numbers: IEC-256/05.05.2017, RP-26/2017, OP13/03.08.2018).

The ethics approval for the project - INHALE: Personal pollution exposure and effects on thelungs in healthy and asthmatic individuals, REC reference: 19/LO/1638, RAS project ID: 270153, was granted on 19 November 2021 by the London - Dulwich Research Ethics Committee.

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

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.

(*Huacheng Song*)

Acknowledgements

I would like to begin by expressing my deepest gratitude to my thesis advisor, Professor D K Arvind, for his unwavering support, encouragement, and guidance throughout this journey. His valuable insights, constructive feedback, and patience have been instrumental in shaping my research and enhancing my academic capabilities. Without his mentorship, this thesis would not have been possible.

I would also like to extend my sincere thanks to Sharan Maiya, who provided me with invaluable advice and assistance every week during our group meetings. His input has been instrumental in refining my research methodology and improving the overall quality of my work.

I would also like to express my thanks to my family and friends for their encouragement and understanding throughout my academic journey.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

In 2016, a study estimated that 4.4% of all deaths, over 2 million, were caused by various lower respiratory tract infections in infants and adults [44]. To prevent such a scenario from happening again in the future, one option is to propose methods that can detect potential respiratory infections early, thereby reducing the likelihood of serious complications later on.

Coughing is an important and fundamental early symptom in most respiratory infections [19], which not only causes discomfort to patients but also impacts their quality of life. Although coughing can be a symptom of many diseases, the most common causes are respiratory infections and allergic reactions. However, with the develoPMent of modern cities, air pollution is becoming increasingly severe, and more and more studies suggest that there is a close relationship between air pollution and coughing [17, 11, 6]. This indicates that coughing can be used for early detection of these respiratory diseases. However, when a person starts coughing, it is almost impossible for an ordinary person to diagnose the underlying disease at home. Unfortunately, when an infected person decides to go to the hospital for clinical testing, it may already be too late [44]. Therefore, understanding the causes of coughing and its relationship with air pollution and human activity levels is very important, not only to help doctors better understand the mechanism of coughing but also to provide important reference for the prevention and treatment of coughing. Studying the causes of coughing and its relationship with air pollution and human activity levels can help us understand how these factors interact with each other, thus improving people's quality of life.

Long-term exposure to air pollution and changes in human activity levels may lead to chronic coughing, while short-term causal relationships refer to changes in air pollution and human activity levels that may cause coughing to occur or worsen in the short term. Therefore, studying the causes of coughing and its relationship with air pollution and human activity levels is of great significance in protecting people's health and improving environmental quality. By delving into the relationship between these factors, we can propose corresponding measures to reduce the risk of coughing in people and improve their quality of life.

1.2 Minf1 Achievements Synopsis

Minf1 project aimed to investigate the exposure-response relationship between respiratory rate and airborne particles of different diameter sizes in adolescents with asthma, using data collected in the "Delhi Air Pollution: Health aNd Effects"(DAPHNE) study. To assess the causal relationship between particle numbers and respiratory rate, the newly published causal discovery method (PCMCI+) was employed [36] [35]. Results showed a robust causal relationship between all 16 different particle diameters and respiratory rate, with a maximum effect observed within 30 minutes of exposure and a second critical exposure at the 7th hour. The conclusions were confirmed by visualising and analysing other factors that affect respiratory rate, such as humidity and temperature.

1.3 Research Aims

This report extends and expands on the Minf1 project, focusing on the relationship between air pollution and health, particularly among vulnerable groups such as adolescents with asthma. The study combines data from the ongoing DAPHNE project and a new project, the INHALE study conducted in London, UK. The study employs AIRSpeck [1] and RESpeck [2] sensors, developed by the Centre for Speckled Computing at the University of Edinburgh, to measure participants' exposure to air pollution and their physiological responses. This study employs PCMCI+ to construct a network in multivariate time series data to investigate the causal relationship between cough activity in adolescent asthma patients and exposure to particles of different diameters and other potential factors. Notably, this study examines, for the first time, the long-term and short-term exposure-response relationship between 16 different diameters (ranging from 0.38-17μm) of particles and coughing in asthma patients while testing for various time lags between exposure and response.

1.4 Report Structure

This report presents a comprehensive study from data to algorithm to conclusion. The structure is as follows: **Chapter 2** provides background knowledge and previous research, analyzes limitations of previous studies, and clarifies research objectives and questions. **Chapter 3** discusses exploratory data analysis, which involves a comprehensive analysis and visualization of the data collected from the sensors worn by the participants. This analysis aims to provide a foundation and direction for subsequent modeling and analysis. **Chapter 4** covers the data cleaning process, which establishes a clean, consistent, and reliable data set for subsequent analysis. **Chapter 5** is divided into multiple sections. Firstly, causal network analysis is used to explore the causal relationship between air pollution data and coughing. Additionally, other confounding factors are analyzed, and the conclusion is validated using DAG and other methods. **Chapter 6** concludes the report by analyzing the strengths and weaknesses of this approach and discussing future improvement methods.

Chapter 2

Background

The impact of air pollution on human health has been widely researched and confirmed. According to the World Health Organization's report, over 7 million people worldwide die each year from diseases caused by air pollution, including respiratory and cardiovascular diseases such as asthma, lung cancer, coronary heart disease, and stroke [32]. Moreover, research has found that air pollutants have a significant impact on the health of asthma patients. A study found that air pollutants such as PM2.5 can exacerbate symptoms in asthma patients, including coughing, shortness of breath, and chest tightness [25]. Another study found that long-term exposure to air pollutants can lead to a decline in lung function in asthma patients, exacerbating symptoms such as coughing [34]. Although some literature has investigated the effects of particulate matter on health, there are still issues such as confounding factors and inappropriate data that prevent determining causality. Furthermore, there is still a research gap on the short-term effects of air pollution particles on asthma patients' conditions, especially cough symptoms. However, in this study, we have filled this gap and investigated the effects of air pollution particles of different sizes on asthma patients' cough symptoms.

Chronic cough is a common disease in primary and secondary healthcare. As of March 2023, a recent study has developed the Leicester Cough Monitor (LCM), which measures cough frequency and is an automated dynamic cough monitoring device based on sound [3]. The study compared the coughs and other sounds manually counted by two observers during a six-hour recording of nine chronic cough patients in a two-hour period to determine the LCM's sensitivity and specificity. However, this method still has limitations. One potential criticism of the cough count obtained from recordings is that they may not accurately reflect the actual cough rate since cough behavior cannot be visually observed, and lack of well-validated outcome measures impedes evaluation and management.

A recent study from 2018 investigated the role of oxidative stress in air particulate matter-induced lung diseases, including asthma [26]. The article pointed out that chemicals in air particulate matter can cause oxidative stress, leading to inflammation and airway constriction, exacerbating cough symptoms in asthma patients. However, it mainly discusses the role of oxidative stress in air particulate matter-induced lung diseases and does not extensively study the relationship between cough symptoms in asthma patients and air pollution. Moreover, the literature mainly focuses on the toxicity of engineered nanoparticles and does not consider the impact of other air pollutants on asthma patients' cough symptoms. Thus, combining other studies is necessary to comprehensively consider the relationship between air pollution and cough symptoms in asthma patients.

A prospective study investigated the relationship between daily symptoms and fine particulate matter (PM) among individuals with chronic obstructive pulmonary disease (COPD) in Guangzhou, China [28]. Results revealed that PM had a significant impact on COPD symptoms, such as coughing and shortness of breath. However, the study has some limitations. Firstly, self-reported symptom surveys were used, which may be subject to recall bias or subjective judgment errors, affecting the accuracy of the study results. Secondly, the study used a cross-sectional design, which only considered the impact of PM on coughing and other symptoms, without taking into account other possible pollutants or environmental factors. Therefore, further research is needed to validate the study results in a wider research context.

In 2019, another study examined daily mortality rates and air pollution in 652 cities across 24 countries or regions, discovering an independent association between short-term exposure to PM_{10} and $PM_{2.5}$ and respiratory symptoms, including coughing [27]. However, as an observational study, it cannot prove air pollution as the cause of increased symptoms without conducting experiments or randomized controlled trials to establish a causal relationship. Furthermore, the study only considered PM concentration's impact on respiratory diseases, without accounting for other confounding factors such as physical activity levels. Thus, the study provides preliminary evidence, and further research is necessary to determine a causal relationship.

The fourth reference is a multicenter study that examines the link between air particulate matter exposure and health outcomes [37]. Data from 14 European cities over a 10-year span was included. The study found a positive correlation between air particulate matter concentration and the risk of asthma and coughing. A 5% increase in the incidence of asthma and coughing was observed for every 10 micrograms/cubic meter increase in PM concentration. However, the study did not account for pollution sources or individual exposure differences within the cities, and did not investigate short- and long-term exposure effects. Therefore, further research is necessary to comprehensively assess the relationship between air pollution and coughing in asthma patients while considering various factors.

To address the aforementioned issues, this project extends previous research in several ways. Firstly, the PM values were refined into 16 different bin sizes, and all influencing factors were studied within the same causal network. The experiment data was collected from two different research locations, and the temporal resolution of all subjects' data was reduced to a minute to ensure maximum accuracy. The latest causal algorithm, which considers factors beyond pollutant particles, such as temperature, humidity, breathing rate, and activity level, was applied to comprehensively analyze their causal relationships with coughing. Overall, the study significantly advances the understanding of the causal factors contributing to coughing by considering various relevant factors and utilizing a refined PM value system.

Chapter 3

Exploratory Data Analysis

Exploratory data analysis (EDA) is a fundamental task in data analysis. It involves the exploration of existing data, especially raw data from surveys or observations, with as few priori assumptions as possible. EDA is a means of identifying patterns and structures in data through graphing, tabulating, equation fitting, and calculating characteristic quantities, among other techniques [18] [45].

This report presents the results of data tracking conducted for two projects: DAPHNE study and INHALE study. Both studies utilized the same sensors, AIRSpeck Personal, which records personal exposure to airborne particulates, and RESpeck, which records various parameters such as respiratory rate, flow/effort, and the intensity and type of physical activity. Detailed description for both sensors listed in Appendix A.2. The DAPHNE study involved 127 adolescents with asthma, generating 222[1] different trials, while the INHALE study recorded 28 older individuals with asthma and control and 30 trials recorded by 15 of the asthmatics are used here. Each trial produced a time series of both sensors data at a time resolution of 1 minute. This chapter provides a detailed analysis of the raw data recorded by the two sensors.

It should be noted that the aim of this chapter is to identify outliers in the raw observations that fall outside the reasonable range of the experiment, as well as to make reasonable inferences about the origin of these outliers. Pre-processing and calibration of the data will be conducted in the next chapter. This chapter aims to develop an understanding of the distribution of complex data from the two studies and provide a comprehensive conclusion.

3.1 AIRSpeck Data

The DAPHNE and INHALE studies utilized data on air pollution obtained from different AIRSpeck sensors, which are categorized into two types: personal and stationary [1]. The personal sensor is specifically designed to be wearable and records air quality

[1] A total of 222 visits were recorded across three AAP visits, with 137 visits on the first, 72 visits on the second, and 13 visits on the third visit.

data, such as PM_1 / $PM_{2.5}$ / PM_{10}, at 30-second interval which is presented as minute-level averages, whereas the AIRSpeck Stationary records data at time intervals ranging between 5 minutes to 30 minutes depending on the season. This type of sensor records information on air quality at the current location and is characterized in Tab 3.1.

Feature	Explanation
Timestamp	UTC timestamp of current observation.
Temperature	The uncalibrated temperature inside the sensor case (°C).
Humidity	The uncalibrated humidity level inside the sensor case.
PMX	The mass of all particles below a size of X µm inside each cubic meter.
bin0-bin15	The count of particles of a certain size range, each bin is mapped to a size range.
Latitude	GPS latitude coordinates.
Longitude	GPS longitude coordinates.
Battery	The battery level. Higher levels mean higher charge of battery.

Table 3.1: Explanation of features measured by AIRSpeck sensors [40].

PMX	Related bin	Particle size(µm)	Related bin	Particle size(µm)
PM_1	**bin0**	0.38 - 0.52	**bin1**	0.52 - 0.75
	bin2	0.75 - 1.0		
$PM_{2.5}$	**bin3**	1.0 - 1.3	**bin4**	1.3 - 1.5
	bin5	1.5 - 2.0	**bin6**	2.0 - 3.0
PM_{10}	**bin7**	3.0 - 4.0	**bin8**	4.0 - 5.0
	bin9	5.0 - 6.5	**bin10**	6.5 - 8.0
	bin11	8.0 - 10.0	**bin12**	10.0 - 12.0
	bin13	12.0 - 14.0	**bin14**	14.0 - 16.0
	bin15	16.0 - Max		

Table 3.2: Particle size distribution mapped to each bin.

Calibration of PM sensors typically consists of two components: zero calibration and span calibration [7]. The former accounts for any offset in the sensor readings, while the latter ensures that the sensor response is consistent across a specified range. However, inconsistencies in calibration standards across companies and regions can lead to errors in the readings. To mitigate this, the original bin values are used, as airborne particulate matter is a fixed value that does not require calibration. This allows subsequent experiments with PM values to be validated. Tab 3.1 stores the number of particles of different sizes in the air, which are accurately measured by particle sensors[2] capable of providing high-quality particle counts and sizes [39]. This sensor records particle diameters ranging from 0.38 to 17µm, utilizing 16 bins to record different sizes, with each bin corresponding to a specific size range. Bin 0 records the smallest particle

[2] Both the DAPHNE and INHALE studies utilize the OPC-R2 particle sensor model. This sensor series provides exceptional performance in a compact device that measures only 70 mm in length and 21 mm in diameter, while also being highly cost-effective. These sensors are widely used in commercial applications within heavily polluted urban environments, and their use in industrial applications is also increasing [39].

diameters, while bin 15 records the largest. A detailed description of this particle sensor is placed in Appendix A.1 for reference.

Tab 3.2 presents the size range associated with each bin, and the PM values are calibrated to these original bin values. For instance, PM_1 is derived from the counts in bins 0, 1, and 2. Of particular note is bin 6, which spans particle sizes from 2.0 to 3.0 µm. This bin encompasses both the $PM_{2.5}$ and PM_{10} size intervals. Previous research has investigated the differences between $PM_{2.5}$ and PM_3, and has found that $PM_{2.5}$ typically contains higher concentrations of organic and elemental carbon, while PM_3 contains higher concentrations of metals such as iron and aluminum. These particles originate from similar sources, including traffic, industry, and dust [23]. Additionally, the World Health Organization reported in 2005 that particles measuring 3.0 µm in size can penetrate deep into the lungs and cause adverse health effects [31]. Therefore, in this study, bin 6 is considered part of the $PM_{2.5}$ category.

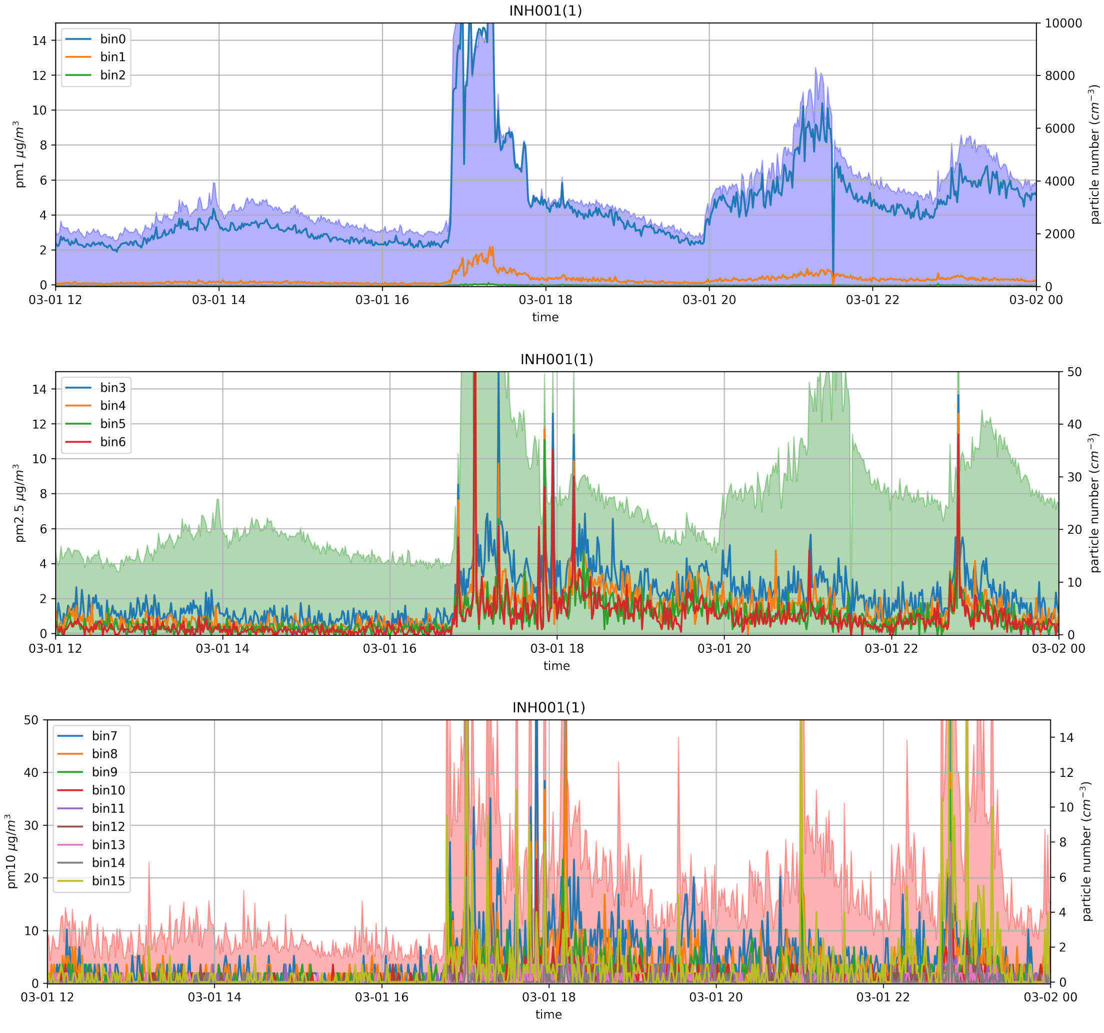


Figure 3.1: Relationships between the three different PM values and the corresponding bin values for trial INH001(1).

Fig 3.1 illustrates the relationship between the three different PM values and the corresponding bin values for trial INH001(1). The figure comprises three subplots, showing PM_1 in blue, $PM_{2.5}$ in green, and PM_{10} in red. For presentation purposes, only the period from 12:00 noon on 1 March until 00:00 the following day, which is the same for all three subplots, is included. It is important to note that the bins corresponding to $PM_{2.5}$ and PM_{10} are 0 to 6 and 0 to 15, respectively. The intersection bins that correspond to the three PM values are omitted in the figure for clarity.

The trends in all three subplots reveal that the PM values and corresponding bins exhibit similar peaks and turning points. Moreover, the trends of the individual bins in the different subplots are also essentially the same, except for the scale. These observations suggest that there is a strong linear relationship not only between PM values and bins but also between the bins themselves. This relationship was found in a total of 210 trials in the DAPHNE and INHALE studies, and it was later verified through kendall rank correlation experiments.

Feature	**Max**	**Min**	**Mean**	**Median**	**Var**	**Std**	**Skewness**[1]	**Kurtosis**[2]	**Autocorr**[3]	**PACF**[4]
Tem	42.5	10.3	30.2	30.5	20.6	4.5	-0.53	2.02	0.99	-9.99e-03
Hum	51.3	12.7	23.5	22.6	25.8	5.1	0.92	1.87	0.99	0.02
PM_1	164.8	4.1e-03	2.7	14.4	1.7	3.7	9.59	228.8	0.88	0.04
bin0	32799	1.0	1381	773	3.7e+06	1.9e+03	5.15	41.2	0.96	-0.02
bin1	20219	0.53	198.8	136.5	2.4e+06	489.3	16.78	401.2	0.87	-0.03
bin2	7187	0.52	31.4	23.0	6743	82.1	46.0	3274.8	0.66	0.01
$PM_{2.5}$	516.0	4.3e-03	5.1	3.6	49.6	7.04	26.4	1540	0.75	2.60e+03
bin3	1036	0.51	10.4	8.0	357	18.8	23.2	864.0	0.87	-0.09
bin4	550	0.51	6.22	4.5	106	10.3	18.7	659.7	0.84	-0.01
bin5	256	0.33	4.06	3.0	38.3	6.18	12.2	311.4	0.77	-0.01
bin6	155	0.01	3.55	2.0	27.4	5.23	8.69	165.3	0.68	0.03
PM_{10}	566.4	4.3e-03	16.4	11.3	475.5	21.8	7.59	110.5	0.61	3.61e+03
bin7	117	0.01	1.77	1.0	8.52	2.91	9.91	238.0	0.56	0.03
bin8	66	0.01	0.98	0.5	3.13	1.77	10.0	240.3	0.48	0.03
bin9	46	0.01	0.54	0.0	1.34	1.15	10.4	251.8	0.40	0.04
bin10	25	0	0.37	0.0	0.63	0.79	8.15	149.0	0.40	0.05
bin11	22	0	0.24	0.0	0.32	0.56	7.97	167.9	0.33	0.04
bin12	10	0.16	0.14	0.0	0.18	0.43	6.22	67.1	0.28	0.03
bin13	10	0	0.15	0.0	0.16	0.40	6.2	76.9	0.30	0.03
bin14	10	0.16	0.09	0.0	0.12	0.34	7.83	110.9	0.21	0.02
bin15	85	0.25	0.91	0.0	5.60	2.38	11.4	242.9	0.46	0.02

[1] Skewness refers to the degree of distortion or asymmetry exhibited by a set of data from a symmetrical bell curve, also known as a normal distribution [9].

[2] kurtosis is a statistical measure used to describe the extent to which scores tend to cluster either in the tails or in the peak of a frequency distribution [29].

[3] Autocorrelation refers to the degree to which a time series is correlated with its own past values, which it is a measure of the similarity between observations at different time points within the same series [4].

[4] The Partial Autocorrelation Function (PACF) is a statistical tool used to measure the correlation between a time series and its lagged values, while controlling for the influence of intermediate time points [8].

Table 3.3: AIRSpeck data statistical metrics from INHALE study.

Fig 3.1 displays significant fluctuations in both PM and bin values, with a sharp increase in the number of polluting particles observed at 17:00. Within one hour, the bin0 value rose from approximately 2,000 to over 10,000, which is a fivefold increase. To avoid any potential subjectivity bias and accurately interpret the data's real changes and trends, it is necessary to perform statistical measurements on all features of the AIRSpeck device. These results are fully documented in Tab 3.3. Since the particle sensor's maximum detectable value is 65,535[3], and its default value is typically 0, these extreme values

[3] 65,535 is the maximum value of a 16-bit binary number and is typically used in digital signal

have been filtered out in advance to ensure the reliability of the analysis.

According to Tab 3.3, the mean value of bin0 for all trials in the INHALE study was 1,381, while its nearest neighbour, bin1, had a mean value of 198.8, a difference of approximately sevenfold. Although the reason for this discrepancy is not yet fully understood, it is reasonable to suggest that it may be due to the significant presence of water molecules in the air, which increases with humidity. The size of water molecules is approximately 0.4 µm, which falls within the bin0 range.

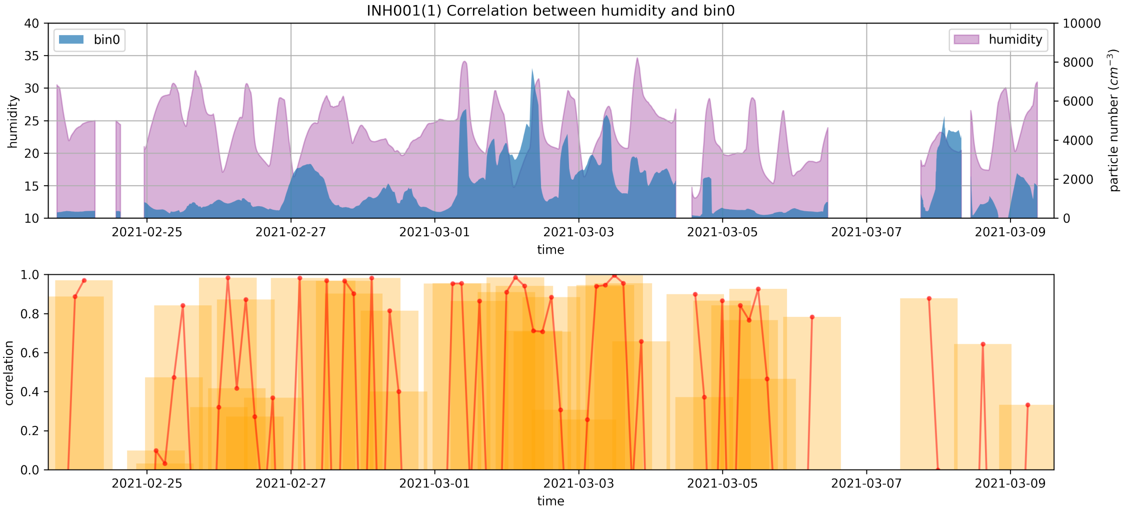


Figure 3.2: Correlations between humidity and bin0 for trial INH001(1). Noted correlations are computed by every 3 hours by rolling mean smoothing technique.

Fortunately, this hypothesis can be confirmed since the AIRSpeck device collected humidity data as well. Fig 3.2 illustrates the relationship between bin0 and humidity. The top graph displays the same-time variations of humidity (purple) and bin0 (blue) over the 24-hour duration of INH001(1), demonstrating a relatively close relationship between the two trends. To compute the correlation between the two, it is necessary to filter out noise and random variation in both datasets. The smoothing technique of rolling mean was applied in this study, which calculates the average of the data over a sliding time window, chosen to be every three[4] hours. The yellow bar plot below shows the linear correlation corresponding to bin0 and humidity, revealing that the correlation coefficients for both are greater than 0.5 over a considerable period. This result explains why the bin0 quantity is much greater than the other bins to a large extent.

Tab 3.3 displays the statistical distribution of each bin under the INHALE study. The results reveal that as the particle diameter increases, the number of particles in the air with the corresponding diameter decreases, and the smaller diameter bins typically exhibit higher variance. For instance, bin 0 has a kurtosis of 41.2, a median of 773,

processing when converting analogue signals to digital signals using a 16-bit analogue-to-digital converter (ADC).

[4] A rolling mean window of every three hours is a common choice for time series analysis as it balances capturing the trend while minimizing noise. It reduces the influence of individual data points affected by errors and allows for the detection of short-term changes with reasonable temporal resolution.

a maximum value of 32,799, and a minimum value of only 1. Such extreme values appear to be inconsistent with objective reality, raising serious suspicions that the data are highly asymmetric. To investigate further, Fig 3.3 provides boxplots that compare the data distribution for all 16 bins. The plot confirms that the excessive kurtosis metric for the smaller diameter bins is attributable to a high degree of data asymmetry caused by a large number of extreme values. Despite the calibration of the three PM values, the presence of these outliers suggests that the data in the AIRSpeck observations may contain anomalies. In conclusion, the data analysis highlights clear anomalies that require cautious interpretation and a reasonable judgement of the data range. A detailed description of the data pre-processing strategies to address these anomalies will be presented in the subsequent chapter.

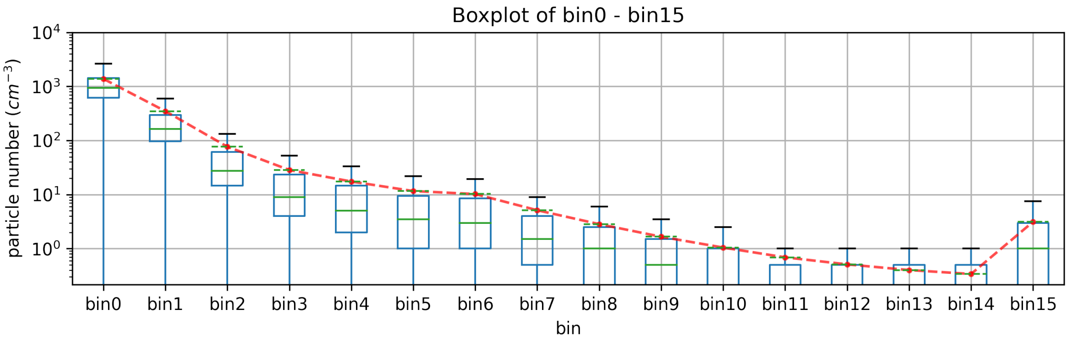


Figure 3.3: Boxplots for all 16 bins, the red line shows how the mean value changes as the particle diameter increases. Note that the y-axis has been set to log scale to eliminate the undesirable effect of large data differences.

Fig 3.4 displays the distribution of three PM observations. The x-axis shows the mass of all particles on a log-scale[5], while the y-axis represents the corresponding density. The PM_1 distribution is relatively concentrated, followed by $PM_{2.5}$, while PM_{10} is the most widely distributed. However, all three PM values exhibit some degree of overlap since PM_{10} includes all particles less than or equal to 10 µm in size, including $PM_{2.5}$ and PM_1. Similarly, $PM_{2.5}$ includes all particles less than or equal to 1 µm in size, including PM_1, which could contribute to a more extended range of concentrations. Additionally, PM_{10} particles can originate from both primary and secondary sources, whereas $PM_{2.5}$ particles are predominantly formed through secondary processes such as gas-to-particle conversion and atmospheric reactions [13] [33]. The Gaussian kernel density estimate line reveals that each of the three PM values has a distinct peak and variance, enabling differentiation through the application of Gaussian Mixture Model (GMM).

To examine the variations in PM levels across different trials and locations, this study selected two trials from each of the first two subjects in the INHALE study for comparison, as illustrated in Fig 3.5. The figure shows the distribution of PM at different trials, where blue corresponds to PM_1, green to $PM_{2.5}$, and red to PM_{10}. As shown in the figure, the distribution of PM levels varies across different trials, depending on the

[5] In order to correctly transform the data range to log-scale, the mode value 0 in the three PM values has been removed in advance.

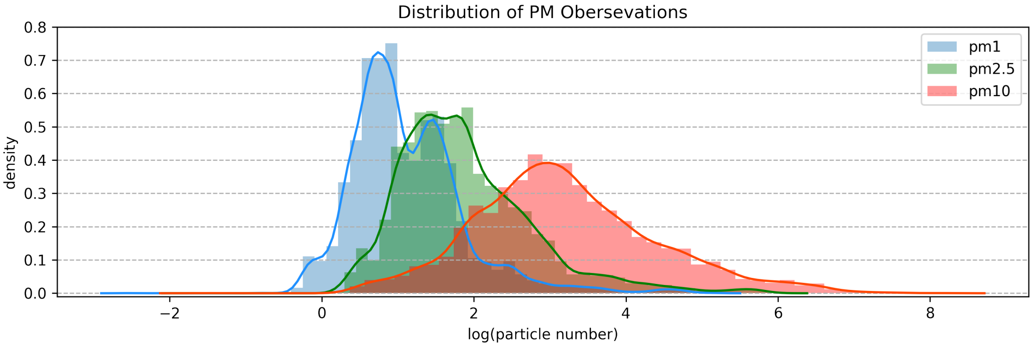


Figure 3.4: Distribution of PM_1, $PM_{2.5}$ and PM_{10} Observations Across All Trials in INHALE Study. Each PM value includes a Gaussian kernel density estimate (line).

location and time of recording. Although all subjects in the INHALE study collected data in London, a tester in a more polluted area or a factory would record higher PM values, resulting in more pronounced peaks for all three PM values, such as INH002(2). Nevertheless, the differentiation of the PM distributions for these different trials within a reasonable range allows for comparability in subsequent experiments. This analysis will investigate whether the intensity of PM exposure has a significant impact on the level of response in humans, by comparing the PM distributions across different trials.

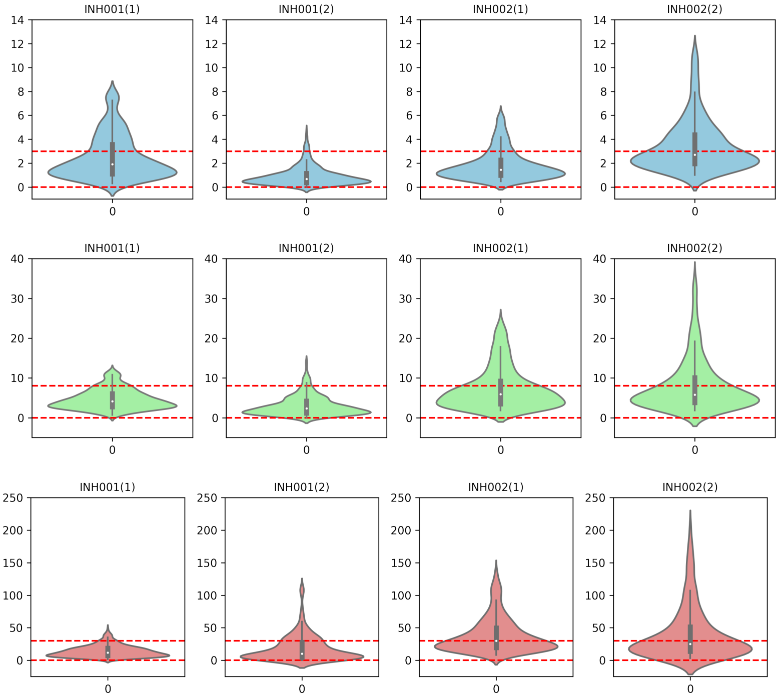


Figure 3.5: The distribution of the different PM levels for the two trials in the first two subjects of the INHALE study. The x-axis is density and the y-axis is PM mass. The peaks for all three PM levels occurs in the approximate range (red dashed line).

3.2 RESpeck Data

In this study, the wearable RESpeck sensor was employed to capture the response levels of DAPHNE - asthmatic adolescsnts and INHALE - older asthamtics. The recorded data, captured at minute intervals, is akin to that of the AIRSpeck [2]. The sensor was placed uniformly on the left side of the subjects' abdomen to obtain accurate readings. By virtue of its built-in 3D acceleration sensor, the RESpeck was able to gather not only the respiratory rate of the wearer but also infer new parameters such as the type of activity being carried out. The specifics of the various parameters measured by the RESpeck are presented in Tab 3.4.

Feature	Explanation
Timestamp	UTC timestamp of current observation.
Respiratory Rate	The breathing rate derived from the breathing signal.
Respiratory Rate std	The standard deviation of the respiratory rate per minute.
Activity Level	A measure for the amount of movement per minute.
Activity Type	The current activity of the subject per minute.
StepCount	The number of steps of the subject per minute.

Table 3.4: Explanation of features measured by RESpeck sensors [40].

In Fig 3.6, the RESpeck device was used to measure the respiratory rate during a single trial INH001(1). The respiratory rate is illustrated with a blue dotted line, which displays changes over time. Additionally, the corresponding activity level was also recorded and is shown in the graph as a coral line for comparison.

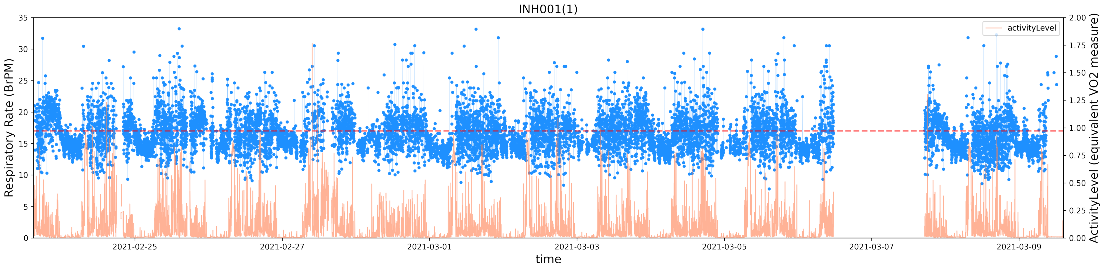


Figure 3.6: The RESpeck device was used to measure the respiratory rate in terms of Breaths per minute (BrPM) during a single trial (INH001(1)). The resulting data was graphically represented by a blue dotted line, which captures changes in the respiratory rate over time. Concurrently, the corresponding activity level was recorded and depicted by a coral line for comparative purposes.

The presented graph depicts the distribution of subject's respiratory rate, with most observations falling between 10-25 breaths per minute, but with a few outliers above 25 BrPM. This finding is consistent with prior research indicating that asthmatics typically have higher respiratory rates than non-asthmatics [20]. In fact, normal adults have a respiratory rate of 12-20 BrPM, whereas asthmatics often have rates exceeding 20 BrPM due to narrowed airways that result in reduced air flow. This reduction necessitates a faster breathing rate to ensure sufficient oxygen intake and carbon dioxide removal, and respiratory rates may also increase during an asthma attack [42].

Moreover, the graph provides initial evidence for a strong positive correlation between respiratory rate and activity level. Specifically, when the activity level surpasses a threshold of 5, the respiratory rate likewise rises, often above 20 BrPM. Conversely, when the activity level is low, hovering near 0, the respiratory rate tends to remain below 17 BrPM, as indicated by the red dashed line on the graph.

Fig 3.6 displays significant intervals of missing data, particularly around March 7, 2021, and during some evening hours. The probable cause of this issue is incorrect sensor placement or body positioning by the user. Specifically, if the RESpeck detects that the user is lying down, or if the user removes the sensor during sleep, the respiration rates recorded during these periods will be filtered out and marked as missing data. Tab 3.5 shows the percentage of missing RESpeck data per subject, score was determined by calculating the ratio of the total number of missing values, including NaN and missing time points, to the total duration of the data. Specifically, the formula used to calculate the percentage was % = (number of NaN + number of missing time points) / total duration.

Subject **Feature**	**INH001(1)**	**INH001(2)**	**INH002(1)**	**INH002(2)**	**INH003(1)**	**INH003(2)**
Respiratory Rate	35.2%	41.3%	47.7%	42.2%	34.4%	39.5%
Activity Level	14.9%	21.2%	19.8%	23.1%	8.4%	13.1%

Table 3.5: The percentage of missing RESpeck data for the first 6 trials. Results were determined by calculating the ratio of the total number of missing values (including NaN and missing time points) to the total duration of the data.

RESpeck reports on ”quiet breathing at rest” meaning that it does not record breathing during movement, speech, or eating. As such, the absence of breathing data during these activities does not necessarily indicate missing data. Moreover, since each trial in the INHALE study spanned a duration of 14 days, the available data could have been substantially improved through judicious trimming and interpolation during data processing. It is important to exercise caution when interpolating some of the remaining missing data, but the high percentage of missing values can actually decrease the impact of outliers on the data to a significant extent. Furthermore, since implausible data has already been filtered out based on the nature of RESpeck, a more reasonable distribution facilitates subsequent interpolation that is both easier and more accurate.

The statistical metrics for respiratory rate and activity level are shown in Tab 3.6. Both have relatively small standard deviations of 2.8 and 0.15, respectively, suggesting a concentrated data distribution. The skewness of respiratory frequency is only 0.98, indicating a generally symmetrical data distribution with approximately the same number of data points on both sides of the distribution. Therefore, the central tendency of the data, either the mean or the median, matches well with the distribution pattern of the data, as determined by the standard deviation, enabling more accurate subsequent analysis and modeling.

Feature	Max	Min	Mean	Median	Quantile(1/4)	Quantile(3/4)
Respiratory Rate	33.2	7.7	16.3	15.8	14.3	18.1
Activity Level	1.76	5.2e-03	0.08	0.24	7.9e-03	8.4e-03

Feature	Variance	Std	Skewness	Kurtosis	Autocorr	PACF
Respiratory Rate	8.3	2.8	0.98	2.07	0.37	0.51
Activity Level	0.02	0.15	3.26	11.6	0.84	0.62

Table 3.6: RESpeck data location measures in INHALE study.

However, the skewness for activity level is 3.6, revealing a clear left tail skew in the data distribution. Previous research suggests that this may be due to the fact that subjects spend most of their time sitting or lying down, and the duration of exercise is significantly less than the time of day when the body is relaxed [40]. Consequently, counts for lower activity levels dominate the data. Nonetheless, the maximum value of 1.76 for activity levels falls within a reasonable interquartile range, indicating that the distribution of activity levels is not significantly abnormal, and there are no erroneous observations.

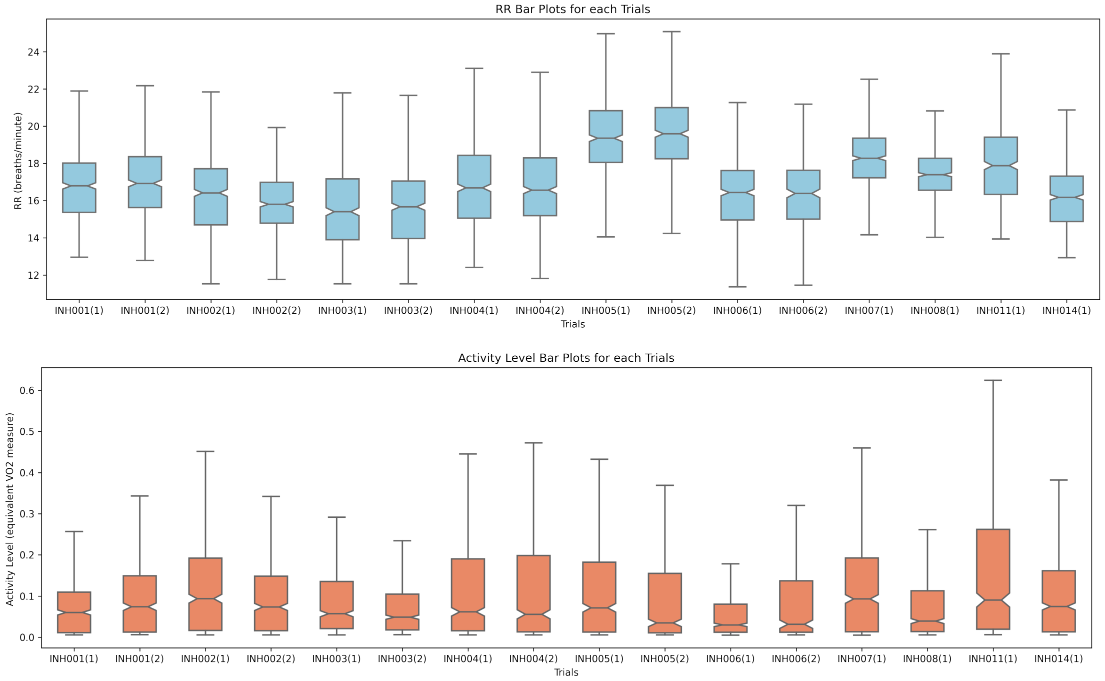


Figure 3.7: Respiratory rate and its corresponding activity level boxplot for each trial. Note that the activity level data are more skewed towards the lower left tail.

It is important to consider that the scores in Table 3.6 are cumulative for all trials and that there may be particular outliers in some specific trials that require special attention. To avoid this, the boxplot for respiratory frequency and activity level for each trial has been presented in Fig 3.7. It can be noticed that the distribution of respiratory rate is relatively similar, but there are still some specific trails with a significantly higher distribution than others, such as IHN005 (1) and IHN005 (2), but the activity level of IHN005 is not significantly different from the other patients. According to a study published in Chest [22], this may indicate that the individual physical condition of IHN005 may differ significantly from the other subjects, suggesting that it requires more respiratory effort to perform daily activities in order to maintain normal gas exchange. However, their activity levels were not significantly affected to overcome the effects of airflow resistance and impaired lung function.

3.3 Kendall Statistical Test

In the previous exploratory data analysis (EDA) of the two sensors' data, certain variables were found to have a strong linear relationship. Therefore, this section aims to conduct a detailed examination of the correlation between these variables. Correlation analysis is a widely used statistical method in scientific research to measure the strength and direction of association between two variables. It helps researchers understand the relationship between variables and make predictions based on that relationship. In particular, Kendall's tau correlation test is a non-parametric method that assesses the degree of association between two variables regardless of their distribution. This test is particularly useful for analyzing correlations between variables that are ordinal or non-normally distributed [21]. Thus, in this section, the Kendall's correlation test will be utilized to assess the degree of association between the two sensor datasets.

This analysis can offer valuable insights into the relationship between the two sensors and enhance our comprehension of their interactions. By conducting a Kendall correlation test, we can ascertain the strength and direction of the relationship between the sensors. This information can be used to improve the accuracy and reliability of future sensor measurements. Furthermore, understanding the correlation between sensors can aid in identifying any possible issues or biases in the data, which can improve the quality of the data analysis. In this way, this analysis serves as a critical step towards advancing our knowledge of the sensors and optimizing their effectiveness.

To conduct the Kendall correlation test, data will be collected from two sensors. A Python statistical package will be utilized to execute the test, which involves calculating the Kendall correlation coefficient (τ) and a corresponding p-value to assess the statistical significance of the results. A p-value below 0.05 will indicate a significant correlation between the two datasets. Fig 3.8 shows the kendall rank correlation coefficient among all features.

The graph illustrates that the coefficient between bin0 and bin6 exceeds 0.6, indicating a stronger correlation between smaller diameter bins. This finding supports the hypothesis in Figure 3.1 that there is a strong linear relationship between the different bins. The high Kendall correlation coefficients observed for the number of particles of different diameters in the air could be due to several factors. One potential explanation is that these particles have similar sources, such as combustion emissions or road dust, causing their number to vary according to environmental factors, including wind speed, air temperature and humidity [47]. Thus, a comparable ordering trend may emerge across all particle sizes, contributing to the large Kendall correlation coefficients observed.

Another possibility is that these different diameters of particles are linked together by physical or chemical processes, creating a comparable ordering trend between them [10]. For example, if all the particles are produced by a single chemical reaction process, they may be linked in a way that produces a similar ordering trend . However, further research is needed to explore this possibility in greater depth.

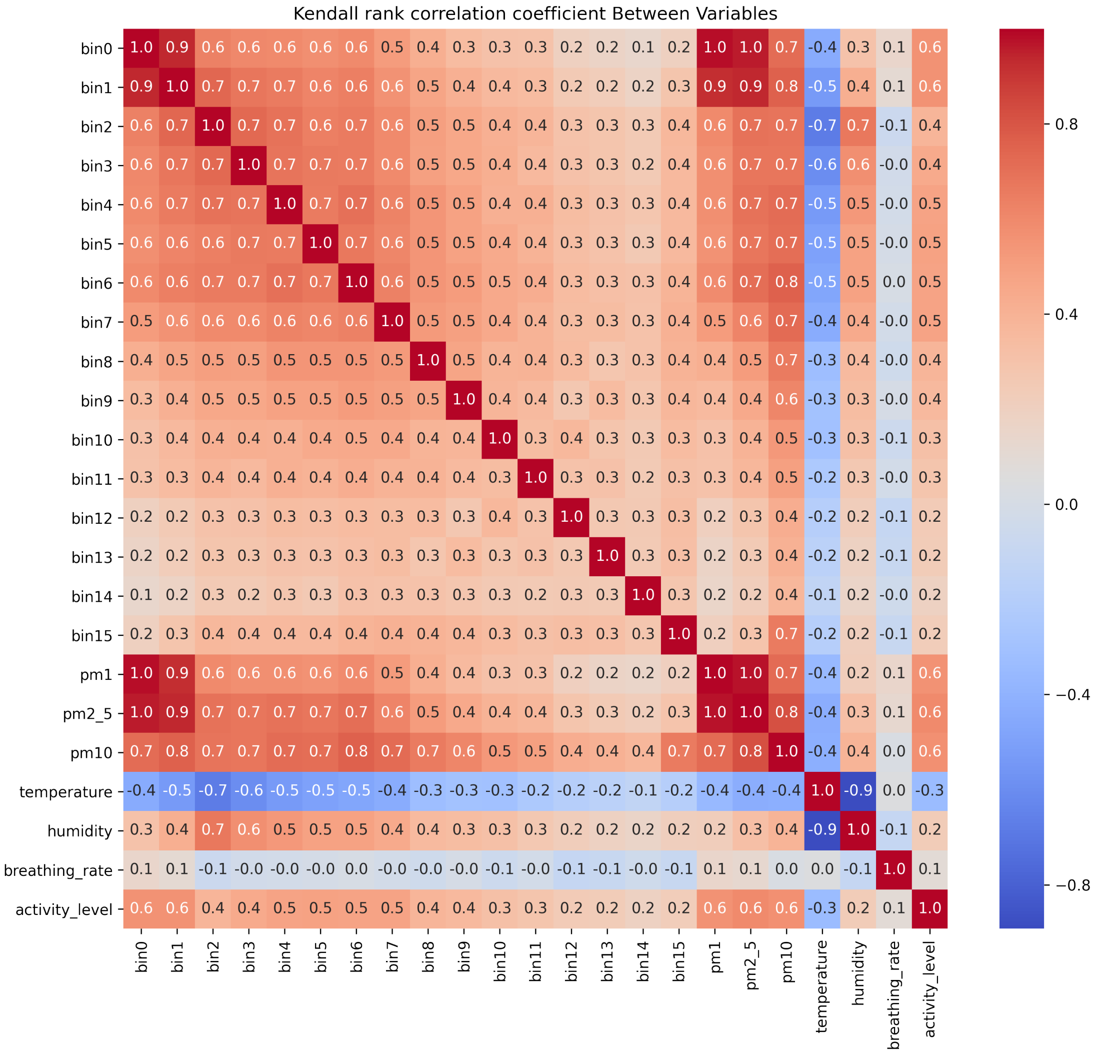


Figure 3.8: Kendall Rank Correlation Coefficient between all Features from both AIRSepeck and RESpeck. The range of values is between -1 and 1, where -1 indicates a complete negative correlation, 0 indicates no correlation and 1 indicates a complete positive correlation.

The graph illustrates a gradual decrease in correlation between particle diameter and other bins, as the diameter increases. For instance, the correlation coefficients between bin11-bin15 and other bins are all less than 0.4. While still positive, the correlation weakens considerably. This phenomenon can be attributed to the high number of 0 values observed in the larger diameter bins from Fig 3.3, leading to a highly skewed data distribution. This may conceal the true relationship between the variables and affect the sensitivity of the Kendall correlation coefficient. Furthermore, the presence of

0 values can also render variable ranking unstable, which may influence the calculation of the Kendall correlation coefficient.

The correlation coefficients of PM_1 with the first three bins are all greater than 0.6 and even above 0.9 for bin 0 and bin 1. $PM_{2.5}$ is above 0.6 with bins 0-6, while PM_{10} has a significantly higher correlation coefficient for the more backward bins than PM_1 and $PM_{2.5}$. Three PM values all have kendall correlation coefficients greater than 0.7 with each other, indicating a strong positive correlation between them. This means that as the level of one pollutant increases, the levels of the other pollutants also tend to increase. This correlation may indicate that the sources of these pollutants are similar and that they have similar transport and deposition patterns in the atmosphere. A reference for this information can be found in a study by Goyal et al. (2016) entitled ”Relationship between PM_1, $PM_{2.5}$ and PM_{10} concentrations in urban, suburban and rural areas of northern India” [24]. In this study, the authors calculated Kendall correlation coefficients between PM_1, $PM_{2.5}$ and PM_{10} concentrations measured at different locations and found that the coefficients were all greater than 0.7, indicating a strong positive correlation between these pollutants.

The graph illustrates an inverse correlation between temperature and all three PM values, with a coefficient of around -0.4 for each bin. Several factors could account for this finding. For instance, during cold weather, the air density increases, and the atmosphere’s stable layer is lower. As a result, PM diffusion is hindered, and it is more prone to being adsorbed by other particulate matter in the atmosphere (such as water vapor or oxygen), leading to higher PM concentrations. Additionally, lower winter temperatures may cause an increase in heating equiPMent usage and higher PM emissions from certain industrial processes. These PM emissions are more likely to be adsorbed by other particulate matter in the atmosphere, further contributing to higher PM concentrations. Another possible explanation is the occurrence of an inversion, which is a phenomenon where atmospheric temperature rises with altitude. This effect is typically observed during autumn and winter, when surface temperatures are low, and the atmosphere is saturated with water vapor, particulate matter, and other substances at lower altitudes. Inversions can create a lower stable layer of the atmosphere, which hinders PM diffusion and increases its likelihood of accumulating in the lower atmosphere, ultimately leading to higher PM concentrations [49].

The Kendall correlation coefficient is a statistical tool used to measure the degree of non-linear correlation between two variables, providing insight into the relationship between them. However, despite its usefulness, the coefficient has certain limitations. Specifically, it fails to take into account non-monotonic correlations, as it can only assess monotonic relationships between variables. In addition, the Kendall correlation coefficient cannot determine causality between variables, as correlations only indicate statistical relationships between variables and do not prove causality. Causality refers to the likelihood that one variable induces a change in another variable. Therefore, when investigating causal relationships between variables, it is important to conduct thorough research and analysis beyond the simple calculation of correlation coefficients. In the next section, the necessary pre-processing of the data before the causal analysis method is carried out.

Chapter 4

Data Pre-Processing

The importance of data pre-processing cannot be over-emphasised in any data analysis task. It is widely accepted that obtaining perfect data is an important prerequisite for the application of subsequent algorithms. However, raw data such as AIRSpeck and RESpeck data are often problematic and incomplete, making pre-processing critical to improving the accuracy and predictive performance of the model. To achieve this, the raw data needs to be cleaned, processed and transformed before subsequent algorithms can be applied. Pre-processing helps to reduce the effects of errors and noise, thereby improving the accuracy and reliability of the model. The operations involved in data pre-processing typically include dealing with missing values, removing outliers, normalising the data, selecting relevant features and transforming the data.

It is worth noting, however, that great care must be taken when carrying out these steps. Improper handling of the data can lead to completely incorrect conclusions and make future analyses unreliable. Therefore, this chapter provides a detailed explanation of each processing step and its purpose to ensure that the data is handled appropriately.

4.1 Common Methods

This section provides a concise overview of the data cleaning techniques utilized in the minf1 project, which have distinct advantages and scenarios in terms of their ability to effectively eliminate invalid or outlier values and impute missing data. The INHALE data will be preprocessed by using the same methods as those employed for the DAPHNE data last year, albeit with different parameters.

4.1.1 Calibration

In this project the OPC-R2 optical particle counter was used to determine the concentration and size distribution of particles in the air. Calibration is essential to ensure the accuracy of the measurement. To achieve this, the manufacturer's calibration guidelines were strictly adhered to in advance, utilising monodisperse particle sources such as polystyrene latex spheres and NIST traceable standards. The calibration method employed was adjusted to match the expected range of particle sizes to be measured, thus minimising experimental error. In addition, the two wearable sensors used in the study, AIRSpeck and RESpeck, also needed to be calibrated. This is necessary because there may be small differences between the sensors assigned to different subjects, which may lead to errors in the recorded data. A study carried out by the Centre for Speckle Computing at the University of Edinburgh showed that if the relative humidity of each sensor was below 80%, a calibration factor could be determined for each sensor by linear scaling. However, if the relative humidity does not meet the requirement, more advanced calibration methods, such as using pre-trained neural networks, are required.

To ensure accurate results from this project, all sensor data was rigorously calibrated before any further cleaning, pre-processing or analysis was carried out.

4.1.2 Anomalous Detection

Detecting outliers in observed data is a challenging task due to their atypical and unusual nature, which may result in incorrect conclusions or predictions. Outliers can occur due to various reasons such as data logging errors, transmission faults, equiPMent failures, or sample bias. Additionally, some outliers may be extreme events with significant value and meaning, and ignoring them may lead to the loss of valuable information and the possibility of identifying a true anomaly. Therefore, it is essential to exercise caution while determining whether an extreme value is an error or an actual occurrence.

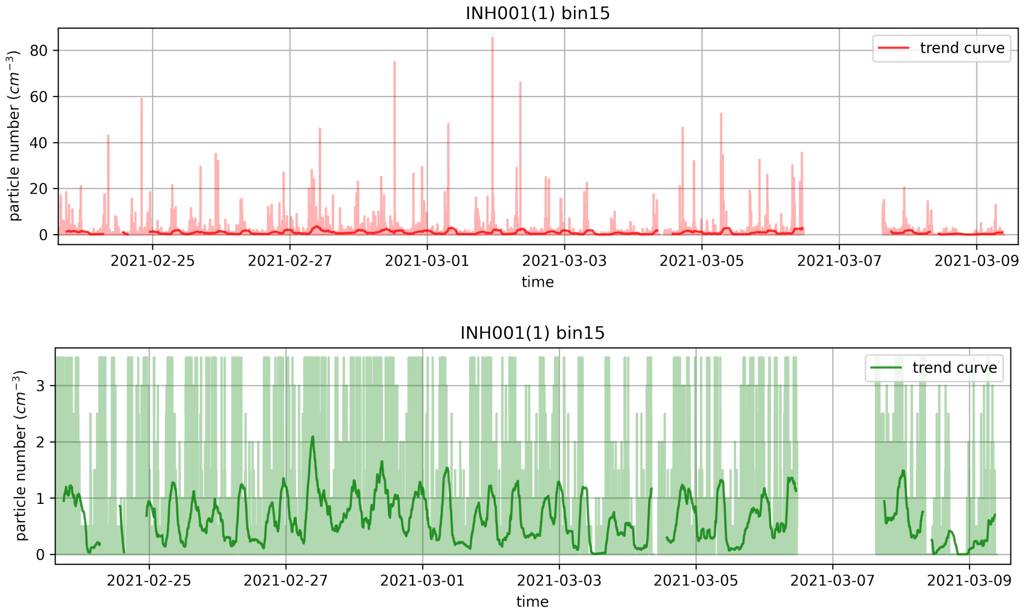


Figure 4.1: bin15 observations with/ignoring massive spikes.

Winsorizing is a commonly used pre-processing technique to handle outliers in a dataset. It involves replacing extreme values in the data with a specified percentile, usually the 5th or 95th percentile. This approach helps to reduce the impact of outliers on statistical analysis by compressing extreme values within a specified range. To implement winsorizing, values in the variable that are less than the lower limit are replaced with the lower limit, while values that exceed the upper limit are replaced with the upper limit. For instance, Fig 4.1 illustrates an example of an unrealistic observation in which individual bin15 for trial INH001(1) exhibits many transient huge peaks. The red line on the graph depicts the trend in the data, and it is apparent that ignoring these extreme values will lead to more realistic values falling within a reasonable range. The lower panel in Fig 4.1 demonstrates the effect of carrying out a 90% winsorization on the dataset, resulting in a more realistic trend curve.

4.1.3 Interpolation

The occurrence of missing values in timing data is prevalent, particularly in the monitoring of physiological signals such as respiratory rate. Missing values can be attributed to equiPMent malfunctions, incomplete signal acquisition, or interruptions in data transmission, such as when the patient removes the RESpeck sensor while sleeping or assumes a lying-down position. The presence of missing values in timing data may negatively impact data analysis and subsequent algorithms. To ensure data completeness and accuracy, missing values need to be interpolated. However, caution is needed when performing interpolation to avoid producing unrealistic or artificial results.

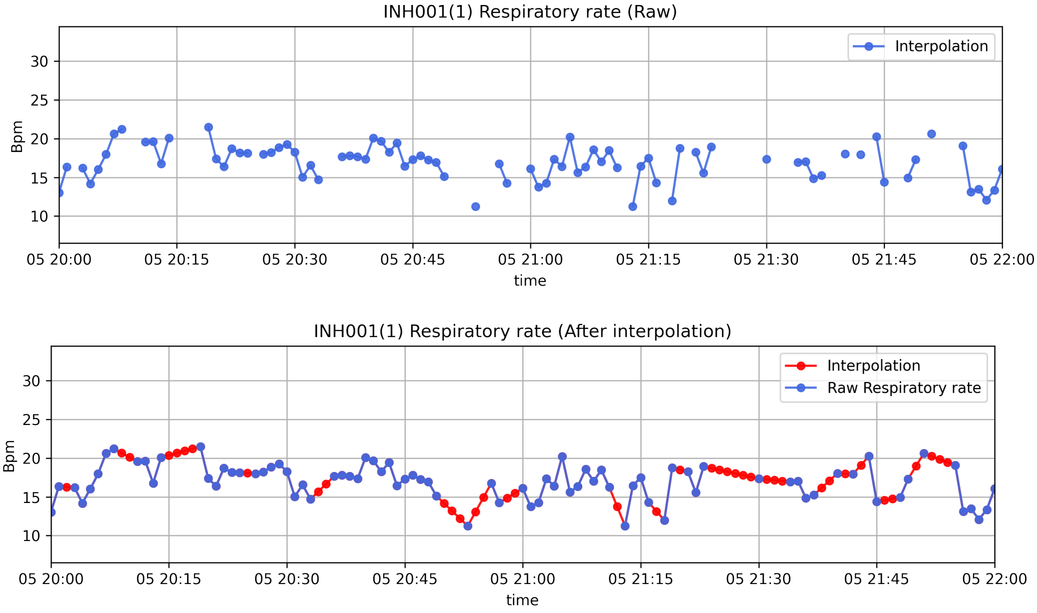


Figure 4.2: Physiological data such as respiratory rate were recorded with missing values, the top graph shows the original observations and the bottom graph shows the results after interpolation.

The choice of interpolation method relies on the data characteristics and missing value distribution. As respiratory frequency is periodic in nature, cyclic median interpolation is often employed to interpolate this physiological signal. In contrast, linear interpolation is often used for air data recorded by AIRSpeck. Nevertheless, it is essential to note that interpolation may introduce errors and, thus, should be approached with caution when analyzing and modeling data. Fig 4.2 displays the missing data for the INH001(1) trail at 2 hours and its corresponding interpolation, which results in continuous data that can be better analyzed and modeled.

4.2 Extraction of Coughing Signal

The RESpeck device's 3D acceleration sensor can detect movement patterns from the recorded timing data. However, it does not include coughing among its 14 different movement patterns. To detect coughing, additional work is necessary. This work builds on two previous tasks: data acquisition and data processing. By analyzing the characteristics of cough signals and training a machine learning model, it is possible to augment the RESpeck's capabilities to detect coughing events in wearers.

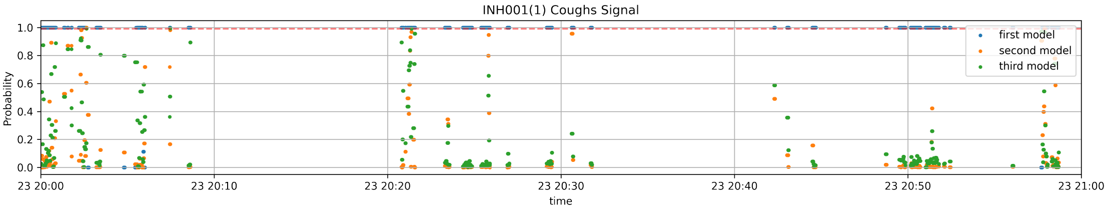


Figure 4.3: Three different models are applied to the recognition of coughs, with the x-axis being time and the y-axis being the corresponding probability. Each scatter represents the probability of a cough occurring in one of the models at a particular moment in time for the wearer.

Accurately identifying coughs is critical, given that they typically last only about one second. To achieve this accuracy, a recent study employed three different models to minimize the number of false positives. In this study, we use an example of a prolonged cough episode (INH001(1) lasting 1 hour) depicted in Fig 4.3. The three models utilized for cough detection were the 1-D-CNNs Social Signals Classification Model, HAR with Ensemble Classification Model (XGBoost, LightGBM, and Random Forest), and the AC-GAN HAR Model. Each model detected coughs using a unique color, with the x-axis representing time and the y-axis representing probability. Each scatter point on the graph represents the probability of a cough occurring at a specific time point. A red dashed line is included to represent a probability of 0.99. To minimize error, a cough is considered to have occurred only when all three models have detected a cough with a probability greater than 0.99 at the same moment.

The RESpeck's built-in 3D acceleration sensor records temporal data at a sampling frequency of 25 Hz, providing 25 data points per second. To determine the number of coughs, we accumulated the identified coughs per minute and analyzed the resulting data at a resolution of one minute. Fig 4.4 illustrates two examples of INH001(1) at different time periods, with the blue vertical lines indicating the number of coughs per minute and the corresponding activity levels shown in coral. This is due to the fact that when the subject coughs violently, the 3D acceleration recorded by the sensor fluctuates more significantly, leading to an increase in the activity level of the detected user.

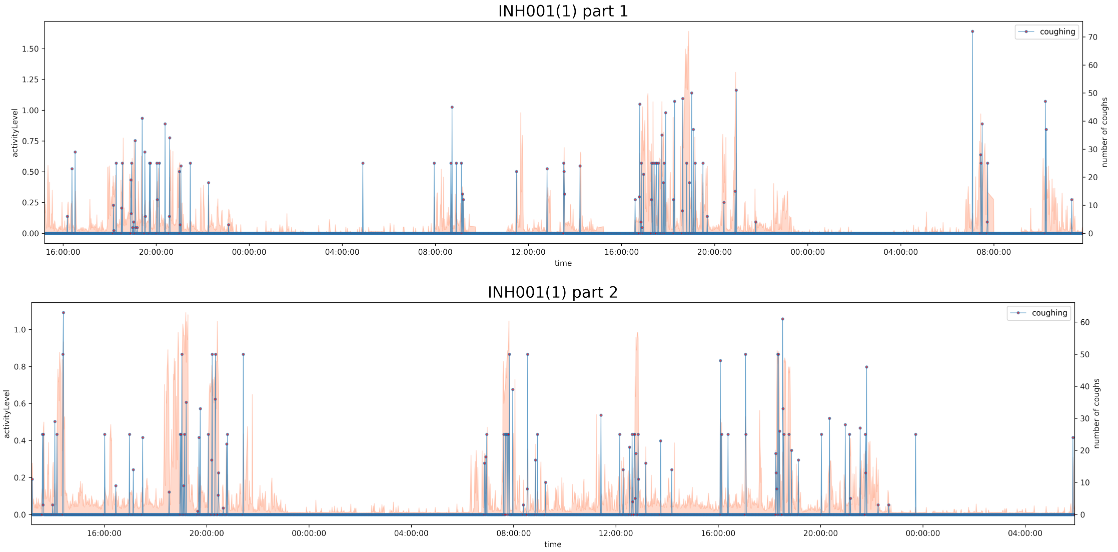


Figure 4.4: Example of the number of coughs at two different time periods for INH001 (1), the blue vertical plots represent the number of coughs in subject at a given minute, and their simultaneous activity levels are plotted in coral colours. The two can be seen to show a more obvious positive association.

4.3 Trimming

The method of data interpolation, while advantageous for preserving the integrity of the data to some extent, may yield inaccurate results when applied to the INHALE study data. The study's long time series and the abundance of missing values may undermine the accuracy of interpolation and, consequently, the validity of subsequent analyses.

To address this issue, we propose the use of data slicing techniques. Data slicing involves dividing long time series data into shorter segments and analyzing each segment separately. This approach has several benefits, including reducing the number of missing values within each segment, thereby improving data completeness, and providing a more precise understanding of data patterns and characteristics. The choice of the appropriate time interval for data slicing should be based on the specific needs and features of the data to ensure optimal accuracy and completeness.

All INHALE data was sliced into 20-hour segments based on the maximum lag time of the subsequent algorithm, allowing for a better balance between runtime and accuracy. However, selecting the appropriate segment to divide the data is crucial, as it should contain minimal missing values, and the allocation of the 14 features' missing values needs careful consideration due to the potential variability across features.

To minimize the number of missing values within each segment, the time points in all 14 features containing missing values were identified and sorted chronologically. A heuristic approach was then employed to distribute these missing values across the 20-hour time slots. Strategies such as evenly distributing the number of missing values per time period and minimizing the proportion of missing values per time period were considered. Ultimately, the goal was to reduce the number of missing values within each time period, thus enhancing the quality of the data and the accuracy of the algorithm.

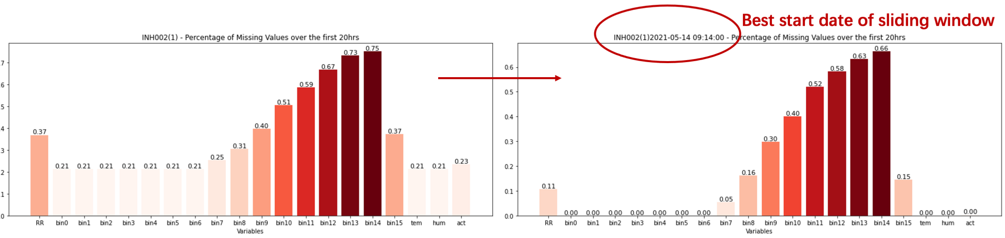


Figure 4.5: To demonstrate the efficacy of this method, a comparison of the missing values before and after the segmentation process was conducted for INH002(1) on a feature-by-feature basis. The results indicate a significant improvement in the quality of the data. Therefore, this method represents a substantial enhancement to the data.

Fig 4.5 illustrates the effectiveness of the proposed approach. Prior to slicing, INH002(1) had all features with missing values exceeding 0.21. Following the segmentation process, the proportion of missing values decreased by an average of 20% across all features, resulting in bin0-bin6, temperature, humidity, and activity levels having no missing values. Fig 4.6 depicts the efficacy of the proposed method in enhancing the quality of data for the initial six trials of the INHALE study. As shown in Table 3.5, the proportion of missing respiratory frequency values was reduced to below 11%. Prior to running subsequent algorithms, a comprehensive data cleaning process was conducted to ensure data quality.

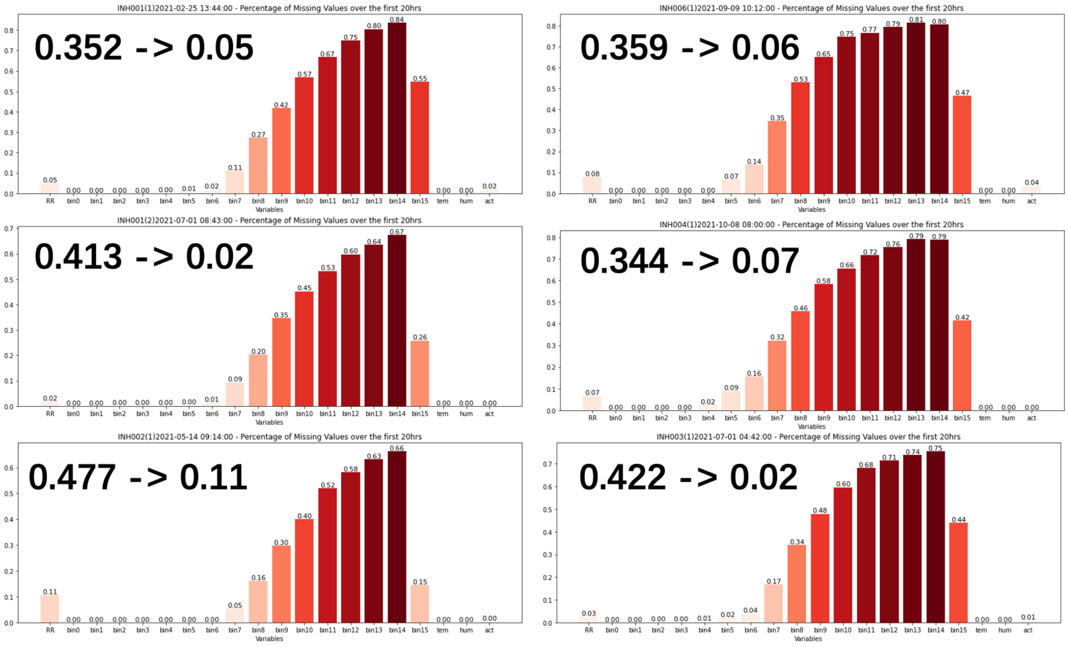


Figure 4.6: The efficacy of the proposed method in enhancing the quality of data for the initial six trials of the INHALE study.

Chapter 5

Causal Discovery Methods

Causal inference techniques are playing an increasingly crucial role in contemporary data-driven science. Advances in data collection and storage technologies have enabled us to work with larger and more complex data sets than ever before. However, the size and complexity of such data also make it more challenging to extract reliable information and insights from it. Consequently, discovering and comprehending causal relationships in data using causal inference techniques can help us better understand the underlying causes and mechanisms of observed phenomena, thereby facilitating scientific research and informed decision-making. PCMCI+ techniques are founded on the principles of graphical causal modelling and causal inference, and can effectively infer causal relationships from data. In this chapter, we apply the method to time series of airborne particulate matter and human cough signals of varying sizes, and analyse the corresponding exposure-response relationships. Through this analysis, we demonstrate the effectiveness of PCMCI+ in uncovering causal relationships in complex systems.

In this chapter, all conclusions are based on the non-linear PCMCI+ algorithm. The traditional PCMCI algorithm operates under the assumption of a linear causal model, which may not accurately capture the non-linear relationships between many variables, such as atmospheric data and human activity levels. Utilizing the non-linear PCMCI+ algorithm is crucial in accurately inferring causal relationships and avoiding misclassification or omission. Moreover, this algorithm not only detects causal relationships, but also provides a more thorough explanation of such relationships. This deeper understanding enables a better grasp of the underlying problem, ultimately improving the overall analysis.

5.1 Causal Results for Airborne Particles

This section aims to perform a step-by-step analysis of the data used in two studies for this project, with a specific focus on observing the exposure-causation relationship between different particles and PM values in the air and patients' cough symptoms. Ultimately, a thorough comparison will be made, considering objective reality and conducting a comprehensive analysis.

5.1.1 INHALE Study

PCMCI+ is a statistical method that aims to identify dependencies between time series data while considering possible confounding variables. It involves testing for linear or nonlinear dependencies between the bin0 and coughs time series for each trial, using different lag lengths.

In this experiment, a maximum lag time of 60 minutes was selected for the algorithm. This decision was based on a consideration of the characteristics of the research object, i.e., that coughing typically does not occur immediately and requires some time to manifest. Therefore, selecting a longer lag time can better reflect the impact of airborne particle concentration on coughing. Additionally, the data collection frequency was taken into account, as data were collected at a high frequency (with a resolution of minutes), allowing for a lag time of up to one hour to be chosen, which can help capture and discover both short-term and long-term causal relationships.

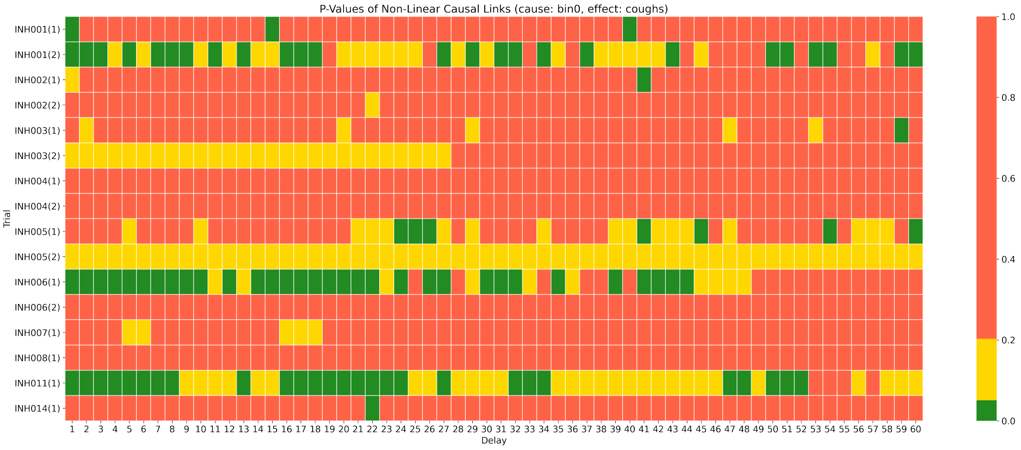


Figure 5.1: The INHALE result heatmap for non-linear dependencies between the bin0 and coughs time series for each trial using lag lengths ranging from 1 to 60 minutes. The resulting *p*-values were color-coded to aid visual analysis, with green denoting statistical significance ($p <$0.05) and strong evidence of non-linear dependency, yellow indicating some evidence of non-linear dependency ($0.05 \leq p \leq 0.2$), and red representing little to no evidence of non-linear dependency ($p >$0.2).

Fig 5.1 shows an example: a heat map of PCMCI+ results for the non-linear dependence between bin0 and coughs time series. The p-value is a statistical measure that provides information on the strength of evidence against the null hypothesis of no linear and non-linear dependency between the two time series. Typically, a p-value less than 0.05, which is depicted in green, is considered statistically significant, indicating strong evidence against the null hypothesis. If the p-value falls between 0.05 and 0.2, which is depicted in yellow, it suggests some evidence against the null hypothesis, but the evidence is not significant. Conversely, a p-value greater than 0.2 indicates little to no evidence against the null hypothesis, which represented in red.

The following are examples of three experiments, each demonstrating a different relationship:

1) **INH001(2)**: The results of this experiment demonstrate that the p-values were significantly maintained in almost every minute during the first hour, indicating not only a significant causal effect of bin0 on the cough response of the subject, but also the possibility of its persistence for a longer period.

2) **INH004(1)**: The p-values of this experiment did not show a significant relationship during the first 0-60 minutes, indicating that the exposure-response relationship of the subject was not apparent.

3) **INH006(1)**: The intriguing results of this experiment show that the subject immediately sensed the effects of the exposure and maintained a strong intensity of response for up to 45 minutes, but these effects rapidly diminished after 45 minutes.

In the current experiment shown in Fig 5.1, a total of 127 meaningful causal relationships were discovered. Interestingly, all 16 trials of the INHHALE experiment could be divided into two categories: one with a total of 127 links discovered within the first hour of exposure (INH001(2), INH005(1), INH006(1), INH011(1)), and the other containing 12 trials with an average of no more than 3 links. This extreme distribution implies that due to individual differences, some trials were more susceptible to the effects of bin0 exposure, resulting in a strong causal relationship with coughing. Given that all experiments were designed as nonlinear, the detection of nonlinear relationships by PCMCI+ is relatively complex as it may involve interactions among multiple variables. Therefore, although the number of causal links detected by nonlinear PCMCI+ may be fewer than that of linear test, it is more effective in distinguishing and avoiding false positives. The significant coughing response observed in these four trials was largely influenced by the exposure to bin0.

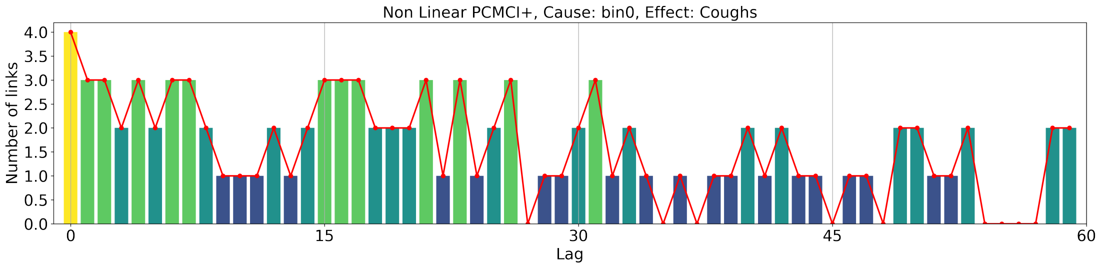


Figure 5.2: Distribution of non-linear causal links across all lag lengths and all trials in INHALE study (Cause: bin0, Effect: coughs).

The distribution of non-linear causal links across all lag lengths and trials, obtained from Fig 5.1, is illustrated in Fig 5.2. The figure indicates that the link strength is highest at minute 0 (instantaneous), reaching a value of 4. As the lag time increases, the number of links gradually decreases, with a link strength of approximately 2.5-3. Furthermore, at minute 30, there is a more prominent weakening of the links, with a strength of approximately 0-2.5.

In time series data analysis, causal links with smaller lag times are generally regarded as short-term causal links, which may be caused by instantaneous and local causal mechanisms. This is consistent with reality since coughing is usually caused by short-term factors [30, 12, 46]. In contrast, causal links with longer lag times may be caused by longer-term and more global causal mechanisms. Therefore, if the analysis results show that the causal links become weaker as the lag time increases, it may be because the effect of the particle count in the air on coughing events is only a short-term causal link, or there is a time delay effect, which means that the effect of the independent variable takes some time to manifest in the dependent variable. It is also important to note that random noise and interference are often significant issues in time series data analysis. With increasing lag time, random noise and interference may gradually accumulate, interfering with the detection and analysis of causal links and leading to a weaker causality.

To enhance the validity of our findings, we will conduct comparisons using the DAPHNE dataset, thereby providing more convincing evidence to support our conclusions.

5.1.2 DAPHNE Study

As of December 2018, the DAPNHE project consisted of a total of 126 trials. For the experiments, we selected 53 trials with missing values of less than 15% across all features. We maintained a high level of consistency with the data preprocessing and model parameters used in the previous INHALE study to ensure accurate conclusions. The experimental results are presented in Fig 5.4 and Fig 5.3.

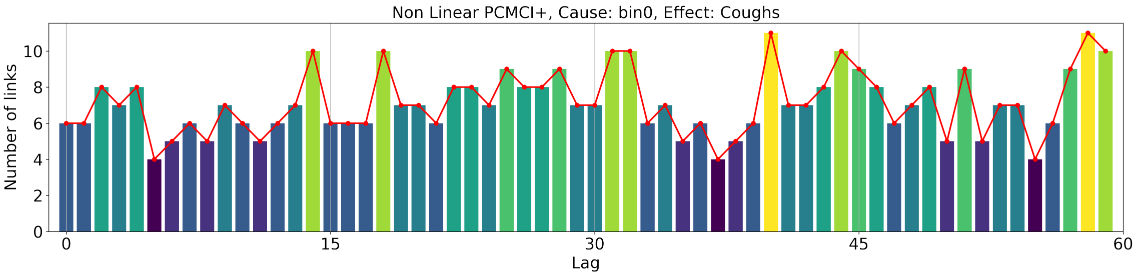


Figure 5.3: Distribution of non-linear causal links across all lag lengths upto 60 minutes and all trials in DAPHNE study (Cause: bin0, Effect: coughs).

The conclusion drawn from the study shows that the results obtained from running the PCMCI+ algorithm on different datasets are varied, despite the similarity in data distribution and sampling frequency. This variability is likely due to the individual differences among subjects, as well as the differences in the sampling locations of INHALE and DAPHNE, with the former conducted in London, UK and the latter in Delhi, India. These variations in the physical condition of subjects and the surrounding environment may have contributed to the observed differences in the results. Nonetheless, the conclusion supports the evidence of a direct and significant causal relationship between the particle count in the air and the number of coughs experienced by asthma patients, as demonstrated by the linkages observed in Fig 5.3, which averaged more than 5 links per hour across an average of 53 subjects. To provide a comparative analysis of the impact of particles of different diameters on human response, a series of experiments will be conducted. The comparison aims to clarify which PM has the most direct impact on human cough response.

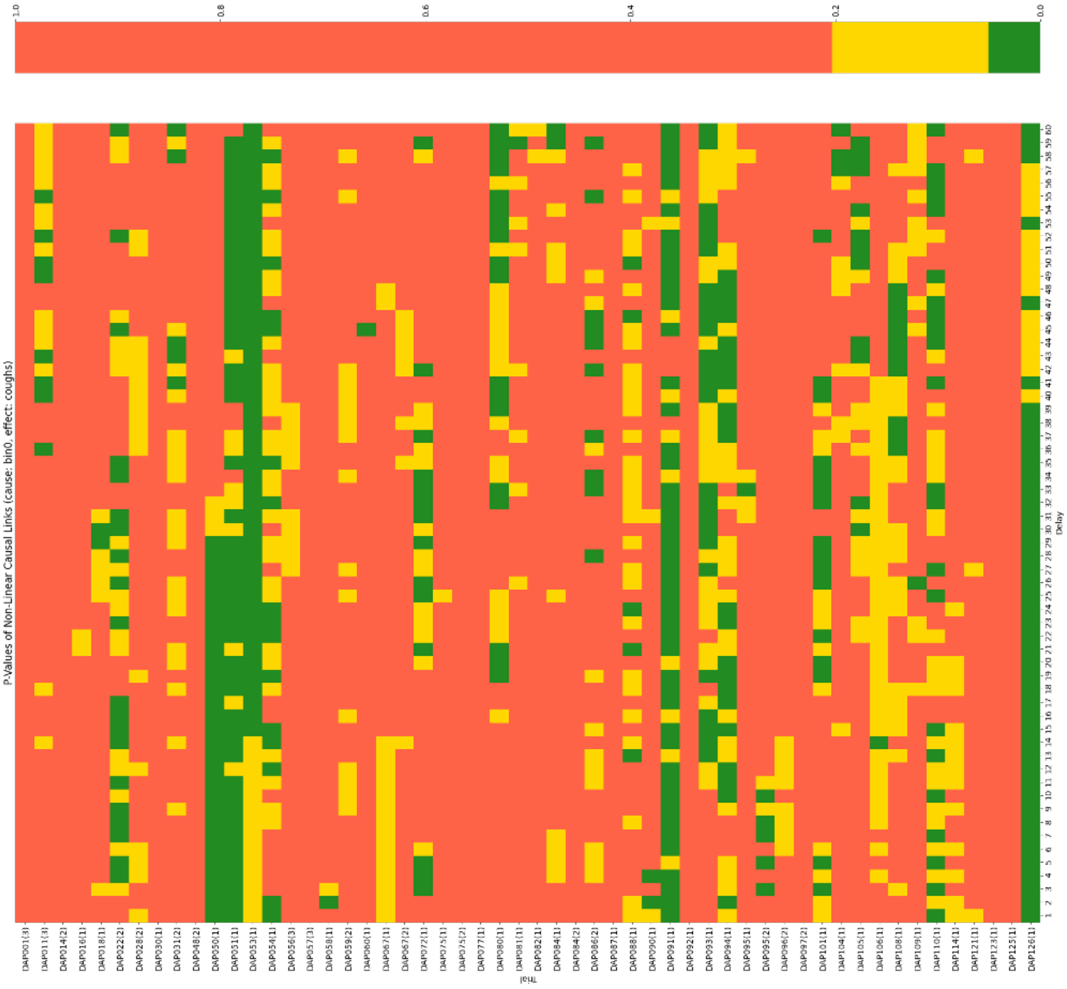


Figure 5.4: The DAPHNE result heatmap for non-linear dependencies between the bin0 and coughs time series for each trial using lag lengths ranging from 1 to 60 minutes.

5.1.3 Accumulation and Comparison

Fig 5.5 displays the average number of causal links of 16 bins with a maximum lag of one hour in both DAPHNE and INHALE studies. It is evident that although there exist individual variations in different regions, the conclusions drawn are very similar. The figure indicates that the smaller the diameter of the particle, the higher its number of causal links for coughing. For the smallest particle bin 0, both studies show more than six causal links. In the DAPHNE data, bins 3 to 12 have causal links below six, while bins 13 to 15 have links that drop to below five. This phenomenon is more pronounced in the INHALE study, where the effect of larger diameter particles on coughing seems weaker. Bin 0 and bin 15 differ almost six-fold (6.0 and 1.1), consistent with the trend observed in last year's project on the exposure response of particle number and respiratory rate. A possible explanation for this is that smaller diameter particles can easily penetrate deep into the respiratory tract, and these particles can stimulate the respiratory tract, causing discomfort symptoms such as coughing.

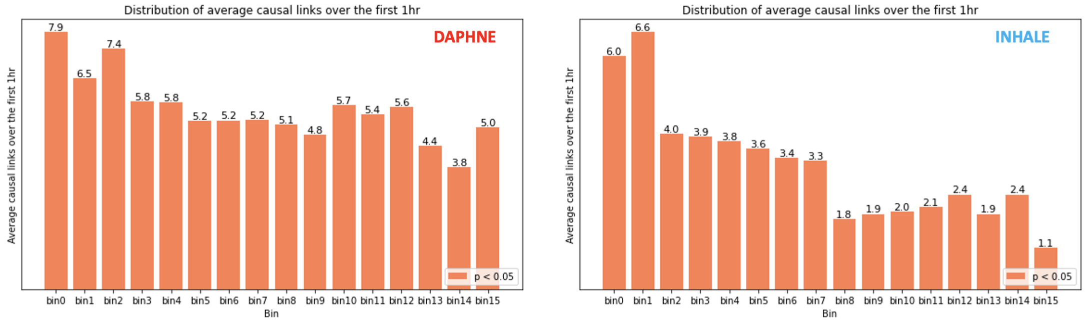


Figure 5.5: For each bin value at a maximum lag time of 1 hour average number of causal links is shown on the left for DAPHNE and on the right for INHALE.

After analyzing the impact strength of particles of different diameters on coughing, it is necessary to consider that particles in the air are observed simultaneously. Thus, aggregating all bins can reveal the most intuitive trend of how air pollutants affect coughing in subjects. This was obtained by averaging the causal distribution of all 16 bins and is presented in Fig 5.6, which includes two subplots: DAPHNE on the left and INHALE on the right. It can be observed that after aggregating the bins, the overall number of causal links becomes more apparent, and the effect progresses from short-term to longer-term impact (with little attenuation). An interesting finding reveals that both the DAPHNE and INHALE studies show a clear peak at around 10 minutes, indicating a relatively short typical delay between exposure and response. Another smaller peak appears at around 45 minutes, suggesting that the body's typical response to exposure occurs in a wave-like manner, possibly corresponding to the extent of particle penetration into the respiratory system [14].

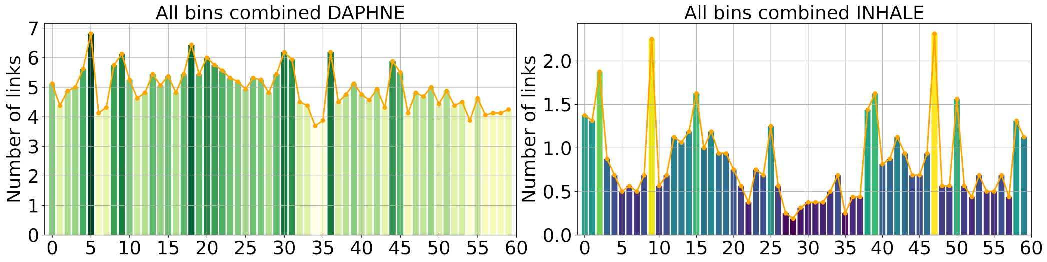


Figure 5.6: All bin combined distribution of non-linear causal links across all lag lengths and all trials (shown on the left for DAPHNE and on the right for INHALE).

The causal links between three different PM values are presented in Fig 5.7, allowing for direct comparison between them. Several conclusions can be drawn from the figure. Firstly, the short-term number of causal links of PM1 is significantly higher than that of PM2.5 and PM10. This observation is consistent with the phenomenon described in Appendix B.1, which suggests that smaller particles have a higher number of short-term links. Secondly, a clear causal lag is observed for PM2.5 and PM10, as the NCL peaks five minutes after subject exposure. This may indicate that coughing reactions only begin to manifest in subjects after the particles have penetrated and been absorbed by the respiratory system. Lastly, although there is some overlap between the particles considered by the three PM values, the overall difference in number of causal links (NCL) is not significant. Among the three PM values, PM2.5 has the greatest impact on coughing, reaching a NCL of 6.3. Meanwhile, PM1 and PM10 have values of 5.2 and 4.5, respectively.

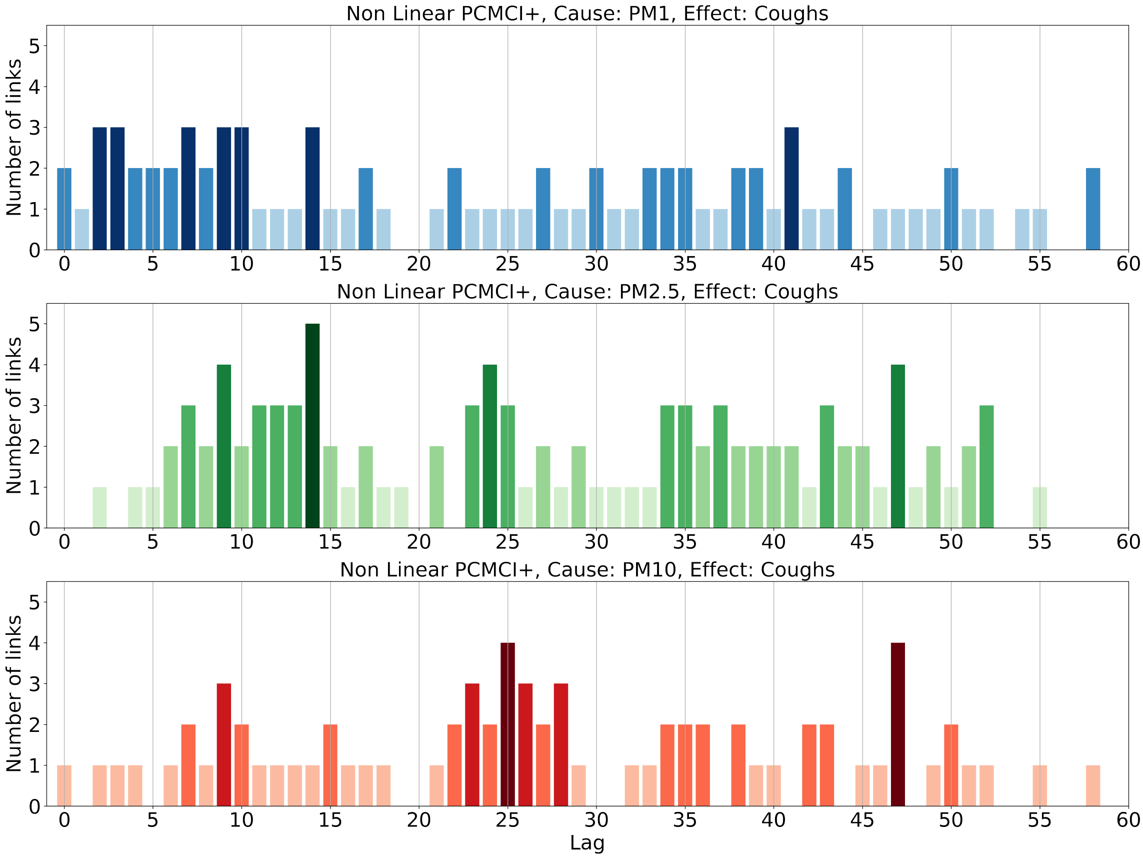


Figure 5.7: Distribution of non-linear NCL values across all lag lengths and all trials (PM_1, $PM_{2.5}$, PM_{10} respectively) for INHALE study.

The study aims to further analyze the relationship between different PM values and their corresponding interval bins. Fig 5.8 displays the cumulative significant p-values of all bins within specific intervals. The subfigures present the results for bins 0 to 6 (corresponding to $PM_{2.5}$), bins 7 to 10 (complementary to $PM_{2.5}$ for PM_{10}), and all 16 bins from 0 to 15. The results indicate a consistent pattern between PM and its corresponding interval bins. Appendix B.3 shows the p-value results of three one-hour PMs. Specifically, the results for PM_1 and $PM_{2.5}$ demonstrate that INH006(1) exhibits significant NCL during the first 20 minutes, with all p-values being less than 0.05. The first subplot of Fig 5.8 (bins 0-6, corresponding to the $PM_{2.5}$ interval) shows that the cumulative significant p-values are consistently greater than 4 out of 6, indicating a completely consistent pattern. These findings confirm the association between PM and bins in terms of their patterns.

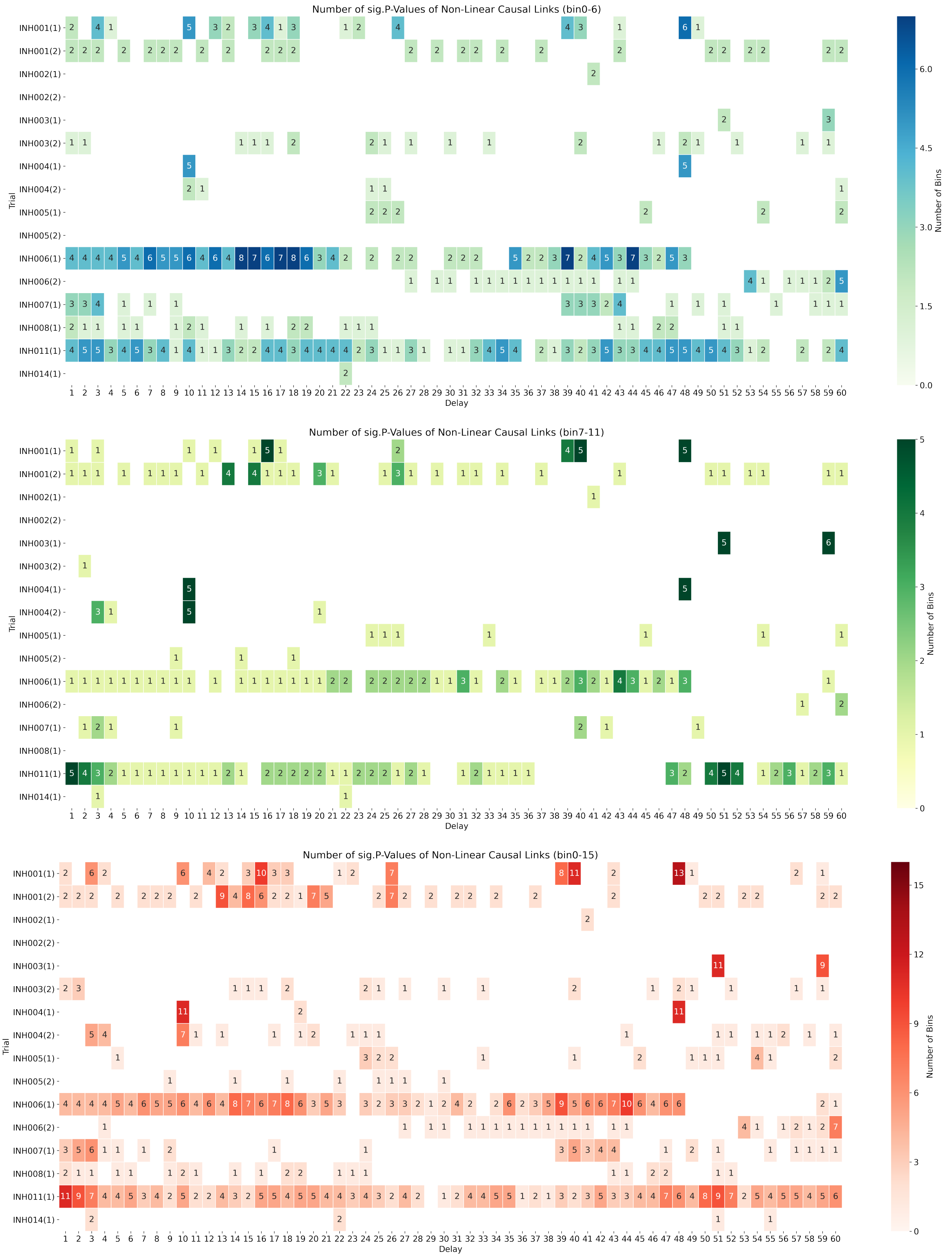


Figure 5.8: Results in INHALE study for the cumulative number of significant p-values. The three plots are cumulative for bin0-bin6 (corresponding to $PM_{2.5}$), cumulative for bin7-bin10 (complement of PM_{10} to $PM_{2.5}$) and cumulative for all bin0-15 for all 16 bins. Results in DAPHNE in Appendix B.4, B.5 and B.6.

5.2 Causal Results for Other Factors

Asthma is a chronic respiratory disease characterized by symptoms such as coughing, shortness of breath, wheezing, and chest tightness. In addition to airborne pollutants, numerous factors may influence the coughing symptoms of individuals with asthma. PCMCI+ algorithm takes all of these factors into account and creates a complex causal network that enables the analysis of the relationships between various variables. However, it is imperative to make reasonable assumptions before applying the PCMCI+ algorithm since unreasonable assumptions may result in biased conclusions. For instance, the causality between individual activity data and air data is unidirectional, i.e., respiratory rate and activity level do not cause changes in air data (such as particle count, temperature, and humidity), but the opposite is true. Before running any algorithm, strict assumptions must be made regarding the parameters of the algorithm and they should be set according to real-world circumstances to obtain objective conclusions.

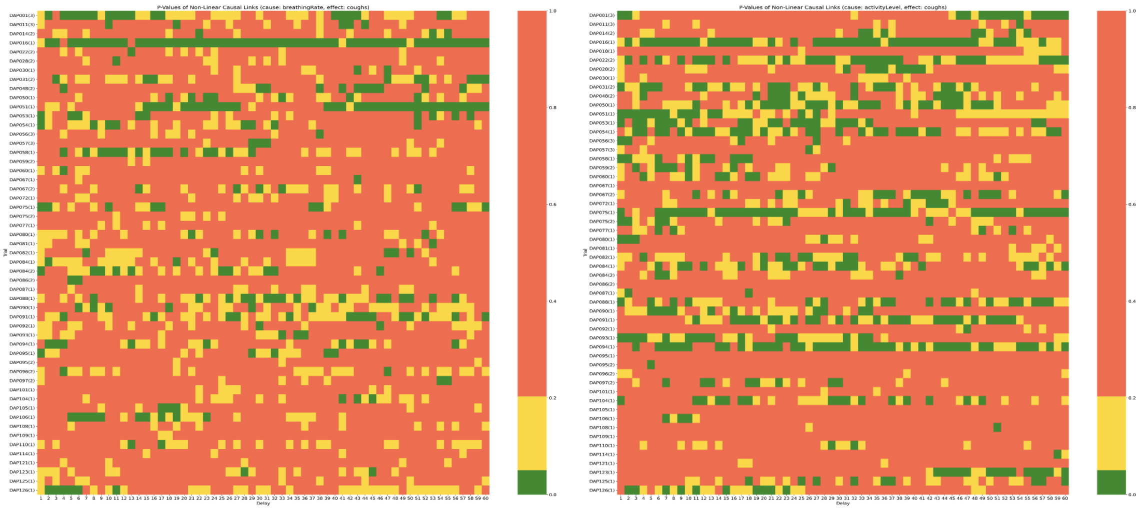


Figure 5.9: The DAPHNE result heatmap for non-linear dependencies between Respiratory Rate(Left) / Activity Level(Right) and coughs for each trial using lag lengths ranging from 1 to 60 minutes.

In this section, a deeper analysis will be conducted on the association between coughing and the activity level and respiratory rate captured by RESpeck, as well as the temperature and humidity measured by AIRSpeck.

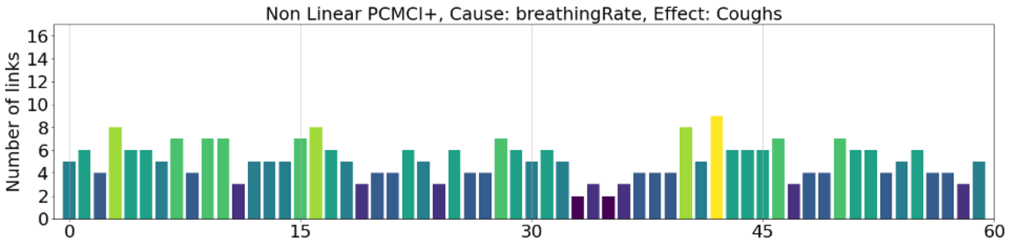


Figure 5.10: Distribution of non-linear causal links across all lag lengths and all trials for DAPHNE study(Cause: Respiratory Rate, Effect: Coughs).

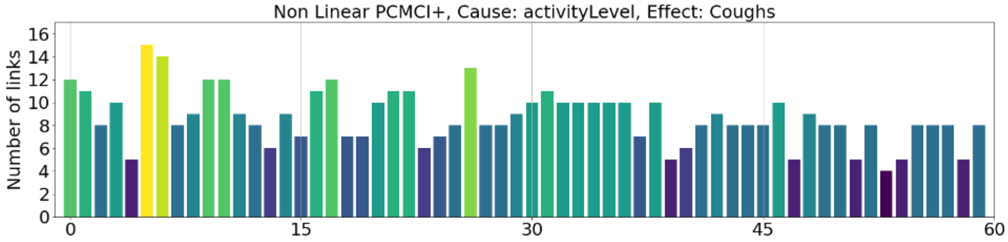


Figure 5.11: Distribution of non-linear causal links across all lag lengths and all trials for DAPHNE study(Cause: Activity Level, Effect: Coughs).

The conclusions of the DAPHNE study are presented in Fig 5.9, where the cause variables, including respiratory rate (left graph) and activity level (right graph), and the effect variable, cough, are depicted. Fig 5.9 significant finding is that the number of causal links to cough associated with activity level (green bars) is substantially greater than those associated with respiratory rate. Fig 5.10 and Fig 5.11 display the distribution of the number of causal links corresponding to each variable. The average number of causal links associated with activity level is over 10 during the first hour, while respiratory rate is only 6. Both variables have a greater number of causal links to cough than air particles, which is likely due to the fact that during high-intensity activity, breathing becomes more rapid to supply more oxygen to the lungs, thereby exacerbating respiratory inflammation and allergic reactions and increasing the incidence of cough. Moreover, a noticeable peak around the fifth minute is evident in Fig 5.11, suggesting that activity level has a relatively high short-term number of causal links to cough, which is consistent with real-world situations [41, 15].

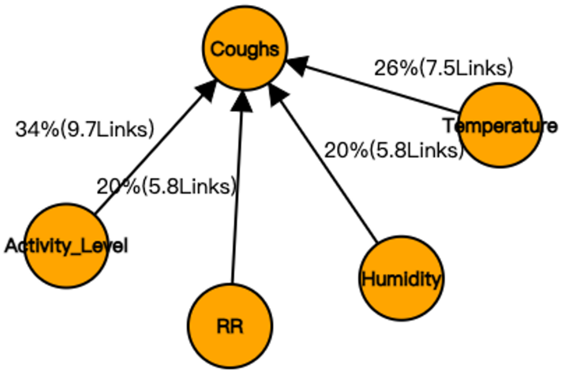


Figure 5.12: Links Intensity for temperature, humidity, activity level and respiratory rate.

In addition to the activity level and respiratory rate recorded by the RESpeck, there are still two features that need to be further analyzed, which are the impact of temperature and humidity on coughing as measured by the AIRSpeck. Fig 5.12 shows the total number of causal links among these four additional features. Among them, the impact of activity level on coughing is the greatest, with an average of 9.7 significant links within one hour, accounting for 34% of the total. Temperature has 7.5 links, accounting for 26%, while humidity and respiratory rate have the same number of causal links, with 5.8 links each, accounting for 20%.

The study presented the number of causal links of 16 bins and other features in Fig 5.13. The research found that activity level and temperature are the most influential factors on coughing. This may be because changes in physical activity and environmental temperature can lead to respiratory system maladaptation and increased load, resulting in coughing [15]. Moreover, previous studies have also shown that environmental factors such as temperature and humidity are related to respiratory system health, and high temperature and low humidity environments may increase the risk of respiratory infections [48].

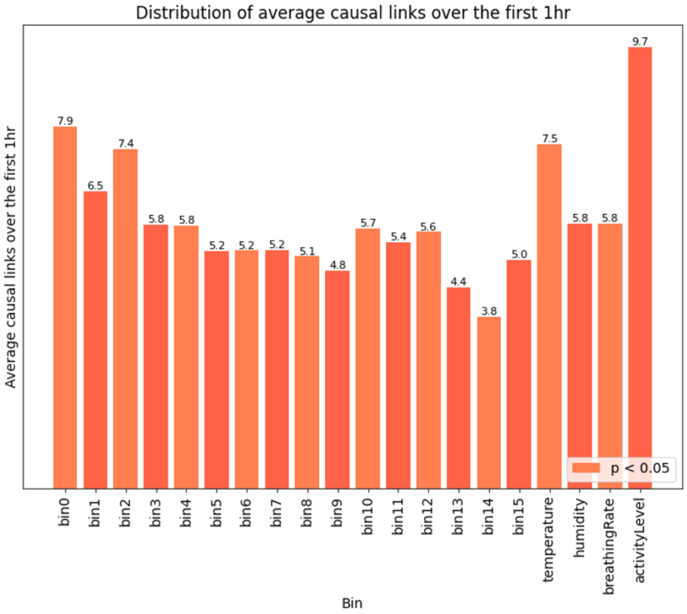


Figure 5.13: Number of causal links for all features.

Furthermore, because small particles have a strong negative correlation with temperature, their number of causal links on coughing is much greater than that of larger particles. Small particles can penetrate the body's defense system and enter the lungs, which can cause inflammation and irritation of the respiratory system, leading to coughing. Therefore, when controlling coughing, it is necessary to consider the impact of environmental factors and particles [5].

In summary, this information is essential for understanding the causes and control factors of coughing. By understanding and controlling environmental factors and physical activity levels, the risk of coughing can be reduced. Additionally, when addressing coughing issues, it is important to consider the impact of small particles and take appropriate prevention and control measures.

5.3 Directed Acyclic Graph Causal Model

The Directed Acyclic Graph (DAG) Causal Model is a graphical representation method used for modeling causal relationships between variables. DAG employs directed acyclic graphs (DAGs) to represent causal relationships between variables, with arrows pointing from the cause to the effect. The purpose of this approach is to explain observed data by identifying causal relationships and to use these relationships for prediction and causal inference.

To investigate the effects of air pollution on human health, specifically on the respiratory system, DAG can be used to represent possible causal relationships. For instance, it may be hypothesized that air pollution causes coughing. In this case, air pollution is considered the cause and coughing the effect, with an arrow pointing from air pollution to coughing to signify the causal relationship. It can be used to observe the overall associations between variables in the PCMCI+ causal network and to discover both short and long-term causal relationships between variables based on different lag times.

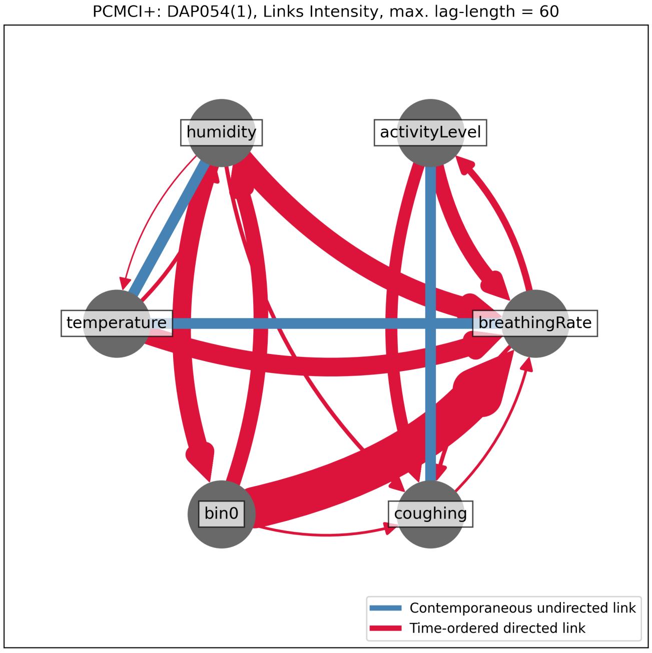

PCMCI+: DAP054(1), Links Intensity, max. lag-length = 60
humidity
activityLevel
temperature
breathingRate
bin0
coughing
Contemporaneous undirected link
Time-ordered directed link

Figure 5.14: DAG containing information about frequency of links in trial DAP054(1) (Maximum lag = 60 mins).

Fig 5.14 shows an example of DAP054(1), which was chosen due to its good data quality (missing values are less than 10%, and there are more cough patterns). The visualization of its time series data is included in Appendix B.7. The figure displays a DAG with a maximum lag time of 60 minutes. To ensure the readability of the figure and avoid excessive linkages, only the causal relationship with the highest strength for the particle count feature (bin0) was selected to represent it, while other features such as temperature, humidity, activity level, respiratory rate, and cough were added as nodes.

The figure contains two types of links. The red time-ordered directed link represents an element that occurs earlier than another element in time. In the DAG, this type of link is typically indicated by an arrow, and in this particular DAG, the thickness of the arrow is adjusted based on the frequency of links. The blue contemporaneous undirected link represents two elements that are not temporally related and can occur independently or simultaneously. This type of link is usually represented by a line segment in the DAG. These links help us understand and analyze the relationships among a set of elements.

From Fig 5.14, we can observe that there are four arrows pointing to cough, which correspond to bin0, humidity, activity level, and respiratory rate. Among them, activity level has the most links, with ten links in the 60-minute interval. To further analyze the distribution of these ten links and identify the short- and long-term effects, the 60-minute interval was divided into four subplots with lag times of 5, 10, 30, and 45 minutes, as shown in Fig 5.15.

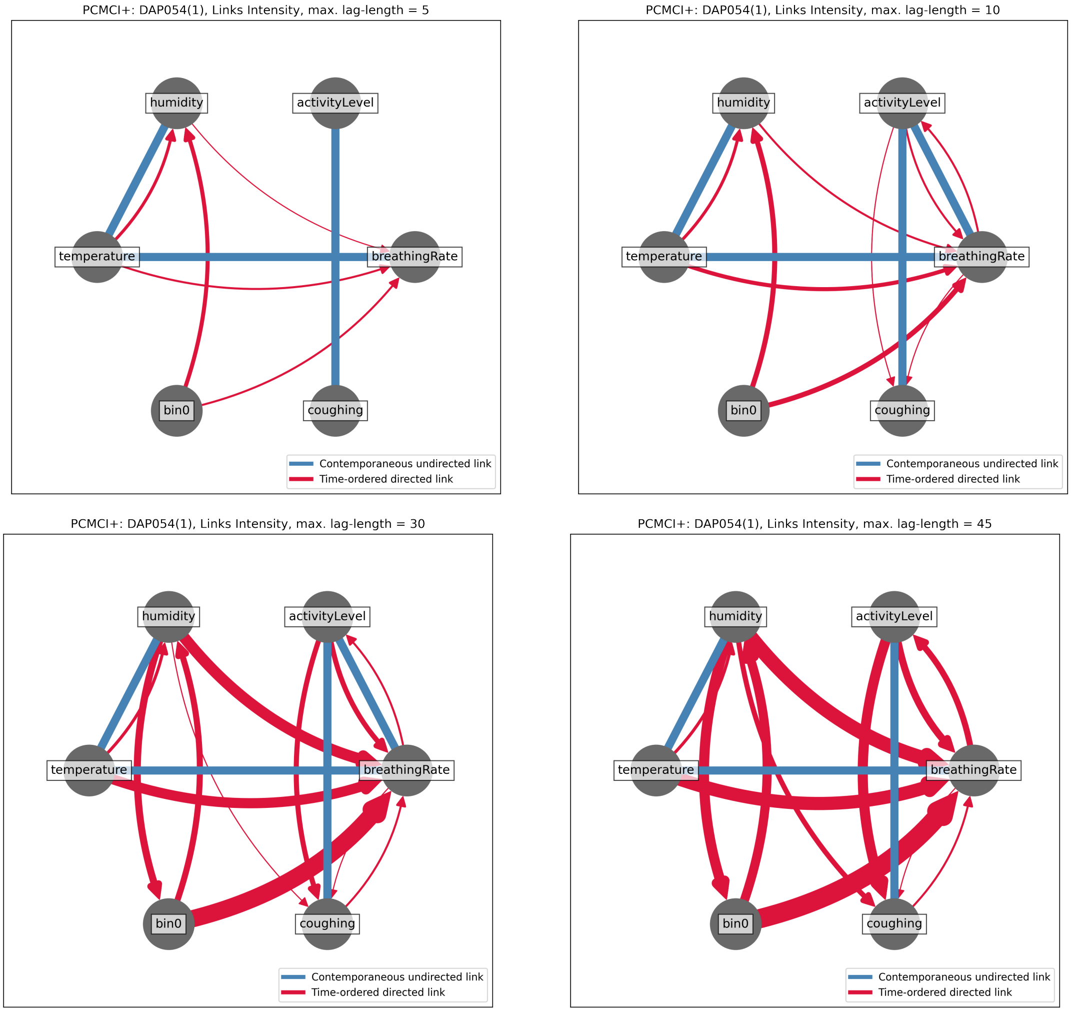

PCMCI+: DAP054(1), Links Intensity, max. lag-length = 5
PCMCI+: DAP054(1), Links Intensity, max. lag-length = 10
PCMCI+: DAP054(1), Links Intensity, max. lag-length = 30
PCMCI+: DAP054(1), Links Intensity, max. lag-length = 45
humidity
activityLevel
temperature
breathingRate
bin0
coughing
Contemporaneous undirected link
Time-ordered directed link

Figure 5.15: DAG containing information about frequency of links in trial DAP054(1) (Lag = 5, 10, 30, 45 mins).

Fig 5.15 illustrates the DAGs for different lag times, which were all generated using the PCMCI+ algorithm once on a single dataset. It is observed that within the first 5 minutes, there is an instantaneous link between coughing and activity level, while there is no apparent factor affecting coughing at other time points, indicating that only vigorous physical activity can influence coughing levels at an instantaneous time scale. As the lag period increases, two links between respiratory rate and coughing are observed within 10 minutes, and humidity levels also begin to affect coughing within 30 minutes. A possible explanation for this observation is that changes in humidity levels can affect the moisture content of the respiratory tract, thereby influencing cough reflex [43]. It is noteworthy that since DAGs are based on specific trials, the causal relationships observed may vary across individuals due to differences in their physiological and environmental conditions. Therefore, the investigation of these causal relationships needs to account for the realistic factors such as individual variability and environmental conditions.

5.4 Validation

In the previous section on preprocessing, it was mentioned that the cough data used in this study was obtained from a previous laboratory experiment. The data was collected by RESpeck sensors and used to train a machine learning model. Coughing can cause intense body tremors in the wearer, and normal intense movements may be wrongly identified as cough signals, leadings to false classification. Despite using three different algorithms to minimize the number of false positives, this issue cannot be entirely resolved.

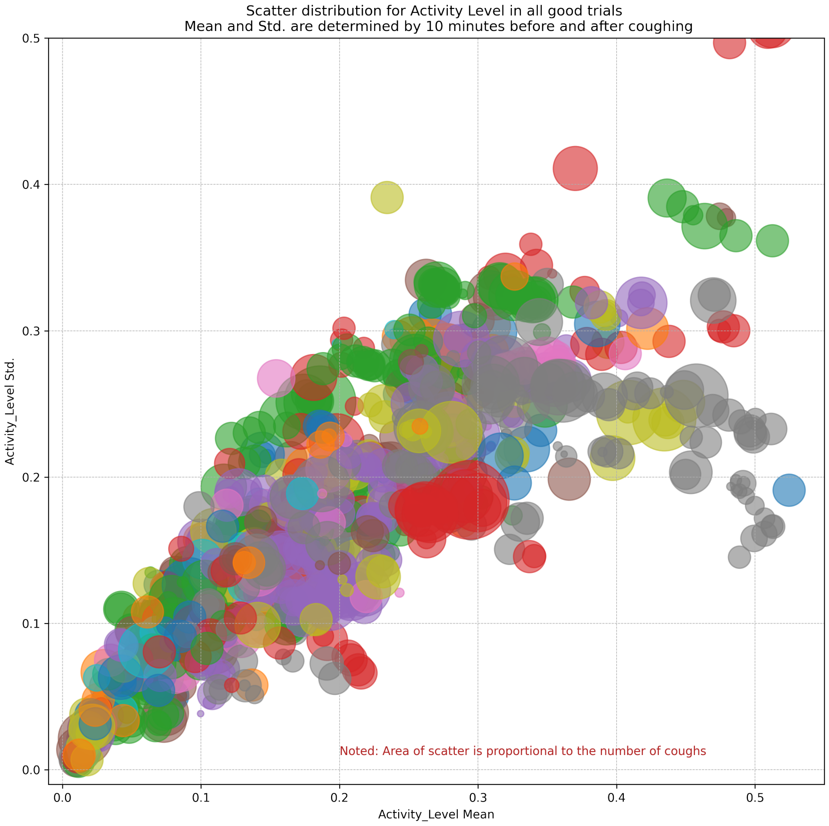


Figure 5.16: Scatter distribution for Activity Level in all good trials. Mean and Std. are determined by 10 minutes before and after coughing (Noted: Area of scatter is proportional to the number of coughs).

One way to validate the experimental results is to observe the wearer's activity level before and after coughing. Typically, the wearer may tend to be calm before coughing, and coughing may cause breathlessness and induce more coughing. Fig 5.16 presents the results of the INHALE study, where different colors represent different participants. Each scatter plot in the figure represents a cough, with larger scatter plots indicating more coughs in that minute. The x and y axes respectively represent the mean and Std. activity levels, which were calculated based on the activity level of the 10 minutes before and after coughing.

The analysis identifies a positive relationship between activity levels and the number of coughs, as evidenced by the increase in scatter size as activity level rises. This finding confirms a link between physical activity and coughing, where higher levels of activity may trigger or exacerbate coughing.

Furthermore, the analysis indicates that as activity level increases, the corresponding standard deviation also increases. This trend suggests that higher levels of physical activity may lead to greater variability in activity levels and consequently increase the likelihood of coughing. Additionally, physical stress and strain associated with higher levels of activity may contribute to coughing.

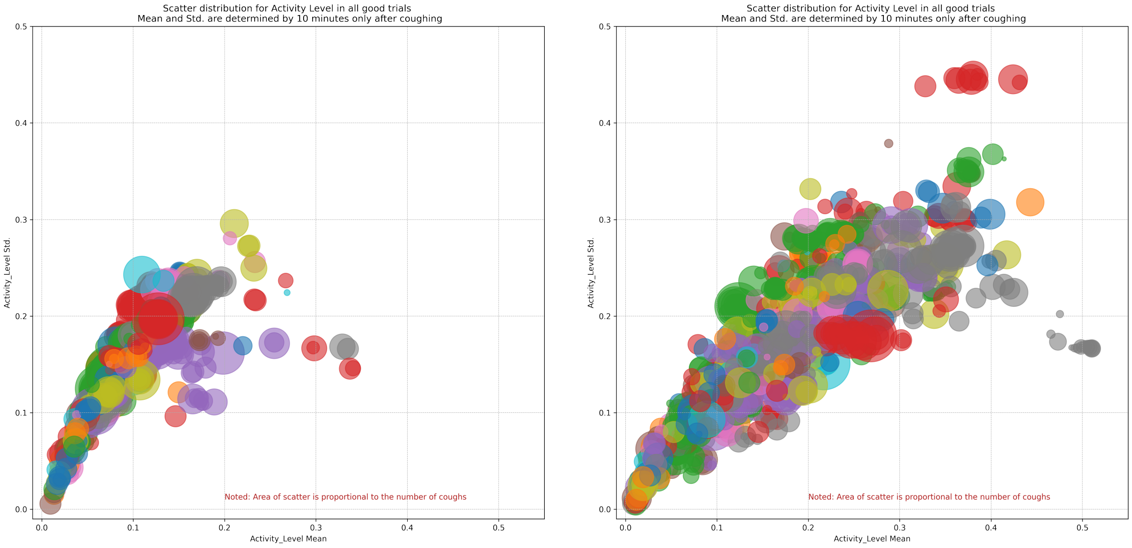


Figure 5.17: Scatter distribution for Activity Level in all good trials. Mean and Std. are determined by 10 minutes only before(left) or after(right) coughing.

To further observe the differences before and after the occurrence of coughing, Fig 5.17 displays the changes in the mean and standard deviation of activity levels corresponding to 10 minutes before and after coughing. It is evident that the activity level is relatively low before coughing, and the subject's state is relatively stable. In contrast, after coughing, the activity level increases significantly, and the standard deviation grows several times, providing strong evidence of the accuracy of cough detection. These findings confirm the PCMCI+ conclusion presented earlier, demonstrating the significant causal impact of activity levels on coughing.

Chapter 6

Conclusions

6.1 Discussion

In this report, a comprehensive analysis of the data provided by the DAPHNE study, consisting of 222 trials from 127 asthmatic adolescents, as well as 30 trials from 15 older asthmatic patients from the INHALE study, was conducted. Both studies utilized the same sensors, AIRSpeck Personal and RESpeck, which record personal exposure to airborne particulates and various parameters such as respiratory rate, flow/effort, and the intensity and type of physical activity.

First, an exploratory data analysis (EDA) was performed on all the datasets. By employing visualization and summary statistics, features, patterns, and anomalies in the sensor data were identified, and each variable was statistically analyzed to guide and support subsequent data analysis and modeling, with the aim of identifying relevant variables for the causal network. Furthermore, a preliminary Kendall correlation experiment was conducted on all features, which confirmed a strong linear correlation between adjacent bins and the PM values matching their diameter range.

Using the statistical features obtained from EDA, the data was preprocessed comprehensively to reduce experimental errors as much as possible. Calibration was performed to ensure comparability of data collected from different regions, Anomalous Detection removed outliers, and data was subjected to strict linear interpolation by applying Winsorizing. In addition, cough signal data was extracted from previous studies and reasonably divided to accelerate subsequent algorithms.

In this project, a newly proposed causal discovery method, PCMCI+, was used to evaluate the causal relationship between particle numbers and coughing in the long and short term. PCMCI+ is a method for discovering causal relationships between time series. Nonlinear tests were used for all experiments in this report because they can more accurately capture the causal relationship between time series, as they can handle nonlinear relationships, any distribution of data, and use more complex and comprehensive methods. This required a large amount of computational power and a significant amount of time. The maximum lag time was selected as the previous 1 hour at 1-minute resolution, to discover the long and short-term effects of coughing. The results showed that all 16 different diameter particles had a strong causal relationship with respiratory rate.

The conclusion of the study demonstrates that there is a significant difference in the overall number of causal links between the 16 bins within one hour, with both the DAPHNE and INHALE studies showing that as particle diameter increases, the overall impact on asthmatic patients becomes weaker. Furthermore, smaller particles have a stronger short-term effect, with their number of causal links gradually decreasing after 20 minutes. A similar trend was observed with PM values, with PM_1 exhibiting the most dramatic short-term causal response (peaking in the first 10 minutes), followed by $PM_{2.5}$ (peaking in 10-15 minutes), while the response of PM_{10} was relatively slower (peaking after 25 minutes). In addition to air particles, other confounding factors were also studied, such as humidity, temperature, and activity level, as well as respiratory rate, to assess their causal impact on coughing. The findings indicate that the causal impact of activity level was the strongest among all features (with an average link count of 9.7 in the first 60 minutes), followed by bin0 (link count of 7.9).

Furthermore, DAGs were drawn for different maximum lag times to compare the short-term and long-term effects of the aforementioned features. Finally, the study confirms the PCMCI+ conclusion presented earlier, demonstrating the significant causal impact of activity levels on coughing, and confirms the conclusions of the case presented in the background through a comparative analysis.

6.2 Limits and Future Works

The background of this report considers numerous potential confounding factors that could affect subjects' responses, such as temperature, humidity, and breathing rate. Despite accounting for these factors, it is still challenging to ensure that other hidden factors do not impact breathing rate [38]. Therefore, including these complex factors in a causal discovery network could result in more accurate conclusions. This approach can help researchers better understand subjects' responses and aid in developing more effective treatment methods.

Certain deep learning techniques, such as Long Short-Term Memory (LSTM), have been specifically developed for processing time series data and can capture long-term dependencies in such data. This makes it an essential tool in causal network analysis. It maintains an internal state vector to handle short- and long-term memory and can selectively forget or store previous states to adapt to different time scales. This enables it to identify causal relationships in sequential data, making it suitable for building causal networks.

Furthermore, as the PCMCI+ method can only handle single time-scale data at present, future work could explore the application of PCMCI+ to multi-scale time series analysis. Data could be divided into different time scales and analyzed at each scale to discover patterns and relationships at different scales. In time series data analysis, multi-scale analysis can be used to explore causal relationships, periodicity, trends, and more at different time scales. Classic methods such as wavelet analysis, time-frequency analysis, and fractal analysis can be used to visualize causal relationships at different scales [16].

Bibliography

[1] 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.

[2] 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.

[3] Surinder S Birring and Arietta Spinou. How best to measure cough clinically. *Current opinion in pharmacology*, 22:37–40, 2015.

[4] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[5] Robert D Brook, Sanjay Rajagopalan, C Arden Pope III, Jeffrey R Brook, Aruni Bhatnagar, Ana V Diez-Roux, Fernando Holguin, Yuling Hong, Russell V Luepker, Murray A Mittleman, et al. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the american heart association. *Circulation*, 121(21):2331–2378, 2010.

[6] Yutong Cai, Tamara Schikowski, Martin Adam, Anna Buschka, Anne-Elie Carsin, Benedicte Jacquemin, Alessandro Marcon, Margaux Sanchez, Andrea Vierkötter, Zaina Al-Kanaani, et al. Cross-sectional associations between air pollution and chronic bronchitis: an escape meta-analysis across five cohorts. *Thorax*, 69(11):1005–1014, 2014.

[7] LY Chan and Helen WY Wu. A study of bus commuter and pedestrian exposure to traffic air pollution in hong kong. *Environment international*, 19(2):121–132, 1993.

[8] Chris Chatfield. *The analysis of time series: an introduction*. Chapman and hall/CRC, 2003.

[9] James Chen. Learn about skewness, Jul 2021.

[10] Ke Chen, Sarah E Metcalfe, Huan Yu, Jingsha Xu, Honghui Xu, Dongsheng Ji, Chengjun Wang, Hang Xiao, and Jun He. Characteristics and source attribution of pm2. 5 during 2016 g20 summit in hangzhou: efficacy of radical measures to reduce source emissions. *Journal of Environmental Sciences*, 106:47–65, 2021.

[11] Renjie Chen, Peng Yin, Xia Meng, Lijun Wang, Cong Liu, Yue Niu, Yunning Liu, Jiangmei Liu, Jinlei Qi, Jinling You, et al. Associations between coarse particulate matter air pollution and cause-specific mortality: a nationwide analysis in 272 chinese cities. *Environmental health perspectives*, 127(01):017008, 2019.

[12] Kian Fan Chung and Ian D Pavord. Prevalence, pathogenesis, and causes of chronic cough. *The Lancet*, 371(9621):1364–1374, 2008.

[13] Douglas W Dockery, C Arden Pope, Xiping Xu, John D Spengler, James H Ware, Martha E Fay, Benjamin G Ferris Jr, and Frank E Speizer. An association between air pollution and mortality in six us cities. *New England journal of medicine*, 329(24):1753–1759, 1993.

[14] Peter Gehr and Joachim Heyder. *Particle-lung interactions*. CRC Press, 2000.

[15] Chinmayee Goda. *Transcriptional regulation of lung diseases by Fox proteins*. PhD thesis, University of Cincinnati, 2020.

[16] Aslak Grinsted, John C Moore, and Svetlana Jevrejeva. Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear processes in geophysics*, 11(5/6):561–566, 2004.

[17] Wei-Jie Guan, Xue-Yan Zheng, Kian Fan Chung, and Nan-Shan Zhong. Impact of air pollution on the burden of chronic respiratory diseases in china: time for urgent action. *The Lancet*, 388(10054):1939–1951, 2016.

[18] Michael Gutmann. Data mining and exploration. 2017.

[19] Richard S Irwin, Cynthia L French, Anne B Chang, Kenneth W Altman, Todd M Adams, Elie Azoulay, Alan F Barker, Surinder S Birring, Fiona Blackhall, Donald C Bolser, et al. Classification of cough as a symptom in adults and management algorithms: Chest guideline and expert panel report. *Chest*, 153(1):196–209, 2018.

[20] Richard L Jones and Mary-Magdalene U Nzekwu. The effects of body mass index on lung volumes. *Chest*, 130(3):827–833, 2006.

[21] Maurice George Kendall. Rank correlation methods. 1948.

[22] Steven Kesten, M Reza Maleki-Yazdi, Bruce R Sanders, Janet A Wells, Susan L McKillop, Kenneth R Chapman, and Anthony S Rebuck. Respiratory rate during acute asthma. *Chest*, 97(1):58–62, 1990.

[23] H. Kim, S. Lee, Y. P. Kim, C. H. Kang, and L. S. Chang. Characteristics and sources of pm2.5 and pm3 in seoul, south korea. *Atmospheric Environment*, 200:44–54, 2019.

[24] Aditi Kulshrestha, P Gursumeeran Satsangi, Jamson Masih, and Ajay Taneja. Metal concentration of pm2. 5 and pm10 particles and seasonal variations in urban and rural environment of agra, india. *Science of the Total Environment*, 407(24):6196–6204, 2009.

[25] Corinne Le Quéré, Robbie M Andrew, Pierre Friedlingstein, Stephen Sitch, Judith Hauck, Julia Pongratz, Penelope A Pickers, Jan Ivar Korsbakken, Glen P Peters,

Josep G Canadell, et al. Global carbon budget 2018. *Earth System Science Data*, 10(4):2141–2194, 2018.

[26] Ning Li, Tian Xia, and Andre E Nel. The role of oxidative stress in ambient particulate matter-induced lung diseases and its implications in the toxicity of engineered nanoparticles. *Free radical biology and medicine*, 44(9):1689–1699, 2008.

[27] Cong Liu, Renjie Chen, Francesco Sera, Ana M Vicedo-Cabrera, Yuming Guo, Shilu Tong, Micheline SZS Coelho, Paulo HN Saldiva, Eric Lavigne, Patricia Matus, et al. Ambient particulate air pollution and daily mortality in 652 cities. *New England Journal of Medicine*, 381(8):705–715, 2019.

[28] Sha Liu, Yumin Zhou, Suixin Liu, Xinyu Chen, Weifeng Zou, Dongxing Zhao, Xiaochen Li, Jinding Pu, Lingmei Huang, Jinlong Chen, Bing Li, Shiliang Liu, and Pixin Ran. Association between exposure to ambient particulate matter and chronic obstructive pulmonary disease: results from a cross-sectional study in china. *Thorax*, 72(9):788–795, 2017.

[29] Saul Mcleod. What is kurtosis?

[30] Alyn H Morice. Epidemiology of cough. *Pulmonary pharmacology & therapeutics*, 15(3):253–259, 2002.

[31] World Health Organization et al. Who. air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulphur dioxide. global update 2005. *World Health Organization. Available from: http://www. euro. who. int/__data/assets/pdf_file/0005/786*, 38:E90038, 2006.

[32] World Health Organization et al. Noncommunicable diseases country profiles 2018. 2018.

[33] C Arden Pope III, Richard T Burnett, George D Thurston, Michael J Thun, Eugenia E Calle, Daniel Krewski, and John J Godleski. Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease. *Circulation*, 109(1):71–77, 2004.

[34] Dou Qiao, Jun Pan, Gongbo Chen, Hao Xiang, Runqi Tu, Xia Zhang, Xiaokang Dong, Yan Wang, Zhicheng Luo, Huiling Tian, et al. Long-term exposure to air pollution might increase prevalence of osteoporosis in chinese rural population. *Environmental research*, 183:109264, 2020.

[35] Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 1388–1397. PMLR, 03–06 Aug 2020.

[36] 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.

[37] Evangelia Samoli, Antonis Analitis, Giota Touloumi, Joel Schwartz, Hugh R Anderson, Jordi Sunyer, Luigi Bisanti, Denis Zmirou, Judith M Vonk, Juha Pekkanen, et al. Estimating the exposure–response relationships between particulate matter and mortality within the aphea multicity project. *Environmental health perspectives*, 113(1):88–95, 2005.

[38] Pablo Andreu Sedeno. A data-driven analysis of the impact of short-term exposure to air pollution on the respiratory rate of asthmatic adolescents.

[39] Particulate sensors: PM sensor. Particulate sensors: Pm sensor, Nov 2021.

[40] Huacheng Song. The health effects of size fractions of airborne particles on asthmatic adolescents. Project Report, The University of Edinburgh, 2022.

[41] Woo-Jung Song, Yoon-Seok Chang, Shoaib Faruqi, Ju-Young Kim, Min-Gyu Kang, Sujeong Kim, Eun-Jung Jo, Min-Hye Kim, Jana Plevkova, Heung-Woo Park, et al. The global epidemiology of chronic cough in adults: a systematic review and meta-analysis. *European Respiratory Journal*, 45(5):1479–1481, 2015.

[42] Laren D Tan, Abdullah Alismail, and Barbara Ariue. Asthma guidelines: comparison of the national heart, lung, and blood institute expert panel report 4 with global initiative for asthma 2021. *Current Opinion in Pulmonary Medicine*, 28(3):234–244, 2022.

[43] Hazel Tapp, Lindsay Kuhn, Thamara Alkhazraji, Mark Steuerwald, Tom Ludden, Sandra Wilson, Lauren Mowrer, Sveta Mohanan, and Michael F Dulin. Adapting community based participatory research (cbpr) methods to the implementation of an asthma shared decision making intervention in ambulatory practices. *Journal of Asthma*, 51(4):380–390, 2014.

[44] Christopher Troeger, Brigette Blacker, Ibrahim A Khalil, Puja C Rao, Jackie Cao, Stephanie RM Zimsen, Samuel B Albertson, Aniruddha Deshpande, Tamer Farag, Zegeye Abebe, et al. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet infectious diseases*, 18(11):1191–1210, 2018.

[45] John W Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67, 1962.

[46] Ir win Rs. Madis on jm. the diagnosis and treatment of cough. *N Engl J Med*, 243(23):1715–1721, 2000.

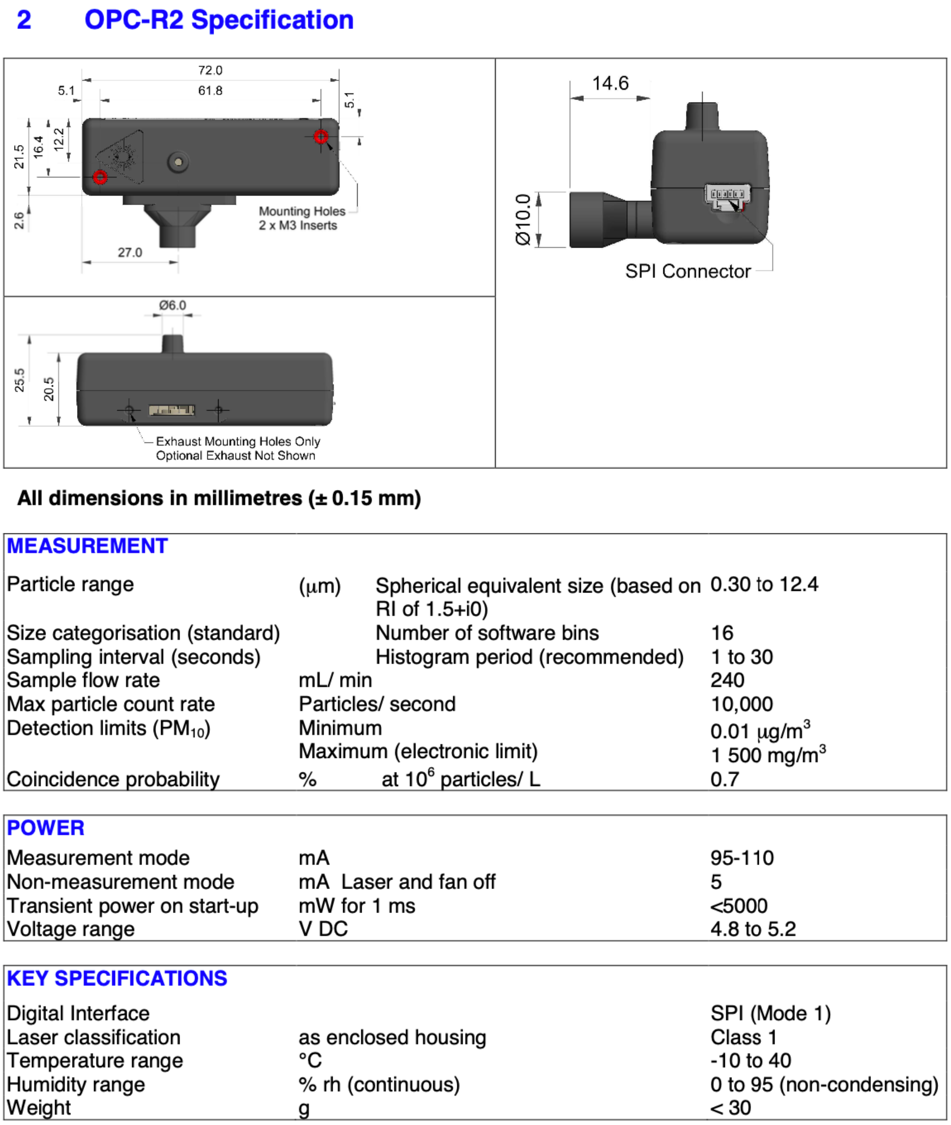
[47] Shao-jin Yang, Jin-quan Dong, and Bing-ru Cheng. Characteristics of air particulate matter and their sources inurban and rural area of beijing, china. *Journal of Environmental Sciences*, 12(4):402–409, 2000.

[48] Liping Yi, Xin Xu, Wenxin Ge, Haibin Xue, Jin Li, Daoyuan Li, Chunping Wang, Haixia Wu, Xiaobo Liu, Dashan Zheng, et al. The impact of climate variability on infectious disease transmission in china: Current knowledge and further directions. *Environmental research*, 173:255–261, 2019.

[49] Junfeng Zhang, Denise L Mauzerall, Tong Zhu, Song Liang, Majid Ezzati, and Justin V Remais. Environmental health in china: progress towards clean air and safe water. *The lancet*, 375(9720):1110–1119, 2010.

Appendix A

Exploratory Data Analysis



72.0
61.8
5.1
5.1
21.5
16.4
12.2
2.6
27.0
Mounting Holes
2 x M3 Inserts
14.6
Ø10.0
SPI Connector
Ø6.0
25.5
20.5
Exhaust Mounting Holes Only
Optional Exhaust Not Shown

2 OPC-R2 Specification

All dimensions in millimetres (± 0.15 mm)

MEASUREMENT			
Particle range	(μm)	Spherical equivalent size (based on RI of 1.5+i0)	0.30 to 12.4
Size categorisation (standard)		Number of software bins	16
Sampling interval (seconds)		Histogram period (recommended)	1 to 30
Sample flow rate	mL/ min		240
Max particle count rate	Particles/ second		10,000
Detection limits (PM_{10})	Minimum		0.01 μg/m^3
	Maximum (electronic limit)		1 500 mg/m^3
Coincidence probability	%	at 10^6 particles/ L	0.7

POWER		
Measurement mode	mA	95-110
Non-measurement mode	mA Laser and fan off	5
Transient power on start-up	mW for 1 ms	<5000
Voltage range	V DC	4.8 to 5.2

KEY SPECIFICATIONS		
Digital Interface		SPI (Mode 1)
Laser classification	as enclosed housing	Class 1
Temperature range	°C	-10 to 40
Humidity range	% rh (continuous)	0 to 95 (non-condensing)
Weight	g	< 30

Figure A.1: OPC-R2 Specification.

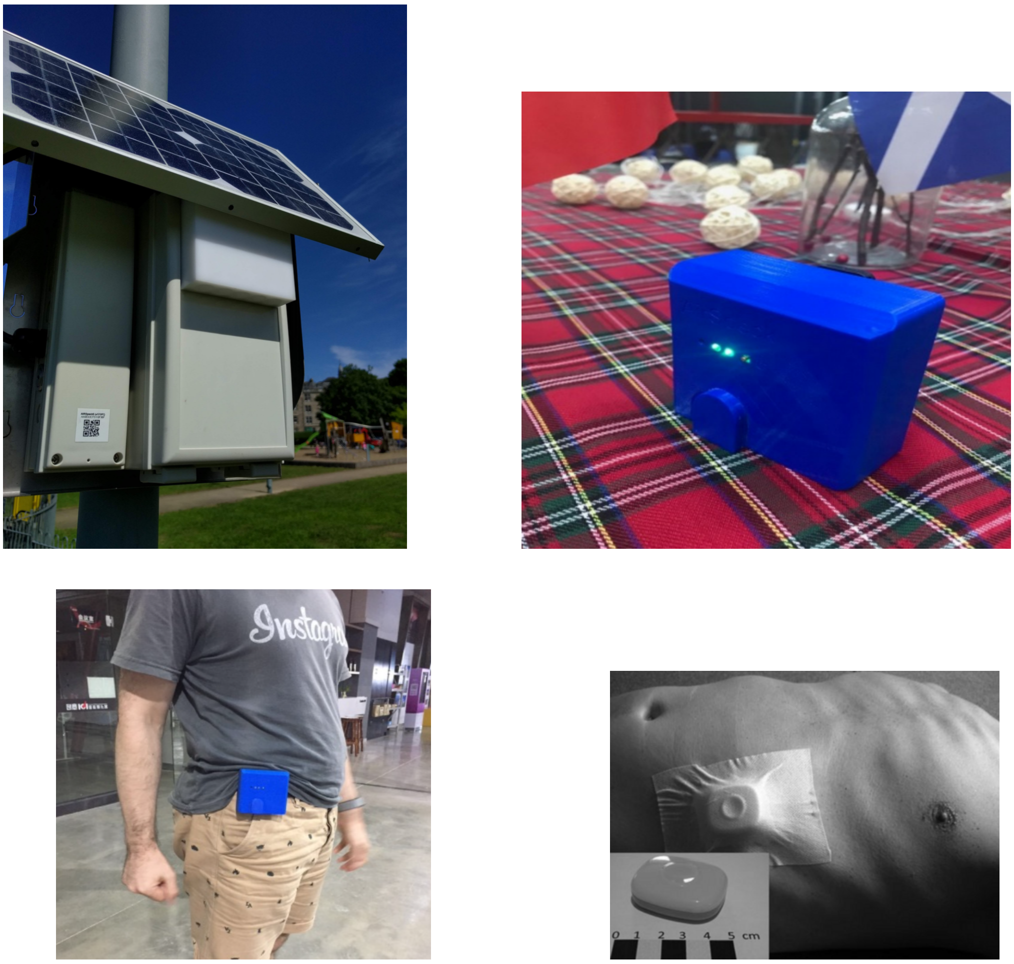


Figure A.2:
Top Left: The stationary AIRSpeck sensor is a stationary device that relies on solar power and can be easily mounted onto typical street furniture, such as poles or lamp posts, as illustrated.

Top Right: The Personal AIRSpeck sensor placed on a table.

Bottom Left: The Personal AIRSPeck sensor is designed to be worn on the belt during observation periods, as depicted in the illustration.

Bottom Right: The RESpeck device is intended to be worn on the skin and affixed with medical tape to facilitate respiratory rate and activity level monitoring.

Appendix B

Causal Discovery Methods

B.1 Causal Results for Airborne Particles

B.1.1 Accumulation and Comparison

B.1.2 Directed Acyclic Graph Causal Model

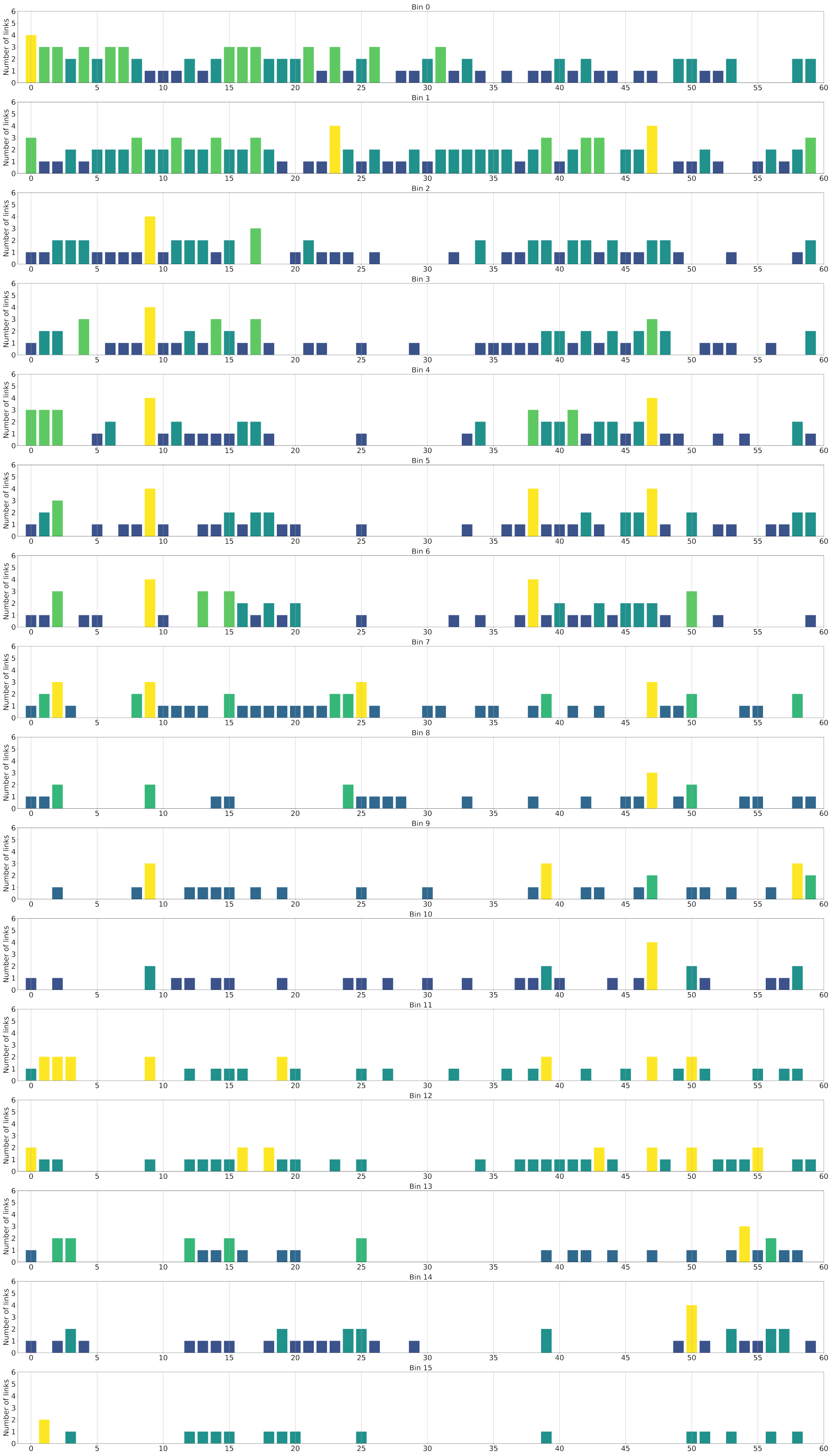


Figure B.1: Distribution of non-linear causal links across all lag lengths and all trials in INHALE study (All 16 bins).

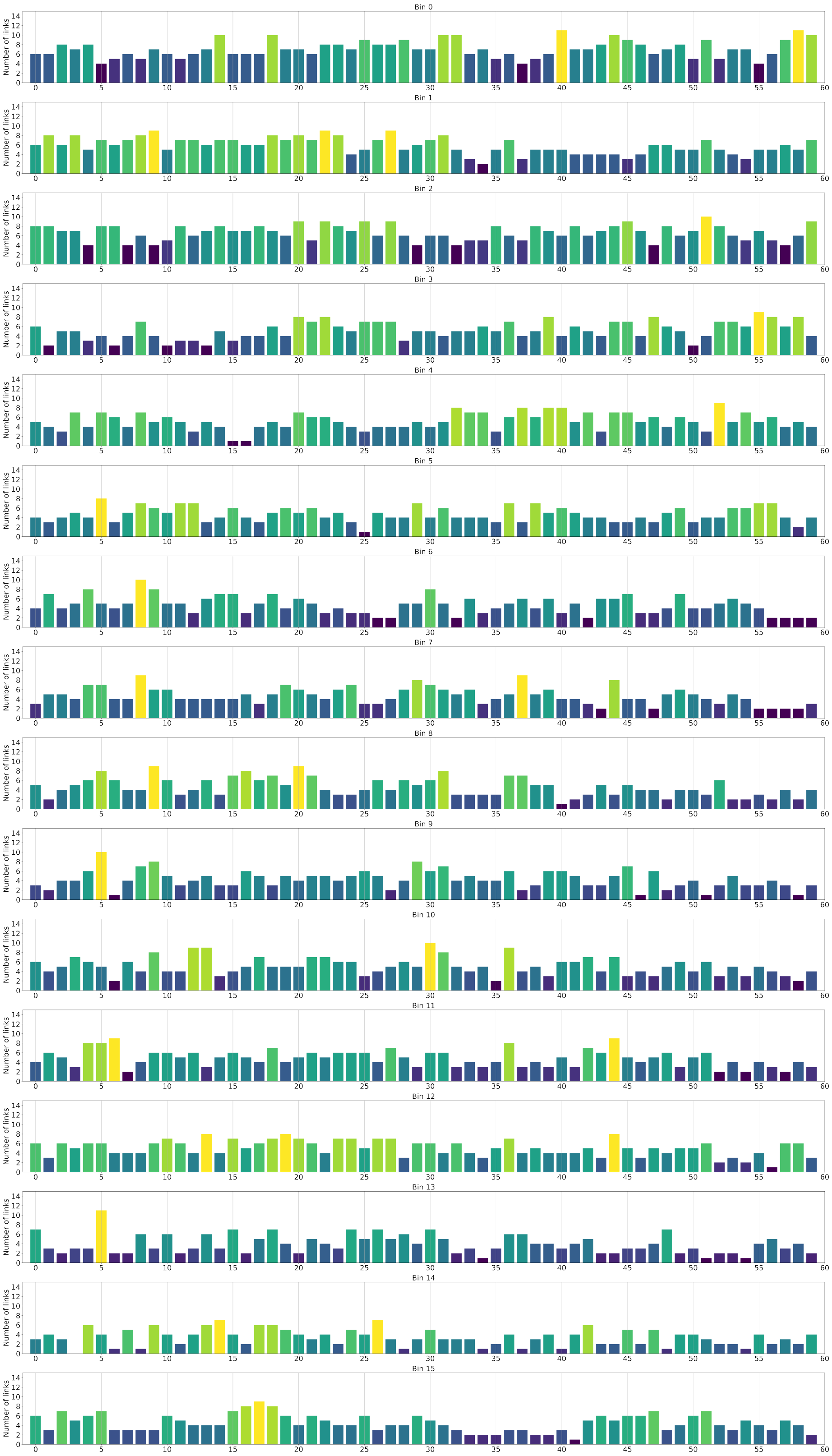


Figure B.2: Distribution of non-linear causal links across all lag lengths and all trials in DAPHNE study (All 16 bins).

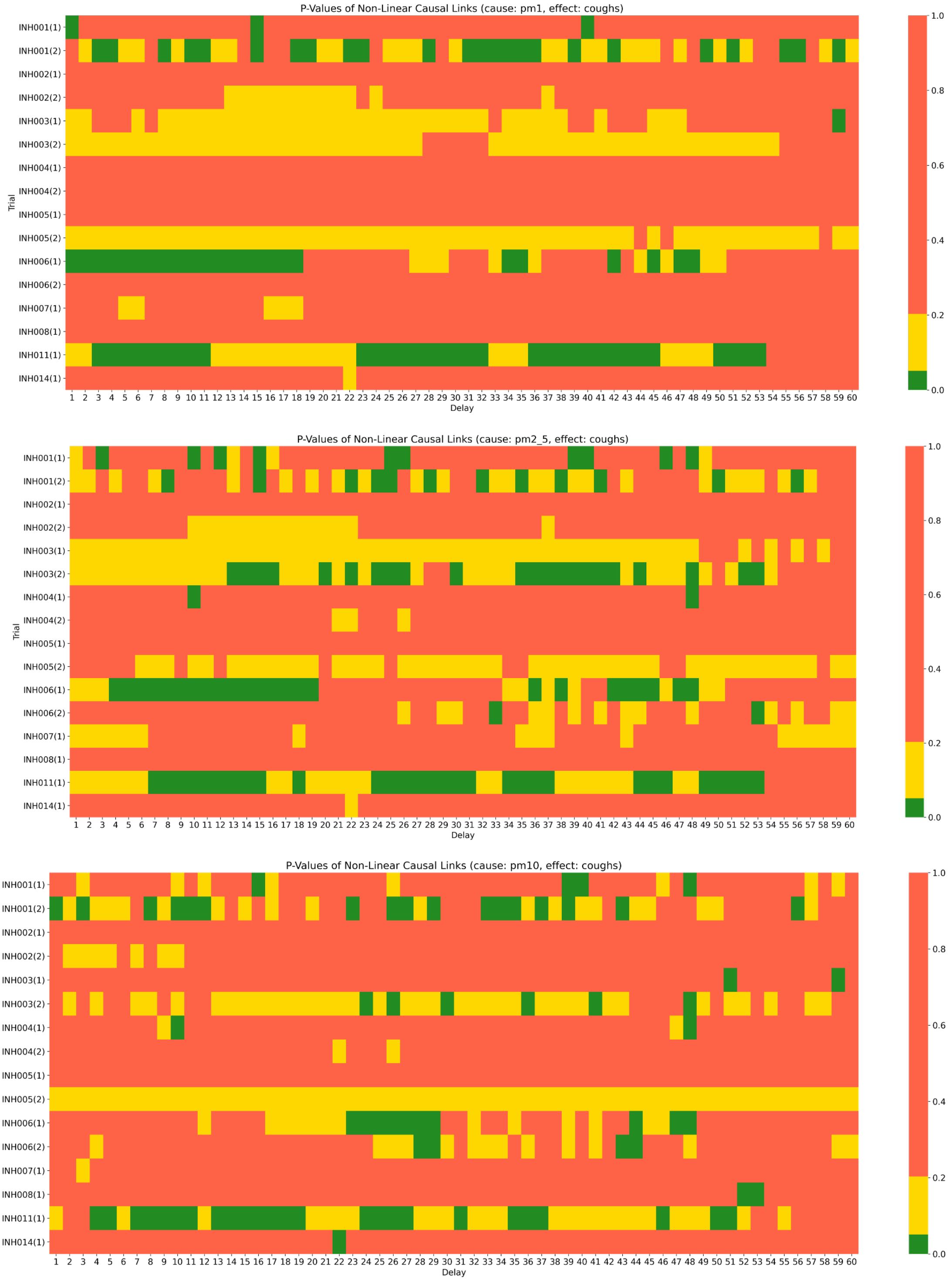


Figure B.3: The INHALE result heatmap for non-linear dependencies between all three PMs (PM$_1$, PM$_{2.5}$, PM$_{10}$) and coughs time series for each trial using lag lengths ranging from 1 to 60 minutes.

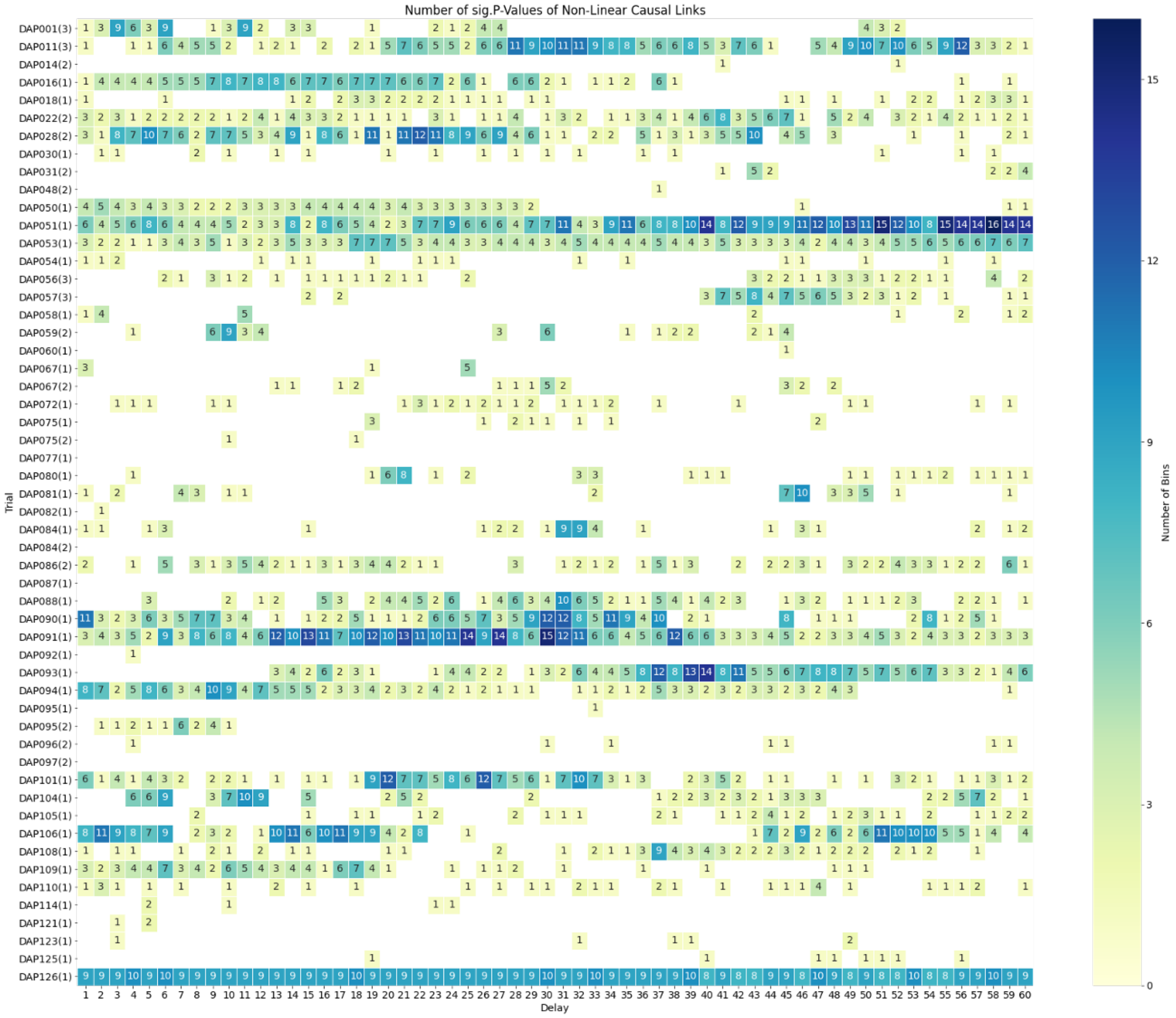


Figure B.4: Results in DAPHNE study for the cumulative number of significant p-values (cumulative for all bin0-15 for all 16 bins).

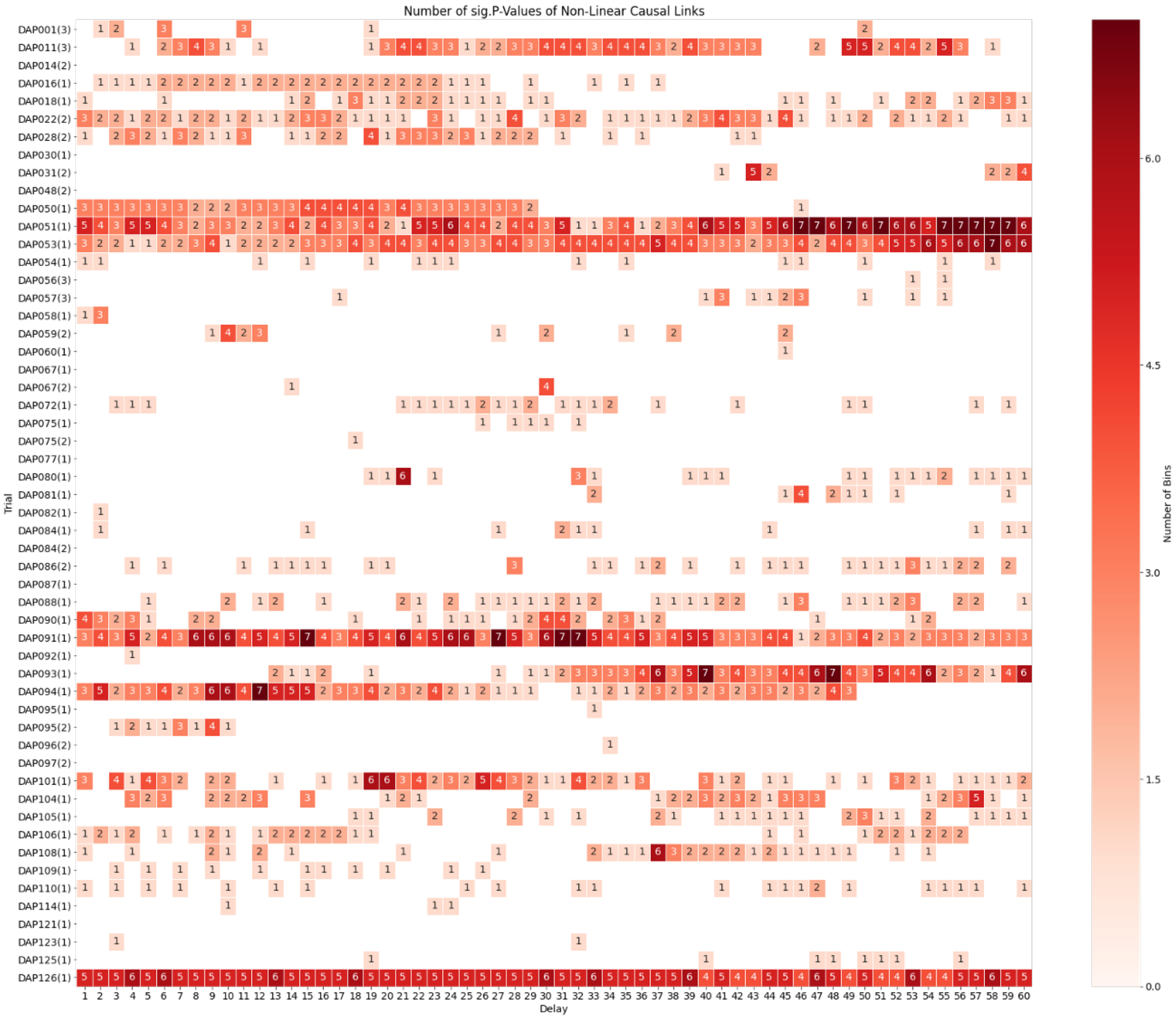


Figure B.5: Results in DAPHNE study for the cumulative number of significant p-values (cumulative for all bin0-6).

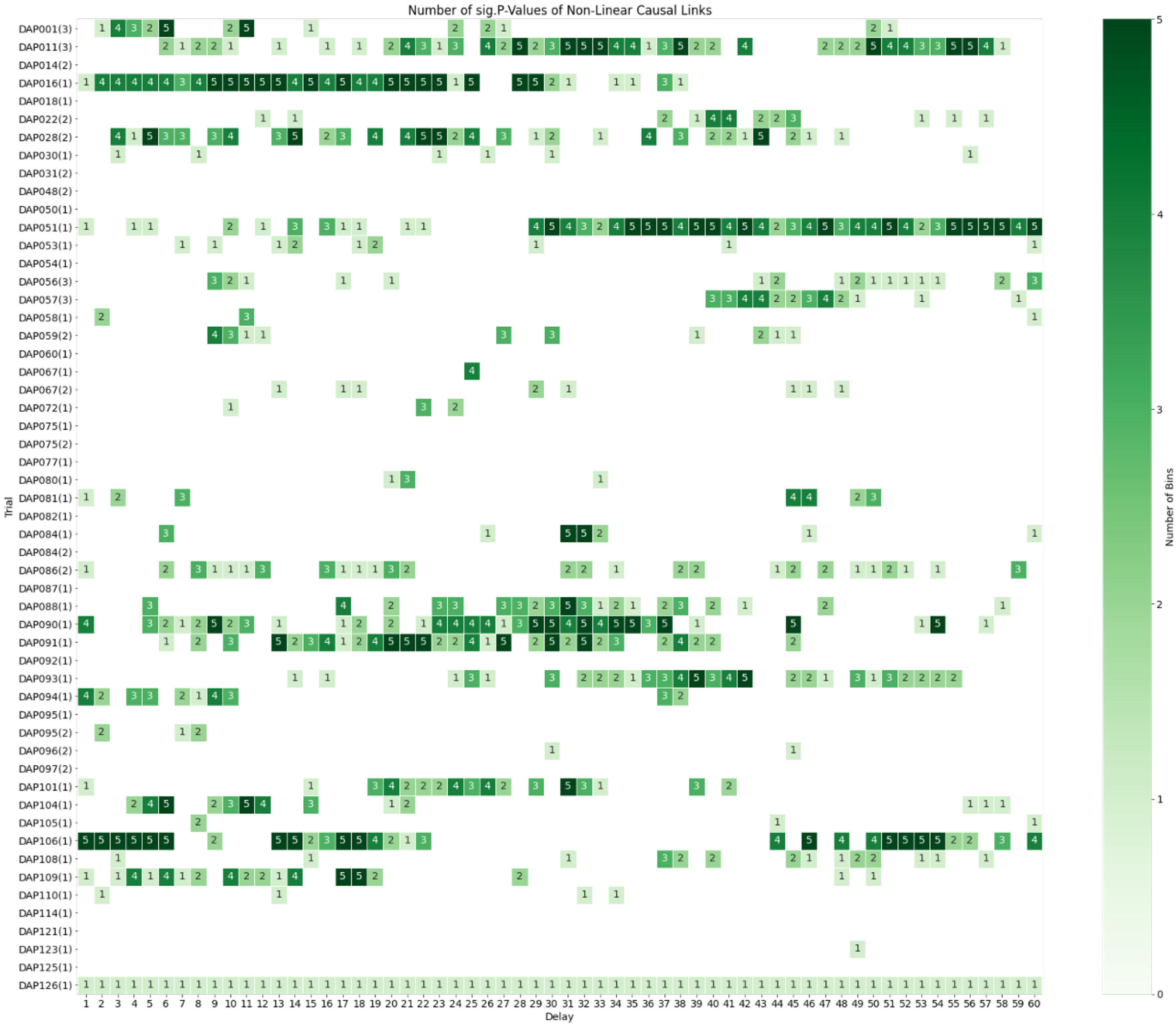


Figure B.6: Results in DAPHNE study for the cumulative number of significant p-values (cumulative for all bin7-10).

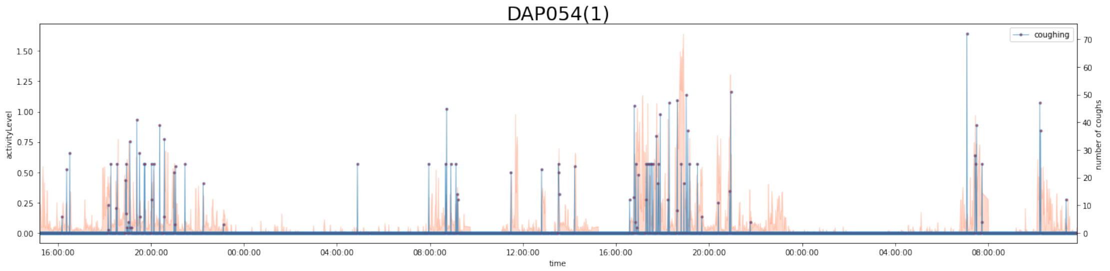


Figure B.7: Example of the number of coughs at two different time periods for DAP054(1), the blue vertical plots represent the number of coughs in subject at a given minute, and their simultaneous activity levels are plotted in coral colours. The two can be seen to show a more obvious positive association.