



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

An Interrupted Time Series Approach to Assess the Effects of Air Pollution on Cardio-Respiratory Activity

TESI DI LAUREA MAGISTRALE IN
MATHEMATICAL ENGINEERING - INGEGNERIA MATEMATICA

Author: **Lucia Gregorini**

Student ID: 10618524

Advisor: Prof. Francesca Ieva

Co-advisors: Dr. Lara Cavinato, Dr. Alessandra Ragni

Academic Year: 2023-24

Abstract

Exposure to air pollution is a major public health concern, with severe consequences on human health and dangerous documented effects on cardiovascular and respiratory systems. This thesis investigates the causal relationship between air pollution exposure and variations in cardio-respiratory activity using an Interrupted Time Series methodology. The study is based on an experimental case conducted in the area of Politecnico di Milano, where a group of participants walked along a predefined path characterized by areas with different environmental conditions. A wearable monitoring system was employed to collect real-time data on physiological parameters, including heart rate and breathing frequency, alongside environmental conditions. Data analysis reveals a significant association between exposure to specific environmental conditions and variations in physiological responses. The application of Segmented Regression allows for the identification of significant changes in physiological trend following exposure to different pollution levels. The findings indicate that individuals experience measurable alterations in cardio-respiratory activity when moving between areas with different environmental conditions. The results of this research demonstrate the effectiveness of Interrupted Time Series analysis in assessing environmental health impacts, and provide strong evidence of the effect of air pollution on human health, highlighting the importance of air quality monitoring and policies aimed at reducing exposure to airborne contaminants.

Keywords: Air Pollution, Interrupted Time Series, Segmented Regression, Cardio-respiratory Health, Environmental Monitoring, Wearable System

Abstract in lingua italiana

L'esposizione all'inquinamento atmosferico rappresenta una delle principali minacce per la salute pubblica, con gravi conseguenze sulla salute umana e pericolosi effetti documentati sul sistema cardiovascolare e respiratorio. Questa tesi analizza la relazione causale tra l'esposizione agli inquinanti e le variazioni nell'attività cardio-respiratoria, utilizzando un approccio basato sulle Interrupted Time Series. Lo studio si basa su un esperimento condotto nell'area del Politecnico di Milano, in cui un gruppo di partecipanti ha seguito un percorso predefinito, attraversando zone con diverse condizioni ambientali. Durante il tragitto, un sistema di monitoraggio indossabile è stato utilizzato per raccogliere dati in tempo reale sui parametri fisiologici, come la frequenza cardiaca e respiratoria, insieme alle condizioni ambientali. L'analisi dei dati rivela una correlazione significativa tra l'esposizione a specifiche condizioni ambientali e le variazioni delle risposte fisiologiche. L'applicazione della Segmented Regression permette di identificare cambiamenti significativi nei trend fisiologici in seguito all'esposizione a diversi livelli di inquinamento. I risultati indicano che gli individui sperimentano alterazioni misurabili nell'attività cardio-respiratoria quando si spostano tra aree con condizioni ambientali differenti. I risultati di questa ricerca dimostrano l'efficacia dell'analisi per Interrupted Time Series nella valutazione degli impatti ambientali sulla salute e forniscono prove concrete dell'effetto dell'inquinamento atmosferico sulla salute umana, sottolineando l'importanza del monitoraggio della qualità dell'aria e di politiche volte a ridurre l'esposizione agli agenti inquinanti.

Parole chiave: inquinamento atmosferico, Interrupted Time Series, Segmented Regression, salute cardio-respiratoria, monitoraggio dell'ambiente, sistema indossabile

Contents

Abstract	i
Abstract in lingua italiana	iii
Contents	v
1 Introduction	1
1.1 Aim of the Analyses	2
1.2 Case Study	4
1.3 Outline of the Thesis	5
2 Literature Review	7
2.1 Air Pollution and Health	7
2.1.1 Common Approaches for Studying Air Pollution and Health	8
2.2 Interrupted Time Series and Segmented Regression in Environmental Health Research	11
2.2.1 Multivariate Interrupted Time Series	12
3 Case Study	15
3.1 A New Wearable System for the Estimation of Personal Air Pollution Exposure	16
3.2 Experimental Campaign	17
3.3 The Dataset	19
4 Methods	23
4.1 Time Series	24
4.2 Interrupted Time Series	24
4.2.1 Critical Issues of Interrupted Time Series	27
4.3 Segmented Regression	29
4.3.1 Limitations of Segmented Regression for ITS	31

4.4	Jittering	33
4.5	Analysis of Outliers	34
4.5.1	Functional Boxplot	36
4.5.2	Outliergram	37
5	Analysis and Results	41
5.1	Explorative Analysis	41
5.2	Outlier detection	50
5.3	Model	55
5.3.1	Choice of the Appropriate Model	56
5.3.2	Jittering Applied over Interrupted Time Series	56
5.3.3	Segmented Regression Applied over Case Study	57
5.3.4	Models with relevant interruptions	64
5.4	Sensitivity analysis	66
6	Discussion and Conclusions	69
	Bibliography	75
A	Appendix A	81
A.1	Evolutions of environmental variables per subject	81
A.2	Comparisons between stations per subject	83
A.3	Traditional boxplots per subject	85
A.4	Magnitude and shape outliers per subject	87
A.5	Models for HR and FB, with single interruption between 2 stations	91
A.6	Sensitivity analysis	93
	List of Figures	97
	List of Tables	99
	List of Acronyms	101
	Acknowledgments	103

1 | Introduction

Atmospheric air pollution has emerged as a significant concern due to its severe impact on both human health and the environment. This issue gained particular attention following several critical pollution episodes that occurred in the mid-20th century such as London's 'Great Smog', which had devastating consequences on the affected population: thousands of people died because of respiratory and cardiovascular disorders caused by the air pollution. An interesting overview of the history of air pollution is provided by [8]. At present, air pollution is recognized as one of the primary environmental factors contributing to premature mortality on a global scale. According to the Organization for Economic Co-operation and Development (OECD) an estimated nine out of ten individuals worldwide are exposed to highly polluted air, with approximately 7 million deaths caused by air pollution every year [2].

Continuous monitoring of air quality is of paramount importance for assessing the composition of the air we breathe daily. This surveillance enables the identification of pollutants and enhances the understanding of their effects on human health. Such insights are crucial for developing new strategies aimed at improving human health and reducing mortality associated with air pollution. For instance, research in this field has contributed to the establishment of regulatory policies by governments, imposing limits on the maximum allowable concentration of pollutant particles in the atmosphere. A notable example is the Clean Air Act in the United States, which has significantly reduced ambient air pollution since its inception in 1970 [1].

Over the years, various technological approaches were introduced in order to monitor

air pollution levels and their components. Among these methodologies, fixed monitoring stations are located on the territory providing measurements that are highly connected to the specific characteristics of the surrounding environment. In addition to these, mobile monitoring stations are employed for short-term campaigns with specific purposes. Satellite-based observations enable to identify larger areas of pollution. Although these technologies provide valuable information regarding outdoor air pollution levels, which are often considered as the primary determinant of personal exposure to pollutants, they are not sufficient to account for precise space-time variability of pollution. Indeed, air quality can drastically change in short distances or short time of intervals owing to its pronounced dependence on both pollution sources and air circulation [12]. For instance, in urban environments characterized by so called "urban street canyons" (narrow streets lined with tall buildings that impede air circulation) pollutant accumulations tends to be significantly higher compared to adjacent areas, which can have completely different pollution levels. As demonstrated by Lu [31], ignoring such spatiotemporal variability of air pollution could lead to exposure misclassification.

In addition to outdoor pollution monitoring, indoor air quality must be monitored as well, as emphasized in the study of Xu et al. [46]. Empirical evidence indicates that people spend most of their time in closed places and the World Health Organization (WHO) reports that 1.6 million premature deaths are caused by indoor air pollution [41]. Given this alarming statistic, it is important to extend air quality assessment efforts to indoor environments too, ensuring a comprehensive understanding of human exposure to air pollution and its associated health risks.

1.1. Aim of the Analyses

Monitoring the quality of the air that individuals breathe is fundamental and the mortality rate associated with elevated levels of air pollution remains alarmingly high, both in indoor and outdoor environments, as previously recalled according to data reported by OECD

[2] and the WHO [41]. Given these critical findings, it becomes essential to focus on the nature of the relationship between what can be defined as an individual's health status and the environment in which they live, in order to determine whether a causal relationship exists between them.

As stated in [17], David Hume provided a definition of the word "cause" that remains one of the most widely adopted today [27][28].

*"We may define a cause to be an object, followed by another, [...] where,
if the first object had not been, the second had never existed."*

- David Hume -

However, it is critical to distinguish causality from correlation. Two variables may exhibit a strong positive relationship, yet this does not necessarily imply a causal relationship between them. In many cases, the observed correlation may be the result of an unobserved third variable, leading to a spurious correlation, which is non-causal [17]. For instance, in his study [32], Messerli identified a correlation between chocolate consumption and the number of Nobel Laureate per country. However, rather than misinterpreting this as a direct causal effect, he attributed the strong association to an unobserved variable, such as the socioeconomic status of individuals, which could simultaneously influence both chocolate consumption and academic achievements leading to a Nobel prize. A truly causal relationship, in contrast, is strictly confined to two variables, where one exerts a direct and significant impact on the other. To establish whether a specific exposure has a causal effect on an outcome of interest we need to know what would happen if a person was exposed, and what would happen if the same person was not exposed. If these outcomes differ, then we can conclude that the exposure is causally related to the outcome [22].

The objective of this study is to determine whether a causal relationship exists between an individual's health status and the environment in which he resides, demonstrating this result through the specific case study: in particular, the evaluation concerns the relationship between air pollution and the cardio-respiratory activity of individuals. Ad-

ditionally, this research aims to identify which specific environmental factors exert the greatest influence on health status, how each of these factor contributes to variations in health-related parameters, and whether health outcomes depend on exposure to specific types of environmental conditions.

1.2. Case Study

This study adopts a case study approach to investigate the relationship between air pollution and cardio-respiratory activity within a controlled setting. This methodology enables a detailed examination of how environmental conditions influence physiological responses by systematically collecting and analyzing relevant data. The study involves monitoring a cohort of individuals as they move through various urban environments, each exhibiting distinct characteristics and differing levels of air pollution. Specifically, the selected walking route comprehends different locations such as parks, streets, gas stations and other urban areas, allowing for an extensive assessment of environmental variability. Throughout the study, physiological parameters, including heart rate and breathing frequency, are continuously recorded alongside environmental variables, such as particulate matter concentration and carbon dioxide levels. Additionally, meteorological parameters including temperature, humidity, and atmospheric pressure, are collected to account for external factors that may influence both environmental conditions and physiological responses. By systematically analyzing these measurements, the study aims to determine whether variations in health-related parameters can be directly correlated with changes in environmental conditions.

Data acquisition is performed using an innovative wearable device, developed and presented by Bernasconi et al. [12], which enables the simultaneous monitoring of both an individual's exposure to air pollution across indoor and outdoor environments and their cardio-respiratory activity. The experimental campaign was conducted in November 2022 within the university district of Politecnico di Milano, following a structured walk-

ing protocol designed to ensure consistency in data collection. Specifically, the walking path includes seven specific stations at which participants are required to pause for five minutes, allowing for controlled measurements under stable conditions.

Due to this structured experimental setup, the collected data naturally take the form of sequential observations recorded over specific time intervals, which can be classified into Time Series (TS) and Interrupted Time Series (ITS). In the case of traditional Time Series, data points are sequentially collected at equally spaced intervals, as observed for environmental parameters, which are recorded continuously both at the designated stations and during the walking segments between consecutive stations. ITS are characterized by the presence of one or more interruptions that segment the sequence into distinct portions. This applies to physiological parameters, which are exclusively recorded at the seven stations, thereby creating clear interruptions in data collection. These interruptions are of particular interest, as they can introduce substantial variations in data trends. Indeed, a primary objective in the analysis of ITS is to evaluate whether the post-interruption data patterns exhibit significant deviations from those observed in the pre-interruption phase, thus assessing the impact of the interruption itself on the recorded physiological responses.

1.3. Outline of the Thesis

This thesis is composed of five chapters beyond this.

Chapter 2 presents a review of the literature, which synthesizes existing research related with the topic of the relationship between air pollution and human health, with a focusing on studies where ITS analysis and Segmented Regression are employed, which is a statistical technique that facilitates the assessment of how an intervention affects the variable of interest, both immediately and over time [45][10]. In Section 2.2.1, an overview of possible methods for Multivariate Interrupted Time Series (MITS) analysis is outlined.

The accurate description of the case study and the acquired dataset is formalized in Chapter 3: it includes the illustration of the device which has recorded the data used in this study in Section 3.1, the characterization of the experimental campaign conducted within the area of Politecnico di Milano in Section 3.2 and the precise description of the dataset obtained after the acquisition of records and a pre-processing step in Section 3.3.

Chapter 4 contains the presentation of the Time Series and Interrupted Time Series frameworks where this work is set, respectively in Section 4.1 and Section 4.2, including the discussion of possible issues deriving from approaching this kind of data structures (Subsection 4.2.1). This presentation is followed by the illustration of Segmented Regression in Section 4.3, which is a powerful method for estimating intervention effects in the analysis of ITS, and the discussion of its limitations. The last part of this chapter includes the explanation of the jittering technique in Section 4.4 and the illustration of some useful tools for the detection of outliers in Section 4.5.

The specific analyses conducted for this work and the related results are presented in Chapter 5. In Section 5.1, an initial explorative analysis of the collected data, including comparative assessments, provides crucial insights into the evolution of various physiological and environmental variables along the path and into the heterogeneity among the stations located along the walk. Section 5.2, the investigation of possible outliers is carried out, aimed at detecting if there are different behaviors of the parameters in different parts of the path. Building on the fundamental observations given from these first analyses and with the aim of addressing the primary research objective outlined above, Segmented Regression is employed: its implementation and consequent results are explained in Section 5.3. In the context of this study, the method is applied to determine whether an environmental transition between two markedly different stations induces a significant and systematic change in a physiological parameter.

Final discussions and conclusions, drawn from the analyses that have been done and the results that have been obtained, are presented in Chapter 6.

2 | Literature Review

2.1. Air Pollution and Health

Air pollution has long been recognized as a significant environmental determinant of health, with extensive research documenting its adverse effects on human well-being and environment. There exist multiple types of pollutant agents which have harmful consequences on human health, particularly on the respiratory and cardiovascular systems. They are produced mostly by vehicles engines and by industrial processes, but their existence can also be caused by natural disasters such as fires and volcanic eruptions. Among the most common and dangerous pollutants there are particulate matter (PM), nitrogen dioxide (NO_2) and ozone (O_3). Particulate matter consists of an heterogeneous mixture of both organic and inorganic substances, existing in the form of solid particles or liquid droplets and covering a wide range of diameters [6]. These particles are classified based on them into distinct categories: Total Suspended Particles, with diameters that are typically $\leq 50 - 100 \mu m$; PM_{10} , which are inhalable particles, with diameters that are generally $10 \mu m$; $PM_{2.5}$, which are finer inhalable particles, with diameters that are generally $2.5 \mu m$; PM_1 , which represents even smaller particles with diameters that are around $1 \mu m$. Nitrogen dioxide (NO_2) is a strong oxidative agent, which reacts with combustible materials. It is produced from all the burning processes at high temperatures involving fossil fuels, such as heating systems, vehicle engines and industrial combustion [4]. Ozone (O_3) is generated by the chemical reaction, facilitated by high temperatures and the presence of solar radiation, between nitrogen oxides and volatile organic compounds (VOCs),

which are usually produced by civil and industrial combustion and by those processes that employ or produce volatile chemicals as solvents and fuels [5]. Exposure to these air pollutants has been associated with increase health sufferings and mortality, particularly due to respiratory and cardiovascular diseases [25][36][13]. Short-term exposure to high pollutant concentrations has been linked to acute health episodes, including the worsening of asthma attack and chronic bronchitis, increased hospital admissions for respiratory and cardiovascular conditions, and higher incidences of strokes and heart attacks [36]. Long-term studies provide strong evidence that chronic exposure to fine particulate matter ($PM_{2.5}$) significantly reduces life expectancy, with estimated reductions of one to two years depending on exposure levels [13]: specifically, some of these pollutants are powerful cancer-causing agents, which could lead to tumors mainly located in the respiratory system [7]. Notably, research suggests that there may be no safe threshold for exposure to $PM_{2.5}$ and O_3 , as even low concentrations contribute to increased health risks.

Vulnerable populations such as children, the elderly and individuals with preexisting respiratory or cardiovascular conditions, face disproportionately higher risks from air pollution exposure compared with the rest of the population. Socioeconomic disparities also play a crucial role in determining health outcomes, with marginalized communities often experiencing higher exposure levels due to residential proximity to industrial or traffic-dense areas. This has significant implications for environmental justice, necessitating targeted policy interventions to mitigate health disparities and reducing health risks for the entire population [13].

2.1.1. Common Approaches for Studying Air Pollution and Health

Since 1997, there has been a substantial amount of research that added to the evidence that breathing fine particulate air pollution is harmful to human health [36]. Various lines of research have helped to understand it, including epidemiological studies, which consist in scientific research aimed at analyzing specific illnesses affecting a population.

Epidemiological studies focused on respiratory and cardiovascular diseases are particularly useful in examining the effect of air pollution on human health. Three of the most common types of studies in this field include time series analyses, cohort studies and case-control studies [36][14].

Time series analyses assess the short-term relationship between air pollution levels and health outcomes, by examining fluctuations in pollutant concentrations over time and correlating them with changes in health indicators, such as hospital admissions and mortality rates. These studies are particularly useful for identifying acute effects of pollution exposure and controlling for confounding factors like seasonal variations and meteorological conditions [14]. They allow researchers to estimate short-term health effects while controlling for seasonal trends. However, time series analyses are limited in their ability to capture long-term health effects and individual level exposure differences [36]. Numerous time-series studies have demonstrated an association between air pollution and daily mortality [40][42][29][43][37][44][24].

Cohort studies follow a group of individuals over an extended period to investigate the long-term health effects of air pollution exposure. These studies provide a strong evidence for causal relationships between pollution and chronic diseases, such as cardiovascular disease and respiratory disorders: exposure to fine particulate matter in the air is associated with life shortening by Dockery et al. [18] and Pope et al. [35], who both investigate the relationship between air pollution and mortality in the US, using observations recorded over a decade. The advantage of cohort studies lies in their ability to account for individual exposure variations and track disease progression over time. However, they require significant resources and long follow-up periods, making them costly and logistically complex. The objective of the work of Dockery et al. [18] is to estimate the effects of air pollution on mortality while controlling for individual risk factors, using data collected within a well-characterized cohort of adults, coming from six different American communities. The study is conducted following a Survival Analysis approach, including Cox

proportional-hazards regression modeling [20]: survival times are calculated by subtracting the date of enrollment from the exact date of death, for people who died during the study, while censored survival times are defined as the date of the end of the study minus the enrollment date for surviving participants. The study provides a significant evidence that the mortality rate, adjusted for health risk factors such as smoking status, sex and average body-mass index, is associated with the level of air pollution: specifically, particulate air pollution is associated with death due to cardiopulmonary causes. Pope et al. [35] conduct a similar study, aimed at evaluating effects of particulate air pollution on mortality using data from a larger cohort drawn from many study areas, as previous studies of particulate pollution and mortality have been limited by small number of subjects or study areas. In this study, adjusted mortality relative risk ratios are estimated using multiple regression analysis based on the Cox proportional hazards model [20], where the time variable used is the survival time from the date of participant's enrollment to the study; survival times of participants who do not die are censored at the end of the study period. The results suggest that the associations observed between particulate air pollution and mortality are not due to confounding by other risk factors: they suggest that air pollutants are the main important contributing factors causing respiratory illness and early mortality due to cardiopulmonary diseases.

Case-control studies compare individuals who have developed a specific disease (cases) with those who have not (controls) to identify differences in past air pollution exposure. These studies are particularly useful to examine rare diseases or conditions with long latency periods, such as lung cancer or chronic obstructive pulmonary diseases [36]. While case-control studies are efficient and cost-effective, they are susceptible to difficulties in accurately reconstructing past exposure levels and to recall bias, which occurs when the participants to the study inaccurately remember or report past exposures, leading to misclassification and potential distortions of study results [13]. Nyberg et al. [33] conduct a population-based case-control study among stable residents of Stockholm who are af-

ected of lung cancer, with the aim of understanding the relationship between lung cancer risk and urban air pollution, specifically detecting which sources of urban air pollution may be of importance. Using geographical information system techniques and dispersion modeling, which is a mathematical simulation that predicts how air pollutants spread, a statistical evidence that air pollution increase lung cancer risk is attested.

2.2. Interrupted Time Series and Segmented Regression in Environmental Health Research

While traditional epidemiological studies have been fundamental in identifying associations between air pollution and health outcomes, they often fall short in evaluating the real-world impact of policy interventions. Measures such as low-emission areas, restrictions on industrial emissions, and traffic congestion charges have been implemented in various regions to mitigate pollution-related health risks. However, assessing their effectiveness requires robust methodological frameworks that go beyond observational studies. ITS analysis offers a valuable approach to quantify the health benefits of these policies by examining changes in disease incidence and mortality before and after their implementation. Various types of interventions have been investigated, as the restriction on the number of prescriptions reimbursed per patient per month [45], policies implemented to study the impact of a program to reduce ADHD medication initiation in children younger than 5 years old [34], or the Italian smoking ban in public places [10].

Segmented Regression analysis of ITS helps in assessing, in statistical terms, how much an intervention changed an outcome of interest, immediately and over time. Specifically, it enables researchers to analyze changes in the outcome before and after policy implementation while accounting for underlying temporal trends. This method is particularly advantageous in air pollution research, where policies such as emission controls or traffic restrictions can be evaluated for their effectiveness in reducing health risks [10].

Bernal et al. [10] present a comprehensive and systematic guide to conduct a robust ITS analysis, starting from the description of its characteristics and the explanation of how to implement the regression model, and emphasizing the strengths of this approach. The study also reviews the main methodological issues associated with ITS studies and the Segmented Regression approach. The illustration of this method is supported by analyses made on a real-world example, evaluating the effects of the Italian smoking ban in public spaces on hospital admissions for acute coronary events. This study offers valuable insights into the practical application of ITS analysis and Segmented Regression.

Wagner et al. [45] place particular emphasis on the interpretability of ITS results. Their study provides a clear and structured explanation of level and trend changes estimated through Segmented Regression, going beyond the statistical computations to explore their practical significance. By following a meticulous step-by-step approach, they highlight both the strengths and limitations of Segmented Regression, also discussing cases involving multiple interventions.

2.2.1. Multivariate Interrupted Time Series

When multiple related ITS are involved in a study rather than a single one, it is appropriate to use a Multivariate Interrupted Time Series analysis, which is an extension of the ITS analysis. One possible scenario involves the simultaneous analysis of several ITS, each representing the temporal evolution of distinct but related variables. In this case, these variables are analyzed to examine and understand their relationships and their response to the intervention. For example, when evaluating the effect of a newly introduced traffic law that reduces speed limits in a city, one might collect time series data on multiple factors, such as the number of car accidents per month, the average driving speed and traffic congestion levels. Analyzing these variables simultaneously allows for a comprehensive assessment of the intervention's effects on the broader traffic system. However, in most applications, the structure of the problem differs. Rather than observing multiple vari-

ables simultaneously, a single variable is recorded multiple times across different units. This occurs, for instance when the variable of interest is measured for each individual participating in a specific study [21], for each patient admitted to a hospital [10], or for each country within a selected subset of nations [15]. In such cases the analysis focuses on managing the heterogeneity existing between the single units, identifying common patterns in the interrupted time series across different subjects, regions or experimental conditions. Several methodologies have been developed to analyze MITS, each with its own strengths and limitations.

Gascoigne et al. [21] propose a Bayesian hierarchical controlled interrupted time series model to evaluate the impact of policy changes, specifically the UKs welfare reform, on participants' mental health, while accounting for spatial and temporal patterns as well as evaluating inequalities indexed by specific variables. Since the policy was introduced at different times across various administrative geographies in England, the study framework incorporates this temporal variation into the analysis by choosing the start of the interventions beforehand in terms of "contextual awareness". The proposed Controlled Interrupted Time Series (CITS) within Bayesian hierarchical framework, follows a typical ITS design. Formally, the response variable is modeled as $y_{it} \sim \mathcal{N}(\mu_{it}, \sigma^2)$, where μ_{it} represents the underlying mean function. This function is modeled through a Segmented Regression which incorporates random effects accounting for residual variations in time and space, as well as for stratification and clustering effects. This method proves naturally effective for evaluating the impact of policy interventions at population level, and it is highly flexible in accounting for spatial and temporal dependencies, and individual and community level characteristics. However, despite its strengths, the method also suffers from some limitations: the accuracy of the results is limited by the choice of the proxy of exposure, and the definition of the start of the intervention remains inherently arbitrary.

Adopting the Bayesian hierarchical methodology for a standard Segmented Regression model from an ITS design, Rana et al. [39] compare the performances of three Bayesian

hierarchical models with a non-hierarchical model (NHM). This comparison is made using real-life data and simulation studies. The hierarchical method has become a standard statistical method for accounting for group-level variability in data analysis. When integrated into ITS designs, it enhances the precision of parameter estimates and improves the statistical power to detect clinically meaningful differences by partitioning the error variance into group-level factors. The study considers three hierarchical models, both with or without measurement error: the Varying Pre-Intervention Intercept Model (VPIM), the Varying Intercept Model (VIM) and the Measurement Error Model (MEM). The comparison among them is made through the Widely Acceptable Information Criteria (WAIC), a standard Bayesian model comparison metric. Simulation results demonstrate that Bayesian hierarchical models for ITS consistently outperform corresponding non-hierarchical Bayesian models. Moreover, incorporating measurement error in the outcome variable results in more precise parameter estimates and improves model performance in most cases. One of the key advantages of the hierarchical modeling approach, which helps providing improved estimates, is its natural ability to mitigate the under-fitting or over-fitting problem of the non-hierarchical model. This characteristic is particularly relevant in MITS analysis, where multiple observations are observed at a particular time point.

Ewusie et al. [19] present a methodological framework within the Segmented Regression modeling strategy, which is called weighted Segmented Regression (wSR). Additionally, they compare it with the traditional Segmented Regression approach. The study is conducted over a real-life dataset, consisting in the measurements of the time spent in out-of-bed by some patients of a hospital. The new model is constructed by introducing weights to account for between-participant variation and the differences across multiple sites: they are calculated as the inverse of the sum of the within-site and between-site variances, and in this way they incorporate both participant-level and site-level variability. The wSR results in having very precise estimates, with very narrow confidence intervals, hence allowing meaningful differences to be detected. Also the statistical power is high.

3 | Case Study

Monitoring air pollution is essential for understanding its impact on human health and several methodologies are available, each offering distinct advantages depending on the scale and purpose of the assessment [12]. Fixed monitoring stations, strategically positioned in urban and industrial areas, provide continuous and reliable measurements over time, making them valuable for long-term trend analysis. For instance, in Italy, within the Lombardy region, the "Agenzia Regionale per la Protezione dell'Ambiente" (ARPA) is responsible for environmental prevention and protection. Among its various duties, it conducts continuous monitoring of the concentrations of major air pollutants, operating 24 hours a day through a network of 85 fixed stations distributed across the entire region [3]. Mobile monitoring stations, often deployed in short-term campaigns, allow for flexible assessments in areas where permanent infrastructure is lacking. ARPA employs 8 mobile laboratories to conduct in-depth assessments of local conditions and specific environmental situations, including those arising from reports submitted by municipal administrations [3]. Additionally, satellite-based remote sensing techniques enable large-scale pollution tracking, offering valuable insights into global and regional air quality patterns. Recent advancements in wearable technology have introduced compact and sensor-equipped devices capable of monitoring personal exposure levels in real-time, capturing the variability in air pollution encountered during daily activities. The integration of these approaches provides a comprehensive understanding of air pollution dynamics and its direct effects on human health.

3.1. A New Wearable System for the Estimation of Personal Air Pollution Exposure

Bernasconi et al. [12] introduce an innovative wearable system designed to monitor an individual's exposure to air pollution across both indoor and outdoor environments, while simultaneously tracking cardio-respiratory activity. The project is conducted within the AdvaNced Technologies for Human-centrEd Medicine (ANTHEM) project, financed by the 'Piano Nazionale di Ripresa e Resilienza'(PNRR): the objective of this project is the contribution to the challenge of reducing the gap between life expectation and healthy life expectation, by developing and implementing new technologies. These tools aim at monitoring patients through new sensors including continuous real-time data collection and integration, at implementing monitoring systems for chronic patients so as to improve early diagnosis by identifying signs of worsening of the health conditions, and at personalizing the medical treatments dedicated to specific target populations and territories.

Complying with the first of these three main objectives, this new system is developed for the accurate assessment of personal pollutant exposure, as it enables real-time monitoring of both the ambient air composition and the respiratory activity. The latter aspect is particularly significant as it allows for a detailed evaluation of the actual pollutants that effectively penetrate the human respiratory system. The newly developed device is designed to be worn either on the wrist or the upper arm, as illustrated in Figure 3.1a from [12].

The more relevant and dangerous air pollutants are particulate matter, nitrogen dioxide and ozone, which have been exhaustively presented in Section 2.1. The newly developed monitoring system is capable of detecting particulate matter, specifically PM_1 , $PM_{2.5}$ and PM_{10} , in addition to other pollutants that significantly impact air quality. These include carbon dioxide (CO_2), total volatile organic compounds ($tVOC$), carbon monoxide (CO) and nitrogen dioxide (NO_2) [12]. Furthermore, key atmospheric parameters are also

detected, providing a more comprehensive evaluation of environmental conditions: in particular they are temperature, humidity and pressure. The environmental monitor node is integrated in a body sensor network (BSN), shown in Figure 3.1b from [12]. This network is specifically designed to collect cardiopulmonary data. The BSN is composed by three electronic device called Inertial Measurement Units (IMUs), positioned one on the thorax, one on the abdomen and one on the back, and a wrist-worn pulse oximeter.

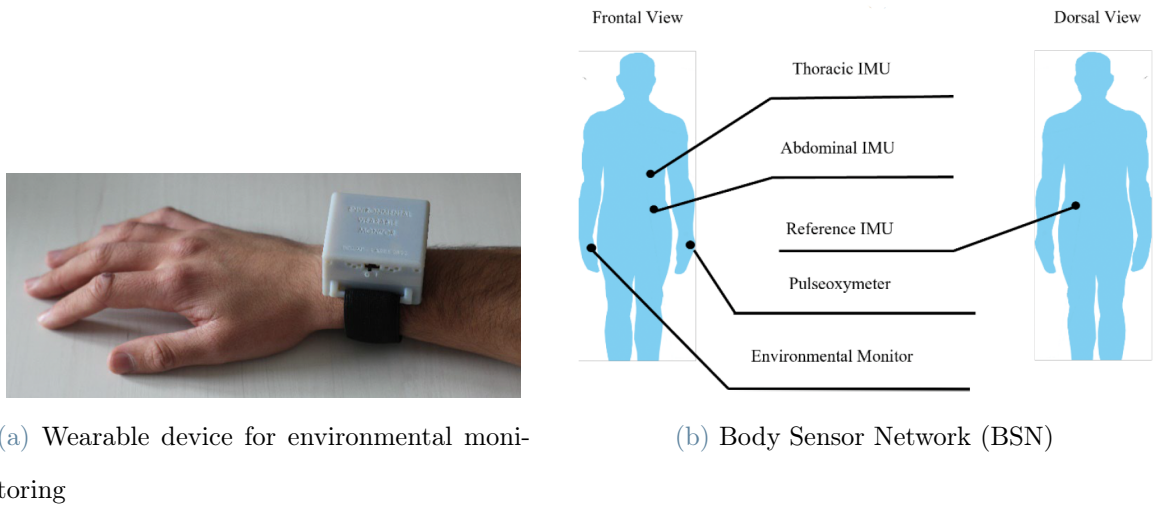


Figure 3.1: Wearable device for personal air pollution exposure estimation

3.2. Experimental Campaign

An experimental campaign is conducted involving a cohort of 20 subjects, during which both environmental and physiological data are systematically collected and saved in an online database (i.e., Google Firebase) for later analyses [12]. The study participants are all in good health and they provide their informed consent prior to participation. The sample population is composed of 14 male and 6 female subjects, with a median of age distribution which is 25.5 with an IQR of 3 years. Three participants exhibit ages classified as outliers, specifically 32, 46 and 55 years. Each subject is asked to go through a predefined route spanning a total distance of 4.5km, within the university area of Politecnico di Milano. The complete path is highlighted in Figure 3.2, as shown in [12].

The walking protocol is intentionally structured as an alternation of specific points along the route at which participants are asked to pause momentarily, and walking segments between a pair of stations. These locations are strategically selected based on their intrinsic environmental characteristics to ensure a diverse representation of urban conditions. The chosen sites include areas presumed to have relatively low levels of air pollution, such as parks, as well as locations hypothesized to have elevated pollutant concentrations, including a gas station, a train station and streets characterized by high vehicular traffic density.

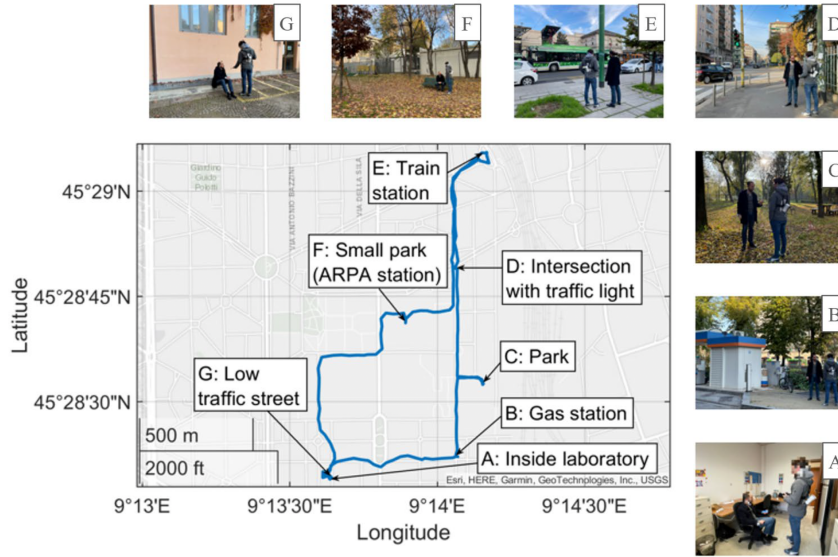


Figure 3.2: Route of the experimental campaign with static acquisition points highlighted (A) laboratory, (B) gas station, (C) park, (D) intersection of roads with traffic lights, (E) train station, (F) small park, (G) low traffic street

As a result of this construction of the walk, the acquisitions made by the monitoring device were of two different types. Static acquisitions consisted of 5 minutes of recording in a static condition in one of the seven acquisition points: both the environmental and physiological parameters were recorded. Specifically, participants were required to be seated at checkpoints A, F and G, whereas they were instructed to stand at points B, C, D and E. Dynamic acquisitions involved only the environmental monitor and consisted in the recording of pollutant concentration while subjects were in motion between the

static points. During this phase, participants were free to walk at their natural pace, without any imposed constraints on walking speed or predetermined travel durations. This condition resulted in different amounts of data available for each subject. A point H was introduced corresponding to point A, in order to obtain dynamic acquisitions along the complete walk including its last segment during which participants returned to their starting position. This addition ensured a complete and consistent dataset for the entire experimental trajectory. Unlike the other acquisition points, no static measurements were recorded at this specific location.

Most of the acquisitions were made in November 2022. Two time slots per day were available, one in the morning around 10:00 a.m. and one in the afternoon around 2:30 p.m. This choice was done to avoid high traffic hours and keep as much as possible the same conditions through all the acquisitions in order to ensure consistency [12].

Regarding the data acquisition rates, the sample frequency of the three IMUs and the pulse oximeter was 10 Hz for each of them. Conversely, the environmental monitor was programmed to transmit data at intervals of 24 s.

3.3. The Dataset

The researchers subsequently decided to retain only a subset of the parameters recorded during the experimental phase to construct the dataset used in this study. Specifically, the selected parameters included ‘Heart Rate’(HR) and ‘Frequency of Breath’(FB), which were chosen to represent the subject’s cardio-respiratory activity. Additionally, a set of environmental parameters was selected to characterize the surrounding conditions in which the subject was walking. These included CO_2 , PM_1 , $PM_{2.5}$, PM_{10} , temperature, humidity, pressure and VOC . Conversely, the parameters CO and NO_2 were excluded from the dataset, as their recorded measurements were considered unreliable and, therefore, unsuitable for a rigorous analysis.

Given that each subject exhibits unique pacing and that an experimental protocol in an

uncontrolled environment is susceptible to random events such as the presence of red traffic lights or crowds affecting walking speed, it was necessary to implement a preprocessing phase on the acquired data. This step was essential to ensure that the collected data were suitable for meaningful analysis and inter-subject comparison. Specifically, a standardized dataset was required to facilitate the comparative assessment of different acquisitions across subjects. Physiological data did not require any preprocessing, as these measurements were obtained only during static acquisition phases and were recorded over a predefined interval of 5 minutes at a fixed frequency. Conversely, environmental data have been resampled with a sampling frequency of 1 Hz. Subsequently, the entire acquisition was segmented using time stamps registered during the acquisition protocol when arriving and leaving each acquisition point. This process generated 7 arrays for static conditions and 7 arrays for dynamic acquisition for each subject, serving as a means of space normalization to facilitate the comparisons. A linear interpolation that took 3 samples per minute was finally applied to obtain in corresponding arrays the same number of samples for each subject [12]. This approach ensured a more consistent basis for comparison, as the number of recorder samples remained consistent across all subjects.

As a result of this pre-processing step, the final dataset is structured as follows. For each physiological parameter, the dataset consists of a 100×7 matrix, where the columns correspond to the 7 acquisitions points (A,B,C,D,E,F,G) and the rows contain 5 measurements for each of the 20 subjects at each acquisition point. Regarding the environmental parameters and air pollution components, a dedicated workspace was created for each subject and for each parameter. This workspace include data from both static and dynamic acquisitions, presented in their original form - prior to pre-processing - and in their interpolated form, which ensures uniform vector dimensions across all subjects. Specifically, for static acquisitions, 7 vectors are obtained, once in their original form and once with a standardized length of $n = 15$: each of them is referred to a single acquisition point. For dynamic acquisitions, the dataset consists of 7 vectors per subject for each parameter,

both before and after the pre-processing step. The standardized dimensions obtained for each segment of the walk, for each subject, are: $n=36$ for the segment between the points A and B, $n=24$ for segment B to C, $n=30$ for segment C to D, $n=27$ for segment D to E, $n=45$ for segment E to F, $n=51$ for segment F to G, and $n=9$ for segment G to H.

4 | Methods

The experimental study described in Chapter 3 has generated a structured dataset consisting of sequential observations collected over time. Given the nature of the data, which includes both environmental parameters recorded continuously along the entire path, and physiological measurements acquired only at the check points, a time-dependent analytical framework is required to investigate potential relationships between environmental conditions and human health status. Two distinct data structures emerge: measurements taken at regular intervals along the entire path can be classified as Time Series, while those exhibiting interruptions between stations, such as physiological parameters, are identified as Interrupted Time Series. These two different structures are explained in Sections 4.1 and 4.2. Their relevance lies in their ability to assess whether physiological responses undergo significant variations following environmental changes. By applying appropriate statistical methodologies, such as Segmented Regression described in Section 4.3, it is possible to evaluate the impact of the transition from a station to a different one on the variable of interest.

This chapter introduces the theoretical foundations and methodologies used to analyze this type of data, including some tools useful for an initial explorative analysis, such as functional boxplots and outliergrams used for outlier detection, presented in Section 4.5. The chapter explores how specific statistical techniques can be used to assess the impact of environmental transitions on physiological responses, trying to go in the direction of a rigorous evaluation of whether changes in air quality and environmental conditions have measurable effects on an individual's health status.

4.1. Time Series

Data obtained from observations collected sequentially over time are pervasive across various fields, including business, meteorology, agriculture, biological sciences, and medicine: the list of areas in which time series are studied is virtually endless [16]. Such sequences of data points are defined as time series when points are collected, observed or recorded over a specified time period. These observations are typically measured at equally spaced time points, usually hourly, daily, monthly or yearly, and are ordered chronologically. Time plays the role of independent variable for the dependent variables of time series data [9].

Mathematically, a time series is written as:

$$X_t = \{x_1, x_2, x_3, \dots, x_n\} \quad (4.1)$$

where t denotes the specific time point at which observations are recorded and X_t represents the corresponding value of the observed variable at that time. There are two characteristics that change a vector of data into a time series. The first is the importance of the temporal order of data that prevents any changes in the index of the variables. The second is the correlation between the data of a time series that enables the development of mathematical models over the data [9]. Outcomes may take various forms such as counts, continuous data or binary variables. Based on the method of data collection, time series can be classified into two categories: discrete and continuous. In discrete time series, data points are recorder at specific and fixed intervals, whereas in continuous time series, data are collected continuously over time.

4.2. Interrupted Time Series

Occasionally, an interruption may occur within a time series, partitioning the sequence of observations into two or more distinct portions, that are referred to as ‘segments’. The

points at which such division occurs are denoted as change points, and a time series that has been partitioned by these points is termed an Interrupted Time Series. Hence, an ITS represents a specific class of time series in which data are systematically collected at regular intervals before and after the occurrence of an interruption, referred to as ‘intervention’. Such interventions may include real-world events, policy changes, technological implementations, economic events or other external factors that influence and have an effect on observed variable. ITS requires a clear differentiation of the pre-intervention period and the post-intervention period. In some evaluations it may be difficult to accurately define the exact moment when the intervention began and to differentiate the effects of different components. This does not necessarily require the intervention to be introduced overnight but the period of implementation should be well defined so that it can be considered separately [10].

The primary objective of ITS analysis is to assess whether the data pattern observed in the post-intervention period exhibit significant deviations compared to the pre-intervention period. ITS analysis is indeed widely regarded as one of the most robust and commonly employed quasi-experimental designs for evaluating the impact of an intervention on the outcome of interest. ITS works best with short-term outcomes that are expected to change either relatively quickly after an intervention is implemented or after a clearly defined lag [10]. Considering public health interventions, their evaluation normally relies on comparing the outcome of interest in a population exposed to the intervention with that in an external control group not subject to the same intervention. A strength of ITS that makes them so popular is that this design adopts a different approach whereby comparisons are instead made across time within a single population [11].

A fundamental assumption underlying ITS analysis is that, in the absence of the intervention, the observed trend and pattern seen in the data would have continued in the same manner. The hypothetical scenario under which the intervention had not occurred and the trend continued unchanged is referred to as the counterfactual [10]. Examining

the counterfactual scenario is crucial for determining whether the intervention has had a meaningful impact. By comparing the ITS before and after the interruption, it is possible to identify some deviations which can be examined and interpreted for the evaluation of the impact of the intervention: indeed, they may suggest a causal relationship between the intervention itself and the observed changes.

Various effect estimates can be employed to describe the impact of the intervention: a change in level corresponds to the difference between the observed value at the time point of interest and the predicted pre-intervention trend, while a change in slope is the difference between the post-and pre-intervention slopes [26] and it represents a gradual change in the value of the outcome during the segment [45]. It is also possible to observe a combination of both level and slope changes. Prior to conducting the analysis, it is important to hypothesize the potential impact of the intervention on the outcome, by selecting an appropriate ‘impact model’. Several examples of possible impact models are illustrated in Figure 4.1, taken from the study conducted by Bernal et al. [10].

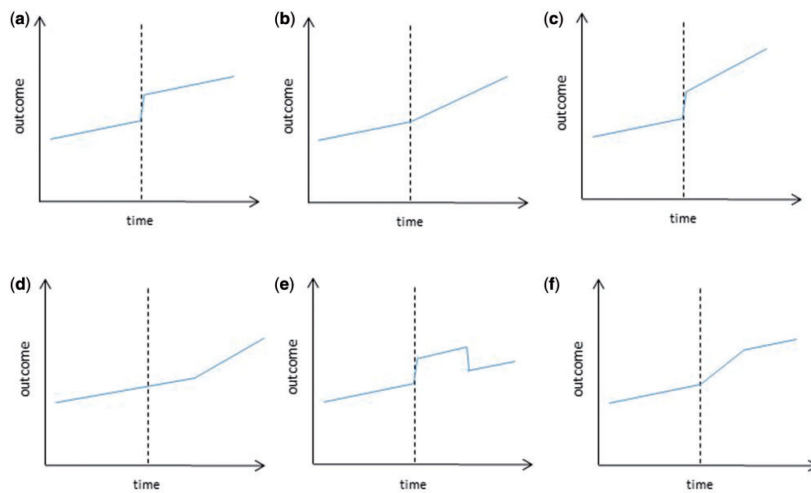


Figure 4.1: Examples of impact models used in ITS

(a) Level change ; (b) Slope change ; (c) Level and slope change ; (d) Slope change following a lag ; (e) Temporary level change ; (f) Temporary slope change leading to a level change

4.2.1. Critical Issues of Interrupted Time Series

While offering significant advantages over simpler before-and-after comparisons, ITS analysis presents several methodological challenges that must be thoroughly addressed to ensure the validity and reliability of the findings. A particularly important consideration is the necessity to account for underlying temporal patterns, such as seasonality and autocorrelation which may inherently exist within the data and influence the observed results. Furthermore, it is also crucial to appropriately identify and manage possible confounders and potential biases that could otherwise distort the interpretation of the intervention's impact. All these challenges require careful consideration. These issues are highlighted here for the reader's awareness, and the following sections will clarify which of them are addressed by the explained techniques, which are not considered, and the reasons for their exclusion.

Seasonality

Seasonal effects are a critical factor in ITS analysis, particularly when observed outcomes exhibit systematic fluctuations over time. Seasonality introduces two main challenges that can affect the reliability of the analysis. First, an uneven distribution of months before and after the intervention (e.g., a higher proportion of winter months) may bias results, especially in short series. Second, outcomes in one month tend to be more similar to those in neighboring months within the same season, leading to autocorrelation (discussed below).

Ignoring seasonal trends can confound the estimated intervention effect, leading to misleading conclusions. Several methods address seasonality, including incorporating Fourier terms (sine and cosine functions of time) or splines. Alternatively, stratifying by calendar months or relevant intervals offers a practical solution. In regression models, adding season-specific terms effectively mitigates confounding effects [10].

Autocorrelation

Time series data often exhibit dependence, where consecutive observations are more similar than those further apart [10]. This phenomenon, known as autocorrelation, is crucial to assess, as ignoring it can lead to underestimated standard errors and overestimated intervention effects.

Autocorrelation can be detected through diagnostic techniques such as visual inspection of residual plots over time or of partial autocorrelation functions. Randomly scattered residuals indicate no autocorrelation, while patterns suggest positive or negative autocorrelation depending on whether consecutive residuals lie on the same or opposite sides of the regression line [45]. Formal statistical tests can further confirm its presence.

Confounders

Confounders are variables or events that may influence the observed outcome independently of the intervention, potentially distorting the causal relationship between an intervention and its effects. In ITS analysis, particular attention must be given to time-varying confounders, which evolve over time and simultaneously affect both the intervention and the outcome. If unaccounted for, they can introduce bias in the analysis. Seasonality is a common example, along with factors like infectious disease outbreaks or weather events. Additionally, events that may occur around the same time as the intervention, such as other interventions or natural events, may influence the outcome [10].

To mitigate their effects, variables representing them directly should be incorporated into the statistical model. When confounders are either unmeasured or unknown, alternative design strategies can mitigate their impact. These include controlled ITS designs with a control group or outcome, multiple baseline designs introducing the intervention at different locations over time, or phase-based approaches where the intervention is introduced and then withdrawn to assess reversibility [11].

4.3. Segmented Regression

Advanced statistical techniques are commonly employed in the analysis of ITS, providing a framework for estimating both the immediate and long-term effects of the intervention while controlling for potential biases. Among the various analytical approaches available, segmented regression analysis stands out as a particularly powerful statistical method for estimating intervention effects and drawing rigorous and more formal conclusions regarding the impact of the intervention on the variable of interest. This method allows researchers to assess, in statistical terms, how much an intervention changed an outcome of interest, both immediately and over time, and whether factors other than the intervention could explain the change [45]. In particular, it is used to measure statistically the changes in level (immediate changes in the rate of the outcome) and slope (long-term changes in the trend) in the post-intervention period compared to the pre-intervention period [34]. By comparing them, segmented regression analysis provides a comprehensive assessment of the intervention's effectiveness and its temporal dynamics.

The typical segmented regression model is widely explained in [10] and [45]. Here a model for data with a single intervention occurring is shown.

Mathematically it is expressed as:

$$Y_t = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot intervention_t + \beta_3 \cdot timeSince_t + \epsilon_t \quad (4.2)$$

where

- Y_t is the outcome variable at time t
- $time$ represents the time elapsed since the start of the study, with the unit representing the frequency with which observations are taken (e.g. month or year)
- $intervention_t$ is the value assumed at instant t by the binary variable $intervention$, which is coded 0 for the pre-intervention period and 1 for the post-intervention period

- $timeSince_t$ is a time variable which corresponds to the interaction between $time$ and $intervention_t$
- ϵ is an error term, which represents the random variability not explained by the model; we assume it to be normally distributed with constant variance σ^2

Assumptions of homoscedasticity, independence, and identical distribution for the residuals have to be taken on.

In this model, the interpretation of the regression coefficients is quite intuitive: β_0 estimates the baseline level of the outcome before the intervention; β_1 estimates the change in outcome associated with a time unit increase; β_2 represents the immediate change in level following the intervention (which can be both a step increase or decrease); β_3 indicates the change in the trend (i.e. difference in slope) following the intervention. Note that the sum of β_1 and β_3 is the post-intervention slope.

One may be interested in the effects of different components of an intervention introduced at different time points. Segmented regression models can specify more than one change point. For instance, a model with two change points would be:

$$Y_t = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot intervention_{1t} + \beta_3 \cdot timeSince_{1t} + \beta_4 \cdot intervention_{2t} + \beta_5 \cdot timeSince_{2t} + \epsilon_t \quad (4.3)$$

For each interruption, variables $intervention_i$ and $timeSince_i$ are created with $i \in \{1, 2\}$, and $intervention_{it}$ is the value at instant t of the variable related to the i -th interruption while $timeSince_{it}$ corresponds to the interaction between $time$ and $intervention_{it}$. In the same way, if N interventions are introduced at N different time points, the model would be:

$$Y_t = \beta_0 + \beta_1 \cdot time + \sum_{i=1}^N (\beta_{2i} \cdot intervention_{it} + \beta_{2i+1} \cdot timeSince_{it}) + \epsilon_t \quad (4.4)$$

Segmented regression analysis allows to control for prior trend in the outcome and to study the dynamics of change in response to an intervention: it visually displays them by showing whether an effect is immediate or delayed, and whether or not an effect persists or is temporary. It can also estimate the size of the effect at different time points, as well as changes in the trend of the effect over time [45].

A significant strength of segmented regression is its ability to provide extremely clear and easy-to-interpret graphical representations of results. Indeed, from graphical outputs it is intuitive to understand whether an intervention resulted in an immediate shift, a gradual trend change or both.

Additionally, segmented regression shows a high flexibility in handling complex data structures and complex scenarios, such as interventions with delayed effects, temporary impacts or multiple interventions.

4.3.1. Limitations of Segmented Regression for ITS

Despite its advantages, Segmented Regression is subject to some methodological issues and certain limitations, that must be considered when interpreting results, together with those issues which are characteristic of ITS and that have been already discussed: among them we find seasonality, autocorrelation and confounders.

A fundamental assumption underlying the models discussed so far is that the outcome follows a linear trend within each segment. However, this assumption may only hold over short intervals, as real-world changes may follow non-linear patterns. In cases where intervention effects diffuse across a population with an increasing or decreasing curvilinear trend, one can accommodate those effects in linear models by using geometrically increasing or decreasing trend terms. Nevertheless, other forms of non-linear trend may require other types of mathematical models, which are more complex.

Another significant limitation of the Segmented Regression approach is that this type of analysis struggles with the control of individual-level covariates. Such covariates are

often considered potential confounders, particularly when they vary over time, thereby complicating the attribution of causal effects to the intervention [45]. While ITS analysis offers some strategies to address these time-varying confounders, effectively accounting for them remains a challenge and may require further methodological adjustments.

Statistical power is another critical aspect of ITS analysis using Segmented Regression. Indeed, segmented regression is highly sensitive to the number of data points collected over time before and after the intervention, which directly affects the ability to detect meaningful changes. A limited number of observations may lead to insufficient statistical power, reducing the ability and the precision needed to detect significant changes in level or trend accurately. A small sample size increases the variance of the estimated regression coefficients, leading to wide confidence intervals and decreased precision. This issue is particularly problematic when dealing with subtle intervention effect, as the limited number of observations may not provide enough resolution to differentiate a true change from random variation. Additionally, a low number of data points in the pre-intervention period may compromise the ability to accurately estimate the underlying secular trend. Without a well-defined baseline, any apparent change observed after the intervention, may be erroneously attributed to the intervention itself rather than reflecting a continuation of a pre-existing trajectory. Conversely, an overly long pre-intervention period can also be problematic. Indeed, it increases the risk of incorporating long-term structural changes or external influences that are unrelated to the intervention and may not be relevant to the evaluation: in these cases pre-intervention trend may not remain stable over time and the counterfactual assumption (i.e. the assumption that the estimated baseline trend would have continued unchanged in the absence of the intervention) becomes weaker. The result is that the true effect of the intervention is obscured and the estimation of its impact is distorted. Taking everything into account, short time series suggest that model fitting is performed with less confidence, that power is reduced, and hence, that there's the possibly to fail to detect secular trends [38].

However there are no fixed limits regarding the number of data points, as the power depends on various other factors including distribution of data points before and after the intervention, variability within the data, strength of effect, and the presence of confounding effects such as seasonality. Zhang et al. [47] conducted simulations with power calculations under different model parameters, and suggest that studies with few time points or with small expected effect sizes should be interpreted with caution as they may be underpowered [10]. The statistical power of Segmented Regression is increased if the numbers of data points are equally distributed before and after the intervention, though this is often not practical [47].

4.4. Jittering

Despite its strengths Segmented Regression lacks in managing Multivariate Interrupted Time Series, but a valid although trivial option is to apply jittering. This technique consists in the introduction of small and randomly generated perturbations to data, aimed at various computational and analytical purposes. In scenarios where multiple measurements are recorded at the exact same time instant, as it could be for a Multivariate Interrupted Time Series, a traditional time series analysis may be hindered: by introducing small perturbations, jittering transforms the dataset into a time series as measurements are re-distributed over time. In this way, more robust analyses are allowed, and vertical patterns are limited, that otherwise would not be captured by the model and would remain in the residuals as an inexplicable variability. This method is also useful in addressing issues related to data granularity, spatial and temporal disaggregation, and ambiguity in sequences of measurements. A notable application is presented in the work of Lovelace et al. [30], where jittering was employed to enable the user to adjust the level of disaggregation and to adjust the location of start and end points of a transport plan, based on variables such as transport network density, residential density or size of commercial buildings acting as trip attractors. By introducing controlled randomness, jittering facilitated a more flexible

and realistic representation of movement patterns within urban environments. Given its wide range of applications, jittering has become a valuable tool in data pre-processing.

Mathematically, jittering can be expressed as:

$$x' = x + \epsilon \quad (4.5)$$

where x is the original data, x' is its jittered version and ϵ is the random perturbation, usually drawn from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ or a Uniform distribution $U(-\delta, \delta)$. The choice of the distribution and scale of ϵ depends on the intended application: both σ and δ should have a very small value.

4.5. Analysis of Outliers

In the field of statistical analysis, the identification of anomalous patterns represent a crucial task, as the presence of outliers can significantly influence forecasting models and statistical inferences. Usually outliers are considered as an error or a source of noise within the dataset. However, they may carry some important information on the phenomenon under study, so they are essential to be detected and analyzed. In the case study, this type of analysis is used to determine if a specific parameter behaves anomalously in one of the seven stations located along the path, and which of these check points has environmental conditions which mostly affect the behaviors of the parameters.

It is important to know that there is no official and general definition of outliers as their presence often depends on the assumptions regarding the hidden structure of data and on the model that is chosen to detect them. Nevertheless, we can refer to the following as an appropriate definition [23].

"[An outlier is] an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism"

- D. M. Hawkins -

The detection of outliers assumes particular importance in those scenarios where contaminations may heavily bias estimation and inference, as functional data analysis (FDA). Under the hypothesis that data can be described by an underlying function, a functional dataset consists of a collection of functional observations. Within this framework, the study of functional data allows for the classification of outliers into two categories: magnitude outliers and shape outliers.

Magnitude outliers are characterized by a consistent deviation from the expected range of values, either by being significantly higher or lower than the majority of the observations at corresponding time points. They deviate in terms of the overall magnitude (intensity or amplitude) of the function. Due to their distinct magnitude differences, they are often easier to identify through visual inspection, as they appear noticeably larger or smaller in magnitude compared to the main cluster of data points. For instance, in a dataset consisting of temperature curves recorded over time, a magnitude outlier would be a curve with significantly higher or lower temperature values compared to the other curves throughout the entire period.

Conversely, shape outliers do not necessarily exhibit extreme magnitudes but rather present an atypical structural pattern that deviates from the expected functional form. These deviations are characterized by differences in the shape of the function, while the overall magnitude remains within the general range observed in the dataset. Such irregularities may manifest as peaks, irregular and unexpected oscillations, localized shifts or distortions that differentiate it from the predominant trends in the dataset. The visual detection of shape outliers can be more challenging, especially with complex functional data where variations in structure are subtle. Returning to the example of temperature curves, a shape outlier could be a curve that follows a different pattern or trend compared to the others, despite having a similar overall temperature range: for instance, it may exhibit a sudden spike or drop in temperature at certain time points that is absent in the other curves.

The distinction between these two types of outliers is fundamental, as their impact on statistical analysis and model performance differs substantially. While magnitude outliers primarily affect measures of central tendency and variability, shape outliers introduce complexities in modeling functional relationships and detecting trends. These types of outliers can be detected respectively by functional boxplots and outliergrams.

4.5.1. Functional Boxplot

The functional boxplot (FBP) is an extension of the traditional boxplot to the domain of FDA. It is a visualization and detection tool, and it is particularly useful for summarizing the distribution of data while identifying potential outliers. Specifically, it is based on functional depths and consists in the following steps. First, a data depth is computed for each function $X_i(t)$ in the given set of functional observations $\{X_i(t)\}_{i=1}^n$: a notable example is the modified band depth (MBD) which ranks functional observations based on how central they are within the dataset. The function with the highest depth value is considered the most central observation, while functions with lower depth values are considered more extreme. Second, functions are ranked based on their depth and the 50% central region is considered, which is:

$$C_\alpha = \left\{ (t, z(t)) : \min_{l=1, \dots, \lceil \alpha N \rceil} X_l(t) \leq z(t) \leq \max_{r=1, \dots, \lceil \alpha N \rceil} X_r(t) \right\} \quad (4.6)$$

It represents the region where the majority of observations lie. Then this region is inflated by a factor $F > 1$ (usually the choice $F = 1.5$ is tuned for the Gaussian case), obtaining fences given by the envelope of functions contained inside the inflated region. Finally, the observations crossing the fences and falling outside the region are flagged as amplitude outliers. An example of the result of this construction is provided by Figure 4.2.

Functional boxplot is a useful tool which provides a compact and interpretable way to visualize a set of functional observations while detecting for amplitude outliers. Since

it is based on depth measures it is robust to noise and effectively capture the central tendency of functional data. They can be used also for time series analysis in order to detect anomalies.

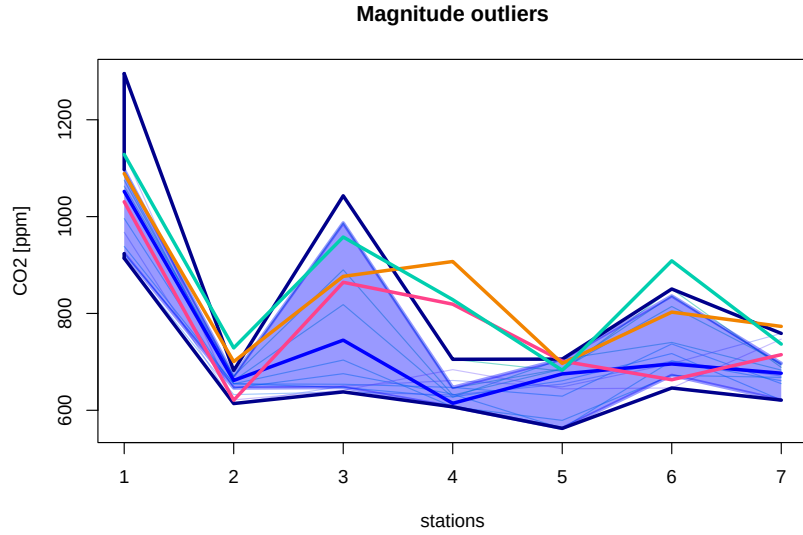


Figure 4.2: Detection of magnitude outliers through functional boxplot

While functional boxplot is effective for detecting magnitude outliers it is less sensitive to shape outliers, which are often detected by a complementary tool: the outliergram.

4.5.2. Outliergram

The outliergram is a recently proposed graphical tool designed to identify shape outliers in functional data. It is based on a scatterplot of the sample modified band depth (MBD) versus the modified epigraph index (MEI), which quantifies the proportion of time that an observed function stays above the median function and helps in differentiating whether an outlier deviates in shape rather than simply in magnitude. They are related by:

$$MBD(f) \leq \alpha_0 + \alpha_1 MEI(f) + \alpha_2 N^2 MEI^2(f) \quad (4.7)$$

For univariate functional data, MBD and MEI of a function f with respect to a sample $f_1, \dots, f_N$ are the following:

$$MBD(f) = \frac{1}{\binom{N}{2}N} \sum_{1 \leq i < j \leq N} \tilde{\lambda}(\{t \in I : f_i(t) \leq f(t) \leq f_j(t)\}) \quad (4.8)$$

$$MEI(f) = \frac{1}{N} \sum_{i=1}^N \tilde{\lambda}(\{t \in I : f_i(t) \geq f(t)\}) \quad (4.9)$$

where $\tilde{\lambda}$ is the normalised Lebesgue measure.

Conversely, for multivariate functional data they are:

$$MBD_{\{f_1, \dots, f_n\}}(f) = \sum_{k=1}^h p_k MBD_{\{f_{1k}, \dots, f_{nk}\}}(f_k) \quad (4.10)$$

$$MEI_{\{f_1, \dots, f_n\}}(f) = \sum_{k=1}^h p_k MEI_{\{f_{1k}, \dots, f_{nk}\}}(f_k) \quad (4.11)$$

with $p_k > 0 \quad \forall k = 1, \dots, h$ and $\sum_{k=1}^h p_k = 1$

Given the relation (4.7), points lying far away from the upper parabolic boundary are likely to be flagged as shape outliers. Specifically, the points flagged as outliers are those such that:

$$d_i \geq Q_3(d) + F \cdot IQR(d) \quad (4.12)$$

with

$$d_i = \alpha_0 + \alpha_1 MEI(f) + \alpha_2 N^2 MEI^2(f) - MBD(f) \quad (4.13)$$

By complementing functional boxplots, outliergrams help in providing a complete outliers detection framework that captures both magnitude and shape outliers, respectively through these two different tools. An example of the result of this construction is provided by Figure 4.3.

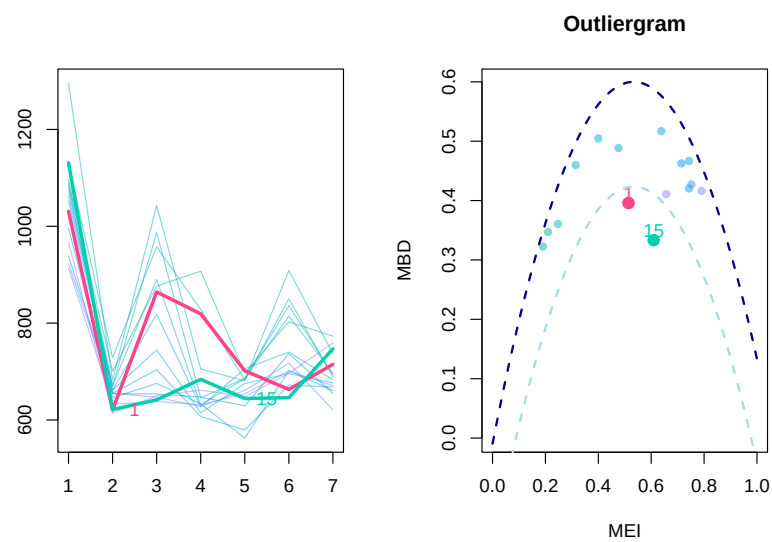


Figure 4.3: Detection of shape outliers through outliergram

5 | Analysis and Results

In the next sections we will present the results of the analysis made with data from the case study presented in Chapter 3. In Section 5.1, following an initial data pre-processing stage aimed at constructing well-structured datasets, we will conduct an exploratory analysis to identify several distinct trends within the time series comprising these dataset. Subsequently, in Section 5.2, employing outlier detection techniques, we will emphasize the differences observed in the data collected from different stations, which are highlighted in the preliminary phases of the explorative analysis and receive here a stronger statistical evidence. Finally (Section 5.3), these discrepancies will be further validated by the outcomes of the Segmented Regression modeling phase, which will enable us to evaluate whether the changes of environments an individual is walking through, would have a significant impact on their health status.

5.1. Explorative Analysis

As a first step, datasets are constructed to provide an appropriate structure for the data collected during the experimental campaign presented by Bernasconi et al. [12] and discussed in Sections 3.2 and 3.3 of this thesis. With respect to the data pertaining to individual's health status, specifically the measurements of HR and FB, these have already been collected into two matrices encompassing all 20 participants to the study, and therefore no reorganization is necessary. For each subject, the sequences of measurements corresponding to these variables can be regarded as Interrupted Time Series, as data are collected exclusively at the 7 stations. Conversely, a reorganization is required for those

data that describe the environment in which the individual is walking, which include measurements of the variables CO_2 , PM_1 , $PM_{2.5}$, PM_{10} , temperature, pressure, humidity and VOC . Since these environmental measurements were collected individually for each subject, it is necessary to consolidate the data across all subjects. This was achieved by creating two distinct datasets for each environmental variable: one containing data collected both at the stations and along the walking segments between the stations, and a second comprising only the data measured at the 7 stations. This distinction is made to ensure uniformity in the treatment of both the physiological and environmental data. For each subject, the sequences of data included in the first dataset are regarded as Time Series, while the ones included in the second dataset are regarded as Interrupted Time Series as for the physiological data.

As an initial analysis, a visual inspection is conducted to determine the trends variables HR and FB. To this end, seven exploratory plots are generated for each variable, corresponding to the seven stations. Each of them represents the levels of the variable against time. The results of this procedure are presented in Figure 5.1 and Figure 5.2.

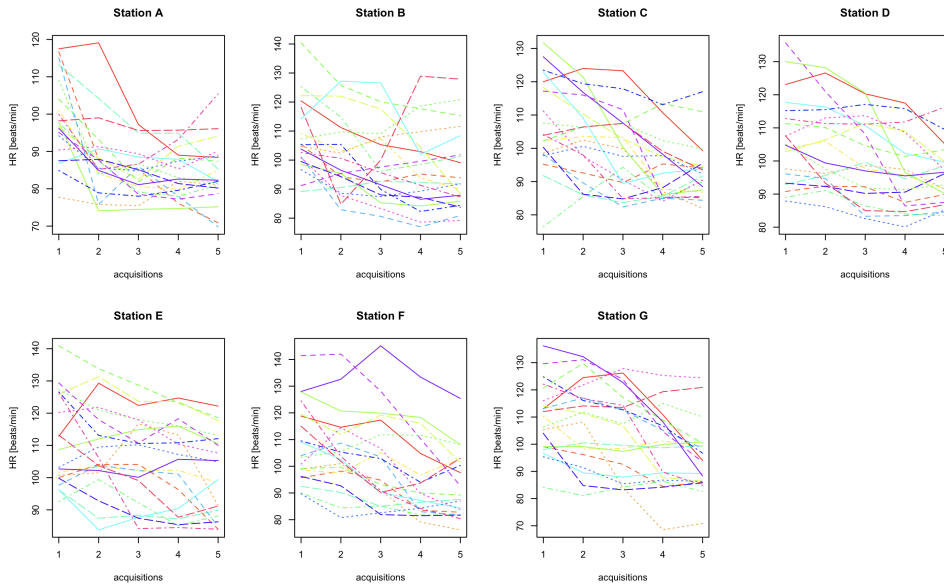


Figure 5.1: Evolutions of variable HR per station

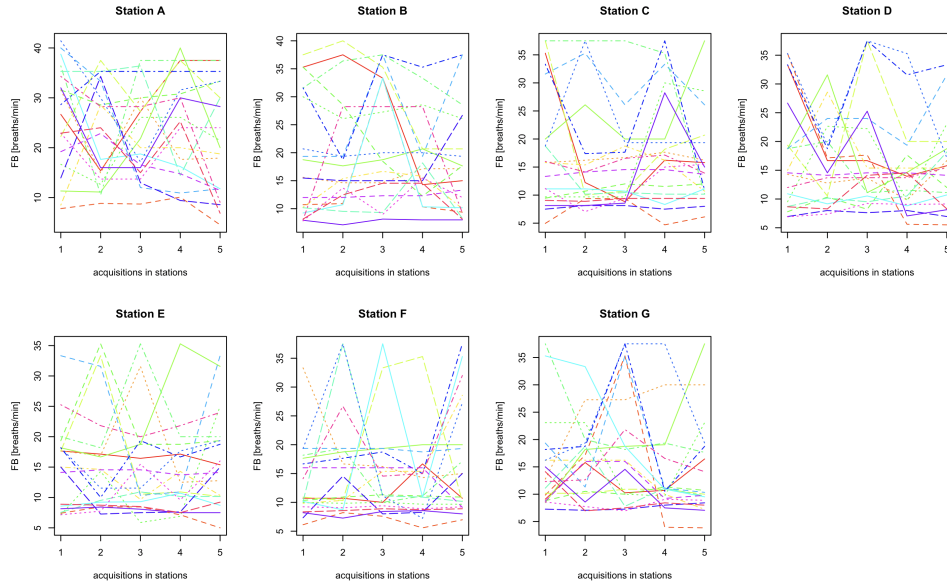


Figure 5.2: Evolutions of variable FB per station

Subsequently, the data are juxtaposed to provide a comprehensive overview while still preserving the distinction between measurements corresponding to different stations. This separation is visually emphasized through vertical lines. The results of this procedure are presented in Figure 5.3 and Figure 5.4.

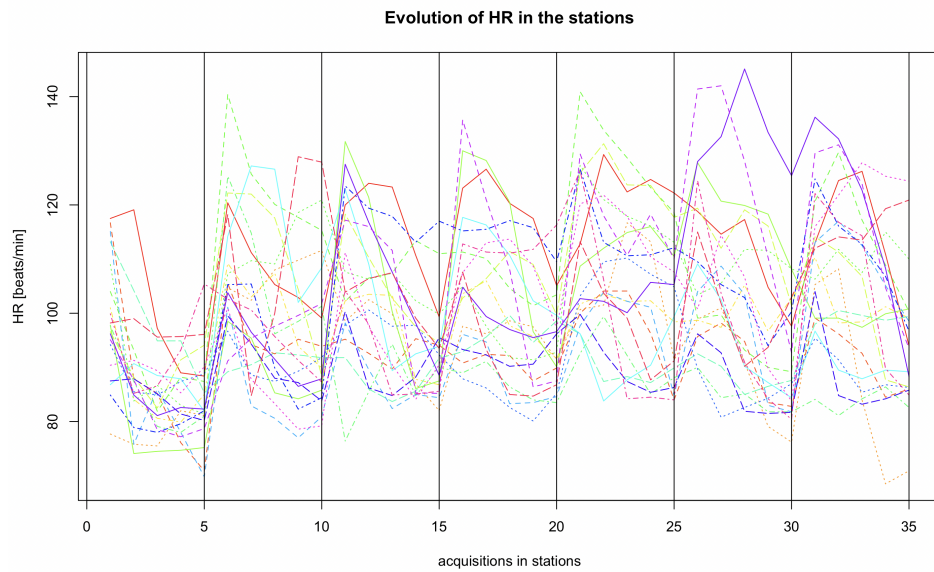


Figure 5.3: Evolutions of variable HR

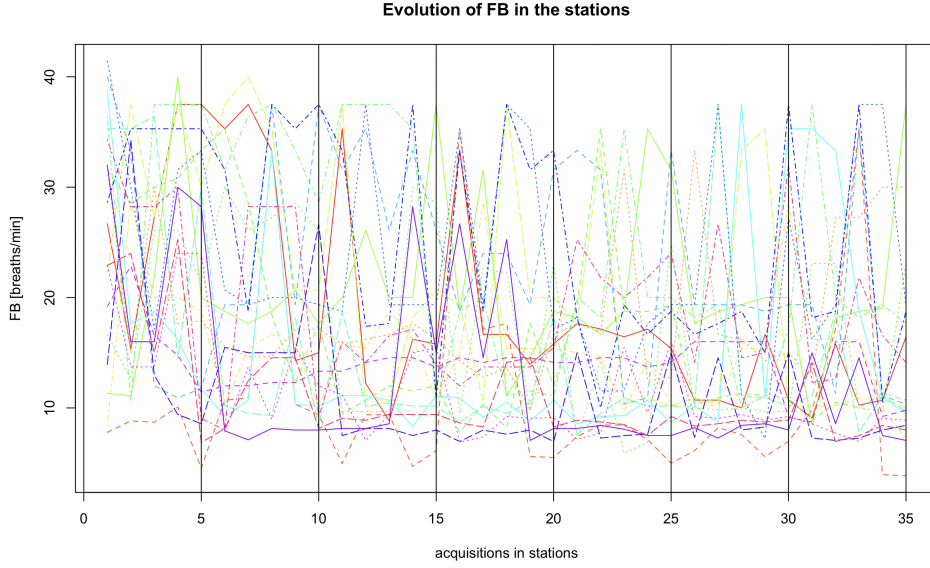


Figure 5.4: Evolutions of variable FB

In the visualization of HR levels, a similar pattern can be observed across the different stations for most of the subjects: in each station, HR values tend to be a little higher in the initial measurements compared to the final ones. This trend is reasonable, considering that at each stations individuals take a rest of 5 minutes, allowing their heart rate to gradually decrease after the physical effort exerted while walking. However, these observations should be interpreted with caution: they are based on only five time-points per station as each participants has to stop in a station only 5 minutes and the device records with a frequency of 1 minute, and may therefore lack reliability and accuracy. Conversely, in the visualization of FB levels, no recurring trend can be identified. The values appear to fluctuate significantly within the range $[0, 40]$, resulting in the presence of numerous peaks but without exhibiting a discernible pattern, such as a consistently increasing or decreasing trend.

As the next step, an analysis is conducted to visualize the trend of the environmental variables. Initially, this is carried out by exploring the data for each individual separately, allowing for a visual identification of potential differences between subjects. As an illustrative example, Figure 5.5 presents the measurements corresponding to the first

subject: in Section A.1 of Appendix A some other examples are provided, but not for all the participants since the results are all quite similar. For each variable, two distinct plots are presented. Taking as an illustrative example the plots related to the parameter CO_2 , the one titled ' CO_2 in stations' displays the variable's values exclusively at the stations; these measurements have been juxtaposed, with different colors used to distinguish between stations. The plot titled ' CO_2 along walk' instead also includes the variable's levels recorded along the walking segments between consecutive stations. This is valid for the construction of the plots related to all the parameters.

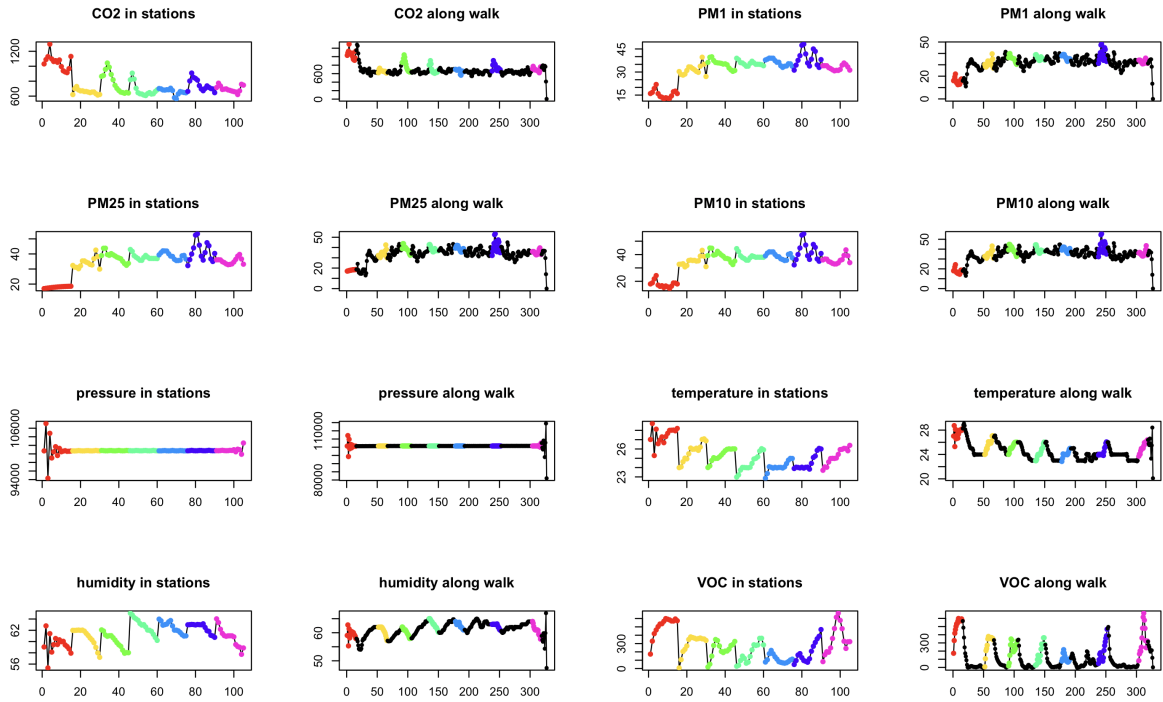


Figure 5.5: Evolutions of environmental variables for subject 1

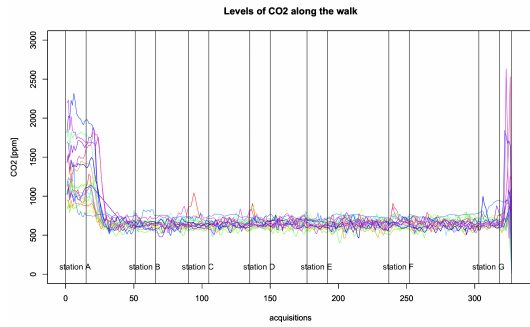
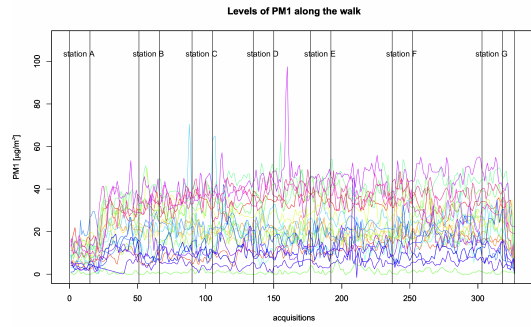
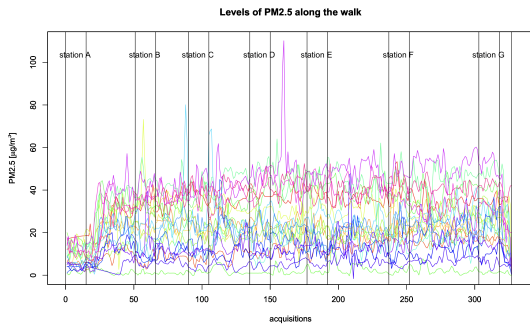
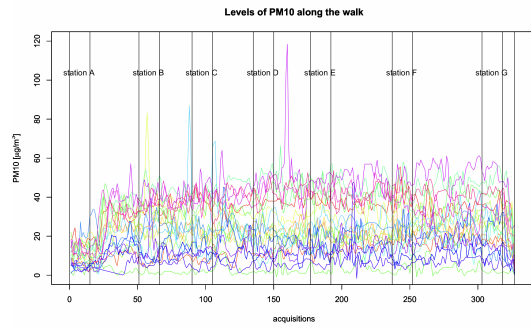
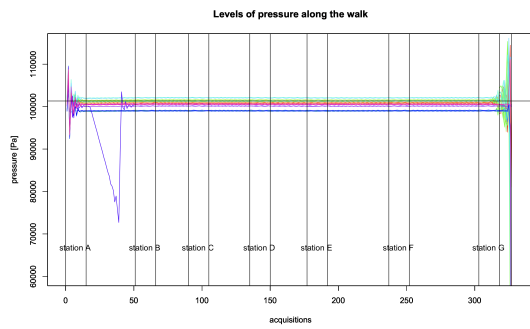
Recurring patterns can be identified across all subjects for each variable, allowing for preliminary conclusions that help characterizing the route walked by individuals. The values of the CO_2 variable are generally higher at station A, highlighted in red, for all subjects. This observation is reasonable, as station A corresponds to an indoor laboratory environment, where the concentration of carbon dioxide from human respiration is expected to be higher. Regarding the concentration of particulate matter, specifically the variables

PM_1 , $PM_{2.5}$ and PM_{10} , their values are lower at station A, while in the other stations, the trends exhibit greater variability. This finding is also consistent with expectations, given that indoor environments typically have lower air pollution levels compared to those found on a busy street. Air pressure is approximately constant, with an average value of 101325 Pa. In this visualization, its behavior aligns with the definition of this variable, except for a few peaks observed in the initial and final measurements. These anomalies are likely due to the subject installing the device on their arm, introducing measurement inaccuracies caused by movement. Temperature exhibits slightly higher values in station A, followed by an increasing trend in measurements taken at the stations and a decreasing trend along the walking segments between stations. Conversely, air humidity follows a decreasing pattern in the station measurements and an increasing trend during the walking segments. Finally, the VOC variable displays a pattern resembling a sequence of small U-shapes, with higher values recorded toward the end of the station measurements and lower values at the conclusion of the walking segments. These conclusions are further supported and more clearly observable through a visual inspection of the trend of each variable across all subjects. The patterns described above can be identified in Figure 5.6.

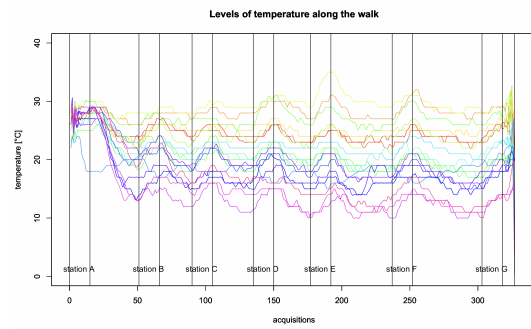
Given the strong similarity in the trends of the particulate matter concentration variables, and considering that PM_1 and PM_{10} were derived from the recorded measurements of $PM_{2.5}$, it is decided to include only the latter in this study, as it serves as a representative indicator of air pollution levels. Additionally, the pressure variable is excluded from the analysis, as its values remained constant across all stations, making it of limited relevance to our objective. The primary aim of this study is to investigate the relationship between an individual's health status and their surrounding environment, particularly in relation to different environmental conditions, and pressure is not informative in this sense. From this point onward, the environmental variables taken into consideration will be: CO_2 , $PM_{2.5}$, temperature, humidity and VOC .

Since station A, which is the only indoor station, behaves differently from the other

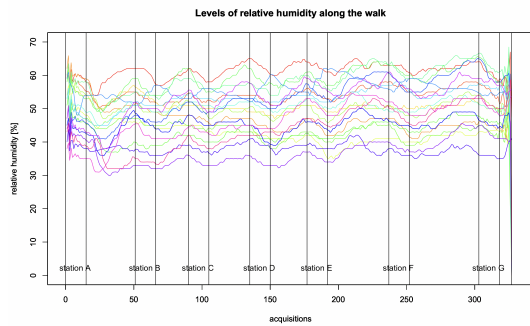
stations for the majority of the variables, it would be of interest to investigate whether this observation can be further validated. To facilitate a more systematic comparison, an alternative visual representation of the data is provided. In Figure 5.7, relative to subject 1, a distinctly different behavior is evident in the measurements recorded at station A compared to those from the other stations, for most of the variables. For other relevant examples, related to other subjects, results can be visualized in Section A.2 of Appendix A.

(a) Evolution of CO_2 (b) Evolution of PM_1 (c) Evolution of $PM_{2.5}$ (d) Evolution of PM_{10} 

(e) Evolution of pressure



(f) Evolution of temperature



(g) Evolution of humidity

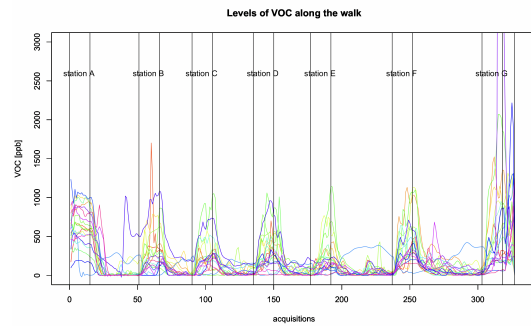
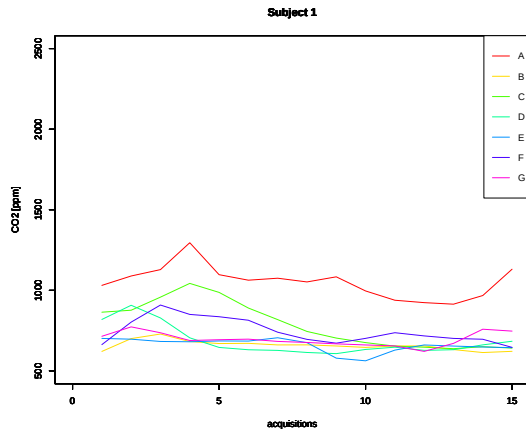
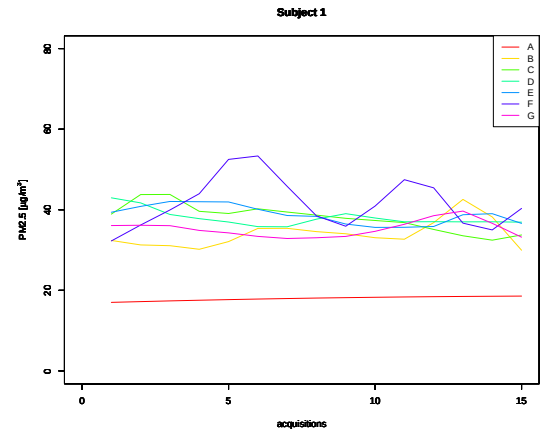
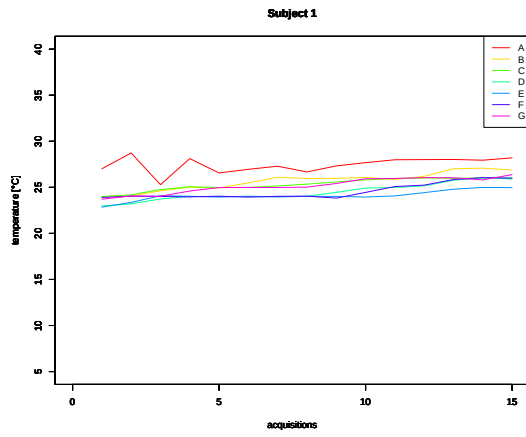
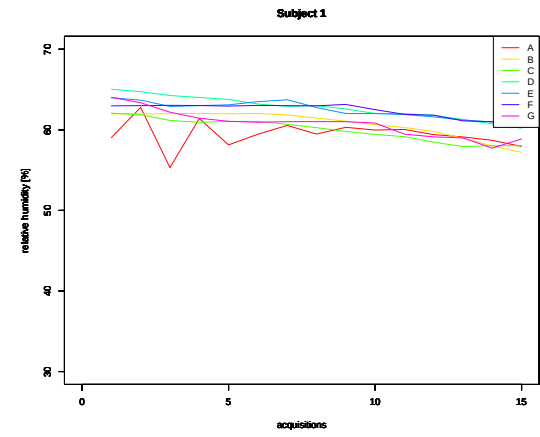
(h) Evolution of VOC

Figure 5.6: Evolutions of environmental variables

(a) Comparison for CO_2 (b) Comparison for $PM_{2.5}$ 

(c) Comparison for temperature



(d) Comparison for humidity

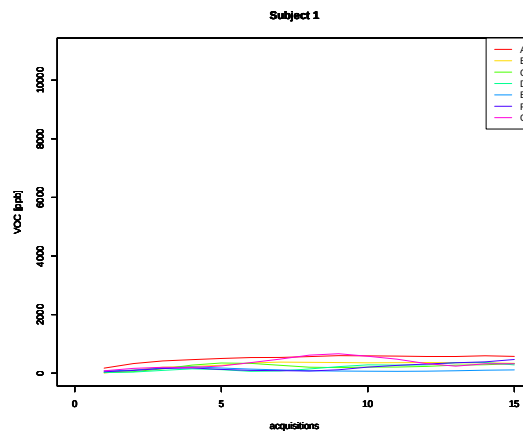
(e) Comparison for VOC

Figure 5.7: Comparisons between stations for subject 1

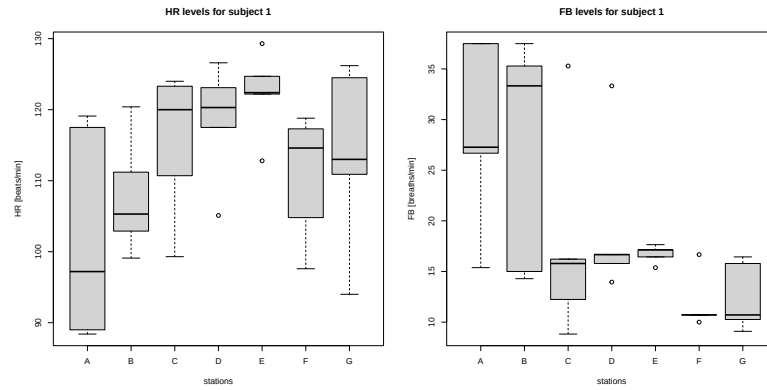
5.2. Outlier detection

To provide statistical evidence for the conclusions drawn from the visual inspection of the plots presented in Section 5.1, a detailed analysis of outliers is conducted. The aim of this analysis is to identify any observation that exhibits anomalous behavior compared to the general trends, as these could potentially lead to inaccurate conclusions.

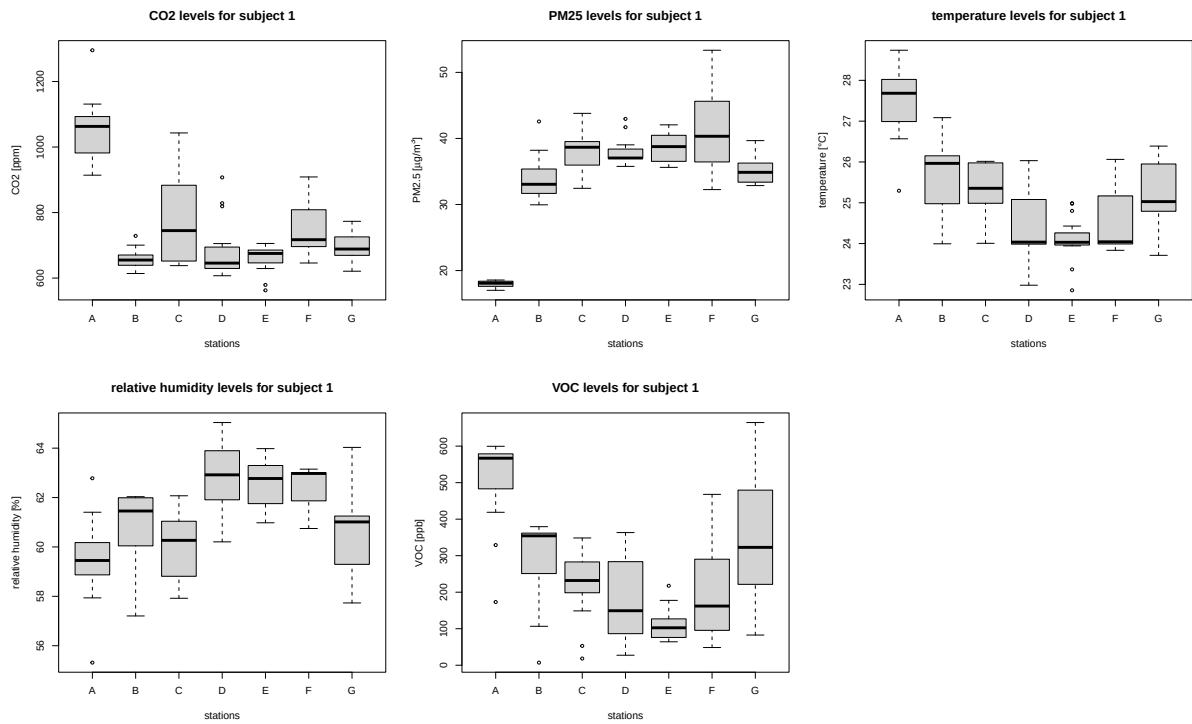
As an initial step, traditional boxplots are examined: they are created to graphically represent the distribution of measurements for each variable at each of the seven checkpoints along the route. This is done both for the HR and FB measurements, as well as for the environmental data. As an illustrative example, those referred to subject 1 are viewable in Figure 5.8. Some other relevant examples, referred to other subjects, are collected in Section A.3 of Appendix A.

It is quite intuitive to observe that the data at station A exhibit a different distribution compared to the other stations, both for physiological variables and environmental variables: in the latter case it is more evident. Additionally, outliers are present at all stations, although their occurrence is not excessive.

However, in order to better identify the outliers and make more informed observations about their behavior, it is decided to conduct an analysis using functional boxplots and outliergrams, in order to identify magnitude and shape outliers which are more informative. To ensure consistency in both the procedure and the conclusions, only the data collected at the seven check points are considered, while measurements from the walking segments between consecutive stations are excluded for the time being. For each subject and each variable, the data are juxtaposed to form a single sequence of observations, which allowed for the application of a functional approach.



(a) Boxplots for variables HR and FB



(b) Boxplots for environmental variables

Figure 5.8: Boxplots for subject 1

Specifically, for each individual and each variable of interest, both a functional boxplot and an outliergram are constructed. These tools enable the identification of stations that exhibit anomalous behavior compared to the others and can therefore be classified as outliers for a given subject and a given variable. Since these methods detect magnitude and shape outliers, a station's behavior is considered anomalous if there is a substantial

deviation from the expected range of value in the former case, or if the curve's shape significantly differs from the others in the latter. Here are some illustrative examples: in Figure 5.9 the functional boxplot for subject 3 is reported and it is clear that there is a single magnitude outlier, station A, highlighted in pink in Figure 5.9a and in black in Figure 5.9b; the outliergram for subject 3 is viewable in Figure 5.10: there is a single shape outlier, station G, which is highlighted in pink in Figure 5.10a and in black in Figure 5.10b. Some other relevant examples of functional boxplots and outliergrams for other individuals are viewable in Section A.4 of Appendix A.

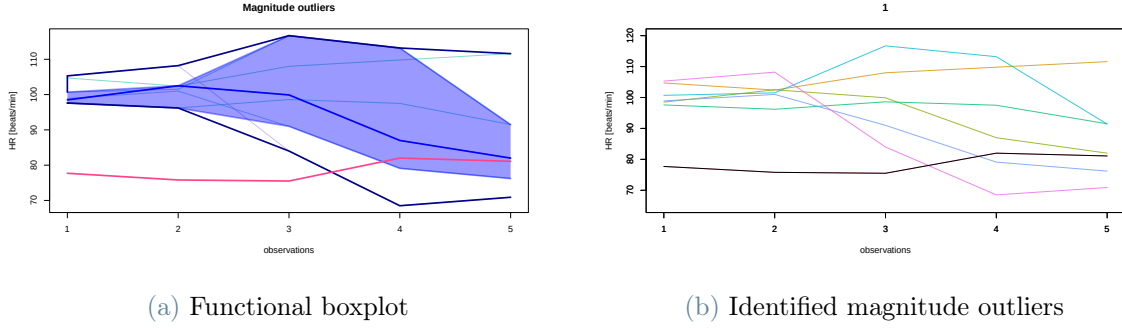


Figure 5.9: Magnitude outliers of HR for subject 3

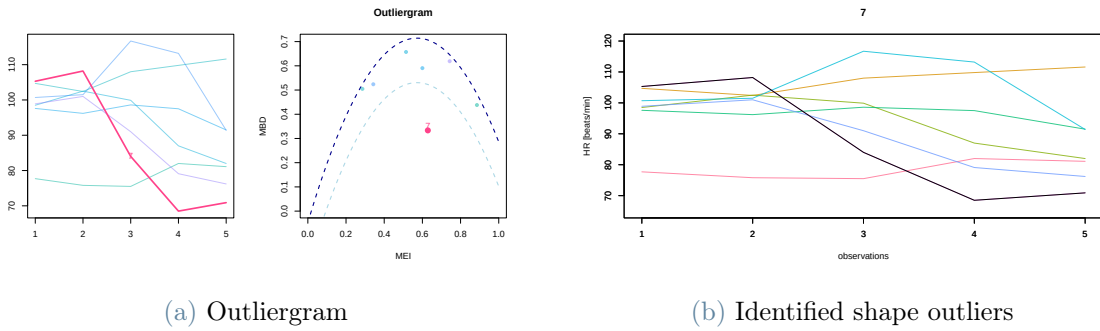


Figure 5.10: Shape outliers of HR for subject 3

What is expected is that station A will frequently exhibit anomalous behavior and thus be classified as an outlier, as already suggested by the previous analyses. However, this does not exclude the possibility that other stations may also display irregular patterns, which

may be less easily detectable through visual inspection alone, as conducted in the earlier stages of the analysis. Indeed, each check point along the path followed by individuals during the experimental campaign exhibits distinct characteristics. Some check points are located in parks, where lower pollution levels are expected, potentially leading to reduced physiological strain on participants. Conversely, other checkpoints are situated in highly trafficked areas, where pollutant concentrations are anticipated to be higher. As a result, it is reasonable to expect, for instance, that individuals arriving at these locations after walking from the previous station may show more labored breathing.

Tables 5.1 and 5.2 summarize the results of this outlier analysis and provide a basis for drawing further conclusions.

	HR magn.	HR shape	FB magn.	FB shape	CO ₂ magn.	CO ₂ shape	PM _{2.5} magn.	PM _{2.5} shape
S1	E		B		A,E		A,C,F	
S2	A,E		B,G	G	A	F	F,G	
S3	A	G	G		A,F		A,D	E
S4	A,F,G	F,G	A,F,G		A,C,F		A,G	
S5				C	A		A,B	
S6			A,C,D		A		A,G	B
S7			A		A		A,B,E	
S8	E			C	A		F	
S9	B				A,B,C	D	A,E,G	
S10	A	A	A,B		A		A,D,G	
S11	B,D		A,F	F	A,F		G	
S12	A,G	A			A		A,C	C
S13	D,E	G	A,D,F		A,G	D	A	
S14	A		C,E,F	C	A,B,C		G	
S15	A		B	E	A	G	A,E	
S16	F		A,D		A		C,D,G	G
S17	A		A,B		A		A,B	B
S18	G	F,G	A		A,E	C	A	
S19	A,D				A,F,G		A,D,E	C
S20	B,G	B	A,B,G		A,E		A	

Table 5.1: Magnitude and shape outliers for HR, FB, CO₂ and PM_{2.5}

	temperature magn.	temperature shape	humidity magn.	humidity shape	VOC magn.	VOC shape
S1	A,D,E		A		A,E	
S2	A,E,F		A		A,C	
S3	B,D		A,B		A,G	D
S4		F	A,B,D		A	F
S5	E	A	A,E		F	
S6	A,C		A		A,C,E	
S7	A,C,D	C,D	A		A,B,C	
S8	A,G	B	A		A,B,G	A,G
S9	A,B,G		G	E,F	A	
S10	A,F,G		F,G		A	G
S11	A,E,G		E	A	B,E,G	F,G
S12	A,E,G	F	A,E,G	A	A	
S13	A,G	C	A	A	A,G	A,G
S14	A	C			A,E,G	
S15	A,B,D		B,D		B,G	A,G
S16	A		A	A	E,G	
S17	A,B,G				A,G	
S18	A				A,B,E	
S19	A				A	
S20	A				A	C

Table 5.2: Magnitude and shape outliers for temperature, humidity and VOC

From these tables, several conclusions can be drawn from two different perspectives: across subjects and across variables. Across subjects, there is strong evidence that station A is the one most frequently identified as a magnitude outlier for all participants. This finding is consistent with the results of previous analyses: as the only indoor environment, station A has specific characteristics that cause the device to record values for various variables that are significantly different from those measured in the other stations. This phenomenon occurs consistently for the majority of the participants. Regarding shape outliers, there is no clear evidence that the curve of a particular station is an outlier for nearly all subjects. However, station G frequently shows anomalous behavior in terms of shape, although this is observed in fewer than a half of the subjects. Across variables, a

similar pattern emerges: station A is the most frequent magnitude outlier across all variables, including both physiological and environmental ones. In contrast, the distribution of shape outliers across variables appears to be rather random, probably because each subject is different from the others and also the measurements were recorded in different days with their specific and characteristic environmental conditions.

The outlier detection process conducted in this study has provided essential insights into the variability and potential anomalies within the dataset. By employing functional box-plots and outliergrams it was possible to identify both magnitude and shape outliers. In particular, the results highlight that outliers are present with certain stations exhibiting more frequent anomalies: the heterogeneity of the environmental characteristics of the stations is evident in the behaviors of measurements recorded there. The presence of outliers is not uniformly distributed across subjects, reinforcing the need for a tailored approach in subsequent analyses, which allows for a better understanding of the real impact of different environmental conditions on physiological outcomes: the reliability of these conclusions regarding this type of impact is enhanced by this outlier analysis.

5.3. Model

To gain a more comprehensive and deeper understanding of how physiological variables are influenced by different environmental conditions, a specific and suitable Interrupted Time Series approach is chosen to be employed. For the subsequent analyses, only the data related to the HR and FB variables will be considered. Indeed, the environmental data have been useful in characterizing the stations and identifying differences in data behavior across locations. However, the focus of the following investigation will be on assessing the impact of environmental conditions changes on physiological responses.

5.3.1. Choice of the Appropriate Model

Segmented Regression is regarded as a suitable and effective technique for addressing the study's research questions. This method allows to assess the impact of transitioning between distinct environments on physiological responses. As previously discussed, data related to the cardio-respiratory activity and collected during the experimental campaign consist solely of static measurements recorded at seven designed stations along the path. As a result, natural interruptions are created between measurements, leading to the identification of this sequences in the Interrupted Time Series. Given this specific data configuration, Segmented Regression is chosen as the most appropriate statistical approach, providing also remarkable flexibility in managing complex analytical scenarios, such as those involving multiple interruptions, which are a defining feature of this study. By incorporating this method it can be rigorously evaluated whether significant shifts in environmental conditions correspond to measurable deviations in physiological responses, further reinforcing the study's objective of assessing the direct impact of air quality and urban exposure on human health.

Seasonality is not considered as a particularly relevant issue in this specific analysis using Segmented Regression, as the measurements included in the dataset are recorded over relatively short time intervals. While similar patterns are repeated at each station, as observed in Figure 5.5, each individual remains at a station for only five minutes with only few records taken during this time, so we cannot really assess seasonality basing this affirmation on sequences consisting of at most 15 values.

5.3.2. Jittering Applied over Interrupted Time Series

Following the pre-processing step discussed in Section 3.3, for each parameter the measurements are taken simultaneously for all participants. Since the measurements recorded for each of the 20 individuals involved in the experimental campaign are independent and

given the need to mitigate the issue of limited data availability caused by the small number of recordings and participants, a single ITS per variable is constructed by merging the individuals series. This procedure is followed for HR and FB, which would be considered as the variables of interest for the next analyses. The resulting ITS would contain multiple values at each time point, with vertical patterns that would not be captured by the model and would remain in the residuals as an inexplicable variability. The jittering technique is applied to address this problem: it redistributes the observations over time in a balanced manner while preserving the structure of the interruptions. As a result of this procedure, a single ITS is obtained for HR and FB respectively, incorporating data from all 20 participants. This approach significantly increases the dataset size, which is of particular interest given the limitations that it could cause to Segmented Regression explained in Section 4.3.1, expanding from 35 to 700 observations per sequence, thereby reducing the risk of inaccurate estimates and enhancing the reliability of the conclusions drawn.

5.3.3. Segmented Regression Applied over Case Study

As a first step, to reinforce the conclusions drawn from previous analyses regarding the differences between stations and the behavior of the data within them, it is decided to treat the stations as independent and interchangeable: this is made to facilitate the analysis, being aware of its limitations, although since the individuals start the walk from the same stations and their walk is continuous from one station to another, they would not be independent. This approach allows for the construction of models based on data from only two stations at a time, focusing exclusively on the transition between them. The reason behind this choice is to develop models capable of quantifying the impact of transitioning between any two distinct environmental conditions on the health status of the individuals involved in the study. In the specific case of this study, all participants follow the same predefined route, starting from the very same initial point. However, this type of single-

interruption analysis could be further refined if participants are assigned different starting points along the same path. This adjustment would mitigate the influence of training effects, ensuring that their performance at a given station is not biased by its position within the sequence. For instance, physiological parameters recorded at the final station of the path would inevitably be higher than those at the first station due to accumulated exertion, potentially distorting the analysis.

For each pair of stations, and consequently for each possible interruption, the variables *time*, *intervention* and *timeSince* are created, following the model described in Equation 4.2 (Section 4.3). In accordance with this formulation, a model is constructed, and its results are analyzed. Considering, for example, station B (gas station) and station C (park), the model with HR as the outcome variable is as follows:

$$HR_t = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot intervention_{BC\ t} + \beta_3 \cdot timeSince_{BC\ t} + \epsilon_t \quad (5.1)$$

where HR_t is the outcome variable at time t , $intervention_{BC} = [0, \dots, 0, 1, \dots, 1]$ is a vector of length 200, with the first 100 occurrences coded 0 and the other half coded 1, and $timeSince_{BC}$, $time$ and ϵ are defined as in Equation 4.2.

In Figure 5.11, the model for the HR variable, including the interruption between stations B and C, is presented along with the counterfactual scenario, which represents the hypothetical case in which the intervention did not occur. As clearly shown in the figure, the transition from station B to station C results in an interruption that induces a level change in HR values, which is evident upon visual inspection. This indicates that a shift occurs, specifically an increase in HR values, which takes place instantaneously. However, no long-term effect is observed, as the slope of the regression line remains consistent before and after the interruption, suggesting that the overall trend does not undergo substantial changes.

These conclusions are not only visually identifiable but are also statistically supported

by the p-values of the regression coefficients derived from the model and displayed in the figure. Indeed, for each coefficient a t-test is conducted in order to verify if the coefficient is significantly different from 0 and consequently has an impact on the outcome variable, or not: the null hypothesis of this test for a generic coefficient β is $H_0 : \beta = 0$, and the alternative hypothesis is $H_1 : \beta \neq 0$. The test statistic is compared with a Student's t distribution, p-values are calculated and a threshold of 0.05 is fixed: if the p-value is below the threshold, the related coefficient is significant and H_0 is rejected, otherwise it cannot be rejected and the coefficient is considered not significant. In the next Tables and Figures p_{step} will indicate a p-value corresponding to the coefficient of an *intervention* variable, while p_{slope} will indicate a p-value corresponding to the coefficient of a *timeSince* variable, recalling the interpretation of the results related to them that we will give later on. In the case of the model presented in Equation 5.1, $p_{step} = 1.103184 \cdot 10^{-4}$, corresponding to the p-value of coefficient β_2 , is below the 0.05 threshold, indicating statistical significance. Conversely, $p_{slope} = 4.050975 \cdot 10^{-1}$, corresponding to the p-value of coefficient β_3 , exceeds the 0.05 threshold, leading to the conclusion that the slope change is not significant.

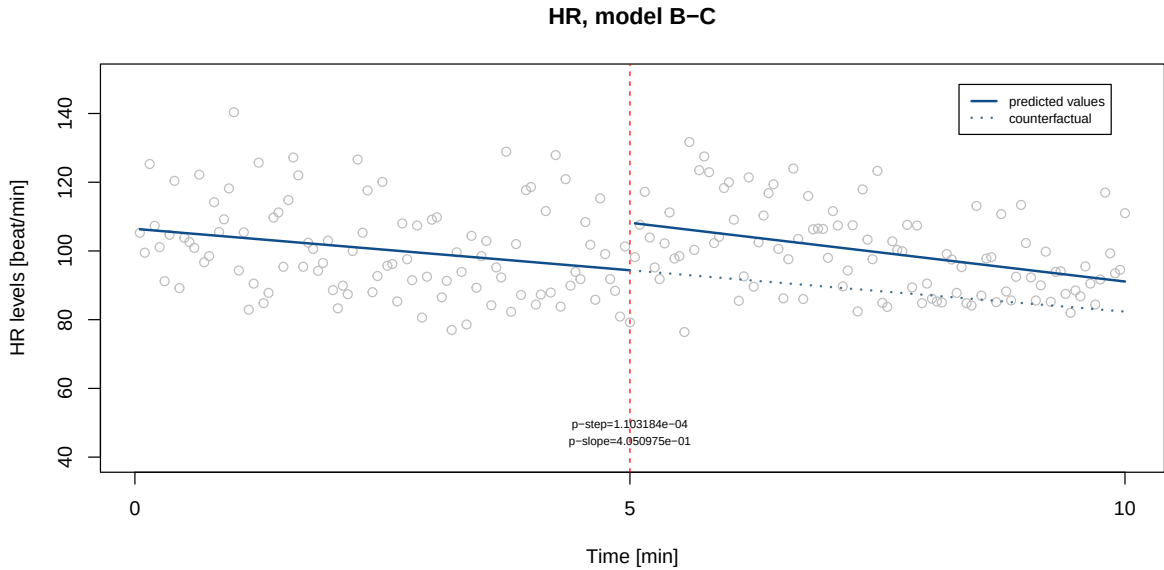


Figure 5.11: Model for HR, with single interruption between stations B and C

After constructing this type of model for each possible pair of stations, including non-consecutive ones, considering HR and FB as outcome variables separately, the corresponding visual representations are generated. Some relevant plots of these can be found in Section A.5 of Appendix A, while the p-values of the regression coefficients β_2 and β_3 are all reported in Table 5.3, together with their level of significance.

	HR - p_{step}		HR - p_{slope}		FB - p_{step}		FB - p_{slope}	
A-B	$2.71 \cdot 10^{-14}$	***	0.442		0.538		0.931	
A-C	$< 2 \cdot 10^{-16}$	***	0.88		0.0335	*	0.5604	
A-D	$< 2 \cdot 10^{-16}$	***	0.774		0.100		0.862	
A-E	$< 2 \cdot 10^{-16}$	***	0.499		0.00387	**	0.51063	
A-F	$3.07 \cdot 10^{-16}$	***	0.549782		0.000921	***	0.265776	
A-G	$< 2 \cdot 10^{-16}$	***	0.854164		0.0234	*	0.8768	
B-C	0.00011	***	0.40510		0.475		0.616	
B-D	0.000107	***	0.643593		0.874		0.786	
B-E	$2.6 \cdot 10^{-6}$	***	0.94030		0.164		0.566	
B-F	$4.57 \cdot 10^{-5}$		0.2510		0.0615	.	0.2988	
B-G	$7 \cdot 10^{-6}$	***	0.4253		0.411		0.948	
C-D	$3.51 \cdot 10^{-7}$	***	0.693		0.584		0.412	
C-E	$5.47 \cdot 10^{-9}$	***	0.455141		0.482		0.976	
C-F	$3.01 \cdot 10^{-7}$	***	0.662081		0.223		0.599	
C-G	$2.47 \cdot 10^{-8}$	***	0.960730		0.882		0.641	
D-E	$1.8 \cdot 10^{-7}$	***	0.70401		0.876		0.350	
D-F	$5.88 \cdot 10^{-6}$	***	0.43606		0.681		0.153	
D-G	$6.4 \cdot 10^{-7}$	***	0.68892		0.492		0.719	
E-F	0.0073	**	0.2842		0.476		0.583	
E-G	0.00202	**	0.47120		0.654		0.588	
F-G	$4.31 \cdot 10^{-7}$	***	0.723		0.921		0.298	

Table 5.3: Models for HR and FB, with a single interruption: p-values of regression coefficients (p_{step} is related to β_2 , p_{slope} is related to β_3)

Legend: *** means $p \leq 0.001$, ** means $p \leq 0.01$, * means $p \leq 0.05$, . means $p < 0.1$

A clear pattern emerges for the FB variable, where statistical significance is observed exclusively in the p-values associated with the regression coefficient β_2 , corresponding to the step change, across some pairs of stations. This suggests that in this cases breathing

frequency undergoes immediate adjustments when environmental conditions change, but does not exhibit long-term variations, as its overall trend remains largely stable. Conversely, for the HR variable, statistical significance is found in the p-values corresponding to β_2 for all the possible pairs of stations, which indicates that heart rate is influenced by changes in environmental conditions immediately after the transitions.

Approaching the analysis through models that incorporate only one interruption at a time has allowed us to further validate several observations from previous analyses and to conclude that changes in environmental conditions have an immediate impact on an individual's health status. Transitioning from a low-pollution area such as a park, to a more traffic-congested zone, for instance, has a measurable effect on physiological responses, leading to increased breathing effort and an elevated heart rate. This effect is observed regardless of the distance already covered before reaching a given station, as the statistical significance of these conclusions has been assessed across all possible station pair combinations.

In order to confirm the homoscedasticity assumption, a visual inspection of the plot of residuals is made for each model: residuals appear randomly scattered without any particular pattern, which suggests that the assumption is respected.

As a next step in this analysis, to identify which environmental transitions have the most significant impact on the 20 participants in this specific case study, a model incorporating all the interruptions between consecutive stations is implemented. Following the formulation in Equation 4.4, the variables *intervention* and *timeSince* are created for interruption, as already explained for the case with one interruption. Selecting HR as the outcome variable for the first model and FB for the second, the resulting models are as follows:

$$\begin{aligned}
HR_t = & \beta_0 + \beta_1 \cdot time + \beta_2 \cdot intervention_{AB\ t} + \beta_3 \cdot timeSince_{AB\ t} + \\
& \beta_4 \cdot intervention_{BC\ t} + \beta_5 \cdot timeSince_{BC\ t} + \beta_6 \cdot intervention_{CD\ t} + \\
& \beta_7 \cdot timeSince_{CD\ t} + \beta_8 \cdot intervention_{DE\ t} + \beta_9 \cdot timeSince_{DE\ t} + \\
& \beta_{10} \cdot intervention_{EF\ t} + \beta_{11} \cdot timeSince_{EF\ t} + \\
& \beta_{12} \cdot intervention_{FG\ t} + \beta_{13} \cdot timeSince_{FG\ t} + \epsilon_t
\end{aligned} \tag{5.2}$$

$$\begin{aligned}
FB_t = & \beta_0 + \beta_1 \cdot time + \beta_2 \cdot intervention_{AB\ t} + \beta_3 \cdot timeSince_{AB\ t} + \\
& \beta_4 \cdot intervention_{BC\ t} + \beta_5 \cdot timeSince_{BC\ t} + \beta_6 \cdot intervention_{CD\ t} + \\
& \beta_7 \cdot timeSince_{CD\ t} + \beta_8 \cdot intervention_{DE\ t} + \beta_9 \cdot timeSince_{DE\ t} + \\
& \beta_{10} \cdot intervention_{EF\ t} + \beta_{11} \cdot timeSince_{EF\ t} + \\
& \beta_{12} \cdot intervention_{FG\ t} + \beta_{13} \cdot timeSince_{FG\ t} + \epsilon_t
\end{aligned} \tag{5.3}$$

The results are then graphically represented and can be viewed in Figures 5.12 and 5.13, while the p-values of the regression coefficients are reported in Table 5.4.

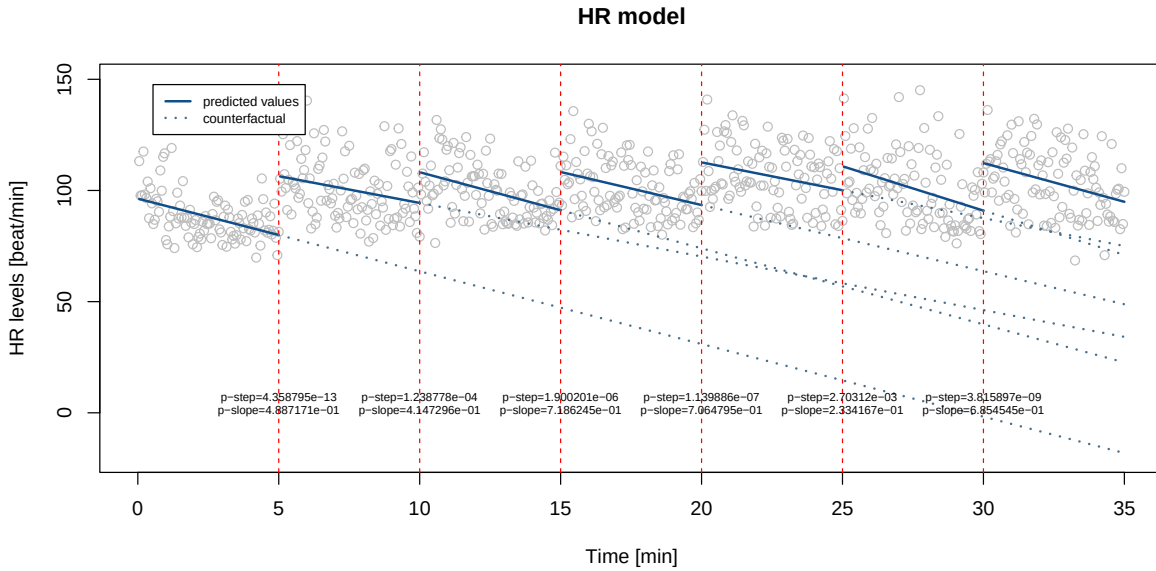


Figure 5.12: Model for HR with all the interruptions

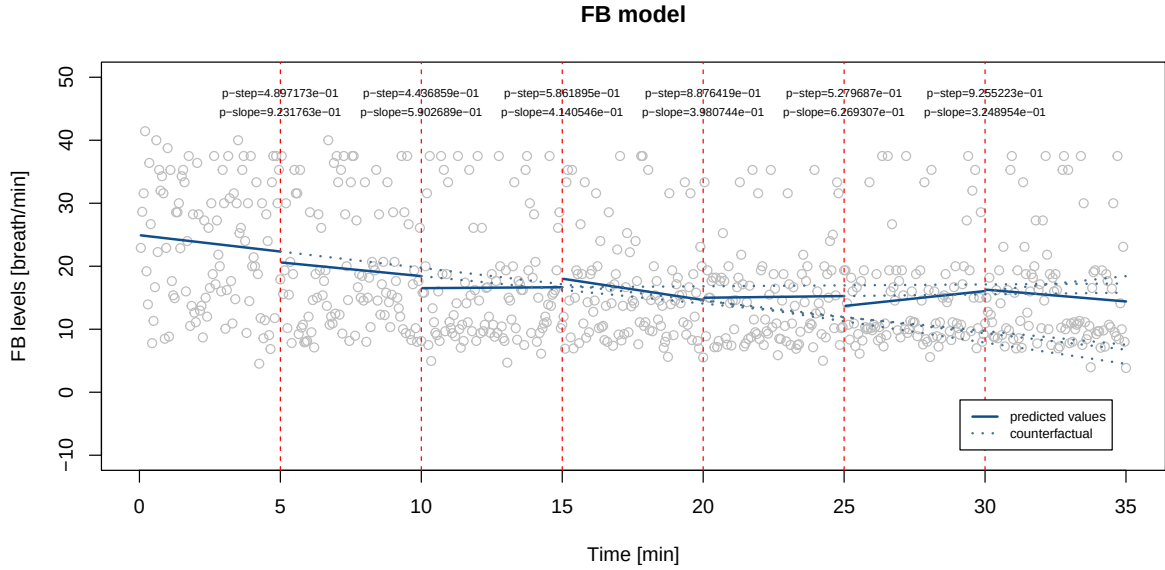


Figure 5.13: Model for FB with all the interruptions

	HR - p_{step}		HR - p_{slope}		FB - p_{step}		FB - p_{slope}	
A-B	$4.36 \cdot 10^{-13}$	***	0.488717		0.490		0.923	
B-C	0.000124	***	0.414730		0.444		0.590	
C-D	$1.90 \cdot 10^{-6}$	***	0.718625		0.586		0.414	
D-E	$1.14 \cdot 10^{-7}$	***	0.706480		0.888		0.398	
E-F	0.002703	**	0.233417		0.528		0.627	
F-G	$3.82 \cdot 10^{-9}$	***	0.685455		0.926		0.325	

Table 5.4: Models for HR and FB, with all the interruptions: p-values of regression coefficients (p_{step} is related to β_2 , p_{slope} is related to β_3) Legend: *** means $p \leq 0.001$, ** means $p \leq 0.01$, * means $p \leq 0.05$, . means $p < 0.1$

Both Table 5.4 and the graph in Figure 5.13 indicate that, for the FB variable, no interruption has a clear and significant impact on its values or overall trend. None of the p-values fall below the 0.05 threshold, and the graph further confirms the absence of level shifts between segments or significant slope changes. This suggests that respiratory frequency is non substantially affected by environmental condition changes along the predefined path but instead exhibits a gradual and uniform increase over time. One might question why models incorporating a single interruption revealed multiple station pairs with highly sig-

nificant level changes, while in this case no pairs exhibit such significance. It is important to note that in Table 5.3, for the FB variable, none of the interruptions between consecutive stations (A-B, B-C, C-D, D-E, E-F, F-G) show statistical significance, either in terms of a level change or a slope variation. These observations suggest that respiratory frequency follows a consistent trend throughout the experimental path, where the effect of prolonged exertion appears stronger than that of environmental transitions. As a result, FB increases steadily, with level shifts emerging only when considering non-consecutive stations. Conversely, for the HR variable, both Figure 5.12 and Table 5.4 indicate that the interruptions have a statistically significant impact on the immediate values of the variable. Specifically the p-values fall below the 0.05 threshold, and this significance is visually evident in the graph from the distinct level shifts between the corresponding segments.

5.3.4. Models with relevant interruptions

As a final step, given the varying significance of the variables included in the models where all the interruptions are considered, it would be reasonable to understand if a model where not all the interruptions are included performs better than the complete one. In order to verify this, the Forward Selection algorithm is employed.

Starting from a model that includes only one interruption, which follows Equation 4.2, the algorithm iteratively adds an additional interruption at each step, introducing *intervention* and *timeSince* variables corresponding to the new interruption added. A comparison is then performed between the current iteration's model with the newly added interruption and that of the previous iteration to determine whether the introduction of the new interruption significantly improves the model. This comparison is conducted using an Anova test. If the added interruption proves the model to be statistically better, it is retained, and the process continues with another iteration. Otherwise, it is removed, and a different interruption is introduced in the next step. At the end of the algorithm, only relevant

interruptions will be included in the model, making it the one that performs better.

Upon completion of this procedure, the final model with HR as outcome is the one considering all the interruptions, meaning that all of them have a significant impact on the value of the physiological parameter. The model can be visualized in Figure 5.14. For the model having FB as outcome, the procedure evaluated the model considering the only interruption between C and D as the one that performs better than the others, meaning the interruption between C and D is the only one that has a really significant impact on parameter FB. The resulting model is viewable in Figure 5.15.

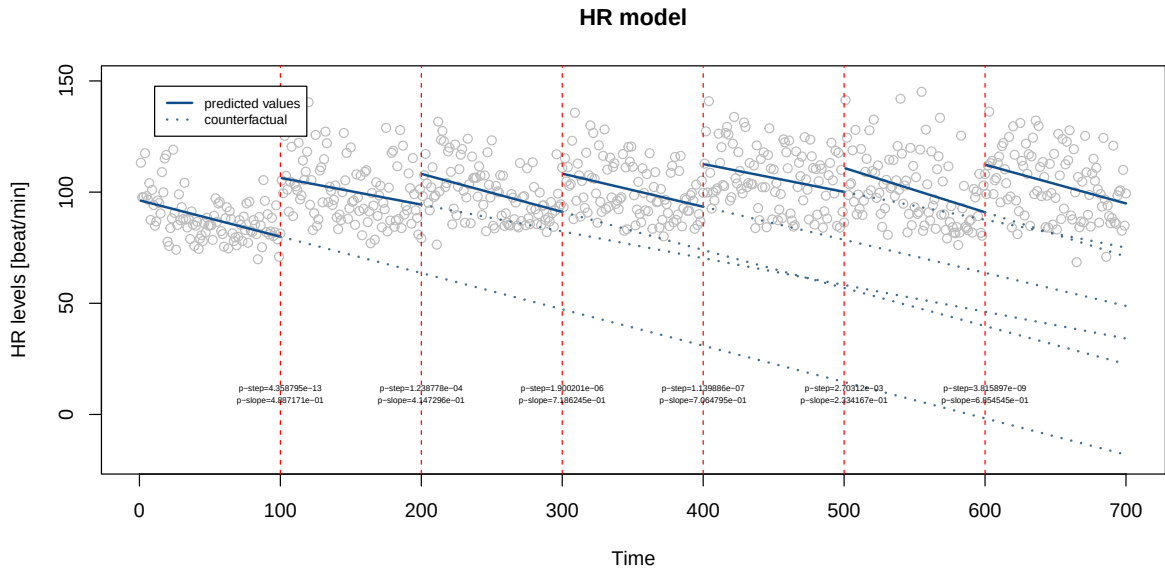


Figure 5.14: Model for HR with the significant interruptions

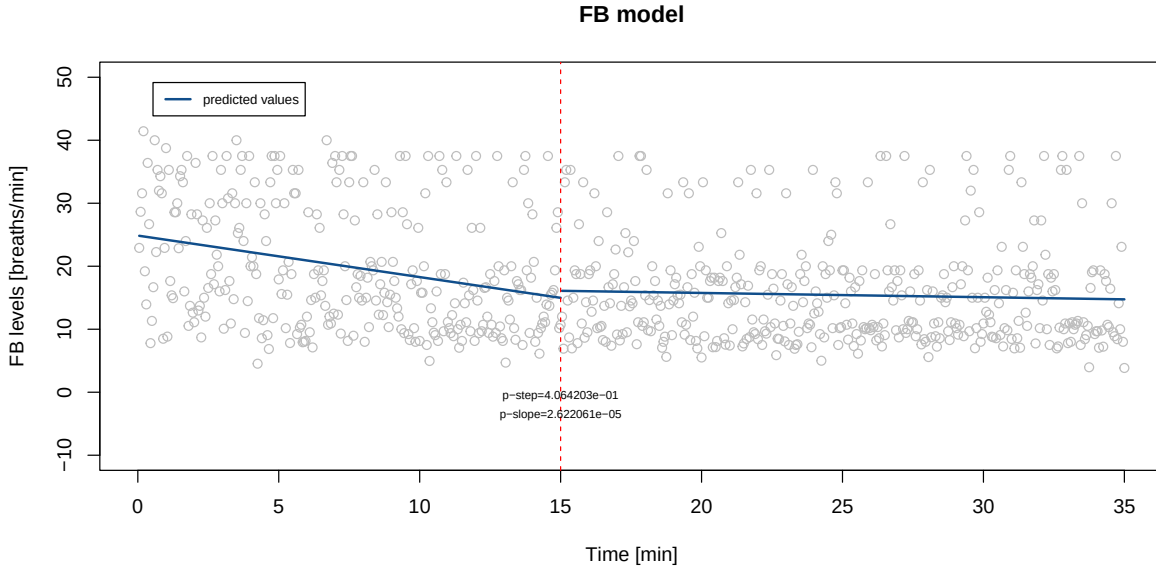


Figure 5.15: Model for FB with the significant interruptions

5.4. Sensitivity analysis

In order to understand if the observations on the results obtained from the conducted analyses and the conclusions drawn from the models are robust and stable with respect to variations in data, a sensitivity analysis is conducted. Indeed, through jittering techniques, new data can be obtained over and over again, which are different from one another only because of the random perturbations, which are very small actually but allows to perform different simulations of the analysis presented in the previous Sections. Specifically, for each simulation all the models are built, both the ones including a single interruption and the ones including all the interruptions, both with HR and FB as outcomes.

After performing $N = 1000$ simulations, and saving the regression coefficients and related p-values for each model, their distributions are visualized through some boxplots: the following figures are used as illustrative examples referring to the coefficients and p-values of the model with all the interruptions with HR as outcome (Figures 5.16-5.17); in Section A.6 of Appendix A other relevant examples can be visualized, specifically the distributions

of coefficients and p-values of the model with all the interruptions with FB as outcome and of the models for HR and FB including only the interruption between stations A and B; for all the other pairs of stations the results are omitted since they are similar to the ones presented in Section A.6.

Considering the graph related to the distribution of the coefficients (Figure 5.16), the variability of the coefficients varies with the size of the boxes: if they are compressed, it means there is a low variability in the coefficients, so the model can be evaluated as robust; conversely, wide boxes refers to high variability of the coefficients and a model sensible to variations of data. It is clear that here there is a very low variability within the regression coefficients, meaning that the model constructed are robust and the conclusions drawn in the previous Sections can be considered reliable conclusions. Regarding the plot of the distribution of the p-values, the narrowness of the boxes corresponds to a low variability of p-values and if the box is narrow and located in the proximity of 0 this corresponds to a high significance of the corresponding variable. It can be noticed that the levels of significance correspond to the ones reported in Tables 5.4 and 5.3.

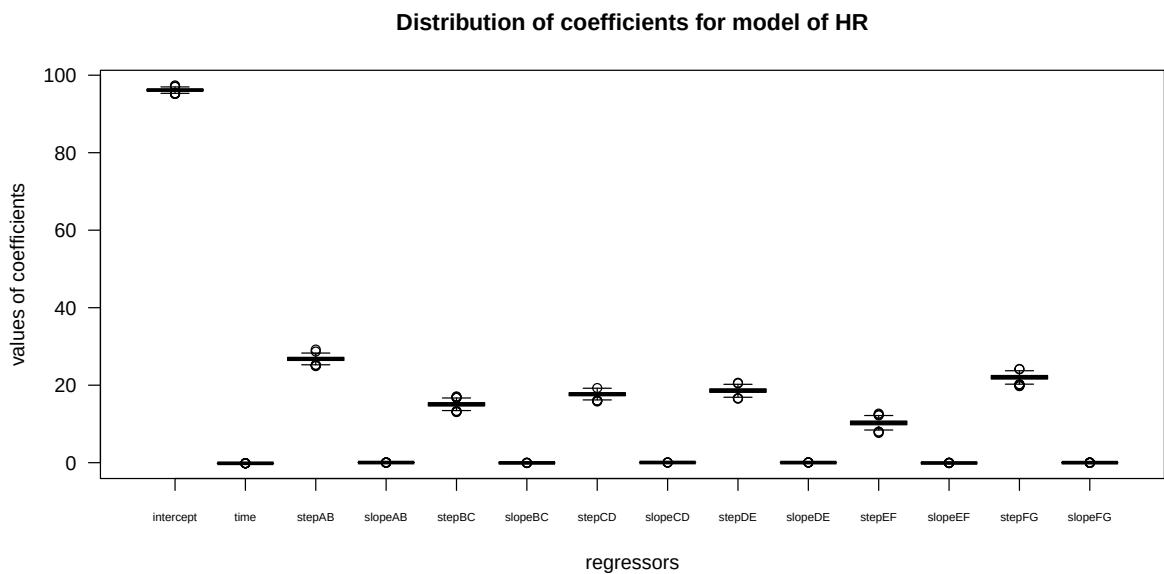


Figure 5.16: Distribution of the coefficients of the model for HR with all the interruptions



Figure 5.17: Distribution of the p-values of the model for HR with all the interruptions

6 | Discussion and Conclusions

Air pollution represents a major and critical global challenge with great implications for human health. For this reason, continuous air quality monitoring is essential, including monitoring of both indoor and outdoor environments. Within the ANTHEM project, which aims at implementing new technologies that could help in reducing the gap between life expectation and healthy life expectation, Bernasconi et al. [12] have introduced an innovative wearable device capable of continuously monitoring of air quality and pollution indicators across various environments while simultaneously tracking physiological parameters of the individual wearing it. This technology enables the collection of data that are ideally suited for an analysis aimed at investigating the relationship between environmental conditions and human health. Once this relationship is clearly identified and wearable devices are fully developed, they could be used by individuals to make decisions in their daily lives—for example, whether to go for a run in a park or in the city neighborhood, or whether to take a walk with an infant or an elderly person, who belong to a vulnerable part of the population. These decisions would be based on environmental conditions and an awareness of their impact on health.

This thesis contributes to the investigation of this relationship with a particular focus on the impact of air pollution and environmental exposure on cardio-respiratory activity.

The analyses conducted in this study are based on physiological and environmental data collected during the experimental campaign carried out by Bernasconi et al. [12]. These data form sequences of observations, classified as Time Series for environmental parameter and Interrupted Time Series for physiological parameters. Initially, a general exploratory

analysis highlights the heterogeneity of environmental conditions among the stations included in the path followed by the study participants, identifying characteristic patterns for each parameter. Subsequently, an outlier detection analysis reinforces the insights gained from the exploratory phase, further validating the variability in environmental exposure and emphasizing the distinct behavior of indoor environments compared to outdoor ones. Finally, Segmented Regression is employed to assess the impact of external interventions on a specific outcome of interest [10][11]. In this study, it enables the evaluation of both short-term and long-term physiological responses to environmental changes. The results demonstrate the ability of Segmented Regression in detecting level and trend variations in physiological variables, caused by the transitions from a location to a different one, characterized by different environmental conditions.

These findings further support the hypothesis of a causal relationship between environmental conditions and human health, particularly demonstrating that changes in environmental exposure have a measurable impact on physiological parameters. Specifically, Heart Rate exhibits significant deviations when transitioning between locations with contrasting pollution levels, such as moving from a gas station to a park. These variations manifest both as immediate shifts and as long-term effects. Conversely, while changes in Breathing Frequency are also observed, they are generally less pronounced and statistically significant only in specific cases. Respiratory frequency appears to be primarily influenced by immediate environmental changes but does not exhibit sustained trend modifications over time.

Overall, these results confirm that transitions between area with differing pollution levels and urban characteristics influence physiological responses, reinforcing the hypothesis that air quality and environmental factors play a fundamental role in shaping human health. Furthermore, they align with and contribute to the scientific understanding of the physiological effects of air pollution exposure.

Despite the robustness of the results, certain limitations of this study must be acknowl-

edged. The collection of data and the design of the study that lead to these drawbacks derive from the necessity of a guarantee of anonymity of data and their reproducibility for the testing of the wearable device. The relatively small sample size, both in terms of the number of individuals involved in the experimental campaign and the number of observations recorded, poses constraints on causal inference, limiting the generalizability of the findings. Expanding the sample size by including a larger and more diverse group of participants, as well as extending the duration of the monitored walk, could enhance the reliability and external validity of the results. A larger dataset would allow for a more comprehensive assessment of the relationship between environmental exposure and physiological responses, reducing potential biases introduced by inter-individual variability.

Another limitation that has to be considered is the outline of the experimental campaign. Indeed, participants all begin their walk in the same station and follow all the same path. Inevitably there would be a light increasing trend both for HR and FB as the activity is continuous, despite the resting phases in the stations which are limited actually. This trend may be managed by the use of more sophisticated statistical methods, which are able to limit or erase it. The assumption made over the independence between the stations that allows their interchangeability has to be put under particular attention: given the fact that the participants starts from the same station, data could not be completely independent, but in the study this is assumed to facilitate the analysis. However, the assumption of homoscedasticity of data is verified.

Another important consideration is the discrete nature of physiological measurements in this study, which are collected only at predefined stationary checkpoints. Implementing continuous physiological monitoring could provide a more precise and dynamic representation of how physiological responses evolve in real time under varying environmental conditions. However, such an approach would make it more challenging to explicitly identify physiological changes attributable to environmental transitions, as interruptions would no longer be clearly defined. The trade-off between increased temporal resolution

and the ability to isolate environmental effects should be carefully evaluated in future research.

A further avenue for improvement involves addressing inter-individual variability more explicitly. In this study, a single Interrupted Time Series is constructed by unifying those belonging to the 20 participants, as their measurements are assumed as independent; then the jittering technique is applied to manage overlapping observations within the Interrupted Time Series framework. While effective in ensuring temporal distribution, this method inevitably leads to a loss of precision in capturing individual-specific responses. A promising alternative would be to explore Multivariate Interrupted Time Series methodologies, which allow for a structures and statistically rigorous assessment of heterogeneous responses across participants. Such an approach could provide deeper insights into how different individuals react to environmental transitions, thereby improving the accuracy and interpretability of the results. Another option would be to add some mixed effect to model, in order to highlight the differences between individual and stations.

Other techniques have been considered to address the same research questions and conduct a more in-depth analysis of the problem of this study. In particular, an assessment was made to determine whether the Bayesian methods described in Section 2.2.1 could offer a more effective and advantageous approach. These methods are highly robust, capable of accounting for spatial and temporal dependencies and proficient in mitigating issues related to over-fitting and under-fitting. However, for the specific dataset and the objectives of this study, Segmented Regression was considered an equally valid yet more suitable technique. It is a computationally lighter and more straightforward approach while still addressing critical aspects in the data. Furthermore, the aforementioned Bayesian methods are generally more appropriate for datasets with a substantial number of observations, whereas their application to a dataset of limited size, such as the one used in this study, could lead to imprecise conclusions. Segmented Regression has to deal with the size of the dataset, but the effects on conclusions are lighter compared to scenarios where a Bayesian

method is used. Finally, Segmented Regression is really performing when data are equally distributed in the pre-intervention and post-intervention period: this is the case for the dataset of this study, following the pre-processing step discussed in Section 3.3.

In conclusion, this thesis presents a comprehensive investigation into the relationship between environmental conditions and physiological responses, demonstrating the effectiveness of statistical methodologies in assessing causal effects within the domain of environmental health research. The findings contribute to the scientific discourse on pollution-related health risks and highlights the critical need for continued research and policy efforts aimed at mitigating environmental health hazards.

Bibliography

- [1] Clean Air Act overview - EPA. [Online; accessed 7-Feb-2025] <https://www.epa.gov/clean-air-act-overview/progress-cleaning-air-and-improving-peoples-health>.
- [2] Air-pollution - OECD. [Online; accessed 5-Feb-2025] <https://www.oecd.org/en/topics/air-pollution.html>.
- [3] ARPA Lombardia, . [Online; accessed 19-Feb-2025] <https://www.arpalombardia.it/chi-siamo/cosa-fa-arpa/aria/>.
- [4] Biossido di azoto - ARPA Lombardia, . [Online; accessed 5-Mar-2025] <https://www.arpalombardia.it/temi-ambientali/aria/inquinanti/biossido-di-azoto/>.
- [5] Ozono troposferico - ARPA Lombardia, . [Online; accessed 5-Mar-2025] <https://www.arpalombardia.it/temi-ambientali/aria/inquinanti/ozono-troposferico/>.
- [6] PM10 and PM2.5 - ARPA Lombardia, . [Online; accessed 5-Mar-2025] <https://www.arpalombardia.it/temi-ambientali/aria/inquinanti/pm10-e-pm25/>.
- [7] Inquinamento atmosferico: gli effetti potenziali sulla salute, . [Online; accessed 5-Mar-2025] <https://www.regione.lombardia.it/wps/portal/istituzionale/HP/DettaglioRedazionale/servizi-e-informazioni/cittadini/salute-e-prevenzione/Sicurezza-negli-ambienti-di-vita-e-di-lavoro/inquinamento-atmosferico/inquinamento-atmosferico>.

- [8] M. Agnoletti and S. Neri Seneri. *The basic environmental history. Environmental history*, volume 4:143-69. Springer, 2014.
- [9] S. Araghinejad. *Time series modeling. In: Data-driven modeling: using MATLAB® in water resources and environmental engineering. Water Science and Technological Library*, volume 67. Springer, Dordrecht, 2014.
- [10] J. L. Bernal, S. Cummins, and A. Gasparrini. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International Journal of Epidemiology*, 2017.
- [11] J. L. Bernal, S. Cummins, and A. Gasparrini. The use of controls in interrupted time series studies of public health interventions. *International Journal of Epidemiology*, 2018.
- [12] S. Bernasconi, A. Angelucci, A. Rossi, and A. Aliverti. A new wearable system for personal air pollution exposure estimation. *IEEE Open Journal of Engineering in Medicine and Biology*, 2023.
- [13] R. D. e. a. Brook. Particulate matter air pollution and cardiovascular disease. an update to the scientific statement from the american heart association. *Circulation*, 2010.
- [14] B. Brunekreef and S. T. Holgate. Air pollution and health. *The lancet*, 360, 2002.
- [15] K. Cowling, E. A. Stuart, R. A. Neff, J. Vernick, D. Magraw, and K. P. Porter. The relationship between joining a us free trade agreement and processed food sales, 2002–2016: a comparative interrupted time-series analysis. *Cambridge University Press*, 2020.
- [16] J. D. Cryer and K. S. Chan. *Time Series Analysis*. Springer, 2008.
- [17] F. Dablander. An introduction to causal inference. *Preprinted by PsyArXiv*, 2020.
- [18] D. W. Dockery, C. A. Pope, III, and X. e. a. Xu. An association between air pollution

- and mortality in six u.s. cities. *The New England Journal of Medicine*, 329:1753–59, 1993.
- [19] J. E. Ewusie, L. Thabane, J. Beyene, S. E. Straus, and J. S. Hamid. Multicenter interrupted time series analysis: incorporating within and between-center heterogeneity. *Clinical Epidemiology*, Dove Press, 2020.
- [20] T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis*. John Wiley & Sons, 1991.
- [21] C. Gascoigne, A. Jeffery, Z. Shao, S. Geneletti, J. B. Kirkbride, G. Baio, and M. Blangiardo. A bayesian interrupted time series framework for evaluating policy change on mental well-being: An application to england’s welfare reform. *Elsevier*, 2024.
- [22] G. Hammerton and M. R. Mufanò. Causal inference with observational data: the need for triangulation of evidence. *Psychological Medicine*, 51:563–78, 2021.
- [23] D. M. Hawkins. *Identification of outliers*. Springer, 1980.
- [24] G. Hoek, J. D. Schwartz, B. Groot, and P. Eilers. Effects of ambient particulate matter and ozone on daily mortality in rotterdam, the netherlands. *Archives of Environmental & Occupational Health*, 52:455–63, 1997.
- [25] G. Hoek, R. M. Krishnan, R. Beelen, A. Peters, B. Ostro, B. Brunekreef, and J. D. Kaufman. Long-term air pollution exposure and cardio-respiratory mortality: a review. *Environmental Health*, 2013.
- [26] J. Hudson, S. Fielding, and C. R. Ramsay. Methodology and reporting characteristics of studies using interrupted time series design in healthcare. *BMC Medical Research Methodology*, 2019.
- [27] D. Hume. *Enquiries Concerning Human Understanding*. 1748.
- [28] D. Hume. *An enquiry concerning the Principles of Morals*. 1751.

- [29] K. Katsouyanni, G. Touloumi, C. Spix, J. Schwartz, F. Balducci, S. Medina, G. Rossi, B. Wojtyniak, J. Sunyer, L. Bacharova, J. P. Schouten, A. Ponka, and H. R. Anderson. Short term effects of ambient sulphur dioxide and particulate matter on mortality in 12 european cities: results from time series data from the aphea project. *British Medical Journal*, 314:1658–63, 1997.
- [30] R. Lovelace, R. Felix, and D. Carlino. Jittering: a computationally efficient method for generating realistic route networks from origin-destination data. *Findings*, 2022.
- [31] Y. Lu. Beyond air pollution at home: Assessment of personal exposure to pm 2.5 using activity-based travel demand model and low-cost air sensor network data. *Environ Res*, 201:111549, 2021.
- [32] F. H. Messerli. Chocolate consumption, cognitive function, and nobel laureates. *The New England Journal of Medicine*, 2012.
- [33] F. Nyberg, P. Gustavsson, L. Jarup, T. Bellander, N. Berglind, R. Jakobsson, and G. Pershagen. Urban air pollution and lung cancer in stockholm. *Epidemiology*, 11: 487–95, 2000.
- [34] R. B. Penfold and F. Zhang. Use of interrupted time series analysis in evaluating health care quality improvements. *Academic Pediatrics*, 2013.
- [35] C. A. Pope, III, M. J. Thun, M. M. Namboodiri, D. W. Dockery, J. S. Evans, F. E. Speizer, and C. W. Heath. Particulate air pollution as a predictor of mortality in a prospective study of u.s. adults. *American Journal of Respiratory and Critical Care Medicine*, 151:669–74, 1995.
- [36] C. A. Pope, III, and D. W. Dockery. Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air & Waste Management Association*, 2006.
- [37] C. A. I. Pope. Mortality and air pollution: associations persist with continued ad-

- vances in research methodology. *Environmental Health Perspectives*, 107:613–14, 1999.
- [38] C. R. Ramsay, L. Matowe, R. Grilli, J. M. Grimshaw, and R. E. Thomas. Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behaviour change strategies. *International Journal of Technology Assessment in Health Care*, 2003.
- [39] M. Rana, J. Kosar, and S. Peermohamed. Bayesian hierarchical models incorporating measurement error for interrupted time series design. *Springer*, 2023.
- [40] W. H. Roemer and J. H. Van Wijnen. Daily mortality and air pollution along busy streets in amsterdam, 1987–1998. *Lippincott Williams & Wilkins*, 2001.
- [41] J. Saini, M. Dutta, and G. Marques. Sensors for indoor air quality monitoring and assessment through internet of things: a systematic review. *Environ Monit Assess*, 193(2), 2011.
- [42] J. Schwartz. Air pollution and daily mortality: a review and a meta-analysis. *Environmental Research*, 64:36–52, 1994.
- [43] S. Vedal. Ambient particles and health: lines that divide. *Journal of the Air & Waste Management Association*, 47:551–81, 1997.
- [44] A. P. Verhoeff, G. Hoek, J. Schwartz, and J. Van Wijnen. Air pollution and daily mortality in amsterdam. *Epidemiology*, 7:225–30, 1996.
- [45] A. K. Wagner, S. B. Soumerai, F. Zhang, and D. Ross-Degnan. Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*, 2002.
- [46] Y. Xu, L. Yi, J. Cabison, M. Rosales, K. O’Sharkey, T. A. Chavez, M. Johnson, F. Lurmann, N. Pavlovic, T. M. Bastain, C. V. Breton, J. P. Wilson, and R. Habre.

The impact of gps-derived activity spaces on personal pm 2.5 exposures in the madres cohort. *Environmental research*, 214:114029, 2022.

- [47] F. Zhang, A. K. Wagner, and D. Ross-Degnan. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. *Journal of Clinical Epidemiology*, 64:1252–61, 2011.

A | Appendix A

A.1. Evolutions of environmental variables per subject

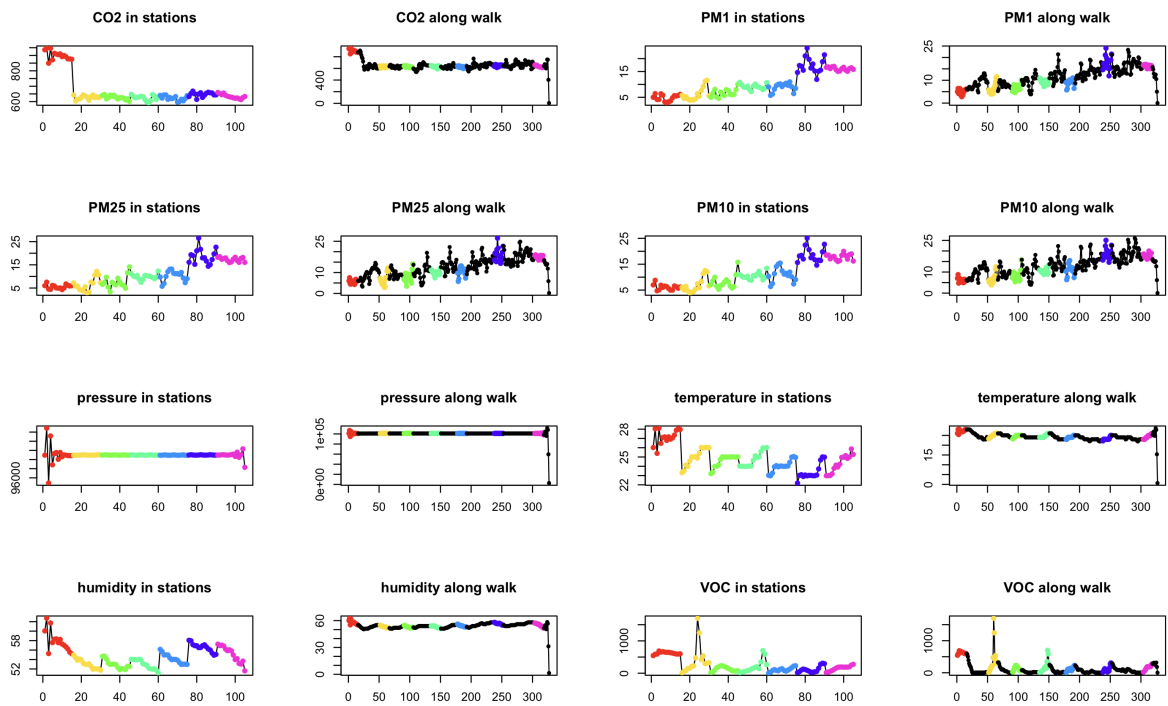


Figure A.1: Evolutions of environmental variables for subject 2

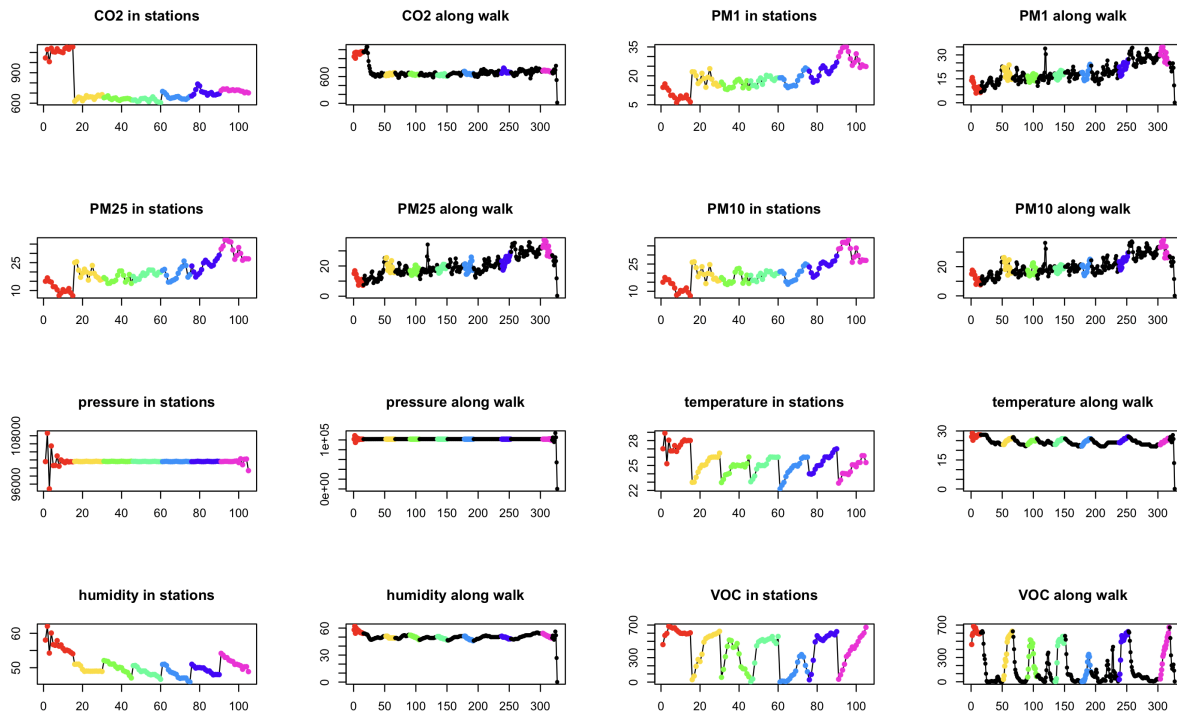


Figure A.2: Evolutions of environmental variables for subject 6

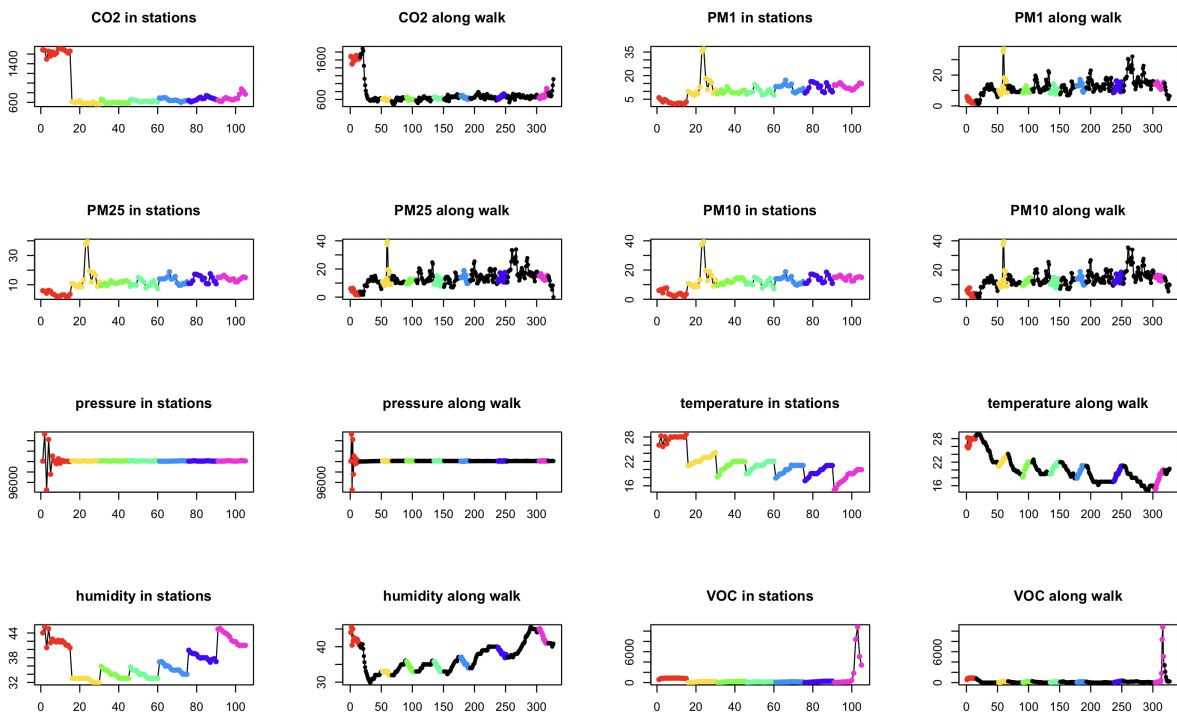
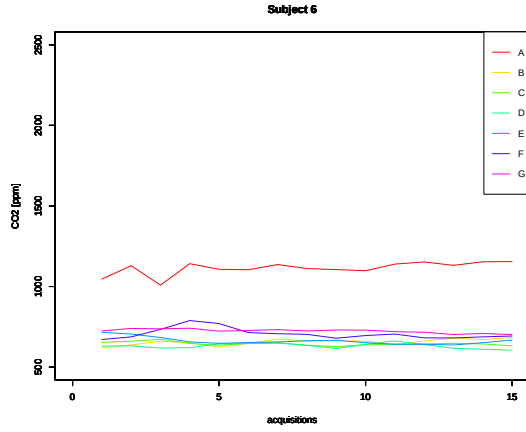
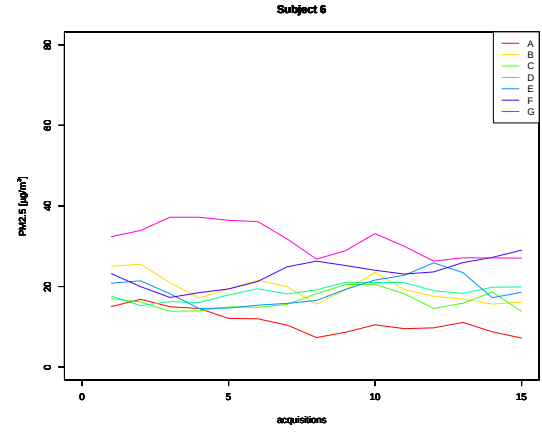
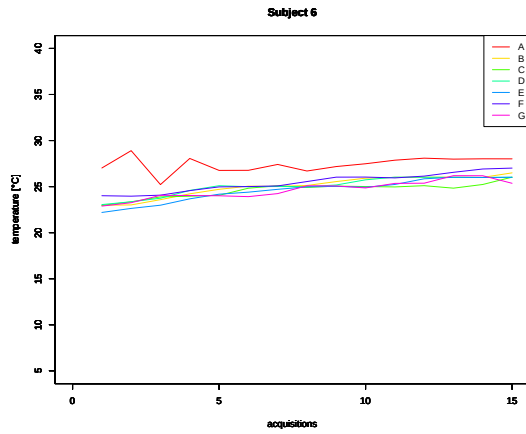
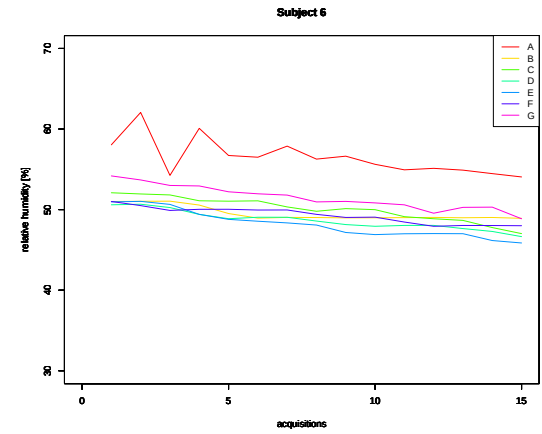


Figure A.3: Evolutions of environmental variables for subject 17

A.2. Comparisons between stations per subject

(a) Comparison for CO_2 (b) Comparison for $PM_{2.5}$ 

(c) Comparison for temperature



(d) Comparison for humidity

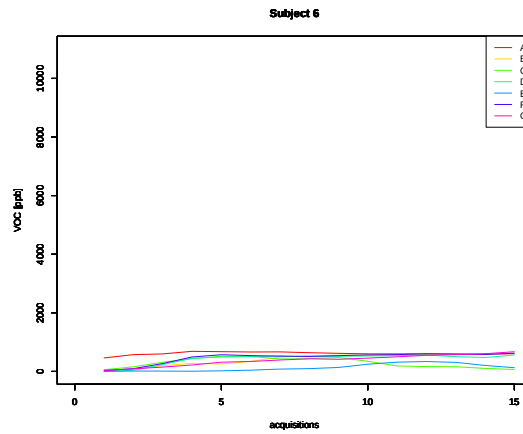
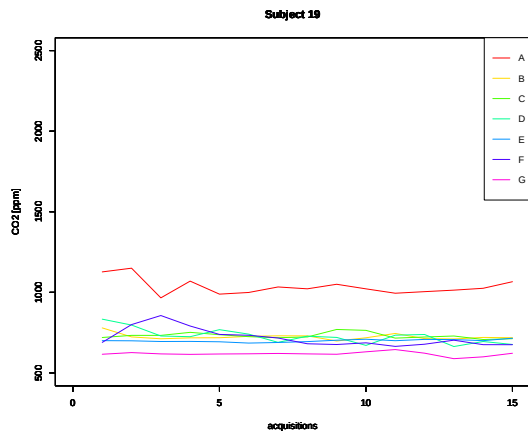
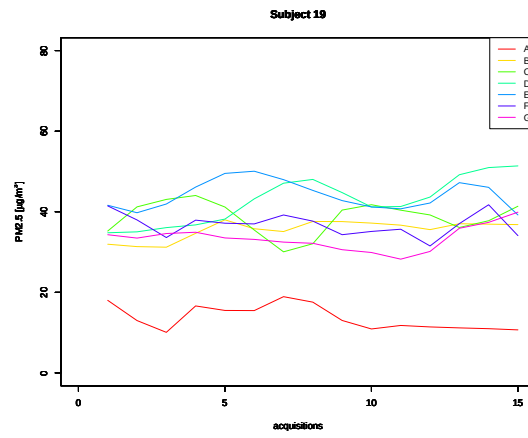
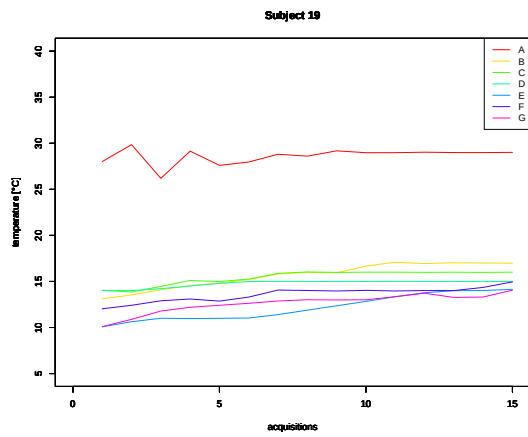
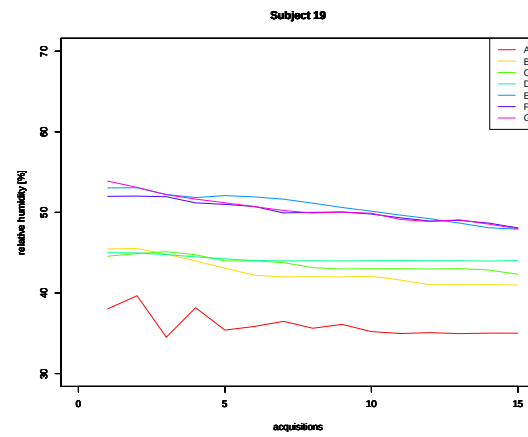
(e) Comparison for VOC

Figure A.4: Comparisons between stations for subject 6

(a) Comparison for CO_2 (b) Comparison for $PM_{2.5}$ 

(c) Comparison for temperature



(d) Comparison for humidity

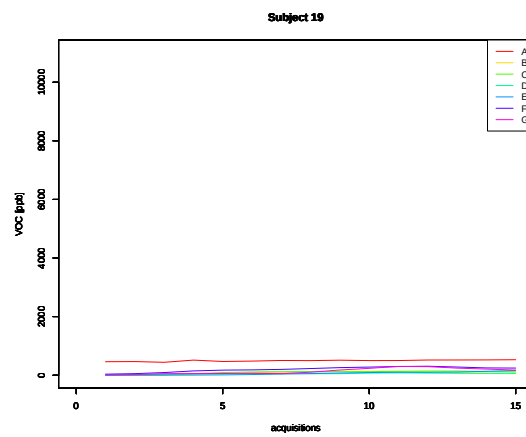
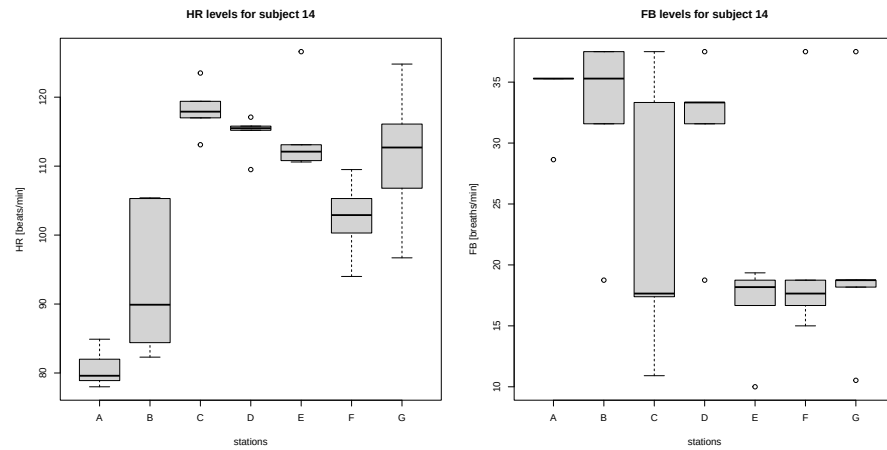
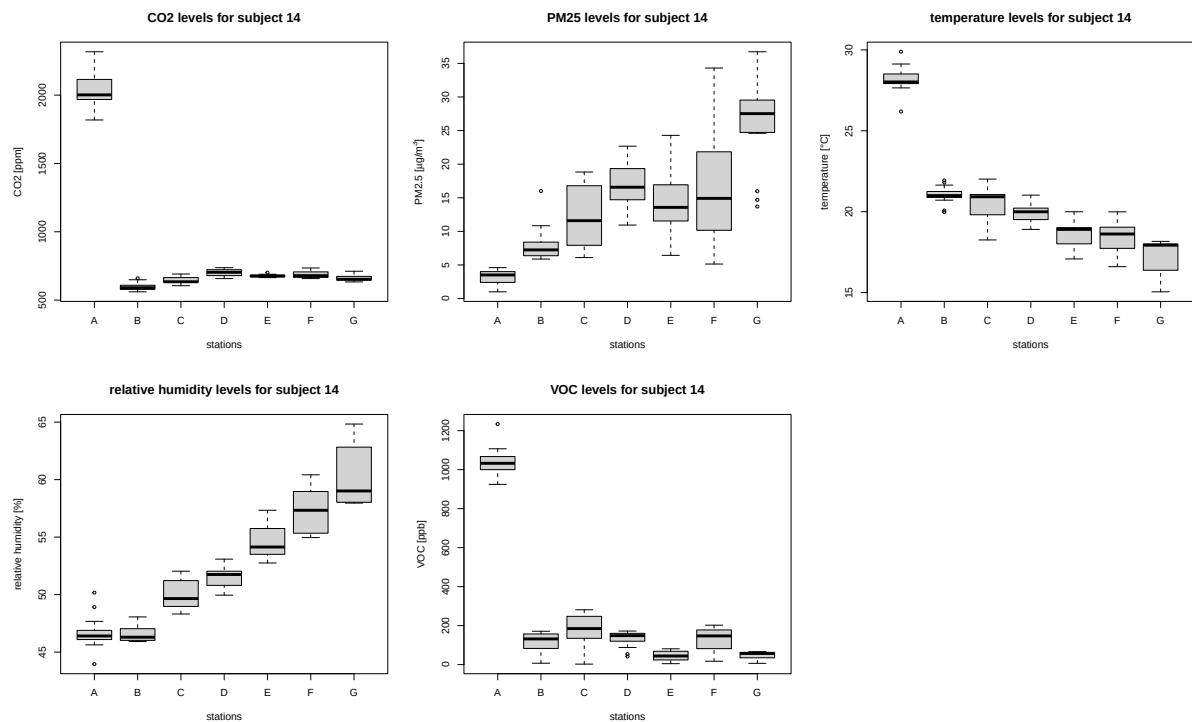
(e) Comparison for VOC

Figure A.5: Comparisons between stations for subject 19

A.3. Traditional boxplots per subject

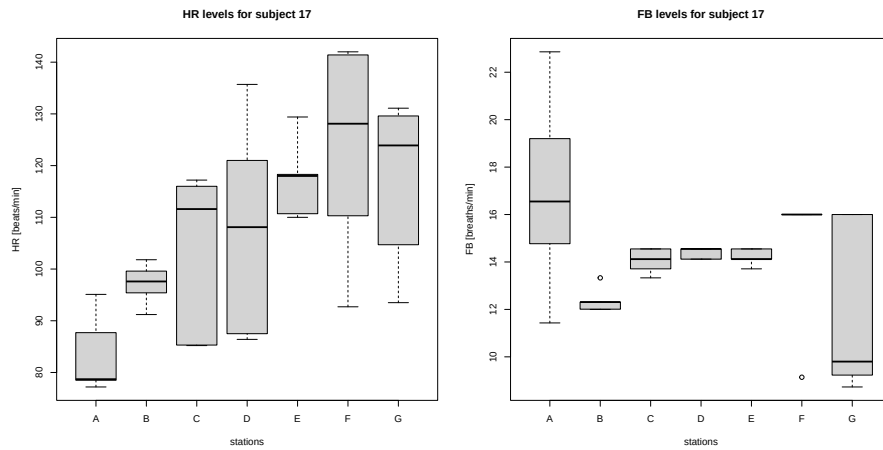


(a) Boxplots for variables HR and FB

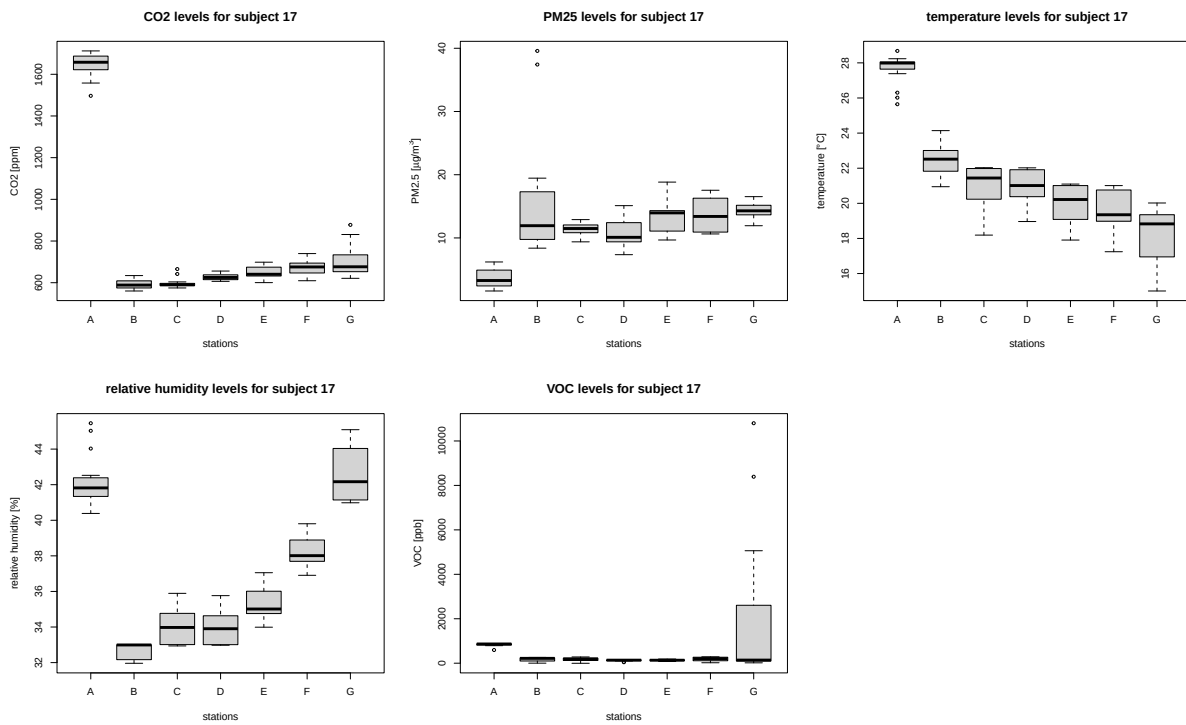


(b) Boxplots for environmental variables

Figure A.6: Boxplots for subject 14



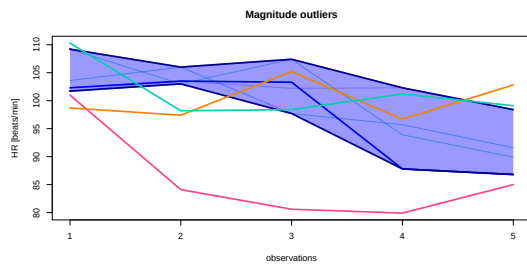
(a) Boxplots for variables HR and FB



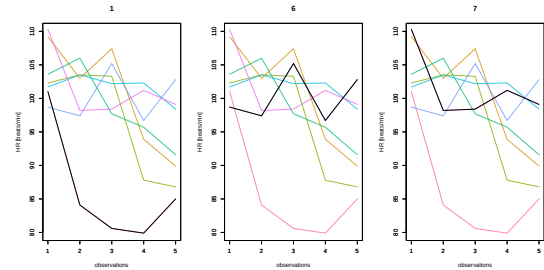
(b) Boxplots for environmental variables

Figure A.7: Boxplots for subject 17

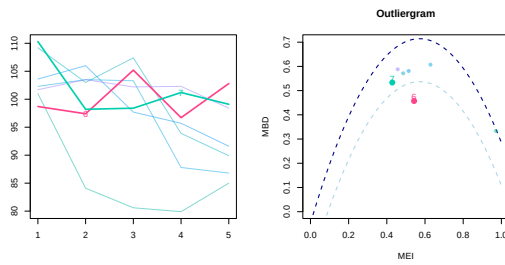
A.4. Magnitude and shape outliers per subject



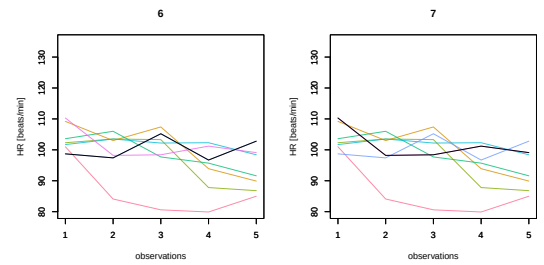
(a) Functional boxplot



(b) Magnitude outliers

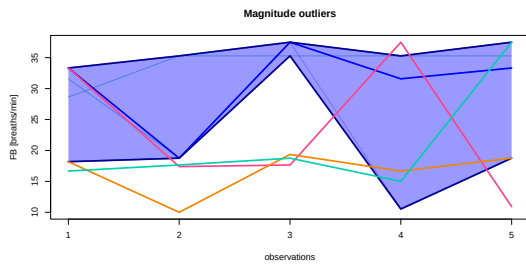


(c) Outliergram

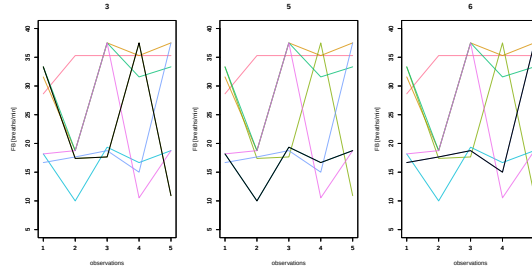


(d) Shape outliers

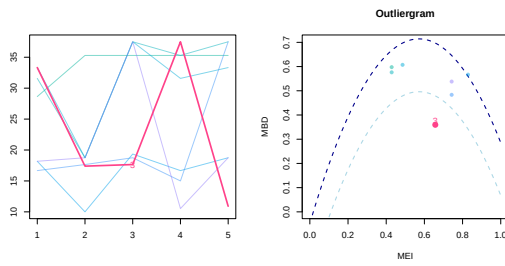
Figure A.8: Magnitude and shape outliers of HR for subject 4



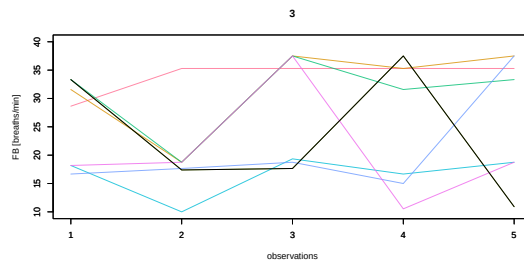
(a) Functional boxplot



(b) Magnitude outliers

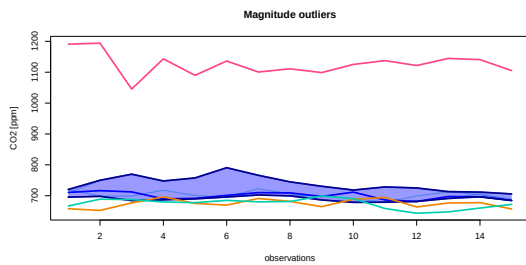


(c) Outliergram

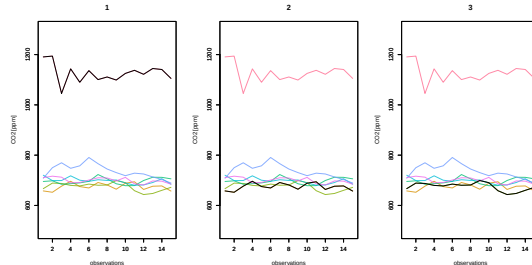


(d) Shape outliers

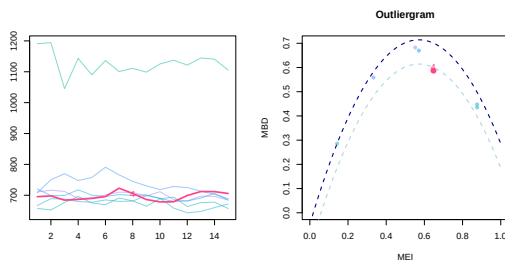
Figure A.9: Magnitude and shape outliers of FB for subject 14



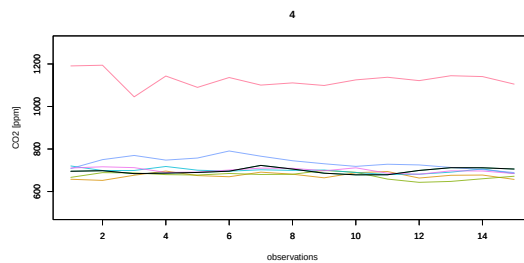
(a) Functional boxplot



(b) Magnitude outliers



(c) Outliergram



(d) Shape outliers

Figure A.10: Magnitude and shape outliers of CO_2 for subject 9

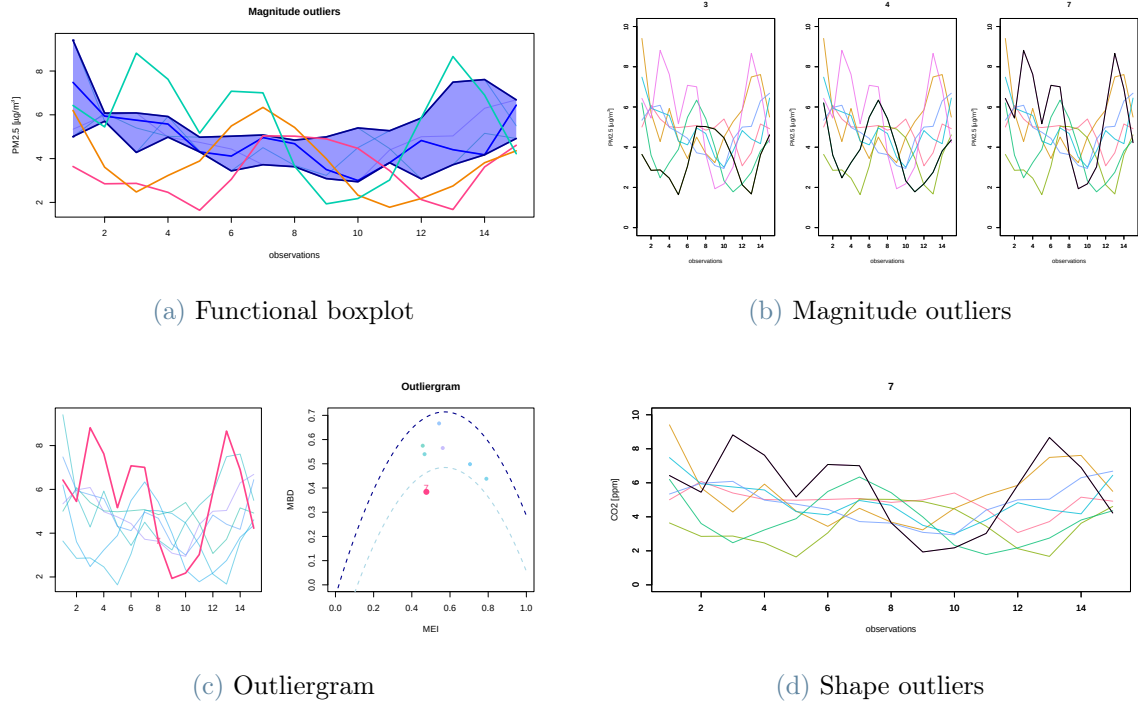
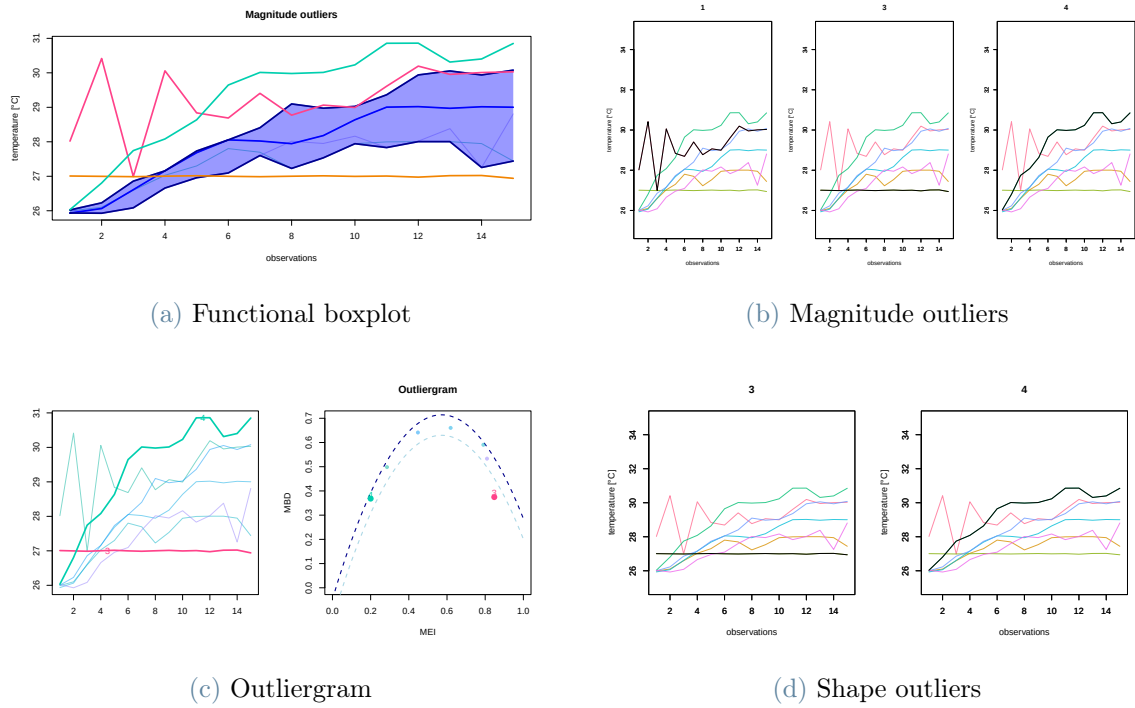
Figure A.11: Magnitude and shape outliers of $PM_{2.5}$ for subject 16

Figure A.12: Magnitude and shape outliers of temperature for subject 7

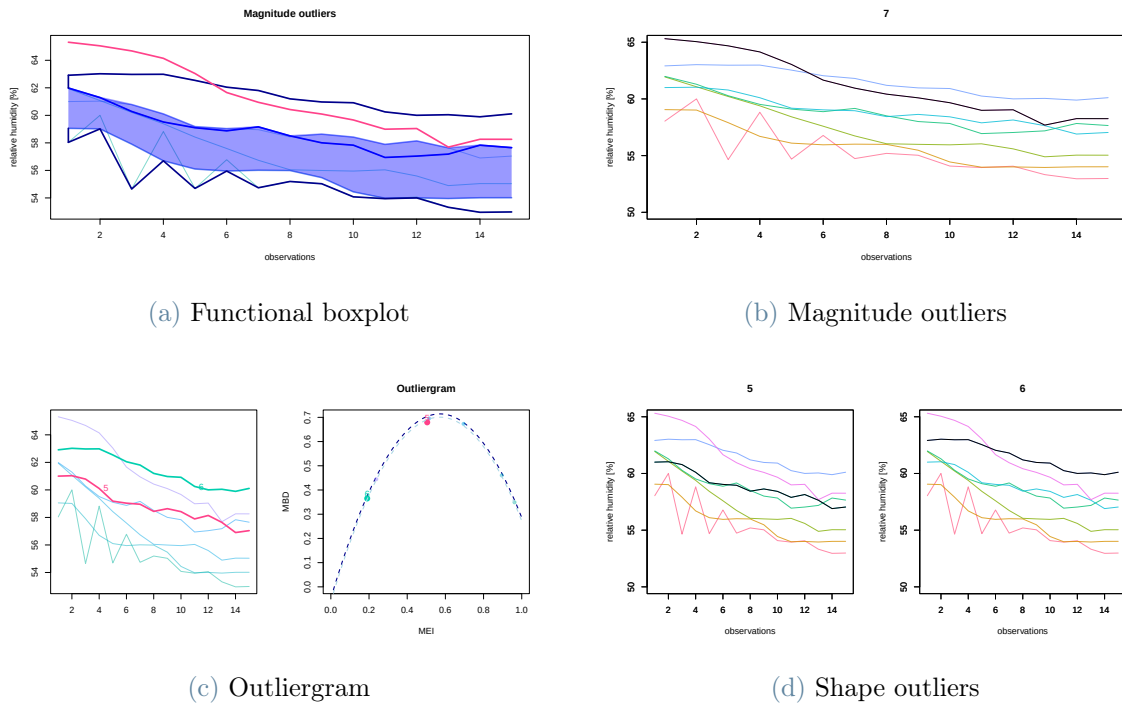


Figure A.13: Magnitude and shape outliers of humidity for subject 9

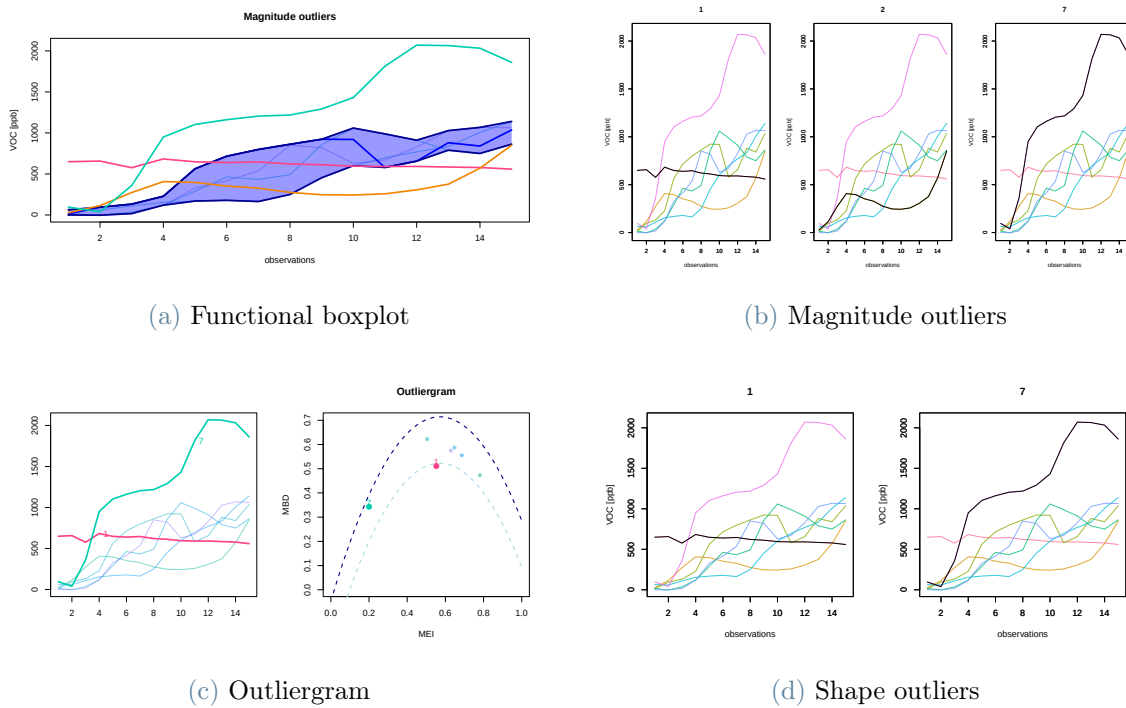


Figure A.14: Magnitude and shape outliers of VOC for subject 8

A.5. Models for HR and FB, with single interruption between 2 stations

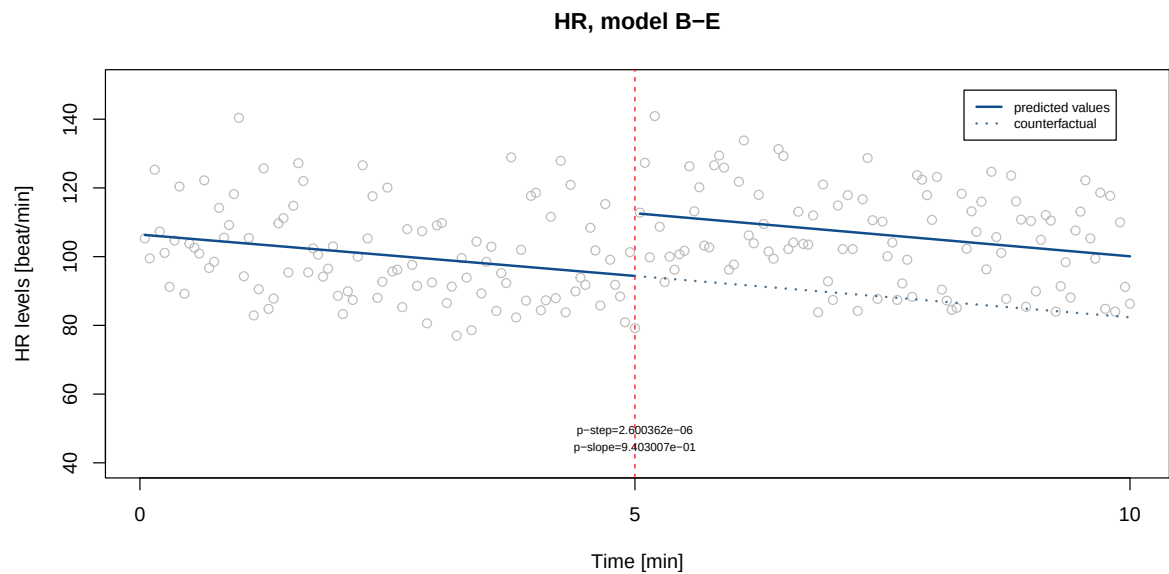


Figure A.15: Model for HR, with single interruption between stations B and E

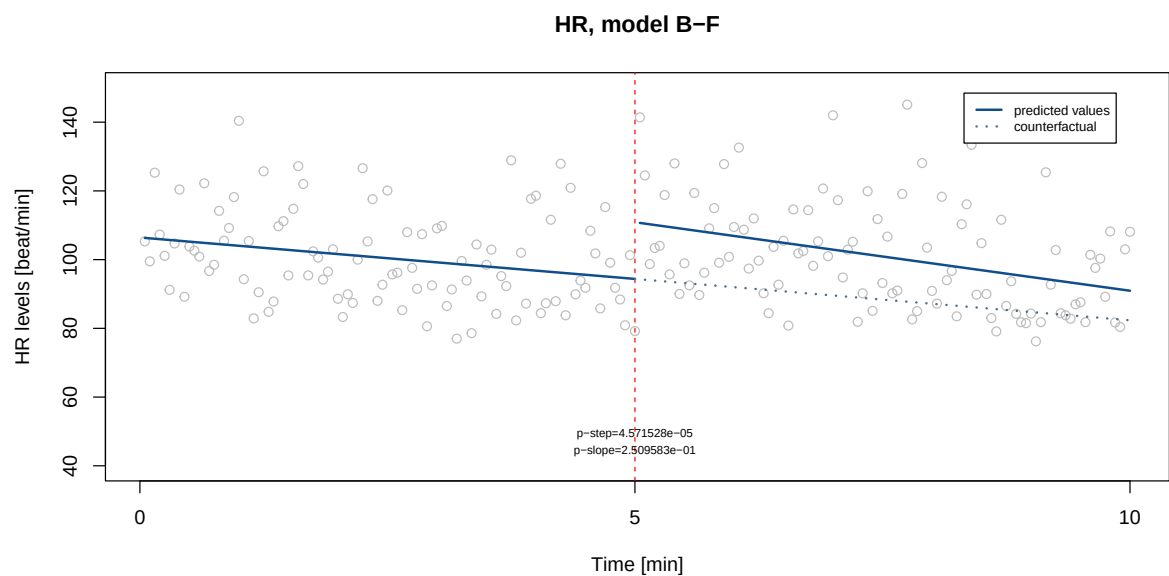


Figure A.16: Model for HR, with single interruption between stations B and F

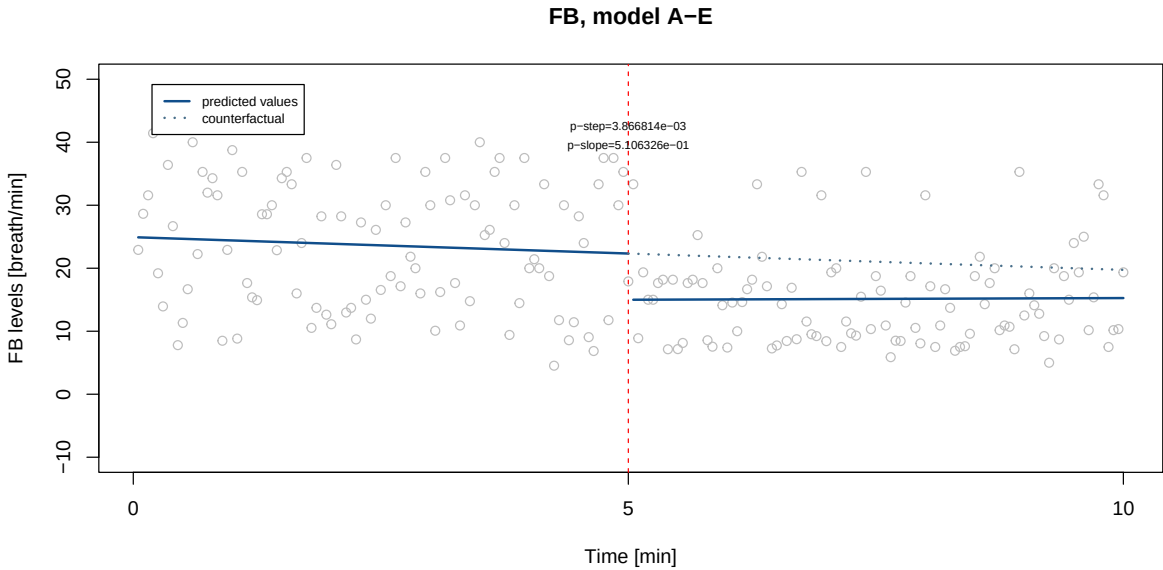


Figure A.17: Model for FB, with single interruption between stations A and E

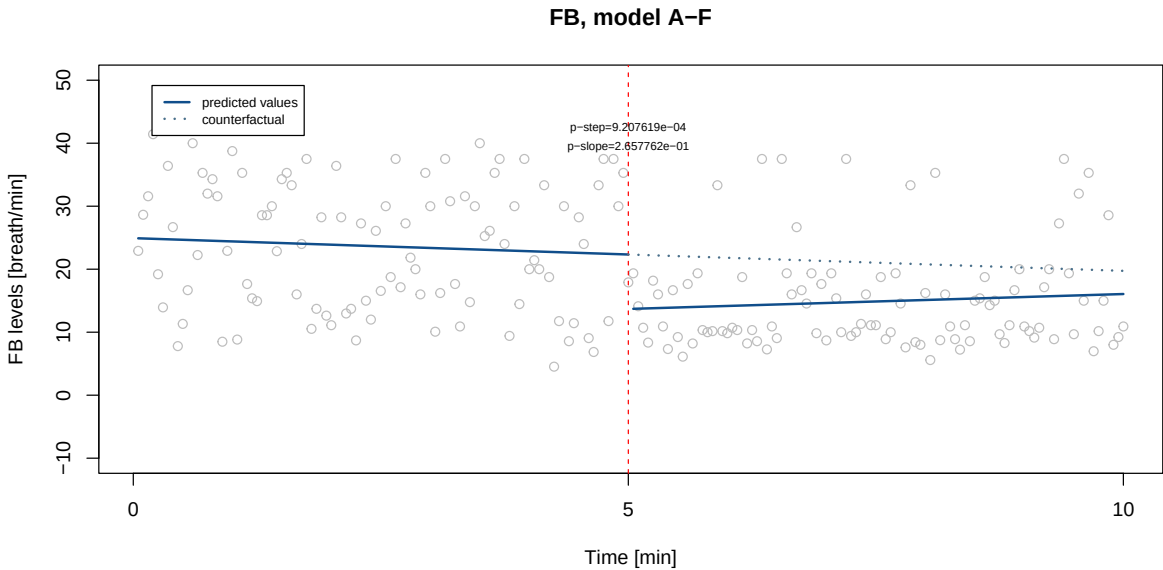


Figure A.18: Model for FB, with single interruption between stations A and F

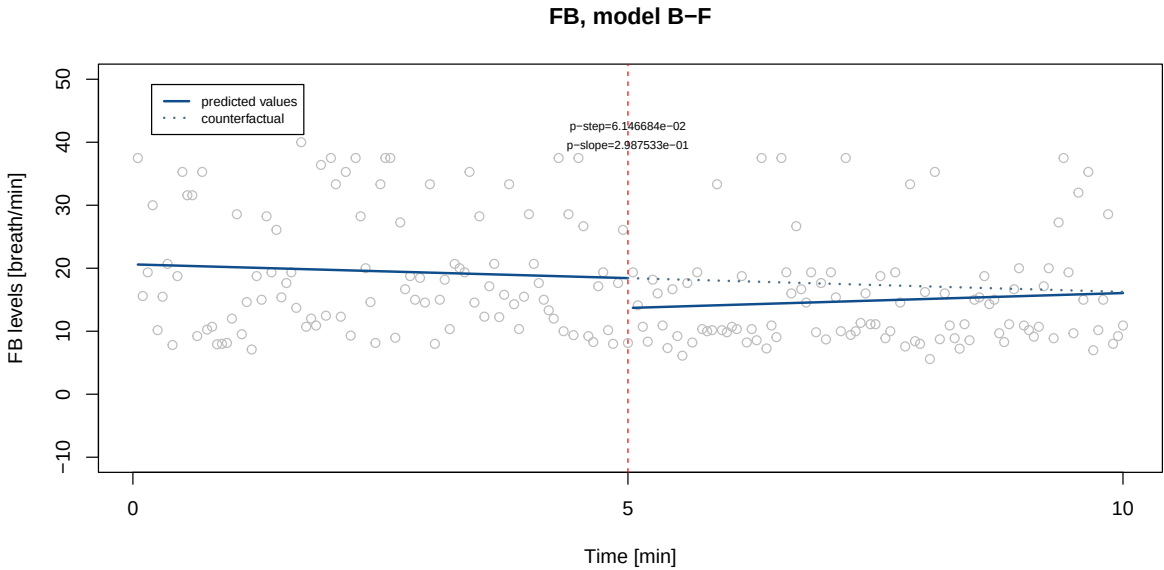


Figure A.19: Model for FB, with single interruption between stations B and F

A.6. Sensitivity analysis

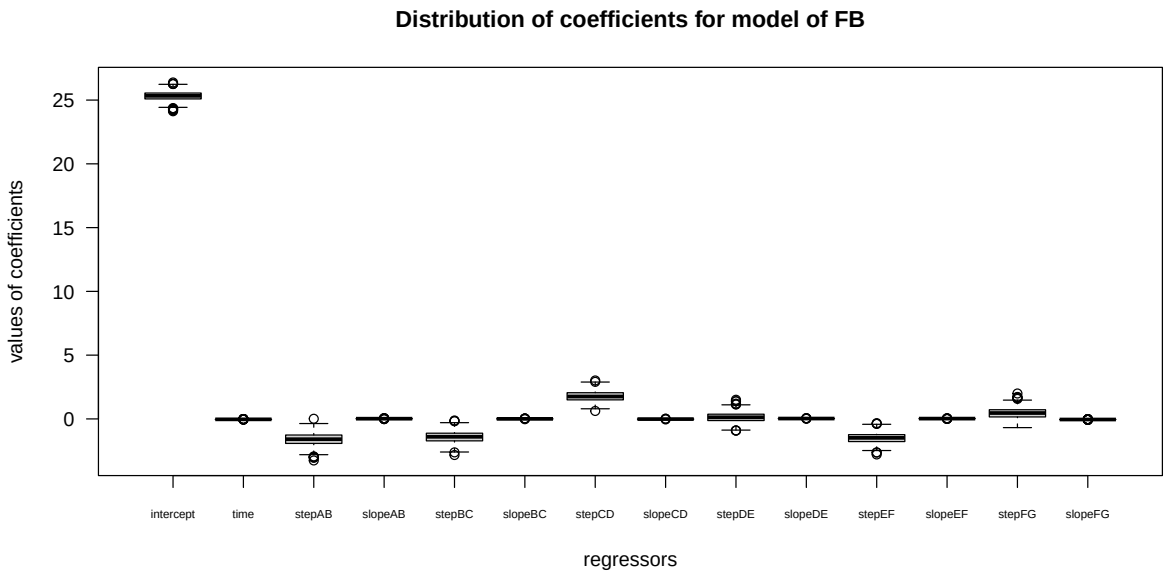


Figure A.20: Distribution of the coefficients of the model for FB with all the interruptions

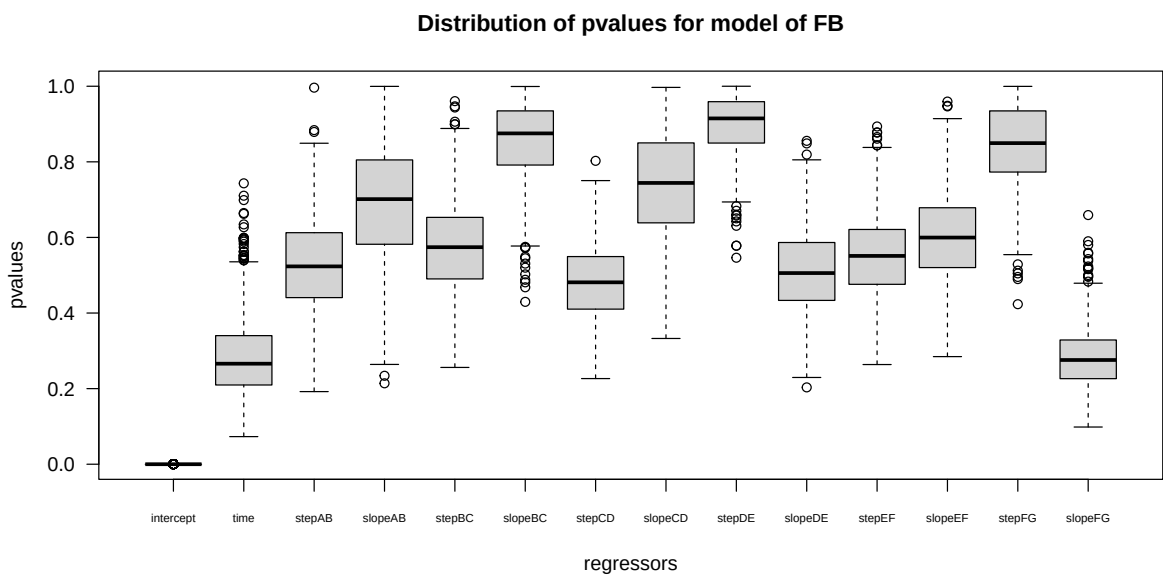


Figure A.21: Distribution of the p-values of the model for FB with all the interruptions

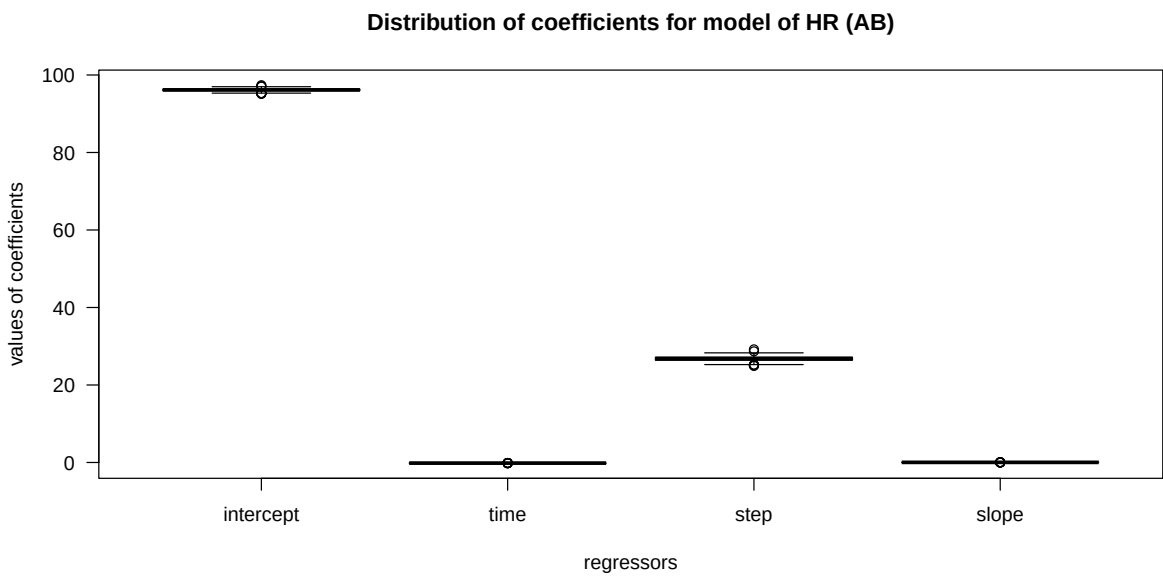


Figure A.22: Distribution of the coefficients of the model for HR with single interruption

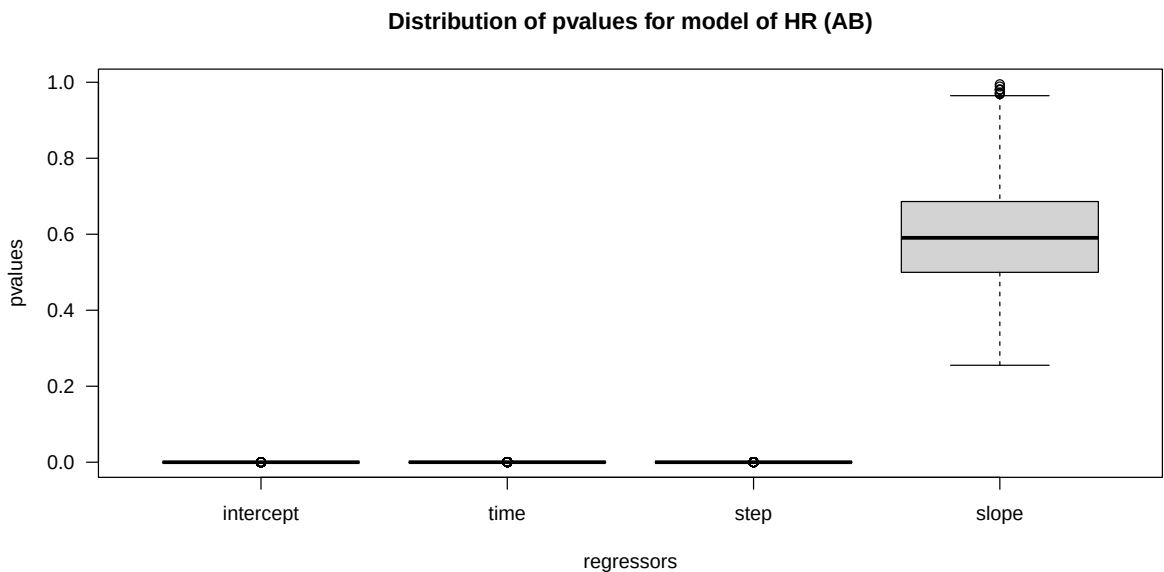


Figure A.23: Distribution of the p-values of the model for HR with single interruption

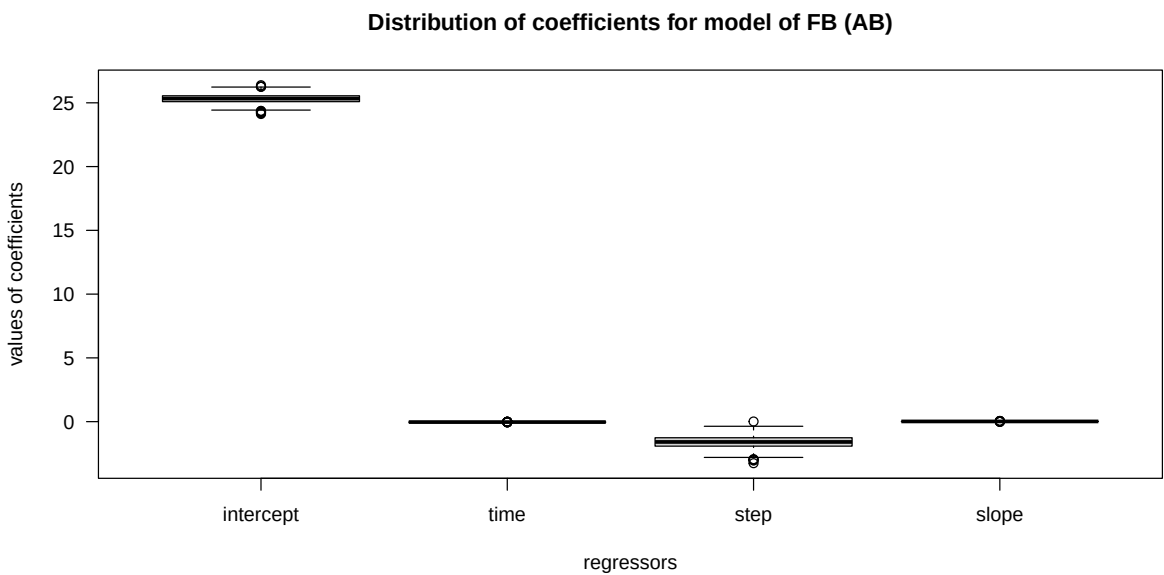


Figure A.24: Distribution of the coefficients of the model for FB with single interruption

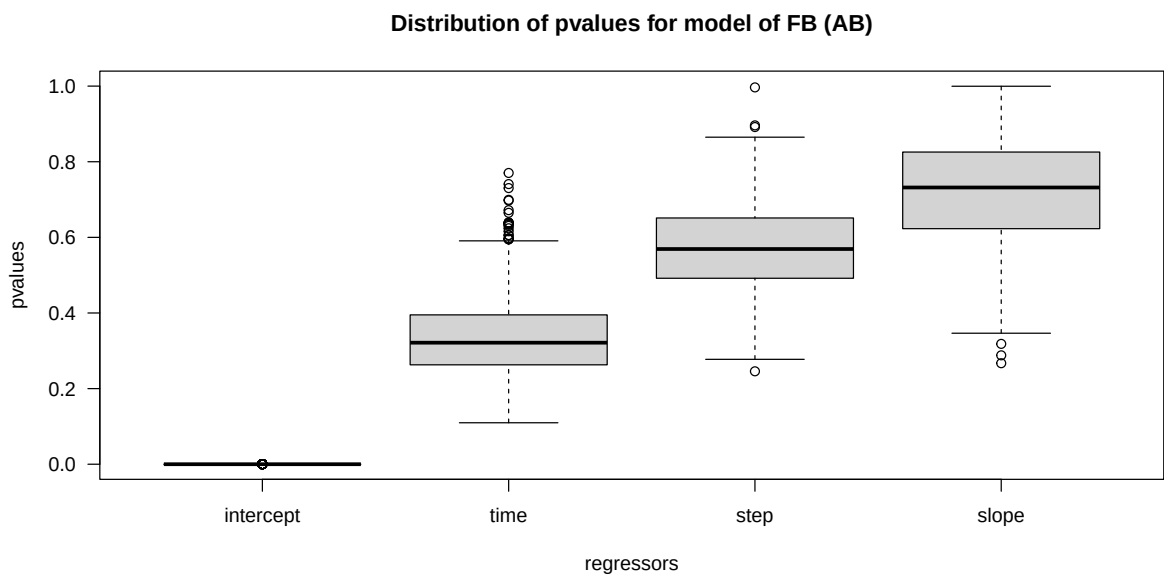


Figure A.25: Distribution of the p-values of the model for FB with single interruption

List of Figures

3.1	Wearable device for personal air pollution exposure estimation	17
3.2	Route with static acquisitions points highlighted	18
4.1	Examples of impact model used in ITS	26
4.2	Magnitude outliers	37
4.3	Shape outliers	39
5.1	Evolutions of variable HR per station	42
5.2	Evolutions of variable FB per station	43
5.3	Evolutions of variable HR	43
5.4	Evolutions of variable FB	44
5.5	Evolutions of environmental variables for subject 1	45
5.6	Evolutions of environmental variables	48
5.7	Comparisons between stations for subject 1	49
5.8	Boxplots for subject 1	51
5.9	Magnitude outliers of HR for subject 3	52
5.10	Shape outliers of HR for subject 3	52
5.11	Model for HR, with single interruption between stations B and C	59
5.12	Model for HR with all the interruptions	62
5.13	Model for FB with all the interruptions	63
5.14	Model for HR with the significant interruptions	65
5.15	Model for FB with the significant interruptions	66
5.16	Distribution of the coefficients of the model for HR with all the interruptions	67

5.17	Distribution of the p-values of the model for HR with all the interruptions	68
A.1	Evolutions of environmental variables for subject 2	81
A.2	Evolutions of environmental variables for subject 6	82
A.3	Evolutions of environmental variables for subject 17	82
A.4	Comparisons between stations for subject 6	83
A.5	Comparisons between stations for subject 19	84
A.6	Boxplots for subject 14	85
A.7	Boxplots for subject 17	86
A.8	Magnitude and shape outliers of HR for subject 4	87
A.9	Magnitude and shape outliers of FB for subject 14	88
A.10	Magnitude and shape outliers of CO_2 for subject 9	88
A.11	Magnitude and shape outliers of $PM_{2.5}$ for subject 16	89
A.12	Magnitude and shape outliers of temperature for subject 7	89
A.13	Magnitude and shape outliers of humidity for subject 9	90
A.14	Magnitude and shape outliers of VOC for subject 8	90
A.15	Model for HR, with single interruption between stations B and E	91
A.16	Model for HR, with single interruption between stations B and F	91
A.17	Model for FB, with single interruption between stations A and E	92
A.18	Model for FB, with single interruption between stations A and F	92
A.19	Model for FB, with single interruption between stations B and F	93
A.20	Distribution of the coefficients of the model for FB with all the interruptions	93
A.21	Distribution of the p-values of the model for FB with all the interruptions	94
A.22	Distribution of the coefficients of the model for HR with single interruption	94
A.23	Distribution of the p-values of the model for HR with single interruption .	95
A.24	Distribution of the coefficients of the model for FB with single interruption	95
A.25	Distribution of the p-values of the model for FB with single interruption .	96

List of Tables

5.1	Magnitude and shape outliers for HR, FB, CO_2 and $PM_{2.5}$	53
5.2	Magnitude and shape outliers for temperature, humidity and VOC	54
5.3	Models for HR and FB, with a single interruption: p-values of regression coefficients	60
5.4	Models for HR and FB, with all the interruptions: p-values of regression coefficients	63

List of Acronyms

Acronym	Description
OECD	Organization for Economic Co-operation and Development
WHO	World Health Organization
TS	Time Series
ITS	Interrupted Time Series
MITS	Multivariate Interrupted Time Series
PM	Particulate matter
VOCs	Volatile Organic Compounds
ADHD	Attention Deficit Hyperactivity Disorder
CITS	Controlled Interrupted Time Series
NHM	Non-Hierarchical Model
VPIM	Varying Pre-Intervention Intercept Model
VIM	Varying Intercept Model
MEM	Measurement Error Model
WAIC	Widely Acceptable Information Criteria
wSR	weighted Segmented Regression

Acronym	Description
ARPA	Agenzia Regionale per la Protezione dell'Ambiente
ANTHEM	AdvaNced Technologies for Human-centrEd Medicine
PNRR	Piano Nazionale di Ripresa e Resilienza
tVOC	total Volatile Organic Compounds
BSN	Body Sensor Network
IMU	Inertial Measurement Unit
IQR	InterQuantile Range
HR	Heart Rate
FB	Frequency of Breath
FDA	Functional Data Analysis
FBP	Functional BoxPlot
MBD	Modified Band Depth
MEI	Modified Epigraph Index

Acknowledgments

I would like to express my gratitude to those who have guided me throughout this work over the past months. I am especially grateful to my supervisor, Francesca Ieva, for proposing the topic of this thesis, for her invaluable expertise and for her support in addressing the challenges related to this research. I extend my sincere thanks to Lara Cavinato for her continuous assistance and willingness to share her knowledge, as well as to Alessandra Ragni for her kindness and availability.

This research is part of the AdvaNced Technologies for Human-centrEd Medicine (AN-THEM) project grant.

