UNDERSTANDING REAL-WORLD PHENOMENA FROM HUMAN-GENERATED SENSOR DATA

by

FANI TSAPELI
SUPERVISOR: MIRCO MUSOLESI, PETER TINO
THESIS GROUP: ATA KABAN, EIKE RITTER

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
College of Engineering and Physical Sciences
The University of Birmingham
February 2018

CONTENTS

Abstract

Nowadays, there is an increasing data availability. Smartphones' and wearable devices' sensors, social media, web browsing information and sales recordings are only few of the newly available information sources. Analysing this kind of information is an important step towards understanding or even predicting human behaviour. The effective utilisation of the rich information that is hidden in such unstructured and noisy datasets remains an open issue. In this dissertation, I propose novel techniques for uncovering the complex dependencies between factors extracted from raw sensor data and real-world phenomena and I demonstrate the potential of utilising the vast amount of human digital traces in order to better understand human behaviour and factors influenced by it. In particular, two main problems are considered: 1) whether there is a dependency between social media data and traded assets prices and 2) how smartphone sensor data can be used to monitor and understand factors that influence our stress level. In the former case, I firstly examine the causal impact of Twitter sentiment on the stock prices of four tech companies. Then, I focus on the detection of Twitter events that are associated with large fluctuations on specific stock markets. In the second case, I attempt to find causal links between daily activities, such as socialising, working and exercising, with stress. In this thesis, I focus on uncovering the structural dependencies among factors of interest rather than on the detection of mere correlation. Special attention is given on enhancing the reliability of the findings by developing techniques that can better handle the specific characteristics and limitations of the examined datasets. In detail, I propose a novel framework for causal inference on observational time-series data that does not require any assumption about the functional form of the relationships among the variables of the study and can effectively control a large number of factors. In addition, I have developed a causal inference method that can handle data with noisy entries and result in more accurate conclusions than existing approaches. Although the approaches developed during this thesis are motivated by specific problems related to human-generated sensor data, they are general and can be applied in any dataset with similar characteristics.

CHAPTER 1

INTRODUCTION

Nowadays, people generate vast amounts of data through the devices they interact with during their daily activities, leaving a rich variety of digital traces. Indeed, our mobile phones have been transformed into powerful devices with increased computational and sensing power, capable of capturing any communication activity, including both mediated and face-to-face interactions. User location can be easily monitored and activities (e.g., running, walking, standing, traveling on public transit, etc.) can be inferred from raw accelerometer data captured by our smartphones [1, 2]. Even more complex information, such as our emotional state or our stress level, can be inferred either by processing voice signals captured by means of smartphone's microphones [3, 4] or by combining information, extracted from several sensors, which correlates with our mood [5, 6, 7, 8, 9]. Moreover, we keep track of our daily schedule by using digital calendars and we use social media such as Facebook, Twitter and blogs to communicate with our friends, to share our experiences and to express our opinion and emotions. Wearable devices that are able to monitor physical indicators with a very high level of accuracy are also increasingly popular [10].

Leveraging this rich variety of human-generated information could provide new insights on a variety of open research questions and issues in several scientific domains such as sociology, psychology, behavioural finance and medicine. For example, several works have demonstrated that online social media could act as crowd sensing platforms [11]; the aggregated opinions posted in online social media have been used to predict movies revenues [12], elections results [13] or even stock market prices [14]. Social influence effects in social networks have also been investigated in several projects either using observational

data [15, 16] or by conducting randomised trials [17, 18]. Other works also use mobility traces in order to study social patterns [19] or to model the spreading of contagious diseases [20]. Moreover, smartphones are increasingly used to monitor and better understand the causes of health problems such as addictions, obesity, stress and depression [21, 22, 9]. Smartphones enable continuous and unobtrusive monitoring of human behaviour and, therefore, allow scientists to conduct large-scale studies using real-life data rather than lab constrained experiments. In this direction, in [23] the authors attempt to explain sleeping disorders reported by individuals, by investigating the correlations between sociability, mood and sleeping quality, based on data captured by mobile phones sensors and surveys. Also, in [24] the authors study the links between unhealthy habits, such as poor-quality eating and lack of exercise, and the eating and exercise habits of the user's social network. However, both studies are based on *correlation* analysis and, consequently, they are not sufficient for deriving valid conclusions about the *causal links* between the examined variables. For example, an observed correlation between the eating and exercising habits of a social group does not necessarily imply that eating and exercise habits of individuals are influenced by their social group. Instead, the observed correlation could be due to the fact that people tend to have social relationships with people with similar habits.

Some recent studies have examined the ability of social media to influence real-world events by applying randomised control trials. For example, authors in [17] examine the effect of political mobilisation messages by using Facebook to deliver them to a randomly selected population; the effect of the messages is measured by comparing the real-world voting activity of this group with the voting activity of a control group. Similarly, in [18] authors use randomised trials in order to examine the social influence of aggregated opinions posted in a social news website. Indeed, randomised control trials are a reliable technique for conducting causal inference studies [25]. However, their applicability is limited since they require scientists to gather data using experimental procedures and do not allow the exploitation of the large amount of observational data. In many cases, it is not feasible to apply experimental designs or it is considered unethical [26]. In addition, experimental procedures might require the recruitment of a potentially large number of participants which incurs additional costs.

Several methods for causality detection in observational data have been proposed

[27, 28, 26]. Causal detection in observational data is based on the assumption that all the factors that influence the relationship between two examined variables are observed. This is a strong assumption which may not hold in some cases. Consequently any causal conclusions could be biased and should be interpreted with caution. Nevertheless, causality studies with observational data could provide useful insights about the dependencies among the examined factors.

When human digital traces are used to analyse complex phenomena such as the impact of social, psychological and emotional factors on stock market prices, a large number of variables need to be included in the study. This results in high-dimensional datasets, which are often characterised by complex dependencies among the included variables. In such cases, conducting explanatory data analysis in order to understand the underlying data structure and the potential causal links is even more challenging. Moreover, human digital traces usually include low level information that requires significant amount of processing in order to extract features deemed important for a study. For example, in [29] a sentiment index is inferred from Twitter data by applying text processing and classification techniques and in [30] factors such as sleeping patterns, social interactions and physical activity are inferred from raw sensor data. When key factors are inferred from other observed characteristics, rather than directly measured, the amount of noise resulted by inaccurate estimations may jeopardise the validity of the study.

1.1 Contribution

Considering the previously discussed opportunities arising from the vast amount of the available human digital traces as well as the open issues on their effective utilisation, the main thesis of this dissertation is that *uncovering the dependencies between factors extracted from human digital traces and real-world events would allow us to better understand human behaviour and events influenced by it.* In order to support this statement I will focus on the following research questions:

1. how can we extract meaningful information from raw sensor data and how can we link this information to real-world phenomena?

2. how can we discover causal links in complex high dimensional observational data?

3. how can we discover causality when there are noisy measurements?

More specifically, I present the design and evaluation of novel techniques that enable researchers to unlock the potential of human generated sensor data by allowing them to analyse not only correlation but also causality relationships. As it was previously mentioned, current studies are mainly based on correlation analysis. However, in many cases the observed correlations may occur incidentally [31] and may not represent the true structural relationships among the examined variables. In this work, I attempt to discover causal links instead of mere correlation. Motivated by the limitations of the existing methodologies for causal inference, (these limitations are further discussed in Chapter 2), I developed a causal inference method for time-series that does not require any assumptions about the statistical relationships among the variables of the study and can effectively handle high-dimensional datasets. Then, I use this framework in order to detect causal links in human digital traces. It should be noted that, any causality study based on observational data is based on the strong assumption that all the factors that influence the examined variables have been included in the study. Since it is usually hard to include all the necessary factors, the term 'causal' should be only interpreted relative to the observed variables of the study.

This dissertation is based on two main data sources: smartphone sensor data and social media data. In particular, I use social media data in order to understand the impact of behavioural factors on stock market prices. Features extracted from social media data are used as indicators of the aggregated people sentiment and opinion about topics of interest. Then, I examine the relationship of the extracted features with financial indicators in two case studies. In the first study, I attempt to measure the causal impact of social media sentiment on traded assets prices of four tech companies. In the second case, I examine whether bursty topics in social media related to politics or finance can be associated with stock market jitters. Finally, I use data captured by smartphone sensors in order to study the impact of daily activities such as socialising, exercising and working on people stress levels.

Finally, as it was previously mentioned, features extracted by human generated sensor

data can be noisy and inaccurate. Motivated by this, I propose a novel approach for causal inference when one or more key variables are observed with some noise. The proposed method utilises the knowledge about the uncertainty of the real values of key variables in order to reduce the bias induced by noisy measurements. Although the methodologies developed for this dissertation aim to tackle specific issues related to causal inference from human digital traces, they are general and they could be applied to other dataset types with similar characteristics.

1.2 Thesis Outline

This dissertation is organised as follows:

- Since significant part of this dissertation focuses on causal inference, in Chapter 2 I provide some background knowledge on the main existing methodologies on causality detection. Special focus is given on Rubin's counterfactuals framework [32] and on quasi-experimental designs for causal inference, since the methodologies proposed at this dissertation is based on these methods. I also present two other widely used approaches on causal inference namely structural equation models (SEMs) [33] and directed acyclic graphs [34]. Finally, I describe current approaches on causal discovery in time-series data, emphasising on Granger-causality based methods and methods based on transfer entropy and I discuss the advantages and limitations of these methods.

- In Chapter 3, I propose a novel method for causal discovery in time-series data based on Rubin's counterfactuals framework [32]. This method is motivated by the need for an approach that can effectively handle high-dimensional data without imposing any restrictions about the form of the dependencies (i.e. linear or non-linear) among the variables of the study. I evaluate this method in comparison with methods based on Granger-causality and transfer-entropy on synthetic data and I demonstrate that the proposed framework is more effective on avoiding false positive causal conclusions.

- In Chapter 4, I investigate the influence of social media on stock market. I first present other works which have demonstrated that features extracted from social media correlate with stock market prices and can be used to improve future prices prediction. Then, I attempt to go one step beyond correlation and study the causal impact of social media sentiment on the stock market prices of four tech companies. I include in my study a large number of factors that could influence both social media sentiment and traded assets such as currency exchange rates, commodities prices and performance of other major companies and I use the method presented in Chapter 3 in order to study the causal link between the examined factors. In this chapter, I also attempt to link bursty topics in social media related to finance or politics with strong stock markets fluctuations. In more detail, I propose a novel event detection method on Twitter, tailored to detect financial and political events that influence a specific stock market and I apply this method on high-frequency intra-day data from the Greek and Spanish stock market. I demonstrate that features extracted from social media data could be useful indicators for the early detection of stock market jitters.

- In Chapter 5, I present the problem of knowledge discovery from raw smartphones' sensor data. Initially, I discuss how smartphone sensor data can be used to monitor and understand human behaviour and how the extracted information can be used to build applications that improve users well-being. I support this statement by presenting existing studies on this domain. Afterwards, I conduct one case study using smartphones sensor data. In more detail, I use StudentLife dataset [35], a dataset containing smartphone data for 48 students from Dartmouth College, in order to study the impact of daily activities such as socialising, exercising and working on the stress level of participants. Information about participants daily social interactions as well as their exercise and work/study schedule is not directly measured; instead, I use raw GPS and accelerometer traces in order to infer high-level information which is considered as implicit indicator of the variables of interest. Also, pop-up questionnaires are used to track participants stress level. The causal inference method presented in Chapter 3 is used to link users activities with their stress level.

- In Chapter 6, I examine the problem of causal inference when one or more key variables are observed with some noise. This is a common problem, especially in studies based on raw sensor data, where important information is often not directly measured and it needs to be inferred from other low level characteristics. I propose a novel causal inference approach, based on Rubin's counterfactuals framework [32], that takes into account the uncertainty about the real values of a noisy variable. Noisy variables are handled as stochastic approaches and the method attempts to maximise the probability that any bias in the study has been sufficiently reduced. I evaluate this method in comparison with existing methods both on simulated and real scenarios and I demonstrate that this approach reduces the bias and avoids false causal inference conclusions in most cases.

- In Chapter 7, I summarise the contributions of this thesis and I discuss future research directions.

1.3 List of Publications

During my PhD I have authored the following papers:

Chapter 3 and Chapter 4

- Fani Tsapeli, Mirco Musolesi and Peter Tino. "Non-parametric causality detection: An application to social media and financial data." Physica A: Statistical Mechanics and its Applications 483 (2017): 139-155.

- Fani Tsapeli, Nikolaos Bezirgiannidis, Mirco Musolesi and Peter Tino. "Linking Twitter Events With Stock Market Jitters." Under review at EPJ Data Science.

Chapter 5

- Fani Tsapeli and Mirco Musolesi. "Investigating causality in human behavior from smartphone sensor data: a quasi-experimental approach." EPJ Data Science 4.1 (2015): 24.

Chapter 6

- Fani Tsapeli, Peter Tino and Mirco Musolesi. "Probabilistic Matching: Causal Inference under Measurement Errors." In Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2017.

Papers not included in this thesis

- Abhinav, Mehrotra, Fani Tsapeli, Robert Hendley and Mirco Musolesi. "MyTraces: Investigating Correlation and Causation between Users Emotional States and Mobile Phone Interaction." Proceedings of the ACM Conference on Interactive, Mobile, Wearable and Ubiquitous Technologies (UbiComp) (2017).

CHAPTER 2

BACKGROUND

In this chapter, I will introduce the two main problems that are considered in this thesis: 1) the influence of social media on financial markets and 2) the analysis of human behaviour using smartphones sensor data. I will discuss how my thesis relates to the previous works on these subjects and I will justify the need of novel methods that would enable us to better utilise these data. Since the main methods developed in this dissertation focus on causal inference, at the third part of this chapter I present an overview of the existing causal inference methodologies as well as my contribution on this domain.

2.1 Influence of Social Media on Financial Markets

According to the efficient market hypothesis (EMH), stock prices are instantaneously reflecting any external information and therefore, fundamental analysis, which utilises information about exogenous factors such as news releases, cannot outperform technical analysis [36]. However, several studies have reported evidence that exogenous factors such as significant political and financial news or macroeconomic releases could cause large fluctuations on stock market prices [37, 38].

Nowadays, online systems such as social media, web blogs, search engines etc., are able to capture the reaction of people on news in real-time and their interest on specific topics. Several works have previously examined the potential of information extracted from social media, search engine query data or other web-related information to predict stock market returns. For example, the correlation between sentiment extracted from Twitter data and

stock market prices is examined in [39, 40]. Similarly, in [41] the authors show that metrics extracted by social media are leading indicators of firms equity value. In [42] the authors examine the mutual information between social media sentiment and hourly traded assets prices. They found a statistically significant association only with a limited set of stocks. The different behaviour that is observed for different companies is attributed mainly to the fact that some company names or ticker symbols attract more message volumes while for others the available information is not sufficient for the analysis. Also in [43] the dependencies between microblogging sentiment and several financial indicators are examined using fuzzy-set qualitative comparative analysis. Other features such us the number of followers are also considered.

Other projects have been focussed on the possible use of sentiment analysis based on social media data for the prediction of traded assets prices by applying a bivariate Granger causality analysis [29, 44, 45] or regression models [46, 47]. In [48] authors propose a prediction method based on machine learning. Also in [49] a support vector machine classifier is trained in order to predict the direction of stock market movement using past prices as predictors and features extracted from text posted on Yahoo Finance message boards. Moreover, in [50] the authors train a classifier to predict daily up and down movements of tech companies traded assets using as features the sentiment of relevant tweets and the degree of stock market confidence. Researchers have also exploited additional information from Twitter to predict stock market movement: examples include information based on influential users [51] or interaction between users [52, 53].

In [54, 55] the authors have also demonstrated that search engine query data correlate with stock market movements. In [56] the authors propose a trading strategy that utilises information about Wikipedia views. They demonstrate that their trading strategy outperforms random strategy.

All the above mentioned studies, are based on bivariate models. Although their results indicate that social media and other web sources may carry useful information for stock market prediction, by using these techniques it is not possible to figure out whether stock markets are actually influenced by this information or whether other factors are influencing both stock market prices and users opinion captured by digital data. In order to examine this, a causality analysis is required.

Trading strategies that utilise both technical analysis and sentiment analysis are discussed in [57, 58]. These works focus on prediction rather than causal inference and they are predominantly base on regression. Applying regression models for causal inference suffers from two main limitations. First, stock market prices can be influenced by a large number of factors such as stock market prices of other companies [59, 60], foreign currency exchange rates and commodity prices. Such factors may also influence people sentiments. Consequently, to eliminate any bias it is required to include a large number of predictors in the regression model. The estimation of regression coefficients in a model with a large number of predictors can be challenging. These issues are discussed in more detail in Section 2.3. Second, most studies on social media and stock market data are based on linear regression models. However, other studies provide evidence of non-linear dependencies [61]. Selecting an appropriate model is usually one of the most difficult aspects of this type of analysis. Inaccuracies in model specification, estimation or selection may result in invalid causal conclusions.

In order to fill this gap, in Chapter 3, I propose a causal inference method for time-series data that overcomes the limitations of existing approaches and can handle more effectively the specific datasets. Then, I apply this method in order to quantify the influence of social media sentiment on traded assets prices [62].

The ability of Twitter sentiment to predict stock market movements has been questioned recently [63]. In particular, authors found no evidence of Granger-causality between Twitter sentiment and stock market prices when they tried to replicate the results of Bollen et al. work [14] for a different time period. They suggest that the results of Bollen's paper are influenced by several biases on the dataset such as small time-period that is considered in the analysis and the limitations of the applied sentiment analysis. In my work, I examine the links between social media and stock market in significantly larger time periods (4 years instead of 11 months). Also, I evaluate the performance of the applied sentiment classification method. Then, the uncertainty about the real sentiment of tweets due to the limitations of the applied sentiment analysis method is included in the study. Finally, I enhance the reliability of the results by conducting a sensitivity analysis.

Moreover, it has been shown in [64] that social media may contain useful information

mainly during the peaks of Twitter volume. In more detail, authors found only a weak correlation between Twitter sentiment and stock market prices of companies from the DJIA index when data from a large time-period are examined. However, they found strong dependency when they examine only periods during which the Twitter volume is high, suggesting that an abnormal Twitter stream may indicate a strong stock market movement. Examining the links between social media and stock market during intervals characterised by abnormal behaviour would enable us to better understand and predict stock market jitters. However, this topic is currently unexplored.

In this dissertation I propose a novel method that detects bursty Twitter topics that are linked with stock market jitters and I demonstrate that Twitter can be used in order to spot early abnormal stock market movements. In addition, although several works have provided evidence of correlation between web-related information sources and stock market prices, little work has been done so far on understanding which features contained in social media could be useful for the understanding of stock market movements. In [42] the strength of the correlation between sentiment extracted from social media and stock market prices of S&P500 companies is quantified. It is also shown that sentiment contains more useful information compared to message volumes. Also, in [65] authors provide useful insights about the correlation of specific text-related features (e.g. the content of the news or the existence of fear/hope sentiment) and author-related features (e.g. author's reputation) with stock market prices. Instead of examining the influence of a specific feature on stock market prices, the event detection method proposed at this thesis combines information on the volume, sentiment and content of the tweets, with information on their authors and geography, and construct feature vectors for each group of tweets associated with an event. I apply a feature selection process in order to understand which of the extracted features contain useful information.

Detecting emerging topics with wide interest on Twitter has been examined by several previous studies. The main approach that is usually followed is the detection of *bursty* terms (i.e., words or segments whose frequency on the Twitter stream is characterised by some unusual pattern during a well-defined time period) followed by a grouping of these terms based on their content similarity or similarities on their arrival patterns. For example, in [66] the authors extract a set of *emerging* terms from the Twitter stream by

assigning weights to each term based on its frequency as well as the *importance* of Twitter users who use it and keeping only the highest weighted terms. The *emerging* terms are grouped together by examining their co-occurrence on the same tweets. EDCoW [67] also constitutes a representative example of this category of event detectors. EDCoW performs a wavelet transformation on the frequencies of words and it uses *vanishing* auto-correlations to eliminate words that do not experience any irregularities on their arrival rates. Then, it creates a graph by using the pairwise correlations among the words wavelets and applies a modularity-based graph partitioning in order to group the words to events. On the same direction, TopicSketch [68] monitors the acceleration of words and pairs of words in order to early detect bursty topics. Twevent [69] proposes the use of word segments instead of single words and detects bursty words by examining the word segments frequency and the number of different users that report these segments. Then, word segments are grouped by examining the content similarity among them.

Other projects focus on detecting events of a specific type. For example, authors in [70] detect local social events by monitoring microblogging activity in geographical regions and reporting any unusual activity. Moreover, in [71] a system for detecting real world events in real-time along with the geographical location of the event is presented. The system uses keywords to detect specific event types. In particular, in [71] the authors consider the problem of earthquakes detection as a case study.

My work substantially differs from the existing event detectors, since my objective is the detection of events that influence a specific stock market rather than the detection of events in general or events of a specific type (e.g., financial events). Thus, my system does not consider real world financial or political incidents that do not impact the examined stock market as *true* events.

Finally, social media studies are not limited in the finance domain. For example, several studies investigate whether election results or box office revenues can be predicted by analysing Twitter sentiment [72, 73, 74]. Others examine if Twitter can influence election results [75]. In [76] authors study the information propagation from other social media to Twitter. Although such studies demonstrate that social media could be useful for understanding and predicting several real-world phenomena, their findings need to be interpreted with caution. Such studies are largely influenced by the efficacy of the applied

sentiment detection method and sometimes their results show a only a weak dependency between the examined factors. Recently the validity of the results of some of these studies has been questioned [77, 78].

2.2 Human Behaviour Analysis based on Smartphone Sensor Data

Nowadays, smartphones and wearable devices have become an indispensable part of our lives. Since we are moving towards the era of the Internet of Things, the amount of human generated sensor data is expected to increase even more. This opens new opportunities to scientists studying human behaviour. So far, studies on human behaviour have relied mainly on paper records. For example, the causal impact of social influence on human behaviour [79, 80, 81, 82, 83], the impact of exercise on mood [84, 85] and factors that influence employees or customers satisfaction [86, 87, 88, 89] are only some of the subjects that have largely concerned sociology researcher. However, results based on questionnaires suffer from several limitations:

1. They may be inaccurate. Participants may intentionally or unintentionally provide inaccurate responses.

2. They cannot capture participants behaviour, emotions or opinions at real-time. Questionnaires usually ask participants to report their behaviour, emotions or opinion about events that happened earlier within the day, week or even month.

3. Filling a questionnaire requires some effort and time from participants and consequently some people may be unwilling to participate to such studies.

On the other hand, smartphones and wearable devices offer continuous, real-time and unobtrusive monitoring of human behaviour. Raw sensor data captured by smartphones or wearable devices can be used to infer users location context (e.g. home, work, restaurant, bar etc.) [90, 91], their emotions [92, 93, 94], their activity level [95, 96, 97, 98] and the level of social interactions that they had [30]. The efficient utilisation of this rich variety of data could revolutionise the current approaches on human behaviour study.

Several studies have utilised smartphone data in order to examine human behaviour and emotions. In [99] authors examine the link between physical activity and happiness. Physical activity is measured using smartphones' accelerometer data and happiness levels are self-reported by participants. Assessing the impact of travels and travel-related activities on participants happiness has also been examined at [100, 101]. In [102] authors present a large-scale study on the correlation of personality, mood and well-being with sociability, activity and mobility. They also demonstrate that mobile sensing data can be used for the prediction of users mood. Similarly, in [103] smartphone sensor data and communication logs are used to assess daily mood patterns. Several works have provided evidence of links between phone usage patterns and mood [104, 105, 106] or other behavioural characteristics such as alertness [107]. In addition, the StudentLife dataset has been used to study the correlation between academic performance and sociability, studying patterns, activity and mobility [108, 109].

The potential of utilising mobile phones in order to monitor people with mental health problems such as depression or bipolar disorder has been examined in several recent studies [110, 111, 112, 113]. Health officers and doctors in the area of mental health care can continuously monitor the behaviour of patients and detect anomalies that may indicate the need of intervention. For example, some studies provide evidence of correlation between activity [112] or mobility [113] and depression levels. Thus, this low-level information could provide useful insights about patients mental health state.

To the best of my knowledge, all the studies on this domain so far are based on mere correlation. However, a correlation analysis is not sufficient in many cases. For example, a correlation between mobility and happiness does not necessarily imply that mobility increases happiness levels. People may visit more places during their leisure time and they may also be happier when they do not have a busy work schedule. Thus, the observed link between mobility and happiness could be due to a less demanding work schedule. In this case, scheduling interventions prompting participants to go for a small trip when they report low happiness level may not be helpful during busy days.

Motivated by the above mentioned limitation of current studies, I propose a method for causal inference on smartphone sensor data. As it will be later discussed in Section 2.3, causal inference in observational data could be biased in case of missing confounding

variables. Since, in many cases, key factors required for the analysis may not be captured by smartphones, any findings should be interpreted with caution. I use this method in order to understand the impact of daily activities, such as exercising and socialising, on the stress level of 48 students.

2.3 Causal Inference

Causal analysis attempts to understand whether differences on a specific characteristic Y within a population of *units* are influenced by a factor X. Y is called *response*, *effect* or *outcome* variable and X *treatment* variable or *cause*. *Units* are the basic objects of the study and they may correspond to humans, animals or any kind of experimental objects. $Y(u)$ and $X(u)$ denote the outcome and treatment values measured for unit u respectively.

In order to claim that a value of a variable Y has been caused by a value of a variable X [26]:

1. The value of variable Y should have occurred after the value of X.
2. There should be an association between the occurrence of these two values.
3. There should be no other plausible explanation of this association.

The key idea of causation theory is that, given a unit u, the value of the corresponding response variable $Y(u)$ can be manipulated by changing the value of the treatment variable $X(u)$ [114, 115]. Initially, I will consider X as a binary treatment variable, although later in this section, the case of non-binary treatments is discussed. According to Rubin's framework [116], the causal impact of a binary treatment on a unit u can be assessed by comparing the outcome $Y_1(u)$, if the unit has received the treatment, with the outcome $Y_0(u)$, if the unit has not received the treatment. The *fundamental problem of causal inference* is that it is not feasible to observe both $Y_1(u)$ and $Y_0(u)$ for the same unit u. Instead, the average treatment effect (ATE) can be estimated as:

$$E\{Y_1\} - E\{Y_0\} \tag{2.1}$$

where $E\{Y_1\}$ and $E\{Y_0\}$ are expectations w.r.t. uniform distribution over treated and untreated units, respectively. The average treatment effect can be estimated only if the following three assumptions are satisfied:

1. The effect variable Y is i.i.d..

2. The observed outcome in one unit is independent from the treatment received by any other unit (*Stable Unit Treatment Value Assumption - SUTVA*).

3. The assignment of units to treatments is independent from the outcome (*ignorability*). For example, let us consider a causal study about the effects of an educational program on students performance. The ignorability assumption is satisfied if the selection of the students that will be involved in the study is independent of the outcome i.e. both strong and weak students are equally likely to be selected for the program. Ignorability can be formally expressed as $Y_1 \perp\!\!\!\perp X$, $Y_0 \perp\!\!\!\perp X$. The assumption of ignorability requires that all the units have equal probability to be assigned to a treatment. If this assumption does not hold, the units that received a treatment may systematically differ from units that did not receive such a treatment. In such a case the average value of the outcome variable of the treated units could be different from that of other units, even if the treatment had not been received at all.

In experimental studies, units are randomly assigned to treatments. Thus, both the SUTVA and the ignorability assumptions are satisfied. However, in many cases it is not feasible to conduct experimental causality studies. Several techniques for causal inference in observational data have been proposed. In this section, I discuss the main methods on this domain. I emphasise on the matching design framework, since the methods developed in this dissertation are based on this approach. The purpose of this section is to provide some background knowledge on the main causal inference methodologies rather than presenting an exhaustive literature review on causality.

Symbol	Description
N	Number of units
P	Number of confounding variables
Y	Outcome variable, described with a $1 \times N$ vector
y_u	The outcome value of unit u
X	Treatment variable, described with a $1 \times N$ vector
x_u	The treatment value of unit u
H	$P \times N$ matrix of confounding variables
h_u	the u^{th} column of H, denoting a $P \times 1$ vector of values of unit u for the P confounders
h^p	the p^{th} row of H, denoting a $1 \times N$ vector of values of the N units for the p^{th} confounder
h_u^p	element in column u and row p of H, denoting the value of unit u for the p confounder
G	Set of matched treated and control units
G_U	Set of matched treated units
G_V	Set of matched control units
$\mathcal{D}(h_u, h_v)$	Distance between vectors h_u, h_v
$\Delta(u, v)$	Distance between units u, v

Table 2.1: Notation.

2.3.1 Matching Design

I describe the treatment and outcome variables X and Y as $1 \times N$ vectors, with N the number of units and x_u, y_u the treatment and outcome values of the unit u respectively (i.e. the u^{th} elements of vectors X and Y). I also define a $P \times N$ matrix of P confounding variables denoted as H. Confounding variables represent baseline characteristics of the units that are considered relevant for the study. For example, in a medical study that examines the impact of a drug, baseline characteristics could be the previous health condition of the units (in this case patients), their age etc.. I denote as h_u the u^{th} column of H, representing a $P \times 1$ vector of values of unit u for the P confounding variables and as h^p the p^{th} row of H, representing a $1 \times N$ vector of values of the N units for the p^{th} confounding variable. For more clarity, I summarise the notation that is used in this chapter in Table 2.1.

As was previously mentioned, in experimental studies, ignorability can be achieved by randomly assigning units to treatments. However, in observational studies this is not feasible. Instead, the average treatment effect can be estimated by relaxing ignorability

to *conditional ignorability* [26]. According to the conditional ignorability assumption, the treatment assignment is independent from the outcome, conditional on a set of confounding variables, represented by matrix H. Thus, conditional ignorability is expressed as $Y_1 \perp\!\!\!\perp X|H$ and $Y_0 \perp\!\!\!\perp X|H$.

Matching methods attempt to achieve conditional ignorability by comparing the outcome values of units with similar observed characteristics. In particular, if U is a set of treated units and V is a set of control units (i.e., units which have not received the treatment), matching methods match each treated unit $u \in U$ with the *"most similar"* control unit $v \in V$. If G is the set of matched pairs of units, the average treatment effect is estimated as

$$E_{(u,v)\in G}\{Y_1(u) - Y_0(v)\} \tag{2.2}$$

where the expectation is with respect to uniform probability distribution over G. The (dis)similarity between units is measured as a distance between their confounding variable values (for some metric).

Several methods for creating pairs of units $(u, v) \in G$ have been proposed. The matching methods involve four steps [117]:

1. **(Dis)similarity Estimation**. In this step, a notion of (dis)similarity between units is defined. The dissimilarity corresponds to a distance metric between the confounding variable values of two units.

2. **Matching Method.** In the matching step, a method that creates pairs of treated and control units $(u, v) \in G$ based on closeness of their confounding variables, as it is defined in step 1, is applied. Units with *similar* values on their confounding variables are matched.

3. **Balance Check.** In the balance check step, the remaining confounding bias due to imperfectly matched units needs to be estimated. The balance can be examined by checking the standardised mean difference between the treated and control units, by applying a t-test or a Kolmogorov-Smirnov test, or by examining the quantiles of the matched units [117]. This checking has to be done for each confounding variable.

If the resulted groups of matched treated and control units are not adequately balanced, the method needs to be revised (i.e., the steps 1 and 2 are modified until sufficient balance between the treated and control units has been achieved).

4. **Treatment Effect Estimation.** When sufficient balance has been achieved, the average treatment effect can be estimated using Eq. (2.2).

2.3.1.1 Distance Metrics

The simplest distance metric is the *exact distance*, according to which the distance between two vectors h_u, h_v is zero only if all their elements are equal. Otherwise, their distance is infinite. Matching with exact distance results in zero confounding bias. However, in most cases, exact matching cannot be applied, especially when the study includes a large number of confounding variables and/or some of them are continuous.

Euclidean or Mahalanobis distances are commonly used for measuring the dissimilarity between two vectors h_u, h_v. Another distance metric that is used to measure the (dis)similarity between the confounding variables values of two units is the absolute difference on the *propensity score* [28]. The propensity score is the probability of a unit to be assigned to a treatment conditional to its confounding variables values. The propensity score is usually estimated using logistic regression, in which the binary treatment is regressed on the confounding variables. The conditional ignorability assumption is satisfied when the matched units have 'approximately' the same probability to be assigned to a treatment.

2.3.1.2 Matching Methods

Several methods have been used for matching. The most straightforward method is *Nearest Neighbor Matching* [117], which matches each treated unit to the control unit with the lowest distance on the corresponding confounding variable values. In each simplest form, each unit can be matched only one time. Some variations of this approach are discussed later in this section.

One-to-one nearest neighbor matching may result in bad pairs when multiple treatment units are 'competing' for a small number of control units. Moreover, the order in which

units are matched may influence the quality of the matched pairs [28]. Optimal matching [118, 119, 120] has been proposed in order to avoid these issues. Optimal matching attempts to optimise a global distance measure among all matched pairs.

Subclassification can also be used for creating pairs of treated and control units [121]. Based on this method, units are split in groups or *subclasses* so that the distribution of the confounding variables values is similar for both treated and control units belonging to the same group. Propensity score quantiles can be used in order to split units into *subclasses*.

Genetic Matching [122] is another popular matching method which uses a generalised weighted Mahalanobis distance and applies an evolutionary search algorithm to determine the weight that needs to be assigned in each confounding variable in order to achieve optimal pairs. Genetic matching uses as distance metric between the confounding variables vectors h_u, h_v the following weighted Mahalanobis distance:

$$d_{u,v,W} = \sqrt{(h_u - h_v)^T \cdot \mathbf{W} \cdot (h_u - h_v)} \tag{2.3}$$

where $\mathbf{W} = (S^{-\frac{1}{2}})^T \cdot W \cdot S^{-\frac{1}{2}}$, with W a $P \times P$ diagonal positive definite weight matrix and $S^{-\frac{1}{2}}$ is the Cholesky decomposition of the sample covariance matrix of $H = [h_1, ..., h_N]$. The diagonal elements of W are selected by applying an evolutionary search algorithm that attempts to find the optimal weights to minimise a loss function. Several loss functions can be used. A commonly used loss is the minimum p-value of a t-test or a Kolmogorov-Smirnov distributional test on the matched pairs of treated and control units resulting from applying a given W in the distance calculations between confounders. The loss is calculated for each confounding variable. Thus, if p_p is the p-value of the p^{th} confounding variable, the objective is to find a matrix W that minimises the $min_p p_p$. Other loss functions are based on comparisons of the quantiles of confounding variables for the matched treated and control units. In detail, denote by G_U and G_V the sets of matched treated and control units, respectively, i.e., for each pair $(u, v) \in G$, $u \in G_U$ and $v \in G_V$. For p^{th} confounding variable, I think of the corresponding values for matched treated units $\{h_u^p : u \in G_U\}$ as realisations of a random variable A^p. Analogously, the values $\{h_v^p : v \in G_V\}$ of matched control units will be considered realisations of a random variable B^p. Given a set of

K quantiles $a^p(k)$ and $b^p(k)$ of A^p and B^p, respectively, I calculate a set of quantile differences $\Delta^p = \{|a^p(k) - b^p(k)|\}_{k=1}^K$. Then, one of the following loss functions can be applied: 1) $mean_p mean\Delta^p$, 2) $max_p\Delta^p$, 3) $median_p\Delta^p$, 4) $mean_p max\Delta^p$, 5) $max_p max\Delta^p$, 6) $median_p max\Delta^p$, 7) $mean_p median\Delta^p$, 8) $max_p median\Delta^p$ and 9) $median_p median\Delta^p$.

Matching Parameters

There are three key parameters that can be adjusted when applying a matching method:

- **Many-to-one Matching:** Instead of 1-to-1 matching (i.e., each treated unit is matched with one control unit), k-to-1 matching can be applied (i.e., k controls are used for each treated unit). In this case, the k control units with the lowest distance are selected.

- **Matching with Caliper Distance (Threshold):** A caliper distance can be used to avoid matching units with large distance (i.e., larger than the caliper distance) in cases that a better match cannot be found.

- **Matching with Replacement:** Matching with or without replacement is another key decision when this method is applied. If matching with replacement is applied, each treatment unit can be matched with multiple control units; otherwise, each control unit can be used only once. Matching with replacement can decrease bias when there are multiple treated units with small distance to a single control unit. Since some control units are used multiple times, the analysis could be dependent on the selected control units. Frequency weights need to be used in order to eliminate this bias.

2.3.1.3 Balance Check

After creating the set of matched units G, it is necessary to assess whether the resulted treated and control groups of units are sufficiently balanced. This means that the distribution of the baseline characteristics, described by the P confounding variables, of the set of matched treated units G_U must be similar to the distribution of the baseline characteristics of the set of control units G_V. The most commonly used metric for the estimation of the balance between the matched treated and control units is the *standardised mean*

difference (SMD). For the confounding variable h^p, the standardised mean difference is estimated as follows:

$$SMD_p = \frac{|\bar{h}_U^p - \bar{h}_V^p|}{\sqrt{(\sigma_{p,U}^2 + \sigma_{p,V}^2)/2}} \tag{2.4}$$

where $\bar{h}_U^p$ and $\bar{h}_V^p$ denote the sample mean of the p^{th} confounding variable in the treated and control groups respectively and $\sigma_{p,U}^2$, $\sigma_{p,V}^2$ their sample variances. The standardised mean difference needs to be estimated for each confounding variable. The groups G_U and G_V are usually considered to be balanced if the standardised mean difference is smaller than 0.1 for each confounding variable. If propensity score matching is used, the balance can be assessed by estimating the standardised mean difference on the propensity score.

The p-values of statistical hypothesis tests, such as t-tests or Kolmogorov-Smirnov distributional tests, have also been used to assess the balance between treated and control groups. However, some studies suggest that statistical tests should not be used for balance check since they depend on the sample size and therefore, low statistical power due to limited number of samples could falsely results in false conclusions about the balance [123].

Graphical balance diagnostics can also be used. For example quantile-quantile plots can be used to assess the balance in each confounding variable [117]. In these plots, the quantiles of the matched treated units are plotted against the quantiles of the control units. When the two groups are balanced the empirical distributions for each confounding variable should be similar for the treated and control subjects, thus the points of the plot should approximately lie on the 45 degrees line. The graphical representation of the standardised mean difference before and after the matching procedure is applied can also be used as a balance diagnostic [117].

2.3.1.4 Matching With Continuous Treatments

Although matching frameworks have been proposed mainly for bivariate treatment variables, some recent studies also consider continuous treatments [124, 125]. In such cases units cannot be split into treatment and control groups. Instead, each unit can be matched to any other unit. The goal of matching is to create pairs of units with similar values on

their confounding variables but different treatment values. In [124] the distance between units u, v is estimated as follows:

$$\Delta(u,v) = \frac{\mathcal{D}(h_u, h_v) + \epsilon}{(x_u - x_v)^2} \tag{2.5}$$

where $\mathcal{D}(h_u, h_v)$ is the distance between the vectors of confounding variables values of units u and v (this can be the Euclidean distance, the Mahalanobis distance or any other distance metric), $\epsilon > 0$ a small constant and x_u, x_v are the treatment values of u and v, respectively. With respect to unit v, unit u will be considered as treated if $x_u > x_v$. The average treatment effect is estimated as follows:

$$E_{(u,v)\in G}\Big\{\frac{y_u - y_v}{x_u - x_v}\Big\} \tag{2.6}$$

2.3.1.5 Unobserved Confounding Variables

Conditional ignorability cannot be achieved when one or more confounding variables are unobserved. The main limitation of all non-experimental causality studies is that the possibility that important confounding variables are missing cannot be eliminated. In case of unobserved confounding variables, the assumption of conditional ignorability is violated. Pre-test post-test designs can be applied in order to handle violations of the ignorability assumption. According to the pre-test post-test design, the value of the variable of interest (effect) is observed both before and after the treatment is applied. Thus, any changes to the observations can be attributed to the treatment. However, the validity of this technique is weak [26]. The observed differences may be due to maturation i.e., the values of the observed variable may change over time. Selection bias could be another threat to the validity of this technique. The examined units may share specific characteristics that cause (or contribute to) the observed difference on the effect variable. These limitation can be eliminated by combining pre-tests designs with matching.

In many cases, pre-test measurements are not available, thus pre-test post-test designs cannot be applied. Another approach for evaluating whether the conditional ignorability assumption is valid is to compare the outcome value on the treated group with the outcome in multiple control groups. If a non-zero average treatment effect is observed, then the

result is less likely to be due to violations on the ignorability assumption.

Finally, a sensitivity analysis can be conducted in order to assess how the results of the study would be influenced in the presence of unmeasured confounding variables [118, 126]. In detail, let us denote with π_u the probability that unit u is assigned to a treatment (i.e., $X(t) = 1$) and $O_u = \pi_u/(1 - \pi_u)$ the odds of u to receive a treatment. Then, I denote with $\Gamma = O_u/O_v$ the ratio of the odds of two units u, v. If $\Gamma = n$, the unit u is n times more likely to receive a treatment than unit v due to unobserved factors. Under the conditional ignorability assumption (i.e., units are equally likely to receive a treatment conditional to their observed characteristics), the ratio Γ should be equal to 1 for two matched time-samples u and v.

In [118], Rosenbaum applies the Wilcoxon's signed rank test [127] for the resulted matched treated and control pairs of a causality study under the null hypothesis that the treatment has no effect on the observed outcome variable. According to this method, for each matched pair (u, v) a rank is assigned to the outcome difference $Y(u) - Y(v)$. The Wilcoxon's signed rank statistic W is estimated as the sum of the ranks of the positive differences (the interested reader can find a detail description of the method in [127]). Under the null hypothesis, the mean value of W is $S \cdot (S + 1)/4$, where S the number of matched samples. When S is sufficiently large, the upper bound of the distribution of W can be approximated by a normal distribution with mean $\Gamma/(1 + \Gamma) \cdot S \cdot (S + 1)/2$. Thus, the sensitivity on unobserved confounding variables can be assessed by computing the upper bounds on the p-values of the Wilcoxon's signed rank test for increasing Γ values.

2.3.2 Directed Acyclic Graphs for Causal Inference

What is missing from the potential outcome framework is a formal language for causal analysis representation. Furthermore, randomised experiments and quasi-experiments do not offer explanatory knowledge of an observed causal relationship. The ability to build a causal explanation model from observation would facilitate the generalization of the causal inference to a larger population than the one that has been tested (given that the conditions that caused the observed effect at the tested group hold also for the larger group) and consequently, enhance the external validity of the study.

In [27] Spirtes, Glymour and Scheines develop a stochastic framework for causal inference based on Directed Acyclic Graphs (DAGs). A DAG is represented by a set of vertices (or nodes) and a set of edges (or arrows). The vertices correspond to the variables of the study and the edges represent the relationships among the variables. A node X is called parent of a node Y if there is an arrow from X to Y. In a probabilistic interpretation of a DAG, an arrow from a node X to a node Y denotes that there is a statistical dependence between them. If a set of nodes S blocks all paths from a node X to a node Y (i.e., there is no path from X to Y that does not pass through a node that belongs to set S) then it is said that S *d-separates* X and Y; in this case Y is independent of X conditional to S and this is written as $X \perp\!\!\!\perp Y|S$.

A DAG can be built by utilising prior knowledge that the researchers may have. In case that there is no adequate knowledge of the causal model, researchers can *guess* a graph and then test their assumptions by using observational data. Arrows in the initial model can be added or deleted by performing conditional independence tests. Discovery algorithms have also been proposed for building DAGs from observational data without the need of prior knowledge. A procedure has been proposed by Spirtes, Glymour and Scheines [128]. The algorithm starts with a complete undirected graph where there are arrows between all nodes. Then, for each pair of variables X and Y and for each set of variables S the statistical dependence of X and Y conditional to S is examined. If S d-separates X and Y, then the arrow between them is removed. Conditional independence tests are also used to orient the direction of the arrows. Discovery algorithms have large computational complexity (depending on the number of variables) and they may also result in more than one equivalent graphs. PC algorithm [129], a modification of this basic algorithm, reduces the computational complexity of the graph discovery by performing the independence tests in some order and skipping unnecessary tests. Several tools have been created for automatic causal discovery based on graph models [130, 131].

Under certain assumptions, a DAG can be interpreted as a causal graph. In a causal graph, an arrow from X to Y denotes that X is a direct cause of Y. Mores specifically, causal graphs are based on the following assumptions:

1. *Causal Markov Condition.* Each variable X in the graph is independent to any other variable, excluding variables that are affected by X, conditional on its direct

causes.

2. *Faithfulness.* Given a graph that satisfies the Causal Markov Condition, any population derived from this graph follows the same independent relationships with the ones obtained by applying d-separation on the graph (as described in the 2nd and 3rd paragraphs of this section), i.e., any observed independences are not due to 'coincidence'. Thus, the sampled data that are used in order to test the dependences among the variables of the study and build the graph should be representative of the real population. This is a reasonable assumption that needs to be made in any causal inference study.

3. *Causal Sufficiency assumption.* All the common causes of any pair of variables (X, Y) of a causal graph are represented in the graph, i.e., there are not any unmeasured confounding variables. It should be stressed that the graph does not have to contain all the causes of any variable in the graph.

2.3.3 Structural Equation Models

There is an alternative causal inference approach which describes each node on a causal graph as a noise variable and derives a set of structural equation, one for each noise variable [33]. For example, consider the causal graph of Figure 2.1. Assuming a linear model, the following structural equations can be derived by the graph:

$$Z = U_Z$$

$$N = \alpha_N \cdot Z + U_N$$

$$X = \alpha_X \cdot Z + U_X$$

$$Y = \alpha_Y \cdot Z + \beta_Y \cdot N + U_Y$$

The variables U_N, U_Z, U_X and U_Y model any unknown independent factors (i.e., latent

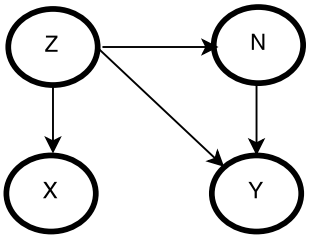

Figure 2.1: Example Graph

variables) that influence the variables N, Z, X and Y respectively and $\alpha_N, \alpha_X, \alpha_Y, \beta_Y$ are real numbers. Non-linear models can also be used in order to model the dependencies among the variables. Normally, researchers use some prior knowledge about the data in order to create some reasonable structures i.e. some structures that are considered to be acceptable based on researchers' knowledge about the field. Then, the model is selected by fitting the data to the candidate models. Goodness of fit measures can be used to evaluate which structural equation model best describes the data.

Structural equation models are strongly related to DAGs. However, the causal structure is learnt by fitting the data into a model rather than by conducting conditional independence tests. Structural equation models can include latent variables, thus they can model the influence of unmeasured factors. On the other hand, they assume a specific functional relationship among the variables. Any model misspecification could result in misleading conclusions. In addition, functional models could be overfitted and thus, the extracted relationships may not accurately describe the real structure of the causal model [132].

2.3.4 Causality on Time-Series Data

All the previously described frameworks assume i.i.d. data, thus they cannot be directly applied to time-series data. In this section, I describe the main approaches for causal inference on time-series and I discuss their limitations. All causal inference methods discussed in this section are based on the following assumptions:

- The data are stationary, i.e., the impact of a time-series X on a time-series Y is independent of the time.

- There is a maximum time-lag L so that there is no influence of any time-sample $Z(t-k)$ on any time sample $R(t)$ for any $t-k \geq L$ for any time-series Z, R.

2.3.4.1 Granger Causality

Causality studies on time-series have been largely based on Granger causality [133]. The Granger causality test examines if past values of one variable are useful in prediction of future values of another variable. In detail, a time-series X Granger causes a time-series Y if modeling Y by regressing it on past values of both Y and X results in reduced residual noise compared to a simple autoregressive model. According to the linear Granger causality, a variable Y is represented as an autoregressive model as follows:

$$Y(t) = \sum_{i=1}^{L} \alpha_i \times Y(t-i) + e(t) \tag{2.7}$$

where $e(t)$ corresponds to random gaussian noise. Then, this model is enhanced by adding lagged values of a variable X as follows:

$$Y(t) = \sum_{i=1}^{L} \alpha_i \times Y(t-i) + \sum_{i=1}^{L} \beta_i \times X(t-i) + e(t) \tag{2.8}$$

Variable X is considered to Granger cause Y, if adding lagged values of X at the autoregressive model of Y improves the model significantly, according to some t-statistic or f-statistic test [134].

A positive result on a Granger causality test does not necessarily imply that there is a causal link between the examined time-series since the conditional ignorability assumption is not satisfied, i.e., the values of both treatment variable X and control variable Y may be driven by a third variable (*confounding bias*). In addition, it considers only linear relationships. Granger causality has been extended to handle multivariate cases [135] as well as non-linear cases [128, 130]. In [131] the authors propose the use of structural equation models for time-series data. An additional model check procedure is applied after fitting a model in order to reduce the amount of false positive causality results. Moreover, in [136] the authors propose a time-series causality framework based on graph models. The main advantage of the proposed method is the ability to model latent variables (i.e.

unobserved confounding variables). However, this method performs worse than Granger causality for large time-series sample sizes.

Causal inference based on functional models suffers from two main limitations. First, estimation of model coefficients in scenarios involving a large number of predictor time-series or when the predictor variables are correlated (multicollinearity) can be challenging. When data dimensionality is comparable to the sample size, noise may dominate the 'true' signal, rendering the study infeasible [31]. Second, it is difficult to select a suitable functional form (i.e. linear or non-linear). Inaccuracies in model specification, estimation or selection may result in invalid causal conclusions.

2.3.4.2 Transfer Entropy

Non-parametric approaches (i.e. approaches that do not require the specification of a model class) for causal inference in time-series based on *transfer entropy* have also been proposed [137]. Transfer entropy is a model-free equivalent of Granger causality [138] and describes the amount of reduction on uncertainty about a time-series Y given the past values of Y when the past values of a time-series X are known and is estimated as follows:

$$T_{X \to Y} = H(Y_t | Y_t^-) - H(Y_t | Y_t^-, X_t^-) \tag{2.9}$$

where $H(Y_t)$ denotes the Shannon's entropy of Y_t and Y_t^-, X_t^- denote the past of time-series Y_t and X_t respectively.

Although transfer entropy is originally designed for bivariate analysis, it has also been extended to multivariate cases [139]. The multivariate case considers a set of time-series $\mathbf{S}$ and examines whether the uncertainty about a time-series $Y \in \mathbf{S}$ is reduced by learning the past of a time-series $X \in \mathbf{S}$, when the past of the other time-series in $\mathbf{S}$ is known. Thus, the transfer entropy on a multivariate analysis is estimated as follows:

$$T_{X \to Y} = H(Y_t | \mathbf{S_t^-} \setminus X_t^-, Y_t^-) - H(Y_t | \mathbf{S_t^-} \setminus Y_t^-) \tag{2.10}$$

The main limitation of this approach is that it requires the estimation of a large number of conditional probability densities, which might be particularly challenging on continuous datasets [128]. Runge et al. [140, 141] propose the combination of transfer

entropy with directed acyclic graphs in order to reduce the number of densities that need to be estimated. In detail, causality is estimated by examining whether uncertainty about time-series Y can be reduced by learning the past of X, when the *parents* of Y and X are known. The parents P_Y of a time-series Y are defined as the minimum set of graph-nodes which separate Y with the past of $\mathbf{S} \setminus \{P_Y\}$. Although this modification reduces significantly the number of density estimations that are required, the dimensionality of the dataset may still be high (i.e. the number of parents of Y and X may be very large) which imposes challenges on the estimation of transfer entropy.

2.3.5 My Contribution

My contribution on the causal inference domain is motivated by specific problems characterising the datasets that are considered in this study. In particular:

1. I propose a novel method for causal inference in time-series based on the matching design framework [62, 142]. The proposed method has two main advantages over the existing approaches:

   (a) In contrast to functional models (described in Sections 2.3.3 and 2.3.4.1), the proposed method does not make any assumptions about the structural relationships among the examined variables and therefore it can handle better non-linear cases. As was previously mentioned, some studies have provided evidence of non-linear dependency between social media data and stock market prices [61, 62]. In such cases, approaches that are based in linear models may fail.

   (b) It can better handle high-dimensional data, compared to the non-parametric approaches described in Section 2.3.4.2. Stock market prices are influenced by a large number of factors such as commodity prices, prices of other traded assets and relevant news. Although studies on human behaviour using smartphone sensor data usually involve a significantly lower amount of variables, they are mostly based on relatively small datasets due to the difficulty of data collection; consequently, the analysis can be challenging even with small number of

variables.

The method is presented in detail in Chapter 3.

2. I propose a causal inference method that can handle noisy data [143]. In most of the studies discussed in Sections 2.1 and 2.2 the variables of interest are not directly measured; instead, they are inferred from other low level information. For example, in [39, 40, 42] the sentiment of tweets is extracted by applying text processing methods and therefore, it is not accurate. This issue has been neglected by previous studies. In Chapter 6, I present a probabilistic causal inference method based on the matching design framework that can handle stochastic variables and maximises the probability that any bias has been eliminated.

2.4 Summary

In this chapter, I have introduced the two main problems that are examined in this thesis i.e., understanding the influence of social media on stock market prices and analysing human behaviour using smartphone sensor data. Currently, most works are based on correlation analysis. Although these works provide valuable insights about the links among the variables of interest, they cannot prove that the predictor variable is actually influencing the response. In this chapter, I have highlighted the importance of detecting causal links in human digital traces. In order to prove causality, an experimental procedure needs to be applied. However, in many cases, conducting randomised trials is not feasible. Consequently, applying causal inference methods based on observational data is the best available option. Although the findings of any causality study based on observational data should be interpreted with caution, such studies comprise a significant step towards understanding human behaviour and phenomena influenced by it.

In addition, in this chapter, I have presented the main methods for causal inference in observational data and I have discussed their strengths and limitations. Functional models (described in Sections 2.3.3 and 2.3.4.1) have been widely used for causal inference in many fields. The main limitation of such models is that they require the selection of a specific functional form. Methods based on conditional independence tests (described in Sections

2.3.2 and 2.3.4.2) are less restrictive. However, they often require large conditioning sets and, therefore, could be unreliable when the size of the dataset is not adequate.

On the other hand, according to the matching design framework the data are adjusted in order to 'mimic' a randomised trial. Matching design does not require any assumptions about the functional form of the data and can better handle high-dimensional data. In addition, the design is separated from the analysis and rules can be applied to assess whether confounding bias has been sufficiently eliminated. However, matching design assumes i.i.d. data and therefore, it cannot be applied for time-series analysis. In the following chapter, I present a novel method for causal inference on time-series based on the matching design framework.

, , , , , 500 , ,

CHAPTER 3

MATCHING DESIGN FOR TIME-SERIES

As was previously discussed in Sections 2.3.4.1 and 2.3.4.2, causal inference in high-dimensional time-series data is a challenging task. Given the limitations of the previously discussed methods, in this chapter I discuss a novel framework for causal inference in time-series that is based on *matching design* method presented in Section 2.3.1. Matching design has several advantages over regression-based methods or those based on information theory. However, it cannot be applied to time-series since it assumes that the objects of the study are realisations of i.i.d variables. I reformulate the concept of matching design to make it suitable for causal inference on time-series data. In this case, the time-series collection includes *treatment* time-series X, *response* time-series Y and a set of time-series $\mathbf{Z}$ which contain characteristics relevant to the study. The units of this study correspond to time-samples; the t^{th} unit is characterised by a treatment value $X(t)$, a response value $Y(t)$ and a set of values representing baseline characteristics $\mathbf{Z(t)}$. The baseline characteristics of the units of the study are any features that can influence the outcome of the study and also their treatment values. For example, let us consider a study about the impact of tourism on the economy of a country; events like natural disasters or turmoils should be considered as baseline characteristic of the study as they may influence both the economy of the country (outcome variable) and the tourism (treatment variable). I assess the causal impact of a time-series X on Y by comparing different units (i.e., time-samples) on Y after controlling for characteristics captured in $\mathbf{Z}$. As explained in the following section, the proposed methodology assures that the objects are uncorrelated, which is a weaker version of the independence assumption requirement of the matching design. The main advantages of the proposed method are:

- It is not based on assumptions about the functional form of the relationships among the examined time-series (i.e., linear or non-linear). As was previously discussed, methods based on matching do not require fitting the data into a model.

- It requires fewer conditional independence tests with significantly smaller conditioning sets, compared to existing approaches based on transfer-entropy, thus it can handle more effectively high-dimensional data.

In the following sections, I will describe the proposed causal inference framework and I will present an evaluation of this approach on synthetic data, in comparison with existing approaches for causal inference in time-series data.

3.1 Mechanism Description

Let us denote by $Y = \{Y(t_i^y) : i = 0, 1..., N\}$ and $X = \{X(t_i^x) : i = 0, 1..., N\}$ the time-series that represent the effect and the cause, respectively and by $\mathbf{Z} = \{\mathbf{Z}(t_i^z) : i = 0, 1..., N\}$ a set of time-series representing other characteristics relevant for the study. In this study, I consider X as a binary treatment variable. As was previously discussed in Section 2.3.1.4, matching design has been proposed also for non-binary treatments [124, 125]. However, in this Chapter, I focus only on binary treatment variables. Let me also denote by $Y^{(l)}$, $X^{(l)}$ and $\mathbf{Z}^{(l)}$ the l-lagged versions of the time series Y, X and $\mathbf{Z}$, respectively (i.e., if $X(t_i^x)$ the i-th sample of X, $X^{(l)}(t_i^x) = X(t_{i-l}^x)$). In Figure 3.1 I provide a graphical representation of time-series X, $X^{(1)}$, ..., $X^{(L)}$. I define a maximum lag value L and a set of time-series $\mathbf{S}$ =$\{Y$, $Y^{(1)}$, ..., $Y^{(L)}$, X, $X^{(1)}$, ..., $X^{(L)}$, $\mathbf{Z}$, $\mathbf{Z}^{(1)}$, ..., $\mathbf{Z}^{(L)}\}$.

As I previously discussed in Chapter 2, the units of a study traditionally correspond to experimental objects and the variables of the study describe the characteristics of the units as well as the treatment they have received and the corresponding outcome. In my framework, the units of the study correspond to time-samples of the set of times-series $\mathbf{S}$. For example, let us consider a study that examines the effects of an industry on the pollution level in a region based on weekly measurements. In this case, a unit of the study corresponds to one week and the variables of the study to weekly pollution measurements,

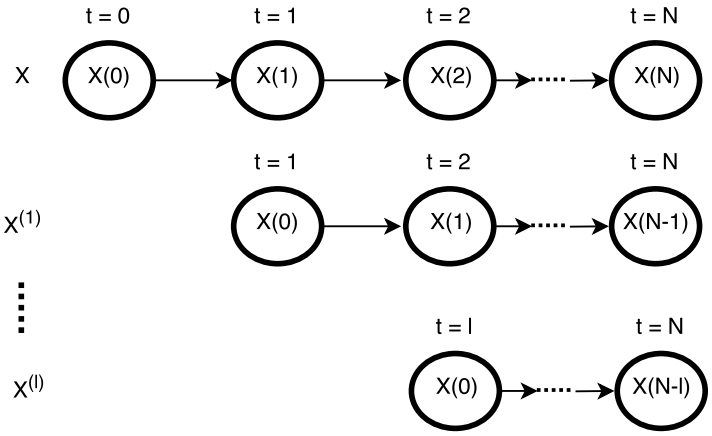

Figure 3.1: Graphical representation of time-series X along with its first l lagged versions.

industrial wastes and other relevant characteristics. In Figure 3.2 I graphically depict the notion of a unit in my time-series matching design framework in comparison with the traditional notion of unit on causality studies. In the rest of this chapter, the terms 'unit' and 'time-sample' will be used interchangeably.

In order to build a graph, I examine the dependencies between the variables X, Y and all the other variables of the set $\mathbf{S}$. In order to examine if two time-series X and Y are independent (assuming that the time-series are stationary in the first two moments) I can estimate the Pearson correlation coefficient as follows:

$$r_{xy} = \frac{\sum_{u=0}^{N}(X(t_u^x) - \bar{X})(Y(t_u^y) - \bar{Y})}{\sqrt{\sum_{u=0}^{N}(X(t_u^x) - \bar{X})^2 \sum_{u=0}^{N}(Y(t_u^y) - \bar{Y})^2}}$$

with $\bar{X}$, $\bar{Y}$ the sample means of X, Y respectively. Vanishing correlation could be considered as indication of independence between the examined time-series. However, a vanishing linear correlation is not always an adequate indication of independence. Alternatively, Spearman rank correlation or mutual information could be used in order to examine the dependencies between time-series.

In a directed acyclic graph representing a Bayesian network, an arrow from a variable W to a variable Q is added only if Q is dependent of W, conditional on all direct predecessors of Q. In our graph representation, I relax this condition as follows:

An arrow from a lagged node $W^{(l)}$ (including lag 0) to a non-lagged node Q exists if:

- $W^{(l)}$ precedes temporally Q, i.e., $t_u^1 < t_u^2$, for any u; and

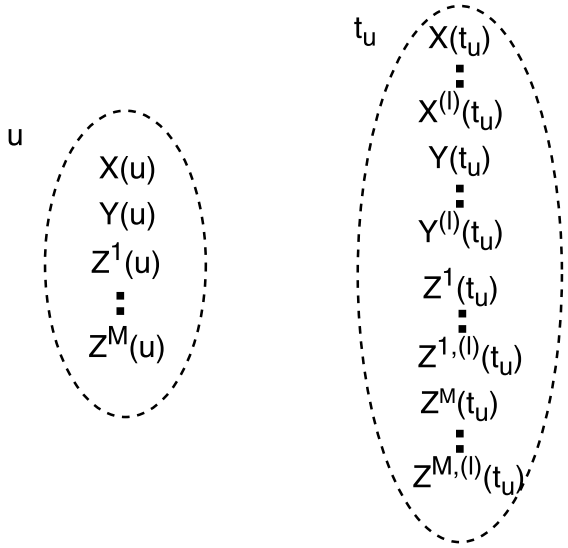

Figure 3.2: Graphical representation of units. On the left side, u represents a unit on a traditional causality study, characterised by its treatment value $X(u)$, its response value $Y(u)$ and M other characteristics $Z^1(u)$, $Z^2(u)$..., $Z^M(u)$. On the right side, t_u represents a unit on our time-series matching design framework. The unit is characterized by the time-series values of set $\mathbf{S}$ at $u-th$ time-sample.

- $Q \not\perp\!\!\!\perp W^{(l)} | \mathbf{P^m} \cap (W, W^{(1)}, ..., W^{(m)})$, where $\mathbf{P^m}$ is the set of the direct predecessors of Q with maximum lag m and $m < l$.

Thus, in my graph representation, a direct edge between two nodes indicates a dependence but not necessarily a causal link. Causality will be examined by applying the matching design framework, where the direct predecessors of the treatment and outcome time-series will serve as the confounding variables of the study. The main advantage of the proposed framework is the requirement of a significantly lower number of densities estimations and conditional independence testing compared to other causal inference approaches on time-series [137, 140, 141]. In detail, since we only need to examine the dependence of a time-series Q with a time-series W conditional to the past of W, the maximum conditioning set is L. In what follows I will discuss the details of my methodology and how the three general assumptions of causality studies (discussed in Chapter 2) are addressed.

Conditional Ignorability Assumption: I apply the Algorithm 1 in order to find the set of time-series $\mathbf{H}$ that need to be controlled in order to satisfy the conditional ignorability assumption. According to the proposed method, the resulted set contains all the direct

Algorithm 1 Defining the set of confounding variables.

Input: The set of time-series **S**
Output: The set of confounding variables **H**

```
{Find the parents of Y.}
P1 :=predecessors(S, Y)
{Find the parents of X.}
P2 :=predecessors(S, X)
{Find the common parents of X, Y.}
H := P1 ∩ P2

{This procedure returns a set P of the direct predecessors of node Q. P is a subset of S.}
predecessors(S, Q)
P := {}
for i=0 to L do
  {For all zero-lagged time-series}
  for all S^(0) ∈ S do
    {Find the lagged versions of S which are also parents of Q.}
    B := (S^(0), ..., S^(i−1)) ∩ P
    if (Q ⫫̸ S^(i)|B) and (S^(i) precedes Q) then
      P := P ∪ S^(i)
    end if
  end for
end for
return P
```

predecessors that nodes X and Y have in common. In Figure 3.3, I depict the resulted set **H** of an example graph comprised by time-series X, Y and W as well as its lagged versions, with maximum lag $L = 2$. The parents of X and Y are selected by conducting conditional independence tests as described in Algorithm 1. For example, the arrow from $X^{(1)}$ to X denotes that $X \not\perp\!\!\!\perp X^{(1)}$ and the arrow from $X^{(2)}$ to X that $X \not\perp\!\!\!\perp X^{(2)}|X^{(1)}$. Similarly, the arrow from W to X denotes that $X \not\perp\!\!\!\perp W$ and the arrow from $W^{(1)}$ to X denotes that $X \not\perp\!\!\!\perp W^{(1)}|W$. The lack of arrow from $W^{(2)}$ to X denotes that $X \perp\!\!\!\perp W^{(2)}|W, W^{(1)}$. **H** includes all the common parents of X and Y. Thus, all the variables that are correlated with both X and Y time-series are included; hence, the set **H** is sufficient. However, **H** may include also redundant time-series, i.e., some of the time-series included in **H** may not correlate with X or Y conditional to a subset of **H**. In causality studies based on regression, including redundant predictors on the model could result in overfitting and would jeopardize the validity of the conclusions. Moreover, the application of methods

based on conditional independence tests using information theoretic approaches would be challenged by the inclusion of redundant covariates since it would require conditioning on large sets of variables. In contrast, studies based on matching are less affected by the inclusion of redundant confounding variables (spurious correlations). Several methods that enable matching on a large number of confounding variables have been proposed [122, 144, 32]. In addition, researchers are able to apply balance diagnostic tests in order to assess if any confounding bias has been adequately eliminated [145]; consequently, false conclusions due to confounding bias can be diminished. Following the matching design, the set of time-series $\mathbf{H}$ is controlled by creating a set of pairs of time-samples G where each u-th time-sample with a positive treatment value $X(t_u^x)$ is matched with a v^{th} time-sample with zero treatment $X(t_v^x)$ such that $\mathbf{H}(t_u^h) \approx \mathbf{H}(t_v^h)$. It should be noted that only factors that precede temporally both Y and Y can influence the study. If a factor precedes only the effect variable (or the treatment variable), it cannot drive the values of both treatment and effect, thus it will not influence the study.

Stable Unit Treatment Value Assumption: Denote by $\mathbf{P}$ the set of time-series that are direct predecessors of the effect variable Y. Assuming $X \in \mathbf{P}$ (if not, X is independent of Y and therefore there is no causation), the assumption is violated if $X^{(l)} \in \mathbf{P}$ and $X^{(l)} \notin \mathbf{H}$, for $l > 0$. Since units correspond to time-samples, $X^{(l)} \in \mathbf{P}$ implies that the outcome value $Y(t_u^y)$ at time t_u^y depends on the value of the treatment time-series X at time t_{u-l}^x. In order to satisfy the assumption, I modify the $\mathbf{H}$ set as follows:

$$\mathbf{H} := ((X^{(1)}, ..., X^{(L)}) \cap \mathbf{P}) \cup \mathbf{H}, \tag{3.1}$$

satisfying $Y(t_u^y) \perp\!\!\!\perp X(t_v^x)|\mathbf{H}(t_u^h)$, $\forall u \neq v$.

I.i.d. assumption: Denote by Y_1 the value of the outcome variable for the time-samples that have a positive treatment value and with Y_0 for time-samples with zero treatment value. The average causal effect is estimated as $\widehat{E}\{Y_1 - Y_0|\mathbf{H}\}$. In order to enable statistical inference, the variable $\Delta Y := Y_1 - Y_0|\mathbf{H}$ needs to be i.i.d.. If $\mathbf{P}$ the set of direct predecessors of Y, the outcome value $Y(t_u^y)$ of each time-sample t_u^y will depend on the outcome value $Y(t_{u-l}^y)$ if there is a time-series $Y^{(l)} \in \mathbf{P}$. In case that $Y^{(l)} \notin \mathbf{H}$, the i.i.d. assumption would be violated. In order to satisfy this assumption, I modify the set

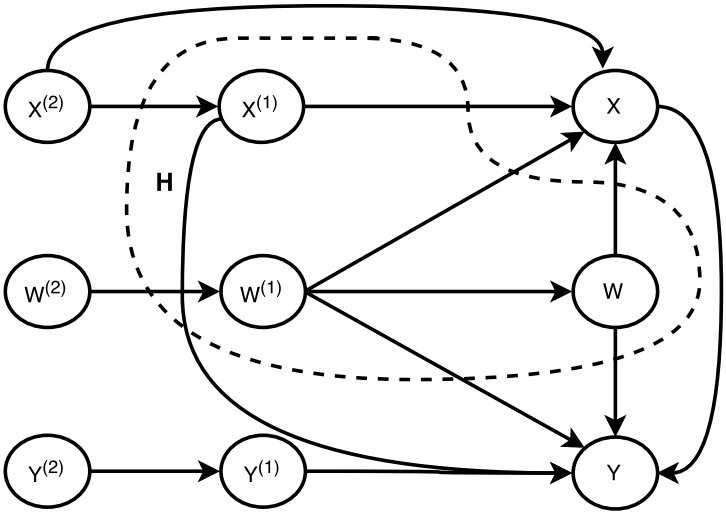

Figure 3.3: Example graph depicting the resulted set **H** when the impact of X on Y is examined. At this example, X precedes temporally Y and W precedes X. The maximum examined time-lag L is 2.

of time-series **H** as follows:

$$\mathbf{H} := ((Y^{(1)}, ..., Y^{(L)}) \cap \mathbf{P}) \cup \mathbf{H} \tag{3.2}$$

Causal inference will be performed by matching on the modified set of time-series **H** thus, the variable $\Delta Y := Y_1 - Y_0|\mathbf{H}$ will be i.i.d.. Any matching method can be applied. Researchers should choose a matching method that achieves sufficient balance between the matched treated and control units by applying the framework described in Section 2.3.1.

3.2 Evaluation on Synthetic Data

In order to demonstrate the potential of this approach I assess its effectiveness in detecting causal relationships on linear and non-linear synthetic data. I also compare my approach with a multivariate Granger causality model and with an information theoretic approach based on Runge's framework [141] and I demonstrate that the proposed method is more efficient on avoiding false causal conclusions. I denote with $X = \{X(t_u^x) : u = 1, 2, ...N\}$ and $Y = \{Y(t_u^y) : u = 1, 2, ...N\}$ the treatment and outcome time-series respectively and with $\mathbf{Z} = \{\mathbf{Z}(t_u^z) : u = 1, 2, ...N\}$ a set of M confounding variables. I also assume that

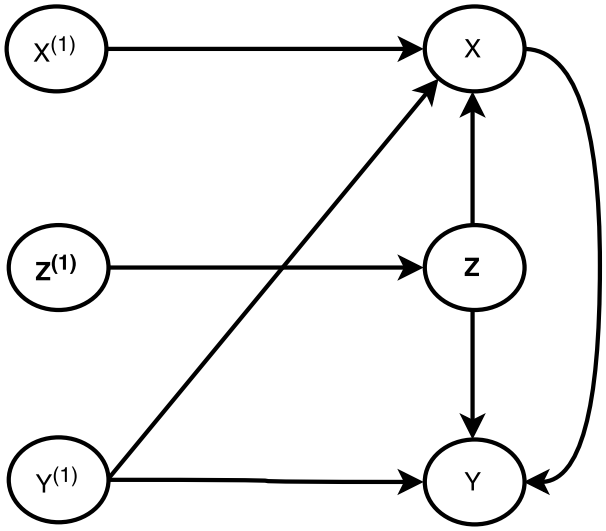

Figure 3.4: Resulting graph after applying Algorithm 1 on the synthetic data when $M = 1$ (i.e. there is only one confounding variable). The graph depicts the direct predecessors of nodes X and Y. The set of nodes $\mathbf{H}$ will contain the direct predecessors that nodes X and Y have in common. In the four examined cases X correlates with Y, though in Case 2 and Case 4, this is a spurious correlation due to the set of confounding variables $\mathbf{Z}$. There is also a spurious correlation of node X with node $Y^{(1)}$. X and Y are independent to $\mathbf{Z}^{(1)}$ conditional to $\mathbf{Z}$ and Y is independent to $X^{(1)}$ conditional to X.

$t_u^z < t_u^x < t_u^y$, $\forall u$. The relationships among X, Y and $\mathbf{Z}$ are described by the following model:

$$X(t_u^x) = h_{xx}(X(t_{u-1}^x)) + f_{xz}(\mathbf{Z}(t_u^z)) + \epsilon_x(t_u^x) \tag{3.3}$$

$$\begin{aligned} Y(t_u^y) &= h_{yy}(Y(t_{u-1}^y)) + f_{yz}(\mathbf{Z}(t_u^z)) \\ &\quad + f_{yx}(X(t_u^x)) + \epsilon_y(t_u^y) \end{aligned} \tag{3.4}$$

$$Z^i(t_u^z) = h_{z^i}(Z^i(t_{u-1}^z)) + \epsilon_{z^i}(t_u^z), \forall Z^i \in \mathbf{Z}, \tag{3.5}$$

where $\epsilon_x(t_u^x)$, $\epsilon_y(t_u^y)$ and $\epsilon_{z^i}(t_u^z)$ are i.i.d. Gaussian noise variables with zero mean and std. dev. equal to $20 + 2 \cdot M$, $10 + 2 \cdot M$ and 10, respectively.

I consider the following four cases:

Case 1. The model is linear. Thus,

$$f_{xz}(\mathbf{Z}(t_u^z)) = \sum_i \alpha_{xz,i} \cdot Z^i(t_u^z)$$

$$h_{xx}(X(t^x_{u-1})) = \alpha_{xx} \cdot X(t^x_{u-1})$$

$$h_{yy}(Y(t^y_{u-1})) = \alpha_{yy} \cdot Y(t^y_{u-1})$$

$$f_{yz}(\mathbf{Z}(t^z_u)) = \sum_i \alpha_{yz,i} \cdot Z^i(t^z_u)$$

$$f_{yx}(X(t^x_u)) = \alpha_{yx} \cdot X(t^x_u)$$

$$h_{z^i}(Z^i(t^z_{u-1})) = \alpha_{z^i} \cdot Z(t^z_{u-1})$$

Case 2. I apply the linear model of Case 1, but I set $f_{yx}(X(t^x_u)) = 0$. In this case the treatment time-series X does not have any causal impact on the outcome time-series.

Case 3. The associations of the confounding variables with the treatment and effect variables are non-linear. In particular, I assume that:

$$f_{xz}(\mathbf{Z}(t^z_u)) = \sum_i \alpha_{xz,i} \cdot (Z^i(t^z_u))^2$$

$$f_{yz}(\mathbf{Z}(t^z_u)) = \sum_i \alpha_{yz,i} \cdot (Z^i(t^z_u))^2$$

I use the linear equations of Case 1 for the rest of the functions.

Case 4. I use the non-linear model of Case 3, but I set $f_{yx}(X(t^x_u)) = 0$. In this case, the multivariate linear Granger causality approach may return positive causality result, even though the treatment time-series $X(t)$ does not have any causal impact on the outcome time-series.

I assume that time-series X is sampled before Y and that all time-series $\mathbf{Z}$ are sampled before X. A unit (i.e. time-sample) t of the study is described by the set of time-series values: $S(t_u) := (X(t^x_u), Y(t^y_u), \mathbf{Z}(t^z_u), X^{(1)}(t^x_u), Y^{(1)}(t^y_u), \mathbf{Z}^{(1)}(t^z_u))$. I apply the following three methodologies on the synthetic data generated using the models above in order to assess the causal impact of variable X on Y:

Multivariate Granger Causality (MGC). I apply stepwise regression in order to fit the data to a multivariate Granger causality model described by the following equation:

$$Y(t^y_u) = a_1 \cdot Y(t^y_{u-1}) + \sum_{l=0}^{(1)} b_l \cdot X(t^x_{u-l}) + \sum_{l=0}^{(1)} \mathbf{c}_l \cdot \mathbf{Z}(t^z_{u-l}) + \delta + \epsilon(t^y_u) \tag{3.6}$$

I conclude that X causes Y if X or any lagged version of X is included in the regression model.

Conditional Mutual Information Tests (CMI). Following Runge's approach [140], a causal graph is created by performing conditional independence tests using conditional mutual information as described in Section 2.3.4.2.

Matching Design for Time-series (MDT). Following the proposed approach, I apply Algorithm 1 in order to find the set of variables $\mathbf{H}$ that needs to be controlled in order to achieve conditional ignorability. The resulted graph is depicted in Figure 3.4. $\mathbf{H}$ includes any $Z^i \in \mathbf{Z}$ that correlates both with X and Y. Moreover, I satisfy the i.i.d assumption by including in $\mathbf{H}$ the time-series $Y^{(1)}$. In order to create groups of treated and untreated units I first transform the time series X into a binary stream $\tilde{X}$: $\tilde{X}(t_u^x) = 0$, if $X(t_u^x) < \mu_X$; $\tilde{X}(t_u^x) = 1$, otherwise, where μ_X is the mean of X (i.e. the u-th time-sample corresponds to a treated unit if $X(t_u^x) > \mu_X$). Then, I create pairs of treated and untreated units (i.e. time-samples) by applying Genetic Matching algorithm [122]. Genetic matching is a multivariate matching method which applies an evolutionary search algorithm in order to find optimal matches which minimise a loss function. Simpler matching approaches (e.g. nearest neighbour matching) were also considered; however genetic matching resulted in more balanced treatment and control groups. As a loss function, I used the average standardised mean difference between the treated and control units for all the confounding variables $H^i \in \mathbf{H}$ which is defined as follows:

$$SMD_H = \sum_{H^i \in \mathbf{H}} \frac{\sum_{(t_u^h, t_v^h) \in G} |H^i(t_u^h) - H^i(t_v^h)|}{|G| \cdot \sigma_{H^i}} / |\mathbf{H}| \tag{3.7}$$

where G corresponds to the set of matched treated and control units (see Table 2.1). Finally, the average treatment is estimated using Equation (2.2) and a t-test is used to examine whether the average treatment effect (ATE) is significantly different from 0.

I generate 100 samples for each time-series. I vary the number of confounding variables M that are included at set $\mathbf{Z}$ from 10 to 50. In detail, I evaluate the three methodologies for $M = \{10, 20, 30, 40, 50\}$. For each M value, I repeat the study for 30 randomly selected sets of model coefficients (αs). All model coefficients are randomly generated from uniform

distribution on $[-4, 4]$ for the linear cases and on $[-1, 1]$ for the non-linear cases. These values are selected so that the resulted signal to noise ratio serves the needs of this study (i.e., if the noise dominates the true signal, then any causality (or correlation) analysis will fail to uncover the true relationships between the variables; on the other hand, if the noise is very small compared to the true signal, any method will result in positive conclusions). Finally, for each one of the 30 sets of model coefficients I repeat each study for 100 different noise realisations. For the n^{th} noise realisation of the k^{th} set of model coefficients, I define:

$$S_{k,n} = \begin{cases} 1 & \text{if } X \text{ was detected as cause of } Y \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

For the k^{th} set of model coefficients I also define $A_k = \sum_{n=1}^{100} S_{k,n}$. In Case 1 and Case 3, A_k denotes the number of times that a causal relationship from X to Y is successfully inferred (*true positive*) for the k^{th} set of model coefficients and different noise realisations, while in Case 2 and 4 it denotes the number of times that a causal relationship is falsely inferred (*false positive*). In Figure 3.5 I present the mean value of A_k, μ_{A_k} along with the standard error of the mean. According to my results, the proposed causal inference technique reduces significantly the number of false positive causality conclusions while it is slightly less successful on detecting real causality for $M = 10$. Multivariate Granger causality achieves almost 100% accuracy on true causality detection both for the linear (Case 1) and non-linear (Case 3) cases. However, it performs poorly in terms of avoiding false positive conclusions. The performance of all the examined methods improves for larger M values (apart from multivariate Granger causality on the linear cases). This is due to the fact that, by adding more variables on the set $\mathbf{Z}$, the dependence of Y and X with each individual $Z^i \in \mathbf{Z}$ is weaker; consequently, although M covariates are used to generate X and Y time-series, for large M values, only a subset of them has significant effect on them. Thus, cancelling out the effect of $\mathbf{Z}$ is easier.

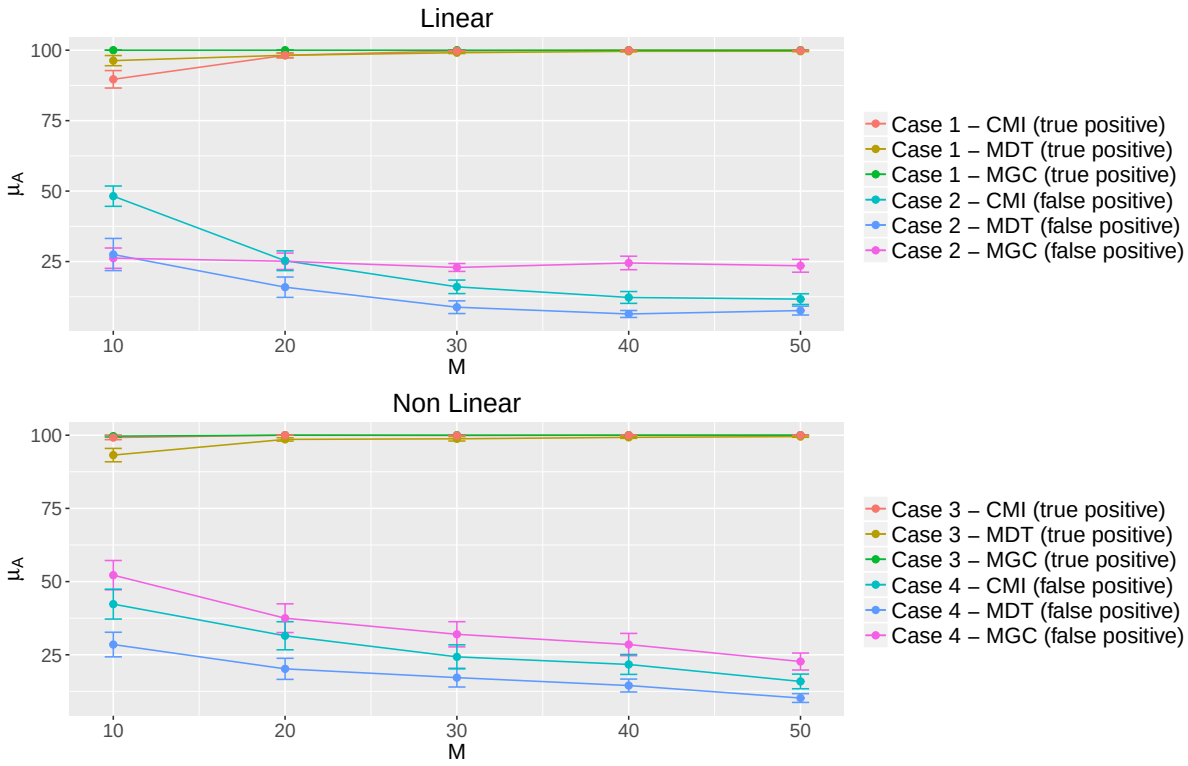

Figure 3.5: Comparison of the MDT, CMI and MGC causality detection methods on synthetic data.

3.3 Discussion

My results on the simulated experiments indicate that the proposed method is more effective on avoiding false positive causality conclusions. I have examined the performance of the proposed method in datasets with up to 50 dimensions and 100 time-samples. Extreme high-dimensional cases with $p > n$ are not considered in this study. In such cases, balancing treatment and control groups for each confounding variable would require a large number of samples. *Propensity score matching* [146] represents an alternative matching method that can effectively handle a large number of confounding variables by performing matching on a single balancing score, i.e. the *propensity score*. The propensity score corresponds to the probability of a unit to be assigned to a treatment and it is usually approximated by applying a logistic regression model of the treatment against the set of confounding variables. *High-dimensional propensity score matching* [147] has also been proposed in order to handle extreme high-dimensional cases with $p >> n$.

One of the main advantages of the proposed MDT (Matching Design for Time-series)

method over multivariate Granger causality and CMI (Conditional Mutual Information tests) is that the design of the study is separated from the analysis. The values of the response time-series Y are not used during the matching process. The causal impact of a time-series X on Y is evaluated only after sufficient balance between the treated and untreated samples has been achieved. In contrast, in a regression-based analysis the response time-series Y is used in order to learn the coefficients of the predictor variables of the study. Many studies suggest that regression-based methods for causal inference are less reliable [148]. Moreover, the proposed method is non-parametric, while Granger causality is based on assumptions about the model class (i.e. linear/non-linear relationships). According to my results, linear Granger causality performs poorly when there are non-linear relationships among the examined time-series.

In addition, as it was previously discussed, MDT requires significantly fewer conditional independence tests and smaller conditioning sets. In detail, the maximum conditioning set of the proposed method is equal to the maximum lag L, while the maximum conditioning set of CMI is $M \cdot L$ (with M the number of confounding variables). Thus, MDT can handle more effectively datasets which include a large number of confounding variables.

Moreover, the computational cost of creating the graph is significantly lower for MDT compared to CMI. Assuming discrete time-series with values in a set V, the computational complexity of creating a graph by applying the proposed method is $O(|V|^L \cdot M \cdot N)$, while the computational cost of CMI is $O(|V|^{M \cdot L} \cdot M \cdot N)$, with $|V|$ the size of set V. However, causal inference with MDT requires an additional matching step, and consequently, its computational complexity largely depends on the matching method that is applied. If a simple nearest neighbour matching method is applied [117], the cost of finding the best match of a single unit is $O(|\mathbf{H}| \cdot N)$, with $|\mathbf{H}|$ denoting the size of set $\mathbf{H}$, and the cost of matching all the units is $O(|\mathbf{H}| \cdot N^2)$. Thus, the overall computational cost of MDT is $O(|V|^L \cdot N + |\mathbf{H}| \cdot N^2)$. However, when more complex matching methods, such as Genetic matching [122], are applied, the computational cost of MDT can be significantly larger. As was previously discussed in Section 2.3.1.2, genetic matching algorithm applies an evolutionary search method in order to find optimal weights for each covariate. In each algorithm iteration, a set of P weights for each one of the variables in set $\mathbf{H}$ is generated.

P corresponds to the *population size* of the genetic algorithm. Then, nearest neighbour matching is applied on the weighted variables of $\mathbf{H}$ for each one of the P weights. A loss function is used to estimate the loss for each of the P resulted sets of matched pairs. If the loss is sufficiently small for any of the P weights, the method terminates; otherwise this process is repeated. A maximum number of iterations I can be used in order to set an upper bound on the computational time of the method. In my experiments, I used as loss function the average standardised mean difference between the treated and control units. The cost of loss estimation for each weights set is $O(|\mathbf{H}| \cdot N)$. Thus, the total computational cost of the matching process is $O(I \cdot P \cdot |\mathbf{H}| \cdot N^2)$. Although the computational cost of MDT could be significantly larger than the cost of CMI, given the availability of advanced computational resources, the computational efficiency can be traded for more reliable results. In my simulated experiments, the running time of CMI was in order of seconds while the running time of MDT was in order of minutes, using a 2.6 GHz quad core CPU and 16 GB RAM.

Finally, I now discuss the assumptions behind my method, thus outlining situations where the method is expected to perform well. There are four key ingredients in my method:

1. I perform independence tests on time series pairs. There is no guarantee that if there were higher-order dependencies among several time-series (e.g. the outcome variable and two confounding variables), they would be detected by the pair-wise tests.

2. The maximum conditioning set of the conditional independence tests that need to be performed is determined by the largest lag L. The value of L is of course upper bounded by the desire to have sample sizes large enough to yield sufficient power to independence tests.

3. The matching procedure assumes that there is an overlap in the confounding variables' values between the groups of treated and control units. If this is not the case, the matching will not achieve sufficient balance.

4. The estimation of the average treatment effect is influenced by the power of the statistical test that is applied.

3.4 Summary

In this chapter, I have presented a novel method for causal inference in time-series data. The proposed method is based on the existing matching framework for i.i.d. data. It provides a methodology for deciding which time-series should be included in the study along with appropriate time lags so that all the necessary conditions for applying the matching framework are satisfied. Then, any of the existing matching methods can be applied and the final conclusion is obtained by examining the average effect of the treatment variable on the the matched treatment and control samples. The main advantages of the proposed method over existing approaches can be summarised as follows:

1. It is non-parametric i.e., it does not require any assumptions about the model class (e.g. linear or non-linear). Thus, it can handle more effectively non-linear cases.

2. It requires fewer conditional independence tests with smaller conditioning sets compared to existing approaches and consequently, it is more effective in high-dimensional datasets.

3. The design of the study is separated from the analysis. The causal link between two variables is examined only when any confounding bias has been sufficiently removed. Thus, the proposed method avoids overfitting.

In order to assess the validity of the proposed method I have conducted an extended evaluation with simulated experiments, in which the *ground truth* is known. In the next chapters, I apply this approach on real datasets and I demonstrate its utility in extracting useful knowledge from human-generated sensor data.

CHAPTER 4

UNDERSTANDING THE IMPACT OF SOCIAL MEDIA ON FINANCIAL MARKETS

In this chapter, I investigate the influence of social media on stock market prices. In the first part, I apply the method presented in Chapter 3 in order to assess the causal impact of social media sentiment on the traded assets of four technological companies. In the second part, I focus on cases characterised by abnormal stock market movements. I propose an event detection method that detects bursty topics on Twitter which are linked with stock market jitters.

4.1 Causal Impact of Twitter Sentiment on Traded Assets

As was previously, many studies so far have focused on using social media data for the prediction of stock market prices. But to what extent do opinions expressed through social media actually have a *causal* influence on stock market? Are stock market prices influenced by the opinions and sentiments that are reported in social media, or is it the case that stock market prices and sentiments are driven only by other (e.g. financial) factors? Would the results have been different if we could manipulate social media data? In order to answer such questions a causality study is required.

I will now discuss the application of the method described in Chapter 3 in order to investigate whether information about specific companies and people reactions, extracted from Twitter data, influence stock market prices. Indeed, Twitter enables us to capture

people opinions about the target companies, the general optimism/pessimism of the public about stock market movements and their reaction to news such as quarterly results announcements or new product launches. Thus, factors related to the company performance and people trust on the company are reflected on Twitter data. My study considers the daily closing prices of four big tech companies based on USA: Apple Inc., Microsoft, Amazon and Yahoo!. I estimate a daily sentiment index for each of these companies by analysing the sentiment of related tweets (the details of this process are presented in Section 4.1.2). My study is based on data gathered for four years, from January 2011 to December 2014. In particular, I examine whether the sentiment of tweets that are posted before stock market closing time influences the closing prices of the target stocks. In order to eliminate any confounding bias, I need to control for factors that may affect both humans sentiments and the target stock prices. Potential influential factors on stocks daily closing prices are their opening prices and their performance during the previous days. Several works have also demonstrated that the performance of other big companies (either local or overseas companies) could influence some stocks (see for example [59, 60]). Foreign currency exchange rates may also cause money flows to overseas markets and consequently influence stocks prices. Finally, commodities prices could affect the earnings of companies and, therefore, their stocks prices. In the next subsections I present the dataset used for this study, the text processing method that was used in order to extract tweets sentiment and the results of the causality analysis. In addition, I conduct a sensitivity analysis, as described in Section 4.1.4 in order to evaluate the sensitivity of the results on missing covariates.

4.1.1 Dataset Description

This study involves the following time-series:

The *response* time-series Y. The difference on the closing prices of the target stocks between two consecutive days. The time-sample t of the time-series corresponds to the closing value of the day t minus the closing value of the previous day.

The *treatment* time-series X. A daily sentiment index that is estimated using tweets related to the target stocks that are posted up to 24 hours before the closing time

of the corresponding stock market. In order to ensure that the values of the treatment variable are driven by information that was available before the closing time of the target stocks, I omit from the study tweets posted up to one hour before the closing time. Thus, the sentiment index of day u is estimated using all the tweets posted from 4:00 p.m. (ET) time (i.e., the NASDAQ closing time) of day $u-1$ to 3:00 p.m. (ET) time of day u. Consequently, our treatment variable captures the people sentiment and reactions to news at any time during the day, up to one hour before the stock market closing time. It should be noted that by excluding tweets posted up to one hour before the stock market closing time, we might exclude important information that comes up just before this time. However, this information will be included in the next day sentiment index. Tweets are filtered using the name of the company and the stock symbol as keywords.

The set of time-series Z. I consider the following time-series which might play a role in this causality study:

1. **The difference between the opening and closing prices of two consecutive days.** This time-series is an indicator of the activity of the target stocks at the start of the trading day.

2. **The stock market prices of several major companies around the world.** In this study I include all the components of the most important stock market indexes such as NASDAQ-100, Dow-30, Nikkei 225, DAX and FTSE. The study could be influenced only by factors that precede temporally both the treatment and effect variables. Thus, I use the difference between the opening and closing prices of two consecutive days for stocks that are traded in the USA exchange markets. The closing time of companies traded in the overseas markets precedes the closing time of the USA stock exchange market, thus the time-series for all the overseas companies stocks correspond to the difference on the closing prices between two consecutive days. Although the values of the treatment variable are driven by tweets that are posted both before and after the corresponding values of the time-series associated to the performance of big companies, for convenience, I consider that the time-sample t of the treatment time-series occurs one hour before the USA stock exchange market closing time on day t. Thus, the time-sample t of any of the time-series that are

used to describe the performance of either a USA-based company or an overseas company temporally precedes the t sample of the treatment time-series.

3. **The daily opening values of foreign currency exchange rates minus the previous day opening values.** I include the exchange rates between Dollar and British Pound, Euro, Australian Dollar, Japanese Yen, Swiss Franc and Chinese Yen.

4. **The difference between the opening values of commodities for consecutive days.** I include the following commodities: gold, silver, copper, gas and oil.

4.1.2 Daily Sentiment Index Estimation

I classify each tweet as negative, neutral or positive using the SentiStrength classifier [149]. SentiStrength estimates the sentiment of a sentence using a list of terms where each term is assigned a weight indicating its positivity or negativity. I updated the list of terms in order to include terms that are commonly used in finance[1]. In total, 39% of the tweets are classified as neutral, 34% as positive and 27% as negative.

Sentiment extraction from text may be inaccurate. Although this issue has been disregarded in previous works [29, 44, 39], here, in order to account for such inaccuracies on sentiment classification, I estimate a probability distribution function of the daily sentiment instead of a single metric. Let us define a set of three objects $S = \{positive, neutral, negative\}$. Each object $i \in S$ denotes a classification category. Let us also define a random variable V_i as follows:

$$V_i = \begin{cases} 0 & \text{if a negative tweet is classified in class i} \\ 1 & \text{if a neutral tweet is classified in class i} \\ 2 & \text{if a positive tweet is classified in class i} \end{cases} \tag{4.1}$$

I derive the probability distribution functions of each random variable V_i, with $i \in S$, based on the classification performance results. I evaluate the performance of the classifier

[1] A list of the words that have been added or modified along with the assigned score is provided at table B.1

	$P(V_i = 0)$	$P(V_i = 1)$	$P(V_i = 2)$
i = positive	0.05	0.27	0.68
i = neutral	0.03	0.91	0.06
i = negative	0.65	0.29	0.06

Table 4.1: Accuracy of the text classification for each classification category (confusion matrix).

by manually classifying 1200 randomly selected tweets (200 tweets for each one of the four examined companies). The probability distribution functions are presented in Table 4.1.

Let us define with N_i the number of tweets posted within a day that are classified in category i. I define a random variable $\mathcal{V}_t$ that corresponds to the sentiment of a day t as follows:

$$\mathcal{V}_t = \sum_{i \in S} N_i \cdot V_i \tag{4.2}$$

Moreover, since 2 is the maximum value of V_i, $\mathcal{V}_t \in \{0, 1, ..., 2 \cdot \sum_i N_i\}$. I estimate the probability distribution of $\mathcal{V}_t$ by deriving the probability-generating function under the assumption that the real sentiment of a tweet is independent of the sentiment of any other tweet conditional to the observed classification of the tweet sentiment (i.e., the inferred sentiment by SentiStrength). Although the sentiment of a tweet may depend on previously posted tweets, given that the probability of correctly inferring the sentiment of a tweet is independent of the sentiment inference of any other tweet, this assumption is realistic. The probability-generating function of $\mathcal{V}_t$ is expressed as follows:

$$G_{\mathcal{V}_t} = \prod_{i \in S} (G_{V_i}(z))^{N_i} = \prod_{i \in S} (\sum_{x=0}^{2} p(V_i = x) \cdot z^x)^{N_i} \tag{4.3}$$

The probability distribution function of $\mathcal{V}_t$ is estimated by taking the derivatives of $G_{\mathcal{V}_t}$. If $\mathcal{N}_t$ the number of tweets posted a day t, then, $\mathcal{V}_t \in \{0, 1, ..., \mathcal{N}_t \cdot M\}$ and the probability that the general sentiment of a day t is positive is given by the probability $P_{pos}(t) = P(\mathcal{V}_t > \frac{\mathcal{N}_t \cdot M}{2})$.

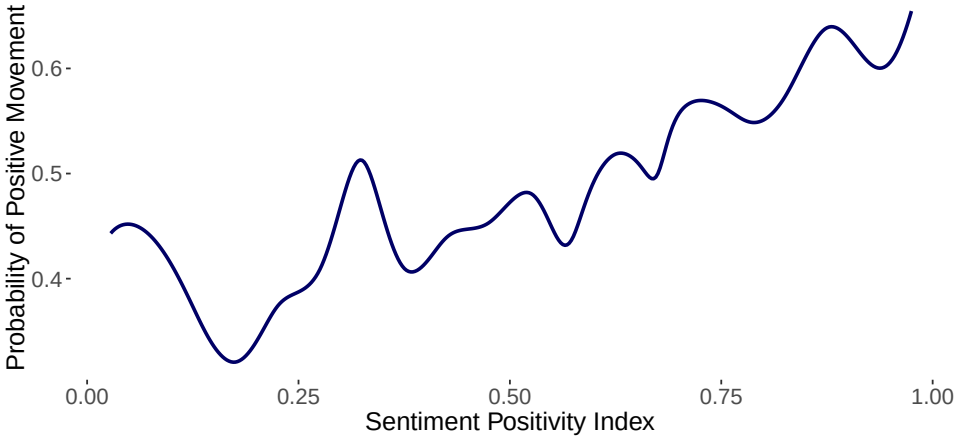

Figure 4.1: Probability distribution function of having a positive movement on the traded assets prices conditional to the sentiment of the tweets.

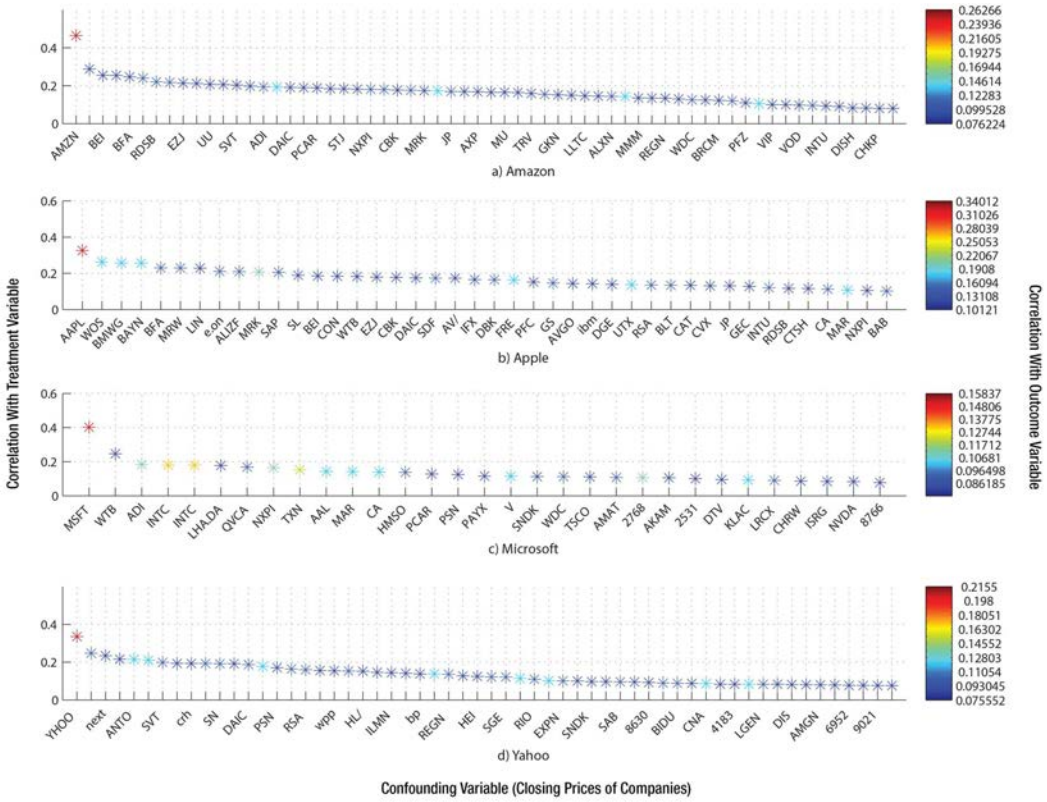

Figure 4.2: Correlation between the confounding variables and the treatment and effect time-series.

4.1.3 Results

I create a binary treatment variable X by applying thresholds on $P_{pos}(t)$. More specifically, a unit t, which describes the t day of the study is considered to be treated (i.e., $X(t) = 1$)

if $P_{pos}(t) \geq P^1_{thresh}$ and untreated (i.e., $X(t) = 0$) if $P_{pos}(t) < P^0_{thresh}$. I conduct my study for three different pairs of thresholds. In detail, I consider a pair of thresholds $T1$, where thresholds P^1_{thresh} and P^0_{thresh} are set to the 50^{th} percentile of X, a pair of thresholds $T2$ where P^1_{thresh} is set to the 60^{th} percentile of X and P^0_{thresh} to the 40^{th} percentile of X and finally a pair $T3$ where P^1_{thresh} and P^0_{thresh} are set to the 70^{th} and 30^{th} percentiles respectively. By increasing the value of P^1_{thresh} and decreasing the value of P^0_{thresh} I eliminate from the study days in which the estimated tweets polarity is uncertain either due to measurement error or because the overall sentiment that is expressed during these days is considered to be neutral. Although discretisation of a continuous variable results in information loss that may jeopardise, in some cases, the reliability of the causal inference, I enhance the validity of my conclusions by considering different threshold values.

I include in this study all the previously mentioned variables. I found that there is no autocorrelation in the time-series, thus, since there is no dependence of our time- series on their past values, I set the maximum lag L, which will be used to create the time-series set $\mathbf{S}$ (see Section3.1), equal to 1 day. For each of the four target stocks, I applied Algorithm 1 in order to find the set of time-series $\mathbf{H}$ that needs to be controlled. I consider a correlation to be statistically significant if the corresponding p-value is smaller than 0.05. I used Spearman's rank correlation in order to capture potentially non-linear relationships among the examined variables. I observed that stock movements are significantly correlated with the sentiment of tweets posted within the same day. These findings are in agreement with results of other studies [29, 54, 39]. I also found that stock prices are independent of past tweets sentiment conditional on more recent tweets. This indicates that any effect of tweets on stock prices is instant rather than long-term. Finally, according to my results, the daily movement of the traded assets for the target companies does not correlate with past days movements. This finding is consistent with the weak-form efficient market hypothesis [36] according to which, it is not feasible to predict stock market movements by applying technical analysis. In Table 4.2 I present the correlation coefficient of the effect variable Y with the treatment variable X and the 1-lagged variables $X^{(1)}$ and $Y^{(1)}$ for each one of the four examined companies. In Figure 4.1 I present the empirical probability distribution function of having a positive movement on the traded assets prices conditional to the sentiment of the tweets $P(Y(t) > 0|X(t))$. The probability distribution

function is estimated using data collectively for the four examined companies. My results indicate that the probability of having a positive movement on the stock market does not increase linearly with the daily tweets positivity index. Stock market movement is quite uncertain when the positivity index of the tweets ranges between 0.35 and 0.65, while the probability of having a positive movement is increasing for positivity index larger than 0.65. Moreover, I notice a relatively high probability of having a positive movement in days with sentiment positivity index lower than 0.1. Considering that daily tweets sentiment captures the current and past stock market trends, this could be attributed to the fact that investors may consider that it is a good time to invest money when assets prices are low; consequently, this could give lead to an increase of stock market prices.

	AAPL	MSFT	AMZN	YHOO
X	0.393	0.155	0.237	0.273
$X^{(1)}$	0.032	0.036	0.012	0.046
$Y^{(1)}$	0.009	-0.003	-0.037	0.031

Table 4.2: Correlation of Y with X, $X^{(1)}$ and $Y^{(1)}$.

Moreover, I found that both the effect and the treatment variables correlate with the most recent stock prices of several local and overseas companies. The daily movements of the target stocks correlate with US dollar exchange rates; however, currency exchange rates do not have any impact on the treatment variable. In Table 4.3 I present the number of variables from each category that will be included in the set **H** for the four target companies and in Figure 4.2, I present the correlation coefficients of the treatment and effect time-series with all variables in set **H**. For all the examined stocks, the strongest confounder is their opening prices.

In order to eliminate the effect of the confounding variables I need to match treated and control units with similar values on their set of confounding variables. I create optimal pairs of treated and untreated units by applying the Genetic Matching algorithm [122]. This is a multivariate matching method that applies an evolutionary search algorithm in order to find optimal matches which minimise a loss function. I use as a loss function the average standardised mean difference between the treated and control units for all the confounding variables $H^i \in \mathbf{H}$ which is defined as follows:

	AAPL	MSFT	AMZN	YHOO
Nasdaq-100 Comp.	6	21	33	7
Nikkei Comp.	1	3	1	13
DAX Comp.	18	2	7	10
FTSE Comp.	10	3	12	26
Dow-30 Comp.	7	3	9	2
FOREX	0	0	0	0
Commodities	0	0	0	0

Table 4.3: Number of variables that are included in the set **H** for each of the four examined companies.

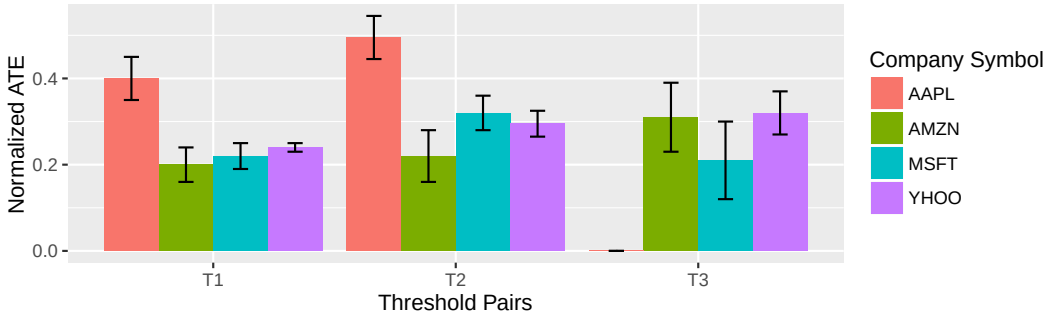

Figure 4.3: Normalized ATE for the three threshold pairs.

$$SMD_H = \sum_{H^i \in \mathbf{H}} \frac{\sum_{(t_u,t_v)\in G} |H^i(t_u) - H^i(t_v)|}{|G| \cdot \sigma_{H^i}} / |\mathbf{H}| \tag{4.4}$$

I check if sufficient balance between treated and untreated subjects has been achieved by analysing the standardised mean difference for each confounding variable. The remaining bias from a confounding variable is considered to be insignificant if the standardised mean difference is smaller than 0.1 [150, 145].

I examine the causal effect of the sentiment of tweets on the target stocks for the three pairs of thresholds. I apply Equation (2.2) in order to estimate the average treatment effect (ATE). Under the assumption that the examined treatment has no impact on the effect variable, the ATE would be equal to 0. I use a t-test to assess how significant is the difference of the observed ATE value from 0. In Figure 4.3, I present the average treatment effect normalised by the variance of the effect variable Y along with the 95% confidence interval values. Confidence intervals are estimated by applying a t-test under the null

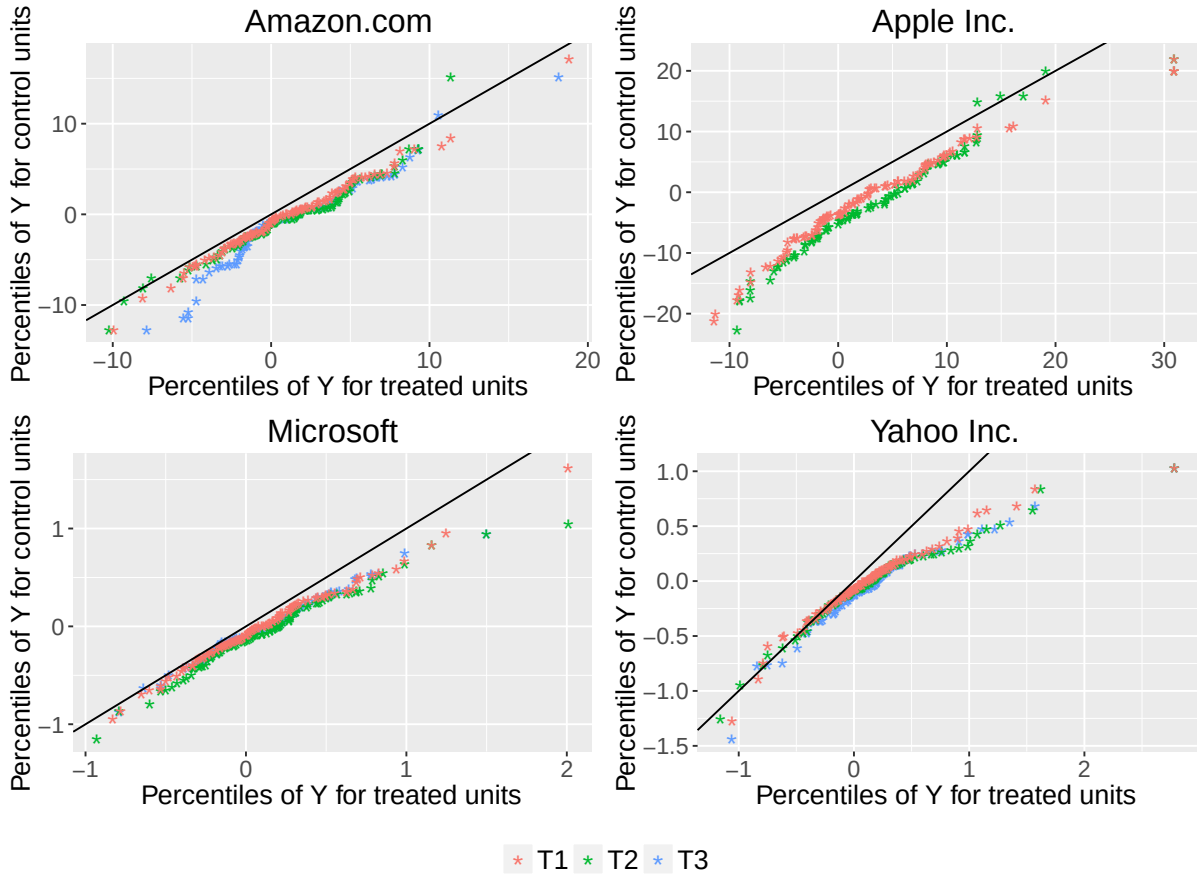

Figure 4.4: Percentiles of treated units versus percentiles of matched control units.

hypothesis that the average effect on the treated units is equal to the average effect on the control units. According to my results, the effect of the tweets sentiment on the stocks prices of all the examined stocks is statistically significant. I also observe that the causal impact is stronger for larger values of the P^1_{thresh} and smaller P^0_{thresh} threshold values, i.e., the observed difference on the effect variable between the treatment and control groups is larger when I consider only days for which there is less uncertainty on the estimated tweets polarity. For Apple, it was not possible to create balanced treated and control groups for the thresholds pair $T3$. This is due to the fact that the opening prices of the AAPL stocks are very strongly correlated with both the effect and treatment variables and, therefore, there were not enough treatment and control units with similar values on their confounding variables. Since causal conclusions are not reliable when the treated and control groups are not balanced, I do not present results for Apple for this pair of threshold values. In addition, I repeat my study for different time-periods using a two-year sliding

window with six-month step. In Figure 4.5 I present the findings for the four examined companies and the first pair of thresholds. According to my results, the difference on the estimated ATE is insignificant for the examined sub-periods. Finally, in Figure 4.4 I compare the distributions of the effect variable Y for the treated and control units by plotting their percentiles against each other. Under the hypothesis that the treatment variable has no effect on variable Y, the curve should be described approximately by $y = x$. However, most of the points of the plot lie below the reference line $y = x$, indicating that the majority of the percentiles of variable Y for the treated units are larger than the corresponding percentiles for the control units.

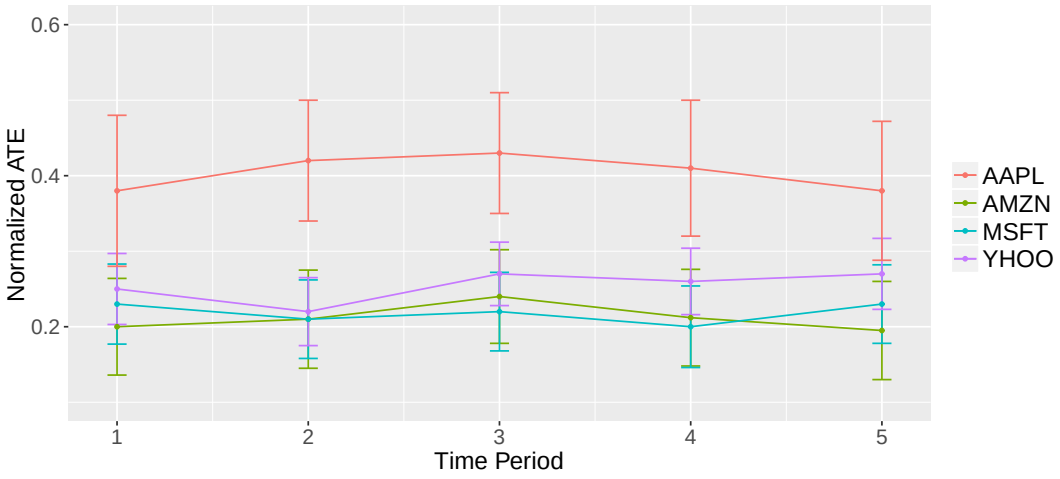

Figure 4.5: Normalised ATE for the threshold pair T1. The analysis is conducted in two-year sub-periods using sliding windows with 6 months step.

4.1.4 Sensitivity Analysis

The main limitation of all non-experimental causality studies is that they are based on the assumption that all confounding variables are known. However, in real scenarios there may be unmeasured factors that influence the assignment of units to treatments. In such cases, the conditional ignorability assumption is violated and consequently, any causal inference result may be biased. In this study, I include a large number of potentially influential factors, such as the performance of other companies traded assets, commodities prices and currency exchange rates. However, there are other factors, such as inflation rates, political changes or economic policy changes that could influence both people sentiment, captured

through Twitter, and traded assets prices. Although such factors may be reflected on the observed confounding variables (e.g. macroeconomic factors such as inflation rates would also affect the prices of other traded assets and consequently, the observed Twitter sentiment may be independent of inflation rates conditional to the performance of other assets included in the study), there may still be some bias due to unobserved factors.

Γ	**Upper bound on *p*-value**			
	AAPL	AMZN	YHOO	MSFT
1.0	0.0000	0.0000	0.0000	0.0000
1.1	0.0000	0.0000	0.0000	0.0000
1.2	0.0000	0.0000	0.0000	0.0005
1.3	0.0000	0.0002	0.0000	0.0027
1.4	0.0001	0.0011	0.0003	0.0109
1.5	0.0003	0.0042	0.0010	0.0327
1.6	0.0009	0.0131	0.0034	**0.0775**
1.7	0.0021	0.0328	0.0091	**0.1519**
1.8	0.0043	**0.0692**	0.0209	**0.2552**
1.9	0.0082	**0.1263**	0.0419	**0.3789**
2.0	0.0142	**0.2050**	**0.0749**	**0.5093**

Table 4.4: Sensitivity Analysis.

In order to evaluate the sensitivity of my results on unobserved confounding variables, I apply Rosenbaum's method [118] described in Section 2.3.1.5. In Table 4.4, I present the results of the sensitivity analysis for $\Gamma \leq 2.0$ and for the $T2$ pair of thresholds. According to my results, the causal influence of Twitter on Apple stock prices would be considered statistically significant (with p-value 0.014) even if some days were twice more likely (i.e., $\Gamma = 2$) to have positive sentiment conditional to the observed confounding variables due to unmeasured factors. Similarly, for Amazon the causal inference results are statistically significant (i.e., p-value < 0.05) for $\Gamma \leq 1.9$, for Yahoo! for $\Gamma \leq 1.7$ and, finally, for Microsoft my conclusion would be invalid for $\Gamma \geq 1.6$.

4.2 Linking Twitter Events with Stock Market Jitters

In the previous section, I have shown that there is strong evidence that there is causal influence of social media on stock market data. Thus, social media may contain useful information for the prediction of events of interest. Given this finding, in this section I examine whether social media data can be used for the early detection of stock market jitters. Identifying factors that could cause big movements on stock prices is very important for financial risk analysis and prediction [151, 152]. As was previously discussed, several studies have shown that collective sentiment extracted from tweets can be linked to traded assets prices and can be used to improve the prediction of stock market movements. However, the exploitation of social media for early prediction of stock market jitters is an interesting, yet unexplored research topic.

I propose a novel framework for detecting financial events on Twitter that impact a specific stock market. Detecting bursts on tweets that correspond to real-world events has been previously discussed in many studies [67, 68, 69]. However, my work substantially differs from these studies, since my objective is to identify Twitter events that are related to large fluctuations on stock market. The proposed financial event detector (FED) monitors the arrival rates of individual words in a stream of tweets related to finance or politics and records an event when an unusual burst is detected. For each event I create a feature vector containing information such as the number and type of words with *unusual* increase on their arrival rates, the volume and the polarity of the related tweets as well as geographical characteristics of the tweets and information about their authors. Then, I exploit stock market data in order to train a classifier to recognise events that influence stock market as *positive* and events with minimum or no impact as *negative*. Thus, my method is trained to detect financial or political events that cause fluctuations on a specific stock market. My method does not require any manual events labelling. Instead, I create a training set by labelling as *positive* event vectors that co-occur with large movements on the examined stock market and as *negative* all the other vectors. The classifier is updated dynamically: after a new event (from the remaining set) is classified, the true event label is learned by examining its impact on the stock market and the classifier is re-trained

accordingly.

Typically, social media information is modelled as one-dimensional time-series, describing the evolution of a feature, such as tweets polarity or volume, over time. However, I show here that using a single feature to model Twitter data results in wasting important information. For example, negative criticism about an educational or health system reform could be a popular topic within a country and result in bursts of tweets with negative sentiment; nevertheless, it will probably have minimum or no impact on the stock market. On the other hand, news related to financial or political instability would probably be commented by a larger number of different people and may have more global interest. Consequently, features such as the number and the profile of different users discussing a topic, the geographical characteristics of tweets (i.e., whether the topic of interest is discussed locally or globally) as well as the individual *bursty* words associated with each event may contain important information. Thus, instead of arbitrarily selecting a single feature for representing Twitter information, I apply feature selection in order to find an optimal subset of features that can be used to identify vectors reflecting important information about the examined stock market. I support my argument by comparing the proposed FED method with a modified version that uses single-feature vectors, which contain only information about events sentiment.

I apply the proposed framework on the detection of events that influence the Greek and Spanish stock markets for the period 03/2015 to 10/2015. I selected these two markets since they were strongly influenced by the European crisis and they experienced high volatility during the examined period. Although stock market jitters constantly occur in many stock markets, the specific study requires the analysis of stock markets in which a large number of strong movements is observed in a relatively small time period. I use intraday 5-minutes returns of the ATHEX and IDEX stock market index. Moreover, I apply a general-purpose state-of-the-art event detector on the Twitter dataset and I demonstrate that such approaches fail to recognise events which influence stock market. In the next sections, I provide a detailed description of the proposed method and I present my results.

4.2.1 Financial Event Detector

In this section I describe the components of the financial event detector (FED). The event detection process is comprised by the following steps:

1. **Bursty Words Detection.** The arrival rate of each word in a stream of tweets is estimated and a set of *bursty* words is extracted.

2. **Events Feature Vectors Extraction.** The busty words as well as information extracted from tweets containing these bursty words (such as information related to tweets polarity, geographical distribution and users characteristics) are used to create feature vectors that represent events.

3. **Events Filtering.** All the detected Twitter events are not necessarily related to stock market jitters. I use stock market data to train a classifier to recognise which event feature vectors do have an impact on the stock market. The financial event detector is trained for a specific stock market. The initial labeled training set is created by utilising solely stock market data, without the need of manual labelling, and the classifier is updated dynamically.

4.2.1.1 Bursty Words Detection

In order to detect bursty topics on a stream of tweets I apply a feature-based event detection method, according to which the arrival rate of each word/feature contained in each tweet is modeled as an inhomogeneous Poisson process. Let me denote by N_w the number of occurrences of each word w in the collection of tweets. I estimate the arrival rate $\lambda_w(t)$ of word w as follows:

$$\lambda_w(t) = \sum_{i=1}^{N} f_\Delta(t - t_i) \tag{4.5}$$

where t_i the time that the i^{th} tweet containing the word w was posted and f_Δ a Gaussian kernel of bandwidth Δ. I characterize a word w as *bursty* during a specific time interval by applying thresholds both on the rate of the word $\lambda_w(t)$ and on the slope $\lambda'_w(t)$ of its rate. In detail, a word w which was not *bursty* at time $t-1$ will be *bursty* at time t

if $\lambda_w(t) > T_R$ and $\lambda'_w(t) > T_S$, for some threshold values $T_R, T_S > 0$. A word w *bursty* at time t will not be *bursty* at time $t' > t$ if $\lambda_w(t') < T_R$. Hence, I examine only the rate of the word in order to change its status from *bursty* to *normal* since, even if the acceleration of the rate of a previously characterized *bursty* word is low, the word should still be considered as *bursty* if it has a sufficiently high arrival rate. The rate of each word is re-computed dynamically every time that a tweet containing that word is posted.

By applying a threshold on the word rates, I detect words with significant *popularity* (i.e., high rate) within a time-period, while by applying a threshold on the rate slope I avoid considering as *bursty* words which are *popular* most of the time. I use the same thresholds for all the words instead of creating word-specific thresholds based on historical rates. In this way, I not only reduce the number of free parameters but also avoid considering as *bursty* words which occur a relatively small number of times during a period while they previously had zero or very low rate.

4.2.1.2 Events Feature Vectors Extraction

An event on Twitter at time t will exist if there is at least one bursty word at time t. I represent events as time-dependent feature vectors (a detailed description of the features is presented in Subsection 4.2.1.3). Only one event feature vector can be active during any time t. Hence, if more than one Twitter events co-occur, they will all be associated with the same feature vector and all bursty words describing these events will comprise different features of the vector. An event starts when a bursty word is detected and it is updated dynamically every time a *significant* change on its characteristics occurs. Thus, multiple feature vectors may be created for the same event. In order to avoid unnecessary overhead, I do not update an event every time an insignificant change occurs in any of the feature values. Instead, I consider that a significant change on the over-all arrival rate of all the words that are associated with the event indicates a noteworthy change on the volume and characteristics of tweets and therefore, the event features need to be re-estimated. I consider an increase in the overall arrival rate of words to be significant if it is at least 10% of the previous value. Thus, let W the set of bursty words associated with an event, a new feature vector for the same event that had its last update at time t_1 will be created at time t_2 if:

$$\sum_{w \in W} \lambda_w(t_2) > 1.1 \cdot \sum_{w \in W} \lambda_w(t_1) \tag{4.6}$$

The feature vector that corresponds to the most recent event update is created so that it represents the *strongest* version of the event. More formally, denote by: $E_{t_s}(t)$ the event started at time t_s and updated at time t, $f_i(t)$ the i^{th} feature of the feature vector $E_{t_s}(t)$, N the number of features, T_u the set of the timestamps at which the event E_{t_s} has been updated and $t_l \in T_u$ the last time that the event has been updated. Then, the feature vector $E_{t_s}(t_l)$ is estimated as follows:

$$E_{t_s}(t_l) = \{\max_{t \in T_u} f_1(t), \max_{t \in T_u} f_2(t), ... \max_{t \in T_u} f_N(t)\} \tag{4.7}$$

The rational behind this idea is that events need to be detected as early as possible, thus the conditions for creating an event should be relatively 'soft'. However, an initially *weak* event may become stronger later and consequently the initial feature vector will not fully represent the strength of the event. On the other hand, if an event occurs when the stock market is closed, its strength may decrease by the time the stock market opens. Nevertheless, a reaction of the stock market to the news is still expected. Thus, I decided to update an event only when its *strength* has increased. If an event is active for more than 24 hours it is updated in order to remove any obsolete information. Since typically, the news circulation is daily, the information contained in an event needs to be re-examined after one day. This may result in discarding terms that are not bursty any more, reducing the 'strength' of the event if the interest of users is gradually fading out, or it may leave the event intact if there is a continuous interest on it.

The process of event detection and update is summarised in Figure 4.6. In detail, I check for new tweets every Δt seconds and I update the rate functions of all the words that are contained in the new tweets. Then, the current set of bursty words is updated accordingly (i.e., words which are not bursty any more are removed from the set and new words which are bursty at the current time are added). While the set of bursty words is empty, I check for new tweets every Δt seconds until at least one word contained in the tweets becomes bursty. Then, if there is no active event (i.e., the set of bursty words was empty Δt seconds before) I create a new event (i.e., a new feature vector $E_t(t)$, where t

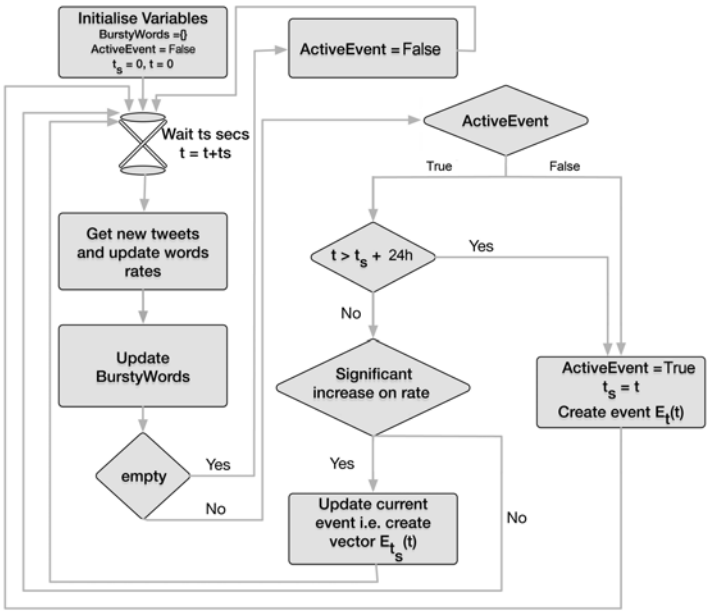

Figure 4.6: Event Detection Process.

the current time). If there is an active event and 24 hours have elapsed since its start time, the current event is inactivated and a new event is created; otherwise, I create a new feature vector, as described in equation (4.7), if the condition of equation (4.6) holds.

4.2.1.3 Features Description

The feature vectors include information about the bursty words that are associated with the event, the tweets polarity and geographical distribution and the reputation and popularity of tweets authors. Instead of using a separate feature for each bursty word, I create categories of words that usually refer to the same subject by estimating the correlation between the words rates. Words with highly correlated rates (i.e., similar arrival patterns) may refer to the same subject [67]. I group words by performing hierarchical clustering, where the 'distance' between two words w_1, w_2 with correlation coefficient c_{w_1,w_2} is equal to $1-c_{w_1,w_2}$. The event $E_{t_s}(t)$ started at time t_s will be described by the following features at time t:

- the *maximum number of bursty words* of each category of words that have been bursty from the start of the event t_s until time t. I denote by $W_i(t)$ the set of bursty

words of the i^{th} category from time t_s until time t and with $|W_i(t)|$ the number of elements in the set. Thus, there is one feature $|W_i(t)|$ for each category i of words.

- the *maximum bursty word rate in each word category i:* $R_i(t) = \max_{w \in W_i(t) t_s \leq t' \leq t} \max \lambda_w(t')$
- the *maximum bursty word rate:* $R(t) = \max_{i \leq N} R_i(t)$.
- the *maximum bursty word rate slope:* $S(t) = \max_{i \leq N} \max_{w \in W_i(t) t_s \leq t' \leq t} \max \frac{d}{dt'} \lambda_w(t')$.
- the *number of verified users* $V(t)$ that have posted a tweet which contains at least one bursty word from the start of the event until time t, normalised by the number of tweets [1]. According to Twitter, verified accounts are highly sought users in interest areas including government, politics, journalism, business, etc., and thus include authenticated accounts of the key players in major political and economic events.
- the *average number of followers* $F_{AVG}(t)$ of users who have posted a tweet containing at least one bursty word. The number of followers of a Twitter user is an indication of the impact his/her tweets may have.
- the *maximum number of followers* $F_{MAX}(t)$ between all the users who have posted a tweet that contains at least one bursty word.
- the *average geographical distance from the examined stock market location* D_{AVG} of users who have posted a tweet associated with the event.
- the *weighted average distance from the examined stock market location* D_{W_AVG}: calculated as D_{AVG}, but this time each user is weighted by the corresponding proportion of the followers, i.e., the number of her followers normalised by the total number of all users followers.
- the *location dispersion* $L(t)$ of the users who have posted a tweet associated with the event. The location dispersion is an indication of whether the topic is discussed mainly locally or whether there is a general interest for the event globally. It is calculated using the coefficient of variation (i.e., the ratio of the standard deviation to the mean) of the distance from the event centre, among all event tweets.

[1] Note that a verified user will be counted for as many tweets as she will post in the event time interval.

- the *sentiment strength index* $SSI(t)$. I use SentiStrength [149] in order to estimate the positivity and negativity index of the tweets that are associated with the event, and calculate the tweet sentiment strength index as the sum of these two indices. The event sentiment strength index is calculated as the absolute value of the average sentiment strength index among all tweets related to that event . SentiStrength has been optimised to detect finance-related sentiment by following the procedure described in [149]. In detail, I have trained SentiStrength with finance-related terms that have been manually assigned with a polarity weight. I used the list of positive and negative finance-related terms that is described in [153]. Each positive term is assigned the maximum SentiStrength weight and each negative term the minimum.

- the *weighted sentiment strength index* $SSI_W(t)$, which is the average sentiment strength index among all tweets of the event, weighted by the number of followers of each tweet author (as in D_{W_AVG}).

Thus, if N the number of word categories, an event will be characterised by a feature vector of $2 \cdot N + 10$ features: $E_{t_s}(t) = \{|W_1(t)|, R_1(t), |W_2(t)|, R_2(t), ... |W_N(t)|, R_N(t), R(t), S(t), V(t), F_{AVG}(t), F_{MAX}(t), D_{AVG}, D_{W\_AVG}, L(t), SSI(t), SSI_W(t)\}$. All features are normalised to zero mean and unit variance. As I will describe in the following section, feature selection will be applied in order to distinguish the features which are actually important for the detection of events that influence the examined stock market.

4.2.1.4 Event Filtering

The last step of the event detection process is filtering out events that do not have any impact on the examined stock market. In order to classify the events to *positive* (i.e., events that have an impact on a stock market) and *negative* (i.e., events that do not influence stock market), I create a labelled training set using data related to a specific stock market. Thus, the classifier is trained to recognise events that influence a specific stock market. In detail, I first define a set of time-slots $\mathcal{T}_{true}$ during which an *unusual* movement on stock market volatility is noticed and a set of time-slots $\mathcal{T}_{false}$ during which the volatility is considered *normal*. In my analysis, I use the historic volatility which is estimated over 1-day windows and corresponds to the variance of the logarithmic returns.

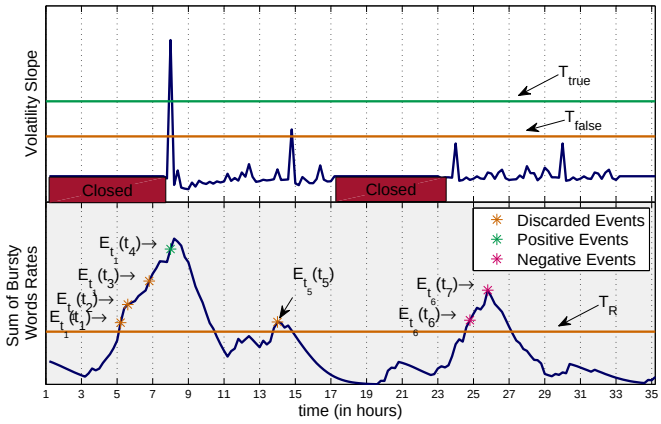

Figure 4.7: Events Labelling Example.

Let us denote the volatility of stock market at time s with $\mathcal{V}(s)$ and the volatility slope with $\mathcal{V}'(s)$. Then, I set $s \in \mathcal{T}_{true}$ if $\mathcal{V}'(s) > T_{true}$ and $s \in \mathcal{T}_{false}$ if $\mathcal{V}'(s) < T_{false}$, where $T_{true} > 0$ and $T_{false} > 0$ are thresholds on the stock market volatility slope. I set $T_{true} > T_{false}$ in order to allow for a *neutral* zone and separate high volatility time-slots belonging to $\mathcal{T}_{true}$ from the normal volatility ones belonging to $\mathcal{T}_{false}$. I also define refer to the set of time-slots that belong to the neutral zone as $\mathcal{T}_{neutral}$.

Afterwords, I need to examine which event feature vectors co-occur with unusual movements on stock market volatility. Since a stock market is open only at specific hours, some events on Twitter may occur when the stock market is closed. However, the effect of these events (if any) will be visible only when the stock market opens. In order to match the time at which an event occurred or was updated with the time that its impact was visible in the stock market, I transform the update time t of each feature vector to the first time $t' \geq t$ that the stock market is open. If multiple update times of the same event match with the same stock market time t', I keep only the most recent feature vector and discard all the previous ones; the rational behind this decision is that the most recent feature vector represents the event on its full strength. If multiple events (i.e., events with different start times) match with the same stock market time, I keep all events. Finally, an event that started at time t_s, matched to the stock market time t', will be assigned a label $C_{t_s}(t')$ as follows:

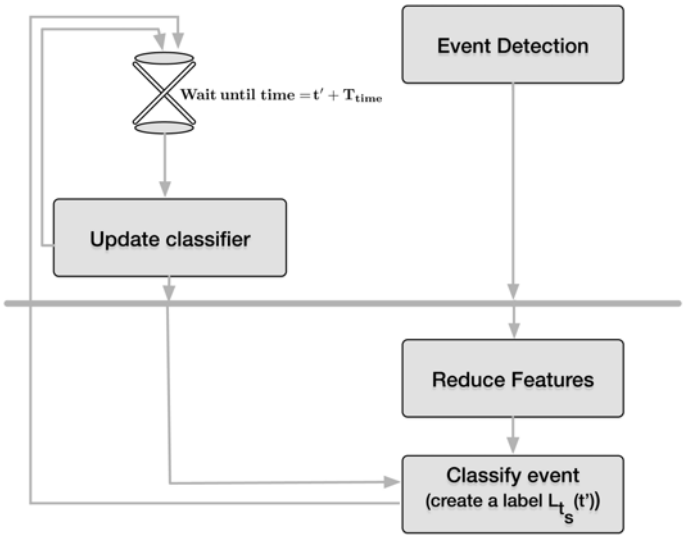

Figure 4.8: Event Classification Process.

$$C_{t_s}(t') = \begin{cases} 1 & \text{if } \min_{s \in \mathcal{T}_{true}} |s - t'| < T_{time} \\ -1 & \text{if } \min_{s \in \mathcal{T}_{neutral}} |s - t'| < T_{time} \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

where T_{time} a threshold that denotes the maximum time distance that a Twitter event may have from a stock market event, value 1 is used to label *positive* events and value 0 *negative*. Event feature vectors with label -1 will be discarded from the training set. In Figure 4.7 I present an events labelling example. The top graph depicts the volatility slope and the bottom one the sum of the bursty words rates. In this example, I set $T_{time} = 1$ hour. Event E_{t_1} is detected when stock market is closed and it is updated at times t_2, t_3 and t_4. I discard the first three event vectors and I set the label of the most recent event update (i.e., $E_{t_1}(t_4)$) equal to 1. Event E_{t_5} will also be discarded since, according to the graph, there is a time sample $s \in \mathcal{T}_{neutral}$ with less than one hour difference from the event detection time t_5. Finally, the event vectors $E_{t_6}(t_6)$ and $E_{t_6}(t_7)$ do not co-occur with any stock market jitter so they will be labeled as negative.

I create a labeled training set by using a subset of the available data. I reduce the feature set by applying a feature selection process on the training set, thus keeping only features useful for distinguishing *positive* events from *negative* events. Finally, I use these reduced-dimensionality feature vectors to train a classifier.

The event filtering is performed dynamically. This process is presented in Figure 4.8. Each time a new feature vector is created (by the process described in Figure 4.6), I firstly reduce the vector dimension by discarding insignificant features, I then match it to the time t' of the next stock market sample and, finally, I use the previously trained classifier to assign a label $L_{t_s}(t')$ to the event.

I am able to determine the actual label $C_{t_s}(t')$ (Equation (4.8)) of the event only at time $t' + T_{time}$. After the true label $C_{t_s}(t')$ becomes available, I dynamically update the classifier based on the estimated and true class labels $L_{t_s}(t')$ and $C_{t_s}(t')$, respectively.

4.2.2 Application to Stock Markets

I apply the proposed framework to verify whether there is a link between detected events in social media (in this case Twitter) and events (large fluctuations) on the Greek and Spanish stock markets. In particular, I estimated the historical volatility of the ATHEX and IBEX stock market index by using 5-minutes intraday data for the period 01/03/2015 to 01/11/2015[1]. I also downloaded Twitter data by tracking terms related to the European financial crisis. In order to select appropriate terms, I applied the RAKE keyword extraction method [154] on the European debt crisis Wikipedia webpage[2]. I set the maximum number of words per keyword equal to 2 and the minimum number of occurrences in the document equal to 4. 171 keywords were extracted in total. The Greek and Spanish Twitter datasets will include only tweets which contain the terms *Greece* or *Greek* and *Spain* or *Spanish* respectively. I use the data of the first 4 months to find the optimal parameters for the classifier and the remaining data for the performance evaluation of FED.

4.2.2.1 Thresholds Selection

One of the key design goals is to minimise the use of thresholds, and in any case to understand and quantify the impact of the choice of different values on the performance

[1]Note that the Greek dataset corresponds to a period of around 7 months given that Greek stock market was closed during July

[2]https://en.wikipedia.org/wiki/European_debt_crisis

of the proposed approach. Thus, in this section I discuss several criteria for the selection of threshold values.

	v = 0.5		**v = 1**		**v = 2**		**v = 4**	
	GR	ES	GR	ES	GR	ES	GR	ES
Missed Events	3	4	3	5	4	6	11	10
False Positives	28	24	27	24	27	23	26	22

Table 4.5: Sensitivity on the selection of the T_R and T_S threshold values.

Thresholds related to Words Arrival Rate, $\mathbf{T_R}, \mathbf{T_S}$. The values of these thresholds regulate how often an event is created or updated based on the process described in Figure 4.6. Large threshold values would result in missing events or in delayed event detection. On the other hand, smaller values would result in more *false positive* events. However, as described in Section 4.2.1.4, a classifier will be used to filter out false positives. I conducted an empirical evaluation in order to assess the sensitivity of the results on the selection of these thresholds. In detail, I set

$$\mathcal{T}_{\mathcal{R}} = \max_w \langle \lambda_w(t) \rangle_t \quad \text{and} \quad \mathcal{T}_{\mathcal{S}} = \max_w \langle \lambda'_w(t) \rangle_t,$$

where $\langle \lambda_w(t) \rangle_t$ and $\langle \lambda_w(t)' \rangle_t$ denote the average values of the functions $\lambda_w(t)$, $\lambda'_w(t)$, corresponding to the word w. The calculation of the maximum value is over all words in the training data. I observe that for words with an unusual increase in their rate, this mostly happens over short time intervals - most other words have just a constant and very low arrival rate. Hence the resulting thresholds $\mathcal{T}_{\mathcal{R}}, \mathcal{T}_{\mathcal{S}}$ are very small. I examine the number of true events missed and the number of false positive events for $\{T_R, T_S\} = v \cdot \{\mathcal{T}_{\mathcal{R}}, \mathcal{T}_{\mathcal{S}}\}$, where $v \in \{0.5, 1, 2, 4\}$. I present the results in Table A.1. My findings indicate that there is a very small variation on the number of false positive events when I increase the values of the T_R and T_S, while the number of true stock market events that are missed increases significantly. In general, 'weak' Twitter events can be easily spotted during the classification process and consequently the number of false positive events is not strongly influenced by the selection of these thresholds. Thus, I set relatively small values on T_R and T_S. In particular, I set $T_R = \mathcal{T}_{\mathcal{R}}$ and $T_S = \mathcal{T}_{\mathcal{S}}$.

Thresholds on Volatility Slope, $\mathbf{T_{true}}, \mathbf{T_{false}}$. The thresholds have to be chosen according to the specific application requirements, i.e., the "intensity" of the event under

Feature	Ranking	
	GR	**ES**
maximum rate value $R(t)$	1	1
maximum slope value $S(t)$	2	2
maximum number of followers $F_{MAX}(t)$	3	4
weighted average distance from stock market location D_{W_AVG}	4	3
weighted sentiment strength index $SSI_W(t)$	5	6
location dispersion $L(t)$	-	5

Table 4.6: Selected non-word features

consideration. I examine the performance of FED for three different threshold values. In detail, I set $T_{true} = \{2 \cdot \langle \mathcal{V}'(s) \rangle_s, 2.5 \cdot \langle \mathcal{V}'(s) \rangle_s, 3 \cdot \langle \mathcal{V}'(s) \rangle_s\}$ and $T_{false} = 0.8\ T_{true}$.

Threshold $\mathbf{T_{time}}$. This threshold denotes the maximum time difference between a Twitter event detection (or update) and a stock market jitter. As mentioned in Section 4.2.1.4, if the stock market is closed at the time of the event detection, the event detection is formally shifted to the closest opening time of the stock market. In my study I set T_{time} equal to 1 hour. The selection of this time threshold is supported by previous works that have shown that traders usually react promptly on news releases [155, 65]. Also, in [65] authors found strong influence between financial tweets and stock market movements on 1-hour intervals suggesting that public traders need more time to evaluate the news.

4.2.2.2 Feature Selection

I use the training data to cluster words into categories, as described in Subsection 4.2.1.3. I apply hierarchical clustering with cut-off distance equal to 0.7, leading to 58 and 79 word categories for the Greek and Spanish datasets respectively. In order to select an optimal cut-off distance, I estimate the silhouette score [156] for different cut-off distances, ranging from 0.4 to 0.9. The silhouette score describes the average distance of the points of each cluster to the points of the neighbouring clusters. For both datasets, the maximum silhouette score is achieved for cut-off distance equal to 0.7. The resulted silhouette scores for both datasets are presented in Figure 4.9.

Only 34 and 41 of these categories contained at least one bursty word respectively for the two datasets. Thus, since I create two features per word category (i.e. number of

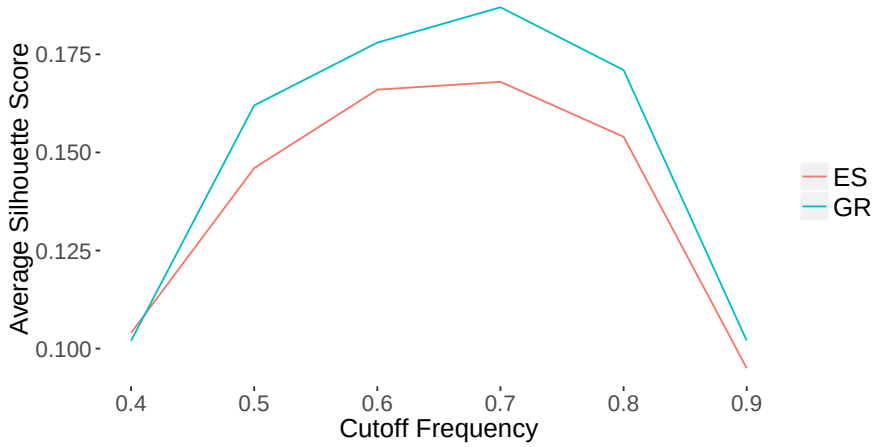

Figure 4.9: Average silhouette score for hierarchical clustering with different cut-off distances.

bursty words in the category and maximum arrival rate among all words in the category), the total number of word-related features for the Greek dataset is 68 and for the Spanish 82. As described in 4.2.1.3, I also use 10 additional features. To select only features deemed important for distinguishing between *positive* and *negative* events I apply two feature selection methods implemented in Weka [157], namely a correlation-based feature selection algorithm [158] and an information gain based feature selection method [159]. Overall, 5 non-word features were selected for the Greek dataset and 6 for the Spanish with both algorithms. The selected attributes along with their ranking based on the correlation-based feature selection are presented in Table 4.6 (the results with the information gain based algorithm are very similar and thus they are omitted). Finally, word features related to 7 and 6 word categories were selected, respectively for the two datasets. In Table 4.7 I present the stemmed words of the selected categories along with the selected features per category.

In this case study, given that the time-period that is considered is relatively short, the feature selection process is conducted one time and the same features will be used for the whole study. In general, we cannot assume that important features will remain stable over long time periods. Thus, the feature selection process might need to be repeated.

	Stemmed Words	Features	
		$\mathbf{R_i(t)}$	$\|\mathbf{W_i(t)}\|$
1	energy, sovereign, pipelin, sanct	Yes	Yes
2	send, reform, troik	Yes	No
3	money, fear, imf, stock, deb, default, pay, repay	Yes	Yes
4	progress, dijsselbloem, finmin, varoufak, min, eurogroup	Yes	Yes
5	press, europ, program, fin, govt, bank, deal, stat, nee, bailout, cris, ecb, let, syriz, country, credit, support, eurozon, meet, grexit, econom, polit, euro, greek, agr, talk	Yes	Yes
6	bahrain, eu, leav	No	Yes
7	bil, rep, loan, germ, germany	No	Yes

(a) Greece

	Stemmed Words	Features	
		$\mathbf{R_i(t)}$	$\|\mathbf{W_i(t)}\|$
1	crit, infl, tax, issu	Yes	Yes
2	govern, ban, black, prep, tsipra, money, england, impact, throughout	Yes	Yes
3	mean, chin, ev, europ	Yes	Yes
4	neg, greec, inform, comp, demand, spend	Yes	Yes
5	reform	Yes	No
6	form, elect, perc, ecb, contribut, podemo, rajoy	Yes	Yes

(b) Spain

Table 4.7: Selected Word-related Features.

4.2.2.3 Evaluation

I apply the methodology discussed in Section 4.2.1.4 to train a support vector machine classifier in an online manner to classify the future Twitter events into *positive* and *negative*. It has been noticed that there is not a clear separation between *positive* and *negative* Twitter events. In such cases, the margin supported by the SVM classifier is essential in order to achieve good performance. The classifier is first trained on an initial data segment (training set), using 10-fold cross-validation to select the kernel type (linear, Gaussian and polynomial), regularisation parameter and loss parameters (to deal with unbalanced class problem - more negative events than positive ones). Polynomial kernels were selected

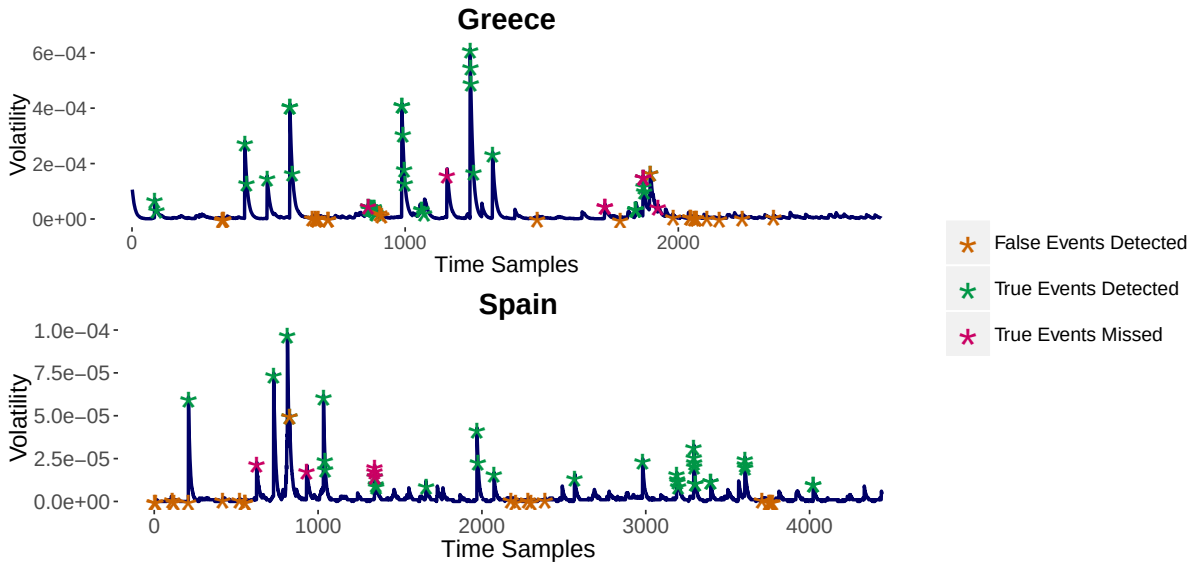

Figure 4.10: Event Detection Results.

for both datasets (order 3 for the Greek dataset and 2 for the Spanish). After that, the classifier is dynamically updated on the remaining data (keeping the hyper-parameters and kernel type fixed) as described in Section 4.2.1.4. All the reported results are based on predictions on unseen Twitter events from this remaining data. Overall 375 Twitter events were detected on the Greek dataset and 349 on the Spanish one.

I estimate the precision, recall and F1 score of the event detection method by comparing the real label $C_{t_s}(t')$ with the predicted label $L_{t_s}(t')$ for each Twitter event. I create two binary streams C and L with all the real and predicted labels of Twitter events, respectively. Since there may be stock market events without any matching Twitter event, I update C and L as follows:

$\forall t' \in \mathcal{T}_{true}$ *and* $t' \notin \mathcal{U}$, *where* $\mathcal{U}$ *is the set of Twitter event times, create a new label* $C_{t'}(t') = 1$ *and* $L_{t'}(t') = 0$.

In Table 4.8 I present the classifier performance for the three different T_{true} thresholds. For both the examined datasets, FED performs better for larger values on T_{true} i.e., when it is trained to detect 'stronger' stock market fluctuations. I also observe slightly improved performance on the Greek dataset. This could be justified considering that during the examined period, Greece was affected by a remarkable financial instability that resulted on several stock market jitters.

In Figure 4.10 I present historical volatility of the ATHEX and IBEX stock market

$\mathbf{T_{true}}$	**Precision**		**Recall**		**F1**	
	GR	ES	GR	ES	GR	ES
$2 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.61	0.52	0.73	0.79	0.66	0.63
$2.5 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.64	0.62	0.84	0.69	0.72	0.66
$3 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.69	0.65	0.81	0.72	0.74	0.68

Table 4.8: Classification Performance

index, for the months after the training period. I also show the correctly and falsely detected events, as well as the missed events for $T_{true} = 3 \cdot \langle \mathcal{V}'(s) \rangle_s$. According to my results, the proposed mechanism successfully detects most of the stock market jitters purely based on Twitter data. Interestingly, although not specifically trained to do so, all detected stock market events were predicted as positive on Twitter before they appeared on the stock market. Finally, there are some Twitter events falsely classified as *positive*. These misclassifications usually occur in bursts. This can be explained by the fact that my approach allows for multiple updated versions of the same event; if one feature vector is misclassified, its subsequent updated versions will be probably misclassified too. One hour after the first misclassified vector occurs, the classifier is updated with the new sample, and consequently avoids repeating the same mistake on any similar subsequent feature vectors.

4.2.2.4 Comparison With Baseline Event Detectors

I compare the performance of my approach with a) a state-of-the-art general-purpose event detector and b) a sentiment-based event detector. For events detected when the stock market is closed, I apply the process used in FED, i.e., such events will be shifted to the opening time of the stock market. If this time is more than 24 hours ahead of the event time (during the weekend), the event will be discarded. If more than one Twitter events are matched to the same stock market time, I keep only the 'stronger' event. The strength of an event is defined based on the event detection method described in the remaining of this section.

General-purpose Event Detector. Although several methods for bursts detection on social media data have been proposed, to the best of my knowledge, this is the first

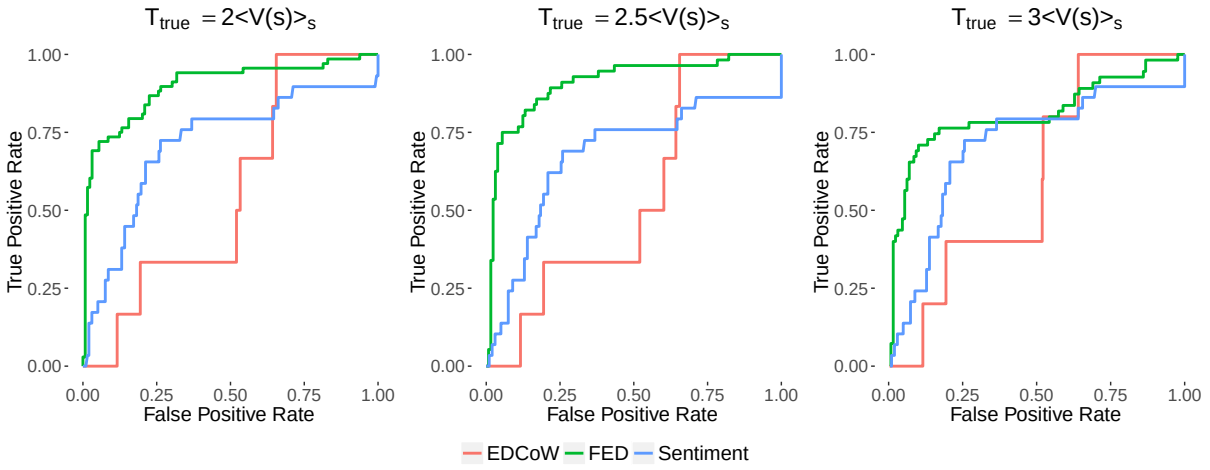

Figure 4.11: ROC Curves.

work that attempts to identify events that influence a specific stock market. The most similar approach to FED is EDCoW. Both FED and EDCoW monitor changes on the arrival rates of individual words in order to trigger the detection of an event and they group bursty words based on the correlations between their arrival patterns. However, in contrast to FED, which uses word groups to construct Twitter event features, EDCoW creates a separate event for each word group. Finally, for each event, EDCoW estimates a value ϵ representing the 'strength' of the event based on the number of its words as well as the correlations among them and filters-out non-significant events (i.e., events with low ϵ value). I perform event detection in 2-hour windows. Similarly to this approach, I assign a label $C(t_w)$ to each event $E(t_w)$ detected during a window started at time t_w as follows:

$$C(t_w) = \begin{cases} 1 & \text{if } \exists s \in \mathcal{T}_{true}, t_w \leq s \leq t_w + 2h \\ -1 & \text{if } \exists s \in \mathcal{T}_{neutral}, t_w \leq s \leq t_w + 2h \\ 0 & \text{otherwise} \end{cases} \tag{4.9}$$

The total number of true positive and false positive events is given by the number of 1- and 0-labels in C, respectively. I also estimate the number of false negative by counting the the number of stock market jitters for which there was no event detected. In Table 4.9 I present the performance of EDCoW for the three T_{true} thresholds and three different values on the γ parameter of EDCoW that is used to define when the correlation between two words (or the autocorrelation of one word) is significant. These results correspond to the optimal threshold on ϵ value, which is used to filter-out non-significant events (i.e., the

γ	$\mathbf{T_{true}}$	**Precision**		**Recall**		**F1**	
		GR	ES	GR	ES	GR	ES
10	$2 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.35	0.39	0.22	0.27	0.27	0.32
	$2.5 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.31	0.33	0.21	0.22	0.25	0.26
	$3 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.31	0.35	0.22	0.22	0.26	0.27
20	$2 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.21	0.24	0.33	0.28	0.25	0.26
	$2.5 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.18	0.22	0.30	0.30	0.23	0.25
	$3 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.18	0.23	0.31	0.32	0.23	0.27
40	$2 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.19	0.25	0.60	0.47	0.29	0.33
	$2.5 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.17	0.22	0.56	0.49	0.26	0.30
	$3 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.17	0.23	0.54	0.49	0.26	0.31

Table 4.9: EDCoW Event Detection.

threshold for which I achieved the highest F1 score). Note, that such an evaluation favors EDCoW method over FED, as the performance estimates will be positively biased. In spite of that, the EDCoW performs poorly for all the examined γ values. This indicates that it is not feasible to detect Twitter events that influence the stock market solely by searching for bursts in the Twitter stream.

Sentiment-based Event Detector. Given that most studies on the influence of social media on the stock market only examine the impact of text sentiment, I compare FED with a sentiment-based event detector. A direct comparison with existing methods is not feasible, since, to the best of my knowledge, their purpose is either to prove a dependency between social media and stock market or to predict future values rather than the detection of jitters. Thus, I adjust FED in order to use only information about tweets sentiment. In detail, I estimate the weighted sentiment strength index $SSI_W(t)$, described in Section 4.2.1.3, using 2-hour sliding windows with 5 minutes step size. I then apply an event detection method similar to the proposed FED approach: I create an event at time t if $\langle SSI_W(t) \rangle_t \leq SSI_W(t)$ and I update the event when there is a 10% increase in its sentiment value. I label events as *positive* or *negative* by applying Equation (4.8) and I train a Support Vector Machine classifier in order to predict the events' classes. In Table 4.10 I present the precision, recall and F1 score of the sentiment-based event detector for the three different T_{true} thresholds. My results indicate that event classification based solely on sentiment performs poorly, since it is not possible to distinguish between events

$\mathbf{T_{true}}$	**Precision**		**Recall**		**F1**	
	GR	ES	GR	ES	GR	ES
$2 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.23	0.21	0.79	0.74	0.37	0.34
$2.5 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.22	0.23	0.76	0.77	0.34	0.35
$3 \cdot \langle \mathcal{V}'(s) \rangle_s$	0.23	0.23	0.79	0.76	0.36	0.35

Table 4.10: Sentiment-based Event Detection.

of negative sentiment that influences stock market (e.g., fears of political instability or bankruptcy) and those that do not (e.g., negative opinions/gossips about politicians).

	$\mathbf{2 \cdot \langle V'(s) \rangle_s}$		$\mathbf{2.5 \cdot \langle V'(s) \rangle_s}$		$\mathbf{3 \cdot \langle V'(s) \rangle_s}$	
	GR	ES	GR	ES	GR	ES
EDCoW	0.0695	0.0724	0.000476	0.000623	0.000736	0.001282
Sentiment	0.1634	0.1921	0.000933	0.001648	0.004825	0.007225

Table 4.11: P-values of the DeLong test under the null hypothesis that the AUC of the FED approach is equal to the AUC obtained with the EDCoW and Sentiment-based method.

Finally, in Figure 4.11 I present the receiver-operating-curves (ROC) for FED, EDCoW and the sentiment-based event detector (with $\gamma = 40$) for the three T_{true} thresholds. The ROC curves are created by applying sequential event evaluation. For the EDCoW method they are created by varying the threshold on ϵ value used in filtering-out non-significant events. In addition, I applied the DeLong method in order to assess whether there is a statistically significant difference between the ROC curves of the examined methods. The p-values under the null hypothesis that there is no difference between the area under the ROC curve (AUC) of the FED approach and the other two examined methods are presented in Table 4.11.

4.2.2.5 Mutual Information Analysis

In this Section I use mutual information to examine the dependence between the Twitter events and the stock market jitters. I represent events in the stock market using the binary stream C of real event classes and Twitter events with the binary stream L of predicted Twitter event classes. The binary streams are not i.i.d.. The probability of a stock market jitter will normally be higher when strong fluctuations have been previously observed and

lower in more 'stable' periods. Thus, I model the binary streams C, L as Markov chains by applying the Causal State Splitting Reconstruction (CSSR) algorithm [160]. CSSR creates a Markov model which best represents the underlying probabilistic model of the streams. The resulted model is a two-state Markov chain (i.e., the probability of having an event labeled as *positive/negative* at time t depends only on the event label at time $t-1$). I denote the probabilities of state i for C and L with π_i^C and π_i^L respectively and the transition probabilities from state j to state i with $p_{i|j}^C$ and $p_{i|j}^L$. Then, the entropy rates of C, L are estimated as follows [161]:

$$H(C) = -\sum_{i=0}^{1} \pi_i^C \cdot \sum_{j=0}^{1} p_{j|i}^C \log p_{j|i}^C$$

$$H(L) = -\sum_{i=0}^{1} \pi_i^L \cdot \sum_{j=0}^{1} p_{j|i}^L \log p_{j|i}^L$$

I measure the reduction of uncertainty about C during a time unit t if I utilise knowledge about L during t by measuring the mutual information rate $MIR(C, L)$ [162] given by the following equation:

$$MIR(C, L) = H(C) + H(L) - H(C, L) \tag{4.10}$$

where $H(C, L)$ denote the joint Shannon entropy of C, L estimated as:

$$H(C, L) = -\sum_{i=0}^{1}\sum_{j=0}^{1} \pi_{i,j}^{C,L} \cdot \sum_{k=0}^{1}\sum_{l=0}^{1} p_{k,l|i,j}^{C,L} \log p_{k,l|i,j}^{C,L} \tag{4.11}$$

where $\pi_{i,j}^{C,L}$ the joint state probability of C and L for states i, j, respectively and $p_{k,l|i,j}^{C,L}$ the joint transition probability of C and L from states i to k and j to l, respectively.

In Figure 4.12 I present the mutual information rate between C and L, when L is estimated by applying a) the proposed FED method, b) the EDCoW method and c) the sentiment-based event detector, for the three different T_{true} threshold values. The estimation of MIR is based only on unseen Twitter events (i.e., I do not use the training set). My results indicate significant dependence between stock market jitters and events detected by the FED approach and much weaker dependence when sentiment-based or

EDCoW event detection is applied.

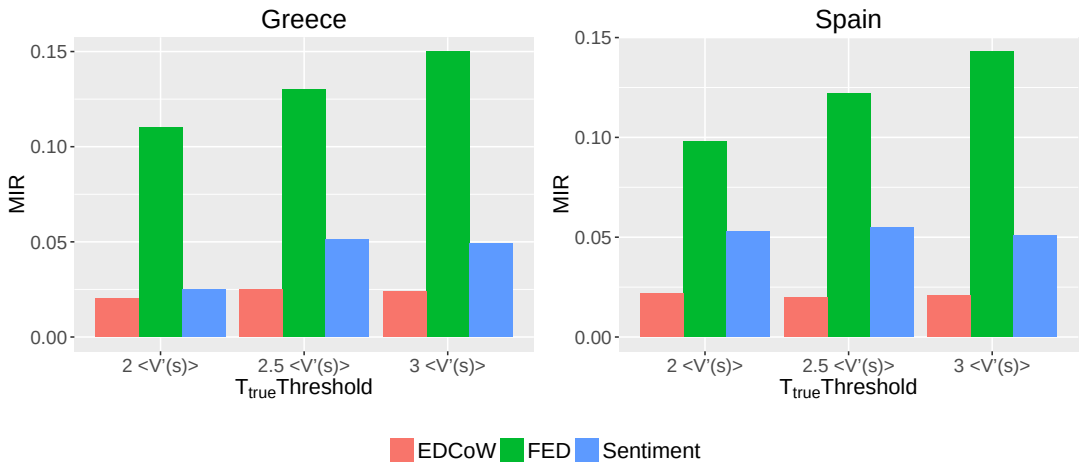

Figure 4.12: Mutual Information Rate between real events and predicted events

Since the sample size is relatively small, I need to examine whether the estimated mutual information is spurious. In order to do this, I shuffle the data 1000 times and I estimate the mutual information between the shuffled time-series. In Table 4.12, I present the 99th percentile of the resulted distribution of the 1000 mutual information values. Based on these results, the 99th percentiles are significantly smaller than the mutual information between our event time-series and the time-series of the stock market jitters.

$2 \cdot \langle \mathcal{V}'(s) \rangle_s$	$2.5 \cdot \langle \mathcal{V}'(s) \rangle_s$	$3 \cdot \langle \mathcal{V}'(s) \rangle_s$
0.0252	0.0256	0.0212

Table 4.12: 99th percentile of the distribution of the mutual information values for the shuffled time-series.

4.3 Summary

In this chapter, I have studied the impact of social media on stock market prices. In the first part of this work, I examine the causal impact of tweets polarity on the traded assets of four companies. A large number of factors, such as commodities prices and prices of other traded assets, has been included in the study. The study is based on observational data rather than experimental procedures. Indeed, causality studies that are based on observational data rather than experimental procedures could be biased in

case of missing confounding variables. However, conducting experimental studies is not feasible in most cases. In this work I have minimised the risk of biased conclusions due to unmeasured confounding variables by including a large number of factors in the study. Additionally, I have conducted an analysis on the sensitivity of my conclusions on missing confounding variables. I have estimated a sentiment index indicating the probability that the general sentiment of a day, based on tweets posted for a target company, is positive. My results show that Twitter data polarity does indeed have a causal impact on the stock market prices of the examined companies. It should be noted that, since all the examined companies belong to the technological sector, my findings cannot be directly generalised for any company. Nevertheless, I believe that social media data could represent a valuable source of information for understanding the dynamics of stock market movements.

On the basis of this conclusion, in the second part of this chapter, I have examined whether social media data can be used for the detection of stock market jitters. I have proposed FED (Financial Event Detector), a novel event detection method which focuses on early detection of events in Twitter that influence a specific stock market. I have modelled Twitter data as multi-dimensional feature vectors by utilising a rich variety of information. I have applied feature selection in order to find which of these features are important for distinguishing between events that influence stock market and insignificant events. I have demonstrated that using only information about tweets sentiment is not adequate for the detection of stock market jitters. I have trained a classifier, solely by utilising stock market data, to recognise which of the detected events will cause strong fluctuations on the examined stock market. I apply this method to the Greek and Spanish stock market and I demonstrate that FED achieves up to 74.32% F1 score. Moreover, I show that general-purpose event detectors fail to recognise events that influence stock market. I have also show the association between strong stock market fluctuations and the detected Twitter events by estimating the mutual information between these two variables.

Finally, this study has been based on news related to the European financial crisis and on two particular stock markets which were strongly influenced by these events. Although these findings provide evidence that information extracted from Twitter could be utilised in order to better understand and detect early factors that influence stock market, the

extrapolation of these findings in more financial markets requires additional work. In addition, the proposed method is based on many free parameters. In detail, researchers need to specify the cut-off frequency for the words clustering step, the thresholds on the words arrival rate, the maximum time difference between a Twitter event and the associated stock market event and the threshold that defines when a change on an event should be considered significant. However, I have assessed the performance of the method with different thresholds and I have justified the selection of some threshold values by conducting additional analysis.

CHAPTER 5

UNDERSTANDING HUMAN BEHAVIOUR USING SMARTPHONE SENSOR DATA

In this chapter, I discuss the benefits of leveraging digital devices in order to continuously and unobtrusively collect data that would facilitate studies on human behaviour. The purpose of this study is to propose a generic causal inference framework for the analysis of human behaviour using digital traces. More specifically, I demonstrate the potential of automatically processing human generated observational digital data in order to conduct causal inference studies based on quasi-experimental techniques. I support this claim by presenting an analysis of the causal effects of daily activities, such as exercising, socialising or working, on stress based on data gathered by smartphones from 48 students that were involved in the StudentLife project [35] at Dartmouth College for a period of 10 weeks. It is also worth noting that although previous studies have provided evidence of peer influence on individuals mood ([82, 80]), the information about the social network of participants is not sufficient to examine the impact of such factors on stress level. The main goal of the StudentLife project is the study of the mental health, academic performance and behavioural trends of this group of students using mobile phones sensor data. To the best of my knowledge, this is the first observational causality study using digital data gathered by smartphones.

Information about participants' daily social interactions as well as their exercise and work/study schedule is not directly measured; instead, I use raw GPS and accelerometer traces in order to infer high-level information that is considered as implicit indicator of the variables of interest.

No active participation of the users is required, i.e., answering to pop-up question-

naires. I automatically assign semantics to locations in order to group them in four categories: home, work/university, socialisation venues and gym/sports centre. By grouping locations into these four categories and continuously monitoring the spatio-temporal traces of users, I can derive high-level information as follows:

- **Work/University.** By analysing the daily time that users spend at their workplace I can infer their working schedule. Prolonged sojourn time at work/university could be considered as an indicator of increased workload.

- **Home.** The time that participants spend at home could serve as a rough indicator of their social interactions. Prolonged sojourn time at home could imply limited social interactions or social interactions with a restricted number of people. In general, spending time outside home usually involves some social interaction. An estimation of the total daily time that participants spend at any place apart from their home and working environment could serve as a rough indicator of their non-work-related social interactions.

- **Socialisation Venues.** By monitoring users visits at socialisation venues, such as pubs, bars, restaurants etc., I can infer the time that they spend relaxing and socialising outside home during a day.

- **Gym/Sports-centre.** Indoor workout can be captured by tracking participants' visits to gyms or sports centres. Outdoor activity can be measured using accelerometer data.

In the following sections, I initially present a general methodology for causal inference based on observational sensor data. Afterwards, I apply the proposed framework to the StduentsLife dataset in order to understand the impact of daily activities on participants stress level. Finally, I discuss the limitations of the proposed approach and I summarise my findings.

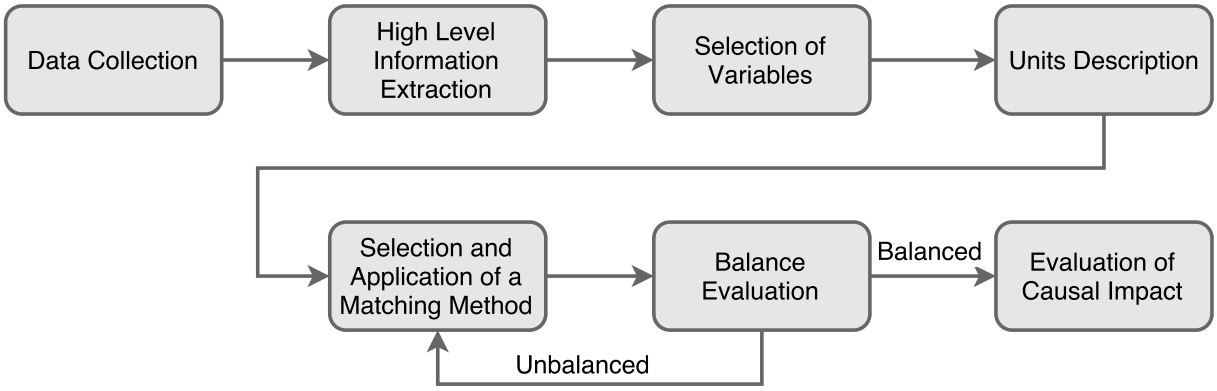

Figure 5.1: Causal Inference Methodology.

5.1 Methodology Description

In this section, I describe the process of causal inference based on observational sensor data. The process is summarised at Figure 5.1. In the following paragraphs each step of this process is analysed.

Data Collection. Smartphones are equipped with a wide range of sensors and are able to capture a rich variety of information. Accelerometer and GPS tracking sensors are able to track our position and movement; communication logs describing our communication through phone calls, SMS messages, emails or through social networks can be captured; photos and video/audio recordings also comprise an important information source. In addition, wearable devices, equipped with biopotential, chemical and stretch and pressure sensors, are able to sense physical and chemical properties of users' bodies and have enabled innovative applications in the domains of health, wellness and fitness.

high-level Information Extraction. Raw sensor data need to be processed in order to extract higher level information. For example, several researchers have used location traces in order to extract significant places, location context (e.g., home, work, restaurant etc.) and high-level activities (e.g., work, sleep, leisure etc.) [163, 90, 164, 91]. In [165] authors use a rich variety of information captured by smartphones, including GPS traces, audio signal and photos, in order to extract location context. Automatically understanding location context as well as the underlying high-level activity would enable a detailed and unobtrusive monitoring of daily human activities and could facilitate many sociological studies. For example, in a study about the influence of exercise on mental health,

labelled location data would enable researchers to infer participants physical activity by examining their location history (i.e., visits to gyms, courts or other similar places). In addition, accelerometer data can be used for activities recognition such as sitting, walking, running [95, 96, 97, 166] or even more complex activities such as eating and sleeping [98] and wearable devices data can be used for emotions recognition [92, 93, 94].

Although the advances on sensors technology have enabled the inference of high-level information, inference methods suffer from limitations and inaccuracies. Inaccurate inference of the information of interest may result in inducing bias in the study. This issue is extensively discussed in Chapter 6. Due to the limitations of inference methods but also due to the lack of suitable sensors for specific types of data, researchers may need to use pop-up questionnaires prompting the user to provide necessary information along with sensor data.

Selection of Variables. After extracting high-level information from the available data, researchers need to define the variables of the causality study. In many cases, the variables of interest are not directly measured; instead, other factors can be used as indicators of the missing variables. For example, in a study measuring the impact of work schedule on participants stress level, the exact work schedule may not be available; instead the time that participants spend at a location labeled as 'work' can be used as indicator of the work schedule. As was previously discussed in Chapters 2 and 3, researchers need to select suitable variables that represent the treatment and the outcome variables of the study as well as the variables that may influence both the treatment and the outcome values.

Units Description. In a causality study based on human-generated sensor data, normally there are multiple time-series for each participant in the study i.e., for each participant there may be a time-series describing his/her location during the period that the data were collected, a time-series describing his/her activities etc.. An object/unit of the study represents the 'state' of a participant during a specific time-frame (i.e., his/her location, activity, emotional state etc.). Thus, there are multiple objects for each participant. Since the 'state' of a participant may depend on his/her past 'state', the objects of the study are not realisations of i.i.d variables and consequently, a traditional matching approach for causal inference, as discussed in Chapter 2 cannot be applied.

Symbol	Description
t_i	Time-period
o	Participant identifier
D	The set of days during which the dataset was collected
$Y^o_{t_i}$	The time-series that describes the outcome values of participant o during the t_i time-periods sampled daily ($Y^o_{t_i} = \{Y^o_{t_i}(d) : d \in D\}$)
$X^o_{t_i}$	The time-series that describes the treatment values of participant o during the t_i time-periods sampled daily ($X^o_{t_i} = \{X^o_{t_i}(d) : d \in D\}$)
$\mathbf{Z^o_{t_i}}$	A set of time-series describing other characteristics relevant for the study for participant o during the t_i time-periods sampled daily ($\mathbf{Z^o_{t_i}} = \{\mathbf{Z^o_{t_i}}(d) : d \in D\}$)
L	The maximum time-lag
$Y^{o,(l)}_{t_i}$	The l-lagged version of time-series $Y^o_{t_i}$
$X^{o,(l)}_{t_i}$	The l-lagged version of time-series $X^o_{t_i}$
$\mathbf{Z^{o,(l)}_{t_i}}$	The l-lagged version of time-series $\mathbf{Z^o_{t_i}}$
$\mathbf{S^o_{t_i}}$	A set of all the time-series for participant o and time-intervals t_i: $\mathbf{S^o_{t_i}} = \{Y^o_{t_i}, ...Y^{o,(L)}_{t_i}, X^o_{t_i}, ...X^{o,(L)}_{t_i}, \mathbf{Z^o_{t_i}}..., \mathbf{Z^{o,(L)}_{t_i}}\}$
$Y^{(l)}$	A set of all the l-lagged version of time-series $Y^o_{t_i}$: $Y = \{Y^o_{t_i}, \forall t_i, o\}$
$Y^{(l)}$	A set of all the l-lagged version of time-series $X^o_{t_i}$: $X = \{X^o_{t_i}, \forall t_i, o\}$
$\mathbf{Z^{(l)}}$	A set of all the l-lagged version of time-series $\mathbf{Z^o_{t_i}}$: $\mathbf{Z} = \{\mathbf{Z^o_{t_i}}, \forall t_i, o\}$
$\mathbf{S}$	A set of all the time-series: $\mathbf{S} = \{\mathbf{Y}, ...\mathbf{Y^{(L)}}, \mathbf{X}, ...\mathbf{X^{(L)}}, \mathbf{Z}..., \mathbf{Z^{(L)}}\}$

Table 5.1: Notation.

In addition, user behaviour and activities' pattern are strongly influenced by the time of the day. For example, a user may work on a normal morning shift and usually has some physical exercise during the evening. He/she may also tend to be more stressed during the working hours (i.e morning) and more relaxed during the evening. A study measuring the causal impact of exercise on participants stress level would be biased if it was based on comparisons between evening time-samples, during which a participant had some exercise, and morning time-samples during which there was no physical exercise. Thus, the time-period is an important confounding variable of the study. In order to avoid any bias induced by matching units describing participants' 'state' at different time-periods of the day (i.e., morning, afternoon etc.), I assign a time-period identifier in each unit. Only units with the same time-period identifier can be matched (i.e., compared). Thus, each day should be split in smaller periods (e.g., early morning, morning, afternoon, etc.). The exact number of time-periods per day is a parameter that should be defined by the

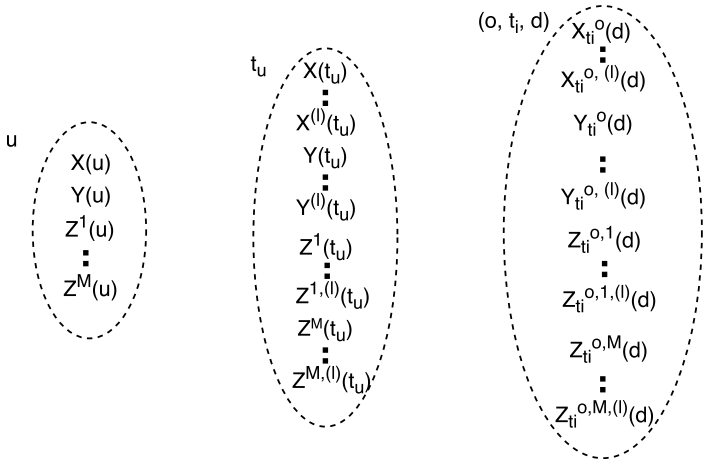

Figure 5.2: Graphical representation of units. On the left side, o represents a unit on a traditional causality study, characterised by its treatment value $X(u)$, its response value $Y(u)$ and M other characteristics $Z^1(u)$, $Z^2(u)$..., $Z^M(u)$. On the middle, t_u represents a unit on the time-series matching design framework introduced in Chapter 3. On the right side, (o, t_i, d) represents a unit on a causality study with smartphones data, based on the proposed framework. It should be noted that u on the left side denotes a unit in the study, which corresponds to a participant. o at the right side, also denotes a participant in the study; however, in this case a unit is not defined only by the participant.

researcher based on the data availability and the type of the study.

The outcome variable of participant o during the t_i time-periods of the study is described by the time-series $Y^o_{t_i} = \{Y^o_{t_i}(d) : d \in D\}$, with D a set of the days of the study. Similarly the time-series $X^o_{t_i} = \{X^o_{t_i}(d) : d \in D\}$ represents the treatment time-series for participant o during the t_i time-periods and $\mathbf{Z^o_{t_i}} = \{\mathbf{Z^o_{t_i}}(d) : d \in D\}$ represents a set of time-series describing other characteristics relevant for the study. Moreover, I define the time-series sets $Y = \{Y^o_{t_i} : \forall t_i, o\}$, $X = \{X^o_{t_i} : \forall t_i, o\}$ and $\mathbf{Z} = \{\mathbf{Z^o_{t_i}} : \forall t_i, o\}$. I use the notation introduced in Chapter 3 in order to describe the lagged versions of the time-series. In detail, I denote by $Y^{o,(l)}_{t_i}$, $X^{o,(l)}_{t_i}$ and $\mathbf{Z^{o,(l)}_{t_i}}$ the l-lagged versions of the time series $Y^o_{t_i}$, $X^o_{t_i}$ and $\mathbf{Z^o_{t_i}}$, respectively (i.e., if $X^o_{t_i}(d)$ is the d-th sample of $X^o_{t_i}$, then $X^{o,(l)}_{t_i}(d) = X^{o,(l)}_{t_i}(d-l)$).

Following the approach introduced in Chapter 3, I define a maximum time-lag L and a set of time-series $\mathbf{S^o_{t_i}} = \{Y^o_{t_i}, ...Y^{o,(L)}_{t_i}, X^o_{t_i}, ...X^{o,(L)}_{t_i}, \mathbf{Z^o_{t_i}}..., \mathbf{Z^{o,(L)}_{t_i}}\}$ for each participant o and each different time-period (i.e., morning time-period, afternoon time-period etc.) t_i. Then, a unit of the study describes the 'state' of a participant o during the t_i time-period of day d and corresponds to the d-th time-sample of the set of time-series $\mathbf{S^o_{t_i}}$. In Figure 5.2, I provide a graphical representation of the notion of unit in comparison with the unit

representation in a traditional causality study and the unit representation based on the time-series framework introduced in Chapter 3.

Algorithm 2 Defining the set of confounding variables.

Input:
The set of time-periods $\mathbf{T}$
The set of participants $\mathbf{P}$
The sets of time-series $\mathbf{S} = \{\mathbf{Y}, ...\mathbf{Y^{(L)}}, \mathbf{X}, ...\mathbf{X^{(L)}}, \mathbf{Z}..., \mathbf{Z^{(L)}}\}$
Output: The set of confounding variables $\mathbf{H}$

```
H := {}
for i=l to L do
  {For all zero-lagged sets of time-series}
  for all S^(0) ∈ S do
    {Find the lagged versions of S which are also parents of Q.}
    B := (S^(0), ..., S^(l−1)) ∩ P
    if (IsIndependent(S^(l), X, B))and (IsIndependent(S^(l), Y, B)) then
      H := H ∪ S^(l)
    end if
  end for
end for
```

```
{This procedure returns TRUE if R is not independent of Q conditional to set B}
IsIndependent(R, Q, B)
pval := {}
for all p ∈ P do
  for all t_i ∈ T do
    Examine the null hypothesis that Q^o_{t_i} ⫫ R^o_{t_i} | B^o_{t_i}
    pval := pval ∪ (p-value of the independence test)
  end for
end for
Combine p-values on the pval set using Fisher's method
if (null hypothesis is rejected) then
  return TRUE
end if
return FALSE
```

Selection and Application of a Matching Method. After defining the units and the variables of the study, the causal impact of a factor X on a factor Y can be assessed by applying the matching design framework. However, as it was previously mentioned, the objects of the study are not realisations of i.i.d. variables and, therefore, the traditional matching framework cannot be applied. In addition, in this case, each variable of the study (e.g., the stress level of the participants) is described by multiple time-series, i.e., one time-series per participant and per time-period. Thus, the matching design for time-

series data framework presented in Chapter 3 is not directly applicable to this scenario. In order to overcome this problem, the process described in Section 3.1 will be applied for each time-series set $\mathbf{S^o_{t_i}}$. In detail, the independence tests need to be applied separately for each set of time-series $\mathbf{S^o_{t_i}}$ and the resulted p-values of the tests are combined using Fisher's method [167] in order to find the final conditioning set $\mathbf{H}$. This process is described by Algorithm 2. Afterwards, the conditioning set $\mathbf{H}$ is updated as described in Section 3.1 in order satisfy the *Stable Unit Treatment Value Assumption* and the *i.i.d. Assumption*.

Then, each treated unit (o, t_i, d) (i.e., $X^o_{t_i}(d) = 1$, assuming a binary treatment variable) needs to be matched with a control unit, which has *similar* values on the variables of the conditioning set $\mathbf{H}$, as described in Section 3.1. The exact matching method that will be applied will be selected based on the dataset characteristics. As was previously mentioned, a treatment unit (o, t_i, d) can be matched only with a control unit (o', t_i, d') (i.e., units must refer to the same time-period t_i). Moreover, researchers may choose to match only units that refer to the same participant (i.e., $o = o'$). However, in studies based on relatively small datasets, this may not be feasible since the statistical power of the test may be significantly reduced.

Balance Evaluation. After applying a matching method, the balance at each variable of the conditioning set $\mathbf{H}$ is evaluated as described in Section 2.3.1.3. If sufficient balance has not been achieved, the matching method is revised and the process is repeated.

Evaluation of Causal Impact. Finally, when sufficient balance has been achieved, the causal impact of X on Y is evaluated using equation (2.2) as described in Section 2.3.1.

5.2 Impact of Daily Activities on Humans Stress Level

5.2.1 Dataset Description

The StudentLife dataset contains a rich variety of information that was captured either through smartphone sensors or through pop-up questionnaires. In this study I use only GPS location traces, accelerometer data, a calendar with the deadlines for the modules that students attend during the term and students responses to questionnaires about their

Algorithm 3 Location clustering

Input: Set of location points $L = \{l_1, l_2, ..., l_n\}$
Output: Set of Clusters $C = \{c_1, c_2, ..., c_m\}$

```
C := {}
for each l ∈ L do
  if accuracy(l)>50 then
    continue
  end if
  locationClusteredFlag := 0
  for each c ∈ C do
    H := {Z^{j,k} : Z^{j,k} ∈ P}
    if distance(l, centroid(c))<50 then
      c := c ∪ {l}
      locationClusteredFlag := 1
      break
    end if
  end for
  if locationClusteredFlag = 0 then
    newCluster := {l}
    C := C ∪ {newCluster}
  end if
end for
```

stress level. Students answer these questionnaires one or more times per day.

I use the location traces of the users to create location clusters. GPS traces are provided either through GPS or through WiFi or cellular networks. For each location cluster, I assign one of the following labels: *home*, *work/university*, *gym/sports-centre*, *socialisation venue* and *other*. Labels are assigned automatically without the need for user intervention. In order to increase the quality of the location estimation, I consider only GPS samples with less than 50 meters accuracy. Moreover, I ignore any sample that was collected while the user was moving. For each new GPS point, I create a cluster only if the distance of this point from the centroid of any of the other existing clusters is more than 50 meters. Otherwise, I update the corresponding cluster with the new GPS sample. Every time a new GPS sample is added to a cluster, the centroid of the cluster is also updated. The pseudo code of the location clustering algorithm is presented in Algorithm 3.

Each location cluster is labeled as *home*, *work/university*, *gym/sports-centre*, *socialisation venue* or *other*. The label socialisation venue is used to describe places like pubs,

bars, restaurants and cafeterias. The label other is used to describe any place that does not belong to the above mentioned categories. I label as *home* the place where people spend most of the night and early morning hours (i.e., the most significant place from 24:00 to 06:00). In order to find clusters that correspond to *gyms/sports-centres* or *socialisation venues* I use the Google Maps JavaScript API [168]. The Google Maps JavaScript API enables developers to search for specific types of places that are close to a GPS point. The type of place is specified using keywords from a list of keywords provided by this API. I use the centroid of each unlabelled cluster to search for nearby places of interest. Places that correspond to *gym/sports-centres* are specified by the keyword *gym* and places that correspond to *socialisation venues* are specified by the keywords *bar*, *cafe*, *movie theatre*, *night club* and *restaurant*. For each unlabelled cluster I conduct a search for nearby points of interest. If a point of interest with distance less than 50 meters from the cluster centroid is found, I label the cluster as *gym/sport-centre* or *socialisation venue* depending on the point of interest type. Otherwise the cluster is labeled as *other*. Any place within the university campus that is not labeled as *gym/sport-centre* or *socialisation venue* is labeled as *work/university*. In Figure 5.3 I present the percentage of time that students spend on average in each of the five location categories that were mentioned above. According to Figure 5.3 students spend the majority of the their time at *home* and at *university*. During night and early morning hours, the location of around 60% of the samples has been labeled as *home* while the majority of samples from 9:00 to 20:00 are labeled as *university*.

I use information extracted from both accelerometer data and location traces to infer whether participants had any exercise (either at the gym or outdoors). The StudentLife dataset does not contain raw accelerometer data. Instead it provides sequences of activities classified by continuously sampling and processing accelerometer data. The activities are classified to *stationary*, *walking*, *running* and *unknown*.

I also use the calendar with students' deadlines, which is provided by the StudentLife dataset, as an additional indicator of students workload. I define the set of all days that the student o has a deadline as $\mathcal{D}^o_{deadline}$. I define a variable $D^o(d)$ that represents how many deadlines are close to the day d for a user o as follows:

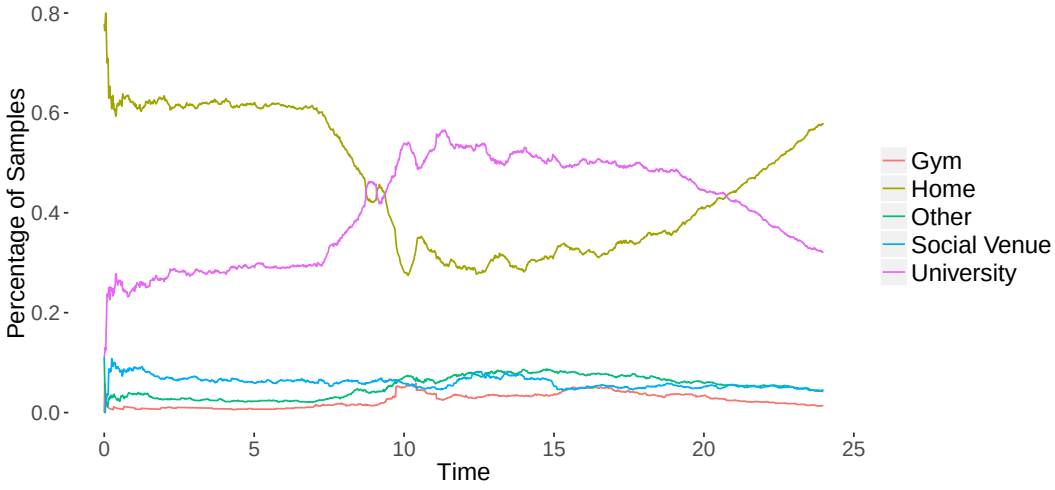

Figure 5.3: Percentage of time that students spend on average in each of the five location categories.

$$D^o(d) = \begin{cases} \sum_{j}^{j \in \mathcal{D}^o_{deadline}} \frac{1}{j-d}, & \text{if } j - T_{days} < d < j \\ 0, & \text{otherwise} \end{cases} \tag{5.1}$$

Thus, $D^o(d)$ will be equal to zero if there are no deadlines within the next T_{days} days, where T_{days} is a constant threshold; otherwise, $D^o(d)$ will be inversely proportional to the number of days remaining until the deadline. In my experiments I set the T_{days} threshold equal to 3. I found that with this value the correlation between the stress level of the participants and the variable $D^o(d)$ is maximised.

Finally, the StudentLife dataset includes responses of the participants to the Big Five Personality test [169]. The Big Five Personality Traits describe human personality using five dimensions: *openness*, *conscientiousness*, *extroversion*, *agreeableness*, and *neuroticism*. The personality traits of participants can be used to describe some baseline characteristics of the units and, for this reason, I include them in the study.

5.2.2 Causality Analysis

I apply the causal inference framework described in Section 5.1 in order to assess the causal impact of factors like exercising, socialising, working or spending time at home on

stress level. [1]

5.2.2.1 Variables

Initially, I define the variables that will be included in the study as follows:

1. $H_t^o(d)$: denotes the total time in seconds that the user o spent at home during day d until time t. I also define $H = \{H_t^o(d) : \forall o, \forall d\}$.

2. $U_t^o(d)$: denotes the total time in seconds that the user o spent at university during day d until time t. I also define $U = \{U_t^o(d) : \forall o, \forall d\}$.

3. $O_t^o(d)$: denotes the total time in seconds that the user o spent in any place apart from his/her home or university during day d until time t. I also define $O = \{O_t^o(d) : \forall o, \forall d\}$.

4. $E_t^o(d)$: denotes the total time in seconds that the user o spent exercising during day d before time t (it is estimated using both location traces and accelerometer data). I also define $E = \{E_t^o(d) : \forall o, \forall d\}$.

5. $C_t^o(d)$: denotes the total time in seconds that the user o spent at any socialisation or entertainment venue during day d before time t. I also define $C = \{C_t^o(d) : \forall o, \forall d\}$.

6. $S_t^o(d)$: denotes the stress level of user o that was reported on day d and time t. Stress level is reported one or more times per day. Thus, in contrast with the above mentioned variables, $S_t^o(d)$ is not continuously measured. I also define $S = \{S_t^o(d) : \forall o, \forall d\}$.

[1] There are some studies that provide evidence that mood of individuals is influenced by the mood of their peers (see for example [82, 80]). However, the dataset limitations do not allow me to investigate whether the stress experienced by a participant could influence his/her social circle. In order to examine this aspect, I create a friendship network based on the the phone calls/SMSes of the users. However, the resulting friendship network is composed of only 19 students out of 48 (i.e., there were only 19 students with at least one friendship link to another student). Moreover, all the users are not active during all the days of the study (e.g., some users do not report their stress level every day). In order to study the impact of friends' stress, I need to consider only samples for which I have information for both the stress level of the student taken into consideration and the stress level of his/her friends. This reduces the size of the dataset by 73%. The sample is not sufficient to derive statistically significant results. For this reason, the impact of the social network of individuals is not considered in this study.

7. $D^o(d)$: represents the upcoming deadlines as described in Equation 5.1. I also define $D = \{D^o(d) : \forall o, \forall d\}$.

8. $\mathcal{E}^o$, $\mathcal{N}^o$, $\mathcal{A}^o$, $\mathcal{C}^o$, $\mathcal{O}^o$: these five variables denote the extroversion, neuroticism, agreeableness, conscientiousness and openness of user o based on his Big Five Personality Traits score respectively. I also define $\mathcal{E} = \{\mathcal{E}^o : \forall o\}$, $\mathcal{N} = \{\mathcal{N}^o : \forall o\}$, $\mathcal{A} = \{\mathcal{A}^o : \forall o\}$, $\mathcal{C} = \{\mathcal{C}^o : \forall o\}$ and $\mathcal{O} = \{\mathcal{O}^o : \forall o\}$.

In this study, I examine the effects of five treatments, denoted by the variables H, U, O, E and C on the stress level of participants, which is described by the variable S.

5.2.2.2 Units

As was previously mentioned in Section 5.1, each unit of the study describes the 'state' of a participant during a time-period of a day. I define a set of time-periods $T = \{4$ am, 8 am, 12 pm, 16 pm, 20 pm, 24 pm$\}$. A time-period t_i corresponds to the i^{th} element of T. Then, I create the time-series: $H^o_{t_i} = \{H^o_{t_i}(d) : d \in D\}$, $U^o_{t_i} = \{U^o_{t_i}(d) : d \in D\}$, $O^o_{t_i} = \{O^o_{t_i}(d) : d \in D\}$, $E^o_{t_i} = \{E^o_{t_i}(d) : d \in D\}$, $C^o_{t_i} = \{C^o_{t_i}(d) : d \in D\}$, $S^o_{t_i} = \{S^o_{t_i}(d) : d \in D\}$, $D^o_{t_i} = \{D^o_{t_i}(d) : d \in D\}$, $\mathcal{E}^o_{t_i} = \{\mathcal{E}^o_{t_i}(d) : d \in D\}$, $\mathcal{N}^o_{t_i} = \{\mathcal{N}^o_{t_i}(d) : d \in D\}$, $\mathcal{A}^o_{t_i} = \{\mathcal{A}^o_{t_i}(d) : d \in D\}$, $\mathcal{C}^o_{t_i} = \{\mathcal{C}^o_{t_i}(d) : d \in D\}$ and $\mathcal{O}^o_{t_i} = \{\mathcal{O}^o_{t_i}(d) : d \in D\}$ by sampling the corresponding time-series at the times indicated by the set T. Since the variable $S^o_t(d)$ is not continuously measured, it is not feasible to sample it for time t_i. Instead, I define $S^o_{t_i}(d)$ as the average stress level of unit o in day d between time t_i and t_{i+1}. Thus, $S^o_{t_i}(d)$ is estimated as follows:

$$S^o_{t_i}(d) = \bar{S}^o_t(d), \text{ for } t_i \leq t \leq t_{i+1} \tag{5.2}$$

In addition, there is one time-sample per day for the variable $D^o_{t_i}$, thus $D^o_{t_i}(d) = D^o_{t_j}(d)$, $\forall t_i, t_j \in T$. Similarly, the extroversion, neuroticism, agreeableness, conscientiousness and openness of each participant is measured one time during the experiment. Thus, $\mathcal{E}^o_{t_i}(d) = \mathcal{E}^o_{t_j}(d')$, $\mathcal{N}^o_{t_i}(d) = \mathcal{N}^o_{t_j}(d')$, $\mathcal{A}^o_{t_i}(d) = \mathcal{A}^o_{t_j}(d')$, $\mathcal{C}^o_{t_i}(d) = \mathcal{C}^o_{t_j}(d')$ and $\mathcal{O}^o_{t_i}(d) = \mathcal{O}^o_{t_j}(d')$, $\forall t_i, t_j \in T$ and $\forall d, d' \in D$.

Finally, I define the maximum time-lag L equal to 1 day. I found that there is no

dependence of the participants stress level on events that happened in the past conditional to their previous day state, thus 1 day time-lag is sufficient. Then, for each participant o and each time-period t_i, I define the set of time-series $\mathbf{S^o_{t_i}} = \{H^o_{t_i}, H^{o,(1)}_{t_i}, U^o_{t_i}, U^{o,(1)}_{t_i}, O^o_{t_i}, O^{o,(1)}_{t_i}, E^o_{t_i}, E^{o,(1)}_{t_i}, C^o_{t_i}, C^{o,(1)}_{t_i}, S^o_{t_i}, S^{o,(1)}_{t_i}, D^o_{t_i}, D^{o,(1)}_{t_i}, \mathcal{E}^o_{t_i}, \mathcal{E}^{o,(1)}_{t_i}, \mathcal{N}^o_{t_i}, \mathcal{N}^{o,(1)}_{t_i}, \mathcal{A}^o_{t_i}, \mathcal{A}^{o,(1)}_{t_i}, \mathcal{C}^o_{t_i}, \mathcal{C}^{o,(1)}_{t_i}, \mathcal{O}^o_{t_i}, \mathcal{O}^{o,(1)}_{t_i}\}$. Then, a unit (o, t_i, d) of the study is described by the d-th time-sample of the set of time-series $\mathbf{S^o_{t_i}}$. Units with missing values are discarded from the study i.e., if for participant o there is no stress level report during the t_i time-period of day d, the unit (o, t_i, d) is discarded.

Units need to be split into *control* and *treatment* groups. I consider binary treatments by applying thresholds to the examined treatment variables. The threshold values are selected so that there is sufficient number of treated and control units. The impact of the threshold selection on the causal inference is evaluated by examining different thresholds. For each of the four examined treatments (i.e., U, O, E, C) the units are split as follows:

1. U: a unit (o, t_i, d) will be a treatment unit if $U^o_{t_i}(d) < \bar{U}_{t_i} - \alpha \cdot \bar{U}_{t_i}$ and control unit if $U^o_{t_i}(d) \geq \bar{U}_{t_i} + \alpha \cdot \bar{U}_{t_i}$, for a constant $\alpha \in [0, 1)$, where $\bar{U}_{t_i}$ is the sample mean value of U over all participants and days for the time-period t_i. Thus, I consider to have a positive treatment value when the university sojourn time is relatively small.

2. O: treatment units are all the units with $O^o_{t_i}(d) > \bar{O}_{t_i} + \alpha \cdot \bar{O}_{t_i}$ and control all the units with $O^o_{t_i}(d) \leq \bar{O}_{t_i} - \alpha \cdot \bar{O}_{t_i}$, where $\bar{O}_{t_i}$ is the sample mean value of O over all participants and days for the time-period t_i. Thus, I consider to have a positive treatment value when the time spent in any non-work-related place outside home is relatively large.

3. E: treatment units are all the units with $E^o_{t_i}(d) > 0$, i.e., all the units that denote that a user o had some exercise at day d before time t_i. In the control group are units with $E^o_{t_i}(d) = 0$.

4. C: similarly to the treatment variable E, treatment units are units with $C^o_{t_i}(d) > 0$ and control units with $C^o_{t_i}(d) = 0$

Thus, when the treatment variables U and O are considered, units are classified as treated and untreated based on the time participants have spent at university or at any

place apart from their home and university, respectively. However, in order to examine the impact of exercising and visiting socialisation venues, the binary treatments are defined by considering only whether there was some exercising activity or a visit to a socialisation place or not. I do not study the impact of these factors by considering also the duration of these events since the amount of the data is not sufficiently large.

5.2.2.3 Matching

As was previously mentioned in Section 2.3.1, in causality studies based on observational data, conditional ignorability need to be achieved by controlling the factors that influence both the treatment and the outcome variables of the study. While there is a large number of factors that could influence the stress level of participants, the study could be biased only by factors that have a direct influence on both the stress level and the variable that is considered as treatment in the study. Thus, in this case I need to specify factors that could influence both the daily activities of participants and their stress level. For example, the workload of students can influence their activities (e.g., in periods with increased workload some students may choose to change their workout schedule, etc.) and their stress level. Since the workload cannot be directly measured using only sensor data from smartphones, I use other variables that provide implicit indicators of workload as confounding variables, such as the time that students spend at home and university and their deadlines. Moreover, participants choice to do an activity may exclude another activity from their schedule and it may also influence their stress level. For example, someone may choose to spend some time in a pub instead of following his/her normal workout schedule. The previous day stress level may also influence both next day's activities and stress level. Finally, several studies have demonstrated that stress level fluctuations are affected by personality traits [8]. In general, more positive and extrovert people tend to be able to handle stress better than people with high neuroticism score. Moreover, personality characteristics may correlate with the daily schedule that people follow. For example, more extrovert people may spend less time at home and more time in social activities.

In order to find the set of variables that need to be controlled in order to reduce confounding bias, I apply Algorithm 2. Independence is tested by estimating the Kendall

rank correlation. In Table 5.2, I present the resulted p-values. Variables that do not correlate with any of the treatment or outcome variables are omitted.

	S	H	U	O	E	C
H	0.3557	0	$6 \cdot 10^{-128}$	$7 \cdot 10^{-182}$	0.0161	$2.7 \cdot 10^{-6}$
U	0.004	$6 \cdot 10^{-128}$	0	$2 \cdot 10^{-6}$	0.042	0.024
O	$6 \cdot 10^{-5}$	$7 \cdot 10^{-182}$	$2 \cdot 10^{-6}$	0	10^{-7}	10^{-13}
E	0.0081	0.0161	0.042	10^{-7}	0	0.222
C	$9 \cdot 10^{-5}$	$2.7 \cdot 10^{-6}$	0.024	10^{-13}	0.222	0
$S^{(1)}$	$2.7 \cdot 10^{-59}$	0.967	0.0071	0.055	0.3897	0.046
D	0.024	$2.5 \cdot 10^{-6}$	0.0014	0.0018	0.002	0.0076
$\mathcal{E}$	$1.69 \cdot 10^{-11}$	$2.27 \cdot 10^{-5}$	0.059	$4.9 \cdot 10^{-4}$	$4.1 \cdot 10^{-5}$	0.0037
$\mathcal{N}$	$1.81 \cdot 10^{-14}$	0.004	$1.2 \cdot 10^{-5}$	$2.3 \cdot 10^{-16}$	0.013	$6 \cdot 10^{-6}$
$\mathcal{A}$	0.007	0.21	0.15	0.047	0.006	0.002
$\mathcal{C}$	0.057	0.078	0.01	0.47	0.352	0.214
$\mathcal{O}$	0.604	0.006	0.005	$2.1 \cdot 10^{-5}$	$4.7 \cdot 10^{-4}$	0.95

Table 5.2: P-values of Kendall correlation under the null-hypothesis that the examined variables are independent.

Finally, in Table 5.3, I present the resulted conditioning set **H** for each treatment variable. Based on the results of Table 5.2, the time that students spend at home does not correlate with their stress level. Thus, the variable **H** will not be included in the causality study. The causal impact of each treatment variable U, O, E and C on the effect variable S will be examined using all the variables that correlate with both the treatment and effect based on Table 5.2 as confounding variables. I consider a correlation to be significant enough if the p-value is smaller than 0.1. While the variables O and C are strongly correlated, I do not include C in the set of confounding variables when the treatment is the variable $O^o_{t_i}(d)$, since my goal is to study the impact of spending time in any place (including socialisation venues) apart from home and working environment. Finally, as discussed in Sections 2.3.1 and 5.1, since $S^{(1)}$ correlates with S, $S^{(1)}$ needs to be included in the conditioning set **H** in order to satisfy the i.i.d. assumption.

As was previously discussed in Section 5.1, units can be matched only if they refer to the same time-period. However, since the dataset is relatively small, matching is allowed between units of different participants. Participants personality characteristics are used

Treatment	Conditioning Set H							
U	S^1	D	O	E	C	$\mathcal{E}$	$\mathcal{N}$	$\mathcal{C}$
O	S^1	D	U	E	$\mathcal{E}$	$\mathcal{N}$	$\mathcal{A}$	-
C	S^1	D	U	O	$\mathcal{E}$	$\mathcal{N}$	$\mathcal{A}$	-
E	D	U	$\mathcal{E}$	$\mathcal{N}$	$\mathcal{A}$	-	-	-

Table 5.3: Confounding variables for the different applied treatments.

as confounding variables in order to reduce the bias induced by differences on participants behaviour due to their personality.

5.2.2.4 Balance Check

In order to create balanced treated and control pairs of units I apply the Genetic Matching method [170]. Other simpler matching approaches (such as nearest neighbour matching and stratification) were also examined, however Genetic Matching reduced the confounding bias more effectively. In order to assess if the treated and control pairs are sufficiently balanced, I check the standardised mean difference for each confounding variables of the study as described with equation (6.8).

5.2.3 Results

I conduct a causal inference study for each one of the four examined treatments that were discussed above. In each study, I use as confounding variables all the variables that are presented in Table 5.3. I report my findings collectively for the whole population. I also repeated these studies separately for participants with high and low extroversion and participants with high and low neuroticism scores in order to investigate whether some of the examined treatments have a different causal impact on these sub-populations. I decided to conduct additional studies separately for these sub-populations because neuroticism and extroversion are strongly correlated with stress level according to Table 5.2. Participants are classified as *highly extroverts* if their extroversion score is higher than the average extroversion score; otherwise, they are classified as member of the *low extroversion* sub-population. Correspondingly, I define two sub-populations of participants with

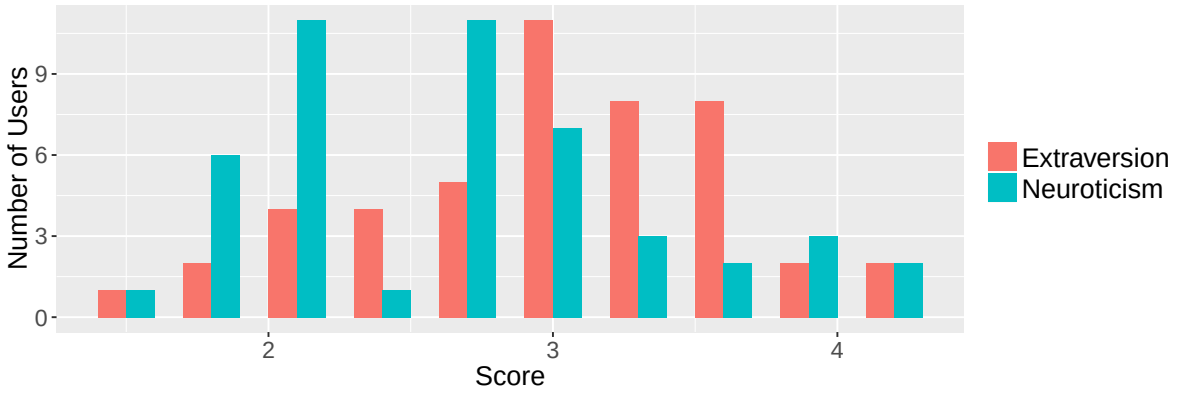

Figure 5.4: Distribution of neuroticism and extroversion scores.

high neuroticism (i.e., participants with neuroticism score higher than the average) and participants with *low neuroticism* scores. In Figure 5.4 I present the distribution of the neuroticism and extroversion scores of the participants.

In Figure 5.5 I show the average treatment effect (ATE) normalised by the average stress level of the control units along with the 95% confidence intervals for each one of the four examined treatment variables. For the treatment variables U and O I present results for α equal to 0, 0.05, 0.1 and 0.15. I do not present results for larger α values since the number of samples that are discarded is large and the remaining data are not sufficient for statistically significant conclusions. In Figure 5.6 and Table 5.4 I present the standardised difference, as described in Equation (6.8), for all the confounding variables that were used in each one of the causation studies. According to my results, the standardised difference between treated and control samples is smaller than 0.1 for all the confounding variables thus any confounding bias has been sufficiently minimised.

My results indicate that the time that students spend at university has only a weak causal impact on the stress level when participants' samples are split into treatment and control groups using an α value equal to 0.15. In detail, participants report 3.1% (with confidence interval $\pm 0.7\%$) lower stress level the days that their sojourn time at university is 15% lower than the average university sojourn time of the whole population compared to days that the university sojourn time is 15% larger than usual. However, when the analysis is limited to people with high extroversion score, there is no statistically significant evidence that the time that students spend at university has any causal effect on stress. When smaller α values are considered, the causality score is close to zero for

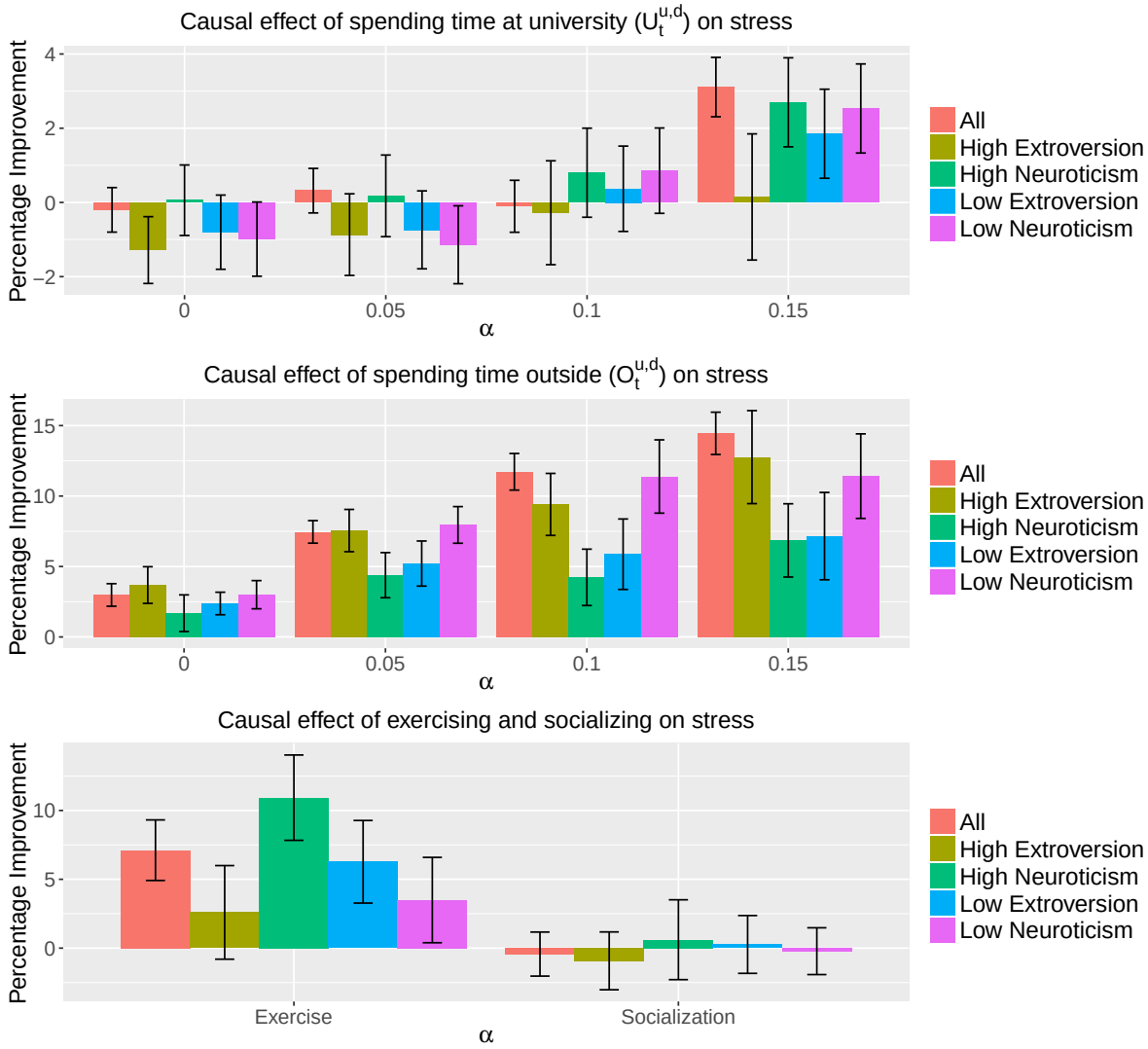

Figure 5.5: Percentage improvement on the stress level of treated units compared to control units when each one of the examined treatments is applied. Results are presented along with the 95% confidence interval. Confidence intervals are estimated by applying a t-test under the null hypothesis that there is no improvement on the stress level.

the examined set of students.

Based on my results, the time that students spend in any place apart from their home and university has a significantly strong causal impact on their stress level. As depicted in the second part of Figure 5.5, students reported around 3% (with confidence interval $\pm 0.65\%$) lower stress level the days that they spend more time outside than the average time compared to days that they spend less time outside (i.e., $\alpha = 0$), when the whole set of participants is considered. Similar results are observed when the study is repeated separately for students with high and low extroversion and students with high and low

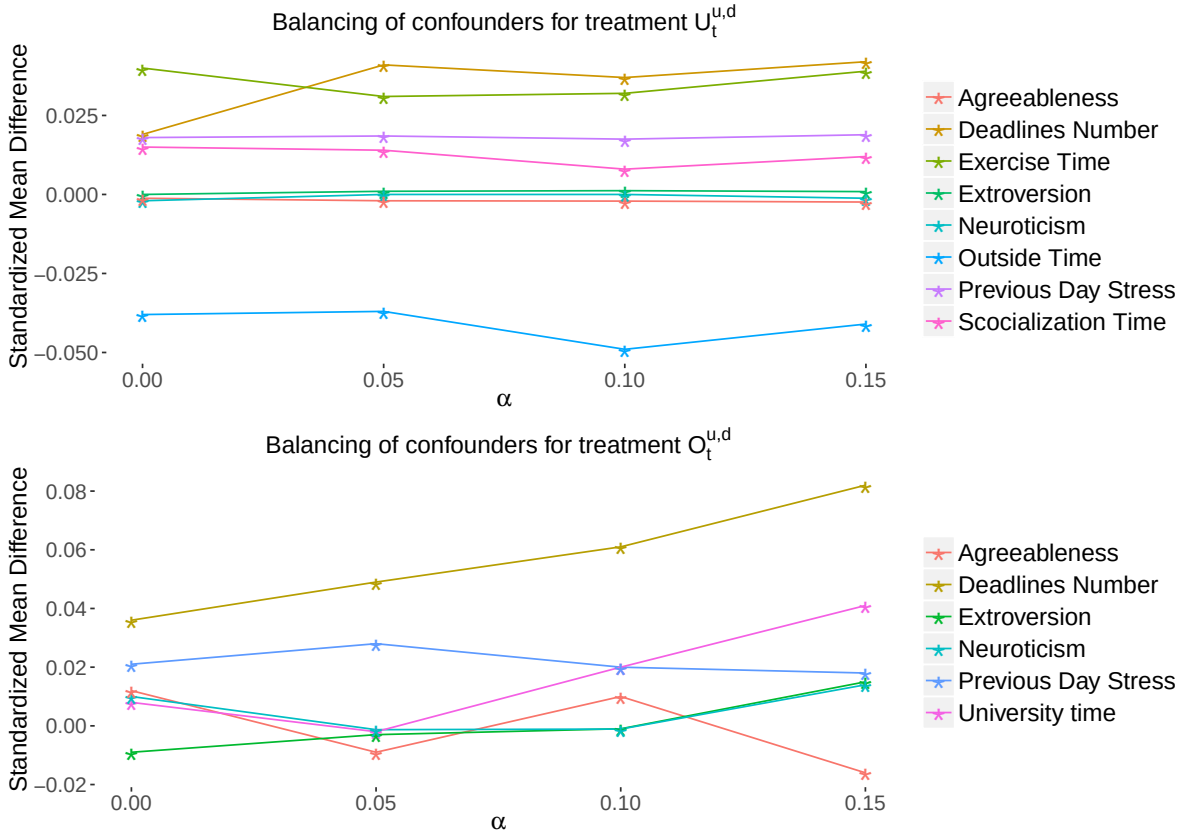

Figure 5.6: Standardized mean difference between treated and control samples for each confounding variable when the applied treatment is the variable U (top figure) and the variable O (bottom figure). The standardized difference for all the confounding variables is less than 0.1, thus the groups are balanced.

neuroticism scores (the observed difference is not statistically significant given the 95% confidence intervals of the study). When the value of α is increased, the causal impact of the examined variable is stronger. For $\alpha = 0.15$, the improvement on the stress level for students who spend more time outside is 14.45% (with confidence interval $\pm 1.5\%$) when the total population is considered. The results are similar when the study is limited to students with high extroversion score and students with low neuroticism scores. However, the examined variable has a significantly lower impact on stress level when only students with high neuroticism score and students with low extroversion score are considered.

In the third part of Figure 5.5, I examine the impact of exercising or visiting socialisation venues on stress level. While the variable C is strongly correlated with the stress level, according to my results, there is no causal link between them. This indicates that, while people benefit from spending time outside home or working environment in general,

there is no statistically significant benefit from visiting specific venues. Finally, exercising has positive effect on the stress level of the examined population. When I examine the four different sub-populations separately, I observe that exercising has a stronger positive effect on the stress level of participants with high neuroticism score while there is no statistically significant benefit for people with high extroversion score. The impact on people with low neuroticism score is also weak.

	$S^{(1)}$	D	U	O	$\mathcal{E}$	$\mathcal{N}$	$\mathcal{A}$
C	-0.0035	0.0442	0.0046	-0.0148	-0.0069	-0.0065	0.0001
E	-	0087	-0.0011	-	0.0047	0	0.0043

Table 5.4: Standardized difference between treated and control samples for each one of the confounding variables when the applied treatments correspond to the variables C and E.

5.2.4 Sensitivity Analysis

In this study, I have considered the most important factors that may influence both participants stress level and their daily activities. However, it is not feasible to assure that there are no other unmeasured factors that may influence the results. For example, some participants could be motivated by their friends to participate in a group exercise. The social interaction with their friends during the training program may also has a positive impact on their stress level. Thus, the observed link between stress and exercise could be actually due to the involved social interactions.

$\mathbf{\Gamma}$	**Upper bound on p-value**		
	U	O	E
1.0	0.032	0.0004	0.007
1.1	0.091	0.0012	0.016
1.2	0.255	0.0023	0.034
1.3	0.347	0.0078	0.061
1.4	0.591	0.023	0.123
1.5	0.784	0.042	0.285

Table 5.5: Sensitivity Analysis.

In order to assess the bias due to unmeasured confounding variables, I conduct a sensitivity analysis based on Rosenbaum's method [118] described in Section 2.3.1.5. In Table 5.5, I present the results of the sensitivity analysis for $\Gamma \leq 1.5$ and for the treatment variables U, O and E. The results are obtained on the whole participants population. For the treatment variables U and O results are obtained with $\alpha = 0.15$. According to my analysis, the observed causal link between U and S is very sensitive in the presence of unobserved confounding variables. The impact of exercise on the stress level could be also biased in case of unmeasured factors while the impact of O on S is more robust.

5.3 Discussion and Limitations

In this Chapter, I have highlighted the opportunities that arise by the utilisation of smartphone sensor data and I have presented a framework for detecting causal links on human behaviour by utilising such datasets. In this section, I discuss the limitations of the proposed approach.

The main limitation of any causal inference study based on observational data is that it could be biased in case of missing confounding variables. However, conducting experimental studies is not feasible in many cases due to either practical or ethical reasons. Smartphones as well as wearable devices can capture a large variety of data and offer useful information about users' daily activities. Additional information that may be needed in a study could be provided by the users through pop-up questionnaires. Thus, by leveraging this technology, scientists could obtain sufficient information in order to conduct reliable causal inference studies. Nevertheless, the possibility that unmeasured factors may influence the study cannot be eliminated, and consequently, any results should be supplemented by a sensitivity analysis.

In addition, in many cases it is hard to prove that the variable representing the *treatment* precedes temporally the variable representing the *outcome*. For example, in this case study, the stress level of users is not measured continuously; instead it is reported up to four times per day. Thus, it is not feasible to accurately know whether an *action* preceded the reported stress level or the stress level simply had not been reported before the *action*. In addition, the emotional state of a participant could be influenced also by

the anticipation of an *action* or *event*. For example, a participant may report increased happiness level when a trip has been scheduled. In this case, the emotional state of the participant is influenced by the upcoming trip although the *outcome*, i.e., the emotional state precedes temporally the *cause*, i.e., the trip. Given these limitations, in this study, the condition that *the cause needs to precede temporally the outcome* is ignored. Thus, although the proposed method uncovers relationships *stronger* than mere correlation, strong claims about the existence of a causal link between the examined variables should be avoided.

In addition, as it was previously discussed in Section 5.1, inferences based on sensor data could also be inaccurate either due to noisy sensor measurements or due to the fact that the variable of interest is inferred by the sensed data rather than directly measured. For example, in this case I assume that a visit to a sports centre implies that the user had some exercise. However, the user may have visited this place to attend a sport event or just to meet friends. This issue is farther examined in Chapter 6. Nevertheless, inferring this high-level information using raw sensor data instead of pop-up questionnaires has several advantages: 1) it offers a more accurate representation of participants activities over time since data are collected continuously; 2) data are collected in an obtrusive way without requiring participants to provide any feedback; this minimises the risk that some users will quit the study because they are dissatisfied because of the amount of feedback that they need to provide; 3) data gathered through pop-up questionnaires may not be objective since participants may provide either intentionally or unintentionally false responses.

It should be noted that the dataset that was used in this study contains sensitive information about participants. It has been shown that location traces contain adequate information in order to identify users [171, 172]. This dataset contains also highly sensitive information about participants mood and personality. The dataset has been released without any obfuscation and can be used solely for research purposes while it is required that researchers will not attempt to identify participants.

Finally, this case study involves a limited number of participants who do not constitute a representative sample of the population; therefore extrapolating general conclusions about the causal impact of the examined factors on stress level is not feasible. However, the purpose of this work is to demonstrate the potential of using smartphones for con-

ducting large-scale studies related to human behaviour, rather than present a thorough investigation on factors influencing the stress level of the participants.

5.4 Summary

In this chapter, I have proposed a framework for causal inference using smartphones sensor data. I demonstrate the potential of utilising this information in order to better understand human behaviour by studying the causal effects of several factors, such as working, exercising and socialising, on stress level of 48 students using data captured by means of smartphone sensors. This study does not consider the impact of social influence on stress level of individuals mainly because of dataset limitations. My results suggest that exercising and spending time outside home or university have a strongly positive causal effect on participants' stress level. I have also demonstrated that the time participants stay at university has a positive causal impact on their stress level only when it is considerably lower than the average daily university sojourn time. However, this impact is not remarkable.

Moreover, I have observed that some of the examined factors have different impact on the stress level of students with high extroversion score and on students with high neuroticism score. More specifically, more extrovert students benefit more from spending time outside home or university, while more neurotic students benefit more from exercising. Investigating whether there is a causal impact between students' personality and the way that different activities impact their stress level is out of the scope of this study.

My study mainly relies on raw sensor data that can be easily captured with smartphones. I have demonstrated that information extracted by simply monitoring users' location and activity (through accelerometer) can serve as an implicit indicator of several factors of interest such as their working and exercising schedule as well as their daily social interactions.

Despite the previously discussed limitations of the proposed method, smartphones sensor data enable the continuous and unobtrusive monitoring of users and could revolutionise the way that causality studies on human behaviour and emotional state are conducted. To the best of my knowledge, this investigation comprises the first step to-

wards this direction.

CHAPTER 6

CAUSAL INFERENCE UNDER MEASUREMENT ERRORS

In the previous chapters, I have discussed how human-generated sensor data can be used to better understand human behaviour and events influenced by it. Significant part of this thesis focuses on discovering causal relationships among factors of interest. For example, in Chapter 4 I have examined the causal impact of Twitter sentiment on the traded assets prices. In Chapter 5 I have also studied the causal impact of daily activities such as exercising, working and socialising, on the stress level as well as the causal links between smartphones usage and mood. In these works, key variables of the causality study are not directly measured. Instead, they are inferred from raw data. In particular, in [142] I use the location context (e.g., home, work, entertainment place etc.) in order to understand the daily time that participants spent working, socialising and exercising. However, the real location context is not known; instead, it is inferred from smartphone sensors and, consequently, it could be inaccurate. Location context could provide valuable information about users' daily schedule and activities and could facilitate many studies on human behaviour. However, requesting users to continuously label their location data would be inconvenient and may discourage them from participating in such studies. Also labelling might not be feasible in commercial applications as well.

Moreover, as it was previously discussed in Chapter 4, several studies attempt to link Twitter sentiment with stock market prices. However, the real text sentiment is not known; instead, it is inferred by applying text processing and classification techniques and consequently, subject to inaccuracies. Moreover, several studies have shown that social media data in some countries have undergone censorship [173]. In such cases, sentiment

or opinion tracking could be biased.

Inaccuracies on the estimation of key variables in causality studies may jeopardise the validity of the results. However, the vast amount of social media and smartphone sensor data contain low level information that requires significant amount of pre-processing in order to extract valuable data. Consequently, it is important to develop new causal inference techniques that could handle unobserved or inaccurately measured data. Latent variable models have been used to handle such cases [174, 175]. Scientists usually attempt to estimate the values of a latent variable from other observed variables by fitting the data in a structural equation model [174]. However, the selection of a proper model is a complex task that may result in misspecification and overfitting.

To the best of my knowledge, the problem of handling noisy variables in causality studies based on the matching design framework has not been addressed so far. In this section, I propose *probabilistic matching*, a matching method that takes into account the uncertainty about the real values of a noisy variable and attempts to find optimal pairs of units in order to maximise the probability that the matched units have similar characteristics. My method is based on the assumption that a probability distribution describing the real values of each unobserved variable is known or can be approximated. Although this assumption may restrict the applicability of the proposed method, it is realistic in many scenarios. For example, when an inference procedure is applied in order to learn the values of an unobserved variable L from some observed attribute C, a probability distribution $Pr(L|C)$ can be approximated, as I discuss later in Section 6.2.

I evaluate the proposed matching framework on two different simulation studies in comparison with a conventional matching method. I demonstrate that probabilistic matching reduces significantly the confounding bias and results in more accurate causal conclusions. I also evaluate my method on a real dataset. In particular, I use the social media dataset described in [176] in order to test whether text messages containing URLs tend to be reposted more often. This dataset includes a rich variety of features extracted from the Weibo microblogging service for 111 users along with a manually assigned binary label for each user indicating whether he/she has been characterised as *spammer* or not. In this scenario, I assume that the *spammer* label is an unobserved confounding variable and I apply a spammer detection method [177] in order to infer a label for each user

Symbol	Description
$\tilde{L}$	Variable with measurement errors, described with a $1 \times N$ vector
L_u	Random variable with $Pr(L_u \vert \tilde{L} = \tilde{l}_u)$
X_u	Stochastic variable describing the treatment of u
$\mathbf{X}$	$1 \times N$ vector of stochastic variables describing the treatments of the N units
$\mathbf{H}$	$P \times N$ matrix of stochastic confounders
H_u	u^{th} column of $\mathbf{H}$, denoting a $P \times 1$ vector of random variables for unit u
H^p	p^{th} line of $\mathbf{H}$, denoting a $1 \times N$ vector of random variables for the p confounding variable
H_u^p	Element in column u and line p of $\mathbf{H}$, denoting a random variable for the p^{th} confounding variable of unit u
$D(H_u^p, H_v^p)$	Distance between random variables H_u^p, H_v^p
$\mathbf{D}(H_u, H_v)$	Distance between random variables vectors H_u, H_v
$\mathbb{D}_{H_u,H_v}$	$P \times 1$ vector of distances between the P random variables H_u^p, H_v^p

Table 6.1: Notation.

from other observed attributes. I map the classification outputs to probability distributions describing the probability of a user to be a *spammer* and I use these probability distributions to the matching framework. I demonstrate that the number of URLs in text messages indeed influences the number of reposts. I repeat the causality study by applying a conventional matching method in two scenarios: 1) the ground truth binary *spammer* identifier is known and 2) only the noisy *spammer* identifier inferred from the data is known. The results of the first scenario serve as the ground-truth. I demonstrate that my results come in agreement with the conclusions of the first scenario, while the examined conventional matching method fails to detect the causal link.

6.1 Probabilistic Matching

Probabilistic Matching is based on the matching with continuous treatments framework discussed in Section 2.3.1.4. In particular, I extend this method in order to handle cases where treatment and/or one or more confounding variables may have noisy or censored measurements. In this chapter I use the notation presented in Table 2.1. The additional notation that is needed is summarised in Table 6.1.

I assume that for each unobserved variable L there is an observed noisy version $\tilde{L}$. For example, $\tilde{L}$ could be a location label inferred from smartphone sensor data (and consequently subject to inaccuracies) and L the real unknown location label. I also assume that for each observation $\tilde{l}_u$ of $\tilde{L}$ the corresponding random variable L_u has known probability distribution $Pr(L_u|\tilde{L} = \tilde{l}_u)$. In the following, I will consider the general case where all the key variables are noisy with the understanding that in the case of no noise the corresponding distribution reduces to the delta function:

$$Pr(L_u|\tilde{L} = \tilde{l}_u) = \begin{cases} 1 & , L_u = \tilde{l}_u \\ 0 & , L_u \neq \tilde{l}_u \end{cases} \tag{6.1}$$

Denote by X_u the random variable describing the treatment of unit u and with $\mathbf{X}$ a $1 \times N$ random vector of treatment variables of all units. I also denote by Z_u^p the random variable describing the p^{th} confounding variable of unit u and with $\mathbf{Z}$ a $P \times N$ matrix of random variables Z_u^p. As before, Z^p will denote the p^{th} row of $\mathbf{H}$ and H_u its u^{th} column. My objective is to find pairs of units with minimum distance Δ as given in Equation (2.5). However, if the treatment and/or any of the confounding variables are noisy, the real distance cannot be calculated. Consequently, the applied matching method may result in poor matches. I attempt to improve the matching by including the knowledge about the uncertainty of the variables into the matching process. Suppose I have a notion of a distance $D(X_u, X_v)$ between random variables X_u, X_v and a distance $\mathbf{D}(H_u, H_v)$ between random vectors H_u and H_v. I need to find pairs of units u, v that minimise

$$\Delta(u, v) = \frac{\mathbf{D}(H_u, H_v)}{D(X_u, X_v)} \tag{6.2}$$

I need to define a suitable distance metric D for the random variables. Commonly used distance metrics for distributions such as f-divergence metrics (e.g., Kullback-Leibler divergence) are not suitable in this case, since my objective is to estimate the probability that the values of two random variables X_u, X_v are close (i.e., $Pr(|X_u - X_v| < \epsilon)$, where ϵ a small positive constant). Since the distance metric needs to measure also the proximity between the values of two random variables, I suggest a metric that is based on comparison of the quantiles of the examined variables. Let me denote by $q_{X_u}(k)$ the k^{th} quantile of

variable X_u, $k = 1, 2, ..., K$. Then, I define $D(X_u, X_v)$ as follows:

$$D(X_u, X_v) = \frac{1}{K} \cdot \sqrt{\sum_k (q_{X_u}(k) - q_{X_v}(k))^2} \tag{6.3}$$

If X is not noisy, the quantile values will be the same and $D(X_u, X_v)$ reduces to the Euclidean distance of x_u and x_v.

6.1.1 Probabilistic Genetic Matching

Although several distance metrics can be used as the distance between random vectors $\mathbf{D}(H_u, H_v)$, in this work, I propose Probabilistic Genetic Matching (*ProbGenMatch*), a modified version of the Genetic Matching distance metric. Before I introduce the Probabilistic Genetic Matching method, I extend the Genetic Matching method [122] described in Section 2.3.1.2 for continuous treatments based on the framework described in Section 2.3.1.4. Although Genetic Matching has been proposed only for binary treatments, it can be extended to continuous treatments by modifying Equation (2.5) as follows:

$$\Delta(u, v) = \frac{d_{u,v,W} + \epsilon}{|x_u - x_v|} \tag{6.4}$$

The loss function also needs to be modified in order to penalise any matrix W that results in matched units with similar treatments. I think of the absolute differences on the p^{th} confounding variable values of the matched treated and control units $\{|h_u^p - h_v^p| : (u, v) \in G\}$ as realisations of a random variable A^p. Then, I define a set of K quantiles $\Delta^p = \{q^p(k)\}_{k=1}^K$. The loss function can be selected based on this quantiles set as described in Section 2.3.1.2.

Then, I further extend Genetic Matching for continuous treatments to Probabilistic Genetic Matching that can handle also stochastic variables. Denote by $\mathbb{D}_{H_u,H_v} = [D(H_u^1, H_v^1),\ D(H_u^2, H_v^2),\ ...,\ D(H_u^P, H_v^P)]^T$ the $P \times 1$ vector of distances $D(H_u^p, H_v^p)$ between the P random variables H_u^p, H_v^p, $p = 1, 2, ..., P$ (see Equation (6.3)). Then, I calculate $\mathbf{D}(H_u, H_v)$ by modifying the Genetic Matching distance of Equation (2.3) as follows:

$$\mathbf{D}(H_u, H_v) = \sqrt{\mathbb{D}^T_{H_u,H_v} \cdot (S^{-\frac{1}{2}})^T \cdot W \cdot S^{-\frac{1}{2}} \cdot \mathbb{D}_{H_u,H_v}} \tag{6.5}$$

The loss function used to select the optimal weight matrix W also needs to be modified. I use the quantiles-based loss functions described in Section 2.3.1.2. In particular, for each pair of units $(u, v) \in G$ I define a random variable:

$$\mathcal{A}^p_{u,v} = \frac{|H^p_u - H^p_u|}{|X_u - X_v|} \tag{6.6}$$

I denote by $a^p_{u,v}(k)$ the k^{th} quantile of $\mathcal{A}^p_{u,v}$. I also define the average k-th quantile for the p^{th} confounding variable, $\bar{a}^p(k) = \frac{1}{|G|} \cdot \sum_{(u,v)\in G} a^p_{u,v}(k)$. Finally, I collect the average quantiles in the set $\Delta^p = \{\bar{a}^p(k)\}^K_{k=1}$ to be used in a quantile-based loss functions described in Section 2.3.1.2.

6.1.2 Implementation

ProbGenMatch has been implemented as an R package and it is based on the *Matching* R package, an open source software which implements several matching methods. *ProbGenMatch* takes as input the probability distributions for all the confounding variables and treatment variable for all the units of the study (along with other optional parameters) and returns the matched pairs according to the previously described framework. If all the variables of the study are observed without any measurement errors, then *ProbGenMatch* is equivalent to the continuous Genetic Matching approach that I presented in Section 2.3.1.4. If also the treatment variable is binary, *ProbGenMatch* is equivalent to the Genetic Matching framework [122].

Many-to-one Matching. The application of many-to-one matching or matching with replacement (discussed in Section 2.3.1.2) in scenarios with continuous treatments is not straightforward since units cannot be grouped to treated and control. Previously presented matching methods with continuous treatments are using one-to-one matching, i.e., each unit can be used only one time. In this implementation, I offer the option to use each unit multiple times. In detail, researchers can decide about the maximum number of times M that units are allowed to be used. Since this may result in some units being

used multiple times while others not, I use frequency weights in order to eliminate the induced bias. Thus, for each matched pair $(u, v) \in G$ I assign a frequency weight:

$$f_{(u,v)} = 0.5 \cdot \left(\frac{1}{n_u} + \frac{1}{n_u}\right)$$

where n_u, n_v the number of times the units u, v have been used respectively.

Caliper Distance. Matching with caliper distance has been previously proposed as a way to impose restrictions on the maximum allowed dissimilarity between the matched units [148]. A caliper distance is simply a threshold that defines the maximum allowed difference of two units on their confounding variable values. In this implementation I also support matching with caliper distance as an optional parameter. For stochastic confounding variables, a probability threshold T_{prob} should be provided along with the caliper distance. This probability threshold allows the matching of two units only if the probability to have a larger difference than the caliper distance on their confounding variable values is smaller than T_{prob}. My implementation also allows users to specify a threshold on the minimum difference between the treatment values of two matched units. For stochastic treatment variables, a probability threshold should be provided along with the minimum treatment difference threshold.

Computational Cost. *ProbGenMatch* requires more computational resources than traditional genetic matching. In detail, the cost of estimating the distance between two confounder vectors, as described at Equation (2.3) is $O(P)$ and the cost of estimating the distances between all units pairs is $O(P \cdot N^2)$. In contrast, *ProbGenMatch* requires $O(K)$ in order to estimate the distance between two random variables, as described by Equation (6.3), $O(K \cdot P)$ for the estimation of the distance between two confounder vectors and $O(K \cdot P \cdot N^2)$ for the estimation of the distances between all units pairs.

6.2 Evaluation

I evaluate the proposed probabilistic matching framework on two synthetic and one real dataset. All the examined scenarios include one unobserved variable L along with an observed noisy version $\tilde{L}$. I use as baselines for the evaluation:

1. the traditional Genetic Matching (*GenMatch*) approach which treats $\tilde{L}$ as the *true* variable.

2. the *optimal* Genetic Matching (*OptGenMatch*), where I assume that L is observed without any noise. The performance of Genetic Matching under this optimal scenario serves as an upper bound to the performance of my method. The results obtained by *OptGenMatch* will be considered as ground-truth.

I use the synthetic datasets to evaluate the performance of the proposed framework on different noise levels. I also examine the sensitivity of my approach to the parameters described in Section 6.1.2 and, in particular, to the minimum allowed treatment difference and to the maximum number of times M that each unit can be used at the matching procedure. Simulated experiments are the most reliable method for assessing the strengths and limitations of the examined methods since the ground-truth is known. In addition, the characteristics of the dataset can be chosen so that the impact of the different method parameters can be evaluated. Finally, I apply my method on the social media dataset described in [176] in order to test whether text messages containing URLs tend to be reposted more often.

The three methods are evaluated using the following criteria:

- **Average Treatment Difference**. The matched units need to have large difference on their treatment values. When the treatment difference on the matched units is small, the impact of the treatment on the outcome variable may fade-out.

- **Remaining Confounding Bias**. As was previously discussed in Section 2.3.1, the resulted groups of matched treated and control units need to have similar distributions on their confounding variables values. I use the standardised mean difference, as described in Section 2.3.1.3, in order to assess whether sufficient balance has been achieved.

- **True/False Positive Causality Conclusions Rate.** When units with similar treatment values are matched and/or when the remaining confounding bias is large, the average treatment effect could be falsely considered as statistically significant (false positive result) or insignificant (false negative result). The ground truth (i.e.,

whether there is a causal link or not between the examined treatment and outcome variables) is known for the synthetic datasets and therefore, the rate of true/false positive causality conclusions rate can be estimated. In the study on the real dataset, I use as ground-truth, the result of the (*OptGenMatch* method.

6.2.1 Synthetic Dataset

I consider a binary variable L describing the class of objects represented by D-dimensional vectors of real numbers. I consider two types of vectors. The first type corresponds to positive examples (i.e., vectors that belong to the class $L = 1$). The data in each of the D dimensions of the first vector type are generated by a Gaussian process with mean value 1 and standard deviation σ_1. The second type of vectors corresponds to negative examples (i.e., $L = 0$) and their values in each dimension are generated by a random Gaussian process with mean -1 and standard deviation σ_2. I train a Support Vector Machine classifier on this synthetic dataset and afterwards I use the classifier on unseen synthetic data (generated with the same procedure) in order to learn a label $\tilde{L}$ for each vector. Then, I map the SVM outputs into probabilities by applying the process described in [178]. For each vector v, the probability distribution of random variable L_v corresponds to the output of this mapping procedure.

In this test case, I consider two-dimensional vectors (i.e., $D = 2$) and I set $\sigma_1 = 1$. I test the performance of my matching framework with different noise levels on the observed variable $\tilde{L}$ by increasing σ_2 from 1 to 2 with step 0.2. By increasing the variance of the second vector type, I make the vectors less separable and consequently, the resulted classes $\tilde{L}$ are less accurate. In all experiments, unless it is differently stated (i.e., Section 6.2.2.2), I set the maximum number of times M that each unit can be used equal to 5 for all the examined methods.

6.2.1.1 Unobserved Treatment Variable

In the first case, I consider L as the treatment variable. I generate two confounding variables $H^1 = \alpha_1 \cdot L + e_1$ and $H^2 = \alpha_2 \cdot L + e_2$, where e_1, e_2 correspond to random Gaussian noise with mean 0 and variance 1 and 2 respectively and α_1, α_2 are model coefficients.

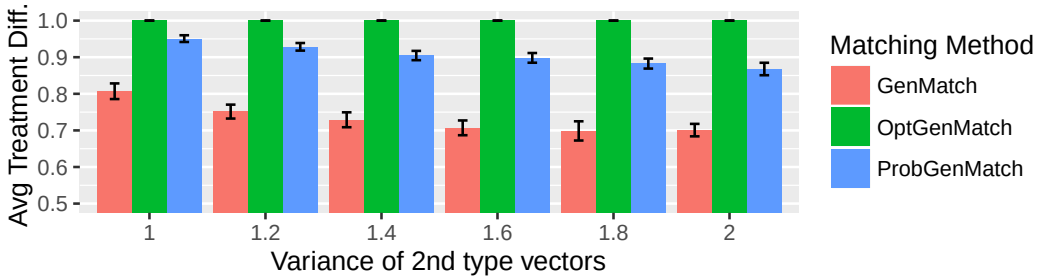

Figure 6.1: Average treatment difference between the matched units.

In the following results, I do not use a caliper distance for the confounding variables. I set the minimum allowed distance between the treatments of matched units equal to 0.1 and the maximum allowed probability that the matched units have a treatment difference larger than 0.1 equal to 0.25. I repeat the study for 10 randomly selected sets of model coefficients (αs). All model coefficients are randomly generated from a uniform distribution on $[0, 1]$. For each one of the 10 sets of model coefficients I repeat each study for 100 different noise realisations. In Figure 6.1 I present the average treatment difference between the matched units for the three examined matching algorithms along with the 95% confidence intervals. The *OptGenMatch* method always avoids matching units with the same treatment value. Thus, given that in this scenario I consider binary treatments, the average treatment difference is always equal to 1. According to my results, the performance of both *GenMatch* and *ProbGenMatch* declines for higher noise levels (i.e., larger σ_2). However, *ProbGenMatch* significantly outperforms *GenMatch* by avoiding matching units with the same treatment for more than 88% of the matched pairs for all examined noise levels.

When the resulted group of matched units contains pairs with the same treatment level, the impact of the examined treatment on the outcome variable cannot be reliably assessed by comparing the matched units on their outcome values. I demonstrate this by generating the following outcome variable:

$$Y = \beta_0 \cdot L + \beta_1 \cdot H^1 + \beta_2 \cdot H^2 + n_u + e_n \tag{6.7}$$

where $\beta_0, \beta_1, \beta_2$ are model coefficients, n_u is a uniform random variable on $[0, 4]$ and e_n is Gaussian noise with mean 0 and variance 1. All β coefficients are randomly generated

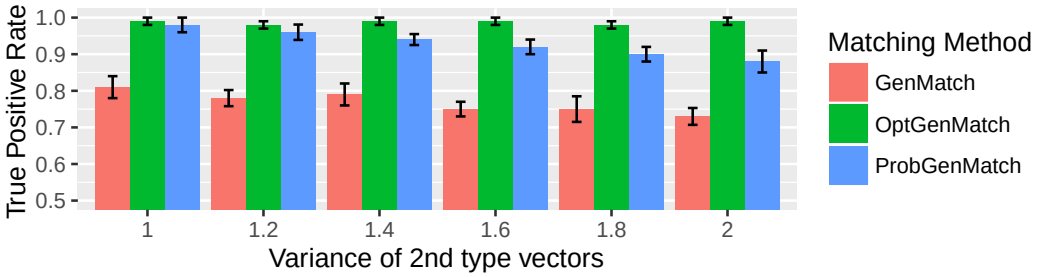

Figure 6.2: Percentage of true positive causality conclusions.

from a uniform distribution on $(0, 1]$. For non-zero β_0, the treatment variable L has a causal impact on Y. I apply a Wilcoxon non-parametric test in order to examine whether the average treatment effect, (Equation (2.6)) is significantly different than zero. When the performance of *OptGenMatch* is examined, I use as treatment (i.e., the variable X of Equation (2.6)) the binary variable L, while for *GenMatch* and *ProbGenMatch* I use the noisy variable $\tilde{L}$. I repeat the study for 10 different sets of model coefficients and 100 realisations of n_u, e_n, for all the groups of matched units resulted after applying the three examined methods, as it was previously described. In Figure 6.2 I depict the average percentage of times that the null hypothesis of the Wilcoxon test (i.e., that the average treatment effect is equal to zero) was rejected with p-value equal to 0.05. *OptGenMatch* successfully detects the causal impact of L on Y in most cases, while *ProbGenMatch* significantly outperforms *GenMatch*.

6.2.1.2 Unobserved Confounding Variable

In the second case, L corresponds to a binary confounding variable. In detail, I consider a continuous treatment variable X that follows a uniform distribution on $[0, 1]$. The binary confounding variable L follows a binomial distribution with success probabilities given by the vector of probabilities $P_S = 1/(1 + e^{-t})$, where $t = \alpha_0 + \alpha_1 \cdot X$. I also create a confounding variable $H^1 = \alpha_1 \cdot X + e_1$. I evaluate the performance of the three examined matching approaches by generating different realisations of the model coefficients and noise e_1, as it was previously described in Section 6.2.1.1. I assess the remaining bias due to imperfect matches by calculating the standardised difference in means for each confounding variable as described in Section 2.3.1.3. In detail, for the binary confounding

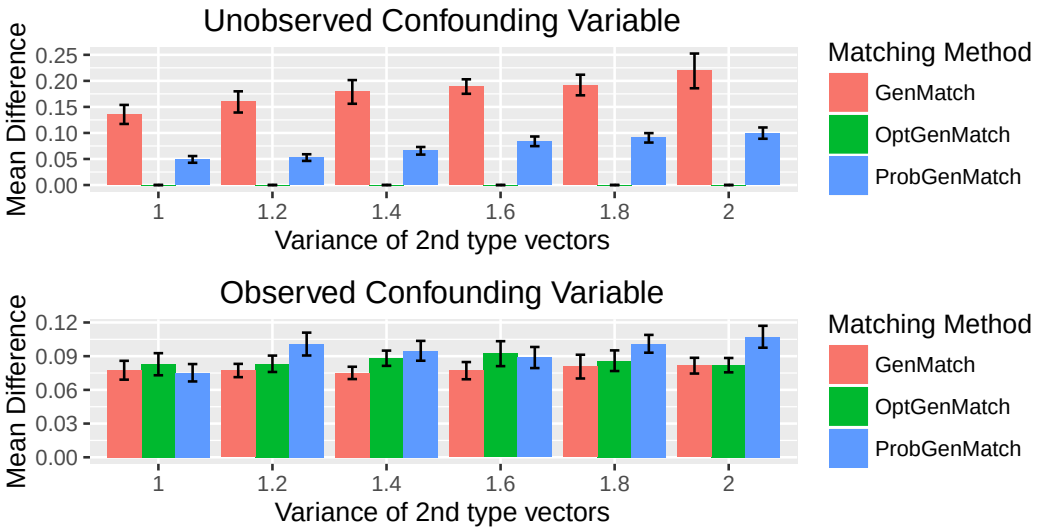

Figure 6.3: Remaining bias for the two confounding variables.

variable L, I consider the values $\{\frac{l_u}{x_u - x_v} : (u,v) \in G\}$ as realisations of a random variable C_U and the values $\{\frac{l_v}{x_u - x_v} : (u,v) \in G\}$ as realisations of a random variable C_V. Then the standardised difference in means for the confounding variable L is estimated as:

$$\frac{|\bar{C}_U - \bar{C}_V|}{\sqrt{(\sigma_U^2 + \sigma_V^2)/2}} \tag{6.8}$$

where $\bar{C}_U$, $\bar{C}_V$ are the mean values of C_U, C_V respectively and σ_U^2, σ_V^2 their variances. The same process is followed for the estimation of the standardised difference in means for H^1.

In Figure 6.3 I present the standardized difference in means for the two confounding variables (i.e., the binary variable L on the top and the continuous H^1 on the bottom). *OptGenMatch* always matches units with the same value on L and therefore, there is zero bias. The proposed *ProbGenMatch* method achieves also low bias, smaller than 0.1 for all the noise levels and significantly outperforms *GenMatch*. Finally, all methods achieve similar performance on the continuous confounding variable H^1, which is considered to be observed without any noise, although the performance of *ProbGenMatch* is slightly worse for large noise levels.

Failing to sufficiently eliminate the bias induced by confounding variables may result in false positive causality conclusions. I demonstrate this by considering again the outcome variable of Equation (6.7). This time I set $\beta_1 = 0$, thus, there is no causal impact of

H^1 on Y. In Figure 6.4 I present the rate of the false positive causality conclusions (i.e., the average percentage of times that the null hypothesis of the Wilcoxon test was not rejected) along with the 95% confidence interval. *ProbGenMatch* achieves up to 8% lower false positive rate than *GenMatch*.

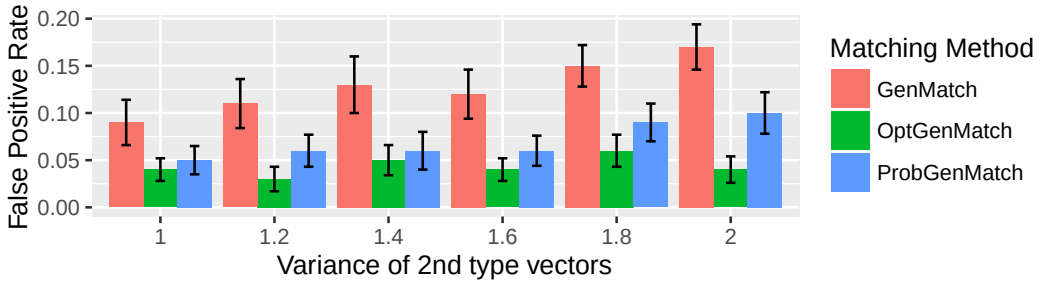

Figure 6.4: Percentage of false positive causality conclusions.

6.2.2 Location-based Synthetic Dataset

In this scenario, the latent variable L represents the daily time that the participants of a study spend in entertainment venues such as pubs, restaurants, bars, etc. I assume a study based on smartphone sensor data, where participants do not report their location; instead, location, along with the underlying context (i.e., work, home, restaurant etc.) is inferred from other raw sensor data. Several methods for automatic location label inference have been proposed [179, 180]. However, the real location context cannot be inferred accurately. As was previously discussed, location context could be very important for studies examining the impact of social behaviour or daily activities (e.g., exercising, socialising etc.) on well-being indicators such as stress level [142] or eating disorders [24].

I synthetically generate a location dataset based on the description of the real dataset presented in [179]. In [179], authors gather several sensor data along with ground truth labels for the locations of 36 participants and they apply a method for automatic location label inference. In order to generate the dataset, I define a variable P denoting a location label. As described in [179], I consider 7 location labels: *home*, *work*, *college*, *entertainment*, *food*, *shops* and *other*. $P_u(t)$ denotes the location of a participant at day u and time t and it is sampled on an hourly basis. The variable $P_u(t)$ is generated so that it simulates the daily human location patterns. In detail, the location pattern that is used

Algorithm 4 Generate Hourly Location Labels

Output: P_u

{Create binary (i.e., True/False) indicators for daily activities using generateBinaryIndicator function (Algorithm 5)}
weekDay := generateBinaryIndicator(0.714) {weekDay is True with probability 0.714 (5 out of 7 days are weekdays)}
lunchOutside := generateBinaryIndicator(0.5)
dinnerOutside := generateBinaryIndicator(0.5)
entertainmentWD := generateBinaryIndicator(0.4)
entertainmentWE := generateBinaryIndicator(0.5)
shop := generateBinaryIndicator(0.4)
other := generateBinaryIndicator(0.3)
{Initialise hourly location labels P_u with label *home*}
for t=1 to 24 **do**
  $P_u(t) = home$
end for
{Use function set (Algorithm 6) to create a label}
currentTime := 8 {Set current time equal to 8 (start of working day)}
if (weekDay) **then**
  {If it is a weekday, set 'work' label for 7-10 hours, starting from currentTime}
  P_u, currentTime := setLabel(currentTime, 7, 10, *work*)
  {if lunchOutside is True, set location label at 12:00 equal to 'food'}
  if (lunchOutside) **then**
    $P_u(12) =$*food*
  end if
  {If entertainmentWD, set 'entertainment' label for 1-3 hours, starting from currentTime}
  if (entertainmentWD) **then**
    P_u, currentTime := setLabel(currentTime, 1, 3, *entertainment*)
  end if
  {If it's weekend and entertainmentWE is true, set 'entertainment' label for 2-5 hours, starting from currentTime}
else if (entertainmentWE) **then**
  P_u, currentTime := setLabel(currentTime, 2, 5, *entertainment*)
end if
{If shop, set the currentTime label equal to 'shop'}
if (shop) **then**
  $P_u(currentTime) = shops$
  currentTime := currentTime + 1
end if
{If other, set 'other' label for 1-3 hours, starting from currentTime}
if (other) **then**
  P_u, currentTime := setLabel(currentTime, 1, 3, *other*)
end if

{If dinnerOutside, set 'food' label for 1-3 hours, starting from currentTime}
if (dinnerOutside) **then**
P_u, currentTime := setLabel(currentTime, 1, 3, *food*)
end if

in this study simulates a user who spends 6-8 hours at his workplace during the weekdays. During the weekdays he has lunch outside his workplace with probability 0.5. He also has dinner outside his home with probability 0.5. After the work he visits an entertainment place with probability 0.4 for 1-3 hours and at the weekends he visits an entertainment place with probability 0.5 for 2-5 hours. The probability to visit a store any day is 0.4 and the probability to visit any other place is 0.3. The process of generating the $P_u(t)$ variable for each day u is described by Algorithm 4.

Algorithm 5 Function **generateBinaryIndicator**.
Input: *probTrue*: the probability that the activity will take place
Output: A binary indicator (True/False) denoting whether or not the activity will take place
generateBinaryIndicator(probTrue)
{Create a random variable S with uniform distribution on [0, 1]}
$S \sim U(0, 1)$
s := random sample from S
{The probability that a random sample s from S is larger than *probTrue* is *probTrue*. So, the binary indicator will be True with probability *probTrue*}
if $s > probTrue$ **then**
return False
else
return True
end if

I also define a variable $E_u(t)$ as follows:

$$E_u(t) = \begin{cases} 1 & , P_u(t) = entertainment \\ 0 & , \text{otherwise} \end{cases} \tag{6.9}$$

Finally, I create a variable L, with values $l_u = \sum E_u(t)$ for each day u. However, in a real study, where participants would probably be unwilling to continuously provide labels for their location data, L would be a latent variable. I generate the discrete variable $\tilde{P}(t)$ denoting the inferred location label based on the method described in [179] by utilising the confusion matrix (Table 3 of [179]) that presents the performance of the proposed location

Algorithm 6 Function **setLabel**.

Input:
P_u: location label per hour
currentTime: the current time
label: the activity label (i.e., work, food etc.)
minActivityTime: the minimum time of the activity
maxActivityTime: the maximum time of the activity
Output: currentTime, P_u
setLabel(P_u, *currentTime*, *minActivityTime*, *maxActivityTime*, *label***)**
{Create a random variable S with uniform distribution on [*minActivityTime*, *maxActivityTime*]}
$S \sim$U(*minActivityTime*, *maxActivityTime*)
{Set the activity time by randomly sampling from S}
activityTime := random sample from S
{Set the location label equal to *label* for time equal to activityTime, starting from currentTime}
for t=*currentTime* to (*currentTime*+*activityDuration*) **do**
$P_u(t)$ =*label*
end for
{Change currentTime}
currentTime := *currentTime* + *activityTime*
return P_u, *currentTime*

inference method. According to this matrix, only 41% of the places with resulted label *entertainment* are correctly labeled while the rest 59% of the places actually correspond to *college* (4%), *work* (4%), *shops* (4%), *food* (33%) and *others* (9%). I create a noisy variable $\tilde{P}(t)$ by randomly inserting bias on $P(t)$ based on these results. Then I define $\tilde{E}_u(t)$ as:

$$\tilde{E}_u(t) = \begin{cases} 1 & , \tilde{P}_u(t) = entertainment \\ 0 & , \text{otherwise} \end{cases} \tag{6.10}$$

I also create $\tilde{L}_u$ with values $\tilde{l}_u = \sum \tilde{E}_u(t)$ for each day u. Finally, based on the performance of the location inference method, I create a random variable L_u with probability distribution $Pr(L_u|\tilde{P}_u(1), \tilde{P}_u(2), ...\tilde{P}_u(24))$. I normalise L, $\tilde{L}$ to $[0, 1]$.

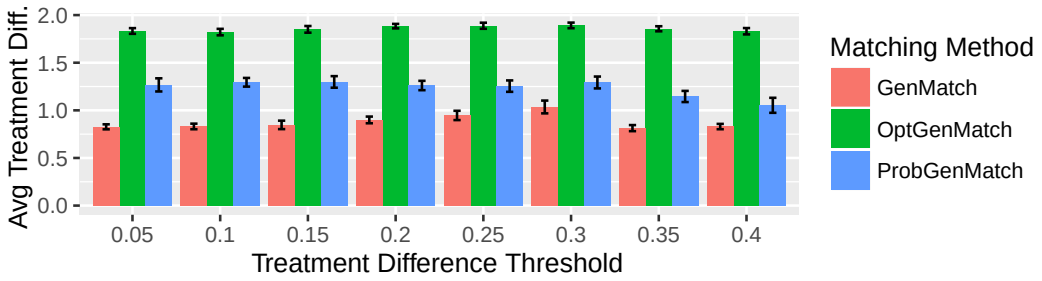

Figure 6.5: Average Treatment Difference between the matched units.

6.2.2.1 Impact of Minimum Treatment Difference Threshold

I use L as the unobserved treatment variable and I generate the confounding variables H^1, H^2 as it is described in Section 6.2.1.1. In this scenario, I examine the impact of the allowed minimum treatment difference on the three examined matching methods. In detail, let me denote with T_{min} the minimum allowed treatment distance. I vary T_{min} from 0.05 to 0.4 with 0.05 step. For *ProbGenMatch* I set the maximum allowed probability that the treatment difference is smaller than T_{min} equal to 0.25. In Figure 6.5 I present the average treatment difference between the matched treated and control groups achieved by the three examined matching algorithms. According to my results, there is not significant impact of the treatment difference threshold on the average treatment difference when the *OptGenMatch* method is applied. There is an improvement on the performance of *GenMatch* for the threshold values 0.2 to 0.3, however its performance is decreased for larger than 0.3 thresholds. Since the threshold is applied on the observed noisy variable $\tilde{L}$ and not on L, large threshold values may prevent the matching of units that are actually good matches. *ProbGenMatch* is not strongly influenced by the treatment difference threshold, however its performance is also decreasing for large threshold values.

Finally, I generate again an outcome variable as described in Equation (6.7) in order to examine the influence that the resulted matching may have on a causality study. I examine the rate of true positive causality conclusions for the three examined methods by repeating the process described in Section 6.2.1.1 and I present my results in Figure 6.6. *ProbGenMatch* achieves a higher rate of true positive conclusions compared to *GenMatch*, however their difference is less significant compared to the binary treatments case examined in Section 6.2.1.1. This is reasonable considering that for binary noisy treatments matching will result more often in pairs with the same treatment value; thus,

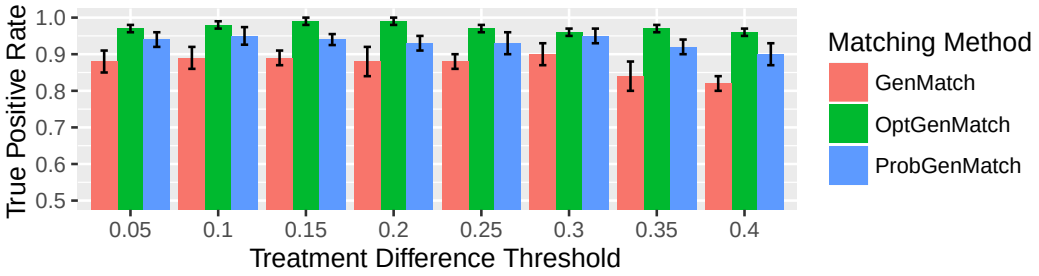

Figure 6.6: Percentage of True Positive Causality Conclusions.

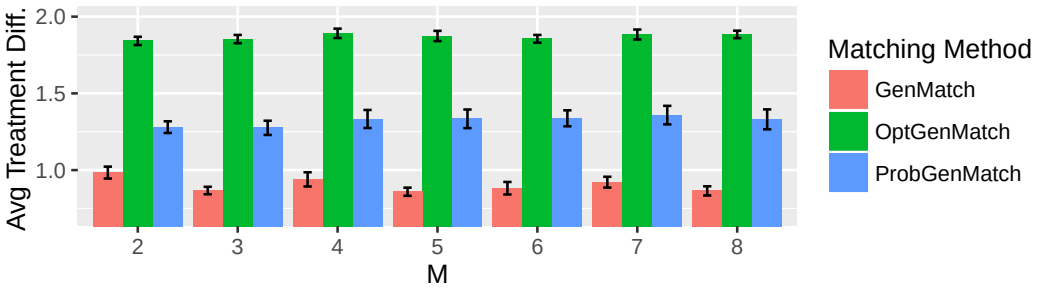

Figure 6.7: Average Treatment Difference between the matched units.

the treatment effect will be weaker.

6.2.2.2 Impact of Parameter M

In this section, I examine the impact of the maximum number of times M that each unit can be used. I set the minimum allowed treatment difference T_{min} equal to 0.1 and I vary M from 2 to 8. All the other settings are the same with those described in Section 6.2.2.1. In Figure 6.7, I present the average treatment difference between the matched treated and control groups for the three examined matching algorithms. According to my results, the M parameter slightly influences only the performance of *GenMatch* while there is no significant difference in the performance of the other two methods. *GenMatch* performs better when M is set equal to 2. Since *GenMatch* uses the noisy treatment variable $\tilde{L}$ for the matching, a unit that is falsely assumed to have large dissimilarity on its treatment value with some others can be matched multiple times when a large M value is used; consequently, the resulted matching could be worse in such cases.

6.2.3 Social Media Dataset

In this section, I evaluate my method on a real dataset. I use the microblogPCU dataset, which is available in the UCI Machine Learning repository [176], in order to examine whether the number of URLs included in microblog messages influences the number of times that these messages are re-posted. The MicroblogPCU dataset has been collected from the Sina Weibo microblog and contains information about the profiles of 782 users, their social network and their microblog activity. It also contains ground-truth binary labels indicating whether a user is a *spammer* or not for 111 users.

I use the ratio of messages with URLs as the treatment variable of the study and the number of re-posts as the outcome variable. Spammers tend to use more URLs in their messages and spammers messages are re-posted less often. Thus the spammer binary indicator should be used as a confounding variable. I also use other indicators that correlate both with the treatment and the outcome variables as confounding variables. In detail, I found that the number of posts, the class level of the user account (this is an indicator assigned by Weibo) and the number of followers correlate with both the treatment and outcome variables.

I assume that the binary spammer indicator is unknown and it needs to be inferred from the data. I apply the method described in [177] in order to classify the users to spammers and non-spammers. I extract attributes from the text content and users profiles as described in [177]. I use all the attributes of [177] apart from the number of times a user replied to a message or received a reply and whether a message is a *reply* message, since this information is not provided in this dataset. Also, instead of the user account age, I use the user account class. The interested reader should refer to [177] for a complete list of all the extracted features. Then, I apply the chi-squared feature selection method in order to find the most important attributes. Six attributes were selected, namely: 1) the fraction of tweets with URLs; 2) the user account class; 3) the average number of URLs per message; 4) the number of followees; 5) the average number of hashtags per message; and 6) the average number of re-posts. Following the procedure of [177], I use Weka [181] to train a support vector machine classifier. I use only the data for the 111 users for which a ground-truth label is available. I use 50% of the dataset for training and the rest for testing. 76% of the spammers and 82% of the non-spammers are correctly classified. The

	Balance on L	Wilcoxon test p-value
OptGenMatch	0.014	0.005
GenMatch	0.36	0.15
ProbGenMatch	0.15	0.041

Table 6.2: Causality Study Results

classifier is more successful on recognizing the spammers and less on recognising the non-spammers compared to [177]. This difference can be attributed to the differences between the dataset characteristics of the two studies. It should be noted that the specific dataset and methodology for classification of users to spammers and no-spammers is used solely to demonstrate the validity of the proposed causal inference framework on real datasets and the purpose of this study is not to conduct an evaluation on the performance of different methods for detecting spammers.

I define as L the ground-truth binary label indicating whether a user is spammer. I also define as $\tilde{L}$ the inferred label based on the above-mentioned process. I also create a random variable L_u for each user u and I obtain a probability distribution for each L_u by mapping the SVM outputs into probabilities. I match the users based on their confounding variables values by applying the three examined approaches. The results obtained by *OptGenMatch* serve as the ground-truth. Finally, I use the Wilcoxon test to examine whether the mean value of the outcome variable for the treated units significantly differs from the mean outcome value of the control units. In Table 6.2 I present the mean difference (see Equation (6.8)) achieved for the binary confounding variable L with the 3 examined methods. I also present the p-values of the Wilcoxon test under the null hypothesis that the treatment variable has no effect. Both *OptGenMatch* and *ProbGenMatch* reject the null hypothesis with p-value smaller than 0.05. However, when the treatment and control pairs are created by applying the *GenMatch* method, the remaining bias on the binary indicator L is large and results in the false conclusion that there is no significant impact of the number of URLs included in text messages to the number of re-posts.

6.3 Discussion

The development of the proposed *probabilistic matching* method has been motivated by the fact that many human generated sensor data contain high-level information that is inferred from lower level data with some degree of uncertainty. In contrast to previous approaches, *probabilistic matching* handles noisy data as stochastic variables and attempts to derive reliable causality conclusions by maximising the probability that confounding bias has been sufficiently eliminated. The method is based on the assumption that a probability distribution for each noisy variable is known or can be approximated. If there is not enough information for the reliable estimation of the required probability distributions, the efficacy of the method might be jeopardised.

I firstly evaluate the performance of the proposed method on simulated datasets. In simulation studies, the ground truth is known, thus it is easier to assess the validity of the examined methods. In addition, simulated studies offer the flexibility to adjust the dataset characteristics so that the impact of the different method parameters can be assessed. I compare *Probabilistic Genetic Matching* with the traditional *Genetic Matching* i.e., the deterministic version of the examined method. I also consider the optimal case where all the variables are observed without any measurement error in order to estimate an upper bound on the performance of the examined methods. The key findings of this evaluation can be summarised as follows:

- When the treatment variable is observed with some noise, we may end up matching units with similar treatment level. When comparing units with similar treatment level, the treatment effect will appear weakened and this may result in false negative causality conclusions. On the other hand, when one or more confounding variables are noisy, the matching method may result in unsuitable pairs of units. Consequently, the confounding bias will be sufficiently eliminated and this may result in false positive causality conclusions. *ProbGenMatch* handles better these two issues, compared to the traditional genetic matching, by incorporating information about the uncertainty about the variables' true values into the matching process.

- *ProbGenMatch* is less effective for more *noisy* variables. This is reasonable, since more noisy data will result in less suitable matches. However, *ProbGenMatch* is more

robust compared to traditional genetic matching and significantly outperforms it for all the examined noise levels.

- *ProbGenMatch* is not strongly influenced by the selection of the matching parameters. In particular, it is moderately influenced by the minimum required treatment difference on the matched units, while there is no significant influence by the selection of the maximum number of times M that a unit is allowed to be used on the matching process. On the other hand, *GenMatch* is more vulnerable on the selection of these parameters. This is due to the fact that by imposing more restrictions on the matching process, suitable matches are more unlikely to be found.

Finally, I demonstrate the applicability of the method by conducting a causality study on a real dataset. Noisy data impose challenges on the causality analysis and every effort should be made in order increase the data quality. However it is often not feasible to obtain accurate datasets. In such cases, the proposed approach results in more suitable matches and consequently, it is more likely to acquire valid causal conclusions.

6.4 Summary

In this Chapter, I have examined the problem of causal inference when key variables of the study are unobserved or noisy. I have proposed *probabilistic matching*, a novel method that utilises the knowledge about the uncertainties on the variables of the study in order to improve matching. I have defined a distance metric, based on Genetic Matching distance, that measures the dissimilarity between units by examining the difference on the quantiles of the stochastic variables of the study. My method is based on the assumption that probability distributions describing the values of the unobserved variables are known or can be approximated. Although this requirement could be restrictive, I have discussed scenarios which satisfy this assumption and I have demonstrated the applicability of my approach using both simulated and real datasets. I have shown how noisy variables can jeopardise the validity of the causality analysis and I have evaluated the performance of the proposed method in datasets with different noise levels. I have also examined the sensitivity of the method on different parameters values. I have compared *probabilistic*

matching with the traditional *Genetic Matching* and I have shown that my approach is able to find better matches and, consequently, achieves more accurate causal conclusions.

CHAPTER 7

CONCLUSIONS AND FUTURE DIRECTIONS

We produce daily vast amounts of data through both our online and offline activities. We use social media and web blogs in order to express our opinion and socialise. Search engines, web blogs and online newspapers are our main source of information and online shopping is getting more and more popular. In addition, smartphones and wearable devices have become an indispensable part of our lives. These devices, equipped with a rich variety of sensors, enable the continuous and unobtrusive monitoring of our offline activities.

The purpose of this thesis was to demonstrate how such information from diverse sources can be combined in order to better understand human behaviour and factors influenced by it. Instead of searching simply for correlation relationships, which may occur incidentally and may not represent the real structure of the data, I have attempted to detect stronger dependencies among the factors of interest which could provide better insights about the underlying mechanisms that influence the values of the examined variables. In this direction, I have developed novel techniques that enable the more effective utilisation of human-generated sensor data. Although the methods developed in this thesis are motivated by problems and properties that characterise this specific kind of data, they are general and could be applied to any dataset with similar properties.

In this dissertation I focused on two case studies. In the first one, I investigated the link between social media and finance and I provided evidence of cause and effect dependencies. In the second case study, I demonstrated how smartphone sensor data can be utilised for the detection of daily activities that influence our stress level.

7.1 Thesis summary and contributions

The contribution of this thesis is twofold. Firstly, I have developed novel methods that facilitate the extraction of useful insights from human-generated sensor data. In detail:

1. In Chapter 3, I have discussed a novel method for causal inference in observational time-series data. The method is based on the matching design framework and it does not require any assumptions about the functional structure of the data. I have conducted an extensive evaluation using synthetic data and I demonstrated that my method is more effective in avoiding false positive causality conclusions compared to widely used existing methods. Then, in Chapter 4 I applied this method in order to study the causal effect of behavioural and emotional factors, captured by social media data, on the traded assets of four companies. In addition, in Chapter 5, the proposed method was modified in order to handle smartphone sensor data. A complete framework for causal inference from smartphone sensor data is discussed.

2. In Chapter 4, I have discussed *FED*, an event detection method tailored to detect bursty Twitter topics that influence a specific stock market. *FED* models bursty Twitter topics as multi-dimensional feature vectors. Then, a classifier is trained to recognise which of the detected Twitter events are linked with stock market jitters on a specific stock market. The classifier is trained by utilising volatility data from the targeted stock market without requiring any manual labelling of the events. The proposed method was tested on real data from the Greek and Spanish stock markets and it successfully predicted the majority of stock market jitters.

3. Finally, motivated by the need for a method that could handle datasets with noisy measurements (such as high-level information that has been inferred from raw sensor data with some uncertainty), in Chapter 6 I have proposed a causal inference method based on the matching design framework that takes into account the uncertainty about the real values of the variables and attempts to improve the reliability of the study by maximising the probability that any confounding bias has been sufficiently eliminated. The method has been tested both on synthetic and real datasets and I have shown that it results in more reliable causal conclusions compared to existing

approaches.

Secondly, the data analysis conducted as part of this dissertation resulted in interesting findings and it featured the potential benefit that the analysis of smartphones and social media datasets could bring on finance and human behaviour studies. In detail:

1. In Chapter 4, I have provided evidence of causal links between the sentiment extracted from social media and prices of traded assets. I have also shown that it is feasible to detect bursty Twitter topics that are linked with strong stock market jitters. These findings demonstrate that social media could contain valuable information for the understanding of the dynamics that drive stock market prices.

2. While most studies so far examine the links between the sentiment extracted by social media and finance, in Chapter 4 I have shown that other features such as the geographical distribution of tweets and information about their authors could contain useful information for the understanding of strong stock market fluctuations.

3. In Chapter 5 I have shown that exercising and spending time outside home or university influences our stress level. Moreover, I have shown that different factors have different impact on the stress level of people with different personality characteristics. For example, I found that students with high extraversion score benefit more from spending time outside home or university, while students with high neuroticism score benefit more from exercising. Although this study was limited to a small population of university students and consequently we cannot derive general conclusions for the whole population, these findings could be a starting point for larger-scale studies.

An important lesson learnt during this thesis is that in order to derive reliable conclusions about the links between different factors of interest it is important to collect a rich variety of information that represents all the factors related to the study. However, sometimes it is not feasible to have access to all the necessary information due to privacy issues, limitations of the data mining methods or due to participants unwillingness to manually provide important information that cannot be collected by other means. In addition, it might happen the analysis is based on a human-generated sensor dataset that

has been collected in the past. Such datasets usually cannot be supplemented with additional information that may be required, since further data collection campaign might not be feasible. However, this does not mean that such datasets are useless and that any conclusions about the dependencies among the examined variables should be avoided. Instead, in this thesis I emphasise the importance of supporting any findings with an additional sensitivity analysis in order to assess their robustness in case of missing confounding variables. In any case, the limitations of the study need to be clearly described and any results should be interpreted with caution.

In addition, a significant insight from this study is that noisy and inaccurate measurements may result in misleading conclusions about the dependencies among the examined factors. Considering that most studies based on human-generated sensor data require the inference of high-level information from raw sensor data, inaccurate values resulted due to the limitations of the applied inference methods are prevalent. Nevertheless, this issue has been largely ignored at previous studies. In this thesis, I have discussed the importance of assessing the impact that uncertainties about the real data values may have on this type of studies.

Overall, the most important lesson learnt during this thesis is that, although the existing datasets and the devices and methods that are used for the data collection and analysis suffer from several limitations, the findings resulted from such studies are still valuable. However, the limitations of any study should be clearly stated and every effort needs to be made in order to improve the reliability of the results.

7.2 Future directions

This thesis raises several interesting questions and the findings of this study could be the starting point of further research in this domain.

Firstly, in order to detect causal links in human-generated sensor data a large number of factors relevant to the study needs to be measured. For example, participants' location context, activities, communication patterns and emotional state are important factors for any study on human behaviour. This information is highly sensitive. Several recent works have shown that users' identity can be inferred simply by analysing his/her location traces

[171, 172]. Obfuscation techniques, according to which noise is intentionally added into the data, have been proposed in order to protect users' identity [182, 183]. Protecting users' privacy while also preserving the validity of the data is an open research issue.

In addition, several factors important for the understanding of human behaviour are inferred from raw sensor data rather than directly measured. Consequently, the performance of the inference methods that are applied has a significant impact on the validity of the results. Accurate inference of the location context, activities, emotions or other high-level information from sensor measurements is a challenging research topic that requires further investigation.

Moreover, the findings presented in Chapter 4 are only a first step towards understanding the influence of social media on traded assets prices. These findings need to be further strengthened by repeating the studies on larger datasets. In particular, in this thesis, I have provided evidence of causal links between Twitter sentiment and the stock market prices of four tech companies. Repeating this study on traded assets of companies from different domains is essential in order to investigate the generalisability of the conclusions. Also, I have shown that it is feasible to detect bursty Twitter topics that influence a targeted stock market. Although my findings are confirmed for two different stock market datasets, an extensive evaluation larger stock market datasets for longer time periods needs to be conducted in order to investigate whether the results are consistent. This study requires several years of data in order to ensure that stock market jitters influenced by different factors are included in the dataset. However, obtaining such large datasets is hard for university-level projects.

Similarly, the findings presented in Chapter 5 about the impact of daily activities on the stress level of 48 participants are based on a relatively small dataset that includes only college students and consequently cannot be generalised for the whole population. Unfortunately, currently, there are no freely-available large datasets suitable for such a study. Recently, an effort to create a rich dataset from sensor data recorded from smartphones and smartwatches of participants with diverse demographic background has started [184]. This initiative could greatly facilitate the research on several domains that depend on this kind of data.

Finally, this work has been focused mainly on causality detection. Detecting causal

links enables us to better understand the human behaviour and other factors of interest that are influenced by it and could be a first step towards the prediction of events or behaviours of interest. Such prediction models would be an interesting future direction.

7.3 Outlook

To summarise, in this thesis I have attempted to utilise smartphone sensor data and social media data in order to understand human behaviour and phenomena influenced by it. I have developed novel techniques that aim at uncovering strong dependencies among the examined factors rather than mere correlation relationships. I hope that the findings of this study will inspire more research in this direction. Moreover, the methods developed during this dissertation are general and can be applied to any dataset with similar properties, thus I hope that they will be useful for researchers and practitioners in a variety of research areas. Although the availability of social media data and sensor data captured by smartphones and wearable devices is gradually increasing, in my opinion there has been little effort on enhancing the reliability of these data and ensuring that a demographically diverse pool of users has been taken into account as well as that the statistical power of the dataset is sufficient in order to derive valid conclusions. In this dissertation I have highlighted the importance of improving the reliability of the collected data and methods that are applied for the data processing. I hope that this work will instigate new studies that will attempt to reveal the structural dependencies among interesting factors and will improve our understanding on a variety of social and economic phenomena.

LIST OF REFERENCES

[1] Miluzzo, E., *et al.*: Sensing meets mobile social networks: the design, implementation and evaluation of the CenceMe application. In: Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems (SenSys'08), pp. 337–350 (2008). ACM

[2] Ravi, N., Dandekar, N., Mysore, P., Littman, M.L.: Activity recognition from accelerometer data. In: Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence (AAAI'05), vol. 3, pp. 1541–1546 (2005)

[3] Lu, H., Rabbi, M., Chittaranjan, G.T., Frauendorfer, D., Mast, M.S., Campbell, A.T., Gatica-Perez, D., Choudhury, T.: StressSense: detecting stress in unconstrained acoustic environments using smartphones. In: Proceedings of the 14th ACM International Conference on Ubiquitous Computing (UbiComp'12), pp. 351–360 (2012). ACM

[4] Rachuri, K.K., Musolesi, M., Mascolo, C., Rentfrow, P.J., Longworth, C., Aucinas, A.: Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In: Proceedings of the 14th ACM International Conference on Ubiquitous Computing (UbiComp'10), pp. 281–290 (2010). ACM

[5] Ma, Y., Xu, B., Bai, Y., Sun, G., Zhu, R.: Daily mood assessment based on mobile phone sensing. In: Proceedings of the Ninth International Conference on Wearable and Implantable Body Sensor Networks (BNS'12), pp. 142–147 (2012). IEEE

[6] Bauer, G., Lukowicz, P.: Can smartphones detect stress-related changes in the behaviour of individuals? In: Proceedings of the International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops'12), pp. 423–426 (2012). IEEE

[7] Bogomolov, A., Lepri, B., Pianesi, F.: Happiness recognition from mobile phone data. In: Proceedings of the International Conference on Social Computing (SocialCom'13), pp. 790–795 (2013). IEEE

[8] Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., Pentland, A.S.: Daily stress recognition from mobile phone data, weather conditions and individual traits. In: Proceedings of Multimedia'14, pp. 477–486 (2014). ACM

[9] Canzian, L., Musolesi, M.: Trajectories of Depression: Unobtrusive Monitoring of Depressive States by means of Smartphone Mobility Traces Analysis. In: Proceedings of 14th ACM International Conference on Ubiquitous Computing (UbiComp'15), pp. 1293–1304 (2015). ACM

[10] Liew, C.S., Wah, T.Y., Shuja, J., Daghighi, B., *et al.*: Mining personal data using smartphones and wearable devices: A survey. Sensors **15**(2), 4430–4469 (2015)

[11] Jaimes, L.G., Vergara-Laurens, I.J., Raij, A.: A survey of incentive techniques for mobile crowd sensing. IEEE Internet of Things Journal **2**(5), 370–380 (2015)

[12] Asur, S., Huberman, B.A.: Predicting the future with social media. In: In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 1, pp. 492–499 (2010). IEEE

[13] Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In: Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM'10), vol. 10, pp. 178–185 (2010). AAAI

[14] Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science **2**(1), 1–8 (2011)

[15] Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in twitter: The million follower fallacy. In: Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM'10), pp. 10–17 (2010). AAAI

[16] Anger, I., Kittl, C.: Measuring influence on twitter. In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'11) (2011). ACM

[17] Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.D.I., Marlow, C., Settle, J.E., Fowler, J.H.: A 61-million-person experiment in social influence and political mobilization. Nature **489**(7415), 295–298 (2012)

[18] Muchnik, L., Aral, S., Taylor, S.J.: Social influence bias: A randomized experiment. Science **341**(6146), 647–651 (2013)

[19] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., Barabási, A.-L.: Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences **104**(18), 7332–7336 (2007)

[20] Colizza, V., Barrat, A., Barthélemy, M., Vespignani, A.: The role of the airline transportation network in the prediction and predictability of global epidemics. Proceedings of the National Academy of Sciences of the United States of America **103**(7), 2015–2020 (2006)

[21] Marsch, L.A.: Leveraging technology to enhance addiction treatment and recovery. Journal of Addictive Diseases **31**(3), 313–318 (2012)

[22] Pejovic, V., Musolesi, M.: Anticipatory mobile computing for behaviour change interventions. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct, pp. 1025–1034 (2014). ACM

[23] Moturu, S.T., Khayal, I., Aharony, N., Pan, W., Pentland, A.S.: Using social sensing to understand the links between sleep, mood, and sociability. In: Proceedings of the International Conference on Social Computing (SocialCom'11), pp. 208–214 (2011). IEEE

[24] Madan, A., Moturu, S.T., Lazer, D., Pentland, A.S.: Social sensing: obesity, unhealthy eating and exercise in face-to-face networks. In: Proceedings of Wireless Health 2010, pp. 104–110 (2010). ACM

[25] Concato, J., Shah, N., Horwitz, R.I.: Randomized, controlled trials, observational studies, and the hierarchy of research designs. New England Journal of Medicine **342**(25), 1887–1892 (2000)

[26] Shadish, W.R., Cook, T.D., Campbell, D.T.: Experimental and Quasi-experimental Designs for Generalized Causal Inference, (2002). Wadsworth Cengage Learning

[27] Spirtes, P., Glymour, C.N., Scheines, R.: Causation, Prediction, and Search, (2000). MIT press

[28] Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research **46**(3), 399–424 (2011)

[29] Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science **2**(1), 1–8 (2011)

[30] Lane, N.D., Lin, M., Mohammod, M., Yang, X., Lu, H., Cardone, G., Ali, S., Doryab, A., Berke, E., Campbell, A.T., *et al.*: Bewell: Sensing sleep, physical activities and social interactions to promote wellbeing. Mobile Networks and Applications **19**(3), 345–359 (2014)

[31] Fan, J., Han, F., Liu, H.: Challenges of big data analysis. National Science Review **1**(2), 293–314 (2014)

[32] Rubin, D.B.: Estimating causal effects from large data sets using propensity scores. Annals of Internal Medicine **127**, 757–763 (1997)

[33] Mouchart, M., Russo, F., Wunsch, G.: Structural modelling, exogeneity, and causality. In: Causal Analysis in Population Studies, pp. 59–82 (2009). Springer

[34] Pearl, J.: Causality: Models, Reasoning and Inference, (2009). Cambridge University Press

[35] Wang, R., *et al.*: Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of 14th ACM International Conference on Ubiquitous Computing (UbiComp'14), pp. 3–14 (2014). ACM

[36] Malkiel, B.G.: Efficient market hypothesis. The New Palgrave: Finance. Norton, New York, 127–134 (1989)

[37] Sornette, D.: Endogenous versus exogenous origins of crises. In: Extreme Events in Nature and Society, pp. 95–119 (2006). Springer

[38] Johansen, A., Sornette, D.: Endogenous versus exogenous crashes in financial markets. Available at SSRN 344980 (2002)

[39] Zhang, X., Fuehres, H., Gloor, P.A.: Predicting stock market indicators through Twitter "I hope it is not as bad as I fear". In: Proceedings of the 2nd Collaborative Innovation Networks Conference, vol. 26, pp. 55–62 (2011). Elsevier

[40] Sul, H.K., Dennis, A.R., Yuan, L.I.: Trading on twitter: The financial information content of emotion in social media. In: Proceedings of the 47th International Conference on System Sciences (HICSS), pp. 806–815 (2014). IEEE

[41] Luo, X., Zhang, J., Duan, W.: Social media and firm equity value. Information Systems Research **24**(1), 146–163 (2013)

[42] Zheludev, I., Smith, R., Aste, T.: When can social media lead financial markets? Scientific Reports **4**, 4213 (2014)

[43] Piñeiro-Chousa, J.R., López-Cabarcos, M.Á., Pérez-Pico, A.M.: Examining the influence of stock market variables on microblogging sentiment. Journal of Business Research **69**(6), 2087–2092 (2016)

[44] Zhang, X., Fuehres, H., Gloor, P.A.: Predicting asset value through Twitter buzz. In: Advances in Collective Intelligence, pp. 23–34 (2012). Springer

[45] Checkley, M., Higón, D.A., Alles, H.: The hasty wisdom of the mob: How market sentiment predicts stock market behavior. Expert Systems with Applications **77**, 256–263 (2017)

[46] Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., Deng, X.: Exploiting topic based twitter sentiment for stock prediction. In: In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, vol. 2, pp. 24–29 (2013). Association for Computational Linguistics (ACL)

[47] Chen, H., De, P., Hu, Y.J., Hwang, B.-H.: Wisdom of crowds: The value of stock opinions transmitted through social media. Review of Financial Studies **27**(5), 1367–1403 (2014)

[48] Porshnev, A., Redkin, I., Shevchenko, A.: Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis. In: Proceedings of the 13th IEEE International Conference on Data Mining Workshops (ICDMW), pp. 440–444 (2013). IEEE

[49] Nguyen, T.H., Shirai, K., Velcin, J.: Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications **42**(24), 9603–9611 (2015)

[50] Vu, T.T., Chang, S., Ha, Q.T., Collier, N.: An experiment in integrating sentiment features for tech stock prediction in twitter. In: Proceedings of the 24th International Conference on Computational Linguistics, p. 23 (2012). Citeseer

[51] Yang, S.Y., Mo, S.Y.K., Liu, A.: Twitter financial community sentiment and its predictive relationship to stock market movement. Quantitative Finance **15**(10), 1637–1656 (2015)

[52] Ruiz, E.J., Hristidis, V., Castillo, C., Gionis, A., Jaimes, A.: Correlating financial time series with micro-blogging activity. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM'12), pp. 513–522 (2012). ACM

[53] Nofer, M., Hinz, O.: Using twitter to predict the stock market. Business & Information Systems Engineering **57**(4), 229–242 (2015)

[54] Preis, T., Reith, D., Stanley, H.E.: Complex dynamics of our economic life on different scales: insights from search engine query data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **368**(1933), 5707–5719 (2010)

[55] Preis, T., Moat, H.S., Stanley, H.E.: Quantifying trading behavior in financial markets using Google Trends. Scientific Reports **3** (2013)

[56] Moat, H.S., Curme, C., Avakian, A., Kenett, D.Y., Stanley, H.E., Preis, T.: Quantifying Wikipedia usage patterns before stock market moves. Scientific Reports **3** (2013)

[57] Schumaker, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems (TOIS) **27**(2), 12 (2009)

[58] Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T., Sakurai, A.: Combining technical analysis with sentiment analysis for stock price prediction. In: Proceedings of IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing (DASC'11), pp. 800–807 (2011). IEEE

[59] Kenett, D.Y., Tumminello, M., Madi, A., Gur-Gershgoren, G., Mantegna, R.N., Ben-Jacob, E.: Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. PloS one **5**(12), 15032 (2010)

[60] Chi, K.T., Liu, J., Lau, F.C.: A network perspective of the stock market. Journal of Empirical Finance **17**(4), 659–667 (2010)

[61] Souza, T.T., Aste, T.: A nonlinear impact: evidences of causal effects of social media on market prices. arXiv preprint arXiv:1601.04535 (2016)

[62] Tsapeli, F., Musolesi, M., Tino, P.: Non-parametric causality detection: An application to social media and financial data. Physica A: Statistical Mechanics and its Applications **483**, 139–155 (2017)

[63] Lachanski, M., Pav, S.: Shy of the character limit:" twitter mood predicts the stock market" revisited. Econ Journal Watch **14**(3), 302 (2017)

[64] Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., Mozetič, I.: The effects of twitter sentiment on stock price returns. PloS one **10**(9), 0138441 (2015)

[65] Strauß, N., Vliegenthart, R., Verhoeven, P.: Intraday news trading: The reciprocal relationships between the stock market and economic news. Communication Research, 0093650217705528 (2017)

[66] Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the 10th International Workshop on Multimedia Data Mining (2010). ACM

[67] Weng, J., Lee, B.-S.: Event detection in twitter. In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM'11), pp. 401–408 (2011). AAAI

[68] Xie, W., Zhu, F., Jiang, J., Lim, E.-P., Wang, K.: Topicsketch: Real-time bursty topic detection from twitter. In: Proceedings of the 13th International Conference on Data Mining (ICDM'13), pp. 837–846 (2013). IEEE

[69] Li, C., Sun, A., Datta, A.: Twevent: segment-based event detection from tweets. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12), pp. 155–164 (2012). ACM

[70] Lee, R., Sumiya, K.: Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, pp. 1–10 (2010). ACM

[71] Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web (WWW'10), pp. 851–860 (2010). ACM

[72] Liu, Y.: Word of mouth for movies: Its dynamics and impact on box office revenue. Journal of marketing **70**(3), 74–89 (2006)

[73] O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. ICWSM **11**(122-129), 1–2 (2010)

[74] Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In: Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM'10), vol. 10, pp. 178–185. AAAI, ??? (2010)

[75] Bae, Y., Lee, H.: Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. Journal of the Association for Information Science and Technology **63**(12), 2521–2535 (2012)

[76] Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., Stringhini, G., Blackburn, J.: The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources. arXiv preprint arXiv:1705.06947 (2017)

[77] Wong, F.M.F., Sen, S., Chiang, M.: Why watching movie tweets won't tell the whole story? In: Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks, pp. 61–66 (2012). ACM

[78] Chung, J.E., Mustafaraj, E.: Can collective sentiment expressed on twitter predict political elections? In: AAAI, vol. 11, pp. 1770–1771 (2011)

[79] Rosenquist, J.N., Murabito, J., Fowler, J.H., Christakis, N.A.: The spread of alcohol consumption behavior in a large social network. Annals of Internal Medicine **152**(7), 426–433 (2010)

[80] Fowler, J.H., Christakis, N.A., *et al.*: Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. British Medical Journal **337**, 2338 (2008)

[81] Cacioppo, J.T., Fowler, J.H., Christakis, N.A.: Alone in the crowd: the structure and spread of loneliness in a large social network. Journal of Personality and Social Psychology **97**(6), 977 (2009)

[82] Rosenquist, J.N., Fowler, J.H., Christakis, N.A.: Social network determinants of depression. Molecular psychiatry **16**(3), 273–281 (2011)

[83] Christakis, N.A., Fowler, J.H.: Social contagion theory: examining dynamic social networks and human behavior. Statistics in Medicine **32**(4), 556–577 (2013)

[84] Hills, P., Argyle, M.: Positive moods derived from leisure and their relationship to happiness and personality. Personality and Individual Differences **25**(3), 523–535 (1998)

[85] Seligman, M.E., Steen, T.A., Park, N., Peterson, C.: Positive psychology progress: empirical validation of interventions. American Psychologist **60**(5), 410 (2005)

[86] Hayes, B.E.: Measuring Customer Satisfaction: Survey Design, Use, and Statistical Analysis Methods, (1998). ASQ Quality Press

[87] Krn, S.: Analysing customer satisfaction and quality in construction the case of public and private customers. Nordic Journal of Surveying and Real Estate Research **2** (2014)

[88] Trivellas, P., Reklitis, P., Platis, C.: The effect of job related stress on employees satisfaction: A survey in health care. Procedia-social and behavioral sciences **73**, 718–726 (2013)

[89] Ibrahim, M.E., Perez, A.O.: Effects of organizational justice, employee satisfaction, and gender on employees' commitment: evidence from the uae. International Journal of Business and Management **9**(2), 45 (2014)

[90] Liao, L., Fox, D., Kautz, H.: Extracting places and activities from gps traces using hierarchical conditional random fields. The International Journal of Robotics Research **26**(1), 119–134 (2007)

[91] Do, T.M.T., Gatica-Perez, D.: The places of our lives: Visiting patterns and automatic labeling from longitudinal smartphone data. IEEE Transactions on Mobile Computing **13**(3), 638–648 (2014)

[92] Li, L., Chen, J.-H.: Emotion recognition using physiological signals. In: Advances in Artificial Reality and Tele-existence, pp. 437–446 (2006). Springer

[93] Lisetti, C.L., Nasoz, F.: Using noninvasive wearable computers to recognize human emotions from physiological signals. EURASIP Journal on Advances in Signal Processing **2004**(11), 929414 (2004)

[94] Hernandez, J., Picard, R.W.: Senseglass: using google glass to sense daily emotions. In: Proceedings of the Adjunct Publication of the 27th Annual ACM Symposium on User Interface Software and Technology, pp. 77–78 (2014). ACM

[95] Lee, Y.S., Cho, S.B.: Activity recognition using hierarchical hidden markov models on a smartphone with 3d accelerometer. In: Hybrid Artificial Intelligent Systems (HAIS 2011). Lecture Notes in Computer Science, vol. 6678, pp. 460–467 (2011). Springer

[96] Su, X., Tong, H., Ji, P.: Activity recognition with smartphone sensors. Tsinghua Science and Technology **19**(3), 235–249 (2014)

[97] Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In: International Workshop on Ambient Assisted Living, pp. 216–223 (2012). Springer

[98] Lara, O.D., Labrador, M.A.: A survey on human activity recognition using wearable sensors. IEEE Communications Surveys and Tutorials **15**(3), 1192–1209 (2013)

[99] Lathia, N., Sandstrom, G.M., Mascolo, C., Rentfrow, P.J.: Happier people live more active lives: Using smartphones to link happiness and physical activity. PLoS ONE **12**(1), 0160589 (2017)

[100] Raveau, S., Ghorpade, A., Zhao, F., Abou-Zeid, M., Zegras, C., Ben-Akiva, M.: Smartphone-based survey for real-time and retrospective happiness related to travel and activities. Transportation Research Record: Journal of the Transportation Research Board (2566), 102–110 (2016)

[101] Ben-Akiva, M., Abou-Zeid, M.: Capturing the relationship between motility, mobility and well-being using smart phones. Technical report (2016)

[102] Servia-Rodríguez, S., Rachuri, K.K., Mascolo, C., Rentfrow, P.J., Lathia, N., Sandstrom, G.M.: Mobile sensing at the service of mental well-being: a large-scale longitudinal study. In: Proceedings of the 26th International Conference on World Wide Web, pp. 103–112 (2017). International World Wide Web Conferences Steering Committee

[103] Ma, Y., Xu, B., Bai, Y., Sun, G., Zhu, R.: Daily mood assessment based on mobile phone sensing. In: Proceedings of the Ninth International Conference on Wearable and Implantable Body Sensor Networks (BSN'12), pp. 142–147 (2012). IEEE

[104] LiKamWa, R., Liu, Y., Lane, N.D., Zhong, L.: Can your smartphone infer your mood. In: PhoneSense Workshop, pp. 1–5 (2011)

[105] Alvarez-Lozano, J., Osmani, V., Mayora, O., Frost, M., Bardram, J., Faurholt-Jepsen, M., Kessing, L.V.: Tell me your apps and i will tell you your mood: correlation of apps usage with bipolar disorder state. In: Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments, p. 19 (2014). ACM

[106] LiKamWa, R., Liu, Y., Lane, N.D., Zhong, L.: Moodscope: Building a mood sensor from smartphone usage patterns. In: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services, pp. 389–402 (2013). ACM

[107] Murnane, E.L., Abdullah, S., Matthews, M., Kay, M., Kientz, J.A., Choudhury, T., Gay, G., Cosley, D.: Mobile manifestations of alertness: connecting biological rhythms with patterns of smartphone app use. In: Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI'16), pp. 465–477 (2016). ACM

[108] Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., Campbell, A.T.: Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'14), pp. 3–14 (2014). ACM

[109] Wang, R., Harari, G., Hao, P., Zhou, X., Campbell, A.T.: Smartgpa: how smartphones can assess and predict academic performance of college students. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'15), pp. 295–306 (2015). ACM

[110] Grünerbl, A., Muaremi, A., Osmani, V., Bahle, G., Oehler, S., Tröster, G., Mayora, O., Haring, C., Lukowicz, P.: Smartphone-based recognition of states and state changes in bipolar disorder patients. IEEE Journal of Biomedical and Health Informatics **19**(1), 140–148 (2015)

[111] Glenn, T., Monteith, S.: New measures of mental state and behavior based on data collected from sensors, smartphones, and the internet. Current Psychiatry Reports **16**(12), 1–10 (2014)

[112] Burton, C., McKinstry, B., Tătar, A.S., Serrano-Blanco, A., Pagliari, C., Wolters, M.: Activity monitoring in patients with depression: a systematic review. Journal of Affective Disorders **145**(1), 21–28 (2013)

[113] Canzian, L., Musolesi, M.: Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'15), pp. 1293–1304 (2015). ACM

[114] Holland, P.W.: Statistics and causal inference. Journal of the American Statistical Association **81**(396), 945–960 (1986)

[115] Morgan, S.L., Winship, C.: Counterfactuals and Causal Inference, (2014). Cambridge University Press

[116] Rubin, D.: Causal inference using potential outcomes. Journal of the American Statistical Association (2011)

[117] Stuart, E.A.: Matching methods for causal inference: A review and a look forward. Statistical Science: a Review Journal of the Institute of Mathematical Statistics **25**(1), 1 (2010)

[118] Rosenbaum, P.R.: Observational studies, pp. 1–17 (2002). Springer

[119] Gu, X.S., Rosenbaum, P.R.: Comparison of multivariate matching methods: Structures, distances, and algorithms. Journal of Computational and Graphical Statistics **2**(4), 405–420 (1993)

[120] Kallus, N.: A framework for optimal matching for causal inference. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (PMLR), vol. 54, pp. 372–381 (2017)

[121] Lunceford, J.K., Davidian, M.: Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in medicine **23**(19), 2937–2960 (2004)

[122] Diamond, A., Sekhon, J.S.: Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. Review of Economics and Statistics **95**(3), 932–945 (2013)

[123] Imai, K., King, G., Stuart, E.A.: Misunderstandings between experimentalists and observationalists about causal inference. Journal of the Royal Statistical Society: Series A (Statistics in Society) **171**(2), 481–502 (2008)

[124] Lu, B., Zanutto, E., Hornik, R., Rosenbaum, P.R.: Matching with doses in an observational study of a media campaign against drug abuse. Journal of the American Statistical Association **96**(456), 1245–1253 (2001)

[125] Hirano, K., Imbens, G.W.: The propensity score with continuous treatments. Applied Bayesian modeling and causal inference from incomplete-data perspectives **226164**, 73–84 (2004)

[126] Rosenbaum, P.R., Rubin, D.B.: Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society. Series B (Methodological), 212–218 (1983)

[127] Lehmann, E.L., D'abrera, H.: Nonparametrics: Statistical Methods Based on Ranks., (1975). Holden-Day

[128] Zhang, K., Peters, J., Janzing, D.: Kernel-based conditional independence test and application in causal discovery. In: In Uncertainty in Artificial Intelligence (2011). Citeseer

[129] Kalisch, M., Bühlmann, P.: Estimating high-dimensional directed acyclic graphs with the pc-algorithm. Journal of Machine Learning Research **8**(Mar), 613–636 (2007)

[130] Nawrath, J., Romano, M.C., Thiel, M., Kiss, I.Z., Wickramasinghe, M., Timmer, J., Kurths, J., Schelter, B.: Distinguishing direct from indirect interactions in oscillatory networks with multiple time scales. Physical Review Letters **104**(3), 038701 (2010)

[131] Peters, J., Janzing, D., Schölkopf, B.: Causal inference on time series using restricted structural equation models. In: Advances in Neural Information Processing Systems (NIPS 2013), pp. 154–162 (2013)

[132] Preacher, K.J.: Quantifying parsimony in structural equation modeling. Multivariate Behavioral Research **41**(3), 227–259 (2006)

[133] Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society, 424–438 (1969)

[134] Stock, J.H., Watson, M.W.: Introduction to econometrics **104** (2003)

[135] Barrett, A.B., Barnett, L., Seth, A.K.: Multivariate Granger causality and generalized variance. Physical Review E **81**(4), 041907 (2010)

[136] Entner, D., Hoyer, P.O.: On causal discovery from time series data using FCI. In: Proceedings of the Fifth European Workshop on Probabilistic Graphical Models, pp. 121–128 (2010)

[137] Schreiber, T.: Measuring information transfer. Physical Review Letters **85**(2), 461 (2000)

[138] Barnett, L., Barrett, A.B., Seth, A.K.: Granger causality and transfer entropy are equivalent for Gaussian variables. Physical Review Letters **103**(23), 238701 (2009)

[139] Lizier, J.T., Prokopenko, M., Zomaya, A.Y.: Local information transfer as a spatiotemporal filter for complex systems. Physical Review E **77**(2), 026110 (2008)

[140] Pompe, B., Runge, J.: Momentary information transfer as a coupling measure of time series. Physical Review E **83**(5), 051122 (2011)

[141] Runge, J., Heitzig, J., Petoukhov, V., Kurths, J.: Escaping the curse of dimensionality in estimating multivariate transfer entropy. Physical Review Letters **108**(25), 258701 (2012)

[142] Tsapeli, F., Musolesi, M.: Investigating causality in human behavior from smartphone sensor data: a quasi-experimental approach. EPJ Data Science **4**(1), 1 (2015)

[143] Tsapeli, F., Peter, T., Musolesi, M.: Probabilistic matching: Causal inference under measurement errors. In: Proceedings of International Joint Conference Of Neural Networks (IJCNN) (2017). IEEE

[144] Sekhon, J.S.: Opiates for the matches: Matching methods for causal inference. Annual Review of Political Science **12**, 487–508 (2009)

[145] Austin, P.C.: Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. Pharmacoepidemiology and Drug Safety **17**(12), 1202–1217 (2008)

[146] Harder, V.S., Stuart, E.A., Anthony, J.C.: Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychological Methods **15**(3), 234 (2010)

[147] Schneeweiss, S., Rassen, J.A., Glynn, R.J., Avorn, J., Mogun, H., Brookhart, M.A.: High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology **20**(4), 512 (2009)

[148] Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research **46**(3), 399–424 (2011)

[149] Thelwall, M.: Heart and soul: Sentiment strength detection in the social web with Sentistrength. Cyberemotions, 1–14 (2013)

[150] Austin, P.C.: The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Medical Decision Making **29**(6), 661–677 (2009)

[151] Rangel, J.G.: Macroeconomic news, announcements, and stock market jump intensity dynamics. Journal of Banking & Finance **35**(5), 1263–1276 (2011)

[152] Kaminsky, G.L., Schmukler, S.L.: What triggers market jitters?: A chronicle of the asian crisis. Journal of International Money and Finance **18**(4), 537–560 (1999)

[153] Bodnaruk, A., Loughran, T., McDonald, B.: Using 10-k text to gauge financial constraints. Journal of Financial and Quantitative Analysis **50**(4), 623–646 (2015)

[154] Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. Text Mining: Applications and Theory, 1–20 (2010)

[155] Luss, R., dAspremont, A.: Predicting abnormal returns from news using text classification. Quantitative Finance **15**(6), 999–1012 (2015)

[156] Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)

[157] Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G., Cunningham, S.J.: Weka: Practical Machine Learning Tools and Techniques with Java Implementations (1999)

[158] Hall, M.A., Smith, L.A.: Feature subset selection: a correlation based filter approach. In: Proceedings of the International Conference on Neural Information Processing and Intelligent Information Systems, pp. 855–858 (1997)

[159] Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: International Conference on Machine Learning (ICML'1997), vol. 97, pp. 412–420 (1997)

[160] Shalizi, C.R., Shalizi, K.L.: Blind construction of optimal nonlinear recursive predictors for discrete sequences. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp. 504–511 (2004). AUAI Press

[161] Ekroot, L., Cover, T.M.: The entropy of markov trajectories. IEEE Transactions on Information Theory **39**(4), 1418–1421 (1993)

[162] Blanc, J.-L., Pezard, L., Lesne, A.: Delay independence of mutual-information rate of two symbolic sequences. Physical Review E **84**(3), 036214 (2011)

[163] Liao, L., Fox, D., Kautz, H.: Location-based activity recognition. Advances in Neural Information Processing Systems **18**, 787 (2006)

[164] Montoliu, R., Blom, J., Gatica-Perez, D.: Discovering places of interest in everyday life from smartphone data. Multimedia tools and applications **62**(1), 179–207 (2013)

[165] Chon, Y., Lane, N.D., Li, F., Cha, H., Zhao, F.: Automatically characterizing places with opportunistic crowdsensing using smartphones. In: Proceedings of the 14th ACM International Conference on Ubiquitous Computing (UbiComp'12), pp. 481–490 (2012). ACM

[166] Lau, S.L., König, I., David, K., Parandian, B., Carius-Düssel, C., Schultz, M.: Supporting patient monitoring using activity recognition with a smartphone. In: Proceedings of the 7th International Symposium on Wireless Communication Systems (ISWCS), pp. 810–814 (2010). IEEE

[167] Fisher, R.: Statistical methods for research workers. In: Breakthroughs in Statistics, pp. 66–70 (1992). Springer

[168] Google Maps Places API. `https://developers.google.com/maps/documentation/javascript/places`. Accessed: 2015-06-21

[169] Bolger, N., Schilling, E.A.: Personality and the problems of everyday life: The role of neuroticism in exposure and reactivity to daily stressors. Journal of Personality **59**(3), 355–386 (1991)

[170] Diamond, A., Sekhon, J.S.: Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. Review of Economics and Statistics **95**(3), 932–945 (2013)

[171] Rossi, L., Walker, J., Musolesi, M.: Spatio-temporal techniques for user identification by means of gps mobility data. EPJ Data Science **4**(1), 11 (2015)

[172] De Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: The privacy bounds of human mobility. Scientific reports **3**, 1376 (2013)

[173] Bamman, D., O'Connor, B., Smith, N.: Censorship and deletion practices in Chinese social media. First Monday **17**(3) (2012)

[174] Fornell, C., Larcker, D.F.: Structural equation models with unobservable variables and measurement error: Algebra and statistics. Journal of Marketing Research, 382–388 (1981)

[175] Bentler, P.M.: Multivariate analysis with latent variables: Causal modeling. Annual Review of Psychology **31**(1), 419–456 (1980)

[176] UCI Machine Learning Repository: microblogPCU dataset. `http://aiweb.techfak.uni-bielefeld.de/content/bworld-robot-control-software/`. Accessed: 2016-10-13

[177] Benevenuto, F., *et al.*: Detecting spammers on twitter. In: In Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS), vol. 6, p. 12 (2010)

[178] Platt, J., *et al.*: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers **10**(3), 61–74 (1999)

[179] Chon, Y., *et al.*: Automatically characterizing places with opportunistic crowdsensing using smartphones. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12), pp. 481–490 (2012). ACM

[180] Zhou, C., *et al.*: Discovering personally meaningful places: An interactive clustering approach. ACM Transactions on Information Systems (TOIS) **25**(3), 12 (2007)

[181] Holmes, G., Donkin, A., Witten, I.H.: Weka: A machine learning workbench. In: Proceedings of the Conference on Intelligent Information Systems, pp. 357–361 (1994). IEEE

[182] Luo, C., Fylakis, A., Partala, J., Klakegg, S., Goncalves, J., Liang, K., Seppänen, T., Kostakos, V.: A data hiding approach for sensitive smartphone data. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'16), pp. 557–468 (2016). ACM

[183] Xiao, Q., Chen, J., Yu, L., Li, H., Zhu, H., Li, M., Ren, K.: Poster: Locmask: A location privacy protection framework in android system. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS'14), pp. 1526–1528 (2014). ACM

[184] Welbourne, E., Tapia, E.M.: Crowdsignals: a call to crowdfund the community's largest mobile dataset. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp'14), pp. 873–877 (2014). ACM

Appendices

APPENDIX A

LISTS OF TWITTER KEYWORDS

The following table presents the keywords that were used for tweets filtering during the financial event detection study presented in Section 4.2. Keywords were extracted from the wikipedia webpage on European Financial crisis using RAKE keyword extractor.

Keyword	**Score**
united kingdom	4.0
european union	4.0
fiscal compact	4.0
odious debt	4.0
euro area	4.0
sovereign debt	4.0
european commission	3.92307692308
gdp ratio	3.90909090909
public debt	3.9
year maturity	3.85714285714
great recession	3.83333333333
citation needed	3.81818181818
spending cuts	3.8125
structural reforms	3.8125
main article	3.81060606061
financial markets	3.79653679654
growth pact	3.79166666667

debt levels	3.78571428571
interest rate	3.78571428571
economic growth	3.72715053763
maastricht treaty	3.7
greek government	3.66920374707
year bonds	3.65217391304
bailout programme	3.57509157509
record high	3.57142857143
european countries	3.56451612903
debt level	3.55555555556
european banks	3.55555555556
deficit spending	3.55324074074
budget deficit	3.54074074074
debt crisis	3.53968253968
financial crisis	3.47907647908
restore competitiveness	3.46666666667
bailout package	3.41025641026
budget deficits	3.38333333333
austerity measures	3.32894736842
eurozone countries	3.00896057348
eurozone crisis	2.98412698413
funds	1.80769230769
budget	1.8
growth	1.79166666667
measures	1.75
deficit	1.74074074074
private	1.73684210526
bailout	1.71794871795
government	1.70491803279

package	1.69230769231
states	1.6875
increase	1.66666666667
bonds	1.65217391304
amount	1.61538461538
announced	1.6
terms	1.6
capital	1.59090909091
deficits	1.58333333333
austerity	1.57894736842
countries	1.56451612903
level	1.55555555556
economies	1.55555555556
banks	1.55555555556
crisis	1.53968253968
default	1.53846153846
governments	1.53333333333
agreed	1.52941176471
proposed	1.5
stability	1.5
including	1.5
taxes	1.5
raise	1.5
economy	1.5
lower	1.5
support	1.46153846154
eurozone	1.44444444444
currency	1.44444444444
provide	1.44444444444

based	1.42857142857
economists	1.42857142857
elections	1.42857142857
progress	1.42857142857
wages	1.42857142857
money	1.41666666667
state	1.41666666667
agreement	1.4
write	1.4
investors	1.375
finance	1.33333333333
found	1.33333333333
series	1.33333333333
suggested	1.33333333333
borrowing	1.33333333333
forecast	1.3
inflation	1.28571428571
prevent	1.28571428571
balance	1.28571428571
required	1.28571428571
years	1.25
months	1.25
businesses	1.25
greece	1.24705882353
country	1.24444444444
collateral	1.22222222222
future	1.22222222222
short	1.22222222222
focus	1.2

benefit	1.2
conditional	1.2
france	1.2
collapse	1.2
portugal	1.2
services	1.2
attempt	1.2
bring	1.16666666667
follow	1.16666666667
downgraded	1.16666666667
ensure	1.16666666667
financed	1.16666666667
proposal	1.16666666667
create	1.16666666667
resulting	1.16666666667
effect	1.16666666667
ireland	1.15384615385
finland	1.15384615385
implementation	1.14285714286
called	1.14285714286
germany	1.13333333333
troika	1.11764705882
return	1.11111111111
order	1.11111111111
result	1.08333333333
october 2012	1.08333333333
june 2012	1.07692307692
spain	1.06666666667
europe	1.05882352941

hoped	1.0
japan	1.0
argued	1.0
cyprus	1.0
issues	1.0
pledged	1.0
condition	1.0
moody	1.0
september 2011	1.0
break	1.0
comply	1.0
netherlands	1.0
borrow	1.0
yield	1.0
continue	1.0
difficult	1.0
reduce	1.0
estimated	1.0
contributed	1.0
formation	1.0
beginning	1.0
figure	1.0
austria	1.0
total	1.0
addition	1.0
leading	1.0
expected	1.0
5 billion	1.0
replaced	1.0

making	1.0
february 2012	1.0
forced	1.0
italy	1.0
currencies	1.0
centre	1.0
response	1.0
purchase	1.0
cutting	1.0

Table A.1: RAKE keywords along with the corresponding score

APPENDIX B

UPDATED SENTISTRENGTH DICTIONARY

Table B.1 presents the list of the keywords added at the SentiStrength classifier that have been used for the Tweets sentiment classification along with the assigned positivity or negativity weight.

Keyword	**SentiStrenth Score**
gain*	5
grow*	3
high*	5
lift*	5
loss*	-5
low*	-4
lunch*	2
miss*	-3
more	2
promising	4
rally	2
reject*	-4
rise	3
sale	-2
sales	3
sell	-2

soar*	5
sour	-5
spike*	5
strong	5
surge*	5
top	5
tough	-4
under	-3
up	2
weak*	-5
win	3
win*	4
wow*	5

Table B.1: Added or modified keywords on SentiStrength along with the assigned positivity/negativity score.