Modelling the impact of climate change on health

Muhammad Saiful Islam

Department of Business Information Systems, Faculty of Science and Technology

Modelling the Impact of Climate Change on Health

Muhammad Saiful Islam

A thesis submitted in partial fulfilment of the requirements of the

University of Westminster for the degree of

Doctor of Philosophy

February 2014

Modelling the Impact of climate change on Health

MUHAMMAD SAIFUL ISLAM

Health and Social Care Modelling Group (HSCMG), Department of Business Information Systems, Faculty of Science and Technology, University of Westminster, 115 New Cavendish Street, London W1W 6UW, United Kingdom

To my parents and wife

Abstract

The main objective of this thesis is to develop a robust statistical model by accounting the non-linear relationships between hospital admissions due to lower respiratory (LR) disease and factors of climate and pollution, and their delayed effects on hospital admissions. This study also evaluates whether the model fits can be improved by considering the non-linearity of the data, delayed effect of the significant factors, and thus calculate threshold levels of the significant climate and pollution factors for emergency LR hospital admissions. For the first time three unique administrative datasets were merged: Hospital Episode Statistics, Met office observational data for climate factors, and data from London Air Quality Network.

The results of the final GLM, showed that daily temperature, rain, wind speed, sun hours, relative humidity, and PM10 significantly affected the LR emergency hospital admissions. Then, we developed a Distributed lag non-linear model (DLNM) model considering the significant climate and pollution factors. Time and 'day of the week' was incorporated as linear terms in the final model.

Higher temperatures around $\geq 27^0$C a quicker effect of 0-2 days lag but lower temperatures ($\leq 0^0$C) had delayed effects of 5-25 days lag. Humidity showed a strong immediate effect (0-3 days) of the low relative humidity at around $\leq$40% and a moderate effect for higher humidity ($\geq$80%) with lag period of 0-2 days. Higher PM10 around $\geq$70-$\mu g/m^3$ has both shorter (0-3 days) and longer lag effects (15-20 days) but the latter one is stronger comparatively. A

strong effect of wind speed around ≥25 knots showed longer lag period of 8-15 days. There is a moderate effect for a shorter lag period of 0-3 days for lower wind speed (approximately 2 knots). We also notice a stronger effect of sun hours around ≥14 hours having a longer lag period of 15-20 days and moderate effect between 1-2 hours of 5-12 days lag. Similarly, higher amount of rain (≥30mm) has stronger effects, especially for the shorter lag of 0-2 days and longer lag of 7-10 days.

So far, very little research has been carried out on DLNM model in such research area and setting. This PhD research will contribute to the quantitative assessment of delayed and non-linear lag effects of climate and pollutants for the Greater London region. The methodology could easily be replicated on other disease categories and regions and not limited to LR admissions. The findings may provide useful information for the development and implementation of public health policies to reduce and prevent the impact of climate change on health problems.

Author's Declaration

I declare that the present work was carried out in accordance with the Guidelines and Regulations of the University of Westminster.

This thesis is entirely my own work and that where any material could be construed as the work of others, it is fully cited and referenced, and/or with appropriate acknowledgement given.

Until the outcome of the current application to the University of Westminster, the work will not be submitted for any such qualification at another university or similar institution.

Signed: Muhammad Saiful Islam **Date:** February, 2014

Acknowledgement

This thesis would not have come to the conclusion as it is today without the help and support of several people at various stages of my thesis via their words and actions. All the help I have received has been warmly appreciated.

First and foremost, I would like to express my sincere gratitude to my supervisors, Professor Thierry J. Chaussalet, Dr. Eren Demir, and Dr. Nazmiye Ozkan for their excellent guidance, mentoring, and immense inspirations throughout my research.

I am grateful to Dr. Salma Chahed, Sarah Dalton, and Philip Worrall for their advice and suggestions. My heartiest thank to Dr. Antonio Gasparrini, London School of Hygiene & Tropical Medicine for his suggestions and analytical discussions regarding the delayed effects of the impact of climate change. I like to thank Patrick F. Lees and Dr. Andrzej Tarczynski for their continuous support and advices regarding my research. I am also acknowledging Met Office, UK and Environmental Research Group (ERG), King's College London for their support related to data.

My parents are the best asset of my life. My mom was always there for me with all her efforts. My whole life is blessed with her care and encouragements that made me move forward at every step of my life. Lastly but most importantly, I like to express deep appreciation to my beloved wife, Dilshad for her tremendous support, care, immense inspiration, and encouragements that helped me complete this thesis.

Table of Contents

List of Figures

List of Tables

Abbreviations

AIC	Akaike information criterion
AIRGENE	Air Pollution and Inflammatory Response in Myocardial Infarction Survivors: Gene-Environment
ANOVA	Analysis of variance
ARIMA	Autoregressive Integrated Moving Average
AURN	Automatic Urban Rural Network
BIC	Bayesian information criterion
CART	Classification and Regression Tree
CET	Central England Temperature
CO	Carbon monoxide
COPD	Chronic obstructive pulmonary disease
DLNM	Distributed Lag Non-Linear Model
DOB	date of birth
ESPERE	Environmental Science Published Everybody Round the Earth
GAM	Generalized Additive Model
GIS	Geographical information system
GLM	Generalized Linear Model
HES	Hospital Episode Statistics
HPA	Health Protection Agency (Public Health England now)

ICD	10th revision of the International Statistical Classification of Diseases and Related Health Problems
IMD	Index of Multiple Deprivations
IPCC	Intergovernmental Panel on Climate Change
LAQN	London AIR Quality Network
LR disease	lower respiratory disease
MCMC	Markov chain Monte Carlo
MET Office	Meteorological Office
NHS	National Health Service
NO	Nitrogen Oxide
NO2	Nitrogen dioxide
ONS	Office for National Statistics
PM10	Particulate matters diameter of 10 micrometres or less
PM2.5	Particulate matters diameter of 2.5 micrometres or less
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QA	Quality assurance
QAIC	Quasi Akaike information criterion
QBIC	Quasi Bayesian information criterion
QC	Quality control
QQ plot	Quantile-Quantile Plots
SO2	Sulphur dioxide
TSP	Total suspended particulate
UKCIP	United Kingdom Climate Impacts

UV	Ultra Violet
VIF	Variation Inflation Factor
WHO	World Health Organization
WMO	World Meteorological Organization

Notations

x, y	Scalar
X, **Y**, **Z, Q**	Matrix
X, Y, Z	Random variable
$N(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$X \sim N(\mu, \sigma^2)$	X follows distribution $N(\mu, \sigma^2)$
$E[X], Var[X]$	Expectation and variance of X
$COV[X, Y]$	Covariance of X and Y
$Corr[X, Y]$	Correlation of X and Y
$\epsilon_{i_{i.i.d}} N(0, \sigma^2)$	$\epsilon_i's$ are error terms or residual, independent of each other and follows a normal distribution with mean 0 and variance σ^2
$L(\beta)$	Likelihood of β
$\mathcal{I}$	Information matrix
D	Deviance of a model
g	Link function
S	Smoother

Chapter 1

Introduction

1.1 Brief background

The ecology and the environment of the world are changing due to shifting patterns of meteorological factors. This is obvious from the most recent but warmest decade (2002-2011) as a succession of the warmest decades: 2000s, 1990s, and 1980s. According to the World Meteorological Office, the 13 hottest years have all occurred in the 15 years between 1997 and 2011 (WMO 2011) and among them 1998 is still the hottest and 2010 is the 2nd hottest years ever (WMO 2011). There is even a clear upward trend in the global temperature anomalies since pre-industrial times on the basis of year to year measurement. The apparent warming of the climate system is inevitable. Therefore, there has been increasing interest in the assessment of the relationships between climate change and health outcomes.

Climate, weather, and climate change

Climate encompasses the statistics of temperature, humidity, atmospheric pressure, wind, precipitation, atmospheric particle count and other meteorological elemental measurements in a given period over long periods. According to Intergovernmental Panel on Climate Change (IPCC) glossary definition (IPCC

2013), ***Climate*** *in a narrow sense is usually defined as the "average weather", or more rigorously, as the statistical description in terms of the mean and variability of relevant quantities over a period of time ranging from months to thousands or millions of years. The classical period is 30 years, as defined by the World Meteorological Organization (WMO). These quantities are most often surface variables such as temperature, precipitation, and wind. Climate in a wider sense is the state, including a statistical description, of the climate system. And* ***Climate change*** *refers to a statistically significant variation in either the mean state of the climate or in its variability, persisting for an extended period (typically decades or longer).* Thus it is measured in terms of years, decades or even centuries. Scientists study climate to look for trends or cycles of variability and also to place cycles or other phenomena into the bigger picture of possible longer term or more permanent climate changes. Since climate is changing rapidly nowadays, climate characteristics are sometimes recalculated every 10 years. However, for special purposes, other climatic time scales are also used (ESPERE 2004). On the other hand, **the weather** is the day-to-day state of the atmosphere, and its short-term (minutes to weeks) variation.

Factors in climate change

Temperature is the most common and influential climate factors impacting health on the top of precipitation, wind speed, humidity, atmospheric pressure, El Nino, UV (ultraviolet) index / solar radiation, cloud cover and so on. Many studies have been conducted on climate change and health related issues using temperature as climate factor (Muggeo and Hajat 2009; Basu and Malig 2011; Pinto, Coelho et al. 2011; Pudpong and Hajat 2011; Vardoulakis and Heaviside 2012). Besides,

levels of pollution are inclined to lead to health hazards during extreme climate events (Rocklöv and Forsberg 2009). Ozone levels, particle matters, / total suspended particulate (TSP), Nitrogen dioxide (NO2), Carbon monoxide (CO) and Sulphur dioxide (SO2) are considered to have a detrimental link with climate change and health. Vardoulakis and Heaviside (2012) mentioned that climate change may result in earlier seasonal appearance of respiratory symptoms due to longer duration of exposure to aeroallergens (pollen, fungal spores, etc.).

Impacts on health

Scientific consensus confirms that the changes in these meteorological variables are already adversely affecting health and such effects will be unevenly distributed throughout the world (WHO 2008). For instance, according to WHO, a one-degree rise in temperature in Europe could increase mortality by 1-4% and 86,000 extra deaths are projected every year, given an expected rise in global mean temperature of 3^0C, by 2071-2100 (Menne, Apfel et al. 2008). The frequency and severity of extreme weather events (e.g., heat waves, flooding and cold winters) are also increasing as an indirect effect of climate change. There were high numbers of excess deaths associated with the European heat wave during August 2003. This number is approximately 2,000 for England & Wales (Johnson, Kovats et al. 2004) and 15,000 for France (Fouillet, Rey et al. 2006). Heat-related mortality is projected to increase steeply in the UK in the 21st century, which is approximately 70% in the 2020s, 260% in the 2050s, and 540% in the 2080s, compared to the 2000s heat-related mortality baseline of around 2,000 premature deaths (Vardoulakis and Heaviside 2012). Various vectors, water, food-borne diseases, and pathogens are directly or indirectly related to

changing behaviour of climate change. The incidence of existing infectious agents, such as Lyme disease transmitted by ticks, is likely to increase in UK (Vardoulakis and Heaviside 2012). The burden of disease during extreme climate events like floods, heat waves, and storms are about to increase because of the associations of climate factors and the vector & waterborne diseases (Parry, Canziani et al. 2007). The river and coastal flood risks are likely to increase in the next decades due to climate change. All populations are at risk of the health effects associated with flooding; however, poorer communities are at higher risk of coastal flooding in the UK, while higher income households tend to be at higher risk of river flooding. According to the HPA (Health Protection Agency) report, such indirect impacts of climate change have wider consequences on existing public health problems during certain occasions related to water availability, nutrition, mental health and well-being, displacement and migration, and health equity (Vardoulakis and Heaviside 2012).

Vulnerable population group

Children, the elderly (especially those living on their own), individuals with pre-existing illness, people living in overcrowded accommodation and socioeconomically deprived are the most at risk due to their frailty (Knowlton, Rotkin-Ellman et al. 2009; Alonso, Achcar et al. 2010; Vardoulakis and Heaviside 2012). The health burdens of the UK may be amplified by an aging population due to climate change, particularly for those over 85 years of age, compared with younger age groups. In the UK, the elderly are the most vulnerable due to flood and climate events.

Challenges in climate change research

The scale of the impact of climate change varies in terms of geographical latitude and climate zone throughout the world. Overall, UK will be negatively affected due to the changing climate and even in the UK the South East, London, the East and West Midlands, the East of England and the South West appear to be more vulnerable to current and future effects of hot weather (Vardoulakis and Heaviside 2012). According to CET (Central England Temperature), there is an increasing trend in the temperature anomalies and a series of warm years since the late 1980s with 2006 as the warmest year on record. Along with this, there have been decreasing numbers of cool and increasing numbers of warm days and night between 1960 and 2010 (Vardoulakis and Heaviside 2012). Rainfall has decreased during the summer and increased during winter (UKCIP trends report). Observations of the English Channel show rises in extreme sea levels at all 16 sites studied (Haigh, Nicholls et al. 2011) and the levels of ultraviolet radiation is also affected due to climate change.

The real cause of the climate change is still a topic of debate though it is admitted by the climate researchers that human anthropogenic activity since 1750 is one of the leading causes of the warming climate (Vardoulakis and Heaviside). Failure to respond now could be very costly in terms of disease, health care expenditure, and lost productivity alongside ecological imbalance and environmental degradation.

Identifying the nature of the relationships between the variations of the climate factors and health is very challenging. Most of the past research works considered this relationships as linear mainly because of the computational

advantages of dealing a linear model. However, recent studies revealed that health or disease exposure generally shows a non-linear U, V, N or even J shaped relationships with the hazard (Braga, Zanobetti et al. 2002; Pattenden, Nikiforov et al. 2003; Pauli and Rizzi 2008; Muggeo and Hajat 2009). Computationally, such nonlinearities are also challenging but provide more efficient results. Moreover, the issues of the delayed effect of sudden climate change and related lag structure of the climate factors and air pollutants are crucial for the efficiency of the modelling. Further elaborative discussion about nonlinearity & smoothing techniques and lag structure & delayed effect can be found in section 2.8 and section 3.6 respectively

To deal with the problem, efficient modelling of this relationship is critical. Unfortunately the full quantitative estimate of the impact of climate change is still not possible due to the lack of reliable exposure-response relationships especially in health. Moreover, the diversified nature of climate and weather made estimating the relationship with the health status of a population extremely complex. Historically this has limited some of the existing plans and policies to face the rapid climate change. Therefore, a number of policies and strategies may need to be revised and/or strengthened under the present levels of risk based on the precise scientific research.

1.2 Main aim of the Thesis

The overall aim of this research is to develop a statistical model to precisely identify and measure the impact of climate change on health (such as daily hospital admissions) by considering non-linear relationships between climate

factors and hospital admission and delayed effects (section 3.6) of the selected climate and pollution factors.

1.3 Specific objectives

- To identify the influential climate and pollution factors in England that may play a significant role in daily admissions.
- Feasibility of the HES for measuring the impact of the climate change on health.
- To illustrate the delayed effect of the significant climate and pollution variables on the hospital admissions.
- To check the efficiency of a proposed structure of the delayed effect (lag structure) of the climate and pollution factors in measuring their impact of hospital admissions.
- To evaluate the efficiency of the non-linear statistical model developed using the proposed lag structure of the selected climate and pollution factors.

1.4 Contributions to knowledge and research

A variety of methods now exist for assessing the impacts of climate change on human health while different approaches for studying the effects of climate factors and extreme climate events on health can result in highly variable estimates (Rocklöv and Forsberg 2009). However, efficient quantitative estimates of the impact of climate change on daily hospital admissions are still limited due to the

lack of reliable disease exposure relationships. The diversified nature of climate and weather made it extremely challenging too.

The lag effect of the factors and exposure are very crucial and a significant amount of climate change health studies conducted using various lag. Despite the efforts no studies have suggested an efficient structure of lag period that can increase the efficiency of any statistical model for measuring the impact of climate factors on health. The same argument goes for devising an efficient threshold limit (section 3.6) for climate variables for any specific region. So far thresholds were mainly estimated for temperature and it is crucial that the calculated threshold is precise and accurate.

This research allows a unique contribution addressing the above mentioned research gaps. We described the contributions of the thesis under two sections: theoretical contributions and applied contributions.

Applied contributions

- Classify the climate, and pollution factors that are significant and should be considered for any specific disease categories (e.g., lower respiratory disease) for a specific region (e.g. Greater London, England).
- Calculate an efficient structure of the lag period of climate-diseases related under the climate change context of the UK.
- Identify the delayed effects of the climate factors for lower respiratory hospital admissions. This will eventually lead towards an efficient threshold climate for emergency hospital admissions of LR disease in Greater London. Moreover this will also lead towards an efficient health alert systems due to sudden climate change.

Theoretical contributions

- Development of an efficient statistical model considering the delayed effect of the significant climate variables and measure the relative efficiency of that model.
- Usefulness of B-Spline smoothing techniques in DLNM model to cover all the non-linearity beyond the boundary knots in the data and thus improve the model efficiency.

This research could enable senior decision makers to adopt more proactive and evidence-based methods in the decision making process, such as future policies based on various climate variable thresholds (e.g. Temperature, rain) which may assist them in finding efficient ways of delivering services.

1.5 Outline of the thesis

Figure 1 illustrates the flow of the remaining chapters and their relations. These chapters, presented in sequence, are grouped with topics of literature reviews, theoretical concepts, and contributions.

Chapter 2: Literature review

In this chapter, we present a systematic review of literature illustrating the nature of the impact of climate change on health, related factors, and research studies with statistical modelling approaches related to climate change and health. The scope of the thesis, namely, the idea of estimating an accurate lag structure, thresholds, and factors for developing the statistical model emerged from this chapter.

Chapter 3: Factors in climate health research

This chapter describes the factors associated with climate health research. This covers meteorological factors (e.g. Temperature), pollutants, demographic factors (e.g., age, sex, and race), lag structure, quality of the data, and geographical factors (latitude, longitude). This chapter can be considered to be an extension of the literature review chapter which later supports us in the selection of relevant factors to include in our statistical models.

Chapter 4: Data sets used

We give a brief overview of the three data sets used in the research: Hospital Episodes Statistics (HES) data, climate data, and pollution data. We also describe the study population and coverage area, data management, and data cleaning process, linking administrative data sets and issues related to missing values.

Chapter 5: Generalized linear modelling

This chapter describes the existing statistical approaches especially the generalized linear modelling and its extension for dealing the count data. We illustrate a brief overview of the theoretical descriptions of the GLM. In addition to that, we describe the extensions of the GLM for the count data and deployed them to our data. Finally, we applied the GLM using the climate and pollution factors for selecting the significant factors in the emergency lower respiratory hospital admissions.

Chapter 6: Modelling with non-linearity and delayed effect of climate factors

In this chapter, we develop our model by considering the non-linear relationships between climate change and emergency hospital admissions. But before

proceeding to the final model, we describe some commonly used smoothing techniques and spline functions for non-linear statistical modelling. We also illustrate the Distributed lag non-linear modelling and develop the final model incorporating the delayed effect and non-linearity of the relationships.

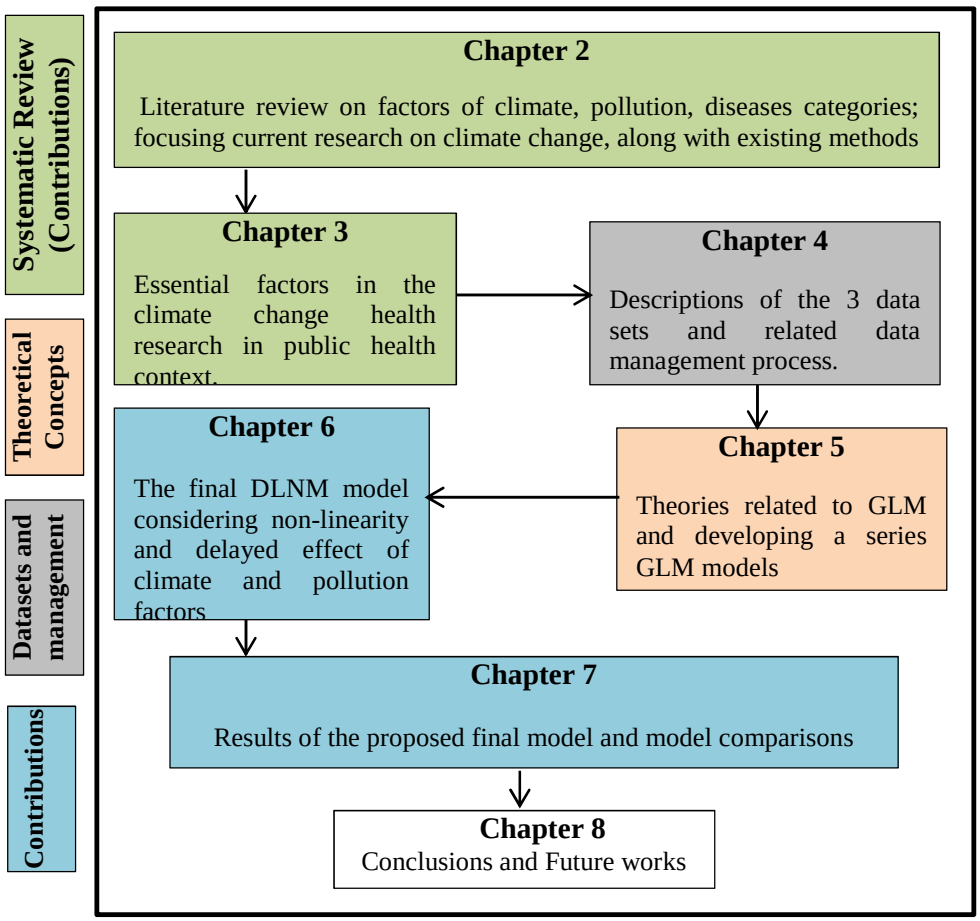

Figure 1: Map of the chapters and their inter-dependency for this thesis

Chapter 7: Results of the final model

In this chapter, we describe and interpret the results of the final model after applying it to our datasets. We also compare the results emerged from the GLM, DLNM model, and the final DLNM model. We show the results that how the final model is providing a better fit to the data. Model comparisons have been done using standard procedures.

Chapter 8: Conclusions and further works

This chapter concludes the thesis and describes some of the limitations of the research. In addition to limitations, we also describe our future plan for extending the model in various aspects of diseases and scenarios.

The systematic literature reviews from chapter 2 and characteristics of factors in chapter 3 form the basis of problem identification and research gap concerning this research. We describe the datasets, missing values, and data management process in Chapter 4. Chapter 5 describes the theory of generalized linear model (GLM) and results from our data sets. In chapter 6, we develop a DLNM model for our problem, followed by the results in chapter 7. We finish this thesis by illustrating the conclusions and future works emerged from this work. In general, the contribution of the thesis lies in the systematic review (chapters 2), linking the three administrative datasets into one platform (chapter 4), devising the significance climate factors other than only temperature (chapter 5) and most importantly, developing a delayed non-linear model (chapter 6 and 7).

1.6 Chapter summary

In this chapter, we provided a brief background of the crucial aspects of climate change and its adverse impact on the environment and health, along with the objective of the thesis, contributions, and the thesis outline. In the next chapter, we present a detailed literature review illustrating some of the key issues and factors associated with the impact of climate change on health.

Chapter 2

Literature review

2.1 Introduction

Climate change has become one of the main areas of research concentration because of its current and future impact on health. As a result, a significant number of diversified research projects have been done recently to deal with this affliction. These studies differ according to their subject areas, objectives, methodologies, population, and disease characteristics, latitudes and climate zone. In this chapter, we conduct a systematic review of the literature on climate change and its impact on health along with the emphasis on statistical modelling adopted in various studies.

Section 2.2 describes the search strategy and selection criteria of the studies, followed by an overall nature of exposure-response relationships of climate change in section 2.3; climate and pollution factors under this context are highlighted subsequently in sections 2.4 and 2.5. Sections 2.6 and 2.7 focus on the sensitive disease categories and most vulnerable cluster of population due to climate change. Finally, the statistical modelling approaches in studies of climate change and health are described in section 2.8.

2.2 Search strategy and selection of articles

A literature review has been carried out with a general quest for examining the methodologies used, to assess the relationships between climate change and health exposure (e.g., hospital admissions, diseases). We followed the guidelines of the PRISMA statement for systematic reviews and meta-analysis (Moher, Liberati et al. 2009). The literature search solely concentrates on studies related to health, health care, and disease epidemiology due to climate change and the methodologies adopted for this purpose, e.g. statistical modelling. In general the following criteria were considered for inclusion:

- Studies examining the relationship between meteorological factors (e.g. Temperature, rainfall) and morbidity or mortality using hospital outcomes or any health statistics data.
- Studies using any statistical modelling approaches related to climate change and health.
- Studies focusing on populations vulnerable to climate change.

Studies without any statistical model, not related to health, health care, disease, climate change or weather variations, published before the year 2000, and not written in English were excluded. We used the ISI web of knowledge (WOK), an academic citation indexing and search service. WOK includes various databases, such as MEDLINE and Web of Science. The review focused on relevant studies published in English since 2000. The keywords used in the search criteria are: climate change, weather, hospital admissions, disease, and health by considering the inclusion criteria mentioned earlier. The search was further

refined based on articles and reviews focused on the subject areas like public health, environmental occupational health, environmental sciences, health care sciences, mathematics, demography, infectious disease, social sciences, and so on. The articles which came out of the above procedure were supplemented by other related articles in the same area, collected previously for research purposes.

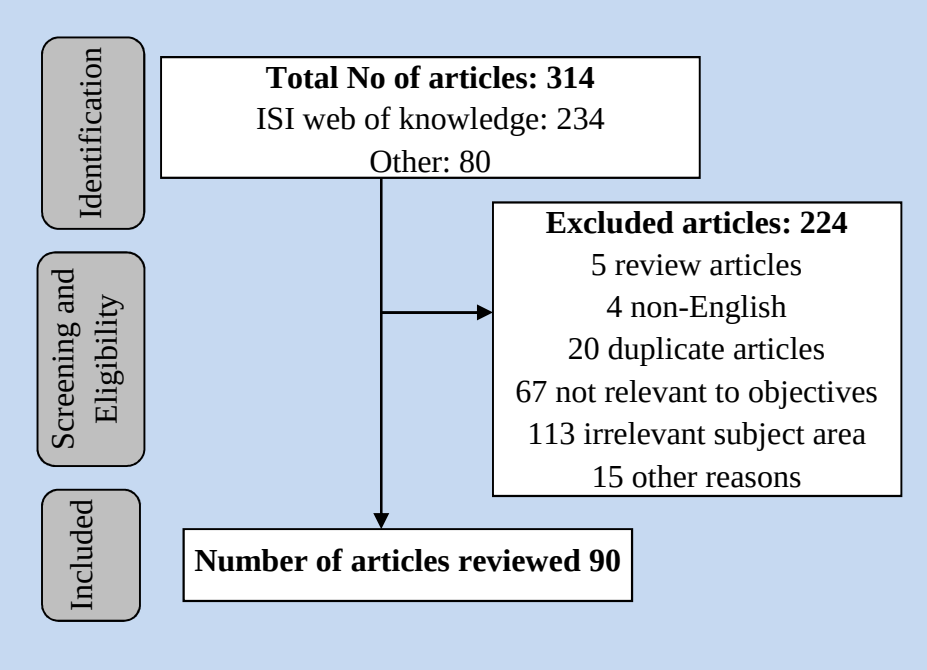

Figure 2: Selection process for articles

We explored all the articles from abstract to conclusion. The screening and eligibility of the articles were based on their objectives, statistical model, assumptions, variables or factors used for measuring climate change, measurements or outcomes used for health exposures, disease categories, study region, time period, individuals studied, bias or limitations of the study and any influential factors in the model. We found 314 citations, 234 from ISI web of

knowledge, and 80 from other sources like references and Google scholar. A total of 90 articles met the inclusion criteria out of 314 citations (**Figure 2**).

Thus the meteorological and pollution factors, disease categories as a consequence of the climate change, the vulnerability of specific populations, and various geographical regions have been reviewed under the climate change health context along with a rigorous evaluation of the existing statistical methodologies that have been developed and applied for modelling the impact of climate change on health. This investigation has strengthened our understanding of the field, enabling us to identify the gaps and challenges to conceptualise the bigger picture of climate and healthcare research. The key findings and contributions to knowledge of this systematic review can be summarized as follows:

- Factors in climate health research should be specific to regions and diseases. A climate index has stronger statistical significance with health than same climate factors used separately.
- Elderly, children, and patients with respiratory diseases are the main groups at risk.
- Non-linearity between climate change and diseases should be considered for model optimisation.
- Lag structures of the factors are very crucial and an efficient climate threshold can lead to an improved health alert system.

2.3 The nature of the exposure-response relationships

The links between weather, climate, and health are still largely unexplored except in recent studies describing their associations (Basu and Samet 2002; Kovats and

Hajat 2008). Most of the observational epidemiology does not show any straightforward linear associations among the considered factors due to the complex multifactorial exposure-response relationships among various factors. Climate change reveals an overall health hazardous picture and literature reviews revealed that a population with a temperate climate generally shows non-linear U, V, N or even J shaped relationships with the hazard (Braga, Zanobetti et al. 2002; Pattenden, Nikiforov et al. 2003; Pauli and Rizzi 2008; Muggeo and Hajat 2009).

2.4 Climatic factors affecting health

Temperature is the most common climate or weather factor in almost all the studies. Apparent temperature, dew point temperature, sea surface temperature, temperature range, and diurnal/ambient temperature are also useful as factors. Besides temperature, wind speed and direction, humidity, rainfall, atmospheric pressure, UV (ultraviolet) index / solar radiation, cloud cover, pressure, El Nino, water vapour pressure have also been used in climate health studies (**Table 1**). Apparently, temperature related factors have been investigated in almost all the reviewed articles concentrating on climate health research (136 times out of 258) followed by factors related to humidity (66 times) and wind (26 times) (**Table 1**). Bartzokas, Kassomenos et al. (2004) used irradiance, water vapour pressure and west-north & south-south wind component along with temperature, wind, humidity, and atmospheric pressure. Nastos and Matzarakis (2006) used UV index along with temperature, wind speed, humidity, and atmospheric pressure. Lam (2007) used temperature, rainfall, relative humidity, and UV index. Rainfall and humidity have also been used by (Pinto, Coelho et al. 2011). Besides temperature

or temperature induced indexes, the use of all other climate factors relevant to disease, climate zone and objective of the research, are crucial to include in the model for model optimisation.

Heat and cold waves

Heat waves cover a huge portion of climate studies because of their catastrophic and sudden impact on health (Huynen, Martens et al. 2001; Díaz, Jordán et al. 2002; Schwartz, Samet et al. 2004; Le Tertre, Lefranc et al. 2006; Medina-Ramón, Zanobetti et al. 2006; Argaud, Ferry et al. 2007; Tan, Zheng et al. 2007; Hansen, Bi et al. 2008; Hansen, Bi et al. 2008; Pauli and Rizzi 2008; Knowlton, Rotkin-Ellman et al. 2009; Tong, Ren et al. 2010; Ma, Xu et al. 2011). They act as a factor in driving adverse health episodes (Pauli and Rizzi 2008), specifically for heatstroke (Argaud, Ferry et al. 2007), mental disorder, morbidity and mortality during summer, which are strongly associated with heat waves (Medina-Ramón, Zanobetti et al. 2006; Tan, Zheng et al. 2007; Hansen, Bi et al. 2008; Hansen, Bi et al. 2008).

The majority of deaths during heat waves appear to be due to pre-existing chronic diseases, especially cardiovascular disease (McGeehin and Mirabelli 2001). The effects of high temperature during heat waves have shown different patterns on hospitalisation (Kovats, Hajat et al. 2004) compared to general summer temperatures (Kovats, Hajat et al. 2004; Michelozzi, Accetta et al. 2009). Climate induced risks are often much higher during heat waves. For example, the heat wave in France during 2003 showed large excess mortality (Le Tertre, Lefranc et al. 2006). Unfortunately, there is a lack of standardised framework for defining **heat waves** because of their variations with respect to the location, time,

subject area, aim, and objectives of the study. Heat waves need to be defined locally and standardise geographically for any homogeneous population (Tong, Ren et al. 2010). Contrary to heat waves, **cold spells** have been given little attention in climate health studies (Huynen, Martens et al. 2001; Revich and Shaposhnikov 2008; Ma, Xu et al. 2011). However, they are also associated to increase hospital admissions (Ma, Xu et al. 2011).

Climate index

Climate index calculated from two or more climate factors is a recent practice in climate health studies. Hartz, Golden et al. (2012) and Tong, Ren et al. (2010) extracted a **heat-index** from temperature and relative humidity. An index of **apparent temperature** was calculated by combining ambient temperature and relative humidity (Green, Basu et al. 2010; Alessandrini, Zauli Sajani et al. 2011; Wichmann, Andersen et al. 2011). **Apparent temperature** has also been calculated using saturated vapour pressure, actual vapour pressure, dew point temperature (Kovats, Hajat et al. 2004; Basu, Feng et al. 2008; Basu and Malig 2011). In general, by using the indexes, stronger statistical significance with health outcomes can be achieved than any single climate factors. This is probably because any combined impact of climate factors on disease exposure is relatively stronger than the effect of any single factor.

2.5 Pollution factors in climate health research

Variations in pollutant levels are found to relate to health hazards during extreme climate events (Rocklöv and Forsberg 2009). Although there is a debate as to whether pollutants should be included in climate research, many studies have

focused on using pollution variables due to their unavoidable link with health (Table 2). Ozone levels and various particle matters are the two dominant measures in current research appearing 57 and 49 times, respectively in our literature review (Table 2). Bhaskaran, Hajat et al. (2010) used these two factors for chronic bronchitis and influenza. Ozone has been used for chronic bronchitis and heart disease and acute myocardial infarction, angina pectoris, pneumonia, diarrheal disease / dehydration by Green, Basu et al. (2010). Besides ozone, Nitrogen monoxide (NO) has been linked with measuring chronic obstructive pulmonary disease (COPD) (Liang et al., 2009), Nitric oxide (NOx) with respiratory and cardiovascular diseases (Díaz, Jordán et al. 2002) and black smoke for the same two diseases (Bartzokas, Kassomenos et al. 2004).

2.6 Disease categories due to climate change

According to IPCC (2007), the association between climate change and health is either direct, e.g. cardiovascular effects of extreme weather or indirect i.e., via pathogens, allergens or vectors (e.g. Vector and waterborne diseases, mould and pollens). Such associations disclose the possibility of burden of disease to increase due to extreme climatic events, e.g. heat waves, floods, cyclone, and storms IPCC (2007).

The literature review revealed specific diseases that are found to be more frequent and influential in climate health research and vary depending on time, place, age, and socioeconomic conditions of the population. **Table 1** (climate-disease) and **Table 2** (pollution-disease) showed the cross tabulation of disease categories by climate and pollution factors that have been considered in the

literature review. In **Table 1** (climate related disease), respiratory (24 times), and cardiovascular diseases (23 times) are the two dominant disease categories followed by COPD (7 times) and diabetes (5 times). Other prominent disease categories in Table 1 are: COPD (7 times), diabetics (5 times), asthma, pneumonia, & atrial fibrillation (4 times each). In Table 2 (pollution related disease), cardiovascular (37 times), respiratory diseases (31 times) are the dominant followed by COPD (13 times), and stroke (9 times). The rest of the highly frequent diseases in this table in descending order are: asthma (8 times), cardiac disease (7 times), cerebrovascular disease (7 times), and 5 times each for diseases in renal system, and kidney & congestive heart failure. It is interesting to see the increased amount of research conducted on asthma and pollution factors associated with climate change (**Table 2**).

Other diseases in climate change studies

Dengue is more frequent in tropical countries where rainfall and humidity play an important role (Pinto, Coelho et al. 2011). Temperature, rainfall, relative humidity and UV index for fever, gastroenteritis and asthma have been considered by Lam (2007). Skin disease has been explored by Mentzakis and Delfino (2010) in relation to factors such as temperature, relative humidity, and atmospheric pressure. The climate factor El Niño has been considered by Ebi, Exuzides et al. (2004) for stroke, congestive heart failure, acute myocardial infarction, and angina pectoris. Ferrari et al. 2011 used UV index, cloud cover, boundary layer height along with temperature, wind speed, and humidity for measuring their impact on COPD. Malignant neoplasm was considered by Huynen, Martens et al. (2001) using temperature. Digestive diseases were examined by Fernández-Raga, Tomás

et al. (2010) using temperature, relative humidity, and atmospheric pressure. Dementia was examined by Hansen, Bi et al. (2008) considering temperature as a factor. Green, Basu et al. (2010) has used apparent temperature as climate factor for measuring its impact on chronic bronchitis or emphysema, intestinal infectious diseases, and acute renal failure. Kawasaki disease has been considered by Checkley, Guzman-Cottrill et al. (2009) for temperature and rainfall. Alonso, Achcar et al. (2010) also considered coronary ischaemic diseases adjusting temperature, humidity, and atmospheric pressure. Temperature and relative humidity have been adjusted for influenza by Bhaskaran, Hajat et al. (2010). Pauli and Rizzi (2006) used UV index for measuring the hospital admissions due to non-accidental causes. UV index (Keatinge and Donaldson 2001; Chang, Zhou et al. 2010), cloud cover (Chang, Zhou et al. 2010), and boundary layer height (Ebi and McGregor 2008) are also used for measuring mortality as an impact of climate change. Mortality data are very good indicator of the impact of the climate change (Table 1 and Table 2), and temperature and relative humidity are predominantly used in such climate research studies compared to any other factors.

Indirect measurements of health outcomes

In addition to disease morbidity and mortality, indirect measurements of health outcomes as a result of climate change are very popular and normally used by adopting various health and administrative terminologies. For instance, emergency call data (Bassil, Cole et al. 2009; Hartz, Golden et al. 2012), emergency hospital admissions / room visits / emergency dispatches (Kovats, Hajat et al. 2004; Argaud, Ferry et al. 2007; Lam 2007; Knowlton, Rotkin-Ellman

et al. 2009; Liang, Liu et al. 2009; Wang, Barnett et al. 2009; Khalaj, Lloyd et al. 2010; Tong, Wang et al. 2010; Alessandrini, Zauli Sajani et al. 2011; Wichmann, Andersen et al. 2011), emergency ambulance data (Dolney and Sheridan 2006; Ferrari, Exner et al. 2012), and hospital discharge / outpatient visits / hospital admission data / Hospital Episode Statistics (Kovats, Hajat et al. 2004; Rudge and Gilchrist 2005; Hansen, Bi et al. 2008; Hansen, Bi et al. 2008; Checkley, Guzman-Cottrill et al. 2009; Pudpong and Hajat 2011; Sung, Chen et al. 2011).

2.7 Vulnerable population and region

The impact of a changing climate can vary due to the variations in human susceptibilities, socioeconomic factors, and population acclimatization to prevailing conditions and other adaptive measures. Elderly and people with cardiovascular & respiratory diseases, mentally ill, people under medications and with diabetes have been identified as disproportionately vulnerable to changing climate (Kaiser, Rubin et al. 2001; McGeehin and Mirabelli 2001; Medina-Ramón, Zanobetti et al. 2006). People with certain psychological or behavioural characteristics are also very sensitive in such situations. Athletes, children, and outdoor workers may likely be affected by heat stroke due to being outdoors longer and exerting themselves, even though they are fit and healthy (Hartz, Golden et al. 2012).

Age is one of the most influential factors and older people are significantly at higher health risk due to global climate change (Huynen, Martens et al. 2001; McGeehin and Mirabelli 2001; Rudge and Gilchrist 2005; Medina-Ramón, Zanobetti et al. 2006; Pauli and Rizzi 2008; Revich and Shaposhnikov

2008; Knowlton, Rotkin-Ellman et al. 2009; Muggeo and Hajat 2009; Alonso, Achcar et al. 2010; Khalaj, Lloyd et al. 2010; Pudpong and Hajat 2011). According to Knowlton, Rotkin-Ellman et al. (2009), people aged 65 or over and children (0-4 years) represent the highest risk group due to their frailty to heat-related causes. Higher sweating thresholds increase the risk of life threatening consequences when body temperatures rise (McGeehin and Mirabelli 2001). Therefore, special attention is required for the elderly and children under the changing climate (Tam, Wong et al. 2009). However, selection of an appropriate age group for such analysis depends on disease category, the aims, and objectives of the research, and availability of quality data. To date, various age groups have been considered in climate health research and 65+ is the most commonly used (Huynen, Martens et al. 2001; Schwartz, Samet et al. 2004; Rudge and Gilchrist 2005; Medina-Ramón, Zanobetti et al. 2006; Kolb, Radon et al. 2007; Hansen, Bi et al. 2008; Hansen, Bi et al. 2008; Qian, He et al. 2008; Tam, Wong et al. 2009; Vaneckova, Beggs et al. 2010; Pudpong and Hajat 2011). Several other elderly age groups have also been considered. For example, >75 years (Díaz, Jordán et al. 2002; Khalaj, Lloyd et al. 2010), 65-74 years (Díaz, Jordán et al. 2002), >= 50 (Keatinge and Donaldson 2001; Basu and Malig 2011), >= 55 (Fouillet, Rey et al. 2007), >=75 (Pauli and Rizzi 2006; Pauli and Rizzi 2008; Pauli and Rizzi 2008), 50-69 and >=70 (Ebi, Exuzides et al. 2004), 75-84 (Bhaskaran, Hajat et al. 2010). On the contrary, very few studies have focused on children (<5 years, Kovats, Hajat et al. (2004) and <6 years, (Lam 2007).

Table 1: Frequency of diseases categories / mortality crossed with meteorological factors in literature review.

Climate Factors / Disease Categories / Mortality	Factors Related To Temperature				Factors related to Wind		Rainfall	Relative Humidity / Humidity	Atmospheric Pressure	Total
	Temp.	Apparent Temp.	Dew Point Temp.	Sea Surface Temp	Wind Speed	Wind Direction				
Respiratory disease	16	4	1		5	1	1	12	5	**24**
Cardiovascular / Circulatory disease	18	3	1		4	1	1	13	4	**23**
Cerebrovascular disease	1	1		1	1			2		**3**
Asthma	2	2			1		1	2		**4**
COPD	6	2			2			4	1	**7**
Heart disease	1	1						1		**1**
Cardiac arrest	2	1			1			1		**2**
Diseases related to renal system, kidney, ureter	3	1			1			2		**3**
Diabetes	4	2			1			3	1	**5**
Dehydration	3	1			1			1		**2**
Mental disorders / Schizophrenia	3	1			1		1	1		**3**
Heat stroke / stroke	6	1		1			1	2		**3**
Congestive Heart Failure	2	1		1			1	1	1	**3**
Cardiac disease	2				1					**1**
Atrial fibrillation	1	1						4		**4**
Acute myocardial infarction, Angina pectoris	4	1		1			1			**1**
Pneumonia	3	1						3	1	**4**
Diarrhoeal disease / Dehydration / Intestinal infectious disease	2	1			1		1	1		**3**
Described as Non-incidental Causes	2				1	2	2	2	1	**8**
Heat-related emergencies	1		1				1	1		**2**
Mortality	21	2	3		1		3	10	2	**16**
Total	**103**	**23**	**6**	**4**	**22**	**4**	**14**	**66**	**16**	

Table 2: Frequency of diseases categories / mortality crossed with pollution factors in literature review.

Climate Factors / Disease Categories / Mortality	Ozone (O_3)	Particulate matter: (e.g.PM_{10} or $PM_{2.5}$ or both) / Total suspended particulate (TSP)	Nitrogen dioxide (NO_2)	Carbon monoxide (CO)	Sulphur dioxide (SO_2)	Total
Respiratory disease	11	10	7		3	**31**
Cardiovascular / Circulatory disease	11	9	9	3	5	**37**
Cerebrovascular disease	2	2	2	1		**7**
Asthma	3	2	2		1	**8**
COPD	4	3	3	1	2	**13**
Cardiac arrest	1	1	1			**3**
Diseases in renal system and kidney	2	2	1			**5**
Diabetes	2	1	1			**4**
Dehydration	1	1	1			**3**
Mental disorders / Schizophrenia	1	1	1			**3**
Heat stroke / stroke	3	2	2		2	**9**
Congestive Heart Failure	2		1	1	1	**5**
Cardiac disease	2	2	2		1	**7**
Atrial fibrillation	1	1	1			**3**
Mortality	11	12	4	2	5	**34**
Total	**57**	**49**	**38**	**8**	**20**	

2.8 Modelling approaches in climate change health research

The review aimed to focus on the various statistical modelling and methodological approaches used in recent studies around the sphere of health care and disease epidemiology. A variety of modelling approaches have been discovered in various climate change health settings. In general, most of the studies involve health exposure as responses (e.g., disease outcomes, morbidity, hospital admissions) and climate variables as explanatory variables.

Generalized linear model (GLM)

The GLM is found to be very useful and frequently used in this context (Sung, Chen et al. 2011; Ferrari, Exner et al. 2012). A standard GLM for normal responses is a multiple regression model in which the dispersion parameter is the error variance (Chandler 2005; Bhaskaran, Hajat et al. 2010). Recently the generalized additive model (GAM) has become one of the main statistical models under the climate change and health framework (Guisan, Edwards et al. 2002). This is because of its nature as a semi-parametric extension of GLM and ability to deal with non-linear and non-monotonic relationships. Khalaj, Lloyd et al. (2010) and Medina-Ramón, Zanobetti et al. (2006) used logistic regression model to determine the health impacts of extreme heat events. Hartz, Golden et al. (2012) used a multivariate analysis using stepwise regression to examine seasonality and identify the statistically significant relationships of selected mortality and meteorological variables. A generalized estimating equation has been used by Wang, Barnett et al. (2009) to investigate the impact of heat and cold on emergency stroke admissions. All are special cases of GLM.

Models with count data

Poisson regressions under the GLM and GAM have been used in various climate change health research studies. Basu, Feng et al. (2008); Alessandrini, Zauli Sajani et al. (2011); Basu and Malig (2011); Vardoulakis and Heaviside (); Liang, Liu et al. (2009); Kovats, Hajat et al. (2004); Hajat, Armstrong et al. (2005); Fouillet, Rey et al. (2007) used Poisson regression with a log link function in either generalized linear model or generalized additive model (GAM) for exploring the relationship between daily emergency ambulance dispatches and apparent temperature, accounting for over dispersion and autocorrelation in the model. Generalized negative binomial regression (Vaneckova, Beggs et al. 2010; Pudpong and Hajat 2011), time series zero-inflated Poisson regression model with classification and regression tree (CART) (Hu, Mengersen et al. 2010) are also applied in climate change research. Poisson regression model with a log link has been used frequently because of the nature of the response variables are counts or rate of disease outcomes (Tam, Wong et al. 2009). Log link is also useful with other modelling practices (e.g., GLM, GAM) (Qian, He et al. 2008).

Exploratory data analysis

Evaluating an exploratory data analysis before fitting any statistical model is a common practice including the area of climate change and health. Test of hypothesis like t-test, 2 sided χ^2-Test or Fisher exact test has been used for measuring the baseline characteristics of the study populations by Argaud, Ferry et al. (2007). Ma, Xu et al. (2011) used rate ratios to estimate the impact of the heat wave and the cold spell on hospital admissions. Rate ratios were also used by Knowlton, Rotkin-Ellman et al. (2009) to investigate highly susceptible age or

race groups to hospitalisations and emergency department (ED) visits during the 2006 California heat wave. Nastos and Matzarakis (2006) used Pearson Chi-Square Test ($\chi 2$) to examine the relationship between meteorological parameter and General Practitioner (GP) consultations as an exploratory analysis.

Case-crossover designs

Case-crossover design to analyse climate health data is also available along with case only study design (Medina-Ramón, Zanobetti et al. 2006). This is equivalent to a matched case-control study where the cases act as their own control. In such case the time-independent factors (e.g., age, sex, race) are unable to confound the observed associations. Kolb, Radon et al. (2007) used a time-stratified case-crossover design considering temperature, pressure, humidity, and adjusting pollutants to determine the associations between weather and daily elderly mortality due to congestive heart failure. The time-stratified approach removes biases from unwanted trends in the mortality time series and leads to unbiased estimates of effect for case-control days selected within specific time windows (Kolb, Radon et al. 2007). Thus it controls for trends and seasonal patterns in the dependent and independent variables (Tong, Wang et al. 2010). Time-stratified case-crossover design has also been adopted by Green, Basu et al. (2010); Wichmann, Andersen et al. (2011); Ostro, Rauch et al. (2010);Tong, Wang et al. (2010).

Data reductions techniques

Statistical data reduction techniques are commonly used for selecting variables or reducing the data dimension in this area. Principal component analysis was used

by Pinto, Coelho et al. (2011) and Fernández-Raga, Tomás et al. (2010) for selecting climate factors (e.g., temperature, rainfall, humidity).

Time series models

Time series analysis became one of the key statistical approaches in the climate change researches. Kaiser, Le Tertre et al. (2007) used advanced time series analysis methods with Poisson regression and penalised regression spline to re-examine the effects of 1995 Chicago heat wave on all-cause, cause-specific mortality, and mortality displacement. Lam (2007) used the ARIMA (Autoregressive integrated moving average) time series model to measure the association between climate factors and childhood illness. Alessandrini et al. (2011) used GAM and time series analysis techniques. Some other recent studies that applied time series modelling approaches in this context are: Chang, Zhou et al. (2010); Basu and Malig (2011); Pudpong and Hajat (2011); Hartz, Golden et al. (2012); Rocklöv and Forsberg (2009); Tong, Ren et al. (2010); Le Tertre, Lefranc et al. (2006); Bhaskaran, Hajat et al. (2010); Kovats, Hajat et al. (2004). Time is also an important factor for statistical modelling like survival analysis or multivariate Cox proportional hazard model (Argaud, Ferry et al. 2007).

Models based on Bayesian approach

A Bayesian approach has been exposed recently in various climate change health research with some promising results. Alonso, Achcar et al. (2010) used a Poisson regression model where the inferences of interest have been obtained using Bayesian methods and the posterior summaries via MCMC simulation methods. The objective of the study was to verify whether climate covariates affect the daily hospitalisation and identify susceptible age groups.

Data structure and dependent variable

The data structures of most of the studies are multidimensional data frequently involved measuring the relationships of climate change over time. Thus nearly all the studies in our literature review used panel or longitudinal data (e.g. (Ferrari, Exner et al. (2012), Fernández-Raga, Tomás et al. (2010), Hajat, Armstrong et al. (2005), Hu, Mengersen et al. (2010), Kalkstein and Davis (2005), Kovats, Hajat et al. (2004), Muggeo and Hajat (2009), Pattenden, Nikiforov et al. (2003), Pauli and Rizzi (2008), Pudpong and Hajat (2011), Schwartz, Samet et al. (2004), (Tam, Wong et al. 2009), Huynen, Martens et al. (2001), Donaldson, Keatinge et al. (2003))). Besides panel data, time series data (e.g., by Basu, Feng et al. (2008), Basu and Malig (2011), Bhaskaran, Hajat et al. (2010), Braga, Zanobetti et al. (2002), Curriero, Heiner et al. (2002), Díaz, García et al. (2005), Hajat, Armstrong et al. (2005)) and case-cross over data (e.g., by Kolb, Radon et al. (2007), Nastos and Matzarakis (2006)) are also found to be used in climate change health researches.

The dependent variables of these studies in the literature reviews are mainly surrounds among the rate of deaths or mortality, number of counts of hospital admissions (inpatient hospital admissions, emergency admissions), GP admissions, morbidity, or disease outcome due to any specific disease with respect to the change in climate factors. Thus most of the dependent variables are in the form of rate of change or count with respect of period of time (e.g. daily or monthly) and for the same reason Poisson regression are one of the most commonly used methods found in the literature reviews (please see the “Models

with count data" at section 2.8). More information about the dependent variable can be also found in section 2.6.

Spatial statistics

Spatial statistics in recent years have received considerable attention. For example, Vaneckova, Beggs et al. (2010) used GLM along with spatial scan statistics and spatial regression to analyse the geographical patterns of heat-related mortality within the metropolitan area of Sydney. Dolney and Sheridan (2006) applied a spatial and temporal analysis using Geographical Information Systems (GIS) to analyse the relationship of extreme heat with the ambulance calls for the city of Toronto, Canada. GIS and geospatial methods were also used by Green, Basu et al. (2010). Davis, Knappenberger et al. (2004) explored the spatial patterns of climate–mortality seasonality in major US cities. Hartz, Golden et al. (2012) used Pearson's correlations and Moran's I index to calculate spatial autocorrelation and thus analyse spatial patterns of heat-related-dispatches. Bassil, Cole et al. (2009) used geospatial methods to map the percentage of heat-related calls (911 medical dispatched data) in each Toronto neighbourhood to demonstrate the potential applications of 911 medical dispatch data due to heat-related illness (HRI), in the summer in Toronto.

Climate threshold

The threshold calculation for any specific climate factors especially for temperature for specific heat wave, region, and health exposure (e.g., disease) is a very useful practice. This is also important for determining a better health alert system. We observed some interesting studies measuring such threshold temperature as an effect of heat waves. Hansen, Bi et al. (2008) analysed the

effect of heat waves and temperature on mental health disorders and mortality. They also calculated related threshold temperature applying Poisson regression accounting for over dispersion and 'hockey stick' method. Beside threshold, extremely hot and cold days were defined using the 99^{th} and 1^{st} percentile, respectively (Medina-Ramón, Zanobetti et al. 2006). Percentiles have been also used by Liang, Liu et al. (2009).

Non-linear models and smoothing

Non-linear relationships of climate and disease exposure are eminent and practically most of them show U- or V-shaped relationships (Muggeo and Hajat 2009). In all the non-linear modelling approaches various types of spline and smoothing techniques are found to be used for measuring precise trends and estimates. For instance, in a multi-lag segmented (piecewise linear) approach, Muggeo and Hajat (2009) used GAM with smooth terms fitted by low-rank penalised splines (B-splines). Spline functions are also used in some other studies by Tong, Ren et al. (2010); Le Tertre, Lefranc et al. (2006); Le Tertre, Lefranc et al. (2006). The smooth and invertible linearizing link function is available both in GLM and GAM models for transforming the expectations of the response variable to the linear predictors. Pudpong and Hajat (2011) used smooth functions of time (b-splines for date) with six degrees of freedom (df) per year were chosen to control for long-term trends and seasonality.

Generalized Additive Model (GAM) is also found to be very useful in various studies: Pauli and Rizzi (2008); Pauli and Rizzi (2006); Tam, Wong et al. (2009); Pauli and Rizzi (2008); Nastos and Matzarakis (2006); Rocklöv and Forsberg (2009); Tong, Ren et al. (2010); Le Tertre, Lefranc et al. (2006). Qian,

He et al. (2008) allowed over dispersion using quasi likelihood in generalized additive models (GAM).

Distributed lag approach and multi-lag segmented modelling approach are found to be efficient for dealing with such non-linear relationships. The latter approach is preferred for considerations of non-linearity and the delayed impact of any climate or pollution factors on health (Muggeo and Hajat 2009).

Limitations and challenges in modelling

While different approaches to studying the effects of an extreme climate event or climate change on health can result in highly variable estimates (Rocklöv and Forsberg 2009), each of these approaches has limitations; collectively they provide information regarding the impacts and can give insight into possible future directions and policies.

Most climate change health research studies have been exploring retrospectively rather than prospectively for future scenarios. A huge portion of the studies has been found to be based on cross-sectional methodologies in spite of their limited power to demonstrate the causality between an exposure variable and the outcome (Lam 2007).

The lag period of climate variables seem to vary in studies and it is fundamental to optimize the length of the lag period for a particular disease, season and country. Any climate change should take into account regional differences (Braga, Zanobetti et al. 2002). Another issue identified in the review is the insufficient duration of the studies to demonstrate the trends and seasonality of the results (Lam 2007).

The ability to make generalisations of existing methodologies is very limited because of the variations of climate exposure relationship across time, region, and populations. Since areas within certain boundaries may have more homogeneous environmental, epidemiologic and demographic characteristics, most of the climate studies are limited to specific regions and populations (Fouillet, Rey et al. 2007; Liang, Liu et al. 2009; Hu, Mengersen et al. 2010; Tong, Ren et al. 2010; Tong, Wang et al. 2010; Ferrari, Exner et al. 2012; Hartz, Golden et al. 2012). Thus research which is not population based could sometimes become confined to generalise the results in all regions (Lam 2007). This introduces uncertainty regarding how to extrapolate from one location or time period to another, given the different population demographics, climate, baseline health status, levels of air pollution, etc.

Measuring the predictive power of the models is found to be very occasional and limited. Many methodologies are found to have weak predictive performance as they are state specific and vary across communities (Chang, Zhou et al. 2010). All these factors make the development of statistical modelling and methodologies more complicated. However, any single model cannot deal and capture the full scenario of climate change and health simultaneously. Therefore, research should focus more on precise locally based modelling approaches with all the influential factors in predictive manners which are essential to improve proactive health measures (Knowlton, Rotkin-Ellman et al. 2009; Tam, Wong et al. 2009).

An efficient health alert system

An efficient health alert system based on a precise methodology to prevent the risk of morbidity and (or) mortality has been a challenge and an earnest quest for researchers to take necessary precautions (McGeehin and Mirabelli 2001; Dolney and Sheridan 2006; Fouillet, Rey et al. 2007; Tan, Zheng et al. 2007). A proper heat mitigation plan for the vulnerable community can also play an important role in this respect (Dolney and Sheridan 2006). This is particularly important for elderly people (Revich and Shaposhnikov 2008).

Conclusion

The diversified nature of climate and its vast associations with the environment and health has made climate health research challenging and complex. Therefore, any recommendations to key policy makers should be given with caution. A general preparedness should be adopted in all countries irrespective of the scenarios and outcomes. Community-wide climate change plans, improved warning systems, better management for facing the impact of climate change are important. Increasing the awareness of people by educating them through community based support and knowledge can play a vital role in improving their adaptive capacity. Better social networking, more informative radio, television and media can be helpful to raise awareness of vulnerable lifestyles and increase the adaptive capacity of the population due to the changing environment.

Although models are useful in conceptualising the dynamic process, more accurate statistical models could achieve a better conceptual representation of an interrelated complex system of climate change and health. No model can completely simulate real life. But such limited empirical studies are the

foundation on which modelling parameters are determined that act as pathways for future research. No doubt that there is always room for improvement, as we progress in time and gain experience to achieve a better understanding of this phenomenon. The impact of climate change on human health has long been a matter of public health and represents a unique different environmental risk factor that will cut across multiple sectors on which human health depends. Therefore, a multidisciplinary approach among health scientists, climatologists, biologists, ecologists and so on is required to face the challenge. Further research is needed to devise and identify the most appropriate statistical approach for both reliability and extrapolative power.

2.9 Chapter summary

In this chapter, we presented an overview of the literature led by a structured search and selection strategy for the sources of articles. We first unveiled exposure-response relations in health care under the climate change context followed by the meteorological, pollution and environmental factors along with the related disease categories that are considered in this research arena. The most susceptible group of people and specific countries and places were reviewed. Most of the studies have dealt with populations in temperate regions which need to expand to other regions around the world. The statistical modelling, related objectives and important results point out the diversified characteristics of the study along with the pros and cons of the existing methods and approaches. The results of the studies reviewed do not cover the whole range of climate change and its impact. Our focus was given to recent work that has focused on disease

outcome and/or hospital admissions. However, it enabled us to gain a sound insight into the issues related to statistical methodologies developed so far, examining the impact of climate on health in this respect.

We separated and extended this chapter to the next, to specify the important factors that are important in modelling and hence should be considered. Thus the next chapter discusses about the important factors that need to be considered in research related to the impact of climate change on health.

Chapter 3

Factors in climate health research

3.1 Introduction

The boundless influence of climate change on ecology and environment is also a multifactorial influence on health. However, the affinity of the factors related to climate, environmental, pollution and health exposures is crucial and failure to properly select these factors may produce conflicting results (Knowlton, Rotkin-Ellman et al. 2009). More information and research is needed surrounding variable selection related to climate and health (McGeehin and Mirabelli 2001). Thus it is imperative to concentrate on the selection of factors for developing any precise model or methodology. In this chapter, we focus on the fundamental factors that need to be considered in modelling.

Section 3.2 indicates the important meteorological factors that need to be inspected in climate change health research. The same insight for pollution factors is given in section 3.3, followed by socio-economic and demographic factors in section 3.4; latitude and regional factors in section 3.5. We describe the lag structure and climate threshold in section 3.6, seasonality of climate change in section 3.7, and conclude the chapter by highlighting other important factors that need to be acknowledged in modelling (section 3.8).

3.2 Climate or meteorological factors

In the literature review (chapter 2), we came across some meteorological factors that should be treated as fundamental for developing any reliable model to measure and predict the impact of climate change in health care. Nonetheless, the inclusion of any meteorological variables depend on the data availability, objective of the study, time, region, disease categories, socioeconomic, demographic factors and so on.

The relationships between temperature and diseases are the main focus of most of the current research (Pauli and Rizzi 2008) and it is also evident in our literature review. In addition to temperature, other influential climate factors should be considered in climate research irrespective of climate zone, time, region, and objective of the studies are: apparent temperature, temperature index, climate index, wind speed, humidity, rainfall, pressure etc. (Table 1).

3.3 Pollution and environmental factors

Pollutants showed significant influences on health in climate change research. Although there is some argument about the inclusion of pollution factors, the literature review proved their importance in modelling the climate health research along with meteorological factors for certain disease categories (Table 2). Thus, pollution factors such as ozone (O_3), particulate matters (PM_{10} or $PM_{2.5}$), Nitrogen dioxide (NO2), and Carbon monoxide (CO) should be considered along with climate factors (Table 2).

3.4 Socioeconomic and demographic factors

Socioeconomic factors, urban living, housing characteristics, including limited access to air conditioning are found to influence health (McGeehin and Mirabelli 2001; O'Neill, Zanobetti et al. 2005; Dolney and Sheridan 2006; Tan, Zheng et al. 2007; Kovats and Hajat 2008; Pauli and Rizzi 2008; Qian, He et al. 2008; Tam, Wong et al. 2009; Ostro, Rauch et al. 2010; Pudpong and Hajat 2011). Along with socioeconomic factors, social network, access to media, various communities, and so on are also important factors for modelling climate change (McGeehin and Mirabelli 2001; Kovats and Hajat 2008). In addition to age, race, and ethnicity are found to be a factor, particularly for the black population (McGeehin and Mirabelli 2001; Basu and Samet 2002; Medina-Ramón, Zanobetti et al. 2006; Kaiser, Le Tertre et al. 2007; Knowlton, Rotkin-Ellman et al. 2009). This could be due to differences in lifestyle, food habit along with socioeconomic conditions.

3.5 Latitude and regional factors

The changes in meteorological variables are already adversely affecting health and environment with different scale and rate in various climate zones (WHO 2008). The articles in the review have focused on various geographical regions, highlighting the effects of different latitudes and climate zone. The results of these studies are compatible to respective location due to socioeconomic factors, lifestyle, and cultural factors which vary in any specific climate zone. We have tabulated the most frequent countries that have come across in our review (Table 3).

Table 3: Number of articles in various countries in the review

Most Frequent countries in the review	Number of articles in the review
United States	14
Australia	7
United Kingdom	5
Italy	5
China	4
Brazil	3
Greece	3
France	3
Canada	2
Taiwan	2

The articles from the USA are more diversified in terms of diseases, climate, and socioeconomic factors, along with methodologies, compared to others. All four articles from China focused mainly on temperature or heat wave using various periods of lag and studied the effect on elderly people. The number of articles from the United Kingdom is very limited and like China almost all have focused only on the impact of temperature. There are also some articles in the review from other countries including Denmark, Chile, Thailand, Russia, Bulgaria, Netherland, Germany, Sweden, India, Singapore, Spain, and Taiwan represents the global interest surrounding this research area.

3.6 Lag structure and climate threshold

Lag structures

The time between the day of disease onset (or mortality) and meteorological exposure is generally termed as the lag period (Hu, Mengersen et al. 2010). The lag effect is important in climate research as the susceptibility rate of a population varies according to disease and geographical area, and exhibits different lag

structures of climate variables depending on the season of the year (Pudpong and Hajat 2011). Hospital admissions predominantly occur within a few days after the exposure of high temperature (Schwartz, Samet et al. 2004; Fernández-Raga, Tomás et al. 2010). The effects of low temperatures appear approximately 10-days after the weather changes, and only after 1 or 2 days for high temperature. Díaz, García et al. (2005) and Kolb, Radon et al. (2007) found an association of hot weather up to 0 to 3 days and cold weather starting after 2 days. Apparently the hot weather has a very quick reaction on health compared to cold weather (Braga, Zanobetti et al. 2002; Pattenden, Nikiforov et al. 2003; Hajat, Armstrong et al. 2005; Nastos and Matzarakis 2006; Muggeo and Hajat 2009; Tam, Wong et al. 2009; Bhaskaran, Hajat et al. 2010). Thus various types of lag period have been used by researchers depending on the nature of disease, seasons, and research characteristics, and to date there is no clear standardised general form and duration of lag structure and period yet.

The most common form of lag measurement is the mean, moving average and cumulative average (Kovats, Hajat et al. 2004; Basu, Feng et al. 2008; Basu and Malig 2011). For this reason, we need to be explicit about the lag structure and duration for efficient results. Some examples of the lag structure we came across include: 0-1, 2-7, 8-14, 15-21, 22-28 days (Bhaskaran, Hajat et al. 2010) , 0-1 to 0-5 days lag (Tam, Wong et al. 2009), 0-1 and 0-13 days lag (Pudpong and Hajat 2011), 0-8 weeks, 0-1 weeks, and 0-4 weeks (Hu, Mengersen et al. 2010), 1-7 days (Ferrari, Exner et al. 2012). It is important to investigate the most appropriate structure of the lag periods in climate health research.

Non-linearity in lag period and climate threshold

Population with a temperate climate generally shows non-linear U, V, N or even J shaped relationships (Braga, Zanobetti et al. 2002; Pattenden, Nikiforov et al. 2003; Pauli and Rizzi 2008; Muggeo and Hajat 2009) and the optimum temperature value(s) corresponding to the lowest point of the U-, V- or J- shaped exposure-disease relationship curve yielded the opportunity for calculating the threshold in climate change health research (Curriero, Heiner et al. 2002). **Threshold temperature** denotes that mortality/morbidity rates are smallest at this temperature and those levels will increase if the temperature increases or decreases from this point (Kalkstein and Davis 2005). Since the related exposure-response relationship is non-linear, the cold (lower than optimum temperature) effects and hot (higher than optimum temperature) effects were usually investigated separately. The threshold or optimum temperature varies according to population, place and disease or 'cause of death'. For example, in the Netherlands between 1979 and 1997, the optimum value was 16.5^0C for total mortality, cardiovascular mortality, respiratory mortality and mortality among those >65 years, whereas for mortality due to malignant neoplasm and mortality in the younger age group, the optimum value was 15.5^0C and 14.5^0C, respectively (Huynen, Martens et al. 2001).

Methods used for climate threshold

Several methods are available in the literature to select the threshold temperature. Kalkstein and Davis (2005) calculated the threshold temperature using the smallest total sum of squares, while Donaldson, Keatinge et al. (2003) calculated it by computing the mean daily mortality over a range of 31^0C at successive

0.11^0C intervals for each year of the data. Recently, smoothing curves were plotted to generate the temperature point at which the minimum mortality occurred (El-Zein, Tewtel-Salem et al. 2004). Percentiles (e.g. 99^{th} or 90^{th}) of temperature have also been used as the threshold temperatures in a meta-analysis (Anderson and Bell 2009). Muggeo developed a segmented approximation to compute the threshold temperature which has been proposed in several studies (Muggeo 2003; Michelozzi, Kirchmayer et al. 2007). Another way to divide hot and cold periods was according to the four seasons where data were analysed for spring, summer, autumn and winter separately (Basu and Samet 2002; Carson, Hajat et al. 2006). A more robust and precise method needs to be developed for calculating the climate and pollutant threshold for specific disease categories.

In general the outcome or event variable of the studies that considered the delayed effect or calculating threshold in research related to climate change and health are disease outcome or mortality. Such disease outcomes are in the form of hospital admissions, GP visits and so on and mortality are described as death due to certain disease or non-accidental death. For example, mortality has been considered as event variable by several studies like (Huynen, Martens et al. 2001; Braga, Zanobetti et al. 2002; Curriero, Heiner et al. 2002; Donaldson, Keatinge et al. 2003; Pattenden, Nikiforov et al. 2003; Díaz, García et al. 2005; Hajat, Armstrong et al. 2005; Kalkstein and Davis 2005; Kolb, Radon et al. 2007; Basu, Feng et al. 2008; Muggeo and Hajat 2009; Tam, Wong et al. 2009; Fernández-Raga, Tomás et al. 2010; Basu and Malig 2011). Among them some studies worked on mortality due to specific disease like cardiovascular mortalities (Tam, Wong et al. 2009), cardiovascular, respiratory, and digestive diseases (Fernández-

Raga, Tomás et al. 2010) and so on. There also studies (e.g. Bhaskaran, Hajat et al. (2010), Ferrari, Exner et al. (2012), Hu, Mengersen et al. (2010), Kovats, Hajat et al. (2004), Nastos and Matzarakis (2006), Pauli and Rizzi (2008), Pudpong and Hajat (2011), Schwartz, Samet et al. (2004)) that considered hospital or GP admissions or morbidity as outcome event and considered delayed effect or (and) threshold calculation. More information can on the event variables can be found in section 2.6.

3.7 Seasonality

Climate variability is the oscillation around the average climate, for various diseases. Therefore, seasonality has become one of the most frequently used terminologies in the climate change health research. The first detectable changes in human health may well be alterations in the geographical range (latitude and altitude) and seasonality of certain vector-borne infectious diseases (McMichael, Haines et al. 1996). A change in the frequency and intensity of heat waves and cold spells would affect seasonal patterns of morbidity and mortality (McMichael, Haines et al. 1996). The amplitude of seasonal variability is generally larger than that of the diurnal cycle at high latitudes and smaller at low latitudes. Many studies considered seasonality in measuring the fluctuations of climate and disease frequencies. The winter dominance of mortality is widely recognised throughout the US and in many other mid-latitude countries that experience some climate seasonality (Davis, Knappenberger et al. 2004). The cases of cardiovascular and respiratory mortality are found to have more seasonal variations than others and their seasonal component so dominate the long-term signal that it is even evident

in plots of daily data (Davis, Knappenberger et al. 2004). Davis, Knappenberger et al. (2004) explored how mortality seasonality has changed over time. The future net mortality changes might arise under different seasonal patterns of climate change. Considering all these facts, seasonality should be treated as a fundamental factor in modelling the impact of climate change on health. However, seasonality of the impact of climate needs to ensure that respective climate factors (e.g., temperature, rainfall) were similar enough to assume linearity within each stratum (Basu and Samet 2002).

3.8 Other factors

Time unit measurement

The correct parameterisation and the time unit (days, weeks, months) for measuring the disease exposures are crucial in climate–health studies. The mean is commonly used for temperature even combined with other factors (Pudpong and Hajat 2011). Other parameterisations have also been used, such as a 3-hour maximum apparent temperature and 5-days cumulative average of the apparent temperature (Wichmann, Andersen et al. 2011), 10-day moving average of the mean temperature, cumulative variable for maximum temperature (Fouillet, Rey et al. 2007). Thus studies concluded using various types of time spans (e.g., days, weeks, months, and so on). Similar to the lag structure, the time unit of explanatory variables also depend on the nature and characteristics of disease, population, seasons, place, data availability, and objective of the study.

Heat wave

The definition of "heat wave" varies in many studies and it is both imperative and challenging to standardise the definition based on correct parameters for specific regions (Pauli and Rizzi 2008; Revich and Shaposhnikov 2008; Tong, Wang et al. 2010; Ma, Xu et al. 2011). This will help to understand its impact on health and develop appropriate public health intervention strategies to prevent and mitigate the impact of climate change following heat waves (Bassil, Cole et al. 2009; Pudpong and Hajat 2011). However heat waves have been defined loosely in most of the studies (Kovats and Hajat 2008) and thus the overall results of any research study using heat wave depend on its definition (Huynen, Martens et al. 2001) along with the reference period of climate health research (Knowlton, Rotkin-Ellman et al. 2009; Ma, Xu et al. 2011). This is also true for "cold-wave."

Quality of data

The lack of good quality data for meteorological factors and pollutants is one of the main difficulties faced by researchers (Bartzokas, Kassomenos et al. 2004; Medina-Ramón, Zanobetti et al. 2006; Qian, He et al. 2008; Mentzakis and Delfino 2010). Missing data are also common along with misclassification (Kolb, Radon et al. 2007; Pudpong and Hajat 2011), measurement errors (Qian, He et al. 2008) and lack of personal health care data due to patient confidentiality, which challenges the precision of results (Pauli and Rizzi 2008; Sung, Chen et al. 2011). For these reasons, current studies are conducted with limited use of climate and pollution variables. These may have produced more reliable results if they had considered all the important factors related to specific diseases (Bartzokas, Kassomenos et al. 2004); the same argument is also true for health outcomes and

disease exposure. Therefore, the quality of data is crucial in climate health research and more efforts are needed to improve this aspect (McGeehin and Mirabelli 2001). Moreover, daily mortality and morbidity data by diseases are required as weather conditions typically vary on a daily basis. One possible surrogate for morbidity is the use of ambulatory medical care. However, data such as number of emergency calls and number of ambulance dispatches often have lots of problems with regard to their accuracy and completeness (Dolney and Sheridan 2006; Alessandrini, Zauli Sajani et al. 2011). Again data need to be standardised with time and locality to improve the quality and precision of climate research (McGeehin and Mirabelli 2001).

Use of hospital admissions data

Hospital admissions data are one of the main identifiers of disease exposure: morbidity and mortality. Lots of studies aimed to measure the relationships of climate and environmental factors with health hazards using hospital outcomes of different forms and in most cases significant relationships have been exposed following a sudden change in climate (Bartzokas, Kassomenos et al. 2004; Pauli and Rizzi 2006; Pauli and Rizzi 2008; Pauli and Rizzi 2008; Liang, Liu et al. 2009; Rocklöv and Forsberg 2009; Wang, Barnett et al. 2009; Alonso, Achcar et al. 2010; Green, Basu et al. 2010; Hu, Mengersen et al. 2010; Khalaj, Lloyd et al. 2010; Ostro, Rauch et al. 2010; Tong, Ren et al. 2010; Ferrari, Exner et al. 2012; Hartz, Golden et al. 2012). However, the use of aggregate hospital admissions data limits the amount of individual-level information (Liang, Liu et al. 2009; Wang, Barnett et al. 2009; Green, Basu et al. 2010; Hu, Mengersen et al. 2010) and in some countries it only covers the people with medical insurance which

brings about the possibility of selection bias (Pudpong and Hajat 2011). Along with this, a huge amount of information is missing in hospital data, for instance data concerning people who are treated in general practices (GP) or outpatient clinics which do not result in hospital admission. Data can also vary between hospitals and physicians due to recording of disease diagnosis, classification, admission criteria, and treatment procedure. Hence, it is important to treat these inconsistencies in hospital data and standardise them based on unique geographical information and other measurement units to avoid biases in the results.

3.9 Chapter summary

This chapter focuses on the factors and issues that need to be considered for developing any model. It is acting as a connection between literature review and thoughts for developing a model. We started with climate and pollution variables followed by socioeconomic and demographic factors. Lag structure, climate threshold, seasonality, and other important issues have been discussed. In the next chapter, we summarise the datasets used in the thesis, missing values, and data management.

Chapter 4

Data sets used

4.1 Introduction

This chapter aims to describe the data sets used in the research. We also described the study population and coverage area, data management & cleaning, linking administrative data sets, and issues related to the missing values. We begin by describing the population covered in the study in section 4.2, followed by the variables of the Hospital Episode Statistics (HES), the core data of this research in section 4.3. We highlight the source and variables related to climate in section 4.4. Section 4.5 covers the data related to the air quality (pollution data) and section 4.6 describes the process and challenges for linking all the data sets. Finally, section 4.7 illustrates data management regarding the missing values and aggregations of the data.

4.2 Study population and catchment area

This research covered the population of Greater London as study population. We considered all age groups for the period 1 January 2000 – 31 December 2009. The main reasons for choosing this:

a) Greater London is the highest density populated area in England (ONS 2012). For this reason, we have more hospital admissions for Greater London compared to other places in England. This is very important if we want to concentrate to any specific disease category.

b) Greater London is more diverse in terms of population characteristics and ethnicity.

c) Air pollution is a big concern for Greater London for the same reason in (a). Thus we will have more opportunity to examine the compounded impact of air pollutants and climate change.

d) We will have the opportunity to use the spatial statistical approach and compare Greater London with the other big metropolitan area in the future (e.g., Greater Manchester).

4.3 Hospital episode statistics

HES[1] is a data warehouse containing details of all admissions, outpatient appointments and A&E attendances at NHS hospitals in England. Along with the admission statistics, it contains all the administrative details of all patients. This data are collected during a patient's time in hospital and are submitted to allow hospitals to be paid for the care they deliver. HES data are designed to enable secondary use, that is used for non-clinical purposes, of this administrative data. HES processes over 125 million admitted patient, outpatient and accident and emergency records each year (HES 2013).

[1] http://www.hscic.gov.uk/hes

HES was originally conceived in 1987 following a report on the collection and use of hospital activity information published by a steering group chaired by Dame Edith Körner (1921-2000) (HES 2013). Initially, data for HES publications were collected annually from provider submissions. After a number of years, the frequency of collections increased to quarterly to allow analysis and investigation (these were not published) and a final annual publication was released at the end of the year. HES data are now collected monthly (HES 2013). It is a record-based system that covers all NHS trusts in England, including acute hospitals, primary care trusts, and mental health trusts. HES information is stored as a large collection of separate records - one for each period of care - in a secure data warehouse. In our research, we have used HES inpatient data for the greater London area for 10 years (2000-2009). We used the episodes of the hospital admissions for our study, not spell. A spell relates to the whole hospital stay of a patient, from admission to discharge. For complex patients the spell may contain many episodes of care under different consultants. We created a database in the university server using the flat files of HES inpatient data.

Variables and factors in HES

HES inpatient or admitted patient data consist of different sections followed by respective subsections. The main sections of the inpatient data are: admissions, augmented/critical care period, clinical, discharges, episodes and spells, geographical, health care resource groups, maternity, organisation, patient, patient pathway, period of care, practitioner, psychiatrist, socioeconomic and system. The name variables and factors from HES inpatient we have used in our research are listed in **Table 4**.

Table 4: Selected variables from HES inpatient data

HES inpatient variables	**HES inpatient variables**
Administrative category of the patient	Hospital provider spell number
Patient age at the end of episode	Primary diagnosis
Patient age at the start of episode	Episode order
Ethnic category of the patient	Current electoral ward
HES generated patient identifier (hesid)	Local authority district
Postcode district of patient's residence	Government office region of residence
Sex of patient	County of residence
Method of admission of the patient	Government office region of treatment
Source of admission of the patient	Regional office of residence
Bed days within the year of the patient	

4.4 Meteorological data

We collected the Met office observational station data sets[2] for meteorological factors from the stations at Heathrow airport and London St. James Park. In both cases, we used the data set for the period 2000-2009.

The Met Office is the UK's national weather service, and deals with weather predictions, forecast, climate change and weather science research. We used the Met office observational data sets from the station at Heathrow airport and St. James Park, London for collecting daily observational data for temperature (maximum, minimum and mean), daily total rainfall, mean wind speed, daily sun hours, radiation, relative humidity, daily mean pressure.

Temperature is the main important meteorological factor because of its quick and detrimental role in the environment and health (Fernández-Raga, Tomás et al. 2010; Khalaj, Lloyd et al. 2010; Tong, Ren et al. 2010; Ferrari, Exner

[2] http://www.metoffice.gov.uk

et al. 2012). Thus almost all the research studies and scientific articles based on health, environment, and climate change have considered temperature (mean, maximum or minimum). However, other climate factors like humidity, wind speed, rainfall showed relationships on some disease exposures (section 2.6). For this reason, we considered these variables to check their impact on hospital admissions besides temperature (**Table 5**).

Table 5: Variables related to meteorological and pollutants

Meteorological variables (Units)		**Variables related to Pollutions (Units)**
Daily maximum Temperature (^{0}C)	Daily Sun hours (hours)	Ozone (µg/m^3)
Daily mean Temperature (^{0}C)	Daily radiation (KJ/sqm)	PM2.5 (µg/m^3)
Daily minimum Temperature (^{0}C)	Daily relative humidity (%)	PM10 (µg/m^3)
Daily Total Rainfall (mm)	Daily mean pressure (hpa/mb)	
Daily mean Wind speed (knots)		* PM: particulate matter

4.5 AIR quality data

Air pollution is assumed to have a significant role in some disease exposures that compound the effect of climate change on health (World Health Organisation (WHO) 2006). For this research we have used London AIR Quality Network (LAQN)[3].

The LAQN is a group of air quality monitoring stations in the 33 London Boroughs, Essex, Kent, and Surrey. Each borough funds the monitoring within its own area, with the exception of eight sites in London, which are funded by the

[3] www.londonair.org.uk

Department of Environmental, Food, and Rural Affairs (Defra) and are affiliated with the Automatic Urban Rural Network (AURN). The LAQN was formed in 1993 to coordinate and improve air pollution in London and operated & managed by the Environmental Research Group (ERG) at King's College London. QA/QC audits are carried out by the National Physical Laboratory (NPL). Each borough funds air quality monitoring in its own area.

Pollutants

There are various types of pollutants that are collected under the LAQN project. These are: Particulate matters (PM10, PM2.5), Ozone, Nitrogen Die Oxide (NO_2), Nitrogen Oxide (NO), Sulphur Die Oxide (SO_2), and so on. All these pollutant factors are ratified after collecting from different types of local stations.

For our study we have used Ozone and PM10 (**Table 5**). According to the literature review, these are the most significant air pollutant on health and since our catchment area (Greater London) do not have a lot of industries we didn't consider Sulphur dioxide for the study. We did not use PM2.5 as air pollutants because of insufficient PM 2.5 observations or huge missing values in LAQN during the study period.

4.6 Linking the three data sets

Linking all three data sets was very important and finding some suitable linking factors or variables for all three datasets was challenging. This was mainly because our HES dataset does not contain the full postcode of the patient (being sensitive). We thus used the HES variable **resro** (indicates the regional office of residence) to identify the patients from Greater London. The **resro** contains the

code for the regional office in which the patient lived immediately before admission. It is derived from the patient's postcode in the field **homeadd** (or home address). We linked the climate variables and the pollutants in the greater London area matching the date of admission and **resro** from the HES inpatient dataset.

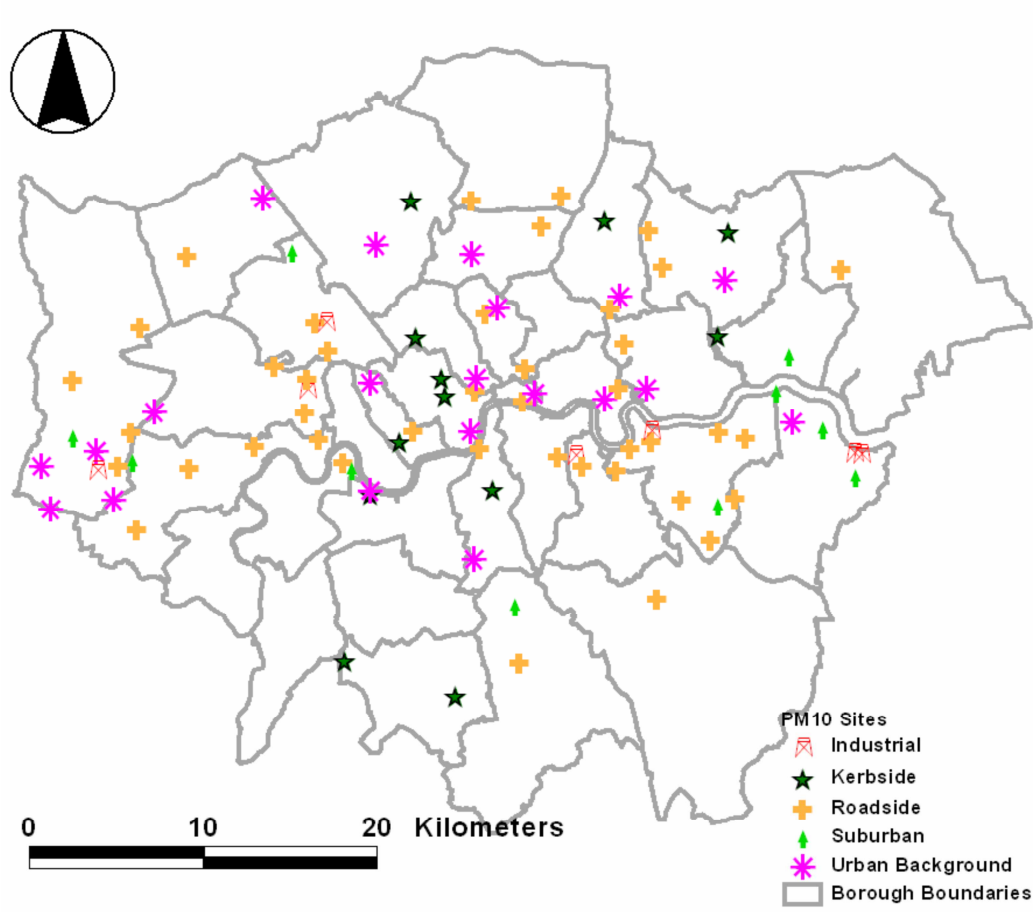

Figure 3: Greater London Air Quality Network

4.7 Data management and cleaning

We found some issues related to the data in Hospital Episode statistics. We cleaned few cases for the invalid date of birth (DOB) recorded as '1582-10-15' (15 October 1582) in the raw data. These are the cases where the data provider has entered an invalid code into a date field (other than one which can be re-derived), i.e. a collection of characters that cannot be recognised as a date by the HES

database software) the date 15th October 1582 (the first date on the Julian calendar) will be substituted. This serves as an indication that the field cannot be used.

There were 1,055,355 rows (episodes) in the raw data for the Greater London the period 2000-2009 including all diseases categories, admissions methods, and age groups. Among them there were 31599 emergency admissions due to lower respiratory diseases in greater London the 10 year period (2000-2009). We choose chronic lower respiratory disease (ICD-10, J40-J47), because this is most climates effected disease category observed in the literature review. This is our main disease exposure data file for the study. We then count the daily number of chronic lower respiratory diseases admissions and link with relevant climate and air pollutants variables.

Data aggregation and missing values

The Met Office observational data were used as a part of the climate information. In Greater London we have two main weather stations: London Heathrow and London St. James's Park. The Heathrow weather station (NRG: 5077E 1767N, altitude: 25 metres, Latitude: 51:48N, Longitude: 00:45W) is more important than St. James Park (NRG: 5298E 1801N, altitude: 5 metres, Latitude: 51:50 N, Longitude: 00:13 W) because of the coverage of area and attributes (**Table 6**). Therefore, we mainly used Heathrow and St. James Park stations to incorporate the missing values of Heathrow. For example, the Heathrow station has 2 missing values for Rainfall, 629 missing values for Wind speed, 1 case for Relative Value, 1 case for mean pressure, 79 for daily radiation. We used the **AIRGENE** algorithm for dealing these missing values (Bhaskaran, Hajat et al. 2010).

Table 6: Properties of the weather stations used

Greater London Weather Stations from Met Office													
Station	**NRG**	**Altitude**	**Latitude**	**Longitude**	**Daily Tem**			**Daily rain**	**Wind Speed**	**Sun hours**	**R. Humidity**	**Radiation**	**Pressure**
					Max	**Min**	**Mean**						
Heathrow	5077E 1767N	25 metres	51:48N	00:45W	Y	Y	Y	Y*	Y*	Y	Y*	Y*	Y*
LONDON, ST. JAMES'S PARK	5298E 1801N	5 metres	51:50N	00:13W	Y	Y	Y	Y	No	No	Y	No	No

AIRGENE algorithm

The AIRGENE algorithm is an improved formula to replace missing values on the aggregate level. The general idea is as follows:

A missing value on day i from monitor j is replaced by the period average of monitor j plus a standardised value of day i over all monitors multiplied by the period standard deviation of monitor j (See the supplemental materials of Bhaskaran, Hajat et al. (2010)). This can be written as follows:

$$\hat{x}_{ij} = \bar{x}_{.j} + \bar{z}_{i.} s_{.j} \qquad (6.1)$$

$$\text{Where, } \bar{z}_{i.} = \frac{\sum_{j=1}^{n} \left(\frac{x_{ij} - \bar{x}_{.j}}{s_{.j}} \right)}{n}$$

In this manner we achieve estimates that consider not only differences in mean values, but also differences in variability between monitors. If all monitors are missing for one day, the averages from the day before and after will be taken.

Figure 4: Dealing with the missing values in air quality data

Mean imputation

We used the mean imputation method for replacing the missing values for the air pollutants (Ozone and PM 10) in the London Air Quality Network data. Mean imputation is popular in this area because of its computational aspects. There are too many missing values in the LAQN network for the PM 2.5 for the study period to make good representative data. For the study period, there are 10 Boroughs (Hammersmith and Fulham, Lambeth, Islington, Merton, Bromley,

Havering, Barking and Dagenham, Waltham Forest, Barnet, Harrow), which have missing values for Ozone and 3 Boroughs (Merton, Sutton, Bromley) have missing values for PM 10. For these cases we used the values from the nearest Boroughs and the average of those Boroughs (mean imputation) for dealing with the missing values. For example, Ozone missing values of Hammersmith and Fulham, we used the average of the 3 nearest Boroughs: Kensington and Chelsea, Wands worth and Brent. For PM10 missing values of Sutton, we used the average of the 2 nearest available Boroughs: Kingston upon Thames and Croydon.

4.8 Chapter summary

Here we describe the data sets used for this research. We also summarised the study population, factors of the three data sets on hospital admissions, climate, and pollution. Furthermore, we described the data aggregations and the techniques used for tackling missing values. The following chapter represents the theories related to Generalized linear model and related results in our context.

Chapter 5

Generalized linear modelling

5.1 Introduction

In this chapter, we summarise various statistical models and their properties related to this research. We begin by reviewing the generalized linear models (GLM) (section 5.2). Here, we illustrate the theories of GLM and their relations to our research. Next, we describe GLM modelling with count data (extension of GLM) in section 5.3, followed by some special circumstances using count data. In section 5.4, we mention about other modelling approaches that are also useful in climate change health research but not directly related to our work. In section 5.5, we describe some GLM modelling approaches using only temperature. Then we illustrate in section 5.6, how multiple climate and pollution factors can improve the model performance.

5.2 Theory of Generalized linear model

5.2.1 The model

A linear model is a statistical model that can be written

$$y_i = X_i\beta + \epsilon_i, \quad \epsilon_{i\,\widetilde{i.i.d}}\, N(0, \sigma^2) \tag{5.1}$$

Where y_i is a response variable and follows independently and identically distributed from the exponential family of distribution, **X** is a model matrix with elements usually depending on some predictor variables (explanatory variables or covariates, X_i's), $\epsilon_i's$ are random variables. $\boldsymbol{\beta}$ is a vector of unknown parameters.

Exponential family of distributions includes distributions such as Poisson, Gaussian (normal), binomial and gamma. A feature of exponential family distributions is that their shape is largely determined by their mean, μ_i ($\mathbb{E}(y_i) = \mu_i$). GLMs are usually written in terms of $link\ function, g$ (the inverse of a smooth monotonic function), as follows

$$g(\mu_i) = X_i\beta, \quad y_i \underset{indep.}{\sim} Exponential\ family\ distribution, \tag{5.2}$$

5.2.2 The exponential family of distributions

The response variable y_i in GLM can have any distribution of the *exponential family*. A distribution belongs to the exponential family of distributions if its probability density function, or probability mass function, can be written as

$$f_\theta(y) = exp\,[\{y\theta - b(\theta)\}/a(\phi) + c(y,\phi)], \tag{5.3}$$

where b, a and c are arbitrary functions, ϕ an arbitrary 'scale' parameter, and θ is known as the 'canonical parameter' of the distribution.

For example, it is easy to see that the normal distribution is a member of the exponential family since

$$f_\mu(y) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \tag{5.4}$$

$$= exp\left[\frac{-y^2 + 2y\mu - \mu^2}{2\sigma^2} - log\left(\sigma\sqrt{2\pi}\right)\right]$$
$$= exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - log\left(\sigma\sqrt{2\pi}\right)\right]$$

which is of exponential form, with $\theta = \mu, b(\theta) = \frac{\theta^2}{2} \equiv \frac{\mu^2}{2}$, $a(\phi) = \phi = \sigma^2$and $c(\phi, y) = y^2/2(\phi) - \log\left(\sqrt{\phi 2\pi}\right) \equiv -\frac{y^2}{2\sigma^2} - \log\left(\sigma\sqrt{2\pi}\right)$.

Similar breakdown for other members of the exponential family of distributions (e.g., Poisson, Bionomial, Gamma, and Inverse Gaussian) is possible and can be found on page 61 of Wood (2006).

The log likelihood of θ, given a particular y, is simply $log[f_\theta(y)]$ considered as a function of θ and can be given as

$$l(\theta) = \{y\theta - b(\theta)\}/a(\phi) + c(y, \phi) \quad (5.5)$$

Based on the log likelihood function above, we can devise the general expressions for the mean and variance of exponential family distributions in terms of a, b and ϕ. The mean of the response variable in GLM can be given as

$$\mu_i = \mathbb{E}(Y) = b'(\theta) \quad (5.6)$$

i.e. the mean, of any exponential family random variable, is given by the first derivative of b w.r.t. θ, where the form of b depends on the particular distribution. This equation is the key to linking the model parameters, β of a GLM to the canonical parameters of the exponential family. In a GLM, the parameters β determine the mean of the response variable, and, via 5.6, they thereby determine

the canonical parameter for each response observation. Similarly, the variance of the response variable in GLM can be given as

$$var(Y) = b''(\theta)a(\phi) \tag{5.7}$$

Here a could in principle be any function of ϕ. Interested readers can go through Wood (2006) and other basic GLM references to find the mathematics for getting the form of mean and variance for GLM.

In equation (5.7), if ϕ is known, normally there is no difficulty in handling any form of a in GLM. However, for unknown ϕ, it might be difficult to work, unless we can write $a(\phi) = \phi/\omega$, where ω a known constant. The expression $a(\phi) = \phi/\omega$ allows the possibility of unequal variances in models based on the normal distribution, but in most cases ω is simply 1. Hence we now have

$$var(Y) = b''(\theta)\phi/\omega \tag{5.8}$$

In is often convenient to consider var(Y) as a function of $\mu \equiv \mathbb{E}(Y)$, and since μ and θ are linked via (5.6), we can always define a variance function $V(\mu) = b''(\theta)/\omega$, such that $\mathrm{Var}(Y) = V(\mu)\phi$.

5.2.3 The canonical link functions

The link function provides the relationship between the linear predictor and the mean of the distribution function and thus links them in one equation. The canonical link g, for a distribution is the link function such that $g(\mu_i) = \theta_i$, where θ_i is the canonical parameter of the distribution. For example, for Poisson distribution the canonical link is the log function (See

Table 7 for other examples).

Table 7: Common distributions and canonical link functions

Distribution	Support of Distribution	Typical Uses	Link Name	Link Function
Normal	Real: $(-\infty, +\infty)$	Linear-response data	Identity	$\boldsymbol{X\beta} = \mu$
Exponential / Gamma	Real: $(-\infty, +\infty)$	Exponential-response data, scale parameters	Inverse	$\boldsymbol{X\beta} = -\mu^{-1}$
Inverse Gaussian	Real: $(0, +\infty)$		Inverse Squared	$\boldsymbol{X\beta} = -\mu^{-2}$
Poisson	integer: $[0, +\infty)$	Count of occurrences in fixed amount of time/space	Log	$\boldsymbol{X\beta} = \ln(\mu)$
Bernoulli	Integer: $[0, 1]$	Outcome of single yes/no occurrence	Logit	$\boldsymbol{X\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$
Binomial	Integer: $[0, N]$	Count of # of "yes" occurrences out of N yes/no occurrences		
Categorical	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1	Outcome of single K-way occurrence		
Multinomial	K-vector of integer: $[0, N]$	Count of occurrences of different types (1 ... K) out of N total K-way occurrences		

The main advantages of the canonical link functions are: μ_i stays within the range of the response variable and provides some mathematical advantages in performing the likelihood maximisation. The canonical link function has many practical uses. For example, for a GLM with an intercept term and canonical link, the residuals will sum to zero. Another one is in categorical data analysis using log linear models; it provides a means of ensuring, via the specification of the

model, that totals which were built into the design of a study can be preserved in any model.

5.2.4 Fitting Generalized linear model

In GLM we have an n-vector of independent response variables, Y_i where $\mu_i \equiv \mathbb{E}(Y_i)$, and $g(\mu_i) = X_i\beta$. Since Y_i are mutually independent, the likelihood of β is

$$L(\beta) = \prod_{i=1}^{n} f_{\theta_i}(y_i) \tag{5.9}$$

and hence the log-likelihood of β is

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} log\left[f_{\theta_i}(y_i)\right] \\ l(\beta) &= \sum_{i=1}^{n} \{y_i\theta_i - b_i(\theta_i)\} / a_i(\Phi) + c_i(\Phi, y_i), \end{aligned} \tag{5.10}$$

where the dependence of the right hand side on β is through the dependence of the θ_i on β. The functions a, b and c may vary with i. But Φ is assumed to be the same for all i. It suffices to consider only cases where we can write $a_i(\phi) = \phi/\omega_i$, where ω_i is a known constant (usually 1), in which case

$$l(\beta) = \sum_{i=1}^{n} \omega_i\{y_i\theta_i - b_i(\theta_i)\}/\Phi + c_i(\Phi, y_i) \tag{5.11}$$

We can maximise the above equation by differentiating l w.r.t. each element of β, setting the resulting expressions to zero and solving for β Wood (2006). Thus

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\Phi}\sum_{i=1}^{n} \omega_i \left(y_i \frac{\partial \theta_i}{\partial \beta_j} - b_i'(\theta_i)\frac{\partial \theta_i}{\partial \beta_j}\right) \tag{5.12}$$

The equations to solve for β are

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \quad \forall\ j \tag{5.13}$$

However, these equations are exactly the equations that would have to be solved in order to find β by non-linear weighted least squares, if the weights $V(\mu_i)$ were known in advance and were independent of β. In this case the least squares objective would be

$$S = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)}, \tag{5.14}$$

where μ_i depends non-linearly on β, but the weights $V(\mu_i)$ are treated as fixed. To find the least squares estimates involves solving $\frac{\partial S}{\partial \beta_j} = 0 \quad \forall\ j$, but this system of equations is easily seen to be (5.12), when the $V(\mu_i)$ terms are treated as fixed.
This correspondence suggests a fitting method. Iterate the following two steps to convergence

i. Given the current $\hat{\mu}_i$ estimates, evaluate the $V(\hat{\mu}_i)$ values

ii. Find a value of β which reduces

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

(the dependence on β is through μ, but not $\hat{\mu}_i$). Let this improved parameter vector be denoted $\hat{\beta}$, and use it to update $\hat{\mu}$.

At convergence $\hat{\beta}$ must satisfy (5.12).To implement this method we need to be able to find the required improved parameter vectors at step 2. To do this, just replace μ_i by its first order Taylor expansion around $\hat{\mu}_i$, so that

$$y_i - \mu_i \simeq y_i - \hat{\mu}_i - \sum_i \frac{\partial \mu_i}{\partial \beta_j}(\beta_j - \hat{\beta}_j)$$

With exact equality at $\hat{\beta} = \beta$ (derivatives evaluated at current $\hat{\beta}$). Now, writing the linear predictor as $\eta_i = X_i\beta$

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{X_{ij}}{g'\mu_i}$$

Hence

$$\sum_i \frac{(y_i - \mu_i)^2}{V(\hat{\mu}_i)}$$

$$\simeq \sum_i \frac{\left(g'(\hat{\mu}_i)\, y_i - g'(\hat{\mu}_i)\mu_i - X_i\beta + X_i\hat{\beta}\right)^2}{g'(\hat{\mu}_i)^2 V(\hat{\mu}_i)} \tag{5.15}$$

$$= \sum_i \omega_i (z_i - \boldsymbol{X}_i\beta)^2 \tag{5.16}$$

where $z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i) + X_i\hat{\beta}$ and $\omega_i = g'(\hat{\mu}_i)^{-2}V(\hat{\mu}_i)^{-1}$. But (5.15) is just a weighted linear least squares problem, which is easily minimized w.r.t. β using standard least squares methods, making it easy to find an improved $\hat{\beta}$. The final expression of $\hat{\beta}$ can be written as:

$$\hat{\beta} = (X'WX)^{-1}X'Wz$$

Hence we arrive at the following GLM fitting algorithm. Iterate the following to convergence. . .

i. Given the current $\hat{\eta}$ and $\hat{\mu}$ estimates, calculate *pseudodata* $\boldsymbol{z}$ and weights $\boldsymbol{w}$, as defined above.

ii. Minimize $\sum_i \omega_i (z_i - \boldsymbol{X_i}\beta)^2$ w.r.t. β to obtain an improved estimate $\hat{\beta}$.

iii. Evaluate a new linear predictor estimate $\hat{\eta} = \boldsymbol{X}\hat{\beta}$ and new fitted values $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

The iteration can be started by setting $\hat{\mu} = y$ (with modification to avoid e.g. $log(0)$). The method is known as *Iteratively Re-weighted Least Squares* (IRLS). McCullagh and Nelder (1989) prove that this algorithm is equivalent to Fisher scoring and leads to maximum likelihood estimates.

5.2.5 The sampling distribution of $\widehat{\boldsymbol{\beta}}$

The maximum Likelihood Estimation $\hat{\beta}$ is

$$\hat{\beta} \sim N(\beta, \mathcal{I}^{-1}) \tag{5.17}$$

where $\mathcal{I}$ is the 'information matrix', with elements $\mathcal{I} = \mathbb{E}\big(\partial l \ /\partial \beta_j \ \ \partial l \ /\partial \beta_i\big)$.

First define vector u such that $\mu_j = \partial l \ /\partial \beta_j$. Then $\mathcal{I} = \mathbb{E}(\mathrm{uu^T})$ and u_j can be written as follows

$$u_j = \frac{\partial l}{\partial \beta_j} = \frac{1}{\Phi}\sum_{i=1}^{n} \frac{X_{ij}(\, y_i - \mu_i)}{V(\mu_i) g'(\mu_i)}$$

If we define diagonal matrices **G** and **V**, where $G_{ii} = g'(\mu_i)$ and $\mathrm{V}_{ii} = V(\mu_i)$, then this last result becomes

$$u = X^T G^{-1} V^{-1} (y - \mu)/\Phi$$

Hence,

$$\mathbb{E}(\mathrm{uu^T}) = \frac{\mathrm{X^T G^{-1} V^{-1}}\mathbb{E}[(\mathrm{Y}-\mu)(\mathrm{Y}-\mu)^{\mathrm{T}}]\mathrm{V^{-1} G^{-1} X}}{\Phi^2}$$

$$= \frac{\mathrm{X^T G^{-1} V^{-1} V V^{-1} G^{-1} X}}{\Phi}$$

$$= \mathrm{X^T W X}/\Phi$$

Since $\mathbb{E}[(\mathrm{Y}-\mu)(\mathrm{Y}-\mu)^{\mathrm{T}}] = \mathrm{V}\Phi$ and $\mathrm{W} = \mathrm{V^{-1} G^{-2}}$

So we end up with

$$\hat{\beta} \sim N(\beta, (X^T W X)^{-1} \Phi) \tag{5.18}$$

For distributions with known scale parameter, Φ, this result can be used directly to find confidence intervals for the parameters, but if the scale parameter is unknown (e.g. for the normal distribution), then it must be estimated, and intervals must be based on an appropriate t distribution.

5.2.6 Calculation of confidence interval

Let $\hat{V}_{\beta} = (X^T W X)^{-1} \hat{\phi}$, the estimated covariance matrix of $\hat{\beta}$ ($\hat{\phi}$ is known to be 1 in some cases). Let $\hat{\sigma}_{\hat{\beta}_i}$ be the square root of the ith diagonal element of $V_{\hat{\beta}}$, that is the estimated standard error of β_j. Using the standard theory for normally distributed estimators, the confidence interval for β_i can be given as below:

i. A $100(1-\alpha)\%$ CI for β_i when ϕ is known (e.g., Poisson or Binomial cases) is

$$\hat{\beta}_i \pm t_{\infty}(1-\alpha/2)\hat{\sigma}_{\hat{\beta}_i}$$

Where $t_{\infty}(1-\alpha/2)$ is the $1-\alpha/2$ critical point of a standard normal distribution.

ii. A $100(1-\alpha)\%$ CI for β_i when ϕ is unknown (e.g., Gaussian or Gamma cases) is

$$\hat{\beta}_i \pm t_{n-\dim(\beta)}(1-\alpha/2)\hat{\sigma}_{\hat{\beta}_i}$$

Where $t_k(1-\alpha/2)$ is the $1-\alpha/2$ critical point of a t_k distribution.

For the normal response and identity link case, both results are only approximate, since they are based on (5.11), which is only approximate.

5.2.7 Model selection

Likelihood ratio test

A likelihood ratio test is a statistical test for making a decision between two hypotheses based on the value of the likelihood ratio (generally denoted by Λ). In statistics the likelihood ratio test is a statistical test used to compare the fit of two models, one of which is the null model (let's say model 1 in **Table 14**) is a special case of the alternative model (say model 2 in **Table 14**). The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other. The likelihood ratio, can then be used to compute a p-value, or compared to a critical value to decide whether to reject the null model (Model 1) in favour of the alternative model (Model 2).

Each of the two competing models, the null model and the alternative model, is separately fitted to the data and the log-likelihood recorded. The test statistic (D) is twice the difference in these log-likelihoods:

$$D = -2\,ln\left(\frac{\text{likelihood for null model}}{\text{likelihood for alternative model}}\right)$$

$$= -2\ln(\text{likelihood for null model}) + 2\ln(\text{likelihood for alternative model})$$

The model with more parameters will always fit at least as well (have an equal or greater log-likelihood). Whether it fits significantly better and should thus be preferred is determined by deriving the probability or p-value of the difference D. Where the null hypothesis represents a special case of the alternative hypothesis, the probability distribution of the test statistic is approximately a chi-squared distribution with degrees of freedom equal to df2 − df1. Symbols df1 and

df2 represent the number of free parameters of models 1 and 2, the null model, and the alternative model, respectively. The test requires nested models, that is: models in which the more complex one can be transformed into the simpler model by imposing a set of constraints on the parameters.

AIC for GLM

Model selection by direct comparison of likelihoods suffers from the problem that, if redundant parameters are added to a correct model, the likelihood almost always increases (and never decreases), because the extra parameters let the model get closer to the data, even though that only means 'modelling the noise' component of the data. As in the linear model case, this problem would be alleviated if we were somehow able to choose models on the basis of their ability to fit the mean of the data, μ, rather than the data, y. In a GLM context, a reasonable approach would be to choose between models on the basis of their ability to maximize $l(\beta, \mu)$, rather than $l(\beta, y)$, but to do so we have to be able to estimate $l(\beta, \mu)$. The required estimator can be written as below: (For calculation please see section 2.1.4 of (Wood 2006).

$$\begin{aligned} \widehat{l(\beta, \mu)} &= k - \frac{1}{2\Phi}\left\|\sqrt{W}\left(z - X\hat{\beta}\right)\right\|^2 + n/2 - tr(A) \\ &\simeq l\left(\hat{\beta}; y\right) - tr(A) + n/2 \end{aligned} \tag{5.19}$$

where $A = X(X^T W X)^{-1} X^T W$ and hence tr (A) = p, the number of (identifiable) model parameters.

Hence, when choosing between models, we would choose whichever model had the highest value of $l\left(\hat{\beta}\right) - p$, which is equivalent to choosing the model with the lowest value of Akaike's Information Criterion (Akaike 1973),

$$AIC = 2\left[-l\left(\hat{\beta}\right) + p\right] \tag{5.20}$$

The foregoing argument assumes that Φ is known. If it is not then an estimate, $\widehat{\Phi}$, will be needed in order to evaluate the AIC, and as a result the penalty term p in the AIC will become $p + 1$.

BIC for GLM

In statistics, the Bayesian information criterion (BIC) or Schwarz criterion is a criterion for model selection among a finite set of models. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC). When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in over fitting. Both BIC and AIC resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC.

The formula for the BIC is:

$$-2ln\,\hat{\mathrm{L}} + \mathrm{k}\ln(n)$$

Where k is the number of parameters to be estimated, $\hat{\mathrm{L}}$ is the maximized value of the likelihood function of the model. The BIC works under the assumption that the model errors or disturbances are independently and identically distributed according to a normal distribution and the derivative of the log likelihood with respect to the true variance is zero.

Given any two estimated models, the model with lower value of BIC is the one to be preferred. Unexplained variation in the dependent variable and the number of explanatory variables increases the value of BIC. Hence, lower BIC implies either fewer explanatory variables, better fit, or both. The BIC generally

penalizes free parameters more strongly than does the Akaike information criterion, though it depends on the size of n and relative magnitude of n and k. It is important to keep in mind that the BIC can be used to compare estimated models only when the numerical values of the dependent variable are identical for all estimates being compared. The models being compared need not be nested, unlike the case when models are being compared using an F or the likelihood ratio test.

5.2.8 Model comparison

By hypothesis testing

For GLM consider testing

$$H_0: g(\mu) = X_0\beta_0$$

against

$$H_1: g(\mu) = X_1\beta_1$$

Let $l(\hat{\beta}_0)$ and $l(\hat{\beta}_1)$ be the maximized log-likelihoods of the two models. If H_0 is true then in the large sample limit,

$$2\left[l(\hat{\beta}_1) - l(\hat{\beta}_0)\right] \sim \chi^2_{p_1-p_0}, \tag{5.21}$$

where p_i is the number of (identifiable) parameters ($\hat{\beta}_i$) in model i. If the null hypothesis is false, then model 1 will tend to have a substantially higher likelihood than model 0, so that twice the difference in log likelihoods would be too large for consistency with the relevant χ^2 distribution.

The approximate result (5.22) is only directly useful if the log likelihoods of the models concerned can be calculated. In the case of GLMs estimated by

Iteratively Re-weighted Least Squares (IRLS), this is only the case if the scale parameter, Φ is known. Hence the result can be used directly with Poisson and binomial models, but not with the normal (for the same normal distribution and identity link), gamma, or inverse Gaussian distributions, where the scale parameter is not known.

By deviance

In GLM practically it is useful to have a quantity (interpreted like residual sum of squares in ordinary linear modelling). This is called deviance and is defined as

$$D = 2\left[l\left(\hat{\beta}_{max}\right) - l\left(\hat{\beta}\right)\right]\Phi$$
$$= \sum_{i=1}^{n} 2\omega_i\left[\, y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right) + b\left(\hat{\theta}_i\right)\right], \qquad (5.22)$$

where $l\left(\hat{\beta}_{max}\right)$ indicates the maximized log-likelihood of the saturated model (model with one parameter per data point). $l\left(\hat{\beta}_{max}\right)$ is the highest value that the log- likelihood could possibly have, given the data, and is evaluated by simply setting $\hat{\mu} = y$ and evaluating the log-likelihood. $\tilde{\theta}$ and $\hat{\theta}$ denote the maximum likelihood estimates of canonical parameters, for the saturated model and model of interest, respectively.

The scaled deviance does not depend on scale parameter and defined as,

$$D^* = D/\Phi, \qquad (5.23)$$

For Binomial and Poisson distributions ($\Phi = 1$), the deviance and scaled deviance are the same, but this is not the case more generally. By the generalized likelihood ratio test result (5.21), we might expect that, if the model is correct, then approximately

$$D^* = \chi^2_{n-p}, \qquad (5.24)$$

in the large sample limit. Given the definition of deviance, it is easy to see that the log likelihood ratio statistic in (5.21) can be re-expressed as $D_0^* - D_1^*$. So under H_0

$$D_0^* - D_1^* \sim \chi^2_{p_1 - p_0} \tag{5.25}$$

(in the large sample limit), where D_i^* is the deviance of the model i which has p_i identifiable parameters. But again, this is only useful if the scale parameter is known so that D^* can be calculated.

Model comparison with unknown ϕ

Under H_0, we have the approximate results

$$D_0^* - D_1^* \sim \chi^2_{p_1 - p_0} \text{ and } D_1^* \sim \chi^2_{n-p}$$

And if $D_0^* - D_1^*$ and D_1^* are treated as asymptotically independent, this implies that

$$F = \frac{(D_0^* - D_1^*)/(p_1 - p_0)}{D_1^*/(n - p_1)} \sim F_{p_1 - p_0,\ n - p_1}, \tag{5.26}$$

in the large sample limit. The useful property of F is that it can be calculated without knowing Φ, which can be cancelled from the top and bottom of the ratio yielding, under H_0, the approximate result that

$$F = \frac{(D_0 - D_1)/(p_1 - p_0)}{D_1/(n - p_1)} \sim F_{p_1 - p_0,\ n - p_1}, \tag{5.27}$$

The advantage of this result is that it can be used for hypothesis testing based model comparison, when Φ is unknown. The disadvantages are the dubious distributional assumption for D_1^*, and the independence approximation, on which it is based.

5.2.9 $\widehat{\phi}$ and Pearson's statistic

As we have seen, the MLEs of the parameters β can be obtained without knowing the scale parameter, Φ but, in those cases in which this parameter is unknown, it must usually be estimated. The approximate result (5.24) provides one obvious estimator. The expected value of a χ^2_{n-p} random variable is $n-p$, so equating the observed $D_0^* = D/\Phi$ to its approximate expected value and re-arranging, we get

$$\widehat{\Phi}_D = \widehat{D}/(n-p) \tag{5.28}$$

The Pearson statistic is defined as

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \tag{5.29}$$

Clearly X^2/Φ would be the sum of squares of a set of zero mean, unit variance, random variables, having $n-p$ degrees of freedom, suggesting that if the model is adequate then approximately $X^2/\Phi \sim \chi^2_{n-p}$. Setting the observed Pearson statistic to its expected value, and re-arranging, yields

$$\widehat{\Phi} = X^2/(n-p) \tag{5.30}$$

It is straightforward to show that

$$X^2 = \sum_{i=1}^{i=n} \omega_i \left(z_i - X_i\hat{\beta}_i\right)^2 \tag{5.31}$$

where ω_i and z_i are IRLS weights and pseudo data, evaluated at convergence.

5.2.10 Residuals and model checking

It is always necessary to check that the model meets its assumptions well enough that the results are likely to be valid, before using the distributional results for

inference. For ordinary linear models, the model checking is based on the examination of the residuals that contain all the information of data and not explained by the systematic part of the model.

For GLM, examination of residuals is also crucial but challenging because we need to standardise the residuals. The main reason for not simply examining the raw residuals ($\hat{\varepsilon}_i = y_i - \hat{\mu}_i$) is the difficulty of checking the validity of the assumed mean variance relationship from the raw residuals. For example, in Poisson model the variance of the residuals should increase in direct proportion to the size of the fitted values ($\hat{\mu}_i$). However, from the raw residuals plotted against fitted values, we can judge whether the residual variability is increasing in proportion to the mean than the square root or the square of the mean (for example). For this reason, GLM residuals are usually standardised so that, if the model assumptions are correct, the standardised residuals should have approximately equal variance, and behave (as far as possible) like residuals from an ordinary linear model.

Pearson residuals:

The Pearson residuals are calculated by dividing the raw residuals by a quantity proportional to their standard deviation from the fitted model. It is defined as

$$\hat{\epsilon}_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{(\hat{\mu}_i)}}, \tag{5.32}$$

This should have approximately zero mean and variance σ^2, if the model is correct. If plotted against the fitted values, or any covariates (whether to include in the model or not), these residuals should not display any trend in mean or

variance. This is called `Pearson residuals' because of the fact that the sum of squares of the Pearson residuals gives the Pearson statistic.

Deviance residuals

The distribution of the Pearson residuals can be quite asymmetric around zero in practice. The *deviance residuals* are often preferable in this respect. The deviance in the *deviance residuals* plays much the same role for GLMs that the residual sum of squares plays for ordinary linear models: indeed for an ordinary linear model the deviance is the residual sum of squares. In the ordinary linear model case, the deviance is calculated from the sum of the squared residuals. That is the residuals are the square roots of the components of the deviance with the appropriate sign attached. So, if d_i indicates the component of the deviance contributed by the $i\ th$ datum, we have

$$D = \sum_{i=1}^{n} d_i \tag{5.33}$$

and from the concept of the ordinary linear model, we can define

$$\hat{\epsilon}_i^d = sign(y_i - \hat{\mu}_i)\sqrt{d_i}. \tag{5.34}$$

The sum of squares of these *deviance residuals* gives the deviance. Now if the deviance were calculated for a model where all the parameters were known, then (5.24) would become $D^* \sim \chi_n^2$, and this might suggest that for a single datum $d_i/\sigma^2 \sim \chi_1^2$, implying that $\epsilon_i^d \sim N(0, \sigma^2)$. Thus from the equation (5.24), we might expect the deviance residuals to behave something like $N(0, \sigma^2)$ random variables, for a well-fitting model, especially in cases for which (5.24) is expected to be a reasonable approximation.

Residual plots

We can use various residual plots by using standardised residuals to find the evidence that the model assumptions are not met. The main useful plots are:

- Standardised residuals against fitted values. A trend in the mean of the residuals violates the independence assumption and often implies that something is wrong with the model from the mean of the response (e.g., perhaps a missing dependence, or the wrong link function). A trend in the variability of the residuals is diagnostic of a problem with the assumed mean variance relationship, i.e. with the assumed response distribution.
- Standardised residuals against all potential predictor variables (selected or omitted from the model). Trends in the mean of the residuals can be very useful for pinpointing missing dependencies of the mean response on predictors.
- Normal QQ plots can be useful for highlighting problems with the distributional assumptions, in cases where the response distribution can be well approximated by a normal distribution (with appropriate non-constant variance). For example Poisson residuals for a response to a fairly high mean fall into this category.
- Plots of standardised residuals against leverage (influential observations) are useful for highlighting single points that have a very high influence on the model fitting. Leverage is a measure of how influential a data point could be, based on the distance of its predictor variables from the predictors of other data.

All plots are useful for spotting potential outliers (points which do not fit well with the pattern of the rest of the data) and deserve special attention. They also check whether the model is erroneous, or the model is not expressing something important about the system that the data relate to.

5.2.11 Quasi-Likelihood

We observed that in GLM the distribution of the response variable follows any distribution from the exponential family and therefore it is better to base models on any particular distribution if there are good reasons to suppose that the response follows that distribution. But in many cases the nature of the response distribution is not known very precisely and it is only possible to specify what the relationship between the variance of the response ($V(\mu_i)$) and its mean should be. The question is whether it is possible to develop GLMs theory for fitting and inference, starting from the position of specifying only the mean-variance relationship.

The concept of **Quasi-likelihood** approach is adequate in such situation. For an observation y_i, of a random variable with mean (μ_i), and known variance $V(\mu_i)$, the log quasi likelihood for μ_i given y_i is defined as:

$$q_i(\mu_i) = \int_{y_i}^{\mu_i} \frac{y_i - z}{\phi V(z)} dz \qquad (5.35)$$

The key feature of this function is that it shares many important properties of the log likelihood l_i, corresponding to a single observation, but only requires knowledge of the variance (V) rather than the full distribution of Y_i. The log quasi likelihood for the mean vector, $\boldsymbol{\mu}$, of all the response data, or any parameter vector

defining $\boldsymbol{\mu}$ (assuming that the data are observations of independent random variables) can be given as:

$$q(\boldsymbol{\mu}) = \sum_{i=1}^{n} q_i(\mu_i) \tag{5.36}$$

The key properties of q is that, for the purpose of the inference of GLMs, it behaves in a very similar manner to the log likelihood, but only requires the knowledge of the variance function to define it. Note that the quasi-likelihood of the saturated model is zero, so the quasi deviance of a GLM is simply

$$D_q = -2q\ (\hat{\mu})\phi \tag{5.37}$$

The calculation of residuals and scale parameter also carries over from the likelihood to the quasi-likelihood without having any change than the replacement of l by q.

The practical use of the quasi-likelihood approach requires that the integral in (5.35) be evaluated, and it is possible for most practical useful forms of V (MacCullagh and Nelder 1989). One of the most common this approach is to provide the means of modelling count data that are more variable than the Poisson or binomial distributions (with their fixed scale parameters). Such 'over-dispersed' data are very common in the environmental and health setting and it is called **Over-dispersion.** Another practical use is for modelling data with a mean variance relationship for which there is no obvious exponential distribution: for example continuous data where variance is expected to be proportional to mean.

5.2.12 QAIC and QBIC

In modelling ecological data, over dispersion is quite common and needs to be included in the model selection procedure. In a typical Generalized Poisson model, the Quasi-AIC or QAIC can be defined as,

$$QAIC = \frac{-2\,L}{\hat{c}} + 2k$$

And the corresponding bias corrected version can be given as,

$$QAIC_c = \frac{-2L}{\hat{c}} + 2k + \frac{2k(k+1)}{n^* - k - 1}$$

$$= QAIC + \frac{2k(k+1)}{n^* - k - 1}$$

Here, L is the log likelihood, n^* is the total number of counts (since Poisson case) or effective sample size, $\hat{c}$ is the parameter for quasi-likelihood or multiplicative factor that represents extra variability due to over dispersion, k is the total number of parameters in the model, which also include α. Typically, $\hat{c}$ take the value 1 which indicates that there is no over dispersion. An estimator of $\hat{c}$ is the deviance divided by its degrees of freedom.

Under the same notation the QBIC can be defined as,

$$QBIC = \frac{-2\,L}{\hat{c}} + k\log(n^*)$$

QAIC and QBIC both have the same interpretations like AIC and BIC. Smaller values indicate better model fit.

5.3 Models with count data

Modelling disease count as response variable is a common task with most data in the environmental, health, and social settings. For this reason, regression models

with count data are also common in climate health research settings. Regression modelling with count data is mainly described under the context of Poisson regression.

A Poisson regression model is a special case of the Generalized Linear model (determined in equation (5.1)). In such case, we consider GLM as regression models for the mean only (as specified by (5.1)) instead of viewing them as models for the full likelihood. However, the classical Poisson regression model for count data is often limited due to over-dispersion and/or an excess number of zeros in the data sets. The quasi Poisson model, the negative binomial (NB) models have been developed to deal with over-dispersion. The Hurdle model and the Zero inflated Poisson can deal with the situation with excess zeros. However, all these models still belong to the GLMs family.

5.3.1 Poisson model

The simplest distribution used for modelling count data is the Poisson distribution. A random variable Y is said to have a Poisson distribution with parameter λ with integer values y= 0, 1, 2, 3, ….. with probability

$$Pr\{Y = y\} = \frac{e^{-\mu}\mu^{y}}{y!} \quad for\, \mu > 0 \tag{5.38}$$

This is a special case of the GLM framework for the count data. The canonical link is $g(\mu) = \log(\mu)$ resulting in a log-linear relationship between mean and linear predictor.

5.3.2 Dealing with over-dispersion

Quasi-Poisson model

One way of dealing with over-dispersion is to use the mean regression function and the variance function from the Poisson GLM but to leave the dispersion parameter ϕ unrestricted. Thus, ϕ is not assumed to be fixed at 1 but is estimated from the data. This strategy leads to the same coefficient estimates as the standard Poisson model but the inference is adjusted for over-dispersion. Consequently, quasi-Poisson models adopt the estimating function view of the Poisson model and do not correspond to models with fully specified likelihoods.

Negative binomial model

A second way of modelling over-dispersed count data is to assume a negative binomial (NB) distribution for $y_i|x_i$ which can arise as a gamma mixture of Poisson distributions. One parameterization of its probability density function is

$$f(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\cdot y!}\cdot\frac{\mu^y\cdot\theta^\theta}{(\mu+\theta)^{y+\theta}}$$

with mean μ and shape parameter θ; $\Gamma(\cdot)$ is the gamma function. For every fixed θ, this is of type (5.2) and thus another special case of GLM framework. It also has $\phi = 1$ but the variance function $V(\mu) = \mu + \frac{\mu^2}{\theta}$.

5.3.3 Dealing with excess zeros

Hurdle model

In addition to the over - dispersion, many empirical count data sets exhibit more zero observations than would be allowed for by the Poisson model. One model class capable of capturing both properties is the Hurdle model. They are two-component models: A truncated count component, such as Poisson, geometric or negative binomial, is employed for positive counts, and a hurdle component model zeroes versus larger counts. In the latter, either a binomial model or a censored count distribution can be employed. More formally, the Hurdle model combines a count data model $f_{\text{count}}(y; x, \beta)$ (that is left truncated at $y = 1$) and zero hurdle model $f_{\text{zero}}(0; z, \gamma)$ (right censored at $y = 1$). Hence the Hurdle model density can be expressed as

$$f_{hurdle}(y; x, z, \beta, \gamma) = \begin{cases} f_{\text{zero}}(0; z, \gamma), & \text{if y} = 0 \\ \left(1 - \left(f_{\text{zero}}(0; z, \gamma)\right)\right) \cdot f_{\text{count}}(y; x, \beta) / \left(1 - \left(f_{\text{count}}(y; x, \beta)\right)\right), & \text{if y} > 0 \end{cases} \tag{5.39}$$

The model parameters β, γ and potentially one or two additional dispersion parameters θ (if f_{count} or f_{Zero} or both are negative binomial densities) are estimated by ML, where the specification of the likelihood has the advantage that the count and the hurdle components can be maximized separately.

Zero-inflated model

Zero-inflated models are another model class capable of dealing with excess zero counts. They are two-component mixture models combining a point mass at zero with a count distribution such as Poisson, geometric or negative binomial. Thus, there are two sources of zeros: zeros may come from both the point mass and from the count component. For modelling the unobserved state (zero versus. count), a binary model is used: in the simplest case only with an intercept but potentially containing regressors.

Thus the, zero-inflated density $f_{zeroinfl}(y)$ is a mixture of a point mass at zero $I_{\{0\}}(y)$ and a count distribution $f_{count}(y; x, \beta)$. The probability of observing a zero count is inflated with probability $\pi = f_{zero}(0; z, \gamma)$, i.e.

$$f_{zeroinfl}(y; x, z, \beta, \gamma) = f_{zero}(0; z, \gamma) \,.\, I_0(y) + \left(1 - f_{zero}(0; z, \gamma)\right) \,.\, f_{count}(y; x, \beta) \tag{5.40}$$

Where $I(.)$ is the indicator function and the unobserved probability π of belonging to the point mass component is modelled by a binomial generalized linear model (GLM) $\pi = g^{-1}(z^T\gamma)$. And the regression equation for the mean is

$$\mu_i = \pi_i \,.0 + (1 - \pi_i) \,.\, exp(x_i^T \beta) \tag{5.41}$$

using canonical log link. Here, z_i is the vector of regressors in the zero-inflation model and x_i are the regressors in the count component. The full set of parameters β, γ can estimate by using maximum likelihood method.

5.4 Other useful modelling approaches

There is also some other modelling approaches that have been used often in environmental and health research settings such climate health research. For example, Generalized additive model, Time series modelling, Spatio-temporal Modelling Approach, Geospatial method, Case-crossover study approaches are some important approaches among them

5.5 GLM results using temperature

We describe here some results from preliminary stage using only the temperature. The results are summarised in the following three sub-sections.

5.5.1 Temperature variations with COPD

An approach to exploring the effect of weather variations on chronic disease incidence rate and potential changes in future health systems (Islam, Chaussalet et al. 2010). (Please see the reference for details).

Many COPD sufferers have their symptoms deteriorate during colder weather; this often leads to an increase in hospital admissions and capacity shortages. In this section, we explore the association between COPD incidence rates and monthly maximum, minimum, mean temperature, and monthly total ruin by using data for April 1997 to March 2004 for the region: England North (the data sets used in the thesis were not acquired during that time). We develop a statistical model (zero-inflated Poisson regression model) to measure the significance of

meteorological variables on COPD admission counts (ICD-10, J40-J44, and J47). Zero-inflated Poisson distribution is useful if the data shows over dispersion or have a higher incidence of zero counts than is expected for the Poisson distribution. Another way of dealing the same situation is to use Zero-inflated negative binomial model.

Three datasets have been used, namely the national Hospital Episodes Statistics (HES) data set, the observational data (monthly maximum, minimum, mean temperature and rain) from the Met Office and mid-year population for a number of years from the Office for National Statistics, UK . We also collected the mid-year population for a number of years from the Office for National Statistics, UK. All these data sets were from April 1997 to March 2004 for the region: England North.

We calculated the person-days of follow up for COPD and COPD incidence rate (per 100 person-days) (O'Loughlin, Robitaille et al. 1993). The percentage of COPD admissions for each month was calculated by using mid-year population for respective years. We found January and February had the highest COPD incidence rate (**Table 8**). We plotted the trends of COPD incidence rate through the trends of maximum temp, mean temp, minimum temp. For each month, we also calculated the correlation between temperature (maximum, mean, and minimum), rain along with the test results for the significance of their correlation and put them in a correlation matrix. For example, the correlation matrix in **Table 9** and **Figure 5-a** shows that the temperature is moderately positively correlated and the rain is moderately negatively correlated to COPD incidence rates. However, none of the correlations is statistically significant.

Table 8: Mean monthly incidence rates

Months	Incidence rate
January	7.15
February	7.51
March	6.75
April	6.96
May	6.68
June	7.0
July	6.65
August	6.93
September	6.65
October	6.58
November	6.81
December	5.97

Table 9: Correlation matrix for July

	COPD Inc rate	Max Temp.	Min Temp.	Mean Temp.	Rain
COPD Inc rate	1.00	0.82	0.68	0.80	-0.69
Max Temp.	0.82	1.00	0.86	0.98	-0.54
Min Temp.	0.68	0.86	1.00	0.95	-0.46
Mean Temp.	0.80	0.98	0.94	1.00	-0.54
Rain	-0.69	-0.54	-0.46	-0.54	1.00
P-Value		0.02	0.09	0.03	0.08

For model fitting we select a random sample of 5000 inpatient COPD admissions from the HES dataset, (England North only) for the year 2003-04. The mean and variance of COPD admission counts are 0.023 and 0.041, respectively. From the histogram of COPD admission counts (**Figure 5**-b), we notice a huge proportion of zeros, and as a result, we used the Zero-Inflated Poisson regression model (section 5.3.3).

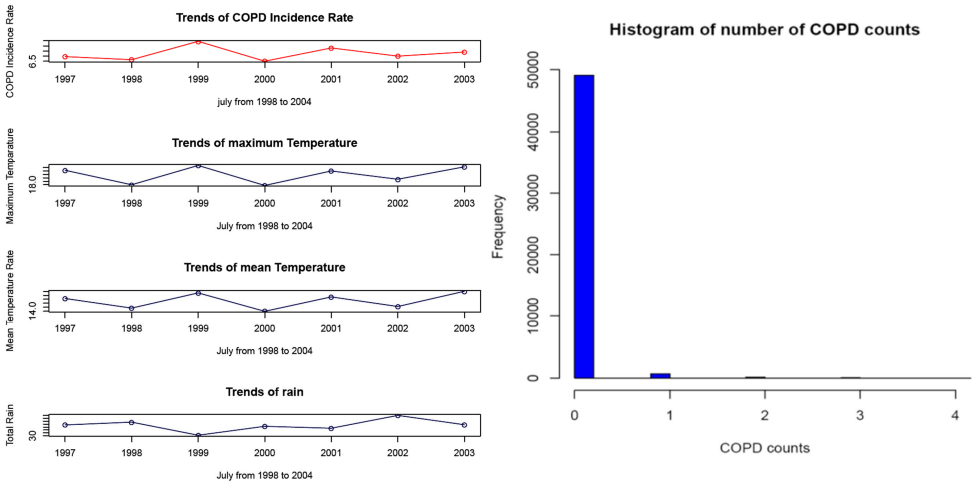

Figure 5: a) Trends of COPD incidence rate, maximum temperature, mean temperature, and total rain for July; b) Histogram of COPD counts

Table 10: Model fitting results

	Estimate	**Std. Error**	**Z value**	**Pr(>|z|)**
Intercept	0.296	0.575	0.515	0.607
Max. Temp.	0.719	1.08	0.663	0.507
Min. Temp.	0.859	1.079	0.797	0.426
Mean Temp.	-1.574	2.156	-0.73	0.465
Rain	-0.005	0.003	-1.71	0.087

The use of Zero-inflated Poisson regression model actually improved the model fit. We perform the Vuong test (test statistics = -10.22 and p-value < 0.0000), which suggests that the zero-inflated model has a significant improvement over Poisson model. However none of the predictor variables are found to be statistically significant for the COPD admissions count with respect to maximum temperature, minimum temperature, mean temperature and total rain. This could be due to the measurement of the data level (considered monthly rather than days), the crudeness of the climate data, also not considering the non-linear behaviour of the climate-disease relationships.

5.5.2 Temperature disparity with COPD readmissions

The impact of temperature disparity on emergency readmissions and patient flows (Islam, Chaussalet et al. 2011). (Please see the reference for details).

Here we explored the impact of temperature variations on COPD hospital readmissions by developing a Frailty model. The time is measured as the "number of days" (difference between previous discharge date and current admission date) and the corresponding event as COPD readmission. We investigated whether there is any relationship of such rehospitalisation time for COPD due to the variability in daily temperature (maximum, minimum, mean) adjusted for gender and age. To highlight the regional heterogeneity among the time of COPD readmissions, we included a random effect term (frailty) in the Cox Proportional Hazard model and fit the frailty model. Here, the Cox-proportional part is quantifying the significance of the explanatory variables (age, gender, various lags of daily temperatures) and frailty term is measuring the regional (Spatial) heterogeneity in this process.

We used two datasets, namely HES (for the COPD hospital episode; ICD-10 codes J40-J44) and temperature (maximum, minimum and mean) from the Met Office. The data were collected for 25 local authorities (seven are from London, six from Cumbria, five from Somerset, and seven from West Sussex) for the financial year of 1997 to 2003. We also calculated lags of temperatures (from lag 1 to lag 5) and 5 days moving averages and exponential moving averages starting from the day of admission to see the effect of various lag values.

The initial number of COPD admission in the selected 25 local authority areas for the period of study was 39980. We cleansed the data for the admissions where discharge date was 'NULL' or episode was not the last of the spell or admissions with unfinished episode or a discharge which indicates that the patient is still in the hospital or discharge date is not available or babies with less than one year. All these above-mentioned events are not mutually exclusive and all together they covered 7458 cases. We selected the patients with more than one admission during the period for model fitting. We had 20496 admissions of this type. To calculate the COPD readmission cases, we subtract the discharge date of a COPD admission from the corresponding next admission date for any specific patient.

The hazard function in the Gamma shared frailty model depends on an unobservable random variable (frailty) which acts multiplicatively on the hazard. The univariate and multivariate Cox proportional hazard model and shared gamma-frailty model used to model the readmission time for each of the COPD patients adjusted for temperature (maximum, minimum, mean), age and gender.

We calculated the hazard ratios (HR) and 95% confidence intervals for each covariate. Using the Wald statistic (0.00121/0.001=1.21), we found that the frailty indicating heterogeneity among the selected counties and/or Boroughs is not significant. Interesting to see that all the variables were significant in the non-adjusted model but only age, gender, and exponential moving average of maximum and mean temperature are significant adjusting for all the variables considered.

The frailty parameter, describing the heterogeneity of selected Boroughs and counties is statistically non-significant, suggesting that there is no variability in terms of risk of readmission among selected counties. The hazard function for the readmission of COPD patients is illustrated with a 'bathtub shape' (**Figure 6-a**). Patients that are readmitted on the same day of discharge have the highest risk of readmission, where the risk gradually decreases up to 100 days. We notice a stable risk of readmission for patients readmitted between 100 to 820 days after discharge and dramatically increase afterwards. From **Figure 6-b**, we can see that men are slightly more susceptible for COPD readmission compared to women.

Table 11: Hazard ratio of readmissions for selected variables

Covariate(s)	Cox model		Shared Gamma Frailty models	
	Univariate model	**Multivariate model**	**Univariate model**	**Multivariate model**
	HR(CI), *p*	HR(CI), *p*	HR(CI), *p*	HR(CI), *p*
Start Age	0.99 (0.99-1.00) <0000*	0.99 (0.99-1.00), <0000*	1(0.99-1.00), <0000*	1(0.99-1.00), <0000*
Sex	0.91(0.88-0.94), <0000*	0.91(0.88-0.94), <0000*	0.91(0.88-0.94), <0000*	0.91(0.88-0.94), <0000*
Maximum Temp	1.01 (1.00- 1.01), <.0000*	1.27(0.92-1.76), .15	1.01 (1.00- 1.01) , <.0000*	1.27(0.93-1.73), .13
M. A. of Max Temp	1.01 (1.01 - 1.01),<.0000*	1.12(0.87-1.43), .37	1.01 (1.00 - 1.01) <.0000*	1.11(0.85-1.45), .45
Exp. Mov. Max Temp.	1.01 (1.00- 1.01), <.0000*	0.69(0.48-1.00), .04*	1.01 (1.00- 1.01), <.0000*	0.69(0.50-0.96), .03*
Minimum Temp	1.01 (1.00- 1.01), <.0000*	1.26(0.91-1.74), .17	1.01 (1.00- 1.01), <.0000*	1.26(0.92-1.72), .15
M. A. of Min Temp	1.01 (1.00- 1.01), <.0000*	1.08(0.84-1.38),.57	1.01 (1.00- 1.01), <.0000*	1.06(0.82-1.39), .64
Exp. Mov. Min Temp.	1.01 (1.00- 1.01), <.0000*	0.74(0.51-1.06), .1	1.01 (1.00- 1.01), <.0000*	0.74(0.53-1.03), .07
Mean Temp	1.01 (1.00- 1.01), <.0000*	0.63(0.33-1.20), .16	1.01 (1.00- 1.01), <.0000*	0.63(0.34-1.16), .14
M. A. of Mean Temp	1.01 (1.00- 1.01), <.0000*	0.81(0.50-1.33), .41	1.01 (1.00- 1.01), <.0000*	0.83(0.49-1.41), .5
Exp. Mov. Mean Temp.	1.01 (1.00- 1.01), <.0000*	2.02(0.98-4.17), .05*	1.01 (1.00- 1.01), <.0000*	2.02(1.05-3.88),.04*
Frailty $\boldsymbol{\theta}$, (S. E. $\boldsymbol{\theta}$)				0.00121(0. 001)

* =Significant

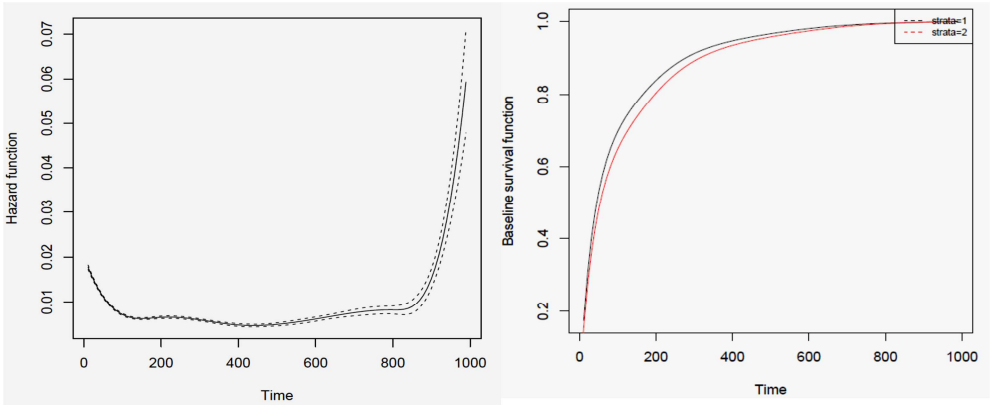

Figure 6: From Left: a) Baseline hazard function for readmission of COPD with 95%; b) Probability of readmission according to sex (strata 2 = female and strata 1 = male)

This paper showed us the evidence of the importance of considering the effect of the lag period in the climate research health study. We also knew that changes in the readmission due to temperature are not significant because of small changes in the areas and COPD readmission is more significant in men than women.

5.5.3 Temperature variations with asthma admissions

Exploring the effect of temperature variations on unplanned hospital admissions for asthma (Islam, Chaussalet et al. 2011). (Please see the reference for details).

Asthma is one of the most effected disease outcome due to climate change and air pollutions (AsthmaUK 2013). The objective of the study is to explore the relationship of temperature varies with the admissions of asthma based on 25 local

authorities (seven are from London, six from Cumbria, five from Somerset, and seven from West Sussex) for the year 1998-2003.

Similar dataset was utilised as in case study 2 except for asthma related admissions. We calculated lags of temperatures (from lag 1 to lag 5) and 5 days moving averages and exponential moving averages starting from the day of admission to see the effect of various lag values. We also considered the lag values for each of these temperatures (e.g., lag 1, lag 2, and lag 5) and calculate 5 days moving and exponential moving averages from the day of admissions.

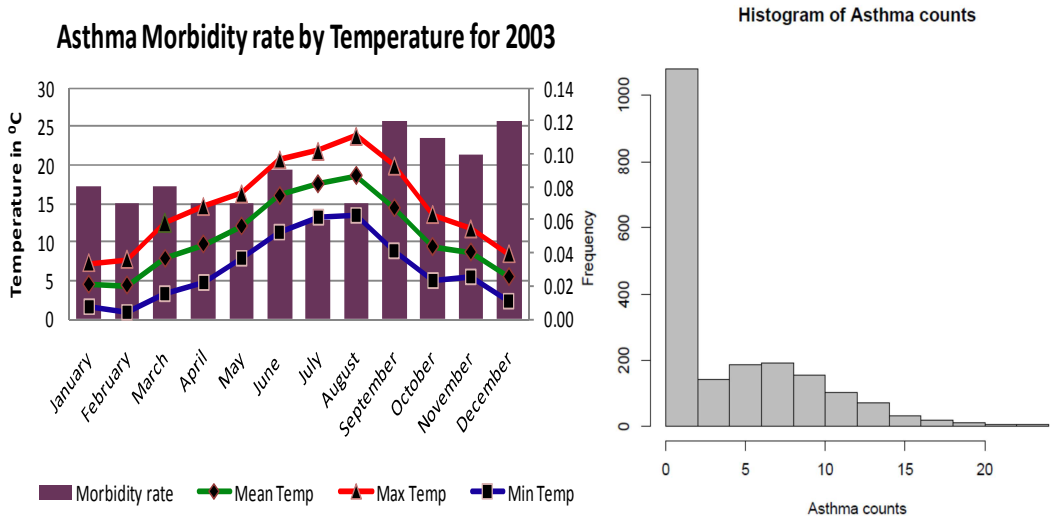

Figure 7: (From left) a) Trends of asthma morbidity rate, mean, maximum, and minimum temperature for 2003, b) Frequency of asthma admission counts

We standardised the morbidity rate of unplanned admissions for selected disease (e.g., asthma) for the whole region (all 25 local authorities) by adopting the respective population estimates (mid-year population) for each of the years (1998 -2003). For each of the years, we explored the trends of the calculated monthly morbidity rate with temperatures to find whether there are any temperature trends and asthma counts. The Poisson regression model and Zero-

Inflated Poisson regression model has been used to highlight whether the relationship between temperature variations and the hospital admission counts are significant for asthma (See section 5.3). We also include the above mentioned temperature lags and moving & exponential to examine the significance on asthma unplanned admissions counts.

Table 12: Zero-inflation model coefficients (binomial with logit link)

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	0.40257	0.20041	2.009	0.0446 **
Max. Temp	2.09332	0.89672	2.334	0.0196 **
Max Temp lag1	-0.02608	0.89424	-0.029	0.9767
Max Temp lag2	1.72513	0.89346	1.931	0.0535 *
Max Temp lag5	-0.35056	0.89043	-0.394	0.6938
Max Temp Mov. Avg	-1.22064	0.65805	-1.855	0.0636 *
Max Temp Exp. Avg	0.51968	0.70869	0.733	0.4634
Min Temp	2.10971	0.89136	2.367	0.0179 **
Min Temp lag1	-0.03758	0.8901	-0.042	0.9663
Min Temp lag2	1.75986	0.89507	1.966	0.0493 **
Min Temp lag5	-0.37711	0.89075	-0.423	0.672
Min Temp Mov. Avg	-1.03175	0.65628	-1.572	0.1159
Min Temp Exp. Avg	0.34071	0.71473	0.477	0.6336
Mean Temp	-4.18968	1.77874	-2.355	0.0185 **
Mean Temp lag1	0.12809	1.77985	0.072	0.9426
Mean Temp lag2	-3.50336	1.7876	-1.96	0.0500 *
Mean Temp lag5	0.72919	1.77873	0.41	0.6818
Mean Temp Mov. Avg	2.29317	1.2831	1.787	0.0739 *
Mean Temp Exp. Avg	-0.97565	1.29276	-0.755	0.4504

** = 0.05, *=0.1

From exploratory data analysis we found clear tendencies to increase the trends of asthma morbidity rate with lower temperatures and vice versa (e.g., **Figure 7**). In terms of months the morbidity rate is higher towards the end of

autumn and the start of winter. Thus it is showing relationships in trends between the temperatures (monthly mean, maximum and minimum) and morbidity rate for asthma

From the results of the Zero-inflated Poisson regression model (**Table 12**) we found that maximum temperature, minimum temperature and mean temperature on the day of admissions are significantly affecting number of unplanned asthma admissions at 5% level of significance. Same results revealed for minimum temperature of 2 days lag. From the result of likelihood ratio test (chi-squared value 56.49), we found that the overall model is significant (p-value < 0.05). We also performed the Vuong test (test statistics = -38.7 and p-value < 0), which suggests that the zero-inflated model has a significant improvement over Poisson model.

In summary, asthma is more significant to lower temperature or during winter and there is some lag effect on asthma hospital counts due to temperature variations. Such results also remind us the importance of lags in climate health research.

5.6 GLM results using climate and pollution factors

In this section we started by some exploratory data analysis to describe the relationships of all the climate and pollution factors followed by (in subsections) developing a series of Generalized Linear Models and then select best model and significant climate and pollutions factors.

5.6.1 Relationships of the factors

Numerical and graphical data analysis was carried out to summarise the main characteristics of the dataset. The objective was to ascertain the distribution of the variables, relationships between the variables, and their trends over time. Furthermore, the analysis could also assist us in deciding the variables for inclusion, smoothing the unusual trends, corresponding lag structure, and lag period.

Scatterplot matrixes with the climate and air pollutants versus daily count of lower respiratory diseases are plotted. In **Figure 8**, the scatter plot matrix shows the nature of the relationships of meteorological variables and pollutants with lower respiratory (LR) disease counts. One observes (**Figure 8**) that LR disease count with climate variables and pollutants are non-linear. It is also evident that some of the explanatory variables such as Sun hours and Radiation are linearly correlated which may suggest the existence of multicollinearity, and as a result it may be possible that one or more of the variables (e.g., sun hours, radiation) may become redundant in the modelling phase.

The second scatter plot matrix (**Figure 9**) explains the relationships among the variables more elaborately by showing the statistical distribution of the variables and trends between the pairs of variables. **Figure 9** also shows histograms, kernel density overlays, absolute correlations, and p-values, i.e. asterisks (0.05, 0.01, and 0.001). For example, we can see that the correlations between a pair of variables: Sun hours versus Radiation ($r = 0.82$), Radiation versus Humidity ($r = 0.78$), and Sun Hours versus Humidity ($r = 0.73$) are very high and significant at the 1 % level of significance. From the histogram, we see

that almost all the variables are skewed or non-normal which is also an indication of the possibility of the over-disperse nature of the data. The kernel density overlays (e.g., 1st column from left) for lower respiratory hospital admission counts show non-linearity with all climate variables and pollutants (**Figure 9**). From both scatter plot matrices, it is evident that the lower-respiratory hospital admissions counts do not vary by seasons (**Figure 9** and **Table 13**).

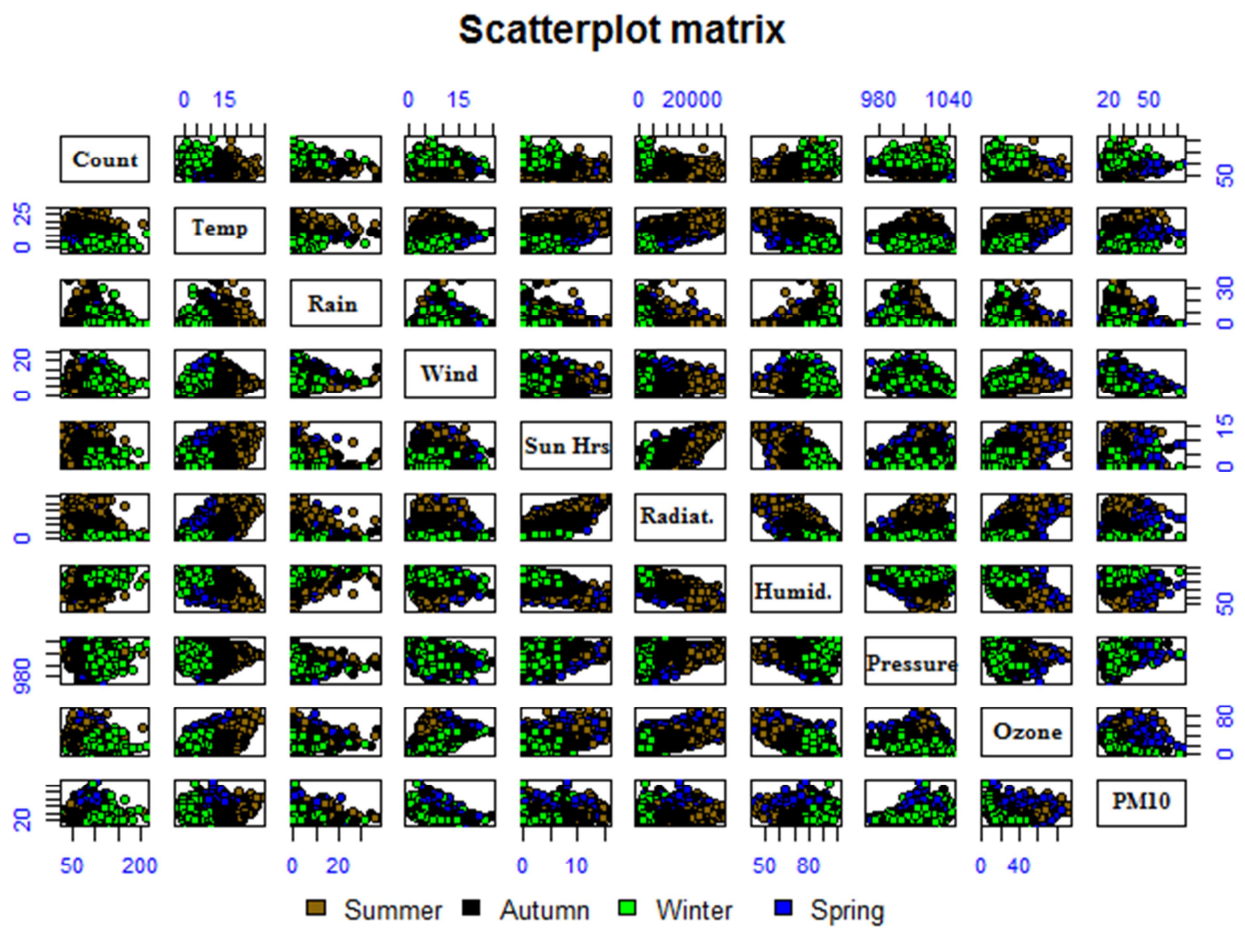

Figure 8: Scatter plot matrix of the disease count, climate variables, and pollutants

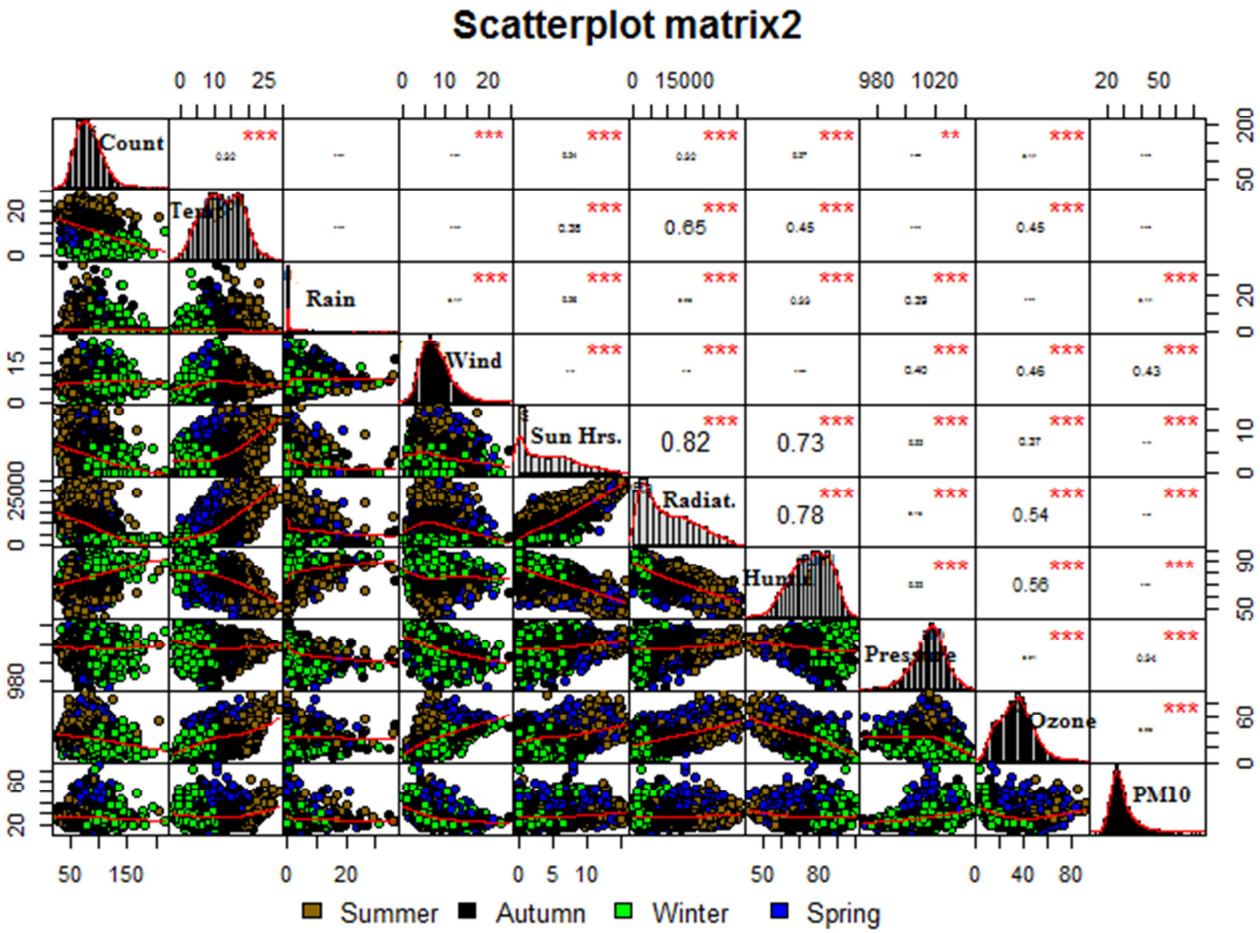

Figure 9: Scatter plot matrix of variables distribution, histograms, kernel density overlays, correlations, and significance

Table 13: Mean seasonal temperature and admissions count

Seasons	Average Temperature	Admissions Count
Summer	18.201	920
Autumn	12.460	910
Winter	5.867	903
Spring	10.766	920

5.6.2 Modelling with Generalized linear model (GLM)

We developed a generalized linear model (GLM) with the climate and pollution variables where 'the daily lower respiratory disease counts' is the response variable. Before developing a full GLM model, we performed and calculated the ANOVA, Akaike information criterion (AIC) (section 5.2.7 and 5.2.12), and Bayesian Information Criteria (BIC) (section 5.2.7 and 5.2.12), to justify the inclusion of all variables in the full model. We did not perform the Likelihood Ratio (LR) test, since the family of the distribution is quasi-Poisson.

Selection of variables

We started a GLM model with the meteorological variable temperature since it is the highest influential factor on health. We then included each of the other variables to form a new GLM, and used ANOVA, QAIC and QBIC (Hastie and Tibshirani 1990; Wood 2006) to check whether the inclusion of that variable actually significantly improves the model or not.

In statistics, the likelihood ratio test is a statistical test to compare the fit of two models, one of which is the null model (let's say model 1 in **Table 14**) is a special case of the alternative model (say model 2 here). The test is based on the likelihood ratio, which expresses how many times more likely the data under one

model fits better than the other. The ANOVA can be used to compute a p-value, or compare to a critical value to decide whether to reject the null model (Model 1) in favour of the alternative model (Model 2).

Table 14: Model check and selection of variables

Models	**Model Form**	**ANOVA for model comparisons : Pr(>F)**	**QAIC**	**QBIC**	**Improved / Significant (YES / No)**
Model1	Count~**Temp**		42925.04	42995.04	
Model2	Count ~ Temp + **Rain**	0.2208	42927.86	43032.86	No
Model3	Count ~ Temp + Rain + **Wind**	3.065e-09 ***	42741.16	42879.99	Yes
Model4	Count ~ Temp + Rain + Wind + **Sunhours**	1.363e-15 ***	42397.84	42568.09	Yes
Model5	Count ~ Temp + Rain + Wind + Sunhours + **Radiation**	5.664e-05 ***	42319.75	42522.87	Yes
Model6	Count ~ Temp + Rain + Wind + Sunhours + Radiation + **Humidity**	8.917e-08 ***	42174.59	42409.71	Yes
Model7	Count ~ Temp + Rain + Wind + Sunhours + Radiation + Humidity + **Pressure**	0.5797	42183.77	42452.52	No
Model8	Count ~ Temp + Rain + Wind + Sunhours + Radiation + Humidity + Pressure + **Ozone**	0.6938	42193.8	42496.25	No
Model9	Count ~ Temp + Rain + Wind + Sunhours + Radiation + Humidity + Pressure + Ozone + **PM10**	0.0129 *	42171.05	42506.85	Yes
Model10	Model9 - Rain	0.002241 **	42211.11	42514.13	No
Model11	Model9 - Pressure	0.9456	42160.22	42462.36	Yes
Model12	Model9 - Ozone	0.79	42160.57	42462.69	Yes
Model13	**Model9 - Radiation**	**0.2898**	**42166.34**	**42468.72**	**Yes**
Model14	Model9-Pressure-Ozone-Radiation	0.6629	42147.14	42382.2	Yes
Model15	Model9-Pressure-Radiation	0.5695	42155.51	42424.23	

Statistically significant at 0.1 %(***), 1 %(**), 5 %(*), 10%(**.**)

In **Table 14**, we can see that including Rain in model 1 does not improve the model (though not statistically significant). However, in model 10 removing the variable rain from the full model does not improve model 9 which is also statistically significant according to ANOVA, AIC and BIC criterion (the lower the better). Similarly we see that inclusion of the variables such as Wind Speed, Sun hours, Radiation, Humidity, and PM10 significantly improves the model results according to ANOVA, AIC, and BIC. On the other hand, we can see from ANOVA, AIC, and BIC results that Pressure and Ozone do not improve the model, i.e., statistically not significant (**Table 14**). For Radiation, the inclusion does significantly improve the model results, but due to multicollinearity (**Table 15**), we remove this variable from the final model. Therefore, Model 13 (**Table 14**) is the final model for the GLM analysis.

Checking multicollinearity

Multicollinearity means that some of the explanatory variables are not independent but correlated. We can check for multicollinearity roughly by means of the correlation matrix (e.g., **Figure 8**, **Figure 9**). In such a matrix, when the correlation coefficient between two explanatory variables is above 0.8, one needs to be aware of possible collinearity. If the correlation coefficient is above 0.95, the problem is really serious. These can be considered as a rule of thumb.

A diagnostic approach to check for multicollinearity after performing regression analysis is to display the Variance Inflation factor (VIF). VIF is a measure of how much the variance of the estimated regression coefficient $\hat{\beta}_i$ is "inflated" by the existence of correlation among the predictor variables in the model. Computationally, the VIF for $\hat{\beta}_i$ is defined as

$$VIF\ (\hat{\beta}_i) = \frac{1}{1 - R_i^2} \tag{5.42}$$

Here R_i^2 is the coefficient of determination of the regression equation. The magnitude of the multicollinearity can be measured by considering the size of $VIF(\hat{\beta}_i)$.

A VIF of 1 means that there is no correlation among the i-th predictor and the remaining predictor variables, and hence the variance of $\hat{\beta}_i$ is not inflated at all. The general rule of thumb is that VIFs exceeding 4 warrants further investigations, >5 indicates multicollinearity is high and exceeding 10 are signs of serious multicollinearity requiring correction.

Table 15: Variation inflation factor: checking multicollinearity

Variable Name	VIF	Multicollinearity (Yes / No)
Temperature	2.040425	No
Rain	1.310265	No
Wind Speed	2.095844	No
Sun Hours	3.748400	No
Radiation	6.971240	Yes
Relative Humidity	3.405485	No
Pressure	1.487464	No
Ozone	2.848567	No
PM10	1.305991	No

From **Table 15**, we can see that the variable Radiation show the existence of multicollinearity in model 9 and thus we finally select model 13 judging by AIC and BIC (**Table 14**).

Table 16: Model fitting results

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	4.080189	0.5054349	8.073	9.26e-16 ***
Temperature	-0.011875	0.0009034	-13.145	< 2e-16 ***
Rain	-0.0040931	0.0013363	-3.063	0.00221 **
Wind Speed	0.0103255	0.0016832	6.135	9.45e-10 ***
Sun Hours	-0.0028628	0.0016581	-1.727	0.08433 ·
Relative Humidity	0.0042411	0.0007459	5.686	1.40e-08 ***
Pressure	0.0000348	0.0004806	0.072	0.94227
Ozone	-0.0003191	0.0004867	-0.656	0.51216
PM10	0.0018279	0.0007151	2.556	0.01062 *

Statistically significant at 0.1 %(***), 1 %(**), 5 %(*), 10%(**.**)

*Model fitting results: (Model 13 in **Table 14**)*

The **Table 16** illustrates model fitting results, and **Table 17** provides the odds and corresponding confidence interval of the estimates. In **Table 16**, we see that **Temperature, Wind Speed, Relative Humidity** are highly significant at 0.1% level of significance ($\alpha = .001$). **Rain, PM10, Sun Hours** are also significant at 1%, 5% and 10% level of significance respectively. **Pressure** and **Ozone** are found to be not significant on the lower respiratory disease count. We observe that **Temperature, Rain,** and **Sun Hours** are negatively affecting daily disease count, and **Relative Humidity, Wind speed,** and **PM10** are showing positive relationships.

Table 17 explains the results of **Table 16** by calculating the odds (exponentiation, since in Poisson we have log-link) of the estimates along with the respective confidence interval of the odds. For example, **Temperature** is significantly affecting (**Table 16**) daily lower respiratory admissions counts and from **Table 17**, we can see that the **odds ratio** corresponding to **mean temperature** is 0.9881950 (95% CI: (0.9864468, 0.9899462)). This implies that if

we fix all other variables (e.g., Rainfall, Humidity), increasing mean temperature by one unit will decrease daily LR emergency admissions count by 0.011805. Similarly, **Wind Speed** is significantly affecting (**Table 16**) daily lower respiratory admissions counts and from **Table 17**, one observes that the **odds ratio** corresponding to **mean Wind Speed** is 1.0103790 (95% CI: (1.0070513, 1.0137177)). This implies that holding all other variables (e.g., Temperature, Rainfall, and Humidity) as constant, increasing mean wind speed by 1 unit will increase the daily emergency LR admissions count by 0.010379.

Table 17: Odds and 95% confidence interval of the estimate

Coefficients	Odds	Confidence Interval of the estimate	
		2.5%	97.5%
Intercept	59.1566487	21.9672943	159.3054215
Temperature	0.9881950	0.9864468	0.9899462
Rain	0.9959153	0.9933102	0.9985271
Wind Speed	1.0103790	1.0070513	1.0137177
Sun Hours	0.9971413	0.9939061	1.0003870
Relative Humidity	1.0042501	1.0027830	1.0057193
Pressure	1.0000348	0.9990933	1.0009772
Ozone	0.9996810	0.9987278	1.0006351
PM10	1.0018296	1.0004264	1.0032348

I calculated the Nagelkerke R-squared (Nagelkerke 1991) for the final model to check the goodness of fit. Nagelkerke R-squared is a modification of the Cox-Snell and ranges between 0 to 1. In our case, the value of the Nagelkerke R-squared is 0.5914073. This indicates that the model has a reasonable predicting power in predicting the emergency lower respiratory hospital admissions given the independent variables (climate and pollution factors). But we still have room to improve the model. Nagelkerke R-squared is a better measurement to compare between models rather than interpreting a specific value of a single model.

Model diagnostics

From the model diagnostic plot (**Figure 10**), we can see that the model fit the data reasonable well, and there is no influential value (plot of cook distance) that may statistically change the results of the model. In the Residual versus Leverage plot (**Figure 10**, and also see section 5.2.10) we can see that almost all the data points are in the non-influential zone (not low or high leverage) except 3 data points which is reasonable for a good model. From the cooks distance we can see that the influence of these 3 data point is very small (<0.02%). From the residual versus fitted plot and the normal Q-Q plot we can also see that the model fit the data reasonably well (**Figure 10**).

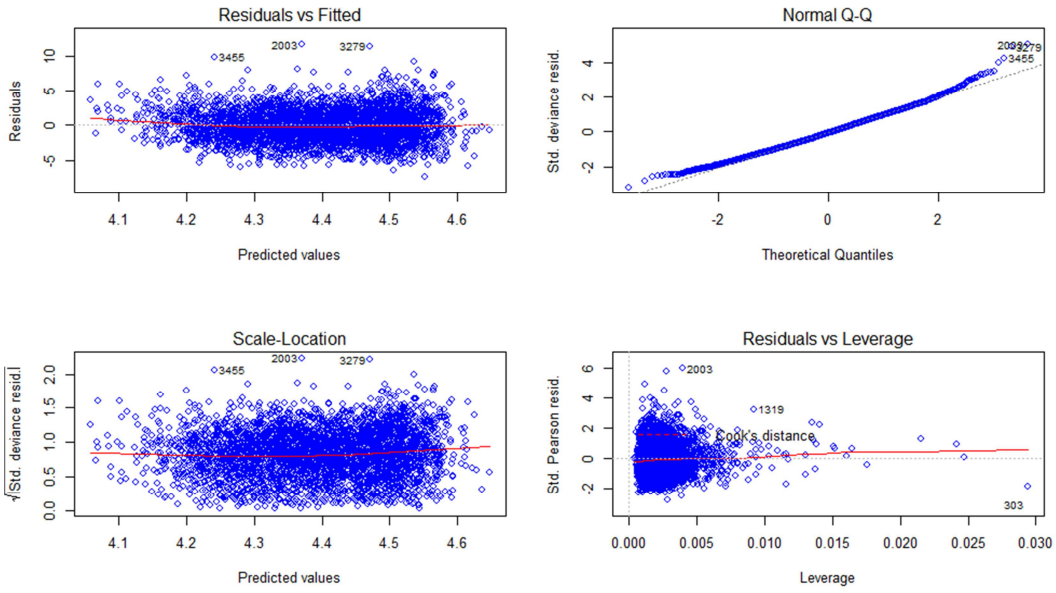

Figure 10: Model diagnostic results from the GLM modelling

In summary, we conclude from the results from the Generalized Linear Model that the **Temperature, Wind Speed, Relative Humidity, Rainfall, PM10, Sun Hours** significantly affects daily lower respiratory hospital admissions. In

contrary, **Pressure** and **Ozone** do not have any significant relationship with LR emergency hospital admissions, and Radiation was removed due to multicollinearity. The diagnostic plots of the final model also reveal that the model fits the data reasonably well.

5.7 Chapter summary

In this chapter, we illustrate the theoretical background of the GLM modelling and developed some GLMs based on our problems and data sets. We also verified how considering more than one climate factor (temperature) can improve the model fitting results. We also selected the significant climate and pollution variables for emergency lower respiratory hospital admissions. In the next chapter, we will demonstrate how considering the delayed effect and non-linearity of the data can improve the model fitting results. In addition to this, we also propose our final model.

Chapter 6

Modelling the non-linearity and delayed effect of climate factors

6.1 Introduction

In this chapter, we propose a new DLNM model considering the non-linear relationships between climate and pollution factors and their delayed effect on the emergency hospital admissions for lower respiratory disease. We also describe the theoretical backgrounds and properties of the model under the context of our problem. In section 6.2, we describe some commonly used smoothing techniques and spline functions for dealing the non-linearity of data. Section 6.3 illustrates the general representations of the distributed lag modelling. Here, we include the basic layout of the model, delayed effect in the model and concept of cross basis which is related to the final model. The framework of the Distributed lag non-linear model has also been described in this section. Finally, we illustrate the proposed final model in this section.

6.2 Smoothing techniques and splines

In this section, we highlight some useful smoothing techniques and spline functions. We describe it since the concept has been adopted later on for dealing

the non-linear nature of the factors of climate and pollutions. We describe only those techniques that relate to this study. For other techniques, we suggest interested readers go through the references (Hastie and Tibshirani 1987; Buja, Hastie et al. 1989; Hastie and Tibshirani 1990).

Smoother

A smoother is a tool for summarising the trend of a response measurement Y as a function of one or more predictor measurements $X_1, X_2, \ldots\ldots\ldots X_p$ (Hastie and Tibshirani 1990). It is called 'Smoother' because of its less variability than Y. Because of its nonparametric nature, it is considered as a tool for **nonparametric regression.** Smoothers can be broadly classified in linear and non-linear. Examples of linear smoothers are: running means, locally-weighted running lines, kernel smoothers, smoothing splines, bin smoothers, and the least square line. The smoother matrix cannot be constructed for non-linear smoother. Most of the linear smoothers depend on a smoothing parameter and if a data oriented technique such as cross-validation is used to select this parameter, they become non-linear smoothers. Examples of non-linear smoothers are running median, robust smoother ("Lowess") and cross-validated variable span smoothers ("Super smoother").

Running means

A running mean smoother produces a fit at x_i by averaging the data points in a neighbourhood N_i around x_i. The neighbourhoods that are commonly used are symmetric neighbourhoods. Assuming, for ω between 0 and 1, that $[\omega n]$ is odd ([.] denoting the integer part), these consists of $[\omega n]$ points, $([\omega n] - 1)/2$ to the

left and right of x_i plus x_i itself. The number ω called the span and controls the smoothness of the resultant estimate – larger spans tend to produce smoother functions.

Running-line smoothers

A running-line smoother fits a line by least square to the data points in a symmetric nearest neighbourhood N_i around each x_i (Buja, Hastie et al. 1989). The estimated smooth at x_i is the value of the fitted line at x_i. This is done for each x_i. The running-line smoother is considered to be the improvement over the running mean because it reduces the biases near the endpoints. Through the use of updating formulas, a running-line smoother can be computed with only $O(n)$ calculations (once the data are sorted). The running-line smoother often produces quite jagged output. When used in an iterative procedure, it is often desirable to re-smooth the final function. Alternatively, it can be modified to produce smoother output, at the cost of increased computations (e.g., adopting the locally-weighted running lines below).

Kernel smoothers

A kernel smoother uses an explicitly defined set of local weights, defined by the kernel, to produce an estimate at each target value (Hastie and Tibshirani 1990). Usually a kernel smoother uses weight that decrease in a smooth fashion as one move away from the target point. Choice of kernel is relatively unimportant compared to the choice of the bandwidth. Kernel smoothers show biases at the end point which can be corrected by using the running lines weighted by a Gaussian kernel.

Locally-weighted running line smoother

This smoother combines the strict local nature of running lines, and the smooth weights of kernel smoothers, in a locally-weighted running-line smoother (Buja, Hastie et al. 1989). The locally weighted smoothers are popular, since they enjoy the best of both of nearest neighbourhood and symmetric neighbourhood. Since the weights have to be recomputed for each neighbourhood, locally-weighted running line smoothers require $O(n^2)$ computations (Hastie and Tibshirani 1990).

Regression splines

Regression spline is a projection method for fitting splines. It can be also projected onto k basis or B-splines placed at judiciously chosen knots in the range of x. Thus, it represents the fit as a piecewise polynomial and the regions that define the pieces are separated by a sequence of knots or breakpoints $\xi_1, \xi_2, \ldots \ldots, \xi_k$ (Buja, Hastie et al. 1989). The number k and positions of the knots are all parameters of the procedure.

Regression splines are attractive because of the computational properties when the knots are given. Fixed knot cubic splines are less appealing than smoothing splines. Although k, the number of knots is usually considered to be the smoothing parameter, one has also to determine the placement of the knots. Thus the difficulty of choosing the number and position of the knots is a drawback of this approach. Another problem is that the smoothness of the estimate cannot be easily verified continuously as a function of a single smoothing parameter (Hastie and Tibshirani 1987).

Quadratic and cubic spline bases

The simple regression splines are not suitable for most applied smoothing problems. It is overly restrictive to only estimate piecewise functions that are linear between the knots during estimating more curvilinear functional forms. The solution is to combine piecewise regression functions with polynomial regression by representing each piecewise regression function as a piecewise polynomial regression function.

Piecewise polynomials offer two advantages. First, piecewise polynomials allow for non-linearity between the knots. Second, piecewise polynomial regression functions ensure that the first derivatives are defined at the knots, which guarantees that the spline estimate will not have sharp corners. It is very simple to alter the regression splines and thus accommodate piecewise polynomials. In any simple model, we can estimate piecewise polynomial fits by adding x^2 to the basis and squaring the results from the basis functions. This alteration forms a quadratic spline basis with a single knot.

Typically, cubic spline bases are used instead of quadratic bases to allow for more flexibility in fitting peaks and valleys in the data. A spline model with a cubic basis and two knots c1 and c2 forms from the following linear regression model:

$$Y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - c_1)^3 + \beta_4 (x - c_1)^3 + \varepsilon \quad (6.1)$$

The spline estimate is again the predictions from the hat matrix applied to the outcome variable. To form the hat matrix, we must first construct a model matrix that contains the correct bases. For this example, the model will contain

$$\begin{aligned} x_1 &= x \\ x_2 &= x^2 \\ x_3 &= x^3 \\ x_4 &= (x - c_1)_+^3 \\ x_5 &= (x - c_1)_+^3 \end{aligned} \tag{6.2}$$

where x represents the original predictor variable. The model matrix will consist of a constant and the above five variables. We use this model matrix to form a hat matrix that is applied to the outcome variable, and the predictions from this model serve as the spline estimate of the possibly non-linear relationship between x and y. The number of parameters used to construct the spline estimate is controlled by the number of knots. If there are k knots, with a cubic basis, the function will require $k + 4$ regression coefficients (including the intercept). The cubic basis allows for flexible fits to non-linearity between the knots and eliminates any sharp corners in the resulting estimate. The latter is true since the first derivative exists for $(x - c_1)_+^3$ and it follows that the first derivative will also exist in any linear combination of the terms in equation 6.2. For cubic regression splines, there are a number of equivalent ways to write the basis.

Natural splines

While cubic splines are widely used, they are often altered slightly to improve the fit. One limitation of cubic splines is that the piecewise functions are only fit between each knot. For data that falls before the first knot and beyond the last knot, we do not fit a piecewise function. Without fits to the boundary of the data, it is possible for the spline fit to behave erratically around the limits of x. Natural cubic splines add two knots to the fit at the minimum and maximum values of x

and fit a linear function between the additional knots at the boundary and the interior knots. This constrains the spline fit to be linear before the first knot and after the last knot. Such enforced linearity at the boundaries avoids any wild behaviour in the spline fit near the extremes of the data. Cubic splines may not display erratic fits at the boundaries, but natural splines can improve the overall spline fit should problems occur. Since little is lost by using natural splines while some gains in model fit are possible, natural cubic splines are generally preferred to cubic splines.

B-splines

There is one further refinement that is also typically applied to cubic splines. For cubic splines (natural or otherwise), the columns of $\boldsymbol{X}$, the model matrix, tend to be highly correlated since each column is a transformed version of x, which can induce considerable collinearity. The collinearity may result in a nearly singular model matrix and imprecision in the spline fit. As a remedy, one can represent the cubic spline (and any other polynomial basis) as a B-spline basis. A k-knot cubic B-spline basis can be represented as:

$$f(x) = \sum_{i=1}^{k} B_i^2\,(x)\beta_i \tag{6.3}$$

Where the B-spline basis are defined as:

$$B_i^2(x) = \frac{x - c_i}{c_{i+2+1} - c_i} B_i^{2-1}(x) + \frac{c_{i+2+1} - x}{c_{i+2+1} - c_{i+1}} B_{i+1}^{2-1}(x) \quad i = 1, \ldots, k \tag{6.4}$$

and

$$B_i^{-1}(x) = \begin{cases} 1, & \text{if } c_i \leq x < c_{i+1} \\ 0, & \text{Otherwise.} \end{cases} \quad (6.5)$$

The B-spline basis function is, in essence, a rescaling of each of the piecewise functions. The idea is similar to rescaling a set of $\boldsymbol{X}$ variables by mean subtraction to reduce collinearity. The rescaling in the B-spline basis reduces the collinearity in the basis functions of the model matrix. The resulting spline model is more numerically stable than the cubic spline. This is especially true if one is using a large number of knots and OLS is used to fit the spline model.

In our study, we used the B-spline basis function for most of the climate and pollution factors. The main causes are: B-spline is data driven and it remains like that after the boundary knots. This is not true for natural spline basis, which becomes linear beyond the boundary knots. Thus B-spline provides a more flexible fit to any non-linear datasets.

6.3 Distributed lag non-linear modelling approach

6.3.1 Introduction

The basic purpose of any generalized linear model or regression model is to define the relationship between predictors and outcome, and estimate the related effect. But sometimes the effect of a specific exposure event is not limited to the period when it is observed, but rather delayed in time. This introduces the problem of modelling the relationship between an exposure occurrence and a sequence of future outcomes, specifying the distribution of the effects at different times after the event (defined **lags**). Ultimately, this step requires the definition of

the additional lag dimension of an exposure–response relationship, describing the time structure of the effect. This situation occurs frequently when assessing the short-term effects of environmental stressors: several time-series studies have reported that the exposure to high levels of air pollution or extreme temperatures affect health for a period lasting some days after its occurrence (Braga, Zanobetti et al. 2001; Gasparrini, Armstrong et al. 2010). Furthermore, the complexity increases in the presence of so-called 'harvesting': the phenomenon that arises when a stressor affects mainly a pool of frail individuals, whose events are only brought forward by a brief period of time by the effect of exposure (Schwartz 2001; Gasparrini, Armstrong et al. 2010). For non-recurrent outcomes, the depletion of the pool following any extreme-event (event) results in some reduction of cases few days later, thereby reducing the overall long-term impact. For both these reasons, the estimate of the effect depends on the appropriate specification of the lag dimension of the dependency, defining models flexible enough to represent simultaneously the exposure–response relationship and its temporal structure.

Distributed lag models (DLM) played a significant role to deal with delayed effects on health. The main advantage of this method is that it allows the model to contain a detailed representation of the time-course of the exposure–response relationship, which in turn provides an estimate of the overall effect in the presence of delayed contributions or harvesting. While conventional DLMs are suitable for describing the lag structure of linear effects, the distributed lag non-linear models (DLNMs) serve to represent non-linear relationships. The

DLNM can describe, in a flexible way, effects that vary simultaneously both along the space of the predictor and in the lag dimension of its occurrence.

6.3.2 The basic model

Distributed lag non-linear models (DLNMs) represent a modelling framework to flexibly describe associations showing potentially non-linear and delayed effects in time series data. This methodology rests on the definition of a *cross basis*, a bi-dimensional functional space expressed by the combination of two sets of basis functions, which specify the relationships in the dimensions of predictors and lags, respectively.

A general representation

A general model representation to describe the time series of outcomes Y_t with $t = 1,2, \dots, n$ is given by

$$g(\mu_t) = \alpha + \sum_{j=1}^{J} S_j(x_{tj}; \beta_j) + \sum_{k=1}^{K} \gamma_k u_{tk}, \qquad (6.6)$$

where $\mu \equiv E(Y)$, g is a monotonic link function, and Y is assumed to follow the exponential family of distribution (MacCullagh and Nelder 1989; Dobson and Barnett 2008).

The functions S_j denote smoothed relationships between the variables x_j and the linear predictor, defined by the parameter vectors β_j. S_j might be also specified through non-parametric methods based on generalized additive models (Hastie and Tibshirani 1990; Wood 2006). The variables u_k include other predictors with linear effects specified by the related coefficients γ_k. The

outcomes Y_t are commonly daily counts (in time series analyses of environmental factors) and assumed to originate from over dispersed Poisson distribution with a canonical log-link. Usually these include a smooth function of time to capture the effect of confounders changing slowly over time, expressed as seasonality or long-time trends. Non-linear effects of meteorological factors such as temperature and humidity are included as well. Categorical variables such as the days of the week or age groups are modelled as factors.

Basis functions

The relationship between x and g(μ) is represented by $s(x)$, which is included in the linear predictor of a generalized linear model as a sum of linear terms. This can be done through the choice of a basis, a space of functions having s is an element (Wood 2006). The related basis functions comprise a set of completely known transformations of the original variable x that generate a new set of variables, termed basis variables. The complexity of the estimated relationship depends on the type of basis and its dimension.

Several different basis functions have been used to describe the potentially non-linear health effects of environmental factors, the choice depending on the assumptions about the shape of the relationship, the degree of approximation required by the specific purposes of the investigation, and interpretational issues.

Among completely parametric methods, the main choices typically rely on functions describing smooth curves, such as polynomials or spline functions, or on the use of a linear threshold parameterization (hockey-stick model), represented by a truncated linear function $(x - k)_+$ which equals $(x - k)$ when

$x > k$ and 0 otherwise. In the hockey-stick model, the effect is likely to exist and be linear only above or below a specific cut-off point (threshold). An extension of this model assumes two distinct linear dependencies below a first threshold and above a second threshold, with a null effect in between them (double threshold).

A general representation of the simple models described above is given by

$$s(x_t; \beta) = z_t^T . \beta \tag{6.7}$$

with $z_t.$ as the tth row of the $n \times V_x$ basis matrix $\boldsymbol{Z}$, obtained by the application of the basis functions to the original vector of exposures x. Z can be then included in the design matrix of the model in equation 6.6 in order to estimate the related unknown parameters β defining the shape of the relationship.

6.3.3 Delayed effects

A delayed (or lagged) effect occurs when for any time series analysis the outcome in a specific time is determined by the level of the predictor in previous times, up to a maximum lag for any given ordered series of predictor values . Therefore, the presence of delayed effects requires taking into account the time dimension of the relationship, specifying the additional virtual dimension of the lags.

A very simple model to deal with delayed effects considers the moving average of the predictor up to a certain lag, specifying a transformed predictor which is the average of the values in that specific lag period. Although simple, this model is limited if the purpose is to assess the temporal structure of the effects. The Distributed lag models (DLMs) addressed these limitations. The main

advantage of this method is the possibility to depict a detailed description of the time-course of the relationship.

An additional dimension

In the presence of delayed effects, the outcome at a given time t may be explained in terms of past exposures $x_t - l$, with l as the lag, representing the period elapsed between the exposure and the response. A comparatively simple approach is to apply a transformation to the original vector of ordered exposures x, deriving the $n \times (L + 1)$ matrix **Q**, such as

$$q_t. = [x_t , \ldots , x_{t-l}, \ldots\ldots , x_{t-L}]^T \tag{6.8}$$

with L defining the maximum lag and $q_{.1}. \equiv$ x (the first column of Q). We can also define $[0, \ldots\ldots, l, \ldots\ldots L]^{\mathrm{T}}$ as vector of lags corresponding to the $L + 1$ columns of **Q**. This step specifies the additional lag dimension of the exposure–response relationship. Ultimately, the aim of the modelling framework proposed here is to simultaneously describe the dependency along two dimensions: the usual predictor space and in the new lag dimension.

Distributed lag models

When a linear relationship is assumed, the delayed effects can be naturally described by distributed lag models (DLM). This methodology allows the effect of a single exposure event to be distributed over a specific period of time, using several parameters to explain the contributions at different lags. The simplest formulation is an unconstrained DLM, specified by the inclusion of a parameter

for each lag (Schwartz 2000). Unfortunately, the precision of the estimates of the effects of specific lags is often very poor, due to the high correlation between exposures in adjacent days and the resulting collinearity in the model (Gasparrini, Armstrong et al. 2010).

To gain more precision in the estimate of the distributed lag curve, it is possible to impose some constraints, for example assuming a constant effect within lag intervals (Gasparrini, Armstrong et al. 2010), or describing a smooth curve using continuous functions such as polynomials (Schwartz 2000) or splines (Zanobetti, Wand et al. 2000). A simple model with the moving average of the exposures in the previous L days as a predictor can be considered as a special case of a DLM. Using the development provided in section 6.3.2 (basis functions) and section 6.3.3 (an additional dimension), it is possible to formulate a simpler and general definition of DLM, in which the shape of the distributed effects along lags is specified by a proper basis. In matrix notation

$$s(x_t;\eta) = q_t^T \cdot C_\eta \tag{6.9}$$

where **C** is an $(L+1) \times V_l$ matrix of basis variables derived from the application of the specific basis functions to the lag vector l, and η a vector of unknown parameters. The addition of the supplementary dimension in equation 6.8 provides a structure for the application of the basis matrix **C**, in order to describe the effects of lagged exposures. All the different DLMs described above can be derived from equation 6.9, by specifying the correspondent basis matrix: $\boldsymbol{C} \equiv 1$ (a vector of ones) for the moving average model, $\boldsymbol{C} \equiv \mathrm{I}$ (an identity matrix) for the

unconstrained DLM, or C defined as a series of polynomial or spline functions of l for DLMs describing the effect as a smoothed curve along lags. From equation 6.9, we can define

$$\boldsymbol{W} = \boldsymbol{QC} \tag{6.10}$$

with **W** the matrix of the V_l transformed variables that are included in the design matrix to allow estimation of the parameters η. The interpretation of the estimated parameters $\hat{\eta}$ is aided by construction from them of the implied linear effects b at each lag, following:

$$\begin{aligned} \hat{\beta} &= C\hat{\eta} \\ V(\hat{\beta}) &= C\, V(\hat{\eta}) C^T \end{aligned} \tag{6.11}$$

Here the choice of the basis to derive **C** can be considered as the application of a constraint to the shape of the distributed lag curve described by $\hat{\beta}$.

Despite the specification of the basis functions in equation 6.9 being slightly different to that in equation 6.6, i.e. being applied to the vector l instead of the exposure series x itself, their goal is conceptually similar to describe the shape of the relationship, the former along distributed lags and the latter in the space of x.

6.3.4 Distributed lag non-linear models

The family of DLNM is achieved through the generation of a new model framework for describing non-linear relationships both in the space of the predictor and along lags. A such model framework based on the concept of the cross-basis.

The concept of cross-basis

The algebraic notation of DLNMs can be quite complex because of its three-dimensional arrays. But the basic concept of a cross-basis on which the DLNMs depend on is straightforward. The cross-basis can be imagined as a bi-dimensional space of functions describing at the same time the shape of the relationship and the distributed lag effects. Thus choosing a cross-basis is based on two sets of basis functions, which will be combined to generate the cross-basis functions. The choice of the two sets of basis functions for each space is perfectly independent, and should be based on a-priori assumptions or on a compromise between complexity and generalizability. Linear, threshold, strata, polynomial or spline functions can be used to define the relationship along the space of predictor, while unconstrained, strata, polynomial or spline functions can be applied to specify the shape along lags.

The algebra of DLNM

To model the shape of the relationship described above, we need to apply simultaneously the two transformations described in section 6.3.2 and section 6.3.3.

First, as in section 6.3.2, we choose a basis for **x** to define the dependency in the space of the predictor, specifying **Z**. Then we create the additional lag dimension, as in section 6.3.3, for each one of the derived basis variables of **x** stored in the **Z**. This produces a $n \times v_x \times (L+1)$ array **R**, which represents the lagged occurrences of each of the basis variables of **x**. The construction is symmetric, in the sense that the order of the two transformations can be reversed, applying the basis functions directly to each column of the matrix **Q**.

Defining **C**, the matrix of basis variables for seen in section 6.3.4, a DLNM can then be specified by

$$g(\mu_t) = \alpha + \sum_{j=1}^{v_x} \sum_{j=1}^{v_l} r_{tj.}^T \, c_{.k} \, \eta_{jk} = w_{t.}^T \eta, \tag{6.12}$$

where $r_{tj.}$ is the vector of lagged exposures for the time t transformed through the basis function j. The vector $w_{t.}$ is obtained by applying the $v_x . v_l$ vx ·v cross-basis functions to x_t , similar to equation 6.10. We keep the same notation to emphasize the fact that the DLM specified in equation 6.9 is a special case of the more general DLNM in equation 6.12. To reach a compact formula for **W** of a similar form to equation 6.10, we need to present it as a tensor product. Defining $P_{i,j}$ as the operator permuting the indexes i and j of an array and assuming a generic $i \times j$ matrix as a $i \times j \times 1$ array, it follows that

$$\dot{\boldsymbol{A}} = \left(\mathbf{1}^T \otimes \dot{\boldsymbol{R}}\right) \odot \left(1 \otimes P_{1,3}(\boldsymbol{C}) \otimes \mathbf{1}^T\right) \tag{6.13}$$

with **1** indicating vectors of ones with appropriate dimensions. The symbols $\otimes$ and $\odot$ represent the Kronecker and Hadamard products, respectively. The $n \times (v_x \cdot v_l) \times (L + 1)$ array $\dot{\mathbf{A}}$ is then re-arranged, summing along the third dimension of lags to obtain the final matrix of cross-basis functions **W**. The equation in equation 6.13 is a modified version of the formula used to implement smoothing on a multidimensional grid through tensor product bases (Gasparrini, Armstrong et al. 2010). The main difference in the cross-basis approach lies in the dimensions considered in the model. While the original method provides a framework to describe a smooth surface in the space of two distinct variables, the

DLNM expresses simultaneously the effects in the space of a variable and in its lag dimension.

Interpreting a DLNM

DLNM raise no more problems than any other generalized linear model, despite its complex parameterization, estimation of and inference about the parameters. It can be carried out with the common statistical software's after the cross-basis variables have been specified. Nonetheless, while the interpretation of the simpler DLM in equation 6.9 is straight forward, consisting in reporting the estimated linear effects $\hat{\beta}$ in equation 6.10 for each lag, the results of a more complex DLNM with smoothed non-linear dependencies are harder to summarise.

One solution is to build a grid of predictions for each lag and for suitable values of exposure, using three-dimensional plots to provide an overall picture of the effects varying along the two dimensions. In addition, it is possible to summarise the relationship at single predictor or lag values, by cutting a "slice" of the grid along specific values. These summaries express a lag-specific association, defined along the predictor space at a given lag value, or a predictor-specific association, defined along the lag space at a given predictor value, respectively. Finally, an estimate of the overall cumulative association can be computed by summing all the contributions at different lags for each predictor value. The associations are usually reported versus a reference value of the predictor, cantering the basis functions for this space to their corresponding transformed values (Gasparrini 2011)

Given a vector x^p of the m exposure values used for prediction and the resultant $m \times v_x$ matrix $\boldsymbol{Z^p}$, the corresponding $m \times v_x \times (L+1)$ array $\dot{\mathbf{R}}^p$ can be

derived by repeating the matrix Z^p $L+1$ times along the dimension of the lags. The computation of $\dot{\mathbf{R}}^p$ is slightly different than for the array $\dot{\mathbf{R}}$ used in the estimation process in equation 6.12. In this case the interest lies in the prediction of the effects at each lag given an exposure, not in the temporal sequence of the exposures themselves. The final array $\dot{\mathbf{A}}^p$ follows simply substituting $r_{\mathrm{t\,j.}}$ with $r_{\mathrm{t\,j.}}^{\mathrm{p}}$ in equation 6.12 or $\dot{\mathbf{R}}$ with $\dot{\mathbf{R}}^p$ in equation 6.13.

The prediction grid, expressed with the $m \times (L+1)$ matrix of predicted effects **E** and related matrix of associated standard errors $\boldsymbol{E}^{sd}$, can be derived using the vector of estimated coefficients $\hat{\eta}$, computed from the model fitted including the matrix of cross-basis functions **W**. For each lag $\boldsymbol{l}$

$$e_{.l} = A^p_{..l}\hat{\eta} \tag{6.14}$$

and, given $V(\hat{\eta}\,)$ the variance–covariance matrix of the estimated coefficients

$$e^{sd}_{.l} = \sqrt{diag\left(\boldsymbol{A}^p_{..l} V(\hat{\boldsymbol{\eta}})\boldsymbol{A}^p_{..l}\right)} \tag{6.15}$$

This grid is useful to compute the estimates of the effects by exposure at lag l_{p} or by lag at exposure x_{p}, simply taking $e_{.l_p}$ and $e_{.x_p}$. , respectively.

Finally, an estimate of the overall effect can be computed by summing all the contributions at different lags. The vector e_{tot} , and associated standard errors e^{sd}_{tot}, obtained summing the contributions at each lag, specify the effects of exposure over the whole lag period. They are obtained from

$$e_{tot} = W^p\hat{\eta} \tag{6.16}$$

and

$$e_{tot}^{sd} = \sqrt{diag\left(W^{p}V(\hat{\boldsymbol{\eta}})\boldsymbol{A}_{..l}^{p^T}\right)} \tag{6.17}$$

6.3.5 The final Model

We see from the section 6.3.4 that the Distributed lag linear and non-linear models are based on two basis function: a lag basis for representing different lags of the explanatory variables and cross-basis function $s(x_t)$ for N-length series of the explanatory variables $x = [x_t, \ldots, x_{t-l}, \ldots\ldots, x_{t-L}]^T$. The definition of $s(x_t)$ first require the derivation of the $N \times (L+1)$ matrix $\boldsymbol{Q}$ of the lagged exposure so that $q_t. = [x_t, \ldots, x_{t-l}, \ldots\ldots, x_{t-L}]^T$. This actually characterizes the new lag dimension identified by the vector $\ell = [0, \ldots\ldots, \ell, \ldots, L]$, having L as the maximum lag.

Now, choosing a first basis with dimensions v_ℓ to represent the association along the new lag space, we can compute a $(L+1) \times v_\ell$ basis matrix $\boldsymbol{C}$ by applying the related functions to ℓ. A compact and general expression for the lag-basis function $s(x_t)$ for DLM is given by:

$$s(x_t; \eta) = \sum_{j=1}^{v_l} q_{t.}^{\mathrm{T}}\, c_{.k}\eta_k = q_t^T \cdot C_\eta = w_{t.}^T\eta \tag{6.18}$$

Note that different models are specified with different choices of the basis to derive $\boldsymbol{C}$. Here, C_η represents the lag specific contributions. The v_ℓ-length parameter vector $\boldsymbol{\eta}$ can be estimated from the equation 6.10.

The non-linear extension to the DLNMs requires the choice of a second basis with dimension v_x to model the relationship along the space of the predictor x, obtaining the $N \times v_x$ basis matrix $\boldsymbol{Z}$ (see equation 6.10) from the

application of the related function to x. Applied together with the transformation which defines the matrix of lagged exposure $\boldsymbol{Q}$ above, this step produces a three-dimensional $N \times v_x \times (L+1)$ array $\dot{\boldsymbol{R}}$.

$$s(x_t;\eta) = \sum_{j=1}^{v_x}\sum_{j=1}^{v_l} r_{tj.}^{T}\, c_{.k}\, \eta_{jk} = w_{t.}^{T}\eta \tag{6.19}$$

To formulate with the cross-basis based on our problem, let us consider the lag period for different climate and pollution variables. For example, if we consider the variable daily mean temperature with maximum lag as *L*=30, then the first basis for temperature from equation 6.18 can be given as

$$s(x_t;\eta) = \sum_{j=1}^{v_{l(L=30)}} q_{t.}^{\mathrm{T}}\, c_{.k}\eta_k \tag{6.20}$$

Where $\ell = [0, \dots\dots, 30]$ and $q_{t.} = [x_t, \dots, x_{t-l}, \dots\dots, x_{t-30}]^T$. Similarly, we can construct the first basis for other variables like daily rain, wind speed, sun hours, relative humidity, pressure, Ozone, and PM10. If we consider a non-linear extension to DLNMs for daily mean temperature, the parameterization of the cross basis function $s(x_t)$ for DLNMs can be give as,

$$s(x_t;\eta) = \sum_{j=1}^{v_{x=temp}}\sum_{k=1}^{v_{l(L=30)}} r_{tj.}^{T}\, c_{.k}\, \eta_{jk} = w_{t.}^{T}\eta.\,(temp, L=30) \tag{6.21}$$

Expanding this procedure by considering all the climate and pollution factors, the DLNM model can be defined as follows,

$$s(x_t;\eta) = \sum_{j=1}^{v_{x=temp}} \sum_{k=1}^{v_{l(L=30)}} r_{tj.}^{T}\, c_{.k}\, \eta_{jk} + \sum_{j=1}^{v_{x=Rain}} \sum_{k=1}^{v_{l(L=15)}} r_{tj.}^{T}\, c_{.k}\, \eta_{jk}$$

$$+ \sum_{j=1}^{v_{x=Wind\ speed}} \sum_{k=1}^{v_{l(L=20)}} r_{tj.}^{T}\, c_{.k}\, \eta_{jk} + \sum_{j=1}^{v_{x=Sun\ hours}} \sum_{k=1}^{v_{l(L=20)}} r_{tj.}^{T}\, c_{.k}\, \eta_{jk}$$

$$+ \sum_{j=1}^{v_{x=R.Humidity}} \sum_{k=1}^{v_{l(L=20)}} r_{tj.}^{T}\, c_{.k}\, \eta_{jk} + \sum_{j=1}^{v_{x=Pressure}} \sum_{k=1}^{v_{l(L=10)}} r_{tj.}^{T}\, c_{.k}\, \eta_{jk}$$

$$+ \sum_{j=1}^{v_{x=Ozone}} \sum_{k=1}^{v_{l(L=30)}} r_{tj.}^{T}\, c_{.k}\, \eta_{jk} + \sum_{j=1}^{v_{x=PM10}} \sum_{k=1}^{v_{l(L=30)}} r_{tj.}^{T}\, c_{.k}\, \eta_{jk}$$

$$= w_{t.}^{T}\eta.(temp, L=30) + w_{t.}^{T}\eta.(rain, L=15)$$

$$+ w_{t.}^{T}\eta.(Wind\ Speed, L=20) + w_{t.}^{T}\eta.(Sun\ Hours, L=20)$$

$$+ w_{t.}^{T}\eta.(R.Humidity, L=20) + w_{t.}^{T}\eta.(Pressure, L=10)$$

$$+ w_{t.}^{T}\eta.(Ozone, L=30) + w_{t.}^{T}\eta.(PM10, L=30) \qquad (6.22)$$

From equation 6.10, in matrix notation this can be written as

$$\boldsymbol{W} = \boldsymbol{QC}$$

These models may be fitted through common generalized linear model techniques with the inclusion of cross-basis matrix $\boldsymbol{W}$ in the design matrix (see equation 6.10 for elaborations). The vector $\widehat{\boldsymbol{\eta}}$ of the estimated parameters of the cross-basis function in (6.22) represents simultaneously non-linear and lagged dependency, and its length $v_x \times v_\ell$ is equal to the product of the dimensions of the bases for two spaces. In completely parametric models as those described

here, the dimensionality is directly associated with the notion of the degrees of freedom (df), related to the flexibility of the function and smoothness of the estimated dependency.

The form of our final model also includes two more factors in the DLNMs model assuming some linear effects of the response variable: daily emergency hospital admissions counts for lower respiratory disease. We add natural cubic splines of time with 7 df to control the secular trends and any additional confounding by seasonally varying factors other than the selected climate and pollution factors in the model. For the same type of confounding factor due to any particular day of a week we included 'day of the week' (DOW) in the model. So eventually our final model takes the form as below:

$$\begin{aligned} s(x_t;\eta) = {} & w_{t.}^{T}\eta.(temp, L=30) + w_{t.}^{T}\eta.(rain, L=15) \\ & + w_{t.}^{T}\eta.(Wind\ Speed, L=20) + w_{t.}^{T}\eta.(Sun\ Hours, L=20) \\ & + w_{t.}^{T}\eta.(R.Humidity, L=20) + w_{t.}^{T}\eta.(Pressure, L=10) \\ & + w_{t.}^{T}\eta.(Ozone, L=30) + w_{t.}^{T}\eta.(PM10, L=30) + S_{ns}(\text{Time}) \\ & + \text{DOW} \end{aligned} \tag{6.23}$$

The interpretations of the final model in equation 6.23, the general form in section 6.3.4 can be followed. We interpret the results of the final model in section 7.3.

6.4 Chapter summary

In this chapter, we developed the mathematical form of the new Distributed Lag Non-linear model. This new model considers the non-linearity in the climate and

pollution datasets and their delayed impact on the emergency hospital admissions for lower respiratory disease counts. We proposed to use a B-spline smoothing technique to deal with the nonlinear relationships. In the following chapter, we are going to apply this proposed DLNM model to our datasets, interpret the results, compare the models and proceed to the conclusions.

Chapter 7

Results of the final model

7.1 Introduction

In this chapter, we begin with exploratory data analysis using the dataset described in chapter 4. In this chapter, we applied the proposed final model in section 6.3.5 of the previous chapter. We also performed the model comparisons based on the results from final model and the results of the final generalized linear model in section 5.6.2. In section 7.2, we perform a graphical exploratory data analysis to visualise the pattern of the non-linear relationships of the emergency LR hospital admissions with climate and pollution factors. Section 7.3 illustrates the results of the final DLNM model along with smoothing techniques and interpretations of the results. Finally, in section 7.4, we compare the results of the various models used throughout this research and benefits of the proposed final model regarding improving the results. The section 7.5 summarises this chapter.

7.2 Exploratory data analysis

We calculated the daily mean emergency admission count for LR disease for the study period (1 January 2000 - 31 December 2009) and performed a visual exploratory data analysis using all the climate and pollution factors to compare their trends with respect to the trends of LR emergency admissions. The results of

this exploratory data analysis will help us understand the nature of the non-linear relationships of the climate and pollution factors with emergency LR admissions. In addition to that, it will also assist us decide about the spline function in the final model.

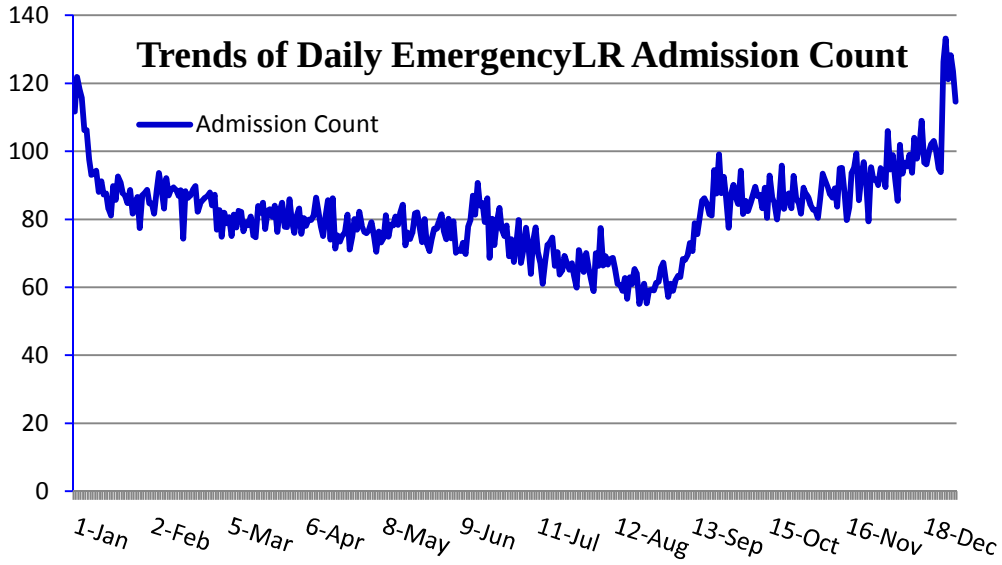

Figure 11: Trends of lower respiratory (LR) disease admissions counts

Figures 11-20 illustrate the trends of the daily mean emergency LR hospital admissions count compared to the mean values of the selected climate and pollution variables, and thus show the seasonality of the rate of change for the admissions count with the rate of change of climate and pollutants throughout the year. In **Figure 11**, it is evident that December has the highest emergency LR hospital admissions compared to any other months. Note that admissions increase from the beginning of autumn (September to November). Interestingly, the emergency LR admissions rate is lower during high temperature (summer) and higher during winter (**Figure 12**). This could be because of the nature data

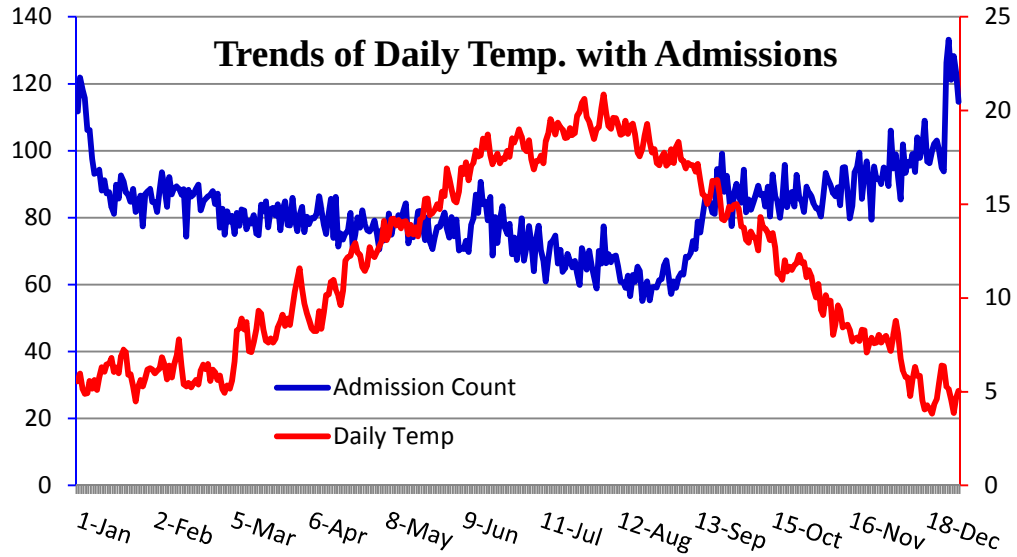

Figure 12: Trends of daily mean temp with LR admissions counts

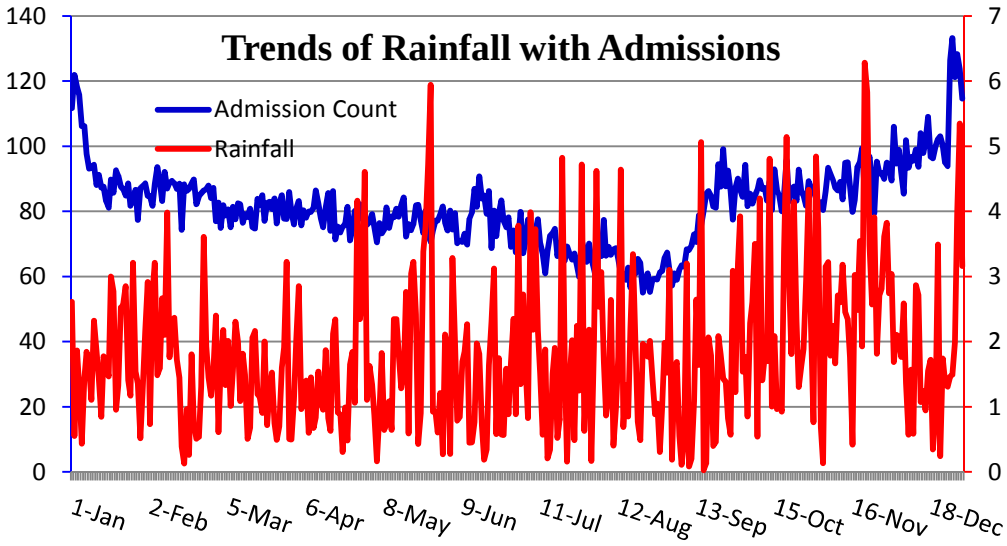

Figure 13: Trends of rainfall with LR admissions counts

(disease categories and hospital admissions data) and regional effects of the north latitude. Since we have considered lower respiratory diseases, we have a huge portion of asthma cases which increases due to increase of pollen in the air during autumn (September-November). In addition to this, because of NHS

administrative & logistic delay, it takes time to get admitted in the hospital and we are also missing the primary outcomes of any sudden climate change since we are not covering GP data (please see limitations of the thesis at section 8.4). Moreover, because of the nature of the north latitude historically winter appear more extremely than compared to summer.

No visible trend is apparent in the rate of change of Rainfall (**Figure 13**).Wind Speed (**Figure 14**) shows an increasing trend at the beginning of autumn (September-November) with increasing emergency LR admissions during the same period. Like the case of temperature, low rate of daily sun hours (**Figure 15**) and daily radiation (**Figure 16**) showed a reciprocal trend in LR admissions and vice versa.

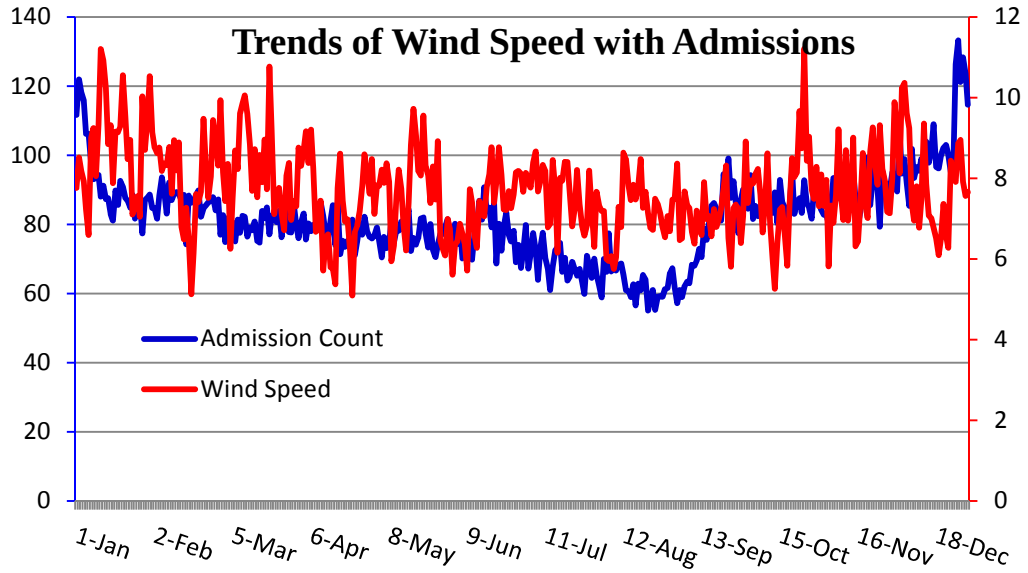

Figure 14: Trends of daily mean wind speed with LR admissions counts

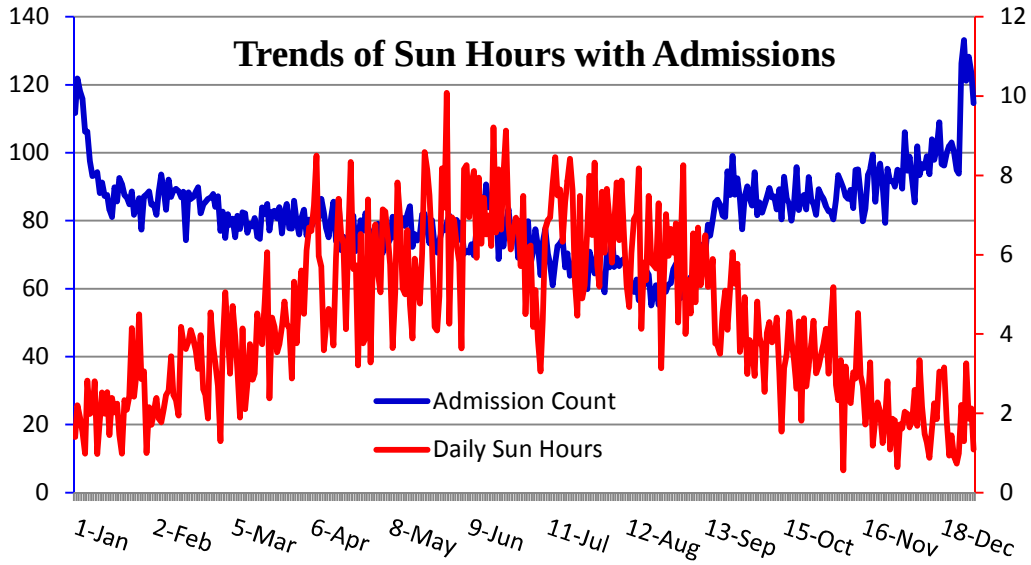

Figure 15: Trends of daily sun hours with LR admissions counts

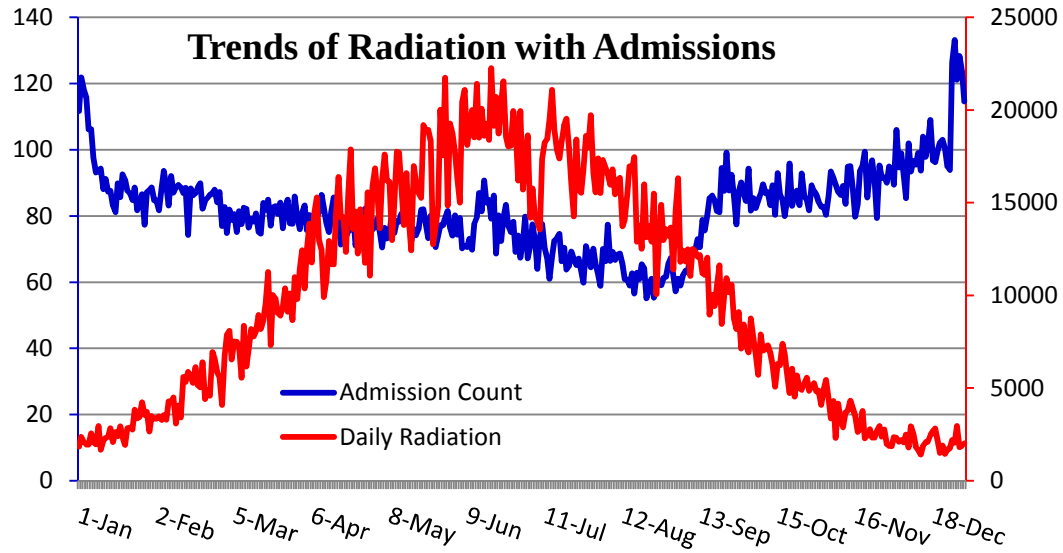

Figure 16: Trends of daily radiation with LR admissions counts

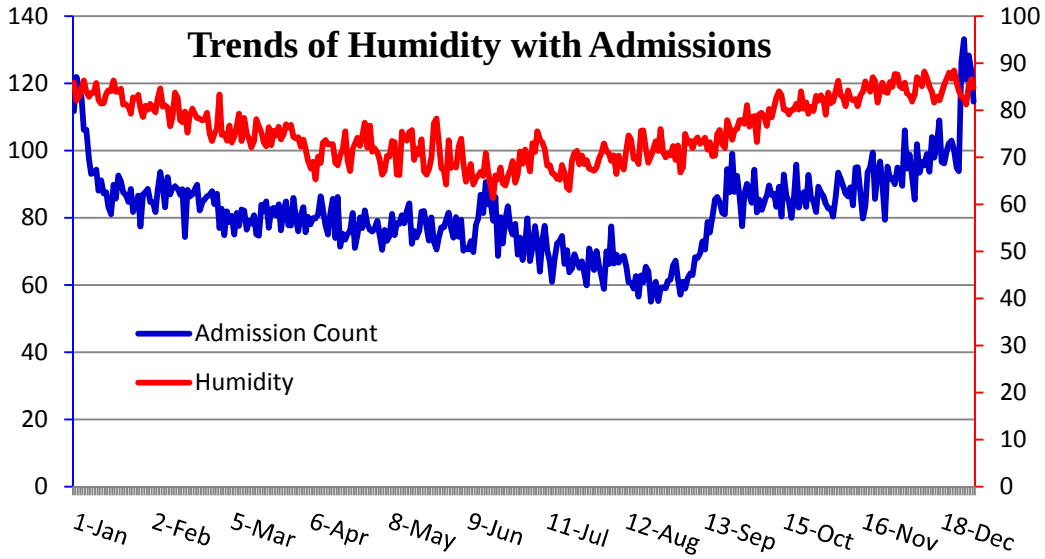

Figure 17: Trends of mean relative humidity with LR admissions counts

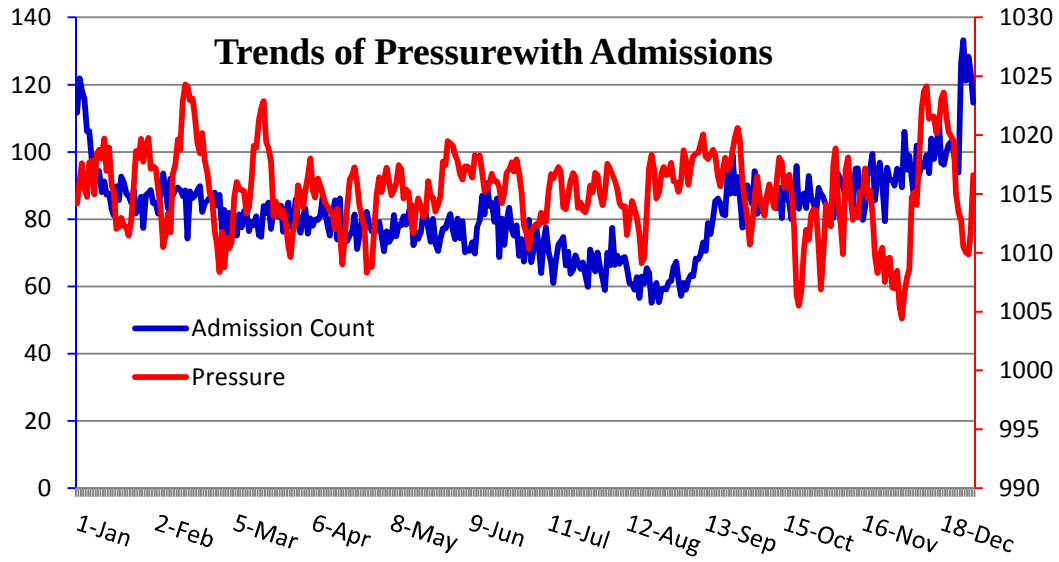

Figure 18: Trends of mean pressure with LR admissions counts

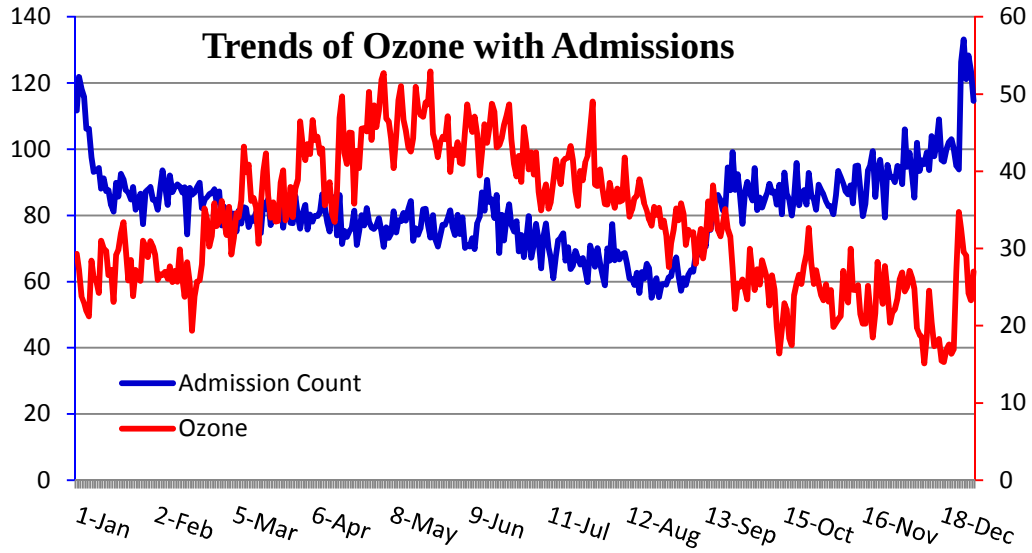

Figure 19: Trends of daily ozone with LR admissions counts

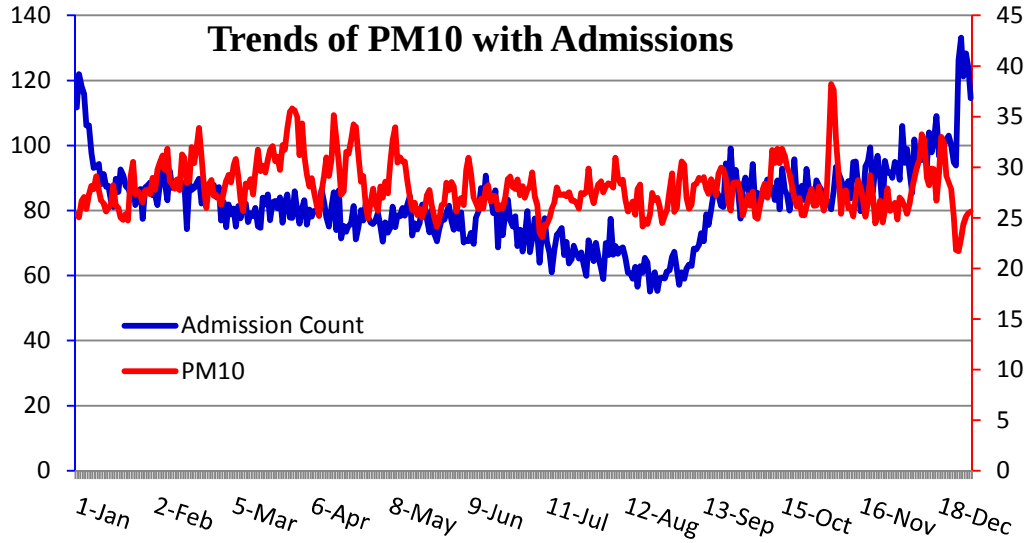

Figure 20: Trends of daily PM10 with LR admissions counts

After visual inspection of Humidity (**Figure 17**) and Pressure (**Figure 18**), no obvious trends are apparent with emergency LR hospital admissions. Similar can be said for Ozone (**Figure 19**) and PM10 (**Figure 20**), except Ozone seems to be higher during spring (March -May). However all of these factors show some spikes in their trends throughout the year.

7.3 Results of the final model

We developed a Distributed Lag Non-linear Model (DLNM), to fit the same dataset with the same climate and pollutions variables used in the exploratory data analysis (section 7.2) and Generalized Linear Model (section 5.6).

The objectives of developing the DLNM model are:

- To justify the impact of Lag-period on the emergency lower respiratory (LR) disease,
- Decide on the precise the structure of the lag-period for different climate and pollutant variables, capturing the non-linear nature of the data by introducing appropriate smoothing techniques.
- To check whether the DLNM fits the data better than the GLM model presented above (section 5.6).

The analysis of the DLNM is based on the model in 6.3.5, fitted through a generalized linear model, and by considering the concept of cross basis and one basis. We considered the quasi-Poisson family in the GLM to deal the over dispersed nature of the data. Along with the climate and pollution variables, we also used a natural cubic spline of '**time'** with 7 degrees of freedom per year (roughly equivalent to a two month moving average). This will allow adequate control for unmeasured confounders (for example long-time trends, seasonality, health related behaviour, diet), while leaving sufficient information from which to estimate the effects of climate and pollutants. From the literature review, we found '**the day of the week**' also affects hospital admissions (**Figure 21**). Thus we also include the 'Day of the week' as a variable in the model.

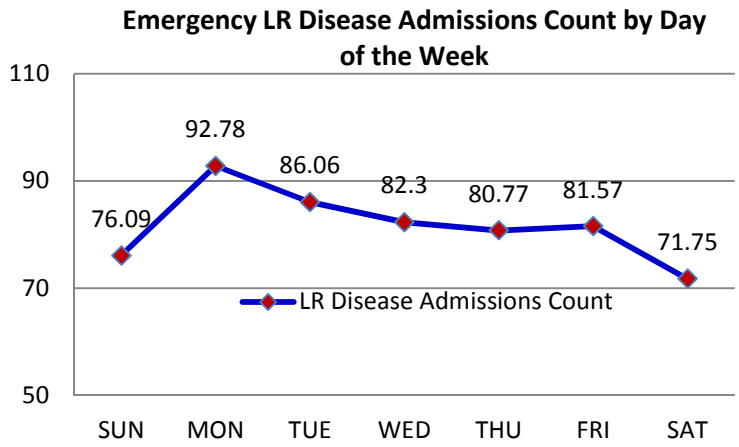

Figure 21: Lower respiratory disease admissions counts by day of the week

The choice of lag period

In the basic formulation, the Distributed lag non-linear model (DLM) is fitted by the inclusion of a parameter for each lagged predictor occurrence (Gasparrini 2011). An estimate of the overall net association is given by cumulating the single lag contributions upon the whole lag period, usually a-priori defined (Schwartz 2000; Hajat, Armstrong et al. 2005). According to Gasparrini (2011), this unconstrained version of DLNM does not require any assumptions of the shape of the association along lags, and consequently on the relationship between the parameters. However, in order to define a more parsimonious model, it is possible to specify some assumptions on the shape of the distributed effects, applying some constraints. The simplest solution is to group the lags in different strata, while a more complex options to force the curve along lags to follow a specific smooth function, for example polynomials or splines.

Table 18: Choice of lag period, variable basis, and lag basis.

Variable Name	Lag Period (days)	Basis for Variable	Basis for Lag
Temperature	30	B-Spline with degree 3 and 5 df	Natural Cubic Spline (ns) with degree 3
Rain	15	B-Spline with degree 3 and 3 df	Polynomial with degree 5
Wind Speed	20	B-Spline with degree 3 and 5 df	Polynomial with degree 3
Sun Hours	20	B-Spline with degree 3 and 5 df	ns with degree 3
Relative Humidity	20	B-Spline with degree 2 and 3 df	ns with degree 3
Pressure	10	B-Spline with degree 3 and 3 df	Polynomial with degree 4
Ozone	30	B-Spline with degree 3 and 10 df	ns with degree 3
PM10	30	B-Spline with degree 3 and 10 df	ns with degree 3

df: degrees of freedom

In our research, the choice of the lag period varied for various climate and pollution factors. We decided the lag period from the literature review and provided the maximum plausible days as lag for all the variables (**Table 18**) to improve the precision of the DLNM model. For example, lower temperature normally shows longer impacts on disease outcome than higher temperature (Hajat, Armstrong et al. 2005; Muggeo and Hajat 2009; Bhaskaran, Hajat et al. 2010). Thus, we adopted a longer 30 days lag period in the model, to cover both the effect of high (summer) and low temperature (winter). In general, the choices of the lag period and spline in **Table 18** are mainly motivated by several methodological and substantive papers on time series analyses in similar applications (Armstrong 2006; Gasparrini, Armstrong et al. 2010; Gasparrini 2011).

Smoothing techniques adopted

In our research, we used B-spline basis for all the variables used in the model instead of natural splines (ns), since natural spline tends to have a linear pattern before and after the boundary knots. So after boundary knots, ns is completely misleading if the data shows more non-linearity. In contrary, B-spline, is more data driven and takes the form according to the data as well. It also works well after the boundary knots. We decided the lag period from the literature review and provided the maximum plausible days as lag for all the variables (**Table 18**). The degree of the polynomial and degrees of freedom for all the variable basis and lag basis are based on the results of exploratory data analysis (as illustrated in section 7.2); previous studies from the literature, and also judging by the AIC/BIC results tested under various values of degrees of freedom (or knots) and degree of polynomials.

Model fitting results

The results of the DLNM model fitting are not possible to describe in a usual way (for example, **Table 16**), simply because of the complexity of the model and its complicated non-linear nature.

Interpretations of DLNM results

DLNM can be interpreted by building a grid of predictions for each lag and for suitable values of the predictor (e.g., Temperature, Rainfall, PM10), using three dimensional plots to provide an overall picture of the association varying along the two dimensions (Gasparrini 2011). In addition, it is possible to summarise the relationship at single predictor or lag values, by cutting a “slice” of the grid along specific values. These summaries express a lag-specific association, defined along

the predictor space at a given lag value, or a predictor-specific association, defined along the lag space at a given predictor values, respectively. Finally, an estimate of the overall cumulative association can be computed by summing all the contribution at different lags for each predictor value. The associations are usually reported versus a reference value of the predictor, centring the basis functions for the space to their corresponding transformed values. For our analysis, we consider the reference values around the corresponding mean of each of the predictor.

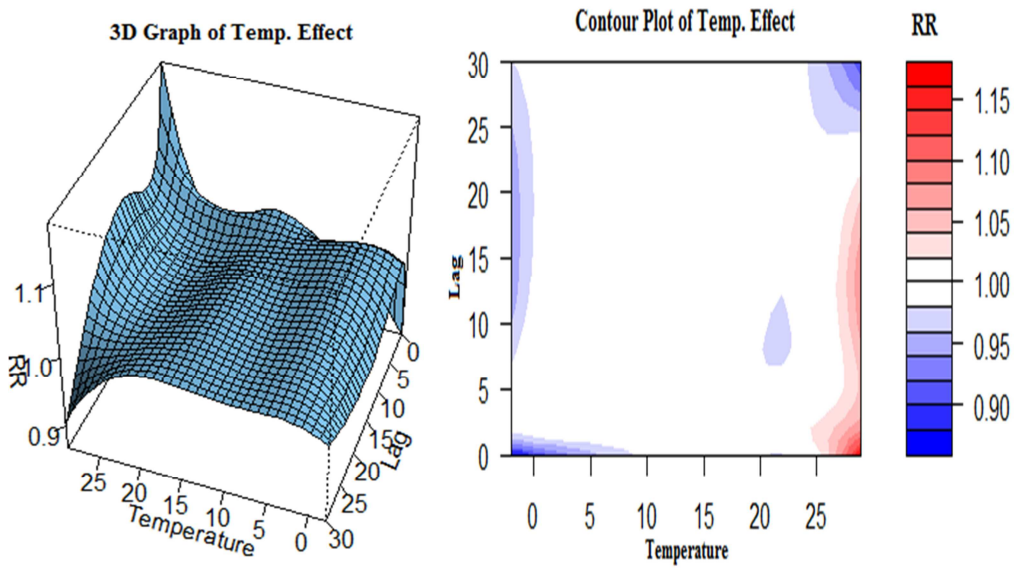

Figure 22: 3D & Contour plot of RR along temperature and lags, with ref. at 12^0C

Figure 22 shows the overall depiction of **temperature** on lower respiratory diseases admissions. Here, it illustrates a 3-D image and corresponding contour plot of the relative risk (RR) along the mean temperature (here 12^0C) and lags. The plot shows a very strong immediate effect of the higher temperature at

around $\geq 27^0$C and lag period of 0-2 days. Higher temperature also seems to have an effect on emergency LR admissions at around 10-15 days lag period. Lower temperature (e.g. 0^0C) seems to have a moderate effect at around 5-25 days lag period.

Figure 23 (top graph), illustrates lag specific associations of different temperatures (-1, 0, 20, 29) ranging from lower to higher temperatures, with reference at 12^0C. One observes that higher temperature has an immediate effect on admissions and longer lag effects up to 2 days and later on a longer effect of 10-15 days lag period. **Figure 23**, also depicts both associations along the predictor range at lag 0 and lag 25 (left column) and associations along lag at temperatures -1^0C and 29^0C. The interpretation of **Figure 24** is twofold: the top curve represents the increase in risk in each future day following an increase of 10^0C in a specific day (forward interpretation), or contributions of each past day with the same temperature increase to the risk in a specific day (backward interpretation). Note that initial increase in risk due to temperature is up to 5 days lag period and then increase of longer days lag of 25 days or over. We also observe the overall cumulative association with a 10^0C over 30 days of lag (summing all the contributions up to maximum lag), together with the 95% confidence interval.

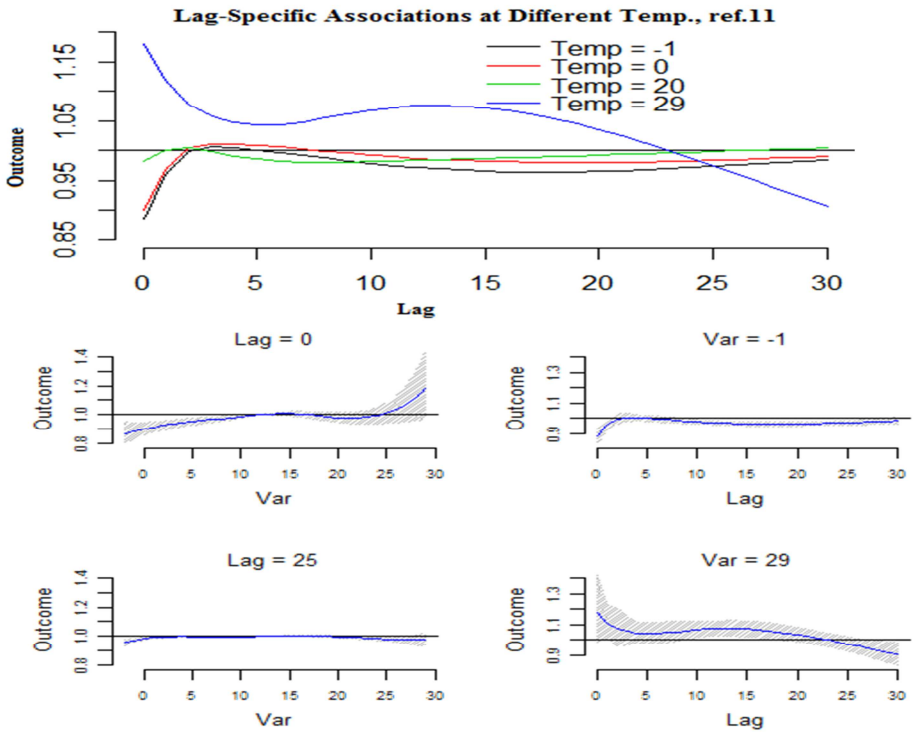

Figure 23: Lag-Specific association at different temperature and lags, ref 12^0C

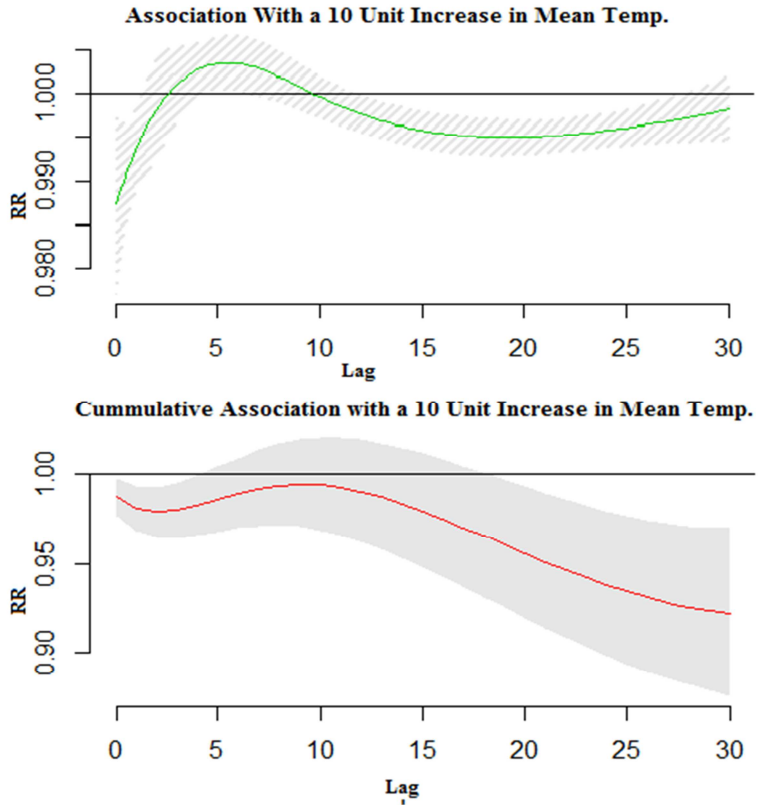

Figure 24: Specific and cumulative association of a 10 unit increase in mean temperature

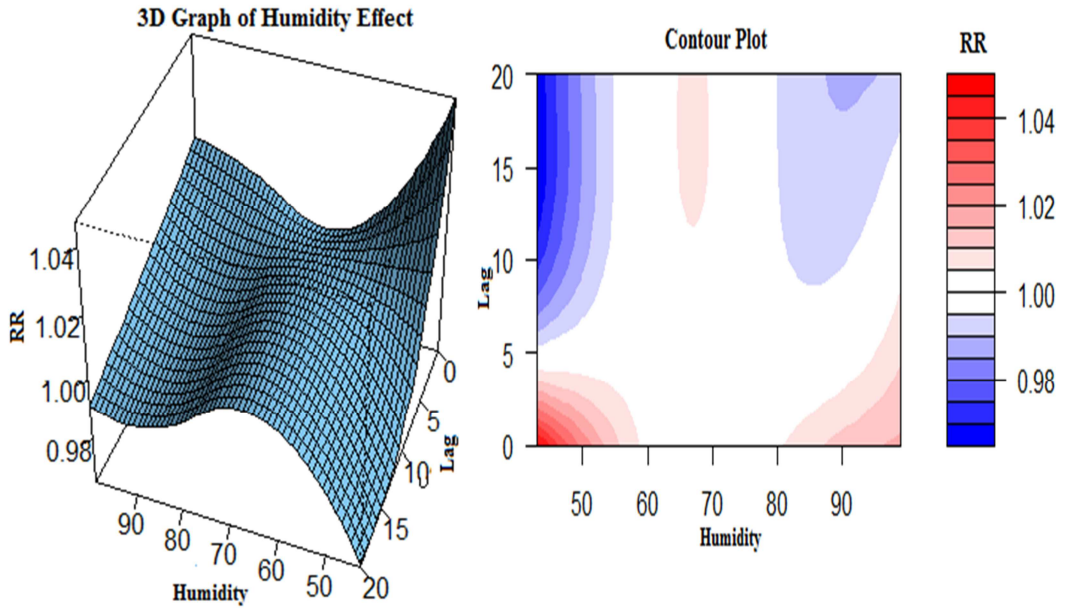

Figure 25: 3D & Contour plot of RR along R.humidity and lags, with ref. at 75.8%

Figure 25 illustrates an overall relationship of **relative humidity** on lower respiratory disease admissions. Both higher and lower humidity show a shorter lag period effect on the emergency LR admissions. The 3-D graph and corresponding contour plot of the relative risk (RR) along the relative Humidity and lags compared with a reference value of 75.8%, shows a very strong immediate effect of the lower relative humidity at around 40% and a lag period of 0-3 days. Similarly, higher relative humidity (80% or more) also seems to have a moderate effect on the admissions at short lag period of 0-2 days.

Figure 26 shows lag specific associations of different relative humidity (45, 60, 85, and 95) % ranging from lower to higher, with reference at 75.8%. We can see that both lower and higher relative humidity (45% and 95%, respectively) have quicker effect on LR admissions and thus shorter lag periods. **Figure 26** also

depicts both associations along the predictor range at lag 0 and lag 15 (left column) and associations along lag at relative humidity 45% and 95%.

The interpretation of **Figure 27** is twofold: the first curve represents the increase in risk in each future day following an increase of 20% relative humidity in a specific day (forward interpretation), or contributions of each past day with the same relative humidity increase to the risk in a specific day (backward interpretation). We can see the initial increase in risk due to relative humidity is up to 3 days lag period. We also observe the overall cumulative association with a 20% increase of relative humidity over 10 days of lag (summing all the contributions up to maximum lag), together with the 95% confidence interval. We can see that cumulative association of relative humidity has a longer term effect of up to 10 days.

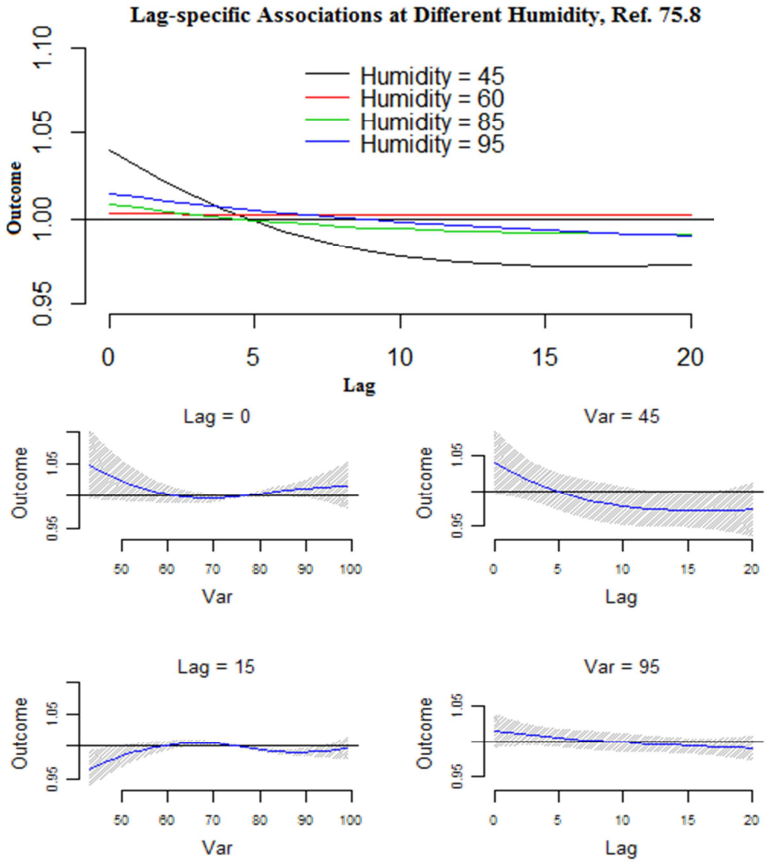

Figure 26: Lag-specific association at different R.humidity and lags, ref 75.8%

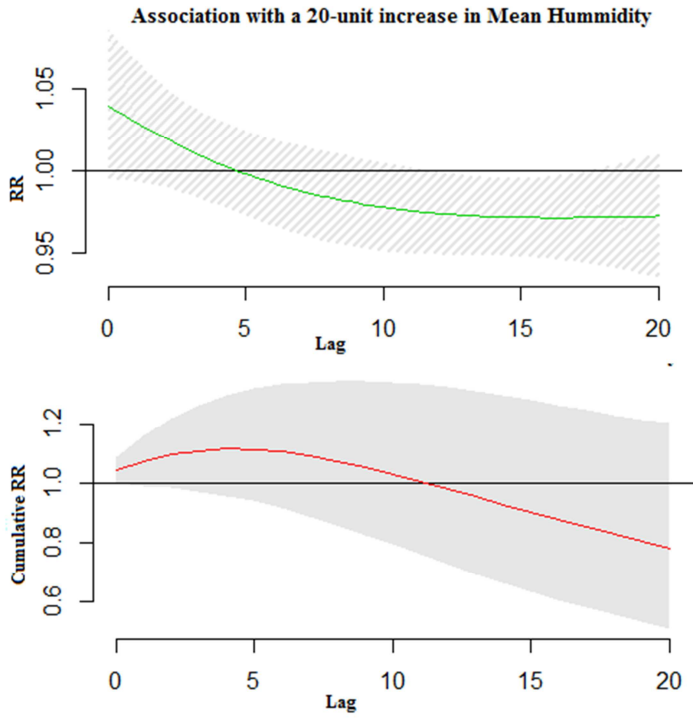

Figure 27: Specific & cumulative association of a 20 unit increase in R.humidity.

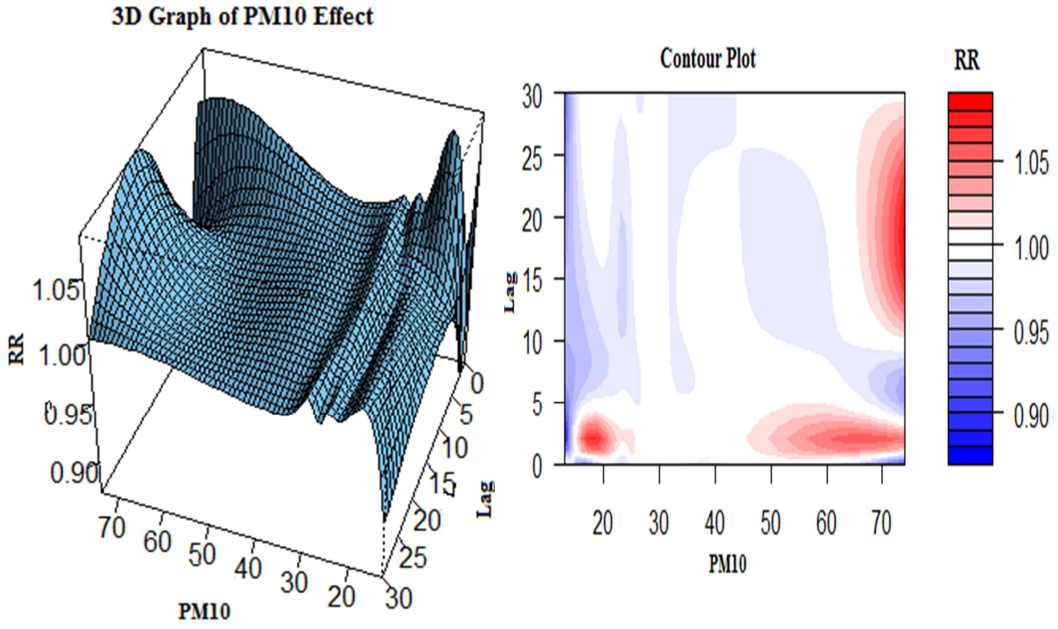

Figure 28: 3D & Contour plot of RR along PM10 and lags, with ref. at 28µg/m^3

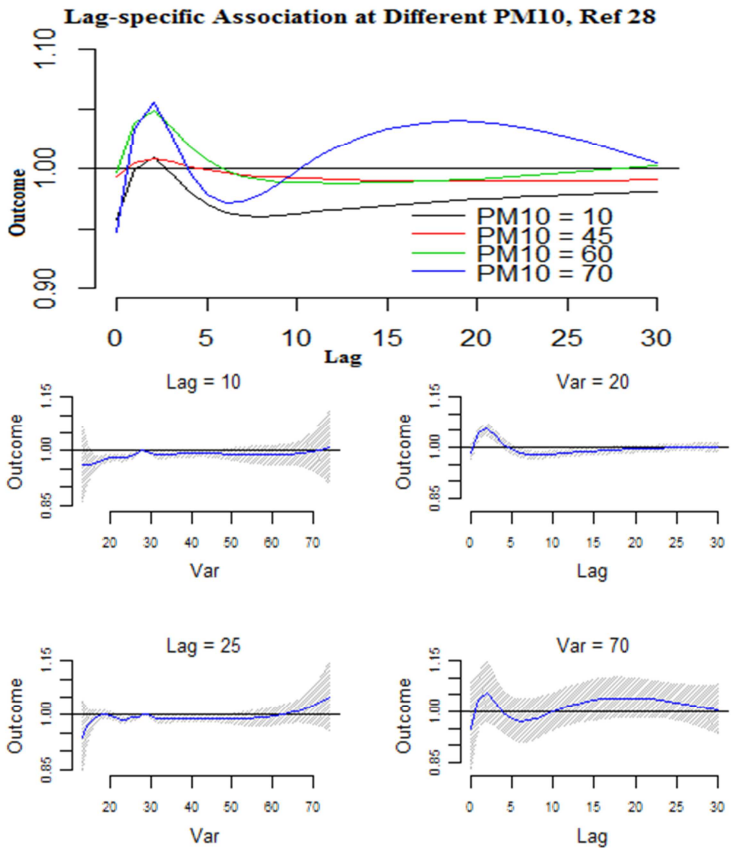

Figure 29: Lag-specific association at different PM10 and lags, ref 28µg/m^3

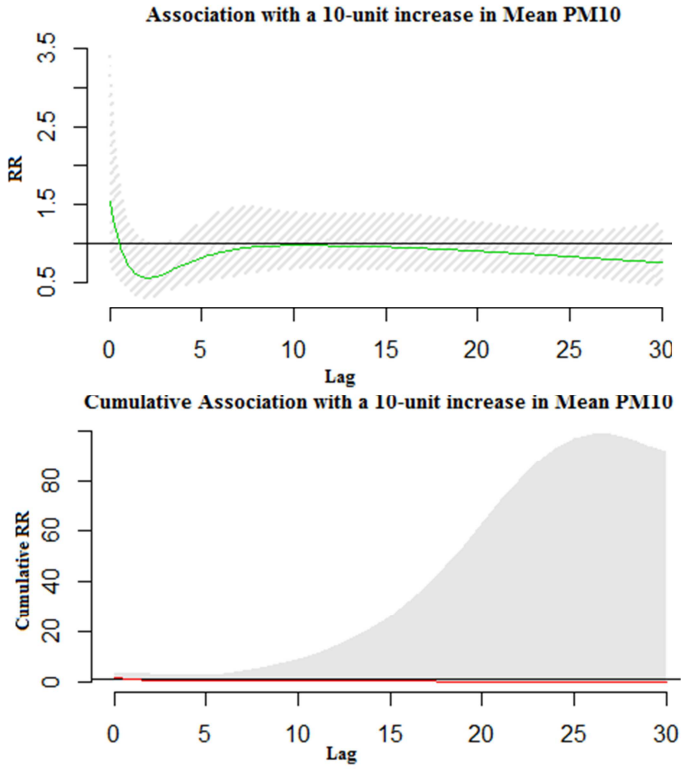

Figure 30: Specific and cumulative association of a 10 unit increase in PM10.

The higher **PM10** shows both longer and shorter lag effects on the emergency LR hospital admissions. The 3-D graph and corresponding contour plot (Figure 28) of the relative risk (RR) along PM10 and lags compared with a reference value of 28µg/m^3, illustrates a strong effect of higher PM10 around 70-µg/m^3 or higher, and longer lag period of 15-20 days. It also shows some immediate short-term effects of 0-3 days lag period. However, the effect related to the longer lag period of 15-20 days of 70-µg/m^3 or more PM10 is comparatively solid (stronger).

Figure 29 (first) illustrates lag specific associations of different PM10 values (10, 45, 60, and 70 µg/m^3) ranging from lower to higher PM10, with reference at 28-µg/m^3. We can see that higher PM10 (blue line) has very quick effect on admissions up to 3 days lag and afterwards longer affects up to 15-20

days lag period. **Figure 29** (second) also depicts both associations along the predictor range at lag 10 and lag 25 (left column) and associations along lag at PM10 20 and 70-µg/m^3. The interpretation of **Figure 30** is twofold: the first curve represents the increase in risk in each future day following an increase of 10-µg/m^3 PM10 in a specific day (forward interpretation), or contributions of each past day with the same PM10 increase to the risk in a specific day (backward interpretation). We also observe the overall cumulative association with a 10-µg/m^3 PM10 over 30 days of lag (summing all the contributions up to maximum lag), together with the 95% confidence interval.

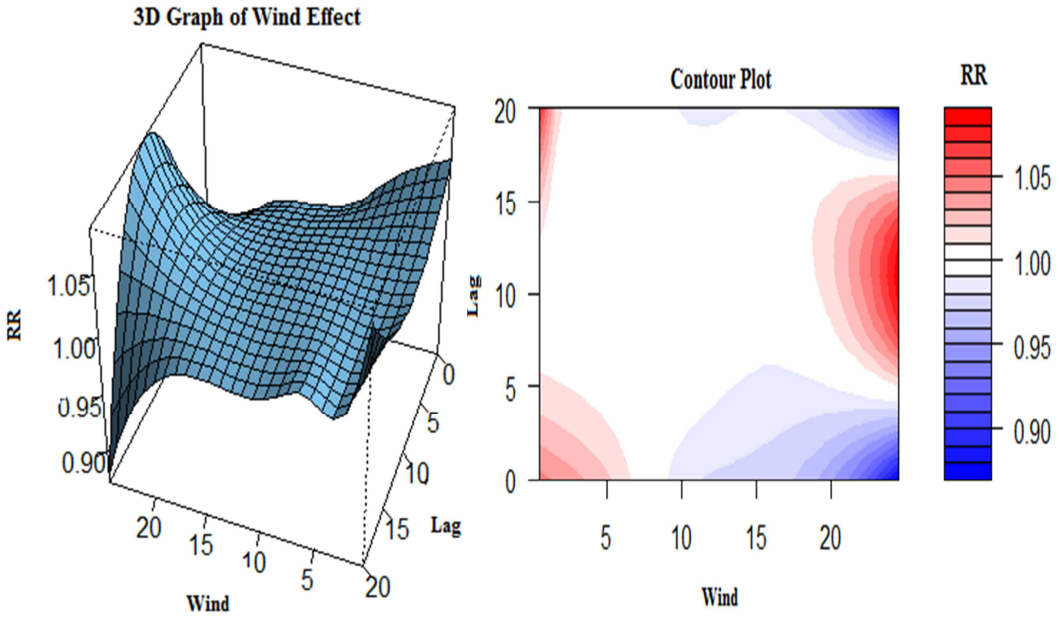

Figure 31: 3D & Contour plot of RR along wind speed and lags, with ref. at 7.7 knots

The relative risk (RR) along **wind speed** and lags compared with a reference value of 7.7 knots, illustrates a strong effect of wind speed around 25 knots or higher, and longer lag period of 8-15 days. At the same time, it shows

moderate effect for a shorter lag period of 0-3 days for lower wind speed (approximately 2 knots). This is shown in the 3-D graph and the corresponding contour plot of wind speed in **Figure 31**.

Figure 32 (first) illustrates lag specific associations of different Wind Speed values (0.5, 4.5, 15.5, and 24.5) knots ranging from lower to higher Wind Speed, with reference at 7.7 knots. We can see that lower wind speed shows moderate effect for shorter day's lag of 0-3 days but higher wind speed (blue line) has delayed effect on admissions up to 8-12 days lag. Figure 32 (second) also illustrates both associations along the predictor range at lag 0 and lag 15 (la wind speed) and associations along lag at Wind speed of 0.5 and 24.5 knots.

The interpretation of **Figure 33** is twofold: the first curve represents the increase in risk in each future day following an increase of 10 knots of Wind Speed in a specific day (forward interpretation), or contributions of each past day with the same Wind Speed increase to the risk in a specific day (backward interpretation). We also observe the overall cumulative association with a 10-knots over 20 days of lag (summing all the contributions up to maximum lag), together with the 95% confidence interval.

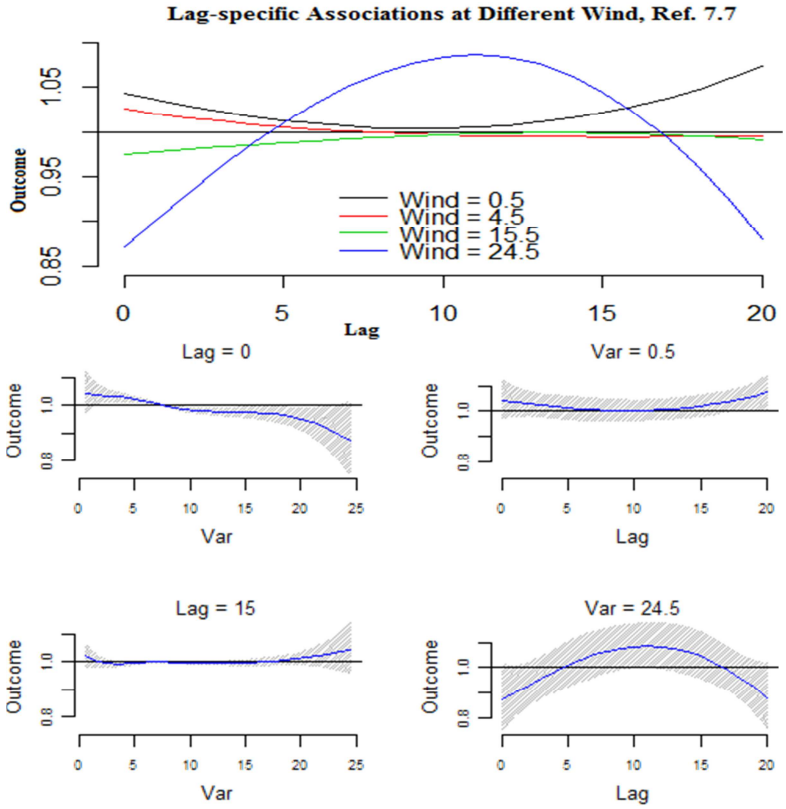

Figure 32: Lag-Specific association at different wind speed and lags, ref 7.7 knots

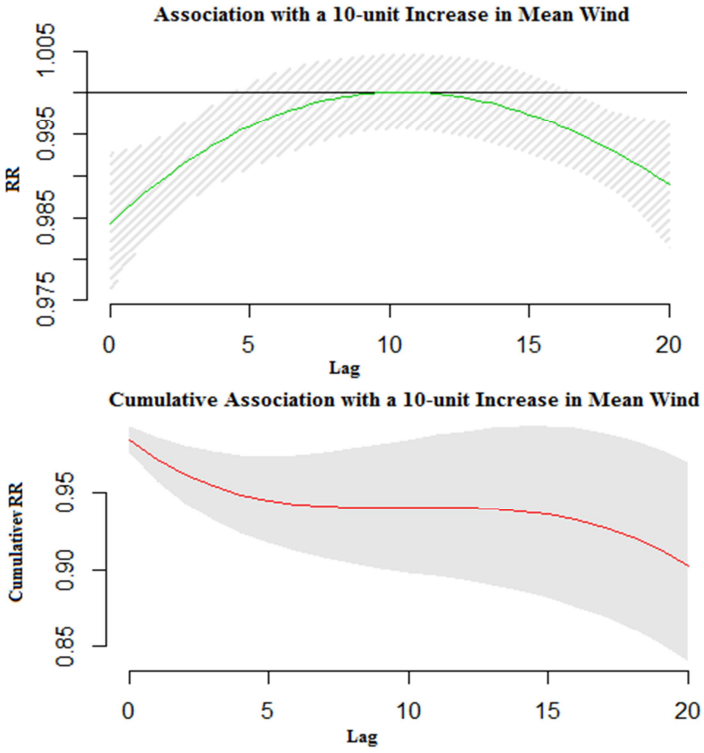

Figure 33: Specific and cumulative association of a 10 unit increase in wind speed

The predictor and lag specific illustrations for the relationships of **sun hours** can be found in **Figure 34** through **Figure 36**. The 3D relationships of Sun Hours and its lag with the emergency LR disease admissions is described in **Figure 34**, followed by lag-predictor specific relationships in **Figure 35** and the specific and cumulative associations of the effect of a 1-hour increase of Sun Hours in admissions in **Figure 36**. These results are based on considering the reference sun hours as 4.4 hours. We can observe a stronger effect of sun hours around 14 hours or more having a longer lag period of 15-20 days and moderate effect between 1-2 hours of 5-12 days lag.

Similarly, we can also see the images of the relationships of **rain** in 3-D and Contour plots in **Figure 37**, Lag and predictor specific in **Figure 38** and finally the specific and cumulative association of Rain and admissions in **Figure 39**. The reference value of rain is 8.8 mm. We can see that higher amount of rain of 30mm or more has a stronger effect on emergency LR hospital admissions, especially for the shorter lag of 0-2 days and longer lag of 7-10 days. The summary of the results from the final model is described in **Table 19**.

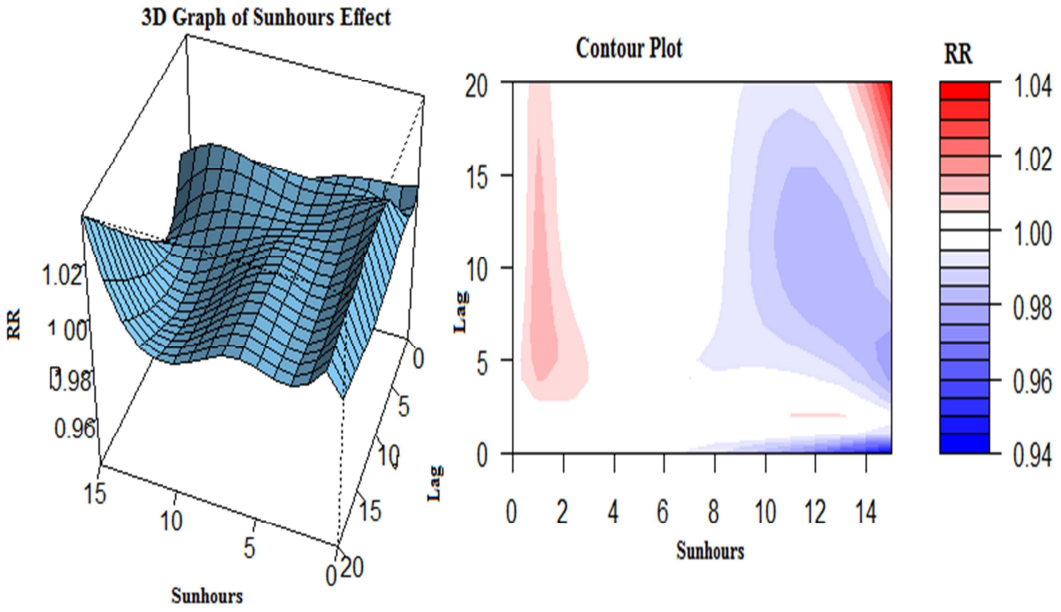

Figure 34: 3D & Contour plot of RR along sun-hours and lags, with ref.at 4.4 hours

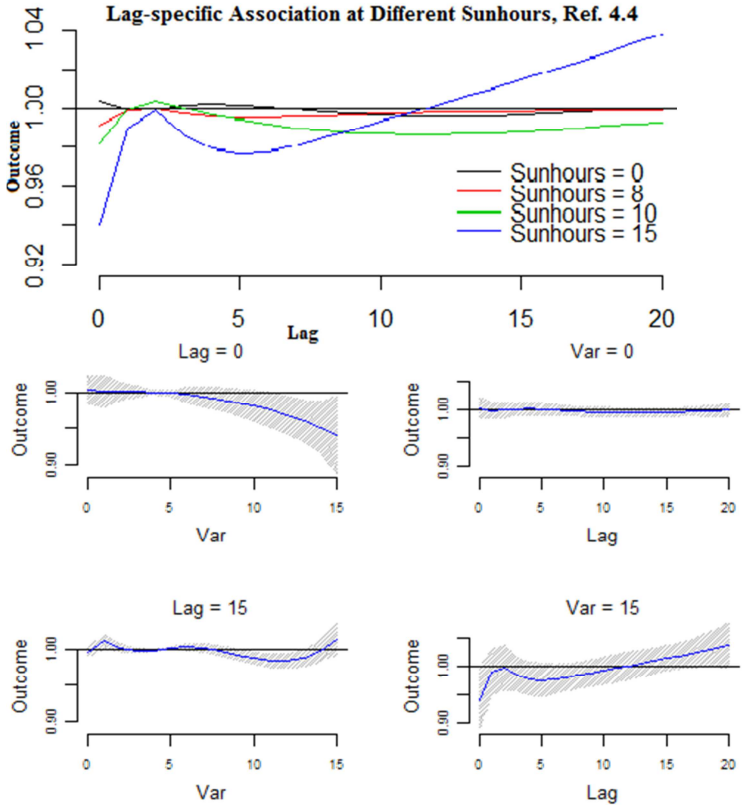

Figure 35: Lag-Specific association at different sun-hours and lags, ref 4.4 hours

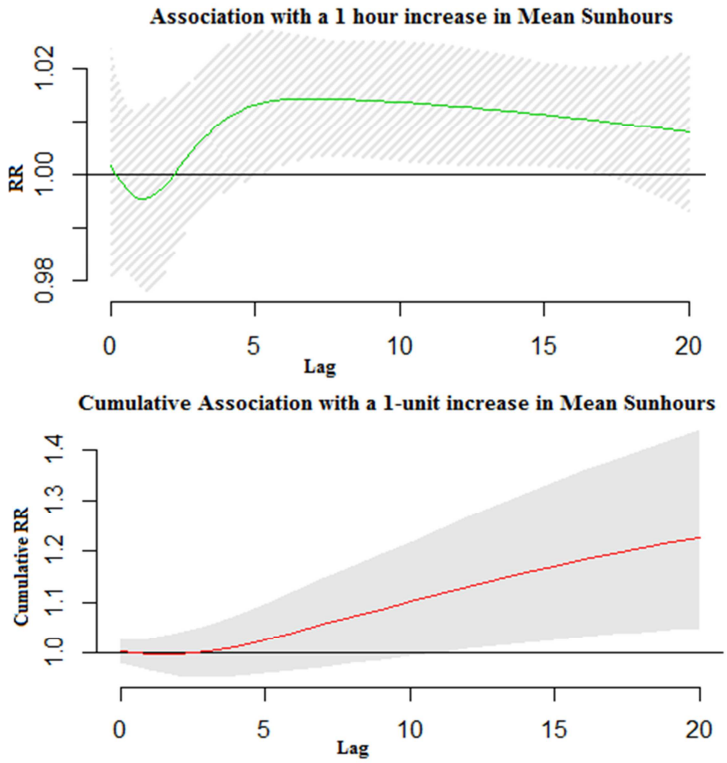

Figure 36: Specific and cumulative association of a 1-hour increase in sun-hours.

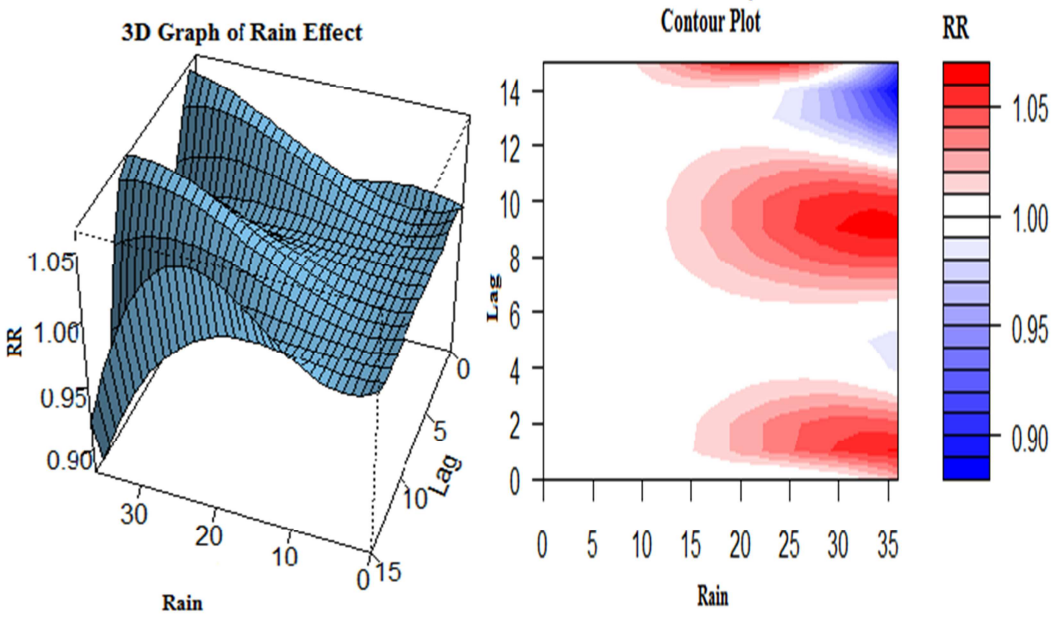

Figure 37: 3D & Contour plot of RR along rain and lags, with ref. at 1.8 mm

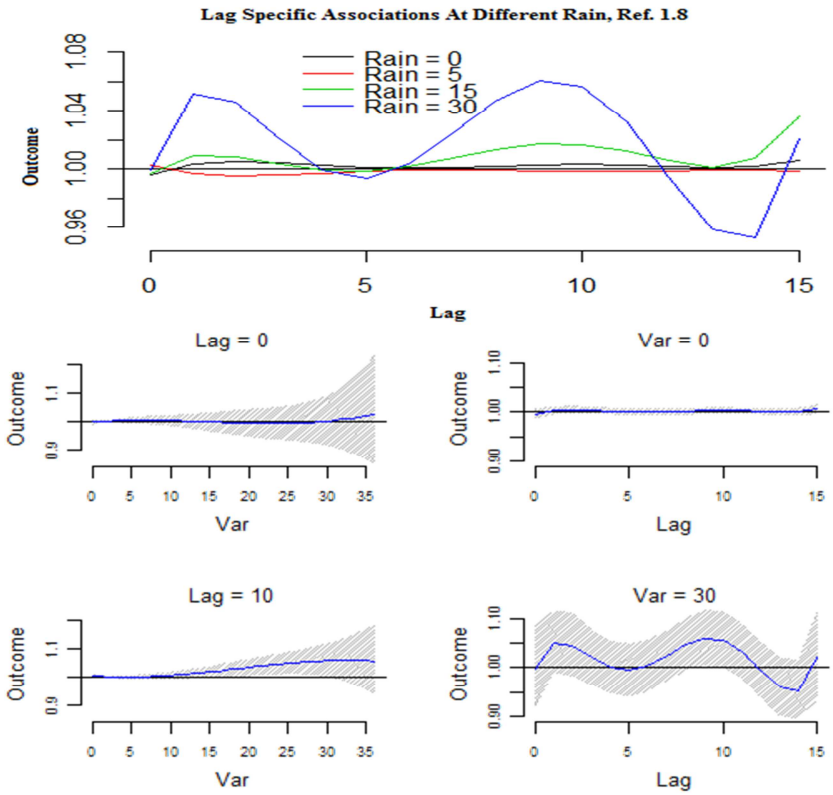

Figure 38: Lag-specific association at different rain and lags, ref 1.8-mm

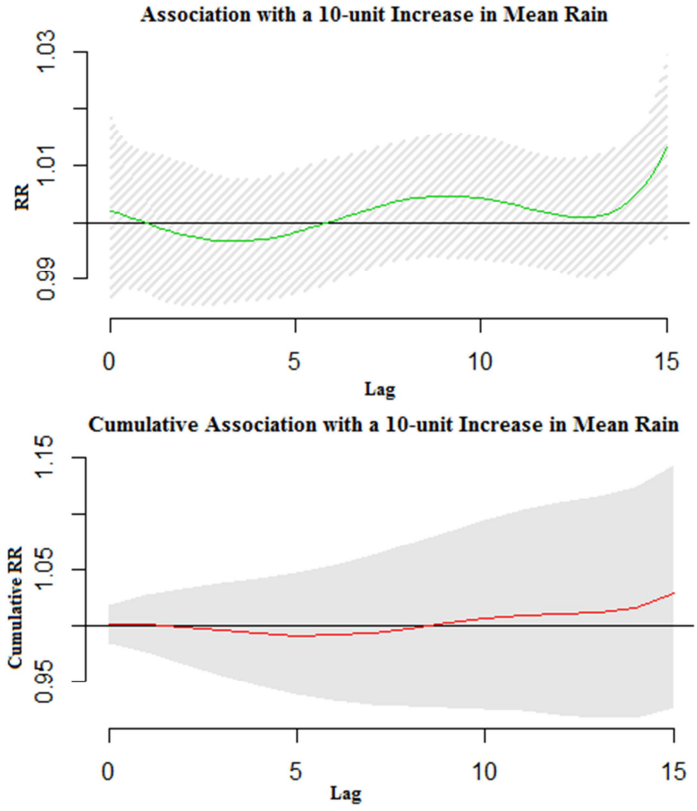

Figure 39: Specific and cumulative association of a 10 unit increase in rain.

Table 19: Climate threshold from the final model

Factors	High or Low (daily average)	Lag period	Strong or moderate
Temperature***	High ($\geq 27^0$C)	0-2 days	Strong
	Low ($\leq 0^0$C)	5-25 days	Moderate
Relative Humidity***	High ($\geq$ 40%)	0-2 days	Moderate
	Low ($\leq$ 40%)	0-3 days	Strong
PM10*	High ($\geq$70-µg/m^3)	0-3 days	Strong
	High ($\geq$70-µg/m^3)	15-20 days	Moderate
Wind Speed***	High (($\geq$25 knots)	8-15 days	Moderate
	Low ($\leq$ 2 knots)	0-3 days	Strong
Sun Hours ·	High (($\geq$ 14 hours)	15-20 days	Strong
	Low (1-2 hours)	5-12 days	Moderate
Rain**	High ($\geq$30mm)	0-2 days or 8-10 days	Strong
	Low (20-25 mm)	0-2 days or 8-10 days	Moderate

Statistically significant at 0.1 %(***), 1 %(**), 5 %(*), 10%(**.**)

We also calculated similar type of relations for other climate and pollution variables but not describe here since no significant relations were found with emergency LR disease admission counts (**Table 16**).

7.4 Model comparison

We compared the models based on the modified Akaike and Bayesian information criteria for models with over dispersed responses fitted through quasi likelihood (Hastie and Tibshirani 1990; Wood 2006), given by:

$$QAIC = -2\mathcal{L}(\hat{\theta}) + 2\hat{\phi}k$$
$$\text{and } QBIC = -2\mathcal{L}(\hat{\theta}) + log(n)\hat{\phi}k \quad (6.22)$$

Where $\mathcal{L}$ is the log-likelihood of the fitted model with parameters $\hat{\theta}$ and $\hat{\phi}$ the estimated overdispersion parameter, whereas k and n are the number of parameters and the number of observations, respectively. The best model is chosen that minimises the above criteria.

Table 20: Model comparison results

Models	**Model Form**	**Model Name**	**QAIC**	**QBIC**	**Nagelkerke R-squared**	**Improved (YES / No)**
Model1 (Originally Model 13 of **Table 14**)	Count ~ Temp + Rain + Wind + Sunhours + Humidity + Pressure + Ozone + PM10	GLM	42166.34	42468.72	0.5914073	
Model2	Count ~ (All the variables in the above model in cross-basis form for DLNM)	DLNM	33997.68	37991.16	0.9674209	Yes
Model3	Model 2 above + **Time** (natural smoothing)	DLNM	31845.22	36075.62	0.9833345	Yes
Model4	Model 2 above + Time (natural smoothing) + **Day of the week**	DLNM	30318.68	33933.83	0.9890791	Yes

From the model comparison results presented in **Table 20**, the Distributed lag non-linear model with the variables: daily mean temp, daily mean rainfall, daily wind speed, daily sun hours, daily relative humidity, daily pressure, daily Ozone, daily PM10, time, and 'day of the week' gave us the best fit (lowest QAIC and QBIC) for lower respiratory disease counts in the Greater London for the year 2000-2009. All the models in **Table 20** are compared based on the results of modified Akaike information criteria (QAIC), modified Bayesian information

criteria (QBIC) and Nagelkerke R-squared. The Nagelkerke R-squared of the final model (model 4 in **Table 20**) is 0.989079 which is a very good indication for goodness of fit. This means that 98.91% of the variation in the response variable (emergency LR admissions counts) can be explained by the explanatory variables (the variables in the final model). The remaining 1.1% can be attributed to unknown, inherent variability.

7.5 Chapter summary

In summary, we can conclude that the idea of Distributed Lag non-linear model, i.e. considering both the current and delayed impact of the predictors on the response variable improves the fit of the data dramatically. For example, we found that the final DLNM model gives the best results according to the Nagelkerke R-squared measurement. And Temperature, Rain, Wind Speed, Sun Hours, Relative Humidity, and PM10 have significant impact on lower respiratory disease admission count with some delayed effects (shorter and longer). Thus DLNM also provides new insights about the coverage of the lag structure and lag period. For example, one observes that higher temperature has an immediate effect on admissions during the summer, whereas in the winter period, due to lower temperature, it shows longer lag effects of up to 20 days. The next chapter concludes the thesis and summarises our thoughts and scope for future works.

Chapter 8

Conclusion and further works

8.1 Summary and conclusions

This thesis, for the first time, focused on the delayed effect of both meteorological and pollution variables on hospital admissions. We considered hospital admissions for lower respiratory diseases since the literature review revealed it as the most climate affected disease category. This study is also specific in terms of using Hospital Episode Statistics (HES) and London Air Quality Network (LAQN) data for Greater London and linking them in one common platform. The main motivation of this research has been the development of statistical models to capture delayed and non-linear effects of climate and pollution variables.

Towards achieving the general and specific objectives of the thesis (section 1.2 and 1.3), we first started a systematic review to explore the current and recent studies and related gaps in the research of climate change and health. The review illustrated some crucial concerns and research gaps in this area.

We linked three administrative data sets: HES, meteorological, and air pollutants into one platform. We observed the results of some case studies (Section 5.5). There we found that considering only temperature in the model is not enough for better model fits and more climate factors along with their delayed effects might provide better modelling results. To deal with the research gaps and

input from the systematic review, we proceeded first by employing the generalized linear model to select the statistically significant factors of climate and pollutants on health exposures (section 5.6). The GLM showed statistically significant relationships of daily mean temperature, wind speed, sun hours, relative humidity, and pressure, and pollution variables: Ozone and PM10 on the daily emergency lower respiratory hospital admissions. However, judging by the generalized R^2 statistics the model fit was poor, even though the model diagnostics results were reasonable.

We performed an exploratory data analysis to check the overall trends and seasonality of the climate and pollution variables (section 5.6.1 and section 7.2). Most of the variables seem to have non-linear relationships with emergency LR disease admissions counts. Quadratic and cubic trends were apparent between some of the factors and the admission counts in the exploratory data analysis. To capture the variations of such non-linear trends and the assumed delayed impact of the climate and pollution variables, we developed the final DLNM model (section 6.3.5) with the same variables that emerged as significant in the GLM (section 5.6). This new approach enabled us to tackle an important gap in research related to non-linearity and the delayed effect of climate factors on health. All the climate and pollution factors showed various delayed effects on LR emergency hospital admissions and the B-Spline was the most plausible smoothing function (**Table 18**). From the results of the final model (**Table 19** and **Table 20**), we can conclude that if we have days with high temperature ($\geq 27^0$C), low relative humidity ($\leq$ 40%), High Pm10 level ($\geq$70-$\mu g/m^3$), low wind speed ($\leq$ 2 knots),

and High rainfall (≥30mm), we can expect a significantly higher number of emergency lower respiratory hospital admissions in the next 2/3 days.

In the following sections, we summarised the important conclusions resulted from this thesis, followed by future directions of this research.

8.1.1 Conclusion-1: A systematic review of impact of climate change

Temperature is the most influential climate factor amongst all the variables in climate health studies. Index of climate factors in model fitting provides a better estimate than modelling with same climate factor separately. The non-linear relationships between climate and health, their delayed effect, and precise lag structure should be considered in climate research for efficient modelling to enable key decision makers develop a robust, reliable and an accurate health alert system. Elderly and children are the highest vulnerable group due to climate change.

In the systematic review, we explored papers published after the year 2000 where there main focus was on climate and pollutions factors, disease categories, and statistical methodologies applied in climate change and health. Temperature was found to be the most influential climate factors. Wind speed, humidity, rainfall, and pollution factors like PM10, ozone were also used in recent studies. They showed compounded effects on a number of disease outcomes. Index of climate and pollution are very useful factors. An index of factors has a stronger statistical significance on health than the same factors used separately. Other factors have

also been considered, such as socioeconomic and demographic factors, latitude-longitude, seasonality, race, and culture. There are spatial, and disease diversities in the impacts of climate change and the factors in climate health research should be specific to regions and diseases. Elderly, children, and patients of respiratory and cardiovascular diseases are the main risk groups due to climate change. COPD, asthma, stroke are also very frequent. Some articles in the literature review described the relationships between climate factors and its impact on health as non-linear, and argued for considering non-linearity for model optimisation. Delayed effects in climate and pollution research have seen considerable attention simply because of its impact on model fit, and thus lag period and thresholds need to be estimated accurately. Lag structures of factors are very crucial to capture both the delayed effects and non-linearity, and an efficient climate threshold can lead to an improved health alert system. Higher temperature tends to have quicker lag effect and vice-versa.

Threshold need to be specific to climate zone and disease. The impact of climate change is quite vast covering all sorts of disciplines like ecology, mathematics economics, hydrology, and so on. Thus the mathematical and statistical modelling approaches and related objectives in different areas are quite diverse. Unfortunately, the ability to make generalisations of most of the existing methodologies is very limited across time, region, and populations. There is a dire need for reliable and accurate models to capture the impact of climate change on health more precisely.

8.1.2 Conclusion-2: Administrative data in climate change research

So far there has never been an attempt to link the three data sets (HES, climate data, LAQN pollution data) to evaluate the impact of climate change and pollution on hospital admissions.

Now-a-days, administrative health care databases play a central role in measuring the exposures of health, disease, and thus evaluation of healthcare systems. Key decision makers within public and private organisations have noticed that priceless information are embedded in routinely collected data, such as HES for informed decision making purposes. Data aggregation and linkage are important steps towards improving the quality of care, explore disease epidemiology, and monitor the system changes (Miriovsky, Shulman et al. 2012) and (Barbieri, Grieco et al. 2010).

We aggregated three administrative datasets based on the date of hospital admissions and the first three characters of each patient's postcode. To the best of our knowledge, this is the first time HES has been linked to climate and pollution factors in England. It has given us the opportunity to measure the impact of both climate variables and pollutants on hospital admissions. This gives us the opportunity to measure the impact for a wide range of disease categories, which could further be investigated based on regional variation, patient types, severity, and many more. To deal with the missing values in the data aggregation, we used mean imputation (for pollution factors) and AIRGENE algorithm (for climate factors) for better representations of the original data.

8.1.3 Conclusion-3: Results from Generalized linear model

Climate change showed a compound effect on hospital admissions. Besides temperature other factors like humidity, wind speed, sun hours, rain, and Pm10 also have significant impact on the emergency lower respiratory hospital admissions.

We developed the GLM model using the climate variables: daily mean temperature, wind speed, sun hours, relative humidity, and pressure, and pollution variables: Ozone and PM10. We used ANOVA, QAIC, and QBIC to select the variables in the final model and calculated the variance inflation factor for all the variables to check for multicollinearity. As a result, radiation was removed from the model. According to our results temperature, wind speed, sun hours, relative humidity, rainfall, and PM10 were statistically associated with lower respiratory emergency hospital admissions. Interestingly, **temperature, rain,** and **sun hours showed** negative relationships with the daily admissions count, whereas **relative humidity, wind speed,** and **PM10** had a positive relationship. For example, keeping all other variables (e.g., rainfall, humidity) fixed, a unit (^{0}C) increase in the mean temperature will increase the daily emergency LR admissions count by 0.9881950 (thus decrease since less than 1). No significant effects of the changes in **pressure** and **ozone** were found on the emergency LR hospital admissions.

The Nagelkerke R-squared for the final GLM model is 0.5914073. This means that 59.14% variation of the lower respiratory hospital admissions can be explained by the considered explanatory variables in the final GLM model. The model diagnostics check (residual plots, section 5.6) results showed that the final GLM model fitted the data reasonably well.

8.1.4 Conclusion-4: Results from the final DLNM

The performance of the model fits reveals a significant improvement after considering the relationships between climate-pollution factors and health as non-linear and existence of their delayed effects. Almost all the factors (related to climate or pollution) showed their respective delayed effects and non-linearity on the emergency hospital admissions.

We performed an exploratory data analysis to check the overall trends and seasonality of the climate and pollution variables. Almost all the variables seem to have non-linear relationships with emergency LR admissions counts. Most of the variables had either a quadratic or cubic trend with daily emergency hospital admissions. To capture the variations of such non-linear trends and the delayed impact of the climate and pollution variables, we developed the DLNM model by using daily mean temperature, daily rain, wind speed, sun hours, relative humidity, pressure, ozone, PM10 along with 'time', and 'day of the week'. To smooth the non-linearity, we used the B-Spline smoothing for most of the variables because of its data driven characteristics after the boundary knots. We illustrated the delayed effect of respective factors and lag period. For instance, for days above 30^0C, we found a quicker but most eminent lag period of 0-2 days and long term moderate effect of 0-15 days. Lower temperatures (0^0C or less) exposed a mild lag period of 5-25 days. Both higher and lower, humidity showed a strong immediate effect or shorter lag period of 0-3 days, stronger for lower humidity. Higher PM10 (70-$\mu g/m^3$ or more) showed a strong effect of 15-20 days lag period compared to the mean reference value of $28\mu g/m^3$. The relative risk (RR) along

wind speed and lags compared with a reference value of 7.7 knots, illustrates a strong effect of wind speed around 25 knots or higher, and longer lag period of 8-15 days. We noticed stronger effect of sun hours around 14 hours or more with a lag period of 15-20 days, compared to the reference sun hours of 4.4 hours. We also observed that higher daily rainfall (e.g., 30mm or more) has a stronger effect on emergency LR hospital admissions, especially for the shorter lag of 0-2 days and longer lag of 7-10 days

8.2 Implications of the research findings

This research has tackled some of the research gaps identified in the systematic review of the literature. First of all, the outcome of the research may enhance our understanding of the relationships between the changing climate and disease epidemiology. This will increase our level of consciousness about climate change in scientific research and daily life, which will ultimately influence our actions towards human induced climate change.

The idea of considering all the significant climate variables in addition to temperature, their non-linearity, and delayed effects can be helpful for policy makers. Hospital managers and commissioners could possibly develop their models to predict emergency admissions for a wide range of disease categories and age group after a sudden change in climate, if we get a better predictive power of the model. Thus, it can improve the understanding of future patient flow related to climate change and help revise seasonal hospital demands. They can also improve patient flow management and review policies to cope with the changing climate. For the same reason, it would be easy to select the most vulnerable

population or disease groups due to climate change. This will also give the opportunity to maintain a proactive special care for these groups.

A better health alert system specifically for vulnerable and elderly people is indispensable due to changing climate for better health care management (IPCC 2007). However, almost all the health alert systems are based on temperature. Such system has thus become very fragile, since other factors like humidity, wind speed also has compounded impact on health and disease frequency. Based on the non-linear model developed in this study, we can calculate regional and disease specific thresholds which can lead towards an efficient and robust health alert system.

8.3 Recommendations and future works

The final model in this study has been developed for specific disease admissions, area, and time period. However, it can be applied and extended in various directions irrespective of time, place, and people.

8.3.1 Disease specific climate threshold and lag period for hospital admissions

This thesis provides the opportunity to calculate the climate threshold for emergency hospital admissions. It is very important to know the threshold level for various climate factors for different disease outcomes. Such threshold will be helpful for policymakers to regulate a “tolerable” amount of climate change for specific disease outcome. Literature review shows that most of the climate

effected disease categories exposed to seasonally in all the climate zones and population. Hospital admissions are also prone to seasonality.

8.3.2 Spatio-temporal modelling with disease specific lag structure and climate threshold

Quantitative description of the space-time effect between climate change and health will enrich the practical implications for the development of a better early warning system. Spatio-temporal modelling is a popular technique in environmental sciences. Identifying spatial hot spot based on an efficient threshold climate and temporal changes of the threshold would be a crucial advancement for determining the most vulnerable areas and population due to climate change. This will eventually lead towards a diversified health warning system, specific to homogeneous climate zone and population. The study in this thesis can also be extended towards a spatio-temporal approach based on lag structure and threshold climate. We have access to climate data provided by the met office for other regions, such as Greater Manchester, Kent, West Sussex, Devon, Dorset, Somerset, and Tyne & Wear.

8.3.3 Extending the DLNM-1

The distributed lag non-linear model developed here is based on the time series design. Theoretically the concept can be extended to other frameworks, such as any family of distribution and link function within the generalized linear model, with extensions to the generalized additive model or models based on generalized

estimating equations. All these theoretical extensions can be tested under the context of climate change and health.

8.3.4 Extending the DLNM-2

The considerations of higher order interactions terms (e.g. temperature*PM10) is important to further improvement of the DLNM model. In addition to this, it is important to check how DLNM model deal with serial autocorrelation between different lags and possibilities of biases in the model because of such serial autocorrelations. Thus we wish to check the feasibility of improving the model by incorporating higher order interactions terms among the exploratory variables (e.g. climate or air pollution factors) and checking possible biasness due to serial autocorrelations.

8.4 Limitations

The lack of quality data aggregated to appropriate levels linked to other sources is one of the toughest challenges in climate change research not mention other challenges such as missing data. Misclassification, measurement errors, and sampling & non-sampling error of the data are also very common in the applied field (e.g., categorising the disease based on the ICD, reporting errors).

Besides, data seem to have in various levels and problematic to aggregate them in more specific and lower level. For example, in our cases we are using patient level hospital admissions data and missing the lower level GP data. So in that sense, we are missing the very primary effected cases due to changing climate. Linking such different administrative data sets (e.g., HES, GP, and

climate data) are very demanding but challenging to deal with. This is also true if we want to consider the variables related to socioeconomic and demographics of climate vulnerable people in all stages.

Lack of connections between GP and hospital admission data means we are missing the part of affected population that visited to GP but not critical enough to admit to Hospitals. However, the good news is that the commissioners and policymakers have decided to link general practice information with secondary care data in NHS England (Davies 2013; Illman 2013).

Chapter 9

Publications during research

Journals

Islam M., Chaussalet T., Demir E., 2013. Modelling the Impact of Climate Change on Health, a Review submitted to the journal of Environmental Research, Elsevier.

Islam M., Chaussalet T., Demir E., 2013. Towards an efficient climate threshold for emergency hospital admissions in Greater London, in preparation for Journal of American Statistical Association.

Proceedings

Islam, M. S., Chaussalet, T., Ozkan, N., Chahed, S., Demir, E., & Sarran, C. (2011). *The impact of temperature disparity on emergency readmissions and patient flows.* Paper in the Proceedings of the 24th IEEE Symposium on Computer Based Medical Systems. DOI: 10.1109/CBMS.2011.5999124 . Pp. 1-6.

Islam, M. S., Chaussalet, T., Ozkan, N., & Demir, E. (2010). An approach to exploring the effect of weather variations on chronic disease incidence rate and potential changes in future health systems. Paper in the Proceedings of the 23rd

IEEE Symposium on Computer Based Medical Systems. DOI:10.1109/CBMS.2010.6042639. Pp. 190 - 196.

Islam, M. S., Chaussalet, T. J., Balta-Ozkan, N., & Demir, E. (2011). *Exploring the effect of temperature variations on unplanned asthma admissions.* Paper at the In: Operational Research Information National Health Policy: proceedings of the 37th ORAHS conference. School of Mathematics, Cardiff University, pp. 74-88. ISBN 9780956915801.

Poster events

The impact of temperature disparity on emergency readmissions and patient flows, 24th International Symposium on Computer Based Medical Systems (CBMS), University of Bristol, UK, 27th June 2011.

Exploring the effect of temperature variations on unplanned asthma admissions, EURO Working Group on Operational Research Applied to Health Services (ORAHS), Cardiff University, UK, 24th July 2011.

Conferences

Islam, M. S., Chaussalet, T., Ozkan, N., Chahed, S., Demir, E., & Sarran, C. The impact of temperature disparity on emergency readmissions and patient flows. 24th IEEE Symposium on Computer Based Medical Systems (CBMS), June 27-30 2011, Bristol, UK.

Islam, M. S., Chaussalet, T., Ozkan, N., & Demir, E. An approach to exploring the effect of weather variations on chronic disease incidence rate and potential changes in future health systems. 23rd IEEE Symposium on Computer Based Medical Systems (CBMS), October 12- 15 2010, Perth, Australia.

Islam, M. S., Chaussalet, T. J., Balta-Ozkan, N., & Demir, E. Exploring the effect of temperature variations on unplanned asthma admissions. EURO Working Group on Operational Research Applied to Health Services (ORAHS), July 24-29 2011, Cardiff University, UK.

Islam, M. S., Chaussalet, T. J., & Demir, E. Impact of Climate Change on Emergency asthma Admission: A case study for Greater London. Facing the Future, CFP, Postgraduate Conference, April 10-12 2013, Dundee, UK.

Islam, M. S., Chaussalet, T. J., & Demir, E. Impact of Climate Change on Emergency Hospital Admission: A case study for Greater London. 7th IMA Conference on Quantitative Modelling in the Management of Health and Social Care, March 25-27 2013, London, UK.

Chapter 10

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Second international symposium on information theory, Akademinai Kiado.

Alessandrini, E., S. Zauli Sajani, et al. (2011). Emergency ambulance dispatches and apparent temperature: A time series analysis in Emilia–Romagna, Italy. Environmental research. **111:** 1192-1200.

Alonso, J. B., J. A. Achcar, et al. (2010). "Climate changes and their effects in the public health: use of poisson regression models." Pesquisa Operacional **30**(2): 427-442.

Anderson, B. G. and M. L. Bell (2009). "Weather-related mortality: how heat, cold, and heat waves affect mortality in the United States." Epidemiology (Cambridge, Mass.) **20**(2): 205.

Argaud, L., T. Ferry, et al. (2007). "Short-and long-term outcomes of heatstroke following the 2003 heat wave in Lyon, France." Archives of Internal Medicine **167**(20): 2177-2183.

Armstrong, B. (2006). "Models for the Relationship Between Ambient Temperature and Daily Mortality." Epidemiology **17**(6): 624-631 610.1097/1001.ede.0000239732.0000250999.0000239738f.

AsthmaUK (2013). "2 million people unaware they are at risk of an asthma attack." Retrieved 1 December 2013, from http://www.asthma.org.uk/News/2-million-people-unaware-they-are-at-risk-of-an-asthma-attack.

Barbieri, P., N. Grieco, et al. (2010). Exploitation, integration and statistical analysis of the Public Health Database and STEMI Archive in the Lombardia region. Complex Data Modeling and Computationally Intensive Statistical Methods, Springer**:** 41-55.

Bartzokas, A., P. Kassomenos, et al. (2004). "The effect of meteorological and pollution parameters on the frequency of hospital admissions for cardiovascular and respiratory problems in Athens." Indoor and Built Environment **13**(4): 271-275.

Bassil, K. L., D. C. Cole, et al. (2009). "Temporal and spatial variation of heat-related illness using 911 medical dispatch data." Environmental Research **109**(5): 600-606.

Basu, R., W.-Y. Feng, et al. (2008). "Characterizing Temperature and Mortality in Nine California Counties." Epidemiology **19**(1): 138-145 110.1097/EDE.1090b1013e31815c31811da31817.

Basu, R. and B. Malig (2011). "High ambient temperature and mortality in California: Exploring the roles of age, disease, and mortality displacement." Environmental research **111**(8): 286-1292.

Basu, R. and J. M. Samet (2002). "Relation between Elevated Ambient Temperature and Mortality: A Review of the Epidemiologic Evidence." Epidemiologic Reviews **24**(2): 190-202.

Bhaskaran, K., S. Hajat, et al. (2010). "Short term effects of temperature on risk of myocardial infarction in England and Wales: time series regression analysis of the

Myocardial Ischaemia National Audit Project (MINAP) registry." BMJ: British Medical Journal **341**.

Braga, A. L., A. Zanobetti, et al. (2002). "The effect of weather on respiratory and cardiovascular deaths in 12 U.S. cities." Environmental health perspectives **110**(9): 859-863.

Braga, A. L. F., A. Zanobetti, et al. (2001). "The time course of weather-related deaths." Epidemiology **12**(6): 662-667.

Buja, A., T. Hastie, et al. (1989). "Linear smoothers and additive models." The Annals of Statistics: 453-510.

Carson, C., S. Hajat, et al. (2006). "Declining vulnerability to temperature-related mortality in London over the 20th century." American Journal of Epidemiology **164**(1): 77-84.

Chandler, R. E. (2005). "On the use of generalized linear models for interpreting climate variability." Environmetrics **16**(7): 699-715.

Chang, H. H., J. Zhou, et al. (2010). "Impact of Climate Change on Ambient Ozone Level and Mortality in Southeastern United States." International Journal of Environmental Research and Public Health **7**(7): 2866-2880.

Checkley, W., J. Guzman-Cottrill, et al. (2009). "Short-term weather variability in Chicago and hospitalizations for Kawasaki disease." Epidemiology (Cambridge, Mass.) **20**(2): 194-201.

Curriero, F. C., K. S. Heiner, et al. (2002). "Temperature and Mortality in 11 Cities of the Eastern United States." American Journal of Epidemiology **155**(1): 80-87.

Davies, M. (2013). "NHS to link up data from GP records and secondary care." Retrieved 1 September 2013, from http://www.pulsetoday.co.uk/your-practice/practice-topics/it/nhs-to-link-up-data-from-gp-records-and-secondary-care/20002260.article#.UieQIu8kwrg.

Davis, R. E., P. C. Knappenberger, et al. (2004). "Seasonality of climate-human mortality relationships in US cities and impacts of climate change." Climate Research **26**(1): 61-76.

Díaz, J., R. García, et al. (2005). "Mortality impact of extreme winter temperatures." International Journal of Biometeorology **49**(3): 179-183.

Díaz, J., A. Jordán, et al. (2002). "Heat waves in Madrid 1986–1997: effects on the health of the elderly." International Archives of Occupational and Environmental Health **75**(3): 163-170.

Dobson, A. J. and A. G. Barnett (2008). "An Introduction to Generalized Linear Models."

Dolney, T. J. and S. C. Sheridan (2006). "The relationship between extreme heat and ambulance response calls for the city of Toronto, Ontario, Canada." Environmental Research **101**(1): 94-103.

Donaldson, G., W. Keatinge, et al. (2003). "Changes in summer temperature and heat-related mortality since 1971 in North Carolina, South Finland, and Southeast England." Environmental Research **91**(1): 1-7.

Ebi, K. L., K. A. Exuzides, et al. (2004). "Weather changes associated with hospitalizations for cardiovascular diseases and stroke in California, 1983–1998." International Journal of Biometeorology **49**(1): 48-58.

Ebi, K. L. and G. McGregor (2008). Climate change, tropospheric ozone and particulate matter, and health impacts.

El-Zein, A., M. Tewtel-Salem, et al. (2004). "A time-series analysis of mortality and air temperature in Greater Beirut." Science of The Total Environment **330**(1): 71-80.

ESPERE (2004). "Environmental Science Published Everybody Round the Earth, Weather.". Retrieved 17 January 2013, from http://www.atmosphere.mpg.de/enid/3se.html.

Fernández-Raga, M., C. Tomás, et al. (2010). "Human mortality seasonality in Castile-León, Spain, between 1980 and 1998: the influence of temperature, pressure and humidity." International Journal of Biometeorology **54**(4): 379-392.

Ferrari, U., T. Exner, et al. (2012). "Influence of air pressure, humidity, solar radiation, temperature, and wind speed on ambulatory visits due to chronic obstructive pulmonary disease in Bavaria, Germany." International Journal of Biometeorology **56**(1): 137-143.

Fouillet, A., G. Rey, et al. (2007). "A predictive model relating daily fluctuations in summer temperatures and mortality rates." BMC Public Health **7**(1): 114.

Fouillet, A., G. Rey, et al. (2006). "Excess mortality related to the August 2003 heat wave in France." International Archives of Occupational and Environmental Health **80**(1): 16-24.

Gasparrini, A. (2011). "Distributed lag linear and non-linear models in R: the package dlnm." Journal of Statistical Software **43**(8): 1.

Gasparrini, A., B. Armstrong, et al. (2010). "Distributed lag non-linear models." Statistics in medicine **29**(21): 2224-2234.

Green, R. S., R. Basu, et al. (2010). "The effect of temperature on hospital admissions in nine California counties." International journal of public health **55**(2): 113-121.

Guisan, A., T. C. Edwards, et al. (2002). "Generalized linear and generalized additive models in studies of species distributions: setting the scene." Ecological modelling **157**(2): 89-100.

Haigh, I., R. Nicholls, et al. (2011). "Rising sea levels in the English Channel 1900 to 2100." Proceedings of the ICE-Maritime Engineering **164**(2): 81-92.

Hajat, S., B. G. Armstrong, et al. (2005). "Mortality displacement of heat-related deaths: a comparison of Delhi, Sao Paulo, and London." Epidemiology **16**(5): 613-620.

Hansen, A., P. Bi, et al. (2008). "The effect of heat waves on mental health in a temperate Australian city." Environmental Health Perspectives **116**(10): 1369.

Hansen, A. L., P. Bi, et al. (2008). "The effect of heat waves on hospital admissions for renal disease in a temperate city of Australia." International journal of epidemiology **37**(6): 1359-1365.

Hartz, D. A., J. S. Golden, et al. (2012). "Climate and heat-related emergencies in Chicago, Illinois (2003–2006)." International journal of biometeorology **56**(1): 71-83.

Hastie, T. and R. Tibshirani (1987). "Generalized additive models: some applications." Journal of the American Statistical Association **82**(398): 371-386.

Hastie, T. and R. Tibshirani (1990). Generalized additive models., Chapman & Hall, CRC Press.

HES (2013). "Hospital Episode Statistics." Retrieved February 2, 2013, from www.hscic.gov.uk.

Hu, W., K. Mengersen, et al. (2010). "The use of ZIP and CART to model cryptosporidiosis in relation to climatic variables." International Journal of Biometeorology **54**(4): 433-440.

Huynen, M. M., P. Martens, et al. (2001). "The impact of heat waves and cold spells on mortality rates in the Dutch population." Environmental health perspectives **109**(5): 463-470.

Illman, J. (2013). "Linked hospital and GP data 'to be available from autumn'.". Retrieved 1 September 2013, from http://www.hsj.co.uk/news/commissioning/linked-hospital-and-gp-data-to-be-available-from-autumn/5059310.article.

IPCC (2007). Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, M.L. Parry, O.F. Canziani, J.P. Palutikof, P.J. van der Linden and C.E. Hanson, Eds., Cambridge University Press, Cambridge, UK, 976pp.

IPCC (2007). Climate Change 2007: Mitigation. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [B. Metz, O.R. Davidson, P.R. Bosch, R. Dave, L.A. Meyer (eds)], Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA., XXX pp.

IPCC (2013). "Intergovernmental Panel on Climate Change. Appendix I: Glossary.". Retrieved 17 January 2013, from http://www.grida.no/publications/other/ipcc_tar/.

Islam, M. S., T. Chaussalet, et al. (2011). The impact of temperature disparity on emergency readmissions and patient flows. Proceedings of 24th IEEE Symposium on Computer Based Medical Systems. DOI: 10.1109/CBMS.2011.5999124 . pp. 1-6.

Islam, M. S., T. Chaussalet, et al. (2010). An approach to exploring the effect of weather variations on chronic disease incidence rate and potential changes in future health systems. Proceedings of 23rd IEEE Symposium on Computer Based Medical Systems. DOI:10.1109/CBMS.2010.6042639. pp.190 - 196.

Islam, M. S., T. J. Chaussalet, et al. (2011). Exploring the effect of temperature variations on unplanned asthma admissions. In: Operational Research Information National Health Policy: proceedings of the 37th ORAHS conference. School of Mathematics, Cardiff University, pp. 74-88. ISBN 9780956915801.

Johnson, H., S. Kovats, et al. (2004). "The impact of the 2003 heat wave on mortality and hospital admissions in England." Epidemiology **15**(4): S126.

Kaiser, R., A. Le Tertre, et al. (2007). "The effect of the 1995 heat wave in Chicago on all-cause and cause-specific mortality." American journal of public health **97**(Supplement_1): 158-162.

Kaiser, R., C. H. Rubin, et al. (2001). "Heat-related death and mental illness during the 1999 Cincinnati heat wave." The American journal of forensic medicine and pathology **22**(3): 303-307.

Kalkstein, L. S. and R. E. Davis (2005). "Weather and human mortality: an evaluation of demographic and interregional responses in the United States." Annals of the Association of American Geographers **79**(1): 44-64.

Keatinge, W. R. and G. C. Donaldson (2001). "Mortality Related to Cold and Air Pollution in London After Allowance for Effects of Associated Weather Patterns." Environmental research **86**(3): 209-216.

Khalaj, B., G. Lloyd, et al. (2010). "The health impacts of heat waves in five regions of New South Wales, Australia: a case-only analysis." International Archives of Occupational and Environmental Health **83**(7): 833-842.

Knowlton, K., M. Rotkin-Ellman, et al. (2009). "The 2006 California heat wave: impacts on hospitalizations and emergency department visits." Environmental health perspectives **117**(1): 61-67.

Kolb, S., K. Radon, et al. (2007). "The short-term influence of weather on daily mortality in congestive heart failure." Archives of environmental & occupational health **62**(4): 169-176.

Kovats, R. S. and S. Hajat (2008). "Heat stress and public health: a critical review." Annual Review of Public Health **29**(1): 41-55.

Kovats, R. S., S. Hajat, et al. (2004). "Contrasting patterns of mortality and hospital admissions during hot weather and heat waves in Greater London, UK." Occupational and environmental medicine **61**(11): 893-898.

Lam, L. T. (2007). "The association between climatic factors and childhood illnesses presented to hospital emergency among young children." International Journal of Environmental Health Research **17**(1): 1-8.

Le Tertre, A., A. Lefranc, et al. (2006). "Impact of the 2003 heatwave on all-cause mortality in 9 French cities." Epidemiology **17**(1): 75-79.

Liang, W. M., W. P. Liu, et al. (2009). "Diurnal temperature range and emergency room admissions for chronic obstructive pulmonary disease in Taiwan." International journal of biometeorology **53**(1): 17-23.

Ma, W., X. Xu, et al. (2011). "Impact of extreme temperature on hospital admission in Shanghai, China." Science of The Total Environment **409**(19): 3634-3637.

MacCullagh, P. and J. A. Nelder (1989). Generalized linear models, CRC press.

McCullagh, P. and J. A. Nelder (1989). Generalized linear model, Chapman & Hall/CRC.

McGeehin, M. A. and M. Mirabelli (2001). "The potential impacts of climate variability and change on temperature-related morbidity and mortality in the United States." Environmental Health Perspectives **109**(Suppl 2): 185.

McMichael, A. J., A. Haines, et al. (1996). Climate change and human health, World Health Organization Geneva.

Medina-Ramón, M., A. Zanobetti, et al. (2006). "Extreme temperatures and mortality: assessing effect modification by personal characteristics and specific cause of death in a multi-city case only analysis." Environmental Health Perspectives **114**(9): 1331.

Menne, B., F. Apfel, et al. (2008). Protecting health in Europe from climate change, World Health Organization.

Mentzakis, E. and D. Delfino (2010). "Effects of air pollution and meteorological parameters on human health in the city of Athens, Greece." International Journal of Environment and Pollution **40**(1): 210-225.

Michelozzi, P., G. Accetta, et al. (2009). "High temperature and hospitalizations for cardiovascular and respiratory causes in 12 European cities." American journal of respiratory and critical care medicine **179**(5): 383-389.

Michelozzi, P., U. Kirchmayer, et al. (2007). "Assessment and prevention of acute health effects of weather conditions in Europe, the PHEWE project: background, objectives, design." Environmental Health **6**(1): 12.

Miriovsky, B. J., L. N. Shulman, et al. (2012). "Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care." Journal of Clinical Oncology **30**(34): 4243-4248.

Moher, D., A. Liberati, et al. (2009). "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement." PLoS medicine **6**(7): e1000097.

Muggeo, V. M. (2003). "Estimating regression models with unknown break-points." Statistics in medicine **22**(19): 3055-3071.

Muggeo, V. M. and S. Hajat (2009). "Modelling the non-linear multiple-lag effects of ambient temperature on mortality in Santiago and Palermo: a constrained segmented distributed lag approach." Occupational and Environmental Medicine **66**(9): 584-591.

Nagelkerke, N. J. (1991). "A note on a general definition of the coefficient of determination." Biometrika **78**(3): 691-692.

Nastos, P. T. and A. Matzarakis (2006). "Weather impacts on respiratory infections in Athens, Greece." International journal of biometeorology **50**(6): 358-369.

O'Loughlin, J. L., Y. Robitaille, et al. (1993). "Incidence of and risk factors for falls and injurious falls among the community-dwelling elderly." American Journal of Epidemiology **137**(3): 342-354.

O'Neill, M. S., A. Zanobetti, et al. (2005). "Disparities by race in heat-related mortality in four US cities: the role of air conditioning prevalence." Journal of urban health : bulletin of the New York Academy of Medicine **82**(2): 191-197.

ONS, E. (2012). "Population Estimates for England and Wales, Mid 2011 (Census Based).". Retrieved 29 May 2013, from http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-262039.

Ostro, B., S. Rauch, et al. (2010). "The effects of temperature and use of air conditioning on hospitalizations." American Journal of Epidemiology **172**(9): 1053-1061.

Parry, M. L., O. F. Canziani, et al. (2007). IPCC, 2007: climate change 2007: impacts, adaptation and vulnerability. Contribution of working group II to the fourth assessment report of the intergovernmental panel on climate change, Cambridge University Press, Cambridge.

Pattenden, S., B. Nikiforov, et al. (2003). "Mortality and temperature in Sofia and London." Journal of epidemiology and community health **57**(8): 628-633.

Pauli, F. and L. Rizzi (2006). "Statistical analysis of temperature impact on daily hospital admissions: analysis of data from Udine, Italy." Environmetrics **17**(1): 47-64.

Pauli, F. and L. Rizzi (2008). "Analysis of heat wave effects on health by using generalized additive model and bootstrap-based model selection." Journal of the Royal Statistical Society: Series C (Applied Statistics) **57**(4): 473-485.

Pauli, F. and L. Rizzi (2008). "Summer temperature effects on deaths and hospital admissions among the elderly population in two Italian cities." Journal of Applied Statistics **35**(3): 263-276.

Pinto, E., M. Coelho, et al. (2011). "The influence of climate variables on dengue in Singapore." International Journal of Environmental Health Research **21**(6): 415-426.

Pudpong, N. and S. Hajat (2011). "High temperature effects on out-patient visits and hospital admissions in Chiang Mai, Thailand." The Science of the total environment **409**(24): 5260-5267.

Qian, Z., Q. He, et al. (2008). "High temperatures enhanced acute mortality effects of ambient particle pollution in the "oven" city of Wuhan, China." Environmental health perspectives **116**(9): 1172-1178.

Revich, B. and D. Shaposhnikov (2008). "Excess mortality during heat waves and cold spells in Moscow, Russia." Occupational and Environmental Medicine **65**(10): 691-696.

Rocklöv, J. and B. Forsberg (2009). "Comparing approaches for studying the effects of climate extremes - a case study of hospital admissions in Sweden during an extremely warm summer." Glob Health Action **2**.

Rudge, J. and R. Gilchrist (2005). "Excess winter morbidity among older people at risk of cold homes: a population-based study in a London borough." Journal of Public Health **27**(4): 353-358.

Schwartz, J. (2000). "The distributed lag between air pollution and daily deaths." Epidemiology **11**(3): 320-326.

Schwartz, J. (2001). "Is there harvesting in the association of airborne particles with daily deaths and hospital admissions?" Epidemiology **12**(1): 55-61.

Schwartz, J., J. M. Samet, et al. (2004). "Hospital admissions for heart disease: the effects of temperature and humidity." Epidemiology **15**(6): 755-761.

Sung, T.-I., M.-J. Chen, et al. (2011). "Relationship between mean daily ambient temperature range and hospital admissions for schizophrenia: Results from a national cohort of psychiatric inpatients." Science of The Total Environment **410–411**(0): 41-46.

Tam, W. W., T. W. Wong, et al. (2009). "Diurnal temperature range and daily cardiovascular mortalities among the elderly in Hong Kong." Archives of environmental & occupational health **64**(3): 202-206.

Tan, J., Y. Zheng, et al. (2007). "Heat wave impacts on mortality in Shanghai, 1998 and 2003." International Journal of Biometeorology **51**(3): 193-200.

Tong, S., C. Ren, et al. (2010). "Excess deaths during the 2004 heatwave in Brisbane, Australia." International Journal of Biometeorology **54**(4): 393-400.

Tong, S., X. Y. Wang, et al. (2010). "Assessment of Heat-Related Health Impacts in Brisbane, Australia: Comparison of Different Heatwave Definitions." PLoS ONE **5**(8): e12155.

Vaneckova, P., P. J. Beggs, et al. (2010). "Spatial analysis of heat-related mortality among the elderly between 1993 and 2004 in Sydney, Australia." Social Science & Medicine **70**(2): 293-304.

Vardoulakis, S. and C. Heaviside (2012). "Health Effects of Climate Change in the UK 2012.". Retrieved 18 January 2013, from http://www.hpa.org.uk/hecc2012.

Wang, X. Y., A. G. Barnett, et al. (2009). "Temperature variation and emergency hospital admissions for stroke in Brisbane, Australia, 1996–2005." International journal of biometeorology **53**(6): 535-541.

WHO (2008). "Taking action to protect health in Europe from Climate Change. Fact Sheet Copenhagen, 4 April 2008, World Health Organization.". Retrieved 3 March 2012, from http://www.euro.who.int/data/assets/pdf_file/0007/95830/fs_4_Apr_08e.pdf.

Wichmann, J., Z. Andersen, et al. (2011). "Apparent Temperature and Cause-Specific Emergency Hospital Admissions in Greater Copenhagen, Denmark." PLoS ONE **6**(7): e22904.

WMO (2011). "Provisional Statement on the Status of the Global Climate, World Meteorological Organization.". Retrieved 19 January 2013, from http://www.wmo.int/pages/mediacentre/press_releases/gcs_2011_en.html.

Wood, S. N. (2006). Generalized additive models: an introduction with R, Chapman & Hall.

World Health Organisation (WHO) (2006). WHO air quality guidelines for nitrogen dioxide, ozone, sulphur dioxide and particulate matter. Global update 2005. Summary of risk assessment. WHO, Geneva.

Zanobetti, A., M. Wand, et al. (2000). "Generalized additive distributed lag models: quantifying mortality displacement." Biostatistics **1**(3): 279-292.