Time series analysis of heat stroke related hospitalizations in the elderly residents in the Boston Metropolitan Statistical Area, from 1991 to 2006

A thesis submitted by

Ruiruo Wu

In partial fulfillment of the requirements
for the degree of

Master of Science

in

Department of Civil & Environmental Engineering

TUFTS UNIVERSITY

August, 2013

Adviser:

Dr. Elena N. Naumova

Abstract

The adverse impact of extreme weather on human health is attracting substantial attentions of public health professionals, environmental scientists, and policy makers. This study uses time series analysis to investigate the direct effects of high ambient temperatures on the health of older residents in the Boston Metropolitan Statistical Area (MSA). This study has three goals: to investigate the seasonality of hospitalizations due to heat stroke (HSH), to describe the association between ambient temperature and HSHs; and to determine the independent effect of heat waves on HSHs in the presence of well-pronounced seasonality. Medical records were abstracted form the Centers of Medicare and Medicaid Services from 1991 through 2006 and regressed against daily meteorological records using Poisson generalized linear regression model. The ambient temperature is positively correlated with HSHs, the relationship exhibits strong exponential growth when temperature exceeds a certain threshold point. The proposed transformation of daily temperature facilitates the capture of rapid increase in HSHs beyond the threshold. Six types of indicator variables were tested to determine the effects of heat waves and their lag structures. The seasonal peaks in HSHs follow the first heat wave each year. The final regression model includes terms for ambient temperatures transformed with respect to threshold points; the best indicator variable for heat waves adjusted for annual, seasonal and weekly fluctuations explains 56% variability in HSHs. The developed methodology for building and validating a regression model sensitive to local temperature features can be further extended to other locations and health outcomes.

Acknowledgement

This thesis could not have been completed without the significant contributions from my committee. Dr. Naumova, your profound knowledge and patient guidance sparks my strong interest in environmental health and encourages me to overcome all the challenges during my study. Thank you Dr. Chui for giving me thorough and detailed suggestions on my thesis editing. I would also like to express my gratitude to Dr. Tanaka for his critical comments and suggestions on my thesis.

This study would not be finished without my supporting and knowledgeable colleague Alex Liss at Tufts Initiative for the Forecasting and Modeling of Infectious Diseases. Thank you Alex for extracting all the data for my study and giving useful suggestions on study design and presentations.

The travel supports given by Tufts University Graduate School and School of Engineering Dean's office allowed me presenting my work and learning from other researchers at 2012 AGU Fall meeting in San Francisco, California.

I am grateful to the writing tutor Jackie O'Dell at Tufts Academic Resource Center for helping me edit my thesis.

I would like to specially thank my parents for supporting my study at Tufts. My father doesn't have the opportunity to witness my success, but I know that he has always been there for me. My boyfriend Hao Zhang, I sincerely appreciate you love and care.

Finally, I want to express my gratitude to all my friends. I truly appreciate all your patience and encouragement.

Contents

List of Figures

List of Tables

List of Abbreviations

Centers for Disease Control and Prevention: CDC

Geographic information system: GIS

Generalized Linear Model: GLM

Generalized Additive Models: GAM

International classification of Disease, 9th Revision, Clinical Modification: ICD-9-CM

Metropolitan Statistical Area: MSA

1. **Introduction**

1.1 General statement of the problem and its significance

Increasing greenhouse gas emissions lead to a rise in average temperature, resulting in a more variable climate system [1-8]. Local events such as heat waves, elevated ambient temperatures, windstorms, droughts, and dust storms are the direct products of climate change, among which heat waves are considered the most threatening one [2, 9]. Elderly people are less capable of adapting to the variable temperatures and more vulnerable to extreme weather related illnesses, such as heat stroke [10, 11]. With the improvement of living standards and medical supports, the elderly population is increasing greatly both in size and proportion. Therefore, it is urgent to understand how heat events affect elderly populations, so that protective strategies can be better formulated.

Studies of heat events and health outcomes have been conducted all over the world: Australia [12-14], Asia [15-17], Europe [18-33] , Latin America [34], and North America [9, 35-40]. Time series analysis [20, 28, 35, 36, 38, 41-47] and case cross-over [34, 39, 48] studies are the most popular study designs. These published studies had some common findings: in each specific location, a "J-shaped" curve can be used to illustrate the relationship between temperature and negative health outcomes [23, 38, 39, 45-47, 49-51]; for each location, there always exists a comfortable temperature region, within which there are fewer health risks. When temperature exceeds a certain threshold, every additional degree in temperature may result in more health risks [35].

This study firstly explains the seasonality of heat stroke related hospitalizations (HSHs) with calendar effects, which enables people to easily predict the timing, intensity and duration of HSHs and guide the planning of community events and daily life. Then this study investigates the association between ambient temperature and HSHs, and finds that there will be more health risks when temperature exceeds the (location-specific) temperature threshold. Applying the location-specific threshold points

and temperature transformations allows this study to extend the model to other locations. The Incidence Rate Ratio (IRR) of temperature exposure and HSHs are proposed as a function of the ambient temperature instead of a constant value, which highlights the role of threshold point in health risk. Lastly, this study proposes an indicator of heat wave for the Boston-Cambridge-Quincy MA NH Metropolitan Statistical Area (Boston MSA), which sufficiently concerns the local climate features, and finds that the annual peak of HSHs follows the first occurrence of the heat wave each year. Knowing the heat wave's starting time will help the public plan accordingly.

1.1.1 Health outcome

Daily hospitalization records from 1991 to 2006 were purchased from the Centers of Medicare and Medicaid Services. This database covers 98% of the population who are 65 year-old or above in the U.S. This comprehensive dataset contains over 100 variables, including age, race, and gender of the patient's; ZIP code and county FIPS code of residence; location of discharge (to home, skilled nursing facility, etc.); date of death if deceased; assigned medical procedures; total charge of the hospitalization; health condition at the time of discharge; and up to 10 diagnostic coding slots with their respective International classification of Disease, 9th Revision, Clinical Modification (ICD-9-CM) code. The study data were extracted for heat stroke related hospitalization (ICD-9-CM 992.0 − 992.9) (Table 1.1, the 2nd column), and there were 48,132 heat stroke related hospitalization cases in the U.S. from 1991 to 2006.

The ZIP codes in the data allow for spatial aggregation up to the level of MSA, which the U.S. Office of Management and Budget defines as "a region that has at least one urbanized area of 5000 or more population, plus adjacent territory that has a high degree of social and economic integration with the core as measured by community ties". There were 701 HSHs in the Boston-Cambridge-Quincy, MA NH MSA from 1991 to 2006. As is shown in Table 1.1, the 3rd column, most HSHs were classified as unspecified heat exhaustion (992.5). The date of admission is recorded in the level of a calendar day,

which is adequate for modeling the effect of heat exposure as it can have lag effects on human health ranging from 2 to 20 days. The daily level data also allow for modeling the accumulated effects of heat exposure on human health. Moreover, the hospitalization peak time usually happens within a week; and the peak of HSHs and its date would be hard to discern if only weekly or monthly data were used.

Table 1.1 The explanation of each ICD code (992.0-992.9) and cases

ICD-9-CM code	Description	Cases
992.0	Heat stroke and sunstroke	115
992.1	Heat syncope	49
992.2	Heat cramps	6
992.3	Heat exhaustion, anhydrotic	40
992.4	Heat exhaustion due to salt depletion	9
992.5	Heat exhaustion, unspecified	465
992.6	Heat fatigue, transient	2
992.7	Heat edema	0
992.8	Other specified heat effects	7
992.9	Unspecified effects of heat and light	8

The Boston-Cambridge-Quincy, MA NH MSA is the 10th most populated MSA in the U.S., according to Census 2010. The total population was 4,552,402, consisted of 74.9%non-Hispanic White population, 9% Hispanic population, 7.4% non-Hispanic Black population, 7.1% Asian population and 1.6% other races population. The proportion of elderly population was 10.1%. The total area of this MSA is 4,674 mi^2.

1.1.2   Temperature exposure

Temperature records of the Boston-Cambridge-Quincy MA NH MSA were obtained from the National Oceanographic and Atmospheric Administration (NOAA). This dataset includes maximum daily temperature, minimum daily temperature and mean daily dewpoint temperature. The daily temperature data were interpolated separately for these centroids of each zip code within the Boston-

Cambridge-Quincy MA NH MSA one the same day using via inverse distance weighting method[1]. Only stations within 120 miles radius from each zip-code centroid were considered in the interpolation.

1.2 Statement of hypotheses

1) The HSHs demonstrate well-pronounced seasonality, and there are some annual variations in intensity and peak time.

2) The ambient temperature highly influences the heat stroke related hospitalization, when ambient temperature exceeds a threshold, the HSHs will increase rapidly.

3) The heat wave independently affects HSHs in the presence of well-pronounced seasonality.

1.3 Organization of this thesis

Chapter 2 is the literature review. Chapter 3 explains the statistical features of HSHs. Chapter 4 investigates the seasonality of HSHs. Chapter 5 investigates the association between temperature and HSHs. Chapter 6 analyzes the heat wave effect on HSHs. Chapter 7 is the study conclusion. Chapter 8 is the study discussion and future study direction.

---

[1] Inverse Distance Weighting (IDW) is a type of deterministic method for multivariate interpolation with a known scattered set of points. The assigned values to unknown points are calculated with a weighted average of the values available at the known points

1. Literature review

2.1 Motivation

A literature review about health outcomes due to heat events was done. The purpose of the review was to understand those studies systematically and thoroughly, and then to compare and contrast the study methods and findings. The literature review contained the health outcome selection, the exposure selection, the application of methodology, seasonality analysis, and the selection of distribution; moreover, focus was also put on the definition of the heat wave in various studies and locations.

This literature review used the "Snowball" method to find articles on the topic about heat events and health outcomes. This study started with the paper "*Ambient Temperature and Cardiorespiratory morbidity: A Systematic Review and Meta-analysis*" [52] published in June 2012. Subsequent studies were then included in the review through tracing the references sections retrospectively.

2.2 Outcome measurement

There is considerable variability in the choice of health outcomes. The health outcomes are usually explained in terms of morbidity and mortality. The concept of heat-related morbidity has a wide coverage of health outcomes, including uncomfortable symptoms, the serious effects on skin or organ, and other life-threatening effects. Table 2.1 summarizes the heat exposure related health symptoms respectively. As is shown in Table 2.1, different types of health outcomes have different severities. This literature review found more studies on heat related mortality [2, 3, 9, 26, 36, 41, 42, 46, 53-55] than those on heat related morbidity [28, 44, 54-57]. The term morbidity in those studies mainly referred to the records of emergency department visits, registry or hospital admission. Using morbidity as the measurement of health outcome can guarantee adequate sample sizes and wider coverage of health outcomes. However, the term morbidity contains quite a wide range of symptoms and conditions, some of which may not be serious enough to require medical attention [23]. Since we only utilized

5

hospitalization records in this study, modeling of the exact morbidity are not fully achievable. The studies on mortality explain the most severe health outcome, but the mortality cases only take up a limited proportion of health outcomes related to a certain cause, therefore utilizing the mortality as the measurement of health outcome in a regression model may limit the sample size, affecting the model quality. In addition, the study on mortality may not give widely applicable suggestions for health protection.

The classification of underlying causes of diseases also significantly affects the study conclusions. Published studies have discussed both all-cause diseases and diseases with specific ICD codes. The published studies applied the all-cause [26, 58, 59], coronary [58, 60, 61], cardiovascular [37, 49, 50, 58, 62-69] and respiratory diseases [30, 42, 49, 50, 67, 70, 71] as the health outcome measurements and only two studies [55, 56] in this review referred to direct health outcomes (ICD-9-CM 992.0 − 992.9) of heat events. Different health outcome extracting criteria have various pros and cons. Applying all-cause mortality or morbidity as health outcomes can avoid bias created by misclassification of causes; however, it may weaken the correlation between heat events and health outcomes. Applying the mortality due to chronic diseases as health outcomes can examine the heat effects on vulnerable people [9, 25, 26, 30, 42, 50, 67, 71], but the harvesting effect [21, 45, 72] will create bias when quantifying the temporal relationship between heat events and health outcomes. Applying the direct health outcomes of heat events can help to understand the independent adverse influence of heat effect, but the sample size of heat related diseases may affect the statistical significance of a study.

Table 2.1 Definitions of heat exposure outcomes and their ICD codes (Data sources are cited in Appendix 1 and 2)

Outcomes	Symptoms	ICD-9-CM code
Heat stroke and sunstroke (Heat apoplexy; Heat pyrexia; Ictus Solaris; Siriasis; Thermoplegia)	•Hot, dry skin or profuse sweating •Hallucinations •Chills •Throbbing headache •High body temperature •Confusion/dizziness •Slurred speech	992.0
Heat syncope (Heat collapse)	•Light-headedness •Dizziness •Fainting	992.1
Heat cramps	•Muscle pain or spasms usually in the abdomen, arms, or legs	992.2
Heat exhaustion, anhydrotic (Heat prostration due to water depletion)	•Heavy sweating •Extreme weakness or fatigue •Dizziness, confusion •Nausea •Clammy, moist skin •Pale or flushed complexion •Muscle cramps •Slightly elevated body temperature •Fast and shallow breathing	992.3
Heat exhaustion due to salt depletion (Heat prostration due to salt (and water) depletion)		992.4
Heat exhaustion, unspecified		992.5
Heat fatigue, transient		992.6
Heat edema		992.7
Other specified heat effects		992.8
Unspecified effects of heat and light		992.9

2.2.1 Vulnerability

Vulnerability refers to the inability to withstand the effects of a hostile environment. People are vulnerable to heat exposure when their physiologic mechanisms for heat loss fail to work, keeping heat from being dissipated through radiation into the air or evaporation in the form of perspiration [73]. The vulnerability to heat can be explained at both the community and individual levels. Identifying vulnerable groups in the community level is an important social and health issue, because the vulnerability is correlated with the housing design, the social economic status of a community. The vulnerability can be explained from both disproportionate exposure and intrinsic biological conditions such as age, gender or health conditions [31, 34, 38, 39, 74]. Immobilization keeps people from looking for a cooler place when exposed to heat; thus elderly people in nursing homes are under this type of threat. It is reported that the elderly are one of the subpopulations at highest risk during heat events [9, 10, 36, 39] , for they have weaker thermoregulatory mechanisms, take other medications, potentially live in social isolation, and are confined to bed [9, 18, 73], thus those communities with high elderly populations may require additional attentions. In addition, people living in regions with higher temperature have different physiological responses to heat due to cultural and social adaptive behaviors [4, 74]. To an individual, when body temperature rises, the blood flow generally shifts from the vital organs to underneath the skin's surface in an effort to cool down [11]. Inadequate thermoregulation may occur when too much blood is diverted, putting increased stress on the heart and lungs [11]. Increased blood viscosity, elevated cholesterol levels associated with higher temperatures, and a higher sweating threshold may also trigger severe heat-related disease or even mortality [11]. A reduced thermoregulatory capacity combined with a diminished ability to detect changes in body temperature may even increase the elderly population's vulnerability to heat [74].

This study selected twenty populated cities with wide spatial coverage across the U.S. (Figure 2.1), collecting the demographic and meteorological information from *Wikipedia* and *Current Results* (data

sources are cited in Appendix 3 and 4.). In Figure 2.2, the x and y-axes stand for the annual low and high temperatures of each city, the size of a bubble is proportional to the number of elderly people (65 and over) (2010 Census data) of each city. The cities with a large group of elderly people may be more vulnerable to heat, like New York and Chicago.
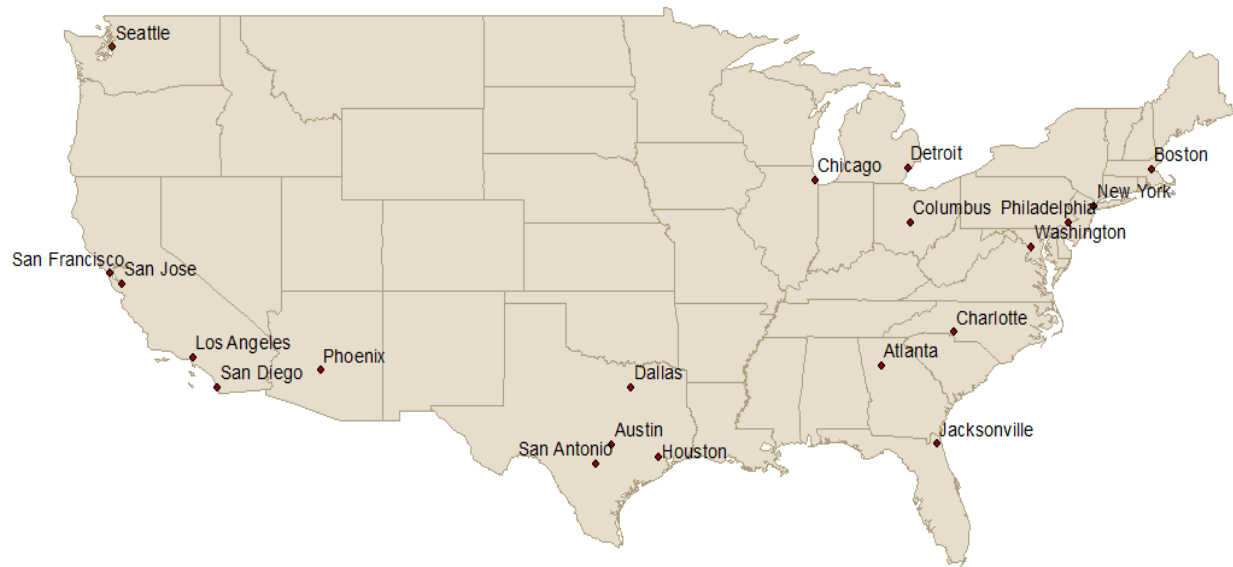
Figure 2.1 The geographic locations of the 20 big cities selected in our study
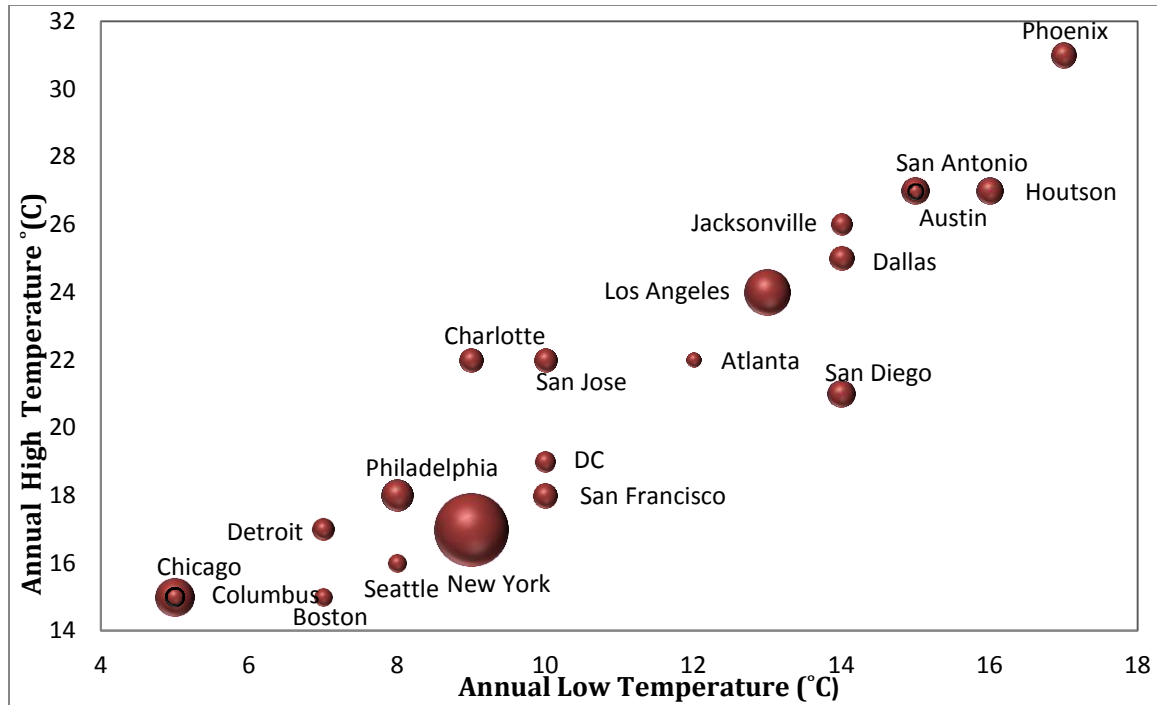
Figure 2.2 The elderly population and temperature features of 20 big cities in the U.S.

The Figure 2.2 indicates that neither temperature features nor the number of elderly people alone can sufficiently explain the heat vulnerability of a city. The city of Chicago and the city of Columbus have very similar temperature features, but Chicago has been more seriously challenged by heat exposure than Columbus has [54, 55]. San Francisco, Los Angeles and San Diego have similar elderly population conditions and are all located along the seashore, but their heat risks vary greatly in terms of intensity and severity [44, 56, 57]. Jacksonville, San Antonio, Houston, Austin and Phoenix are much warmer than Philadelphia, and they have similar elderly populations, but the warmer cities are reported to have less vulnerability to heat than Philadelphia [38, 64, 75], because the adaption and acclimatization also directly affect people's vulnerability to heat. Generally, people in cooler regions tend to be more vulnerable to heat.

2.3 Exposure measurement

10

The Centers for Disease Control and Prevention (CDC) defines heat exposure as "leaving the whole or part of the body uncovered or unprotected from heat conditions", which is the direct cause of health outcomes.

One measurement of heat exposure is the ambient temperature, while different types of temperatures were applied in published studies. Mean temperature [23, 60, 64] , minimum [24, 57] and maximum temperatures [12, 24] are straightforward and simple temperature indicators. For patients suffering from cardiovascular and respiratory diseases [60, 76], diurnal temperature range can be used as a risk factor. Some other studies used bio-meteorological indices such as apparent temperature [44, 49] and humidex [77]. The combination of air temperature and humidity is a better explanation of the effect of heat exposure on the human body, but it remains difficult to identify a single variable's impacts on health [78]. There exists no absolutely superior temperature measurement so far [79]. The widely used exposure measurements [2, 35, 44, 49, 77-80] are summarized in Table 2.2.

Table 2.2 Types of temperature exposure cited from previous studies

Single measurement	Composed measurement	Other indexes
Ambient Temperature [79]	Apparent Temperature[2] (combination of temperature and dew-point temperature) [44, 49, 79]	Minimum Mortality Threshold temperature (MMT) [4] [2]
Skin Temperature [79]	Minimum Temperature index (high night- time temperature) [78]	Physiological equivalent temperature (PET)[3] [80]
Maximum Temperature [78]	Heat index (air temperature and humidity) [77, 78]	
Minimum Temperature [78]	Humidex[1] (temperature and moisture content) [35, 79]	
Average Daily Temperature [79]		
Outdoor Temperature [79]		

1. Humidex (Canadian Centre for Occupational Health and Safety 2009)

$$Humidex(\deg C) = Mean\ Temperature(\deg C) + 0.5555(6.11E - 10) \qquad (1.1)$$

$$E = \exp(5417.753(\frac{1}{273.16} - \frac{1}{dew\ point\ temperature\ (\deg K)} \qquad (1.2)$$

2. Apparent temperature

$$Tapp = -2.653 + (0.994 \times air\ temperature) + (0.0153 \times (dew\ point\ temperature)^2)\ (1.3)$$

3. Physiological equivalent temperature (PET): the air temperature at which, in a typical indoor setting (without wind and solar radiation), the heat budget of human body is balanced with the same core and skin temperature as under the complex out door conditions to be assessed.

4. Minimum mortality threshold (MMT): temperature with the lowest average mortality rates for a given city.

The heat wave effect is a sub topic of heat exposure measurement. There are three elements to characterize a heat wave: duration, intensity and timing. The heat wave is generally defined according to these elements: 1) duration: a single day [31, 32, 43, 81] or consecutive days' climate conditions [5, 10, 13, 17, 20, 31, 77, 78]; 2) intensity: absolute temperature metric or relative (percentile) temperature threshold for a certain location [5, 13, 17, 20, 31, 32, 43, 77, 78, 81] ; 3) timing: specific seasons and periods of time [5, 13, 18, 21, 23, 54, 56, 82, 83] (Table 2.3 shows some definitions of heat wave in published studies). In most cities, the longer duration and the higher the intensity of heat events, the more severe the observed health outcomes [5, 31], but the exact effects of heat event's intensity and length vary from one location to another. This study will work out the most explanatory indicator of heat wave for the Boston MSA.

Table 2.3 Definitions of heat wave cited from published studies

Heat wave definition (Criteria)	Reference	Study location(city/region, country) and time
August, 2003 (timing)	S. Vandentorren 2006 [18]	France (August 8 - August 13)
Hottest summer ever observed in this area (timing)	J Kyselý 2004 [21]	Czech Republic (1982-2000)
Summer 2003 (timing)	P Michelozzi 2004 [83]	Rome, Italy (June-August, 2003)
June, 2003 (timing)	B Cerutti 2006 [82]	Ticino, Switzerland (2003)
July 29 to August 3, 1995 (timing)	RS Kovats 2004 [23]	Greater London, UK (April 1,1994 - 1 March 31, 2000)
Any day on which the maximum temperature exceeded 36.5 °C (density)	J Diaz 2002 [43]	Madrid Autonomous Region (Jan 1, 1986 - Dec 31, 1997)
The dates of the first and last reported heat-related deaths in California associated with the 2006 heat wave (timing)	Knowlton K 2009 [56]	California, USA (July 15 -August 1, 2006)
July 13 through 19,1995 (timing)	Semenza JC 1999 [54]	Chicago, US (1995)
A time span of at least five days of which each day has a maximum temperature of at least 25°C and of which at least three days reach a maximum temperature of at least 30°C (duration and density)	S Hajat 2002 [20]	Great London, UK (1976-1996)
Three or more consecutive days with Humidex above 40°C (duration and density)	Mastrangelo G 2007 [77]	Veneto Region, Italy (June 1 - August 31, 2002-2003)
Periods of 2 or 4 or more days of continuous temperatures more than 98.5th, 99th, or 99.5th percentile of the community's temperature distribution (duration and density)	GB Anderson 2009 [10]	107 US communities (198-2000)
With a daily minimum temperature higher above 99th percentile (duration and density)	Medina-Ramón M 2007 [81]	50 cities ,US (1989-2000)

Periods of at least two days with Tappmax exceeding the 90th percentile of the monthly distribution or periods of at least two days in which Tmin exceeds the 90th percentile and Tappmax exceeds the median monthly value (duration and density)	D'Ippoliti D 2010 [31]	Europe (Athens,Barcelona, Budapest,London,Milan,Munich,Paris, Rome,Valencia) (1990-2004)
A minimum of 2 or 4 consecutive days with temperature above 98th,99th or 99.5th percentile of daily temperature in whole dataset (duration and density)	Shakoor Hajat 2006 [78]	London,UK(1976-2003), Budapest,Hungary(1970-2000), Milan,Italy(1985-2002)
Apparent temperatures exceeding 40.6°C (duration and density)	Smoyer KE 1998 [32]	St. Louis, Missouri,US,(1980 - 1995)
More than 2 consecutive days with daily mean temperature at or above 98th percentile for the warm season in each city (duration, density and timing)	Ji-Young Son 2012 [17]	Seoul,Bushan,Incheon, Daegu,Daejeon,Gwangju, Ulsan,Overall;Korea,(2000-2007)
Summer with temperatures often exceeding 40°C (duration, density and timing)	AL Hansen 2008 [13]	Adelaide, Australia (January 1, 1995 - December 31, 2006)
More than 2 days with temperature above 95th percentile for the community for 1May through 30 September (duration, density and timing)	Anderson GB 2011 [5]	43 cities, US (1987-2005)

The lag effect [34, 45, 49] explains the association between exposure over previous days and health effect on a particular day. The lag effect of heat exposure plays a big role in detecting the occurrence time of heat events related diseases, because the adverse health outcome of heat exposure may not appear instantly. The length of the lag ranges from one day to a month [25, 44, 50, 61, 64, 66, 84], with shorter lags during warmer seasons and longer lags during cooler seasons [25, 61, 64]. The lengths of the lag also vary according to different types of diseases. For example, a study [50] showed that the majority of hospital admissions for respiratory and cardiovascular diseases happened following the sharp increase in temperature in 0 to 3 days; and a 7-day' lag was applied to assess the impact of temperature on hospital admissions for cardiovascular diseases [57]. This study will select the length of lag according to the heat wave definition and considering the accumulated effect of heat.

2.4 Methodology

2.4.1 Trend analysis

Trend analysis aims at extracting an underlying pattern of behavior in a time series. Identifying the trend of the heat related disease can indicate its variation over long time, because it is hard to conclude whether there exists a definitive upward or downward trend of heat related diseases. On one hand, the potential global warming [81] trend may result in an elevated ambient temperature and increase the intensity and frequency of heat events; therefore there would be more heat related diseases over time. On the other hand, the improving technical support, such as the air conditioning system or housing design, may better protect human health from the fluctuation in climate conditions.

The actual trend of heat related disease is likely to be determined by the comprehensive competition between climate condition and human acclimation. Studies [7, 51, 85] in Europe predicted that around 2 Centigrade warming would occur in the next half century, but the populations' acclimatization to heat are regarded to be incomplete so far [81], thus the warmer climate will threaten human health. The

16

potential acclimatization mechanisms comprise physiological, infrastructural and technological acclimations (e.g. air conditioning and heating system) [85, 86], as well as proper prevention mechanisms (e.g. heat wave warning system etc.). Some studies hypothesized that the mortality associated with warm and humid days would systemically decline over decades [87, 88] due to technical improvements and acclimation [38, 67, 89] . The ultimate impact of climate change will depend upon the extent of population adaptation and the effectiveness of implementation [6, 8, 87, 90].

2.4.2 Association between temperature exposure and health outcomes

The association between temperature exposure and health outcomes [9, 34, 45, 47, 51, 72] can be graphically illustrated by a J-shaped curve, quantitatively explained by the relative risk, and temporally demonstrated by the harvesting effect.

The association between temperature exposure and health outcomes can be illustrated as a J-shaped curve. These J-shaped curves are characterized by their local climate features, while their patterns are similar. There always exists a comfortable zone for each location, within which heat related diseases are less likely to happen [38-40, 46, 47, 59, 91]. Another element of a J-shaped curve is the threshold point. Studies [23, 26, 49, 50] working on heat events related health outcomes in multiple cities found that there is an upper temperature threshold point for each city, beyond which there will be more health risks. A case-crossover study [55] which took the year 1994 as the reference year and the year 1995 as the case year, looked at the 1995's heat disaster in Chicago, pointing out in the reference year, heat related emergency department visits were not significantly correlated with ambient temperature, while in the case year, the heat related emergency visits were highly correlated with ambient temperature, because the ambient temperature usually exceeded the threshold. This finding explains the significance of threshold point in health outcomes. Figure 2.3 was adopted from a study conducted from 1973-1994, in Boston, Chicago, New York, Philadelphia, Baltimore, Washington DC, Charlotte, Atlanta, Jacksonville,

Tampa and Miami [38]. The threshold points vary greatly from the north to the south and from inland cities to coastal cities.

The association between temperature and its negative health effects can be mitigated by technical strategies [4, 92] and affected by other social economic factors [4, 38, 47, 73, 87, 89, 93] . As is shown in Figure2.4, the relative risk for each location may not necessarily increase along with the increase of maximum temperature, because the relative risk is affect by other factors such as people's acclimatization, social status, educational levels, medication, and health conditions [4, 38, 74].
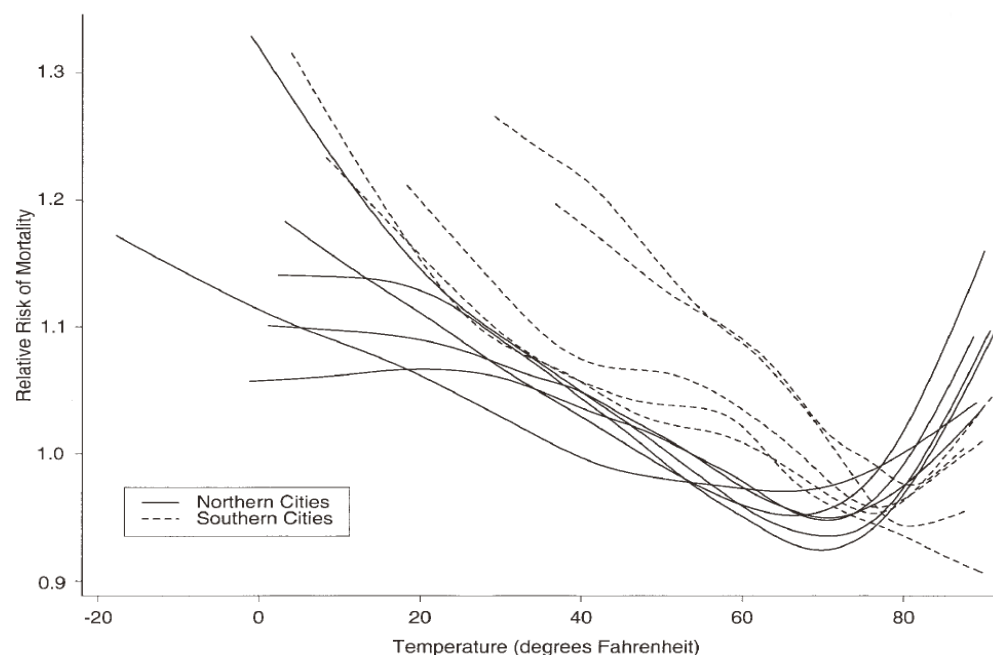
Figure 2.3 Temperature-mortality relative risk functions for 11 US cities, 1973–1994. Northern cities: Boston, Massachusetts; Chicago, Illinois; New York, New York; Philadelphia, Pennsylvania; Baltimore, Maryland; and Washington, DC. Southern cities: Charlotte

The relative risk is the ratio of two risks with the same health outcome but different exposure statuses, which quantifies the association between temperature exposure and health outcomes. Table 2.4 summarizes the relative risk of elevated temperature and the duration of heat wave [34, 36, 38, 39, 44-46, 57, 77, 81, 90].

Table 2.4 Summary of the relative risk of health outcomes cited from previous studies

Location(city/region, country) and time	Relative Risk (95% CI)	Reference
20 US metropolitan areas, seasonal analysis 1992	Per 10°F, **1.15 (1.07-1.24), 1.10 (0.96-1.27), 1.08(0.92-1.26), 1.08 (1.02-1.15)**, and **1.01 (0.92-1.11)** mortality in the Southwest, Southeast, Northwest, Northeast, and Midwest (Mean daily temperature per 10°F adjusted for dew point temperature).	Basu 2005 [36]
3 California regions, 1983-1998	3 °C decrease in maximum temperature or a 3 °C increase in minimum temperature **(1.06-1.11)** for acute myocardial infarction and congestive heart failure, **(1.10-1.18)** for stroke.	Ebi 2004 [57]
9 California counties, May to September 1999-2003	Per 10°F increase in mean temperature, **2.3 (1.0-3.6)** mortality.	Basu 2008 [44]
Veneto Region, Italy, June 1 to August 31 in 2002 and 2003	**1.16 (1.12, 1.20)** hospitalization for heat disease, **1.05(1.03, 1.07)** for respiratory diseases with each additional day of heat wave duration.	Mastrangelo 2007 [77]
43 US cities, 1987-2005	**3.74% (2.29–5.22%)** increase in mortality during heat waves compared with non-heat wave days. Every 1°F increase in heat wave intensity **2.49%** increase in mortality. Every 1-day increase in heat wave duration **0.38%** increase in mortality. **5.04% (3.06–7.06%)** increase in mortality during the first heat wave of the summer ; **2.65% (1.14–4.18%)** increase in mortality during later heat waves, compared with non-heat wave days.	Anderson 2008 [90]
50 US cities in cold (November to March) and warm (May to September) seasons	**5.74 (3.38-8.15)** mortality for extreme heat (Binary variable as extreme heat (range 22-32°C)).	Medina-Ramon 2007 [81]
Sao Paulo, Brazil, Santiago, Chile and Mexico City, Mexico, 1998-2002	**2.69% (-2.06, 7.88)** mortality for Santiago, **6.51% (3.57, 9.52)** for Sao Paulo and **3.22% (0.93, 5.57)** for Mexico City (Same day apparent temperature compared with days at 75th percentile).	Bell 2008 [34]

15 European cities, April-September 1990-2000 (5-11 years depending on data availability for city)	1°C increase above threshold **3.12 (0.60-5.72)** mortality in Mediterranean and **1.84 (0.06-3.64)** in north continental region (Maximum apparent temperature (threshold 29.4°C Mediterranean cities and 23.3°C north-continental cities)).	Baccini 2008 [45]
11 Eastern US cities, 1973-1994	Per 10°F above minimum mortality temperature (65.2-90.3°F) **(1.4-6.7)** mortality.	Curriero 2002 [38]
Bologna, Milan, Rome, Turin, 1997- 2003	(Odds ration) **1.34 (1.27, 1.42)** mortality (30°C mean apparent temperature relative to 20°C).	Stafoggia 2006 [39]
4 Italian cities, June to September, 2003- 2004 and reference period (Roma, Torino, Milano: 1995-2002 and Bologna: 1996-2002)	1°C above threshold **3.2 (1.9-4.6), 5.0 (3.8-6.1), 5.4 (4.3-6.5), 3.8 (2.5-5.0)** mortality for Bologna, Milano, Roma, and Torino, respectively (Daily maximum apparent temperature thresholds (28-32°C)).	Michelozzi 2006 [46]
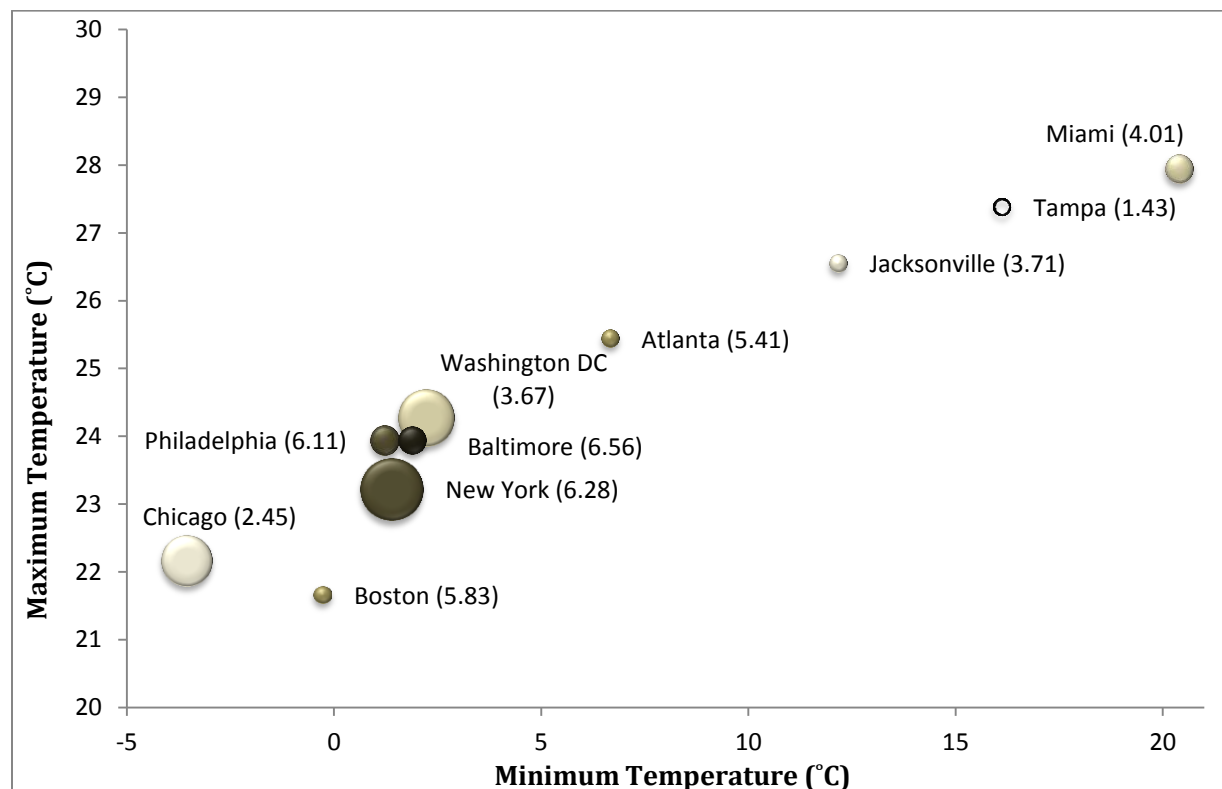
Figure 2.4 The vulnerability of 10 US cities (The data are cited from study [38]: the sizes of bubbles are proportional to their populations in 1980; the bubble's darkness is adjusted according the value of the relative risk, which is the average slope of the J-shaped curve of each city in Figure 2.3)

Comparing Figure 2.4 and Figure 2.1, this study found that the temperature-mortality relative risk was determined by many factors, such as climate patterns and demographics. For example, in Figure 2.4, Chicago and New York have similar population conditions but their relative risks vary greatly. Washington DC and Baltimore have almost identical weather patterns but their relative risks still differ from each other.

Another temporal feature of negative health outcome is the harvesting effect, which is an immediate increase in mortality followed by a decrease in mortality among vulnerable people. Some studies [33, 40, 54-56, 64, 67, 72, 82, 94] focusing on a harvesting effect found that despite the increase of death was consistent with the extreme temperature, the exposure to heat was not the real underlying cause of death. The exact harvesting effect on morbidity has not been fully understood so far.

2.4.3 Time series analysis

A time series is a sequence of observations ordered in time. The distinguishable feature of time series analysis is the explicit recognition of the importance of the observations' order. Time series analysis is useful for identifying seasonal variations, like extreme weather events. Time series method has been widely used in temperature-morbidity related studies [23, 24, 41, 49, 64], applying morbidity counts or rates as the health outcomes and different types of temperature measured at corresponding intervals as the exposure indicators. The daily data was applied more commonly, though weekly or monthly data were also used in some studies [20, 23, 26]. Results are often presented as the change in health outcomes with per unit change in temperature.

The Poisson regression through generalized additive models (GAM) [10, 27, 37, 38, 47] or generalized linear models (GLM) [16, 47, 66] were widely applied to assess the relationship between temperature and morbidity, by adjusting trend effects, seasonality and other periodicity varying factors [14, 61, 64].

Some studies indicated summer or winter and warm or cold seasons to stratify their analyses in order to remove seasonal patterns included in the data [12, 50, 66, 70].

Time series analysis is also a powerful tool to investigate the health effects of air pollutants [3, 16, 35, 66, 70, 71] and their interactions with temperature. The conclusions about the relationship between air pollution, high temperature and health effects are currently inconclusive.

2.5 Seasonality of heat related disease

The seasonality is generally defined as systematic, or repetitive, periodic fluctuations in a variable within the course of a year [95]. Specifically, the disease seasonality is the periodic surge in disease incidence over the course of a year. Adequate understanding of seasonality can improve the accuracy of forecasting systems. Seasonal patterns vary according to types of diseases, locations and subpopulations of interest.

Disease seasonality is characterized by the timing, magnitude and duration [95]. The timing is measured by the position of the maximum and minimum points on the seasonal curve, while the magnitude is correlated with the maximum and minimum values on the seasonal curve. The duration is defined by a shape of a seasonal pattern which reflects how fast a temporal curve reaches its peak and declines to nadir over a course of a full cycle.

The seasonality of a certain disease is correlated with other natural processes. Generally, meteorological factors appear to be critically linked to seasonal patterns of a disease. In fact, the periodicities of environmental factors and disease incidences may not perfectly synchronize and the difference between the time point in which exposure and disease incidence peak is called the lag.

The occurrence of a heat related disease is naturally and directly correlated with ambient temperature and summer seasons, but duration of heat wave is much shorter than the length of a summer season. In addition, there may be some lags between the occurrence of heat wave and the occurrence of negative

health outcomes. Therefore, it is necessary to independently discuss the seasonality of heat related disease.

The calendar reflects the cyclic rhythm of nature and synchronizes with social events. Therefore, describing the seasonality of heat related disease with a calendar can give suggestions on the planning of social events, on the operation of engineering facilities and on the special medical preparations for social events during the hot periods of the year.

2.6 Selection of distribution

The distribution is the probability of a particular value or value range of a variable. The health outcomes are mainly explained as either hospitalizations or mortality counts in past studies, and both of them are discrete count data. Many studies assumed that their health outcome data [3, 20, 64, 81] followed a Poisson distribution and very few studies applied negative binomial distribution to explain their data [22]. It is worth pointing out that the software development limits the selection and application of certain distributions in practice. Selecting a certain distribution without verifications may affect the width standard error and significance level of the estimate coefficients for a regression model; therefore this study will evaluate the selection of statistical distribution in detail.

2.7 Problem discussion and the goal of this study

2.7.1 Selection of study location

The study location is the Boston-Cambridge-Quincy MA NH Metropolitan Statistical Area (Boston MSA). Selecting a location with a standard and official definition can facilitate the comparability and applicability of the study conclusion. What's more, the metropolitan statistical area represents the principal population center, which is the prerequisite for study significance. Boston's cold winters and susceptibility to hot, humid summers [35] make it an exemplary location to investigate the association

23

between high ambient temperature and its negative health effects. As is shown in Figure1.5, the areas in light blue in the map stand for the Boston MSA.
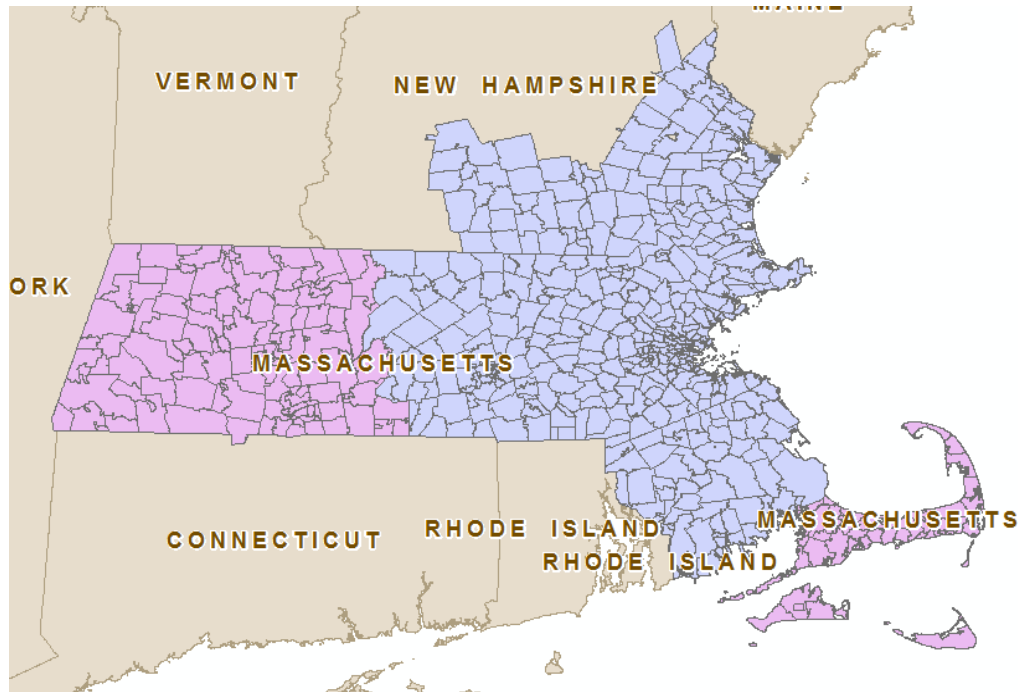
Figure 2.5 The map of study location (Data source: GIS online database. Other states surrounds Massachusetts are in light grey; the State of Massachusetts is mapped in light purple and the Boston-Cambridge-Quincy MSA is mapped in light blue)

2.7.2 Selection of health outcome

The hospitalization is an appropriate health outcome to analyze in order to detect the negative effect of heat events, though other health outcomes are also epidemiologically meaningful. The strengths of utilizing the hospitalization as the measurement of health outcomes are: 1) the hospitalization can cover more cases than mortality, thus the correlation between ambient temperature and hospitalization is expected to be more statistically significant; 2) the record of hospitalization is readily accessible, and our familiarity with it allows for a better understanding of the data quality, reliability and potential bias.

This study extracted hospitalizations with diagnostic codes ICD-9-CM 992.0 – 992.9 from 1991 to 2006 from the CMS database. ICD-9-CM 992.0 – 992.9 explains the health effects of heat and light. Diseases such as burns (ICD-9-CM 940.0 – 949.5), disease of sweat glands due to heat (ICD-9-CM 705.0 – 705.9),

malignant hyperpyrexia following anesthesia (ICD-9-CM 995.86) and sunburn (ICD-9-CM 692.71, 692.76 – 692.77) were not included. Selecting hospitalization records with ICD-9-CM 992.0 – 992.9 as health outcome is expected to provide large enough sample sizes and help investigate a significant correlation between heat events and their direct health outcomes.

2.7.3 Selection of exposure measurement

In this study, both daily maximum and minimum temperatures are applied as predictors of exposure, though they are highly correlated. The estimate coefficients for these two indicators are in the same order of magnitude and quite close to each other, and can be explained as the weights given to these two variables, therefore a weighted average daily temperature is applied in the regression model. The weighted average daily temperature is a more accurate measurement of daily temperature exposure than arithmetic average temperature, because the maximum temperature and minimum temperature are good indicators of the daytime and nighttime exposures respectively; and the proportions of people's daytime temperature exposure and their nighttime temperature exposure are not identical.

2.7.4 Selection of statistical method

This study is about building up a regression model of the daily hospitalization counts over 16 years (1991 –2006). Time series analysis was used to adjust the linear trend effect, yearly effect, seasonality, weekday effect and to investigate the effects of ambient temperature on HSHs.

Regression method is an integral part of time series analysis, and mostly deals with linear models and continuous values. However, in real-life cases, some records are not continuous and a linear model is no longer appropriate. Therefore, it is necessary to extend a linear model to a generalized linear model (GLM) which has a conditional distribution from exponential family of the response variable, a linear function of explanatory variables, and a link function transforming the expectation of the response

variable to the linear predictor. The GLM is applied to detect the relationship between heat related

hospitalization counts and ambient temperature in this study.

2. **Statistical features of hospitalization records**

In probability theory, the distribution is defined by its sets of possible values and the probability assigned to each of those possible values, and each specific family of distribution can help understand different processes and reveal and estimate information conveyed by the data.

3.1 Discrete distribution

A discrete distribution is described by a set of random variables that take only finite set of values. A discrete probability distribution comes into play when there are a finite number of discrete possible events.

Count response models are a subset of discrete response models, aiming at explaining the number of occurrences, or counts, of an event. Two examples of count models are the Poisson and negative binomial model.

3.1.1 Poisson model

The Poisson distribution is a discrete probability distribution that describes the probability of observing a given number of events in a fixed interval of time or space if these events occur with a known average rate and independent of the time since the last event.

The Poisson distribution defines a probability distribution function for non-negative counts or outcomes. The Poisson distribution for the number of occurrences of the event, with density

$$\Pr[Y = k] = \frac{e^{-\mu}\mu^k}{k!} \quad (3.1)$$

Where Y is the number of events in a unit interval of time or in a unit distance $k = 0,1,2\cdots$, μ is the intensity or rate parameter.

The expectation (Equation 3.2) and variance (Equation 3.3) of a Poisson distribution are equal to μ, which is one of the characteristics of a Poisson distribution.

27

$$E[Y] = \mu \qquad (3.2)$$

$$Var[Y] = \mu \qquad (3.3)$$

The regression method is an analysis process, which utilizes an equation to relate the mean response of a variable to a set of explanatory variables. As is explained in part 2.7.4, the Generalized Linear model (GLM) is a generalization of ordinary linear regression that allows for response variables to have an exponential distribution. The GLM generalizes a linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. The term "Poisson regression model" in the following chapters is referred to a GLM with response variables coming from a Poisson distribution.

For count data, the Poisson regression model is considered to be the base count response regression model [96]. A key assumption of the Poisson regression model is that the variance equals the mean. However, the actual data often exhibit overdispersion, with a variance larger than the mean. Overdispersion might be due to unobserved heterogeneity: the data may be generated from mixture distributions; and it might be due to violation of the assumption of independence of events which is implicit in the Poisson process. Overdispersion may cause standard errors of the estimates to be underestimated, thus a variable may appear to be a significant predictor while it is in fact not.

Another limitation of the application of the Poisson model is that real-life count data have far more zeros than expected by the Poisson model's distributional assumptions. The negative binomial model is expected to deal with overdispersed data, while the hurdle model and the zero-inflated model are expected to cope with the dataset with excess zeros.

3.1.2 Negative binomial model

The negative binomial distribution is a discrete probability distribution describing the number of successes in the sequence of Bernoulli trials before a specified numbers of failures occur. The name

"negative binomial" comes from applying the general form of the binomial theorem with a negative exponent. The negative binomial distribution can be seen as an extension of the Poisson model that allows for a greater variance.

There are two methods to derive negative binomial distribution: 1) the negative binomial distribution can be arisen as a gamma mixture of Poisson distribution; 2) the negative binomial probability function can be thought of as the probability of observing y failures before the rth success in a series of Bernoulli trials.

$$\Pr[Y = y] = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\cdot y!} \cdot \frac{\mu^{y}\cdot\theta^{\theta}}{(\mu+\theta)^{y+\theta}} \quad (3.4)$$

where μ is the mean value and θ is the shape parameter.

The negative binomial distribution allows the expectation (Equation 3.5) and variance (Equation 3.6) to be different. That's why the purpose of negative binomial model is assumed to deal with overdispersed Poisson data.

$$\mathrm{E}[Y] = \mu \qquad (3.5)$$

$$V[Y] = \mu + \frac{\mu^{2}}{\theta} \quad (3.6)$$

The negative binomial distribution is a member of the two-parameter exponential family, and was not considered suitable for entry into the mainstream family of GLM because of the complexity of the ancillary parameter [97], and until 1989 McCullagh and Nelder recognized that the negative binomial distribution could be considered in a GLM.

3.1.3 Hurdle model (zero altered models)

The hurdle model, developed by Mullahy in 1986, is capable of dealing with excess zero counts, which are not allowed by the distributional assumptions of a Poisson model and negative binomial model [96,

97]. A hurdle model is to partition a model into two parts: 1) a binary process generating positive counts and zero counts; 2) a process generating only positive counts. There have been nine commonly used hurdle models: the binary part modeled by logit, probit, or complementary loglog, and the count part modeled using Poisson, geometric, or negative binomial. The notion of hurdle comes from considering the data as being generated by a process that commences generating positive counts only after crossing a zero barrier or hurdle. Until the hurdle is crossed, the process generates a binary response. The nature of the hurdle is left unspecified, but may simply be considered as the data having a positive count. In this case, the hurdle is crossed if a count is greater than zero.

3.1.4 Zero-inflated (Poisson) model

Like the hurdle model, the zero-inflated model introduced by Lambert in 1992 provided another approach to accounting for excess zero counts [96, 97]. Like the hurdle mode, the zero-inflated model also considers two distinct sources of zeros: 1) generated from individuals who do not enter the counting process; 2) generated from individuals who do enter the counting process but result in a zero outcome. The zero-inflated Poisson model assumes that some zeros occur through a Poisson process, but others are not even eligible to have the event occur. Therefore there are two processes at work: one is to determine whether the individual is eligible for a non-zero response, and the other that is to determine the count of that response for eligible individuals. The tricky part is that both processes can result in a zero count, so it is hard to distinguish the coming process of zeros. The zero-inflated Poisson model simultaneously fits two separate regression models. One is a logistic or probit model that models the probability of being eligible for a non-zero count. The other models the size of that count.

Unlike the hurdle model, the zero-inflated model provides for the modeling of zero counts using both binary and count processes. The hurdle model separates the modeling of zeros from the modeling of counts, entailing that only one process generates zeros.

Zero inflated models should be tested to determine whether they are statistically different from their base model (i.e. Poisson or negative). A Vuong test will be applied to evaluate the model, whether the data are Poisson, negative binomial or if the excess zeros come from a different generating process.

3.2 Statistical features of actual data

3.2.1 Statistical features of the raw data

It is helpful to investigate the distribution of actual data by plotting them on a histogram of daily hospitalizations counts (Figure 3.1). Random Poisson values were generated using the mean values of the actual data and its subsets, in order to compare how much the actual data derive from the data set following a Poisson distribution.
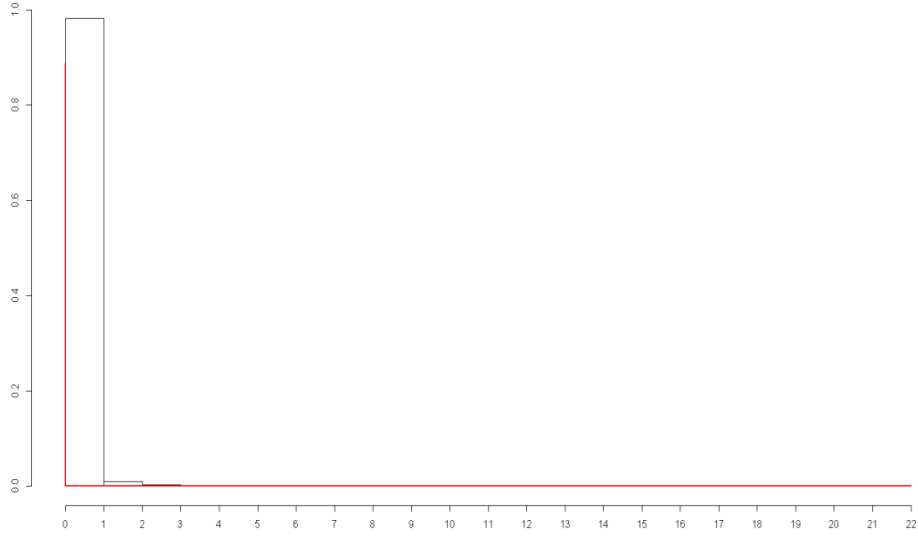
Figure 3.1 Numbers of cases on a given day

Table 3.1 The summary of statistical features of actual data and simulated data

	data	Mean	Median	Variance	Standard deviation	Skewness	Kurtosis	Var/mean
whole data (5844)	actual	0.12	0	0.46	0.68	15.56	370.00	3.83
	simulated	0.13	0	0.12	0.36	2.75	7.11	0.92
counts>=1 (426)	actual	1.65	1	3.84	1.96	6.12	48.16	2.33
	simulated	1.55	1	1.48	1.22	0.76	0.46	0.95
counts>=3 (49)	actual	5.45	4	16.3	4.04	2.51	6.22	2.99
	simulated	5.31	5	5.13	2.27	0.12	-0.80	0.97
counts>=5 (20)	actual	8.50	7	24.3	4.93	1.51	1.06	2.85
	simulated	8.65	8.5	12	3.47	0.65	0.00	1.39

From the table 3.1 above, this study summarizes the features of hospitalization data below:

- They are highly skewed to the right; the skewness is as high as 15.56. This phenomenon is closely related to the 5417 zeros in the dataset, which takes up 92.7% of the data. When excluding the zero counts, the skewness value became smaller, but it still greater than the theoretical value of a Poisson distribution. The proportion of extremely high HSHs is small, while its magnitude is great, which can be explained by the long-tailed shape of the histogram in Figure 3.1.

- The Kurtosis value is extremely high (as high as 370), which can be explained by the large proportion of zeros. The kurtosis of the sub- dataset with more than 3 counts is a negative value, which indicates that the distribution of this sub-dataset shows a rectangle shape instead of a sharp peak.

- The actual data's variance is much greater than the mean value, which provides evidence of overdispersion phenomena. A negative binomial model may handle the overdispersion better than a Poisson model.

- The hospitalizations have more zeros that can be explained by a Poisson model. A zero-inflated (Poisson) model may be an alternative to explain the excess zeros.

- The smaller the sample size, the less stable the data might be: the mean value of twenty Poisson data is smaller than the variance.

3.2.2 Statistical features of the transformed data

Taking logarithm transformations of HSHs data with different bases (i.e. log10(), ln(), log2())to test whether the dataset can be stabilized, whether the transformed data set can be better fitted by a certain distribution and whether one base works better than the others. For the whole data set, a constant (i.e. 1) has been added to the raw records (Table 3.2) to mathematically avoid the zeros in the logarithm transformations. The transformation formulae are log2(x+1), log2(x+1), and ln(x+1).

Table 3.2 The summary of statistical features of transformed data

	days	Mean	Median	Variance	Standard deviation	Skewness	Kurtosis	Var/mean
log2(x+1)	5844	**0.091**	0	**0.128**	0.400	4.837	29.700	1.410
Ln(x+1)	5844	**0.063**	0	**0.061**	0.200	4.837	29.700	**0.970**
log10(x+1)	5844	0.028	0	0.012	0.100	4.837	29.700	0.430
log2(x)	426	0.388	0	0.61	0.800	2.282	5.453	1.570
Ln(x)	426	**0.269**	0	**0.293**	0.500	2.282	5.453	**1.090**
log10(x)	426	0.117	0	0.055	0.200	2.282	5.453	0.470
log2(x)	49	2.213	2	0.551	0.700	1.310	1.116	0.250
Ln(x)	49	1.534	1.390	0.265	0.500	1.310	1.116	0.170
log10(x)	49	0.666	0.600	0.050	0.200	1.310	1.116	0.080
log2(x)	20	2.920	2.810	0.450	0.700	0.980	-0.300	0.150
Ln(x)	20	2.020	1.950	0.220	0.500	0.980	-0.300	0.110
log10(x)	20	0.880	0.850	0.040	0.200	0.980	-0.300	0.050

- From the table above, this study finds the whole data set and the subset with one count and more, may be fitted by a Poisson distribution after natural log transformation, because their mean values and variances are quite close. However, the equidispersion feature cannot be found in other subsets of data.

  The data set has excess zeros and I predict that not a single distribution can completely explain the sources of zeros included in the dataset. Thus, the zero-inflated Poisson model or hurdle model may be an alternative.

- The natural log-transformed data may be better fitted by a Poisson distribution than the raw dataset. This tells us that the application of a GLM for detecting the relationship between temperature and HSHs is more appropriate than a simple liner model.

- When daily hospitalization counts are greater than three, and the data's mean and variance is quite different, this study predicts the sample size is also a factor that affects the stability of the data set.

- This study finds another interesting phenomenon; the skewness and Kurtosis values are constant no matter what kinds of transformations have been taken. This study cannot find evidences from the theoretical formula to support the "transformation-free" feature of skewness and Kurtosis.

3.3 Potential applications of discrete distribution

3.3.1 Potential application of Poisson model

Poisson regression is the base count response regression model and it is quite natural to start with Poisson regression, when discussing the hospitalization counts. However, the actual data may violate the assumptions of the Poisson Model. The actual data are highly skewed to the right and the number of

zeros included in the data is more those can be explained by a Poisson model. The equidispersion assumption of a Poisson distribution would be violated. The overdispersed data may be better fitted by a negative binomial model. The excess zeros may be fitted by other mixed models such as the hurdle or zero-inflated model. This study starts regression analyses with a Poisson regression model.

3.3.2 Potential application of negative binomial model

Theoretically, the overdispersed data could be better fitted by a negative binomial model. However, published studies applied Poisson regression models to explain their data far more frequently than a negative binomial regression model. I don't think that those actual data can successfully support all the assumptions of a Poisson distribution [97]: 1) The probability of observing a single event over a small interval is approximately proportional to the size of that interval; 2) The probability of two events occurring in the same narrow interval is negligible; 3) The probability of an event occurring within a certain interval does not change over different intervals; 4) The probability of an event occurring in one interval is independent of the probability of an event occurring in any other non-overlapping interval. If either of these last two assumptions is violated, the extra variation (overdispersion) will take place. The availability of software support may limit the selection and application of certain distributions in practice. Fewer software programs support the negative binomial regression; what's worse their performances cannot guarantee the model quality. Maybe that's why a negative regression model was applied less frequently. This study compares the Poisson and negative binomial models step by step and explains their performances in the R software environment.

3.3.3 Potential application of hurdle model

The hurdle model may better accommodate the dataset with excess zeros than a Poisson model can explain. The application of a hurdle model means assuming that the big set of zero hospitalization comes from a separate distribution, therefore the hurdle model may better fit our actual data mathematically.

The zero and non-zero counts explain the probability of occurrence of HSHs. When considering the causal factor of heat stroke related hospitalizations, the hurdle can be illustrated as a temperature threshold: no hospitalization will occur if temperature is below this threshold. In other words, the distribution from which the hospitalizations come was determined by the ambient temperature conditions, thus a hurdle model might be a biologically acceptable alternative. There are two challenges to applying the of hurdle model: 1) it is hard to verify which distribution to use on a certain day: 2) there is a lack of adequate software support. This study applied hurdle regression package in R software, even though this package has some bugs that greatly affect the model quality.

3.3.4 Potential application of zero-inflated model

The zero inflated models apply the same logic as a hurdle model to categorize the counts into zero and non-zero parts, and try to explain the sources of zeros. Detecting the sources of zeros in actual data is a theoretical challenge. Actually, we can never know the real distribution of actual data, thus this study still regards the zero-inflated model as a mathematical method to improve the regression model.

This study applied the hurdle and zero-inflated models to conduct regression analyses using the same combination of predictors as that in a Poisson model to check how much improvement could be achieved with the application of the mathematical tools.

3. The time series analysis of hospitalizations and calendar effects

4.1 Background

Time series analysis is a powerful tool for detecting the variation of health outcomes according to time. A time series can be decomposed into three elements: 1) Trend: long term movements of the mean; 2) Seasonal effects: cyclical fluctuations related to the calendar; 3) Weekday and holiday: the effect of daily life on heat stroke related hospitalizations (HSHs); and 4) Residuals: other random and systematic fluctuations.

This chapter aims to explain the time series of heat stroke related hospitalizations (1991 to 2006) with calendar information to test the hypothesis: whether or not the HSHs demonstrate an evident trend, well-pronounced seasonality, or some annual variations in intensity and peak time.

4.2 Methodology

Regression analysis is an integral part of time series analysis. The ordinary linear regression relates the mean response of a variable to a set of explanatory variables by means of a linear equation. In many cases, the linear regression is based on the assumption of data's normality and independence. As is explained in Chapter 3, the HSHs are non-normal, highly skewed and zero-excessed, thus applying the ordinary linear regression may lead to highly inaccurate results. The challenge may be resolved by extending the generalized linear model methodology to time series analysis.

The generalized Linear Model (GLM), introduced by Nelder and Wedderburn (1972), hypothesizes that the response variable of the model has the variance that is reflected by a member of the single-parameter exponential family of probability distribution [96-98]. The GLM can be understood as an extension of the linear modeling process which allows models to be fitted to data following non-normal distributions, such as the Poisson and Binomial distributions. A GLM is consisted of three components: a probability distribution from the exponential family, a linear predictor; and a link function which

associates the mean of the distribution to the linear predictor [98]. The GLM relaxes the requirement of equality or constancy of variances that is required for the hypothesis test in traditional linear models and is particularly useful when the errors are not normally distributed and the variance is not constant.

In this chapter, this study conducts time series analysis of HSHs, from Jan 1st 1991 to Dec 31st 2006, using a generalized linear model with both Poisson and Negative binomial distribution with the adjustment for calendar effects (i.e. the linear trend, the year effect, the seasonality, and the weekday effect).

4.3 Calendar effects

A calendar is a system of organizing units of time for the purpose of reckoning time over extended periods. The principal astronomical cycles form the astronomical bases of a calendar's unit: 1) year: based on the revolution of the Earth around the Sun; 2) month: based on the revolution of the Moon around the Earth; 3) day: based on the rotation of Earth on its axis.

The seasonality of HSHs (ICD-9-CM 992.0 – 992.9) can be explained by calendar information: the linear (long-term) trend, the year effect, the harmonic terms of seasonality, and the weekday and holiday effects. Both Poisson and negative binomial distributions are compared in those following steps.

The Poisson Generalized Linear Regression Model with calendar effects is explained in Equation 4.1:

$$Y_t = exp^{(\beta_0 + \beta_1 \, (linear \, trend) + \beta_2 \, (year \, effect) + \beta_3 \, (seasonality) + \beta_4 \, (weekday \, effect))} \quad (4.1)$$

In Equation 4.1: Y is the daily hospitalization count from Jan 1st, 1991 to Dec 31st 2006; t is the date of the hospitalization record; β_0 is the intercept for the regression model; β_1 is the coefficient for the linear trend; β_2 is the vector of coefficients for the year effect; β_3 is the vector of coefficients for the seasonality; β_4 is the vector of coefficients for the weekday and holiday effects.

All data analyses were conducted with R 12.5.2. Statistical significance was based on α=0.05.

4.3.1 Linear trend

A trend reflects the properties of a time series at the highest level of aggregation for a given period. The simplest form is linear trend, which is described as a straight line along the points in a time series plot. In this study, the linear trend is expressed as the sequence of day from January 1st, 1991 to December 31st 2006, 5844 days in total. The sequence variable ranging from 1 to 5844 was added to the regression model (Equation 4.2), the linear trend term is expected to minimize the residual deviance.

$$Y_t = exp^{(\beta_0 + \beta_1 \, (linear \; trend))} \; (4.2)$$

Table 4.1 The estimate coefficients of the indicator of linear trend

	Model	Estimate (β_1)	Std.Error	P-value	Null deviance	Residual deviance	Explained variability	AIC
day	Poisson	-0.0001	2.25E-05	0.0002	4139.3	4124.9	0.35%	5085.1
size=5844	Negative binomial	-7.93E-05	3.47E-05	0.0223	1366.4	1360.8	0.41%	3929.6
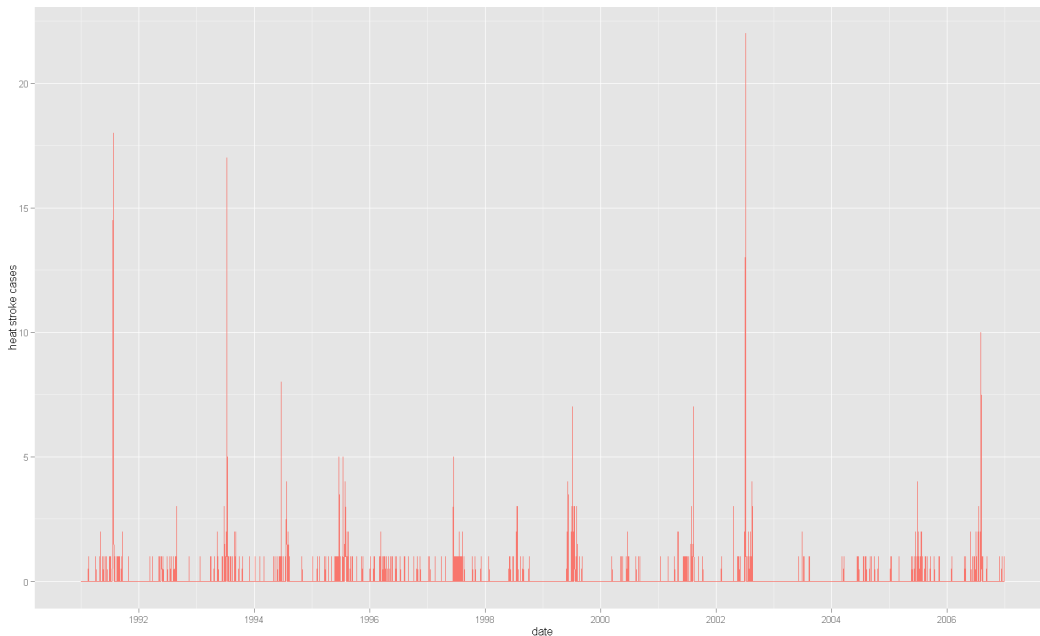
Figure 4.1 The time series plot of daily HSHs from Jan. 1st 1991 to Dec. 31st 2006
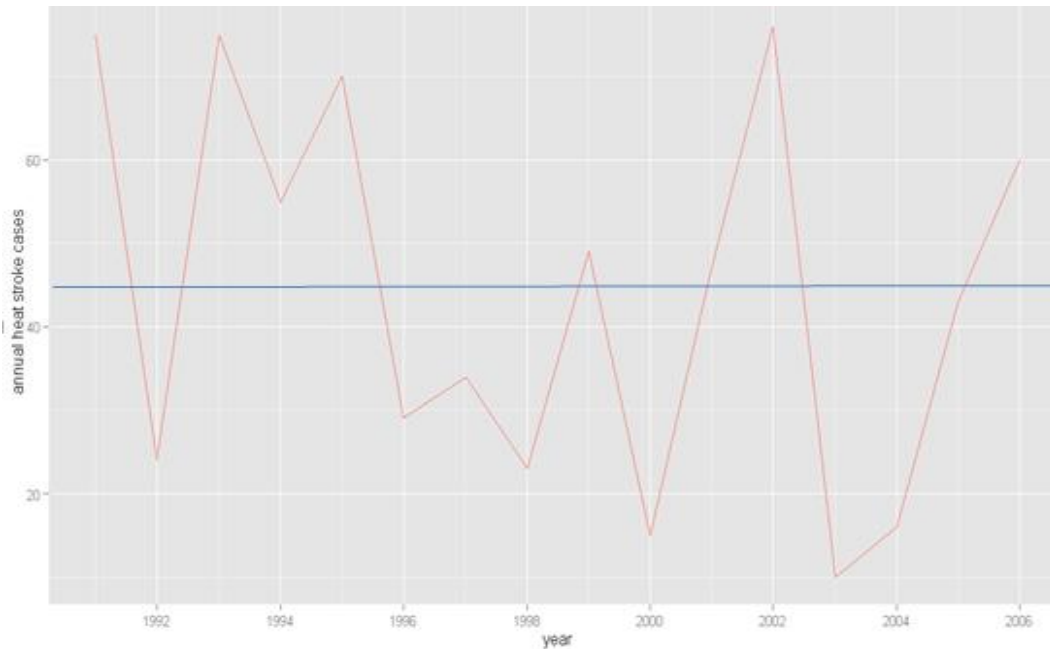
Figure 4.2 The time series plot of yearly HSHs from 1991 to 2006

As is shown in Table 4.1, the estimate coefficients for the linear term are close to zero, which indicate that choosing January 1st, 1991 as the reference day, the daily hospitalizations exhibits very weak downward trend. In Figure 4.1, it is hard to visualize an evident trend for the 16 years' HSHs on the daily basis. The heat stroke related hospitalization counts vary greatly when aggregated to the year level (Figure 4.2). It is more proper to add an indicator variable of each year to the daily based regression model than conducting the analyses on the year level, because the yearly aggregation will greatly mask the seasonality of HSHs in the time course of a year and shrink the sample size.

4.3.2 Year effect

The analysis of year effect is a good adjustment for systematic fluctuation. As is shown in Figure 4.2, HSHs show evident yearly variation when they are aggregated to year level, therefore it is necessary to have an indicator of each year to describe the yearly variation which may be brought by some unknown systematic errors.

41

$$Y_t = exp^{(\beta_0 + \beta_2 (1991,1992,1993 \cdots,2005))} \quad (4.3)$$

Dummy variables of year 1991-2005 (i.e. taking year 2006 as the reference year) were added to the daily

based regression model (Equation 4.3), where "1" represents the presence of the year of interested and

"0" represents this year's absence. In Table 4.2, the signs of estimate coefficients for each year are not

uniform (i.e. both negative and positive signs exist), which indicate that there is no monotonic upward

or downward trend for heat stroke related hospitalizations even on year level. Dummy variables of years

indicate some potential yearly systematic errors and trend effect, even though this study is unable to

detect those bias right now.

Table 4.2 The estimate coefficients of the regression models with the indicator of year effect

	Poisson Model			Negative Binomial Model		
	Estimate	Std.Error	P-value	Estimate	Std.Error	P-value
constant	-1.7973	0.1291	< 2e-16	-1.7973	0.2085	< 2e-16
1991	0.2149	0.1732	0.2147	0.2149	0.2888	0.4568
1992	**-0.9273**	0.2415	0.0001	-0.9273	0.3342	0.0055
1993	0.2149	0.1732	0.2147	0.2149	0.2888	0.4568
1994	-0.0953	0.1867	0.6098	-0.0953	0.2971	0.7485
1995	0.1601	0.1754	0.3613	0.1601	0.2901	0.5811
1996	-0.738	0.2262	0.0011	-0.738	0.3232	0.0224
1997	-0.6061	0.2167	0.0052	-0.6061	0.3168	0.0557
1998	**-0.9671**	0.2452	8.03E-05	-0.9671	0.337	0.0041
1999	-0.2108	0.1926	0.2737	-0.2108	0.3008	0.4835
2000	**-1.3973**	0.2887	1.30E-06	-1.3973	0.3697	0.0002
2001	-0.2525	0.1948	0.195	-0.2525	0.3022	0.4036
2002	0.2281	0.1727	0.1864	0.2281	0.2885	0.4291
2003	**-1.8000**	0.3416	1.36E-07	-1.8000	0.4124	1.27E-05
2004	**-1.3327**	0.2814	2.17E-06	-1.3327	0.364	0.0003
2005	-0.3414	0.1998	0.0875	-0.3414	0.3055	0.2638
Null deviance			4139.3		1483.9	
Residual deviance			3941.4		1386.7	
Variability explain by the model			4.78%		6.55%	
AIC			4929.5		3870.4	

As is highlighted in Table 4.2, the year 1992, 1998, 2000, 2003, and 2004 have greater estimate coefficients, which indicate that these years' hospitalizations are sharply different from their previous years. The great negative sign indicates a sharp decrease in HSHs from the previous year, while the great positive sign indicates a sharp increase in HSHs comparing with the previous year on the year level.

4.3.3 Seasonality

The seasonality of a disease refers to the systematic periodic fluctuations in disease counts within the course of a year. Like other diseases, the seasonality of heat stroke related disease can also be characterized by the timing, magnitude and duration of a seasonal increase. The analysis conclusion of seasonality is greatly influenced by the selection of temporal resolution. The use of daily time series enables the investigation of significant differences in the seasonal peaks of disease. The effective analysis of seasonality also requires data collected over a long period with sufficient frequency. A parametric analysis using harmonic regression method is conducted below. Generally, harmonic regression helps reveal the point in time when a seasonal curve reaches its maximum; absolute and relative amplitudes of the peak; and the duration of a seasonal increase with a defined shape of a curve.

$$Y_t = exp^{(\beta_0 + \beta_3(sin(2\pi\omega t), cos(2\pi\omega t), sin(4\pi\omega t), cos(4\pi\omega t))}$$ (4.4)

In Equation 4.4, π is the constant; ω is the frequency-sequence of the day divided by the number of cycles. This study uses 365.25 as the number of cycle in order to consider the effect of a leap year. The terms $sin(2\pi\omega t)$ and $cos(2\pi\omega t)$ model the annual oscillation cycle, while the $sin(4\pi\omega t)$ and $cos(4\pi\omega t)$ model the half-year oscillation cycle.

Table 4.3 The estimate coefficients of regression models with the indicator of harmonic term

	Poisson Model			Negative Binomial Model		
	Estimate	Std.Error	P-value	Estimate	Std.Error	P-value
$sin(2\pi\omega t)$	-2.30E+00	1.06E-01	< 2e-16	-2.0396	0.114	< 2e-16
$cos(2\pi\omega t)$	-3.22E-01	6.77E-02	1.99E-06	-0.2121	0.084	0.0116
Null deviance			4139.3		2037.4	
Residual deviance			3198.7		1466.8	
Variability explain by the model			22.72%		28.01%	
AIC			4160.8		3472.1	
	Poisson Model			Negative Binomial Model		
	Estimate	Std.Error	P-value	Estimate	Std.Error	P-value
$sin(4\pi\omega t)$	1.2059	0.0689	< 2e-16	1.1653	0.089	< 2e-16
$cos(4\pi\omega t)$	0.5064	0.0604	< 2e-16	0.4574	0.0823	2.68E-08
Null deviance			4139.3		1662.4	
Residual deviance			3685.4		1426.1	
Variability explain by the model			10.97%		14.21%	
AIC			4647.6		3725.1	
	Poisson Model			Negative Binomial Model		
	Estimate	Std.Error	P-value	Estimate	Std.Error	P-value
$sin(2\pi\omega t)$	-1.6062	0.1003	1.39E-10	-1.591	0.1068	< 2e-16
$cos(2\pi\omega t)$	0.1014	0.1141	< 2e-16	0.0756	0.1177	0.521
$sin(4\pi\omega t)$	0.7511	0.0809	< 2e-16	0.6979	0.0944	1.44E-13
$cos(4\pi\omega t)$	0.5554	0.0866	< 2e-16	0.4825	0.0988	1.03E-06
Null deviance			4139.3		2142.8	
Residual deviance			3067.7		1462.9	
Variability explain by the model			25.89%		31.73%	
AIC			4033.8		3397.5	

When $sin(2\pi\omega t), (2\pi\omega t)$, $sin(4\pi\omega t)$ and $cos(4\pi\omega t)$ are applied together, the seasonality of heat stroke related hospitalizations can be depicted at high significance level. The next step is to test the weekday and holiday effects on hospitalizations.

Using the estimates of the amplitude and the phase angle can express the proposed characteristics of the seasonality, such as the magnitude and the peak time [67]:

The average maximum value on the seasonal curve of HSHs:

$$max\{Y(t)\} = exp\{\beta_0 + \gamma\} \quad (4.5)$$

The average minimum value on the seasonal curve of HSHs:

$$min\{Y(t)\} = exp\{\beta_0 - \gamma\} \quad (4.6)$$

The average peak timing, the position of the maximum point on the seasonal curve of HSHs:

$$P = 365(1 - \psi/\pi)/2 \quad (4.7)$$

In Equation 4.5, 4.6, and 4.7, β_1 and β_2 are the estimated coefficients for the terms $sin(2\pi\omega t)$

and $cos(2\pi\omega t)$, $\gamma = a(\beta_1^2 + \beta_2^2)^{\frac{1}{2}}$, where $a = -1$ when $\beta_2 < 0$ and $a = 1$ otherwise;

$\psi = -arctan(\beta_1/\beta_2)$, with $-\frac{\pi}{2} < \psi < \frac{\pi}{2}$.

Table 4.4 The characteristics of seasonality estimated by the amplitude and the phase angle

	Poisson Model	Negative binomial Model
β_0	-3.1767	-3.0343
Average maximum value	1.0227	1.0110
Average minimum value	0.0041	0.0062
Average peaking time (Julian day/Date)	191/July 9th	189/July 7th

From Table 4.4, this study finds that the regression model with harmonic terms (e.g. $sin(2\pi\omega t)$ and $cos(2\pi\omega t)$) cannot explain the actual magnitude of HSHs. The estimated average peak time is July 9th, which can explained as the average peak time for the 16 years, because this is the overall model for the 16 years' records without including information about year variations and the estimate coefficients are adjusted by all 16 years' records.

The regression model with $sin(2\pi\omega t)$ and $cos(2\pi\omega t)$ was run for each year separately in order to detect the more accurate peak time and intensity for each year. As is explain in Table 4.5, the annual peak time and intensity vary greatly from one year to another.

Table 4.5 The annul peak and intensity of HSHs estimated by regression models for each year

Year	Julian day	Date	Minimum value	Maximum value
1991	198	17-Jul	0.0006	0.9455
1992	184	3-Jul	0.0090	0.1781
1993	190	9-Jul	0.0023	0.8361
1994	182	1-Jul	0.0018	0.6075
1995	193	12-Jul	0.0091	0.6441
1996	104	14-Apr	0.0338	0.1441
1997	187	6-Jul	0.0121	0.2563
1998	199	18-Jul	0.0006	0.2613
1999	186	5-Jul	0.0000	1.0044
2000	166	15-Jun	0.0013	0.1469
2001	198	17-Jul	0.0059	0.4362
2002	190	9-Jul	0.0001	1.0642
2003	192	11-Jul	9.04E-08	0.1813
2004	204	23-Jul	0.0058	0.1198
2005	190	9-Jul	0.0069	0.3828
2006	201	20-Jul	0.0044	0.6029

4.3.4 Weekday and holiday effects

A week is defined based on days and also the standard time period used for cycles of work days and rest days, therefore a week can be explained as the link between natural phenomena and daily life. The weekday effects may have substantial and multifaceted effects on the public's health, especially in the U.S., because a majority of the holidays in the US have floating dates: they occur on the same day of the week each year. Holidays are usually correlated with influx of travel around long weekends and changes of regular food consumption patterns. Take the example in Massachusetts: the Patriots' Day, the 3rd Monday in April; there is a Marathon race on that day, when the temperature turns to be warmer in April. All the above factors may result in the increase in heat stroke and dehydration cases that day.

$$Y_t = exp^{(\beta_0 + \beta_4(Monday, Tuesday, Wednesday \cdots Saturday))}$$ (4.8)

Dummy variables of Monday to Saturday were added to the daily based regression model and Sunday is selected as the reference day.

Table 4.6 The estimate coefficients of the regression models with the indicator of weekday effect

	Poisson Model			Negative Binomial Model		
	Estimate	Std.Error	P-value	Estimate	Std.Error	P-value
Monday	0.2231	0.2196	0.3100	0.2231	0.143	0.1187
Tuesday	0.2763	0.2186	0.2060	0.2763	0.1414	0.0507
Wednesday	0.1659	0.2208	0.4530	0.1659	0.1448	0.2522
Thursday	0.1464	0.2212	0.5080	0.1514	0.1455	0.3142
Friday	-0.0126	0.2250	0.9550	-0.0126	0.1512	0.9334
Saturday	0.0660	0.2231	0.7680	0.0660	0.1483	0.6566
Null deviance			4139.3	1362.2		
Residual deviance			4132	1359.2		
Variability explain by the model			0.18%	0.22%		
AIC			5102.2	3942.2		

The P-values of each estimate coefficients are close and none of these weekdays is more statistically significant than the other days. The weekday effect does not significantly affect the HSHs, because the heat stroke related diseases are so acute and urgent that other social factors cannot influence the "see doctor decision".

Some holidays have fixed dates, like New Year's Day (Jan 1st), Labor Day (June 19th), Independence Day (July 4th), Veterans Day (Nov 11st), and Christmas Day (Dec 25th). This study adds dummy variables for each of these fixed holidays to test whether the holiday effect can explain the occurrence of HSHs.

This study tested and found that the dummy variable of fixed holidays are statistically significant, but only 0.3% variability in hospitalizations can be explained. In fact, it is not worth adding indicators of fixed holidays, because HSHs are unlikely to occur in November, December and January, and the HSHs occurring in June and July can be explained by the seasonality and temperature exposure (in next chapter).

4.4 Results

The regression models with the calendar effects (the linear trend, the year effect, the seasonality, and the weekday effect) can explain around 30% variability in the HSHs.

When putting all the calendar effects variables into the same model, the dummy variables of each year lose their significance, because those predictors are not independent and convey overlapping information.

Table 4.7 The estimate coefficients of the regression models with the indicator of calendar effects

	Predictors	Poisson Model			Negative Binomial Model		
		Estimate	Std.Error	P-value	Estimate	Std.Error	P-value
Linear trend	day	0.0003	0.0017	0.8359	-0.0004	0.0018	0.8149
Year effect	1991	2.1110	9.1273	0.8170	-2.0987	9.7466	0.8295
	1992	0.8460	8.5205	0.9209	-2.7911	9.0982	0.7590
	1993	1.8599	7.9106	0.8141	-1.7536	8.4478	0.8356
	1994	1.4238	7.3028	0.8454	-1.9131	7.7991	0.8062
	1995	1.5530	6.6943	0.8165	-1.4097	7.1497	0.8437
	1996	0.5310	6.0877	0.9305	-1.7653	6.5010	0.7860
	1997	0.5348	5.4793	0.9222	-1.9	5.8526	0.7454
	1998	0.0479	4.8729	0.9922	-2.1756	5.2056	0.6760
	1999	0.6782	4.2628	0.8736	-1.3819	4.5553	0.7616
	2000	-0.6322	3.6613	0.8629	-2.1414	3.9117	0.5841
	2001	0.3843	3.0476	0.8997	-0.8921	3.2588	0.7843
	2002	0.7389	2.4392	0.7619	-0.521	2.6114	0.8419
	2003	-1.4152	1.8564	0.4459	-2.2291	1.9895	0.2625
	2004	-1.0716	1.2484	0.3907	-1.4778	1.3425	0.2710
	2005	-0.2086	0.6398	0.7444	-0.4187	0.7084	0.5545
Seasonality	$sin(2\pi\omega t)$	-1.6043	0.1001	< 2e-16	-1.5874	0.1068	< 2e-16
	$cos(2\pi\omega t)$	0.1429	0.2323	0.5385	0.0271	0.2438	0.9113
	$sin(4\pi\omega t)$	0.7518	0.0809	< 2e-16	0.6877	0.0937	2.13E-13
	$cos(4\pi\omega t)$	0.5697	0.1112	3.02E-07	0.4356	0.1260	0.0005
Weekday effect	Monday	0.2230	0.1430	0.1190	0.1465	0.2003	0.4646
	Tuesday	0.2761	0.1414	0.0508	0.2469	0.1976	0.2115
	Wednesday	0.1664	0.1448	0.2506	0.1823	0.1993	0.3605
	Thursday	0.1476	0.1455	0.3104	-0.0395	0.2059	0.8478
	Friday	-0.0115	0.1512	0.9394	-0.1701	0.2103	0.4185
	Saturday	0.0658	0.1483	0.6573	0.0451	0.2033	0.8245
Analysis	Null deviance			4139.3			2297.7
	Residual deviance			2863.1			1477.6
	Variability explain by the model			30.83%			35.69%
	AIC			3873.3			3351.7

This study superimposed the predicted daily time series of hospitalization for each year with the overall regression model for 16 years, and the intensity of HSHs shows strong annual variability (Figure 4.3). Year 1991, 1993, 1995 and 2002 are predicted to have high hospitalization intensity. The predicted average peak time is July 9th. The predicted intensity exhibited an almost 7-fold difference between the severest year 1995 and the least severe year 2003.

By superimposing the predicted daily time series of HSHs for each year using the separate regression model each year, this study investigates the variability in annual intensity and peak time (Figure 4.4).
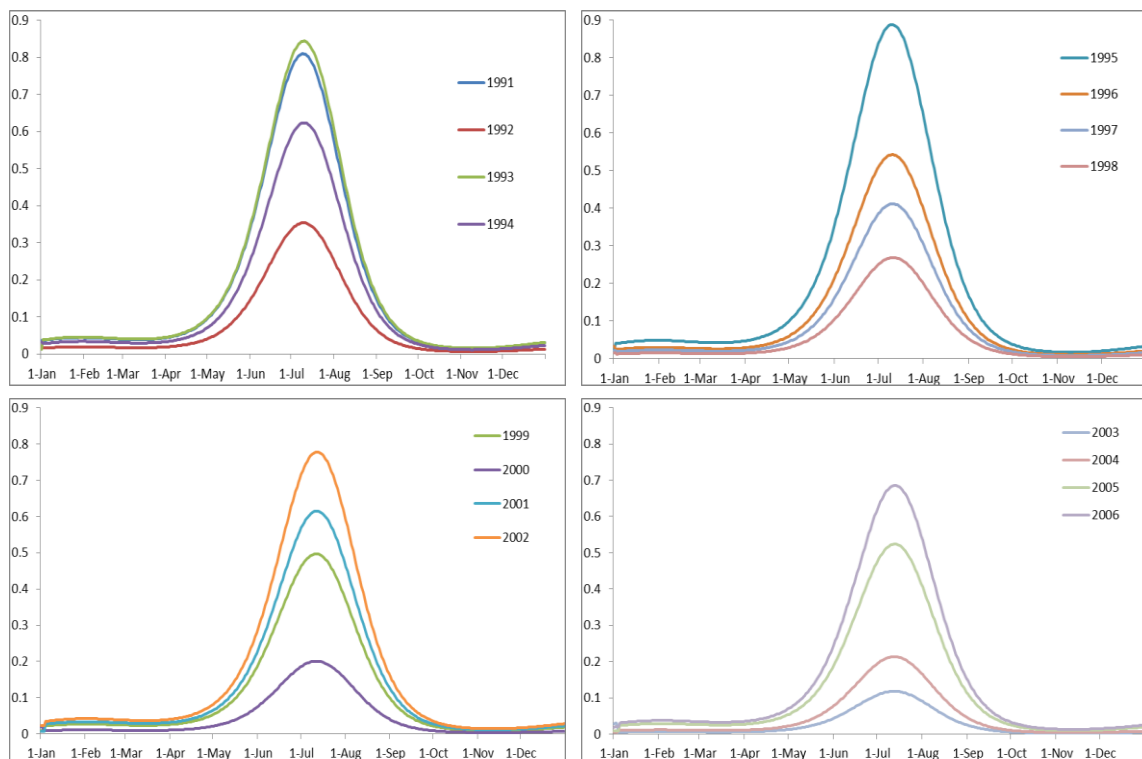
Figure 4.3 Superimposed daily time series of HSHs for elderly people (1991-2006) as predicted from the overall regression model for 16 years with calendar effects

50
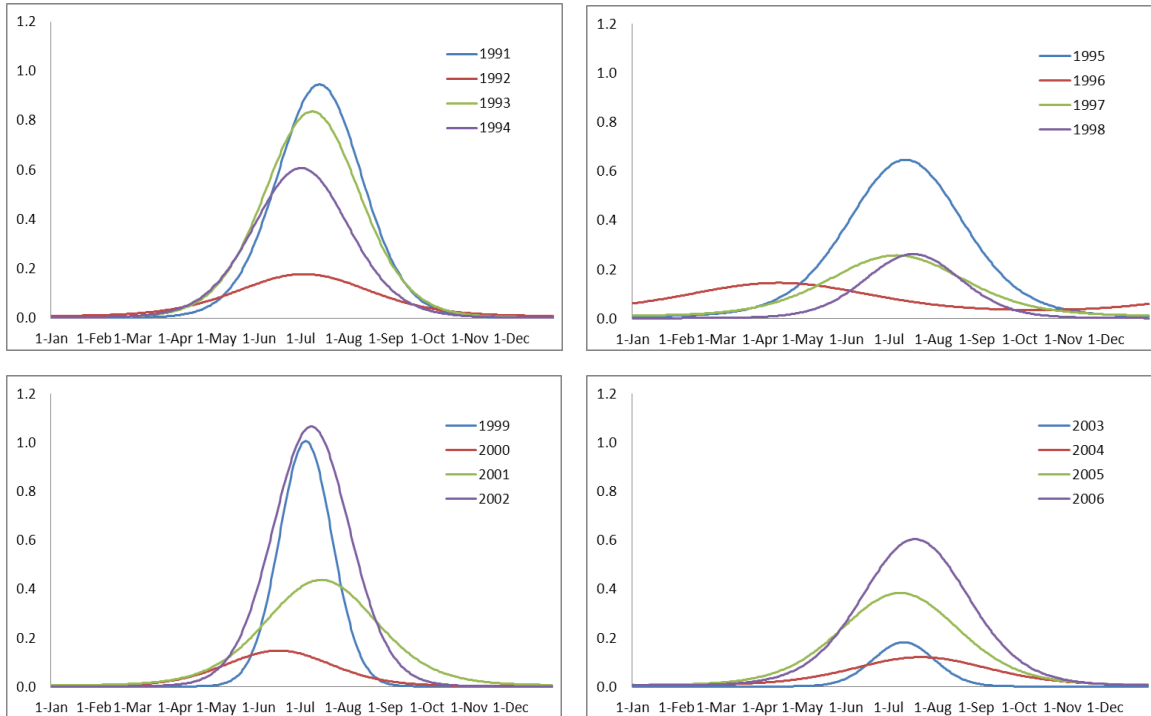
Figure 4.4 Superimposed daily time series of HSHs for elderly people (1991-2006) as predicted from the separate regression models for 16 years with calendar effects

51

Evaluating the accuracy of the prediction with calendar effects from 1) timing: the average peak time of each year is July 9th, and the predicted peak time of each year using separate regression model for each year ranging from April to July, while the actual peak time of each year varies ranging from June to August; 2) magnitude: the predicted absolute magnitudes of hospitalizations are smaller than 1, which greatly derivate from the actual magnitudes; the predicted relative magnitudes: the magnitude ratio of the highest year 1995 to the lowest year 2003 is 7:1; while the actual magnitude ratio of the highest year 2002 to the lowest year 2003 is 7.6:1; 3)duration: the predicted duration of HSHs is as long as the whole month, while the actual duration of HSHs is less than a week. Calendar information alone cannot provide enough information to predict the accurate hospitalization date and the actual numbers of HSHs either.

4.5 Conclusion

The HSHs don't demonstrate an evident long term trend, but do show strong seasonality and annual variations in intensity and peak time. Calendar effects can explain 30% of the variability in HSHs and regression models for each separate year can provide important information to describe the annual peak time of HSHs. The calendar effects cannot sufficiently explain the intensity and duration of HSH, because the calendar effect cannot provide adequate biological and physiological information to explain the mechanism of heat related disease. The calendar information is the most straightforward predictors of HSHs, but not their real underlying factors. The dummy variables of fixed holidays are not included in the final model, because the information expressed by the dummy variable is incorporated by other variables. In fact, it is still necessary to give heat warnings on holidays, especially if there are some celebration events outdoors in summer seasons.

This chapter supports the hypothesis that the HSHs demonstrate well-pronounced seasonality, and there are some annual variations in intensity and peak time. Next chapter will explain and analyze the

temperature effects on HSHs. Temperature effects are the direct measurements of heat exposure and

can also provide information to discuss the lag effect and the accumulated heat effect on human health.

4.   The association between ambient temperature and heat stroke related hospitalization

5.1 Background

Being aware of the calendar effects' inadequacy in the explanation of heat stroke related hospitalizations (HSHs), this chapter evaluates the impact of ambient temperature on HSHs, testing the hypothesis that the ambient temperature highly influences the heat stroke related hospitalization, when ambient temperature exceeds a threshold, the HSHs will increase rapidly.

Projected climate change scenarios lead to a more variable temperature system [81] , two of whose manifestations are the increasing ambient temperatures and heat wave effects. Human beings have a strong physiological capability of adjusting exterior heat effects, but human beings have limited tolerance of the amount of heat exposure and feel uncomfortable outside the temperature range of 17-30 ˚C [93]. High temperature causes the clinical syndromes of heat stroke, heat exhaustion, heat syncope, and heat cramps. Elderly people have limited heat regulation ability and cannot sufficiently cope with the variation in temperature; therefore they are more vulnerable to the variation ambient temperature than healthy adults.

Understanding the relationship between high ambient temperature and its negative effects on human health can assist predicting the influence of climate change on human health in a big scenario, and can provide more information for public health interventions; moreover can guide allocating material and facilities for vulnerable subgroups.

5.2 Methodology

This chapter aims to explain the association between ambient temperature and HSHs. This study uses generalized linear regression analysis of time series data from 1991 to 2006 to quantitatively illustrate the ambient temperature-HSHs association for the Boston-Cambridge-Quincy MA NH Metropolitan

Statistical Area (Boston MSA). This study also assesses the selection of predictor variables, the lag effect and the transformation of temperature in following sections.

5.3 Statistical features of daily temperature

The Gaussian (bell-shaped) distribution is the most commonly used continuous distribution because most phenomena are a result of many forces, like the daily temperature is the comprehensive result of particles' motion. The Central Limit Theorem states that any large sum of independent, identically distributed random variables are approximately Normal. In a long period of time, daily temperature can be regarded as a random variable though it principally correlates with a series of complicated physical and chemical reactions.

5.3.1 The distribution of daily temperature
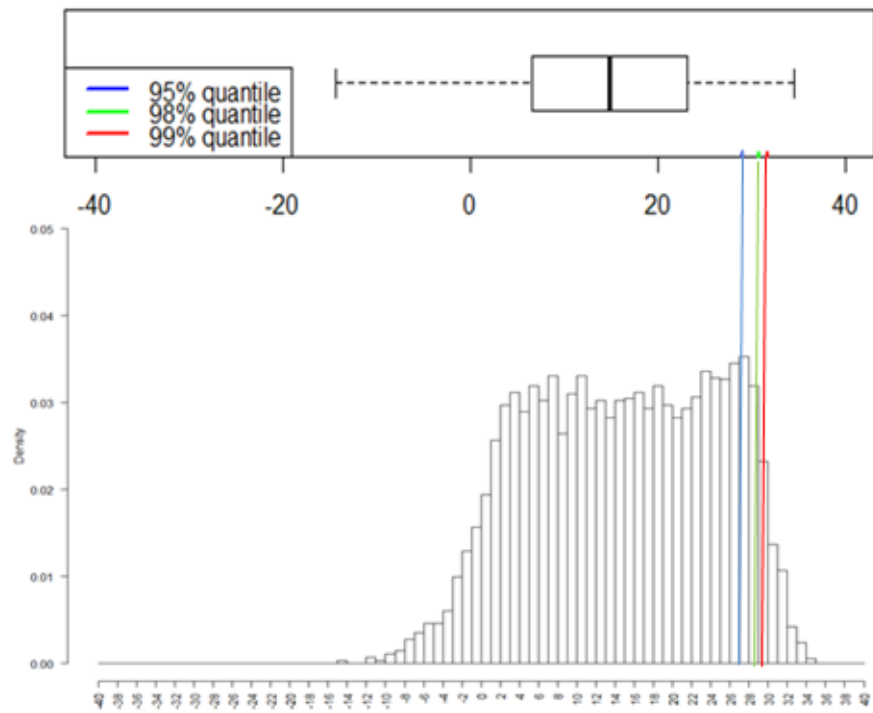
Figure 5.1 The distribution density plot of daily maximum temperature
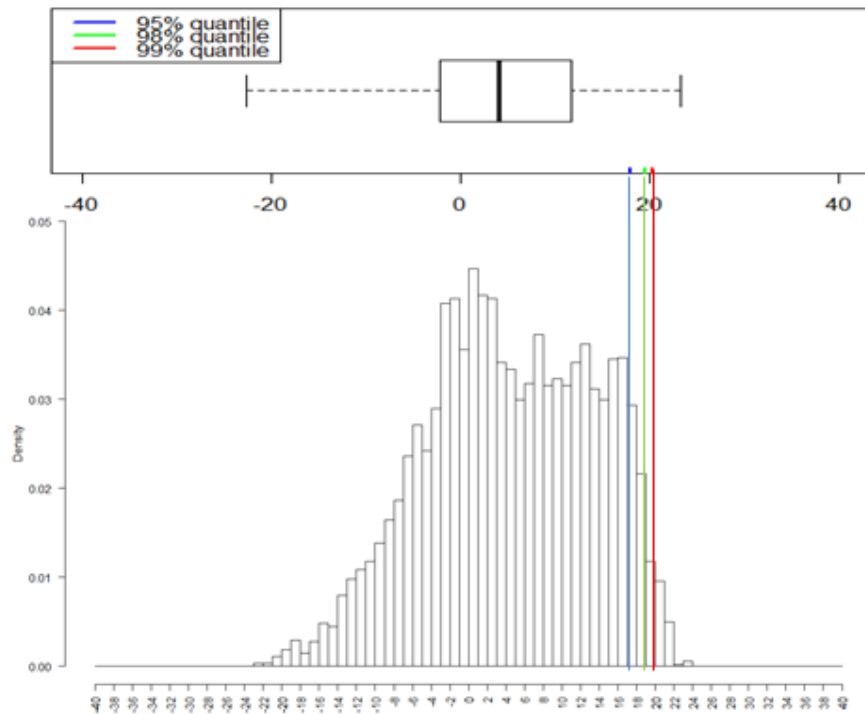
Figure 5.2 The distribution density plot of daily minimum temperature

Table 5.1 the summary of statistical feature of temperature

	Maximum temperature (°C)	Minimum temperature(°C)	Mean temperature (°C)
Total days	5844	5844	5844
Mean	15.46	5.07	10.26
Variance	103.20	85.68	90.69
Standard Deviation	10.16	9.26	9.52
Minimum	-14.28	-22.62	-17.70
Maximum	36.54	24.75	30.63
25th percentile	7.07	-1.71	2.83
50th percentile	15.78	5.04	10.41
75th percentile	24.27	12.82	18.56
95th percentile	30.42	18.89	24.40
98th percentile	32.26	20.50	25.84
99th percentile	33.40	21.35	26.83
Skewness	-0.14	-0.21	-0.18
Kurtosis	-0.98	-0.72	-0.88

5.4 Analysis of the temperature-hospitalization association

As is explained in Chapter 2.4.2, the ambient temperature directly affects human health and the occurrence of adverse health outcomes may occur several days follow the increase in temperature. This study uses a generalized linear model with Poisson distribution or Negative binomial distribution to explain the association between daily ambient temperature, lag effect and HSHs on the MSA level (Equation 5.1).

$$Y_t = exp^{(\beta_0 + \beta_1\,(T_{min}) + \beta_2\,(T_{max}) + \beta_3\,(Lag))} \quad (5.1)$$

where Y is daily HSHs; t is the date from January 1st, 1991 to December 31st 2006; β_0 is the intercept for the regression model, β_1 is the coefficient for daily minimum temperature; β_2 is the coefficient for daily maximum temperature; β_3 is vector of coefficients for lag effect. All regression analyses were conducted using R (Version 12.5.2) and the statistical significance was based on α=0.05.

5.4.1 T_{min} and T_{max}

The ambient temperature is the straightforward measurement of heat exposure, while there has been no superior type of temperature measurement so far. This study applies both the daily minimum (T_{min}) and maximum temperature (T_{max}) in the same regression model, even though they are highly correlated. This study tests and finds that the estimate coefficients for daily maximum and minimum temperature are of the same order of magnitude and are close, thus coefficients can be explained as the weight assigned to these two parameters. Applying both the daily maximum and minimum temperature in the same model allows the utilizing of "a weight-averaged daily temperature" to conduct regression analysis. Moreover, the daily maximum and minimum temperatures can be explained as the daytime temperature and the nighttime temperature respectively, therefore the weights before them can illustrate the different proportions of people's exposure to daytime and nighttime temperatures accordingly.

$$Y_t = exp^{(\beta_0 + \beta_1 (T_{min}) + \beta_2 (T_{max}))} \quad (5.2)$$

Table 5.2 summary of the regression model with maximum and minimum temperatures

Model	Poisson Model			Negative binomial Model		
Variables	Estimate	Std.Error	P-value	Estimate	Std.Error	P-value
T_{max}	0.1708	0.0127	< 2e-16			
T_{min}	0.0770	0.0132	< 2e-16			
Null deviance			4139.3			
Residual deviance			2543.1			
Variability explain by the model			38.56%			
AIC			3505			
	Estimate	Std.Error	P-value	Estimate	Std.Error	P-value
T_{max}	-0.2765	0.0434	1.79E-10	-0.2612	0.0444	3.92E-09
$(T_{max})^2$	0.0093	0.0008	< 2e-16	0.0088	0.0009	< 2e-16
T_{min}	0.1249	0.0325	0.0001	0.1225	0.0011	0.3365
$(T_{min})^2$	-0.0005	0.001	0.6144	-0.0011	0.0323	0.0001
Null deviance			4139.3		3248.3	
Residual deviance			2134.4		1704.3	
Variability explain by the model			48.44%		48%	
AIC			3100.5		3004.9	

The T_{min} and T_{max} are both statistically significant in the regression model. The estimate coefficient for maximum temperature is greater than that of minimum temperature, which indicates every degree increase in T_{max} will lead to more health risks than that brought by every degree increase in T_{min}. The T_{min} and T_{max} together can explain 38.56% of the variability in HSHs. The regression model with a linear term of temperature is not adequate to detect the association between temperature and HSHs. Adding quadratic indicators of maximum and minimum temperature to the regression model can explain another 10% of the variability in HSHs. The quadratic indicator of T_{min} is not statistically significant and can explain dispensable variability in HSHs, which indicates that HSHs are more closely related with daily maximum temperatures.

The software support limits the application of a negative binomial model. The MASS package in R software (Version 2.15.2) has some bugs and it cannot model the association between T_{min}, T_{max} and

HSHs using generalized linear model with negative binomial distribution, warning the non- divergence of HSHs value. Theoretically, the null deviance should be constant for all negative binomial models if explaining the same dataset. The changing null deviance of a negative binomial model also creates doubt about the quality of the negative binomial model.

5.4.2 Threshold of temperature effect

Human beings have strong and subtle heat adjustment capabilities, and in a certain location, there exists an upper limit of temperature, exceeding which people's thermoregulation system will fail to work and more adverse health outcomes will occur. This temperature limit is called the threshold of temperature [2, 24, 26]. The temperature threshold is a location-specific parameter and also correlated with other social-economic factors.

This study applies a data-driven process to investigate the temperature threshold of HSHs in the Boston MSA. This study categorized daily temperature according to the HSHs happened on each day, and then observed the maximum and minimum records of maximum and minimum temperatures on those days in each subgroup. As is shown in Table 5.3, the variances of temperature records for those days with more than 5 daily hospitalization counts are smaller than those of other groups, which indicate that the high daily maximum and minimum temperatures are the prerequisite for HSHs, though the high temperatures won't necessarily lead to hospitalizations. The minimum records of daily maximum and minimum temperatures (i.e.30.02°C and 16.72 °C) of those days with five or more HSH counts are selected as the threshold points of temperatures. The odds ratio for temperature exposure beyond the threshold is calculated in Table 5.4. As is illustrated in Figure 5.3 and Figure 5.4, the horizontal lines classify all the days into two groups: days with less than 5 daily cases and days with 5 or more daily cases. The vertical lines indicate the lower band of the daily minimum and maximum temperatures, and all those days with 5 or more HSHs are located beyond these two lines.

Table 5.3 The summary of temperature features according to the hospitalizations per day

		HSHs	T_{max} (°C)	T_{min} (°C)
all data (5844)	Mean	0.12	15.46	5.07
	Median	0	15.78	5.04
	Variance	0.46	103.2	85.68
	Standard Deviation	0.68	10.16	9.26
	Minimum	0	-14.28	-22.62
	Maximum	22	36.54	24.75
	1st Quartile	0	7.07	-1.71
	3rd Quartile	0	24.27	12.82
	Skewness	15.56	-0.14	-0.21
	Kurtosis	370	-0.98	-0.72
days with more than 1 count/day (426)	Mean	1.65	25.55	13.92
	Median	1	28.93	16.52
	Variance	3.84	84.24	61.63
	Standard Deviation	1.96	9.18	7.85
	Minimum	1	-3.62	-13.98
	Maximum	22	36.54	24.75
	1st Quartile	1	22.43	10.92
	3rd Quartile	1	31.83	19.37
	Skewness	6.12	-1.37	-1.32
	Kurtosis	48.16	1.03	1.14
days with more than 3 counts/day (49)	Mean	5.45	33.14	20.43
	Median	4	34.05	21.12
	Variance	16.29	6.35	8.12
	Standard Deviation	4.04	2.52	2.85
	Minimum	3	26.11	9.13
	Maximum	22	36.54	24.75
	1st Quartile	3	31.28	19.67
	3rd Quartile	6	34.73	22.23
	Skewness	2.51	-0.81	-1.89
	Kurtosis	6.22	-0.04	4.52
days with more than 5 counts/day (20)	Mean	8.5	34.19	21.7
	Median	7	34.36	21.65
	Variance	24.26	3.89	2.83
	Standard Deviation	4.93	1.97	1.68
	Minimum	5	30.02	16.72
	Maximum	22	36.5	24.75
	1st Quartile	5	33.85	21.18
	3rd Quartile	9.25	35.75	22.42
	Skewness	1.51	-0.84	-0.88

	Kurtosis	1.06	-0.38	1.76

Table 5.4 The odds ratio of temperature exposure beyond the threshold point

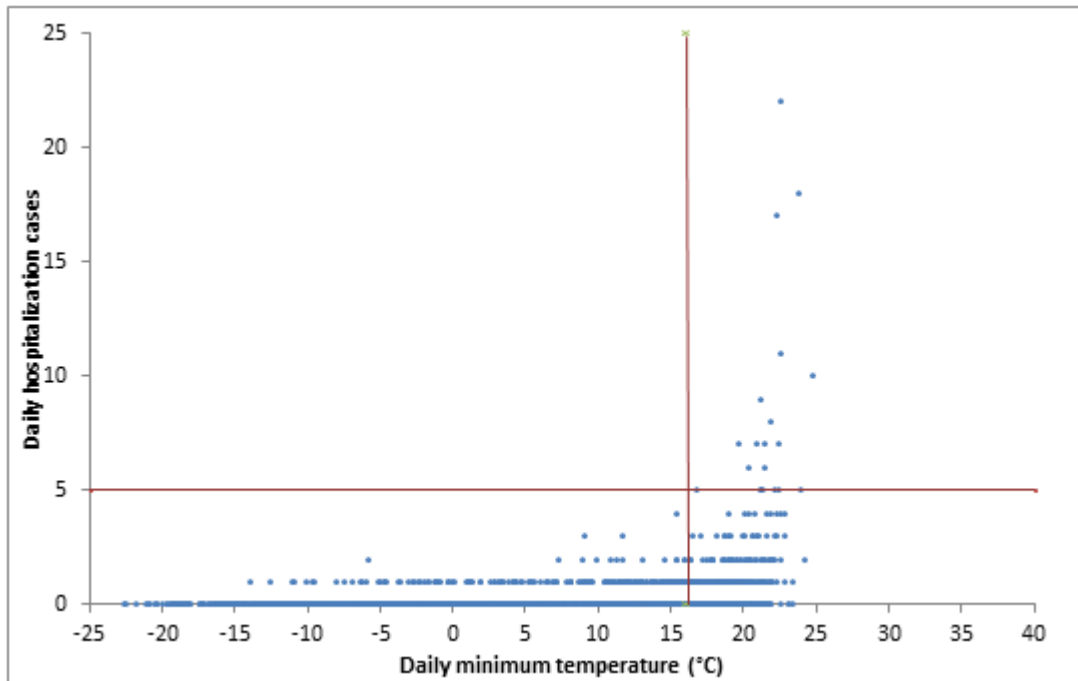|	(T_{max}, T_{min}) > (30.02°C, 16.72 °C)	
HSHs on that day	YES	NO
YES	131	295
NO	102	5316
OR (95% CI)	23.14 (17.42, 30.75)	

Figure 5.3 the association between daily hospitalization cases and minimum temperature
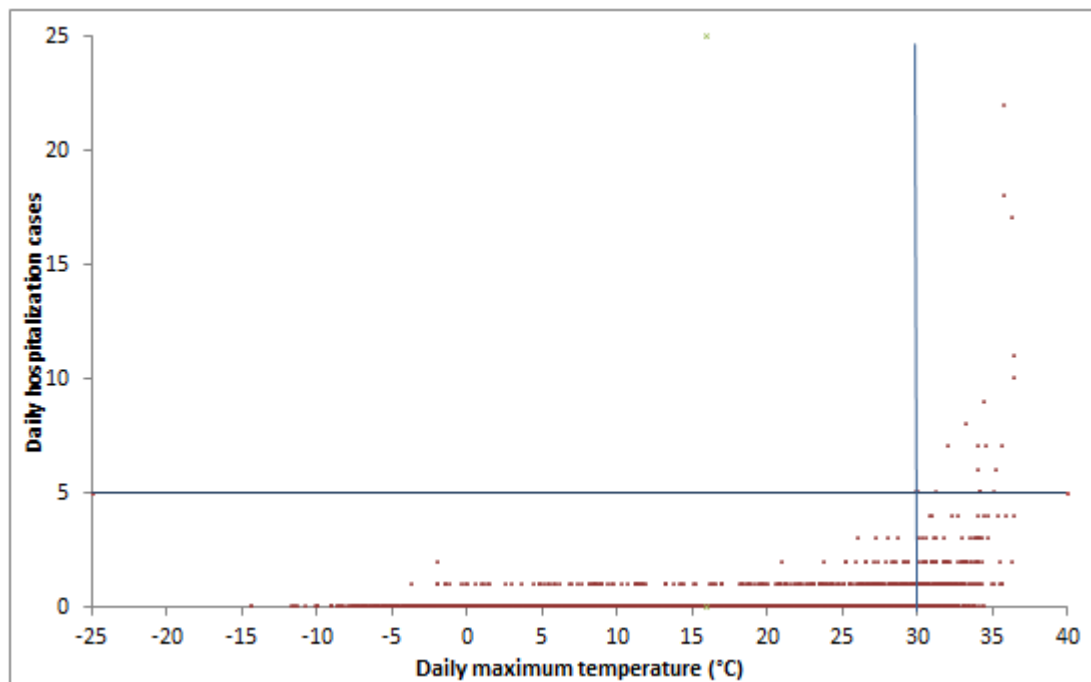
Figure 5.4 the association between daily hospitalizations and maximum temperature

5.4.3 Temperature transformation

With the understanding of threshold temperature, this study enhances the analysis of the relationship between daily temperature parameters and HSHs. Based on the biochemical kinetic principles, the effects of temperature on human health can be explained by the temperature effects on metabolic rate [99]. The metabolism is an energy and material transformation process which is closely correlated with temperature. The process of metabolism is the occurrence of countless chemical reactions in cells, therefore it is logical to borrow some concepts from the chemical field to explain and quantify the relationship between temperature and its health outcomes.

The Arrhenius equation (Svante Arrhenius, 1889) explains the energy required to start a chemical reaction. The reaction rate depends on temperature, because which offers the activation energy. A reaction cannot start until the atom can leap over the activation energy threshold.

$$k = Ae^{-\frac{E_a}{RT}} \quad (5.3)$$

where k is the rate constant of a chemical reaction; T is the reaction temperature; A is a pre-factor; R is the Universal gas constant; E_a is a factor that can affect the chemical reaction rate in the existence of catalyst, which is also influenced by temperature.

The enzyme is a specific type of biological catalyst that keeps the chemical reactions in the human body occurring quickly enough to keep the body alive. Enzymes are sensitive to their environmental temperature. Up to a threshold, the rate of chemical reaction in the human body increases as a function of temperature because the substrates collide more frequently with the enzyme active site. However, at both high and low temperature extremes, the native structure of the enzyme will be compromised and the molecule becomes inactive (as is shown in Figure 5.5). The convex curve of enzyme reactions together with regression model in Table 5.2 indicates that the association between temperature and HSHs is not a simple linear one.
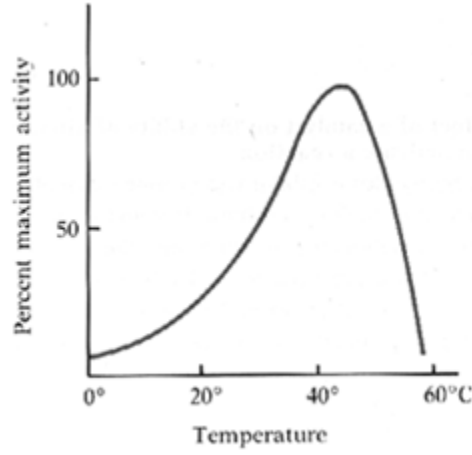
Figure 5.5 The relationship between temperature and activity of enzyme reactions (The data source is cited in Appendix 5)

This study illustrates the heat stroke related disease as the macro manifestation of the enzyme action: when people are exposed to extremely high temperature environments, their heat regulation mechanism may fail to control and keep the constant status of the interior temperature environment, and then the enzymes cannot work within the optimal temperature range; consequently, the heat stroke-related disease occurs. This study demonstrates the mechanism in a quantity form:

$$Transformed\ (T_{ij}) = e^{\frac{T_{ij} - Threshold_j}{R}} \quad (5.4)$$

Equation 5.4 is formula of temperature transformation, T_{ij} is the transformed daily maximum or minimum temperature, where i is the date from January 1st ,1991 to December 31st ,2006 ; $j = 1$ indicates daily maximum temperature and $j = 2$ indicates daily minimum temperature; $Threshold_{ij}$ are the thresholds of maximum temperature ($j = 1, Threshold_1$ =30.02 °C) and minimum temperature ($j = 2, Threshold_2$ =16.72 °C) (explained in chapter 5.4), exceeding which, there will be more HSHs; R is the median level (10.00 °C) of daily average temperature, which is used to adjust the order of the magnitude of the exponential term. The temperature transformation aims to emphasize the exponential increase of negative health outcomes after the threshold and detect those extreme values in HSHs.

65

Table 5.5 the regression model with transformed daily temperature

Model	Poisson Model			Negative binomial Model		
Variables	Estimate	Std.Error	P-value	Estimate	Std.Error	P-value
Transformed (T_{max})	1.1857	0.0595	< 2e-16	1.3763	0.0906	< 2e-16
Transformed (T_{min})	0.4485	0.0464	< 2e-16	0.5359	0.0681	3.72E-15
Null deviance			4139.3		3232.7	
Residual deviance			2145.4		1679.5	
Variability explained by the model			48.17%		48.05%	
AIC			3107.6		2983.5	

Comparing table 5.5 and table 5.2, this study finds that the transformed daily temperatures are statistically significant and can explain 10% more variability in HSHs than the untransformed variables.

5.4.4 Lag effect

The lag effect explains the temporal pattern of association between temperature exposure over previous days and negative health outcomes on another particular day. The lag effect of heat events plays a crucial role in the severity of health outcomes, because the length of lag is correlated with the duration of the heat event. This section tests and selects the length of lag according to the AIC value of each model. As is shown in Table 5.6, adding 1 and 2 days' lag can explain another 4% of the variability in HSHs. These two lag terms are both statistically significant and the maximum temperature of two previous days greatly affects human health.

Table 5.6 the regression model with transformed daily temperature and two days' lags

Model	Poisson Model			Negative binomial Model		
Variables	Estimate	Std.Error	P-value	Estimate	Std.Error	P-value
Transformed (T_{max})	0.8289	0.0665	< 2e-16	0.9965	0.0930	< 2e-16
Transformed (T_{min})	0.2538	0.0525	1.31E-06	0.2062	0.0706	0.0035
Transformed (T_{max}(lag1))	0.8034	0.2079	0.0001	0.7235	0.2554	0.0046
Transformed (T_{max}(lag2))	0.9620	0.1672	8.70E-09	0.9555	0.2040	2.80E-06
Null deviance			4139.3		3539.1	
Residual deviance			1971.7		1726.2	
Variability explain by the model			52.37%		51.22%	
AIC			2937.9		2899.4	

Chapter 6.4 will verify the selection of length of lag by considering the duration of heat wave and explain the lag effect as the manifestation of the accumulative heat effect.

5.5 Seasonality of maximum temperature

In chapter 4.3, the seasonality of HSHs was detected using a harmonic function, and this chapter aims at modeling the seasonality of maximum temperature for each year with harmonic functions and investigating whether the annual peak time of HSHs is correlated with annual peak time of maximum temperature.

$$Y_t = exp^{(\beta_0 + \beta_1(sin(2\pi\omega t), cos(2\pi\omega t)))} \quad (5.5)$$

where Y is the daily maximum temperature; t is the date from January 1st, 1991 to December 31st, 2006; β_1 is the vector of coefficients for the seasonality; π is the constant; ω is the frequency-sequence of the day divided by the number of cycles, whose value is 365 for the average year and 366 for the leap year. The terms $sin(2\pi\omega t)$ and $cos(2\pi\omega t)$ model the annual oscillation cycle.

Applying the same formula in Equation 4.5-4.7 to estimate annual peak time of the daily maximum temperature. The estimated results are shown in Table 5.7. Recalling the peak time of HSHs shown in Table 4.3, the scatter plot of the annual peak time of HSHs and the annual peak time of the daily maximum temperature is shown in Figure 5.6. There is no strong correlation between the annual peak time of the daily maximum temperature and the peak time of HSHs.

Year	1991	1992	1993	1994	1995	1996	1997	1998
Julian day	199	203	205	205	204	203	205	206
Date	18-Jul	22-Jul	24-Jul	24-Jul	23-Jul	22-Jul	24-Jul	25-Jul
Year	1999	2000	2001	2002	2003	2004	2005	2006
Julian day	205	201	209	200	208	208	204	208
Date	24-Jul	20-Jul	28-Jul	19-Jul	27-Jul	27-Jul	23-Jul	27-Jul
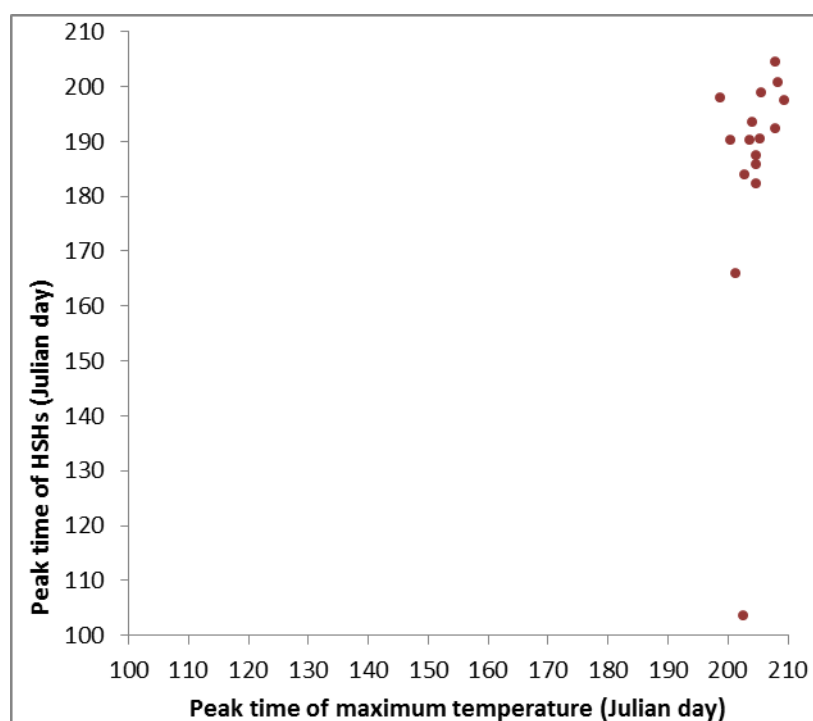
Figure 5.6 The scatter plot of annual peak time of HSHs and daily maximum temperature

5.6 Result

This study started the analysis of the association between ambient temperature and HSHs with linear indicators of daily maximum and minimum temperatures, and then created the transformation formulae for the daily temperatures to model the different behaviors of HSHs when whose daily maximum and minimum temperature are below or beyond the local temperature thresholds. Applying the transformed temperature and two days' lag can explain 52.37% of variability in HSHs.

This study superimposed the daily time series of HSHs predicted from the overall model for 16 years with temperature effects, and the variability in annual intensity and peak time are plotted in Figure 5.7. Year 1991, 2006, 1993, and 2002 are predicted to have high hospitalization intensity. The model with temperature effects can detect the peak time, magnitude and duration of HSHs much better than the model with calendar effects. The model with temperature effects can predict the variation in annual peak time quite close to the observed dates, ranging from June to July, can predict the HSHs are in the same order of magnitude as the observed records and can predict the accurate length of heat wave duration.

To visualize the distances between observed hospitalization dates and predicted hospitalization dates, the hospitalization calendar is built for June, July and August from 1991 to 2006. The daily hospitalization counts are sorted and the top 100 daily hospitalizations and their dates are selected to plot in the HSHs calendar (Figure 5.8).
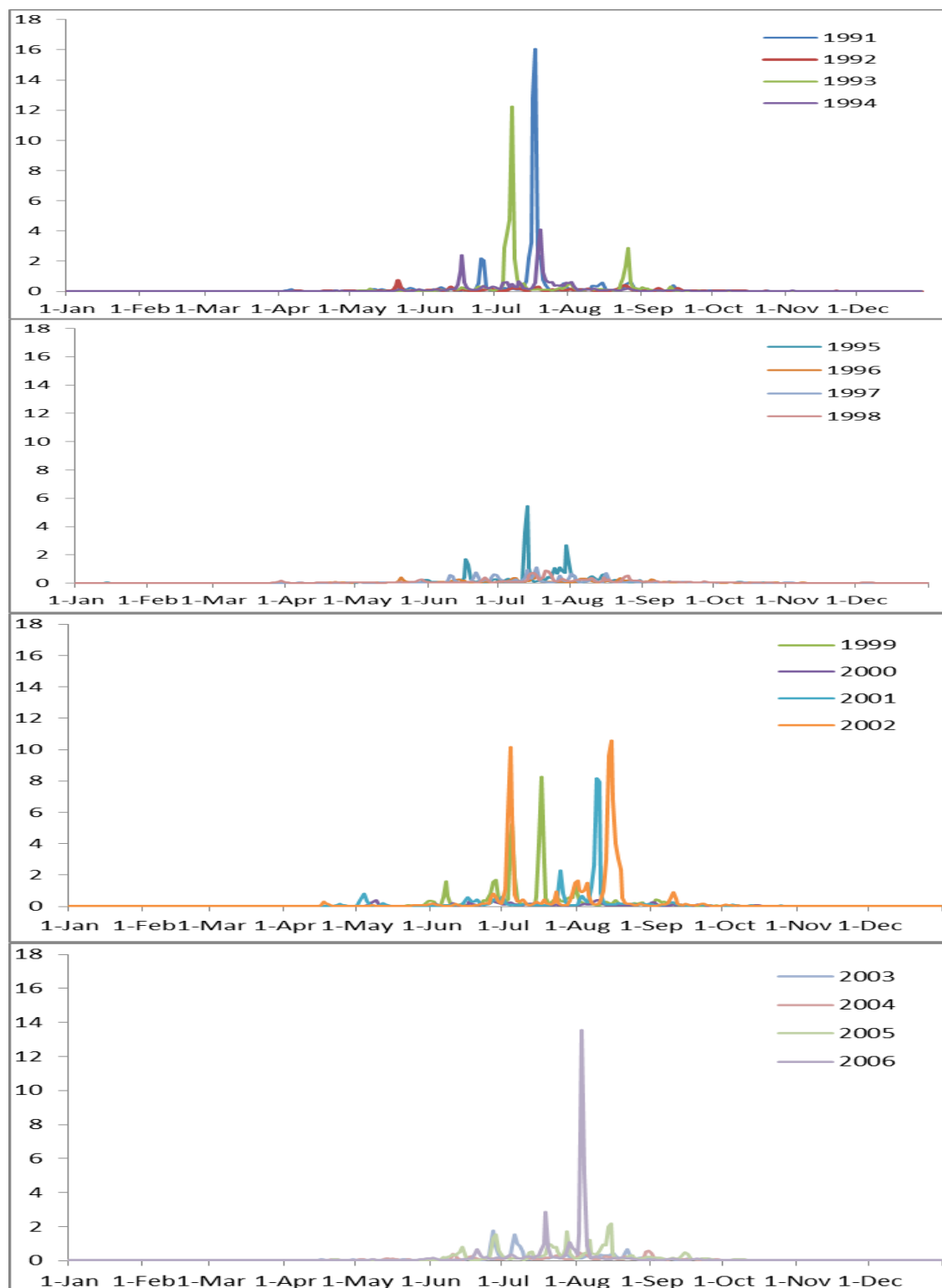
Figure 5.7 Superimposed daily time series of HSHs for elderly people (1991-2006) as predicted from the model with temperature effects

70
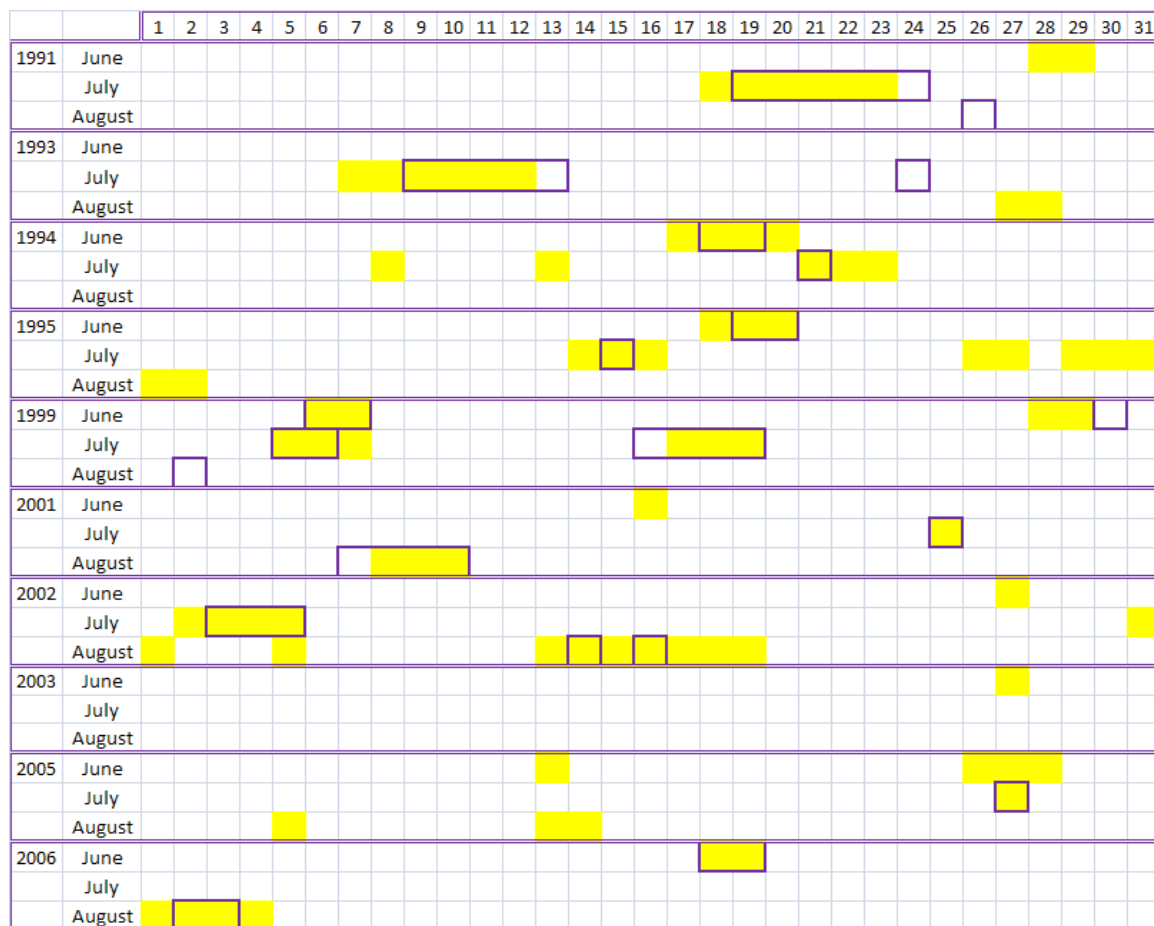
Figure 5.8 HSHs calendar (The window with purple border stands for the observed HSHs dates; the window in yellow stands for the HSHs dates predicted by model with temperature effects)

5.7 Conclusion

The elevated ambient temperature exposure is the direct cause of HSHs and there exists a non-linear relationship between temperature and heat HSHs. The temperature transformation formula takes the temperature threshold idea into account and describes the exponential increase tendency of hospitalizations when the temperature threshold is exceeded. The transformed temperature can accurately detect the magnitude of HSHs.

This chapter investigates the general temperature effects and supports the hypothesis that the ambient temperature highly influences the heat stroke related hospitalization, when ambient temperature exceeds a threshold, the HSHs will increase rapidly, and in next chapter, this study will focus on those prolonged extremely hot temperature evens — heat waves.

5. Heat wave and heat stroke related hospitalizations

Chapter 5 investigated the association between ambient temperature and heat stroke related hospitalizations (HSHs) and found that there are some extremely high hospitalizations on some "extreme days". This chapter aims to test the hypothesis that the heat wave independently affects HSHs in the presence of well-pronounced seasonality, detecting how those "extreme days" can be distinguished from the common days, and how to utilize those "extreme days" to better understand the occurrence of HSHs

6.1 Background

The heat wave and its severe negative health outcomes attracted study attentions all over the world, especially after the 1995 Chicago and 2003 European heat wave events. What's worse, the potential climate change may increase the frequency, duration, and intensity of the heat wave. Published studies have pointed out that more severe health effects would be resulted from the prolonged duration of heat wave, but very few studies investigated how the timing of first occurrence of the heat wave each year affect the negative health effects due to the heat wave.

All the further analyses about intensity, duration and timing should conduct based on the definition of heat wave; however, the biggest challenge of heat wave study is the lack of standard and consistent definitions of heat wave.

6.2 Definition of heat wave

The heat wave can be conceptually explained as a type of extremely uncomfortable and life-threating heat event, which is beyond people's heat adjustment capacity and may lead to lift-threating health outcomes. The health outcomes are associated with the intensity, duration and timing of the heat wave, as well as people's adaptation and the availability of technical facilities [86]. The definitions of heat wave applied in published studies were summarized in Table 1.3.

This study investigates the best indicator of the heat wave for the Boston-Cambridge-Quincy MA NH Metropolitan Statistical Area (Boston MSA) by referring to published studies and sufficiently considering the Boston MSA's local temperature features. This study expects the selected indicator of the heat wave to be a proper warning signal of the arrival of the heat-related disease.

In published studies, heat waves were defined according to both absolute and relative temperature metric, and the duration of heat events. It makes more sense to define the heat wave by its relative intensity in order to consider the local temperature characteristics and long-time records. What's more, the duration of heat wave, describing the accumulative heat effect on human health is highly determined by the geographical location of each city, for example the heat wave lasts longer in locations surrounded by rivers and mountains, like the city of Chongqing in China. This study evaluates the indicator of the heat wave by analyzing relative temperature conditions for several consecutive days.

Table 6.1 Application of various definitions of heat wave in the Boston MSA

Definition of heat wave	Days that are defined as heat wave	Variability explained by the Dummy variable	Variability explained in the presence of seasonality
Days with minimum temperature above 99th percentile of daily mean temperature in this location[81]	28	15%	33%
Days with daily maximum temperature above 99th percentile of 16 years' daily maximum temperature[78]	28	21%	37%
2 consecutive days with mean temperature above 95th percentile of warm season (May 1 to September 30) daily mean temperature for this location (community)[5]	168	19%	36%
2 consecutive days with daily maximum temperature above 98th percentile of 16 years' daily maximum temperature*	68	21%	38%
2 consecutive days with daily maximum temperature above 95th percentile of 16 years' daily maximum temperature*	170	21%	41%
3 consecutive days with daily maximum temperature above 95th percentile of 16 years' daily maximum temperature*	158	30%	43%

 * Original definition

According to Table 6.1, utilizing various definitions of heat wave may categorize different days into the heat wave period accordingly, directly affecting the study results and hindering comparison or synthesis of results across studies.

This study added a dummy variable to indicate heat wave days which are defined using different definitions (Table 6.1) and conducted regression analyses of HSHs with the indicators of heat waves. This study selected the definition "3 consecutive days with daily maximum temperature above 95th percentile of 16 years' daily maximum temperature", for it can explain more variability in HSHs than any other definitions.

Another issue resulting from the definition of heat wave is whether or not to include the first two days in the 3-day heat wave period. (In Table 6.1, the first two days of these 3 consecutive days with daily maximum temperature above 95th percentile of 16 years' daily maximum temperature are included in those 170 heat wave days.)

6.3 Analysis of heat wave and hospitalizations

The annul peak time of HSHs usually occurs during the first heat wave period. As is shown in Figure 6.1, the annul peak time of HSHs is closely correlated with the occurrence of the first heat wave each year, following the occurrence of the first heat wave in seven days.
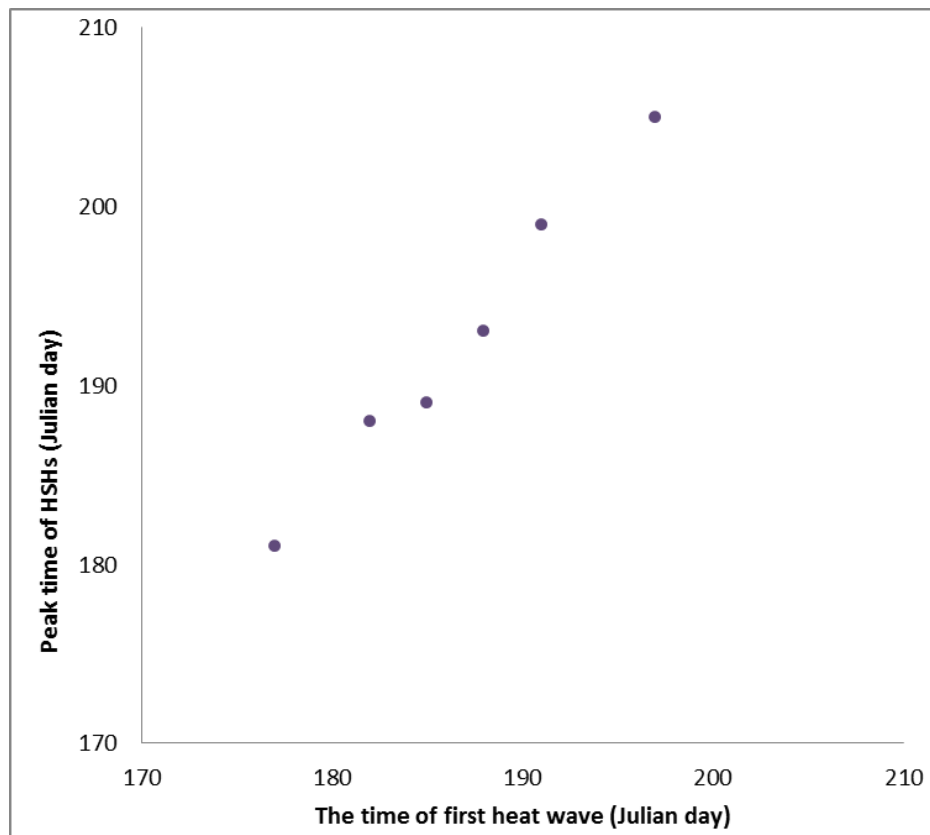
Figure 6.1 The scatter plot of the peak time of HSHs and the time of first heat wave (No points stand for year 1996, 1998, 2003, and 2004, because no heat wave happened in those years.)

Adding an indicator variable of the heat wave to the regression model with calendar effects can be explain the effect of heat wave on HSHs (ICD-9-CM 992.0 – 992.9) in the presence of seasonality. Generalized linear regression with Poisson or negative binomial distribution is conducted in the following sections.

$$Y_t = exp^{(\beta_0 + \beta_1 (linear\ trend) + \beta_2 (year\ effect) + \beta_3 (seasonality) + \beta_4 (weekday\ effect) + \beta_5 (heatwave))} \quad (6.1)$$

In Equation 6.1: Y is the daily hospitalization count from January 1st, 1991 to December 31st, 2006; t is the date of the hospitalization record; β_0 is the intercept for the regression model; β_1 is the coefficient for the linear trend; β_2 is the vector of coefficients for the year effect; β_3 is the vector of coefficients for the seasonality; β_4 is the vector of coefficients for the weekday and holiday effects; β_5 is the coefficient for the indicator of the heat wave.

Table 6.2 The results of two regression models with heat wave effect

		Poisson Model			Negative Binomial Model		
	Predictors	Estimate	Std.Error	P-value	Estimate	Std.Error	P-value
Linear trend	day	-0.0014	0.0017	0.4068	-0.0015	0.0018	0.4173
	1991	-7.9313	9.5238	0.4050	-7.9846	9.8381	0.4170
	1992	-8.0704	8.8918	0.3641	-8.0595	9.1842	0.3802
	1993	-6.7648	8.2551	0.4125	-6.8826	8.5274	0.4196
	1994	-6.7913	7.6201	0.3728	-6.7165	7.8723	0.3936
	1995	-5.9321	6.9852	0.3958	-5.8393	7.2164	0.4184
	1996	-5.6709	6.3519	0.3720	-5.5520	6.5613	0.3975
	1997	-5.1178	5.7169	0.3707	-5.0739	5.9060	0.3903
Year effect	1998	-4.8485	5.0843	0.3403	-4.8846	5.2526	0.3524
	1999	-4.4187	4.4484	0.3205	-4.4937	4.5979	0.3284
	2000	-4.2812	3.8203	0.2624	-4.2979	3.9476	0.2763
	2001	-3.1521	3.1804	0.3216	-3.1968	3.2886	0.3310
	2002	-2.3953	2.5462	0.3468	-2.5790	2.6351	0.3277
	2003	-3.7120	1.9379	0.0554	-3.7363	2.0099	0.0630
	2004	-2.1158	1.3020	0.1042	-2.0652	1.3492	0.1259
	2005	-1.2815	0.6670	0.0547	-1.0102	0.7048	0.1518
	$sin(2\pi\omega t)$	-1.2965	0.1024	< 2e-16	-1.2994	0.1061	2.00E-16
Seasonality	$cos(2\pi\omega t)$	-0.0253	0.2337	0.9136	-0.0539	0.2403	0.8225
	$sin(4\pi\omega t)$	0.5110	0.0821	0.0000	0.4870	0.0905	0.0000

	$cos(4\pi\omega t)$	0.1847	0.1122	0.0998	0.2046	0.1217	0.0928
Weekday effect	Monday	0.2069	0.1432	0.1485	0.1053	0.1828	0.5645
	Tuesday	0.2806	0.1415	0.0474	0.1716	0.1810	0.3430
	Wednesday	0.1976	0.1453	0.1737	0.0567	0.1846	0.7588
	Thursday	0.1616	0.1461	0.2687	-0.1963	0.1932	0.3094
	Friday	0.0174	0.1517	0.9085	-0.2888	0.1974	0.1435
	Saturday	0.1113	0.1487	0.4544	0.0123	0.1861	0.9473
Heat wave	Indicator	2.2104	0.0940	< 2e-16	2.2551	0.1450	2.00E-16
Analysis	Null deviance	4139.1			2877.6		
	Residual deviance	2339.1			1597.6		
	Variability explain by the model	3351.1			3129.7		
	AIC	43.49%			44.48%		

6.4 Heat wave and lag effect

The lag effect of temperature on HSH in the regression model is expected to reflect the accumulated heat effect and indicate the lack of relief during that heat wave period. The study proposes the heat wave definition "3 consecutive days with daily maximum temperature above 95th percentile of 16 years' daily maximum temperature", indicating the heat wave effect will become risky with two previous days accumulation. What's more, four or more consecutive days with daily maximum temperatures all above the 95th percentile are less likely to happen in the Boston MSA, if more than two days' lag were added to the regression model; the cool temperature might alleviate accumulated heat effect.

Comparing the Figure 6.2 and Figure 4.3, I can see the significant contribution of heat wave effect in the explanation of daily HSHs.
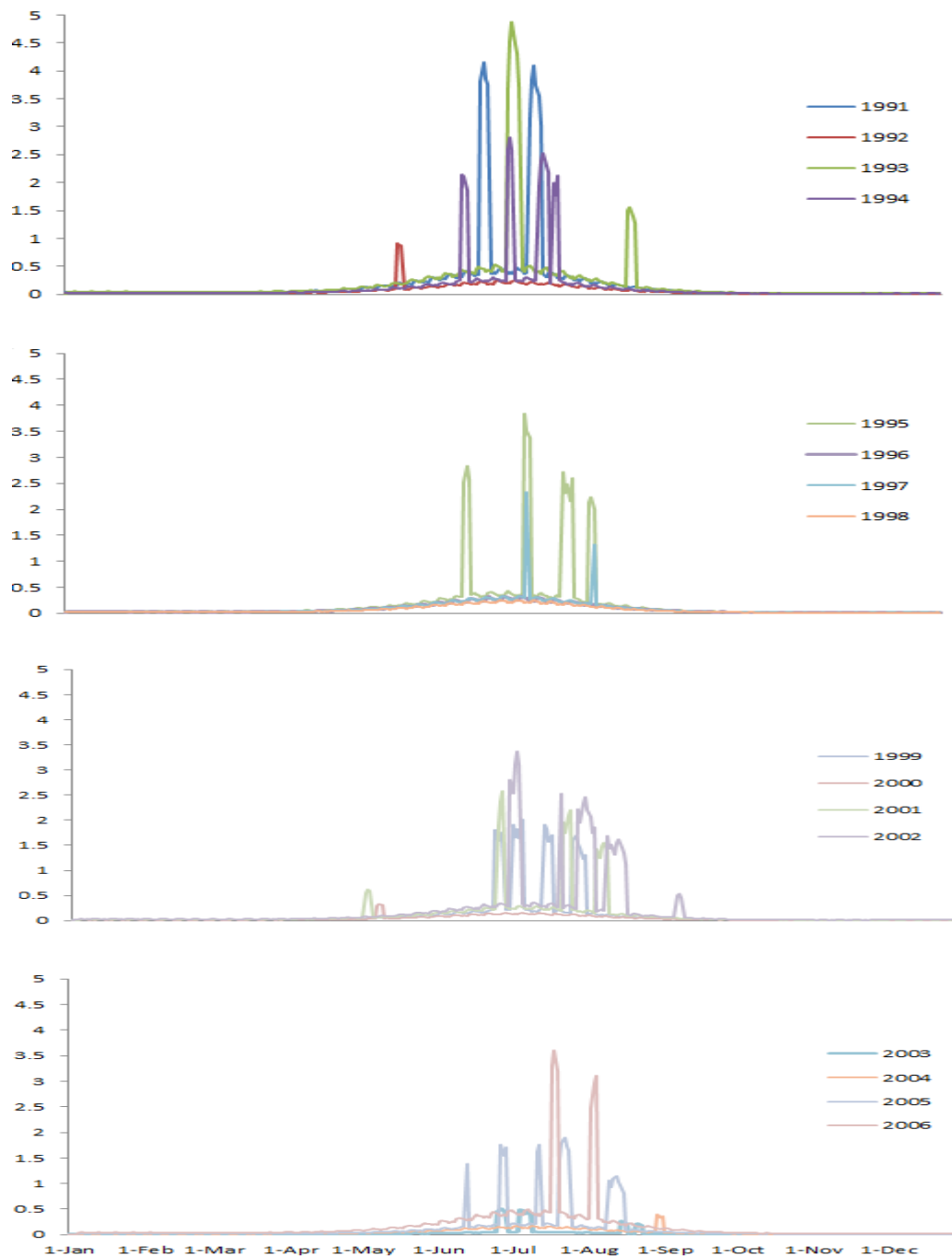
Figure 6.2 Superimposed daily time series (predicted with calendar and Heat wave effects) of HSHs for elderly people for 16 years (1991-2006)

80

6.5 Conclusion

Understanding the correlation between the heat wave effect and HSHs can provide suggestions on the prevention heat wave's adverse health effects effectively and HSHs preparations timely.

The first occurrence of the heat wave each year is significant in guiding the heat wave prevention: it can serve as a proceeding warning signal to of the steep increase in HSHs. If the warning system or engineering control system can be operated instantly when the heat wave begins each year, people can be better protected from the heat wave. However, the warning system operation needs financial support from the local government, and starting the warning and other public cooling system instantly after the arrival of heat wave may cost more money.

This chapter supports the hypothesis that the heat wave independently affects HSHs in the presence of well-pronounced seasonality, and in next chapter, this study will explain the model based on the combination of calendar effects, temperature effects and the heat wave effect.

6. Model explanation

This study investigated the calendar effect, temperature effects and the heat wave effect on HSHs respectively in previous chapters, and now aims to build a model to explain HSHs with both calendar and temperature effects.

7.1 Model building

The calendar effects alone can explain 30% of the variability in HSHs, and the temperature effects can explain 52% of the variability in HSHs. However, when integrating the temperature effect and calendar effect in the same regression model, they can explain 56% of the variability in HSHs in total, which is smaller than the additive variability explained by calendar and temperature effects, because the information expressed by calendar effects is overlapping with that conveyed by temperature effects. Naturally, the calendar is created according to the rotation regularity of Sun and Earth and the variation of temperature and seasons are also the direct manifestations of relative position between Sun, Moon and Earth.

Even though calendar effects and temperatures convey somewhat repeated information, it is still meaningful to consider these two effects in the same regression model: 1) calendar effect links the natural phenomena to social events, and it is an easy way to record the occurrence of HSHs; 2) temperature effect explains the environmental trigger of HSHs, and it is a straightforward way to understand and evaluate the occurrence of HSHs. What's more, temperature adjustment is the main element considered in the heat prevention strategy.

Table 7.1 The results of two regression models with calendar effects and temperature effects

	Predictors	Poisson Model			Negative Binomial Model		
		Estimate	Std.Error	P-value	Estimate	Std.Error	P-value
Linear trend	day	-0.0026	0.0018	0.1367	-0.0025	0.0018	0.1602
Year effect	1991	-14.6383	9.7242	0.1322	-14.0735	9.8552	0.1533
	1992	-13.6121	9.0810	0.1339	-13.0590	9.2018	0.1558
	1993	-12.5456	8.4295	0.1367	-11.9813	8.5425	0.1607
	1994	-11.5672	7.7809	0.1371	-11.1398	7.8847	0.1577
	1995	-10.3851	7.1329	0.1454	-9.9296	7.2285	0.1695
	1996	-9.6084	6.4873	0.1386	-9.2023	6.5738	0.1616
	1997	-8.7864	5.8379	0.1323	-8.4492	5.9158	0.1532
	1998	-8.1334	5.1900	0.1171	-7.8840	5.2593	0.1339
	1999	-7.4421	4.5428	0.1014	-7.2264	4.6043	0.1165
	2000	-6.4551	3.9021	0.0981	-6.2613	3.9543	0.1133
	2001	-5.1631	3.2476	0.1119	-5.0446	3.2918	0.1254
	2002	-4.4130	2.6021	0.0899	-4.4090	2.6378	0.0946
	2003	-4.3106	1.9750	0.0291	-4.2640	2.0029	0.0333
	2004	-2.5461	1.3291	0.0554	-2.5127	1.3478	0.0623
	2005	-1.2857	0.6818	0.0593	-1.3067	0.6949	0.0600
Seasonality	$sin(2\pi\omega t)$	-0.0564	0.1222	0.6448	-0.0706	0.1269	0.5781
	$cos(2\pi\omega t)$	0.2794	0.2375	0.2393	0.2687	0.2391	0.2612
	$sin(4\pi\omega t)$	0.1909	0.0850	0.0246	0.1608	0.0878	0.0672
	$cos(4\pi\omega t)$	-0.2398	0.1171	0.0406	-0.2023	0.1202	0.0922
weekday effect	Monday	0.0829	0.1451	0.5681	0.1082	0.1582	0.4940
	Tuesday	0.0926	0.1440	0.5205	0.1078	0.1586	0.4969
	Wednesday	-0.0308	0.1474	0.8346	0.0159	0.1610	0.9215
	Thursday	-0.1781	0.1502	0.2358	-0.1645	0.1678	0.3271
	Friday	-0.1455	0.1539	0.3446	-0.1807	0.1718	0.2927
	Saturday	0.1372	0.1490	0.3569	0.1177	0.1619	0.4671
Temperature effect	Trans (T_{max})	0.8976	0.0701	2.0E-16	0.9930	0.0870	2.0E-16
	Trans (T_{min})	0.1611	0.0634	0.0110	0.1334	0.0758	0.0785
	Trans(T_{max}(lag1))	0.8686	0.2082	3.0E-05	0.8273	0.2380	0.0005
	Trans(T_{max}(lag2))	1.2026	0.1763	9.0E-12	1.1617	0.1986	5.0E-09
Analysis	Null deviance			4139.3			3781.9
	Residual deviance			1801.6			1679.2
	Variability explain by the model			56.48%			55.60%
	AIC			2819.7			2811.1

As is shown in Table 7.1, the variables of calendar effects lose their significance when combined with temperature effects in the same regression model, but the calendar effect can deal with the autocorrelation in model residuals. The significance of calendar effects also lies that they can adjust for the annual, seasonal and weekly variations, some of which are resulted by social factors. Therefore, the calendar effect plays a significant role in explaining the social factor of HSHs.

The lag effects of temperature on HSHs are statistically significant. As is explained by the estimate coefficients, the two previous day's temperatures greatly affect HSHs. The exact health effect of temperature on HSHs within each temperature range can be explained as the incidence rate ration in chapter 7.3.

The Poisson regression model can produce more reliable results under the R software (Version 2.15.2) environment. The null deviances keep constant in all the Poisson regression models, which makes sense, for the null deviance is calculated from an intercept-only model, and it should be constant when using the same dataset and should not be affected by the combination of predictor variables. However, the null deviances of negative binomial models change all the time. Moreover, the negative model cannot cope with some data sets, which are thought to be divergent. The quality of a negative binomial regression model in R is doubtful. The software development is one of this study's, reading its explanation in Chapter 8.1.3

7.2 Model comparison

In this part, this study tests whether a zero inflated Poisson model can be a better fit of the HSHs data. The zero-inflated Poisson model can be a good alternative to explain HSHs, because zeros take up 93% of the dataset.

The **Vuong test** (Equation 7.1) is used to test the zero-inflated model can bring significant improvement in the explanation of HSHs comparing to the Poisson model**.**

$$V = (sqrt(N) * mean(m))/Sm \ (7.1)$$

In equation 7.1, m is the result of $ln(\mu_1/\mu_2)$ where μ_1 is the predicted probability of HSHs for the zero

inflated model; μ_2 is the predicted probability of HSHs for the base model (Poisson); Sm is the standard

deviation of m; N is the number of observations in each model. The test statistic V is asymptotically

normal. If *V>1.96*, the zero-inflated model is preferred. If *V<-1.96*, the base model is preferred. Values of

V between -1.96 and 1.96 indicate that neither model is preferred.

As is tested in R:

V= 4.517918 with p-value 3.12e-06

So the zero-inflated model is statistically preferred over the Poisson model.

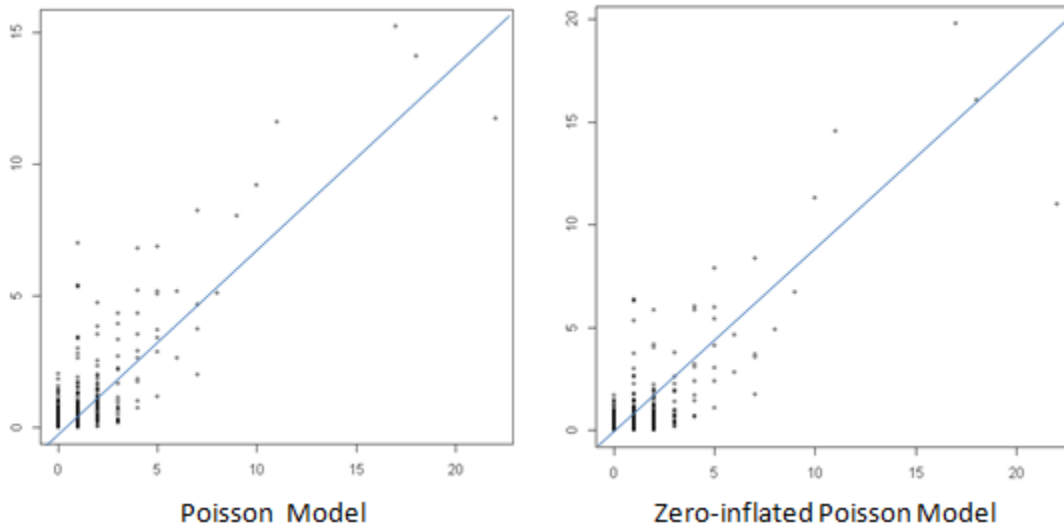Poisson Model   Zero-inflated Poisson Model

Figure 7.1 The scatterplot of observed HSHs and predicted HSHs with Poisson and zero-inflated Poisson Model (x-axis is the observed HSHs and y-axis is the predicted HSHs)

Comparing these two scatter plots in Figure 7.1, the zero-inflated Poisson model can detect those

extremely high HSHs more accurately than a Poisson model.

7.3 Interpretation of coefficients

Interpreting the coefficients under the log link function can be difficult; therefore this study uses the coefficients in the transformed form to explain the change of HSHs in the presence of temperature change.

The measurement of Incidence Rate Ratio (IRR) was defined as the rate of change in the health outcome (106).

$$IRR = \frac{exp\{\beta_0 + \beta_1(f(x+a))\}}{exp\{\beta_0 + \beta_1 f(x)\}} \quad (7.2)$$

In Equation 7.2, β_0 is the intercept estimated by a regression model; β_1 is the coefficient of exposure term estimated by a regression model; $f(x)$ is the expression of exposure, a is the measurement of difference in exposure.

For example, the exposure can be the ambient temperature:

When $f(x) = t$, $f(x+a) = t + a$

$$IRR = \frac{exp\{\beta_0 + \beta_1(t+a)\}}{exp\{\beta_0 + \beta_1(t)\}} = \exp(\beta_1 \times a) \quad (7.3)$$

Therefore, for every degree increase in ambient temperature, i.e. $a = 1$

$$IRR = \frac{exp\{\beta_0 + \beta_1(t+a)\}}{exp\{\beta_0 + \beta_1(t)\}} = \exp(\beta_1) \quad (7.4)$$

In our study, where $f(x) = \exp(\frac{T_{i,j} - Threshold_j}{R})$ (7.5)

In Equation 7.5, $T_{i,j}$ is the transformed daily maximum or minimum temperature, where i is the date from January 1st ,1991 to December 31st ,2006 ; $j = 1$ indicates daily maximum temperature and $j = 2$ indicates daily minimum temperature; $Threshold_{ij}$ are the thresholds of maximum temperature ($j = 1, Threshold_1$ =30.02 °C) and minimum temperature ($j = 2, Threshold_2$ =16.72 °C) (explained in chapter 5.4), exceeding which, there will be more HSHs; R is the median level (10.00 °C) of daily average

86

temperature, which is used to adjust the order of magnitude of the exponential term. The temperature transformation aims to emphasize the exponential increase of negative health outcomes after the threshold and detect those extreme values in HSHs.

Then Equation 7.2 becomes

$$IRR =$$

$$\frac{exp\left\{\beta_0+\beta_1\left(exp(\frac{T_{i+1,j}-Threshold_j}{R})\right)\right\}}{exp\left\{\beta_0+\beta_1\left(exp(\frac{T_{i,j}-Threshold_j}{R})\right)\right\}} = exp\left\{\beta_1\left[exp\left(\frac{T_{i+1,j}-Threshold_j}{R}\right) - exp\left(\frac{T_{i,j}-Threshold_j}{R}\right)\right]\right\} \quad (7.6)$$

The Taylor series of a real or complex-valued function $f(x)$ that is infinitely differentiable in a neighborhood of zero is the power series:

$$f(x) = f(0) + \frac{f'(0)}{1!}(x) + \frac{f''(0)}{2!}(x)^2 + \frac{f'''(0)}{3!}(x)^3 + \cdots \quad (7.7)$$

When $f(x) = e^x$ the approximate result estimated by a Taylor Series is:

$$f(x) = e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24}\cdots \quad (7.8)$$

In our study, Equation (7.5) can be approximated as

$$exp\left(\frac{T_{i,j} - Threshold_j}{R}\right) =$$

$$1 + exp(\frac{T_{i,j}-Threshold_j}{R}) + \frac{exp(\frac{T_{i,j}-Threshold_j}{R})^2}{2} + \frac{exp(\frac{T_{i,j}-Threshold_j}{R})^3}{6} + \frac{exp(\frac{T_{i,j}-Threshold_j}{R})^4}{24}\cdots \quad (7.9)$$

For daily maximum temperature, Equation (7.9) becomes

$$exp(\frac{T_{i,1}-30.02}{10}) = 1 + (\frac{T_{i,1}-30.02}{10}) + \frac{(\frac{T_{i,1}-30.02}{10})^2}{2} + \frac{(\frac{T_{i,1}-30.02}{10})^3}{6} + \frac{(\frac{T_{max}-30.02}{10})^4}{24}\cdots \quad (7.10)$$

For daily minimum temperature, Equation (7.9) becomes

$$\exp(\frac{T_{i,2}-16.72}{10}) \quad = 1 + (\frac{T_{i,2}-16.72}{10}) + \frac{(\frac{T_{i,2}-16.72}{10})^2}{2} + \frac{(\frac{T_{i,2}-16.72}{10})^3}{6} + \frac{(\frac{T_{i,2}-16.72}{10})^4}{24} \cdots \quad (7.11)$$

Equation (7.4) for every degree variation in daily maximum temperature becomes:

$$IRR = \exp\left\{\beta_{T_{max}}\left[(1 + (\frac{T_{i+1,1}-30.02}{10}) + \frac{(\frac{T_{i+1,1}-30.02}{10})^2}{2} + \frac{(\frac{T_{i+1,1}-30.02}{10})^3}{6} + \frac{(\frac{T_{i+1,1}-30.02}{10})^4}{24}) - (1 + \right.\right.$$

$$\left.\left.(\frac{T_{i,1}-30.02}{10}) + \frac{(\frac{T_{i,1}-30.02}{10})^2}{2} + \frac{(\frac{T_{i,1}-30.02}{10})^3}{6} + \frac{(\frac{T_{i,1}-30.02}{10})^4}{24})\right]\right\} \quad (7.12)$$

Equation (7.4) for every degree variation in minimum maximum temperature becomes:

$$IRR = \exp\left\{\beta_{T_{min}}\left[(1 + (\frac{T_{i+1,2}-16.72}{10}) + \frac{(\frac{T_{i+1,2}-16.72}{10})^2}{2} + \frac{(\frac{T_{i+1,2}-16.72}{10})^3}{6} + \frac{(\frac{T_{i+1,2}-16.72}{10})^4}{24}) - (1 + \right.\right.$$

$$\left.\left.(\frac{T_{i,2}-16.72}{10}) + \frac{(\frac{T_{i,2}-16.72}{10})^2}{2} + \frac{(\frac{T_{i,2}-16.72}{10})^3}{6} + \frac{(\frac{T_{i,2}-16.72}{10})^4}{24})\right]\right\} \quad (7.13)$$

In Equation 7.12 and 7.13, the confidents $\beta_{T_{max}}$ and $\beta_{T_{min}}$ are the estimates in Table 7.1 for Trans (T_{max}) and Trans (T_{min}) respectively.
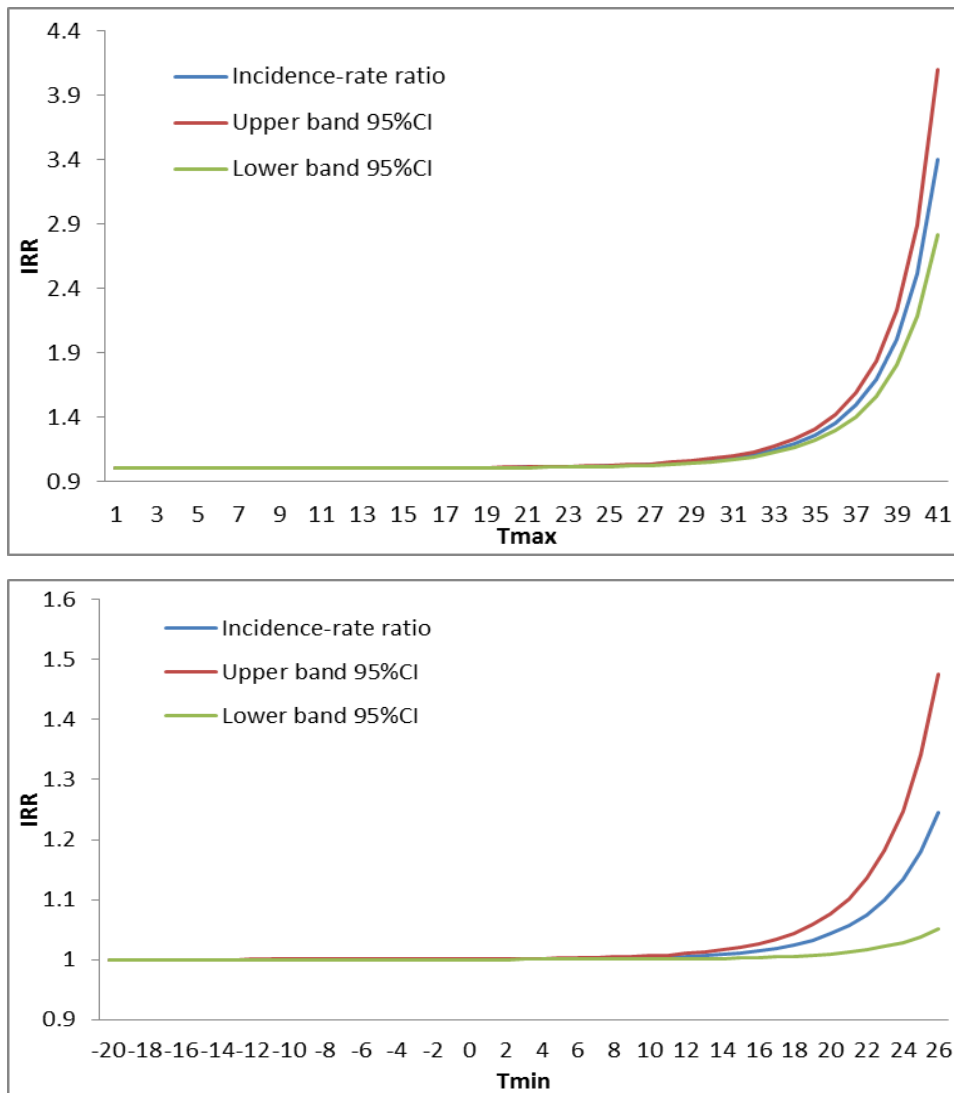
Figure 7.2 IRR of HSHs explained by maximum and minimum temperature (CI: 95%)

When maximum temperature is as low as -20 ˚C, the IRR of HSHs is close to 1, indicating that there is less probability of the occurrence of HSHs. The IRR increases along with the increase of daily maximum temperature and when maximum temperature exceeds the threshold (30.02 ˚C) point, the increase of IRR tends to be much steeper. The IRR is a good explanation of the association between temperature and HSHs: temperature directly affects human health, but exact health risk increases exponentially with the increase of temperature, especially when temperature exceeds the threshold point.

7.4 Conclusion

This study aims to understand the occurrence of HSHs rather than only investigate how much variability can be explained by the model.

The model with both calendar effects and temperature effects can explain 56% variability in HSHs which indicates that firstly, the occurrences of HSHs show regularity and can be partly detected by temperature exposure and calendar effects; secondly, the occurrences of HSHs are also affected by other random factors which cannot be explained by any model systematically; moreover, a statistical model can only summarize the general behaviors of HSHs rather than the accurate reflection of all the extreme events and details.

Specifically speaking, the model can give general guidance on the community level, but it is not adequate enough to provide sufficient suggestions to every individual: the personal health condition, social economic status and adaption to heat events greatly affect the individual's health outcome, even if their exposures are identical. The heterogeneity in exposures will be explained in Chapter 8.

The IRR is a quantitative way to explain the health risk brought on by heat exposure. Yet, it is meaningless to emphasis the exact value of IRR. This study aims to warn elderly residents in the Boston MSA that HSHs are directly correlated with ambient temperature, especially when temperature exceeds

the threshold, because temperature will lead to more acute and dangerous health outcomes with every single degree increase of temperature beyond the threshold.

This study focuses not only on the definition of heat wave, but on how the heat wave's timing and intensity affect human health. This study finds that the annual peak of HSHs follows the first occurrence of the heat wave in about seven days. The first day with daily maximum temperature above the 95th percentile of 16 years' records of each year is the signal of the arrival of the heat wave. Maybe it is hard to persuade the government to start to operate cooling facilities at the occurrence of the first heat wave each year, but we can persuade individuals to gradually adjust their clothing, foods and exercise intensity at least. During that time, the hospitals may need to make special preparations for HSHs and individuals may need to be aware that they should seek medical cares without delay if they have some symptoms of the heat stroke related disease. Earlier and more instant medical care may be the last opportunity to protect lives during the presence of heat events.

7. Discussion

8.1 Study limitation

8.1.1 Potential uncertainty

This study was conducted based on three general assumptions: the heat exposure and HSHs happened at the same time; the individuals' heat exposure happened around their residential area; the ICD-9-CM 992.0 – 992.9 included all the adverse health outcome of heat exposure. In reality, the real life situations may not perfectly support these assumptions.

Temporal factors:

- Delay in HSHs registration date: people may not start feeling sick the exact day when temperature achieves a high level; and people may not go to hospital the day they start feeling sick; or the urgent care patients may not be registered in the insurance system timely when they were sent to hospital. Those last two types of delay may partly explain the temperature's lag effect of adverse health outcomes.

Spatial factors:

- The billing zip code of each beneficiary may not be the location in which the heat exposure happened; what's more the zip code may be incorrectly recorded. And for such a long time interval (16 years), the areas covered by a zip code may change.

Individual and diagnostic:

- Individual have quite different heat vulnerability, and people with chronic disease are quite vulnerable to heat exposure, however when they are sent to hospital during the heat wave period; the heat stroke is less likely to be diagnosed as their primary cause of hospitalization.

what's more, there is no standard definition for heat events related disease, and this study

extracted cases with diagnostic codes ICD-9-CM 992.0 – 992.9, which explain the effect of heat

and light, but sun burns, burns, diseases of sweat glands due to heat, malignant hyperpyrexia

following anesthesia are not included. Thus the records may not completely reflect all the

negative health outcomes caused by heat events.

8.1.2 Heterogeneity

This utilized individual records to analyze the health outcomes in the Boston-Cambridge-Quincy MA NH

Metropolitan Statistical Area (Boston MSA) without taking the heterogeneity into account. Adequate

considerations of heterogeneity may facilitate the understanding of health risk on the individual level.

- Heterogeneity in individual exposure

The individual's exact temperature exposure may be significantly various or may deviate greatly from

the entire community' exposure. I conducted the regression analysis using the average temperature for

the whole Boston MSA, which is an approximate exposure measurement for each sub-location within

this MSA. What's more, the exact individual exposure is actually unknown, because the insurance

system only recorded the billing zip codes of patients and lacked the records of exact locations in which

the hospitalization happened; the insurance system only recorded the HSHs date and lacked the exact

time. Even though we may apply remote sensing tools to extract temperature records on a smaller

temporal and spatial resolution, the exact individual temperature exposure is still immeasurable.

- Heterogeneity in individual status

Even if the temperature exposure were assumed to be identical for every individual, the heterogeneity

of individual status such as health condition, demographical status, social economic status and the

adaptation to heat may also lead to various health outcomes. The health condition of each individual

greatly determines his vulnerability to heat; gender and race also indicate different adaptation abilities

93

to heat. The personal exposure to heat is closely associated with individuals' job types and educational levels. The exterior temperature exposure may not really affect human health if they live or work in houses with good air conditioning systems; therefore the housing and facility types also affect people's heat risk.

8.1.3 Understanding and application of statistical theory

The negative binomial regression is thought to be a useful accommodating of otherwise overdispersed Poisson data. However, within the GLM framework, the link of negative binomial regression attracted continuing discussion: the canonical link (i.e. $\eta = -\ln(\frac{1}{\alpha\mu} + 1)$) is stable, but it has properties that often result in non-convergence [96].

The development of software may limit the application of statistical theory in practice, and the regression results are determined by the application of algorithms and the setting of upper limit of iterations. In this study, the negative binomial is expected to be a better fit of the hospitalization data theoretically, but when I try to conduct the regression analysis using generalized linear model with negative binomial distribution, I am always warned that the upper limits of iterations have been achieved, thus the negative binomial regression doesn't show its superiority. Moreover, the MASS package in R software (Version 2.15.2) applies a log function $\eta = \log(\mu)$)to link the linear predictor and mean parameter in the generalized negative binomial regression model. The log link doesn't sufficiently consider the dispersion parameter in a negative binomial distribution; therefore the superiority of a negative binomial model does not appear either. In terms of algorithm, maximum likelihood (ML) and an iteratively re-weighted least squares (IRLS) are two main approaches to estimating models of count data. IRLS is intrinsic to the estimation of generalized linear models, as well as to certain extensions to the generalized linear model algorithm [96]. In reality, the strengths of one algorithm over the other decide on the dataset, thus it is worth comparing more algorithms to improve the regression result.

8.2 Future direction

8.2.1 A more refined spatial resolution

Aggregation to HSHs data to the MSA level can guarantee adequate elderly population base, while this study applied single temperature measurement to represent the average heat exposure in a sufficiently large area and didn't consider the heterogeneity in demographics and social-economic status in each sub location.

Urban areas may have quite different temperature patterns from those of rural patterns, which may lead to various health outcomes. If this study defines the urban areas according to local meteorological features and then investigate the association between ambient temperature and HSHs in the rural and urban areas respectively, a more specific conclusion with a sufficiently consideration of the social factor as a modifier of health outcomes can be achieved.

The contextual demographic and social-economic status may not always accurately reflect individual variables. If this study can be conducted on a more refined spatial resolution (zip code or county), the roles of demographic and social economic statuses in HSH in the presence of the similar temperature exposures will be investigated.

8.2.2 The role of technical adaptation

This study currently applies the ambient temperature recorded by ground monitor stations as the measurement of temperature exposure; however, the exact temperature exposure of each individual varies greatly. Elderly people's exposure to ambient temperature is even more limited; instead, the housing temperature plays a more significant role in their health outcomes.

The technical adaptation embraces the design of indoor ventilation systems, the efficiency of air conditioning systems, and the covering of green plants surroundings. Those factors are correlated with

an individual's social-economic status; therefore an adequate consideration of the technical adaptation's effects on human health can facilitate our understanding of HSHs.

8.2.3 Evaluation of the health protection system

A complete health protection system includes both prevention strategies before the occurrence of heat related diseases and health care after the occurrence of negative health risks. The main issue facing us is not the lack of such a health protection system, but its financial support and efficacy.

We should realize that the interplay between the financial support and efficacy of health protection systems: a system that fails to get financial support from government will not operate efficiently, while a system without adequate scientific evidence to verify its efficacy will never be accepted by policy makers.

We are most concerned with the quality of health protection information and it is urgent to provide health protection suggestions with solid scientific evidence and propose the exact local suggestion with a full consideration of local economic and climate characteristics. From the national and global level, it is quite significant to analyze the contradictory information, such as the application of fans and consumption of caffeinated drinks, which may either ease or increase people's heat burden, depending upon people's living regions and long term adaption and acclimation. It is worth considering and evaluating the effectiveness of health protection strategy for each region, and then specifying the effective region of each suggestion.

In reality, it is worth evaluating a health protection system's costs and benefits, for example, the health care savings and potential decreases in mortality and morbidity. What's more, targeting the most vulnerable population may greatly increase the efficiency of a health protection system.

8.2.4 Vulnerability map

The vulnerability of each community is affected by its demographics, economic level, housing type, accessibility to medical care, and climate characteristics. Visualizing the heat risk within a certain community is a vivid way to warn people of the potential adverse health effects. Vulnerability is defined in one instance as the summation of all risk and protective factors that ultimately determine whether an individual or subpopulation experiences adverse health outcomes. The vulnerability maps comprise two parts: historical evidences and vulnerable factors. Mapping the already existed adverse health outcomes for each community can help individuals to realize their community's risk history, and also provide information for new settlers to select their new houses. Mapping the risk factors for each community can provide information for residents to decide what they can do to protect themselves from heat risk if they decide to live in the community.

Appendix

1.   Data source of explanations of the heat exposure's symptom:

"HEAT STRESS," last modified May 28, 2013. http://www.cdc.gov/niosh/topics/heatstress/

2.   Data source of explanations of ICD codes:

"ICD-9-CM," http://icd9cm.chrisendres.com

3.   Data sources of demographic information for each city:

"Los Angeles," last modified May 28, 2013. http://en.wikipedia.org/wiki/Los_Angeles#2010_Census

"CENSUS 2010: Changes in the Elderly Population of New York City, 2000 to 2010," last modified July 16,

2012. http://www.nyc.gov/html/dfta/downloads/pdf/demographic/elderly_population_070912.pdf

"Chicago," last modified May 28, 2013. http://en.wikipedia.org/wiki/Demographics_of_Chicago

"Houston," last modified May 28, 2013. http://en.wikipedia.org/wiki/Demographics_of_Houston

"Jacksonville," last modified May 28, 2013. http://en.wikipedia.org/wiki/Jacksonville,_Florida

"Atlanta," last modified May 28, 2013. http://www.citytowninfo.com/places/georgia/atlanta

"Washington D.C.," last modified May 28, 2013. http://quickfacts.census.gov/qfd/states/53000.html

"Detroit," last modified May 28, 2013. http://en.wikipedia.org/wiki/Detroit#Demographics

"Philadelphia," last modified May 28, 2013. http://en.wikipedia.org/wiki/Philadelphia#Demographics

"Phoenix," last modified May 28, 2013. http://en.wikipedia.org/wiki/Phoenix,_Arizona#Demographics

"San Antonio," last modified May 28, 2013.   http://en.wikipedia.org/wiki/San_Antonio#Demographics

"San Diego," last modified May 28, 2013.   http://en.wikipedia.org/wiki/San_Diego#Demographics

"Dallas," last modified May 28, 2013.   http://en.wikipedia.org/wiki/Dallas#Demographics

"San Jose," last modified May 28, 2013.

http://en.wikipedia.org/wiki/San_Jose,_California#Demographics

"Austin," last modified May 28, 2013.   http://en.wikipedia.org/wiki/Austin,_Texas#Demographics

"San Francisco," last modified May 28, 2013.

http://en.wikipedia.org/wiki/San_Francisco#Demographics

"Columbus," last modified May 28, 2013.   http://en.wikipedia.org/wiki/Columbus,_Ohio#Demographics

"Charlotte," last modified May 28, 2013.

http://en.wikipedia.org/wiki/Charlotte,_North_Carolina#Demographics

"Boston," last modified May 28, 2013.   http://en.wikipedia.org/wiki/Boston

"Seattle," last modified May 28, 2013.   http://en.wikipedia.org/wiki/Seattle

4.   Data sources of climate information for each city:

"Current Results," last modified May 28, 2013.   http://www.currentresults.com/lather/US/average-

annual-temperatures-large-cities.php

5.   Data sources of the enzyme action curve:

"How Temperature Affects Reaction Rate of Enzyme Reactions," last modified May 28, 2013.

http://science.halleyhosting.com/sci/soph/enzyme/enzfactorstemp.htm

Reference

1. Pachauri, R.K., *Climate change 2007. Synthesis report. Contribution of Working Groups I, II and III to the fourth assessment report.* 2008.
2. Kinney, P.L., et al., *Approaches for estimating effects of climate change on heat-related deaths: challenges and opportunities.* environmental science & policy, 2008. **11**(1): p. 87-96.
3. Hales, S., et al., *Daily mortality in relation to weather and air pollution in Christchurch, New Zealand.* Australian and New Zealand Journal of Public Health, 2000. **24**(1): p. 89-91.
4. O'Neill, M.S., et al., *Preventing heat-related morbidity and mortality: new approaches in a changing climate.* Maturitas, 2009. **64**(2): p. 98.
5. Anderson, G.B. and M.L. Bell, *Heat waves in the United States: mortality risk during heat waves and effect modification by heat wave characteristics in 43 US communities.* Environmental Health Perspectives, 2011. **119**(2): p. 210.
6. McGeehin, M.A. and M. Mirabelli, *The potential impacts of climate variability and change on temperature-related morbidity and mortality in the United States.* Environmental Health Perspectives, 2001. **109**(Suppl 2): p. 185.
7. Warrick, R.A. and E.M. Barrow, *Climate change scenarios for the UK.* Transactions of the Institute of British Geographers, 1991: p. 387-399.
8. Kalkstein, L.S. and J.S. Greene, *An evaluation of climate/mortality relationships in large US cities and the possible impacts of a climate change.* Environmental Health Perspectives, 1997. **105**(1): p. 84.
9. Basu, R. and J.M. Samet, *Relation between elevated ambient temperature and mortality: a review of the epidemiologic evidence.* Epidemiologic reviews, 2002. **24**(2): p. 190-202.
10. Anderson, B.G. and M.L. Bell, *Weather-related mortality: how heat, cold, and heat waves affect mortality in the United States.* Epidemiology (Cambridge, Mass.), 2009. **20**(2): p. 205.
11. Bouchama, A. and J.P. Knochel, *Heat stroke.* New England Journal of Medicine, 2002. **346**(25): p. 1978-1988.
12. Wang, X.Y., et al., *Temperature variation and emergency hospital admissions for stroke in Brisbane, Australia, 1996–2005.* International journal of biometeorology, 2009. **53**(6): p. 535-541.
13. Hansen, A.L., et al., *The effect of heat waves on hospital admissions for renal disease in a temperate city of Australia.* International journal of Epidemiology, 2008. **37**(6): p. 1359-1365.
14. Ren, C. and S. Tong, *Temperature modifies the health effects of particulate matter in Brisbane, Australia.* International journal of biometeorology, 2006. **51**(2): p. 87-96.
15. Ohshige, K., et al., *Influence of weather on emergency transport events coded as stroke: population-based study in Japan.* International journal of biometeorology, 2006. **50**(5): p. 305-311.
16. Piver, W.T., et al., *Temperature and air pollution as risk factors for heat stroke in Tokyo, July and August 1980-1995.* Environmental Health Perspectives, 1999. **107**(11): p. 911.
17. Son, J.-Y., et al., *The impact of heat waves on mortality in seven major cities in Korea.* Environmental Health Perspectives, 2012. **120**(4): p. 566.
18. Vandentorren, S., et al., *August 2003 heat wave in France: risk factors for death of elderly people living at home.* The European Journal of Public Health, 2006. **16**(6): p. 583-591.
19. Abrahamson, V., et al., *Perceptions of heatwave risks to health: interview-based study of older people in London and Norwich, UK.* Journal of Public Health, 2009. **31**(1): p. 119-126.
20. Hajat, S., et al., *Impact of hot temperatures on death in London: a time series approach.* Journal of Epidemiology and Community Health, 2002. **56**(5): p. 367-372.

21.   Kyselý, J., *Mortality and displaced mortality during heat waves in the Czech Republic.* International journal of biometeorology, 2004. **49**(2): p. 91-97.

22.   Dawson, J., et al., *Associations between meteorological variables and acute stroke hospital admissions in the west of Scotland.* Acta Neurologica Scandinavica, 2008. **117**(2): p. 85-89.

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

24.   Linares, C. and J. Díaz, *Impact of high temperatures on hospital admissions: comparative analysis with previous studies about mortality (Madrid).* The European Journal of Public Health, 2008. **18**(3): p. 317-322.

25.   Hajat, S. and A. Haines, *Associations of cold temperatures with GP consultations for respiratory and cardiovascular disease amongst the elderly in London.* International journal of Epidemiology, 2002. **31**(4): p. 825-830.

26.   Ballester, F., et al., *Mortality as a function of temperature. A study in Valencia, Spain, 1991-1993.* International journal of Epidemiology, 1997. **26**(3): p. 551-561.

27.   Wichmann, J., et al., *Apparent temperature and cause-specific emergency hospital admissions in Greater Copenhagen, Denmark.* PloS one, 2011. **6**(7): p. e22904.

28.   Bhaskaran, K., et al., *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, 2010. **341**.

29.   Juopperi, K., et al., *Incidence of frostbite and ambient temperature in Finland, 1986-1995. A national study based on hospital admissions.* International Journal of Circumpolar Health, 2002. **61**(4).

30.   Hajat, S., W. Bird, and A. Haines, *Cold weather and GP consultations for respiratory conditions by elderly people in ⌐ 16 locations in the UK.* European journal of epidemiology, 2004. **19**(10): p. 959-968.

31.   D'Ippoliti Daniela, M.P., et al., *The impact of heat waves on mortality in 9 European cities: results from the EuroHEAT project.* Environmental Health. **9**.

32.   Smoyer, K., *A comparative analysis of heat waves and associated mortality in St. Louis, Missouri– 1980 and 1995.* International journal of biometeorology, 1998. **42**(1): p. 44-50.

33.   Johnson, H., et al., *The impact of the 2003 heat wave on mortality and hospital admissions in England.* Health Statistics Quarterly, 2005(25): p. 6-11.

34.   Bell, M.L., et al., *Vulnerability to heat-related mortality in Latin America: a case-crossover study in Sao Paulo, Brazil, Santiago, Chile and Mexico City, Mexico.* International journal of Epidemiology, 2008. **37**(4): p. 796-804.

35.   Rainham, D.G. and K.E. Smoyer-Tomic, *The role of air pollution in the relationship between a heat stress index and human mortality in Toronto.* Environmental research, 2003. **93**(1): p. 9-19.

36.   Basu, R., F. Dominici, and J.M. Samet, *Temperature and mortality among the elderly in the United States: a comparison of epidemiologic methods.* Epidemiology, 2005. **16**(1): p. 58-66.

37.   Bayentin, L., et al., *Spatial variability of climate effects on ischemic heart disease hospitalization rates for the period 1989–2006 in Quebec, Canada.* International journal of health geographics, 2010. **9**(5).

38.   Curriero, F.C., et al., *Temperature and mortality in 11 cities of the eastern United States.* American journal of epidemiology, 2002. **155**(1): p. 80-87.

39.   Stafoggia, M., et al., *Vulnerability to heat-related mortality: a multicity, population-based, case-crossover analysis.* Epidemiology, 2006. **17**(3): p. 315-323.

40.	Muggeo, V.M. and S. Hajat, *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, 2009. **66**(9): p. 584-591.

41.	Filleul, L., et al., *The relation between temperature, ozone, and mortality in nine French cities during the heat wave of 2003.* Environmental Health Perspectives, 2006. **114**(9): p. 1344.

42.	O'Neill, M.S., et al., *Impact of control for air pollution and respiratory epidemics on the estimated associations of temperature and daily mortality.* International journal of biometeorology, 2005. **50**(2): p. 121-129.

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

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

45.	Baccini, M., et al., *Heat effects on mortality in 15 European cities.* Epidemiology, 2008. **19**(5): p. 711-719.

46.	Michelozzi, P., et al., *Temperature and summer mortality: geographical and temporal variations in four Italian cities.* Journal of Epidemiology and Community Health, 2006. **60**(5): p. 417-423.

47.	Braga, A.L.F., A. Zanobetti, and J. Schwartz, *The time course of weather-related deaths.* Epidemiology, 2001. **12**(6): p. 662-667.

48.	Kyobutungi, C., et al., *Absolute temperature, temperature changes and stroke risk: a case-crossover study.* European journal of epidemiology, 2005. **20**(8): p. 693-698.

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

50.	Lin, S., et al., *Extreme high temperatures and hospital admissions for respiratory and cardiovascular diseases.* Epidemiology, 2009. **20**(5): p. 738-746.

51.	Keatinge, W., et al., *Heat related mortality in warm and cold regions of Europe: observational study.* BMJ: British Medical Journal, 2000. **321**(7262): p. 670.

52.	Turner, L.R., et al., *Ambient temperature and cardiorespiratory morbidity: a systematic review and meta-analysis.* Epidemiology, 2012. **23**(4): p. 594-606.

53.	Bell, M.L., et al., *Ozone and short-term mortality in 95 US urban communities, 1987-2000.* JAMA: the journal of the American Medical Association, 2004. **292**(19): p. 2372-2378.

54.	Semenza, J.C., et al., *Excess hospital admissions during the July 1995 heat wave in Chicago.* American journal of preventive medicine, 1999. **16**(4): p. 269-277.

55.	Rydman, R.J., et al., *The rate and risk of heat-related illness in hospital emergency departments during the 1995 Chicago heat disaster.* Journal of medical systems, 1999. **23**(1): p. 41-56.

56.	Knowlton, K., et al., *The 2006 California heat wave: impacts on hospitalizations and emergency department visits.* Environmental Health Perspectives, 2009. **117**(1): p. 61.

57.	Ebi, K., et al., *Weather changes associated with hospitalizations for cardiovascular diseases and stroke in California, 1983–1998.* International journal of biometeorology, 2004. **49**(1): p. 48-58.

58.	Danet, S., et al., *Unhealthy effects of atmospheric temperature and pressure on the occurrence of myocardial infarction and coronary deaths: a 10-year survey: the Lille-World Health Organization MONICA project (Monitoring Trends and Determinants in Cardiovascular Disease).* Circulation, 1999. **100**(1): p. e1-e7.

59.	Zeka, A., A. Zanobetti, and J. Schwartz, *Individual-level modifiers of the effects of particulate matter on daily mortality.* American journal of epidemiology, 2006. **163**(9): p. 849-859.

60.	Liang, W.-M., et al., *Ambient temperature and emergency room admissions for acute coronary syndrome in Taiwan.* International journal of biometeorology, 2008. **52**(3): p. 223-229.

61.   Barnett, A.G., et al., *Cold periods and coronary events: an analysis of populations worldwide.* Journal of Epidemiology and Community Health, 2005. **59**(7): p. 551-557.

62.   Creason, J., et al., *Particulate matter and heart rate variability among elderly retirees: the Baltimore 1998 PM study.* Journal of exposure analysis and environmental epidemiology, 2001. **11**(2): p. 116.

63.   Gold, D.R., et al., *Ambient pollution and heart rate variability.* Circulation, 2000. **101**(11): p. 1267-1273.

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

65.   Hong, Y.-C., et al., *Ischemic stroke associated with decrease in temperature.* Epidemiology, 2003. **14**(4): p. 473-478.

66.   Koken, P.J., et al., *Temperature, air pollution, and hospitalization for cardiovascular diseases among elderly people in Denver.* Environmental Health Perspectives, 2003. **111**(10): p. 1312.

67.   Braga, A.L., A. Zanobetti, and J. Schwartz, *The effect of weather on respiratory and cardiovascular deaths in 12 US cities.* Environmental Health Perspectives, 2002. **110**(9): p. 859.

68.   Bhaskaran, K., et al., *Effects of ambient temperature on the incidence of myocardial infarction.* Heart, 2009. **95**(21): p. 1760-1769.

69.   Cheng, X. and H. Su, *Effects of climatic temperature stress on cardiovascular diseases.* European journal of internal medicine, 2010. **21**(3): p. 164.

70.   Ye, F., et al., *Effects of temperature and air pollutants on cardiovascular and respiratory diseases for males and females older than 65 years of age in Tokyo, July and August 1980-1995.* Environmental Health Perspectives, 2001. **109**(4): p. 355.

71.   Dominici, F., et al., *Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases.* JAMA: the journal of the American Medical Association, 2006. **295**(10): p. 1127-1134.

72.   Toulemon, L. and M. Barbieri, *The mortality impact of the August 2003 heat wave in France: Investigating the 'harvesting'effect and other long-term consequences.* Population studies, 2008. **62**(1): p. 39-53.

73.   Hajat, S., M. O'Connor, and T. Kosatsky, *Health effects of hot weather: from awareness of risk factors to effective health protection.* Lancet (London, England), 2010. **375**(9717): p. 856.

74.   Medina-Ramón, M., et al., *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, 2006. **114**(9): p. 1331.

75.   Kalkstein, L.S., et al., *The Philadelphia hot weather-health watch/warning system: development and application, summer 1995.* Bulletin of the American Meteorological Society, 1996. **77**(7): p. 1519-1528.

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

77.   Mastrangelo, G., et al., *Pattern and determinants of hospitalization during heat waves: an ecologic study.* BMC Public Health, 2007. **7**(1): p. 200.

78.   Hajat, S., et al., *Impact of high temperatures on mortality: is there an added heat wave effect?* Epidemiology, 2006. **17**(6): p. 632-638.

79.   Barnett, A.G., S. Tong, and A. Clements, *What measure of temperature is the best predictor of mortality?* Environmental research, 2010. **110**(6): p. 604-611.

80.   Höppe, P., *The physiological equivalent temperature–a universal index for the biometeorological assessment of the thermal environment.* International journal of biometeorology, 1999. **43**(2): p. 71-75.

81. Medina-Ramón, M. and J. Schwartz, *Temperature, temperature extremes, and mortality: a study of acclimatisation and effect modification in 50 US cities.* Occupational and environmental medicine, 2007. **64**(12): p. 827-833.

82. Cerutti, B., et al., *Temperature related mortality and ambulance service interventions during the heat waves of 2003 in Ticino (Switzerland).* Sozial-und Präventivmedizin, 2006. **51**(4): p. 185-193.

83. Michelozzi, P., et al., *Impact of heat waves on mortality-Rome, Italy, June-August 2003.* MMWR, 2004. **53**(17): p. 369-371.

84. Lan Chang, C., et al., *Lower ambient temperature was associated with an increased risk of hospitalization for stroke and acute myocardial infarction in young women.* Journal of clinical epidemiology, 2004. **57**(7): p. 749-757.

85. Davis, R.E., et al., *Changing heat-related mortality in the United States.* Environmental Health Perspectives, 2003. **111**(14): p. 1712.

86. Höppe, P., *Aspects of human biometerology in past, present and future.* International journal of biometeorology, 1997. **40**(1): p. 19-23.

87. Donaldson, G., W. Keatinge, and S. Näyhä, *Changes in summer temperature and heat-related mortality since 1971 in North Carolina, South Finland, and Southeast England.* Environmental research, 2003. **91**(1): p. 1-7.

88. Davis, R.E., et al., *Decadal changes in heat-related human mortality in the eastern United States.* Climate Research, 2002. **22**(2): p. 175-184.

89. Hajat, S. and T. Kosatky, *Heat-related mortality: a review and exploration of heterogeneity.* Journal of Epidemiology and Community Health, 2010. **64**(9): p. 753-760.

90. Chestnut, L.G., et al., *Analysis of differences in hot-weather-related mortality across 44 US metropolitan areas.* environmental science & policy, 1998. **1**(1): p. 59-70.

91. Michelozzi, P., et al., *Air pollution and daily mortality in Rome, Italy.* Occupational and environmental medicine, 1998. **55**(9): p. 605-610.

92. Ostro, B., et al., *The effects of temperature and use of air conditioning on hospitalizations.* American journal of epidemiology, 2010. **172**(9): p. 1053-1061.

93. Kovats, R.S. and S. Hajat, *Heat stress and public health: a critical review.* Annu. Rev. Public Health, 2008. **29**: p. 41-55.

94. Oberlin, M., et al., *Heat-related illnesses during the 2003 heat wave in an emergency service.* Emergency medicine journal, 2010. **27**(4): p. 297-299.

95. Naumova, E.N. and I.B. MacNeill, *Seasonality assessment for biosurveillance systems*, in *Advances in Statistical Methods for the Health Sciences*2007, Springer. p. 437-450.

96. Hilbe, J.M., *Negative binomial regression*2011: Cambridge University Press.

97. Hardin, J.W. and J. Hilbe, *Generalized linear models and extensions*2007: Stata Corp.

98. McCullagh, P. and J.A. Nelder, *Generalized linear model*. Vol. 37. 1989: Chapman & Hall/CRC.

99. Gillooly, J.F., et al., *Effects of size and temperature on metabolic rate.* Science, 2001. **293**(5538): p. 2248-2251.