UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

The impact of heatwaves on mortality in the Lisbon district – ICARO system revisited

Maria Carolina Serrão Nogueira Nascimento Bulhosa

Mestrado em Bioestatística

Relatório de Estágio orientado por:
Professora Doutora Marília Antunes
Doutor Baltazar Nunes

2019

Agradecimentos

Gostaria de agradecer, em primeiro lugar, aos meus orientadores, Professora Marília Antunes e Doutor Baltazar Nunes, por me terem possibilitado a realização deste trabalho de estágio e guiado ao longo de todo o processo. Não posso deixar de agradecer à Susana Silva, que não sendo oficialmente minha orientadora, foi sempre de uma dedicação e ajuda preciosas.

Gostaria de agradecer também a todos os colegas do Departamento de Epidemiologia do Instituto Nacional de Saúde Doutor Ricardo Jorge, por me terem acolhido tão amigavelmente durante estes meses. Um especial muito obrigada à Inês Batista. Ao Constantino, o meu obrigada por toda a ajuda na programação.

Quero ainda agradecer aos meus colegas de curso, por toda a amizade e colaboração ao longo destes dois anos.

Por fim, mas de extrema importância, o meu eterno obrigada à minha família e amigos mais próximos, por todo o apoio emocional e não só. Aos meus pais, Márcio, Marta, Isabel, Sofia, Rita, Pedro e João, muito obrigada!

Resumo

A temperatura é um fator ambiental com influência direta no conforto e saúde humana. Sabe-se que diferentes populações estão adaptadas a diferentes climas, consoante as regiões onde vivem, e que se adaptam inclusive às alterações sazonais que ocorrem ao longo do ano, tornando-se mais tolerantes ao frio ao longo do inverno e ao calor ao longo do verão. Ainda assim, temperaturas extremas (tanto quentes como frias) podem causar um aumento da mortalidade. Normalmente a mortalidade é mais elevada no inverno do que no verão, embora esporadicamente ocorram picos de mortalidade durante alguns verões, que podem ser explicados por calor extremo. O foco deste trabalho será este excesso de mortalidade associada a calor extremo no distrito de Lisboa.

O estudo do efeito que o excesso de calor tem sobre a mortalidade é de extrema importância hoje em dia, tendo em conta o período de alterações climáticas em que vivemos - que está a provocar o aumento das temperaturas e a intensidade e frequência de fenómenos extremos, como subidas súbitas da temperatura e ondas de calor – e também devido ao aumento de certos fatores de risco que aumentam a vulnerabilidade da população ao calor, nomeadamente a crescente urbanização e envelhecimento da população portuguesa. Assim, a relação entre a temperatura e a mortalidade em Portugal deve ser profundamente estudada, de modo a tentar prever e atenuar as consequências do aquecimento global, com especial atenção à população mais vulnerável e com menos capacidades de adaptação.

Em Portugal, o excesso de calor nos verões de 1981 e 1991 provocou um número de mortes muito acima do número de mortes esperadas, que poderia ter sito evitado, pelo menos em parte, caso existisse algum sistema de aviso e resposta a emergências relacionadas com o calor. Surgiu assim a necessidade de se desenvolver um sistema de alerta e vigilância de calor, com o objetivo de mitigar esse efeito. Desta forma, criou-se em 1999 o sistema ÍCARO – Importância do Calor: Repercussão sobre os Óbitos -, que funciona desde então entre os meses de maio e setembro, visando monitorizar e alertar para possíveis aumentos de mortalidade causados por calor extremo. Para isso foi inicialmente criado um modelo estatístico de séries temporais com um limiar dinâmico, calibrado para o distrito de Lisboa, considerando os dados relativos às ondas de calor de 1981 e 1991: modelo ÍCARO Lisboa. Este modelo identifica bem os períodos de calor extremo com impacto sobre a mortalidade. No entanto, por um lado, sobrestima alguns dos aumentos de mortalidade, dando origem a falsos alarmes e, por outro lado, subestima os impactos tardios de ondas de calor mais prolongadas.

O objetivo deste trabalho é, assim, formular um novo modelo estatístico que relacione o calor e a mortalidade com o mesmo grau de sensibilidade, mas maior especificidade que o modelo ÍCARO Lisboa. Esse novo modelo será então desenvolvido e depois comparado com o modelo ÍCARO, com vista a optimizar/atualizar o sistema de vigilância.

Sendo que o efeito das temperaturas adversas sobre a mortalidade não é imediato, podendo prolongar-se no tempo ao longo de vários dias, este deve ser explicado pela combinação de exposições ao longo de vários momentos do passado, dependendo simultaneamente da intensidade e do momento das exposições ao calor. Desenvolveu-se então um modelo não linear de desfasamentos distribuídos (DLNM – *dis-*

tributed lag non-linear models). Esta família de modelos tem em conta os diferentes efeitos que as temperaturas de cada dia vão ter na mortalidade de um dia futuro, já que inclui uma dimensão temporal que permite predizer de forma mais eficaz o efeito atrasado que o calor pode causar na mortalidade, bem como o seu efeito cumulativo. Os DLNM baseiam-se numa função bidimensional, denominada "*cross-basis*", que descreve simultaneamente as dependências de exposição-resposta ao longo do espaço da temperatura e ao longo da sua estrutura temporal. Deste modo, o efeito de cada exposição até um determinado atraso – *lag* - máximo pode ser somado para obter o efeito cumulativo da temperatura sobre a mortalidade. Tal obtém-se tendo em conta a interação entre duas "funções-base", para modelar cada uma destas duas dimensões (uma função para a exposição-resposta e outra para o desfasamento no tempo-resposta, que juntas originam a função *cross-basis*, que é o núcleo dos DLNM). A função *cross-basis* origina uma série de transformações sobre a variável preditora, originando novas variáveis, conhecidas por variáveis-*cross-basis*, que irão entrar no modelo de regressão em vez das variáveis originais. As duas funções-base que compõem a *cross-basis* podem ser escolhidas independentemente uma da outra a partir de um variado leque de opções. Neste trabalho foram testadas várias hipóteses, tendo-se usado o critério de informação de Akaike modificado para escolher o melhor modelo.

Diversos estudos previamente publicados comprovam a eficácia de modelos da família dos DLNM para estudar a relação entre temperaturas extremas e a mortalidade em Lisboa. No entanto, este projeto distingue-se por ser o primeiro que usa este método para estudar apenas o período de verão, considerar a mortalidade por todas as causas e usar uma base de dados tão alargada (correspondente a 38 anos). Assim, o modelo escolhido foi calibrado com dados restritos aos meses mais quentes do ano (maio a setembro), referentes ao distrito de Lisboa no período entre 1980 e 2017, relacionando as temperaturas máximas diárias com a mortalidade total diária, que se assume proveniente de uma distribuição de *Poisson* sobredispersa.

Decidiu-se ajustar um modelo com a função *cross-basis* composta por uma função *B-spline* de grau 2 com 5 nós na dimensão do preditor e uma função *B-spline* de grau 3 com 3 nós na dimensão dos desfasamentos no tempo, permitindo um desfasamento máximo de 10 dias. A sazonalidade e tendência foram controladas através de funções do tempo, usando *B-splines* cúbicos com 5 e 4 graus de liberdade, respetivamente, e ainda se considerou o dia da semana como variável categórica e a população média anual como variável linear. Este modelo foi então usado para prever o risco relativo de morte, considerando a mortalidade esperada com 25.7 ^{o}C como referência.

Obteve-se um aumento acentuado do risco relativo para temperaturas superiores a 40 ^{o}C entre *lags* 0 e 4. Ou seja, temperaturas assim tão elevadas têm impactos negativos sobre a mortalidade no próprio dia, que se prolongam até 4 dias depois. Atingindo-se o máximo risco relativo para 43 ^{o}C com o desfasamento de 1 dia.

Por outro lado, no verão, temperaturas máximas inferiores a 15 ^{o}C parecem ser protetoras no próprio dia em que ocorrem, efeito que se mantém até 3 dias depois. Atingindo-se o mínimo risco relativo para 11 ^{o}C com o desfasamento de 1 dia.

Tendo em conta o efeito cumulativo, temperaturas menores que 15 ^{o}C são protetoras enquanto que temperaturas superiores a 30 ^{o}C já aumentam o risco de morte, risco este que vai aumentando cada vez mais para temperaturas mais altas.

A avaliação do DLNM final e a sua comparação com o modelo ÍCARO para Lisboa foi feita através de validação cruzada. Este é um método através do qual se avaliam as capacidades preditivas dos modelos, testando-os em novos dados que não tenham sido tidos em conta quando se ajustaram os modelos.

Comparando as predições feitas por cada um dos modelos e as contagens de mortes diárias observadas nos verões testados, percebe-se que ambos têm uma boa capacidade de detetar os maiores picos

de mortalidade relacionados com o calor extremo. Porém, os seus comportamentos diferem em relação às alterações menos acentuadas da mortalidade. O DLNM é capaz de acompanhar melhor todas as alterações no número de mortes (tanto os aumentos como as diminuições) do que o ÍCARO, mas tende a subestimar a magnitude dos picos de mortalidade menos acentuados. Esta característica tem a vantagem de provocar menos falsos alarmes que o modelo ÍCARO, no entanto, pode não detetar alguns excessos de risco, para os quais deveria ser dado um alerta que permita uma resposta adequada para proteção da população.

Este modelo tem ainda outras limitações, nomeadamente não ter em conta algumas possíveis variáveis de confundimento, como a poluição atmosférica e outras variáveis meteorológicas, e fatores de risco individuais, como a idade, o género e doenças pré-existentes, entre outras. Apesar disso, acredita-se que o modelo não linear de desfasamentos distribuídos obtido capta pelo menos as principais características da relação entre o calor extremo e a mortalidade.

Surgem como possíveis futuros desenvolvimentos deste trabalho a escolha de diferentes opções metodológicas para a construção da função *cross-basis*, como o uso de funções mais simples, por exemplo, e a inclusão de alguns fatores de risco que alteram o efeito adverso que a temperatura provoca na mortalidade.

Palavras-chave: Modelos de desfasamento distribuído; ÍCARO; Calor; Lisboa; Mortalidade

Abstract

Temperature is an environmental factor that influences human comfort and health, so much that both extreme heat and cold increase mortality. Studying the effect that extreme heat has on mortality is of utmost importance, in order to try to predict and mitigate the consequences of global warming, with special attention to the most vulnerable and least adaptive population.

In 1991, a heat health warning system that monitors possible increases in mortality due to extreme heat was created - the ICARO system. It was initially developed based on a time series statistical model using dynamic regression techniques and a dynamic threshold, which were calibrated for Lisbon data concerning the 1981 and 1991 heatwaves.

The purpose of this work is to formulate a new kind of model to study the heat-mortality relation, aiming to optimise/update the ICARO system. Since the effect of extreme heat on mortality is not limited to the time period when it occurs but is delayed in time, using a model from the family of distributed lag non-linear models (DLNM) seems to be appropriate. A DLNM is based on a bidimensional function, called "cross-basis" function, which describes the shape of the relationship simultaneously along the space of the predictor - temperature -, and along its lag dimension. Therefore, this type of functions allows to explain an exposure-response effect, considering both the intensity and timing of a combination of several past exposures, up to a determined maximum lag.

The model proposed here, was calibrated with data from the district of Lisbon from 1980 to 2017, restricted to the months between May and September. The total counts of daily deaths were explained in function of the daily maximum temperatures, through a cross-basis function allowing for a maximum lag of 10 days. The model also accounted for two time functions to control for seasonality and trend. The day of the week and annual average population entered the final model, as well.

The results revealed that heat has a sustained effect up to 4 days, causing an overall increase in the relative risk of death for temperatures above 30 ^{o}C. However, temperatures below 15 ^{o}C during summer confer some protection.

The predictive performance of the DLNM and the ICARO model were assessed and compared, through a cross-validation method. It revealed that both models have a good capacity to predict the highest peaks of mortality, but the DLNM tends to underestimate the magnitude of the lower ones. Overall, the DLNM obtained is considered a good model, since it seems to capture at least the main features of the studied relationship.

There are some possible future developments for this theme, such as simpler modelling choices for the cross-basis function and accounting for some of the known risk factors for the heat-related mortality.

Keywords: Distributed lag models; ICARO; Heat; Lisbon; Mortality

Acronyms

AE Absolute Error.

bs Natural quadratic B-spline.

df Degrees of freedom.

DLM Distributed lag model.

DLNM Distributed lag non-linear model.

DOS Day of the season.

DOW Day of the week.

Exc Excessive heat.

FunLag Basis-function for the lag dimension.

FunVar Basis-function for the predictor dimension.

GATO Generalised accumulated thermal overload.

GHLen Generalised heatwave length.

ICARO Portuguese heat surveillance and warning system.

II Icaro Index.

INE Statistics Portugal.

INSA Portuguese national health institute.

IPMA Portuguese meteorology institute.

MAE Mean Absolute Error.

MSE Mean Squared Error.

nk Number of knots.

ns Natural cubic B-spline.

Acronyms

QAIC Modified Akaike information criteria.

QBIC Modified Bayesian information criteria.

RMSE Root Mean Squared Error.

RR Relative Risk.

SD Standard Deviation.

SE Squared Error.

Index

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

1.1.1 The impact of weather on health

It is well known that temperature influences health and human comfort and it is currently accepted that mortality rates describe seasonal patterns, being higher in the winter and lower in the summer, with the exception of some rare peaks during summer that may be explained by the exposure to excessive heat [1, 2, 3]. So, one may say, temperature and death correlations are positive in summer and negative in winter [4], i.e., during summer, mortality increases because of very high temperatures, while during winter, it increases because of very low temperatures.

Also, the cold is known to provoke more adverse health effects in warmer countries and, contrariwise, heat is more dangerous in colder countries [5]. Thus it is perceived, there is no global optimal temperature for human comfort. Each population is adapted to its usual local climate [3], so the temperature associated to better health and less mortality in each region is a determined percentile of temperatures felt in this location [6]. For this reason, this type of impacts must be studied locally, since temperatures that may have an adverse effect in one region, may have no effect at all in other regions, where people are adapted to those particular levels of cold or heat. More than that, the same population may adapt seasonally to different weather, becoming more tolerant to heat along summer and the inverse for winter [7].

Both extreme cold and extreme heat cause several adverse effects on human health:

On the one hand, cold has been associated with deaths in different countries, and some studies show that its impact is greater in poorer countries and countries with warmer winters [8], like Portugal which is one of the European countries with the highest excess winter mortality rate [1]. The exposition to low temperatures affects the human cardiovascular system, since it provokes vasoconstriction and increases blood pressure and cholesterol, which may cause heart attacks and strokes. At a respiratory level, cold may cause nose congestion, rhinorrhea and bronchoconstriction and exacerbate preexisting conditions of people who suffer from asthma or chronic obstructive pulmonary disease, leading to death in the worst case scenarios [5].

On the other hand, heat is also responsible for a large number of hospitalisations every year [9], once it can cause several malaise, like severe dehydration, exhaustion, cramps, oedema, heatstroke, acute cerebrovascular accidents ant it is responsible for the aggravation of preexisting illnesses, as pulmonary chronic diseases, cardiac conditions and kidney disorders [10]. Unfortunately, many of such problems may evolve to death, a risk which is higher for people who have some of the identified risk factors: cog-

nitive limitations, isolation, low socioeconomic status, advanced age and living in urban areas, among many other [7, 9]. Although, the older individuals suffer more with heat adverse effects, there are registries of severe heatwaves with some impacts on all group ages mortalities [7].

This work was intended to focus on the effects of heat on mortality rates, even though mortality is higher during the cold months; because its correlation with heat is stronger [4] and also because the global average temperatures are increasing due to climate changes. Hence, it is important to study phenomena like heatwaves with impact on health, even though it has no consensual definition on the literature [3, 9]. From a strictly meteorological point of view, the World Meteorological Organization has defined a heatwave as a period of at least six consecutive days with maximum temperatures 5 ^{o}C above the average daily temperatures for the respective time period [11]. However, this definition is not directly related to the heat impacts on human health. From a public health point of view, a heatwave is usually described as several consecutive days above a determined air temperature threshold, but the number of days to be considered is not clear and the temperature threshold changes for different regions and populations [7]. Therefore, in this work the term "heat wave" will be used to refer to a period of excess heat that impacts human mortality.

Nevertheless, extreme heat is associated with an increase in mortality [3, 7] and morbidity, which impacts health care systems, being considered a public health problem for those reasons [7]. It is important to note that heat related deaths could be decreased with a suitable warning and response system to heat emergencies [9].

In Portugal, several heatwaves and their impacts in mortality have been registered since the final decades of last century [7, 12, 13]:

- June 1981 - excess mortality estimated in 1900 deaths
- July 1991 - excess mortality estimated in 1000 deaths
- July/August 2003 - excess mortality estimated in 2000 deaths
- July 2006 - excess mortality estimated in 1100 deaths
- June/July 2013 - excess mortality estimated in 1700 deaths

Due to climate change, both minimum and maximum temperatures are increasing [1]. Besides, most scientists expect extreme weather phenomena, like heat and coldwaves, with sudden drops or increases in temperature, to happen more intense and frequently in the next few decades [1, 9, 10, 14]. These facts, together with ageing and ever more urbanised Portuguese population (two risk factors for heat vulnerability [7, 9]), suggest strongly that summer mortality will increase in the coming years, eventually becoming an urgent public health problem. As such, temperature-mortality relation must be deeply studied, in order to predict the consequences of global warming, particularly for those most vulnerable and least able to adapt [9].

1.1.2 The ICARO surveillance system

In Portugal, the heatwaves of 1981 and 1991 were responsible for a very high number of deaths in a short time, as described previously. As those deaths were clearly related to an excessive heat, it is safe to say that at least part of them would have been avoidable [12], showing that extreme heat is an important health problem in the country [15]. Hereupon, there was a need to create a heat surveillance and warning

system, in order to mitigate this effect, which led to the ICARO surveillance system being implemented in 1999 [7, 12, 15].

ICARO is an acronym for *Importância do CAlor: Repercussão sobre os Óbitos*, which means *Importance of Heat and its Repercussions in Deaths* [15]. It is a heat health warning system that monitors possible increases in mortality due to extreme heat and is set in motion every year between May and September [7, 12]. It was constructed by the portuguese national health institute (INSA) - that developed statistical models in order to understand how weather variables relate to heat associated mortality -, in a partnership with the portuguese meteorology institute (IPMA) that provides registries and three-day predictions of temperatures [12].

INSA initially developed a time series statistical model using dynamic regression techniques, which were calibrated for Lisbon data concerning the 1981 and 1991 heatwaves [12, 15] and it was later updated to account for the prolonged heatwave of 2003 with the implementation of a dynamic threshold [7].

In this work, the model that will be considered as the ICARO model is the model for the district of Lisbon, proposed by Nogueira & Paixão [7], which will be refered to as ICARO Lisbon. It takes as independent variable the generalised accumulated thermal overload (GATO) - a combination of two weather direct functions [7]:

1. GHLen - a generalised function of the heatwave length, that counts the number of consecutive days in which the maximum temperature went above the considered threshold. This function was constructed in such way that keeps considering the effect of several days with excessive heat even if the temperature lowers under the defined threshold for a few days in between [7].

2. Exc - corresponds to the observed excess temperature above the considered threshold [2, 7].

The threshold used - τ - is a dynamic threshold that depends on the week, beginning as $\tau = 29$ ^{o}C in May and then increasing 1 ^{o}C every week since week 22 (end of May/ beginning of June) until week 28 (second week of July), remaining $\tau = 35$ ^{o}C until the end of September [7]:

$$\tau_{w(t)} = \begin{cases} 29 + (w(t) - 22) & , \quad \text{if week } 22 \leq w(t) < \text{week } 28 \\ 35 & , \quad \text{if} \quad w(t) \geq \text{week } 28 \end{cases} , \tag{1.1}$$

with $w(t) =$ isoweek of the day t.

Then, $\tau_{w(t)}$ enters the two weather direct functions, considering the daily maximum temperatures ($MaxT_t$):

Number of consecutive days above the threshold $\tau_{w(t)}$:

$$GHLen_t = \begin{cases} GHLen_{t-1} + 1 & , \quad \text{if} \quad MaxT_t \geq \tau_{w(t)} \\ GHLen_{t-1} - 1 & , \quad \text{if} \quad MaxT_t < \tau_{w(t)} \ \ \& \ \ GHLen_{t-1} > 0 \\ 0 & , \quad \text{if} \quad MaxT_t < \tau_{w(t)} \ \ \& \ \ GHLen_{t-1} = 0 \end{cases} \tag{1.2}$$

Excess temperature above the threshold $\tau_{w(t)}$:

$$Exc_t = \begin{cases} MaxT_t - \tau_{w(t)} & , \quad \text{if} \quad MaxT_t \geq \tau_{w(t)} \\ 0 & , \quad \text{if} \quad MaxT_t \leq \tau_{w(t)} \end{cases} \tag{1.3}$$

Neither $GHLen_t$ and Exc_t, if considered on their own, are highly related with mortality, but their combination into the main variable $GATO_t$, that enters the final model, is proportionally related with

daily mortality [2, 7]:

$$GATO_t = GHLen_t \times Exc_t \tag{1.4}$$

GATO enters the model so to predict the expected daily mortality taking into account the high temperatures effect [2, 7, 12]. The generic model tested for the ICARO Lisbon is a regression linear model following:

$$Y_t = C + \alpha \times Year + \beta \times GATO_{t-1} + \gamma \times GATO_t + \delta \times Exc_t + \varepsilon_t \quad , \tag{1.5}$$

where

- Y_t is the number of deaths in day t;
- C is the regression constant;
- *Year* is a linear term of the year;
- α, β, γ and δ are regression parameters;
- ε_t is a random process.

This model is based on the observed and predicted temperatures for 3 days (given by IPMA) [7, 15] and showed to be well adjusted to the data, to easily detect the extreme heat impacts on mortality. However, it overestimates some of the mortality increases [2, 12] and, on the other hand, it seems to underestimate the later observed impacts of a long heatwave [7].

Even so, the estimate daily mortality with heat effect is then used to calculate the ICARO Index (II) [12]:

$$II = \frac{\text{Number of expected deaths with heat effect}}{\text{Number of expected deaths without heat effect}} - 1 \tag{1.6}$$

The *II* gives the excess relative risk of deaths due to heat effect and will be zero if there is no expected effect [12].

Nowadays, there is in place a system with mainland Portugal coverage, which uses a generic model similar to the one described. However it considers predefined thresholds, instead of dynamic ones, like the Lisbon model does. Everyday between May and September, a daily report - the ICARO Index Report - is built and sent to the Portuguese National Health Directorate and the Portuguese National Service of Fireman and Civil Protection, among others [12]. The report presents:

- the 48 most recent predictions of the *II* (which includes the adjusted value for the previous day and the values for the current day and the 2 following days), with information about the statistical significance of the predictions;
- the highest *II* registered, as a comparison term;
- the predicted maximum and minimum temperatures for the 2 following days in the 18 districts of mainland Portugal;
- several technical notes.

1.2 Objectives

As mentioned before, the model used on ICARO Lisbon overestimates the mortality increases caused by heatwaves [2, 12] and tends to underestimate their later impact [7]. For that reason, we propose to formulate a new statistical model with equal sensitivity but better specificity, i.e., a model that keeps ICARO's good capacity to detect heatwaves with expected impact on mortality, but that also accounts for later impacts and prevent false alarms, which may discredit future warnings. To formulate such a new model we have tried to take into account the different features of the heat-mortality relationship, described in the literature:

- **Non-linear relation -** various studies investigated the relation between temperature and mortality and showed that it is U-, V- or J-shaped, with mortality increasing for extremely low or high temperatures [7, 14, 16, 17, 18, 19, 20]. These shapes indicate a non-linear heat-mortality relation, as already described in the literature [4, 7, 16, 17].

- **No immediate effect -** it has also been reported that the effect of extreme heat is sustained in time [3, 14, 16, 21], so the rise in mortality does not just occur immediately when the temperature increases, but has been described to last 1 to 5 days after the excessive heat day [7, 9, 14, 22].

- **Harvesting effect -** there is evidence that heat causes the death of people who are in poor health conditions and who would probably die otherwise a few days or weeks later, even without the high temperatures [9, 16]. In other words, heat may just anticipate unavoidable deaths, a phenomenon called harvesting, which increases the complexity of modelling the heat-death effect [21].

- **Isolated and sustained heat effect -** there are commonly two different ways for approaching the effects of heat on health: one is to take warm days as single episodes, considering isolated daily temperatures; the other is to consider temperature as a continuous risk factor, taking into account series of daily temperatures [3]. Both can be combined, describing the mortality risk to increase due to the independent effect of isolated days with high temperatures and also with the duration of consecutive days with high temperatures - sustained effect [3, 16].

That being said, we propose to consider a model capable of combining non-linear exposure-response functions (in this case, temperature-mortality) with non-linear distribution of the time - lags - effect in order to account for the different weight of each past days' temperatures effect in the mortality of the present day. For that reason, the main goal of this work is to develop a distributed lag non-linear model (DLNM) to monitor the mortality increases caused by heat. This model will include a time dimension in order to better predict the delayed effect of heat in mortality, as well as its cumulative effect. It will be used to study the extreme heat effects on Lisbon's mortality, aiming to optimise/update the model used in the ICARO system.

1.3 Work Structure

This Masters dissertation consists of 4 chapters: Introduction, Materials and Methods, Results and Discussion.

The first and present chapter presents the background of the research's theme. The adverse effects of heat on human health and mortality are explained as well as how the ICARO surveillance and alert system works. Lastly, the goals of the research are enunciated.

Chapter two presents all the data accessed and its transformations. Besides, the statistical methods that will be used to model the daily mortality in the district of Lisbon as a function of daily temperatures will be referenced and it will be explained what is a distributed lag non-linear model. Moreover, the method used to compare the new model obtained (DLNM) with the model currently used by the ICARO system - cross validation - will be presented.

The third chapter presents the results of the exploratory data analysis of the variables that will enter the proposed model. The chapter also shows the results of the adjustment of the new model (DLNM) and the results of the assessment and comparison of DLNM against the ICARO model, as well as the analysis of these results.

Finally, chapter four consists of the discussion and conclusions of this work, considering its limitations, as well as possible future developments of this theme.

Chapter 2

Materials and Methods

2.1 Data Collection

In this work, it was decided to study the heat-mortality relation for the district of Lisbon.

For this purpose, we accessed Lisbon's daily total death toll numbers between 1980 and 2017 and stratified these deaths (by all causes) into two age groups: population under and 65 years of age and above. All this data was provided by the Statistics Portugal (INE).

The weather data was provided by the Portuguese Institute for Sea and Atmosphere (IPMA) and consists of the Gago Coutinho meteorological station daily records of minimum and maximum air temperature in Celsius degrees for the period between 1980 and 2018. Yet, it was decided to use only the data until year 2017 since the counts of deaths for 2018 were not available.

INE also provided the average annual resident total population in Lisbon's district for the years between 1980 and 2017 and the data concerning the district's over 65 years old population between 1992 and 2017. The total resident population will be used to control for possible confounding.

2.1.1 New Variables

The collected weather data was used to create some more variables that, according to the literature, might relate with mortality. Namely, the daily thermal amplitude, calculated as the difference between the maximum and minimum air temperatures for each day, and an approximation to the daily mean temperature, calculated as the average of the minimum and maximum air temperatures [14].

Also, the average resident population under 65 years of age for the years between 1992 and 2017 was calculated as the difference between the total population number and the number of the population aged 65 and above.

This way, there are a total of 10 variables (all referring to the district of Lisbon) that can be used in the statistical analysis:

- Daily counts of total number of deaths (TMort)
- Daily counts of number of deaths of people with 65 years old or more ($\geq$ 65Mort)
- Daily counts of number of deaths of people under 65 years old ($<$ 65Mort)
- Daily maximum temperature (MaxT)
- Daily minimum temperature (MinT)

2. MATERIALS AND METHODS

- Daily mean temperature (MeanT)
- Daily thermal amplitude (TA)
- Annual average of the total resident population (TPop)
- Annual average of the 65 years old or more resident population ($\geq$ 65Pop)
- Annual average of the under 65 years old resident population ($<$ 65Pop)

From these ten variables, the data that will enter the modelling is: the daily counts of total number of deaths, the daily maximum temperatures and the annual average of the total resident population. The rest of the collected data will only be taken into account in the exploratory analysis, in order to provide a broader picture of the whole Lisbon's mortality-temperature-population scenario.

2.2 Statistical Methods

In this work all the statistical analysis will be undertaken in several different R 3.5.2 software packages. R is a free software environment for statistical computing and graphics (`http://www.r-project.org`) [23]. In particular, the *dlnm* package, developed by Gasparrini (2015) [24], will be used, as it offers some functions to run DLNMs in a simple way.

2.2.1 Exploratory data analysis

In order to become more familiar with the data set, statistical analysis should always start with an exploratory data analysis. Exploratory data analysis consists of collecting basic descriptive statistics of the variables and visualising the data trough simple exploratory graphs.

The descriptive statistics will be obtained using the R functions *summary()* and *sd()* from the package *stats*. The first one gives measures of location and number of missing values and the second one gives a measure of dispersion or variability - the standard deviation.

The initial exploratory analysis will also employ time series graphs, i.e., graphic representations of repeated measurements taken over regular time intervals (daily counts of deaths and temperatures, and yearly population densities, in what concerns this thesis), as well as boxplots. Both aim to facilitate the perception of the numeric variability, trend and seasonality of the different variables. Moreover, correlation graphs of daily mortalities *vs.* daily temperatures will be presented.

2.2.2 Distributed lag non-linear models

Oftentimes, the effect of exposure to environmental stressors or other events is not limited to the time period when it occurs, but is delayed in time [21]. That is the case of climatic variables effects (as adverse temperatures) that may last in time, spreading over several days [8, 17, 21, 25], as mentioned in the introduction chapter. That being said, the impact of the environmental stressors at a given time may be explained by a combination of past exposures over several time lags [21], once it depends simultaneously on the intensity and timing of the exposures [26]. A common way to model a non-linear effect with this additional time dimension is through distributed lag non-linear models (DLNMs) [1, 6, 9, 14, 17, 27].

The family of distributed lag non-linear models was developed to simultaneously estimate the non-linear dose-response dependencies and the delayed effects of temperature on mortality [21]. It is based on a bidimensional space of functions, called “cross-basis”, that describes the shape of the relationship simultaneously along the space of the predictor - temperature, in this case - and along its lag dimension, i.e., the time structure of the exposure–response relationship [21]. This way, each exposure up to a maximum lag may be summed to obtain the cumulative effect of temperature [17].

Initially, distributed lag models (DLMs) were developed for time series analysis and extensively used in econometric and social sciences before being adapted to epidemiology research [28, 29]. This family of models used to account for linear dependencies only, so Armstrong [16] extended this methodology to distributed lag non-linear models (DLNMs), and it has since been used to simultaneously estimate the non-linear and delayed effects of temperature and air pollution on mortality or morbidity [1, 16, 17, 21, 26, 29]. Hence, to understand a DLNM well one must first understand DLMs.

A DLM is a dynamic model that estimates the effect of a regressor x on a response y over different

time moments t. It can generally be represented as follows:

$$\begin{aligned} y_t &= \alpha + \beta_0 x_t + \beta_1 x_{t-1} + ... + \beta_l x_{t-l} + ... + \beta_L x_{t-L} + u_t \\ &= \alpha + \sum_{l=0}^{L} \beta_l x_{t-l} + u_t \quad , \end{aligned} \tag{2.1}$$

where α is the intercept, u_t is a stationary error term and L is the maximum lag allowed (L $\geq$ 1) [30].

Each of the coefficients β_l stands for the weight of the respective lag l ($l = 0, 1, ..., L$). β_l coefficients may be interpreted either from a backward standpoint - the effect of the past exposition x_{t-l} on the present moment response y_t, i.e., the effect felt today due to the exposition l days before, $lag = l$ -, or from a forward standpoint - the effect of the current exposition x_t on the future response l moments later, y_{t+l}, i.e., the effect that today's exposure will cause l days from now [26, 31]. Thus, the coefficients β_l represent the lag weights and all together define the lag distribution [30].

Assuming there is a temporary change on the regressor variable x, which increases one unit only in moment t (x_t), then its immediate effect, y_t, will have an increase equal to the value of β_0. On the next moment (t+1) the effect y_{t+1} will increase β_1 units. After that the effect y_{t+2} will increase β_2 units and so on, until the maximum lag, L, when the effect y_{t+L} increases β_L units. This is called the marginal effect of x on y.

Another hypothesis to consider would be a permanent change in the regressor variable. Assuming it increases one unit in moment t and remains that high in all future moments, then its immediate effect y_t will also have an increase equal to the value of β_0. However, in the future moment (t+1) the effect y_{t+1} will increase $\beta_0 + \beta_1$ values, after which the effect y_{t+2} will increase $\beta_0 + \beta_1 + \beta_2$ values and so on, until the maximum lag L, when the effect y_{t+L} increases $\beta_0 + \beta_1 + \beta_2 + ... + \beta_L$ values. This is called the cumulative effect of x on y [30].

Variable x may change more or less than one unit and increase or decrease. These changes may occur at any moment, be temporary or permanent. Therefore, from (2.1), we understand that the change on y_t depends on the sum of the products of each past moment value of x_{t-l} and the weight of its moment β_l, with $0 \leq l \leq L$. As such, the combination of all lag weights (the cumulative and marginal effects) defines the pattern of how the magnitude and timing of x influences y over time.

In this modelling approach that includes one parameter β_l to each lag, each individual coefficient is highly correlated to the others, which raises a problem of collinearity. That problem, however, has been overcome with the imposition of some restrictions to β_l, originating a constrained distributed lag model [16, 21]. According to Armstrong (2006) [16], the constraints may consist in compelling the coefficients β_l to be equal for a determined range of lags - which is called a lag-stratified distributed lag model - or in forcing β_l to follow some smooth curves, as a polynomial, a natural cubic spline or a penalised spline function, among other options.

For now, we will focus only on the relation of the regressor x and the response y, defined as $s(x,t)$.

According to Gasparrini (2014) [26] and assuming there is a linear exposure–response relationship, a general notation to describe the dependency in terms of exposure history to x evaluated at time t is:

$$\begin{aligned} s(x,t) &= \int_{l_0}^{L} x_{t-l}\, w(l) \quad dl \\ &\approx \sum_{l=l_0}^{L} x_{t-l}\, w(l) \quad , \end{aligned} \tag{2.2}$$

where $w(l)$ represents the weighting basis-function applied to constrain the coefficients β_l. $w(l)$ is directly defined in the lag dimension and determines the *lag-response function* that models the lag–response curve associated with exposure x, [26].

$s(x,t)$ will then be included in a generalised linear model as a sum of linear terms, with related parameters η.

The function $s(x,t)$, defined in (2.2), may be rewritten following a matrix notation, by applying the basis transformation over the lags - $w(l)$ - and then combining it with the vector of expositions $\mathbf{q}_t$ [21, 26]:

$$s(x_t;\eta) = \mathbf{q}_t \mathbf{C}\eta = \mathbf{W}_t^T \eta \quad , \tag{2.3}$$

where

- $\mathbf{q}_t = [x_t, ..., x_{t-l}, ..., x_{t-L}]^T$ is an original vector of ordered exposure histories, corresponding to a column of the $n \times (L+1)$ matrix $\mathbf{Q}$ [21]. As such $\mathbf{q}_t$ changes along time, depending on the moment t when it is defined [26];
- $\mathbf{l} = [0, ..., l, ..., L]^T$ is a vector of lags corresponding to the L+1 columns of $\mathbf{Q}$ [21];
- $\mathbf{C}$ is a $(L+1) \times \upsilon_l$ matrix of lag-basis variables originated from the application of the basis-function to the lag vector $\mathbf{l}$ (where the basis-function is defined as $w(l)$ with dimension υ_l);
- η is a vector of unknown parameters.

Hereupon, $\mathbf{W}_t^T$ is a vector from the matrix $\mathbf{W} = \mathbf{QC}$, which is the matrix of the υ_l transformed variables (obtained from the application of a basis lag function to the original lag vectors $\mathbf{l}$ - $w(l)$ - combined with the original exposure histories, $\mathbf{q}_t$). This υ_l basis variables from the matrix $\mathbf{W}$ will be included in the design matrix to allow the estimation of the unknown parameters η. The estimated parameters $\hat{\eta}$ define the previously mentioned coefficients β_l [21]:

$$\hat{\beta} = \mathbf{C}\hat{\eta} \tag{2.4}$$

The extension from DLM to distributed lag non-linear models (DLNM) was achieved by adding an exposure–response non-linear function along the dimension of the predictor x [26]:

$$s(x_t;\beta) = \mathbf{Z}_{t.}^T \beta \tag{2.5}$$

On this function, $\mathbf{Z}_{t.}^T$ is the t^{th} line of the matrix $\mathbf{Z}$ - a $n \times \upsilon_x$ basis matrix resulting from the application of the basis-function - called $f(x)$, of dimension υ_x -, to the original vector of exposures x [21].

Therefore, a generalisation of (2.2) to DLNM will be adding a basis-function along the dimension of the predictor x to the already mentioned basis-function along the dimension of the lag l:

$$\begin{aligned} s(x,t) &= \int_{l_0}^{L} f(x_{t-l})\, w(l) \quad dl \\ &\approx \sum_{l=l_0}^{L} f(x_{t-l})\, w(l) \quad , \end{aligned} \tag{2.6}$$

where $f(x)$ is the *exposure-response function* with dimension υ_x and $w(l)$ is the previously mentioned *lag-response function*. These are the two basis-functions, which may be chosen independently of each

other [16, 26]. The basis-functions impose a set of completely known transformations of x, generating new variables, called *basis variables* [21].

The representation in (2.6) assumes that the functions $f(x)$ and $w(l)$ are independent, i.e., that the *exposure-response function* is the same along all the lag space and that the lag structure is equal for all values of x [26]. If we relax this assumption, admitting an interaction between the value of the predictor and its timing, then (2.6) may be more flexibly represented as:

$$\begin{aligned} s(x,t) &= \int_{l_0}^{L} f \cdot w(x_{t-l},\, l) \quad dl \\ &= \sum_{l=l_0}^{L} f \cdot w(x_{t-l},\, l) \end{aligned} \tag{2.7}$$

Following this notation, $f \cdot w(x,\, l)$ is a bivariate function, that models simultaneously the exposure-response structure along x and the the lag-response structure along l, defining the *exposure–lag–response function* [26]. In other words, $s(x,t)$ is a linear combination of the basis-functions $f(x)$ and $w(l)$, integrated over the lag dimension, which defines a bidimensional space of functions that Armstrong (2006) [16] called *cross-basis function* and which represent the core of DLNMs [16, 26, 32].

Each basis-function ($f(x) = \sum_{b=1}^{B} \beta_b x_b$ and $w(l) = \sum_{p=1}^{P} \beta_p l_p$) is fitted following the chosen function distribution, $f()$ and $w()$ respectively, originating B/P basis variables from the original ones. Then the new transformed variables will enter the regression model, instead of the original variables. The number of basis variables, B and P, determines the degrees of freedom (df) of the respective curves [16]. B df for the dimension of the predictor and P df for the dimension of the lags. So the degrees of freedom of the cross basis-function is given by the product of the number of basis variables from $f(x)$ and from $w(l)$, $B \times P$.

The cross-basis function is better represented using matrix notation as Gasparrini *et al.* [21].

Let $\dot{R}$ be a $n \times \upsilon_x \times (L+1)$ array of the lagged occurrences of each of the basis variables of x, keeping $\mathbf{C}$ as the matrix of basis variables for the lag dimension,

$$s(x_t;\eta) = \sum_{j=1}^{\upsilon_x} \sum_{k=1}^{\upsilon_l} \mathbf{r}_{tj}^T \mathbf{c}_{.k} \eta_{jk} = \mathbf{w}_t^T \eta \quad , \tag{2.8}$$

where $\mathbf{r}_{tj}$ is the vector of lagged exposures for the time t transformed through the basis-function j and $\mathbf{c}_{.k}$ is the vector of lags transformed trough the basis-function k.

Now, $\mathbf{w}_t^T$ will be defined as the vector obtained by applying the $\upsilon_x \cdot \upsilon_l$ cross-basis functions to x_t. As such, both basis-functions are then simultaneously used to create the $\upsilon_x \times \upsilon_l$ basis variables, stored in the $\mathbf{W}$ matrix.

According to Gasparrini *et al.* [21], in order to reach a compact formula for $\mathbf{W}$, there is a need to represent it as a tensor product, $\dot{A}$, defined as a $n \times (\upsilon_x \cdot \upsilon_l) \times (L+1)$ array, that will be rearranged summing along the third dimension of lags, obtaining $\mathbf{W}$, the final matrix of the cross-basis variables values and η, the associated parameters, expressed in (2.8) [26].

In conclusion, the cross-basis flexibly describes the relation along x, allowing for linear and non-linear exposure-responses, combining it with the distributed lag-effects (an additional time-dimension) [26]. So, the $\upsilon_x \times \upsilon_l$ basis variables originated from the cross-basis function will enter the regression model instead of the original variables.

We may now rewrite (2.1) and generally represent a basic DLNM, using a cross-basis function, as

defined in (2.8), to express the non-linear relation between the predictor variable x and the response variable y along time. Following the notation of Gasparrini *et al.* [21], we obtain:

$$g(\mu_t) = \alpha + \sum_{j=1}^{J} s_j(x_{tj}; \eta_j) + \sum_{k=1}^{K} \gamma_k u_{tk} \quad , \tag{2.9}$$

where $\mu_t \equiv E(Y_t)$ and g is a monotonic link function; α is the intercept; s_j is the cross-basis function that denotes smoothed relationships between the variables x_j and the linear predictor - defined by the parameter vectors η_j -, and u_k are other predictors variables with linear effects over Y, measured by the coefficients γ_k.

This type of functions (DLNM) should be more adequate for this kind of modelling than simpler ones, because they reduce the risk of confounding by a residual main effect of heat on consecutive days [3].

In this family of models, Y is assumed to follow a distribution from the exponential family [21]. The two basis-functions used in the cross-basis s_j may be chosen from a wide variety of modelling options [16, 21, 26]. Armstrong (2006) [16] has reviewed the options suitable to model the relationship between ambient temperature and daily mortality. Those options are:

- for the shape of the temperature–mortality curve at a given lag:
  - a polynomial function, whose order must be determined;
  - a stratified model, with chosen strata intervals;
  - a spline function, for which the number and placing of knots must be chosen;
  - linear thresholds;
- for the change in coefficients by lag:
  - a polynomial function, whose order must be determined;
  - a stratified model, with chosen strata intervals;
  - a spline function, for which the number and placing of knots must be chosen;
  - the coefficients may be unconstrained;

The options selected will determine the flexibility of the model. Simpler models (as linear thresholds) are usually less flexible but easier to interpret than the more complex ones (as natural cubic splines functions), while the latter may better adjust to the data and capture most of the relationship details and are less likely to leave residual confounding [16].

Other predictor variables may be included in the model. For instance, as happens in order to control for confounding, a smooth function of time to capture long-time trends and/or seasonality and some categorical variables, as day of the week [21], are applied.

Also, the analysts must determine the maximum lag, L, which will depend on how long they believe an effect may be sustained in time. For instance, the maximum lag allowed should be higher if harvesting effects are expected, which may reflect on negative coefficients for longer lags [16],.

All options have advantages and disadvantages, so the choice will depend on the purpose of the analysis, on *a priori* assumptions and/or the fitting of the model. The model fitting criteria mostly used in this context are the modified Akaike's information criteria (QAIC) and the modified Bayesian

information criteria (QBIC) for models with overdispersed responses fitted through quasi-likelihood [21], with the following expressions:

$$QAIC = -2\mathscr{L}(\hat{\theta}) + 2\hat{\phi}k \tag{2.10}$$

and

$$QBIC = -2\mathscr{L}(\hat{\theta}) + log(n)\hat{\phi}k \quad , \tag{2.11}$$

where $\mathscr{L}$ is the log-likelihood of the fitted model with parameters $\hat{\theta}$; $\hat{\phi}$ is the estimated overdispersion parameter; k is the number of parameters and n is the number of observations.

These two information criteria are goodness-of-fit measures that evaluate the quality of the models. They equally assess the sum of squared residuals through the log-likelihood function, but penalise differently large numbers of estimated parameters [30]. As both criteria estimate the amount of information lost by a model, it is aimed to minimise them. The model with the lowest QAIC or QBIC should be the one which lost less information, i.e., should be the higher quality model.

Previous studies found that the QAIC provides better results than the QBIC in this context [16, 26]. Therefore, along with other previous approaches, we will only consider the QAIC in this work [1, 19, 31].

2.2.2.1 Heat-mortality modelling

The purpose of this work is to model the impact of heat on mortality in the district of Lisbon, so the main focus, data wise, are days with excessive heat and heatwaves, that usually occur during or around summer. For that reason, the modelling will be restricted to the months from May to September, as in the current model used by the ICARO system [12]. This way, the time series model becomes simpler, with less auto-correlated errors [7]. From now on, this five months period will be generally referred to as the "summer period".

The data available spans from January 1^{st} of 1980 to December 31^{st} of 2017, so there will be a total of 38 years. Since this work will focus only in the summer period, the data to be used is assumed to come from several time series, each one corresponding to the period between May and September of each year, i.e., 38 groups/sections from the original complete time series.

The impact of heat on mortality will be studied considering a maximum lag of 10 days, as the literature shows that 10 days are enough to capture the delayed and harvesting effects of heat [3, 33]. Thus, to assess the effect that the 10 previous days have on the first day of May, we must account for the last 10 days of April. This way, there will be a total of 6194 observations, each one corresponding to a day between April 21^{st} and September 30^{th} from the years 1980 to 2017.

The model employed will be a DLNM that follows (2.9). The response variable, Y_t, will be daily counts of deaths, which is assumed to arise from an *overdispersed Poisson* distribution with a canonical log-link function: $E(Y) = \mu$ and $V(Y) = \theta\mu$, where θ is the dispersion parameter.

As mentioned in section 2.1, the average annual total resident population in Lisbon is available for the all period between 1980 and 2017. However, resident population data stratified by age is only available from 1992. For that reason, it was decided that the total daily counts of deaths would be used as the response variable in the model (instead of counts of deaths stratified by age) to better control for this possible confounding of resident population during a larger time period.

The predictor variable, X_t, will be maximum daily temperature, once it is intended to study how extreme heat affects mortality rates and to better compare the results with the current ICARO model, which uses maximum temperatures.

The main focus of this type of models (DLNMs) is the cross-basis function, for which we must choose two functions: one to the exposure-response curve and another to the lag-response curve. As previously mentioned, there is a large amount of options for these two functions. Hence, we will construct several models using natural cubic B-splines (ns) and quadratic B-splines (bs) as the basis-functions, which are the type of functions more frequently mentioned in the literature for modelling the effects of extreme temperatures on mortality [3, 6, 14, 17, 19, 21, 27, 32].

It must be noted that the larger number of knots chosen, larger the number of df, the more flexible but less smooth will the shape of the function be [17, 21]. Therefore, the choice of number of knots, which determines the smoothness of the curve, is more important than precise knot placement, with many knots tending to show spurious wiggles and too few knots losing real wiggles [16]. For this reason, we will try different number of knots (3 to 5) for both basis variables, allowing them to be equally spaced with no intercept, following the chosen spline function for the predictor dimension, and equally spaced in the log scale with an intercept in the lag dimension.

The QAIC will be used to chose each basis-function's most appropriate degree of the spline (quadratic or cubic) and the number of knots (3 to 5), which determines the degrees of freedom:

- Natural cubic B-spline: Number of knots = df - 1 - intercept
- Quadratic B-spline: Number of knots = df - 2 - intercept

Regression diagnostics such as residual plots are also presented, as they may be helpful in the selection of different models [21].

To control for seasonality, a natural cubic B-spline of day of the season with 4 internal knots, defining 5 degrees of freedom (one for each month - May to September) will be used. In its term, the long-term trend will be controlled for with a natural cubic B-spline of time with 3 internal knots, defining 4 degrees of freedom (approximately one for each decade). Moreover, the model will include day of the week as a categorical variable and population as a linear variable (entering he model as an offset). All these decisions were based on previous work restricted to one season [1, 33].

The selected model will then be used to predict the relative risk (RR) associated to determined combinations of temperature and lag values *vs.* a reference temperature associated to the minimum mortality. As effects on health are more associated with temperature percentiles than absolute temperature [6], which is due to climate adaptation, we chose as minimum mortality temperature the 75^{th} percentile of the yearly maximum temperatures registered in Lisbon for the period in analysis (1980-2017), which is 25.7 ^{o}C. This methodological option is consistent with previously published work [3, 6].

2.2.3 Models assessment and comparison

Finally, the chosen DLNM will be compared with the model considered for Lisbon in the ICARO system. Given that the statistical model used by ICARO is a linear regression model which does not account for overdispersion, it does not make sense to calculate its QAIC, so this criteria cannot be used for comparison with our DLNM. Hence, another method to evaluate both models and allow a fair comparison between them is needed. As both models will be used to forecast future response values (number of daily deaths), one common method to assess and compare them is to evaluate their predictive performances. The model that better predicts future response variables will be considered the best model. This may be achieved through cross-validation.

Cross-validation is a common technique to estimate how accurately a predictive model will perform on new (unseen) data. It consists in splitting the data set in two complementary groups. One subset is

called *training data* and will be used to perform the analysis, adjusting the estimates parameters of the model. The other subset, called *test data*, will be used to validate the analysis, making predictions for this new data subset and obtaining some measures of fitness to estimate the predictive performance of the model. This process allows to detect problems such as overfitting or bias, once the forecast will be made in data that had not been taken into account during the adjustment of the model. There is a broad range of data partitioning methods and usually multiple rounds of cross-validation are performed with different partitions of the data set in each round, to reduce variability.

Given that we are working with time series, it is important to be particularly careful when choosing the partitioning method. Most partitioning methods are random processes that split the data into two subsets without keeping the order of the observations. However, in time series analysis the observations follow a specific sequence and may not be considered as independent. For that reason, the most common approach of cross-validation for time series analysis is the *last block evaluation* [34]. It consists of continuous forecasting of upcoming values, where the origin of the test data follows the last known value of the train data. This way the dependencies of the data are respected, assuming that the future (that which we want to predict) depends on the past (which we used for modelling) [34]. Once again, there is more than one method for partitioning the data for last block evaluation. Following Bergmeir's (2012) nomenclature [34], this research work will perform two different partitions, the *rolling-origin-update* and the *rolling-window evaluation*.

In order to perform the *rolling-origin-update*, we will start by adjusting the models to data corresponding to the first 10 years of our complete data set (1980 to 1989) - train data - and then evaluate their predictive performance to the subsequent year (1990) - test data. The process will then be repeated, including the previous test data in our training data (which will now be from 1980 to 1990) and readjusting the model for the new training data as well as to test it on the new test data (1991) and so on. This process will occur time and again, until the "final round", when the training data will be from 1980 to 2016 and the test data will correspond to the year of 2017.

As a result, predictions of daily deaths to all summer periods from the year 1990 to 2017 will be obtained by the two models (DLNM and ICARO). The predictions obtained for each year will be based on the models adjusted by all the previous years' data (since 1980). This method simulates real-life application, once it used all past information available to predict each upcoming summer heat-related mortality and then the data from each new year integrates the data set used for adjusting the model, which will again be used for forecasting the new summer period and so on, repeating the whole process every year.

As the training data grows bigger, this proceeding may have the disadvantage of old values disturbing the present model forecast. This possibility is quite strong in this scenario because population changes along the years, and it is likely to be more protected from heat in the present moment, due to better adaptation habits and socioeconomic as well as housing conditions, among other factors that may decrease the adverse effect of heat. Therefore, the relationship between adverse temperature and deaths in 2019 is not the same as it was in 1980. Hereupon, it may be wise to consider a fix length window for the training data.

Following this rational, plots of the overall RR along temperatures estimated by the DLNM adjusted for different time periods (1980-1989; 1990-1999; 2000-2009 and 2010-2017) will be present. Those will show if there were any evolution of the heat-mortality relationship along time. If so, it makes sense to also consider *rolling window evaluation*.

To perform *rolling window evaluations*, the first step is to choose the length of the window, i.e., the dimension of the time period considered to adjust the models. Lets consider a train data length of 10

years. In this case, as in the *rolling-origin-update* approach, we would start by adjusting the models to data corresponding to the first 10 years of our data set (1980 to 1989) - train data - and then we would evaluate their predictive performance to the subsequent year (1990) - test data -. Then we would repeat the process, but now moving the all window forward one year. The test data will enter our training data, which will lose the first year. Now the new train data would become from 1981 to 1990. Then we readjust the model for the new training data and test it on the new test data (1991). And so on, keeping always a train data with 10 years, until the final round, when the training data will be from 2007 to 2016 and the test data will correspond to the year 2017. As a result we will also obtain the two models (DLM and ICARO) predictions of daily deaths to all summer periods from the year 1990 to 2017, but now the predictions obtained for each year will be based on the models adjusted only by the data of the 10 previous years.

This procedure will be repeated using different lengths for the window, to check if the estimated effect of heat on mortality changes with different number of years entering the modelling. We will empirically use 10 and 5 years windows.

As a result of each round from both methods, error measures will be obtained to evaluate the predictive performance of both ICARO and DLNM models. An error is the difference between an observed response value on moment t (y_t) and the value predicted by the model for this same moment ($\hat{y}_t$). So the absolute error (AE) is given by $|y_t - \hat{y}_t|$ and the squared errors (SE) is given by $(y_t - \hat{y}_t)^2$. Then the errors in each round are averaged for an easier interpretation.

The error measures used in this work are the mean of the absolute errors (MAE), the mean of the squared errors (MSE) and the root of the mean squared errors (RMSE), obtained by the following expressions:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{2.12}$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \tag{2.13}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{2.14}$$

The comparison of the two models will be based on these error measures. The model with the lowest error measures is the one whose predictions for the unseen data are closer to the real observed values, which means that it is the model that better generalised to new data. In summary, the lowest error measures, the better is the predictive performance of the model.

This comparison will also allow to choose how many years should be accounted in the estimation of the models parameters to allow a better prediction of the upcoming summer. The cross-validation method with lower error measures, is the one with the best training data choice.

Also, to better understand the different performances of the ICARO and the distributed lag non-linear models (using the best train data length for each model, chosen by the cross-validation), there will be constructed some time series graphs of the observed daily deaths and the respective predicted number of daily deaths by each of the models (ICARO and DLNM) for the years tested with known heatwaves (1991, 2003, 2006, 2013 and 2017). This will allow an easy interpretation of the performing differences between the two models.

Chapter 3

Results

3.1 Exploratory Data Analysis

In this section, the results of the exploratory data analysis described in section 2.2.1 are presented.

3.1.1 Descriptive statistics

The descriptive statistics were obtained for all the variables described before, in section 2.1. Tables 3.1 and 3.2 present the mean; standard deviation (SD); minimum (Min); 25^{th}, 50^{th} and 75^{th} percentiles; maximum (Max) and number of missing values (NA) of the 10 previously described variables. All the values were rounded to one decimal place. Those statistics were obtained for the complete data set (all year) - Table 3.1 - and just for the summer period (May to September) - Table 3.2 -, both for the period between 1980 and 2017.

Table 3.1: Descriptive statistics of year-round data.

Variables	Mean	SD	Min	25^{th}	50^{th}	75^{th}	Max	NA
MaxT	21.3	6.1	6.5	16.2	20.5	25.7	43.0	12
MeanT	17.1	5.1	3.3	13.1	16.7	21.1	33.9	16
MinT	12.9	4.4	-1.0	9.7	13.0	16.4	26.4	11
TA	8.4	3.2	0.3	6.0	8.0	10.4	22.7	16
TMort	56.2	12.2	20.0	48.0	55.0	63.0	172.0	-
≥65Mort	37.8	10.9	9.0	30.0	37.0	44.0	126.0	-
<65Mort	18.4	5.6	3.0	14.0	18.0	22.0	56.0	-
TPop	2144173.0	69552.8	2061150.0	2083090.0	2102360.2	2223396.0	2253911.5	-
≥65Pop	375180.4	62529.5	278941.0	322507.5	365336.0	430008.5	487967.0	4383
<65Pop	1793241.9	24253.4	1763790.0	1768603.5	1784199.5	1818875.5	1828691.5	4383

The maximum temperature achieved in this period was 43 ^{o}C which occurred on June 14 of 1981, the minimum temperature was -1 ^{o}C on January 12 of 1985 and the thermal amplitude varied from 0.3 ^{o}C (7/12/1991) to 22.7 ^{o}C (15/9/1992).

All mortality statistics are higher for the 65 years old or more population than the under 65 years old population; and the highest count of deaths, 172, occurred on June 15 of 1981, one day after the highest temperature registered.

The population statistics are very similar in both tables, because this data corresponds to annual average, so the values are the same for all days of each year.

Table 3.2: Descriptive statistics of only May to September data.

Variables	Mean	SD	Min	25^{th}	50^{th}	75^{th}	Max	NA
MaxT	26.4	4.7	14.4	23.4	26.2	29.4	43.0	-
MeanT	21.5	3.4	10.8	19.4	21.5	23.4	33.9	-
MinT	16.5	2.7	7.0	14.9	16.7	18.1	26.4	-
TA	9.9	3.2	1.1	7.5	9.6	12.2	22.7	-
TMort	50.5	9.5	20.0	44.0	50.0	56.0	172.0	-
$\geq$65Mort	33.1	8.5	9.0	27.0	33.0	38.0	126.0	-
$<$65Mort	17.4	5.3	3.0	14.0	17.0	21.0	56.0	-
TPop	2144175.0	69555.4	2061150.0	2083090.0	2102360.2	2223396.0	2253911.5	-
$\geq$65Pop	375183.0	62531.2	278941.0	322507.5	369014.0	430008.5	487967.0	1836
$<$65Pop	1793242.3	24255.8	1763790.0	1768603.5	1787830.5	1818875.5	1828691.5	1836

The analysis of both tables show that the standard deviations are less in the one restricted to the warmer months, indicating less variability in this case, as expected.

The data restricted to months between May and September have no missing values, except for the stratified by age population variables, for which there are no values corresponding to the first twelve years (1980 to 1991) - Table 3.2 - as mentioned in section 2.1.

3.1.2 Exploratory graphical analysis

Figures 3.1, 3.2 and 3.3 present time series obtained for the total daily counts of deaths, daily maximum temperatures and the annual average resident population between 1980 and 2017, respectively; all for the complete data set (all months of the year).

The analysis of Figure 3.1 shows that mortality in Lisbon has a seasonal pattern, describing a sinusoidal curve, being higher in the winter and lower in the summer. With the exception of some peaks at the middle of some years, as stands out the one corresponding to the day with maximum count of deaths (172) in the summer of 1981, previously mentioned.

Figure 3.2 also shows a seasonal pattern of the daily maximum temperatures, that are much higher in summer and lower in winter, as expected.

The analyses of Figure 3.3 lets us know that, even though the total resident population in Lisbon suffered two small drops (one in the beginning of the 90's and another in the beginning of this decade), it shows a positive trend, with a general significant increase since 1980 until 2017.

The evolution of Lisbon's resident population is better understood in Figure 3.4, where we can see the annual average of total, under 65 and above 65 years old population, since 1992. This graph shows that the population under 65 years old is the vast majority and did not change much since 1992. However it is noticeable a growth in the oldest population, which may explain the increase in the total population (clearly seen in Figure 3.3).

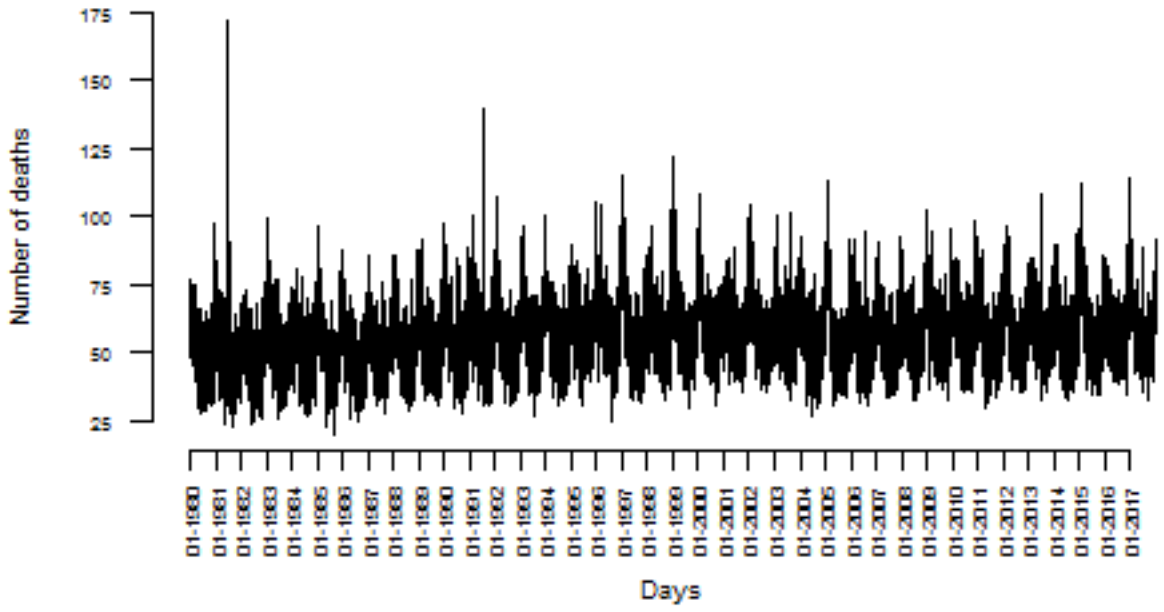

Figure 3.1: Time series of daily mortality.

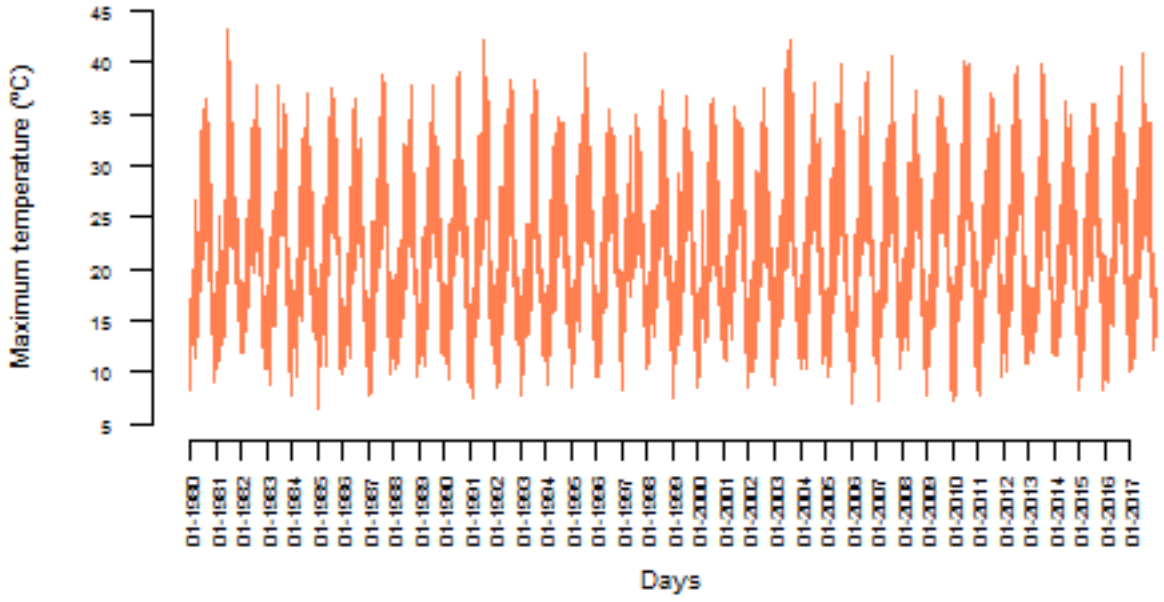

Figure 3.2: Time series of daily maximum temperature.

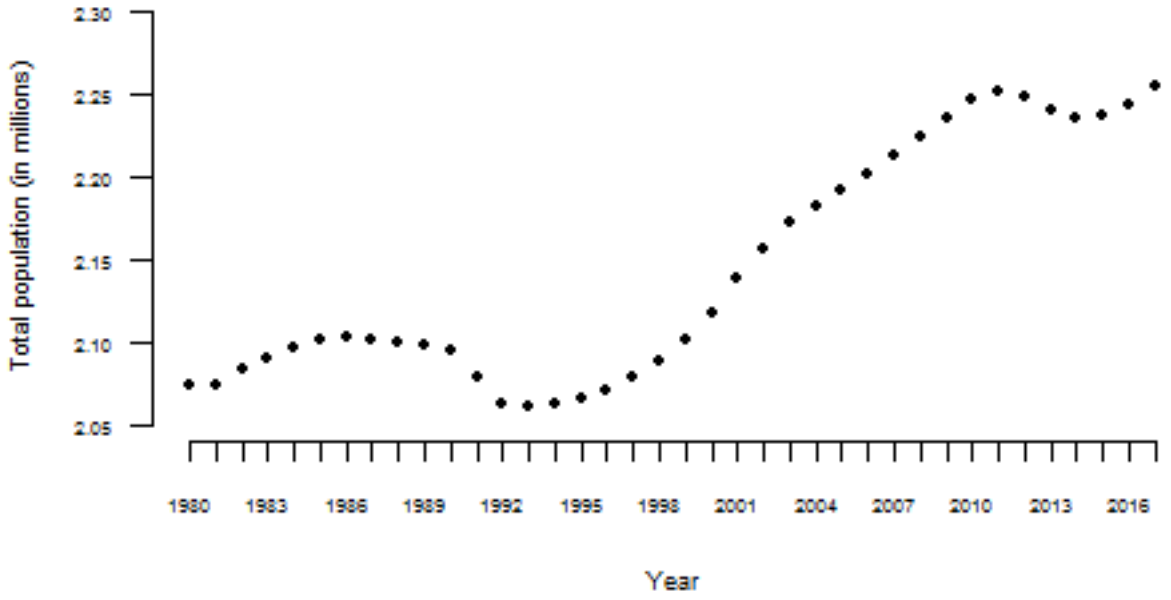

Figure 3.3: Time series of annual average of Lisbon's total population.

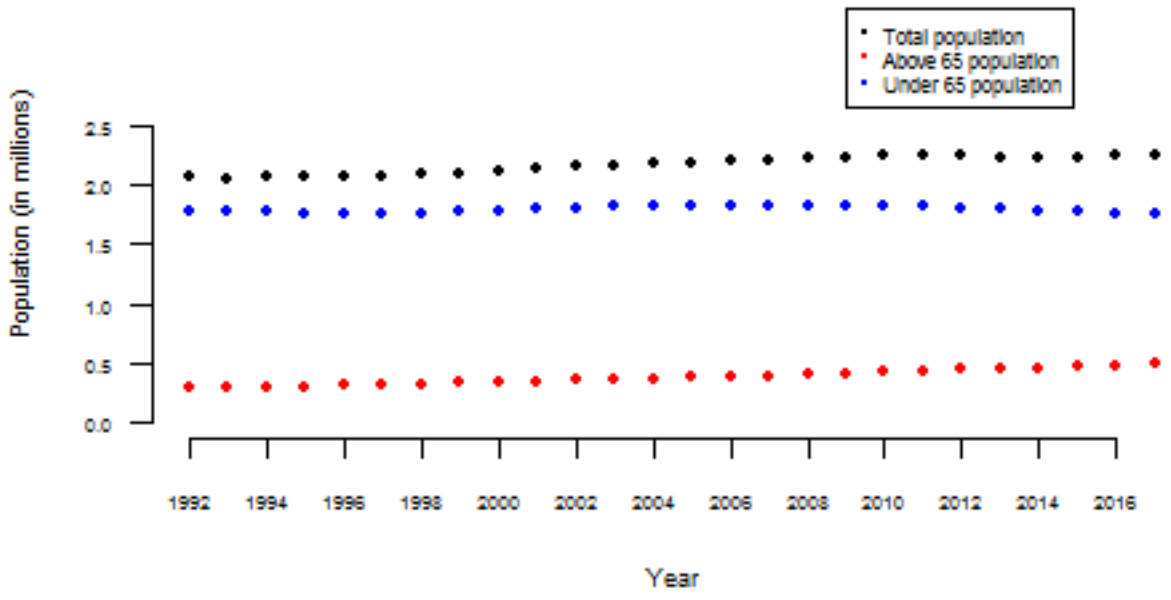

Figure 3.4: Time series of annual average of Lisbon's population (total and stratified by age).

To better understand the trend and seasonality of daily mortality, we present its distribution for each year, month and day of the week, through boxplots.

Figures 3.5, 3.6 and 3.7 show the distribution of daily total counts of deaths in the district of Lisbon, between 1980 and 2017 for each year, month and day of the week, respectively.

Figure 3.5 suggests there was a slight increase in mortality in the first two decades and that then it remained almost constant in the last years, revealing no significant trend or seasonality along the years. Figure 3.6 shows a clear seasonality, with counts of deaths higher in cold months and lower in warmer months (May, June, July, August and September) with the exception of some outliers during summer, as was expected. And finally, Figure 3.7 shows an apparent uniform distribution of mortality over the different days of the week.

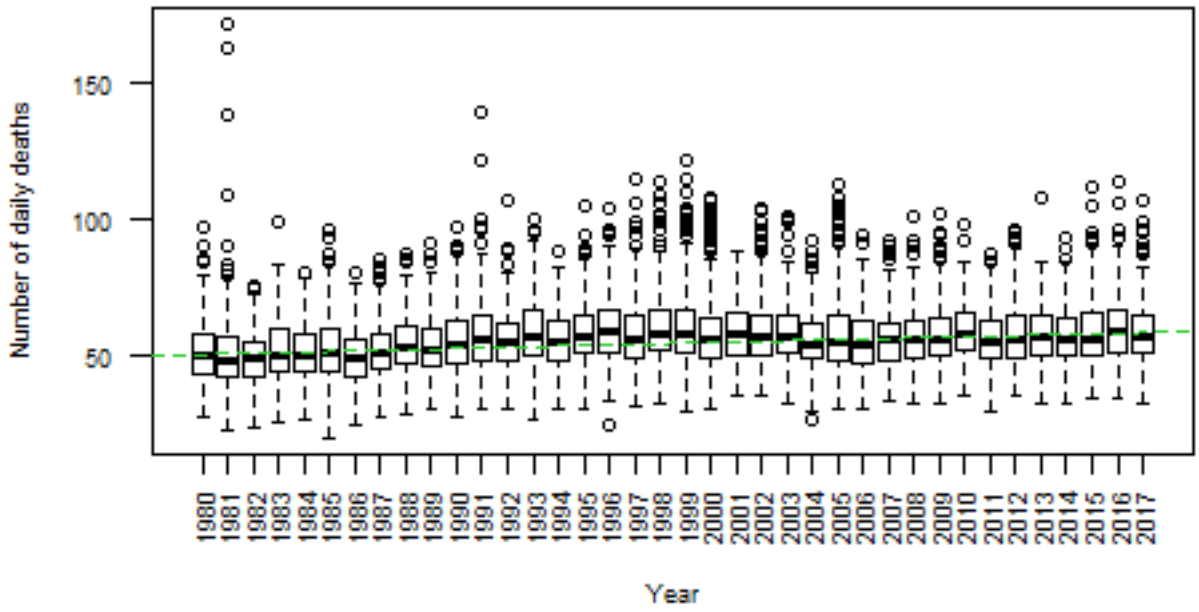

Figure 3.5: Boxplots of total daily deaths by year. And a regression line adjusted to the medians (green).

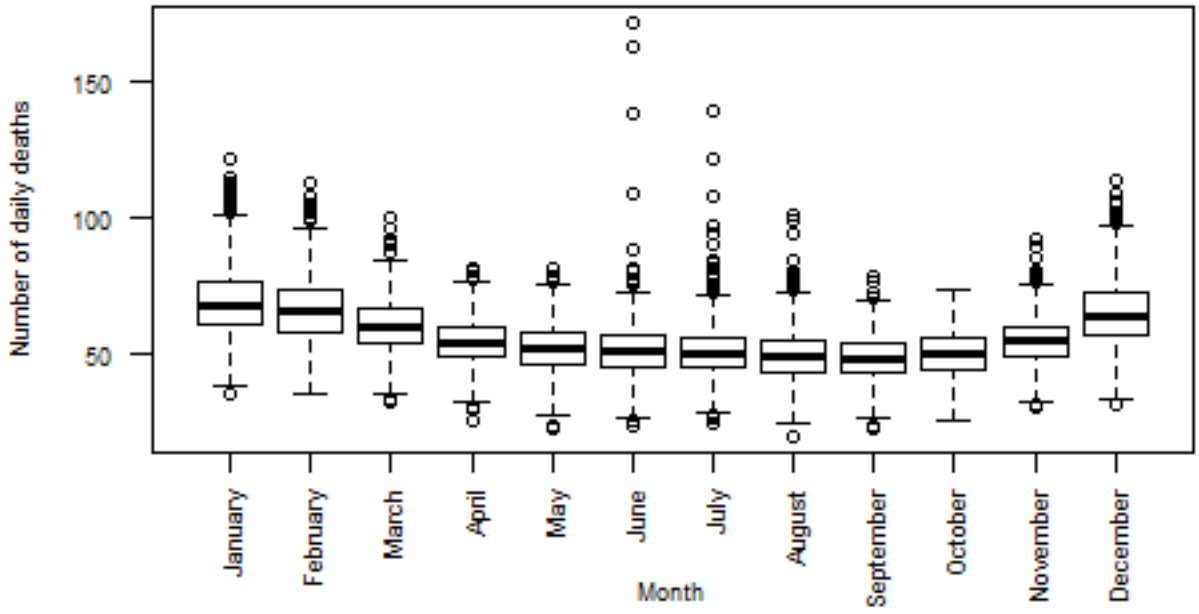

Figure 3.6: Boxplots of total daily deaths by month.

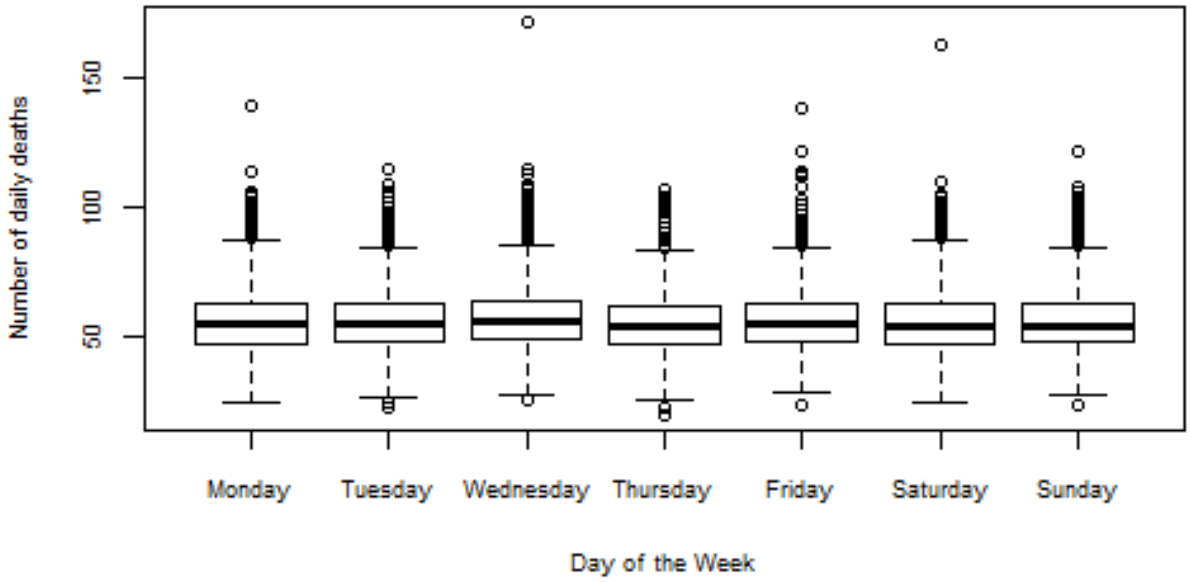

Figure 3.7: Boxplots of total daily deaths by day of the week.

In order to get better familiarised with the evolution of mortality numbers, it is further present a time series of daily counts of deaths as in Figure 3.1, but also with lines representing the stratified by age population (total, under 65 years old and 65 or more years old) - Figure 3.8 -. Also, a graph with boxplots of the distributions of the mortalities by the two age groups and its total is presented - Figure 3.9 -.

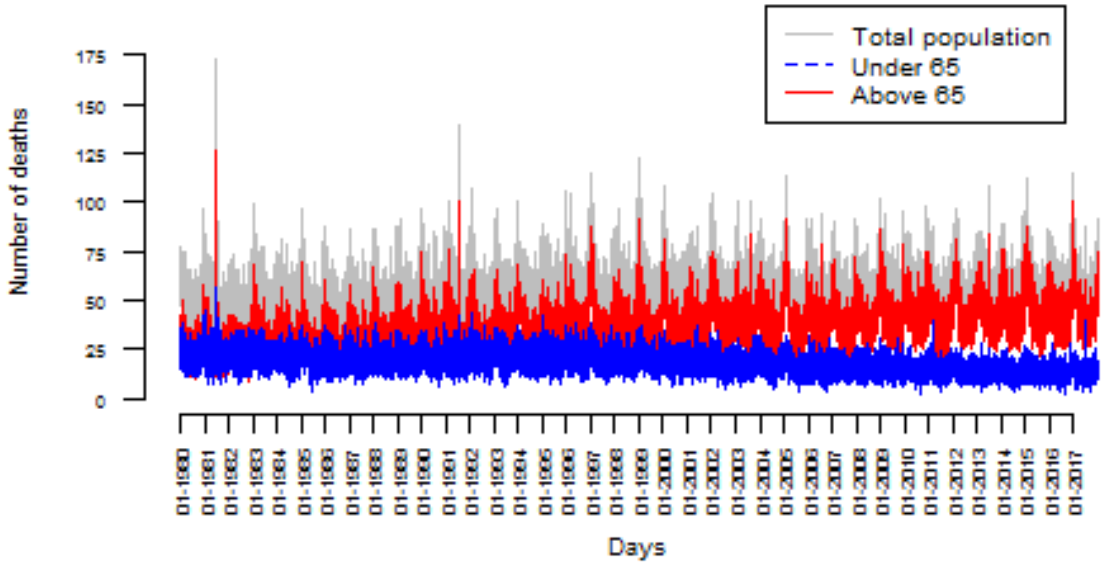

Figure 3.8: Time series of daily mortality (total and stratified by age population).

Figures 3.8 and 3.9 highlight that the number of daily deaths is much higher for the older population, than the younger one. Also they show a larger variability in the daily counts of deaths for the population over 65 years old, comparing with the under 65 years old population. Also Figure 3.8 apparently shows a trend of increase of daily deaths along the years for the oldest, while the youngest population deaths seems to be decreasing in this last decades.

Therefore, it seems that most of the seasonality and growing trend of total counts of deaths is explained by the mortality of the population over 65 years old. This reveals that elderly people are more susceptible to external factors, as mentioned in the introduction of this work.

To study trend and seasonality of daily maximum temperatures in the district of Lisbon, its distribution between 1980 and 2017 is presented through boxplots for each year and month - Figures 3.10 and 3.11, respectively.

Figure 3.10 suggests a very slight trend of increasing maximum temperatures over the years and Figure 3.11 reveals a relevant seasonality along the year, with increasing temperatures from January to July/August and then a decrease until the end of the year.

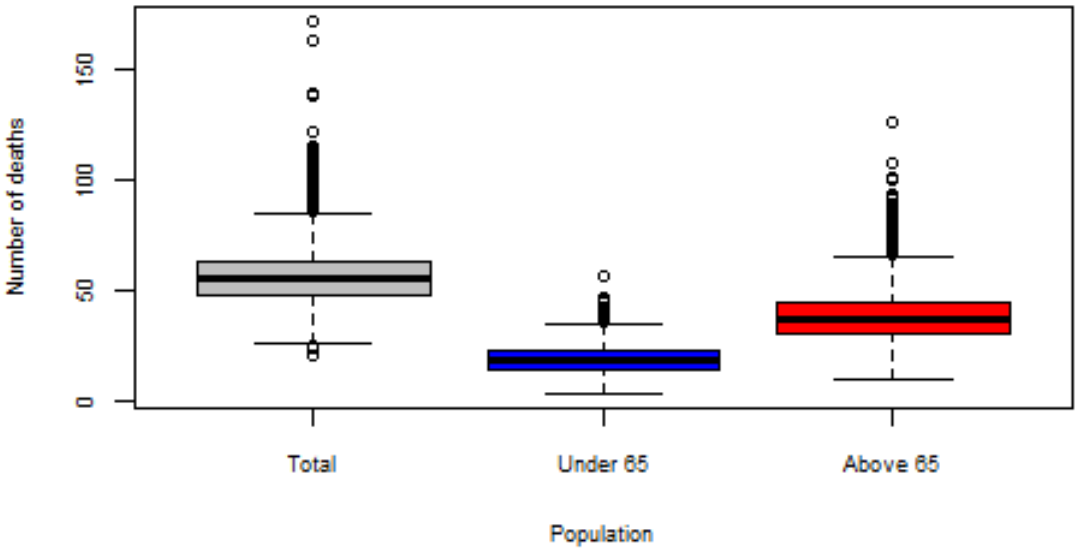

Figure 3.9: Boxplots of daily mortality (total and stratified by age population)

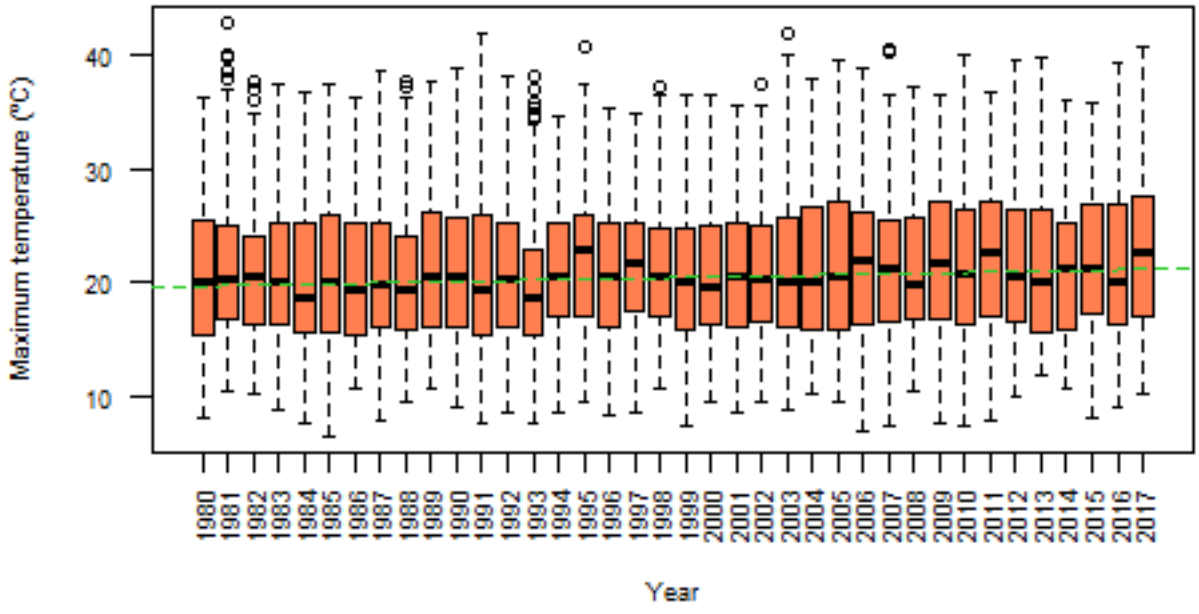

Figure 3.10: Boxplots of daily maximum temperatures by year. And a regression line adjusted to the medians (green).

Figure 3.12 puts the time series of daily counts of deaths together with the time series of daily maximum temperatures for the all period in study (1980 - 2017).

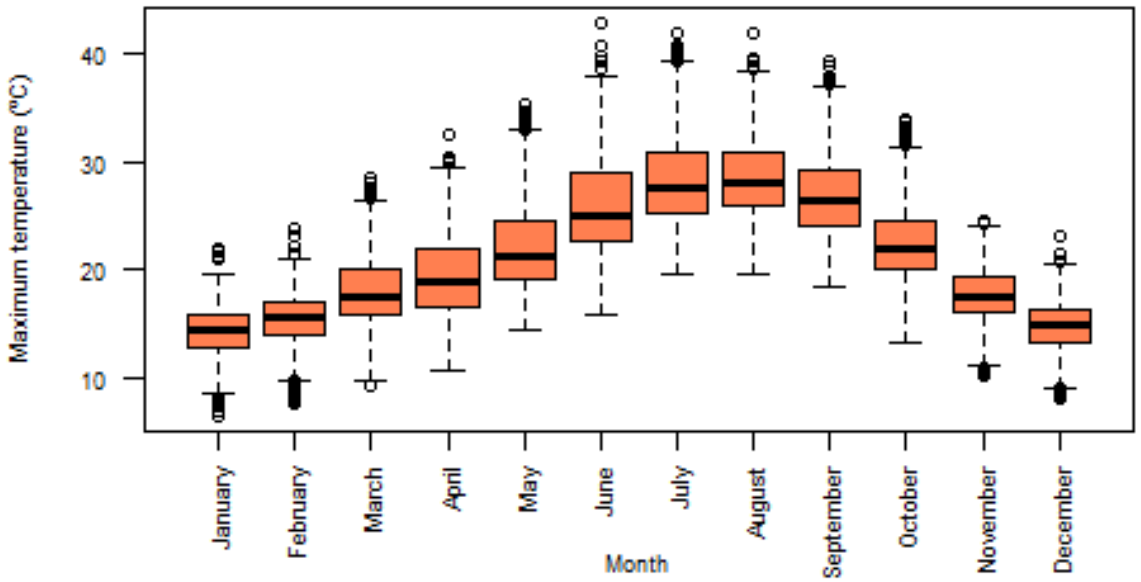

Figure 3.11: Boxplots of daily maximum temperatures by month.

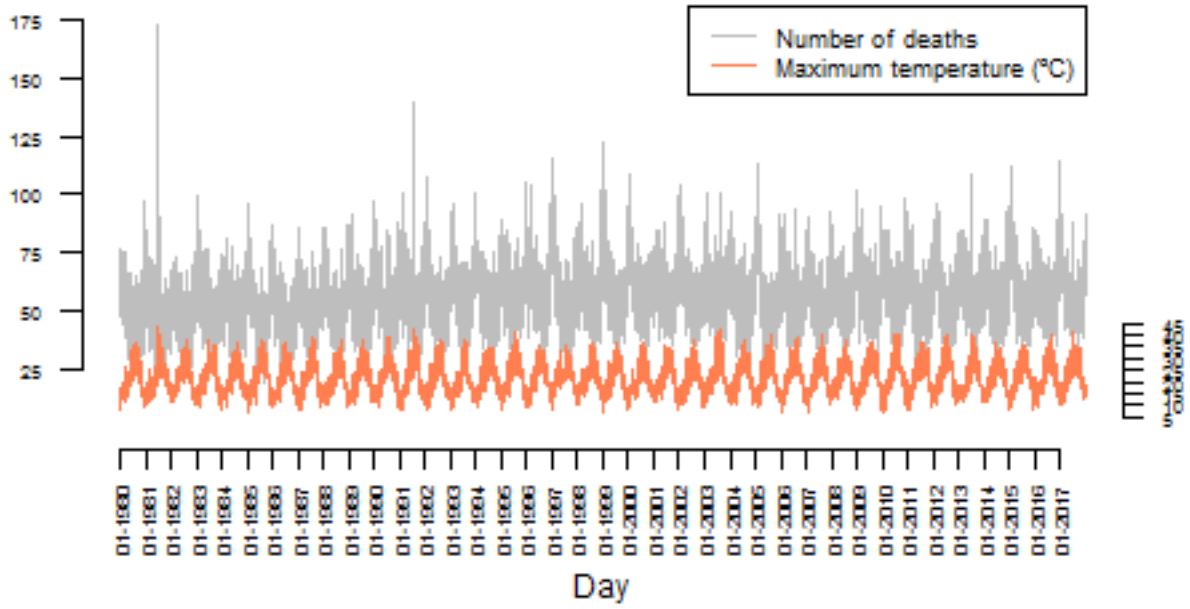

Figure 3.12: Time series of daily maximum temperatures and total mortality.

This graph (Figure 3.12) shows that the seasonality of deaths is inverse to the seasonality of maximum temperatures in Lisbon: the lowest counts of deaths occur in warmer days and vice-versa, as expected.

3. RESULTS

With the exception of same peaks of deaths occurring in periods with very high temperatures.

To better recognise this kind of temperature-mortality relation in extremely warm days, time series of daily deaths *vs.* time series of daily temperatures of some years will be presented. The graphs presented correspond to the years when heatwaves occurred in Lisbon (1981, 1991, 2003, 2006, 2013 and 2017 - Figures 3.14, 3.15, 3.16, 3.17, 3.18 and 3.19, respectively). But first one year without heatwave (2016), will be presented as a comparison term - Figure 3.13.

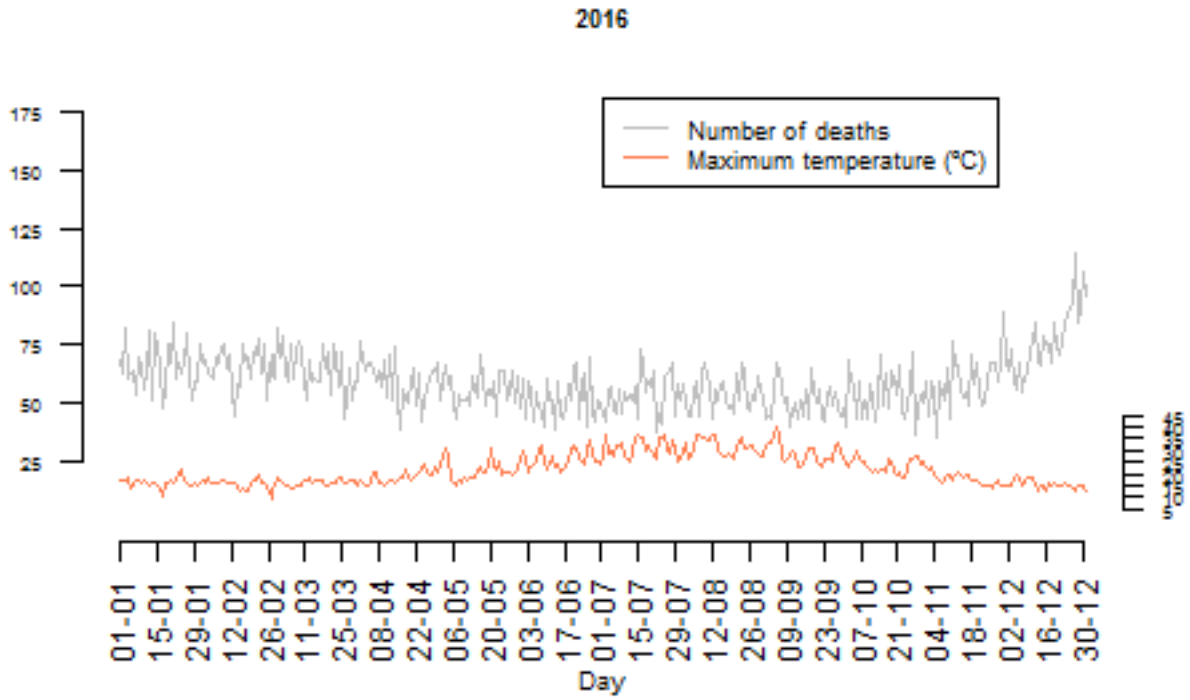

Figure 3.13: Time series of daily maximum temperatures and total mortality of 2016.

From Figure 3.13, one may easily understand the negative relation between temperature and mortality and notice that mortality is lower during the warmer months. However Figures 3.14 to 3.19 show some extreme peaks in the number of deaths during summer. Some of them even exceeding the maximum number of deaths during the winter. All of this excessive summer deaths occurred very close to a heat peak.

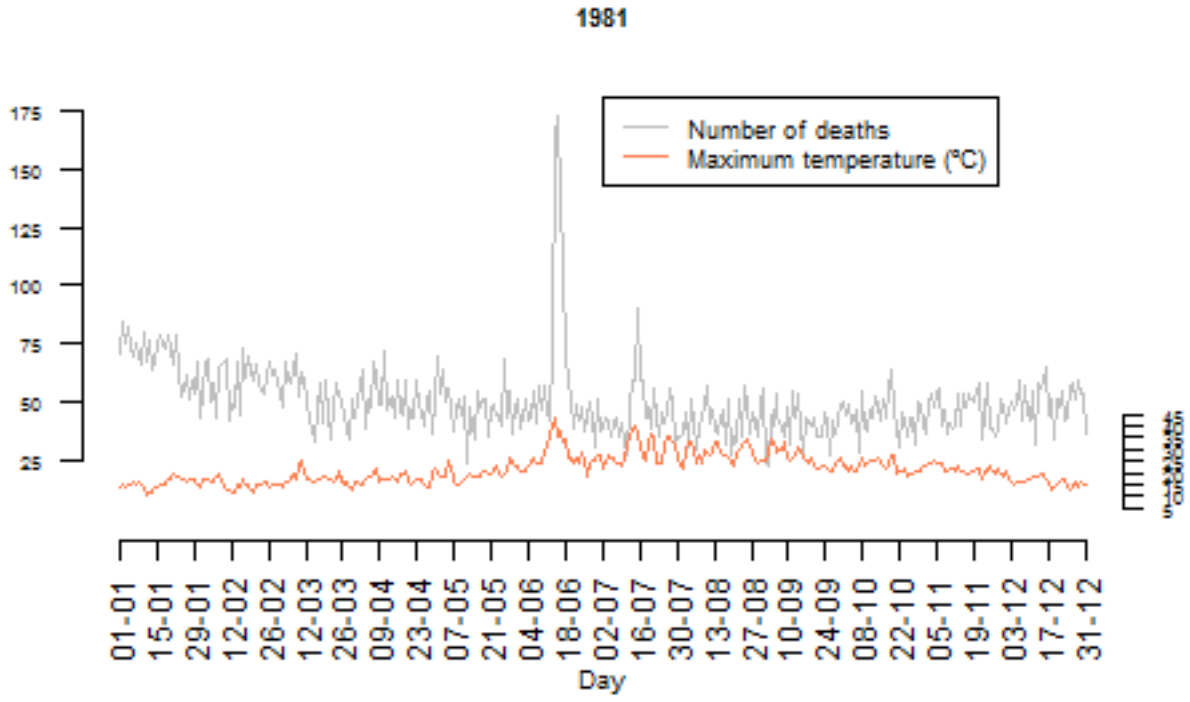

Figure 3.14: Time series of daily maximum temperatures and total mortality of 1981.

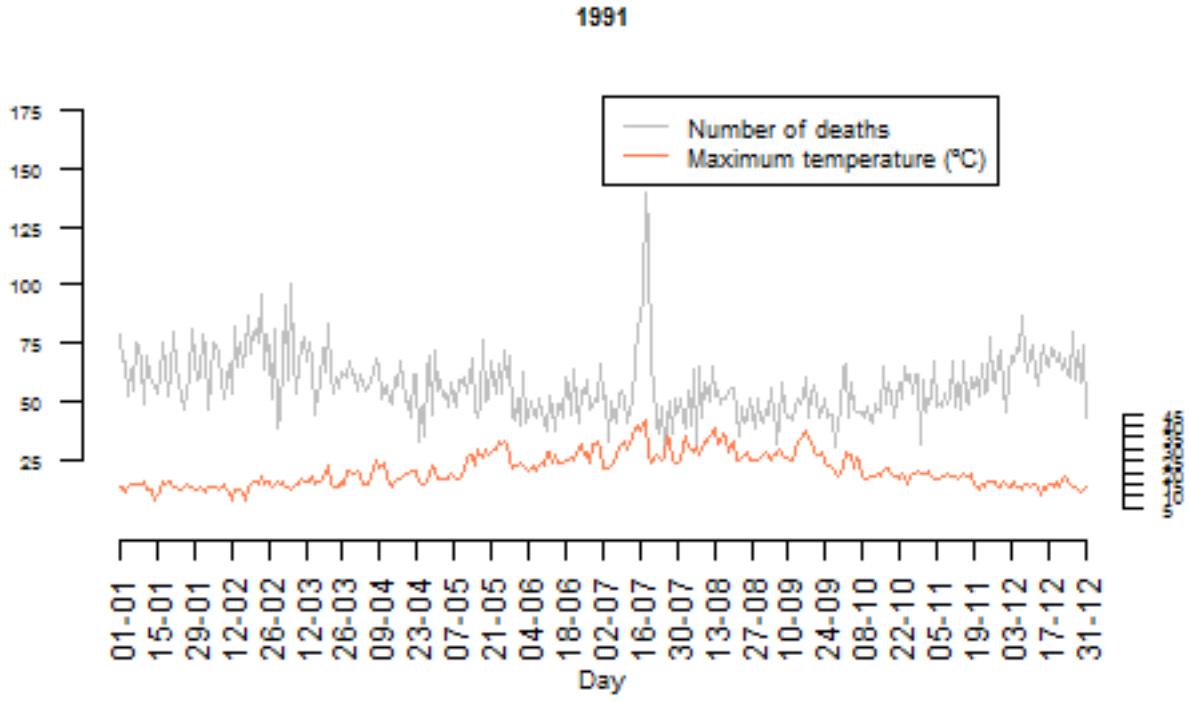

Figure 3.15: Time Series of daily maximum temperatures and total mortality of 1991.

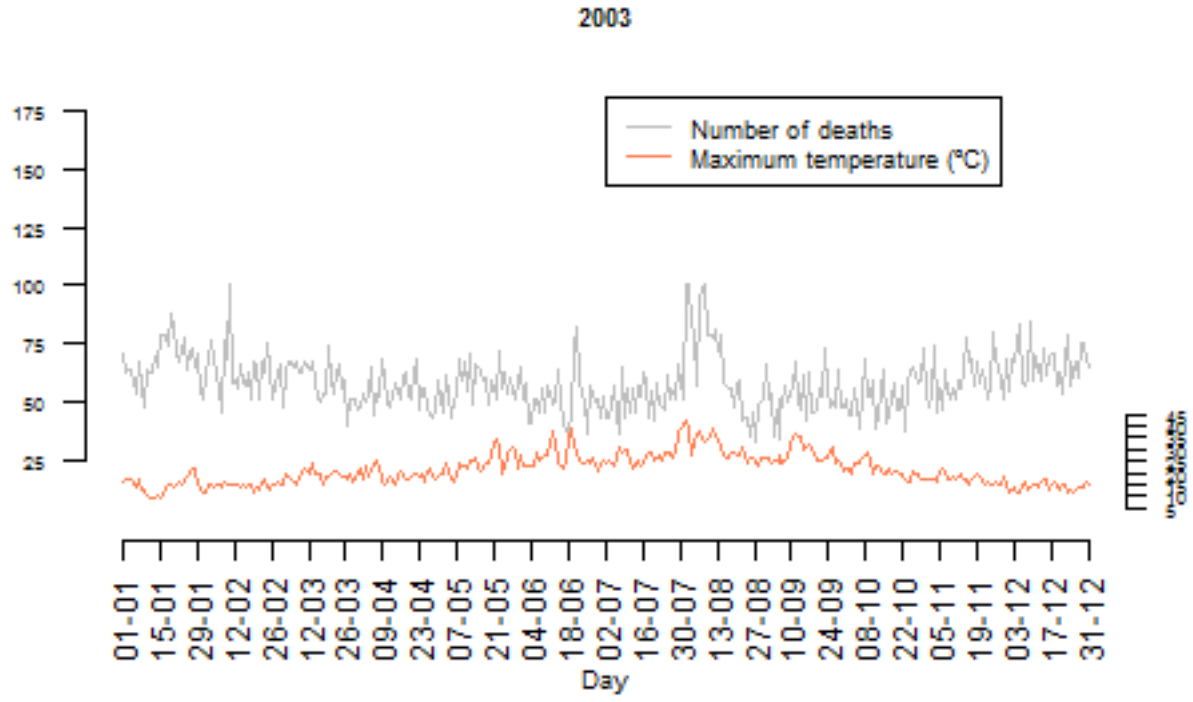

Figure 3.16: Time series of daily maximum temperatures and total mortality of 2003.

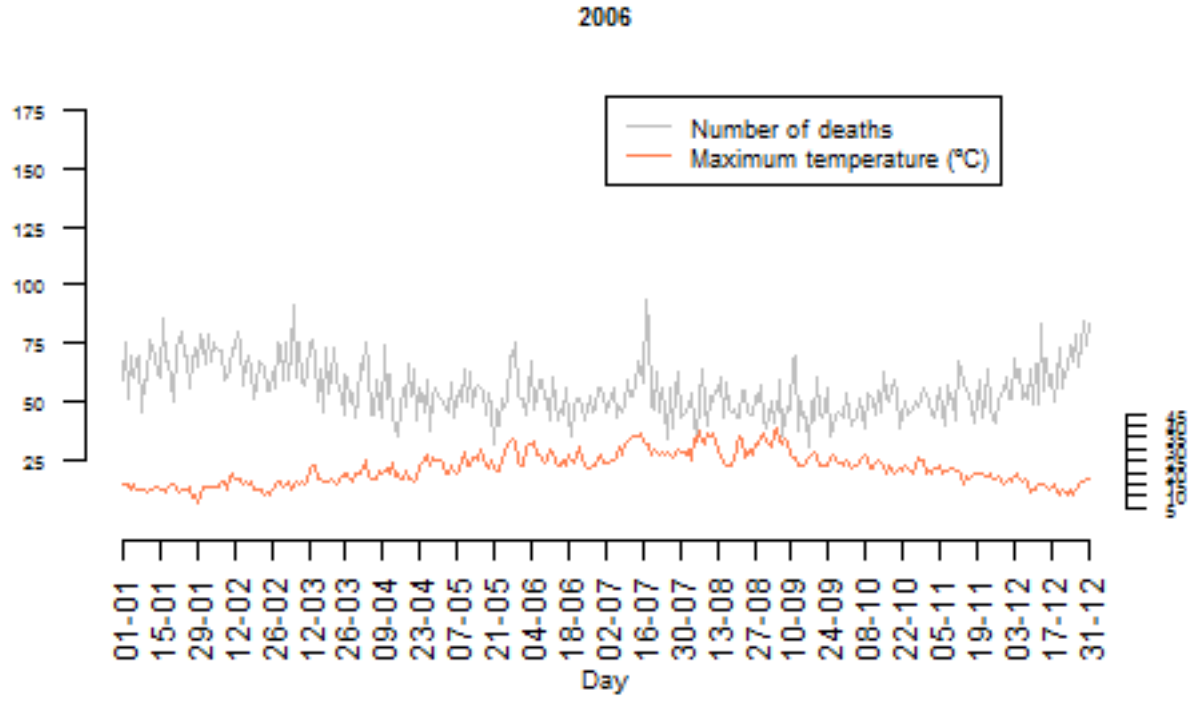

Figure 3.17: Time Series of daily maximum temperatures and total mortality of 2006.

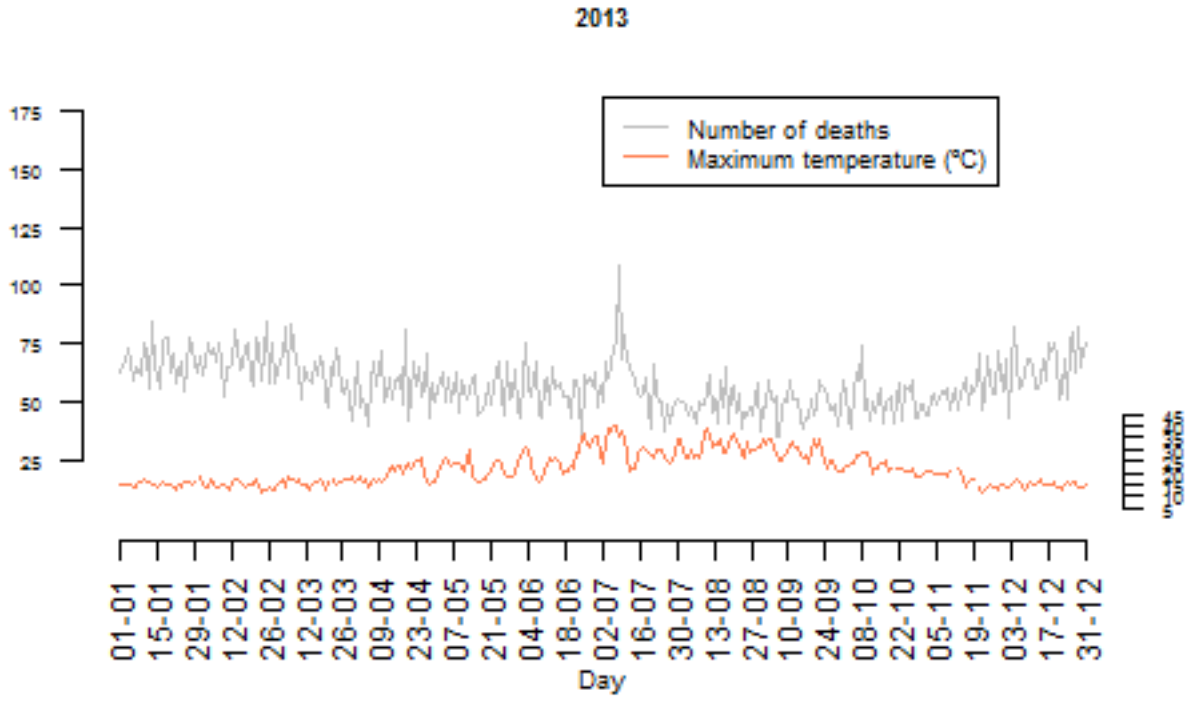

Figure 3.18: Time series of daily maximum temperatures and total mortality of 2013.

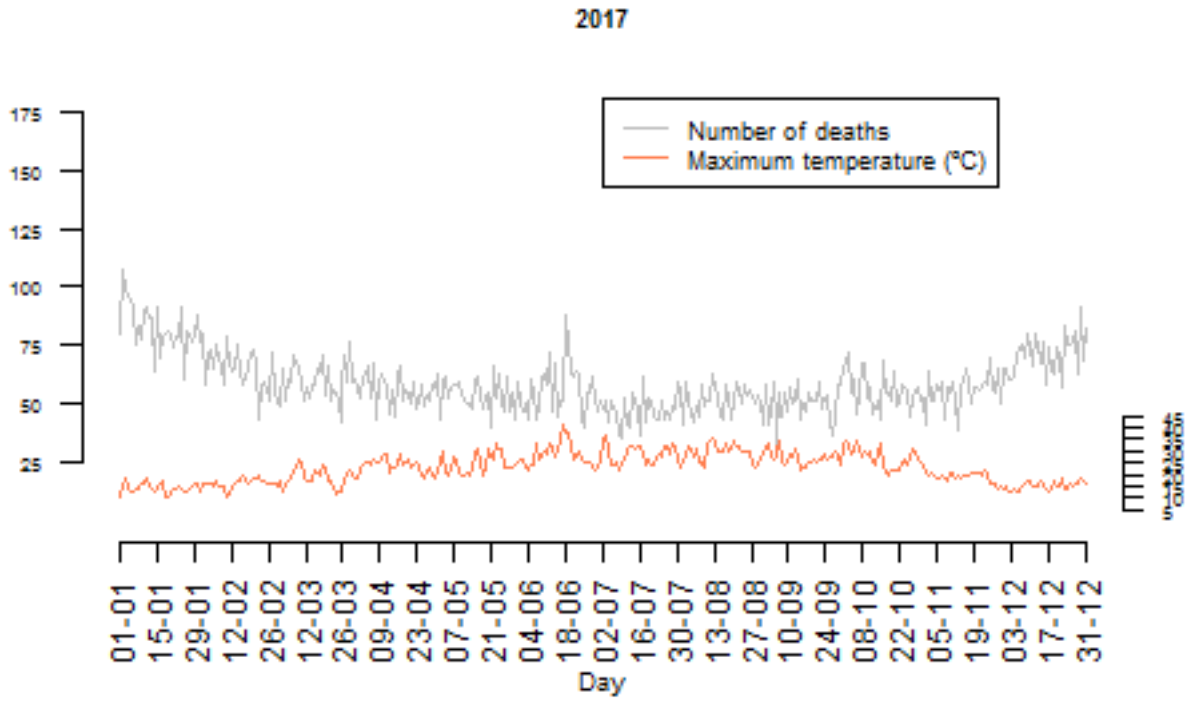

Figure 3.19: Time series of daily maximum temperatures and total mortality of 2017.

3. RESULTS

Finally, the correlation between mortality and temperature in Lisbon is shown: in Figure 3.20 each point corresponds to one day, while in Figure 3.21 the data was aggregated by mean, to simplify the interpretation of the graph. Both figures suggest that extreme temperatures (either cold or hot) are associated with higher counts of deaths. One may say that the minimum mortality in Lisbon is associated with maximum temperatures between 20 ^{o}C and 30 ^{o}C.

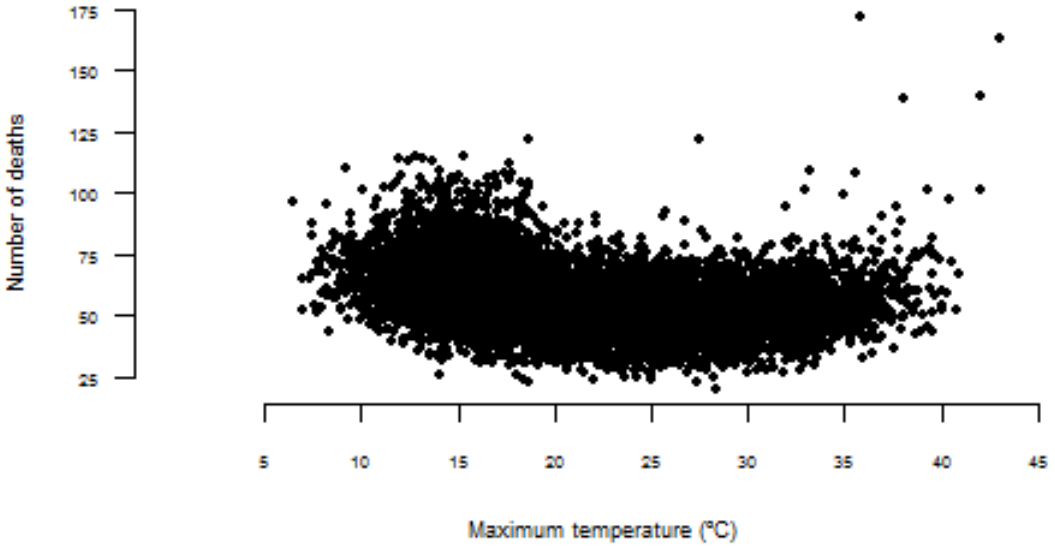

Figure 3.20: Scatter plot of the crude relationship of daily total mortality with daily maximum temperature.

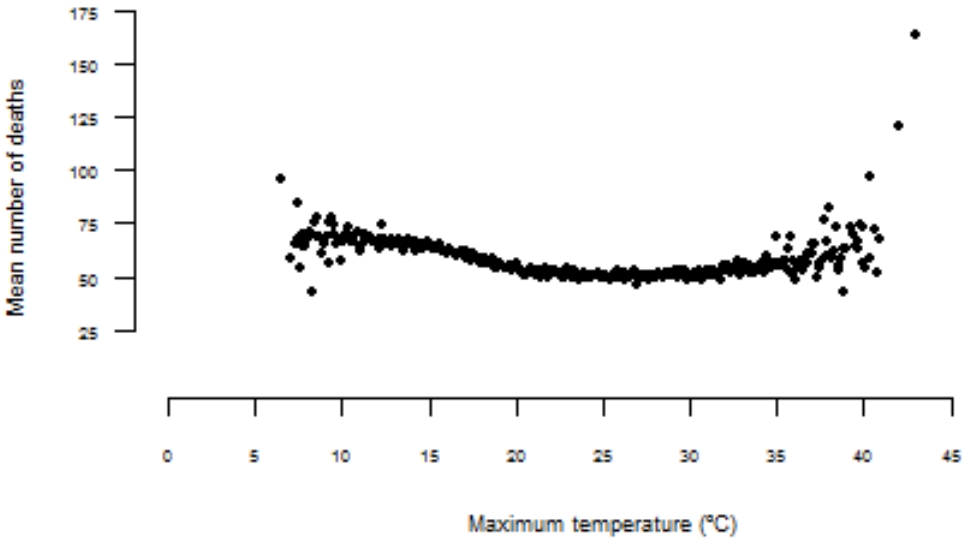

Figure 3.21: Scatter plot of the crude relationship of daily total mortality with daily maximum temperature aggregated by mean.

Also, the analyses of both graphs (3.20 and 3.21) suggests the existence of more outliers associated with extreme hot days, than associated with extreme cold days. Therefore, there is evidence that, even though there are more people dying during winter, the extremely high temperatures seem to be more strongly associated with death, as anticipated in Chapter 1.

Once the main focus of this work is the summer period and the modelling will be restricted to the months from May to September, there are presented some of the previous exploratory graphs restricted to these months, of the years 1980 until 2017 in Lisbon district:

First is presented the distribution of daily counts of deaths of the total population and stratified by age (under and above 65 years old) - Figure 3.22-. This graph keeps the same pattern as Figure 3.9, revealing that summer mortality is also much higher in the older population, which is consistent with older people being more susceptible to the adverse effects of heat. Such suggests that the total summer mortality variability is well explained by the mortality of population above 65 years old.

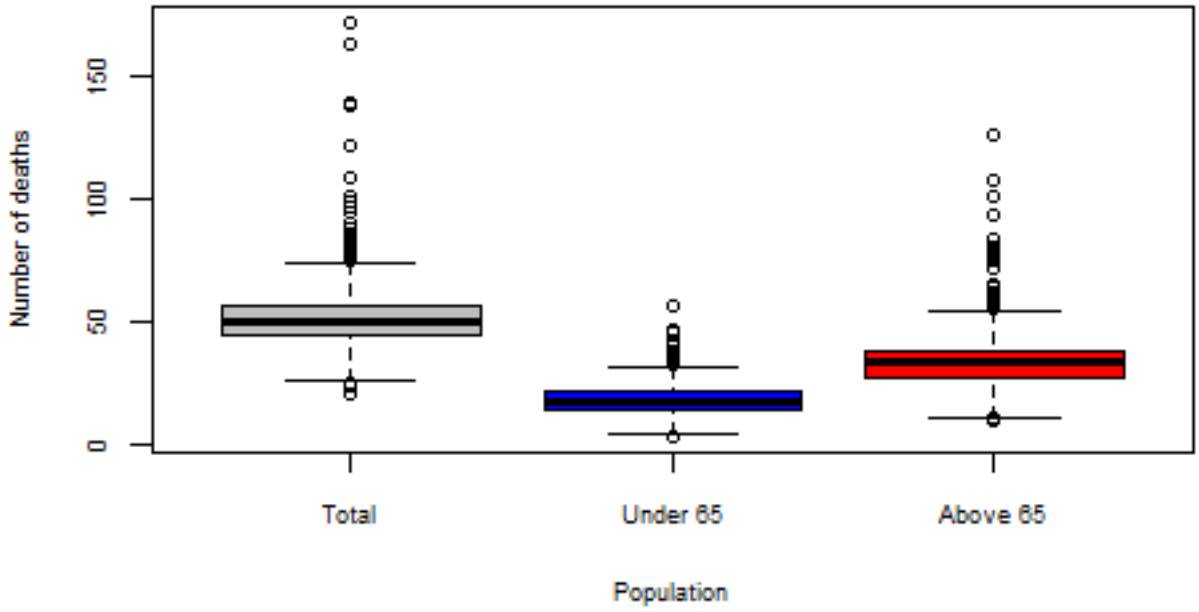

Figure 3.22: Boxplots of daily mortality from May to September (total and stratified by age population).

Then the distribution of total daily counts of deaths for each year, month and day of the week is presented - Figures 3.23, 3.24 and 3.25, respectively-. The boxplots of total daily deaths by year show a smooth growing trend in the number of total deaths from May to September along the last decades. The boxplots by month reveal that the distributions of this counts are very similar in each of this 5 months, suggesting no seasonality nor trend of mortality during each summer (opposing to what was seen considering the all year data in Figure 3.6). And the boxplots by day of the week (DOW) also show a common distribution in the number of deaths for each DOW (like in Figure 3.7, referring to all year data).

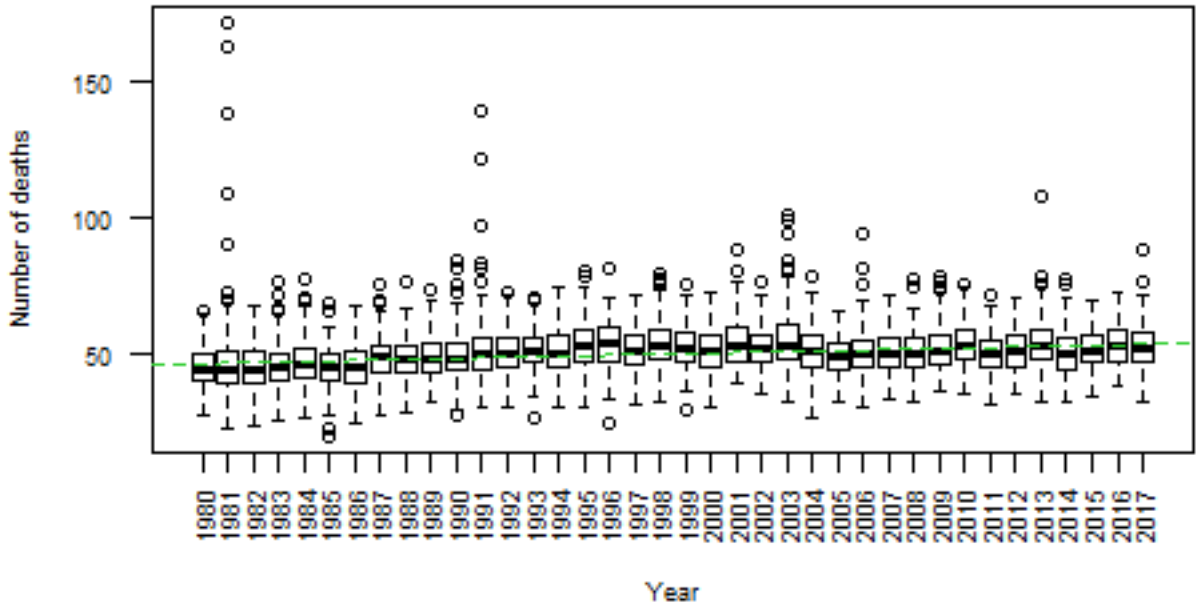

Figure 3.23: Boxplots of total daily deaths by year, from May to September. And a regression line adjusted to the medians (green).

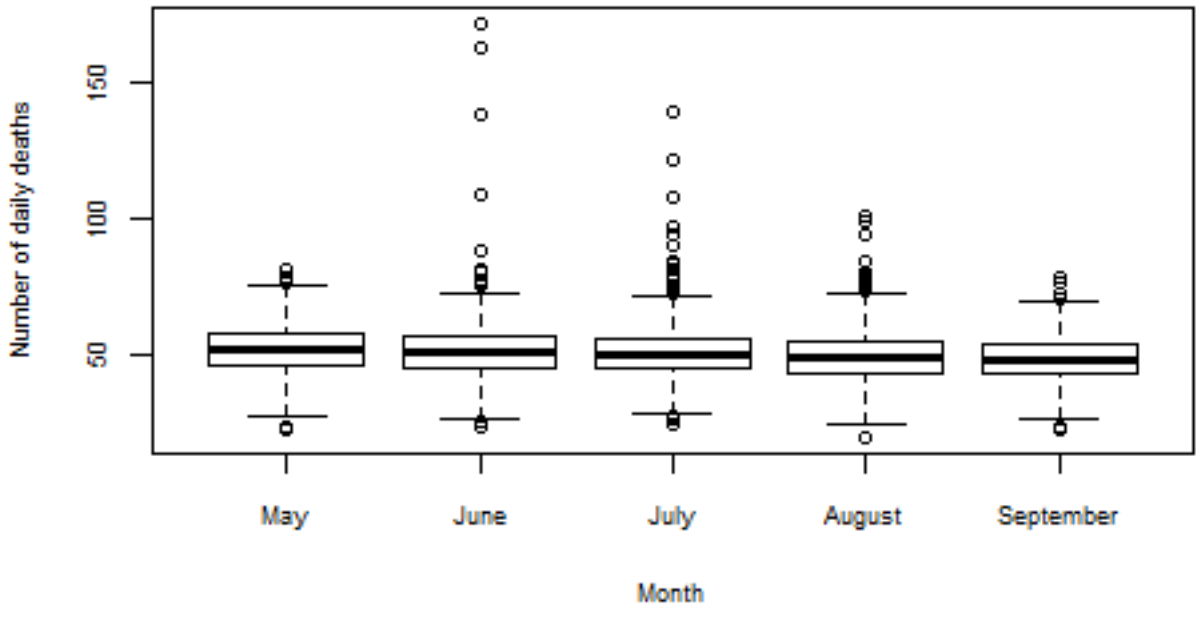

Figure 3.24: Boxplots of total daily deaths by month.

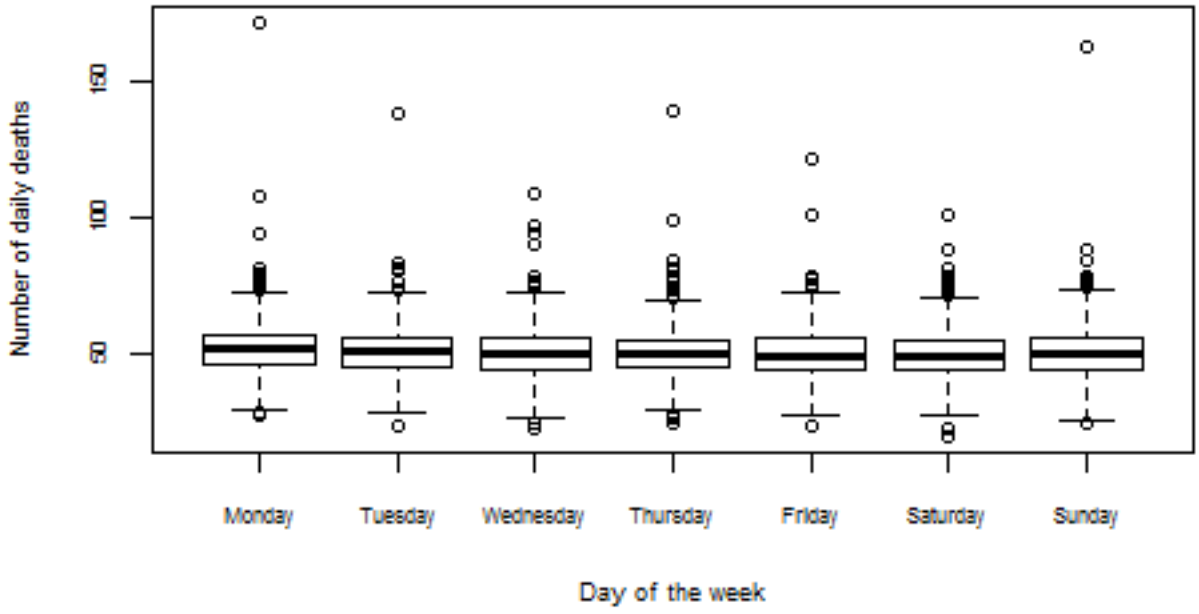

Figure 3.25: Boxplots of total daily deaths by day of the week, from May to September.

Also, the distribution of daily maximum temperatures for each year and month, restricted to data from May to September of 1980 to 2017, are represented in Figures 3.26 and 3.27, respectively. The boxplots of maximum temperatures by year reveal that, even restricting the data to the warmer months, there was still a smooth growing trend of temperatures along the years (as observed in Figure 3.10). It also shows that some years, or in this case some summers, had quite different distributions of daily maximum temperatures from others. The boxplots by month show a clear seasonality of maximum temperatures in each summer, with growing temperatures from May to July and then a slight decrease in September.

Finally, Figures 3.28 and 3.29 show the crude relationship of daily total mortality and maximum temperatures in the district of Lisbon, only for the summer period data (similar to Figures 3.20 and 3.21 for the complete data set). From both this graphs, one may understand that, when restricting the analyses to summer period, the number of deaths shows no relevant association with temperature, except for extreme high temperatures (more than 30 ^{o}C).

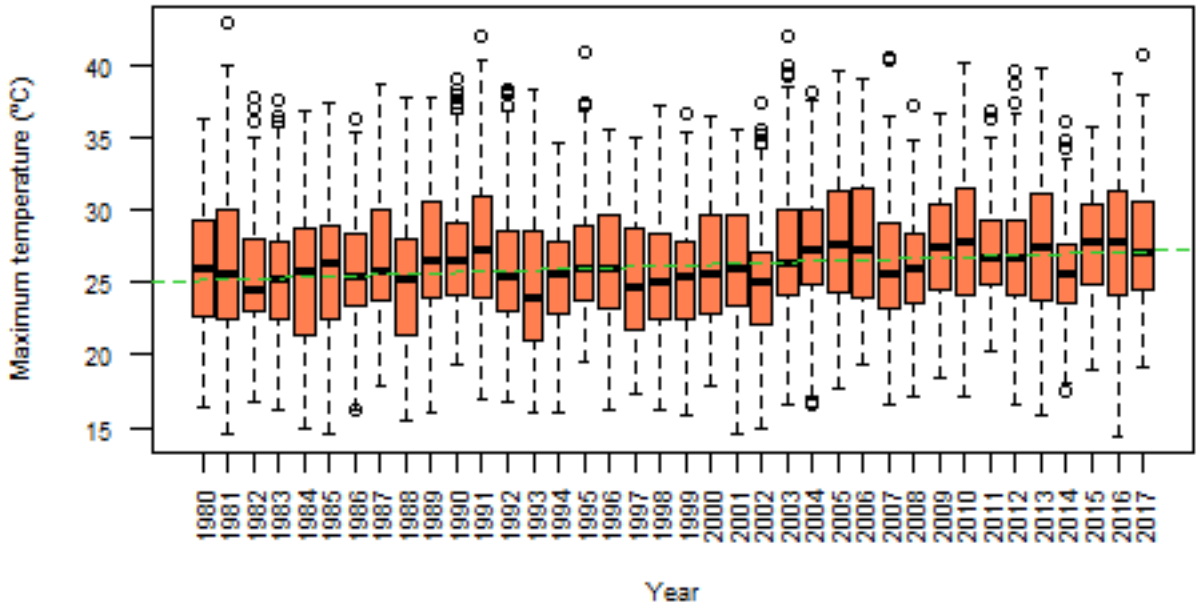

Figure 3.26: Boxplots of daily maximum temperatures by year, from May to September. And a regression line adjusted to the medians (green).

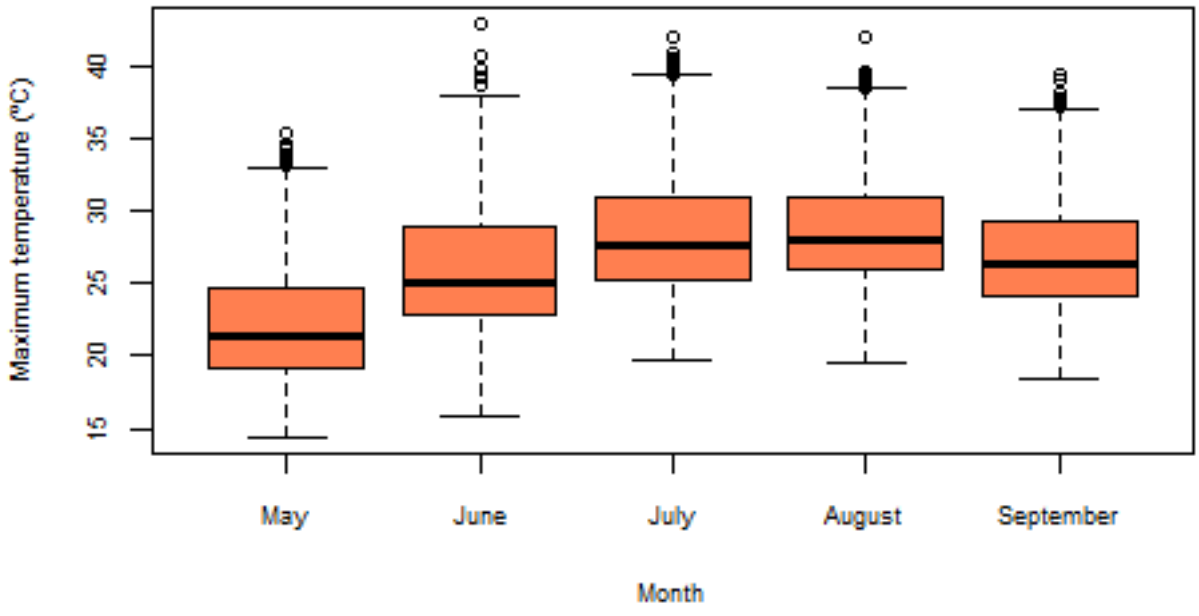

Figure 3.27: Boxplots of daily maximum temperatures by month.

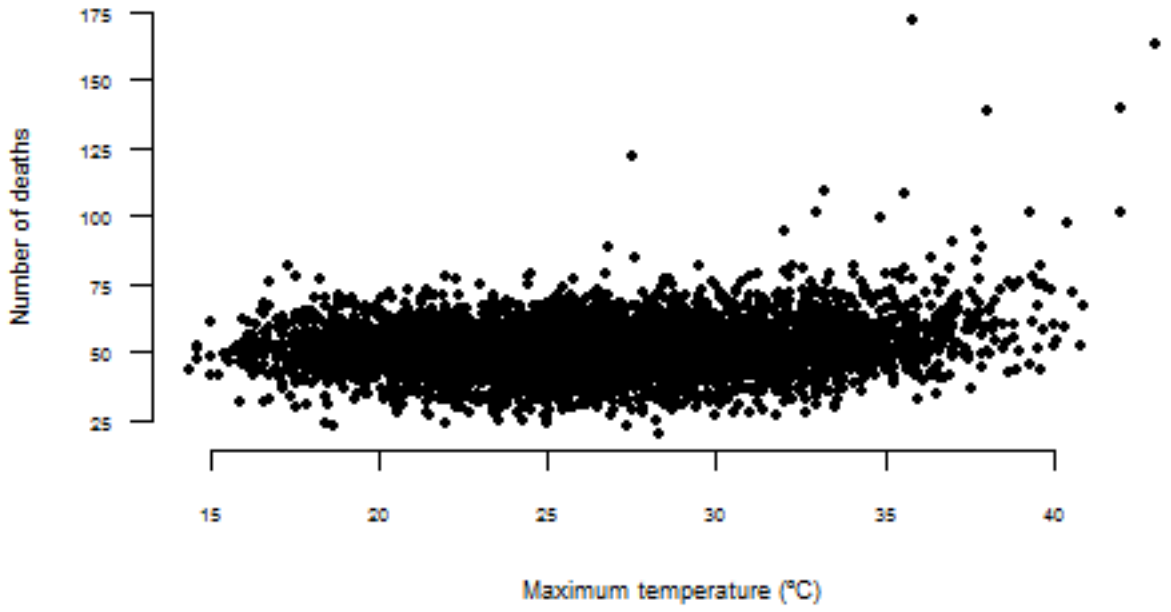

Figure 3.28: Scatter plot of the crude relationship of daily total mortality with daily maximum temperature, using May to September data.

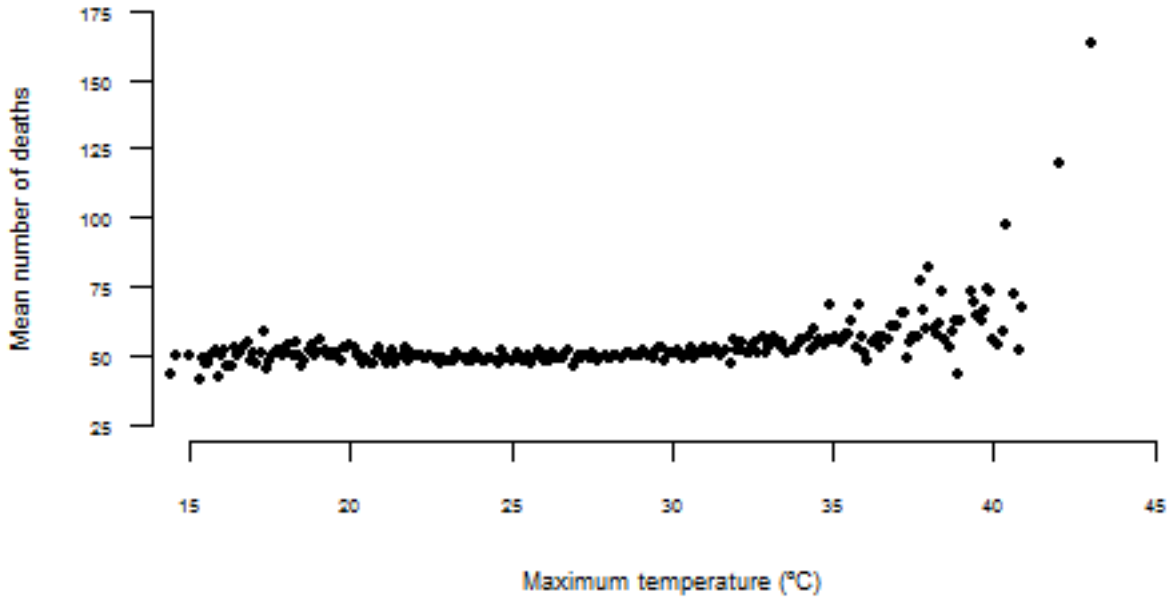

Figure 3.29: Scatter plot of the crude relationship of aggregated by mean daily total mortality with daily maximum temperature, using May to September data.

3.2 Distributed lag non-linear models to describe the heat-mortality relation

This section presents the results of the distributed lag non-linear modelling described in section 2.2.2. Thirty six different models were constructed following the general expression:

$$\log(E(TMort_t)) = \alpha + cb(MaxT_t, MaxT_{t-1}, ..., MaxT_{t-10}) + ns(Time_t, 4) + ns(DOS_t, 5) + DOW_t + \log(TPop_t) \quad , \tag{3.1}$$

where

- $E(TMort_t)$ is the expected value of total mortality on day t;
- α is the intercept;
- $cb(MaxT_t, MaxT_{t-1}, ..., MaxT_{t-10})$ is the cross-basis function of the maximum temperatures, as defined in (2.7) as $s(x,t)$ up to maximum lag of 10 days;
- $ns(Time_t, 4)$ is a natural cubic B-spline of the calendar time with 4 df, to control for long-time trend;
- $ns(DOS_t, 5)$ is a natural cubic B-spline of day of the season with 5 df, to control for seasonality;
- DOW_t is a categorical term of the day of the week;
- $\log(TPop_t)$ is a linear term to control for total population evolution, which enters the model as an *offset*.

Table 3.3 shows all the combinations of the two basis-functions tested for the cross-basis, as well as the number of df and knots: FunVar - basis-function for the predictor dimension (temperature), $f(x)$ - and FunLag - basis-function for the lag dimension, $w(l)$-. Each of the 36 cross-basis functions, resulting from these different combinations, entered a model following (3.1) and were adjusted for the summer period of the years 1980 to 2017. Their respective QAIC are also present in Table 3.3.

According to the QAIC, the model that better adjust to the data is Model 31. This is also the model associated with the the lowest value of QBIC. Hereupon, that is the one to be consider as the final model.

Table 3.3: All basis-functions tested for the cross-basis (FunVar and FunLag), with respective number of degrees of freedom (df), number of knots (nk) and the respective model quasi-AIC (QAIC).

	FunVar			FunLag			
Models	fun	df	nk	fun	df	nk	QAIC
Model 1	ns	4	3	ns	5	3	40595.61
Model 2	ns	4	3	ns	6	4	40594.50
Model 3	ns	4	3	ns	7	5	40602.55
Model 4	ns	4	3	bs	6	3	40594.41
Model 5	ns	4	3	bs	7	4	40602.57
Model 6	ns	4	3	bs	8	5	40612.53
Model 7	ns	5	4	ns	5	3	40575.74
Model 8	ns	5	4	ns	6	4	40576.38
Model 9	ns	5	4	ns	7	5	40584.40
Model 10	ns	5	4	bs	6	3	40575.88
Model 11	ns	5	4	bs	7	4	40584.50
Model 12	ns	5	4	bs	8	5	40597.31
Model 13	ns	6	5	ns	5	3	40559.23
Model 14	ns	6	5	ns	6	4	40560.99
Model 15	ns	6	5	ns	7	5	40569.16
Model 16	ns	6	5	bs	6	3	40560.36
Model 17	ns	6	5	bs	7	4	40569.44
Model 18	ns	6	5	bs	8	5	40584.92
Model 19	bs	5	3	ns	5	3	40527.73
Model 20	bs	5	3	ns	6	4	40527.01
Model 21	bs	5	3	ns	7	5	40533.02
Model 22	bs	5	3	bs	6	3	40526.35
Model 23	bs	5	3	bs	7	4	40533.21
Model 24	bs	5	3	bs	8	5	40546.28
Model 25	bs	6	4	ns	5	3	40491.22
Model 26	bs	6	4	ns	6	4	40492.90
Model 27	bs	6	4	ns	7	5	40502.44
Model 28	bs	6	4	bs	6	3	40492.33
Model 29	bs	6	4	bs	7	4	40503.13
Model 30	bs	6	4	bs	8	5	40518.86
Model 31	**bs**	**7**	**5**	**ns**	**5**	**3**	**40450.04**
Model 32	bs	7	5	ns	6	4	40452.76
Model 33	bs	7	5	ns	7	5	40462.01
Model 34	bs	7	5	bs	6	3	40452.17
Model 35	bs	7	5	bs	7	4	40463.48
Model 36	bs	7	5	bs	8	5	40482.43

3.2.1 Final model

The chosen model (model 31 of Table 3.3) was fitted with a quadratic B-spline with 5 equally spaced internal knots, defining 7 df for the dimension of the variable; and a natural cubic B-spline with 5 df, 3 internal knots equally spaced in the log scale and an intercept for the lag dimension. So the two basis-functions combined in the crossbasis matrix describe the overall effect over these two dimensions, with 35 df ($7 \times 5 = 35$). This, together with the other terms mentioned before (3.1), creates a model with a total of 51 df (i.e., 51 estimated parameters). The summary of the model is:

```
Call:
glm(formula = TMort ~ cb + ns(Date, df = 4) + ns(dos, df = 5) +
    offset(log(TPop)) + dow, family = quasipoisson(), data = dados_Summer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1254  -0.7567  -0.0318   0.6934   4.0582

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -11.564377   1.173095  -9.858  < 2e-16 ***
cbv1.l1             0.584519   0.738202   0.792 0.428501
cbv1.l2             0.306717   0.373437   0.821 0.411489
cbv1.l3            -0.430934   0.455012  -0.947 0.343636
cbv1.l4             0.099147   0.950104   0.104 0.916892
cbv1.l5             0.279934   0.578540   0.484 0.628502
cbv2.l1             0.552454   0.714990   0.773 0.439747
cbv2.l2             0.302286   0.350252   0.863 0.388145
cbv2.l3            -0.461489   0.438934  -1.051 0.293125
cbv2.l4             0.078937   0.930671   0.085 0.932409
cbv2.l5             0.222011   0.563840   0.394 0.693782
cbv3.l1             0.585608   0.718353   0.815 0.414986
cbv3.l2             0.285762   0.354252   0.807 0.419894
cbv3.l3            -0.477333   0.441914  -1.080 0.280121
cbv3.l4             0.122726   0.933641   0.131 0.895425
cbv3.l5             0.179762   0.566469   0.317 0.750998
cbv4.l1             0.604659   0.717421   0.843 0.399362
cbv4.l2             0.287306   0.352853   0.814 0.415543
cbv4.l3            -0.484752   0.440917  -1.099 0.271632
cbv4.l4             0.115366   0.932543   0.124 0.901549
cbv4.l5             0.176320   0.565567   0.312 0.755236
cbv5.l1             0.694030   0.718201   0.966 0.333912
cbv5.l2             0.293629   0.354030   0.829 0.406917
cbv5.l3            -0.507460   0.441738  -1.149 0.250695
cbv5.l4             0.211745   0.933358   0.227 0.820538
cbv5.l5             0.119234   0.566230   0.211 0.833227
cbv6.l1             0.827697   0.718843   1.151 0.249603
```

```
cbv6.l2             0.296268   0.354510   0.836 0.403352
cbv6.l3            -0.506636   0.442076  -1.146 0.251826
cbv6.l4             0.205581   0.933124   0.220 0.825634
cbv6.l5             0.151534   0.566280   0.268 0.789020
cbv7.l1             1.515879   0.724814   2.091 0.036535 *
cbv7.l2             0.630603   0.366358   1.721 0.085255 .
cbv7.l3            -0.781625   0.453506  -1.724 0.084849 .
cbv7.l4             1.162567   0.936464   1.241 0.214493
cbv7.l5            -0.388693   0.572482  -0.679 0.497190
ns(Data, df = 4)1   0.251641   0.010234  24.588  < 2e-16 ***
ns(Data, df = 4)2  -0.020080   0.010302  -1.949 0.051318 .
ns(Data, df = 4)3   0.122800   0.021182   5.797 7.09e-09 ***
ns(Data, df = 4)4   0.061084   0.009244   6.608 4.25e-11 ***
ns(dos, df = 5)1   -0.083418   0.019584  -4.260 2.08e-05 ***
ns(dos, df = 5)2   -0.117728   0.024964  -4.716 2.46e-06 ***
ns(dos, df = 5)3   -0.116789   0.017288  -6.756 1.56e-11 ***
ns(dos, df = 5)4   -0.164181   0.044145  -3.719 0.000202 ***
ns(dos, df = 5)5   -0.072925   0.012472  -5.847 5.27e-09 ***
dowTuesday         -0.019145   0.007496  -2.554 0.010678 *
dowWednesday       -0.026172   0.007527  -3.477 0.000511 ***
dowThursday        -0.030359   0.007544  -4.024 5.80e-05 ***
dowFriday          -0.029967   0.007537  -3.976 7.10e-05 ***
dowSaturday        -0.038969   0.007550  -5.162 2.53e-07 ***
dowSunday          -0.031600   0.007528  -4.198 2.74e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.195976)

    Null deviance: 9924.6  on 5813  degrees of freedom
Residual deviance: 6916.3  on 5763  degrees of freedom
  (380 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 4
```

The diagnostic plots associated to this model are represented in Figure 3.30. The Residuals *vs.* Fitted plot show no particular pattern of the residuals, which are equally spread around the horizontal line. Therefore, it does not seem to exist any non-linear relationship not captured by the generalised linear model. The top right panel shows the Normal Q-Q plot, which shows there are no reason to doubt about the normal distribution of the residuals. In the Scale-Location (or Spread-Location) plot the values of the residuals appear to be randomly spread evenly for small predicted values, but increasing for predicted values larger than 4.2. Values for which the data is more sparse, once most observations are restricted to smaller predicted values (big concentration of dots on the left side). And finally, the Residuals *vs.* Leverage plot does not identify any outlier with possible influence on the regression results.

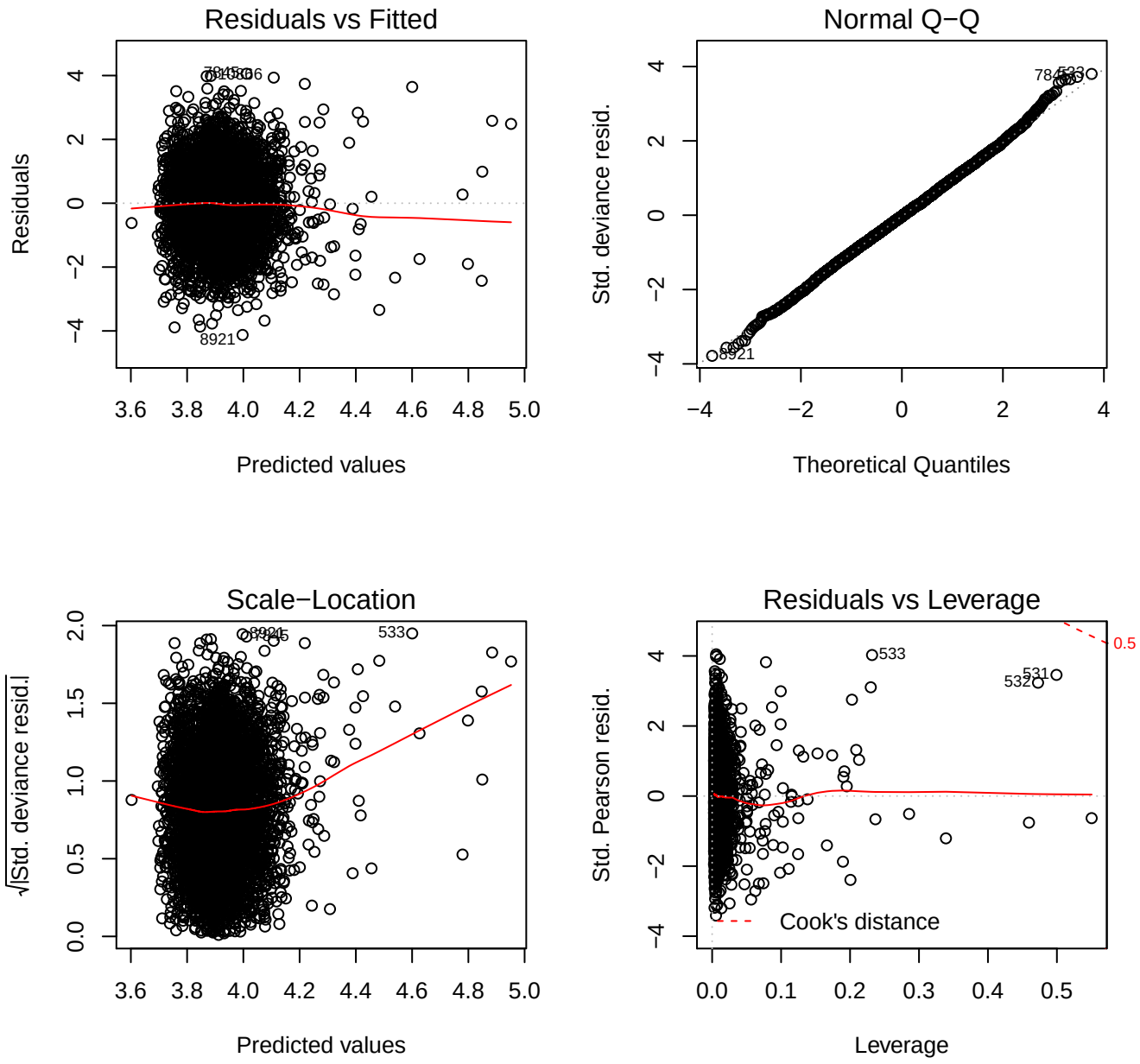

Figure 3.30: Diagnostic plots of the final model.

This model will then be used to predict the global effect of the maximum temperatures on mortality in the district of Lisbon. Figure 3.31 shows a three dimensional graph of the RR along the daily maximum temperatures and lag, with reference at 25.7 ^{o}C (overall minimum mortality temperature). The same is presented in Figure 3.32, as a contour plot.

Both figures show an accentuated increase in the RR for temperatures above 40 ^{o}C between lags 0 and 4. This means that such high temperatures increase mortality on the same day and in the 4 following days. Also, the graphs show a RR $<$ 1 for temperatures under 15 ^{o}C over lags 0 to 3 and 10, indicating some possible protective effect of maximum temperatures under 15 ^{o}C, preventing deaths for a few days and 10 days later.

According to this model, during summer, temperature has its maximum adverse effect at 43 ^{o}C and lag 1, corresponding to a RR of 2.47. Opposing to the lowest RR of 0.66 for maximum temperatures of 11 ^{o}C also at lag 1.

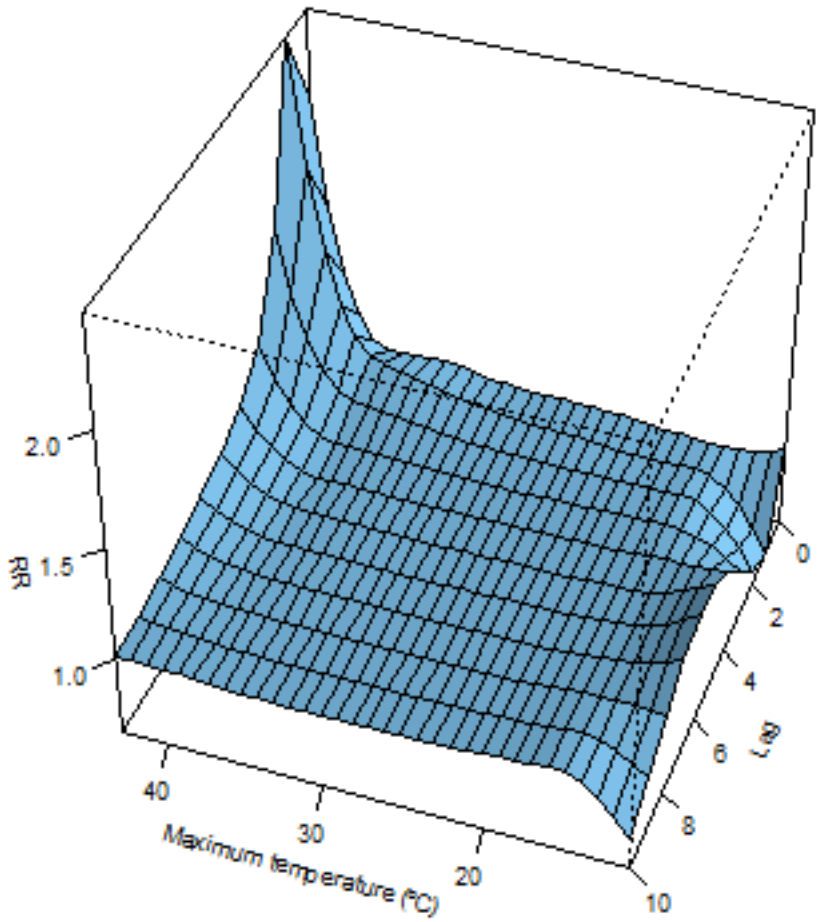

Figure 3.31: 3D plot of RR along temperature and lag.

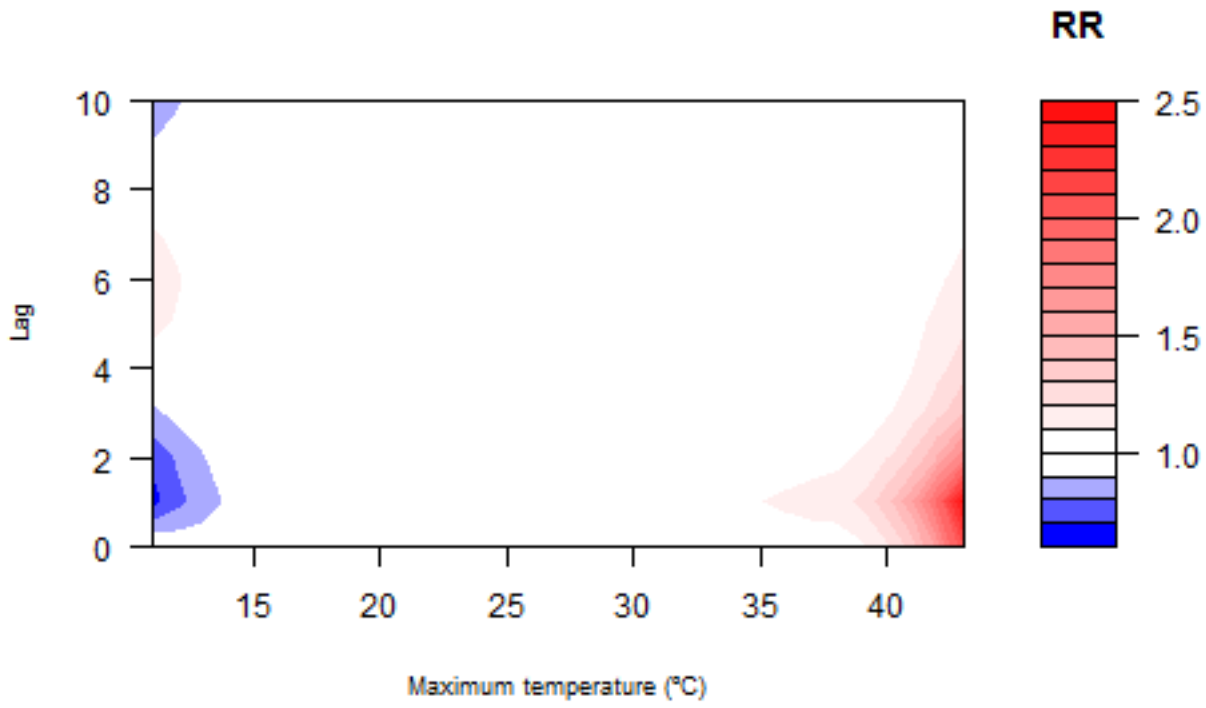

Figure 3.32: Contour plot of RR along temperature and lag.

To better understand the relationship along each of the two dimensions and also account for uncertainty of the estimates, Figure 3.33 presents slices of the 3D plot for specific values of both dimensions (temperature and lag), with the respective 95% confidence interval of the estimates. It shows the effect by temperature at specific lags (0, 2, 4 and 10) and the effect by lag at specific temperatures (15 ^{o}C, 33 ^{o}C, 35 ^{o}C and 41 ^{o}C).

Finally, Figure 3.34 shows the overall effect of temperature, summing up the 10 days of lag considered. It reveals an overall protective effect of temperatures lower than 15 ^{o}C and a growing RR as temperatures grow higher than 30 ^{o}C. Also it is noticed that the uncertainty of the estimates are much higher for lower temperatures, which are not the focus of this work.

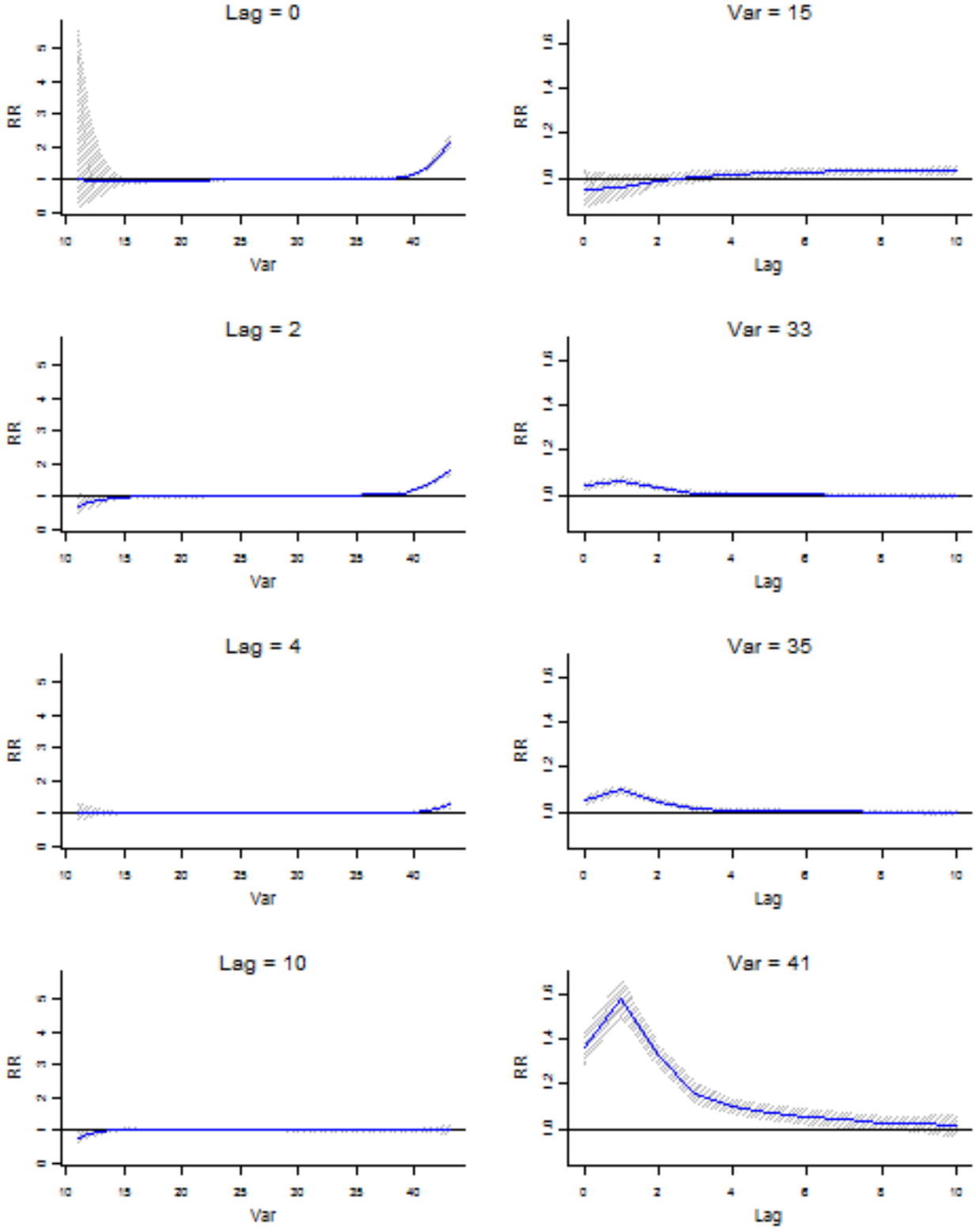

Figure 3.33: Plot of RR by the variable temperature (Var) at specific lags (left) and RR by lag at 0.1^{th}, 90^{th}, 95^{th} and 99.9^{th} percentiles of summer daily maximum temperatures distribution (right).

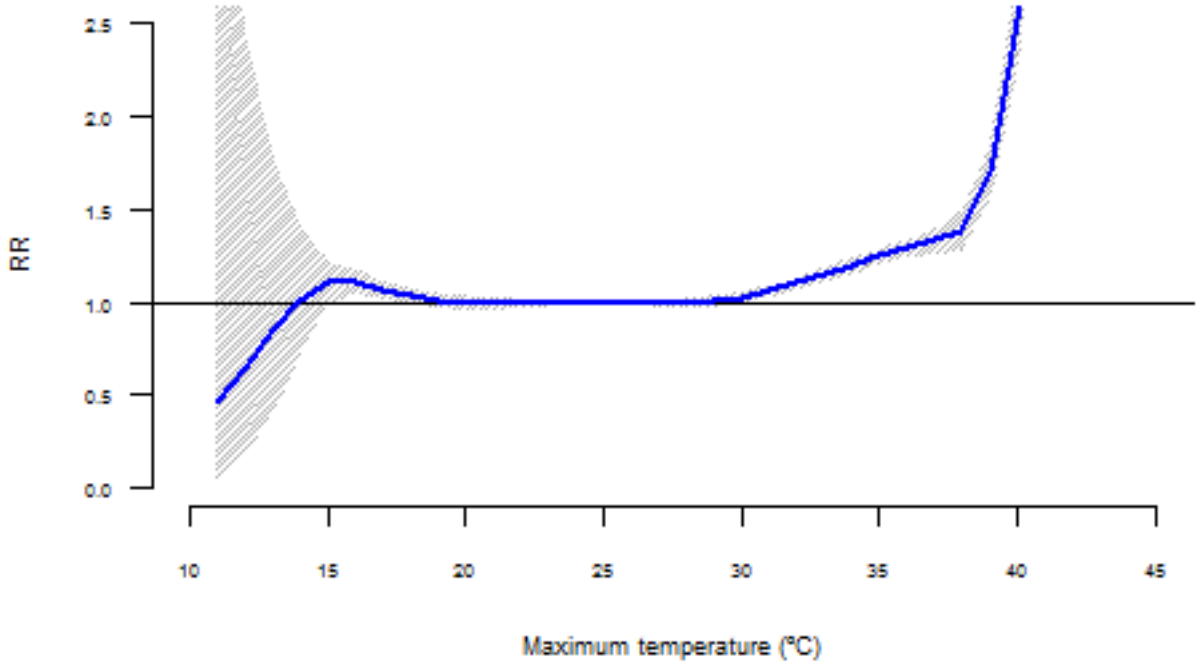

Figure 3.34: Plot of overall RR along temperature.

3.3 Comparison Between the DLNM and the ICARO Model

This section presents the results obtained from the assessment and comparison between the DLNM and the ICARO model by the cross-validation processes described in section 2.2.3.

Table 3.4 shows the MAE, MSE and RMSE obtained with the final DLNM and the ICARO model for each year tested through a *rolling-origin update* cross-validation. Using this method, the daily numbers of deaths were predicted for each year summer since 1990 until 2017 by both models adjusted with the data set of all previous years (since 1980).

Table 3.4: Error Measures (MAE - mean absolute error; MSE - mean squared error; RMSE - root mean squared error) obtained for each year tested using the *rolling-origin update* method of cross-validation.

Error Measures	MAE		MSE		RMSE	
Year	DLNM	ICARO	DLNM	ICARO	DLNM	ICARO
1990	7.88	5.90	213.97	53.74	14.63	7.33
1991	7.99	7.08	181.16	93.82	13.46	9.69
1992	8.58	6.35	182.61	64.44	13.51	8.03
1993	8.46	6.38	196.47	67.91	14.02	8.24
1994	8.44	6.91	196.33	75.76	14.01	8.70
1995	7.81	6.47	150.86	69.65	12.28	8.35
1996	6.78	6.66	72.06	69.40	8.49	8.33
1997	6.70	6.32	70.92	63.43	8.42	7.96
1998	6.35	6.82	61.70	77.50	7.85	8.80
1999	6.50	6.55	71.93	65.32	8.48	8.08
2000	7.36	7.72	104.16	97.69	10.21	9.88
2001	7.15	8.22	84.74	122.35	9.21	11.06
2002	6.46	5.96	80.83	57.25	8.99	7.57
2003	9.68	8.50	285.21	129.71	16.89	11.39
2004	7.95	7.77	151.77	91.84	12.32	9.58
2005	7.40	8.49	195.37	120.38	13.98	10.97
2006	7.51	8.02	164.38	103.00	12.82	10.15
2007	9.09	6.81	382.59	71.21	19.56	8.44
2008	6.86	6.70	75.03	72.32	8.66	8.50
2009	6.10	7.20	61.10	87.25	7.81	9.34
2010	8.01	6.61	171.61	69.25	13.10	8.32
2011	5.74	6.40	51.06	65.60	7.15	8.10
2012	7.55	6.78	159.88	70.53	12.64	8.40
2013	10.39	6.33	667.07	64.86	25.83	8.05
2014	6.55	7.02	65.41	73.87	8.09	8.59
2015	6.32	6.30	61.08	59.23	7.82	7.70
2016	8.40	6.41	191.19	59.59	13.83	7.72
2017	7.09	6.41	126.13	62.78	11.23	7.92

Following this procedure, the three error measures used are lower for the ICARO than the DLNM for most years tested. The analysis of Table 3.4 shows only 9 of the 28 years tested with better performance of DLNM according to the MAE (1998, 1999, 2000, 2001, 2005, 2006, 2009, 2011 and 2014) and 5 years according to the MSE and RMSE (1998, 2001, 2009, 2011 and 2014).

Considering all years tested, the average of MAE obtained with the DLNM is 7.53857, higher than the one obtained with the ICARO model, which is 6.898424. The average of the MSE is 159.8793 for

the DLNM, which is much higher than the 77.84594 obtained for the ICARO model and the average of all the RMSE obtained are 11.97441 for the DLNM and 8.757503 for the ICARO.

However, as explained in section 2.2.3, the relationship between heat and mortality may have changed along time. Thus, using such a large period of time in the modelling may be causing some disturbance in the analyses. To assess if that is happening, Figures 3.35, 3.36 and 3.37 present the RR across temperatures estimated by the DLNM adjusted for the last three decades: time periods 1988-1997, 1998-2007 and 2008-2017, respectively.

These three figures show some differences of the relative risk for each temperature along the years considered. Note that all three periods show an increase in the RR for daily maximum temperatures above approximately 30 ^{o}C, however the curve seems to be smoother for the later period (Figure 3.37).

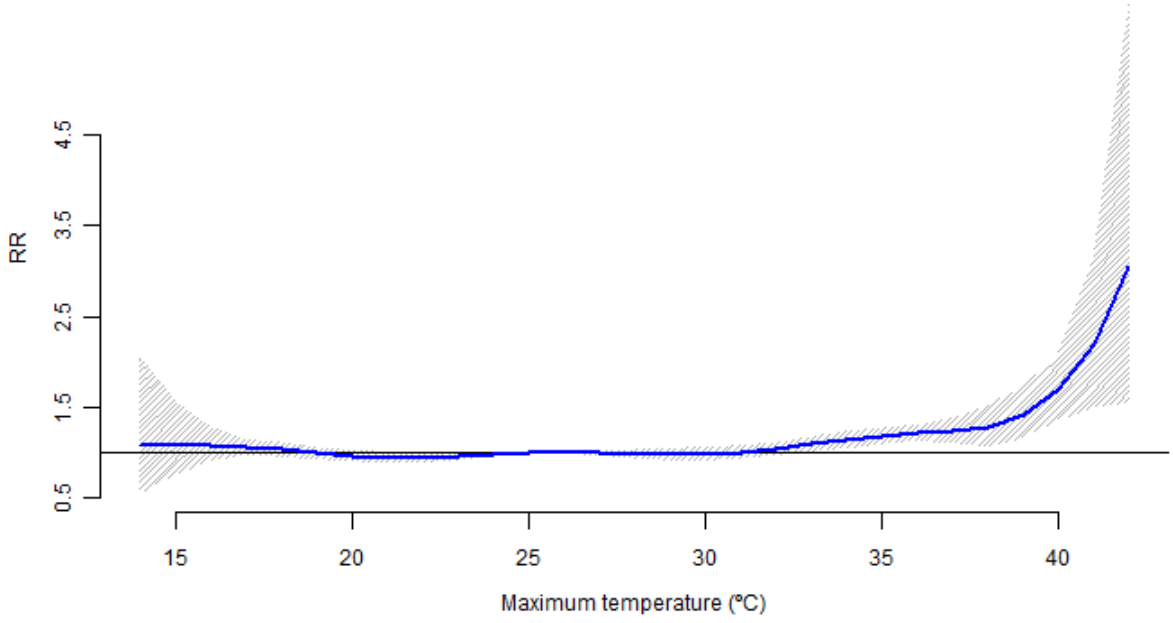

Figure 3.35: Plot of overall RR along temperature, predicted for years 1988 to 1997.

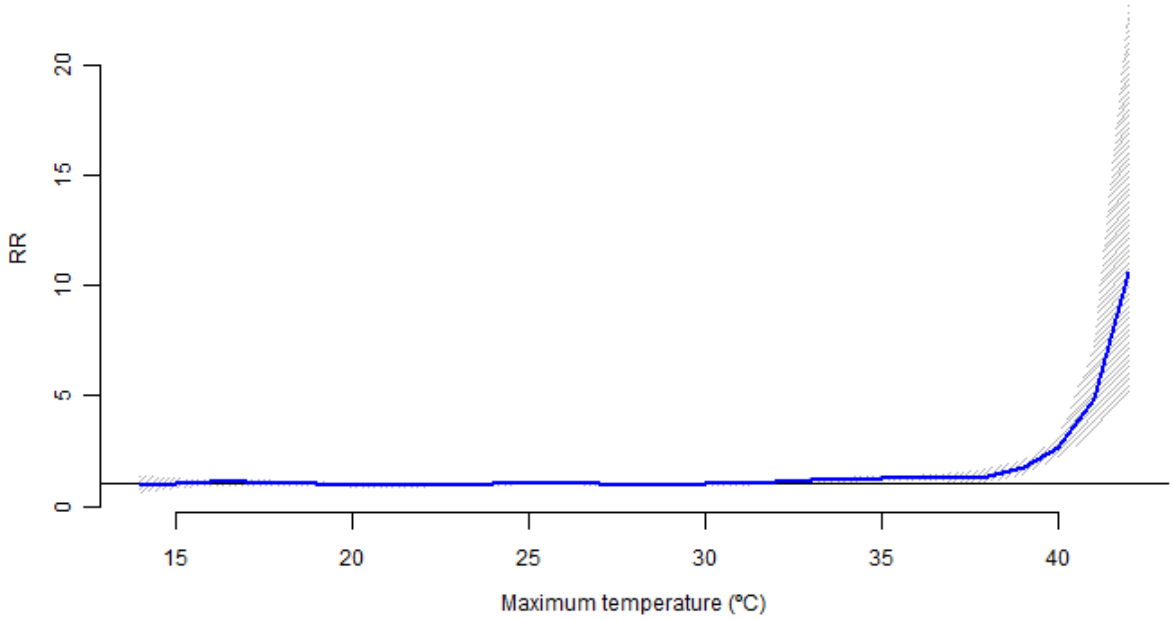

Figure 3.36: Plot of overall RR along temperature, predicted for years 1998 to 2007.

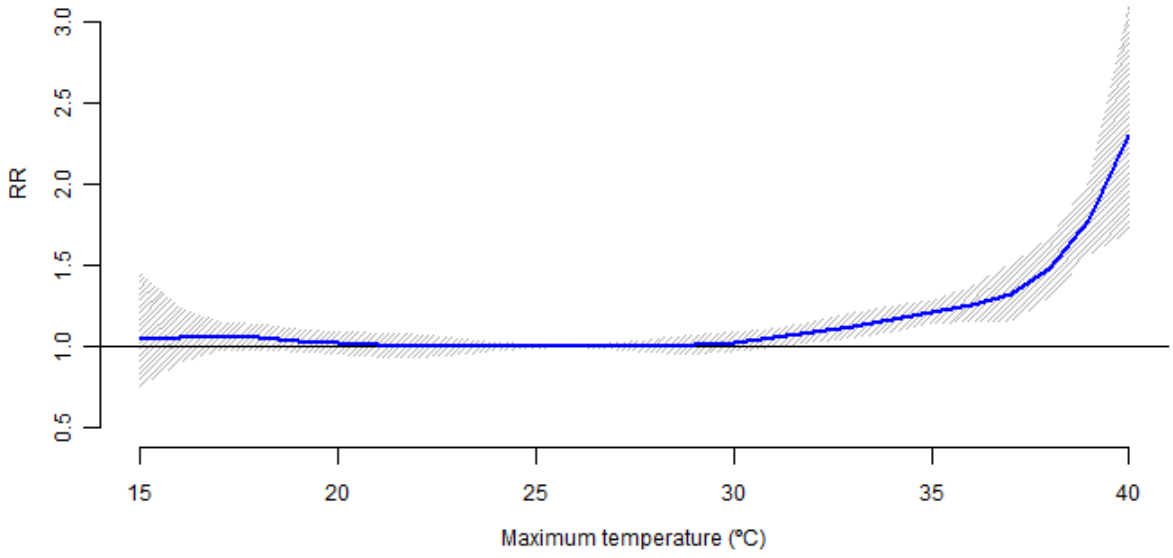

Figure 3.37: Plot of overall RR along temperature, predicted for years 2008 to 2017.

For that reason, we also preformed the cross-validation following the *rolling-window evaluation* method. Tables 3.5 and 3.6 shows the MAE, MSE and RMSE obtained with both models (DLNM and ICARO), using 10 and 5 years length window, respectively. In those cases, the number of deaths for each year tested were predicted by both models adjusted with the data set of the previous 10 or 5 years.

The three Tables 3.4, 3.5 and 3.6 present the error measures values rounded to two decimal cases.

Table 3.5: Error Measures (MAE - mean absolute error; MSE - mean squared error; RMSE - root mean squared error) obtained for each year tested using 10 years *rolling-window evaluation* method of cross-validation.

Error Measures	MAE		MSE		RMSE	
Year	DLNM	ICARO	DLNM	ICARO	DLNM	ICARO
1990	7.88	5.90	213.97	53.74	14.63	7.33
1991	8.30	7.24	215.79	97.50	14.69	9.87
1992	7.42	6.50	107.57	66.16	10.37	8.13
1993	7.31	6.51	106.28	68.04	10.31	8.25
1994	7.15	7.66	81.28	85.12	9.02	9.23
1995	7.01	6.85	88.89	76.95	9.43	8.77
1996	6.78	7.34	70.88	82.83	8.42	9.10
1997	6.59	8.15	67.47	98.75	8.21	9.94
1998	6.41	7.96	63.85	92.51	7.99	9.62
1999	6.20	8.81	64.38	108.72	8.02	10.43
2000	6.54	10.20	67.00	150.58	8.19	12.27
2001	6.77	8.57	73.12	113.19	8.55	10.64
2002	5.89	6.57	58.43	66.84	7.64	8.18
2003	7.86	8.80	99.90	138.16	9.99	11.75
2004	7.32	7.73	85.59	91.33	9.25	9.56
2005	5.90	6.00	58.30	59.22	7.64	7.70
2006	6.82	6.94	87.16	92.40	9.34	9.61
2007	7.60	7.79	98.98	92.10	9.95	9.60
2008	6.86	10.94	78.82	173.67	8.88	13.18
2009	6.69	9.26	70.16	135.82	8.38	11.65
2010	7.03	8.98	87.01	122.76	9.33	11.08
2011	5.99	6.13	54.59	56.10	7.37	7.49
2012	6.84	7.00	74.38	77.58	8.62	8.81
2013	6.81	6.81	86.96	79.55	9.33	8.92
2014	6.46	9.60	62.60	132.22	7.91	11.50
2015	6.34	8.75	60.42	110.51	7.77	10.51
2016	6.18	6.68	55.76	64.60	7.47	8.04
2017	5.68	7.75	50.01	90.76	7.07	9.53

Using the 10 years rolling-window method, Table 3.5 shows that the error measures are mostly higher for the ICARO model than the DLNM, opposing to what is seen in Table 3.4. Actually, there are only 5 years for which the ICARO seems to preform better than the DLNM according to the MAE (1990, 1991, 1992, 1993 and 1995) and 7 years, according to MSE and RMSE (1990, 1991, 1992, 1993, 1995, 2007 and 2013).

Considering all years tested by this method, the average of MAE obtained with the DLNM is 6.807961, lower than the one obtained with the ICARO, which is 7.765035. The average of the MSE is 85.33083 for the DLNM and 95.63294 for the ICARO. And the average of all years tested RMSE is 9.062904 and 9.666991 for the DLNM and the ICARO model, respectively.

Table 3.6: Error Measures (MAE - mean absolute error; MSE - mean squared error; RMSE - root mean squared error) obtained for each year tested using 5 years *rolling-window evaluation* method of cross-validation.

Error Measures	MAE		MSE		RMSE	
Year	DLNM	ICARO	DLNM	ICARO	DLNM	ICARO
1985	6.80	6.44	75.40	66.21	8.68	8.14
1986	7.24	7.07	92.66	74.48	9.63	8.63
1987	6.28	6.78	58.08	70.54	7.62	8.40
1988	6.13	5.80	64.18	57.22	8.01	7.56
1989	6.20	6.41	65.90	60.82	8.12	7.80
1990	5.62	7.79	51.32	94.19	7.16	9.70
1991	7.24	8.52	114.95	118.58	10.72	10.89
1992	7.68	6.85	108.29	71.81	10.41	8.47
1993	7.72	7.25	131.37	79.10	11.16	8.89
1994	7.75	9.15	99.15	116.98	9.96	10.82
1995	7.09	8.05	91.62	101.45	9.57	10.07
1996	7.96	9.54	119.97	136.46	10.94	11.68
1997	6.59	12.43	88.46	201.20	8.31	14.18
1998	7.24	7.46	6059	84.91	9.34	9.21
1999	7.04	9.67	76.94	129.72	8.77	11.39
2000	6.50	7.76	74.57	87.68	8.64	9.36
2001	7.87	7.70	104.78	99.50	10.24	9.98
2002	7.11	5.64	76.48	53.84	8.75	7.34
2003	8.32	10.66	119.97	206.52	10.95	14.37
2004	7.45	10.99	88.46	170.34	9.41	13.05
2005	6.02	6.53	60.59	68.06	7.78	8.25
2006	6.96	23.02	90.91	604.69	9.53	24.59
2007	8.05	27.58	135.23	815.64	11.63	28.56
2008	6.44	23.57	65.13	620.79	8.07	24.92
2009	6.47	6.87	64.75	69.95	8.05	8.36
2010	6.22	16.77	59.48	345.35	7.71	18.58
2011	6.26	22.18	61.59	541.64	7.85	23.27
2012	6.55	7.01	66.81	73.11	8.17	8.55
2013	6.68	8.21	73.29	110.79	8.56	10.53
2014	6.21	7.57	60.83	85.90	7.80	9.27
2015	6.37	6.15	62.21	58.39	7.89	7.64
2016	6.82	8.70	66.26	108.69	8.14	10.43
2017	5.59	11.95	49.50	192.88	7.04	13.89

Finally, using the 5 years rolling-window method, each year were tested by both models estimated with the 5 previous years data, so there are error measures from 1985 until 2017 (33 years on total). Table 3.6 shows that the error measures are mostly higher for the ICARO model than the DLNM, as was found using the 10 years window. In this case, according to the MAE, there are only 8 years for which the ICARO seems to preform better than the DLNM (1985, 1986, 1988, 1992, 1993, 2001, 2002 and 2015) and 10 years, according to the MSE and RMSE (1985, 1986, 1988, 1989, 1992, 1993, 1998, 2001, 2002 and 2015).

Considering all years tested by this method, the average of MAE obtained with the DLNM is 6.862639, lower than the one obtained with the ICARO, which is 10.24484. The average of the MSE

is 81.4115 for the DLNM and 175.0734 for the ICARO. And the average of all years tested RMSE is 8.936216 and 12.02372 for the DLNM and the ICARO model, respectively.

The averages of all years tested error measures obtained through the 3 previously mentioned cross-basis methods are present in Table 3.7 for the ICARO model and in Table 3.8 for the DLNM. All values were rounded to two decimal cases.

Table 3.7: Averaged Error Measures (MAE - mean absolute error; MSE - mean squared error; RMSE - root mean squared error) obtained for all years tested by the 3 methods of cross-validation used, considering the ICARO model.

ICARO model

Cross-basis method	MAE	MSE	RMSE
Rolling-origin	6.90	77.85	8.76
Rolling-window (10 years)	7.77	95.63	9.67
Rolling-window (5 years)	10.24	175.07	12.02

Table 3.8: Averaged Error Measures (MAE - mean absolute error; MSE - mean squared error; RMSE - root mean squared error) obtained for all years tested by the 3 methods of cross-validation used, considering the DLNM.

DLNM

Cross-basis method	MAE	MSE	RMSE
Rolling-origin	7.54	159.88	11.97
Rolling-window (10 years)	6.81	85.33	9.06
Rolling-window (5 years)	6.86	81.41	8.94

Table 3.7 shows that, concerning the ICARO model, the *rolling-origin* method produces the lowest error measures. This means that the ICARO model performs better when considering all past years available data to the adjustment. For that reason, the *rolling-origin update* (with the growing length window) appears to be the best method to estimate the ICARO model parameters, being from now on the one to be considered.

Contrariwise, Table 3.8 shows that the chosen DLNM performed by the *rolling-origin* method produces higher error measures than when using the *rolling-window* method. Revealing that this last performance predictions are closer to the real observed values, than predictions using the rolling-origin method. However, the length of the window to be used is not clear. The window with 10 years produces, in average, lower MAE, but higher MSE and RMSE, than the 5 years window. Hence, to better understand the distribution of the error measures obtained with each of the *rolling-window evaluation* methods, Figures 3.38 and 3.39 are presented.

Figure 3.38 compares the distribution of the MAE obtained with the 10 years length window and the 5 years length window, while Figure 3.39 does the same for the RMSE.

From Figure 3.38, we may understand that the distribution of the MAE is quite similar for the two window dimensions. However, Figure 3.39, reveals that the RMSE distribution is lower for the 10 years length window, than the 5 years length window, even though the average of those values being higher in the first case (Table 3.8). Such may be due to some outliers present in the 10 years length window that are disturbing the mean of the error measures, increasing it.

Taking all of these information into account and also the fact that 5 years is a short period of time in which may not occur any heatwave, the 10 years *rolling-window* will be considered as the best performance to estimate the DLNM parameters.

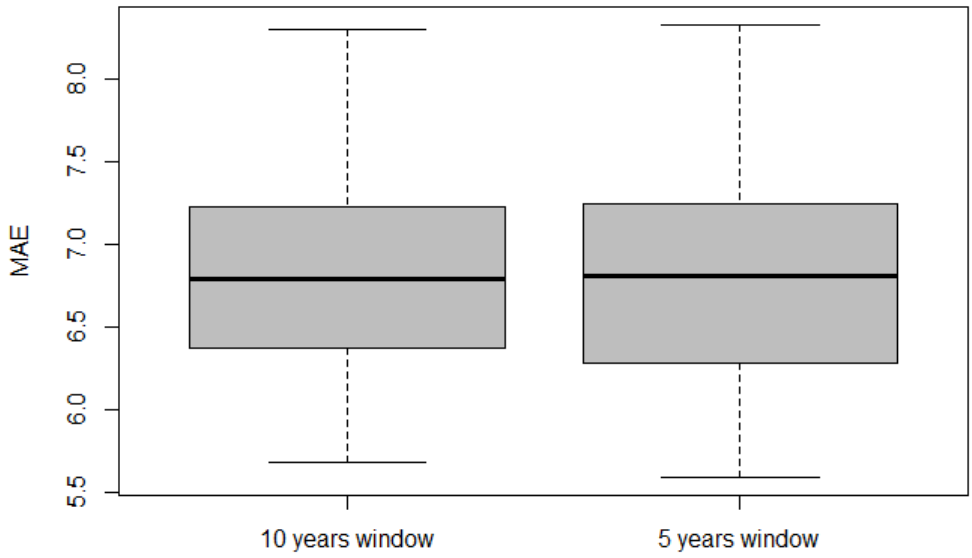

Figure 3.38: Boxplot of MAE obtained using 10 years length *rolling-window evaluation vs.* 5 years length *rolling-window evaluation*.

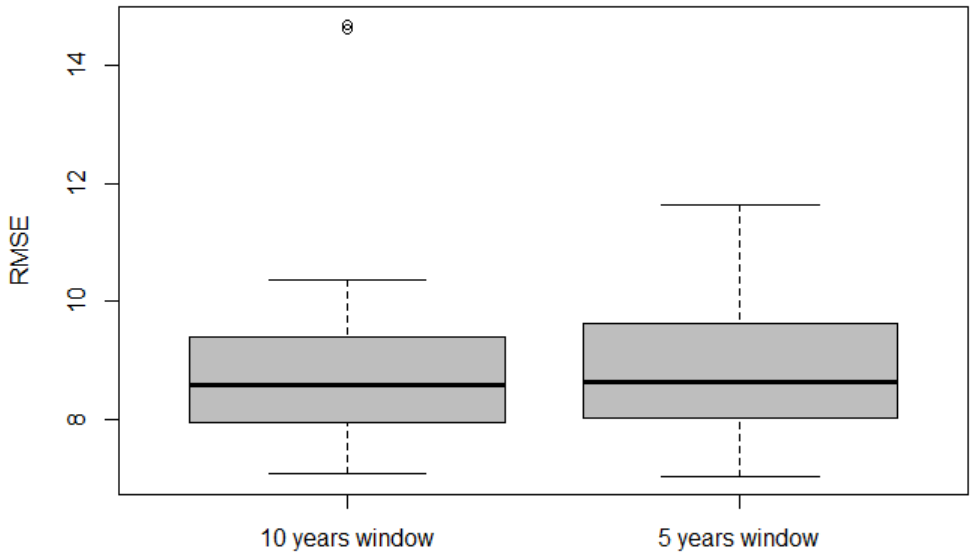

Figure 3.39: Boxplot of RMSE obtained using 10 years length *rolling-window evaluation vs.* 5 years length *rolling-window evaluation*.

3. RESULTS

In order to better compare the predictive performance of the ICARO and the DLNM, Figures 3.40 to 3.44 present the observed number of daily deaths *vs.* the predicted number of daily deaths by each model for the summer periods of years when known heatwaves with impact on mortality occurred in the district of Lisbon.

For each year tested, the ICARO model was adjusted using the data of summers from 1980 to the previous year, while the DLNM was adjusted using only the 10 previous years summer data.

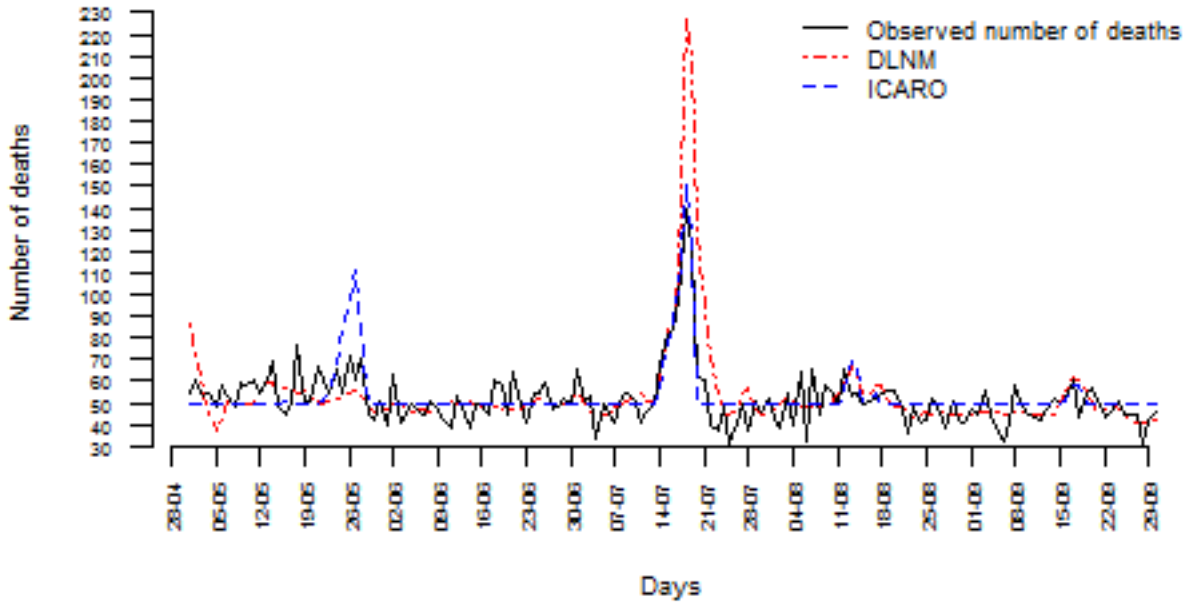

Figure 3.40: Time series of the observed number of daily deaths *vs.* the predicted number of daily deaths by the ICARO model and the DLNM - year 1991.

Figure 3.40 shows the summer of 1991 in which occurred a big heatwave with the highest peak corresponding of 139 deaths on July 18^{th}. As may be seen from the graph, both models detect and overestimate this biggest peak, with the DLNM overestimating it even more than the ICARO model. However, in the rest of this summer the behaviour of the two models were different. In May there were some increases in the observed mortality that were predicted by the DLNM, even though it underestimated their magnitude. On the other hand, ICARO only predicted an increase in the end of this month, overestimating it a lot. Also, immediately after the big heatwave in July, the DLNM predicted a new small increase in mortality, which did not really occurred, and then both models were successful to detect the two last increases of deaths in the middle of August and September.

In Figure 3.41, the focus is the long heatwave in August of 2003. During that month there were 3 consecutive peaks of deaths. All of them were detected by both models. However, while ICARO overestimated the first and last one, DLNM underestimated the magnitude of all the 3 increases. Out of this heatwave period, both one and the other models predicted almost all of the small mortality peaks, but the DLNM also underestimated most of them.

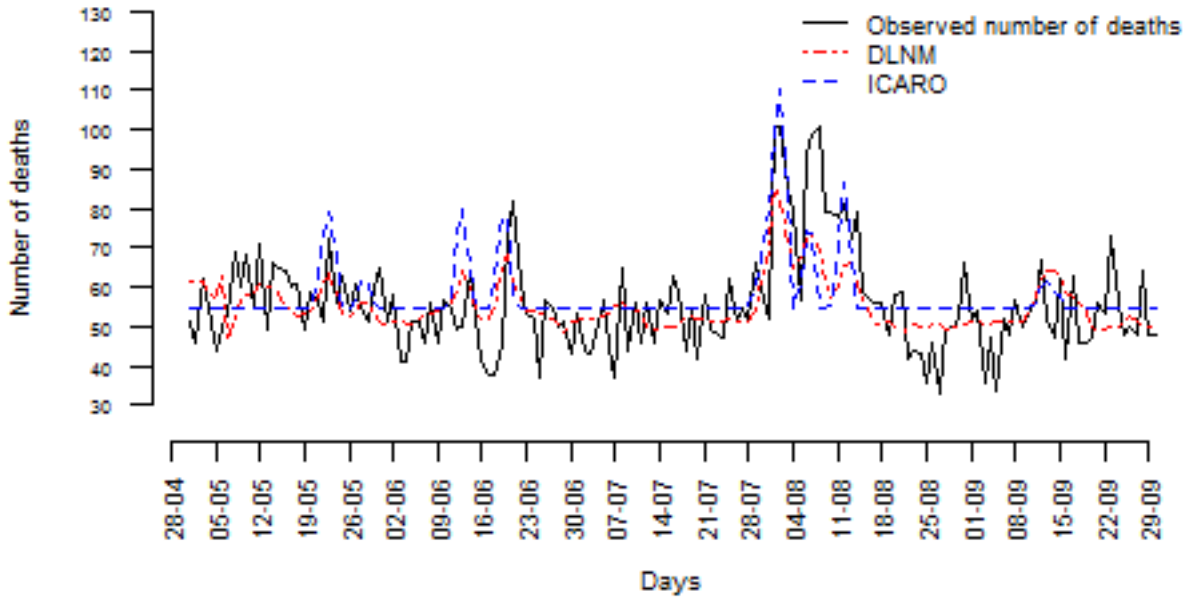

Figure 3.41: Time series of the observed number of daily deaths *vs.* the predicted number of daily deaths by the ICARO model and the DLNM - year 2003.

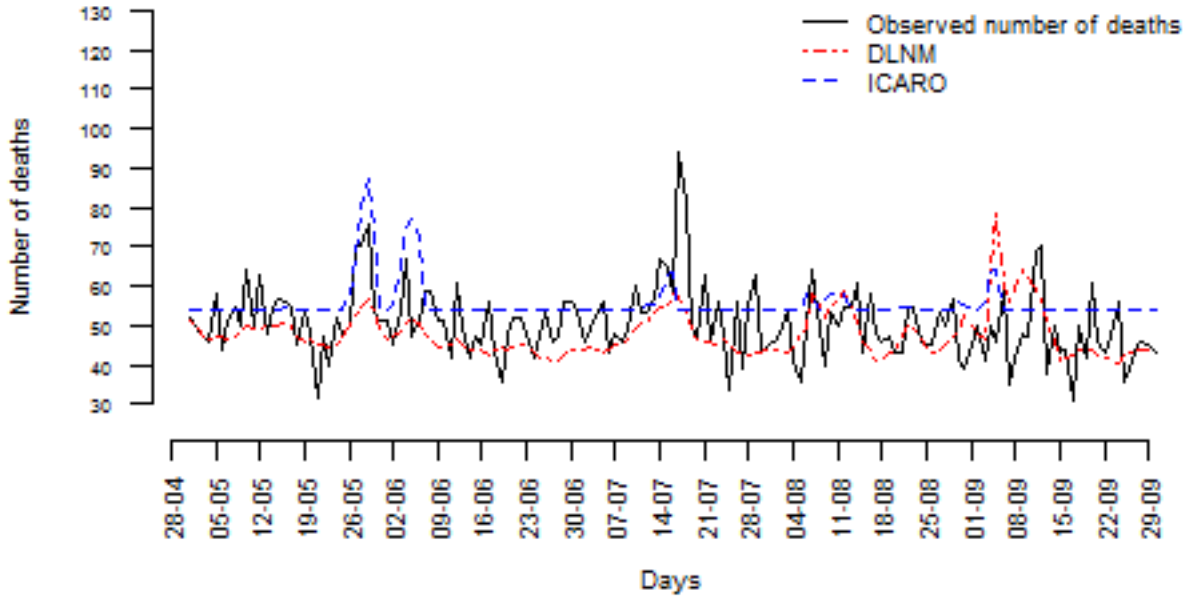

Figure 3.42: Time series of the observed number of daily deaths *vs.* the predicted number of daily deaths by the ICARO model and the DLNM - year 2006.

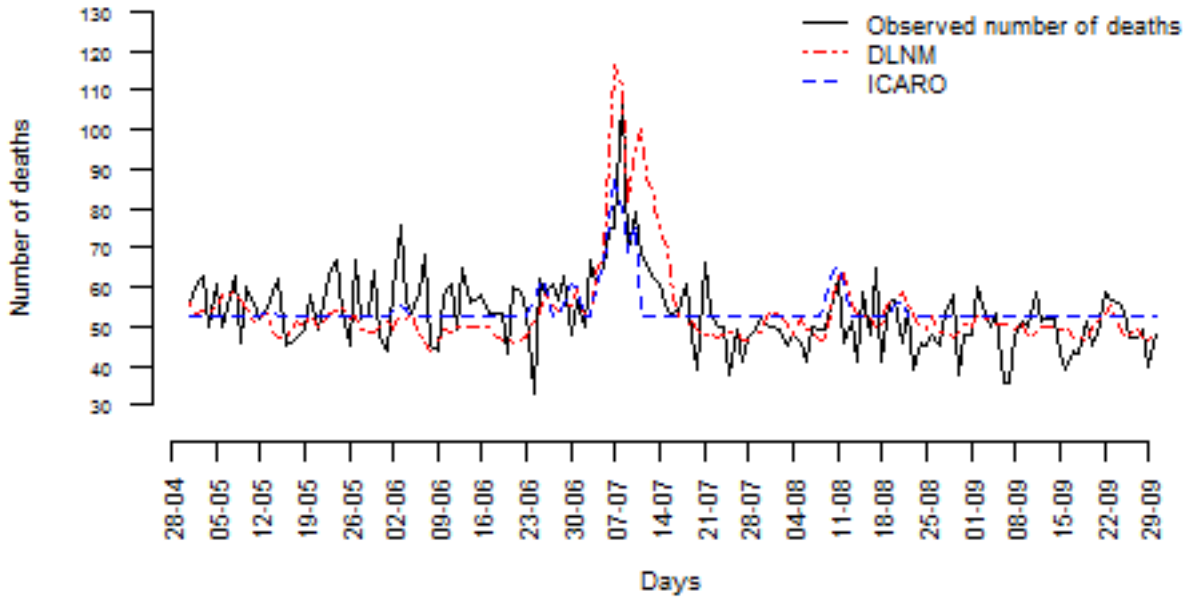

Figure 3.43: Time series of the observed number of daily deaths *vs.* the predicted number of daily deaths by the ICARO model and the DLNM - year 2013.

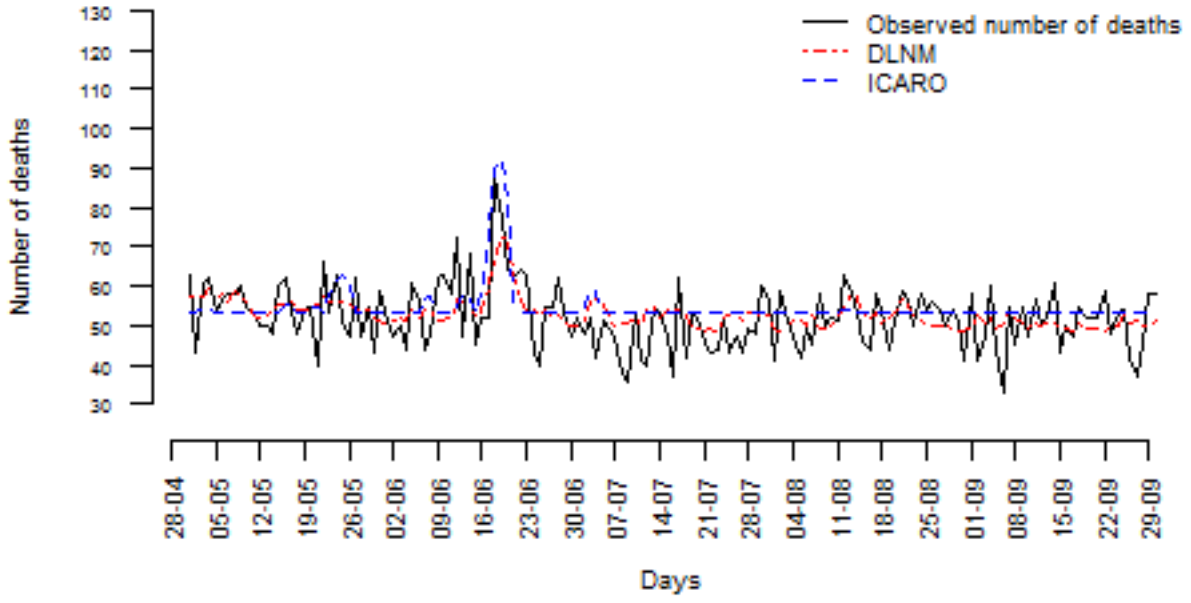

Figure 3.44: Time series of the observed number of daily deaths *vs.* the predicted number of daily deaths by the ICARO model and the DLNM - year 2017.

In the summer of 2006 (Figure 3.42), there were registered two small increases in the daily mortality in the end of May/beginning of June that were well predicted by the ICARO model, but completely missed by the DLNM. Still in this year,a quite big increase in the number of deaths occurred in the middle of July, but it was not detected by any of the models. Later, in September, the two models predicted an increase in mortality that did not occur and then DLNM seems to have predicted in advance the last peak of mortality of that summer.

Figure 3.43 shows the summer of 2013. During the month of June both ICARO and DLNM failed do predict the small death peaks observed during this month. Later, during the month of July, a heatwave with accentuated impacts over mortality occurred. The two models tested were successful to forecast the increased counts of deaths during those days and the DLNM showed an advantage by detecting the length of this period better than the ICARO.

Finally, Figure 3.44 shows the daily number of deaths observed and predicted by the two different models for the year of 2017. In this graph the focus is the heatwave that occurred in the middle of the month of June. This heatwave's impact on mortality was well predicted by the two models, but its magnitude were underestimated by the DLNM.

All these 5 time series of observed and predicted daily deaths show that the two models are good detecting the biggest peaks of mortality. However, their performance on the forecast of smoother increases of deaths are poorer and different from each other.

On the one hand, the ICARO model remains unchanged for most part of the days in all years tested, missing some of the smaller increases in the number of deaths. However, when it detects the smaller peaks, it does not underestimate their magnitude, on the contrary, it tends to overestimate them.

On the other hand, the DLNM follows better all the changes in the number of daily deaths (increases and decreases). However, it tends to underestimate most of the mortality peaks, which might be a dangerous characteristic for an alert system.

Chapter 4

Discussion

This dissertation project studied the relationship between extreme high temperatures and mortality in the district of Lisbon. Daily maximum temperatures and counts of total deaths were used to model this relationship through a distributed lag non-linear model, considering only the 5 warmer months of the year (May, June, July, August and September). The data used corresponds to the years of 1980 until 2017.

There are some similar works that have studied the impact of extreme temperatures on mortality in Lisbon using DLNM [1, 17]. However, as far as we know, this is the first time this methodology is used restricted to the summer period, considering mortality by all causes and with such a big data set (38 years).

The main goal of this work was to study if a DLNM could have a better predictive performance than the regression linear model currently used by the portuguese early heat health warning system for extreme heat with impact in mortality - ICARO. Published studies revealed that the relationship between temperature and mortality fits a non-linear shape [4, 7, 14, 17, 18, 19, 20] and that the heat adverse effect is not immediate but stays sustained in time [3, 7, 9, 14]. Therefore, DLNM seems to be an adequate modelling strategy for this subject, once it may account for both these characteristics. Also, it is a particular good solution to consider both the isolated effect of each day temperature and the sustained and accumulated effect of several days temperatures (consecutive or not) on mortality, which might be highly correlated.

Actually, DLNM were already used and proved to be an adequate statistical method to describe temperature-mortality relations by several previous works [1, 2, 3, 6, 16, 17, 29].

The core of a DLNM is the cross-basis matrix, obtained by two basis-functions that may be chosen from a broad range of modelling options [16, 21]. This framework achieves a compromise between enough complexity and flexibility to capture in detail the heat-mortality relation and enough simplicity to an easy interpretation of the results.

During this study, it was decided to try B-splines as both basis-functions, testing several combinations of the spline degree (quadratic or cubic) and the number of equally spaced knots (3, 4 or 5) for each of the two basis-functions, allowing a maximum lag of 10 days. These options were chosen based on some literature references on this theme, however it was possible to chose different types of functions (as polynomials, stratified or linear functions), or using this ones with different number and placement of knots and even consider a different number of maximum lag. It is difficult to decide what criteria to use for choosing the wright options between this wide range of alternatives [21].

I chose *a priori* information for the type of functions and maximum lag, and information criteria (QAIC) for the spline degree and number of knots, following Gasparrini *et al.* [21] approach. The

cross-basis functions used by the elected DLNM consist of a quadratic B-spline with 7 df, determined by 5 equally spaced knots with no intercept, as the basis-function used for the variable - temperature - dimension; and a natural cubic spline with 5 df, determined by 3 internal knots equally spaced in the logarithmic scale with an intercept, as the lag-basis-function. The cross-variables matrix, obtained by applying these two basis-functions over the original temperature variable, entered the model as the independent variable. The previously presented summary of the final model reveals that only one of the thirty five resulting cross-basis terms is statistically significant in the model, considering a significance level of 5 %. This suggests that, even though these model choices were considered the best from the ones tested considering QAIC, they may not be the most appropriates. In future developments, simpler functions to the cross-basis, as using less degrees of freedom or linear basis-functions, should be considered. Even so, we believe this DLNM is capable of predict the main effects that extreme heat provokes on mortality.

This model predictions of the RR show that extreme heat have an immediate increasing effect over the mortality in Lisbon that keeps sustained up to 4 days, and hits its maximum one day after the exposition to heat. These results are consistent with other studies [6, 9, 22]. The overall effect of temperature (obtained by summing up all the past lags contributions) reveals that, during summer in Lisbon, the risk of dying starts increasing for temperatures above 30 ^{o}C and the increase of the risk gets stronger for the higher temperatures. Also, maximum temperatures below 15 ^{o}C have shown to be protective.

While interpreting this results, it is important to keep in mind that the coefficients were estimated through a constrained DLNM. Hence, they reflect information from both the data and the used constrains [16]. Even so, constraining the model is important to reduce the existing collinearity caused by consecutive observations, in order to get more precise estimates [16, 35]. Furthermore, overall effects estimated from a constrained DLNM or from an unconstrained DLNM are not significantly different [16, 35].

The evaluation of the final model predictive performance through cross-validation reveals that the selected DLNM is a good predictor for all the heat-related mortality. In all the years tested, its predictions did closely follow the increases and decreases of the summer daily deaths and did capture well the highest peaks of deaths associated with heatwaves. However, it tends to underestimate some of the smaller increases in mortality.

Comparing this DLNM with the ICARO model performance highlights the fact that DLNM accounts for the "harvesting" effect, contrarily to the ICARO model. It predicts the decreases of expected mortality after a high peak of deaths, which may be due to the anticipated deaths of ill people, who would have died anyway a few days later (without the heat effect). This phenomenon is ignored by ICARO, which never predicts lower counts of deaths than the expected with the minimum mortality temperature.

More than this, in the years tested, DLNM would have underestimated the smaller peaks, but overestimated the magnitude of the heatwaves with greater impact on deaths. This features, have the advantage of producing less false alarms than ICARO, but still catch the most dangerous peaks of heat. Such might be important, in the way that false alarms overcharge the competent authorities with useless/confusing information and could cause a decrease in their commitment with the alert system. However, this characteristic might leave some of the risk increases to occur with no warning and, for that reason, with no adequate response or precautions from the society.

The assessment and comparison of the two models through cross-validation showed that ICARO has a better performance when adjusting the model with data from all years available before the year tested (since 1980) and DLNM is better adjusted using a fixed 10 years window. Thus, when comparing the two models for the same year tested, the data taken into account by each model is different. For instance, when the models were tested for the summer of 2017 (Figure 3.44), ICARO had been adjusted with 37 years data (1980 to 2016), while DLNM had been adjusted only with 10 years data (2006 to 2016). This

means that each model is forecasting the future based on different information, which makes it difficult to do an unrestricted comparison.

There is no universal criteria to choose the length of the window to be used as training data when performing cross-validation. In this work, it was empirically chosen to adjust the DLNM using data from 10 years. On the one hand, longer time periods could introduce some disturbance, because of changes in the heat-mortality relation along time due to evolution of several population conditions (age, gender, socioeconomic conditions, house conditions, use of air conditioning, clothing, behavioural habits, adaptation to climate changes, among many others). On the other hand, shorter time periods may not capture any heatwave, since it is a relatively rare phenomenon. Also, such a small specific set of the data could originate an overfitted model. Therefore it is important to try different lengths for the training data, in order to get confident that it is being used the best data set possible to adjust the model.

The approach taken in this work has some acknowledged limitations in what concerns the modelling choices. One is the use of mortality by all causes as the response variable. It may be disturbing the magnitude of the heat-mortality relationship, due to the fact that some deaths not related to extreme temperatures may be taken into account. This would be the case of accidental deaths and deaths caused by some diseases not provoked neither aggravated by heat, among others. However, once there is no widely accepted criteria for heat-related mortality [9], using the total counts of deaths seems to be the best way not to lose heat-related deaths due to miss-classification. More than that, the fact that it was used daily counts of deaths of the resident population in the district of Lisbon, instead of counts of deaths that occurred in the district, may also be a factor of confounding. Some of the deaths entering the modelling may not be related to the meteorological conditions felt in Lisbon, because some people with official residence in Lisbon may die somewhere else and still be wrongly accounted in the model.

Using only the daily maximum ambient temperatures as the independent variable may also be raising some limitations. Firstly, it does not account for other meteorological conditions, as humidity and wind, for example, which may interact with the ambient temperature effects over human health [36]. Secondly, it ignores the daily thermal amplitude, which means that two days with the same maximum temperature will be considered as causing the same impact over mortality, regardless of the temperature distribution along the day, i.e., the model will not consider, for instance, if there was a significant cooling over the night or not. Thirdly, since in Lisbon most people spend long time indoor, ambient temperature may not be the most suited to the analysis, as it does not take into account the confounding caused by house conditions and the use of air conditioning.

Even so, the maximum ambient temperature has been identified as the meteorological variable most strongly related with mortality [4] and the best determinant of populations exposure to heat [9]. Hereupon, we believe that maximum temperature may be used to generally represent the heat exposure at a population level.

Also, there are several individual risk factors for mortality associated to high temperatures that were not taken into account by this approach, in particular the advanced age [7, 9]. The relationship between heat and mortality is stronger for people over 65 years of age and infants. Therefore, the heat-mortality relationship should, ideally, be modelled separately for the different age strata.

Besides that, the fact that the portuguese population has aged in the last decades may be increasing the magnitude of the relation in study along the decades. On the other hand, the population socioeconomic conditions have been improving, which should confer some protection from heat. Hereupon, using a data set with so many years to adjust the model may introduce some disturbance, as in the previous years the risk of dying due to extreme high temperatures were different than in the most recent years, because of the evolution of all population conditions. Therefore, a natural cubic B-spline of time with 4 df was used

in the model, as an attempt to control for this possible long-term trend.

Finally, several studies showed that populations are able to adapt to different weather conditions, getting more resistant to heat along summer [7]. Hence, the use of a dynamic threshold, as in the ICARO model should be an advantage, in relation to the fixed minimum mortality temperature used as a reference to estimate the RR of heat-related death by the DLNM. Hereupon, in the DLNM proposed, a natural cubic B-spline of day of the season with 5 df was used as an attempt to work around this seasonality problem.

Despite all these acknowledged limitations, it is believed that the framework of the distributed lag non-linear model proposed in this work captures at least the main characteristics of the heat-mortality relationship. Even so, there are some possible future developments that could improve this model.

There are several risk factors for heat-related mortality that were not taken into account in this study. Namely, air pollution and other weather variables [36], individual characteristics (as age, gender, pre-existing illness, socioeconomic conditions, etc) [7, 9] and population conditions (as policy measures to reduce heat impact, for example) [10]. If some of this data could be obtained, the model used here could be upgraded to a multilevel design that would consider much more information and hopefully be more accurate.

More than that, considering climate change, it is likely that heatwaves will happen with more frequency and intensity and it is expected an increase of the adverse effects of high temperatures as well [9, 10]. Moreover, it will probably occur a geographically change on the occurrence of such phenomena [10]. For that reason, it is of major importance to study more deeply the heat-mortality relationship and all its confounding, mitigating and exacerbating variables. Gathering all the possible information about this subject is essential to keep improving the existent warning systems around the globe and create new ones, where it does not yet exist and are likely to be needed in a near future.

Bibliography

[1] Liliana Antunes, Susana Pereira Silva, Jorge Marques, Baltazar Nunes, and Sílvia Antunes. The effect of extreme cold temperatures on the risk of death in the two major portuguese cities. *International Journal of Biometeorology*, 61(1):127–135, 2017.

[2] Baltazar Nunes and Luísa Canto e Castro. Não Morrer de Calor!... Será uma Questão de Habituação? *V Congresso Anual, Sociedade Portuguesa de Estatística*, pages 457–463, 1997.

[3] Antonio Gasparrini and Ben Armstrong. The impact of heat waves on mortality. *Epidemiology (Cambridge, Mass.)*, 22(1):68–73, 2011.

[4] Jorge Marques and Sílvia Antunes. A perigosidade natural da temperatura do ar em portugal continental: a avaliação do risco na mortalidade. *Territorium*, 16:49–61, 2009.

[5] Niilo R.I. Ryti, Yuming Guo, and Jouni J.K. Jaakkola. Global association of cold spells and adverse health effects: A systematic review and meta-analysis. *Environmental Health Perspectives*, 124(1):12–22, 2016.

[6] Yuming Guo, Antonio Gasparrini, Ben Armstrong, Shanshan Li, Aurelio Tobias, Eric Lavigne, Micheline De Sousa, and Zanotti Stagliorio. Global variation in the effects of ambient temperature on mortality: a systematic evaluation. *Epidemiology*, 25(6):781–789, 2014.

[7] Paulo Nogueira and Eleonora Paixão. Models for mortality associated with heatwaves: update of the portuguese heat health warning system. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 28(4):545–562, 2008.

[8] S. Pattenden, B. Nikiforov, and B. G. Armstrong. Mortality and temperature in Sofia and London. *Journal of Epidemiology and Community Health*, 57(8):628–633, 2003.

[9] Rupa Basu and Jonathan M Samet. Relation between elevated ambient temperature and mortality: a review of the epidemiologic evidence. *Epidemiologic reviews*, 24(2):190–202, 2002.

[10] Glenn R McGregor, Pierre Bessemoulin, Kristie L Ebi, and Bettina Menne. *Heatwaves and health: guidance on warning-system development*. World Meteorological Organization Geneva, Switzerland, 2 edition, 2015.

[11] Instituto Português do Mar e da Atmosfera. Enciclopédia IPMA.PT o que é a onda de calor? `https://www.ipma.pt/pt/enciclopedia/clima/index.html?page=onda.calor.xml`, 2019. Accessed: 2019-09-18.

[12] Paulo Nogueira, Baltazar Nunes, Carlos Dias, and José Matias Falcão. Um Sistema de Vigilância e Alerta de Ondas de Calor com Efeitos na Mortalidade: o Índice Ícaro. *Revista Portuguesa de Saúde Pública*, 1(6):78–84, 1999.

[13] Carla Selada, Graça Freitas, and Ana Leça. Plano de Contigência para Temperaturas Extremas Adversas - Módulo calor 2013. Relatório final de acompanhamento e avaliação 15 de Maio a 30 de Setembro. Technical report, Direção de Serviços de Prevenção da Doença e Promoção da Saúde, DGS, 2013.

[14] Ana M Vicedo-Cabrera, Bertil Forsberg, Aurelio Tobias, Antonella Zanobetti, Joel Schwartz, Ben Armstrong, and Antonio Gasparrini. Associations of inter-and intraday temperature change with mortality. *American journal of epidemiology*, 183(4):286–293, 2016.

[15] PJ Nogueira, A Machado, E Rodrigues, B Nunes, L Sousa, M Jacinto, A Ferreira, JM Falcão, and P Ferrinho. The new automated daily mortality surveillance system. *Eurosurveillance*, 183(4):1–14, 2010.

[16] Ben Armstrong. Models for the relationship between ambient temperature and daily mortality. *Epidemiology*, 17(6):624–631, 2006.

[17] Mónica Rodrigues, Paula Santana, and Alfredo Rocha. Effects of extreme temperatures on cerebrovascular mortality in lisbon: a distributed lag non-linear model. *International journal of biometeorology*, 63(4):549–559, 2019.

[18] Brooke G Anderson and Michelle L Bell. Weather-related mortality: how heat, cold, and heat waves affect mortality in the united states. *Epidemiology (Cambridge, Mass.)*, 20(2):205–213, 2009.

[19] Jun Yang, Chun-Quan Ou, Yan Ding, Ying-Xue Zhou, and Ping-Yan Chen. Daily temperature and mortality: a study of distributed lag non-linear effect and effect modification in guangzhou. *Environmental Health*, 11(1):63, 2012.

[20] Aurelio Tobías, Ben Armstrong, and Antonio Gasparrini. Brief report: investigating uncertainty in the minimum mortality temperature: methods and application to 52 spanish cities. *Epidemiology (Cambridge, Mass.)*, 28(1):72–76, 2017.

[21] Antonio Gasparrini, Ben Armstrong, and Mike G Kenward. Distributed lag non-linear models. *Statistics in medicine*, 29(21):2224–2234, 2010.

[22] R. García-Herrera, J. Dìaz, R. M. Trigo, and E. Hernández. Extreme summer temperatures in Iberia: health impacts and associated synoptic conditions. *Annales Geophysicae*, 23(2):239–251, 2005.

[23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[24] Antonio Gasparrini. Distributed Lag Linear and Non-Linear Models in R : The Package dlnm. *Journal of Statistical Software*, 43(8):1–11, 2015.

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

[26] Antonio Gasparrini. Modeling exposure-lag-response associations with distributed lag non-linear models. *Statistics in Medicine*, 33(5):881–899, 2014.

[27] A. Gasparrini, Y. Guo, M. Hashizume, Eric Lavigne, Antonella Zanobetti, Joel Schwartz, and Aurelio Tobias. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. *The Lancet*, 386(6):369–375, 2015.

[28] Antonio Gasparrini, Fabian Scheipl, Ben Armstrong, and Michael G. Kenward. A penalized framework for distributed lag non-linear models. *Biometrics*, 73(3):938–948, 2017.

[29] Alfésio L.F. Braga, Antonella Zanobetti, and Joel Schwartz. The Effect of Weather on Respiratory and Cardiovascular Deaths in 12 U.S. Cities. *Environmental Health Perspectives*, 110(9):859–863, 2002.

[30] Jeffrey Parker. Distributed-Lag Models. Reed Collge, unpublished work, Portland.

[31] Antonio Gasparrini and Michela Leone. Attributable risk from distributed lag models. *BMC Medical Research Methodology*, 14(55):1–8, 2014.

[32] Antonio Gasparrini. Modelling lagged associations in environmental time series data: A simulation study. *Epidemiology*, 27(6):835–842, 2016.

[33] Antonio Gasparrini, Yuming Guo, Masahiro Hashizume, Patrick L Kinney, Elisaveta P Petkova, and Eric Lavigne. Temporal variation in heat–mortality associations: a multicountry study. *Environmental Health Perspectives*, 123(11):1200–1207, 2015.

[34] Christoph Bergmeir and José M Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

[35] Krishnan Bhaskaran, Antonio Gasparrini, Shakoor Hajat, Liam Smeeth, and Ben Armstrong. Time series regression studies in environmental epidemiology. *International Journal of Epidemiology*, 42(4):1187–1195, 2013.

[36] Krzysztof Błazejczyk, Gerd Jendritzky, Peter Bröde, Dusan Fiala, George Havenith, Yoram Epstein, Agnieszka Psikuta, and Bernhard Kampmann. An introduction to the Universal thermal climate index (UTCI). *Geographia Polonica*, 86(1):5–10, 2013.