Faculty of
Environmental Sciences

MASTER THESIS

Mahulena Kořistková

Comparison of methodological approaches for modeling the effect of heat alerts on mortality

Department of Water Resources and Environmental Modeling

Supervisor of the thesis: Mgr. Aleš Urban, Ph. D.

Advisor of the thesis: Veronika Huber, Ph. D.

Study programme: Environmental Modelling

Prague 2025

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Environmental Sciences

DIPLOMA THESIS ASSIGNMENT

Bc. Mahulena Kořistková

Environmental Modelling

Thesis title

Comparison of methodological approaches to modelling the protective effect of heat alerts on mortality

---

Objectives of thesis

Adaptation measures to mitigate the adverse effects of heat on human health have been implemented in many countries around the world. Among the most prominent adaptation measures are heat early warning systems (HEWSs), consisting of alerts issued based on weather forecasts. The main aim of this thesis is to review and compare selected statistical approaches to modelling the links between heat alerts and heat-related mortality in selected location.

Methodology

The study will use the access to daily counts of all-cause mortality from selected European locations, combined with temperature data from relevant weather stations, and heat alert data collected by the author of the thesis. The new data to be collected consists of daily heat alerts issued by national meteorological services based on weather forecasts for every year since the implementation of the warning system. The main task of the thesis will be to compare selected statistical approaches (such as random forest classification) to heat alert data treatment and their subsequent use in the analysis of links between heat alerts and heat-related mortality using time series regression models.

The proposed extent of the thesis

50

Keywords

statistical analysis, random forest, time series regression, heat alerts, heat-related mortality

---

Recommended information sources

Biecek P, Burzykowski T (2021) Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models (1st ed.). Chapman and Hall/CRC. ISBN 9780429027192

McGregor GR, Bessemoulin, P., Ebi K, Menne B. Heatwaves and Health: Guidance on Warning-System Development.; 2015. http://www.who.int/globalchange/publications/WMO_WHO_Heat_Health_Guidance_2015.pdf

Sera F, Armstrong B, Blangiardo M, Gasparrini A (2019) An extended mixed-effects framework for meta-analysis. Stat Med 38:5429–5444. https://doi.org/10.1002/sim.8362

Expected date of thesis defence

2024/25 SS – FZP

The Diploma Thesis Supervisor

Mgr. Aleš Urban, Ph.D.

Supervising department

Department of Water Resources and Environmental Modeling

Advisor of thesis

Dr. Veronika Huber, PhD.

Electronic approval: 12. 12. 2024

prof. Ing. Martin Hanel, Ph.D.

Head of department

Electronic approval: 12. 12. 2024

prof. RNDr. Michael Komárek, Ph.D.

Dean

Prague on 17. 02. 2025

I hereby declare that I have independently elaborated the diploma thesis with the topic of: *Comparison of methodological approaches for modelling the protective effect of heat alerts on mortality*, and that I have cited all the information sources that I used in the thesis and that are also listed at the end of the thesis in the list of used information sources.

I am aware that my diploma thesis is subject to Act No. 121/2000 Coll., on copyright, on rights related to copyright and on amendment of some acts, as amended by later regulations, particularly the provisions of Section 35(3) of the act on the use of the thesis.

I am aware that by submitting the diploma thesis I agree with its publication under Act No. 111/1998 Coll., on universities and on the change and amendments of some acts, as amended, regardless of the result of its defence.

With my own signature, I also declare that the electronic version is identical to the printed version and the data stated in the thesis has been processed in relation to the GDPR.

In ..... date ....

Author's signature

I would like to express my sincere gratitude to my supervisors, Mgr. Aleš Urban, Ph.D., and Veronika Huber, Ph.D., who kindly guided and supported me throughout this research. I am thankful for the countless hours they devoted to sharing their expertise and previous work, and allowing me to build upon the foundations they established. I am truly very grateful for their help, and their patience.

I am also grateful for the lifelong support of my late father, who was eager to see me start this journey, but unfortunately won't be there to meet me at the finish line. And lastly, I would like to thank my mother for being a great role model, and for showing me how to be resilient.

Title: Comparison of methodological approaches for modeling the effect of heat alerts on mortality

Author: Bc. Mahulena Kořistková

Department: Department of Water Resources and Environmental Modeling

Supervisor: Mgr. Aleš Urban, Ph.D.

Advisor: Veronika Huber, Ph.D.

Abstract

Background: Heat Early Warning Systems (HEWSs) are essential tools for alerting the public and healthcare professionals to forecasts of extreme heat. However, assessing their life-saving potential is challenging, as estimating excess heat-related mortality requires advanced statistical methods. This thesis applies a difference-in-differences (DID) study design to isolate the effect of implementing HEWSs by comparing mortality on heat alert and non-alert days, while accounting for unobserved factors influencing all-cause mortality. Five Polish cities were analyzed.

Methods: Mortality on heat alert and non-alert days was compared across pre- and post-implementation periods using DID. To retrospectively assign heat alert eligibility prior to HEWS implementation, two forest classifiers were tested for sensitivity. Distributed lag models (DLMs) were used to capture delayed effects of heat alerts.

Results: The method of eligibility assignment had little effect on DID estimates. However, the year 1994, marked by extreme heat, significantly influenced results by inflating pre-treatment mortality, thus affecting the apparent protective effect of HEWS. The challenge of verifying core methodological assumptions was also noted.

Conclusions: Applying DID to evaluate HEWS effectiveness requires careful sensitivity analysis to detect years with disproportionate influence. Machine learning-based eligibility assignment offered a flexible and informative approach for defining heat alert days.

Keywords: difference-in-differences, forest model, DLM, Poland

Název práce: Porovnání metedologických přístupů k modelování ochranných účinků výstrah proti vysokým teplotám na úmrtnost

Autor: Bc. Mahulena Kořistková

Katedra: Katedra vodního hospodářství a environmentálního modelování

Vedoucí práce: Mgr. Aleš Urban, Ph.D.

Konzultant: Veronika Huber, Ph.D.

Abstrakt

Motivace: Systémy včasného varování před horkem jsou důležitým nástrojem pro zvýšení povědomí o předpovědi vysokých teplot, a to jak pro širokou veřejnost, tak zdravotní personál. Kvantifikace jejich přínosu je obtížná, jelikož stanovení počtu úmrtí majících přímou souvislost s vysokými teplotami vyžaduje pokročilé statistické analýzy. V této práci je aplikována metoda rozdílů v rozdílech (difference-in-differences, DID), která umožňuje izolovat efekt výstrah pomocí srovnání úmrtnosti ve dnech s výstrahou a bez ní, a zároveň ošetřit dodatečné faktory ovlivňující celkovou úmrtnost. Metoda byla aplikována pro pět polských měst.

Metody: Pomocí DID analýzy byla porovnána úmrtnost ve dnech s výstrahou a bez ní, a to v období před i po implementaci výstražného systému. K určení dnů, kdy by byla vydána výstraha v období před implementací, byly použity dva různé klasifikační modely využívající náhodné lesy. Pro zachycení zpožděných účinků výstrah byly využity tzv. distributed lag models (DLM).

Výsledky: Volba modelu pro přiřazení dnů způsobilých k vydání výstrahy před vedrem měla zanedbatelný vliv na výsledky. Naopak rok 1994, poznamenaný extrémními vedry, měl výrazný dopad, neboť zvýšená úmrtnost v tomto roce ovlivnila zdánlivý ochranný efekt výstrah v období po jejich zavedení. Zároveň bylo poukázáno na obtížnost ověření některých klíčových metodologických předpokladů.

Závěr: Při použití metody DID je nezbytné provádět citlivostní analýzy, které odhalí roky s nepřiměřeným vlivem na výsledky. Přiřazování způsobilých dnů pomocí strojového učení se ukázalo jako přínosný způsob identifikace dnů vhodných pro vydání výstrahy.

Klíčová slova: difference-in-differences, náhodné lesy, DLM, Polsko

Contents

List of Figures

List of Tables

Introduction

The discourse surrounding changing climate and weather patterns becomes particularly prominent during the summer months, especially in regions which historically did not experience repeated heatwaves, as well as in countries where the rising temperatures have lead to an increased incidence of heat-related disasters, such as wildfires.

A literature review conducted by Hondula et al., 2015, on the impact of rising temperatures on human health, reveals a dual nature of the reported conclusions: while short-term projections suggest an increase in heat-related mortality, historical evidence indicates a gradual decline in human susceptibility to heat, suggesting the possibility of human adaptation to higher temperatures.

Although overall human acclimatization may, statistically, reduce the health impacts of heatwaves, the absolute number of heat-related casualties remains considerably high, and some studies suggest increasing heat-related mortality in the most recent years (Urban et al., 2022). According to the UN, 2022 overview of natural disasters in the 2000–2019 period, extreme temperature events (both hot and cold) were responsible for 13% of all disaster-related deaths worldwide, with 91% of these deaths attributable to heatwaves, and 88% of the extreme temperature deaths allocated in Europe. Robine et al., 2008 reported 70 000 excess deaths in 16 European countries in 2003, and Guha-Sapir et al., 2011 reported over 55 000 excess deaths from July to August 2010, as an aftermath of a heat wave and subsequent wildfires in Russia. Ballester et al., 2023 estimated around 62 000 heat-related deaths in summer 2022 in Europe, and Gallo et al., 2024 approximately 48 000 heat-related deaths in 2023.

Numerous protective and educational measures have been implemented around the world, particularly targeted at the susceptible groups, e.g., children

and the elderly. In many places, systemic measures have also been set in place: following the 2003 heatwave, multiple European regions implemented a Heat Early Warning Systems (HEWS). The purpose of HEWS is to alert the public to a high temperature forecast and to raise awareness about the potential effects of very high temperatures on the human health (WMO, 2015). Casanueva et al., 2019 provided an overview of 16 European HEWS, the trigger criteria for issuing a heat alert, implemented intervention strategies, and regional weather forecast systems, suggesting that HEWSs have been developed mainly in western and southern European countries that were hit by the 2003 heatwave most significantly.

Research on the actual effectiveness of HEWS remains fairly limited. Furthermore, there is no universally accepted methodology for assessing the effectiveness of heat alert systems; instead, most studies adopt approaches tailored to the available data, local conditions, and research goals. The objective of this thesis is to examine the application of a difference-in-differences (DID) study design to assessing the relationship between the implementation of HEWS and mortality, as employed by Benmarhnia et al., 2016 and Feldbusch, 2023. Individual steps of the analysis were refined, and different approaches to partial solutions applied in the study design were compared.

As part of a data collection effort aimed at assembling data on heat alerts and all-cause mortality in Europe, access to data for 5 Polish cities (Krakow, Lodz, Poznan, Warsaw, and Wroclaw) were facilitated. This data was chosen for this thesis because of the comparatively long time-series available and the fact that it had not been analysed previously. For ease of reading and consistency, the city names in this thesis are written without diacritical marks (e.g., “Poznań” is written as “Poznan”).

1. Background and theoretical framework

1.1 Heatwaves and Public Health Impact

According to the definition given by the World Meteorological Organization (WMO, 2025), "A heatwave can be defined as a period where local excess heat accumulates over a sequence of unusually hot days and nights." This definition implies that a region-specific approach to the topic must be taken, since what is considered "unusual" varies across continents and latitudes. The research subjects of this thesis are 5 Polish cities, therefore, a mostly Eurocentric view on the effects of heat — and their prevention — will be considered.

The United Nations Intergovernmental Panel on Climate Change (IPCC) issued their latest assessment report (IPCC, 2022) in 2022: the document provides a detailed insight into the current effects of climate change on biodiversity, water cycle, and humans across continents. It also introduces multiple detailed projections of climate change impacts, and among others, the socio-economic effects of said projection scenarios. One of the key indicators of climate change is the rising global mean temperature (GMT). The chapter on Europe (Bednar-Friedl et al., 2022) refers to studies reporting increased number of hot days in capitals, and heatwaves directly attributable to climate change: combined with the effect of urban heat island, the thermal comfort in cities is notably reduced during extreme heat events. The report highlights that countries in Western and Central Europe will, in the future, be as much at risk of extreme heatwaves as Southern Europe.

1.1.1 Physiological Effects of Heat

WHO, 2024 describes the causal factors leading to heat-related illness, which fundamentally stem from the inability to regulate internal body temperature due to a combination of factors: namely, the environmental heat stress caused by high humidity and temperature, a barrier effect of clothing, and heat gain from external sources.

Some mild heat illnesses which can be treated at home are listed in a booklet by WHO, 2011: heat rash, cramps or exhaustion, as well as moderately severe heat illnesses which should be treated at hospital: hyperthermia, hypotension or respiratory failure. Kilbourne, 1997 summarizes the effects of heat on human health and mortality, referring to multiple US-based statistics. The author lists 4 illnesses which can be recognized as a direct consequence of exposure to high environmental temperatures. Heat stroke (with the highest death-to-case ratio), heat exhaustion, heat syncope, and heat cramps: these causes of deaths will result in a death being recorded as heat-related in the death certificate. However, as Kilbourne, 1997 points out, these directly attributable casualties make up only a fraction of heat-related deaths, and the overall heat-related excess deaths can only be estimated using statistical approaches.

1.1.2 Vulnerable Groups

A publication by WHO, 2011 provides a comprehensive list of risk factors for heat illness and mortality. Infants and the elderly are particularly at risk due to reduced thermoregulatory functions. People with pre-existing cardiovascular medical conditions are also at higher risk due to the stress on heart when the body tries to cool itself. Further risk factors include social isolation, medication use, air pollution, or living in urban areas.

A thesis by Vésier, 2022 focuses on the social inequalities in heat-related mortality in the Czech Republic: multiple demographic groups were compared, and it was shown that women are at higher risk than men. Figure 1.1 further highlights some important factors contributing to heat vulnerability.

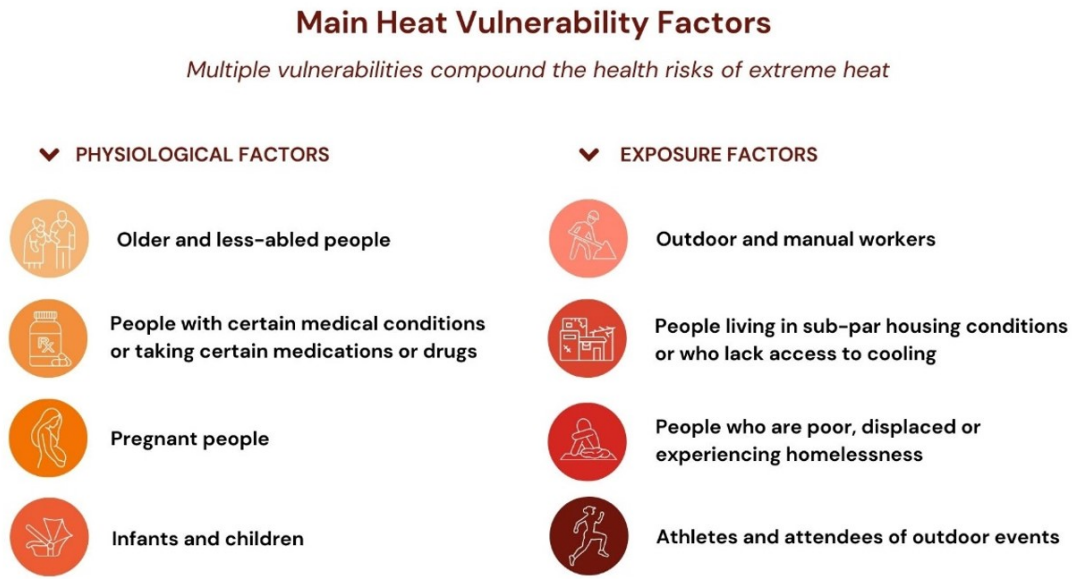


Figure 1.1 Some of the factors contributing to heat vulnerability — an infographic by WHO, 2025.

1.2 Heat-Health Action Plans and Warning Systems

In order to reduce overall heat related health risks, the World Health Organization suggests that countries should implement a structured Heat-Health Action Plan (HHAP) as a portfolio of measures to be taken before and during a heatwave (WHO, 2008). Some attributes of HHAP are:

- heat-health information plan: detailing how to coordinate key communication between interested institutions,
- a plan to increase preparedness in the health and social care system: training and education,
- conceptual urban planning: keeping heat exposure reduction in mind,
- a Heat-Health Warning System.

Overall, HHAP represents a long-term strategy aiming to decrease the impact of heat on human health. An important element of a HHAP is the Heat-Health Warning System (HHWS), or Heat Early Warning System (HEWS). In contrast

to HHAP, the warning system serves as an immediate response to incoming hot weather. It is a weather-based alert component, and WMO, 2015 provides guidelines on how to develop such system. The key components of HEWS are weather forecasts and definitions of the action triggers. An action trigger might be simply temperatures expected to exceed a certain threshold, or an arbitrary index including additional factors besides temperature contributing to thermal discomfort, like relative humidity.

There are multiple terms used to describe HEWS in literature, some of them are country-specific. Table 1.1 provides a brief summary of some common terms.

Table 1.1 A list of commonly used labels for a heat early warning system.

Shortcut	Meaning
HHWS	Heat-Health Warning System
HARS	Heat Alert Response System, Canada
HHAS	Heat-Health Alert Service, England
HWWS	Heat Wave Warning System
	Heat Watch-Warning System
HHEWS	Heat-Health Early Warning System
HEWS	Heat Early Warning System

1.3 Poland: Climate and Heat-Related Risks

1.3.1 Heatwave Trends and Characteristics

Wibig, 2017 and Wibig, 2021 identified major heatwaves which occurred in Poland since 1951. For heatwave definition, Wibig, 2017 and Wibig, 2021 adopts the one used by Huth et al., 2000: the longest continuous period during which the daily maximum air temperature is equal to or higher than 30 °C in at least three days, the mean maximum temperature during the whole heat wave is equal to or higher than 30 °C, and the maximum temperature does not drop below 25 °C during the whole period.

In Table 1.2, 11 major country-wide heatwaves recorded by at least 10 meteorological stations across Poland are listed. While the longest heatwaves were

recorded in 1992 and 1994, the majority of the heatwaves occurred at the very end of the observation period. Further analysis by Wibig, 2021 shows an increasing trend in the number of hot days registered at meteorological stations between 1990–2020, compared to the 1950–1990 period.

In the previous study, Wibig, 2017 lists sixteen longest heatwaves in Poland between 1951 and 2015, measured at any given station. Major heatwaves registered in cities subject to analysis within this thesis are given in Table 1.3. During the study period, Poznan experienced the second-highest number of heatwaves (54) among the 24 stations analyzed (Wibig, 2017). A thorough analysis of the 2015 heatwave over Poland is provided by Krzyżewska et al., 2018, using data from meteorological stations in Wroclaw, Poznan and Warsaw. The heatwave lasted for two weeks, and maximum temperatures exceeding 35 °C were recorded.

Table 1.2 The longest heatwaves observed simultaneously at least at 10 stations in Poland (1951–2019). **HW Days** is the number of heatwave days. Wibig, 2021

HW Days	Period	HW Days	Period
7	1. 8.–7. 8. 1963	8	2. 8.–9. 8. 2013
13	30. 7.–11. 8. 1992	13	3. 8.–15. 8. 2015
16	23. 7.–7. 8. 1994	7	30. 7.–5. 8. 2017
12	18. 7.–29. 7. 2006	8	28. 7.–4. 8. 2018
9	5. 7.–13. 7. 2006	7	26. 8.–1. 9. 2019
9	9. 7.–17. 7. 2010		

Table 1.3 Major heatwaves which affected Wroclaw, Lodz and Poznan between 1951 and 2015, as reported by Wibig, 2017 (abbreviated).

City	Length	Period
Wroclaw	21	22. 7.–10. 8. 1994
Lodz	17	23. 7.–08. 8. 1994
Poznan	19	24. 7.–11. 8. 1994
Poznan	18	2. 8.–19. 8. 2015

Regarding long-term trends in the monthly numbers of hot days, the statistics presented by Wibig, 2021 (see Figure 1.2) show that until 1990, the number of hot days in July always surpassed the number of hot days in August. From 1991 forward, the average number of hot days in August almost doubled, and in two decades following 1991 (1991–2020 and 2011–2019) the number of hot days in August surpassed July. Furthermore, the number of hot days in September and May is gradually increasing from 2001 onward, suggesting an increasing dispersion of hot days into the previously colder months.

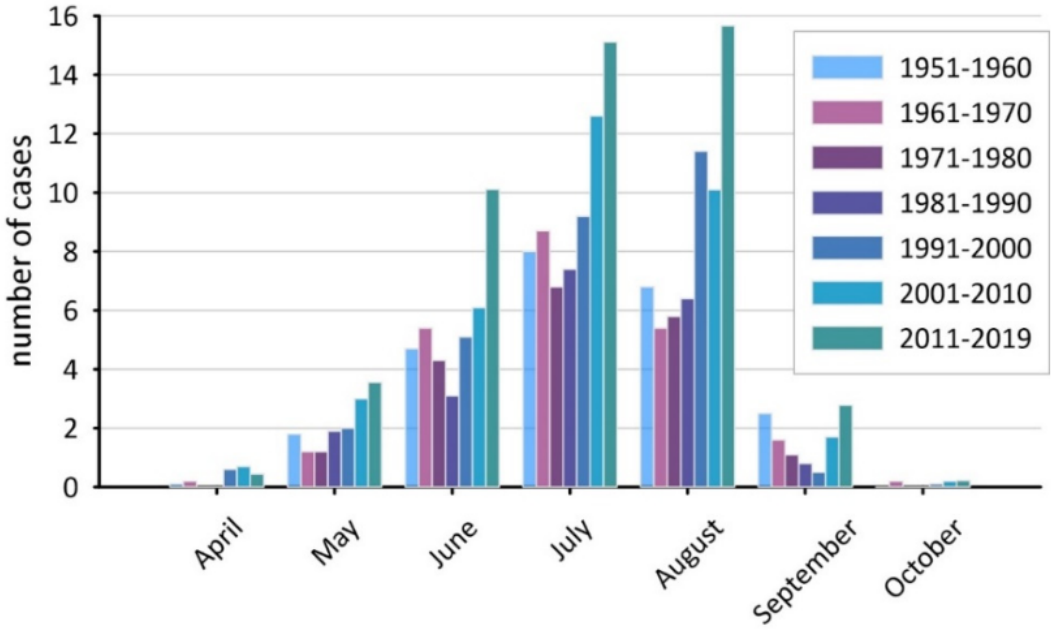


Figure 1.2 Average monthly numbers of hot days from April to October in 7 decades from 1951 to 2019, Poland. Image from Wibig, 2021.

A further evidence of the changing temperature trends and an overall increase of mean temperatures can be illustrated on the comparison of mean annual air temperatures between (1981–2010) and (1961–1990), published by Ustrnul et al., 2014. Figure 1.3 highlights areas where the increase of mean temperature exceeded 0.5 °C. Poznan, Warsaw and Wroclaw lie within a larger area with an increase in the recorded mean temperature, while Lodz and Krakow lie more on a border, which is not easily readable due to the spatial resolution of the map.

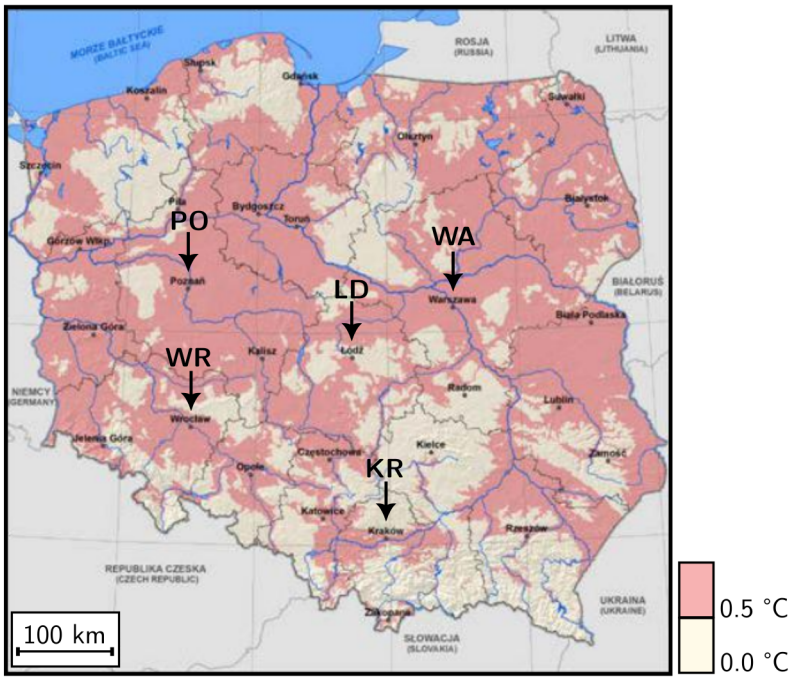


Figure 1.3 Deviations of the mean annual air temperature (1981–2010) with respect to the (1961–1990) period. Arrows point to cities which were subject to analysis in the thesis: Poznan (PO), Warsaw (WA), Lodz (LD), Wroclaw (WR) and Krakow (KR). Image from Ustrnul et al., 2014, edited.

Kuchcik et al., 2021 examines the temporal and spatial distribution of heat and cold stress conditions across Poland from 1951 to 2018. In order to quantify temperature stress, the Universal Thermal Climate Index (UTCI) is utilized. UTCI is a measure of human physiological response to thermal environment (European Environment Agency, 2020); in the study, days with UTCI exceeding 32 °C were labeled as "heat stress" days. Figure 1.4 shows the average numbers of heat stress days across Poland.

The study by Tomczyk et al., 2020 assessing the impact of heat waves on human comfort and health in Poland highlights an escalating severity of heat waves, and emphasizes the importance of developing strategies to reduce their adverse effects on human health. Similarly, Błażejczyk et al., 2022 investigated a heatwave that occurred in Poland in June 2019: the authors report a 10% increase in total mortality, and a fivefold increase in heat-related mortality compared to the previous decade. Overall, these studies along with the previous findings on changing temperature conditions in Poland highlight the necessity of public health interventions aimed at reducing heat-related harm in Poland.

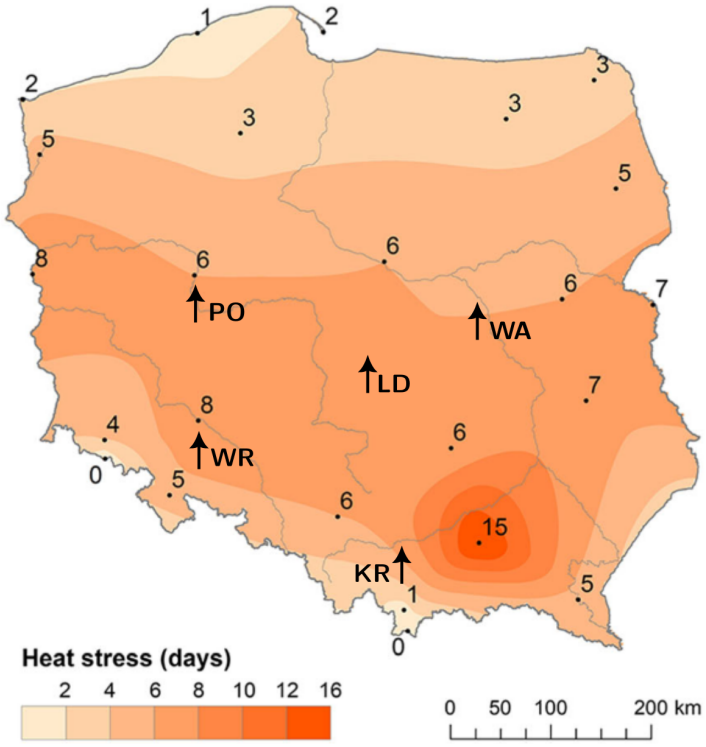


Figure 1.4 Yearly average number of heat stress days, 1951–2018. Arrows point to the approximate location of cities which were subject to analysis in the thesis: Poznan (PO), Warsaw (WA), Lodz (LD), Wroclaw (WR) and Krakow (KR). The cities had a comparable number of heat stress days, with the possible exception of Krakow, which lies in the UTCI gradient between Tarnow (inside the hotspot) and the Western Carpathains in the south. Image from Kuchcik et al., 2021, edited.

1.3.2 Crisis Preparedness

The fundamental document on emergency preparedness in Poland is the National Crisis Plan (Government of Poland, 2022), which comprises of two parts:

1. threat identification, characterization, and assessment of the risks, and
2. tasks and responsibilities for crisis management infrastructure.

The following text will focus on content relevant to heatwaves, however, the National Crisis Plan (NCP) covers numerous possible threats, including epidemiological threats, security threats and natural disasters.

In the first part of the document, the action triggers for a heat alert are defined, along with supplementary information on the effects of heat and drought. In the second part, specific bodies, offices and institutes are assigned with tasks relevant

to heat and drought emergency prevention (ahead of a crisis) and preparation (during an ongoing crisis).

The second part of NCP authorizes the Institute of Meteorology and Water Management (IMGW) to maintain a HEWS[1]. Similarly, the Ministry of Health is named as a support body, which should prepare work forces and resources to provide medical care during an ongoing heat crisis. However, even the IMGW does spread health awareness (see Figure 1.5).

The National Crisis Plan applies to all provinces (voivodships) within the country. Each province then maintains their own Province Crisis Plan (Wojewódzki Plan Zarządzania Kryzysowego), which mostly adopts the structure of NCP, adds province-specific information, and names the province-specific authorities responsible for crisis management. To the best of the author's knowledge, there is currently no dedicated national Heat-Health Action Plan, and heat-related crisis management is organized on a province level, with respect to the responsibilities defined within the Province (and by extension, the National) Crisis Plan.

1.3.3 Heat Events and the Warning System

Information about the history of HEWS were kindly provided by the Institute of Meteorology and Water Management, Warsaw.

At the time of writing the thesis, the latest update of NCP was issued in 2022. The NCP defines criteria for HEWS alert triggers, which are regularly subject to changes. However, the trigger thresholds have always been based on the forecast temperatures. Since 2011, an additional requirement on the forecast probability was implemented, setting the minimum forecast probability to 80 %. In 2017, this requirement was lowered to 70 %.

On province level, HEWS in Poland was available since 2003, and in 2009, city-specific HEWS were implemented. Furthermore, since 2018, it is possible to issue an alert on a county (powiat) level. Cities, counties and provinces all follow the alert trigger temperature thresholds defined within the National Crisis Plan.

[1]The document (Government of Poland, 2022) does not explicitly use any variation of the term "Heat Early Warning System". Instead, it tasks IMGW with "the development and transmission of warnings to public administration bodies", which is, in nature, equivalent to HEWS.

Stańczyk Jan Matejko, 1862.
Oil on canvas, 88 x 120 cm, realism, National Museum, Warsaw.

Limit physical activity

To get rid of heat excess from the body, in a relatively short timescale the circulation system raises the pulse rate and diastole the tiny blood vessels to increase blood circulation in upper skin layers. It is estimated that a rise of internal temperature by 1 degree requires an increase of pulse rate by 10 beats/minute in the first 5 minutes and then even by 20 beats/minute (the sauna experiment). Thus the heart is much more loaded than in neutral conditions. Physical activity leads to additional internal temperature growth (due to the work of muscles and the increased necessity for oxygen). Taking into account the enormous load of circulation and respiratory systems every additional physical effort can be excessive and lead to adverse effects such as heat cramps of muscles, heat stroke or even arythmia or circular system insufficiency. Even healthy persons who adapted to seasonal heat must remember that their body stamina depends on many external and internal factors, so we cannot estimate the level of risk of the particular situation. In the case of severe heat, no additional physical activity is recommended. It is better to give up running, or to shift that to the late evening hours.

MODELE IMGW-PIB
modele.imgw.pl

04

Figure 1.5 An example of a campaign run by IMGW to raise awareness on the effects of heat, providing health advice in combination with classic art. This page features a piece from the Polish painter Jan Matejko. The jester, sitting isolated and being pensive, contrasts with the ongoing ball celebrations in the background. This particular figure illustrates the need to take rest and *Limit physical activity* during excessively hot days. The historical context of the painting is, however, slightly less calm, as it reflects the loss of Smolensk during the Muscovite–Lithuanian War (CMM IMGW-PIB, 2024; Meyer, 2024).

1.4 Effectiveness of Heat Warning Systems

1.4.1 Economical Benefits

One of the first papers detailing the effectiveness of heat warning systems has been published by Ebi et al., 2003, who focused on the economical benefits of a heat wave warning system in Philadelphia, USA. The study concluded that during 3 years of operation between 1995–1998, 117 saved lives were attributable to the warning system, thus more than justifying the costs associated with running the system. Another study by Williams et al., 2022 looked into the cost effectiveness of a heat warning system in Adelaide, Australia: considering the attributable reductions in hospital admissions and ambulance call-outs, the cost of activating the HHWS was also deemed effective. More studies estimating the economical benefits of heat warning systems were published for Madrid (Chiabai et al., 2018), Madrid, London and Prague (Hunt et al., 2017), and most recently an overall assessment of socioeconomic benefits of HHWS by Rao et al., 2025.

1.4.2 Heat Warning Systems and Health Protection

Toloo et al., 2013 made a review of 15 existing studies assessing the effect of heat warning systems on saving lives and reducing harm. Out of the 15 studies, 7 addressed the effectiveness in reducing health impacts, rather than an overall response to a heat warning. Three of these seven studies inspected US cities, two were based in China, and two processed data from France, and Italy. Since then, more research has been done in the USA (Wellenius et al., 2017 and Weinberger et al., 2018), Spain (Martínez-Solanas et al., 2019) or Australia (Nitschke et al., 2016).

The study conducted by Wellenius et al., 2017 questioned the local guidelines in New England and their ability to reduce adverse heat-related health impacts. Based on the gathered evidence, the authors suggested that lowering the thresholds of local criteria for issuing a heat advisory might substantially reduce the observed heat-related mortality.

The paper by Weinberger et al., 2018 pooled results from 20 US cities, docu-

menting any decrease of mortality associated with the heat alerts issued by the US National Weather Service (NWS). Except for Philadelphia, no statistically significant association between NWS heat alerts and reduction of heat-related mortality was found.

Martínez-Solanas et al., 2019 examined the temperature-related mortality during a 20-year period in Spain, and looked into the possible impact of regional actions implemented as part of the heat health prevention plan. While some decrease in mortality attributable to extreme heat was reported, more so in regions implementing more preventative actions, the study concluded that more public actions aimed at reducing heat-related mortality are needed.

A comparison of two extreme heat events in Adelaide, Australia was published by Nitschke et al., 2016. A heat wave of 2009 was compared to that of 2014, with the latter being accompanied by a preventative programme. The effect of the preventative measures was quantified based on ambulance call-outs and emergency presentations. In the year exposed to a heatwave warning system, reduced incidence of cardiac, renal and heat-related diagnoses was observed, and the authors conclude that the reduction of specific morbidity is likely attributable to the intense preventative measures.

Most of the aforementioned studies relied on time series analysis, some employed non-linear distributed lag models as introduced by Gasparrini et al., 2010. Numerous other papers on heat alert effectiveness were published (Kovats et al., 2006; Sheridan, 2007; Chau et al., 2009; Schifano et al., 2012; Hess et al., 2018; Wu et al., 2023; and others). However, only a few of the previously cited studies utilize a quasi-experimental approach, which allows for a more rigorous assessment of the heat alert system effectiveness.

1.4.3 Studies Utilizing a DID Design

In a study by Benmarhnia et al., 2016, the quasi-experimental difference-in-differences approach was first introduced as an applicable method for determining the effect of heat alert system implementation. The study assessed the effect of a Heat Action Plan on heat-related mortality in Montreal, Quebec, focusing on the differences between various demographic groups. A positive effect of the interven-

tion was reported, particularly for the elderly and low-education neighborhoods.

In the paper by Heo et al., 2019, a slight protective effect was reported for some specific age groups in Korea, however, no evidence for decreased all-cause mortality was found. Feldbusch, 2023 used a DID methodology to assess the effects on mortality in 15 German cities. The pooled results indicated a slight protective effect of the implementation, however, for individual cities, the effectiveness varied significantly.

2. Methods

2.1 Location and Data

Figure 2.1 A map of Poland with the 5 analyzed cities — Krakow, Lodz, Poznan, Warsaw, and Wroclaw. Data source Eurostat, 2024.

Data Sources and Providers

All cause mortality data was obtained from the National Institute of Public Health – National Institute of Hygiene, Warsaw, Poland. The heat alert and meteorological data were kindly provided by the Centre of Numerical Weather Prediction, Institute of Meteorology and Water Management – National Research Institute, Warsaw, Poland.

2.1.1 Mortality

Daily all-cause mortality from 1991 to 2020 in five Polish cities was analyzed. As illustrated in the time series plot shown in Figure A.1 provided in the Appendix, there was a 2 year gap in the dataset between 1. 1. 1997 – 31. 12. 1998.

The monthly distributions of mortality (Figure A.2) follow an annual cycle, with the maximum mortality observed in the winter months: this behaviour is typical for countries with moderate climate. The distribution of daily mortality per 100 000 inhabitants shown in Figure 2.2 demonstrates a slight positive offset of relative mortality observed in Lodz, which might be explained by population age: the median of ages in Lodz was 4-5 years higher than in the other cities for at least the last 10 years (Eurostat, 2025b).

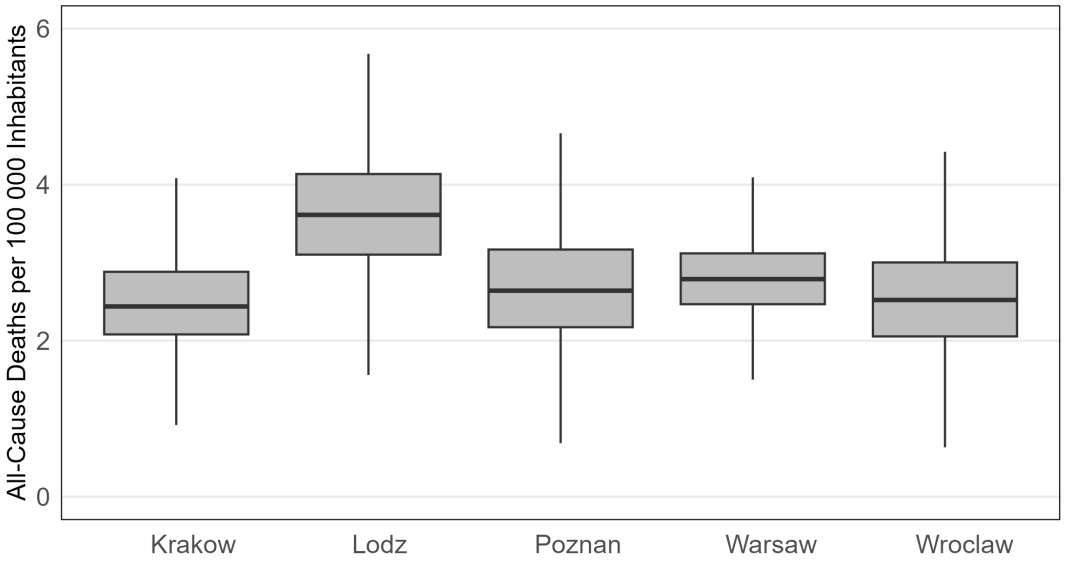


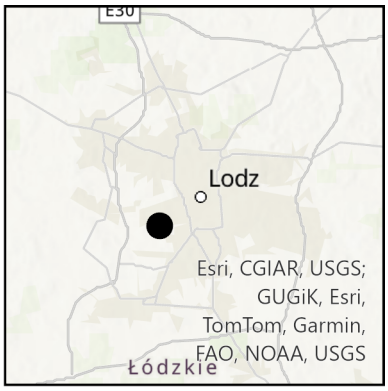
Figure 2.2 Daily mortality per 100 000 inhabitants in the 5 analyzed Polish cities. Yearly population data were taken from Eurostat, 2025a.

2.1.2 Temperature

The locations of the meteorological stations which collected the temperature data are shown in Figure 2.3, all are located in a vicinity of an airport. Table 2.1 presents a summary of the meteorological dataset attributes. Figure A.3 displays the complete city-specific time series for maximum, minimum and mean temperature, which were the only meteorological indicators included in the dataset. Figure A.4 shows the monthly temperature distributions in the warm season from May to September.

Table 2.1 A list of attributes of the meteorological dataset and their description.

Attribute	Description
`date`	Dates spanning from 1. 1. 1991 to 31. 12. 2020
`tmean`	Daily mean temperature
`tmin`	Daily minimum temperature
`tmax`	Daily maximum temperature
`cityname`	City – Krakow, Lodz, Poznan, Warsaw or Wroclaw



City	WMO
Wroclaw	12424
Poznan	12330
Warsaw	12375
Krakow	12566
Lodz	12465

Figure 2.3 Locations of the stations providing meteorological data, and their World Meteorological Organization (WMO) Codes.

2.1.3 Heat Alerts

The data were provided in an Excel file, with a separate sheet for each city containing entries defined by the attributes listed in Table 2.2. Additionally, a separate sheet detailing the development of the Polish HEWS was provided, and changes in the alert triggers relevant to the analyzed time period are summarized in Table A.1.

The original data required modifications, specifically a fragmentation into daily records. Furthermore, changes in heat alert duration based on entries modifying existing heat alerts had to be accounted for. A comprehensive description of the dataset preparation is provided in Section A.2.2.

Figure 2.4 shows the annual number of heat alert days in the five Polish cities. In 2009, no heat alert days were recorded in Lodz, Poznan and Warsaw. This could be due to HEWS not yet being implemented in those cities. Table A.3 provides a more in-depth overview of the city-specific numbers of heat alerts.

The temperatures on days with heat alerts are shown in Figure 2.5. Out of the 694 days with heat alerts, 179 (25%) days recorded maximum temperatures below 30 °C.

Table 2.2 A list of attributes of the heat alert dataset and their description.

Attribute	Description
`year`	
`city`	
`area`	In most cases, `area = city`
`warning issued`	Date + time corresponding to the moment of entering the data into the system
`warning starts`	The warning takes effect (date + time)
`warning ends`	the warning ends (date + time)
`type`	Alert categorization. This attribute took 3 values: *warning, warning change* and *cancellation of warning*
`level`	Alert level (1–3); 0 for alert cancellation

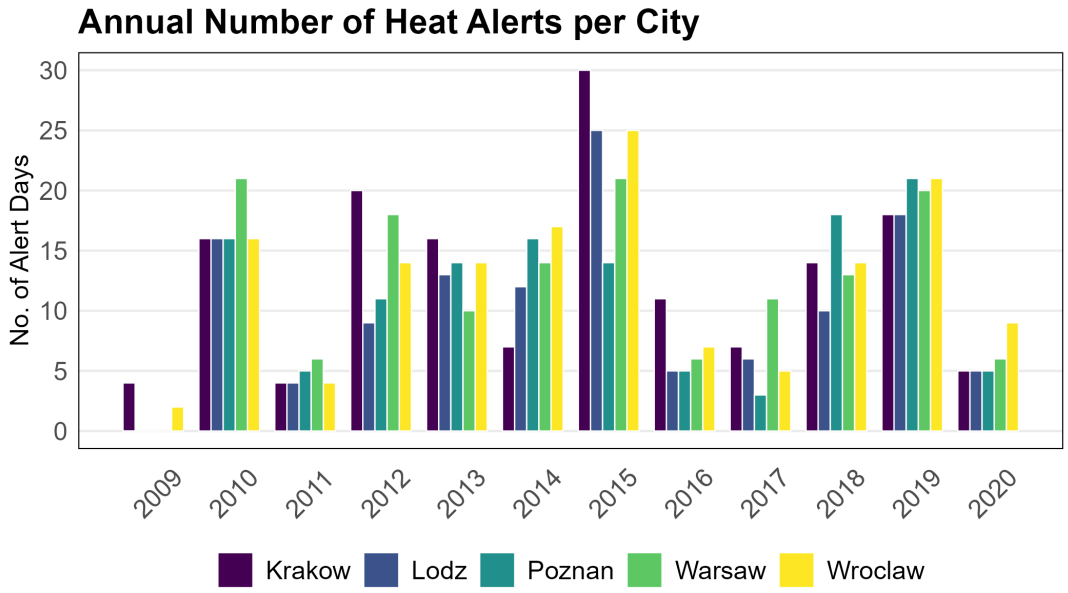


Figure 2.4 Annual number of days with an active heat alert in each city. In 2009, no city-specific heat alerts were recorded for Lodz, Poznan and Warsaw.

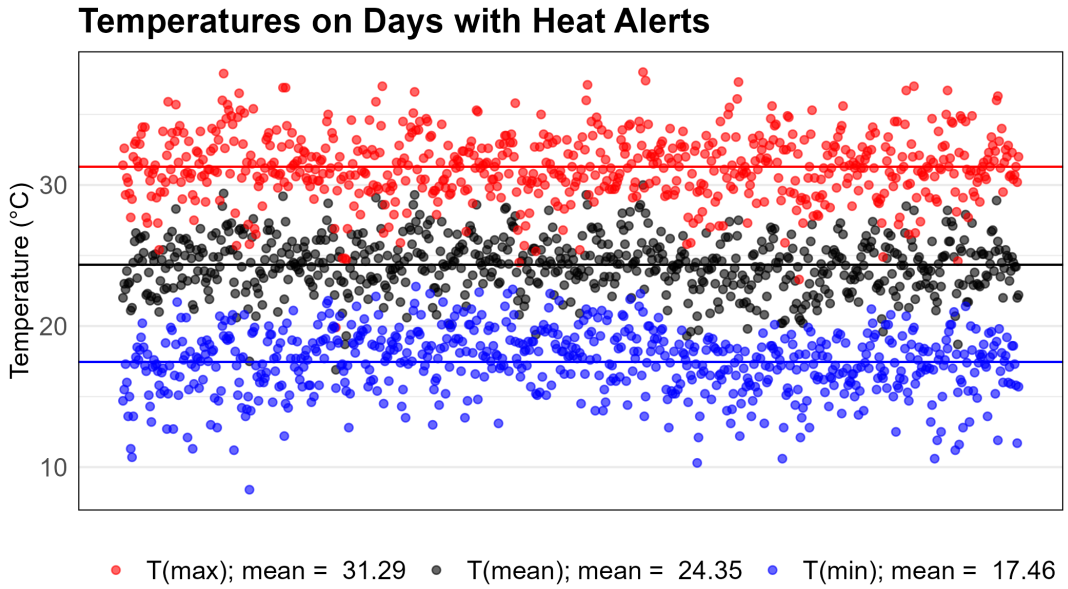


Figure 2.5 Temperatures recorded on days with heat alerts, 2009–2020. The mean values are represented by a straight line. The x-axis represents the sequence number of data points rather than specific dates.

2.2 Study Design

In order to assess the protective effect of heat alert implementation, a quasi-experimental approach was employed. A quasi-experiment is well-suited for the task at hand, as the goal is to observe the effect of treatment (i.e., heat alert implementation) on the outcome variable (mortality), but a random assignment of the treatment to selected study groups is not possible and moreover, the treatment was already applied in the past (Shadish et al., 2002).

Since the treatment is not assigned randomly, the study groups might differ in ways beyond the presence of the treatment. The goal of a quasi-experimental study is to rule out the possible alternative explanations of the observed effect (i.e., the alternative hypotheses) (Shadish et al., 2002). An applicable quasi-experimental method is the Difference-in-Differences (DID) approach, which was already utilized for heat alert effect quantification in the past (Benmarhnia et al., 2016; Feldbusch, 2023).

The analysis was restricted to months from May to September, as these were the months with the earliest and latest occurrences of recorded heat alerts. Furthermore, the levels of individual heat alerts were not taken into account and the presence of heat alert on any given day was represented by a binary attribute.

Figure 2.6 presents a visual guide on the individual steps of the analysis. The following text goes into detail explaining each step.
All statistical analyses were performed using the R software (R Core Team, 2021).

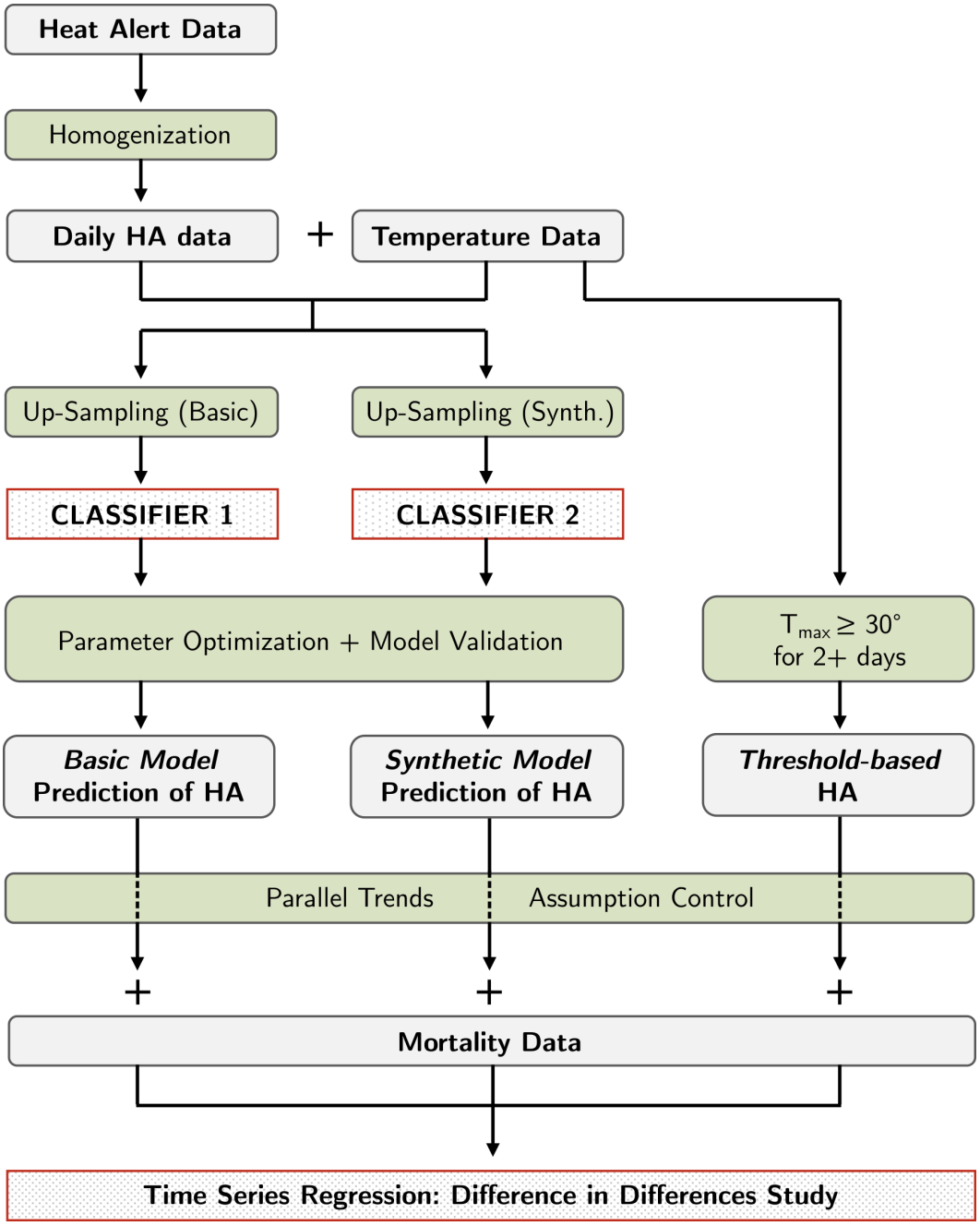


Figure 2.6 A flowchart highlighting key steps of the analysis. Grey boxes represent data, green boxes represent data processing, and boxes with an orange outline represent models. First, days preceding HEWS implementation were classified as either eligible or non-eligible for a heat alert: heat alert data coupled with temperature data, both from the post-implementation period, were treated as to address the imbalance between the number of days with a heat alert, and days without. Two methods of data up-sampling were employed, thus creating two training datasets, and two forest model classifiers were trained and optimized. The two models were applied to pre-implementation data to predict heat alert eligibility. For comparison purposes, pre-implementation data were also classified based on the Polish fundamental alert trigger. Second, the parallel trends assumption of a DID study was investigated. Third, a time series regression employing the DID formula was employed.

2.2.1 Difference-in-Differences Approach

The DID method quantifies the difference in the differences between trends observed for separate groups in separate time periods, split by application of some treatment. In principle, multiple groups and multiple time periods can be observed. However, since this study falls into the category of DID scenarios with two groups and two time periods, the following text will assume this simple scenario, too.

A straightforward way of deducing the magnitude of a treatment effect in an intervention group could be a simple observation of the values for the outcome variable in the pre-intervention and post-intervention periods. However, this approach would fail to reflect any external, unmeasured variables — the inclusion of a comparison group makes it possible to control for influences common to both groups (Wing et al., 2018).

The unmeasured variables which can be accounted for by using the comparison group are restricted to

- time-invariant group attributes (i.e., attribute specific to a single group which does not change over time), and
- time-varying, group invariant attributes (i.e., attribute common to both groups, which develops over time) (Wing et al., 2018).

These restrictions are called the *common* trends assumption. In other words, the assumption states that in the absence of treatment, the difference between the treatment and control groups remains constant over time. If satisfied, the time series of the outcome variable for both groups will be parallel — therefore, this requirement is also referred to as the *parallel* trends assumption (Wing et al., 2018). Effects which do not satisfy the above mentioned requirements must be controlled for. Figure 2.7 shows the principle of assessing the intervention effect via DID, illustrating the need for satisfying the parallel trends assumption.

The parallel trends assumption is fundamental to the validity of DID analyses. However, this assumption is inherently untestable because it pertains to a counterfactual scenario: we cannot observe what would have happened to the treatment

group had they not received the treatment. Nevertheless, pre-treatment trend analyses can be conducted as a form of diagnostic check. Examining whether the treatment and control groups exhibited similar trends prior to intervention can provide evidence supporting the plausibility of the parallel trends assumption. It is important to note that even if pre-treatment trends appear parallel, this does not conclusively confirm that the assumption holds post-treatment. In order to support the validity of the assumptions, robustness checks of the DID models should be performed (Rambachan et al., 2023; Roth, 2022). In order to explore the validity of the common trends assumption in the daily mortality data, a generalized linear model was employed (see Section B.5).

DID Equation

Let $g \in \{1, 2\}$ be a group index and $t \in \{1, 2\}$ a time index, where

$$g = \begin{cases} 1 & \text{for control group,} \\ 2 & \text{for intervention group,} \end{cases}$$

$$t = \begin{cases} 1 & \text{for pre-intervention period,} \\ 2 & \text{for post-intervention period.} \end{cases}$$

Then, let T_g be a binary indicator for observations in the intervention group:

$$T_g = \begin{cases} 1 \Leftrightarrow g = 2 \\ 0 \Leftrightarrow g = 1 \end{cases} \tag{2.1}$$

and P_t a binary indicator for observations for the post-intervention time period:

$$P_t = \begin{cases} 1 \Leftrightarrow t = 2 \\ 0 \Leftrightarrow t = 1 \end{cases} \tag{2.2}$$

Then, the basic DID estimating equation for this two-group, two-period scenario,

where Y_{gt} denotes the outcome in group g and period t, can be expressed as

$$Y_{gt} = \beta_0 + \beta_1 T_g + \beta_2 P_t + \beta_3 (T_g \times P_t) + \epsilon_{gt} \quad \text{(Wing et al., 2018)}, \tag{2.3}$$

where β_0 is the baseline outcome, β_1 is the group effect (i.e., the permanent difference between groups), β_2 is the time trend (i.e., a change that both groups are subject to), β_3 is the treatment effect — also called the DID estimator — and ϵ_{gt} is the error term.

Equation (2.3) describes a simple model using only predictors (2.1) and (2.2). However, to control for confounding effects affecting daily mortality, supplementary terms for the day of week, month, etc. were introduced, extending the relationship for the outcome variable as follows:

$$Y_{gt} = \beta_0 + \beta_1 T_g + \beta_2 P_t + \beta_3 (T_g \times P_t) + \sum_{i=4}^{k} \beta_i X_i + \epsilon_{gt}, \tag{2.4}$$

where β_i is the effect of confounding variable X_i.

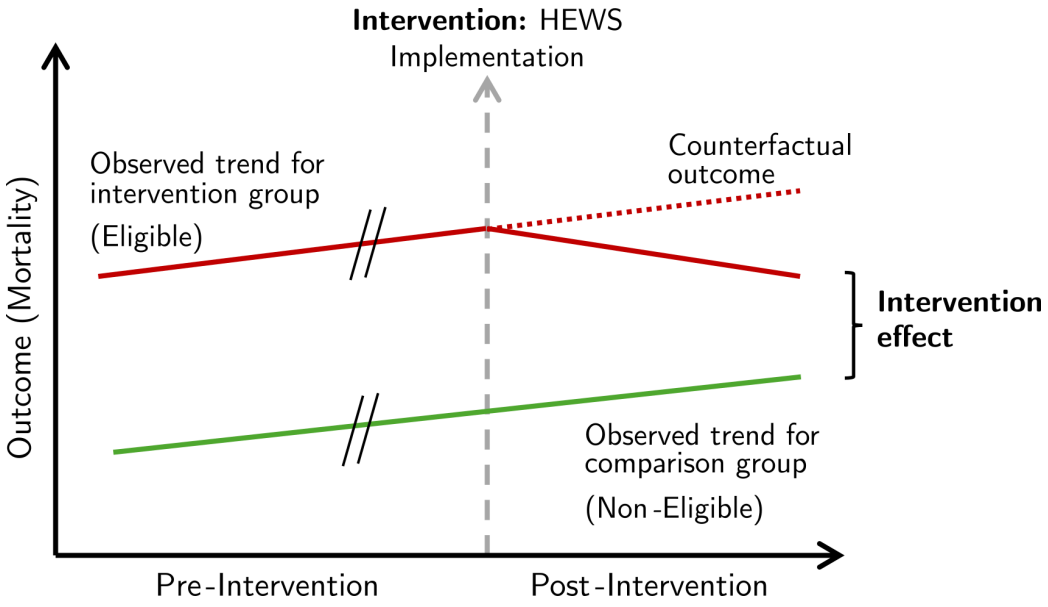


Figure 2.7 The premise of a difference-in-differences quasi-experimental study is to compare the difference between trends observed in the intervention group versus the comparison group in both the pre-intervention and post-intervention period. The difference in the trends — after controlling for influences other than the intervention — can then be attributed to the intervention effect. The pre-intervention trends are assumed to be parallel. The compared study groups are the *Eligible* and *Non-Eligible* for a heat alert.

Intervention Group in a Heat Alert Study

In the economy sector, DID was utilized by Card et al., 1994, who proved that increasing minimum wage in New Jersey led to an increase in fast-food sector employment. However, unlike fast food sector employees, the intervention group in a heat alert study (i.e., days with heat alerts) did not actually exist prior to HEWS implementation. However, days from the pre-intervention period which would be candidates for a heat alert[1] may still be recognized: Benmarhnia et al., 2016 labels these days as *eligible* days. The methodology of sorting pre-implementation days into *eligible* and *non-eligible* groups is discussed in the following Section.

2.3 Pre-Implementation Heat Alert Days

The heat alert intervention group consists of two distinct subsets: first, the definitive heat alert days in the post-implementation period, and second, days potentially *eligible* for a heat alert in the pre-implementation period. In order to produce the binary indicator defined in Equation (2.1), two approaches for determining eligibility in the pre-implementation period were considered:

1. marking eligible days based on known heat alert criteria, and
2. training a binary classifier based on post-intervention data.

Neither of these methods can guarantee completely reliable results. In practice, heat alerts are issued based on weather forecast models, and retrospective assignment of heat alerts using empirical meteorological data would not accurately reflect the forecast conditions available in the days preceding an anticipated heat event. Furthermore, the measured data might originate from a remote meteorological station, i.e., from an airport. These values might not be representative of the conditions within the city.

Applying the criteria-based approach to the post-intervention period failed to reproduce a significant portion of actual heat alert days (as detailed in Section B.1). Therefore, the alternative method, training a binary classifier, was employed. Nevertheless, the criteria-based approach was employed for comparison purposes in the final time regression model.

[1] If there was a HEWS in place at that time.

2.3.1 Classifier Training Data

The model intended for labeling *eligible* days should reliably identify days which would be candidates for a heat alert. In the context of binary classification, heat alert days will be designated as the **positive class**, whereas days without a heat alert as the **negative class**. In the post-implementation period, restricted to the warm months from May to September, a total of 694 days with a heat alert were recorded, versus 8486 days without — thus bringing the positive to negative class ratio of the training dataset close to 0.08.

The challenges of training a classifier on an imbalanced dataset are illustrated by Lantz, 2019 using an example of identifying birth defects. To paraphrase the example, a classifier that correctly classifies 92 out of 100 days as either eligible or not eligible for a heat alert achieves an accuracy of 92 %. However, given that only 8 out of 100 days are actually eligible for a heat alert, this particular classifier might not be the most reliable tool for identifying heat alert days. Therefore, this imbalance should be addressed.

The two general approaches to dataset balancing are **up-sampling** (increasing the number of instances in the less represented class) and **down-sampling** (decreasing the number of instances in the more represented class). Since down-sampling would significantly downsize the training dataset, up-sampling was employed and the following approaches were compared:

1. **Instance Replication**

   A straightforward way to enhance the less represented class is to replicate its instances. This method will be referred to as "basic" up-sampling. This method was previously employed by Feldbusch, 2023.

2. **Generating Synthetic Data**

   Duplicating random instances might accidentally amplify the importance of certain outliers. Additionally,no new variability is introduced that way. However, utilizing distribution-based algorithms to create new data points addresses both issues. This approach will be referred to as "synthetic" up-sampling.

Before applying any up-sampling procedures, outliers within the distribution

of maximum temperatures on heat alert days were identified and reassigned: ten data points from the positive class with T_{max} values below the interquartile range were reclassified as non-heat alert days. Additional details, and a list of these data points, are provided in Section B.3.

Basic Up-Sampling

Instances of the positive class were randomly duplicated to match the number of negative class instances. The input for training the classifier was a union of the original positive class, the duplicated positive class instances, and the negative class.

Synthetic Up-Sampling Using SMOTE

A Synthetic Minority Over-Sampling Technique (SMOTE) algorithm from the `smotefamily` package (Siriseriwan, 2024) available from CRAN was employed for enhancing the minority class. SMOTE uses a K-nearest neighbor (KNN) algorithm to generate new, synthetic data to match the number of the majority class instances.

2.3.2 Forest model classifier

As input variables for the classification model, daily values of mean, minimum and maximum temperatures were used, as well as their 2- and 3-day rolling means. Similarly to the work done by Feldbusch, 2023, a forest model was chosen as a best fit for the limited training data on hand. A comprehensive list of all variables used throughout the analysis is provided in Table A.4.

Forest models

Forest models are ensemble type models introduced by Leo Breiman (Breiman, 2001). They consist of multiple decision trees, and each tree is trained using a different random subset of the training data. The final prediction of the ensemble model is decided either by voting (in case of classifiers), or averaging the predictions of individual decision trees.

The process of choosing a random subset of training data for each decision tree is called as bagging, short for *bootstrap aggregating*. The subsets are chosen uniformly and with replacement, i.e., with duplicates. Therefore, each tree is built from a different subset (subsample) of the training data (Flach, 2012).

Each subsample leaves out about one third of the training data points (Flach, 2012). These are called out-of-bag (OOB) instances, and they can be utilized as testing data for estimating the so-called OOB error. A data point is passed to each decision tree which did not use this particular point for its training — i.e., to each tree for which this particular point is an OOB instance. Once a decision has been made by all trees for which the data point is an OOB instance, a vote is cast. Once all data points have been evaluated, the OOB error estimate of the forest model can be derived from its prediction accuracy on OOB instances (Breiman, 2001). With a growing number of trees, the OOB error of the forest model will gradually converge.

The number of trees in a forest model does not affect the value to which the OOB error converges. Therefore, some random forest implementations (e.g., the one in the `caret` package) do not make it possible to adjust this parameter, and usually fix the number of trees to 500. The number of trees in this analysis was left default, i.e., 500 trees.

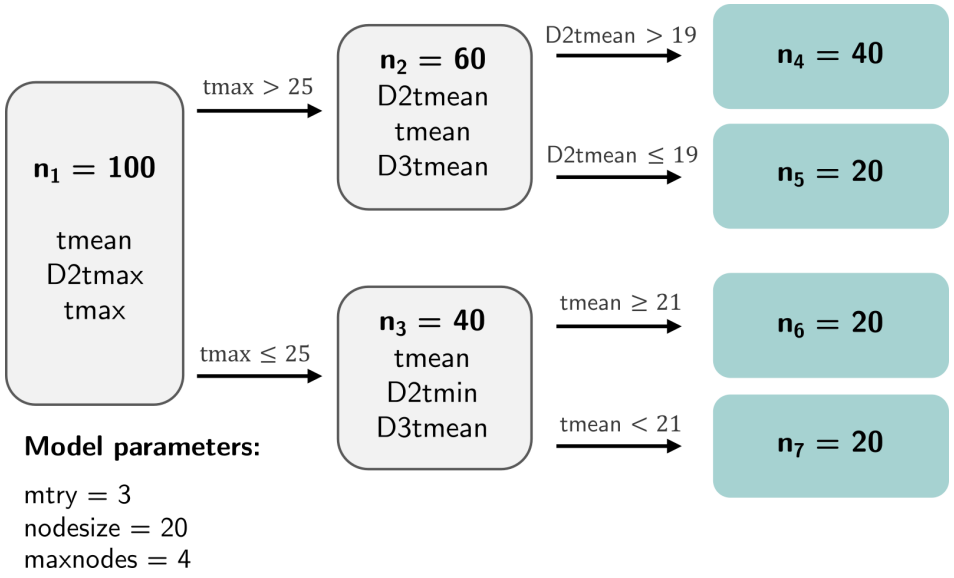


Figure 2.8 A graphic representation of a decision tree. A certain number of instances is passed to the parent node $\mathbf{n_1}$. Then, **mtry** predictor variables are randomly selected as candidates for a splitting criterion, applied to the $\mathbf{n_1}$ instances. Whichever combination of variable and its value results in the most homogeneous subsets gets selected, and the instances are split accordingly. This process repeats at each node, until one of the tree growth limiting criteria is met. Either the maximum number of nodes, **maxnodes**, is reached, or there are **nodesize** instances left in the node.

Hyperparameter Optimization

The parameters subject to optimization in a random forest model primarily define the structure of individual decision trees. Three hyperparameters, `nodesize`, `maxnodes` and `mtry` were optimized.

`Nodesize` sets the minimum number of instances each node should work with. Figure 2.8 illustrates a simplified decision tree, where a subset of $n_t = 100$ instances from the original training dataset is used to train the tree. This subset is gradually split at each node, based on a threshold value of a selected predictor, until `nodesize` instances are left.

Aside from `nodesize`, `maxnodes` is another growth-limiting criterion defining the maximum permissible number of terminal nodes, and node splitting stops once either of the conditions is met. In the example depicted in Figure 2.8, nodes $n_5 - n_7$ can not further split (they already reached the limit for `nodesize`), and therefore become terminal nodes (or leaf nodes). If `nodesize` was the only criterion, node n_4 could still split. However, this is prevented by the `maxnodes` parameter constraining the maximum number of terminal nodes. Splitting the n_4

node would create 5 terminal nodes, therefore breaking the `maxnodes` criterion.

At each node, a random subset of `mtry` predictor variables is selected. Out of the `mtry` variables, the optimal splitting variable is selected using an objective improvement criterion, which prefers a variable that increases homogeneity in the following nodes: this is further discussed in Section 2.3.2. Additionally, the node also selects the optimal threshold value of the chosen variable. The `caret` package provides a `train` function that performs automated tuning of the `mtry` parameter using k-fold cross-validation (CV).

Optimization Algorithm

For each combination of `nodesize` and `maxnodes` in the hyperparameter space, the optimal value for `mtry` was chosen by `caret`'s `train` function by maximizing the k-fold cross-validation accuracy.

During k-fold cross-validation, the training data are split into k random parts called folds. Then, k models are built. Each model is trained on data from which a different fold has been withdrawn, and the data points within the fold are used for model evaluation (testing). Then, the average accuracy across all folds is reported — this will be referred to as the CV accuracy. A 10-fold cross validation was employed (Lantz, 2019, Kuhn et al., 2013). The optimization process is described in Algorithm 1.

Algorithm 1 Tuning the forest model hyperparameters: looping through models with different tree settings. For evaluation purposes, the prediction accuracy on the up-sampled train and test sets were reported, as well as the CV accuracy for the `mtry` parameter.

for `ns` $\in \{1, 10, 20, 30, 50\}$ **do**
  for `mn` $\in \{5, 10, 50, 100, 200, 500, 1000, 2000, 5000, 7500, 10000\}$ **do**
    Train *model_forest* with:
      - **mtry**: Selected from $\{3, 4, 5\}$ using 10-fold cross-validation
      - **nodesize**: `ns`
      - **maxnodes**: `mn`

    Evaluate Training Performance:
      - Report training set prediction accuracy
      - Report selected `mtry` value
      - Report cross-validation accuracy for selected `mtry`

    Test Model:
      - Predict test set using *model_forest*
      - Report test set prediction accuracy

    Apply Model to Original Data:
      - Predict original, imbalanced data using *model_forest*
      - Extract the p-value
  end for
end for

Feature importance

Two common approaches to assessing feature importance in forest models is either evaluating their impact on model accuracy, or their ability to produce homogeneous subsamples during node splitting. Homogeneity equals class purity: each tree aims to have samples belonging to a single class at its terminal nodes.

The change in sample homogeneity before and after split is measured via the Gini index, which can be attributed to each node, and compared in-between them. When choosing the optimal splitting variable, the one causing the biggest decrease of Gini index, i.e., causing the maximum possible decrease of impurity, is selected (Kuhn et al., 2013).

Once all the trees are grown, each feature can be assessed with regard to the mean decrease in Gini index it made throughout the whole forest: this is referred to as the Mean Decrease Gini (MDG). The other method of assessing feature importance is calculating the Mean Decrease Accuracy (MDA). MDA quantifies how much the model accuracy drops when a specific feature is permuted while keeping everything else the same (Liaw et al., 2002).

2.4 Time-series Regression

The time series regression was based on the architecture of non-linear Distributed Lag Models (DLM), and implemented through the `dlnm` package in R (Gasparrini, 2011; accessible via CRAN). While the package was created with Distributed Lag Non-linear Models (DLNMs) in mind, as introduced by Gasparrini et al., 2010, allowing for a *simultaneous* non-linear and lagged behaviour of the predictor, the option of adding non-linearity to the lagged predictors was not utilized. Nevertheless, the `dlnm` package provides a useful framework for working with lagged predictors.

In relation to modeling HEWS effects, DLNMs were employed by Martínez-Solanas et al., 2019, who used them to link mortality counts with temperature. Their study, however, did not follow a DID design. In the DID study by Benmarhnia et al., 2016, the time series analysis was performed using standard Poisson models, while Feldbusch, 2023 employed quasi-Poisson models – however, neither study used DLNMs or DLMs to any extent. To the best of the author's knowledge, no previous studies have specifically used DLM models for a DID study of heat alert effectiveness. In this study, the lagged effect is applied for the heat alert itself (i.e., for the indicator defined in Equation (2.1)).

Foundations of DLMs

The key features of non-linear DLMs are:

- **Additivity:** the effects of predictors can be expressed as smooth functions, allowing for capturing non-linear behaviour.
- **Lag:** adding a lagged effect to a variable makes it possible to capture delayed effects (e.g., increased mortality following a heatwave).

Additionally, instead of directly linking the response variable to the predictors, the response variable was transformed using a so-called link function, allowing for error distributions other than normal. This is the principle of generalized models, and it allows for considering a Poisson distribution of a positive, discrete outcome variable like mortality. A Poisson non-linear DLM can be expressed using the following equations:

$$g(\mu) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_k(X_k) \tag{2.5a}$$

$$g(\cdot) = \log(\cdot) \tag{2.5b}$$

$$\mu = \mathbb{E}[Y|X_1,\, X_2,\, \ldots, X_k] \tag{2.5c}$$

where $g(\cdot)$ is the response variable link function, α is the offset, f_1, f_2, $\ldots$, f_k are arbitrary smooth functions and μ is the mean of the Poisson-distributed response variable Y, i.e., the average expected value of the outcome variable Y based on predictors X_1, X_2, $\ldots, X_k$ (Hastie et al., 2009).

Let Z be a predictor with delayed effect on outcome variable Y. At a given time t, the contribution of Z to the outcome Y can be expressed as follows:

$$Y_t \sim \sum_{\ell=0}^{L} \xi_\ell \cdot Z_{t-\ell}, \tag{2.6}$$

where ℓ indicates the lag, $Z_{t-\ell}$ is the value of Z at time $t-\ell$, L is the maximum lag and ξ_ℓ is a coefficient proportional to the impact $Z_{t-\ell}$ has on the outcome variable Y_t.

2.4.1 General Model Design

The extended form of a DID model presented in Equation (2.4) can be adapted to include parameters represented by smooth functions as in Equation (2.5a), as well as lagged variables as specified in Equation (2.6). The only variable for which a lagged effect will be considered is the group indicator T_g (2.1): this setup allows heat alerts to affect mortality in the follow-up days. For the pre-implementation period, T_g was determined via the forest models. For the post-implementation period, T_g reflected the actual issued heat alerts.

For clarity, the subscripts g and t in the indicators T_g and P_t may be omitted, given that their primary purpose is to improve readability. Using a log link function (2.5b) for the expected mean value of outcome variable Y at a given time t, the final model can be expressed as

$$g(\mu_t) = \alpha + \sum_{\ell=0}^{L} \xi_\ell \cdot T_{t-\ell} + \beta \cdot P + \sum_{\ell=0}^{L} \zeta_\ell \cdot (T_{t-\ell} \times P) + \sum_{i=1}^{m} f_i(X_i) + \sum_{i=m+1}^{k} \gamma_i \cdot X_i + \epsilon. \quad (2.7)$$

The effect of the treatment will be expressed as cumulative lag effect. Since the link function is a natural logarithm, the overall effect, or the relative risk associated with the treatment, is given by

$$\zeta = \exp\left(\sum_{\ell=0}^{L} \zeta_\ell\right). \quad (2.8)$$

Since $\zeta < 1 \Leftrightarrow \sum \zeta_\ell < 0$, values of ζ below 1 indicate a protective effect of the treatment.

Quasi-Poisson Models

A Poisson model is applicable if the distribution of the mean expected outcome satisfies the Poisson relationship for variance given by

$$\text{var}(Y) = \mathbb{E}[Y|X_1, X_2, \ldots X_k]. \quad (2.9)$$

However, real-life data are usually more dispersed, thus making the variance (2.9) greater than the mean expected value, and the error estimates of a simple

Poisson model are not representative of the overdispersed data (McCullagh et al., 1989). Quasi-Poisson models relax the variance-mean equality by introducing a dispersion parameter[2] σ^2:

$$\text{var}(Y) = \sigma^2 \cdot \mathbb{E}[Y|X_1,\, X_2,\, \ldots\, X_k], \tag{2.10}$$

while the log link function (2.5b) remains the same. Therefore, the mean expected value also remains the same, as well as the coefficients of predictor variables $X_1,\, X_2,\, \ldots\, X_k$. Introducing the dispersion parameter does, however, expand the standard error of both the prediction and the coefficients. Therefore, quasi-Poisson distribution was utilized as a conservative choice without performing additional analysis of the overdispersion extent.

2.4.2 Implemented models

Multiple models were implemented and compared, all adhering to the general relationship defined by Equation (2.7), and the effect was observed a 3-day lag. The components describing the seasonal variations were adapted from Feldbusch, 2023, and the general formula for the seasonal component was

$$ns(DoS,\, df = 4) : year + ns\left(date,\, df = round\left(\frac{n_Y}{10}\right)\right) + DoW \tag{2.11}$$

where ns denotes a natural spline with df degrees of freedom, DoS is Day of Season, n_Y is the number of years in the whole modeled period and DoW is the Day of Week. All models were implemented using the `glm` function native to R.

Table 2.3 presents a summary of the implemented models. Models 0-1 differed in the way the T_g indicator for group membership in the pre-implementation period was assessed. In Model 2, threshold criteria for heat alert day eligibility assignment were applied, mirroring the approach taken by Benmarhnia et al., 2016. Model 3 excluded the 2003–2008 period, where heat alerts were issued on province-level instead of city-level. Additionally, year 2009 was omitted too, as there are were no data on heat alerts in 3 out of the 5 cities (see Figure 2.4).

[2]Some literature denotes the dispersion parameter as Φ. However, adhering to the notation given by McCullagh et al., 1989 emphasizes the non-negativity of the parameter.

Models 4-5 excluded years 1994 and 2015 in order to observe the sensitivity of the DID estimator on years with major heatwaves (see Tables 1.2 and 1.3).

Table 2.3 Characteristics of four models compared within the DID study design. $\mathbf{T_g}$ is the method of group affiliation assessment and **X** denotes the predictor formula. All models contain the fundamental DID formula (2.3), with $\mathbf{T_g}$ having a 3-day lagged effect. Time trends are represented either as natural splines $\mathbf{ns_{df}(predictor)}$, where **df** denotes the degrees of freedom, or as linear predictors. **DoS** stands for Day of Season, **Yr** for year, **M** for Month and **DoW** for Day of Week.

$$\mathbf{X \sim lag_3(T_g) + lag_3(T_g) \times P_t + P_t + ns_4(DoS):Yr + ns_2(date) + DoW}$$

0	$\mathbf{T_g}$	Synth. Up-sampling (< 2009) + Real Data (≥ 2009)
1	$\mathbf{T_g}$	Basic Up-sampling (< 2009) + Real Data (≥ 2009)
2	$\mathbf{T_g}$	Temp. Threshold (< 2009) + Real Data (≥ 2009)
	Note	Threshold: $T_{max} \geq 30°C$ for 2+ days
3	$\mathbf{T_g}$	Synth. Up-sampling (< 2003) + Real Data (> 2009)
	Note	Excluded 2003–2009
4	$\mathbf{T_g}$	Synth. Up-sampling (< 2009) + Real Data (≥ 2009)
	Note	Excluded 1994
5	$\mathbf{T_g}$	Synth. Up-sampling (< 2009) + Real Data (≥ 2009)
	Note	Excluded 2015

3. Results

3.1 Eligibility Classifiers

3.1.1 Data Up-Sampling

The effects of heat alert data treatment prior to training the classifier model are illustrated in Figure 3.1. Points corresponding to heat alert days are shown within the $T_{max} \times T_{min}$ space, as those variables are most indicative of the conditions on heat alert days.

The "Original dataset" panel in Figure 3.1 shows the original heat alert days. Following outlier removal described in Section B.3, days with temperatures on the lower end of the distribution were reduced, as shown in the top right panel. The bottom panels show the effects of the two up-sampling techniques - with "basic" up-sampling, the reduced transparency of the points indicates their multiplication. On the other hand, the K-nearest neighbor algorithm generated new points in areas potentially corresponding to heat alert days, while not overfilling areas between outliers.

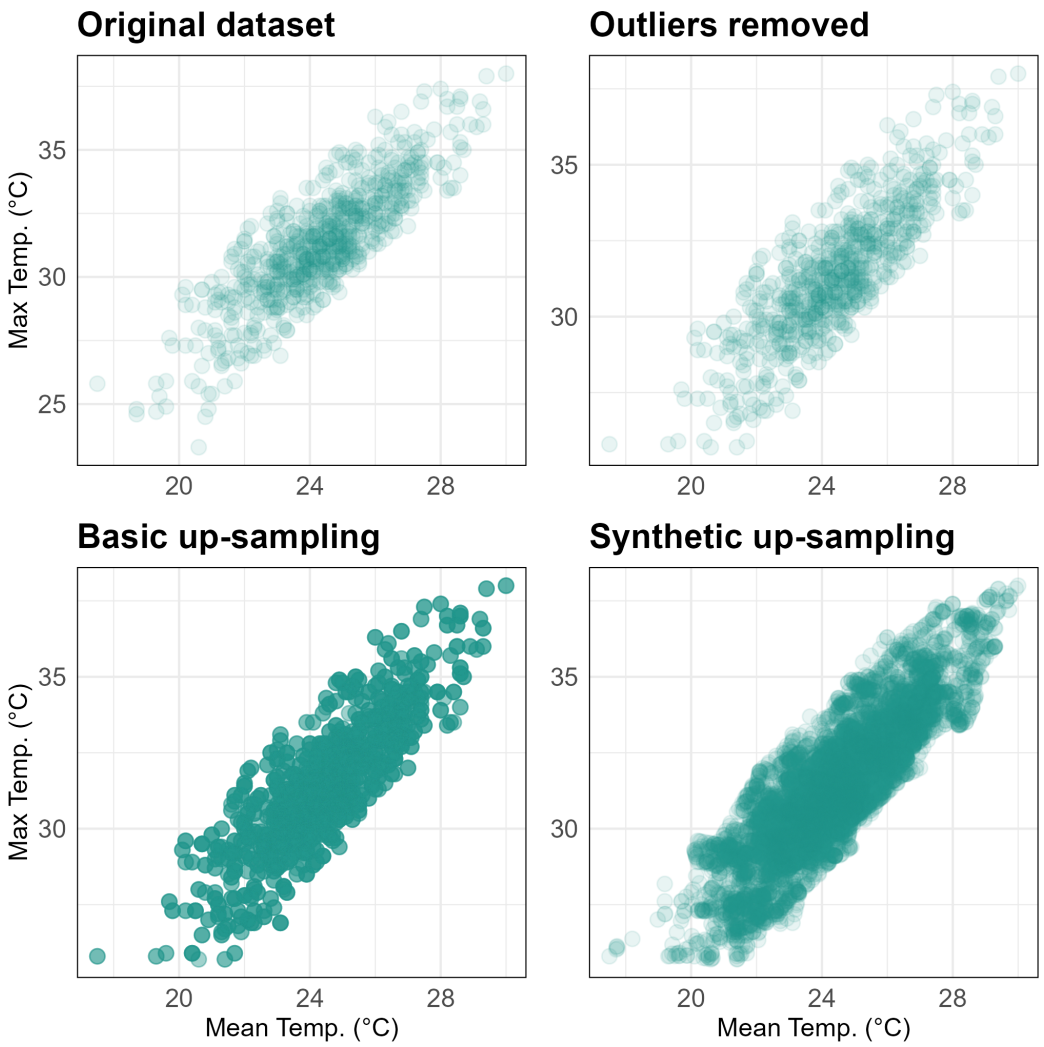


Figure 3.1 Days with heat alerts, shown in the $\mathbf{T_{max}} \times \mathbf{T_{min}}$ space. From the original dataset, outliers were removed based on IQR criterion. To counteract the imbalance between days with heat alerts and days without, two up-sampling techniques were employed. **"Basic"** up-sampling stands for a dataset where minority instances were duplicated. **"Synthetic"** up-sampling applied a KNN algorithm to generate new data points.

3.1.2 Parameter Tuning

The hyperparameter search space was defined as follows:

$$(\mathbb{P}_{ns} \times \mathbb{P}_{mn}) \times \mathbb{P}_{mtry} \tag{3.1a}$$

$$\mathbb{P}_{ns} = \{1,\ 10,\ 20,\ 30,\ 50\} \tag{3.1b}$$

$$\mathbb{P}_{mn} = \{5,\ 10,\ 50,\ 100,\ 200,\ 500,\ 1000,\ 2000,\ 5000,\ 7500,\ 10000\} \tag{3.1c}$$

$$\mathbb{P}_{mtry} = \{3,\ 4,\ 5\} \tag{3.1d}$$

where $\mathbb{P}_{ns}$ are `nodesize` values, $\mathbb{P}_{mn}$ are `maxnodes` values and $\mathbb{P}_{mtry}$ are `mtry` values. The optimal `mtry` value is bound to each combination of `nodesize` and `maxnodes`: therefore, this parameter is not subject to standalone tuning.

The tuning Algorithm 1 was applied separately to find optimal parameters for both classifiers: the "basic" one, and the "synthetic" one. The values of `maxnodes` and `nodesize` were selected based on model performance on train and test data. Parameter `mtry` was determined for each (`nodesize` $\times$ `maxnodes`) combination based on 10-fold cross validation.

The possible value of `mtry` was picked from a limited set: $\{3,\ 4,\ 5\}$, with 3 being the default in the `randomForest` (Liaw et al., 2002) function, and 5 just exceeding half of the total number of predictor variables. Most models benefited from higher `mtry`, with only a few opting for `mtry` = 3. The optimal values of `mtry` in the $\mathbb{P}_{ns} \times \mathbb{P}_{mn}$ space are listed in Table B.2 (Basic up-sampling) and Table B.3 (Synthetic up-sampling).

Figures B.2 ("Basic" up-sampling) and B.3 ("Synthetic" up-sampling) show heatmaps of Train, Test and CV accuracy in the hyperparameter space. The CV accuracy applies for the corresponding value of `mtry` given in Tables B.2 ("Basic" up-sampling) and B.3 ("Synthetic" up-sampling).

Given the same combination of `maxnodes` and `nodesize`, the model based on synthetic up-sampling yields overall lower accuracy. For lower `maxnodes` (100 and less), both models are insensitive to the value of `nodesize`. For `nodesize` equal to 1, i.e., when splitting was limited solely by the maximum permitted number of terminal nodes, train accuracy of 100 % was reached by both models

for `maxnodes` exceeding 500.

Finally, the "basic" model exhibits higher sensitivity to `maxnodes`: its maximum difference in train accuracy, given constant `nodesize`, was 5.7 %, while for the "synthetic" model, the difference in accuracy between 5 and 10 000 leaf nodes was 3.9 %

All models throughout the hyperparameter space were validated using the original (imbalanced) post-implementation daily data: each model was used to predict the true heat alerts, and p-values of the predictions were recorded. Tables B.4 and B.5 list the p-values of the "basic" models and the "synthetic" models, respectively. All models where the tree growth was limited to 10 and less leaf nodes provided unreliable results, with p-values equal to 1. Equally, both models improved with 50 leaf nodes, and the p-values decreased with increasing `nodesize`. For the same combination of hyperparameters, the "synthetic" models yielded lower p-values than the "basic" ones. Furthermore, the decrease of p-value with increasing `nodesize` was much steeper for the "synthetic" models.

Based on the aforementioned indicators, the same setting of parameters was chosen for both models:

$$\texttt{nodesize} = 30 \tag{3.2a}$$

$$\texttt{maxnodes} = 500 \tag{3.2b}$$

$$\texttt{mtry} = 5 \tag{3.2c}$$

with `mtry` being inherent to the combination of `nodesize` and `maxnodes`. Further discussion on the model selection is provided in the Discussion.

3.1.3 Model Performance

Figures 3.2 ("basic" model) and 3.3 ("synthetic" model) show the progress of forest model error rates with increasing number of trees. The converged error rates of correctly assigning days without heat alerts are similar, while the error rate of correctly assigning heat alert days is near zero for the "basic" model and about 2 % for the "synthetic" model: however, the "synthetic" model seems to converge slightly earlier than the "basic" one. The out-of-box error is approximately 3 % for the "basic" model, and almost 4 % for the "synthetic" one.

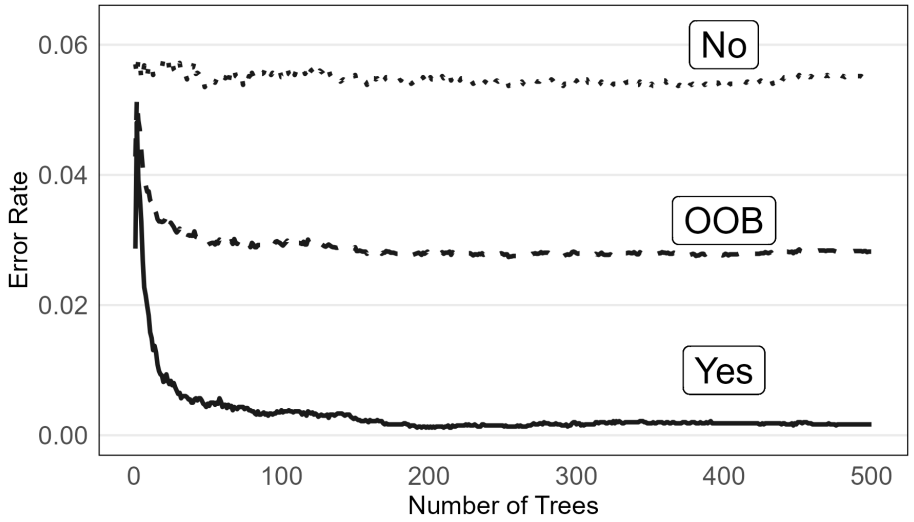


Figure 3.2 Class-specific and OOB errors of random forest classifier trained on data up-sampled using the **basic** approach.

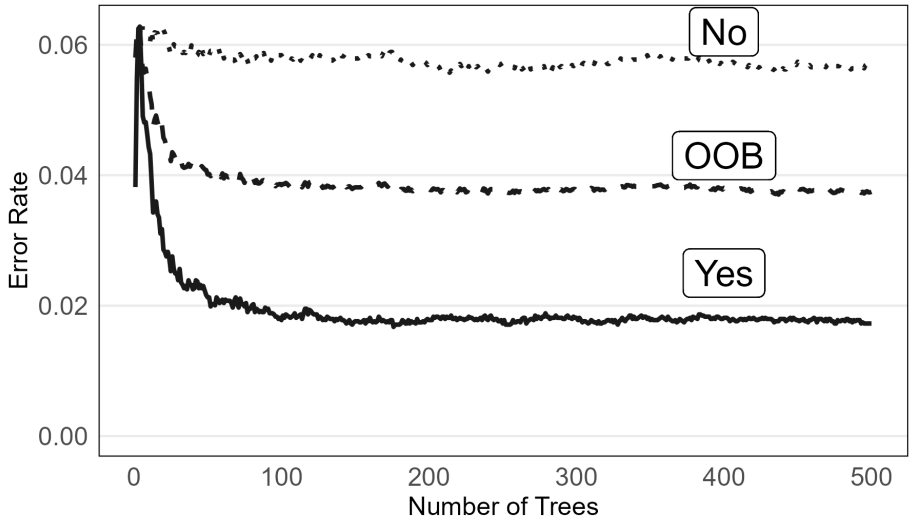


Figure 3.3 Class-specific and OOB errors of random forest classifier trained on data up-sampled using the **synthetic** approach.

Figures 3.4 ("basic" model) and 3.5 ("synthetic" model) present the variable importance plots, featuring the MDA and MDG of all predictors. Description of the variables is provided in Table A.4. The "basic" model exhibited higher MDG for its most impurity reducing variable, T_{mean}, with the second most reducing variable, T_{max}, scoring approximately half the value for T_{mean}. Regarding accuracy, the most important variable was also T_{mean}, with T_{max} second in place.

The order of the first two most important variables was switched for the "synthetic" model. Using T_{max} to split nodes was most beneficial in reducing impurity, and the model was most sensitive to its permutation.

The daily minimum temperatures and their 2- and 3-day rolling means exhibited the lowest ability to produce pure nodes on split in both models, however, permuting the values of T_{min} and $D3T_{min}$ had major effect on the accuracy of both models.

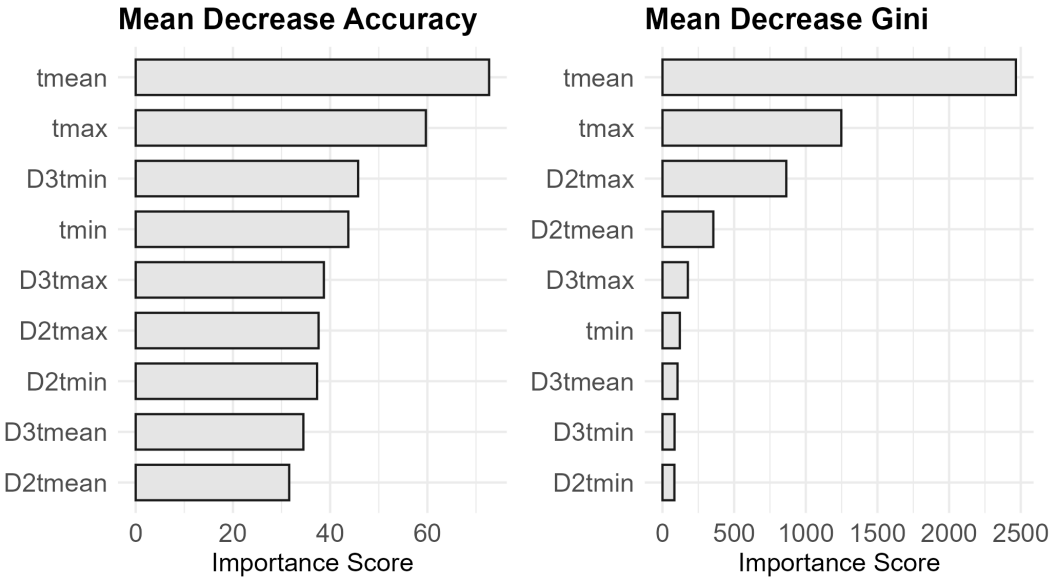


Figure 3.4 Variable importance plots for the forest model trained on data up-sampled using the **basic** approach.

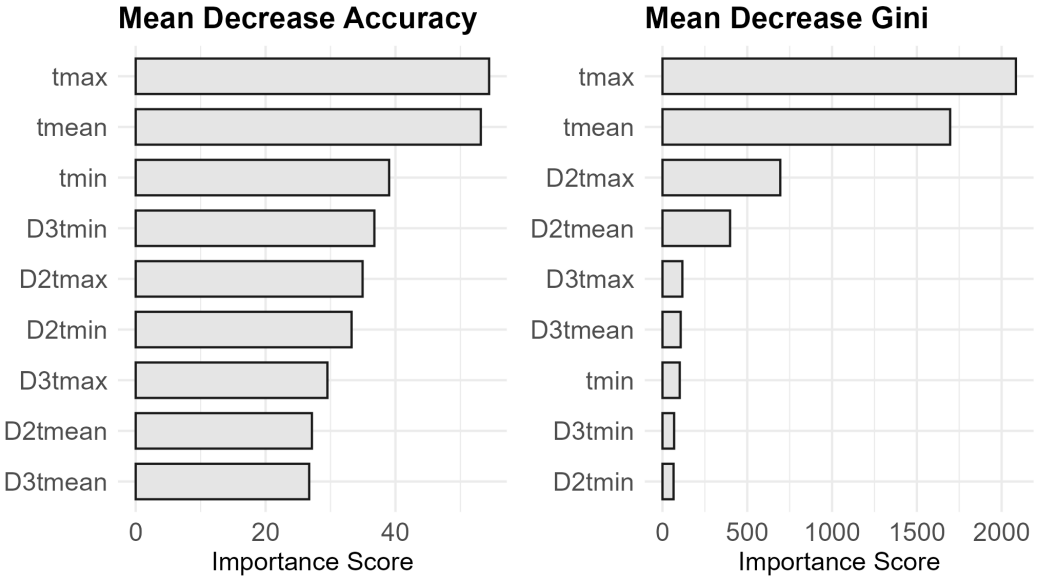


Figure 3.5 Variable importance plots for the forest model trained on data up-sampled using the **synthetic** approach.

3.1.4 Performance on Known Data

Applying the models on the original dataset produced the confusion matrices in Table 3.1. Due to random effects introduced during random forest training, these specific numbers only correspond to a specific seed.

The "synthetic" model produced slightly more false positives (FP) and multiple false negatives (FN), while the "basic" model only yielded one FN. Figures 3.6 to 3.8 show the temperatures on days corresponding to FP or FN predictions. No plot is provided for the single FN day produced by the "basic" model.

The mean of T_{max} on false positive days exceeded 30 °C for both models. While the mean values of T_{max}, T_{mean} and T_{min} were comparable between the models, the means of the "synthetic" one were slightly lower. The measured maximum temperatures on false negative days produced by the "synthetic" model were below 30 °C and the mean of T_{mean} was slightly above 21 °C.

Table 3.2 provides a city-specific overview of FP and FN days predicted by both models. Additionally, the results obtained by applying the elementary heat alert condition ($T_{max} \geq 30$ °C for 2+ days) are presented.

Table 3.1 Confusion matrices for final models applied to original post-2009 dataset.

	Reference	
Prediction	NO	YES
NO	8204	1
YES	292	683

(a) Model based on **basic** up-sampling: Performance on the original, post-implementation dataset.

	Reference	
Prediction	NO	YES
NO	8173	15
YES	323	669

(b) Model based on **synthetic** up-sampling: Performance on the original post-implementation dataset.

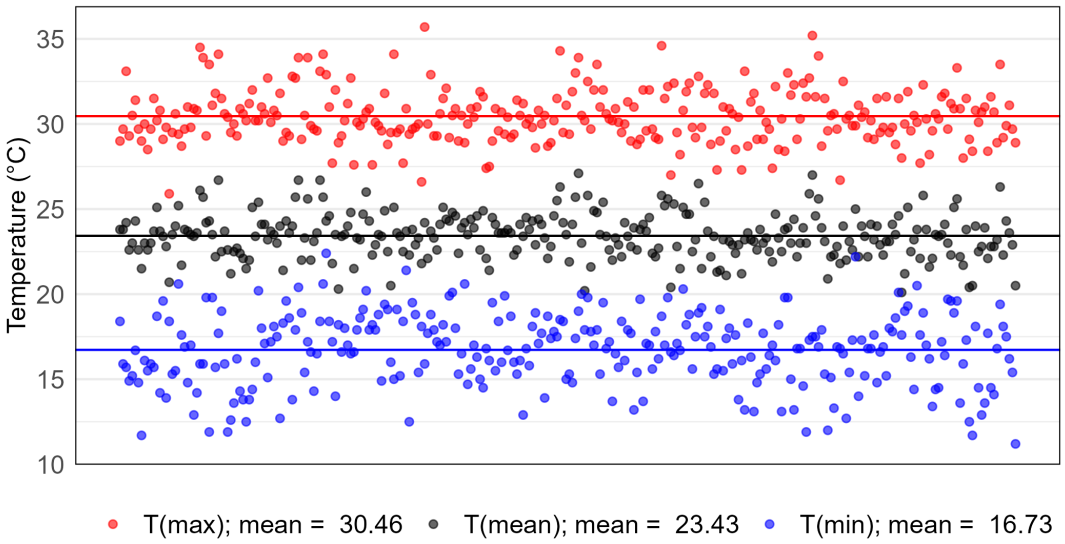


Figure 3.6 Temperatures on false positive days predicted on the original dataset by the classifier trained on data up-sampled using the **basic** approach. The mean values of temperatures recorded on these days are marked by a solid line. The x-axis represents the sequence number of data points rather than specific dates.

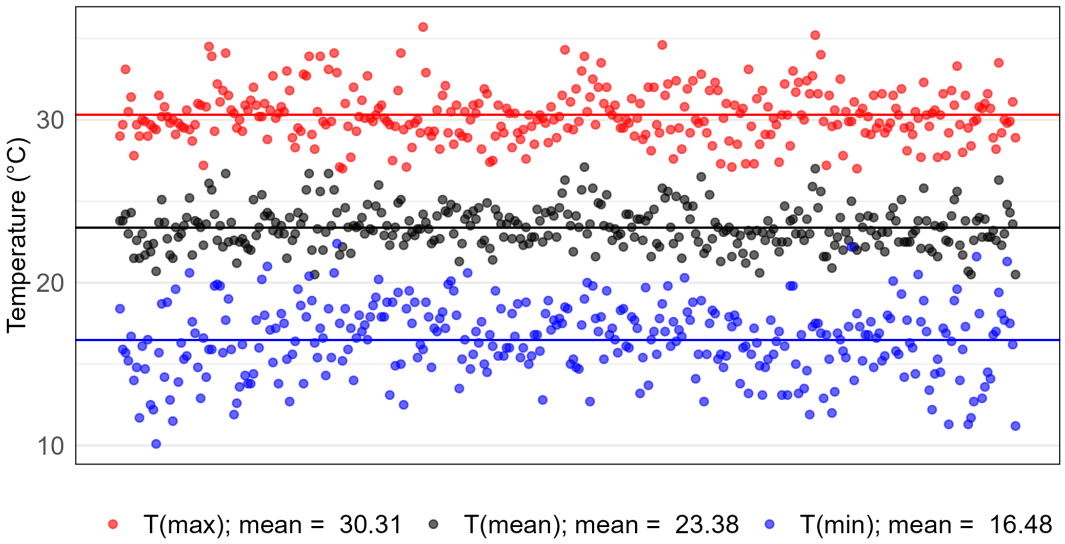


Figure 3.7 Temperatures on false positive days predicted on the original dataset by the classifier trained on data up-sampled using the **synthetic** approach. The mean values of temperatures recorded on these days are marked by a solid line. The x-axis represents the sequence number of data points rather than specific dates.

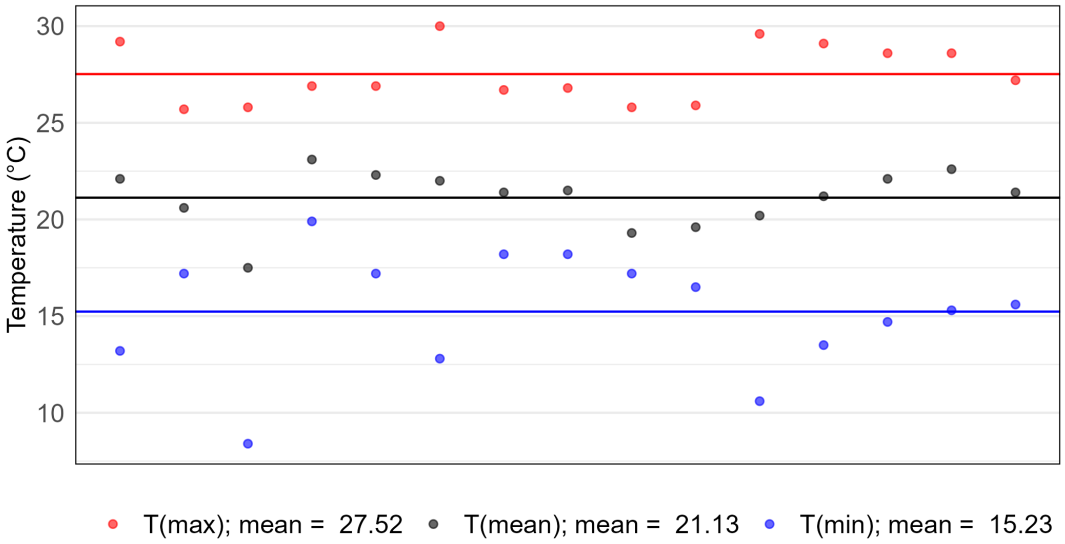


Figure 3.8 An overview of temperatures on false negative days predicted on the original post-implementation dataset by a random forest classifier trained on data up-sampled using the **synthetic** approach. The mean values of temperatures recorded on these days are marked by a solid line. The x-axis represents the sequence number of data points rather than specific dates.

Table 3.2 City-specific prediction performance on the original post-implementation dataset. For each city, the total number of heat alerts (Total HA) is reported, as well as the number of days with maximum temperatures exceeding 30°C. The number of matching predictions as well as the false positives (FP) and false negatives (FN) are reported for three methods of heat alert day determination — applying threshold criteria for heat alerts and training a forest model classifier based on measured temperatures.

		Krakow	Lodz	Poznan	Warsaw	Wroclaw
Total HA		150	121	125	141	147
Days with $T_{max} \geq 30°C$		154	143	153	136	192
Criteria	**Match**	82	91	87	90	114
	FP	15	16	26	7	32
	FN	69	32	41	54	34
Basic	**Match**	150	121	125	141	146
	FP	60	42	61	52	77
	FN	0	0	0	0	1
Synthetic	**Match**	144	120	124	138	143
	FP	64	48	65	56	90
	FN	6	1	1	3	4

3.1.5 Predicting Pre-Implementation Eligible Days

Table 3.3 provides an overview of the number of days predicted as eligible for a heat alert in each city. Maximum temperatures on individual days with heat alerts are shown in Figure 3.9.

Since the time series regression model will be applied for a 3-day lag, mortality distribution on eligible and the three following days is visualized for the "synthetic" model in Figure 3.10 (pre-implementation period). Additionally, the distribution of mortality for non-eligible days is included for comparison. Since the predictions of the "basic" and "synthetic" models were similar, as shown in Figure 3.9, no stand-alone figure for the "basic" model is provided. A similar comparison is shown in Figure 3.11 for the post-implementation period. In this case, real heat alert data were used.

During the pre-implementation period, the 1994 distribution for eligible days exhibits particularly wide dispersion in Lodz, Poznan, and Warsaw. In contrast, while Krakow generally exhibited more dispersed mortality values for eligible days throughout the years, 1994 saw a noticeable reduction in this spread–though the median value that year was slightly higher than in other years. In Lodz, the medians of the mortality counts were similar for both groups across the pre-implementation period.

Table 3.3 City-specific prediction on the pre-implementation dataset. S=1 $\wedge$ B=1 denotes the number of days predicted as eligible by both models, S=1 $\wedge$ B=0 is the number of days predicted as eligible only by the "synthetic" model, and S=0 $\wedge$ B=1 is the number of days predicted as eligible only by the "basic" model. For each city, the number of days with maximum temperatures exceeding 30°C is reported.

	Krakow	**Lodz**	**Poznan**	**Warsaw**	**Wroclaw**
Days with $T_{max} \geq 30°C$	65	69	74	67	77
S=1 $\wedge$ B=1	88	77	87	81	89
S=1 $\wedge$ B=0	10	8	14	14	14
S=0 $\wedge$ B=1	9	2	11	5	4

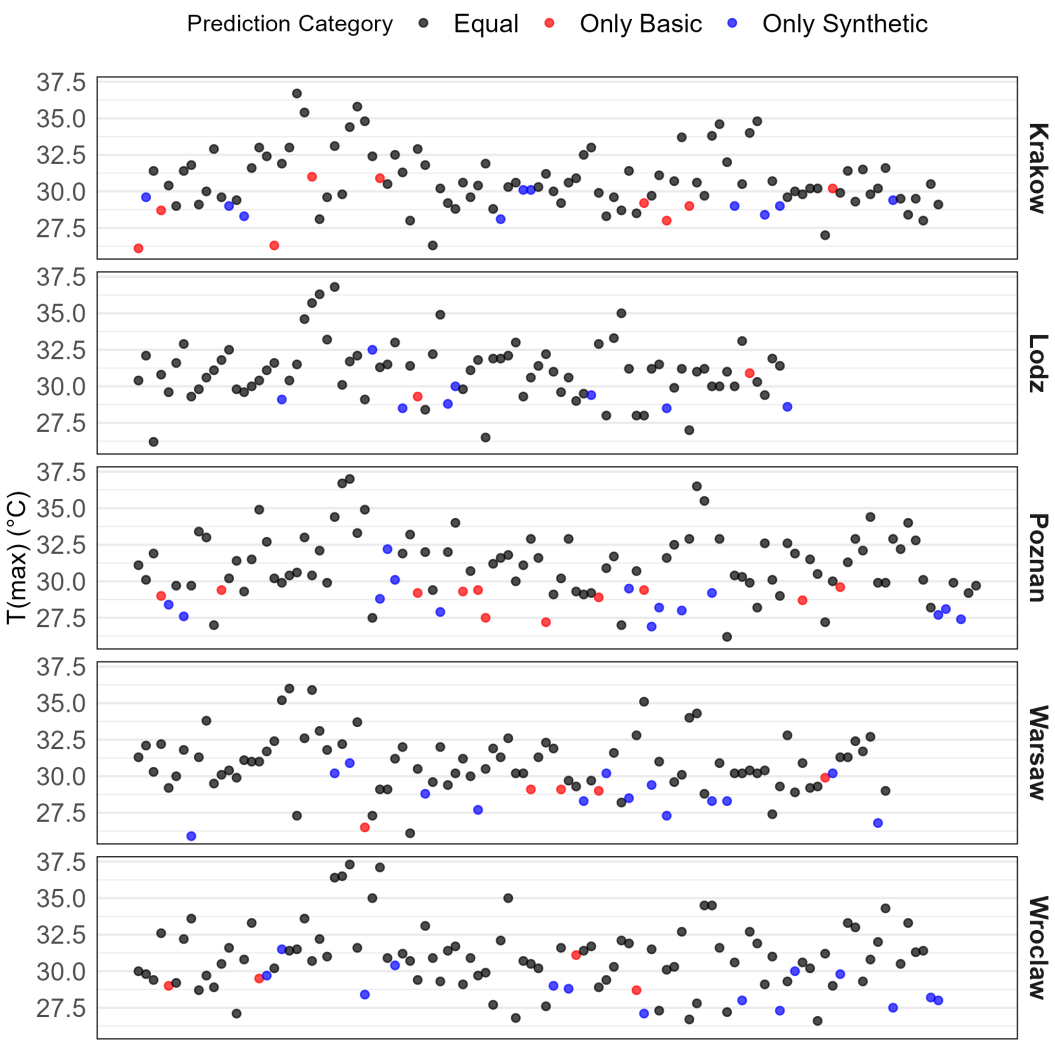


Figure 3.9 The maximum temperatures recorded on days assigned eligible for a heat alert either by both models, or only the "basic" or "synthetic" one. The x-axis represents the sequence number of data points rather than specific dates.

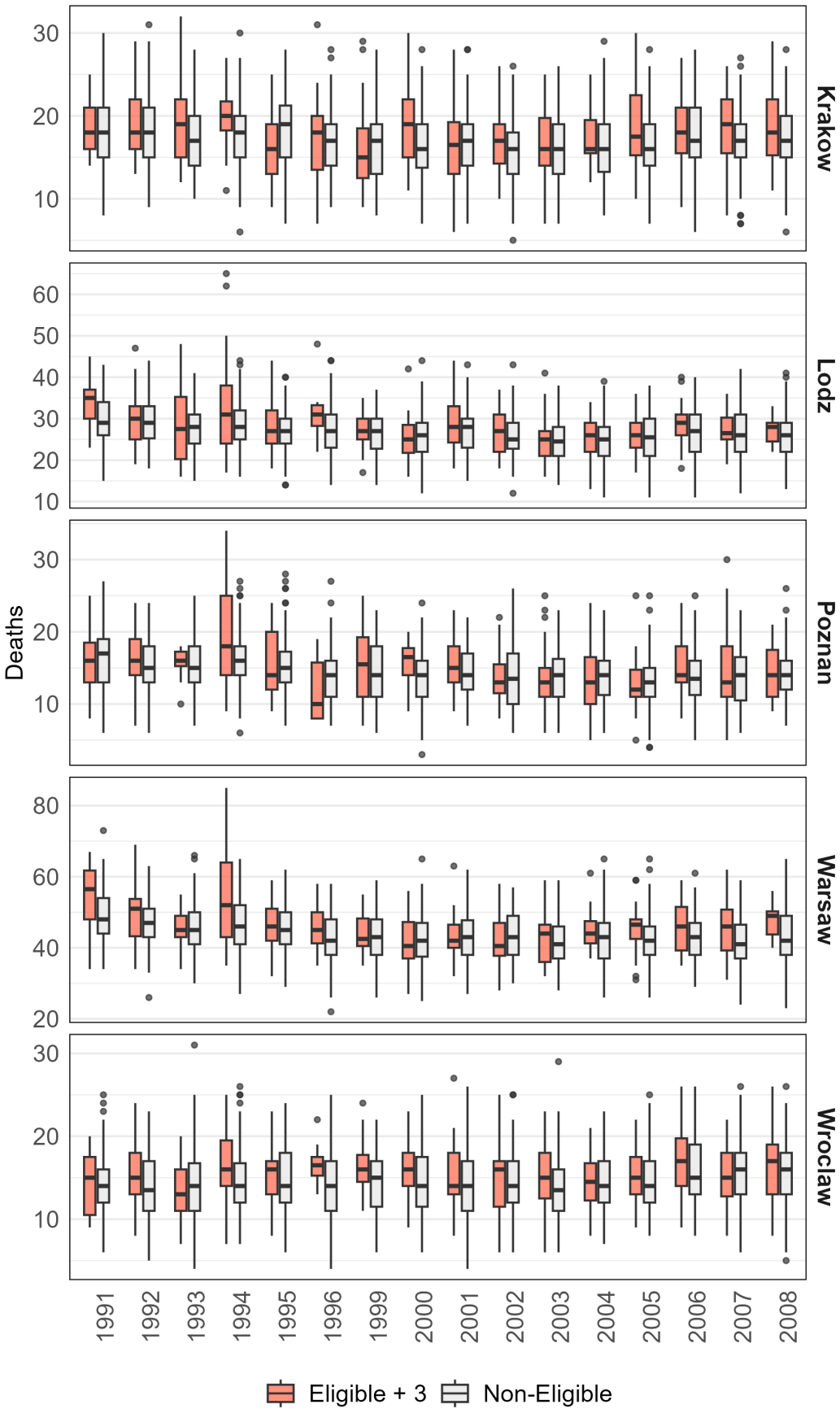


Figure 3.10 Mortality on eligible days and the three days after, versus mortality on non-eligible days. Pre-implementation period, eligibility based on "Synthetic" model.

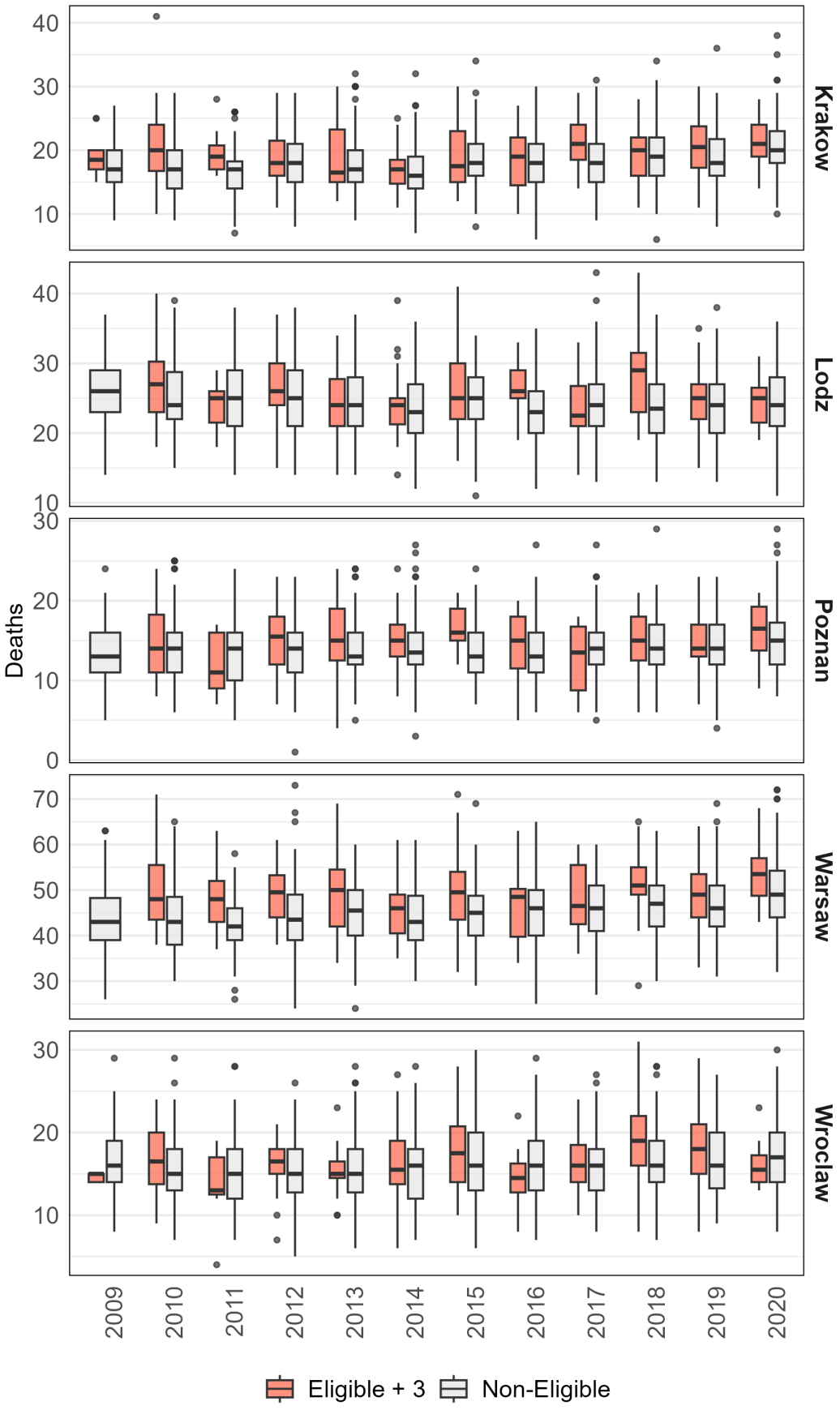


Figure 3.11 Mortality on eligible days and the three days after, versus mortality on non-eligible days. Post-implementation period, eligible days correspond to real heat alert days.

3.2 Time Series Regression

The models introduced in Table 2.3 were applied for a 3-day lag period. Figure 3.12 shows the values of the ζ coefficient (the DID estimator), defined in Equation 2.8, for the relative risk (RR) associated with the implementation of a heat alert. The individual values are also listed in Table 3.4. Values of $\zeta < 1$ indicate a protective effect.

The Effect of Eligibility Assignment Method

The relative risk estimates produced by Model 0 and Model 1 were statistically comparable within the bounds of uncertainty, with no consistent positive or negative bias. The greatest discrepancy between these models was noted in Wroclaw, with Model 1 predicting a protective effect even within the bounds of uncertainty, and Model 0 predicting no effect.

Model 2 employed a temperature threshold criterion for eligibility (i.e., temperatures exceeding 30 °C). With the exception of Poznan, this model yielded results similar to those of Models 0 and 1 within the range of uncertainty. However, its mean RR estimates were significantly different for Poznan and Warsaw. Furthermore, the values were burdened with very high uncertainty.

Sensitivity Analyses

Models 3–5 used the same eligibility assignment method as Model 0. In the pre-implementation days, eligibility was assigned based on the "synthetic" model, and in the post-implementation days, real heat alert data were utilized.

In Model 3, the period between 2003–2008 was omitted. While there was a province-level heat alert in place, the data on heat alerts are not available for this period. Additionally, the year 2009 was left out in this model for a similar reason: there are no available data on heat alerts in 3 out of the 5 cities for this year. Excluding this period did not yield significantly different results to Model 0, but a slight increase in the predicted protective effect is observable.

In order to test the impact of years affected by significant heatwaves (see Tables 1.2 and 1.3), years 1994 and 2015 were omitted in Models 4 and 5, respec-

tively. Model 5 yielded almost the same results as Model 0, indicating no impact of the 2015 heatwave on the DID estimator. However, excluding 1994 resulted in a noticeable difference in the DID estimator, particularly for Lodz and Poznan: the apparent protective effect of heat alerts for those two cities was noticeably reduced. Similarly, the DID estimator was reduced for both Warsaw and Wroclaw. On the other hand, the opposite can be seen for Krakow.

Table 3.4 Median values for the DID estimator ζ representing the relative risk associated with the intervention (HEWS implementation). Results given with confidence intervals as ζ **(95% CI (Lower) – 95% CI (Upper))**. **RD** denotes Real Data.

	Model 0	**Model 1**	**Model 2**
Lodz	0.961 (0.898–1.030)	0.948 (0.883–1.016)	0.973 (0.835–1.134)
Krakow	0.944 (0.876–1.017)	0.951 (0.883–1.025)	0.936 (0.776–1.128)
Poznan	0.908 (0.833–0.989)	0.912 (0.836–0.994)	0.754 (0.622–0.914)
Warsaw	1.006 (0.959–1.056)	1.010 (0.960–1.063)	1.069 (0.949–1.204)
Wroclaw	1.000 (0.926–1.080)	0.931 (0.860–1.009)	0.947 (0.774–1.159)

	Model 3	**Model 4**	**Model 5**
Lodz	0.935 (0.866–1.008)	1.000 (0.932–1.075)	0.947 (0.880–1.018)
Krakow	0.920 (0.844–1.004)	0.933 (0.864–1.008)	0.924 (0.855–0.998)
Poznan	0.878 (0.796–0.969)	0.958 (0.876–1.047)	0.900 (0.823–0.985)
Warsaw	0.986 (0.933–1.042)	1.026 (0.976–1.079)	1.000 (0.952–1.052)
Wroclaw	0.986 (0.901–1.080)	1.016 (0.939–1.100)	0.985 (0.910–1.067)

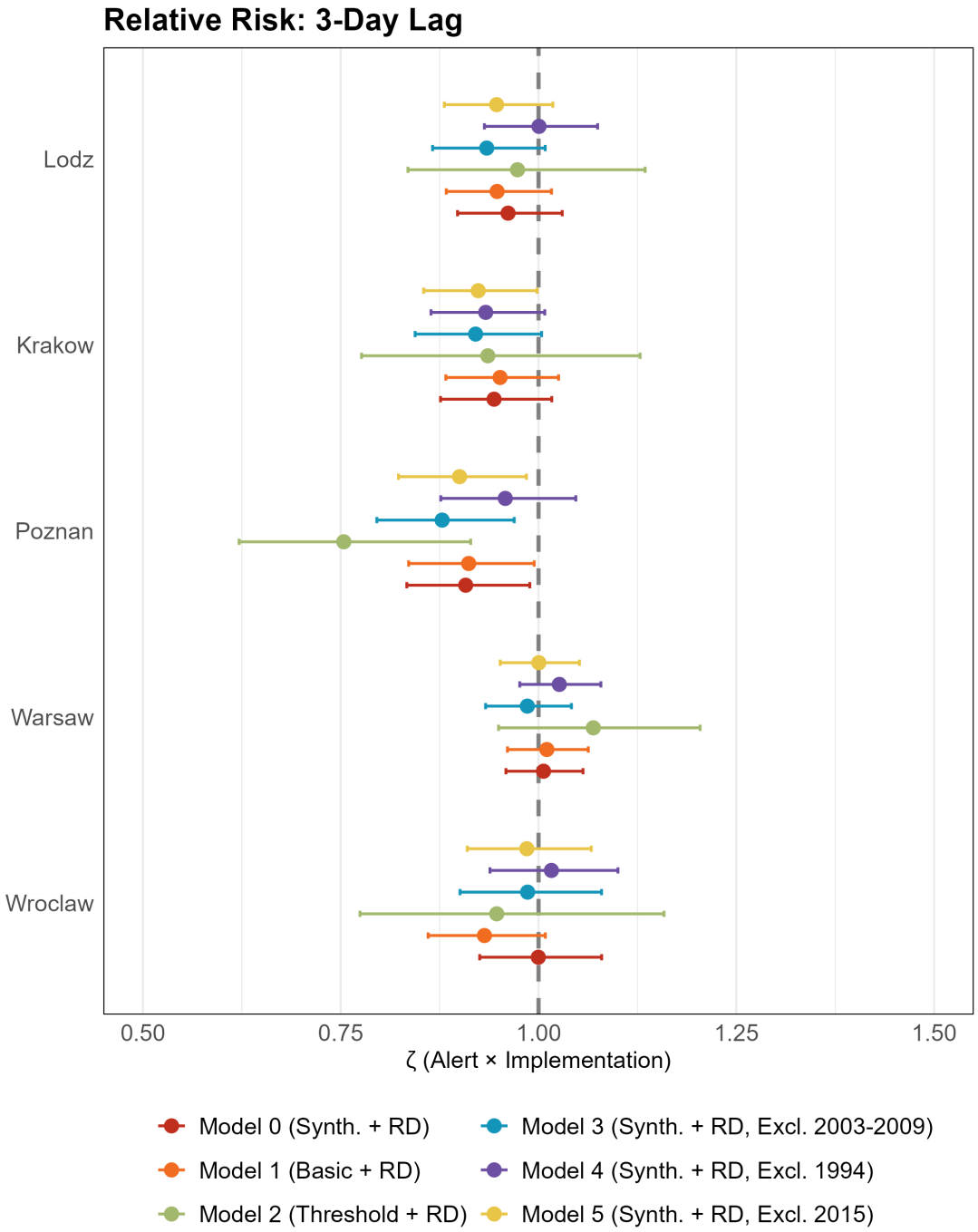


Figure 3.12 The values of the reduced coefficient ζ, indicating relative risk associated with the implementation of heat alerts, for five models defined in Table 2.3. ζ represents the cumulative effect for 3 days following a heat alert.

4. Discussion

This study investigated the implementation of a difference-in-differences study design to asses the effect of heat alert implementation on mortality. While this particular study design is mostly used in econometrics, some previous research has utilized this approach to investigate the heat alert – mortality relationship (Benmarhnia et al., 2016; Feldbusch, 2023). Heo et al., 2019 also employed a DID study design, however, their methodological approach to defining pre- and post-implementation periods differed substantially. Heat alert and non-alert days were treated simultaneously as both treatment groups and time periods, without incorporating any pre-implementation data. As this design diverges significantly from the approach adopted in this thesis, a direct comparison with Heo et al., 2019 is not supplied.

The classic DID approach compares trends between a treatment and control group, in a period prior to implementing an intervention (pre-implementation period), and a period following the implementation (post-implementation period). In this particular context, the treatment is the implementation of a Heat Early Warning System, the treatment group are days with a heat alert, and the control group are days without a heat alert.

In order to implement the DID model, a quasi-Poisson time series regression model was used, including a Distributed Lag Model (DLM) for the heat alert indicator. The previous studies following a DID study design (Benmarhnia et al., 2016; Feldbusch, 2023; Heo et al., 2019) analyzed the immediate, same-day protective effect of heat alerts. However, using DLMs makes it possible to assess the cumulative protective effect in the span of several days.

The DID study was implemented for five Polish cities: Krakow, Lodz, Poznan, Warsaw, and Wroclaw. The pre-implementation period spanned from 1991 to

2009, and the post-implementation period from 2009 to 2020. While a HEWS on a province level existed since 2003, a city-level HEWS was implemented in 2009, and there are no heat alert data available for the period between 2003 and 2008. The sensitivity of the model on this particular period (2003–2008) was tested, as well as its sensitivity to years affected by major heatwaves (1994 and 2015). Furthermore, mortality data were missing for 1997 and 1998: these years could not be included in the analysis. The analysis was restricted to the warm months from May to September.

4.1 Treatment Group Assignment

One of the challenges of determining heat alert effect via DID is the proper definition of the pre-implementation treatment group. In principle, heat alert days did not exist prior to HEWS implementation. However, days *eligible* for a heat alert can be identified and used as the pre-implementation treatment group. Benmarhnia et al., 2016 used available meteorological data to retrospectively assign heat alert days based on alert trigger criteria. For Poland, this would be maximum temperatures over 30 °C for 2+ days. Applying this threshold approach failed to reproduce the majority of heat alerts in the post-implementation period (Section B.1), therefore, a machine learning approach to classifying eligible days was employed.

A random forest classifier was trained in order to assign eligibility flags to pre-implementation days, following the approach utilized by Feldbusch, 2023. The assignment procedure proposed by Feldbusch, 2023 was extended, with a thorough model optimization, and an alternative suggestion for resolving a key issue with the heat alert data: their imbalance. Throughout the analyzed period restricted to months May–September, there was an average of 8 days with warning for 92 days without. In order to balance the dataset for the purpose of training a random forest, Feldbusch, 2023 used an up-sampling method based on random replication of heat alert days, thus bringing the class ratio of the dataset to 1:1. This method is referred to as basic up-sampling, and the corresponding forest model as "basic" model. In this thesis, an alternative, statistical approach to up-

sampling is proposed: utilizing the SMOTE (synthetic minority over-sampling technique) function in R, which uses a K-nearest neighbor algorithm to generate new points within the minority class. This method is referred to as synthetic up-sampling, and the corresponding forest model as "synthetic" model.

4.1.1 Hyperparameter Tuning

The up-sampled datasets were split into training and testing portion in a 70:30 ratio. An algorithm searching for optimal hyperparameters defining the maximum number of terminal leaf nodes (`maxnodes`) and the minimum number of data points at any given node (`nodesize`) was applied for both models (Algorithm 1). For each combination of `maxnodes` and `nodesize`, the optimal value for `mtry`, defining the number of predictor variables randomly selected at each node, was determined using 10-fold cross validation via `caret`'s `train` function.

Both models exhibited signs of overfitting for `nodesize`=1 and `maxnodes` exceeding 500, with the train accuracy equal to 100% (Figures B.2 – "basic" model, and B.3 – "synthetic" model). This strongly suggests that the models memorized the training dataset and would not be able to generalize. The accuracy was still very high (over 98%) for `nodesize` equal to 10 and 20 and `maxnodes` exceeding 500.

The accuracy on the testing datasets was very high throughout the hyperparameter space, with the "synthetic" model scoring higher accuracy for lower `maxnodes`, but slightly lower accuracy for high `maxnodes`. The lower range of the test accuracy values of the "synthetic" model suggests better stability of the model with respect to the `maxnodes` parameter. The cross-validation accuracy was similar to the test accuracy for both models, indicating reliable cross-validation results.

Throughout the hyperspace, both models preferred higher values of `mtry` (Tables B.2 – "basic" model and B.3 – "synthetic" model). This might be due to both models strongly preferring a limited number of variables, and a higher `mtry` increases the probability that the preferred variable will appear in the random selection.

The final selection of the hyperparameters was as follows: `nodesize` = 30,

`maxnodes` = 500, and the corresponding value for `mtry` was 5. Multiple considerations were taken into account: first, a singular *best* combinations of parameters does not exist, and multiple combinations will yield very similar results. Second, the risk of model overfitting increases for smaller `nodesize` and higher `maxnodes`. Third, all models were used to predict the original, not up-sampled post-implementation dataset, and the p-values of the predictions were examined (Tables B.4 and B.5), thus eliminating low values of `maxnodes`. Since the above mentioned characteristics applied to both "basic" and "synthetic" models, the final choice of hyperparameters was the same for both.

4.1.2 Final Forest Models

Once the final classifiers were trained, the class-specific and out-of-bag errors (Figures 3.2 and 3.3) were examined. Both the class-specific and OOB errors were comparable for both models, however, the "basic" model displayed very low error rate for the positive class (heat alert days), which converged to nearly zero.

Figures 3.4 and 3.5 show the variable importance plots of the "basic" and "synthetic" model, respectively. The strong preference of T_{mean} in the "basic" model and T_{max} and T_{mean} in the "synthetic" model reflect the reliance of heat alert triggers on the maximum temperatures. The two-day rolling mean of T_{max} scored approximately the same MDG importance, placing it third in importance for both models. This again mirrors the actual heat alert trigger, which requires at least 2 consecutive hot days. The low importance score of the MDG for minimum temperatures might be simply due to the variables not being selected for node splitting, therefore having no contribution to the overall MDG. In comparison, the forest classifier implemented by Feldbusch, 2023 for German data yielded the apparent temperature and its 2-day mean as the most important with respect to MDG. Unfortunately, humidity data for Poland was not available, and the apparent temperature could not be determined.

The prediction reliability of both models on the post-implementation period data was compared via confusion matrices, and by examining the days marked as false positives or false negatives. The "basic" model yielded a singular false negative, which translates to roughly 0.15% of the total heat alert days. On the

other hand, approximately 3.5% of non-heat alert days were wrongly predicted as having one. This is indicative of the model simply memorizing the heat alert days, which were present in multiple copies in the training dataset. The "synthetic" model marked 2% of heat alert days as false negatives and 3.8% of non-heat alert days as false positives, thus yielding more class-balanced predictions. However, the false positives overestimated the total number of heat alerts by almost 50% with respect to the actual value. Although the purpose of the model is not to replicate post-implementation heat alerts, this signals that the models were not able to capture the distinction between a hot day, and a hot day with a heat alert.

A closer look on the FP days (Figures 3.6 – "basic" model and 3.7 – "synthetic" model) reveals that the mean T_{max} on false positive days slightly exceeded 30 °C for both models. While the mean T_{max} on true heat alert days was 31.29°C (Figure 2.5), therefore almost a degree higher, one might still expect a heat alert on a day with temperatures exceeding 30 °C. These false positive days therefore cannot be simply attributed to inaccuracy of the model. This illustrates the disparity between the weather forecast, which triggers a heat alert, and the actual recorded temperatures.

Table 3.2 shows the city-specific prediction accuracy, as well as the accuracy of the threshold-based aproach to assigning eligible days. In contrast with the forest models, the threshold approach ($T_{max} > 30$ °C for 2+ days) yielded a significantly higher number of false negative days. This can likely by attributed to the weather stations located at remote locations (see the map in Figure 2.3), where the recorded temperature might me lower than within the city.

4.1.3 Parallel Trends Assumption

After both models were applied to the pre-implementation data, the validity of the parallel trends assumption was explored (see Section B.5). The parallel trend seemed to be disturbed in Lodz (both models) and Krakow ("basic" model). The trend disturbance for Warsaw ("synthetic" model) can likely be neglected, as it only concerns a specific day of the week, and no other coefficients were statistically significant. Furthermore, the disturbance only concerns selected seasonal

patterns, and the coefficient for the date variable was not significant: this might indicate that the long-term trend was not disturbed.

The approach taken in this study is different to the one employed by Feldbusch, 2023: in her thesis, the parallel trends check was based on linear regression of mortality on eligible and non-eligible days. While this approach is straightforward and does provide a visualization of the overall mortality trends, the parallel trends assumption should be verified on a scale reflecting the resolution of the outcome variable: in other words, the trends should be compared on a day-to-day basis for models predicting daily mortality. In contrast, the approach described in Section B.5 takes daily variability into account. However, it is important to note that neither approach offers an objective criterion to test the parallel trends assumption.

4.2 Time Series Regression

Six DLMs (Table 2.3) were implemented, all following the general formula set in Equation (2.7). The protective effect over a 3-day lag period was analyzed. The models using eligibility assignment from the "basic" and "synthetic" models yielded similar results, with the only major difference observable for Wroclaw. In Wroclaw, at least 5 years from the post-implementation period recorded lower mortality medians on heat alert days plus the three following days compared to non-alert days (see Figure 3.11), while in the other cities, usually only one year with such discrepancy occurred. This might explain the lower value of ζ for Wroclaw, indicating a less protective effect of heat alerts.

Applying threshold criteria to assign eligibility ensures homogeneity of the pre-implementation treatment group and follows the study by Benmarhnia et al., 2016. However the ζ coefficients are burdened with a high uncertainty: applying such broad criterion inherently includes a higher number of eligible days, thus bringing in more days with both lower and higher mortality. Furthermore, this approach does not address the discrepancy between the forecast and measured variables in any way, and applying this criterion to post-implementation days failed to reproduce a significant portion of true heat alert days, as described in

Section B.2.

Excluding years between 2003–2009 had negligible effect on the results, despite some form of heat alerts on a province level being in place. Omitting 2003–2008 from the pre-implementation period without any observable effect points to several conclusions: one, this period does not drive the pre-implementation trend. Second, the parallel trends assumption is likely satisfied, as both the treatment and control group behaved similarly prior to intervention. Furthermore, omitting 2009 without an effect points to the year not being a *shock* year – no sudden reaction to heat alert implementation can be observed in cities which did record several heat alerts (Krakow and Wroclaw).

Excluding 1994, when Poland experienced a major heatwave, had a noticeable effect. For all cities except Krakow, the protective effect of heat alerts was reduced. This can likely be attributed to increased mortality as an aftermath of the heatwave: this would increase the overall pre-treatment mortality, and exaggerate the treatment effect of HEWS implementation. Finally, excluding 2015 from the post-treatment period had an almost negligible effect on the estimators. This indicates that the observed treatment effect is robust against this potential outlier.

4.3 Suggestions for Future Research

The difference-in-differences study design is particularly effective when confounding variables are shared across two groups. DID controls for such confounders without relying on absolute trends. However, applying this approach to heat alert studies presents challenges, notably in assigning pre-implementation eligibility. The uncertainty of this assignment arises not only from discrepancies between weather forecasts and actual recorded meteorological data but also from the inherent uncertainty in model predictions. For Poland, no information on daily relative humidity was available, and this variable could not be included in the forest models predicting the eligible days. However, in the analysis performed for German data by Feldbusch, 2023, the apparent temperature derived from the daily temperature and humidity proved to be the most important within the

model: testing the sensitivity of DID results against inclusion of humidity would be beneficial.

The protective effect of heat alerts over the 3 days following its trigger were observed. A supplementary sensitivity analysis examining the impact of lag length choice on both the predicted protective effect and its uncertainty would be suitable in order to find a balance between capturing the temporal span of the protective effects while not unnecessarily inflating the uncertainty.

This thesis focused on predicting the protective effect of HEWS on mortality, and for most cities, minimal protective effects were observed, particularly after omitting 1994 which was shown to significantly overestimate the effect. It would be helpful to apply a similar sensitivity analysis for a different outcome, e.g., ambulance call-outs or number of hospitalizations. Similarly, applying this methodology and sensitivity analyses to data from a different country would be also useful.

The issue of not being able to objectively test for the parallel trend assumption is resolved within the event study design (Miller, 2023). This study is based on observing the treatment effect at multiple time points by splitting both the pre- and post-implementation period into subperiods. Therefore, a gradual and changing behaviour between the groups in both periods can be observed, instead of reducing the overall effect to one single coefficient. Furthermore, the parallel trend assumption for each pre-treatment subperiod can be tested. Applying this method for the Polish data could verify the results obtained within this thesis.

5. Conclusion

A difference-in-differences (DID) study design was employed to observe the treatment effect of heat early warning system implementation in five Polish cities, using distributed lag models to examine lagged effects of heat alerts on mortality. Two forest model classifiers were trained to identify days eligible for heat alerts, in order to assign treatment group affiliation in the pre-implementation period. The training data for the classifiers came from the post-implementation period, and two approaches to solving strong imbalance between days with heat alerts and days without heat alert in the training data were compared. First, a simple up-sampling method relying on random data replication was employed, and second, an interpolation method based on K-nearest neighbors algorithm. While both models yielded slightly different predictions of the pre-implementation eligible days, the up-sampling method had a negligible effect on the final DID estimator representing the protective effect of heat alerts. Using heat alert trigger criteria instead of training a model yielded similar results in most cases, but were burdened with a considerably higher uncertainty.

Aside from the eligibility assessment method, robustness checks were conducted in order provide validation for the estimated treatment effects. The exclusion of years 2003–2009, despite the presence of limited regional heat alert systems during that period, had minimal impact on the DID estimates, suggesting that this interval did not influence the pre-implementation trend and supporting the parallel trends assumption. In contrast, excluding 1994, a year marked by an extreme heatwave, significantly influenced the results in most cities, likely due to elevated baseline mortality inflating the estimated protective effect of heat alerts. Finally, excluding 2015, a year when most cities also experienced a major heatwave, had no meaningful effect on the estimators.

Overall, the robustness of the quasi-experimental DID study was tested, and while certain methodological choices were shown to have minimal effect on the results, mortality outliers had a noticeable effect on the DID estimates. Several supplementary sensitivity tests were proposed, providing opportunities for future research.

Bibliography

BALLESTER, Joan; QUIJAL-ZAMORANO, Marcos; TURRUBIATES, Raúl Fernando Méndez; PEGENAUTE, Ferran; HERRMANN, François R.; ROBINE, Jean Marie; BASAGAÑA, Xavier; TONNE, Cathryn; ANTÓ, Josep M.; ACHEBAK, Hicham, 2023. Heat-related mortality in Europe during the summer of 2022. *Nature Medicine*. Vol. 29, pp. 1857–1866. Available from DOI: https://doi.org/10.1038/s41591-023-02419-z.

BEDNAR-FRIEDL, B.; BIESBROEK, R.; SCHMIDT, D.N.; ALEXANDER, P.; BØRSHEIM, K.Y.; CARNICER, J.; GEORGOPOULOU, E.; HAASNOOT, M.; COZANNET, G. Le; LIONELLO, P.; LIPKA, O.; MÖLLMANN, C.; MUCCIONE, V.; MUSTONEN, T.; PIEPENBURG, D.; WHITMARSH, L., 2022. Europe. In: *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by PÖRTNER, H.-O.; ROBERTS, D.C.; TIGNOR, M.; POLOCZANSKA, E.S.; MINTENBECK, K.; ALEGRÍA, A.; CRAIG, M.; LANGSDORF, S.; LÖSCHKE, S.; MÖLLER, V.; OKEM, A.; RAMA, B. Cambridge, UK and New York, NY, USA: Cambridge University Press, pp. 1817–1927. Available from DOI: https://doi.org/10.1017/9781009325844.015.

BENMARHNIA, Tarik; BAILEY, Zinzi; KAISER, David; AUGER, Nathalie; KING, Nicholas; KAUFMAN, Jay S., 2016. A Difference-in-Differences Approach to Assess the Effect of a Heat Action Plan on Heat-Related Mortality, and Differences in Effectiveness According to Sex, Age, and Socioeconomic Status (Montreal, Quebec). *Environmental Health Perspectives*. Vol. 124, no. 11, pp. 1694–1699. Available from DOI: https://doi.org/10.1289/EHP20.

BREIMAN, Leo, 2001. Random Forests. *Machine Learning*. Vol. 45, no. 1, pp. 5–32. Available from DOI: https://doi.org/10.1023/A:1010933404324.

BŁAŻEJCZYK, Krzysztof; TWARDOSZ, Robert; WAŁACH, Piotr; CZARNECKA, Kaja; BŁAŻEJCZYK, Anna, 2022. Heat strain and mortality effects of prolonged central European heat wave—an example of June 2019 in Poland. *International Journal of Biometeorology*. Vol. 66, pp. 149–161. Available from DOI: 10.1007/s00484-021-02202-0.

CARD, David; KRUEGER, Alan B., 1994. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *The American Economic Review*. Vol. 84, no. 4, pp. 772–773.

CASANUEVA, Ana; BURGSTALL, Annkatrin; KOTLARSKI, Sven; MESSERI, Alessandro; MORABITO, Marco; FLOURIS, Andreas D.; NYBO, Lars; SPIRIG, Christoph; SCHWIERZ, Cornelia, 2019. Overview of Existing Heat-Health Warning Systems in Europe. *International Journal of Environmental Research and Public Health*. Vol. 16, no. 15:2657, p. 2657. Available from DOI: https://doi.org/10.3390/ijerph16152657.

CHAU, P. H.; CHAN, K. C.; WOO, Jean, 2009. Hot weather warning might help to reduce elderly mortality in Hong Kong. *International Journal of Biometeorology*. Vol. 53, pp. 461–468.

CHIABAI, Aline; SPADARO, Joseph V.; NEUMANN, Marc B., 2018. Valuing deaths or years of life lost? Economic benefits of avoided mortality from early heat warning systems. *Mitigation and Adaptation Strategies for Global Change*. Vol. 23, pp. 1159–1176. Available from DOI: https://doi.org/10.1007/s11027-017-9778-4.

CMM IMGW-PIB, 2024. *The Art of Heat*. Centrum Modelowania Meteorologicznego, Instytut Meteorologii i Gospodarki Wodnej, Państwowy Instytut Badawczy. Available also from: https://cmm.imgw.pl/?p=42258.

EBI, Kristie L.; TEISBERG, Thomas J.; KALKSTEIN, Laurence S.; ROBINSON, Lawrence; WEIHER, Rodney F., 2003. Heat Watch/Warning Systems Save Lives: Estimated Costs and Benefits for Philadelphia 1995–98. *Bulletin of the*

American Meteorological Society. Vol. 85, no. 8, pp. 1067–1074. Available from DOI: https://doi.org/10.1097/00001648-200309001-00064.

EUROPEAN ENVIRONMENT AGENCY, 2020. *Thermal Comfort Indices – Universal Thermal Climate Index, 1979-2020*. Available also from: https://climate-adapt.eea.europa.eu/en/metadata/indicators/thermal-comfort-indices-universal-thermal-climate-index-1979-2019. Accessed: 2025-03-18.

EUROSTAT, 2024. *GISCO Geodata, Administrative units - Countries* [https://ec.europa.eu/eurostat/web/gisco/geodata/administrative-units/countries].

EUROSTAT, 2025a. *Population on 1 January by age groups and sex - functional urban areas* [https://ec.europa.eu/eurostat/databrowser/view/urb_lpop1/default/table?lang=en]. Available from DOI: https://doi.org/10.2908/URB_LPOP1. Accessed: 2025-03-07.

EUROSTAT, 2025b. *Population structure indicators by NUTS 3 region* [https://ec.europa.eu/eurostat/databrowser/view/demo_r_pjanind3/default/table?lang=en]. Available from DOI: https://doi.org/10.2908/DEMO_R_PJANIND3. Accessed: 2025-03-07.

FELDBUSCH, Hanna, 2023. *Assessing the effects of the heat health warning system on mortality in 15 German cities: A difference-in-differences approach*. Master's thesis. Ludwig-Maxmilians-Universität München, Faculty of Medicine.

FLACH, Peter, 2012. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press. ISBN 978-1-107-42222-3.

GALLO, Elisa; QUIJAL-ZAMORANO, Marcos; TURRUBIATES, Raúl Fernando Méndez; TONNE, Cathryn; BASAGAÑA, Xavier; ACHEBAK, Hicham; BALLESTER, Joan, 2024. Heat-related mortality in Europe during 2023 and the role of adaptation in protecting health. *Nature Medicine*. Vol. 30, pp. 3101–3105. Available from DOI: https://doi.org/10.1038/s41591-024-03186-1.

GASPARRINI, A., 2011. Distributed lag linear and non-linear models in R: the package dlnm. *Journal of Statistical Software*. Vol. 43, no. 8, pp. 1–20. Available from DOI: 10.18637/jss.v043.i08.

GASPARRINI, Antonio; ARMSTRONG, Ben; KENWARD, Michael G., 2010. Distributed lag non-linear models. *Statistics in Medicine*. Vol. 29, no. 21, pp. 2224–2234. Available from DOI: https://doi.org/10.1002/sim.3940.

GOVERNMENT OF POLAND, 2022. *Krajowy Plan Zarządzania Kryzysowego*. Available also from: https://www.gov.pl/web/rcb/krajowy-plan-zarzadzania-kryzysowego. Accessed: 2025-03-17.

GUHA-SAPIR, Debby; VOS, Femke; BELOW, Regina; PONSERRE, Sylvain, 2011. Annual Disaster Statistical Review 2010. *Centre for Research on the Epidemiology of Disasters* [https://www.cred.be/sites/default/files/ADSR_2010.pdf].

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome, 2009. *The Elements of Statistical Learning*. 2nd ed. Springer New York, NY.

HEO, Seulkee; NORI-SARMA, Amruta; LEE, Kwonsang; BENMARHNIA, Tarik; DOMINICI, Francesca; BELL, Michelle L., 2019. The Use of a Quasi-Experimental Study on the Mortality Effect of a Heat Wave Warning System in Korea. *International Journal of Environmental Research and Public Health*. Vol. 16, no. 12. Available from DOI: https://doi.org/10.3390/ijerph16122245.

HESS, Jeremy J.; SATHISH, L. M.; KNOWLTON, Kim; SAHA, Shubhayu; DUTTA, Priya; GANGULY, Parthasarathi; TIWARI, Abhiyant; JAISWAL, Anjali; SHEFFIELD, Perry; SARKAR, Jayanta; BHAN, S. C.; BEGDA, Amit; SHAH, Tejas; SOLANKI, Bhavin; MAVALANKAR, Dileep, 2018. Building Resilience to Climate Change: Pilot Evaluation of the Impact of India's First Heat Action Plan on All-Cause Mortality. *Journal of Environmental and Public Health*. Vol. 2018, pp. 1–8. Available from DOI: 10.1155/2018/7973519.

HONDULA, David M.; JR., Robert C. Balling; VANOS, Jennifer K.; GEORGESCU, Matei, 2015. Rising Temperatures, Human Health, and the Role of Adaptation. *Climate Change and Human Health*. Vol. 1, pp. 144–154. Available from DOI: https://doi.org/10.1007/s40641-015-0016-4.

HUNT, Alistair; FERGUSON, Julia; BACCINI, Michela; WATKISS, Paul; KENDROVSKI, Vladimir, 2017. Climate and weather service provision: Economic appraisal of adaptation to health impacts. *Climate Services*. Vol. 7, pp. 78–86. ISSN 2405-8807. Available from DOI: https://doi.org/10.1016/j.cliser.2016.10.004. IMPACT2C - Quantifying projected impacts under 2°C warming.

HUTH, Radan; KYSELÝ, Jan; POKORNÁ, Lucie, 2000. A GCM Simulation of Heat Waves, Dry Spells, and Their Relationships to Circulation. *Climatic Change*. Vol. 46, pp. 29–60. Available from DOI: 10.1023/A:1005633925903.

IPCC, 2022. *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK and New York, NY, USA: Cambridge University Press, International Panel on Climate Change. Available from DOI: https://doi.org/10.1017/9781009325844.

KILBOURNE, Edwin M., 1997. *The public health consequences of disasters*. Heat waves and hot environments. Oxford University Press. ISBN 0-19-509570-7.

KOVATS, R Sari; KRISTIE, L Ebi, 2006. Heatwaves and public health in Europe. *European Journal of Public Health*. Vol. 16, no. 6, pp. 592–599. Available from DOI: https://doi.org/10.1093/eurpub/ckl049.

KRZYŻEWSKA, Agnieszka; DYER, Jamie, 2018. The August 2015 mega-heatwave in Poland in the context of past events. *Weather*. Vol. 73, no. 7, pp. 207–214. Available from DOI: https://doi.org/10.1002/wea.3244.

KUCHCIK, Magdalena; BŁAŻEJCZYK, Krzysztof; HALAŚ, Agnieszka, 2021. Long-term changes in hazardous heat and cold stress in humans: multi-city study in Poland. *International Journal of Biometeorology*. Vol. 65, no. 9, pp. 1567–1578. Available from DOI: 10.1007/s00484-020-02069-7.

KUHN, Max; JOHNSON, Kjell, 2013. *Applied predictive modeling*. Springer New York, NY. ISBN 978-1-4614-6848-6.

LANTZ, Brett, 2019. *Machine learning with R: expert techniques for predictive modeling*. Third. Packt. ISBN 978-1-78829-586-4.

LIAW, Andy; WIENER, Matthew, 2002. Classification and Regression by randomForest. *R News*. Vol. 2, no. 3, pp. 18–22. Available also from: `https://CRAN.R-project.org/doc/Rnews/`.

MARTÍNEZ-SOLANAS, Èrica; BASAGAÑA, Xavier, 2019. Temporal changes in temperature-related mortality in Spain and effect of the implementation of a Heat Health Prevention Plan. *Environmental Research*. Vol. 169, pp. 102–113. Available from DOI: `https://doi.org/10.1016/j.envres.2018.11.006`.

MCCULLAGH, P.; NELDER, John A., 1989. *Generalized linear models*. Second. Chapman & Hall/CRC. ISBN 0-412-31760-5.

MEYER, Isabella, 2024. Stańczyk by Jan Matejko – A Detailed Artwork Analysis. *Art in Context*. Available also from: `https://artincontext.org/stanczyk-by-jan-matejko/`.

MILLER, Douglas L., 2023. An Introductory Guide to Event Study Models. *Journal of Economic Perspectives*. Vol. 37, no. 2, pp. 203–230. Available from DOI: `10.1257/jep.37.2.203`.

NITSCHKE, Monika; TUCKER, Graeme; HANSEN, Alana; WILLIAMS, Susan; ZHANG, Ying; BI, Peng, 2016. Evaluation of a heat warning system in Adelaide, South Australia, using case-series analysis. *BMJ Open*. Vol. 6, no. 7. Available from DOI: `https://doi.org/10.1136/bmjopen-2016-012125`.

R CORE TEAM, 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available also from: `https://www.R-project.org/`.

RAMBACHAN, Ashesh; ROTH, Jonathan, 2023. A More Credible Approach to Parallel Trends. *Review of Economic Studies*. Vol. 90, no. 5, pp. 2555–2591. Available from DOI: `10.1093/restud/rdad020`.

RAO, Shilpa; CHAUDHARY, Prayash; BUDIN-LJØSNE, Isabelle; SITOULA, Susan; AUNAN, Kristin; CHERSICH, Matthew; DONATO, Francesca de'; KAZMIERCZAK, Aleksandra, 2025. Evaluating the socioeconomic benefits of heat-health warning systems. *European Journal of Public Health*. Vol. 35, no. 1, pp. 178–186. Available from DOI: https://doi.org/10.1093/eurpub/ckae203.

ROBINE, Jean-Marie; CHEUNG, Siu Lan K.; ROY, Sophie Le; OYEN, Herman Van; GRIFFITHS, Clare; MICHEL, Jean-Pierre; HERMANN, Francois Richard, 2008. Death toll exceeded 70,000 in Europe during the summer of 2003. *Nouveautés en cancérogenèse / New developments in carcinogenesis*. Vol. 331, no. 2, pp. 171–178. Available from DOI: https://doi.org/https://doi.org/10.1016/j.crvi.2007.12.001.

ROTH, Jonathan, 2022. Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends. *American Economic Review: Insights*. Vol. 4, no. 3, pp. 305–22. Available from DOI: 10.1257/aeri.20210236.

SCHIFANO, Patrizia; LEONE, Michela; SARIO, Manuela De; DE'DONATO, Francesca; BARGAGLI, Anna Maria; D'IPPOLITI, Daniela; MARINO, Claudia; MICHELOZZI, Paola, 2012. Changes in the effects of heat on mortality among the elderly from 1998-2010: results from a multicenter time series study in Italy. *Environmental Health*. Vol. 11, no. 58. Available from DOI: https://doi.org/10.1186/1476-069X-11-58.

SHADISH, William R.; COOK, Thomas D.; CAMPBELL, Donald T., 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company. ISBN 978-0395615560.

SHERIDAN, Scott C., 2007. A survey of public perception and response to heat warnings across four North American cities: an evaluation of municipal effectiveness. *International Journal of Biometeorology*. Vol. 53, pp. 3–15. Available from DOI: https://doi.org/10.1007/s00484-006-0052-9.

SIRISERIWAN, Wacharasak, 2024. *smotefamily: A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE* [https://CRAN.R-project.org/package=smotefamily]. R package version 1.4.0.

TOLOO, Ghasem; FITZGERALD, Gerard; AITKEN, Peter; VERRALL, Kenneth; TONG, Shilu, 2013. Evaluating the effectiveness of heat warning systems: systematic review of epidemiological evidence. *International Journal of Public Health*. Vol. 58, no. 5, pp. 667–681. Available from DOI: https://doi.org/10.1007/s00038-013-0465-2.

TOMCZYK, Arkadiusz M.; BEDNORZ, Ewa; MATZARAKIS, Andreas, 2020. Human-biometeorological conditions during heat waves in Poland. *International Journal of Climatology*. Vol. 40, no. 12, pp. 5043–5055. Available from DOI: 10.1002/joc.6503.

UN, 2022. *The human cost of disasters: an overview of the last 20 years (2000–2019)*. United Nations Office for Disaster Risk Reduction.

URBAN, Aleš; FONSECA-RODRÍGUEZ, Osvaldo; DI NAPOLI, Claudia; PLAVCOVÁ, Eva, 2022. Temporal changes of heat-attributable mortality in Prague, Czech Republic, over 1982–2019. *Urban Climate*. Vol. 44, p. 101197. ISSN 2212-0955. Available from DOI: https://doi.org/10.1016/j.uclim.2022.101197.

USTRNUL, Zbigniew; WYPYCH, Agnieszka; HENEK, Ewelina; CZEKIERDA, Danuta; WALAWENDER, Jakub; KUBACKA, Danuta; PYRC, Robert; CZERNECK, Bartosz, 2014. *Meteorological Hazard Atlas of Poland*. Instytut Meteorologii i Gospodarki Wodnej Państwowy Instytut Badawczy and Wydawnictwo Attyka. ISBN 978-83-64979-05-7.

VÉSIER, Chloé, 2022. *Social inequalities in heat-related mortality in the Czech Republic*. Master's thesis. Czech University of Life Sciences Prague, Faculty of Environmental Sciences.

WEINBERGER, Kate R.; ZANOBETTI, Antonella; SCHWARTZ, Joel; WELLENIUS, Gregory A., 2018. Effectiveness of National Weather Service heat alerts in preventing mortality in 20 US cities. *Environment International*. Vol. 116, pp. 30–38. Available from DOI: https://doi.org/10.1016/j.envint.2018.03.028.

WELLENIUS, Gregory A.; ELIOT, Melissa N.; BUSH, Kathleen F.; HOLT, Dennis; LINCOLN, Rebecca A.; SMITH, Andy E.; GOLD, Julia, 2017. Heat-related morbidity and mortality in New England: Evidence for local policy. *Environmental Research*. Vol. 156, pp. 845–853. Available from DOI: https://doi.org/10.1016/j.envres.2017.02.005.

WHO, 2008. *Heat-Health Action Plans*. Ed. by MATTHIES, Franziska; BICKLER, Graham; MARÍN, Neus Cardenosa; HILES, Simon. World Health Organization. ISBN 978-92-890-7191-8.

WHO, 2011. *Public Health Advice on Preventing Health Effects of Heat: New and Updated Information for Different Audiences* [https://iris.who.int/bitstream/handle/10665/341580/WHO-EURO-2011-2510-42266-58691-eng.pdf?sequence=1]. World Health Organization.

WHO, 2024. *Heat and health*. World Health Organization, 2024-05-08. Available also from: https://www.who.int/news-room/fact-sheets/detail/climate-change-heat-and-health.

WHO, 2025. *Main Heat Vulnerability Factors*. World Health Organization. Available also from: https://www.who.int/multi-media/details/main-heat-vulnerability-factors.

WIBIG, Joanna, 2017. Heat waves in Poland in the period 1951-2015: trends, patterns and driving factors. *Meteorology Hydrology and Water Management*. Vol. 6, no. 1, pp. 37–45. Available from DOI: https://doi.org/10.26491/mhwm/78420.

WIBIG, Joanna, 2021. Hot Days and Heat Waves in Poland in the Period 1951–2019 and the Circulation Factors Favoring the Most Extreme of Them. *Atmosphere*. Vol. 12, no. 3, p. 340. Available from DOI: 10.3390/atmos12030340.

WILLIAMS, Susan; NITSCHKE, Monika; WONDMAGEGN, Berhanu Yazew; TONG, Michael; XIANG, Jianjun; HANSEN, Alana; NAIRN, John; KARNON, Jonathan; BI, Peng, 2022. Evaluating cost benefits from a heat health warning system in Adelaide, South Australia. *Australian and New Zealand Journal of Public Health*. Vol. 46, no. 2, pp. 149–154. ISSN 1326-0200. Available from DOI: https://doi.org/10.1111/1753-6405.13194.

WING, Coady; SIMON, Kosali; BELLO-GOMEZ, Ricardo A., 2018. Designing Difference in Difference Studies: Best Practices for Public Health Policy Research. *Annual Review of Public Health*, pp. 453–469. Available from DOI: https://doi.org/10.1146/annurev-publhealth-040617-013507.

WMO, 2025. *Heatwave*. World Meteorological Organization, 2025-03-16. Available also from: https://wmo.int/topics/heatwave.

WMO, WHO, 2015. *Heatwaves and Health: Guidance on Warning-System Development*. Ed. by MCGREGOR, Glenn; BESSEMOULIN, Pierre; EBI, Kristie; MENNE, Bettina. ISBN 978-92-63-11142-5.

WU, Xiao; WEINBERGER, Kate R; WELLENIUS, Gregory A; DOMINICI, Francesca; BRAUN, Danielle, 2023. Assessing the causal effects of a stochastic intervention in time series data: are heat alerts effective in preventing deaths and hospitalizations? *Biostatistics*. Vol. 25, no. 1, pp. 57–79. Available from DOI: 10.1093/biostatistics/kxad002.

Appendix

A. Data Overview and Preparation

A.1 Mortality

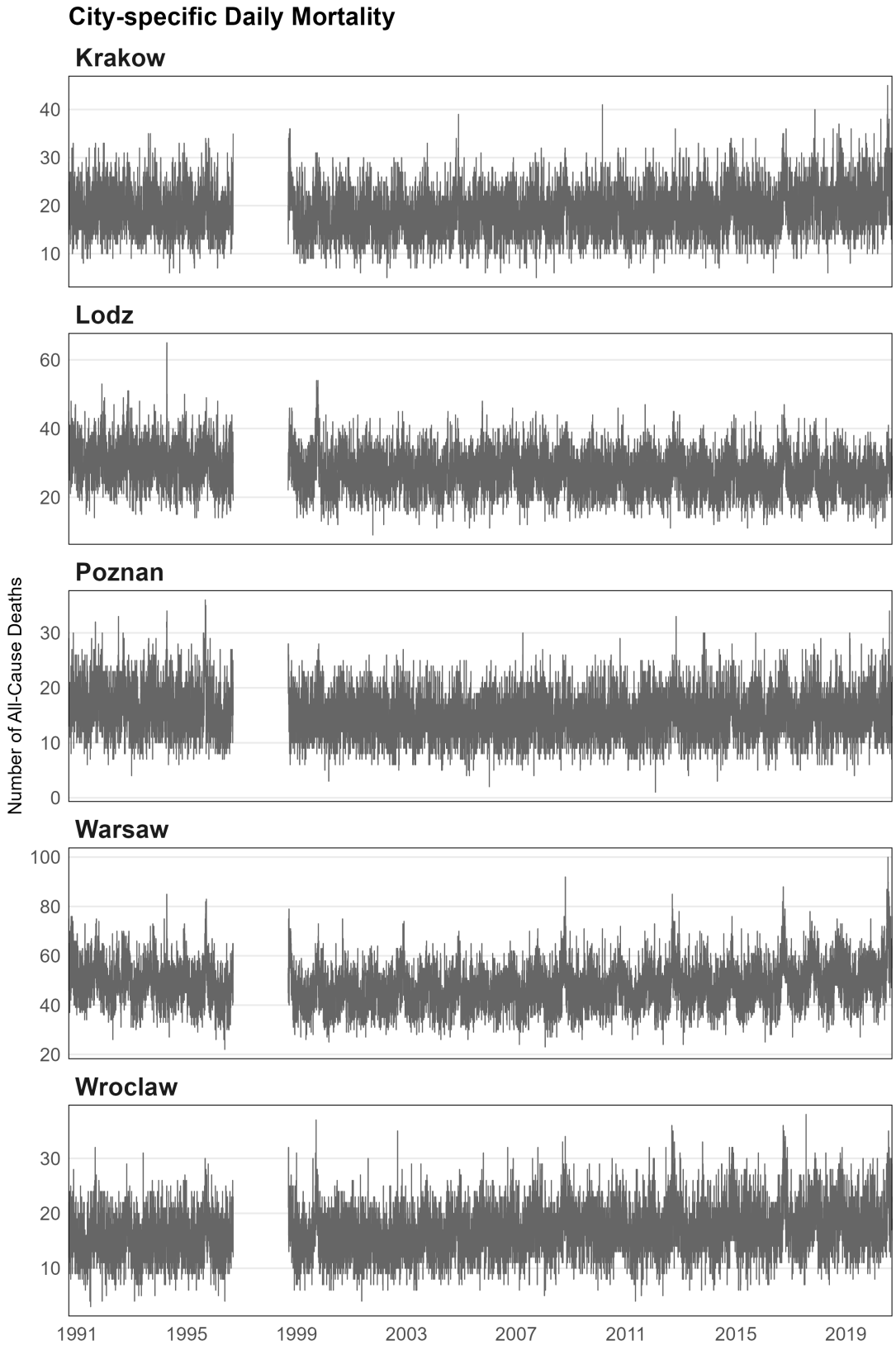


Figure A.1 Complete time-series of daily all-cause mortality for 5 Polish cities. Data spanning 1. 1. 1991 – 31. 12. 2020, with a two-year gap between 1997–1998.

A.2 Temperatures

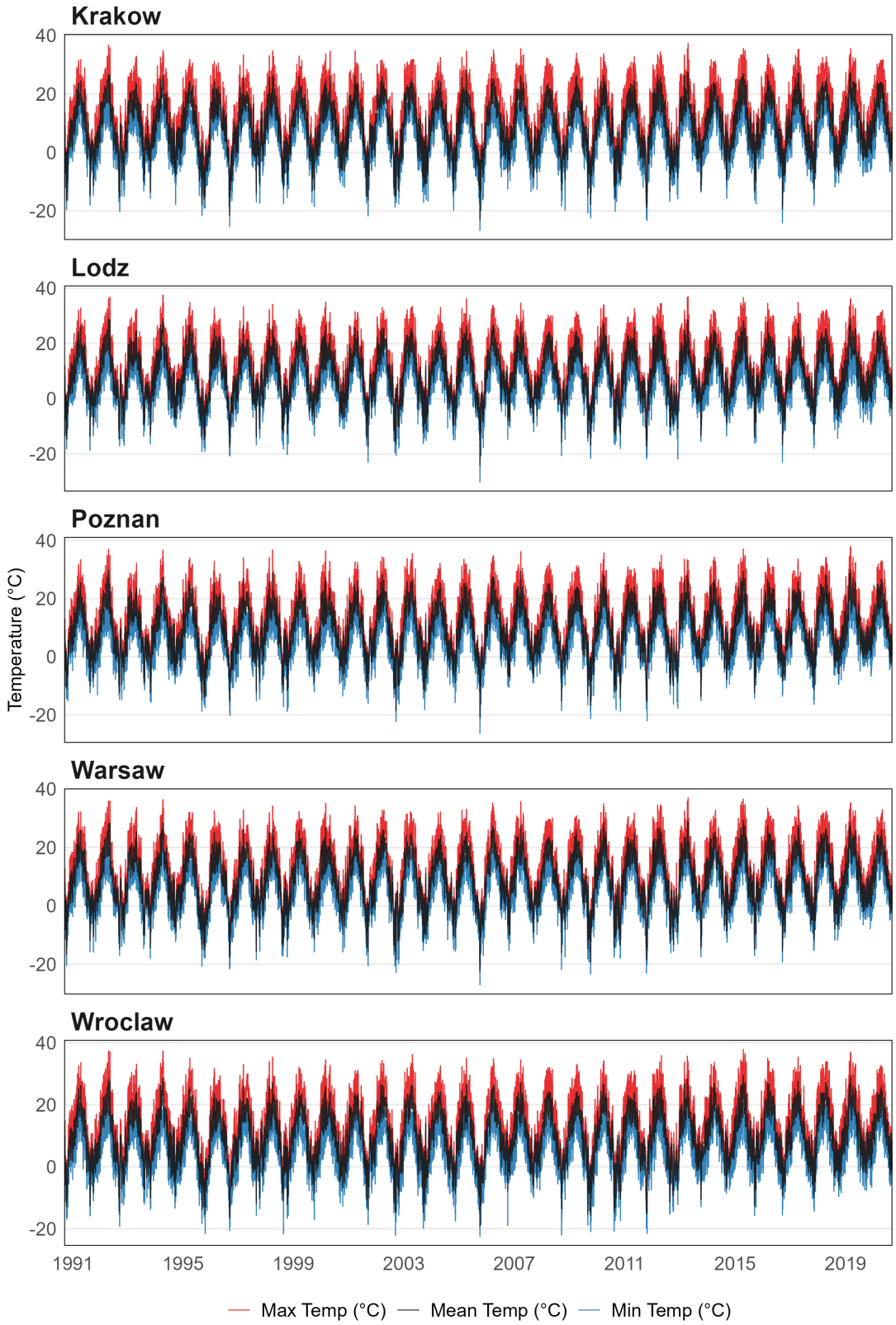


Figure A.3 A time-series of maximum, mean and minimum temperatures for 5 Polish cities cities. Data spanning 1. 1. 1991 – 31. 12. 2020.

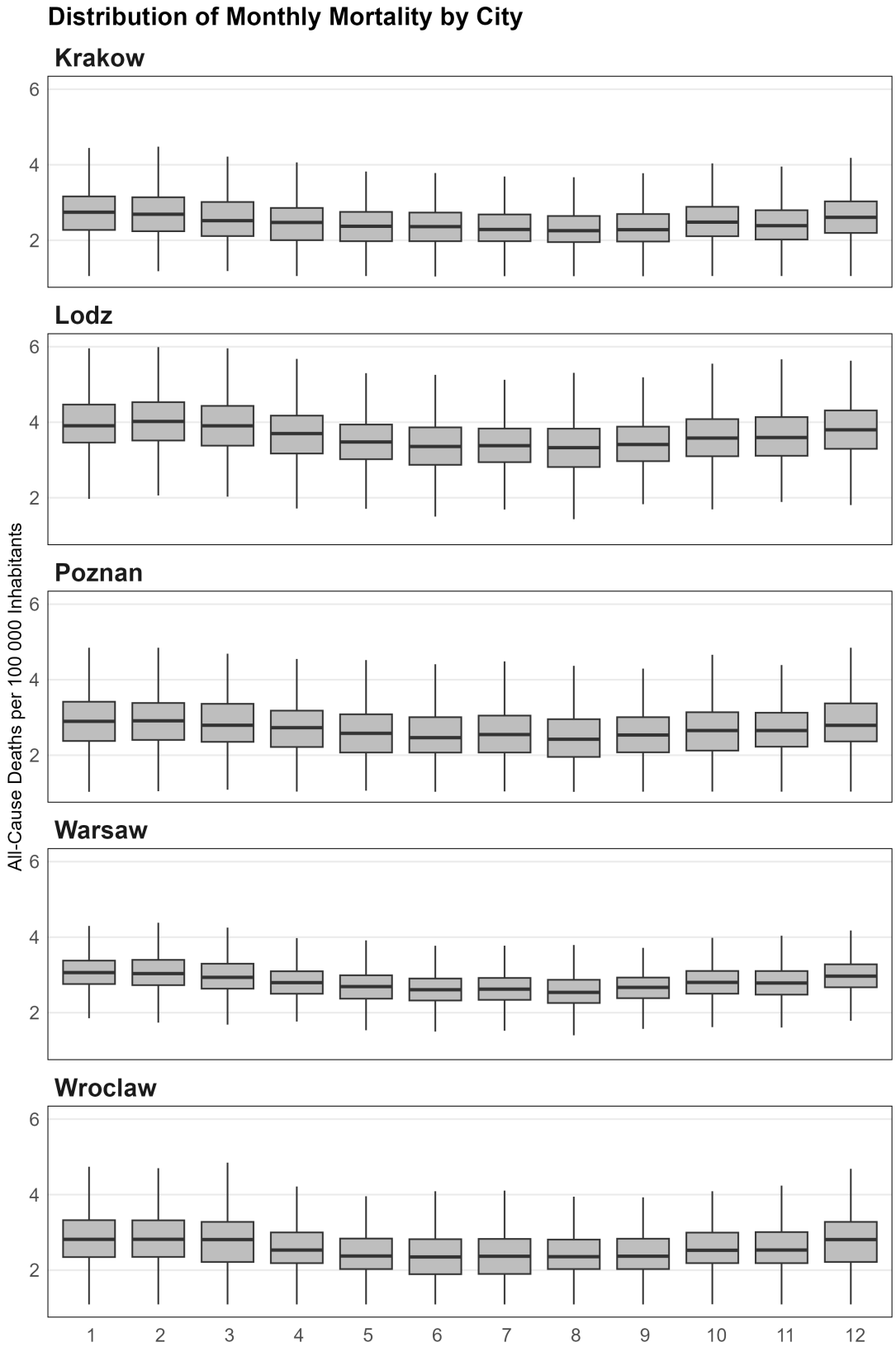


Figure A.2 Distributions of all-cause mortality per 100 000 inhabitants for each month. Yearly population data were taken from Eurostat, 2025a.

A.2.1 Heat Alert Criteria

Table A.1 An overview of requirements for Level 1–3 heat alerts in 6 subperiods following the implementation of city-specific HEWS. T_{max} and T_{min} denote the thresholds for maximum and minimum temperatures, respectively. D represents the minimum number of consecutive days that must satisfy the temperature criteria for a heat alert to be issued.

	Level 1			*Level 2*			*Level 3*		
Period	T_{max}	T_{min}	D	T_{max}	T_{min}	D	T_{max}	T_{min}	D
1. 1. 2009 – 12. 5. 2009	(30, 35]	–	2+	≥ 35	–	1+	–		
				>30	≥ 18	2+			
13. 5. 2009 – 26. 1. 2010	(30, 35]	< 18	2+	(30, 35]	≥ 18	2+	> 35	–	2+
27. 1. 2010 – 10. 5. 2011	(30, 35]	< 20	2+	(30, 35]	≥ 20	2+	> 35	–	2+
11. 5. 2011 – 18. 7. 2017	> 30	–	2–3	> 30	–	4–5	> 30	–	6+
19. 7. 2017 – 5. 5. 2019	≥ 30	< 18	2+	[30, 34]	≥ 18	2+	> 34	≥ 18	2+
6. 5. 2019 – 31. 12. 2020	≥ 30	< 18	2+	[30, 34]	≥ 18	2+	> 34	–	2+
	≥ 35	–	1+						

A.2.2 Heat Alert Data Processing

The original data sheets comprised of 3 types of entries. Entry type *warning* defined a period with an active heat alert (Record 1 in Table A.2). Entry type *warning change* either changed the warning level, or the duration of the heat alert. The new heat alert days were specified by the `warning starts` and `warning ends` (Records 2 and 3 in Table A.2). A combination of changing both the level and the duration of the heat alert was also possible. Entry type *cancellation of warning* canceled all heat alerts in a period defined by `warning starts` and `warning ends` (Record 4 in Table A.2).

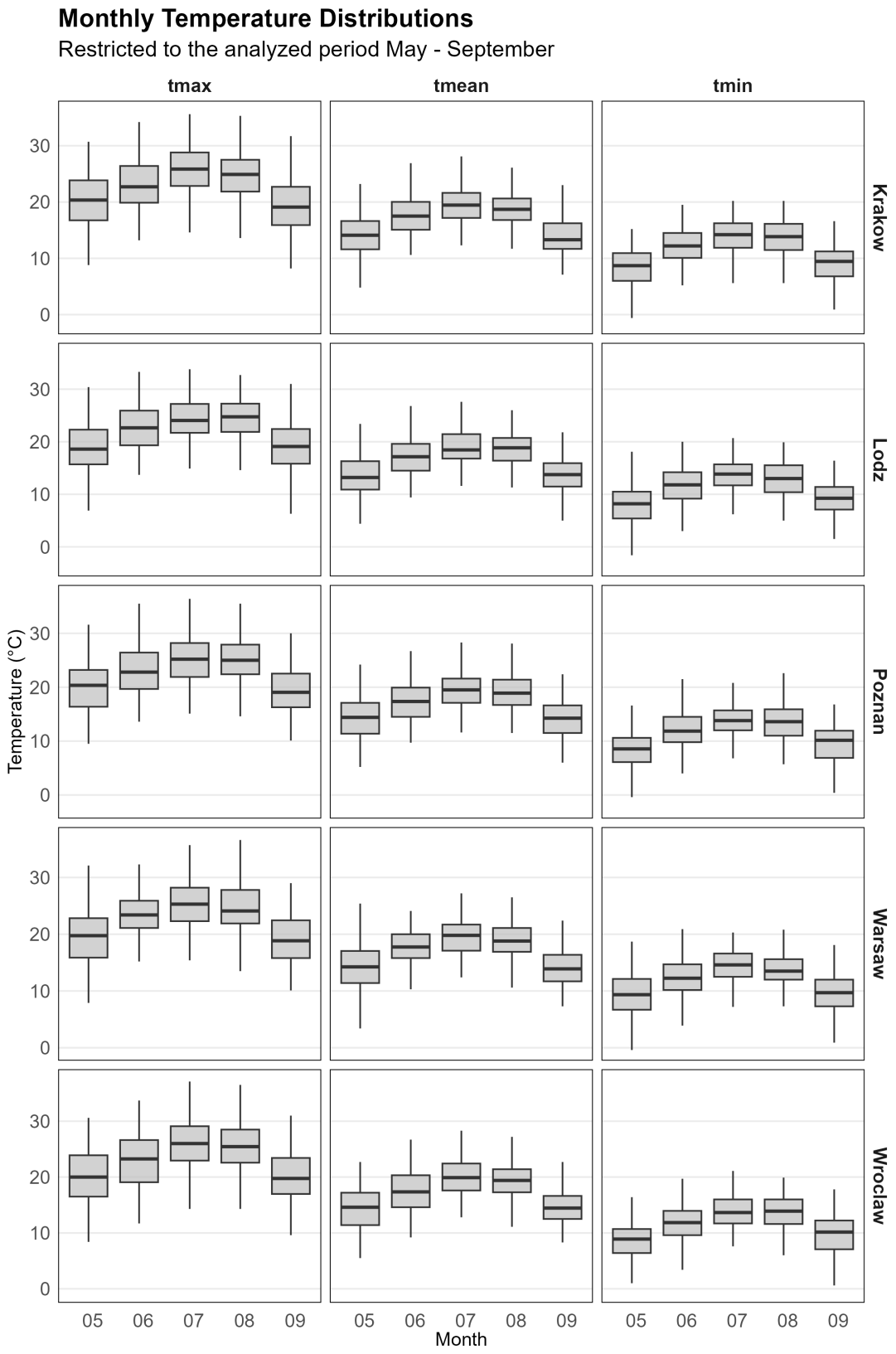


Figure A.4 Monthly distributions of maximum, mean and minimum temperatures for 5 Polish cities. A subset of months from May to September, which were used for the heat alert analysis.

Table A.2 Data on historical heat alerts in Poland were provided in the form of Excel spreadsheets. The table shows an example of four records relating to one streak of heat alert days in Warsaw. Record *1* represents a nationwide heat alert (`area` = Krajowy) and, therefore, appears in the records for each city. Records *2-4* are city-specific (`area` = Warszawa). Records *2* and *3* prolong the alert until the 13th and 16th June, respectively. Record *4* cancels the alert on the 16th. In conclusion, this particular heat alert in Warsaw lasted from the 10th to the 15th of July 2019, and the heat alert level remained unchanged.

Record	***1***	***2***	***3***	***4***
`year`	2019	2019	2019	2019
`area`	Krajowy	Warszawa	Warszawa	Warszawa
`warning issued`	9.6.2019 13:33	11.6.2019 11:53	12.6.2019 14:45	16.6.2019 6:18
`warning starts`	10.6.2019 13:00	11.6.2019 11:53	12.6.2019 14:45	16.6.2019 6:18
`warning ends`	12.6.2019 20:00	13.6.2019 20:00	16.6.2019 20:00	16.6.2019 6:18
`type`	warning	warning change	warning change	cancellation of warning
`level`	2	2	2	2
`city`	Warszawa	Warszawa	Warszawa	Warszawa

All *warning change* and *cancellation of warning* entries were processed, and the duration or level of the corresponding heat alert was changed.

In some records, the `area` and `city` did not match, and both stated a different city[1]. It was possible to attribute most of the mismatched entries to the correct city based on context (i.e., the entry was of type *warning change* and corresponded to an existing heat warning in one of the cities). Entries that were not possible to conclusively attribute to either city were omitted.

Lastly, the tables had to be rearranged in order to obtain a list of individual dates corresponding to active heat alerts. In the analysis, the specific level of the heat warning was not taken into account.

[1]Some records, like Record 1 in Table A.2, were issued as a country-wide alert (`area` = Krajowy). This was not the case.

A.2.3 Heat Alert Data Statistics

Table A.3 Monthly and annual numbers of heat alerts sorted by city. The mean values of the maximum temperatures $\overline{T_{max}}$, mean temperatures $\overline{T_{mean}}$ and minimum temperatures $\overline{T_{min}}$, recorded on heat alert days, are provided.

	Krakow	**Lodz**	**Poznan**	**Warsaw**	**Wroclaw**
Total HA	150	121	125	141	147
May	2	0	2	1	3
June	28	19	29	24	24
July	50	43	38	51	56
August	68	57	54	63	60
September	2	2	2	2	4
2009	4	0	0	0	2
2010	16	16	15	20	16
2011	4	3	5	3	4
2012	19	9	10	17	13
2013	16	13	14	10	14
2014	7	12	15	14	17
2015	30	24	14	21	25
2016	10	5	5	6	7
2017	7	6	3	11	5
2018	14	10	18	13	14
2019	18	18	21	20	21
2020	5	5	5	6	9
$\overline{\mathbf{T_{max}}}$	31.16	31.49	31.51	31.27	31.64
$\overline{\mathbf{T_{mean}}}$	23.65	24.41	25.00	24.72	24.45
$\overline{\mathbf{T_{min}}}$	16.56	17.17	18.46	18.40	16.93

A.3 Dataset Variables

Table A.4 Variable names used throughout the analysis.

Variable name	Description
`date`	
`year`	
`month`	
`day`	
`yday`	Day of Season
`dow`	Day of Week
`tmean`	Mean Daily Temperature (°C)
`tmax`	Maximum Daily Temperature (°C)
`tmin`	Minimum Daily Temperature (°C)
`death`	Daily All-Cause Deaths
`hw`	Heat Alert Indicator (yes/no)
`D2tmean`	2-day rolling mean of T_{mean}
`D2tmax`	2-day rolling mean of T_{max}
`D2tmin`	2-day rolling mean of T_{min}
`D3tmean`	3-day rolling mean of T_{mean}
`D3tmax`	3-day rolling mean of T_{max}
`D3tmin`	3-day rolling mean of T_{min}
`basic.predict`	`year` $<$ 2009: "basic" model prediction `year` $\geq$ 2009: Equal to `hw`
`synth.predict`	`year` $<$ 2009: "synthetic" model prediction `year` $\geq$ 2009: Equal to `hw`
`threshold.predict`	`year` $<$ 2009: threshold-based eligibility `year` $\geq$ 2009: Equal to `hw`

B. Supplementary Analyses

B.1 Eligibility Based on Heat Alert Threshold Criteria

The reliability of flagging pre-implementation days as eligible for a heat alert based on heat alert criteria was tested on post-implementation data. Between 2009–2020, the criteria changed multiple times. The information on heat alert thresholds were provided together with the heat alert archive (see Table A.1).

In the five Polish cities, a total of 694 heat alerts were issued between 2009–2020. Applying the threshold criteria successfully matched 464 heat alert days, leaving more than 30 % of actual heat alert days unidentified. Moreover, 96 days were marked as heat alert days despite an actual heat alert not being in place. The city-specific predictions, along with predictions made by forest model classifiers, are stated in Table 3.2.

Due to the high number of unidentified (false negative) heat alert days, the alternative approach, training a classifier model to flag days eligible for a heat alert, was selected.

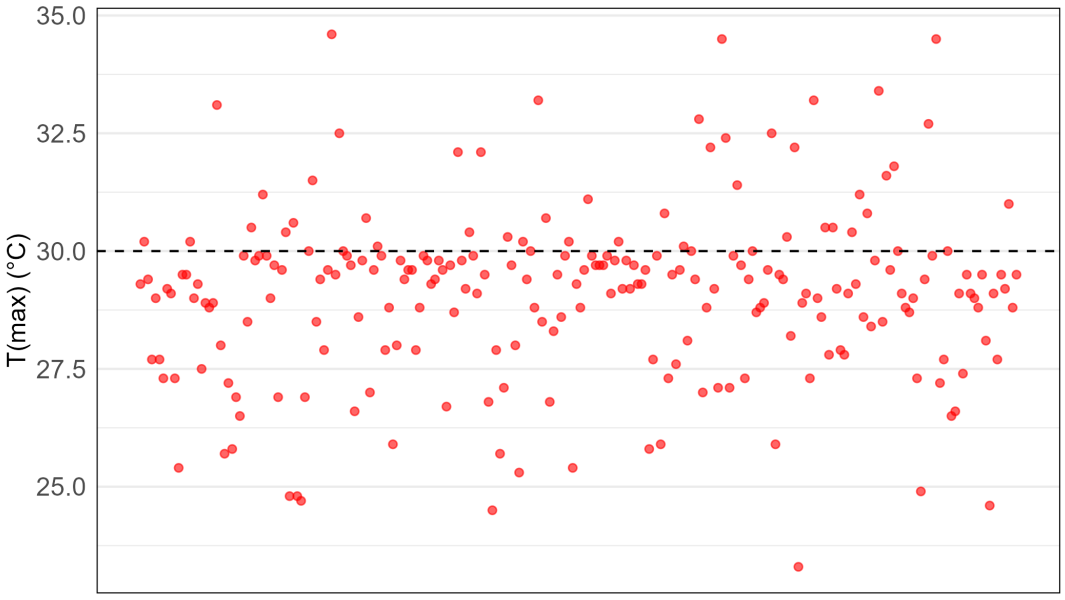


Figure B.1 Maximum temperatures observed on days with heat alerts that were not recognized as such by the criteria-based method. In most instances, the recorded maximum temperature remained below 30 °C. The x-axis represents the sequence number of data points rather than specific dates.

B.2 Controlling for the changing heat alert criteria

In Poland, the criteria for issuing heat alerts are based on forecast temperatures. The default condition remained consistent in all periods: a heat alert was triggered if maximum temperatures exceeded 30 °C for two consecutive days. The classification into various levels of heat alerts was based on additional conditions.

The individual periods defined in Table A.1 differ in whether $T_{max} = 30\,°C$ is included in the default heat alert condition, a difference that is likely more administrative rather than practical.

Two periods included a supplementary criterion for days where the maximum temperatures were expected to exceed 35 °C despite not being part of a 2+ day series of hot days. However, in the analyzed period, such a day was always either preceded or followed by a day where $T_{max} \geq 30\,°C$. Based on these findings, the pattern in assigning heat alert days can be considered consistent during the whole period after HEWS implementation, and no further control for a specific sub-period is needed.

B.3 Heat Alert Outliers

Outliers identified based on the interquartile range criterion (Eq. B.1) were omitted from the group of post-implementation eligible days, and were marked as not having a heat alert. The criterion was applied to T_{max} and points below the IQR threshold were omitted.

The range between the 25$^{\text{th}}$ and 75$^{\text{th}}$ percentile was

$$\text{IQR} = 2.9\,^{\circ}\text{C},$$

and the 25$^{\text{th}}$ percentile corresponded to a temperature of

$$q_{0.25} = 29.9\,^{\circ}\text{C}.$$

Therefore, the IQR criterion

$$T_{max} \leq q_{0.25} - 1.5 \cdot \text{IQR} \tag{B.1}$$

set the lower cutoff value as 25.55 °C. Table B.1 provides a complete list of entries reclassified as not having a heat alert.

Table B.1 Days with heat alerts where T_{max} was below the set interquartile range. These points were labeled as not having a heat alert.

City	Date	Deaths	T_{max} [°C]	T_{mean} [°C]	T_{min} [°C]
Wroclaw	25. 7. 2012	15	25.4	21.0	15.8
Warsaw	7. 6. 2011	49	24.8	20.9	17.3
Warsaw	24. 8. 2011	37	24.8	18.7	16.0
Warsaw	25. 8. 2011	56	24.7	19.3	15.4
Poznan	14. 8. 2010	9	24.5	20.8	18.8
Poznan	4. 8. 2012	15	25.3	19.4	15.2
Poznan	4. 8. 2014	8	25.4	20.9	18.1
Krakow	28. 7. 2016	26	23.3	20.6	18.0
Lodz	25. 8. 2011	25	24.9	19.6	16.3
Lodz	29. 8. 2015	22	24.6	18.7	15.3

B.4 Hyperparameter Optimization

B.4.1 Decision Tree Parameters

Table B.2 Optimization of the model trained on data up-sampled using the **basic** up-sampling method (duplication): values of `mtry` with the highest 10-fold cross-validation accuracy.

		maxnodes										
		5	**10**	**50**	**100**	**200**	**500**	**1000**	**2000**	**5000**	**7500**	**10^4**
nodesize	**1**	3	5	5	5	5	5	3	3	4	3	3
	10	4	5	5	5	5	5	5	4	4	4	3
	20	4	5	5	5	4	5	5	5	5	5	4
	30	4	5	4	5	5	5	5	5	5	5	5
	50	3	5	5	5	5	4	5	5	5	4	4

Table B.3 Optimization of the model trained on data up-sampled using the **synthetic** up-sampling method (using K nearest neighbors): values of `mtry` with the highest 10-fold cross-validation accuracy.

		maxnodes										
		5	**10**	**50**	**100**	**200**	**500**	**1000**	**2000**	**5000**	**7500**	**10^4**
nodesize	**1**	3	5	5	5	5	4	4	4	5	3	3
	10	3	4	5	5	5	3	5	4	4	5	4
	20	3	4	5	5	5	4	4	3	5	4	4
	30	3	5	5	5	5	5	5	5	5	5	5
	50	3	5	5	5	5	4	5	5	4	5	5

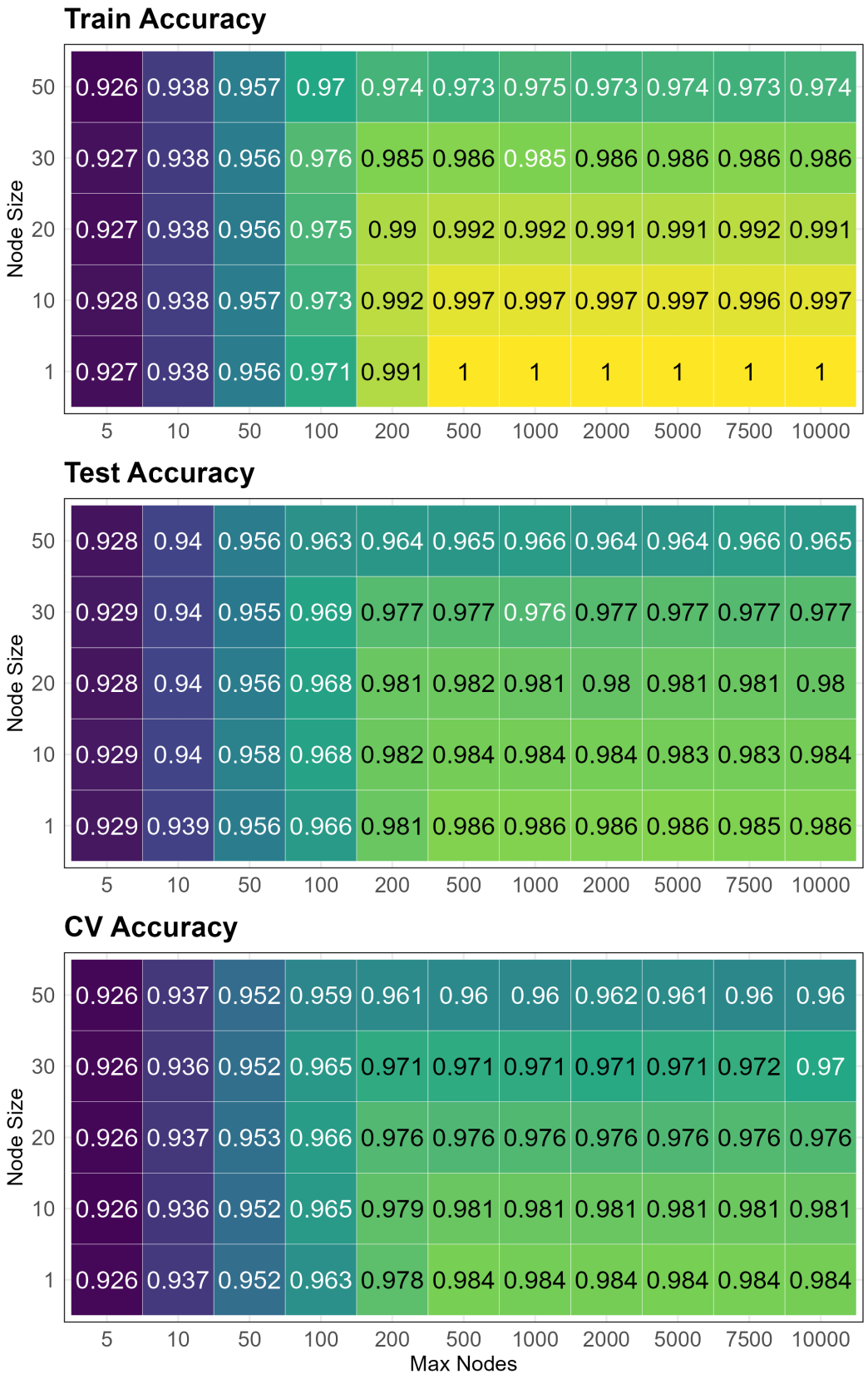


Figure B.2 The dataset for training the eligibility classifier, up-sampled using the **basic** duplication method, was partitioned into training and testing subsets in a 70:30 ratio. The reported training and testing accuracy values for different `nodesize` and `maxnodes` settings correspond to the model with the best performing `mtry` parameter. This parameter was selected using the `train` function based on cross-validation (CV) accuracy.

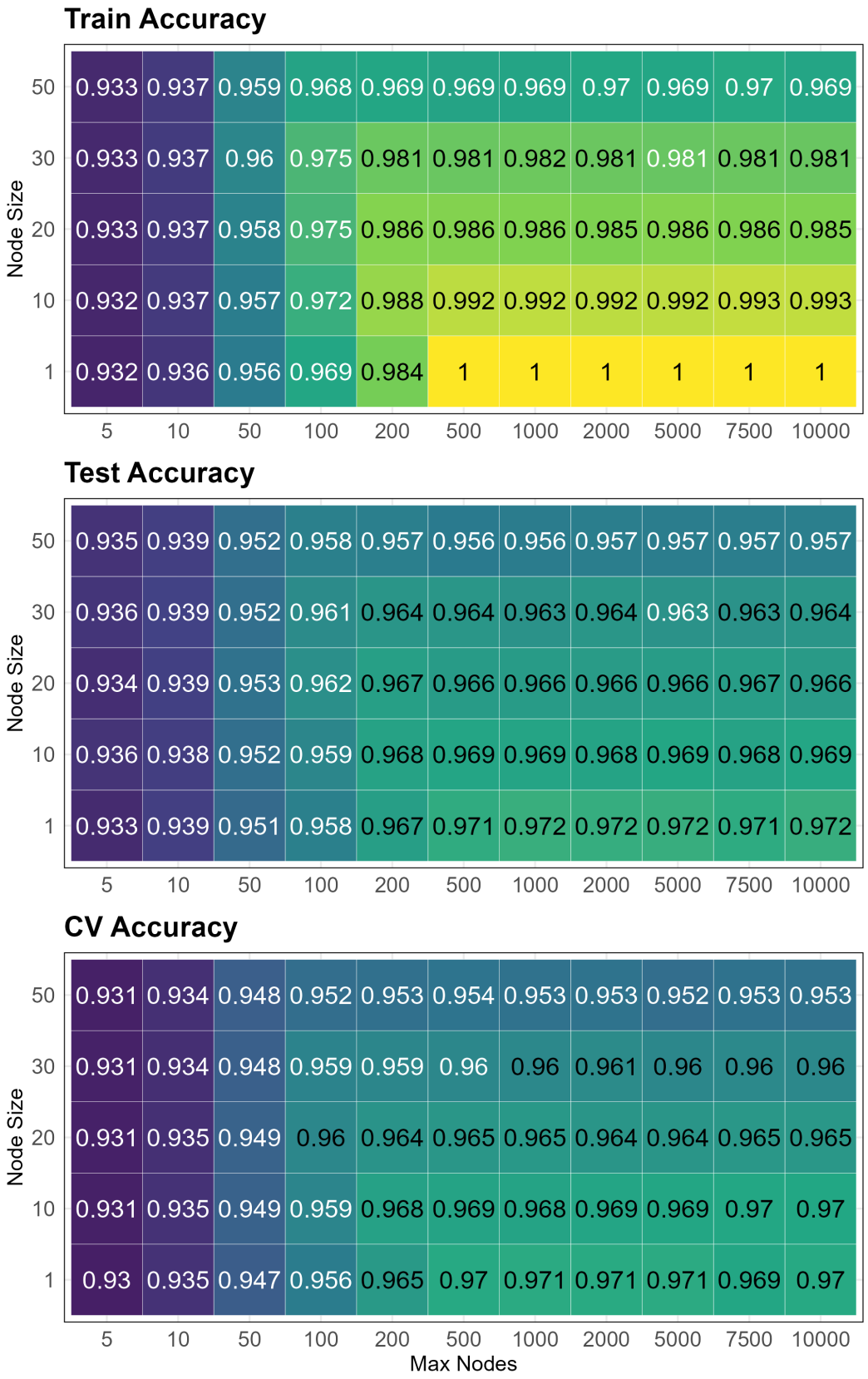


Figure B.3 The dataset for training the eligibility classifier, up-sampled using the **synthetic** duplication method, was partitioned into training and testing subsets in a 70:30 ratio. The reported training and testing accuracy values for different `nodesize` and `maxnodes` settings correspond to the model with the best performing `mtry` parameter. This parameter was selected using the `train` function based on cross-validation (CV) accuracy.

B.4.2 P-Values

Table B.4 The p-values of predictions made on the original post-implementation dataset by a random forest classifier trained on data up-sampled using the basic approach. Results for different combinations of decision tree parameters. All models where `maxnodes` ≥ 100 had p-value < 0.05.

		maxnodes									
		5	**10**	**50**	**100**	**200**	**500**	**1000**	**2000**	**5000**	**7500**	**10^4**
nodesize	**1**	1	1	0.97								
	10	1	1	0.96								
	20	1	1	0.84					<0.05			
	30	1	1	0.73								
	50	1	1	0.51								

Table B.5 The p-values of predictions made on the original post-implementation dataset by a random forest classifier trained on data up-sampled using the synthetic approach. Results for different combinations of decision tree parameters. All models where `maxnodes` ≥ 100 had p-value < 0.05.

		maxnodes									
		5	**10**	**50**	**100**	**200**	**500**	**1000**	**2000**	**5000**	**7500**	**10^4**
nodesize	**1**	1	1	0.89								
	10	1	1	0.92								
	20	1	1	0.37					<0.05			
	30	1	1	0.07								
	50	1	1	<0.05								

B.5 Parallel Trend Control

For the test of parallel trends between the eligible and non-eligible days in the pre-implementation period, a GAM quasi-Poisson model was applied. The model for trends in daily mortality was formulated as follows:

$$\begin{aligned} Mortality \sim\ & T_g \\ & + T_g \times ns_2(Date) + T_g \times ns_4(DoS) \times Year + T_g \times DoW \\ & + ns_2(Date) + ns_4(DoS) \times Year + DoW, \end{aligned} \tag{B.2}$$

where T_g is the group indicator defined in Equation (2.1), ns_k indicates a natural spline with k degrees of freedom, DoS is Day of Season and DoW is Day of Week. The first term in Equation (B.2) represents the offset in trends between eligible and non-eligible days. The interaction terms between the time predictors and group indicator reflect additional, variable contributions to mortality on eligible days. If the parallel trend assumption is satisfied, the coefficients of these terms should be near zero and low in significance reflected by a high p-value. The time variables without interaction terms reflect the periodical variation in mortality, and the same time trends were used in the final DID models. All notably significant coefficients are listed in Table B.6.

Table B.6

T_g Assignment	City	Coefficient	p-value
Basic up-sampling	*Krakow*	$T_g \times 1999$	0.03
		$T_g \times DoS_4 \times 1999$	0.01
	Lodz	$T_g \times DoS_1 \times 1993$	0.01
		$T_g \times DoS_2 \times 1993$	0.01
Synth. up-sampling	*Lodz*	$T_g \times 1993$	$\ll 0.05$
		$T_g \times 2001$	0.04
		$T_g \times$ Monday	0.02
		$T_g \times$ Tuesday	0.02
		$T_g \times DoS_1 \times 1993$	$\ll 0.05$
		$T_g \times DoS_2 \times 1993$	$\ll 0.05$
		$T_g \times DoS_3 \times 1993$	$\ll 0.05$
		$T_g \times DoS_1 \times 2001$	0.04
		$T_g \times DoS_2 \times 2001$	0.04
		$T_g \times DoS_3 \times 2001$	0.04
		$T_g \times DoS_4 \times 2001$	0.04
	Warsaw	$T_g \times$ Monday	0.02