

UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S
THESIS

Developing an Early-warning system for Acute Myocardial Infractions in Catalonia

Author:

Gabriela ZEMENČÍKOVÁ

Supervisor(s):

Prof. Xavier Rodó, PhD,
Alejandro Fontal, MSc,
Laura Igual Muñoz, PhD

*A thesis submitted in partial fulfilment of the requirements
for the degree of MSc in Fundamental Principles of Data Science*

in the

Facultat de Matemàtiques i Informàtica

June 30, 2024

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Developing an Early-warning system for Acute Myocardial Infarctions in Catalonia

by Gabriela ZEMENČÍKOVÁ

We aim to investigate the association between temperature, humidity, pollution and the incidence of acute myocardial infarction (AMI) in Catalonia, Spain and set up a new predictor scheme. While it is established that both hot and cold temperatures influence AMI incidence, the relationships between other environmental variables and AMI, particularly across different seasons, are less well understood. Furthermore, a predictive modelling scheme is yet lacking. A dataset with 22,812 hospital admissions at the scale of the 948 municipalities in Catalonia stratified by province, sex and age is available for analysis during the interval 2010-2018, together with daily average temperature, pollution and humidity values. We employ two modelling approaches: Seasonal Autoregressive Integrated Moving Average (SARIMA) models and Long Short-Term Memory (LSTM) neural networks. The analysis identifies key environmental predictors of AMI incidence. Our findings underscore the importance of considering environmental factors in public health strategies to reduce AMI risk, offering a foundation for future studies and policy-making.

Acknowledgements

Finishing this project would have been impossible without my supervisors, who have provided amazing guidance, support, and interesting insights throughout the project; Nacho, who provided the best coffee and climbing sessions; Ro, who dragged me to the beach occasionally so I wouldn't get too crazy; Karolina, who has been very kind to listen to my rants; my family, who still has no clue what I am actually studying; and everyone else, who puts up with me barely being a person these days. I want to remember a song when I find a printed version of this report in 30 years: *You took your time* by Mount Kimbie and King Krule.

Contents

Abstract	ii
Acknowledgements	iii
Contents	iv
List of Figures	vi
List of Tables	vii
List of Abbreviations	viii
1 Introduction	1
2 Data Analysis	3
2.1 Data Sources	3
2.2 Data Exploration	4
2.2.1 Sex and Age as Significant Factors	5
2.2.2 Choice of Age-Standardized Incidence Rates (ASIR)	6
2.2.3 Spatial Distribution of AMI Incidence	8
2.2.4 Seasonal Variations and Weather Patterns	12
3 Methodology	15
3.1 ARIMA, SARIMA and SARIMAX Models	15
3.2 The Recurrent Neural Network	17
3.2.1 Structure of an LSTM Cell	18
3.2.2 Working Mechanism	19
3.2.3 LSTM in practice	20
3.2.4 Hyperparameters optimisation	21
4 Results	23
4.1 Model Comparison and Results	23
4.1.1 SARIMA/SARIMAX Model Results Summary	23
4.1.2 LSTM Model Results Summary	25
4.2 Feature Importance	26
4.2.1 Permutation Feature Importance	26
4.2.2 Feature Importance from Model Coefficients	28
4.3 Prediction skill vs Lead-time	29
4.4 Limitations	30
5 Conclusion and future directions	31
5.1 Conclusion	31
5.2 Future Directions	32

A Code	33
B Map of Catalonia split by Regions	34
C Descriptive Statistics	35
Bibliography	39

List of Figures

2.1	Missing data in pollutants Over time	4
2.2	Spatial Distribution of Monitoring Stations	4
2.3	Age and sex distribution of all AMI alerts (2010-2018)	6
2.4	Development of Population (2010 - 2018)	7
2.5	Development of Incidence Rate per 100k (2010 - 2018)	8
2.6	Spatial Development of AMI alerts per 100k	10
2.7	Spatial Development of Age-corrected Population Risk Factor	11
2.8	Moving Averages of ASIR Over Time	12
2.9	Correlation plot	13
3.1	ACF and PACF plots	17
3.2	A single RNN Cell	18
3.3	A single LSTM Cell	20
4.1	SARIMA vs SARIMAX predictions	24
4.2	LSTM predictions with exogenous variables	27
4.3	Feature Importance in LSTM Model with Exogenous Variables	28
4.4	Prediction Skill vs Lead-time	30
B.1	Map of Catalonia split by Regions	34
C.1	Decomposition of ASIR	35
C.2	Development of variables over years for Catalonia on monthly basis	37

List of Tables

2.1	AMI alerts by AT and year	12
2.2	Variables for Modeling	14
4.1	Performance Comparison between SARIMA and SARIMAX Models .	25
4.2	Model Architecture	26
4.3	Evaluation metrics for LSTM (Exogenous Variables)	26
4.4	Feature Importance based on Mean Absolute Error (MAE)	28
4.5	Prediction Skill vs Lead Time	29
C.1	Age and Sex-Specific Trends in AMI Incidence with Statistical Significance	36
C.2	Descriptive statistics for daily levels of meteorological variables and air pollutant levels (Lag0) in Catalonia.	38

List of Abbreviations

ADF	Augmented Dickey-Fuller
AMI	Acute Myocardial Infarction
ASIR	Age-Standardized Incidence Rates
CVD	Cardiovascular Disease
ERF	Exposure-Response Function
EWS	Early Warning System
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
NO₂	Nitrogen Dioxide
PM	Particulate Matter
RNN	Recurrent Neural Networks
RMSE	Root Mean Square Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SARIMAX	Seasonal Autoregressive Integrated Moving Average with Exogenous Variables
SHAP	Shapley Additive Explanations
SMBO	Sequential Model-Based Optimization

Chapter 1

Introduction

Cardiovascular diseases (CVDs) remain a significant public health concern globally. They contribute significantly to mortality rates, causing almost one-third of all deaths (Cesare et al., 2024; Lindstrom et al., 2022). Among the various manifestations of CVDs, acute myocardial infarction (AMI) is one of the leading causes of mortality, requiring prompt medical attention and intervention. It is defined as myocardial necrosis, which occurs due to decreased coronary blood flow, leading to insufficient oxygen supply to the heart and cardiac ischemia (Mechanic, Gavin, and Grossman, 2024). The interaction between environmental factors having an effect on AMI incidence has gained increased attention. Understanding the underlying physiological causes of mortality rates is crucial for developing additional effective strategies.

In regions characterised by diverse climatic conditions, where temperature fluctuations, humidity levels, and pollution levels exhibit considerable variability across different seasons and geographical locations, exploring the nexus between environmental parameters and the occurrence of AMI becomes especially relevant. Catalonia, with its coastal and inland areas, urban centres, rural landscapes, mountains, and sea, provides a compelling setting for investigating these associations.

The objective of this study is to explain the association between temperature, humidity, pollution, and the incidence of AMI in Catalonia, Spain. The analysis uses a dataset that includes hospital admissions from 2010 to 2018 in 948 municipalities, categorised by province, sex, and age. It also includes daily meteorological and pollution data to help understand how these variables interact.

While existing literature indicates that both hot and cold temperatures have an impact on the incidence of AMI (Mohammad et al., 2018; Bhaskaran et al., 2010), the relationship between other environmental variables and AMI, particularly across different seasons, remains less clear. Furthermore, despite advancements in statistical modelling techniques, a comprehensive predictive modelling scheme well-fitted for Catalonia's context has yet to be established.

We contribute to the existing literature by addressing these gaps through a multifaceted approach that incorporates two different methodologies, specifically Seasonal Autoregressive Integrated Moving Average Models (SARIMA) and Long Short-Term Memory (LSTM) machine learning algorithms. Combining these models with a detailed exploratory analysis of the given dataset, we aim to develop a robust predictive model that captures the relationships between environmental factors and the incidence of AMI in Catalonia.

Considering climate change, there has been a discussion about the relationship between environmental factors and the incidence of AMI. As changing climatic conditions impact cardiovascular health, through this interdisciplinary approach, we

aim to enhance our understanding of the environmental determinants of AMI. Moreover, this thesis should provide policymakers and healthcare practitioners with valuable insights to mitigate the issue of cardiovascular diseases in Catalonia and address the dynamic challenges posed by climate change on public health.

Furthermore, Catalonia's diverse geographical and demographic landscape provides an opportunity to investigate potential spatial and demographic variations in the relationship between environmental factors and AMI incidence. By stratifying our analysis by province, sex, and age groups, there is a potential to explore any disparities or differential susceptibility to environmental variables and potential triggers across different population segments. As a result, understanding the dynamics can help create interventions aimed at the most susceptible group and reduce the incidence and impact of AMI in Catalonia. Additionally, looking closer at the predictive power of models and assessing the reliability of the predictions with increasing lead-time is crucial for developing effective early warning systems (EWS), which would require that the predictions remain reliable during the period.

This paper is structured to provide a comprehensive analysis of the relationship between environmental factors and AMI incidence. A section 2 begins with a detailed data analysis, describing the sources of our hospital admissions, meteorological, and pollution datasets, followed by the data wrangling process and exploratory data analysis to uncover initial patterns and relationships. A section 3 is a methodology section that outlines the theoretical framework of the SARIMA and LSTM models used in this study, along with the hyperparameter optimisation process to enhance model performance. In the results section, section 4, we present the evaluation of the models, including a comparison of suitable models based on various performance metrics, and discuss the predictive capabilities of each model while identifying key environmental variables influencing AMI incidence and look closer at the predictive power of models with increasing lead-time. We then acknowledge the limitations of our study, such as potential data quality issues and the assumptions inherent in our modelling approaches. Finally, in section 5, the paper summarises our findings, their implications for public health strategies, and suggestions for future research to further investigate the dynamics between environmental factors and cardiovascular health outcomes.

Chapter 2

Data Analysis

2.1 Data Sources

The dataset used in the analysis combines data on meteorological, air quality, and population variables. Various data sources were used for the study. The environmental dataset under investigation is composed of Meteorological Data - temperature, relative humidity; AirQuality Data - air pollution data, and Census Data - population estimates; all were obtained from the Statistical Institute of Catalonia. The census data was obtained from the population estimates, which provide updates every six months at the county level for 5-year age groups stratified by sex.

Air pollution metrics were collected from a network of 90 monitoring stations dispersed across Catalonia, as shown in Figure 2.2a. These stations provide comprehensive coverage of atmospheric conditions across the regions. Hourly readings of each pollutant were aggregated to daily observations for analysis. There were various contaminants measured, specifically benzene (C₆H₆), chlorine (Cl₂), carbon monoxide (CO), hydrogen sulfide (H₂S), mercury (Hg), nitric oxide (NO), nitrogen dioxide (NO₂), nitrogen oxides (NOx), ozone (O₃), particulate matter (PM), and sulfur dioxide (SO₂). However, only O₃, NO₂, CO, SO₂ and PM10 were selected for the study due to the inconsistent measurement of all contaminants across stations and a significant amount of missing data for some pollutants (for more details, see Figure 2.1).

Similarly, meteorological observations, including outdoor temperature and relative humidity, were collected from 239 meteorological stations distributed throughout Catalonia, as illustrated in Figure 2.2b. Initially recorded at half-hour intervals, raw measurements were aggregated to daily data points to perform the analysis. There were no missing data.

The AMI dataset was obtained from ten main hospitals across Catalonia. Even though the dataset contains detailed spatial information, including the postal code of residence for each patient, it is necessary to aggregate the data to a higher level of spatial resolution to recognise meaningful patterns. Furthermore, the data is stratified by age and sex of each hospitalised patient. Our initial approach involved aggregating the data to the county-level (comarques) and higher-level territories (àmbits territorials) within Catalonia. However, we encountered limitations when working with county-level data, as some counties yielded sparse observations, often with only one or no cases of AMI recorded per day for many counties. This data density limited our ability to extract meaningful insights and detect significant trends. As a reason, we used higher-level regional territories. For a more detailed explanation, see section 2.2.3.

FIGURE 2.1: Missing data in pollutants Over time

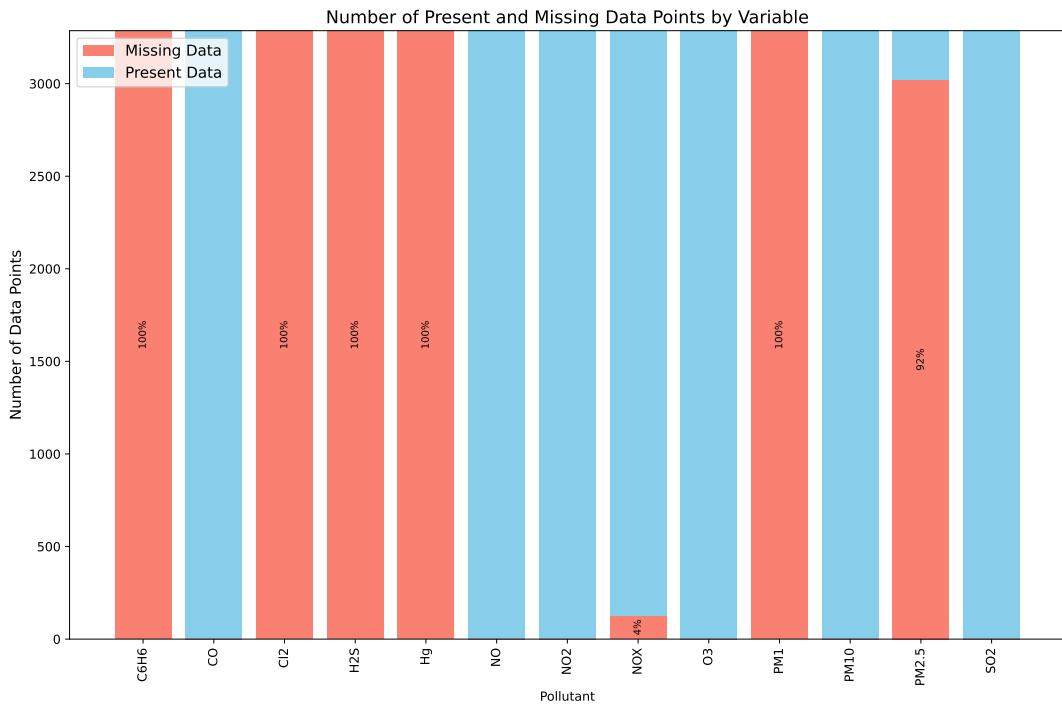
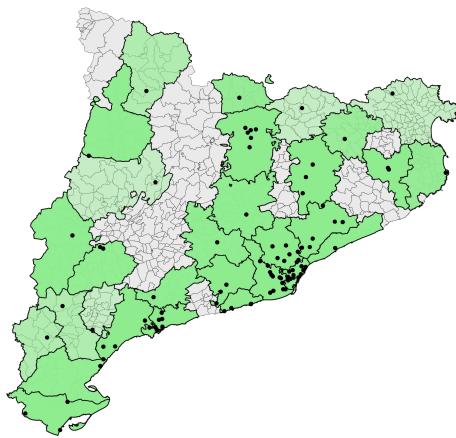
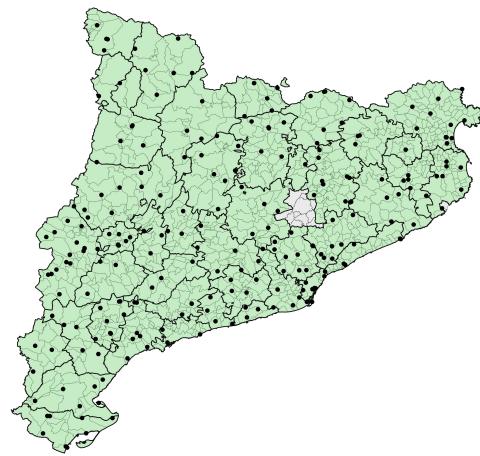


FIGURE 2.2: Spatial Distribution of Monitoring Stations

(A) Spatial Distribution of Air Pollution



(B) Spatial Distribution of Meteorological variables



2.2 Data Exploration

The complex dynamics of factors influencing the incidence of AMI necessitates a multifaceted approach. This exploratory analysis delves into several key factors that may contribute to variations in AMI incidence within Catalonia, Spain. As noted by Mechanic, Gavin, and Grossman (2024), there are many risk factors potentially contributing to the occurrence of AMI. Considering the set of non-modifiable risk factors such as sex and age. The associations of these variables were further analysed

by Canto et al. (2012).

2.2.1 Sex and Age as Significant Factors

The role of sex as an essential factor in the epidemiology of AMI is well established. Sexual dimorphism in cardiovascular physiology and pathophysiology underpins the differential risk profiles observed in males and females (Zhang et al., 2012). While men traditionally exhibit a higher overall incidence of AMI, women often present with AMI at older ages (as they tend to live longer) and experience worse outcomes, including higher mortality rates. Furthermore, women are more susceptible to the effects of air pollution, given sex differences in air pollution lung deposition, and because a more significant proportion of women have airway hyperresponsiveness than men. Hormonal influences, anatomical differences, and variations in risk factor prevalence contribute to these sex-specific disparities (Čulić et al., 2002). To explore the role of sex in AMI incidence within Catalonia, it is necessary to aggregate data by sex and examine potential differences in risk factor profiles between males and females.

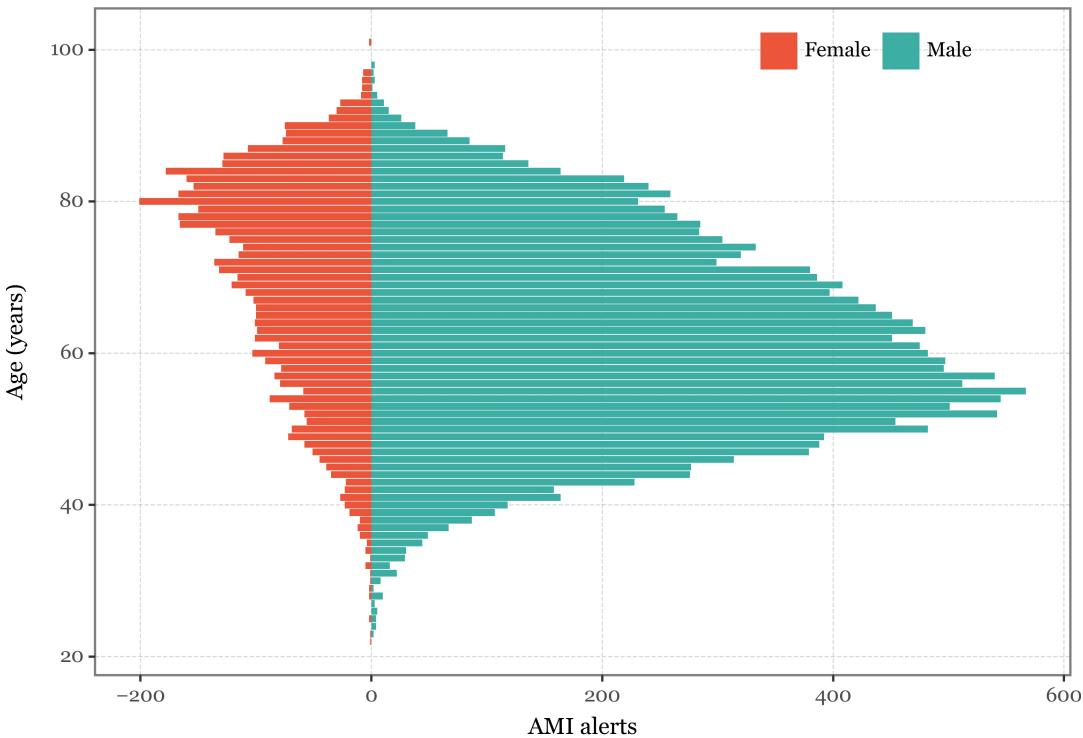
Age is a significant predictor of AMI incidence. Exploratory analysis, together with previous academic findings, suggests that certain age groups are more susceptible to the effects of environmental factors. The risk increases with advancing age. Age-related changes in cardiovascular structure and function, coupled with the cumulative impact of traditional risk factors such as hypertension, diabetes, and dyslipidemia, contribute to the elevated AMI risk observed in older adults (Mehta et al., 2001). In order to explore age-specific trends in AMI incidence within Catalonia, it is necessary to stratify the population into distinct age groups and analyse age-specific incidence rates, temporal trends, and clinical characteristics. This will enable the identification of age-specific risk factors and vulnerabilities.

Figure 2.3 displays the AMI alerts distribution by sex and age. We can observe that the distribution differs when comparing both genders. The highest number of AMI alerts for females is around their eighties. On the other side, when it comes to men, their peak is around the fifties. Both of these results are consistent with the previous paragraph. It suggests that there is some dependence on population structure. Thus, the population structure needs to be considered when disaggregating spatial data for any given area.

Since we are dealing with a dataset spanning various years, there is a need to examine the overall trend of the series. As we assume that both sex and age may be a potential factor for AMI alerts, we should consider the changes in population structures in the examined period. In other words, in dealing with population structure, there is a potential for an ageing population to be present in the data. Figure 2.4 shows the development of AMI alerts in 2010 and 2018 and the population structure. It can be observed that the overall increasing long-term trend of AMI alerts and population ageing are somehow correlated. Combining both variables, we calculate the AMI incidence rate as plotted in Figure 2.5. By examining the age-sex specific incidence rates, we can observe that the incidence of AMI in Catalonia has increased over the years, indicating complex dynamics of various factors beyond changes in population structure alone. This means that, even after controlling for age, we should observe an increase in the incidence of age-group specific AMI events over the years.

Upon examining the yearly incidences for age and sex groups, we proceeded to fit a linear regression model for the annual incidence of each group. We plotted the slope of the regression line along with the 95% confidence interval. While the

FIGURE 2.3: Age and sex distribution of all AMI alerts (2010-2018)



slope for many age and sex groups is not significantly different from zero, certain groups show a significant increase. Specifically, we found that males aged 50 to 64 and 75 to 94, and females aged 55 to 64 and 85 to 89, exhibited a significant trend at a 95% confidence level (see Table C.1 in Appendix B). Combining all information, we proceed to define our target variable.

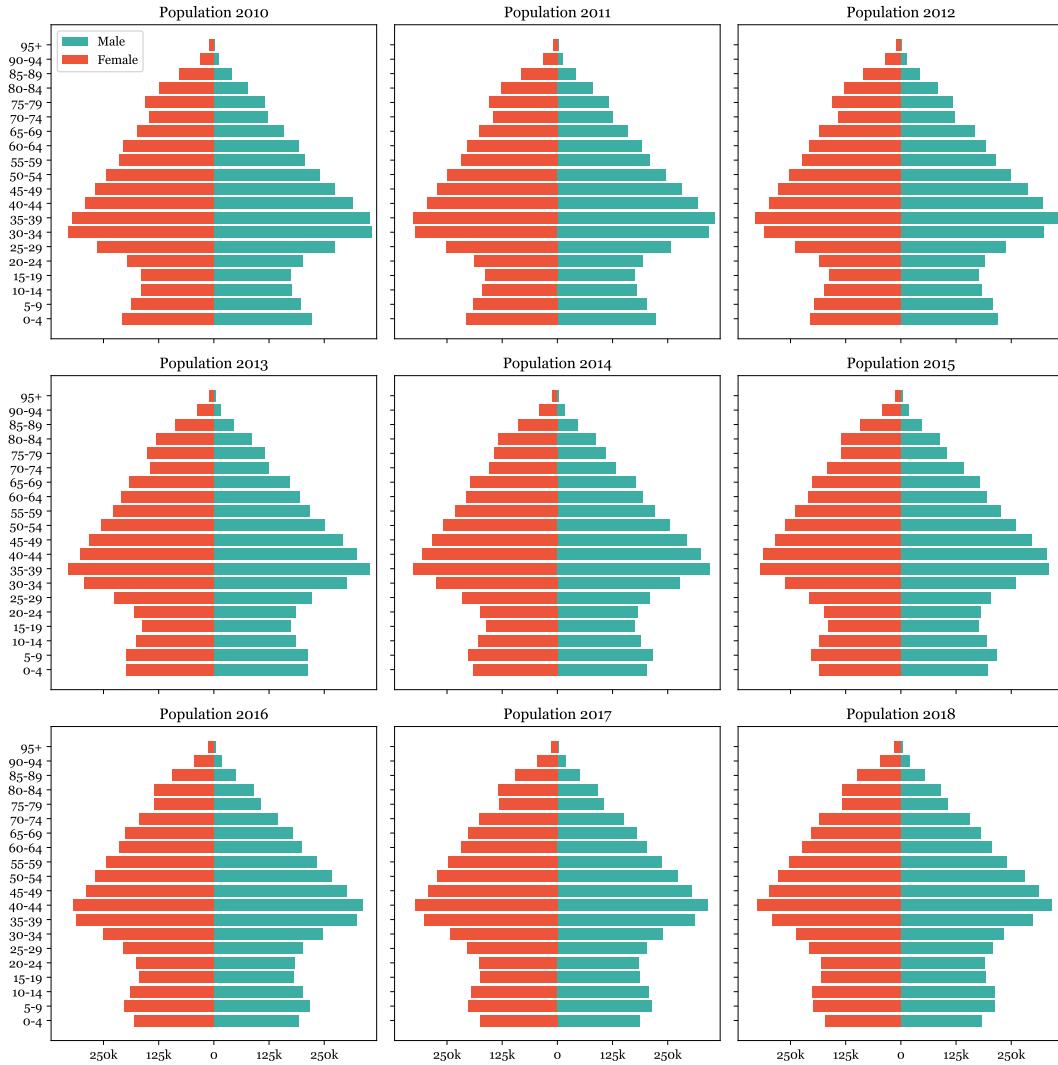
2.2.2 Choice of Age-Standardized Incidence Rates (ASIR)

In analysing AMI incidence, researchers often face the challenge of comparing incidence rates across populations with different age distributions. As mentioned in the previous subsections, the raw incidence rate of AMI may be influenced by demographic factors, i.e. differences in the age structure of the populations under study, making direct comparisons problematic. To address this issue, we use Age-Standardized Incidence Rates (ASIR), which adjusts for differences in age distribution across populations.

ASIR is a valuable metric for comparing disease incidence rates between populations or over time while accounting for differences in age structure. By standardising incidence rates to a reference population with a standard age distribution, ASIR enables fair comparisons across populations with different age profiles. This adjustment helps to isolate the underlying differences in disease burden attributable to factors other than age distribution, facilitating more meaningful comparisons and interpretations.

In the context of our analysis of the association between environmental factors and AMI incidence in Catalonia, Spain, the use of ASIR offers several advantages. Catalonia encompasses diverse demographic profiles across its municipalities, with variations in age distribution that may influence the observed AMI incidence rates.

FIGURE 2.4: Development of Population (2010 - 2018)

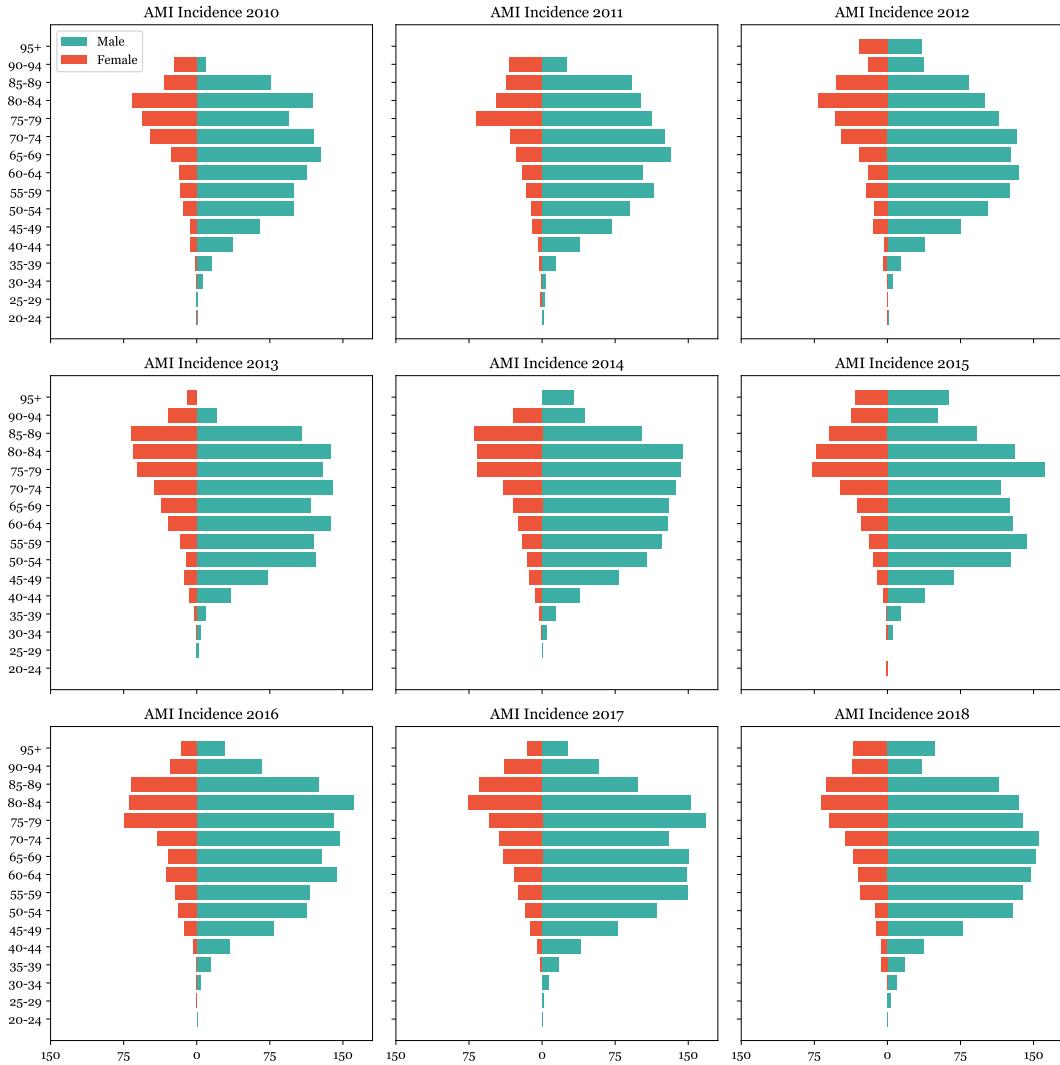


By standardising AMI incidence rates to a reference population, such as the World Health Organization (WHO) standard population, we can obtain ASIR estimates that account for differences in age distribution, enabling more robust comparisons of AMI incidence across municipalities and over time.

Our analysis begins with estimating AMI incidence over a year, stratified by 5-year age groups and sex, utilising comprehensive datasets. Subsequently, population distributions sourced from IDECAT are employed to compute the expected number of AMI events for each region and year. By comparing these expected values with the actual observed events, we can discern AMI incidence trends independent of demographic influences.

The calculation of ASIR represents a pivotal step in standardising incidence rates to a reference population, facilitating equitable comparisons across regions and temporal intervals. This standardisation process effectively mitigates the confounding effects of age and sex distributions, enabling a more accurate assessment of actual changes in AMI incidence. ASIR serves as a robust analytical tool, offering insights into the temporal evolution of AMI incidence while controlling for demographic heterogeneity.

FIGURE 2.5: Development of Incidence Rate per 100k (2010 - 2018)



Visualisation of raw AMI case data reveals intricate temporal patterns, including daily, weekly, monthly, and yearly fluctuations. While indicative of potential seasonal variations and long-term trends, these patterns may also be influenced by shifts in population demographics. By employing ASIR, we aim to mitigate the changes in AMI incidence from demographic effects.

In summary, the choice of ASIR over raw AMI incidence rates in our analysis helps to mitigate the confounding effect of age distribution on the observed incidence rates. By standardising incidence rates to a reference population, ASIR facilitates more accurate comparisons of disease burden across populations with different age structures, enhancing the validity and interpretability of our findings. Before examining ASIR closely, it is crucial to analyse data spatially.

2.2.3 Spatial Distribution of AMI Incidence

In this subsection, we explore the spatial distribution of AMI incidence across counties and territorial regions in Catalonia. By examining the geographical variability

of AMI incidence rates and adjusting for population differences, we aim to identify patterns and trends that can help target public health interventions and allocate resources effectively. Many factors, including urbanisation, socioeconomic status, healthcare infrastructure, and environmental exposures, influence the geographical distribution of AMI incidence within Catalonia. Urban centres may exhibit higher AMI incidence rates due to the concentration of risk factors such as sedentary lifestyles, unhealthy dietary habits, and air pollution. Conversely, rural areas may face challenges related to limited access to healthcare services and longer transport times to medical facilities.

Counties

First, we examine the AMI alerts per 100,000 inhabitants for each county and year, displayed in Figure 2.6. This analysis reveals substantial variability across counties, indicating heterogeneous AMI incidence rates within Catalonia. Moreover, the observed trends in AMI incidence appear consistent across all counties, suggesting the presence of region-wide patterns. Subsequently, we calculate the ASIR for each county and year, providing a more accurate representation of AMI incidence that adjusts for demographic variations in Figure 2.7. High-risk areas are displayed in red.

Visualising the ASIR for each county and year offers insights into the spatial distribution of AMI incidence rates, clear comparisons across regions and revealing areas of elevated risk. By accounting for population differences, ASIR enables a more precise assessment of AMI incidence trends, aiding in identifying high-risk areas and implementing targeted interventions.

Àmbits Territorials

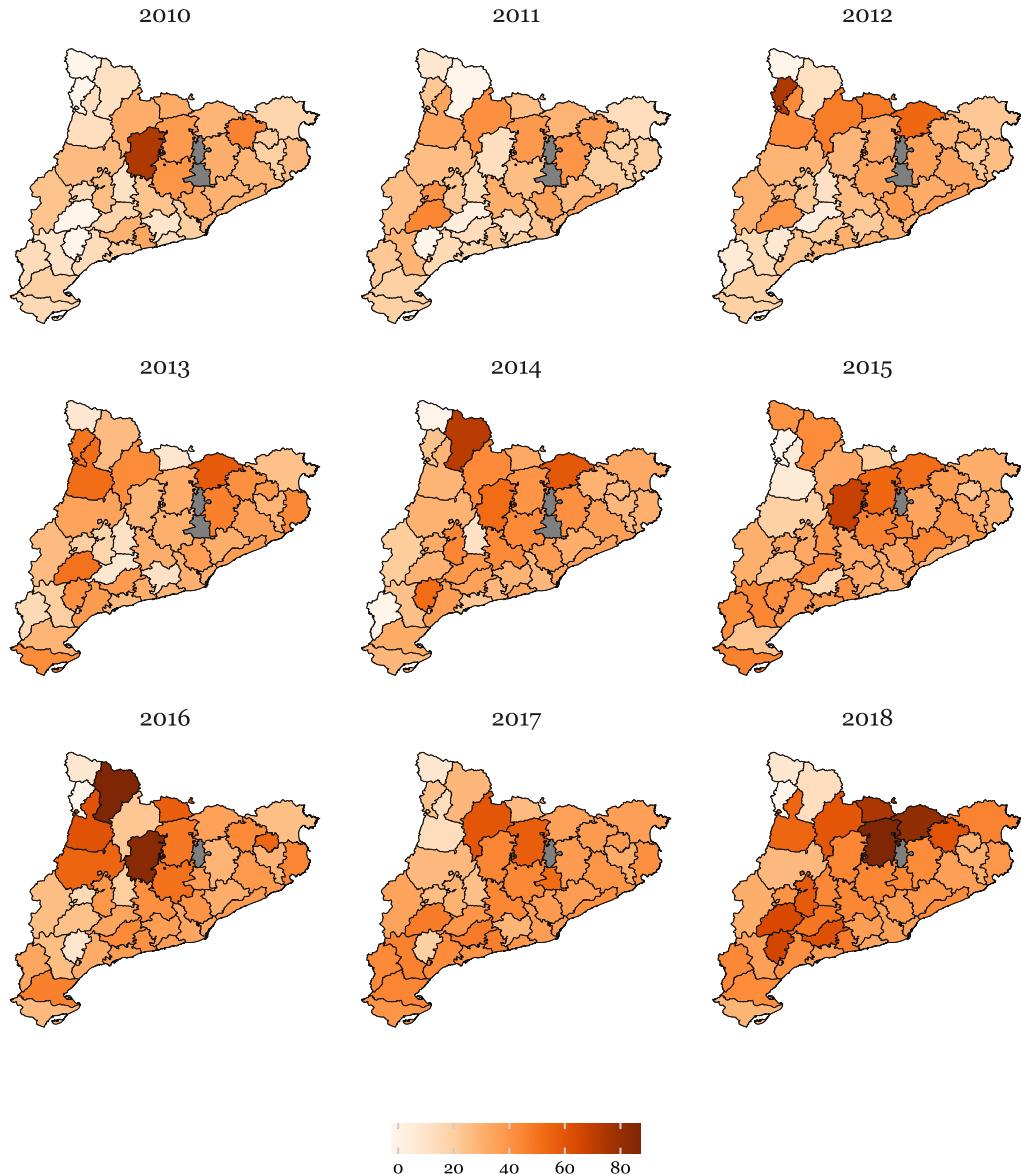
Next, we extend our analysis to Catalonia's territorial regions (àmbits territorials), which offer more significant and stable geographical units for examination. Like the county-level analysis, we compute AMI incidence rates per 100,000 inhabitants for each territorial region and year.

The analysis reveals an increase in AMI incidence rates across all territorial regions, accompanied by significant spatial variability within each year. Subsequently, we calculate the ASIR for each territorial region and year. The analysis highlights areas with elevated ASIR, indicating regions of heightened AMI risk.

When exploring detailed temporal patterns in Age-Standardized Incidence Rates (ASIR) across territorial regions (ATs), we encounter challenges due to the varying population sizes and the resulting noise in the data. With the metropolitan area of Barcelona housing a significant portion of Catalonia's population, the distribution of AMI alerts across ATs differs significantly, leading to discrepancies in signal strength. Given that the yearly incidence of AMI alerts ranges from 30 to 40 cases per year per 100,000 inhabitants, the signal-to-noise ratio in the least populated areas becomes notably skewed, rendering analysis at the daily scale impractical. For reference, Table 2.1 illustrates the total number of AMI alerts for each AT and year.

To address this issue of high variability, we utilise moving averages to smooth the ASIR data and extract hidden trends while minimising the impact of short-term fluctuations. By interpolating the expected daily AMI alerts based on population estimates for every six months in each AT and computing daily ASIR moving averages, we obtain a cleaner signal that enables a more precise assessment of temporal patterns in AMI incidence across territorial regions.

FIGURE 2.6: Spatial Development of AMI alerts per 100k



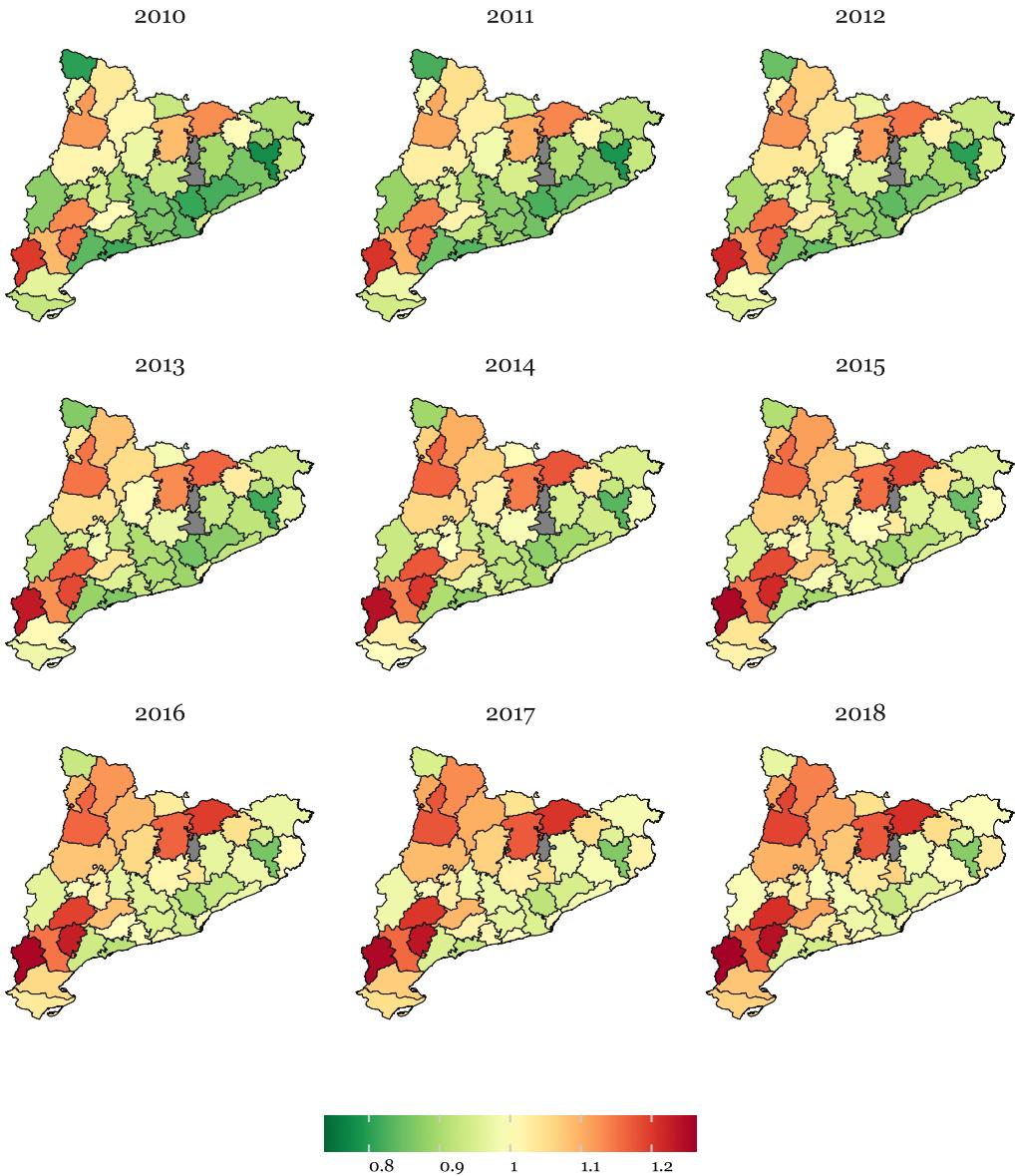
The moving average indicates that the current value is linearly dependent on the following terms - the series mean, the current and previous error terms (Peixeiro, 2022). This approach allows us to discern underlying trends and variations in AMI occurrence, facilitating a more comprehensive understanding of spatial and temporal patterns in AMI epidemiology. Mathematically, it is expressed as

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$$

where:

- y_t represents the value of the time series at time t ,
- μ is the mean of the time series,
- ε_t denotes the error term or random shock at time t ,

FIGURE 2.7: Spatial Development of Age-corrected Population Risk Factor



- $\theta_1, \theta_2, \dots, \theta_q$ are the parameters of the model representing the coefficients of past error terms,
- q is the order of the moving average model, indicating how many past error terms are included.

For smoothing ASIR by using moving averages of different numbers of days, see Figure 2.8. The figure shows a long-term trend with several peaks within the examined period and an apparent cyclical pattern in the data.

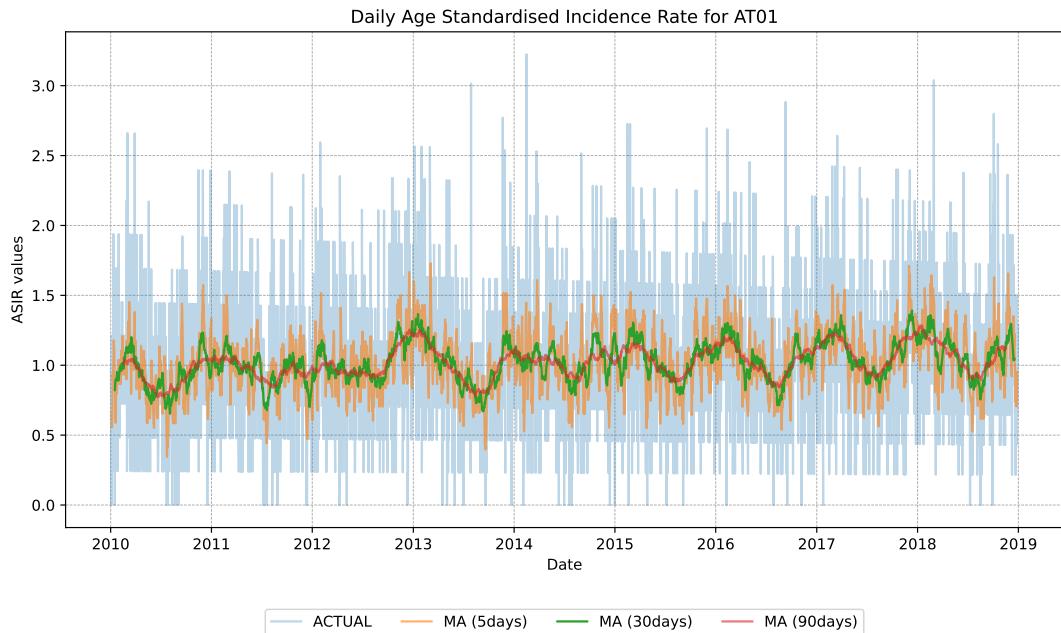
There is a clear seasonal pattern in the series. The ASIR is higher during the beginning and end of the year and lower in the summer months. Also, we decomposed the series to make sure the seasonal component was present (Figure C.1 in Appendix C. We can observe a long-term increase in the series over time, as well as

TABLE 2.1: AMI alerts by AT and year

AT		Year								
		2010	2011	2012	2013	2014	2015	2016	2017	2018
AT01	Metropolità	1401	1469	1574	1582	1629	1663	1697	1826	1820
AT02	Comarques Gironines	192	165	218	265	245	258	268	277	306
AT03	Camp de Tarragona	103	87	134	164	161	190	181	214	216
AT04	Terres de l'Ebre	34	42	37	61	52	66	65	77	66
AT05	Ponent	82	103	107	99	95	112	107	120	135
AT06	Comarques Centrals	152	138	144	154	171	172	172	175	201
AT07	Alt Pirineu i Aran	16	21	29	23	25	19	30	23	34
AT08	Penedès	98	108	133	127	162	157	181	190	189

an annual seasonal pattern with repeated fluctuations. It is important to note that when smoothing the data, there needs to be a balance between reducing noise and maintaining the integrity of the original data. Thus, the data will be aggregated on a weekly level.

FIGURE 2.8: Moving Averages of ASIR Over Time



2.2.4 Seasonal Variations and Weather Patterns

A growing body of evidence indicates that seasonal variations in weather conditions, such as temperature, humidity, and air pollution levels, are associated with fluctuations in AMI incidence. These fluctuations are particularly evident between winter and summer months. Cold temperatures and winter-related factors, including respiratory infections, holiday stress, and changes in physical activity and dietary habits, may contribute to a higher incidence of AMI during winter months (Vaiciulis et al., 2021; Vassalle et al., 2024). Conversely, high temperatures and summer-related factors, such as dehydration, outdoor physical exertion, and increased air pollution

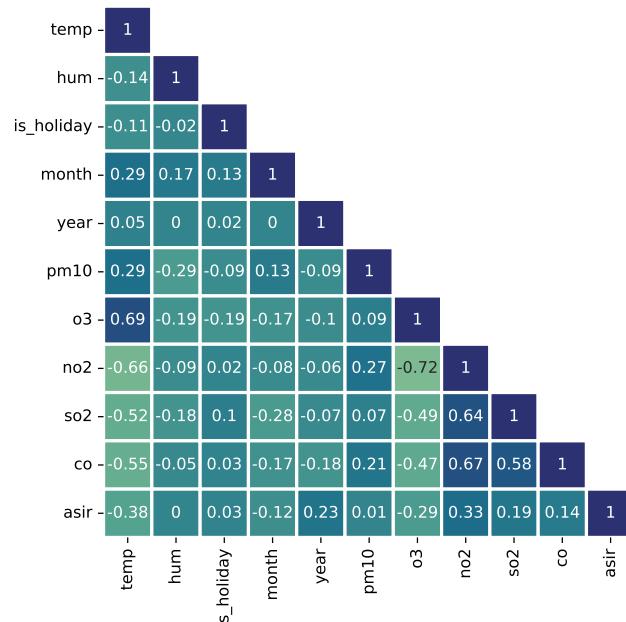
levels, may exacerbate cardiovascular risk during summer months (Chaseling et al., 2021).

After analysing seasonal trends in AMI incidence within Catalonia and its temporal patterns, we explore potential interactions between weather variables and cardiovascular risk factors. In order to proceed, the data has been split into two seasons, specifically—winter and summer—based on their average temperatures. The summer season is from May to October, and the rest of the months are winter months.

The Table C.2 in Appendix C provides a summary of crucial weather variables, including maximum and minimum temperatures, humidity, and air pollutants (CO, NO₂, NO, PM10, SO₂, O₃), during different periods, such as the entire year, winter season, and summer season. Overall, we observe significant differences in weather conditions between the winter and summer seasons. During the winter, maximum and minimum temperatures are lower than in the summer, resulting in lower humidity levels and air pollutants such as ozone (O₃) and nitrogen dioxide (NO₂). Conversely, during the summer, we see higher temperatures, increased humidity levels, and elevated concentrations of air pollutants, mainly ozone and nitrogen dioxide. These findings highlight a similar pattern when considering seasonal variations in weather conditions and the cardiovascular variable. For the monthly development of variables, see Figure C.2 in Appendix C.

Lastly, we constructed a correlation plot (Figure 2.9) that visually represents the pairwise correlations to analyse the interdependencies among variables. We can identify potential multicollinearity among predictors. Ozone levels exhibit a high positive correlation of 0.69 with temperature and a negative correlation of -0.72 with nitrogen dioxide (NO₂). Additionally, nitrogen dioxide (NO₂) shows a positive correlation of 0.67 with carbon monoxide (CO) and a correlation of 0.66 with temperature. Sulfur dioxide (SO₂) correlates positively with carbon monoxide (CO) at 0.58. Due to the high values of O₃ and potential multicollinearity issues, we decided to drop it.

FIGURE 2.9: Correlation plot



To understand the impact of environmental factors on forecasting accuracy, we selected a set of variables for our models (Table 2.2).

TABLE 2.2: Variables for Modeling

Variable	Description	Units
Temperature	Temperature readings	Celsius ($^{\circ}\text{C}$)
Humidity	Humidity levels	Percentage (%)
SO2	Sulfur dioxide concentration	$\mu\text{g}/\text{m}^3$
NO2	Nitrogen dioxide concentration	$\mu\text{g}/\text{m}^3$
CO	Carbon monoxide concentration	mg/m^3
PM10	Particulate matter (PM10) concentration	$\mu\text{g}/\text{m}^3$
Is holiday	Number of holidays within the week	Count
ASIR	Age-Standardized Incidence Rate of AMI alerts	Cases per 100,000 inhabitants

Furthermore, we decided to concentrate our analysis on Barcelona, mainly due to the availability of reliable data. Air quality monitoring stations cover the whole region, suggesting more robust outcomes. By focusing on Barcelona, we can mitigate the challenges associated with data sparsity and variability across different ATs, ensuring a more robust analysis of temporal and spatial patterns in AMI incidence.

To evaluate the influence of environmental variables on forecasting accuracy, we explored several models, specifically SARIMA (Seasonal Autoregressive Integrated Moving Average) to capture seasonal patterns, SARIMAX (SARIMA with exogenous variables) to incorporate external factors like weather data, and LSTM (Long Short-Term Memory) networks for their ability to learn from sequential data and capture complex temporal dependencies. Both are described in detail in the next section.

Chapter 3

Methodology

3.1 ARIMA, SARIMA and SARIMAX Models

ARIMA, or Autoregressive Integrated Moving Average, is a widely used time series forecasting method that models the next step in the sequence as a linear function of the observations and their lagged values, trends, and stationarity through differencing. It combines autoregression (AR), differencing (I), and moving average (MA) components to capture the temporal dependencies and patterns present in the data. As described in book by Shumway and Stoffer, 2011, the following components are described as:

Autoregressive (AR) Component (p): The AR part of the model captures the relationship between an observation and a number of lagged observations (i.e., the series itself). It expresses the current value of the series as a linear function of its past values, weighted by coefficients. The AR(p) model can be formulated as:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t$$

where $\phi_1, \phi_2, \dots, \phi_p$ are the parameters of the model, and ϵ_t is the white noise error term at time t .

Integrated (I) Component (d): The I part of ARIMA involves differencing the time series data to achieve stationarity. Stationarity means that the statistical properties of the series, such as mean and variance, do not change over time. Differencing involves subtracting the previous observation from the current observation. For example, $\Delta y_t = y_t - y_{t-1}$ for first-order differencing. The parameter d represents the number of differencing steps needed to achieve stationarity.

Moving Average (MA) Component (q): The MA part of the model captures the relationship between an observation and a residual error from a moving average model applied to lagged observations. The MA(q) model can be expressed as:

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}$$

where $\theta_1, \theta_2, \dots, \theta_q$ are the parameters of the model, and ϵ_t is the white noise error term at time t .

Seasonal ARIMA (SARIMA) extends ARIMA to incorporate seasonal variations alongside non-seasonal components, addressing periodic patterns in the data.

Seasonal Autoregressive (SAR) Component (P): Models the relationship between an observation and its seasonal lagged observations.

The SAR(P) component is formulated as:

$$y_t = \phi_1^* y_{t-s} + \phi_2^* y_{t-2s} + \dots + \phi_P^* y_{t-Ps} + \epsilon_t$$

where $\phi_1^*, \phi_2^*, \dots, \phi_P^*$ are the seasonal AR parameters, and s is the seasonal period.

Seasonal Moving Average (SMA) Component (Q): Captures the dependency between an observation and its seasonal lagged forecast errors.

The SMA(Q) component is represented as:

$$y_t = \epsilon_t + \theta_1^* \epsilon_{t-s} + \theta_2^* \epsilon_{t-2s} + \dots + \theta_Q^* \epsilon_{t-Qs}$$

where $\theta_1^*, \theta_2^*, \dots, \theta_Q^*$ are the seasonal MA parameters.

Seasonal Differencing (D): Extends the non-seasonal differencing d to account for seasonal variations and achieve stationarity across seasonal cycles.

Seasonal differencing can be applied to the series y_t as:

$$\Delta_s y_t = y_t - y_{t-s}$$

where s is the seasonal period, and Δ_s denotes the seasonal differencing operator. The parameter D specifies the number of seasonal differencing steps needed.

SARIMAX extends SARIMA by incorporating exogenous variables (denoted as X_t) that are external to the time series but may influence it. Exogenous variables are additional factors that may impact the time series but are not influenced by it. These variables can improve the model's forecasting accuracy by accounting for external influences on the target variable.

Before including any exogenous variable, specifically environmental data, we closely looked at the target variable. Starting by evaluating the stationarity of the time series data by performing the Augmented Dickey-Fuller (ADF) test to ensure the series is suitable for modelling. Furthermore, we can observe the temporal dependencies and seasonality by plotting ACF and PACF (Figure 3.1). There is an annual seasonality present, specifically the ASIR increases during the winter months and is more subtle during the hot months and also that there are some statistically significant lagged values present.

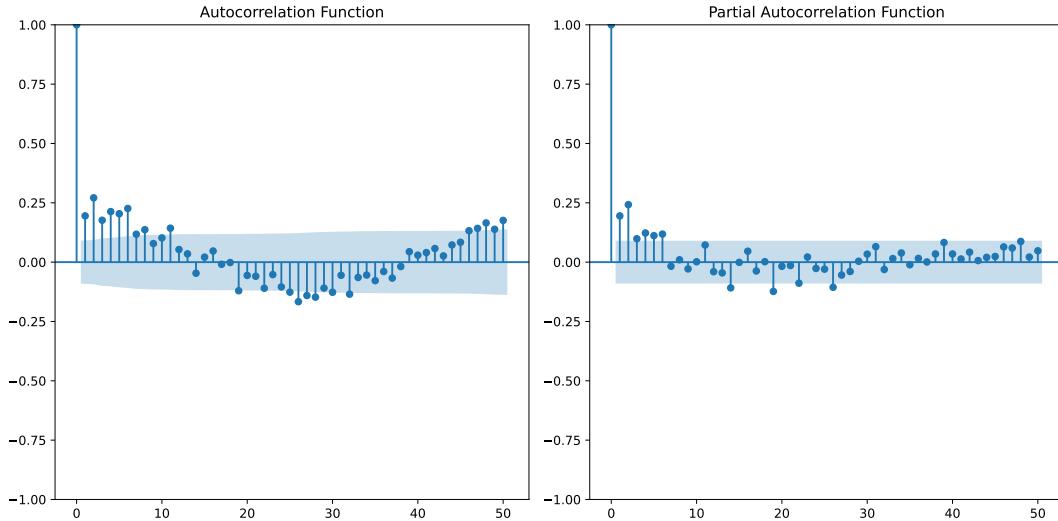
Due to the data exhibiting seasonal patterns, ARIMA model may not be the best model, therefore we may consider SARIMA, which adds seasonal parameters to ARIMA model. Using the `pmdarima` library in Python, we employed an automated approach to determine the optimal SARIMA model parameters.

After fitting the SARIMA model, we evaluated its performance and diagnostic checks. Diagnostic plots, including residual analysis and model diagnostics such as the Ljung-Box test for autocorrelation in residuals, were examined to ensure the model's adequacy. These visual and statistical tests verified that the SARIMA model effectively captured the underlying patterns in the data.

We explored the SARIMAX model to evaluate whether environmental variables can improve prediction accuracy. This extension allows for modelling additional factors that could impact the seasonal patterns observed in the time series.

In the context of time series analysis, the equation for a Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) model is given by:

FIGURE 3.1: ACF and PACF plots



$$\begin{aligned} y_t = & \phi_p y_{t-1} + \cdots + \phi_1 y_{t-p} + \theta_q \varepsilon_{t-q} + \cdots + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\ & + \theta_{s \cdot q} \varepsilon_{t-s \cdot q} + \cdots + \theta_s \varepsilon_{t-s} + \cdots + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} \end{aligned}$$

where:

- y_t represents the value of the time series at time t ,
- $\phi_1, \phi_2, \dots, \phi_p$ are the autoregressive (AR) parameters of the model representing the coefficients of past values of the series,
- p is the order of the autoregressive process,
- $\theta_1, \theta_2, \dots, \theta_q$ are the moving average (MA) parameters of the model representing the coefficients of past errors,
- q is the order of the moving average process,
- ε_t denotes the error term or random shock at time t ,
- s is the seasonal period,
- $\theta_{s \cdot q}, \theta_{s \cdot (q-1)}, \dots, \theta_s$ are the seasonal moving average parameters,
- $\beta_1, \beta_2, \dots, \beta_k$ are the coefficients of exogenous variables $x_{1,t}, x_{2,t}, \dots, x_{k,t}$.

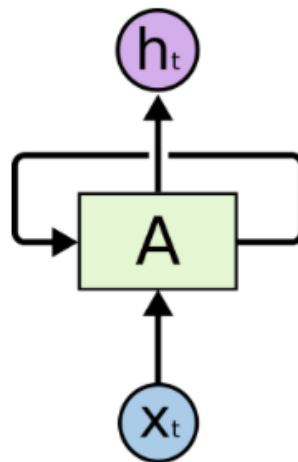
3.2 The Recurrent Neural Network

In deep learning, Recurrent Neural Networks (RNN) are defined as bi-directional artificial neural networks. As part of supervised learning, their usage is universal and multidisciplinary. They are commonly used for time-series data analysis and forecasting, classification, regression, image/video processing and many other problems. The central concept of the RNN is that the neural network is based on cycles, meaning that the network depends not only on the current data but also on the previous data it used (Hrnjica and Bonacci, 2019). In other words, the output

from nodes can have an effect on other inputs provided to the same nodes. Figure 3.2 shows an example of a single RNN cell. The hidden state is looped back and used on other inputs. The main setback of RNNs is that they are computationally very costly, meaning they are slow. Furthermore, they can only carry short-term information, and lastly, they tend to have vanish gradient problems (when a change in gradient is close to 0 or very small), implying that the model stops learning.

FIGURE 3.2: A single RNN Cell

Source: Olah, 2015



To overcome most of the problems, Long Short-Term Memory Networks (LSTM) are a special kind of RNN that were designed. LSTM has an extra cell state added. The networks are able to keep information over long sequences and can overcome the vanishing gradient problem. When it comes to the architecture of LSTM, which has a chain-like structure, as a network, it has a unique design. It contains only four neural networks and different memory cells that regulate the flow of information. Each contains three main components - input, forget and output gates, which are multiplicative units that help to store, update and retrieve information over long sequences. As described in Peixeiro, 2022, the structure of LSTM cells is as follows.

3.2.1 Structure of an LSTM Cell

- *Cell State (C_t):*

The cell state is a key feature of LSTMs, allowing information to be carried across long sequences. It acts as a memory that can maintain relevant information throughout the processing of sequences. Furthermore, the cell state is updated with the help of the forget and input gates, which are both described below. For now, it is important to note that both gates contain weights that help to decide which time steps to incorporate into the cell state. They are not all equally incorporated, making LSTM interesting in their ability to dynamically and adaptively decide what periods to include.

- *Hidden State (h_t):*

The hidden state contains the output of the LSTM cell at each time step. It is also passed to the next cell and is used to compute the network output. In other words, it can be seen as encoding the latest time step.

- *Forget Gate (f_t):*

The forget gate decides which information from the cell state should be discarded. It takes the previous hidden state (h_{t-1}) and the current input (x_t). It is further multiplied by the weight matrices and additional bias. The output is between 0 and 1 for each number in the cell state C_{t-1} . Outputs close to 0 mean that the information is forgotten; close to 1, the information is retained and very important.

Mathematically, $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$, where σ is the sigmoid activation function, W_f are the weights, and b_f is the bias.

- *Input Gate (i_t):*

The input gate determines which new information will be stored in the cell state. It consists of two parts: an update vector (\tilde{C}_t) and the gate itself. The sigmoid function regulates and filters information using inputs h_{t-1} and x_t .

The gate, i_t , is calculated similarly to the forget gate: $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$.

The update vector, \tilde{C}_t , uses the activation *tanh* function which output ranges between -1 and 1 containing all possible values of h_{t-1} and x_t . Then the values of the vector and the regulated values are multiplied to create a vector of new candidate values: $\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$.

- *Cell State Update:*

The cell state is updated by combining the old cell state and the new candidate values. The forget gate f_t controls what proportion of the old cell state should be kept, and the input gate i_t controls how much of the candidate values should be added.

Mathematically, $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$.

- *Output Gate (o_t):*

The output gate determines the next hidden state, which will be used in the next time step and as the output of the current time step. It combines the previous hidden state and the current input to decide which parts of the cell state will be output. In other words, we generate a vector by applying *tanh* function and regulate and filter the information using sigmoid function, inputs h_{t-1} and x_t . One of the functions of the *tanh* function is to normalise data encoding.

The output gate is calculated as: $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$.

The new hidden state h_t is then calculated as: $h_t = o_t * \tanh(C_t)$.

3.2.2 Working Mechanism

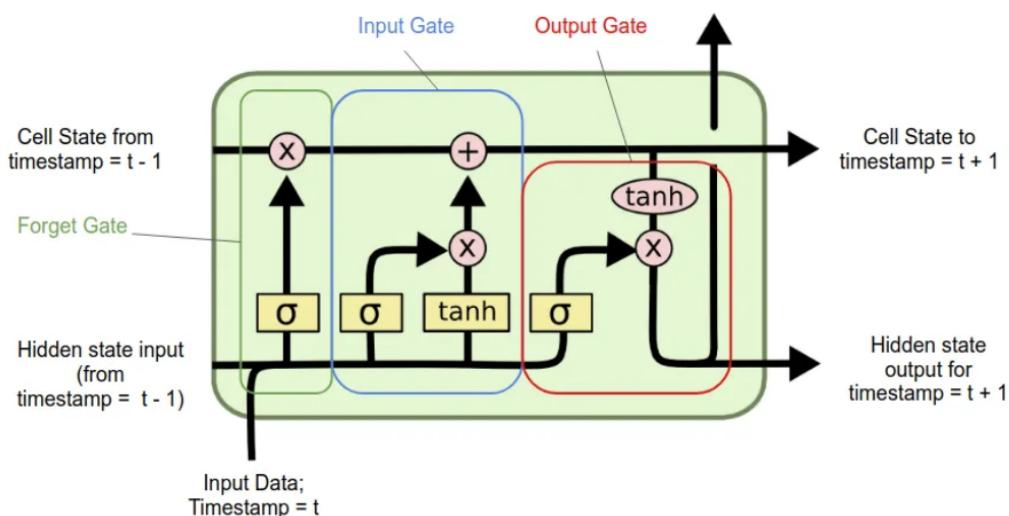
At each time step, the LSTM cell takes in the current input x_t and the previous hidden state h_{t-1} . The forget gate determines which parts of the previous cell state C_{t-1}

should be retained. The input gate decides which new information from the current input x_t should be added to the cell state. The cell state is then updated accordingly. Finally, the output gate computes the new hidden state h_t , which is the output for the current time step and will be passed to the next LSTM cell.

LSTM cells, with their unique gating mechanisms, provide a powerful way to capture long-term dependencies and temporal patterns in time series data. By carefully controlling the flow of information through forget, input, and output gates, LSTMs effectively address the limitations of traditional RNNs, making them suitable for complex predictive modeling tasks.

FIGURE 3.3: A single LSTM Cell

Source: Olah, 2015



Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) units to predict the incidence of AMI based on environmental factors. RNN-LSTM is chosen for its capability to model time series data with complex temporal dependencies and non-linear relationships thanks to LSTMs feedback connections.

3.2.3 LSTM in practice

Data preprocessing was involved prior to training the models so the data would be suitable for LSTM. Normalisation was performed using MinMaxScaler to scale the feature variables to a range of 0 to 1. The target variable, representing ASIR, was similarly scaled.

To prepare the data for the LSTM model, the time series data was converted into sequences. This involved creating fixed-length sequences of input features and corresponding target values. A sequence length of 12 days was chosen based on domain knowledge and preliminary experiments.

The data was divided into training and testing sets, with around 85% of the data allocated for training and the remaining 15% for testing, more specifically train set was years 2010-2017 and test set was year 2018. This split ensured that the model had sufficient data for training while still providing a robust test set for evaluation.

3.2.4 Hyperparameters optimisation

Hyperparameter optimisation (also called hyperparameter tuning) is a process of selecting the optimal configuration of model hyperparameter values such that the model performance is optimised with respect to some performance metric (Wu et al., 2019). Unlike model parameters, which are estimated from the data, model hyperparameters cannot be estimated from the data, and their values are selected prior to model training. A few standard methods of performing hyperparameter optimisation include Random Search, Bayesian optimisation, and Sequential Model-Based Optimization (SMBO); the last one uses a surrogate model to iteratively explore the hyperparameter space and find the configuration that maximises the model's performance.

Random search is a search algorithm that iterates over some pre-specified number n of combinations of hyperparameter values to find the optimal setting (De Sa, 2020). First, a distribution of values is defined for every hyperparameter being optimised, and then n combinations of hyperparameter values are sampled randomly from the joint distribution of the hyperparameter values. For each of these n settings, a model is fit and evaluated. The best setting is the one that minimises the metric of interest the most, i.e. results in a minor loss. In random search, each iteration's combinations of hyperparameter settings are independent. All these aspects need to be considered when fine-tuning the model and balanced with trade-offs of having too much complexity.

This is not the case for Bayesian optimisation (Brochu, Cora, and De Freitas, 2010). Bayesian optimisation is an optimisation strategy that performs an 'informed' search through the parameter space in the sense that it incorporates knowledge about previous combinations to sample the following combination. It aims to find the extrema of a function – in the context of hyperparameter optimisation, the objective of Bayesian optimisation is to find a minimum of the loss function. The algorithm starts by specifying the probability distributions of the hyperparameter values and sampling multiple points from the joint distribution of the hyperparameter values. Then, a surrogate function (approximating the loss function that can be expressed as a probability distribution) is formed based on the sampled points. The algorithm then identifies the following combination of hyperparameter values such that it compromises between searching previously unexplored areas of the surrogate function with high uncertainty and regions, which leads to further improvement of the surrogate function. The surrogate function then gets updated by making use of Bayes Theorem, which is given by:

$$P(f | \mathcal{D}) \propto P(\mathcal{D} | f)P(f),$$

where the prior distribution (the surrogate function) $P(f)$ is multiplied by the likelihood function $P(\mathcal{D} | f)$, and $P(f | \mathcal{D})$ represents the resulting posterior distribution.

The parameters that can be tuned in an LSTM include:

Number of units in each LSTM layer: This determines the dimensionality of the output space. More units can capture more complex patterns but might lead to overfitting.

Number of LSTM layers: Multiple layers can capture hierarchical patterns in the data, but adding too many layers can make the model unnecessarily complex.

Dropout rate: Dropout is a regularisation technique that prevents overfitting by randomly setting a fraction of input units to 0 at each update during training. A higher dropout rate means more units are dropped, leading to more robust regularisation.

Learning rate: This controls how much the model is adjusted in response to the estimated error each time the model weights are updated. A lower learning rate means the model learns more slowly but can converge to a better solution.

Batch size: The number of samples processed before the model is updated. Smaller batch sizes can lead to more stable learning but can be computationally expensive.

Activation function: Determines the output of a node. Common choices for LSTM include tanh and ReLU.

Optimizer: The algorithm is used to change the attributes of the neural network, such as weights and learning rate, to reduce the losses. Standard optimisers include Adam, RMSprop, and SGD.

Chapter 4

Results

4.1 Model Comparison and Results

In this section, we present a comparative analysis of the SARIMA and SARIMAX models, along with the Long Short-Term Memory (LSTM) model. We summarize the performance of each model, highlighting their respective strengths and limitations in forecasting AMI incidences. Furthermore, we take a closer look at the optimisation processes for these models, examining how different hyperparameters and exogenous variables impact predictive accuracy. Finally, we explore feature importance and assess how the models' prediction skill varies with lead time, providing a comprehensive evaluation of their applicability for an EWS for AMI.

4.1.1 SARIMA/SARIMAX Model Results Summary

The SARIMAX model has successfully converged, indicating a stable optimisation process. The optimal model suggested was specified as SARIMAX(1, 1, 1)x(2, 0, 0, 52), indicating one autoregressive term, one differencing term, one moving average term, and two seasonal autoregressive terms with an annual periodicity of 52 weeks.

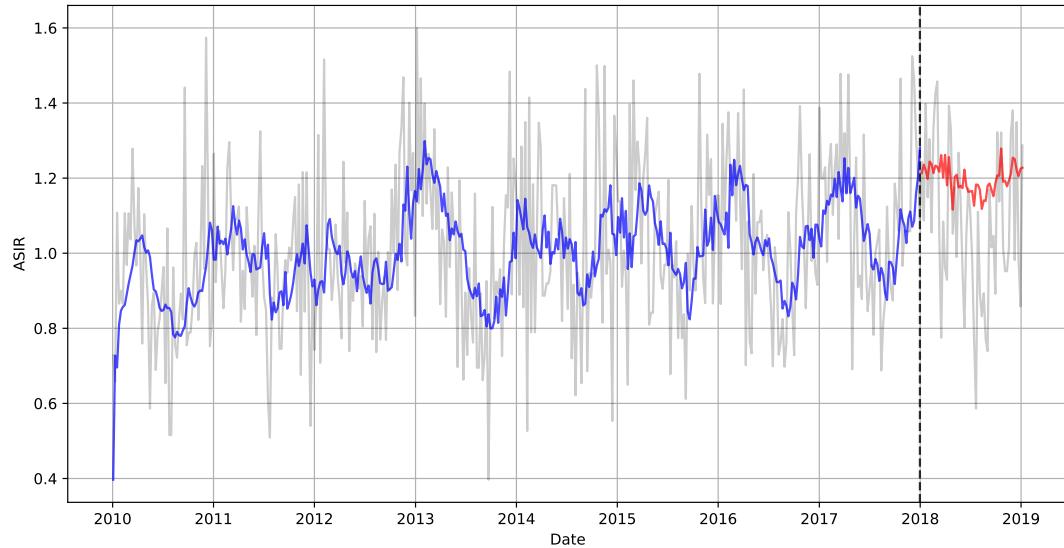
We constructed models with and without exogenous variables and plotted the predictions in Figure 4.1. It can be observed that the SARIMAX model, which incorporates exogenous variables, demonstrates better performance when it comes to capturing dynamic changes, especially sudden drops and increases throughout the year. This model exhibits less smoothness and closely follows the observed trend, highlighting its ability to incorporate external factors effectively. Furthermore, our selection of exogenous variables suggests that environmental factors correlate with ASIR fluctuations and capture additional sources of variability. Specifically, SARIMA exhibits a delayed response to these fluctuations, causing predictions to be shifted to the right.

The performance comparison between the SARIMA and SARIMAX models is summarised in Table 4.1. The SARIMAX model outperforms the SARIMA model in both training and validation datasets. Specifically, the SARIMAX model achieved a Training MAE of 0.148 compared to 0.158 for the SARIMA model and a Validation MAE of 0.151 compared to 0.184 for the SARIMA model. Similarly, the Training MAPE for the SARIMAX model was 15.7% compared to 16.7% for the SARIMA model, and the Validation MAPE was 15.2% compared to 19.9% for the SARIMA model. These results clearly demonstrate that incorporating exogenous variables significantly improves the model's accuracy.

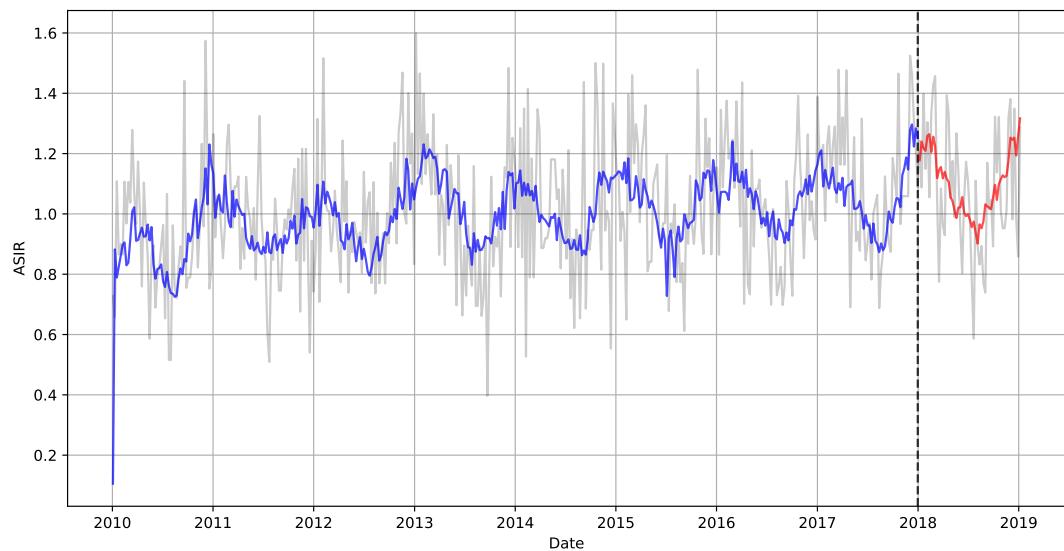
This table highlights the improvements made by the SARIMAX model in both the training and validation phases, emphasising the importance of incorporating exogenous variables for better predictive performance.

FIGURE 4.1: SARIMA vs SARIMAX predictions

(A) Time Series Fitted by SARIMA



(B) Time Series Fitted by SARIMAX



— Actual Time Series --- End of Training — Training Predictions — Test Predictions

TABLE 4.1: Performance Comparison between SARIMA and SARI-MAX Models

Model	Train		Test	
	MAE	MAPE (%)	MAE	MAPE (%)
SARIMA	0.158	16.7	0.184	19.9
SARIMAX	0.148	15.7	0.151	15.2
<i>Improvement %</i>			17.93%	19.60%

Moreover, we conducted a parameter tuning process for the SARIMAX model using a grid search approach. This involved evaluating multiple combinations of parameters through cross-validation, aiming to identify the optimal set that maximises the model's predictive performance. Using TimeSeriesSplit cross-validation and tracking performance metrics, such as mean squared error (MSE), across different parameter configurations, we identified and selected the most suitable SARI-MAX model for forecasting.

4.1.2 LSTM Model Results Summary

Keras Tuner, frequently used to find optimal parameters for neural networks (O'Malley et al., 2019), was utilised to optimise the hyperparameters of the LSTM model. The following hyperparameters were determined using the Keras Tuner, which involved a random search over the specified hyperparameter space. The search process used a maximum of 30 trials, each executed three times to ensure robustness in the evaluation. Early stopping was employed during training to prevent overfitting and to restore the best weights based on validation loss.

Best Hyperparameters and Model Description

The hyperparameter optimisation process identified the following optimal hyperparameters for the LSTM model and together with the model is described below:

The model was compiled with the Adam optimiser and mean squared error as the loss function. The training configuration included:

Epochs: 100

Batch size: 32

Validation split: 0.1

Model Training

The training process was monitored using a validation split of 10%, and early stopping was applied with patience of 10 epochs. The early stopping mechanism monitored the validation loss, with a minimum delta of 0.01 for improvements to be considered significant.

The final model, trained with the best hyperparameters, exhibited robust performance with the specified architecture and training configuration. The use of dropout layers helped mitigate overfitting, and the choice of the Adam optimiser ensured efficient and effective convergence of the model parameters.

TABLE 4.2: Model Architecture

```
Model: "sequential"
```

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(30, 8)]	0
lstm_1 (LSTM)	(200,)	167200
dropout_1 (Dropout: 0.2)	(200,)	0
dense_1 (Dense)	(1,)	201

Total params: 167401
Trainable params: 167401
Non-trainable params: 0

Optimiser: Adam
Activation Function: ReLU

The LSTM model, which incorporates exogenous variables, performs very similarly to the SARIMAX model. Table 4.3 displays the evaluation metrics in the LSTM model for both the training and testing periods. The predictions are visually presented in Figure 4.2.

TABLE 4.3: Evaluation metrics for LSTM (Exogenous Variables)

Metric	Train	Test
MAE	0.156	0.159
MSE	0.039	0.041
RMSE	0.203	0.203
MAPE	16.93%	15.69%

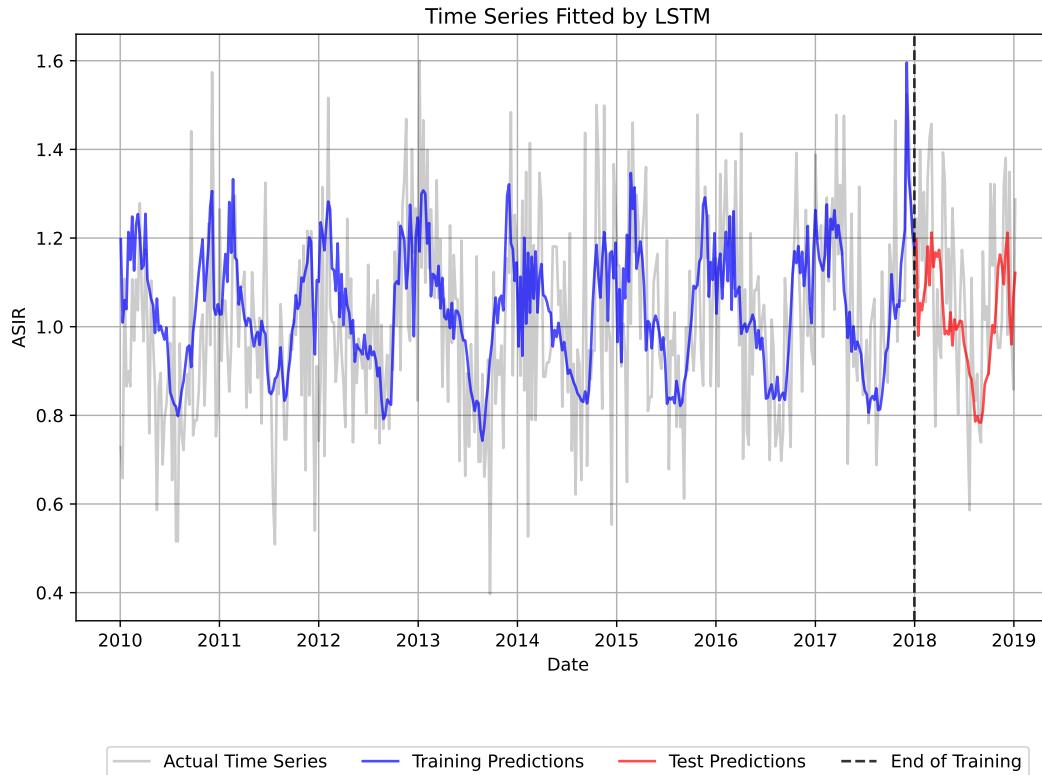
4.2 Feature Importance

Feature importance is a technique for obtaining the impact of variables on model predictions. Various methods extract and analyse the importance, such as permutation feature importance, feature importance from model coefficient, Shapley Additive Explanations (SHAP) values, or Local Interpretable Model-agnostic Explanations (LIME). In this analysis, we will examine the first two methods more closely.

4.2.1 Permutation Feature Importance

The permutation feature importance is a method that measures the feature's importance by comparing the model's error with and without permuting the feature's values. A feature is considered significant if the model error increases after the values

FIGURE 4.2: LSTM predictions with exogenous variables



are permuted, suggesting the model’s reliance on the feature Molnar (2022).

In our case, the analysis aims to understand the contribution of individual environmental variables—temperature, humidity, and pollution—to the predictive performance of the LSTM model with exogenous variables. The model’s ability to capture and utilise these features is crucial for interpreting its predictions. Together with the importance of features, we can improve the understanding of a “black box” prediction model.

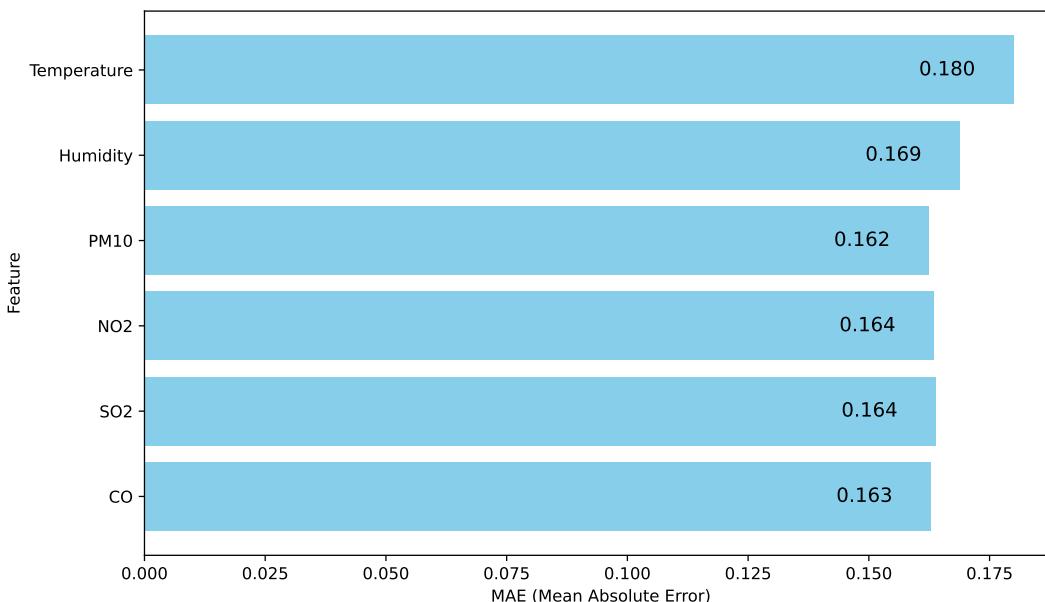
We assessed the importance of each feature by evaluating the mean absolute error (MAE) when each feature was individually shuffled while keeping others unchanged. Other techniques could be used, such as added noise, integrated gradients or ablation Freeborough and Zyl (2022). This method allows us to quantify the impact of each feature on the accuracy of the model’s predictions. The results of this analysis are summarised in Table 4.4. To ensure robustness in our feature importance analysis, we employed K-fold cross-validation. This method involves splitting the dataset into K folds, training the model on K-1 folds, and validating it on the remaining fold. We repeated this process K times, each time with a different validation fold, to obtain a comprehensive assessment of the model’s performance.

Figure 4.3 illustrates the outcomes of the feature importance analysis, visually depicting the MAE values associated with each environmental variable. These findings highlight the relative importance of temperature, humidity, and various pollutants in influencing the model’s predictive performance. Among the evaluated features, the MAE of temperature has the highest value, i.e., 0.18, which indicates it to be the highest of importance, as the second most crucial variable is humidity, with MAE = 0.169. The pollutants have approximately the same importance, around 0.163.

TABLE 4.4: Feature Importance based on Mean Absolute Error (MAE)

Feature	MAE	% Increase
<i>Baseline MAE</i>	0.1590	–
Temperature	0.1800	13.21%
Humidity	0.1689	6.23%
SO2	0.1640	3.14%
NO2	0.1635	2.83%
CO	0.1629	2.45%
PM10	0.1624	2.14%

FIGURE 4.3: Feature Importance in LSTM Model with Exogenous Variables



The analysis shows that temperature exhibits the highest influence on AMI incidence predictions, followed by humidity and pollution levels. These findings suggest that temperature fluctuations play a significant role in the occurrence of AMI, aligning with previous research highlighting the impact of weather on cardiovascular health.

4.2.2 Feature Importance from Model Coefficients

Another commonly used method involves assessing feature importance based on model coefficients, particularly in linear models, where higher magnitude coefficients indicate more significant contributions to predictions. It is important to note that these coefficients represent linear dependencies within the SARIMAX framework.

We observe that temperature exhibits the most significant negative coefficient magnitude (-0.362), indicating a strong negative relationship with AMI incidence predictions in the SARIMAX model. Higher temperatures are associated with lower predicted AMI incidence, aligning with the established physiological effects of heat

on cardiovascular health. In contrast, humidity (coefficient -0.068) shows a smaller effect size, suggesting a weaker impact on AMI incidence in this model context. The presence of pollutants such as PM10 (coefficient 0.153) and NO2 (coefficient 0.1632) also demonstrates significant coefficients, with positive signs indicating potential positive associations with AMI risk.

4.3 Prediction skill vs Lead-time

A critical issue to explore in this context is the prediction skill versus lead time. While the current approach predicts the entire validation period of one year as a whole, in practice, an Early Warning System (EWS) would be continuously updated as new data becomes available. Therefore, it is crucial to assess whether there is a degradation in the performance of the models as the lead time increases. This is particularly relevant since the models rely on both forecasts of the target variable and the exogenous variables.

To investigate this, we can compare the prediction skill at different lead times, ensuring that the forecasts are updated continuously using the most recent observations. Precisely, we can measure the model's performance at various points within the validation period, examining how the accuracy of predictions changes as the forecast horizon extends.

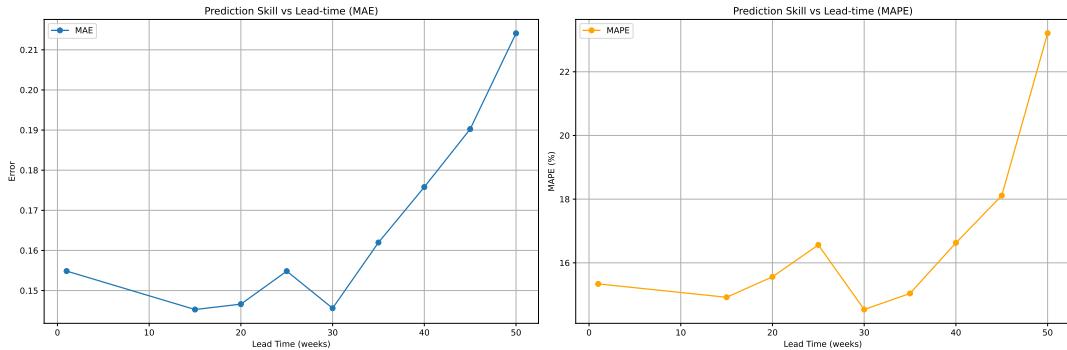
We implemented a rolling forecast approach, where the model is updated with new data and forecasts are made for lead times ranging from 1 to 50 weeks. The model's performance was evaluated using Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) for these different lead times. The results in Table 4.5 show how the prediction accuracy degrades as the lead time increases, which is crucial for understanding the reliability of the EWS over extended forecast horizons.

TABLE 4.5: Prediction Skill vs Lead Time

Lead Time	MAE	MAPE (%)
1	0.1549	15.34
15	0.1453	14.92
20	0.1466	15.56
25	0.1548	16.56
30	0.1456	14.53
35	0.1620	15.04
40	0.1758	16.63
45	0.1902	18.11
50	0.2141	23.21

We observe that the model achieves relatively accurate forecasts for shorter lead times, such as one week, with a Mean Absolute Error (MAE) of approximately 0.155 and a Mean Absolute Percentage Error (MAPE) of about 15.34%. However, we observe a decline in prediction accuracy as the lead time extends. For instance, at a lead-time of 25 weeks, the MAE increases to around 0.155, and the MAPE rises to 16.56%. This trend continues, with the MAE and MAPE reaching 0.214 and 23.21%, respectively, at a lead time of 50 weeks. For graphical visualisation, see Figure 4.4. These findings highlight that to maintain prediction accuracy over a more extended forecast, it is essential to update the model and add new data regularly to maintain the accuracy and effectiveness of the EWS.

FIGURE 4.4: Prediction Skill vs Lead-time



4.4 Limitations

The SARIMAX model, while useful for modelling time series data, has several limitations. Firstly, it assumes linear relationships between the predictor and response variables. However, this may not accurately capture the complex nonlinear relationships in the data. Additionally, SARIMAX models are sensitive to outliers and extreme values, which can lead to biased estimates if not appropriately addressed. Moreover, the chosen seasonal and non-seasonal orders may not fully capture the seasonal patterns and trends present in the data, potentially resulting in model misspecification. Furthermore, SARIMAX models rely on the assumption of stationarity, which may not hold for all time series data, further impacting model performance. Finally, SARIMAX models require accurate values of exogenous variables for prediction, and any inaccuracies in these variables can affect the accuracy of the forecasts.

Similarly, LSTM models have their own set of limitations. One major limitation is the requirement for a large amount of data for effective training. LSTM models may struggle to learn complex patterns without sufficient data and suffer from poor performance. Additionally, hyperparameter tuning for LSTM models can be computationally intensive and may require extensive experimentation to identify optimal configurations. Furthermore, LSTM models are prone to overfitting, especially when trained on noisy or high-dimensional data, leading to poor generalisation of unseen data. Lastly, the interpretability of LSTM models can be challenging due to their black-box nature, making it difficult to understand the underlying mechanisms driving predictions. Similar to SARIMAX models, LSTM models also require accurate values of exogenous variables for prediction, and any inaccuracies in these variables can impact the accuracy of the forecasts. Moreover, to accurately predict the models using exogenous variables, there is also a need to know those values for forecasting, making it one of the crucial limitations of both models.

Furthermore, our analysis utilised linear interpolation to derive daily population data estimates from six-month county-level estimates. This approach ensured the dataset's comprehensibility; however, it also introduced potential minor disruptions. Although these shifts are generally minimal due to the gradual nature of population changes from year to year, they could still influence model performance.

Chapter 5

Conclusion and future directions

5.1 Conclusion

Understanding the environmental triggers of AMI can build better prevention strategies and potentially decrease mortality rates by providing early warnings. As climate change continues to alter weather patterns, the incidence of AMI may be influenced by new and changing environmental factors, making timely and accurate predictions even more critical.

In summary, this study aimed to investigate the relationship between environmental factors and the incidence of AMI in Catalonia, Spain, and set the foundations for a potential EWS for AMI. Encompassing hospital admissions, meteorological, and pollution data, we employed distinct modelling approaches: SARIMA, SARIMAX, and LSTM.

Our results demonstrated that the SARIMAX model significantly outperformed the SARIMA model by incorporating exogenous variables by around 18%, highlighting the importance of external factors in improving forecast accuracy. Furthermore, when assessing prediction skill versus lead time, the SARIMAX model maintained low errors, up to 30 predictions ahead of time. Additionally, we examined feature importance and found that temperature and humidity were the most significant factors, followed by pollution.

Despite the promising results, the SARIMAX model has several limitations. It assumes linear relationships between predictor and response variables, which may not fully capture the complex nonlinear relationships in the data. Additionally, SARIMAX models are sensitive to outliers and extreme values, potentially leading to biased estimates. The model's reliance on accurate exogenous variable values for prediction poses a significant limitation, as inaccuracies in these variables can adversely affect forecast accuracy. The chosen seasonal and non-seasonal orders might also not fully capture all the seasonal patterns and trends, leading to model misspecification. Moreover, SARIMAX models assume stationarity, which may not hold for all time series data, further impacting model performance.

The LSTM model also has its limitations. It requires a large amount of data for effective training and is prone to overfitting, especially with noisy or high-dimensional data. Hyperparameter tuning for LSTM models is computationally intensive and requires extensive experimentation. The interpretability of LSTM models is challenging due to their black-box nature, making it difficult to understand the mechanisms driving predictions. Like SARIMAX, LSTM models require accurate values of exogenous variables for prediction, with any inaccuracies potentially impacting forecast accuracy.

The results from the study are highly relevant for public health planning and resource allocation. By identifying main environmental predictors of AMI, health authorities can implement interventions that would be targeted together with resource

allocation efficiently during high risk periods. Additionally, the techniques used in this research can be generalized and expanded to other areas and health conditions, thus making this work more impactful.

5.2 Future Directions

Future research could explore several avenues to address the limitations identified in this study and further improve the predictive performance of time series models. One potential direction is to develop or integrate models that capture complex non-linear relationships between predictor and response variables, such as generalized additive models (GAMs) or machine learning approaches like random forests and gradient boosting. Implementing robust statistical methods to handle outliers and extreme values, such as robust regression or outlier detection and treatment techniques, could mitigate their impact on model performance.

Future studies could incorporate additional variables, such as socioeconomic factors and lifestyle variables, to improve the predictive accuracy and robustness of the models. Enhancing the accuracy and reliability of exogenous variable forecasts could also improve the overall model performance. Moreover, advanced modelling techniques could be explored to capture complex temporal dependencies and enhance prediction performance, including hybrid models that combine SARIMAX and LSTM or other deep learning architectures such as Transformer models. Conducting a spatial analysis of AMI incidences to identify geographical hotspots and understand regional variations in the impact of environmental factors could also be beneficial.

Additionally, future research could leverage ERA5 data for grid-level temperature and relative humidity to improve the granularity and precision of the environmental variables used in the models. Incorporating this high-resolution climate data could enhance the understanding of how microclimate variations influence AMI incidences and improve the predictive accuracy of the models.

Appendix A

Code

The code of this project is available in the following link: [Master Project GitHub repository](#).

Appendix B

Map of Catalonia split by Regions

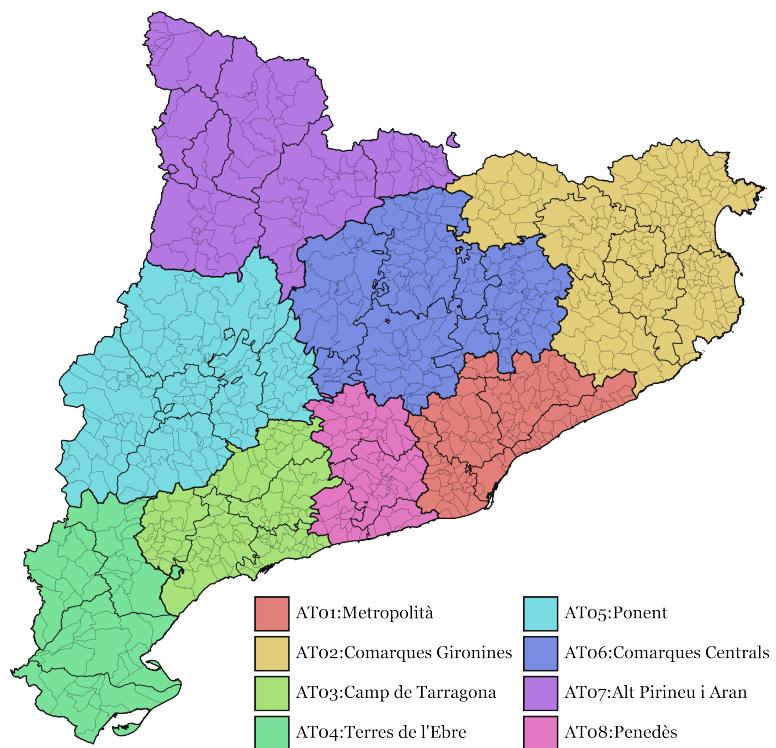


FIGURE B.1: Map of Catalonia split by Regions

Appendix C

Descriptive Statistics

FIGURE C.1: Decomposition of ASIR

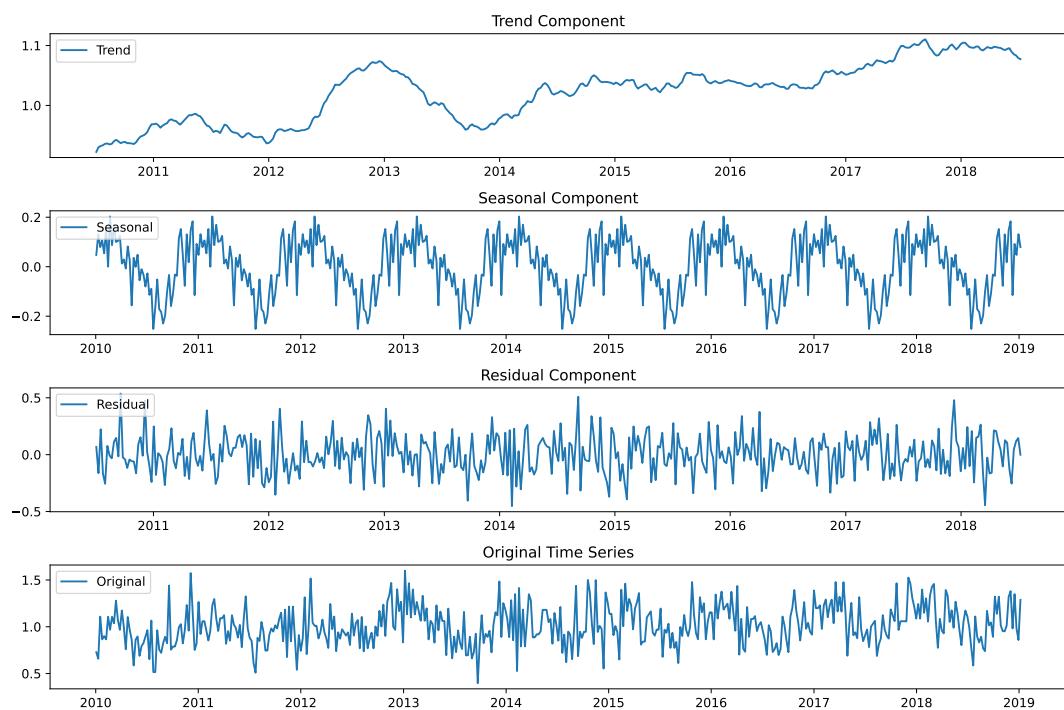


TABLE C.1: Age and Sex-Specific Trends in AMI Incidence with Statistical Significance

Age	Sex	Trend	T (0.025)	T (0.975)	R2	p-value	SE
20-24	Female	-0.033	-0.163	0.097	0.049	0.567	0.055
	Male	-0.063	-0.180	0.054	0.187	0.245	0.050
25-29	Female	-0.057	-0.180	0.066	0.146	0.310	0.052
	Male	0.059	-0.293	0.411	0.022	0.704	0.149
30-34	Female	-0.047	-0.182	0.087	0.091	0.431	0.057
	Male	0.323	-0.185	0.832	0.244	0.176	0.215
35-39	Female	0.125	-0.447	0.697	0.037	0.621	0.242
	Male	0.439	-0.259	1.137	0.240	0.180	0.295
40-44	Female	0.004	-0.456	0.463	0.000	0.985	0.194
	Male	0.031	-0.610	0.672	0.002	0.912	0.271
45-49	Female	0.316	-0.361	0.993	0.148	0.306	0.286
	Male	1.147	-0.145	2.439	0.386	0.074	0.546
50-54	Female	0.500	-0.296	1.297	0.240	0.181	0.337
	Male	3.770	1.052	6.488	0.606	0.013	1.149
55-59	Female	1.237	0.551	1.922	0.722	0.004	0.290
	Male	4.408	1.087	7.729	0.585	0.016	1.404
60-64	Female	1.525	0.752	2.298	0.757	0.002	0.327
	Male	4.698	1.955	7.441	0.701	0.005	1.160
65-69	Female	1.156	-0.001	2.313	0.444	0.050	0.489
	Male	2.760	-0.164	5.685	0.416	0.061	1.237
70-74	Female	0.160	-1.386	1.706	0.008	0.814	0.654
	Male	2.676	-0.707	6.058	0.333	0.104	1.431
75-79	Female	0.600	-2.114	3.315	0.038	0.617	1.148
	Male	7.172	2.852	11.492	0.688	0.006	1.827
80-84	Female	1.720	-0.539	3.980	0.317	0.115	0.955
	Male	5.457	0.616	10.298	0.504	0.032	2.047
85-89	Female	3.734	0.830	6.639	0.569	0.019	1.228
	Male	4.006	0.461	7.550	0.505	0.032	1.499
90-94	Female	1.512	-0.138	3.162	0.401	0.067	0.698
	Male	4.919	0.753	9.084	0.527	0.027	1.762
95+	Female	3.004	-0.823	6.831	0.330	0.106	1.618
	Male	5.388	-0.161	10.937	0.430	0.055	2.347

FIGURE C.2: Development of variables over years for Catalonia
on monthly basis

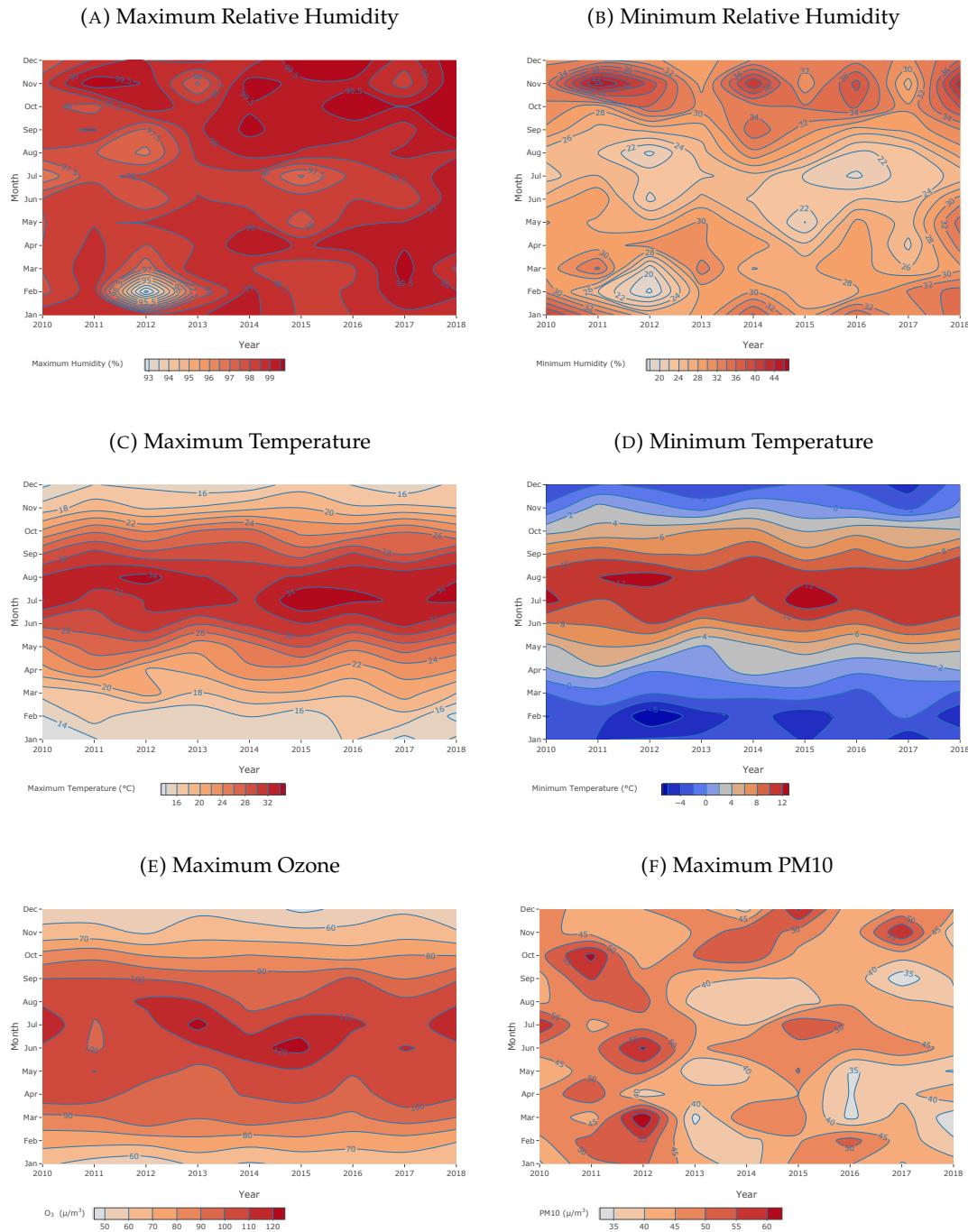


TABLE C.2: Descriptive statistics for daily levels of meteorological variables and air pollutant levels (Lag0) in Catalonia.

Variable	Mean	SD	Minimum	Q25	Median	Q75	Maximum	IQR
<i>All - 3287 days</i>								
Max Temperature	23.50	7.51	1.70	17.40	23.20	29.70	43.00	12.30
Min Temperature	3.30	7.03	-23.20	-1.80	3.10	8.70	20.70	10.50
Max Humidity	98.63	3.08	62.00	99.00	100.00	100.00	100.00	1.00
Min Humidity	29.23	13.66	0.00	20.00	28.00	37.00	89.00	17.00
Max CO	0.52	0.28	0.13	0.32	0.45	0.62	4.60	0.30
Max NO2	50.13	19.41	6.20	35.91	48.50	62.42	222.00	26.50
Max NO	43.37	39.66	1.00	14.10	31.46	60.19	531.00	46.09
Max PM10	44.60	22.29	4.00	31.50	40.75	52.55	658.00	21.05
Max SO2	5.94	5.52	0.67	3.00	4.50	7.00	110.80	4.00
Max O3	87.96	24.51	17.75	70.29	88.67	104.83	184.00	34.55
<i>Winter - 1631 days</i>								
Max Temperature	17.53	4.41	1.70	14.60	17.50	20.40	34.90	5.80
Min Temperature	-1.71	4.89	-23.20	-4.60	-1.40	1.60	12.70	6.20
Max Humidity	98.70	3.21	62.00	99.00	100.00	100.00	100.00	1.00
Min Humidity	31.14	15.33	0.00	20.00	30.00	41.00	89.00	21.00
Max CO	0.60	0.31	0.13	0.40	0.54	0.73	4.60	0.33
Max NO2	56.81	19.27	6.50	43.62	56.20	69.00	165.00	25.38
Max NO	59.37	45.81	1.20	23.54	49.26	84.90	531.00	61.36
Max PM10	44.50	21.05	4.00	30.71	41.00	53.67	317.50	22.95
Max SO2	6.64	5.22	0.67	3.50	5.18	8.00	89.29	4.50
Max O3	74.45	21.42	17.75	58.67	72.80	89.71	152.00	31.05
<i>Summer - 1656 days</i>								
Max Temperature	29.34	4.84	7.40	26.00	29.60	33.00	43.00	7.00
Min Temperature	8.21	5.08	-14.40	5.00	8.60	12.00	20.70	7.00
Max Humidity	98.55	2.95	71.00	98.00	100.00	100.00	100.00	2.00
Min Humidity	27.35	11.50	1.00	19.00	26.00	34.00	81.00	15.00
Max CO	0.43	0.22	0.13	0.30	0.40	0.50	2.90	0.20
Max NO2	43.59	17.19	6.20	31.29	41.74	53.36	222.00	22.08
Max NO	27.68	23.70	1.00	10.09	21.00	38.45	214.00	28.36
Max PM10	44.70	23.44	6.00	32.00	40.50	51.50	658.00	19.50
Max SO2	5.26	5.73	0.67	2.67	3.83	6.00	110.80	3.33
Max O3	101.19	19.65	18.89	87.44	100.00	114.00	184.00	26.56

Bibliography

- AirQuality, Data (2023). *Statistical Institute of Catalonia*. Accessed: June 30, 2024. URL: https://analisi.transparenciacatalunya.cat/en/Medi-Ambient/Qualitat-de-l-aire-als-punts-de-mesurament-manuals/qg74-87s9/about_data.
- Bhaskaran, Krishnan et al. (2010). "Short term effects of temperature on risk of myocardial infarction in England and Wales: time series regression analysis of the Myocardial Ischaemia National Audit Project (MINAP) registry". In: *Bmj* 341.
- Brochu, Eric, Vlad M Cora, and Nando De Freitas (2010). "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning". In: Preprint, Viewed 16 April 2022. URL: <https://arxiv.org/abs/1012.2599>.
- Canto, John G. et al. (2012). "Association of Age and Sex With Myocardial Infarction Symptom Presentation and In-Hospital Mortality". In: *JAMA* 307.8, pp. 813–822. DOI: [10.1001/jama.2012.199](https://doi.org/10.1001/jama.2012.199).
- Census, Data (2023). *Statistical Institute of Catalonia*. Accessed: June 30, 2024. URL: <https://www.idescat.cat/pub/?id=ep&lang=en>.
- Cesare, Mariachiara et al. (Jan. 2024). "The Heart of the World". In: *Global Heart* 19. DOI: [10.5334/gh.1288](https://doi.org/10.5334/gh.1288).
- Chaseling, Georgia et al. (Nov. 2021). "Extreme Heat and Cardiovascular Health: What a Cardiovascular Health Professional Should Know". In: *Canadian Journal of Cardiology* 37. DOI: [10.1016/j.cjca.2021.08.008](https://doi.org/10.1016/j.cjca.2021.08.008).
- De Sa, Christopher (2020). *Lecture 14: Hyperparameter Optimization*. Lecture notes, CS4787 — Principles of Large-Scale Machine Learning Systems, Cornell University, Delivered 2020. URL: <https://www.cs.cornell.edu/courses/cs4787/2020sp/lectures/Lecture14.pdf>.
- Freeborough, Warren and Terence van Zyl (Jan. 2022). "Investigating Explainability Methods in Recurrent Neural Network Architectures for Financial Time Series Data". In: *Applied Sciences* 12, p. 1427. DOI: [10.3390/app12031427](https://doi.org/10.3390/app12031427).
- Hrnjica, Bahrudin and Ognjen Bonacci (May 2019). "Lake Level Prediction using Feed Forward and Recurrent Neural Networks". In: *Water Resources Management*, pp. 1–14. DOI: [10.1007/s11269-019-02255-2](https://doi.org/10.1007/s11269-019-02255-2).
- Lindstrom, Megan et al. (2022). "Global burden of cardiovascular diseases and risks collaboration, 1990–2021". In: *Journal of the American College of Cardiology* 80.25, pp. 2372–2425.
- Mechanic, OJ, M Gavin, and SA Grossman (2024). "Acute Myocardial Infarction". In: *StatPearls*. [Updated 2023 Sep 3]. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK459269/>.
- Mehta, Rajendra H et al. (2001). "Acute myocardial infarction in the elderly: differences by age". In: *Journal of the American College of Cardiology* 38.3, pp. 736–741. DOI: [10.1016/S0735-1097\(01\)01432-2](https://doi.org/10.1016/S0735-1097(01)01432-2). eprint: <https://www.jacc.org/doi/pdf/10.1016/S0735-1097%2801%2901432-2>. URL: <https://www.jacc.org/doi/abs/10.1016/S0735-1097%2801%2901432-2>.

- Meteorological, Data (2023). *Statistical Institute of Catalonia*. Accessed: June 30, 2024.
URL: https://analisi.transparenciacatalunya.cat/en/Medi-Ambient/Dades-meteorologiques-de-la-XEMA/nzvn-apee/about_data.
- Mohammad, Moman A et al. (2018). "Association of weather with day-to-day incidence of myocardial infarction: a SWEDEHEART nationwide observational study". In: *JAMA cardiology* 3.11, pp. 1081–1089.
- Molnar, Christoph (2022). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. URL: <https://christophm.github.io/interpretable-ml-book>.
- Olah, Chris (2015). *Understanding LSTM Networks*. Accessed: 2024-06-19. URL: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>.
- O'Malley, Tom et al. (2019). *KerasTuner*. <https://github.com/keras-team/keras-tuner>.
- Peixeiro, Marco (2022). *Time Series Forecasting in Python*. Shelter Island, NY: Manning Publications. ISBN: 9781617299889.
- Shumway, Robert and David Stoffer (Jan. 2011). *Time Series Analysis and Its Applications With R Examples*. Vol. 9. ISBN: 978-1-4419-7864-6. DOI: [10.1007/978-1-4419-7865-3](https://doi.org/10.1007/978-1-4419-7865-3).
- Vaicius, Vidmantas et al. (Aug. 2021). "Association between winter cold spells and acute myocardial infarction in Lithuania 2000–2015". In: *Scientific Reports* 11. DOI: [10.1038/s41598-021-96366-9](https://doi.org/10.1038/s41598-021-96366-9).
- Vassalle, Cristina et al. (Apr. 2024). "Environmental Temperature, Other Climatic Variables, and Cardiometabolic Profile in Acute Myocardial Infarction". In: *Journal of Clinical Medicine* 13, p. 2098. DOI: [10.3390/jcm13072098](https://doi.org/10.3390/jcm13072098).
- Wu, Jia et al. (2019). "Hyperparameter optimization for machine learning models based on Bayesian optimization". In: *Journal of Electronic Science and Technology* 17.1, pp. 26–40.
- Zhang, Zefeng et al. (2012). "Age-Specific Gender Differences in In-Hospital Mortality by Type of Acute Myocardial Infarction". In: *The American Journal of Cardiology* 109.8, pp. 1097–1103. ISSN: 0002-9149. DOI: <https://doi.org/10.1016/j.amjcard.2011.12.001>. URL: <https://www.sciencedirect.com/science/article/pii/S0002914911035119>.
- Čulić, Viktor et al. (2002). "Symptom presentation of acute myocardial infarction: Influence of sex, age, and risk factors". In: *American Heart Journal* 144.6, pp. 1012–1017. ISSN: 0002-8703. DOI: <https://doi.org/10.1067/mhj.2002.125625>. URL: <https://www.sciencedirect.com/science/article/pii/S0002870302002259>.