

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:

Xavier Rodó,
Alejandro Fontal,
Laura Igual Muñoz

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

in the

Facultat de Matemàtiques i Informàtica

June 17, 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. It is established that both hot and cold temperatures influence the incidence of AMI, while the relationship between humidity and humidity during different seasons remains not so well understood. However, 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 modeling approaches: Distributed Lag Non-Linear Models (DLNM) 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 aimed at reducing 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; my classmates, who always provided hugs; Karolina and Katarina, who have 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. A song I want to remember when I find a printed version of this report in 30 years time: *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	3
2.2.1 Sex and Age as Significant Factors	4
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	11
3 Methodology	14
3.1 Distributed Lag Non-Linear Models	14
3.1.1 Model Structure	14
3.1.2 Basis Functions and Non-Linearity	14
3.1.3 Lag Structures	15
3.1.4 Cross-Basis Functions	15
3.1.5 Poisson and Negative Binomial Distributions - Families	16
The Poisson distribution	16
The Negative Binomial distribution	16
3.1.6 Estimation and Inference with Cross-Basis Functions	17
3.1.7 Applications and Considerations	17
3.1.8 Inference and Interpretation	18
3.2 The Recurrent Neural Network	18
3.2.1 Structure of an LSTM Cell	19
3.2.2 Working Mechanism	20
3.2.3 LSTM in practice	20
3.2.4 Hyperparameters optimisation	21
4 Results	23
LSTM Model Results Summary	23

5 Conclusion and future directions	25
5.1 Conclusion	25
5.2 Future Directions	25
A Map of Catalonia split by Regions	27
B Descriptive Statistics	28
Bibliography	31

List of Figures

2.1	Missing data in pollutants Over time	4
2.2	Spatial Distribution of Monitoring Stations	5
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	Population Estimates for Pendès (AT 08)	9
2.7	Moving Averages of ASIR Over Time	12
2.8	Decomposition of ASIR	12
2.9	Weekly and Monthly ASIR Over time	13
3.1	A single RNN Cell	19
3.2	A single LSTM Cell	21
A.1	Map of Catalonia split by Regions	27
B.1	Development of variables over years for Catalonia on monthly basis . .	29

List of Tables

2.1	AMI alerts by AT and year	11
B.1	Age and Sex-Specific Trends in AMI Incidence with Statistical Significance	28
B.2	Descriptive statistics for daily levels of meteorological variables and air pollutant levels (Lag0) in Catalonia.	30

List of Abbreviations

AMI	Acute Myocardial Infarction
ASIR	Age-Standardized Incidence Rates
CVD	Cardiovascular Disease
DLNM	Distributed Lag Non-Linear Models
ERF	Exposure-Response Function
LASSO	Least Absolute Shrinkage and Selection Operator
LSTM	Long Short-Term Memory
NO₂	Nitrogen Dioxide
PM	Particulate Matter
RNN	Recurrent Neural Networks
SARIMA	Seasonal Autoregressive Integrated Moving Average
SMBO	Sequential Model-Based Optimization

Chapter 1

Introduction

Cardiovascular diseases (CVDs) remain a significant public health concern globally, contributing substantially to morbidity and mortality rates. Among the various manifestations of CVDs, acute myocardial infarction (AMI) is not only one of the leading causes of mortality but also it stands out as a critical condition 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 the incidence of AMI gained an increased attention in recent years, urging researchers to delve deeper into understanding the complexity involved.

In regions characterised by diverse climatic conditions, such as Catalonia, Spain, where temperature fluctuations, humidity levels, and pollution levels exhibit considerable variability across different seasons and different geographical locations, exploring the nexus between environmental parameters and the occurrence of AMI becomes especially relevant. Catalonia, with its unique blend of coastal and inland areas, urban centres, rural landscapes, mountains and sea provides a compelling place 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 is performed using a dataset encompassing hospital admissions over a time-frame between years 2010 to 2018 across 948 municipalities stratified by province, sex, and age, and daily meteorological and pollution data, that is used to understand the dynamics and interaction of the variables.

While existing literature indicates that both hot and cold temperatures have an impact on the incidence of AMI, the relationship between humidity 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 is yet to be established.

To address these gaps, this study proposes a multifaceted approach that incorporates two different methodologies, specifically Seasonal Autoregressive Integrated Moving Average (SARIMA) models, and Long Short-Term Memory (LSTM) machine learning algorithms. Combining these models with 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 the climate change, there has been a discussion about the relationship between environmental factors and the incidence of AMI. As the changing of 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 to 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 segments of the population. As a result, understanding the dynamics can help to create interventions that would be aimed at specific most susceptible group and reduce the incidence and impact of AMI in Catalonia.

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 process of hyperparameter optimization 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. We then acknowledge the limitations of our study, such as potential data quality issues and the assumptions inherent in our modeling approaches. Finally in section 5, the paper concludes with a summary of 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

For the purposes of the analysis there have been various data sources used. The environmental dataset under investigation composed of temperature, relative humidity, air pollution data, and census data was obtained from the Statistical Institute of Catalonia. The census data was obtained from the population estimates, which provides 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, chlorine, carbon monoxide, hydrogen sulfide, mercury, nitric oxide, nitrogen dioxide, nitrogen oxides, ozone, particulate matter and sulfur dioxide. However, only ozone 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).

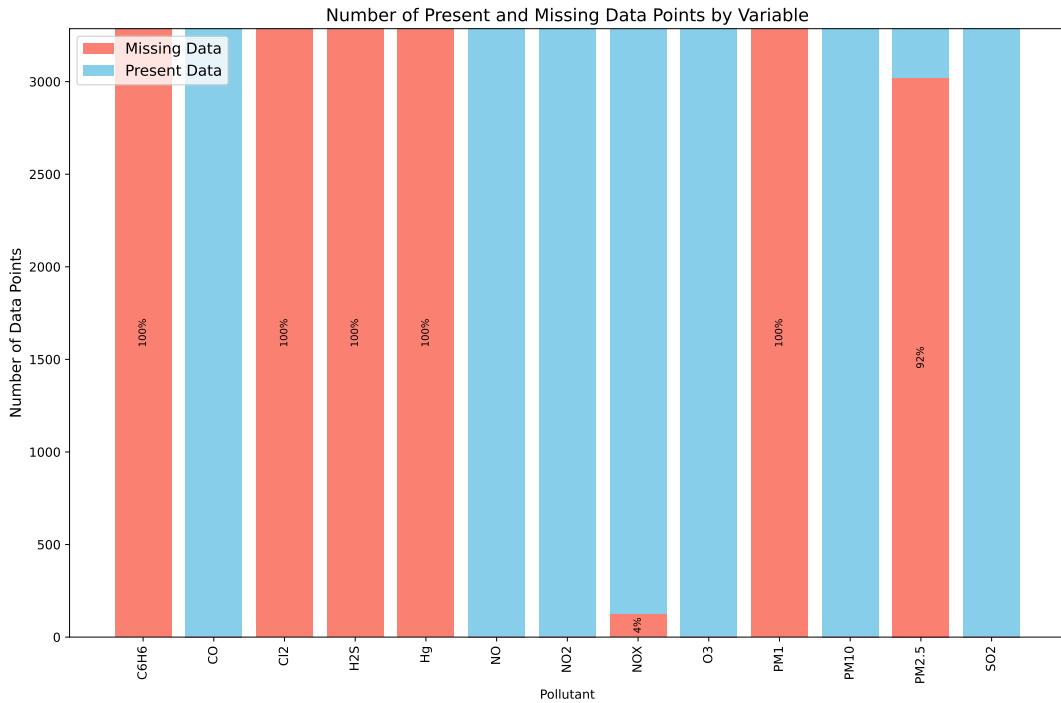
Similarly, meteorological observations, including outdoor temperature and relative humidity, were collected from 239 meteorological stations distributed throughout Catalonia, as illustrated in Figure 2.2b. Raw measurements, initially recorded at half-hour intervals, were aggregated to daily data points 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 limited data density hindered our ability to extract meaningful insights and detect significant trends. As a reason, we used higher-level regional territories.

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

FIGURE 2.1: Missing data in pollutants Over time



may contribute to variations in AMI incidence within Catalonia, Spain. In this section, the exploratory analysis is shown on aggregated data for the separate high-level regions.

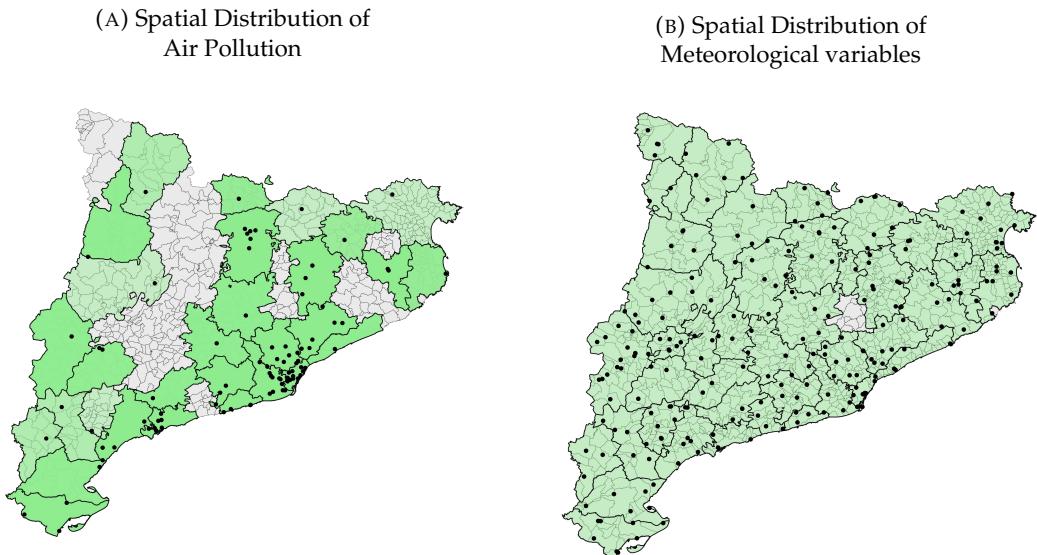
As noted by Mechanic, Gavin, and Grossman (2024) there are many risk factors potentially contributing to the occurrence of AMI. Considering the set of non-modifiable risk factors that include factors such as sex and age. The further analysis of associations of these variables was conducted by Canto et al. (2012).

2.2.1 Sex and Age as Significant Factors

The role of sex as an important 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. While men traditionally exhibit a higher overall incidence of AMI, women often present with AMI at older ages 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 greater proportion of women have airway hyper-responsiveness than men. Hormonal influences, anatomical differences, and variations in risk factor prevalence contribute to these sex-specific disparities. In order to explore the role of sex in AMI incidence within Catalonia, it is necessary to disaggregate data by sex and examine potential differences in risk factor profiles, symptom presentation, and healthcare utilisation patterns between males and females.

Age is a significant predictor of AMI incidence. Exploratory analysis together with previous academic findings suggest that there are certain age groups which are more susceptible to the effects environmental factors. The risk increases with advancing age. Age-related changes in cardiovascular structure and function, coupled

FIGURE 2.2: Spatial Distribution of Monitoring Stations



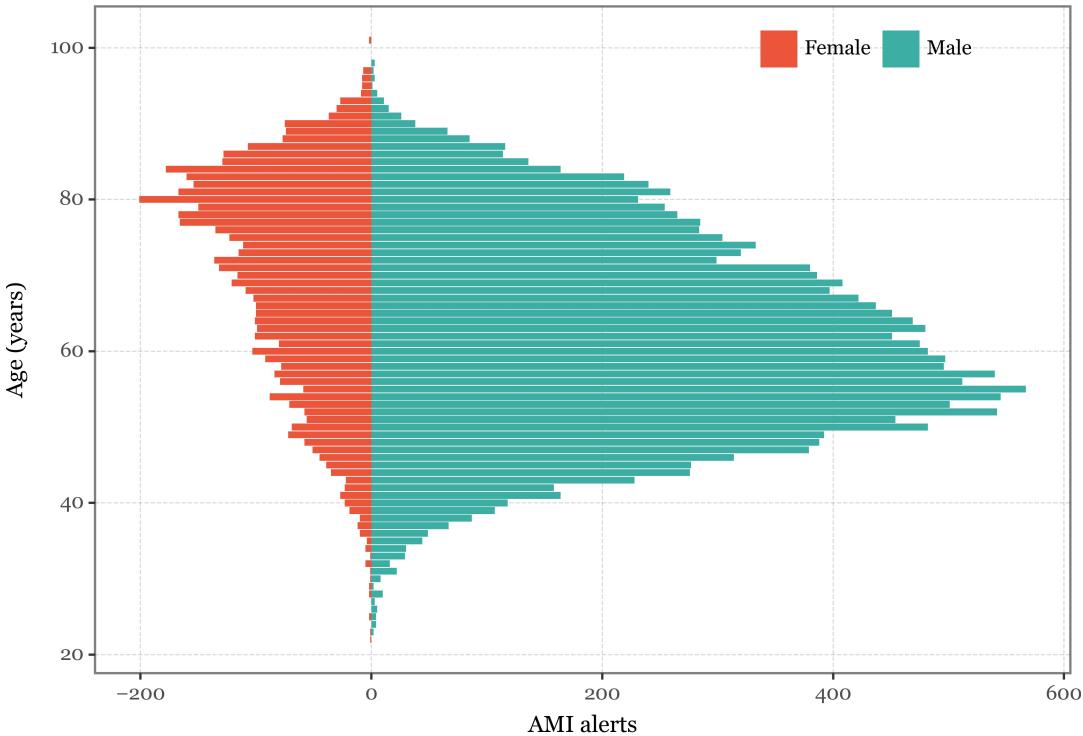
with the cumulative effects of traditional risk factors such as hypertension, diabetes, and dyslipidemia, contribute to the elevated AMI risk observed in older adults. In order to explore age-specific trends in AMI incidence within Catalonia, it is necessary to stratify the population into distinct age 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 when it comes to females is around their eighties. On the other side, when it comes to men their peak is around fifties. Both of these results are consistent with the previous paragraph. It suggests that there is some dependence on population structure. Thus, when dis-aggregating data spatially for any given area, there is a need to take into account the population structure.

Since we are dealing with dataset during 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 factors of AMI alerts, we should take into consideration the changes in population structures in the examined period. In other words, dealing with population structure, there is a potential of ageing population that could 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 ageing of the population are somehow correlated. By combining both variables, we calculate 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 exhibited an increasing trend over the years, indicating a 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 yearly incidence of each group and plotted the slope of the regression line along with the 95% confidence interval. While the slope for many age and sex groups is not significantly different from zero, certain groups

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



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 B.1 in Appendix B). Combining all information together, we proceed and define our target variable.

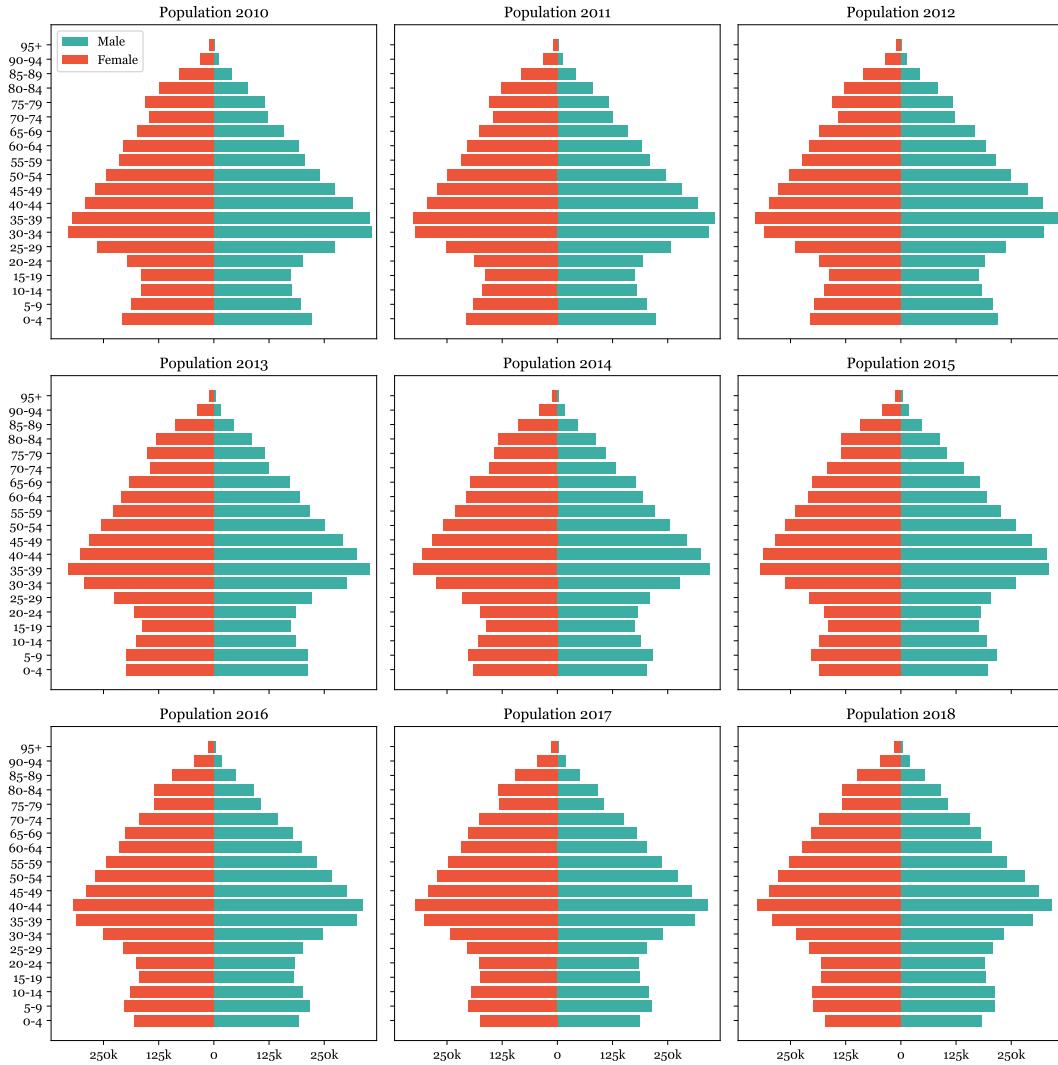
2.2.2 Choice of Age-Standardized Incidence Rates (ASIR)

In the analysis of 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 choose to use Age-Standardized Incidence Rates (ASIR), which adjust for differences in age distribution across populations.

ASIR is a useful metric for comparing disease incidence rates between populations or over time while accounting for differences in age structure. By standardizing 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 on 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. By standardizing AMI incidence rates to a reference population, such as the World

FIGURE 2.4: Development of Population (2010 - 2018)

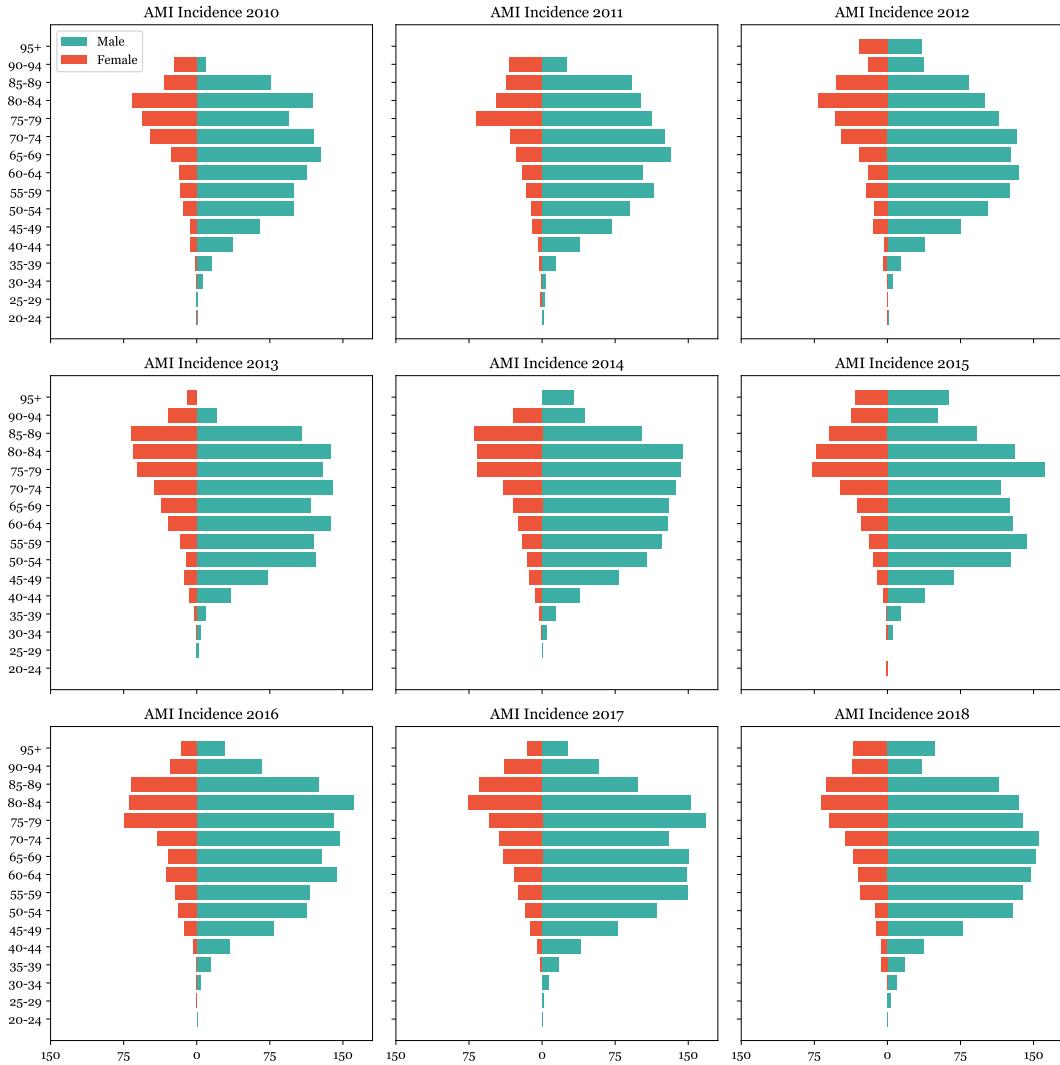


Health Organization (WHO) standard population, we can obtain ASIR estimates that account for differences in age distribution, enabling more robust comparisons of AMI incidence across municipalities and over time.

Our analysis begins with the estimation of AMI incidence over a year, stratified by 5-year age groups and sex, utilizing comprehensive datasets. Subsequently, population distributions sourced from IDECAT are employed to compute the expected number of AMI events for each region and year. By juxtaposing these expected values with the actual observed events, we can discern trends in AMI incidence that are independent of demographic influences, providing a nuanced understanding of underlying patterns.

The calculation of ASIR represents a pivotal step in standardizing incidence rates to a reference population, thereby facilitating equitable comparisons across regions and temporal intervals. This standardization process effectively mitigates the confounding effects of age and sex distributions, enabling a more accurate assessment of true 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)



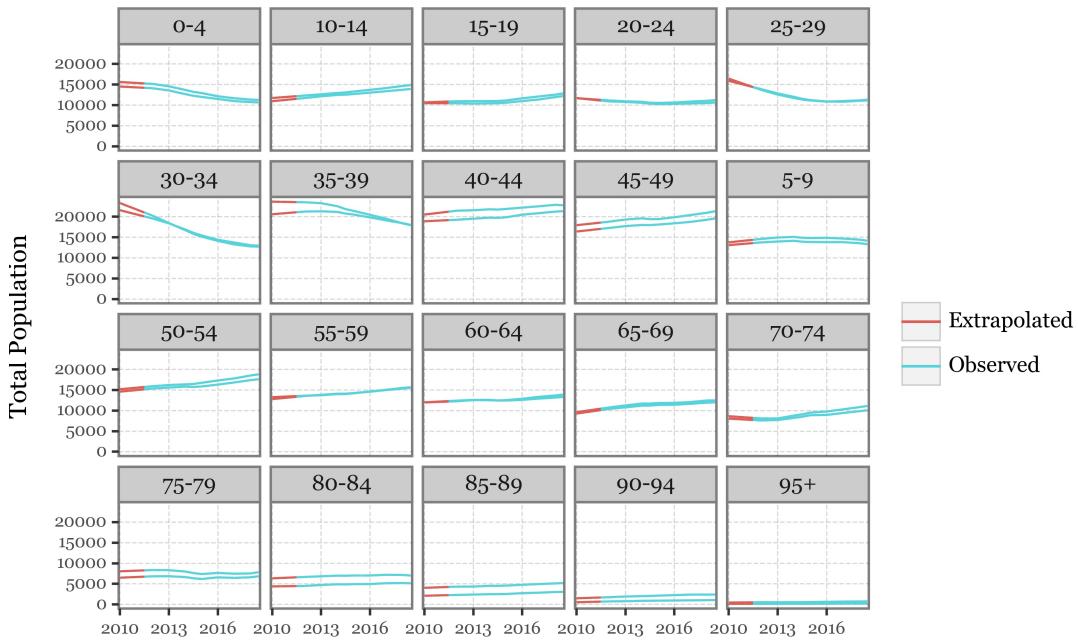
Visualization of raw AMI case data reveals intricate temporal patterns, including daily, weekly, monthly, and yearly fluctuations. These patterns, while indicative of potential seasonal variations and long-term trends, may also be influenced by shifts in population demographics. By employing ASIR, we 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 standardizing 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

FIGURE 2.6: Population Estimates for Penedès (AT 08)



AMI incidence rates and adjusting for population differences, we aim to identify patterns and trends that may inform targeted public health interventions and resource allocation strategies. The geographical distribution of AMI incidence within Catalonia is influenced by a myriad of factors, including urbanisation, socioeconomic status, healthcare infrastructure, and environmental exposures. 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.

Data Wrangling

For the purposes of the analysis, there is a need to obtain a daily measurements of the data, similarly as with air pollution, AMI and meteorological variables. In order to assign the number of cases for each area we have to account for population structure of the given high-level region. In other words, it is needed to take the differences in population demographics into account.

When dealing with the population data on high-level region, Penedès as the last territorial ambit, has not been officially established until late 2011. As a consequence, the population data on this spacial level are inconsistent. Since we assigned each municipality (and thus every AMI alert) to its current territorial ambit, our incidence estimates for all those territorial ambits in 2010 would be inaccurate as well. To address this challenge, considering that the population changes linearly year-on-year, we will exclude the data for 2010 and extrapolate from the period between 2011 and 2022 (see Figure 2.6).

Lastly, to obtain daily population data by linearly interpolating the population estimates that were on the six month county level to fill in missing values and ensure the comprehensibility of the dataset for analysis.

Counties

First, we examine the AMI alerts per 100,000 inhabitants for each county and year. 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.

However, reporting AMI incidence per 100,000 inhabitants may skew results if not accounting for population structure differences among counties. To address this limitation, we compute the expected number of cases for each county and year using the respective population structures. Subsequently, we calculate the ASIR for each county and year, providing a more accurate representation of AMI incidence that adjusts for demographic variations.

Visualizing the ASIR for each county and year offers insights into the spatial distribution of AMI incidence rates, facilitating 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 the identification of high-risk areas and the implementation of targeted interventions.

Àmbits Territorials

Next, we extend our analysis to the territorial regions (àmbits territorials) in Catalonia, which offer larger and more stable geographical units for examination. Similar to 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. However, differences in population structure among territorial regions may influence observed incidence rates. To address this issue, we compute the expected number of cases for each territorial region and year using population data.

Subsequently, we calculate the ASIR for each territorial region and year, enabling a comprehensive assessment of AMI incidence trends while adjusting for demographic differences. The analysis highlights areas with elevated ASIR, indicating regions of heightened AMI risk.

By comparing expected and observed AMI incidence rates, we identify deviations from expected trends, providing valuable insights into spatial patterns of AMI occurrence. This information can inform targeted public health interventions and resource allocation efforts, ultimately contributing to the prevention and management of AMI at a regional level.

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 greatly, 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, the Table ?? illustrates the total number of AMI alerts for each AT and year.

To address this issue of high variability, we utilize moving averages to smooth the ASIR data and extract hidden trends while minimizing the impact of short-term fluctuations. By interpolating the expected daily AMI alerts based on population

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	

estimates for every 6 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.

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

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

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 ,
- $\theta_1, \theta_2, \dots, \theta_q$ 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 in the model.

For smoothing ASIR by using moving averages of different number of days see Figure 2.7. From the figure, we can slightly observe a long-term trend with several peaks within the examined period and also there is 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 2.8). We can see the long-term change in the series which presents increase over time, the seasonal pattern which we can see an annual repeated fluctuations. It is important to note that when smoothing the data there is a need for a balance between the noise reduction and still maintaining the integrity of the original data. Thus, the data will be aggregated on a weekly level. For the comparison see Figure 2.9.

2.2.4 Seasonal Variations and Weather Patterns

There is a growing body of evidence indicating that seasonal variations in weather conditions, such as temperature, humidity, and air pollution levels, are associated

FIGURE 2.7: Moving Averages of ASIR Over Time

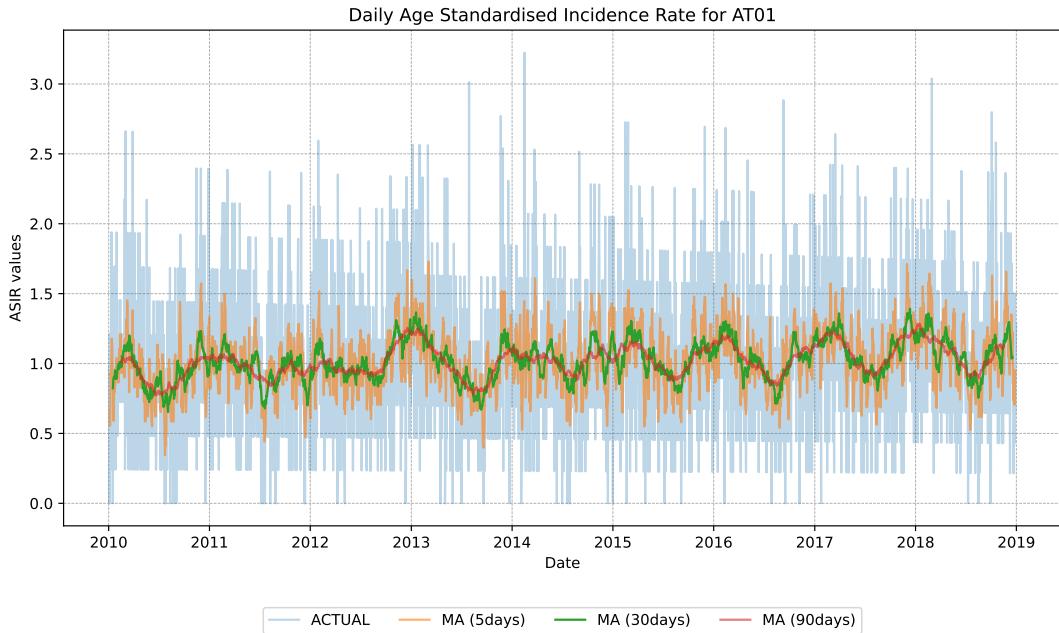
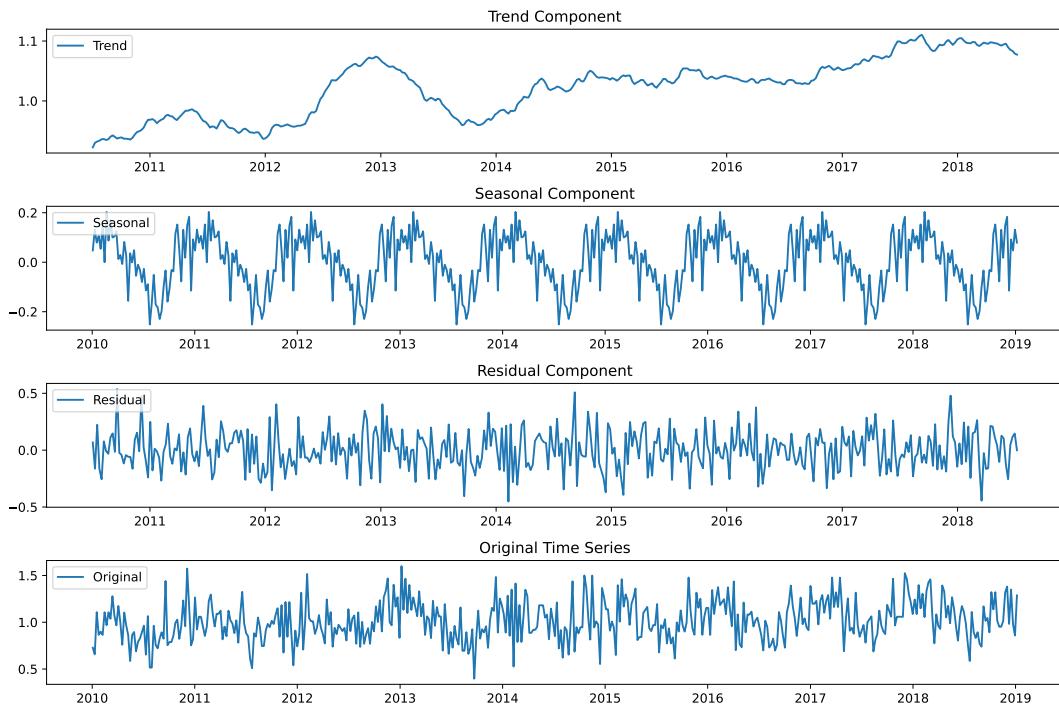
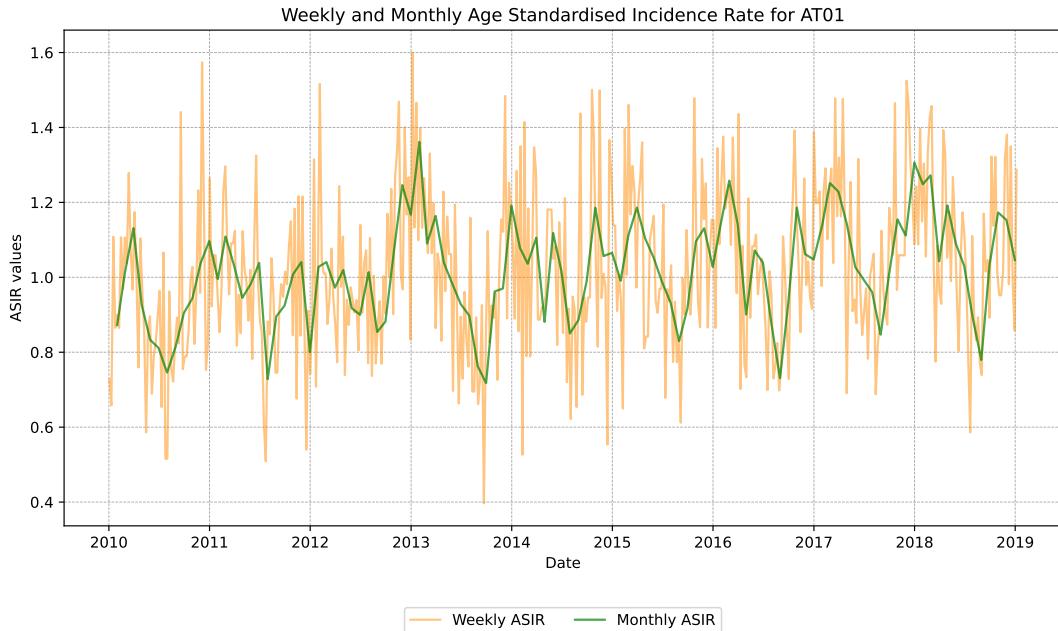


FIGURE 2.8: Decomposition of ASIR



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. Conversely, high temperatures and summer-related factors, such as dehydration,

FIGURE 2.9: Weekly and Monthly ASIR Over time



outdoor physical exertion, and increased air pollution levels, may exacerbate cardiovascular risk during summer months.

To analyze seasonal trends in AMI incidence within Catalonia, it is necessary to assess the temporal patterns of AMI occurrence, identify seasonal peaks and troughs, and 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 period based on their average temperatures. The summer season is from May - October and the rest of the months are winter period.

The Table B.2 in Appendix B provides a summary of key 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 season, maximum and minimum temperatures are lower compared to the summer season, resulting in lower levels of humidity and air pollutants such as ozone (O₃) and nitrogen dioxide (NO₂). Conversely, during the summer season, we see higher temperatures, increased humidity levels, and elevated concentrations of air pollutants, particularly ozone and nitrogen dioxide. These findings highlight the importance of considering seasonal variations in weather conditions when analyzing cardiovascular health outcomes and planning public health interventions. For the monthly development of variables see Figure B.1 in Appendix B.

Chapter 3

Methodology

3.1 Distributed Lag Non-Linear Models

Distributed Lag Non-Linear Models (DLNMs) are a class of statistical models used to analyze time series data where both the predictor and the outcome can exhibit non-linear and delayed relationships. These models are particularly useful in fields such as environmental epidemiology, econometrics, and climate science, where understanding the time-varying impact of an exposure or intervention on an outcome is crucial. Gasparrini (2011) describes the dynamics as an outcome that is explained by multiple exposure events in the past. In other words, the framework leverages by their bi-dimensional relationship along space of the predictors and time (dimension of lags).

3.1.1 Model Structure

DLNMs extend traditional linear models by incorporating distributed lag effects and allowing for non-linear relationships between predictors and outcomes. By relaxing the linearity assumption, it offers an approach to modelling the nonlinear and lagged effects of exposures. The core idea is to model the effect of a predictor variable not only at the current time point but also at several preceding time points, capturing the delayed impact over time.

The general form of a DLNM can be represented as:

$$Y_t = \alpha + \sum_{j=0}^L f_j(X_{t-j}) + \epsilon_t$$

where:

Y_t is the outcome variable at time t .

X_{t-j} is the predictor variable at time $t - j$, with j ranging from 0 to L , the maximum lag.

f_j represents a non-linear function describing the relationship between X_{t-j} and Y_t .

α is the intercept term.

ϵ_t is the error term at time t .

3.1.2 Basis Functions and Non-Linearity

A key component of DLNMs is the use of basis functions, sometimes referred to as the exposure-response functions ERF, to model non-linear relationships. They

can be characterised as the functional form of the relationship between the exposure variable X and the health outcome Y . Basis functions transform the predictor variables into a set of terms that can capture complex, non-linear patterns. Common choices for basis functions include polynomials, splines (such as B-splines or natural splines), and step functions.

For instance, using B-splines, the function $f_j(X_{t-j})$ can be expressed as:

$$f_j(X_{t-j}) = \sum_{k=1}^K \beta_{jk} B_k(X_{t-j})$$

where:

B_k are the B-spline basis functions.

β_{jk} are the coefficients to be estimated.

3.1.3 Lag Structures

DLNMs incorporate lag structures to account for delayed effects. These formulations allow for flexible modeling of the temporal dynamics of exposure effects, capturing both short-term and long-term lagged effects. The lag dimension is typically modeled using a separate set of basis functions, allowing the lagged effects to vary flexibly over time. This can be represented as:

$$\sum_{j=0}^L f_j(X_{t-j}) = \sum_{j=0}^L \sum_{k=1}^K \beta_{jk} B_k(X_{t-j})$$

Alternatively, a tensor product of basis functions can be used to jointly model the non-linear and lagged effects, providing a more comprehensive framework:

$$Y_t = \alpha + \sum_{k=1}^K \sum_{m=1}^M \beta_{km} B_k(X_t) \cdot L_m(j) + \epsilon_t$$

where:

$L_m(j)$ are the basis functions for the lag structure.

β_{km} are the coefficients for the interaction between the predictor and lag basis functions.

3.1.4 Cross-Basis Functions

Cross-basis functions are a fundamental extension in DLNMs that allow for the simultaneous modeling of non-linear relationships in both the predictor and lag dimensions. This approach uses a tensor product of basis functions to capture the complex interactions between the exposure levels and their distributed lag effects.

A cross-basis function can be represented as:

$$CB(X_{t-j}, j) = \sum_{k=1}^K \sum_{m=1}^M \beta_{km} B_k(X_{t-j}) L_m(j)$$

where:

$CB(X_{t-j}, j)$ is the cross-basis function evaluating the effect of the predictor X_{t-j} at lag j .

$B_k(X_{t-j})$ are the basis functions for the predictor.

$L_m(j)$ are the basis functions for the lag.

β_{km} are the coefficients to be estimated.

The tensor product of basis functions allows the model to capture interactions where the effect of the predictor varies smoothly over both the value of the predictor and the lag.

3.1.5 Poisson and Negative Binomial Distributions - Families

In the analysis of AMI incidence and its relationship with environmental factors, such as temperature, humidity, and pollution levels, the choice of probability distribution family is crucial for accurately modeling count data and estimating exposure effects. Both the Poisson and Negative Binomial distributions offer valuable frameworks for modeling count data, each with its own strengths and applicability depending on the characteristics of the data and the underlying mechanisms driving AMI incidence.

The Poisson distribution

The Poisson distribution is well-suited for modeling count data where events occur independently and at a constant rate over time or space. The probability mass function of the Poisson distribution is given by:

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

where: - X is a random variable representing the number of events, - k is a non-negative integer representing the number of events, - λ is the mean or rate parameter of the distribution, and - e is the base of the natural logarithm.

In the context of AMI incidence and environmental exposures, the Poisson distribution may be appropriate when the incidence of AMI events is relatively low, and events occur randomly and independently. This distribution provides a straightforward and interpretable framework for modeling the count of AMI cases, particularly when investigating short-term effects of environmental exposures on AMI incidence.

The Negative Binomial distribution

On the other hand, the Negative Binomial distribution is more appropriate for modeling overdispersed count data, where the variance exceeds the mean. The probability mass function of the Negative Binomial distribution is given by:

$$P(X = k) = \frac{\Gamma(k + \theta)}{\Gamma(k + 1)\Gamma(\theta)} \left(\frac{\mu}{\mu + \theta} \right)^k \left(\frac{\theta}{\mu + \theta} \right)^\theta$$

where: - X is a random variable representing the number of events, - k is a non-negative integer representing the number of events, - μ is the mean parameter of the distribution, - θ is the dispersion parameter, and - Γ denotes the gamma function.

In the context of AMI incidence and environmental exposures, overdispersion may arise due to various factors, including unobserved heterogeneity, spatial clustering of cases, and temporal trends. If the variance in AMI incidence exceeds what

would be expected under a Poisson distribution, indicating greater variability or dispersion in the data, the Negative Binomial distribution offers a more suitable framework. This distribution allows for greater flexibility in capturing the extra variability in AMI incidence attributable to environmental exposures, facilitating more accurate estimation of exposure effects and uncertainty.

In summary, the choice between the Poisson and Negative Binomial distributions depends on the characteristics of the AMI data and the underlying mechanisms driving AMI incidence in response to environmental factors. While the Poisson distribution is suitable for modeling equidispersed count data with a constant rate of event occurrence, the Negative Binomial distribution offers greater flexibility in capturing overdispersion and variability in event counts, making it more appropriate for modeling AMI incidence in the presence of complex environmental exposures. By carefully selecting the appropriate probability distribution family, researchers can ensure the validity of statistical inference and gain valuable insights into the relationship between environmental factors and AMI incidence.

3.1.6 Estimation and Inference with Cross-Basis Functions

The parameters of DLNMs with cross-basis functions are estimated using methods such as maximum likelihood estimation (MLE) or Bayesian approaches. These methods involve specifying the appropriate basis functions and their corresponding degrees of freedom, as well as the maximum lag L .

Inference in models with cross-basis functions includes evaluating the overall effect of the predictor across different lags and exposure levels. This includes calculating overall cumulative effects over the lag period, generating plots of the lag-response and exposure-response surfaces and assessing the significance of these effects using confidence intervals and hypothesis tests to evaluate the uncertainty associated with the model estimates.

When applying cross-basis functions, it is crucial to consider:

- The selection and number of basis functions for both dimensions to balance flexibility and overfitting.

- The interpretability of the resulting surfaces and the practical implications of the estimated effects.

- The computational complexity, as the tensor product increases the number of parameters to be estimated.

Once the cross-basis functions are specified, DLNM is estimated using standard regression techniques, such as generalized linear models (GLMs) or generalized additive models (GAMs). The parameters of the cross-basis functions, including the coefficients β and γ , are estimated from the data using maximum likelihood estimation or other optimization algorithms. Model estimation involves fitting the DLNM to the observed data, adjusting for potential confounders and covariates.

3.1.7 Applications and Considerations

DLNMs are widely used in applications where delayed and non-linear effects are expected. For example, in environmental health studies, DLNMs can assess the impact of air pollution on health outcomes over time. In finance, they can model the effect of economic indicators on stock market returns with delayed reactions.

When applying DLNMs, it is crucial to consider issues such as:

Selection of the appropriate basis functions and the number of knots.

Determining the maximum lag L based on subject-matter knowledge or data-driven methods.

Checking for model assumptions and potential overfitting.

The theoretical framework of DLNMs provides a robust and flexible approach for analyzing complex time series data with non-linear and lagged effects. By incorporating advanced techniques such as basis functions and tensor products, DLNMs enable a nuanced understanding of the dynamic relationships between predictors and outcomes.

3.1.8 Inference and Interpretation

Once validated, DLNM models are used for inference and interpretation to quantify the effects of environmental exposures on health outcomes. Researchers can estimate exposure-response curves, identify lagged effects, and quantify the magnitude and uncertainty of exposure effects. Statistical inference, such as hypothesis testing and confidence interval estimation, allows for the assessment of the significance and robustness of the estimated effects. The results provide valuable insights into the temporal dynamics of exposure-response relationships, informing public health interventions and policies aimed at mitigating the health impacts of environmental exposures.

By leveraging DLNM methodology, researchers can uncover complex and delayed relationships between environmental exposures and health outcomes, facilitating a deeper understanding of disease etiology and guiding evidence-based interventions for disease prevention and control.

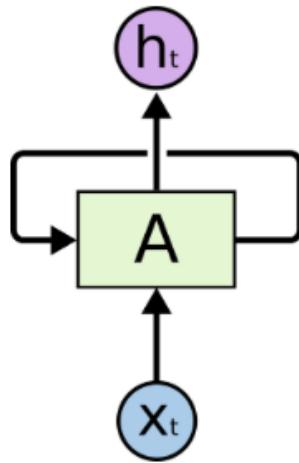
3.2 The Recurrent Neural Network

In the field of deep learning, Recurrent Neural Networks (RNN) are defined as artificial neural networks that are bi-directional. As being part of a supervised learning their usage is universal and also multidisciplinary. They are commonly used for time-series data analysis and forecasting, classification, regression, image/video processing and many other problems. The main 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 affect on other input provided to the same nodes. Figure 3.1 shows an example of a single RNN cell. The main setbacks of RNN are 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 problem implying that the model stops learning.

To overcome most of the problems, Long Short-Term Memory networks (LSTM), are a special kind of RNN, that were designed. 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, has a chain like structure, as a network it has a special design. It contains only four neural networks and different memory cells, that regulate flow of information. Each contain three main components - input, forget and output gates, which are multiplicative units that help to store, update and retrieve information over long sequences.

FIGURE 3.1: A single RNN Cell

Source: Medium



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 which makes 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 output of the network. In other words, it can be seen as an encoding of 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 a number between 0 and 1 for each number in the cell state C_{t-1} . Outputs close to 0 mean that the information is forgotten and close to 1 the information retains and is 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 with the use of 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 *tanh* function is to normalise encoding of the data.

- 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 serves as 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.

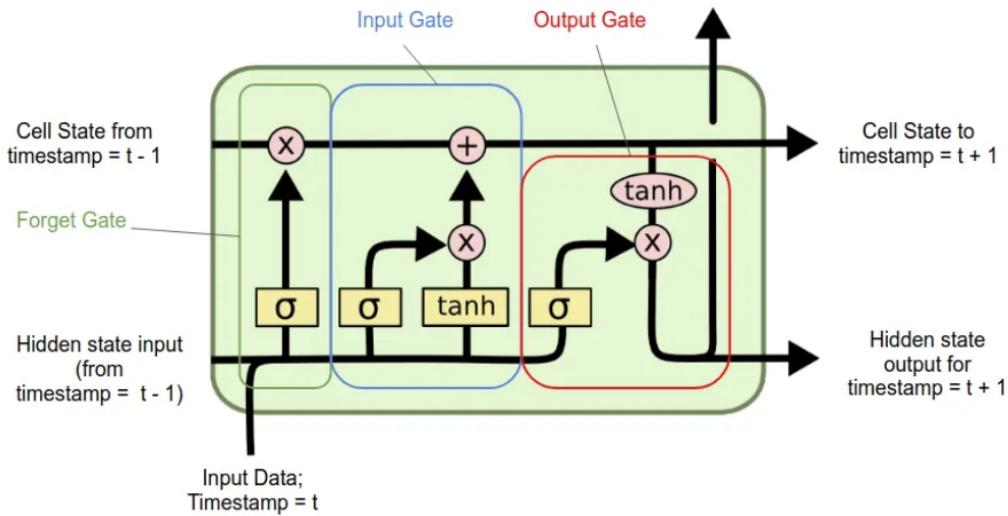
Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) units to predict the incidence of AMI based on environmental factors. RNN-LSTM is chosen for its capability to model time series data with complex temporal dependencies and non-linear relationships thanks to LSTMs feedback connections.

3.2.3 LSTM in practice

Prior to training the models, there has been data preprocessing involved so the data would be suitable for LSTM. Normalization was performed using MinMaxScaler to

FIGURE 3.2: A single LSTM Cell

Source: Medium



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 common 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 maximizes 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 the smallest loss. In random search, the combinations of hyperparameter settings at each iteration are independent.

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 next 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 (which is an approximation of the loss function that can be expressed as a probability distribution) is formed based on the sampled points. The algorithm then identifies the next combination of hyperparameter values such that it compromises between searching previously unexplored areas of the surrogate function with high uncertainty, and areas which lead 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 regularization technique used to prevent 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 stronger regularization.

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 used to change the attributes of the neural network such as weights and learning rate in order to reduce the losses. Common optimisers include Adam, RMSprop, and SGD.

Chapter 4

Results

LSTM Model Results Summary

Keras Tuner, which is frequently used to find optimal parameters for a neural networks (O'Malley et al., 2019), was utilized to optimize the hyperparameters of the LSTM model. The following hyperparameters were determined using Keras Tuner, which involved a random search over the specified hyperparameter space. The search process used a maximum of 30 trials, with each trial executed 3 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:
[colback=white,colframe=black, title=Model Architecture]

```
Model: "sequential"
-----
Layer (type)           Shape          Param #
-----
lstm (LSTM)            (10, 150)      91200
-----
dropout (Dropout: 0.2) (10, 150)      0
-----
lstm_1 (LSTM)          (10, 150)      180600
-----
dropout_1 (Dropout: 0.2)(10, 150)      0
-----
lstm_2 (LSTM)          (150,)         180600
-----
dense (Dense)          (1,)           151
-----
Total params: 452,551
Trainable params: 452,551
Non-trainable params: 0
-----
Optimizer: Adam
Activation Function: ReLU
```

The model was compiled with the Adam optimizer 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 a 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 in mitigating overfitting, and the choice of the Adam optimizer ensured efficient and effective convergence of the model parameters.

Chapter 5

Conclusion and future directions

5.1 Conclusion

In summary, this study aimed to investigate the relationship between environmental factors and the incidence of AMI in Catalonia, Spain, through the lens of time series analysis. Encompassing hospital admissions, meteorological, and pollution data, we employed two distinct modeling approaches: Seasonal Autoregressive Integrated Moving Average (SARIMA) models and Long Short-Term Memory (LSTM) neural networks.

Our analysis revealed insights into the temporal dynamics of AMI incidence, with both SARIMA and LSTM models demonstrating their efficacy in capturing and predicting patterns within the data. SARIMA models, leveraging their ability to model seasonal variations and autocorrelation, provided valuable insights into the short-term fluctuations and seasonal trends in AMI occurrences. Conversely, LSTM networks, with their capacity to capture complex non-linear relationships, offered a deeper understanding of the underlying dynamics and long-term trends in AMI incidence.

The important variables identified in our analysis include:

- Maximum temperature (max_temp): A decrease of 1°C in maximum temperature is associated with a decrease of approximately 0.345 units in the Age-Standardized Incidence Rate (ASIR) of AMI.
- Mean PM10 concentration (mean_PM10): An increase of 1 $\mu\text{g}/\text{m}^3$ in mean PM10 concentration is associated with an increase of approximately 0.146 units in ASIR of AMI.

Through model evaluation metrics such as AIC, BIC, MAE, MSE, RMSE, we assessed the performance of each model, considering their ability to accurately predict AMI incidence while balancing model complexity. The results underscored the utility of both SARIMAX and LSTM models in capturing the dynamics between environmental factors and cardiovascular health outcomes. Our findings help with setting up interventions aimed at mitigating the risk of AMI by considering the impact of these important environmental factors.

5.2 Future Directions

This project has laid the foundation for understanding the impact of environmental factors on the incidence of AMI using time series analysis and machine learning techniques. However, several avenues for future research and improvement exist. Future studies could incorporate additional variables, such as socioeconomic factors, and lifestyle variables, to improve the predictive accuracy and robustness of

the models. Advanced modeling techniques, including hybrid models that combine SARIMAX and LSTM, or other deep learning architectures such as Transformer models, could be explored to capture complex temporal dependencies and enhance prediction performance. 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. Developing a real-time prediction system that continuously updates with new data would provide timely alerts and forecasts to healthcare providers and policymakers for proactive intervention and resource allocation.

Appendix A

Map of Catalonia split by Regions

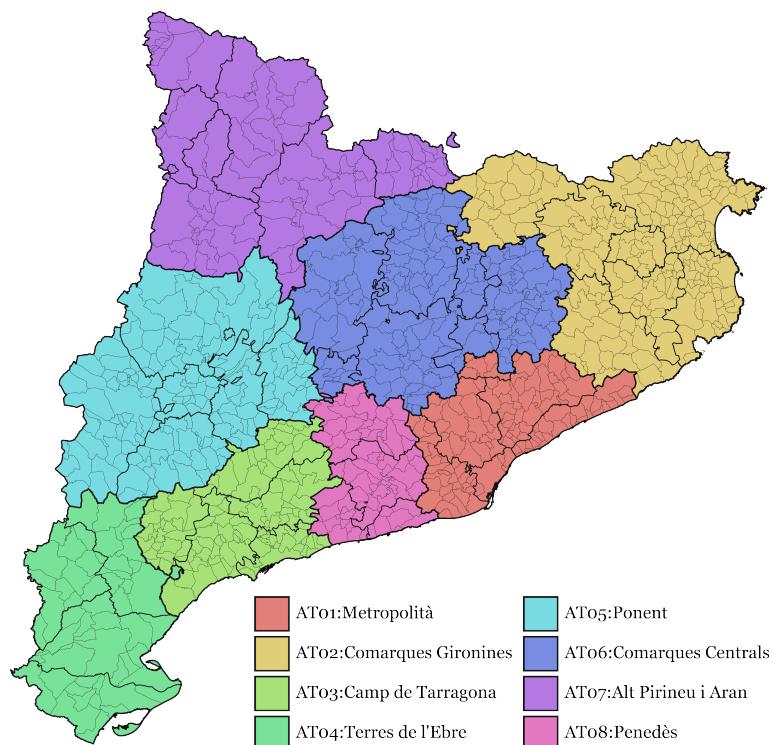


FIGURE A.1: Map of Catalonia split by Regions

Appendix B

Descriptive Statistics

TABLE B.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 B.1: Development of variables over years for Catalonia
on monthly basis

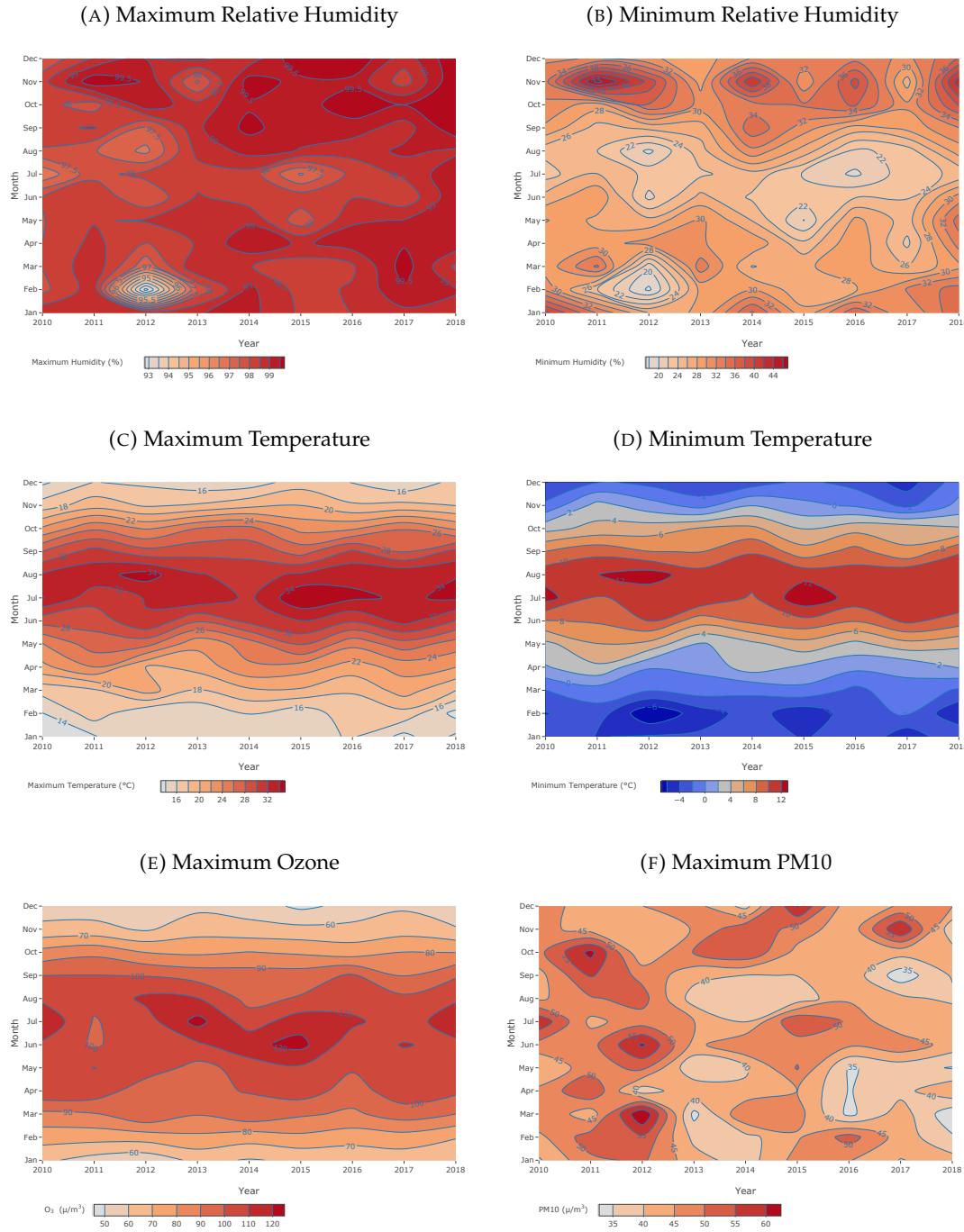


TABLE B.2: Descriptive statistics for daily levels of meteorological variables and air pollutant levels (Lag0) in Catalonia.

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

- 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).
- 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>.
- Gasparrini, Antonio (July 2011). "Distributed Lag Linear and Non-Linear Models in R: The Package dlnm". In: *Journal of statistical software* 43, pp. 1–20. DOI: [10.18637/jss.v043.i08](https://doi.org/10.18637/jss.v043.i08).
- 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).
- 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/>.
- 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.
- 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.