UNIVERSITAT DE BARCELONA

Impact of Environmental Factors on Acute Myocardial Infarction: A Time Series Analysis

(TIME SERIES FINAL PROJECT 2024)

Gabriela Zemencikova
NIUB: 21447473

Faculty of Mathematics & Computer Science
University of Barcelona

May, 2024

Contents

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 occurs due to decreased coronary blood flow, leading to insufficient oxygen supply to the heart and cardiac ischemia (Mechanic et al., 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.

There are two main objectives of this study. Firstly, to predict AMI so there can be prevention strategies in-placed by policy makers. Secondly, 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.

This study will try two different models, including Seasonal Autoregressive Integrated Moving Average (SARIMA) models and Long Short-Term Memory (LSTM) machine learning algorithms. 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. The study primary focuses on Metropolita region.

1.1 Data Sources

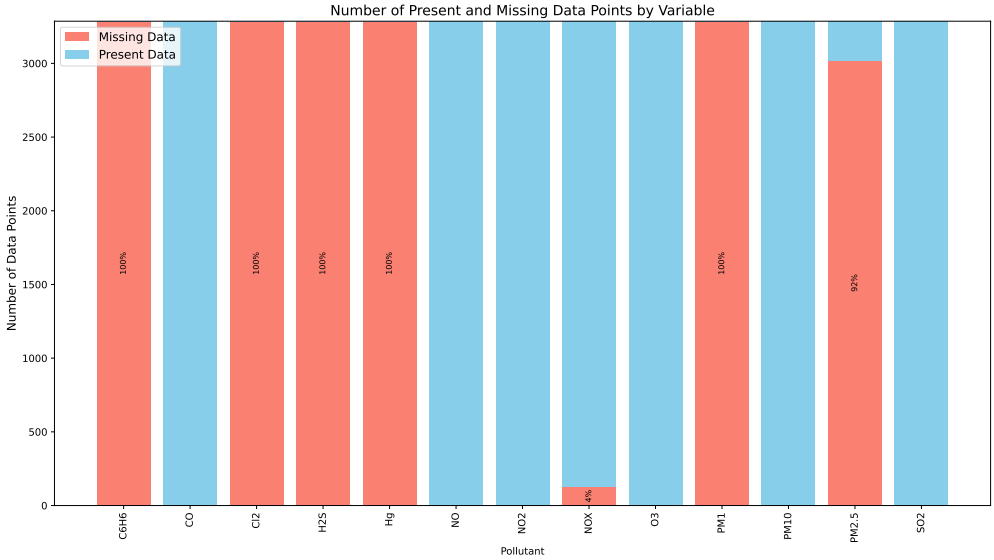
For the purposes of the analysis there have been various data sources used. The environmental dataset under investigation composed of temperature, relative humidity, air pollution data, and census data was obtained from the Statistical Institute of Catalonia. The census data was obtained from the population 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 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, particular 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 1).

Similarly, meteorological observations, including outdoor temperature and relative humidity, were collected from 239 meteorological stations distributed throughout Catalonia, as illustrated in Figure 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

Figure 1: Missing data in pollutants Over time



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.

1.2 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. In this section, the exploratory analysis is shown on aggregated data for the separate high-level regions below is also explained why such a choice.

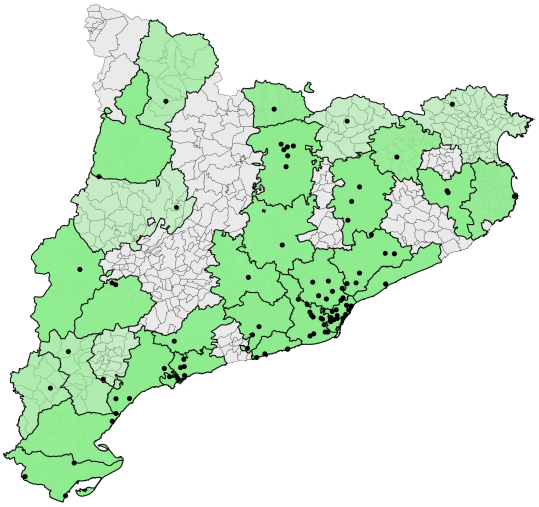
As noted by Mechanic et al. (2024) there are many risk factors potentially contributing to the occurrence of AMI. Considering the set of non-modifiable risk factors that include factors such as sex and age. The further analysis of associations of these variables was conducted by Canto et al. (2012).

1.2.1 Spatial Distribution of AMI Incidence

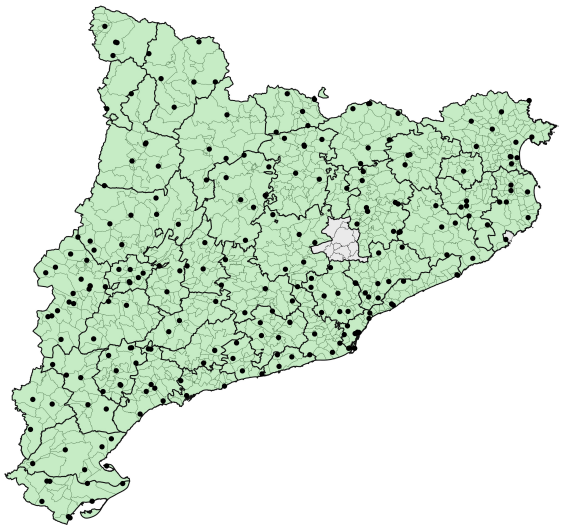
In this subsection, we explore the spatial distribution of AMI incidence across counties and territorial regions in Catalonia. By examining the geographical variability of AMI incidence rates and adjusting for population differences, we aim to identify patterns and trends that may inform targeted public 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. Geospatial analysis techniques, such as mapping and spatial clustering analysis, can help identify hotspots of AMI incidence and elucidate the underlying determinants of geographic disparities in cardiovascular health outcomes.

Figure 2: Spatial Distribution of Monitoring Stations

(a) Spatial Distribution of Air Pollution



(b) Spatial Distribution of Meteorological variables



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. 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.

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. When exploring detailed temporal patterns in AMI 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 1 illustrates the total number of AMI alerts for each AT and year.

The analysis reveals several problems. Firstly, there is an increase in AMI incidence rates across all territorial regions, accompanied by significant spatial variability within each year and later we encounter issue of daily variability in the data. 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. Once computing ASIR, another potential issue

Table 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

that arose was the noise in the data. Having raw daily ASIR measurements produced unwanted noise and variability.

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 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,
- $\theta_1, \theta_2, \ldots, \theta_q$ 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 4. From the figure, we can slightly observe a long-term trend with several peaks within the examined period and also the is an apparent cyclical pattern in the data.

There is a clear seasonal patter 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 3). 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 5.

Figure 3: Decomposition of ASIR

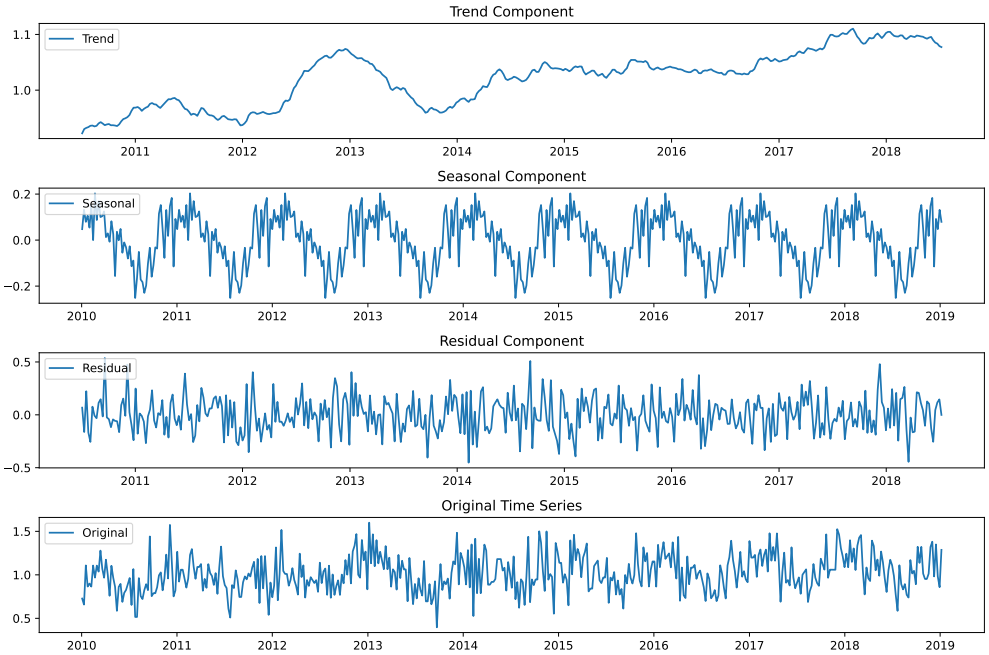
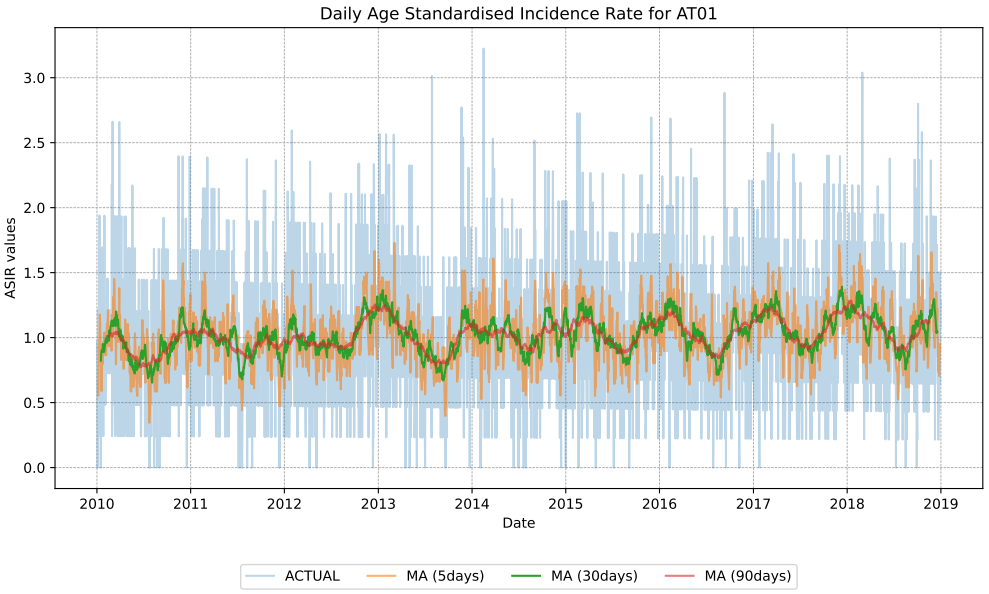


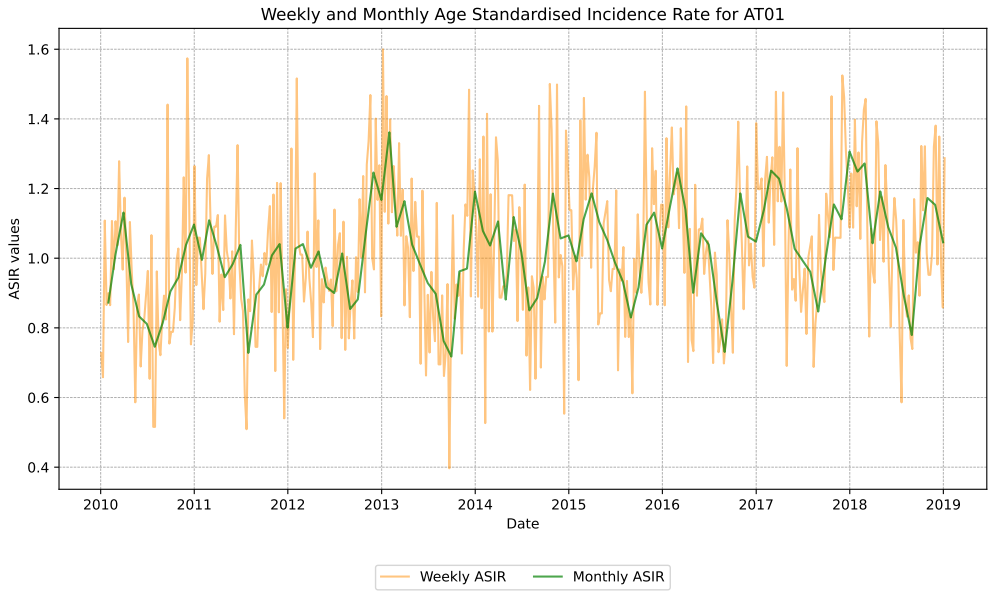
Figure 4: Moving Averages of ASIR Over Time



1.2.2 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 before, 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

Figure 5: Weekly and Monthly ASIR Over time



(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 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.

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

The Table 2 below 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,

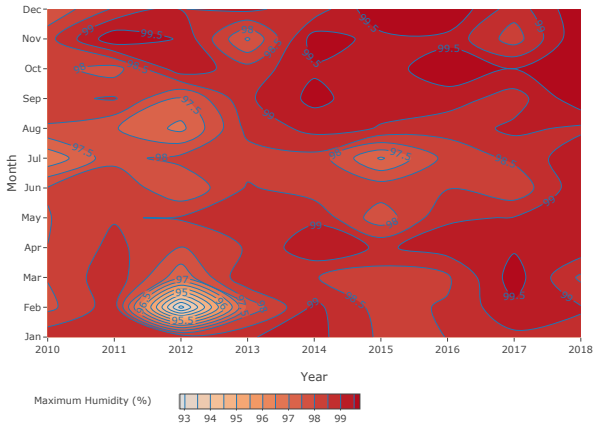
increased humidity levels, and elevated concentrations of air pollutants, particularly ozone and nitrogen dioxide. These findings highlight the importance of considering seasonal variations in weather conditions when analyzing cardiovascular health outcomes and planning public health interventions. For the monthly development of variables see Figure 6.

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

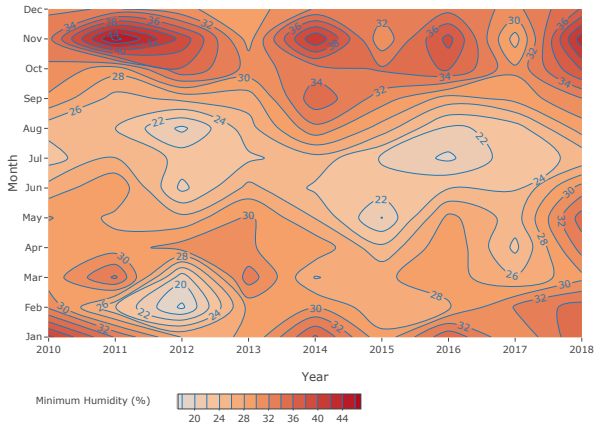
Variable	Mean	SD	Minimum	Q25	Median	Q75	Maximum	IQR
All - 3287 days								
Max Temperature	23.50	7.51	1.70	17.40	23.20	29.70	43.00	12.30
Min Temperature	3.30	7.03	-23.20	-1.80	3.10	8.70	20.70	10.50
Max Humidity	98.63	3.08	62.00	99.00	100.00	100.00	100.00	1.00
Min Humidity	29.23	13.66	0.00	20.00	28.00	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

Figure 6: Development of variables over years for Catalonia on monthly basis

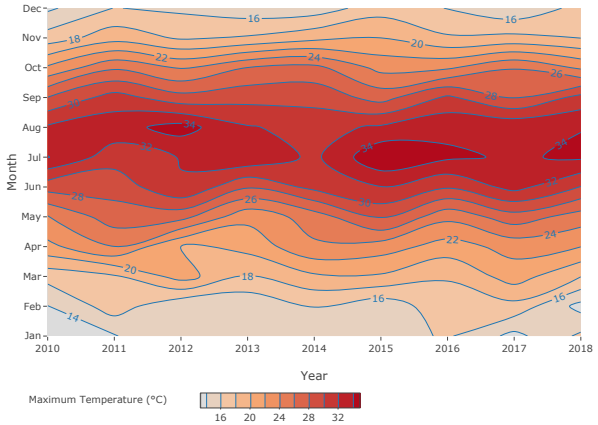
(a) Maximum Relative Humidity



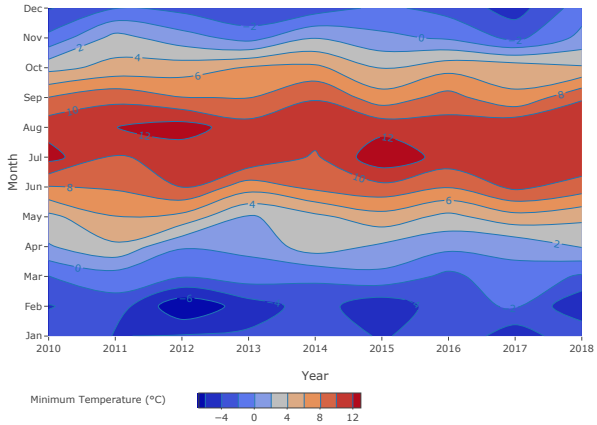
(b) Minimum Relative Humidity



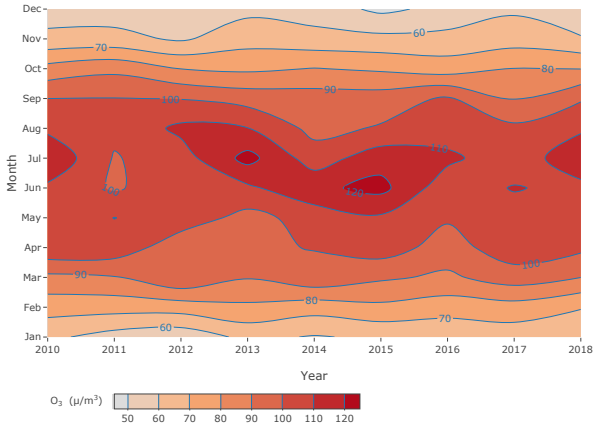
(c) Maximum Temperature



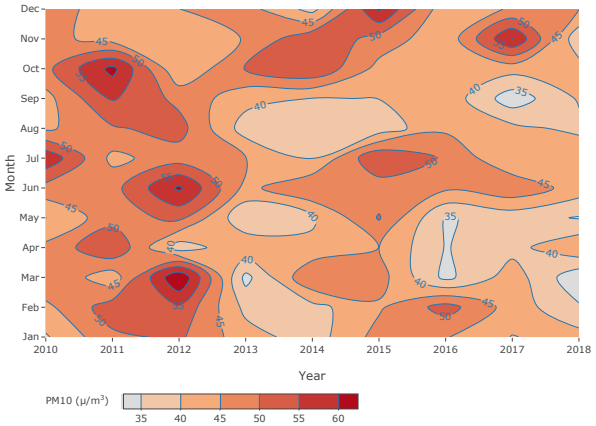
(d) Minimum Temperature



(e) Maximum Ozone



(f) Maximum PM10



2 Time series Analysis

In our analysis, we examined AT code, AT01, based on the distribution of the stations from which the data has been observed as well as number of cases present. This uniformity in station coverage ensures that our observations are representative and reliable across the entire period. The variables that are observed are the following: ASIR, Temperature, Humidity, PM10, O3 and Public Holidays.

Considering the nature of our time series data and the need for accurate forecasting, we will employ Seasonal AutoRegressive Integrated Moving Average (SARIMA) model. This models is well-suited for capturing linear dependencies and seasonal patterns in time series data, making it a suitable candidate for our analysis.

In addition to SARIMA, we will also consider Long Short-Term Memory (LSTM) networks. LSTM networks are a type of recurrent neural network (RNN) capable of capturing long-term dependencies and non-linear patterns in sequential data. Given the potentially complex and non-linear nature of our time series data, LSTM models offer a promising approach for accurate forecasting and modeling of the underlying patterns.

During model evaluation, we considered a range of indicators to assess the effectiveness of each approach:

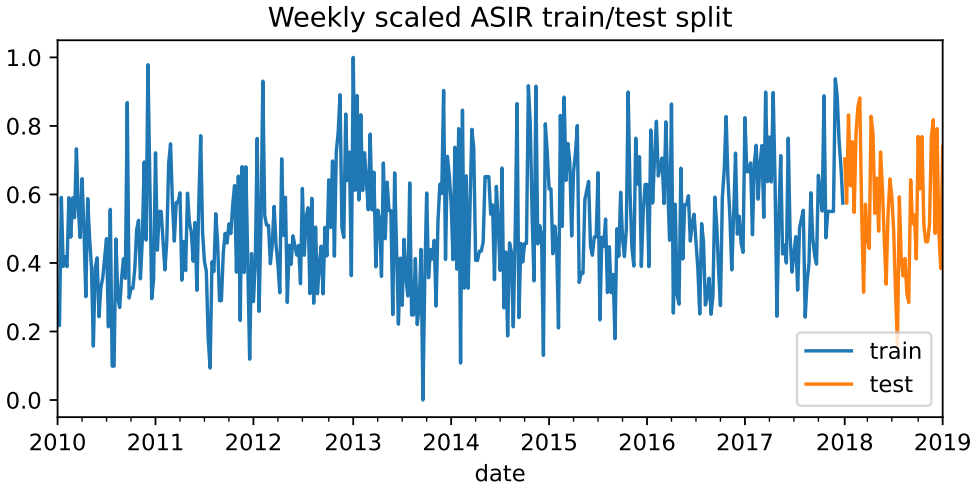
AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion): These metrics gauge the goodness of fit of the model while penalizing for complexity, aiding in the selection of the most appropriate model.

MAE (Mean Absolute Error), MSE (Mean Squared Error), and RMSE (Root Mean Squared Error): These metrics measure the accuracy of the model's predictions, providing a quantitative assessment of its performance.

p-value (Box-Ljung Test): A p-value greater than 0.05 indicates independence of residuals, suggesting a satisfactory fit of the model to the data.

To prepare our data for modeling, we first split it into training and test sets. The training set comprises data from the years 2010 to 2017, while the test set includes data from the year 2018. As a final step, we applied scaling to both sets. Refer to Figure 7 for a visualization of the train/test split.

Figure 7: Train/Test scaled split



We applied scaling to both the training and test sets to ensure that all variables have a similar scale. This helps in improving the performance and convergence of many machine learning algorithms, especially those that are sensitive to the scale of the input features, such as support vector machines (SVM), k-nearest

neighbors (KNN), and neural networks. By scaling the variables, we make the optimization process more efficient and prevent certain features from dominating others simply because of their larger scale. This ensures that the model is better able to learn from the data and generalize well to unseen examples.

2.1 ARIMA/SARIMAX

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. In our dataset, by plotting ACF and PACF we can observe, there is an annual seasonality present, specifically the ASIR increases during the winter months and is more subtle during the hot months. Furthermore, there is an increasing overall trend. Due to the data exhibiting seasonal patterns, ARIMA model may not be the best model, therefore we will use SARIMA, which adds seasonal parameters to ARIMA model. It is frequently used for non-stationary series and the data does not fluctuate around the same mean, variance and co-variance. Note that in our case since we are working with weekly components and observing a annual seasonality, we choose $s = 52$.

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

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

where:

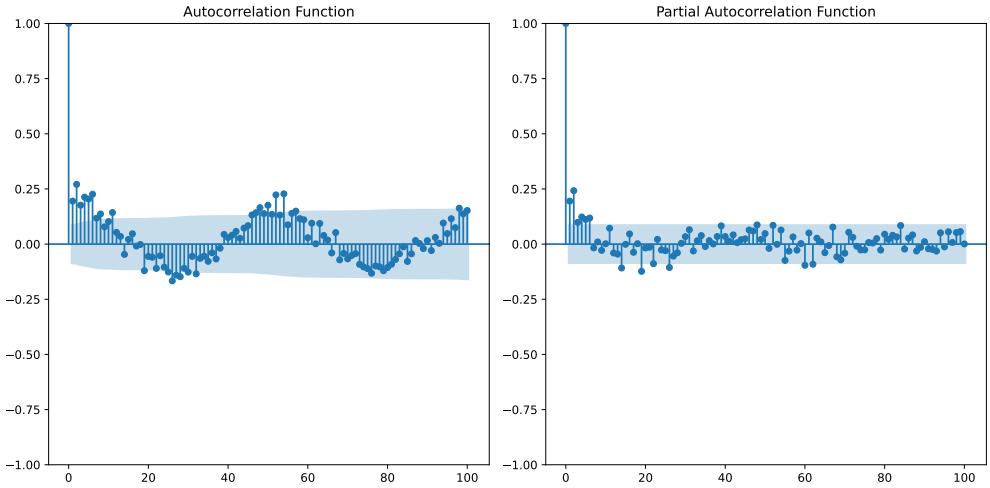
- y_t represents the value of the time series at time t,
- $\phi_1, \phi_2, \ldots, \phi_p$ are the autoregressive (AR) parameters of the model representing the coefficients of past values of the series,
- p is the order of the autoregressive process,
- $\theta_1, \theta_2, \ldots, \theta_q$ are the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

2.2 LSTM

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.

Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) units ware chosen in the project 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.

Figure 8: ACF and PACF plots



2.2.1 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 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 weeks 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.

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.

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.

2.3 Model Comparison and Results

2.3.1 SARIMAX Model Results Summary

The SARIMAX model has successfully converged, indicating a stable optimization process. The optimal model is specified as SARIMAX(0, 1, 1)x(0, 0, 1, 52), indicating the absence of autoregressive terms, a first-order moving average term, and a seasonal moving average term with an annual periodicity of 52 weeks. With the log-likelihood value (172.78) indicating the model fits the data well.

Parameter Estimates:

- 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.
- `min_hum`: There is no significant association between minimum humidity and ASIR of AMI (p=0.521).
- `mean_PM10`: An increase of 1 µg/m³ in mean PM10 concentration is associated with an increase of approximately 0.146 units in ASIR of AMI.
- `mean_O3`: There is no significant association between mean O3 concentration and ASIR of AMI (p=0.484).
- `is_holiday`: There is no significant association between holidays and ASIR of AMI (p=0.896).
- `ma.L1`: The first-order moving average term coefficient indicates a strong negative association between the lagged residual and the ASIR of AMI, suggesting a high dependency on past observations.
- `ma.S.L52`: The seasonal moving average term coefficient suggests a weak positive association with the ASIR of AMI (p=0.080), indicating a potential weekly seasonal pattern in AMI incidence.

Residual Variance (σ^2): The residual variance (σ^2)) is estimated to be 0.0254, representing the unexplained variability in the ASIR of AMI after accounting for the predictor variables.

2.3.2 LSTM Model Results Summary

Keras Tuner 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:

Units: 150

Activation: ReLU

Dropout rate: 0.2

Number of layers: 2

Optimizer: Adam

LSTM layer 0 units: 150

LSTM layer 1 units: 150

The LSTM model built with the optimal hyperparameters is described below:

First LSTM Layer: 150 units, ReLU activation, return sequences set to True

Dropout Layer: 0.2 dropout rate

Second LSTM Layer: 150 units, ReLU activation, return sequences set to True

Dropout Layer: 0.2 dropout rate

Output Layer: Dense layer with 1 unit

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.

2.3.3 Comparison

The performance of the SARIMA and LSTM models was evaluated based on several metrics. Table 3 presents a summary of the results obtained from both models.

Comparing the performance metrics between the SARIMA and LSTM models, SARIMA exhibits a lower Mean Absolute Error (MAE) of 0.12 compared to LSTM's 0.18, indicating that SARIMA forecasts are closer to the actual values on average. Additionally, SARIMA achieves a lower Mean Squared Error (MSE) of 0.02,

suggesting better overall accuracy in predicting the squared differences between forecasted and observed values compared to LSTM's MSE of 0.05. These results suggest that SARIMA outperforms LSTM in terms of both MAE and MSE, making it a more suitable choice for this forecasting task. However, the LSTM model exhibits substantially lower AIC and BIC values compared to SARIMA, indicating a potentially better fit to the data and superior long-term forecasting capabilities. In addition to the tabulated results, Figures 9 and 10 depict the forecasted values of the SARIMA and LSTM models, respectively. These plots provide visual representations of the model predictions.

Figure 9: SARIMA results

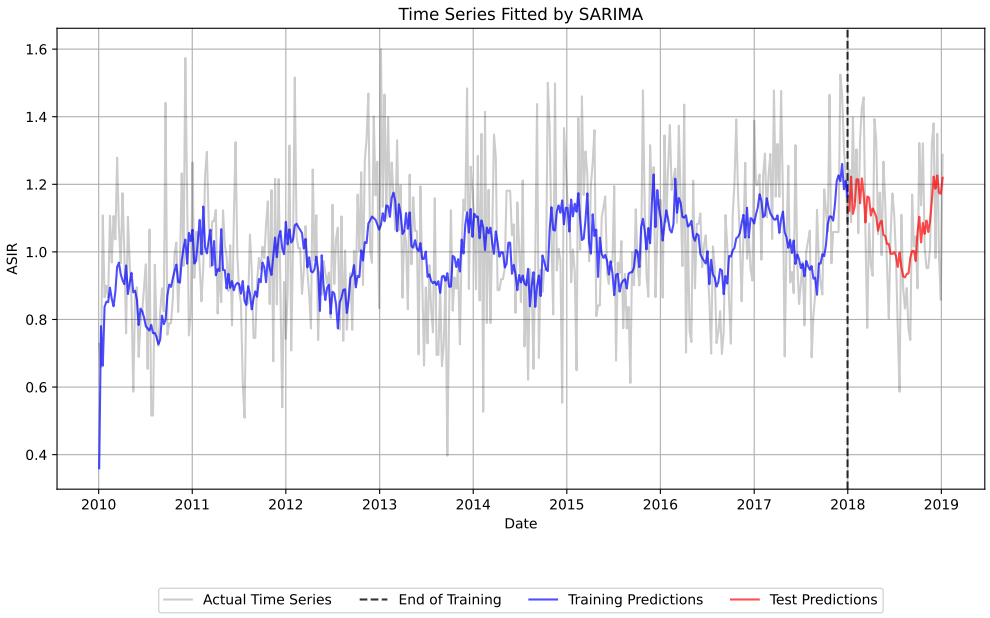
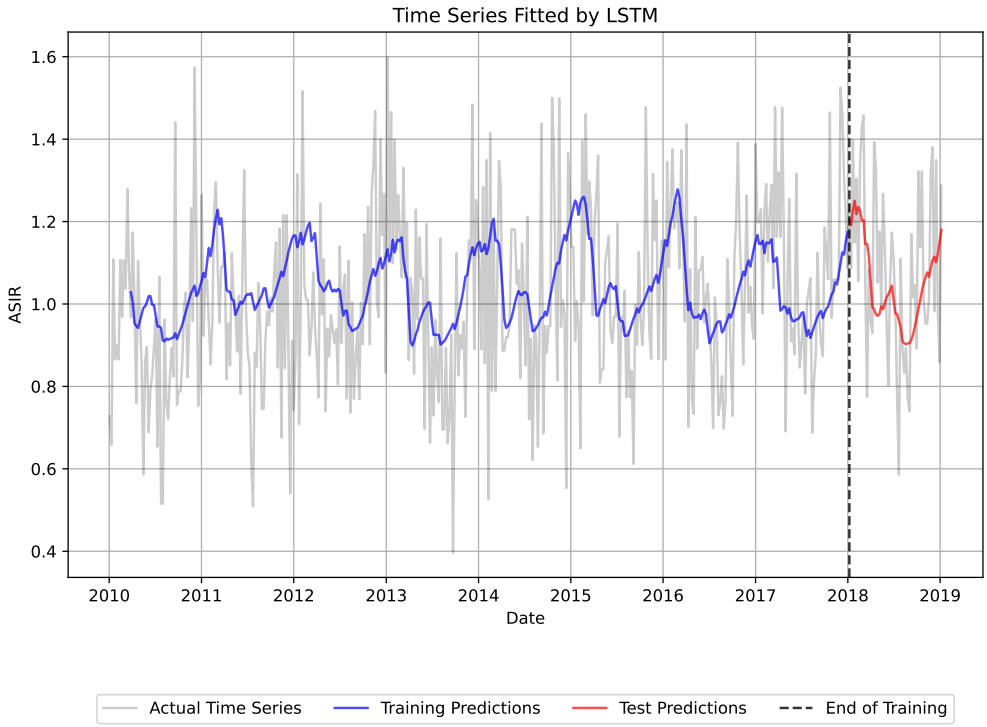


Table 3: Results

	SARIMAX	LSTM
MAE	**0.12**	0.15
MSE	**0.02**	0.03
RMSE	**0.15**	0.19
MAPE	26.60	14.53
AIC	-327.86	909872.17
BIC	-291.56	1806258.44

Figure 10: LSTM results



3 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 $\mu g/m^3$ 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.

References

J. G. Canto, W. J. Rogers, R. J. Goldberg, E. D. Peterson, N. K. Wenger, V. Vaccarino, C. I. Kiefe, P. D. Frederick, G. Sopko, Z.-J. Zheng, and N. Investigators. Association of age and sex with myocardial infarction symptom presentation and in-hospital mortality. *JAMA*, 307(8):813–822, Feb 2012. doi: 10.1001/jama.2012.199.

C. De Sa. Lecture 14: Hyperparameter optimization, 2020. URL https://www.cs.cornell.edu/courses/cs4787/2020sp/lectures/Lecture14.pdf. Lecture notes, CS4787 — Principles of Large-Scale Machine Learning Systems, Cornell University, Delivered 2020.

O. Mechanic, M. Gavin, and S. Grossman. Acute myocardial infarction. *StatPearls*, Jan 2024. [Updated 2023 Sep 3]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK459269/.

M. Peixeiro. *Time Series Forecasting in Python*. Manning Publications, Shelter Island, NY, October 2022. ISBN 9781617299889.

J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng. Hyperparameter optimization for machine learning models based on bayesian optimization. *Journal of Electronic Science and Technology*, 17 (1):pp. 26–40, 2019.