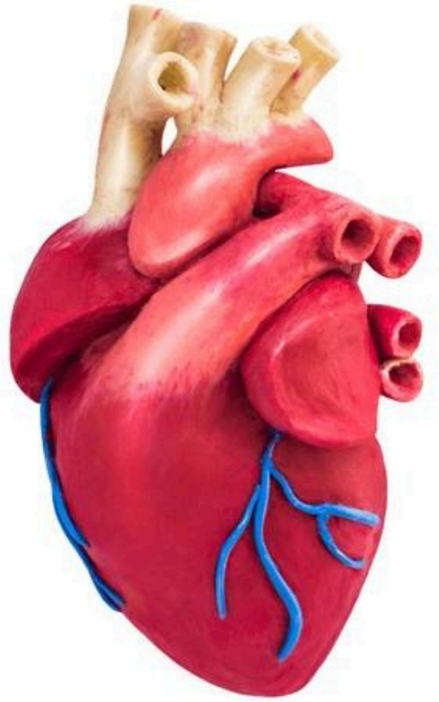
UNIVERSITAT DE BARCELONA

DEVELOPING AN EARLY-WARNING SYSTEM FOR ACUTE MYOCARDIAL INFARCTIONS IN CATALONIA



THESIS DEFENCE

GABRIELA ZEMENČÍKOVÁ

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

July 2024

TABLE OF CONTENT

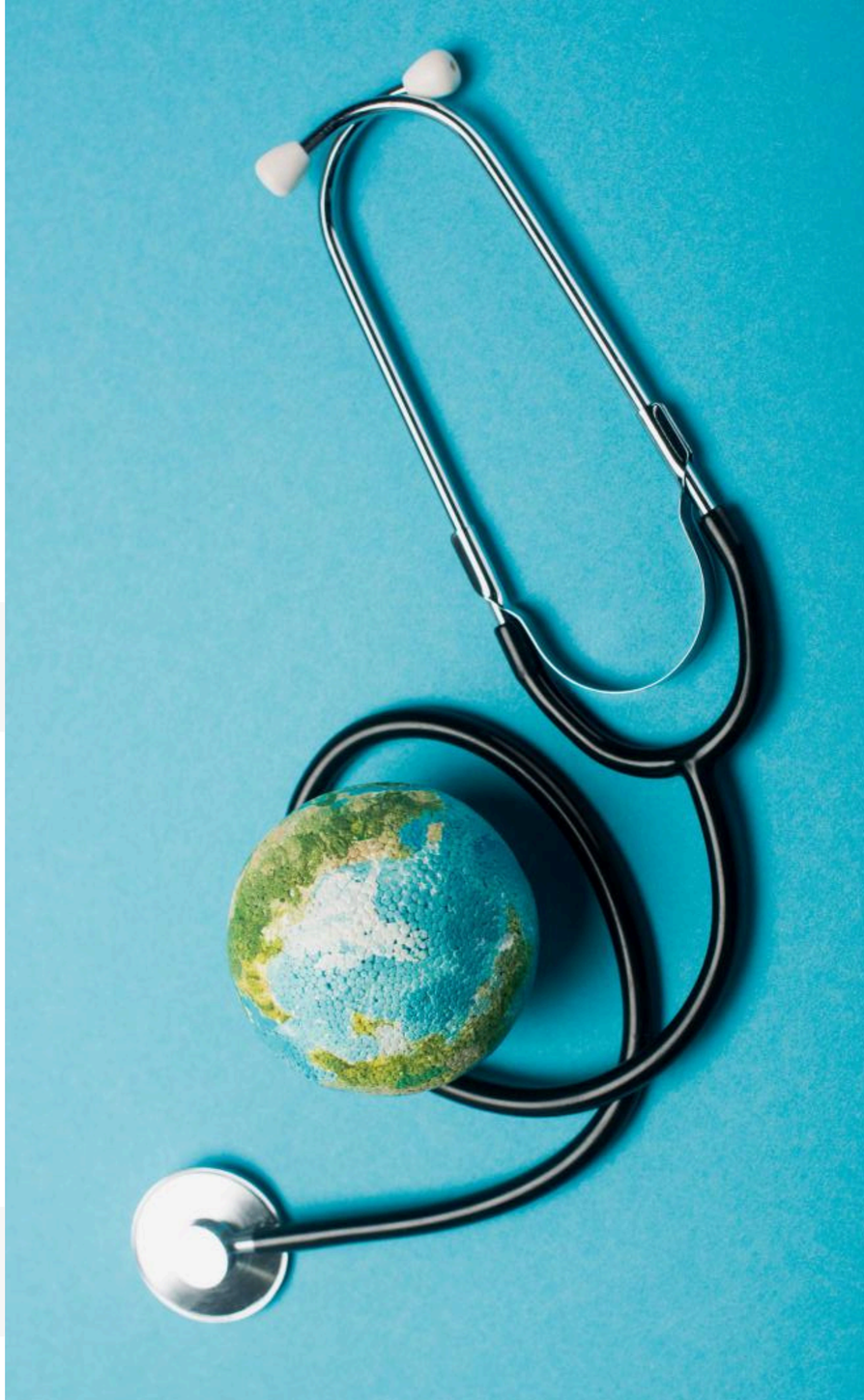
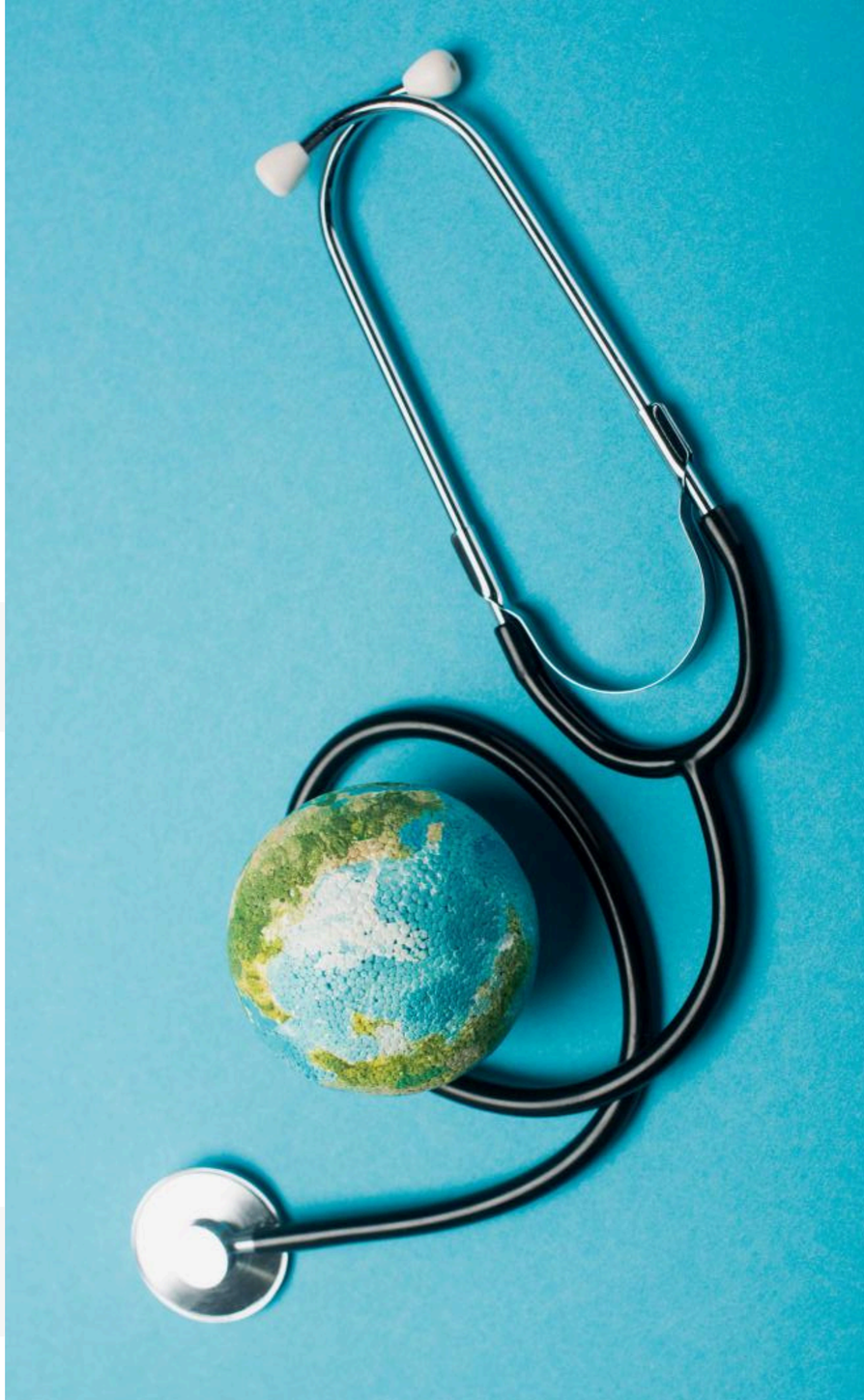


TABLE OF CONTENT

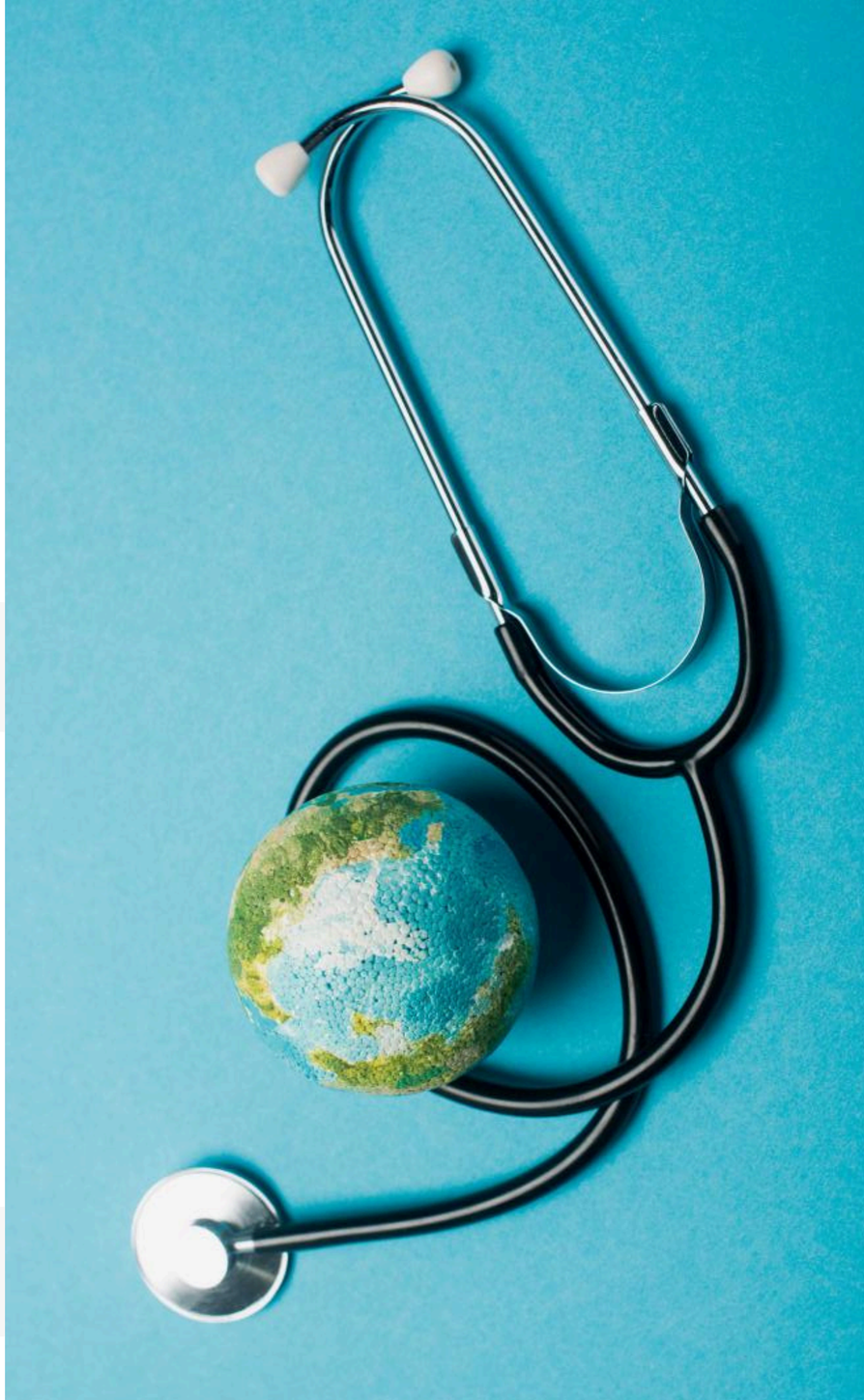


BACKGROUND AND AIMS

"ACUTE MYOCARDIAL INFARCTION IS ONE OF THE LEADING CAUSES OF MORTALITY."

- Explain the association between environmental variables and the incidence of AMI
- Identify disparities or differential susceptibility to environmental variables across different population segments.
- Implementation and comparison of models for series prediction.
  - Seasonal Autoregressive Integrated Moving Average (SARIMAX)
  - Long Short-Term Memory (LSTM)
- Assess the reliability of the predictions with increasing lead-time for effective EWS

TABLE OF CONTENT

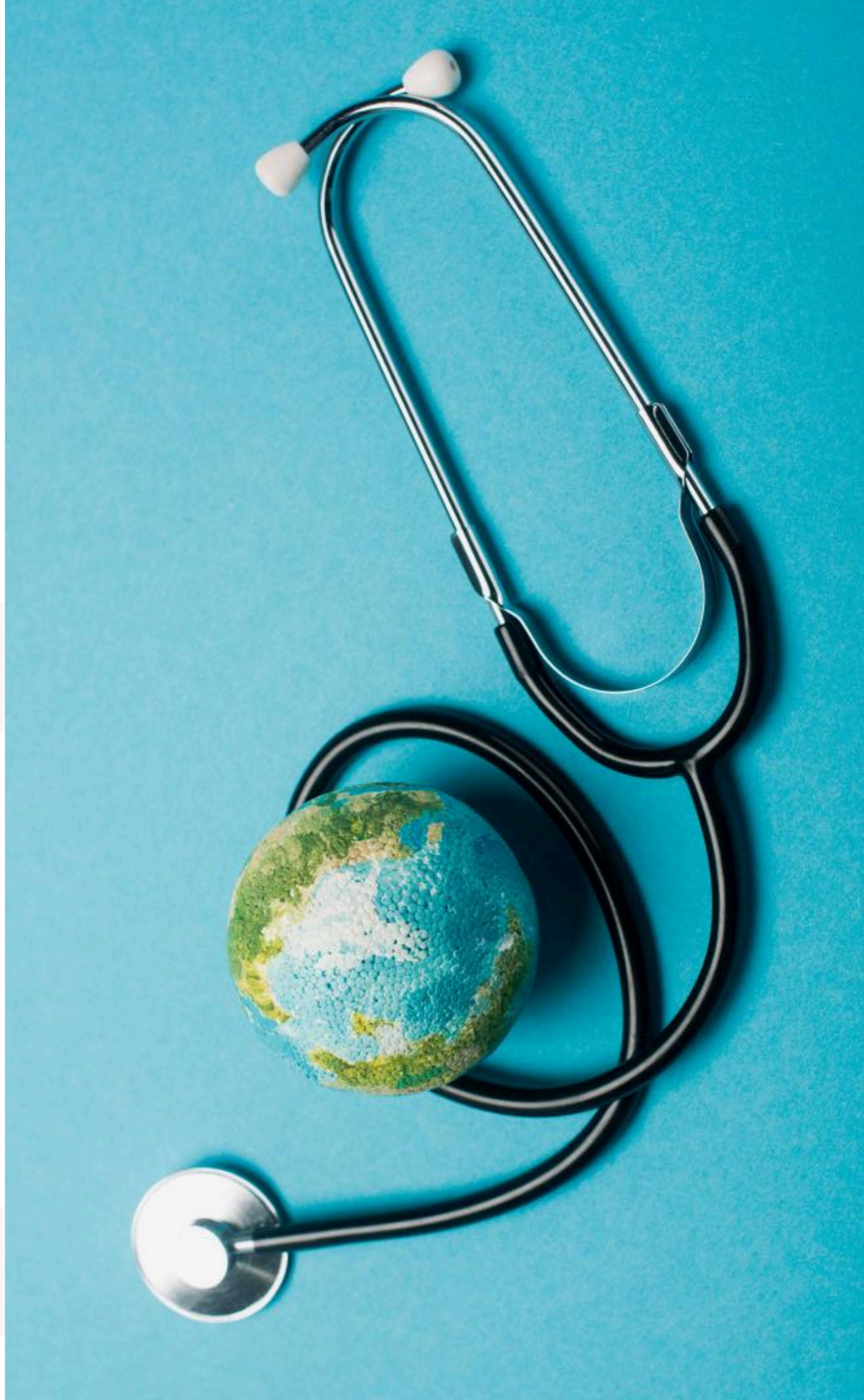


INTRODUCTION

"THE INTERACTION BETWEEN ENVIRONMENTAL FACTORS HAVING AN EFFECT ON THE INCIDENCE OF AMI GAINED AN INCREASED ATTENTION IN RECENT YEARS."

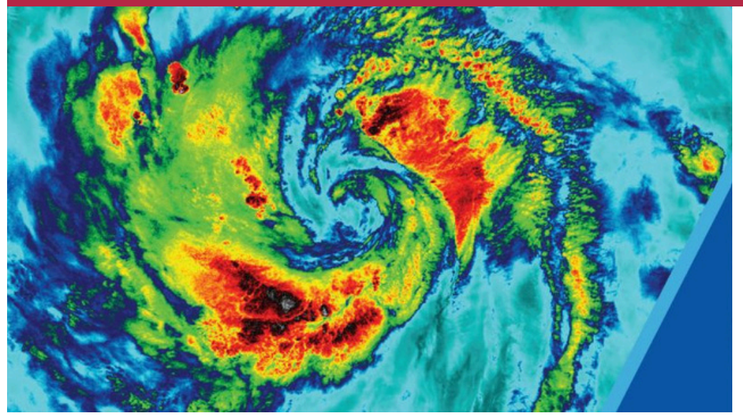
- Myocardial necrosis
- Existing literature indicates that both hot and cold temperatures have an impact on the incidence of AMI
- Catalonia's diverse geographical and demographic landscape
  - temperature fluctuations,
  - humidity levels, and
  - pollution levels

TABLE OF CONTENT



DATASET & DATA PREPARATION

Statistical Institute of Catalonia



METEOROLOGICAL

- **239 meteorological stations**
- 30min readings
- Relative Humidity
- Temperature

AIRQUALITY

- **90 monitoring stations**
- Hourly readings
- Carbon monoxide (CO),
- Nitrogen dioxide (NO2),
- Ozone (O3),
- Particulate matter (PM),
- Sulfur dioxide (SO2).

Hospitals

HEALTH

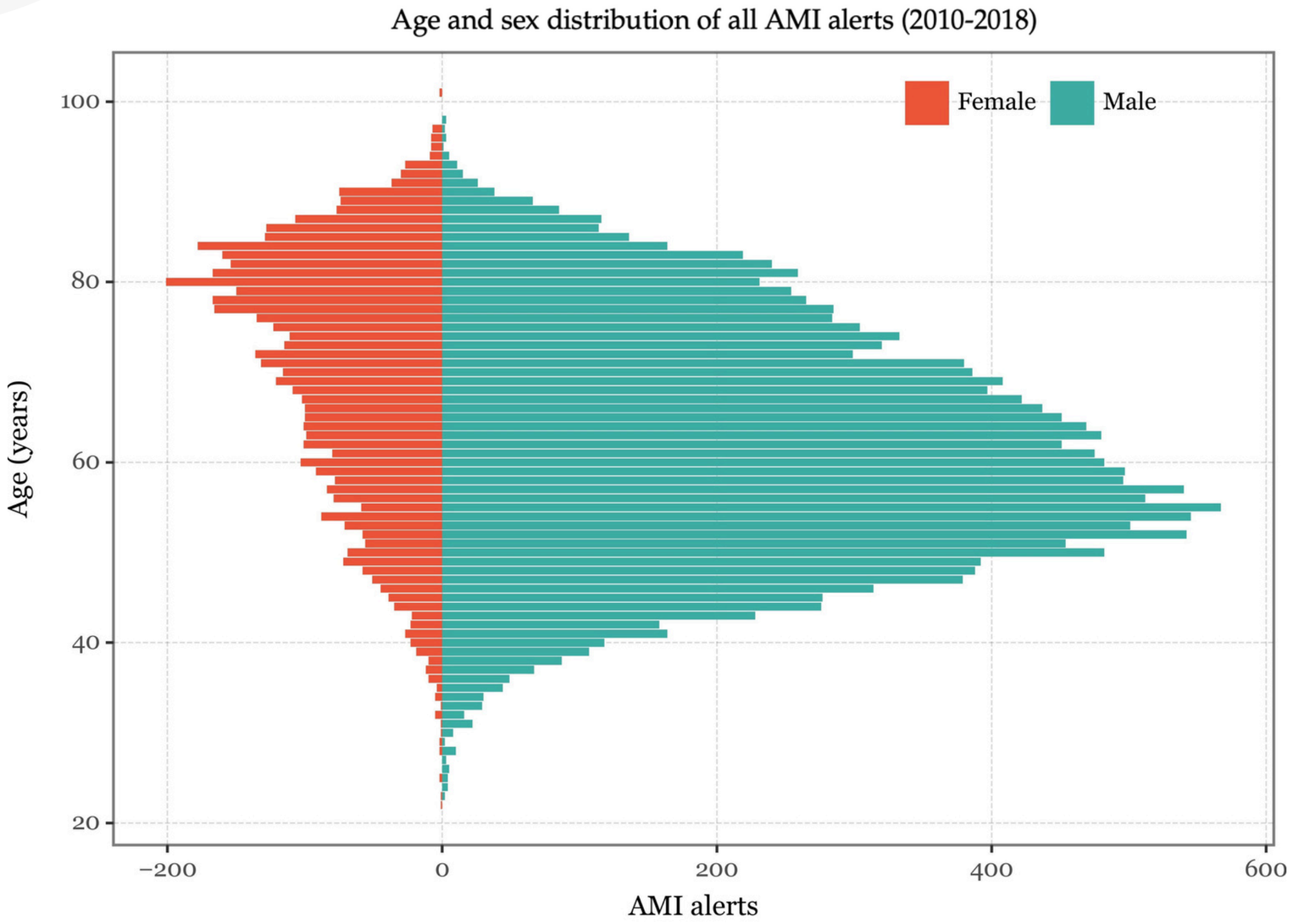
- **10 hospitals**
- Daily counts
- 22,812 admissions
- 948 municipalities
- stratified - province,
  sex
  age

THE CHOICE OF ASIR

> *"Researchers often face **the challenge of comparing** incidence rates **across populations with different age distributions**."*

- Ensures comparability of events **across populations**.
- Ensures comparability **across different age distributions**.
- Adjusts for age as a confounding factor, providing **a more accurate representation** of AMI incidence.
- **Allows comparisons between regions or over time periods.**

THE CHOICE OF ASIR



THE CHOICE OF SPATIAL LEVEL

Higher regions offer more stable geographical units for examination than counties level.

High sparsity even on higher region level on daily basis

Yearly 35 cases per year per 100k inhabitants, the signal-to-noise ratio in the least populated areas becomes skewed, rendering analysis at the daily scale impractical.

> *"THE GEOGRAPHICAL DISTRIBUTION OF AMI INCIDENCE WITHIN CATALONIA IS INFLUENCED BY A MYRIAD OF FACTORS, INCLUDING URBANISATION, SOCIOECONOMIC STATUS, HEALTHCARE INFRASTRUCTURE, AND* ***ENVIRONMENTAL EXPOSURES****."*

THE CHOICE OF TARGET

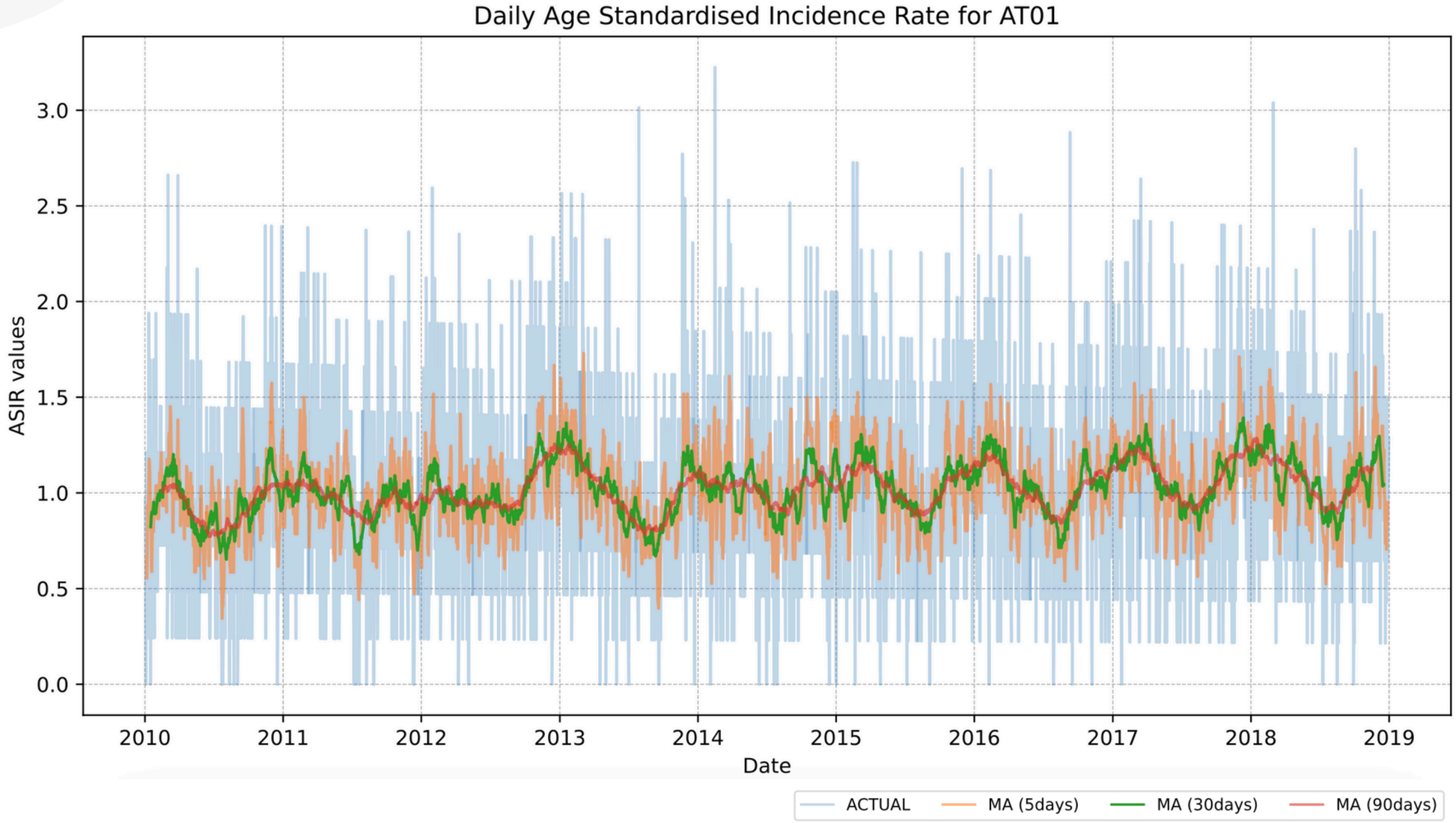
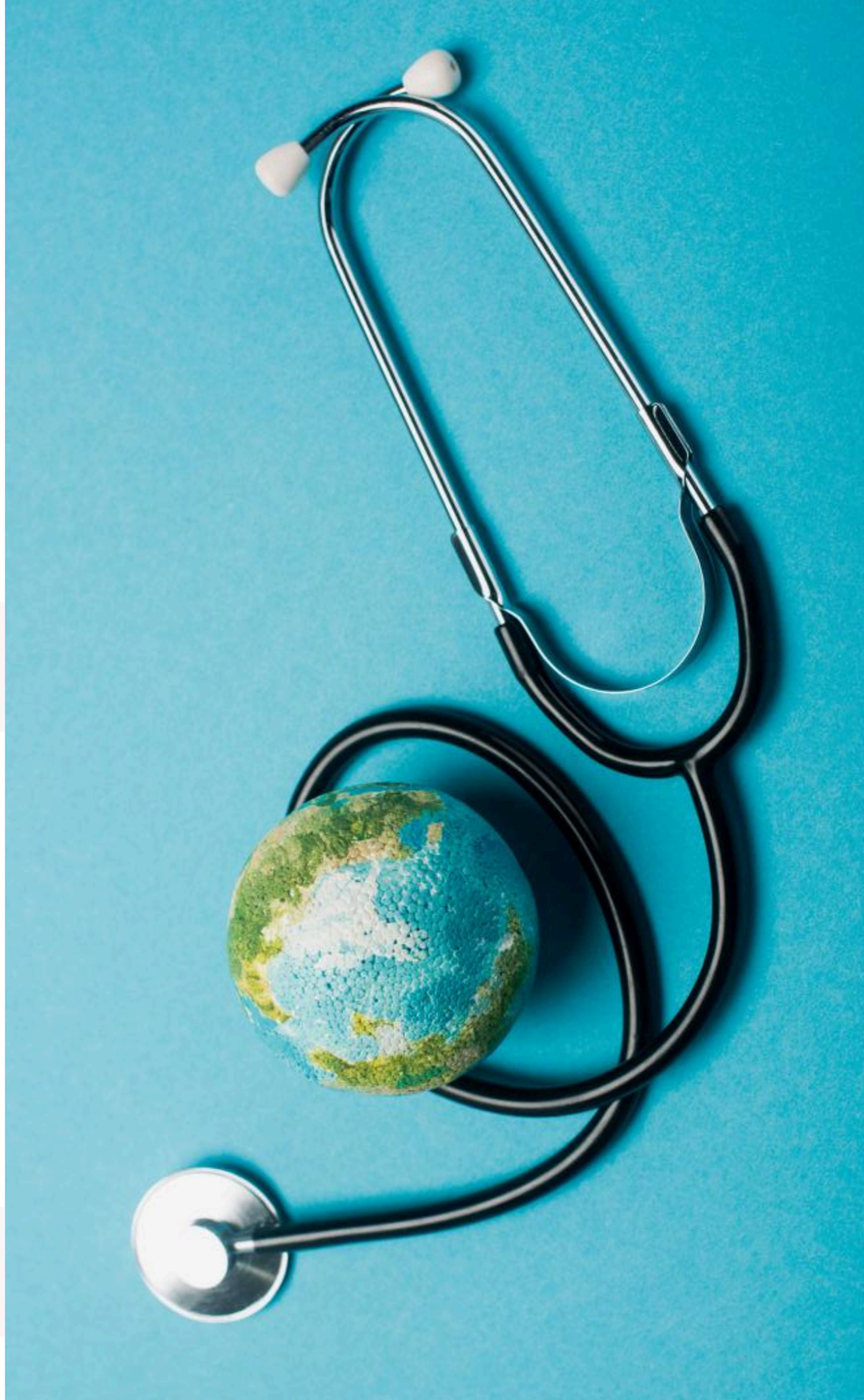


TABLE OF CONTENT



METHODOLOGY

AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

- Generalisation of ARMA
- Used for non-stationary data
- Dependent on past values and past errors
- Instead of using the original series uses the differenced series

$$y'_t = c + \sum_{i=1}^{p} \phi_i y'_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

where $y'_t = (1 - B)^d y_t$.

AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

AUTOREGRESSIVE COMPONENT

- Captures the relationship between **the current observation** and a **number of lagged observations**

$$y'_t = c + \sum_{i=1}^{p} \phi_i y'_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

where $y'_t = (1 - B)^d y_t$.

AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

ERROR COMPONENT

- **Unexplained variation** (random noise)

$$y'_t = c + \sum_{i=1}^{p} \phi_i y'_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

where $y'_t = (1 - B)^d y_t$.

AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

MOVING AVERAGE COMPONENT

- Captures the relationship between **the current observation** and **the residual errors** from a MA of lagged observations

$$y'_t = c + \sum_{i=1}^{p} \phi_i y'_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

where $y'_t = (1 - B)^d y_t$.

AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

DIFFERENCING COMPONENT

- Differencing the series certain number of times to **make it stationary**

$$y'_t = c + \sum_{i=1}^{p} \phi_i y'_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

where $y'_t = (1 - B)^d y_t.$

SEASONAL AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

$$y'_t = c + \sum_{i=1}^{p} \phi_i y'_{t-i} + \sum_{i=1}^{P} \Phi_i y'_{t-is} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \sum_{i=1}^{Q} \Theta_i \epsilon_{t-is}$$

where $y'_t = (1-B)^d (1-B^s)^D y_t$.

SEASONAL AUTOREGRESSIVE COMPONENT

- Captures the relationship between **the current observation** and a **number of lagged observations seasonally**

SEASONAL AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

$$y'_t = c + \sum_{i=1}^{p} \phi_i y'_{t-i} + \sum_{i=1}^{P} \Phi_i y'_{t-is} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \sum_{i=1}^{Q} \Theta_i \epsilon_{t-is}$$

where $y'_t = (1-B)^d (1-B^s)^D y_t$.

SEASONAL MOVING AVERAGE COMPONENT

- Captures the relationship between **the current observation** and **the residual errors** from a MA of lagged observations **seasonally**

SEASONAL AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

$$y'_t = c + \sum_{i=1}^{p} \phi_i y'_{t-i} + \sum_{i=1}^{P} \Phi_i y'_{t-is} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \sum_{i=1}^{Q} \Theta_i \epsilon_{t-is}$$

where $y'_t = (1-B)^d (1-B^s)^D y_t$.

SEASONAL DIFFERENCING TERM

- Seasonal differencing the series certain number of times to **make it stationary at seasonal intervals**

SEASONAL AUTOREGRESSIVE INTEGRATED MOVING AVERAGE WITH EXOG. VARIABLES

$$y'_t = c + \sum_{i=1}^{p} \phi_i y'_{t-i} + \sum_{i=1}^{P} \Phi_i y'_{t-is} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \sum_{i=1}^{Q} \Theta_i \epsilon_{t-is} + \sum_{i=0}^{k} \beta_i X_{t,i}$$

where $y'_t = (1-B)^d (1-B^s)^D y_t$ and $X_{t,i}$ are the exogenous variables.

EXOGENOUS COMPONENT

- Covariates are **external variables** that can influence the time series.

METHODOLOGY

LONG SHORT-TERM MEMORY

- A type of RNN architecture specifically designed to model sequence data while **addressing the vanishing gradient problem.**
- The LSTM cell has a memory cell and **three gates**: input gate, forget gate, and output gate
- Capable of learning **long-term dependencies**
- Capture **complex patterns** and relationships

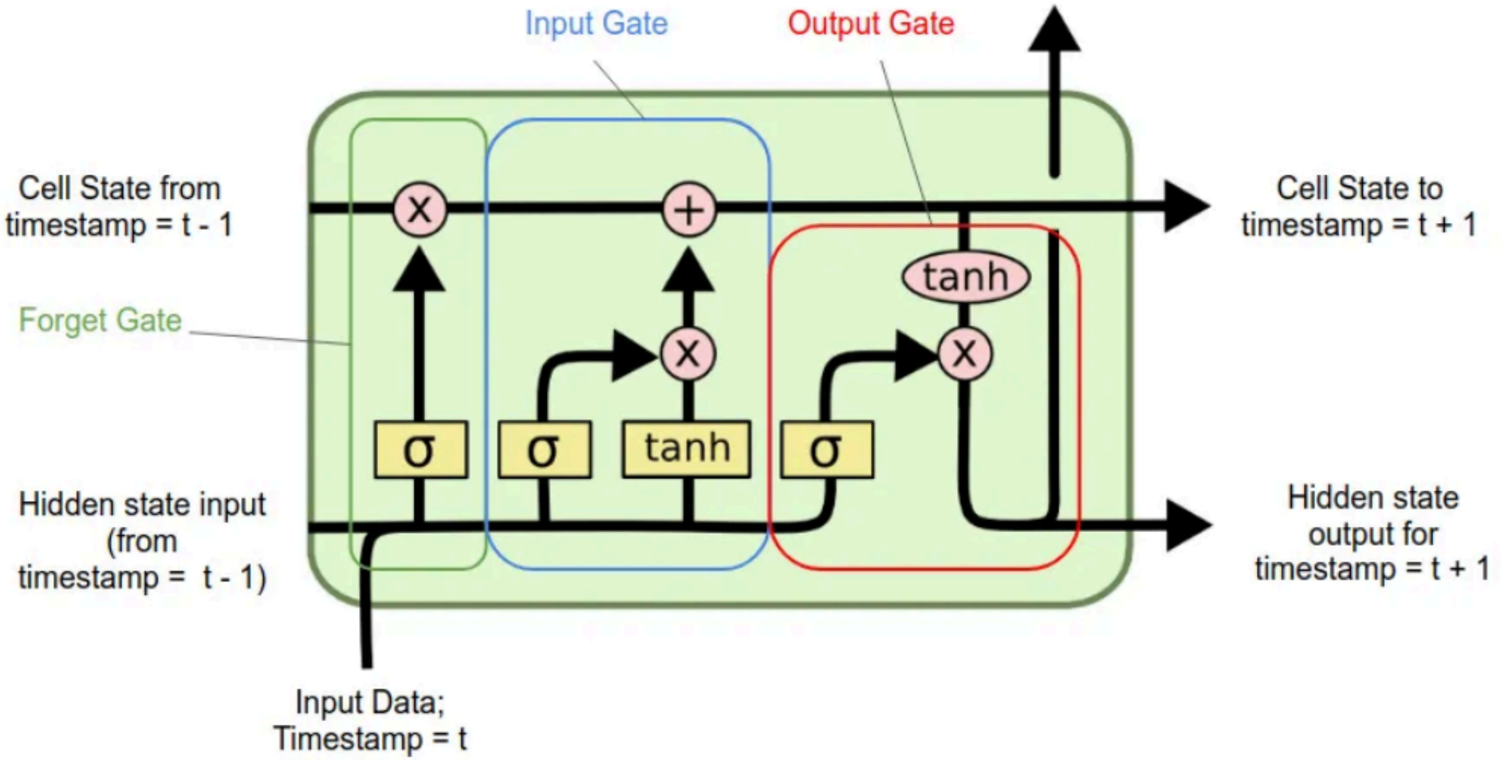
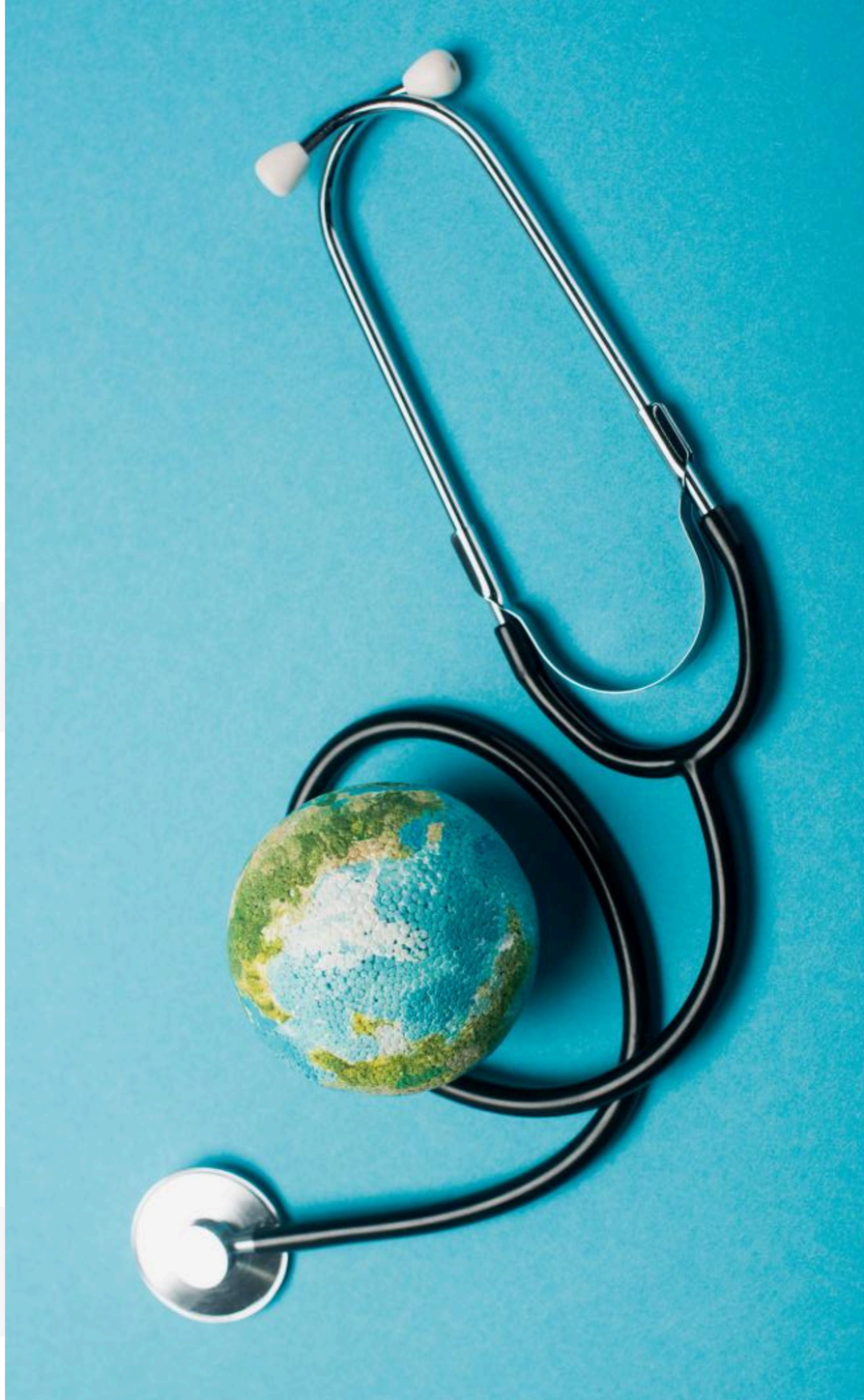


TABLE OF CONTENT



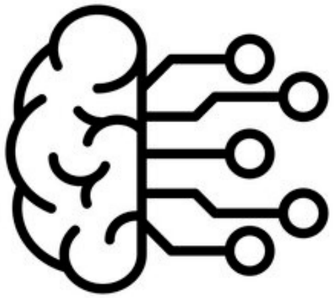
RESULTS & LIMITATIONS

THE OPTIMAL MODEL

SARIMAX(1, 1, 1) x (2, 0, 0, 52)

EXOGENOUS VARIABLES

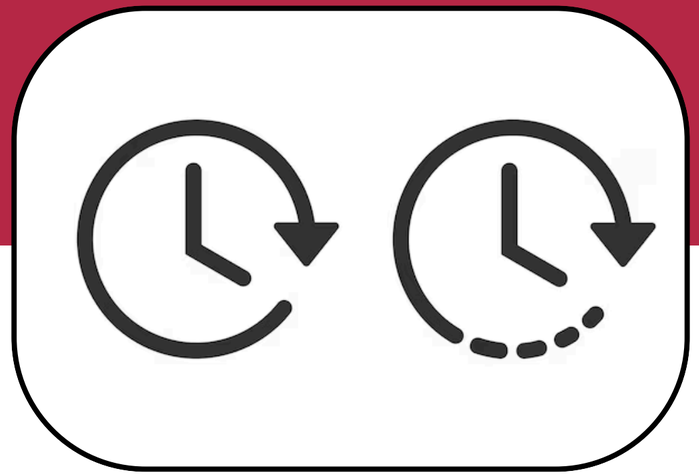
Model comparison with SARIMA



LSTM

Hyperparameter tunning
& performance

RESULTS & LIMITATIONS



LEAD-TIME

Rolling forecast predictions

FEATURE IMPORTANCE

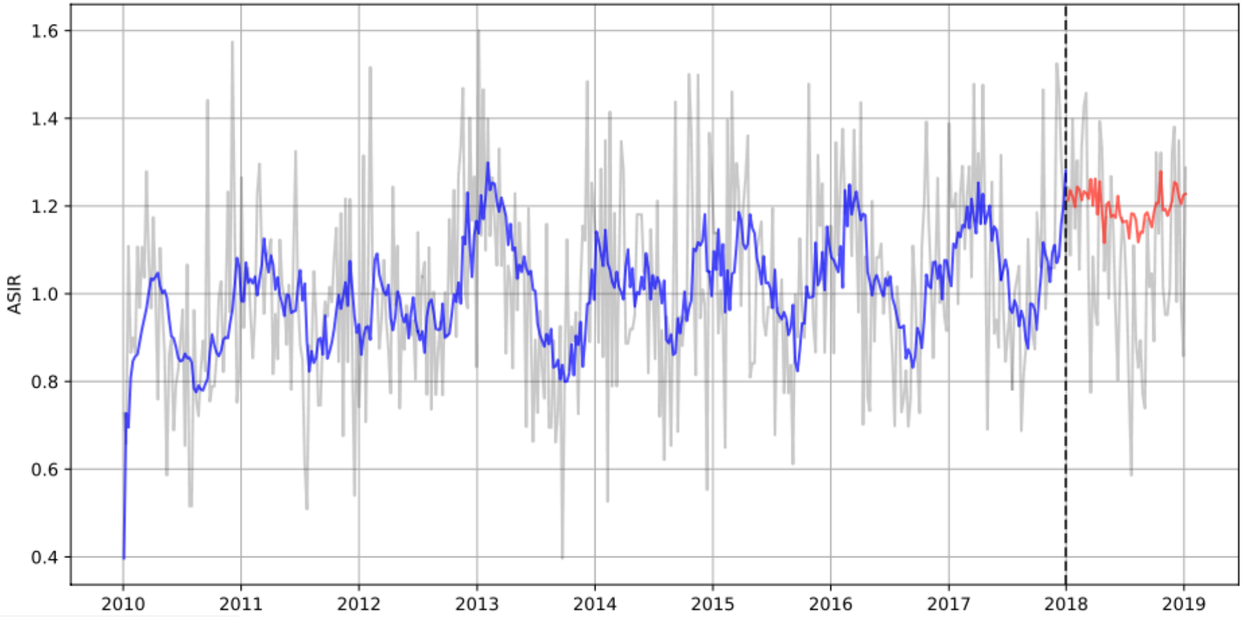
Permutation Feature Importance

LIMITATIONS

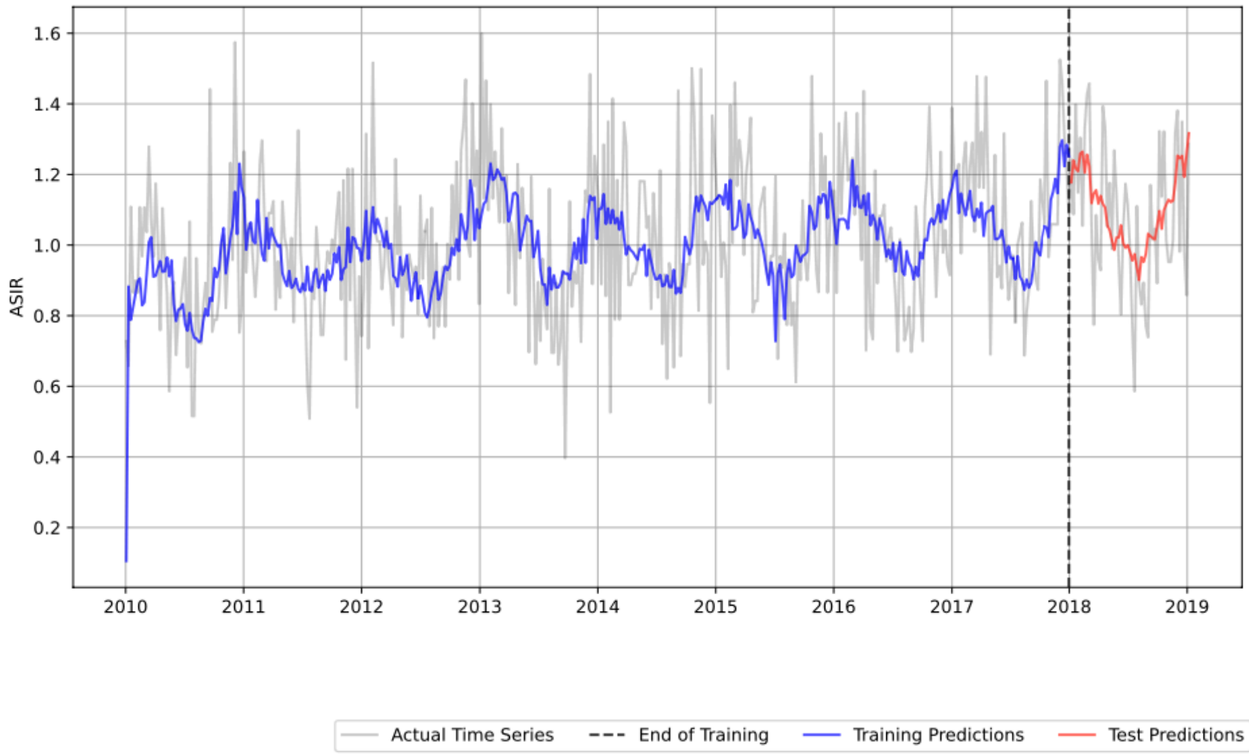
Potential data/model problems

EXOGENOUS VARIABLES

(A) Time Series Fitted by SARIMA



(B) Time Series Fitted by SARIMAX



LONG SHORT-TERM MEMORY

Similar performance as SARIMAX

Hyperparameter tunning: 30 trials, each executed three times for robust evaluation.

Early stopping with patience 10 epochs

Validation split 10% and validation loss with a minimum delta of 0.01

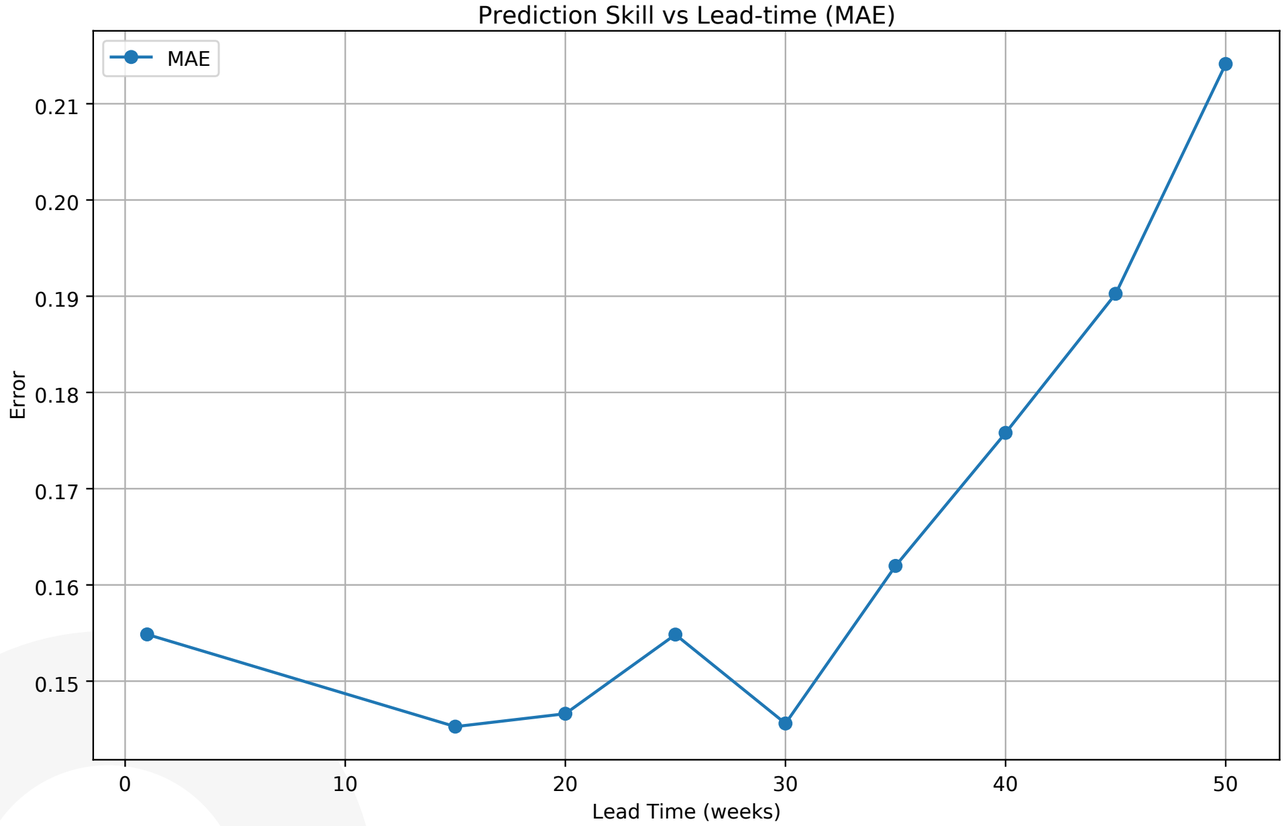
Batch size 32

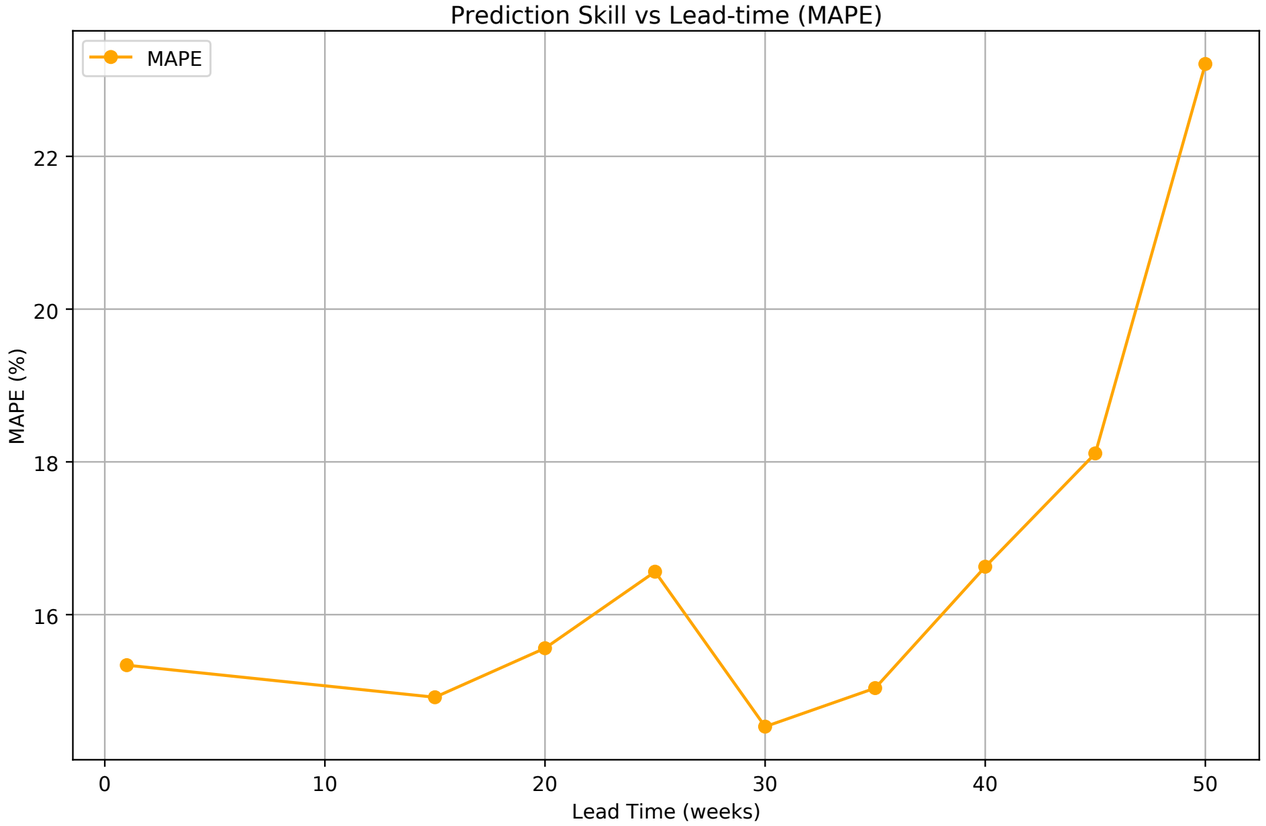
PARAMETERS

- **2 layers,**
- **dropout rate 0.2,**
- **Adam,**
- **RELU activation**

PREDICTION SKILL VS. LEAD-TIME

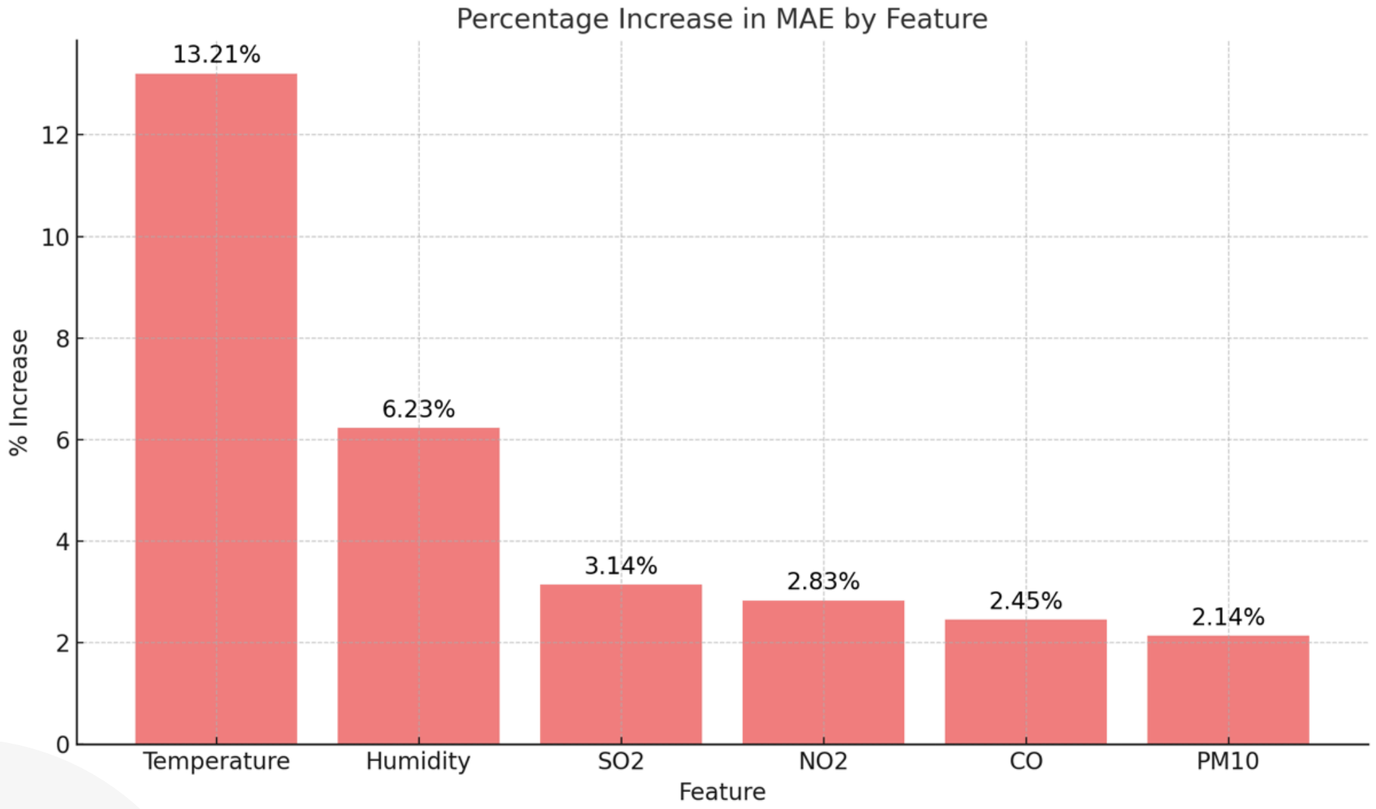
> *"TO MAINTAIN PREDICTION ACCURACY OVER A MORE EXTENDED FORECAST, IT IS ESSENTIAL TO UPDATE THE MODEL AND **ADD NEW DATA REGULARLY TO MAINTAIN THE ACCURACY** AND EFFECTIVENESS OF THE EWS."*





FEATURE IMPORTANCE

> *"THERE ARE VARIOUS METHODS THAT EXTRACT AND ANALYSE THE IMPORTANCE SUCH AS PERMUTATION FEATURE IMPORTANCE, FEATURE IMPORTANCE FROM* ***MODEL COEFFICIENT, SHAPLEY ADDITIVE EXPLANATIONS (SHAP) VALUES OR LOCAL INTERPRETABLE MODEL- AGNOSTIC EXPLANATIONS (LIME).*** *"*



- The **permutation feature importance** measures the importance of the feature by comparing the error of the model with and without permuting the feature's values.
- To ensure robustness in our feature importance analysis, we employed **K-fold cross-validation**
- Other techniques that could be used are **adding noise, integrating gradients or ablation**

LIMITATIONS

01 Models require accurate values of exogenous variables for prediction

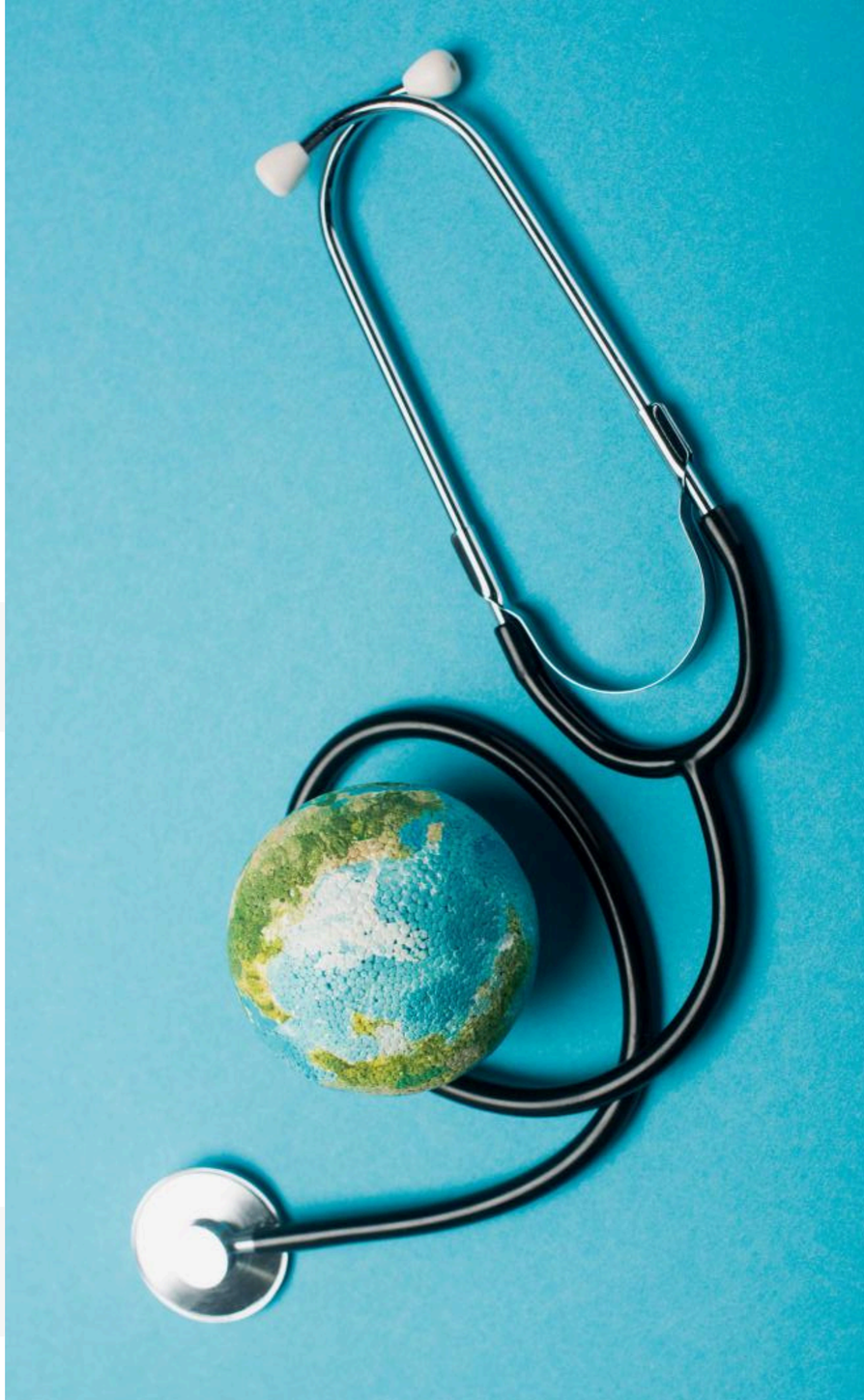
02 Interpretability of LSTM models and requirement of large amount of data for tunning

03 SARIMAX captures only linear relationship between the predictor and response variables

04 Use of linear interpolation can cause minor disruptions

05 Precision of the environment data

TABLE OF CONTENT



CONCLUSION

> *AS **CLIMATE CHANGE CONTINUES TO ALTER WEATHER PATTERNS,** THE INCIDENCE OF AMI MAY BE INFLUENCED BY NEW AND CHANGING ENVIRONMENTAL FACTORS, MAKING TIMELY AND ACCURATE PREDICTIONS EVEN MORE CRITICAL.*

01 DIFFERENT SUSPECTABLE GROUPS

02 IMPORTANCE OF ENVIRONMENTAL FACTORS

- **18% improvement** of forecast accuracy
- By identifying main environmental predictors of AMI, health authorities can implement interventions that would be targeted **on high-risk areas and time**
- **Temperature** being the most crucial variable

03 LEAD TIME UP TO HALF YEAR AHEAD

UNIVERSITAT DE BARCELONA

THANK YOU

FOR YOUR ATTENTION

GABRIELA ZEMENČÍKOVÁ

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

JULY 2024