Multivariate Time Series:
Challenges, missing data, and forecasting

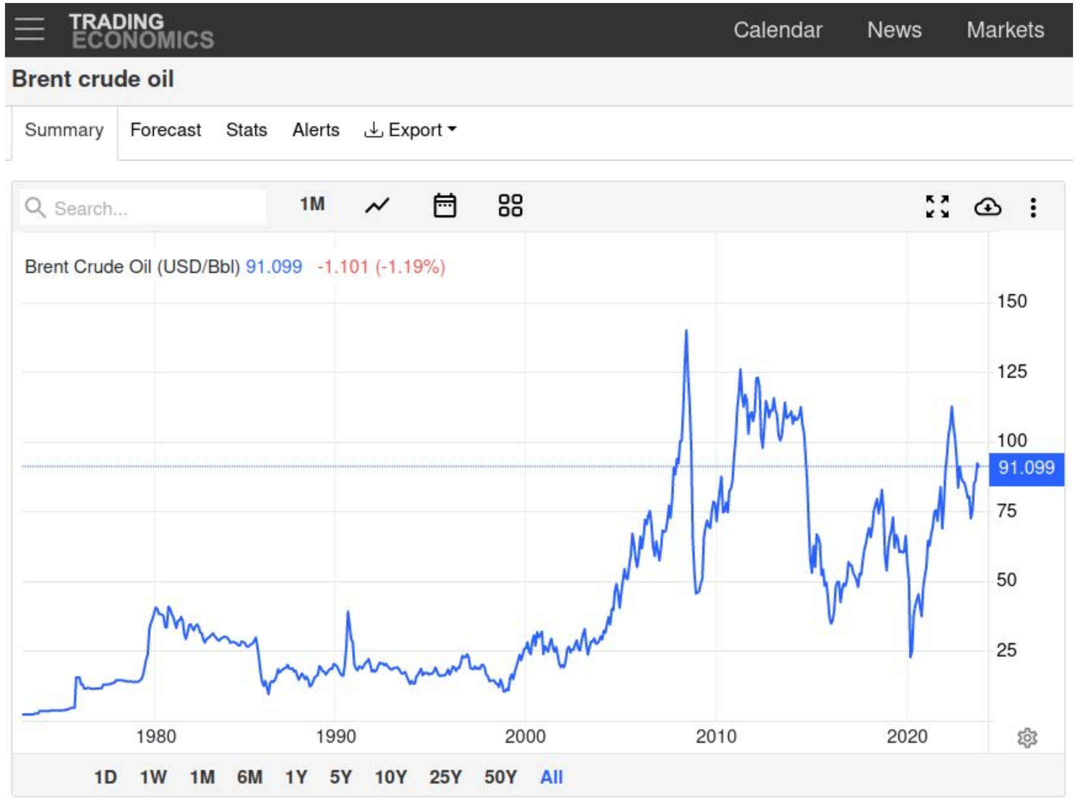
Ana Filipa Almeida
anaa@ua.pt

Introduction

Brief overview of time series data

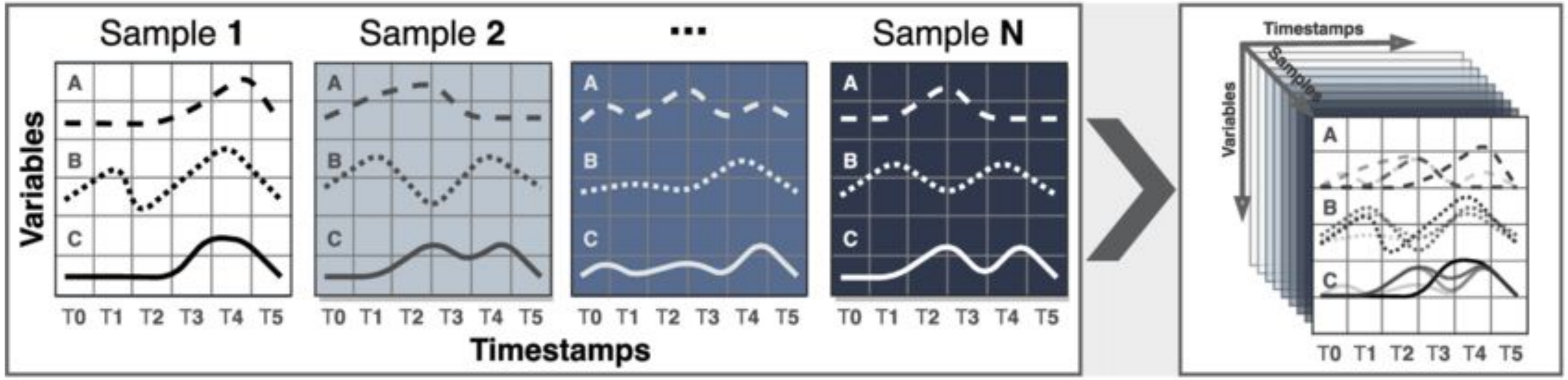
- Time series data:
  - describes how something changes over time,
  - is a sequential set of data points,
  - can present temporal patterns.

- Understanding time series data enables us to make predictions, identify patterns, and make informed decisions.



What is a multivariate time series?

- Multivariate time series data:
  - Evolution of several variables over time
  - Different variables might influence each other
  - We might find more complex patterns, such as spatio-temporal patterns



Challenges in analyzing time series data (1)

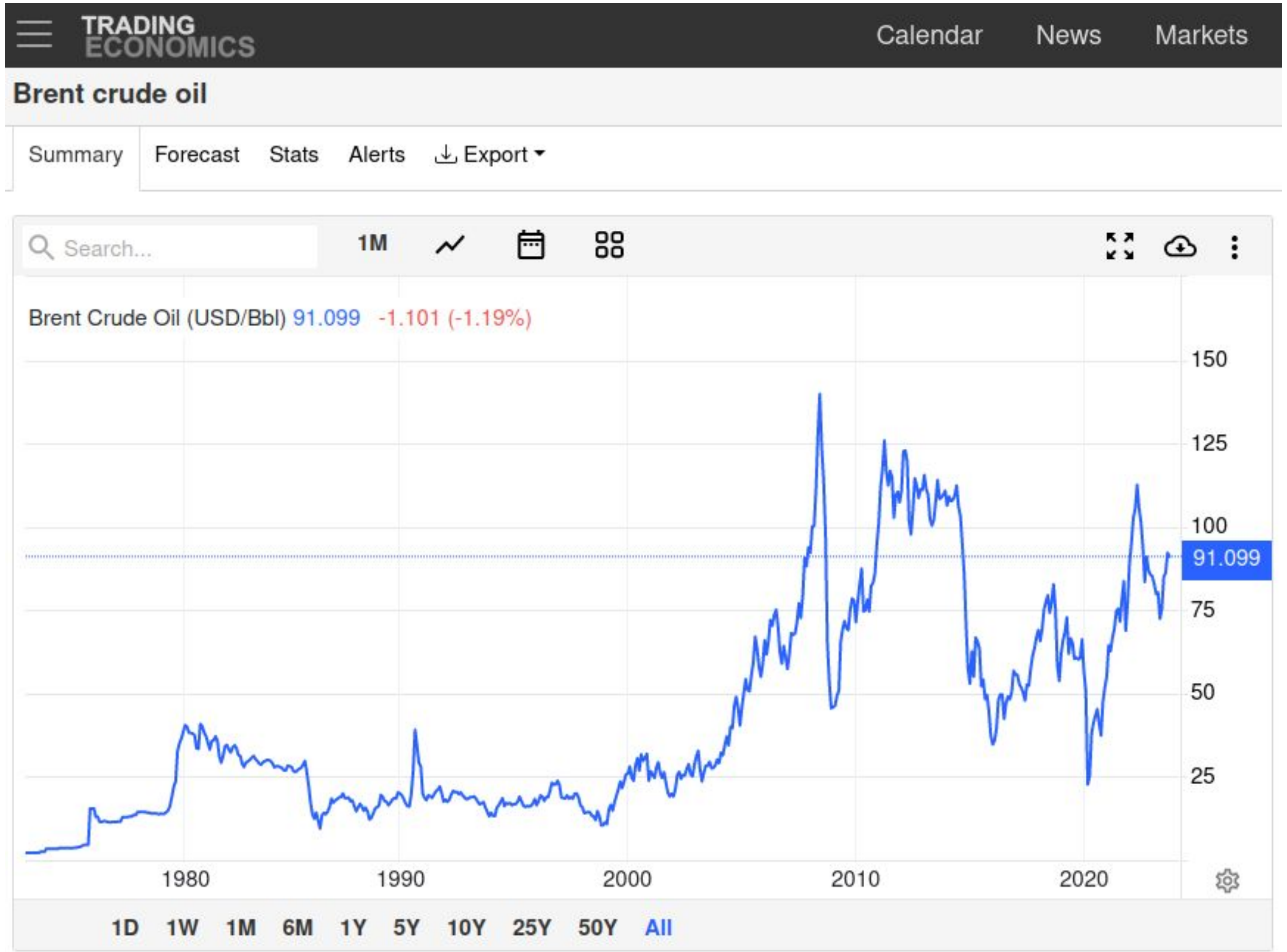
- Complexity
  - Complex systems
  - Complex relationships between features

- Data quality
  - Missing data
  - Outliers
  - Noise
  - …

- Non-stationary time series

- High-Dimensionality

Challenges: Complexity



- ❑ Supply and demand
- ❑ Geopolitical events
- ❑ Financial crises
- ❑ Pandemic
- ❑ Weather conditions
- ❑ Seasonal patterns
- ❑ Cyclic patterns
- ❑ Currency fluctuations

Challenges: Data Quality

- The quality of data significantly impacts the reliability of data analysis, modeling, and decision-making

- Problems commonly found:
  - Missing data
  - Anomalies
  - Inconsistent sampling rates
  - Duplicate data
  - Lack of consistency in units
  - Data drift
  - Noise
  - Data quality degradation over time

Use cases

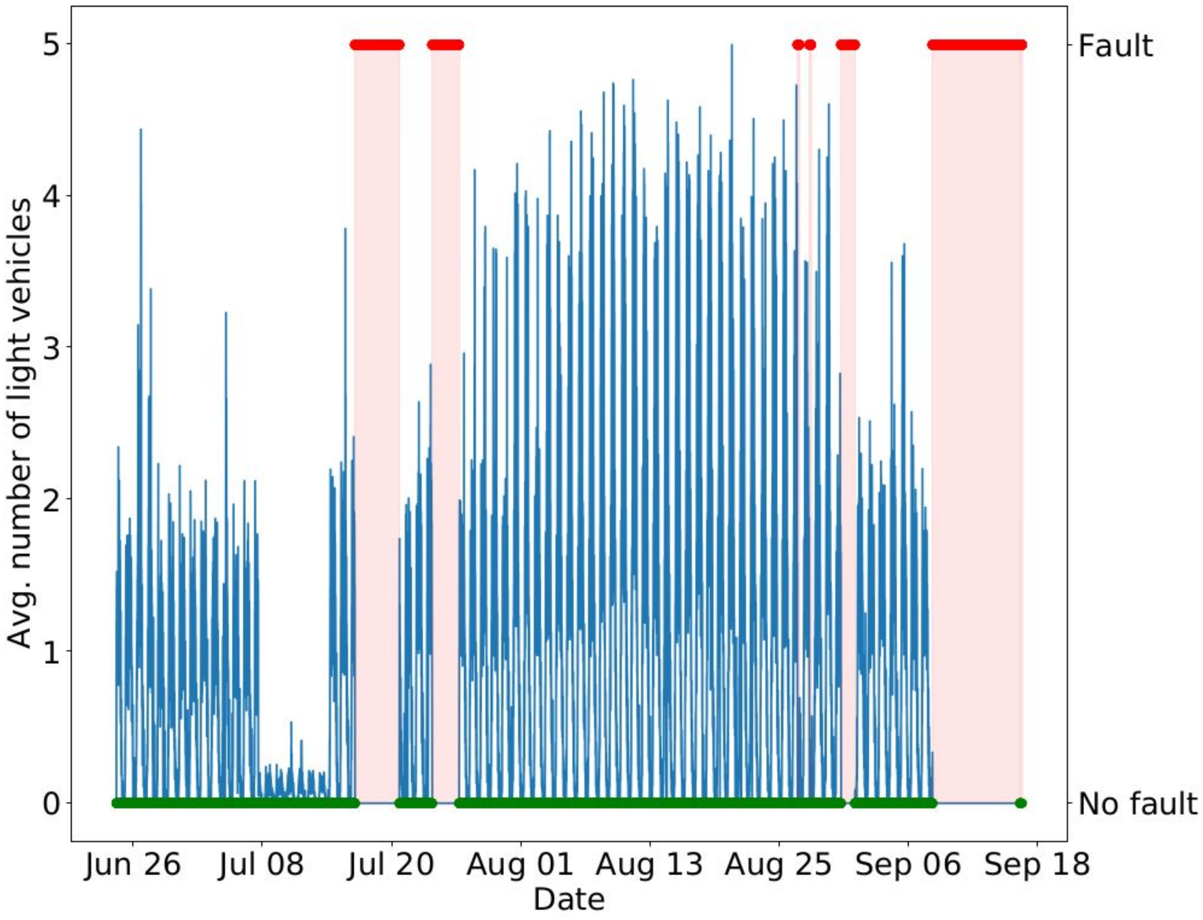
- Finances
  - Stock Market Analysis
  - Forecasting stock prices
- Healthcare
  - Patient monitoring
  - Detecting diseases
- Urban mobility
  - Traffic Management
  - Real-time traffic forecasting
  - Optimize traffic signal timing

My research aims to **forecast traffic metrics** even in the presence of **high ratios of missing data**.

Dealing with Missing Data

Time series missing data

- Missing data in time series can hide existing patterns and trends
- Simple imputation methods can fail in the presence of long intervals of missing data
- Data analysis and algorithms can be affected by the existence of missing data



How can we notice missing data?

- '*Nan*' values
- Masked missing data
  - Default values out of range/ that are not possible

    The number of people in Hamburg yesterday was -1

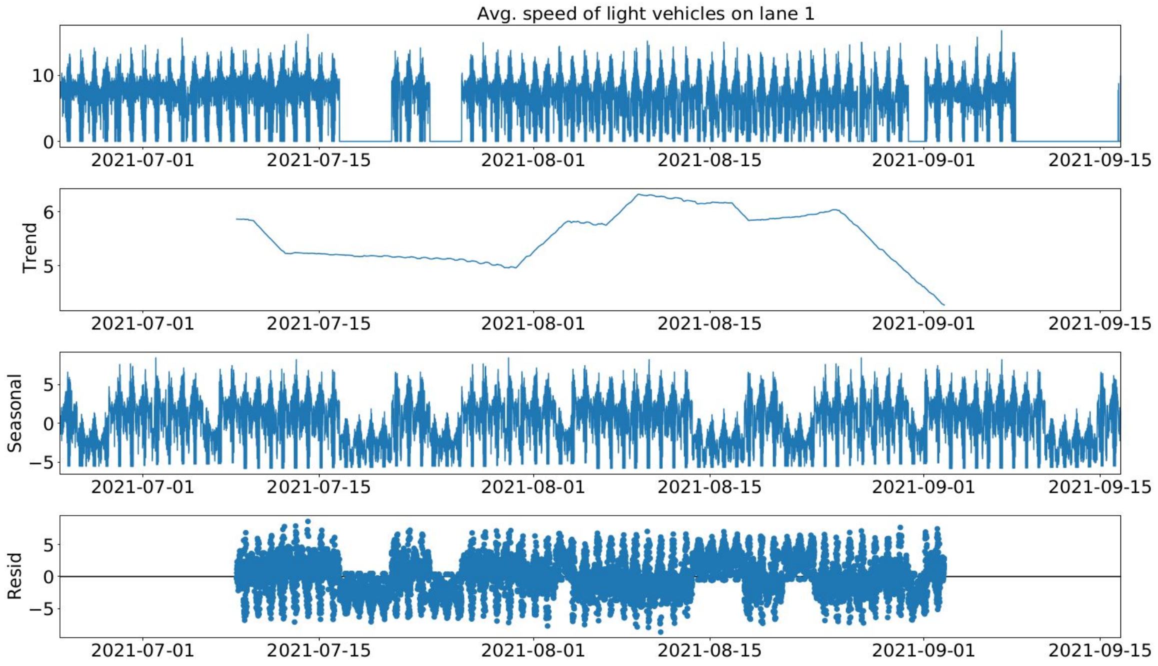
  - Default values inside the range/ are possible
    - This can be dangerous depending on the context
    - This can make it difficult for us to identify missing data

    The number of people in Hamburg yesterday was 0

    The speed of cars between 9 a.m. and 10 a.m. yesterday was 0 km/h

What is the impact of missing data?

- We can have **different percentages** of missing data
- Missing data can **affect one or more features**
- We can have problems that are more tolerant to missing data than others
- We can have models that deal better with lower rates of missing data, while other may deal better with higher rates of missing data
- We can have different scenarios of missing data



Missing data can affect data analysis, models, and algorithms. Having a negative impact on its applications in business and research.

How can we handle missing data?

- Before it happen
  - We can prevent it from happen by **monitoring our system and preventing conditions that lead to missing data**
  - However, this is not always possible!!!!
- After it happen
  - Ignore missing data (not a very good solution)
  - Delete observations with missing data (not adequate for time series, even if we have few observations with missing data)
  - Replace missing data for a value
    - 0 or values out of range (however, this can bring additional problems)
    - Mean, median, mode…
    - Simple univariate imputation techniques
      - e.g., Moving Average
    - Multivariate imputation techniques

Case study: OpenWeather dataset

- Dataset provided by OpenWeather
- Data from sensors such as temperature, pressure, humidity, wind speed, wind direction, wind gust, and cloudiness
- Data from 20 cities
- Hourly data from 2022

Workflow


Time series
Generate synthetic data
Time series with missing data
Missing data imputation
Reconstructed time series
Evaluation and model selection

Generation of synthetic missing data

	f_0	f_1	f_2	f_3	f_4		f_{n-2}	f_{n-1}
t_0	2	9	5	8	2	...	6	3
	...	...	...	...	...	...	...	...
t_i	5	X	6	X	X	...	X	X
t_{i+1}	7	X	8	X	X	...	X	X
	...	...	...	...	...	...	...	...
t_{i+p}	3	X	3	X	X	...	X	X
	...	...	...	...	...	...	...	...
t_{m-1}	8	9	7	0	2	...	9	4

p consecutive rows missing per column

k columns with missing data

Overlap

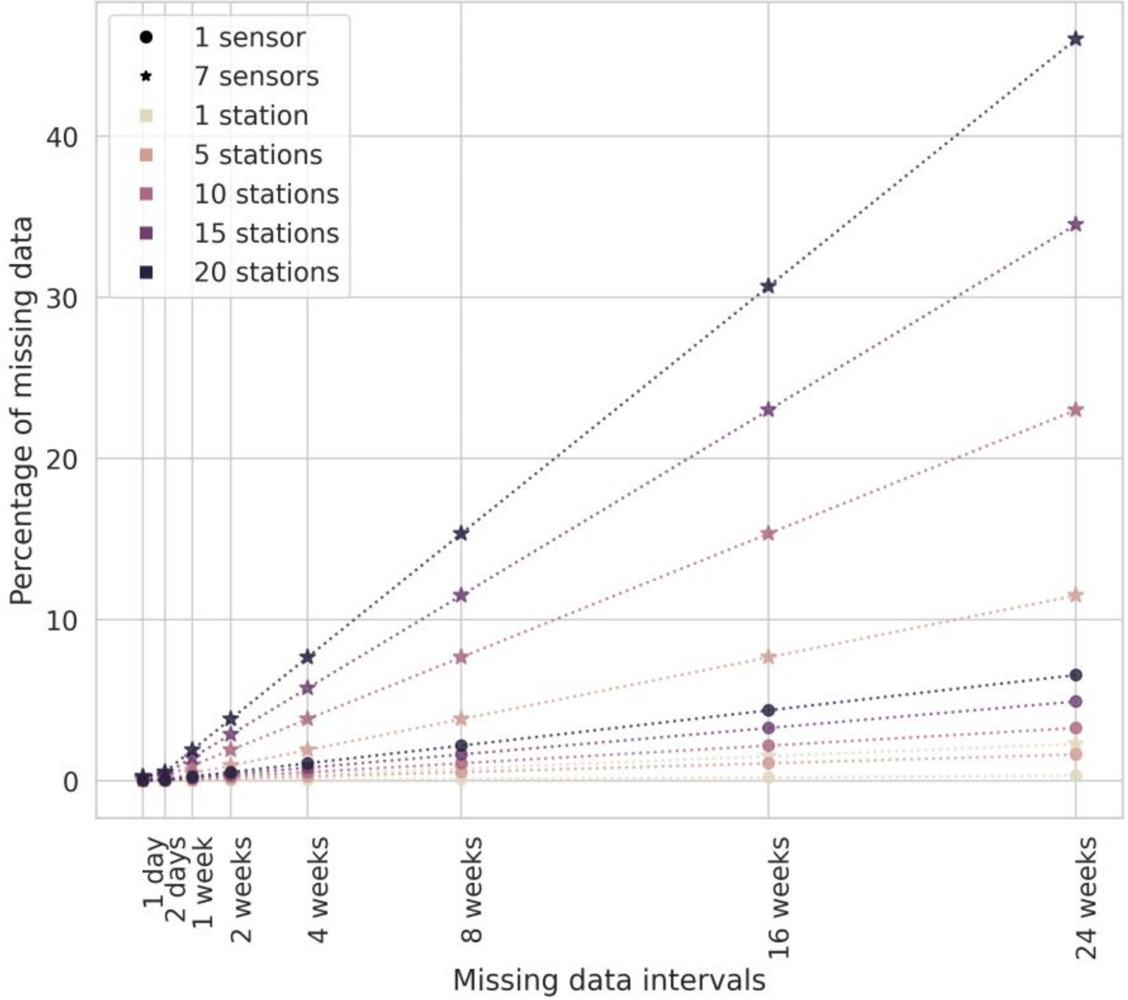
	f_0	f_1	f_2	f_3	f_4		f_{n-2}	f_{n-1}
t_0	2	X	5	8	2	...	6	3
	...	...	...	...	...	...	...	...
	5	4	6	X	X	...	X	6
	7	9	8	X	7	...	X	5
	...	...	...	...	...	...	...	...
	3	7	3	X	4	...		X
	...	...	...	...	...	...	...	...
t_{m-1}	8	9	7	0	2	...	9	4

p consecutive rows missing per column

k columns with missing data

Disjoint

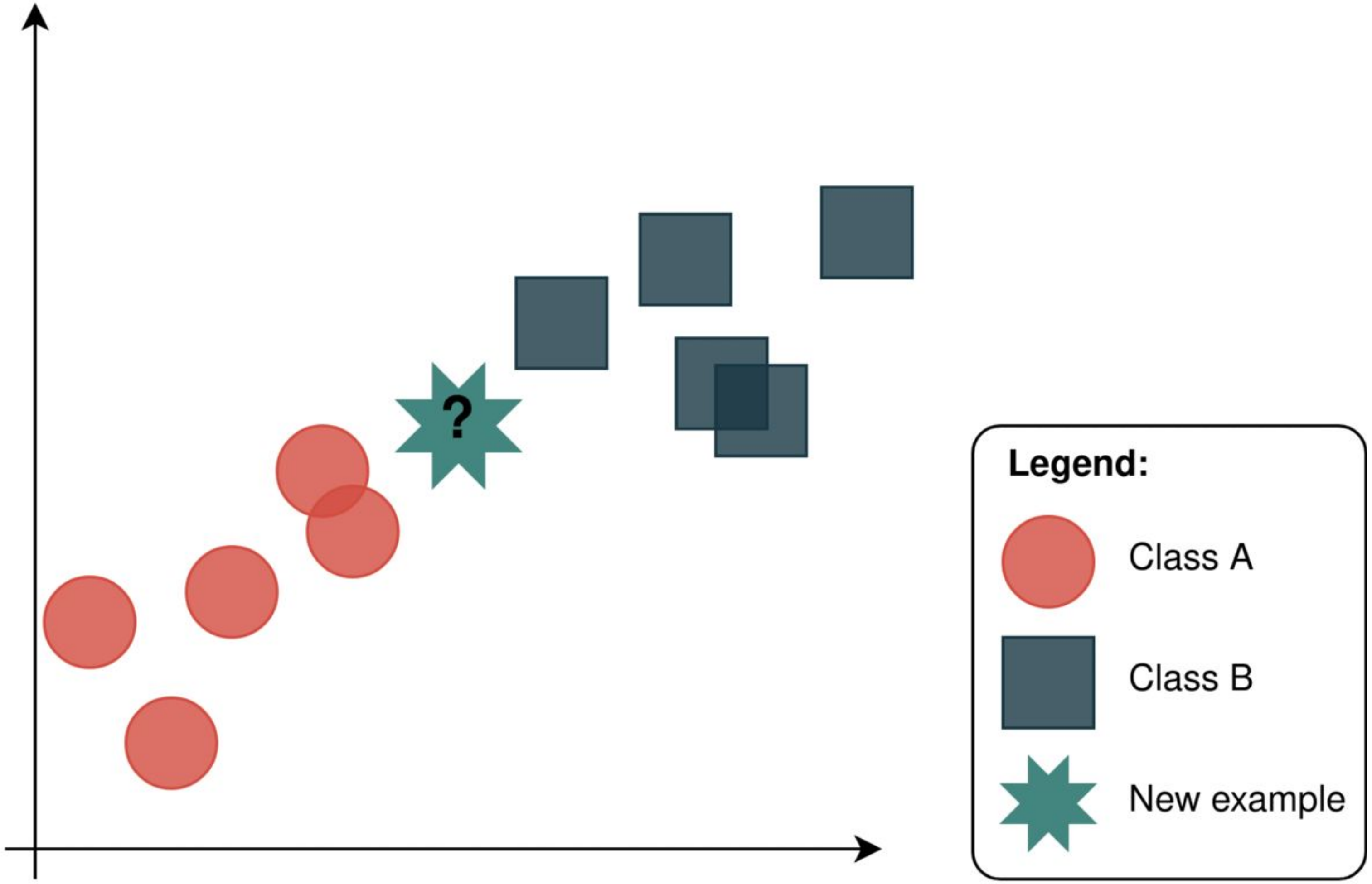
Generation of synthetic missing data



Algorithms

- We selected the top 3 from 20 statistical methods as baseline to evaluate our algorithm:
  - Replaced missing data using a specific value:
    - Mean, median, last value, previous value, the nearest value, zero
  - Interpolation techniques:
    - Barycentric, pchip, splines, polynomial functions, piecewise polynomial, akima
- Experimented with the KNN imputer
- Propose the Focalized KNN algorithm:
  - Based on KNN imputer

k-NN Algorithm

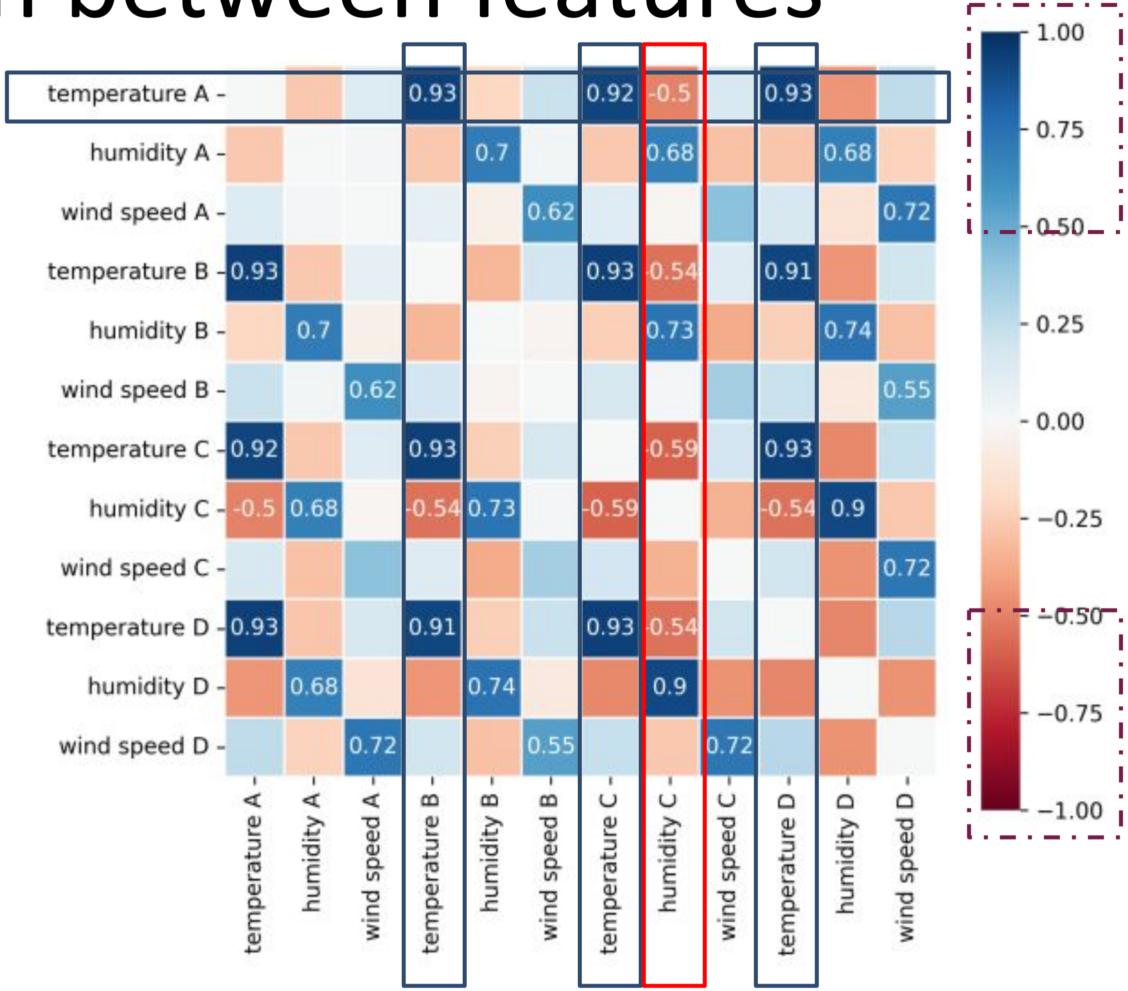


k-NN imputer

- Based on k-NN
- It can be used for imputation by using similar points to guess the missing data
- However, k-NN has some problems…
  - Suffers from the curse of high dimensionality
  - Stores the complete dataset in memory

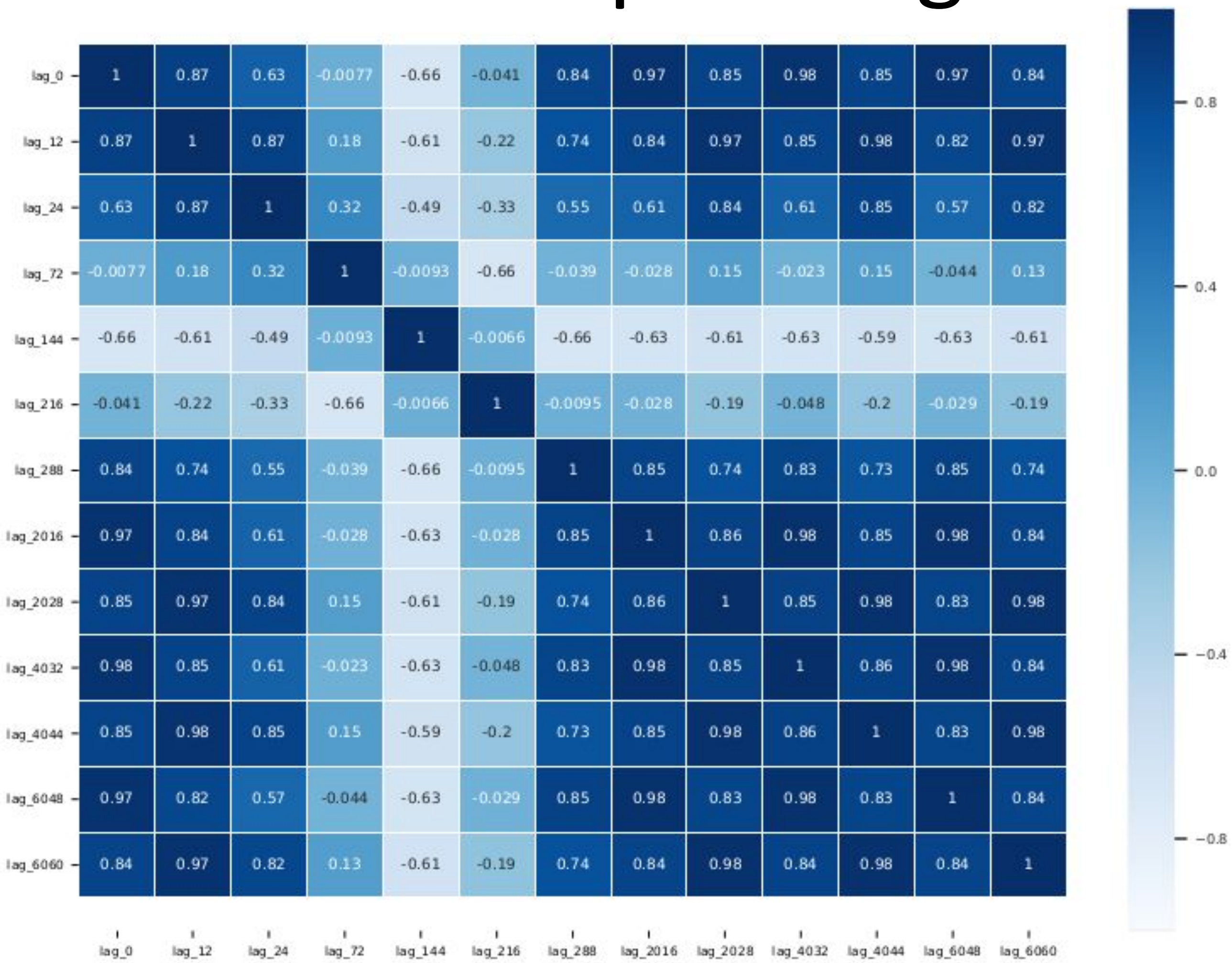
- KNN imputer can be used for time series datasets; however, it does not take advantage of time series properties!!!

Correlation between features



Spatio-temporal patterns!

Correlation between temporal lags



Focalized KNN

- Select the most correlated features
- Select the most relevant temporal lags
- Select the column with less missing data
  - Apply the KNN imputer for the matrix composed with:
    - the column c with missing data + correlated features + relevant temporal lags
  - Replace column c with the column with the imputed data
  - Repeat until there is no more missing data

Evaluation metrics

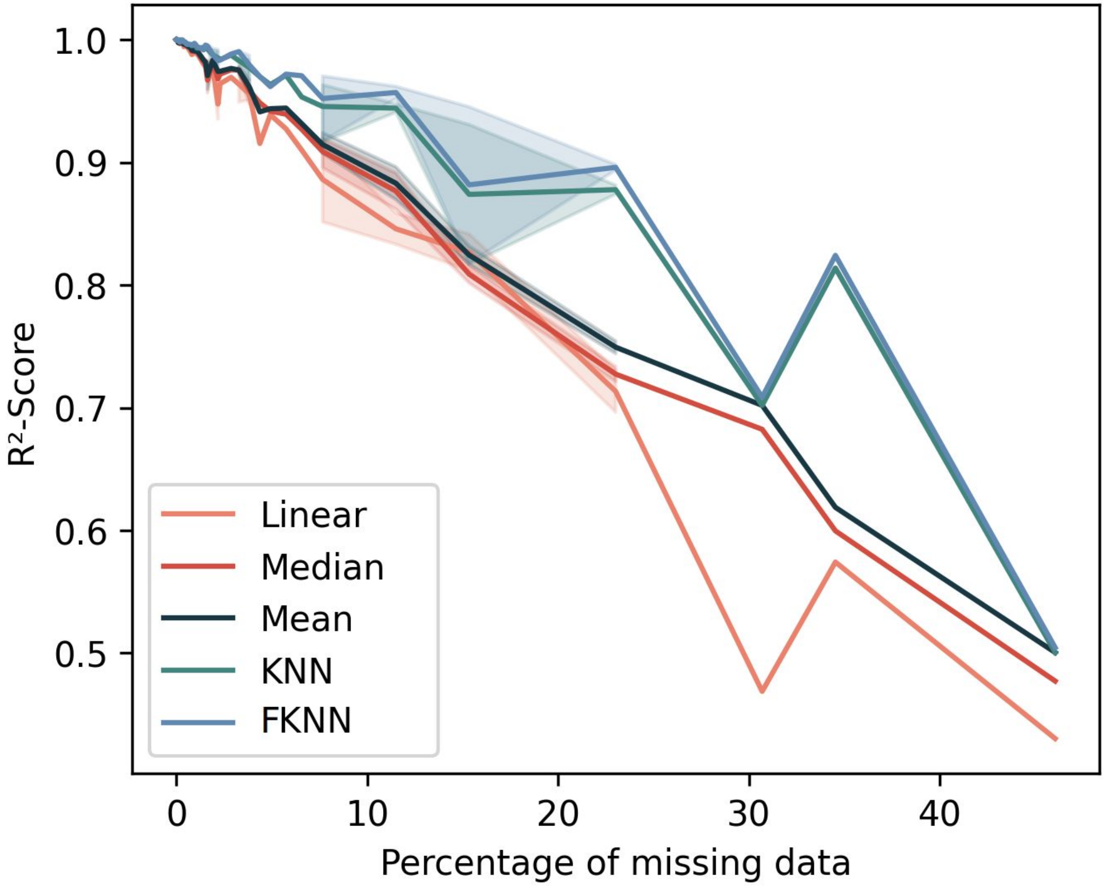
- Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

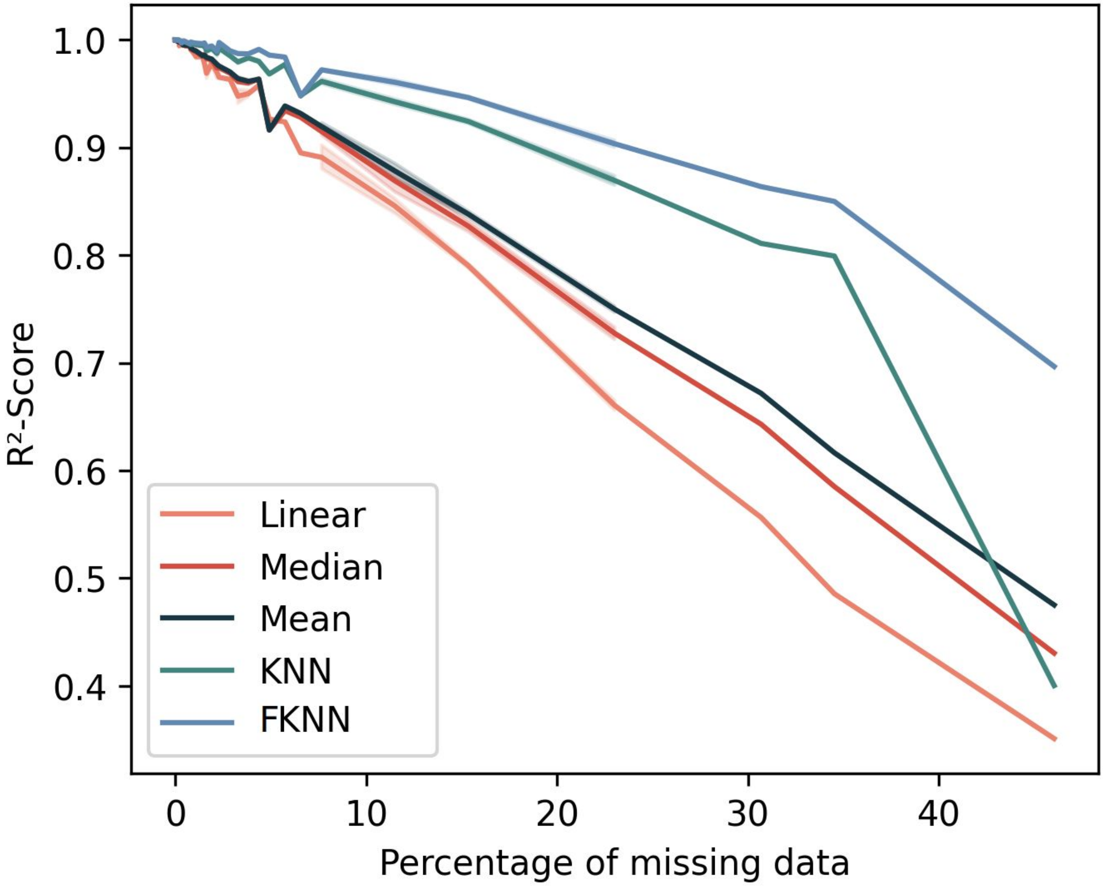
- R^2-Score

$$R^2 - Score = 1 - \frac{MSE}{MSE_{baseline}}$$

Overlap versus Disjoint missing pattern



Overlap



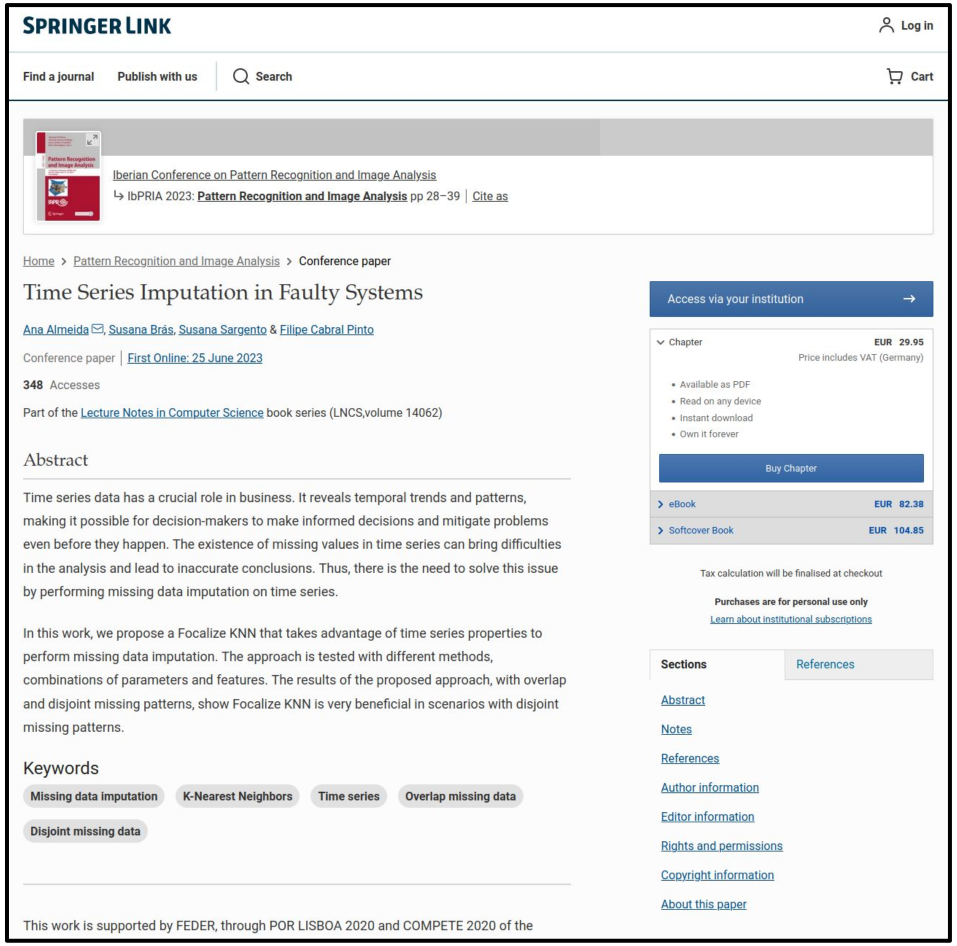
Disjoint

Discussion

- ❑ Pros
  - ❑ We do not need to train our algorithm
  - ❑ Good with disjoint patterns
  - ❑ Our solution helps with the curse of high dimensionality
- ❑ And Cons
  - ❑ Usually, it takes more time than regular KNN, especially if we have missing data affecting several columns
  - ❑ Not very good with overlap patterns

In the future, we would like to:

- ❑ Develop other methods to perform imputation in time series, such as methods based on Autoencoders
- ❑ Create more patterns of missing data to evaluate our models



Forecasting

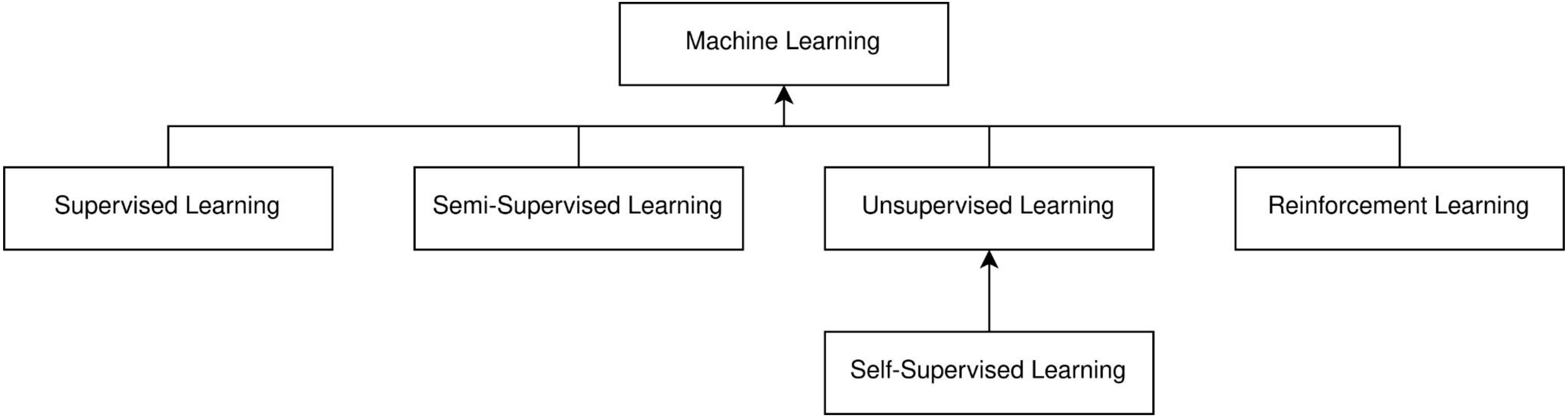
Brief overview of forecasting

- Forecasting is the process of making predictions about future events.
- Forecasting can be very beneficial in decision-making, planning, and risk management.
- There are different time horizons for forecasting, such as short, medium, and long-term.
- Before applying forecasting techniques, we should analyze the time series.

Forecasting methods

- Naïve methods
  - Use the last value to forecast the next one
- Statistical methods
  - AutoRegressive (AR)
  - Moving Average (MA)
  - Autoregressive Integrated Moving Average (ARIMA)
  - Seasonal ARIMA (SARIMA)
- Machine Learning methods
  - SVM
  - kNN
  - LightGBM
- Deep Learning methods
  - FNN
  - GRU
  - LSTM
  - CNN

What type of learning problem is forecasting?

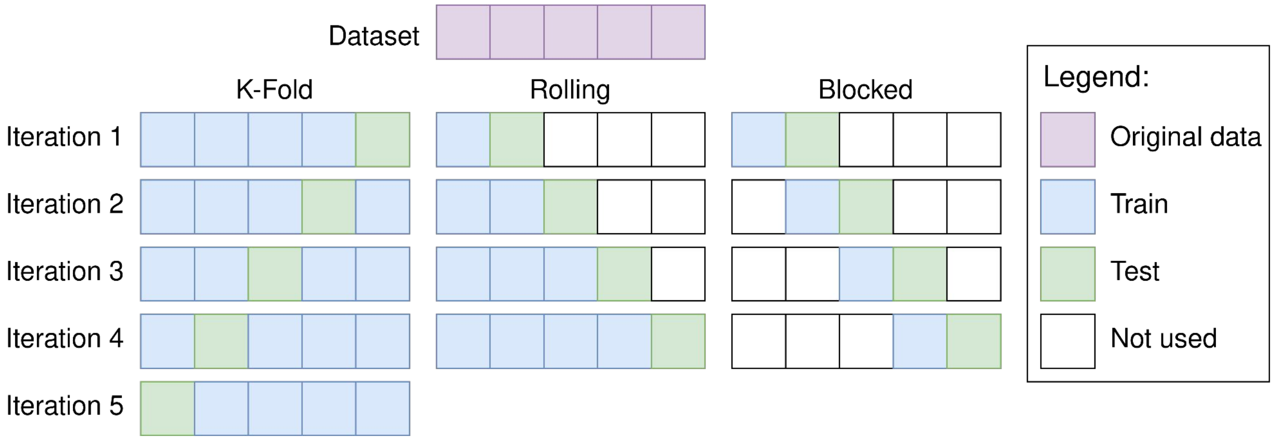


Choosing the best model for forecasting methods

- Most used evaluation metrics in the literature:
  - Mean Absolute Error (MAE)
  - Mean Absolute Percentage Error (MAPE)
  - Root Mean Squared Error (RMSE)
  - R^2-Score

Metrics based on comparing the **observed value** with the **predicted value**

- Cross-Validation for Time-series



Case study: Forecasting traffic flow

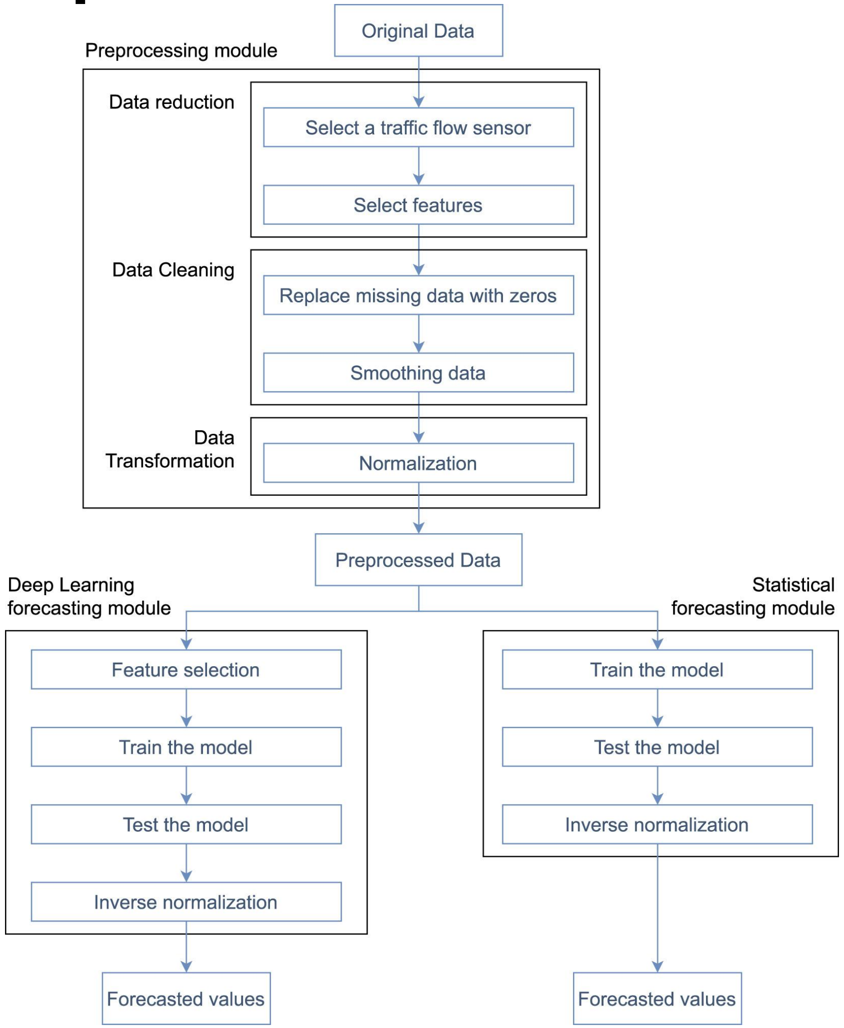
Dataset:

- Traffic counters deployed in Oporto, Portugal
- Data from September 30 to November 3 of 2019
- 5 minutes interval
- More than 100 sensors

Approach and Methods

- SARIMA
- FNN models
- LSTM-based models
- CNN-based models
- Hybrid LSTM-CNN models

Forecasting Pipeline



Results and Discussion

- Short-term forecasting
  - SARIMA achieves a good performance
  - Computationally light
  - (We can also use deep learning strategies)
- Long-term forecasting
  - SARIMA is not suitable for long-term forecasting
  - Deep learning strategies achieve good performance
  - The best model was based on CNNs
  - LSTMs take more time to train than CNNs or FNNs

ELSEVIER
Future Generation Computer Systems
Volume 128, March 2022, Pages 429-442

Vehicular traffic flow prediction using deployed traffic counters in a city ☆

Ana Almeida [a b], Susana Brás [a b], Ilídio Oliveira [a b], Susana Sargento [a c]

Show more

+ Add to Mendeley   Share   Cite

https://doi.org/10.1016/j.future.2021.10.022

Get rights and content

Highlights

- Development of a pipeline to forecast the traffic flow observed through both statistical and deep learning methods.
- Research approaches for anomaly detection and traffic prediction upon anomalies, through the counting traffic data.
- Platform for dynamic and stream-based traffic flow forecasting.

Abstract

The sustainable growth of cities created the need for better informed decisions based on information and communication technologies to sense the city and quantify its pulse. An important part of this concept of “smart cities” is the characterization of vehicular traffic flows and the prediction of urban mobility. Although there are several sensors that are able to infer the traffic flows in the city, road-mounted traffic counters can measure the number of vehicles in different parts of the roads. However, they are not usually used in

Future Work

Future Work

- Develop a model able predict future values even in the presence of missing data
  - Imputation + Forecasting

Q&A

Thank you!

anaa@ua.pt