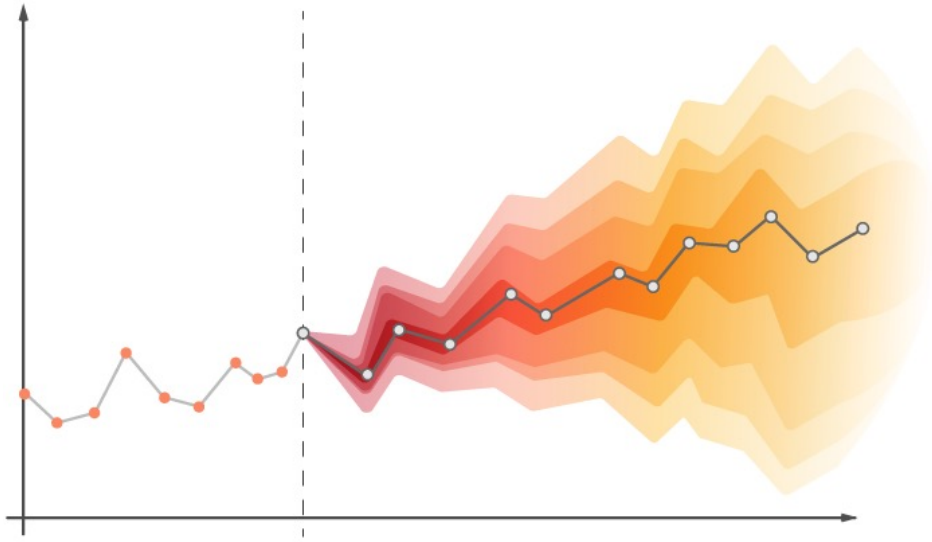
# Previsione Serie Storica di Monossido di Carbonio



Lorgna Lorenzo 829776

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

# Contenuti

# 01

# Introduzione

# Introduzione

## Obiettivo

Il progetto si pone come obiettivo l'implementazione di diversi algoritmi con lo scopo di effettuare una **previsione di valori di monossido di carbonio (CO)**

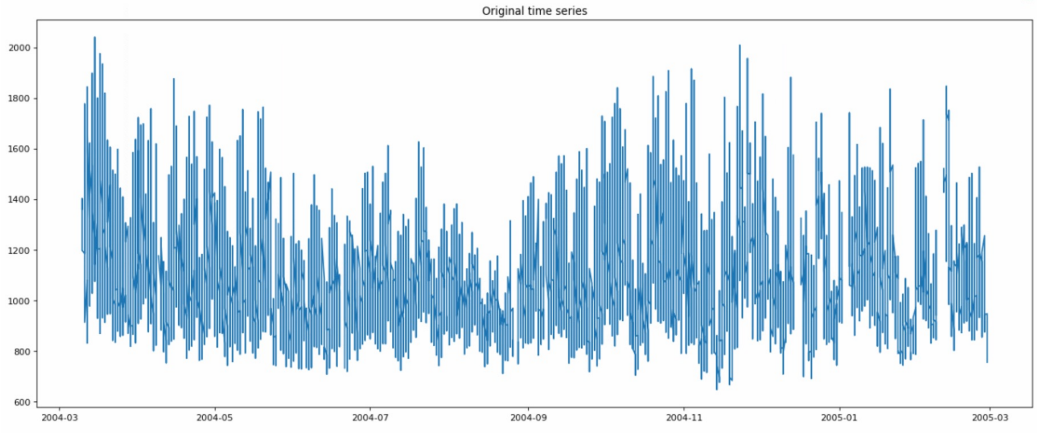
- ARIMA
- UCM
- MACHINE LEARNING

# Introduzione

## Dataset

Il dataset fornito rappresenta una **serie storica univariata** relativa a misurazioni orarie di CO. I dati sono organizzati nelle seguenti 3 colonne:

- **Date**
- **Hour**
- **CO**




I dati a disposizione si riferiscono al periodo: 10 marzo 2004 00:00 - 28 febbraio 2005 23:00

L'intento è prevedere il mese di **marzo 2005**

02

# Preprocessing e Analisi esplorativa

# Preprocessing

**Valori duplicati**

Si è verificata la presenza o meno di record duplicati evidenziandone l'**assenza**

**Valori mancanti**

Sono presenti **365 valori mancanti** per la feature CO. I valori mancanti sono stati sostituti con una **media** dei due valori precedenti e i due valori successivi, presi alla medesima ora e giorno della settimana

**Sistemazione feature**

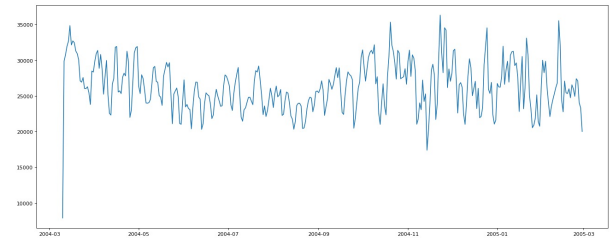
Le due feature **Date e Hour** sono state accorpate. Sono state generate altre fetaure quali ad esempio: **dayofweek**, **month**, **year** e **dayofmonth**

**Train e Test split**

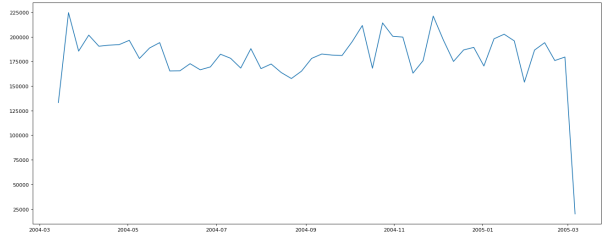
Il dataset è stato diviso in train e test, come mostrato. Si è deciso di considerare come **test set** febbraio 2005, l'ultimo mese a disposizione (**7.9% delle osservazioni**). Tutti i dati precedenti per il **training** del modello (**92.1% delle osservazioni**)

# Analisi esplorativa

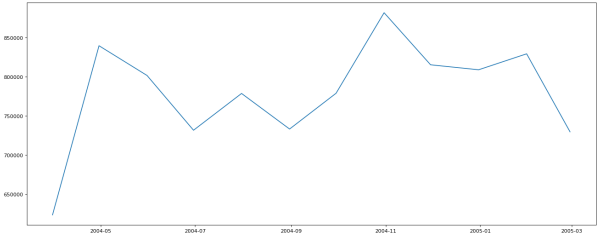
Livello giornaliero




Livello settimanale

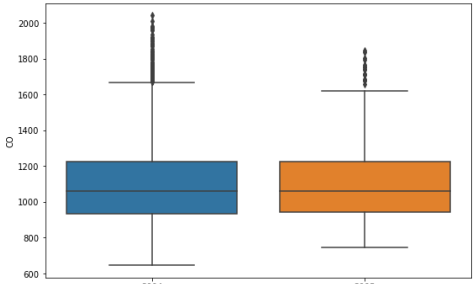



Livello mensile

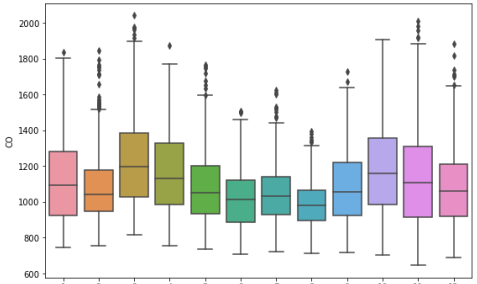



Assenza di un trend evidente

# Analisi esplorativa



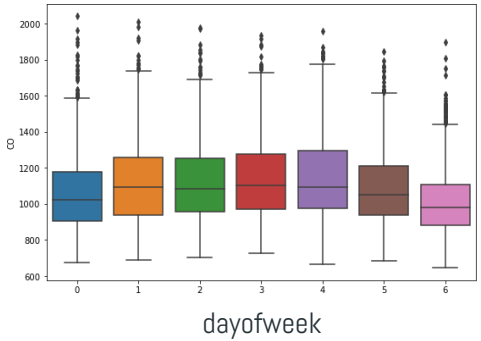

year



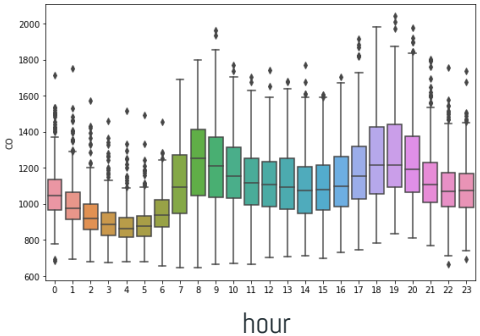

month

Distribuzione annuale e mensile

Distribuzione settimanale e giornaliera




dayofweek




hour

# 03

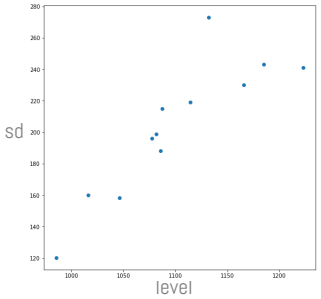
# Modelli sviluppati

# ARIMA

## Procedura di **Box e Jenkins**

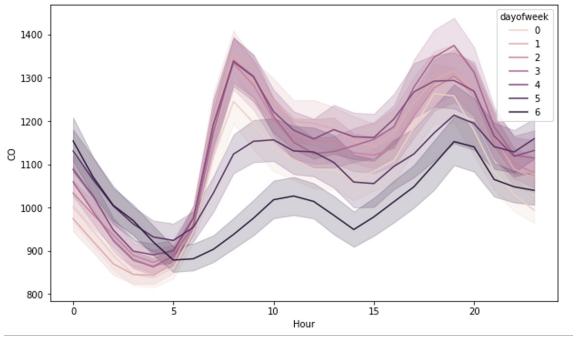
- verifica della **stazionarietà** della serie
- identificazione del modello ARIMA
- stima dei parametri
- verifica del modello tramite analisi dei residui

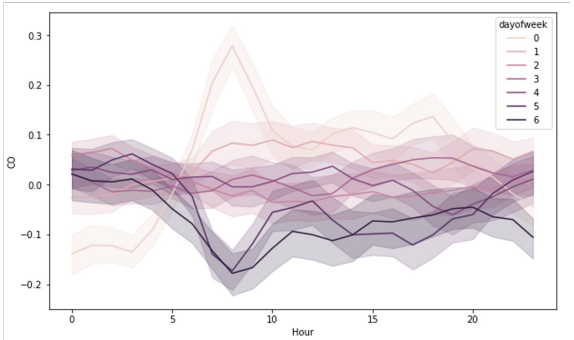
Analisi non stazionarietà in varianza




Analisi non stazionarietà in media

*Augmented Dickey-Fuller test*
*p-value: 0.197302*




Applicazione logaritmo e differenza stagionale

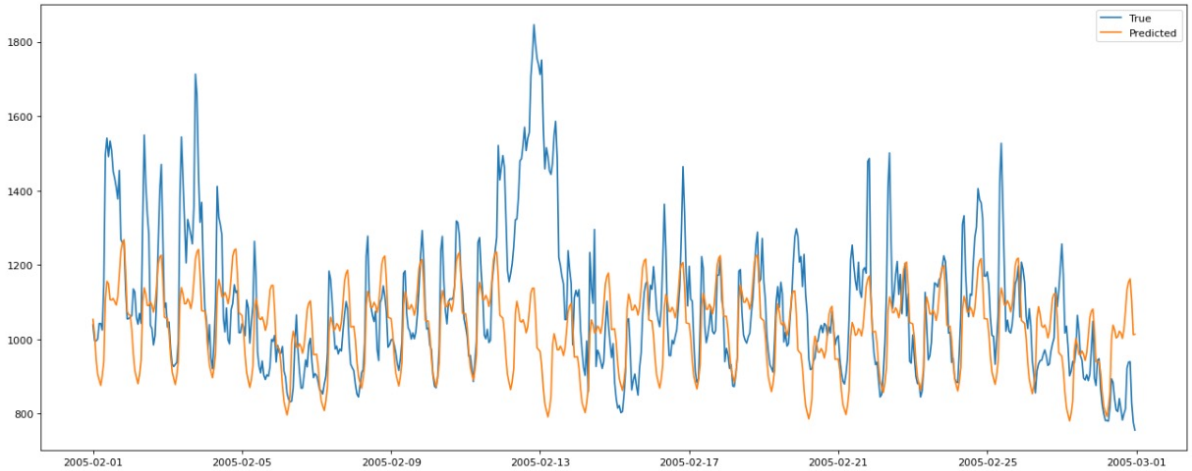
# ARIMA

I **parametri** considerati sono:

- p, d, q
- P, D, Q

| Modello | AIC |
|---|---|
| ARIMA (0, 1, 0)(0, 1, 0)[24] | -15544 |
| ARIMA (0, 1, 0)(0, 1, 1)[24] | -19685 |
| ARIMA (0, 1, 0)(1, 1, 1)[24] | -19742 |
| ARIMA (2, 1, 2)(1, 1, 1)[24] | -20260 |
| ARIMA (2, 1, 2)(1, 1, 1)[24] con regressori | -20297 |

Previsioni sul test con il modello migliore **ARIMA (2, 1, 2)(1, 1, 1)[24] con 3 sinusoidi a periodo 168**



# UCM

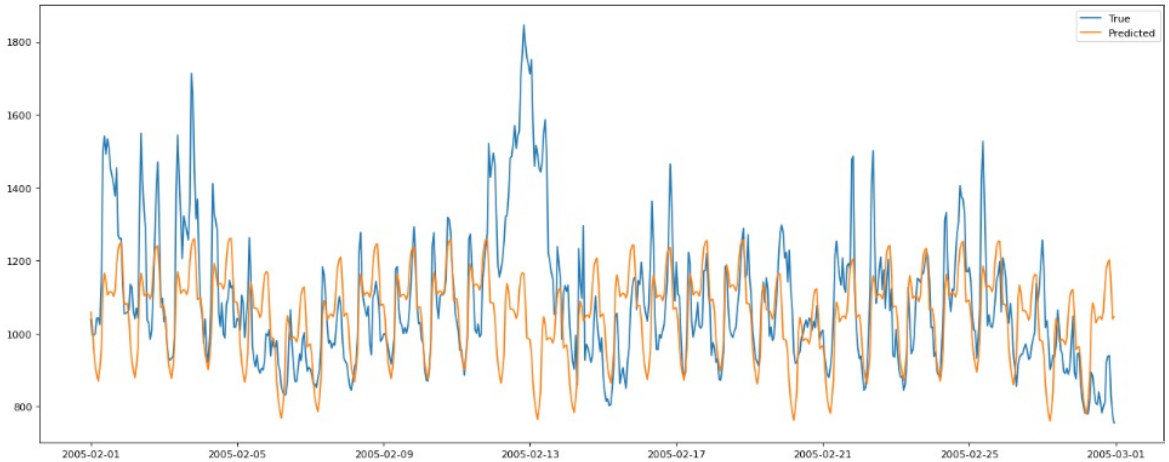
I **parametri** considerati sono:

- LEVEL
- SEASONAL

| Modello | MAPE (%) | AIC |
|---|---|---|
| No trend | 100.73 | 132294.86 |
| Deterministic constant | 12.16 | 104027.66 |
| Local level | 12.20 | 91661.99 |
| Random walk | 12.20 | 91569.99 |
| Deterministic trend | 11.95 | 104028.77 |
| Local linear deterministic trend | 12.04 | 91654.41 |
| Random walk with drift | 12.04 | 91652.40 |
| Local Linear trend | 12.04 | 91656.40 |
| Smooth trend | 2412.27 | 94506.95 |
| Random trend | 1487.71 | 96516.59 |

Previsioni sul test con il modello migliore con componenti:

- **LEVEL**: Random walk with drift
- **SEASONAL**: 3 sinusoidi di periodo 24 e 2 sinusoidi di periodo 168



# MACHINE LEARNING – KNN

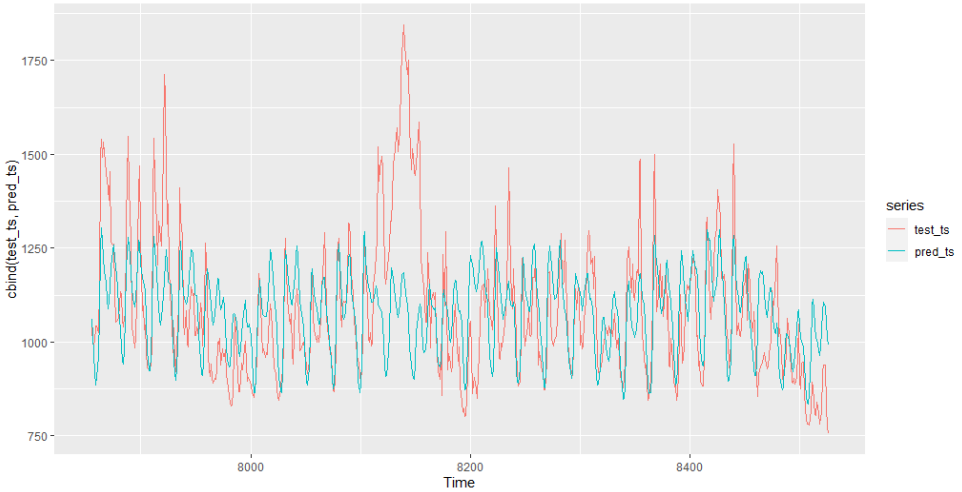
Strategia multi-step multi-output (MIMO)

I **parametri** considerati sono:

- **p**: numero di valori passati che vengono considerati. Diverse prove con (2,4,6, … ,36,38,40)
- **k**: numero di sottosequenze simili da prendere in considerazione. Diverse prove con (24*7, 24*14, 24*21)

Previsioni sul test con il modello migliore con parametri:

- **p**: 504
- **k**: 18



# MACHINE LEARNING – LSTM

I **parametri** considerati sono:

- lookback
- Numero di layer e numero di neuroni
- Numero di epoche
- batchsize

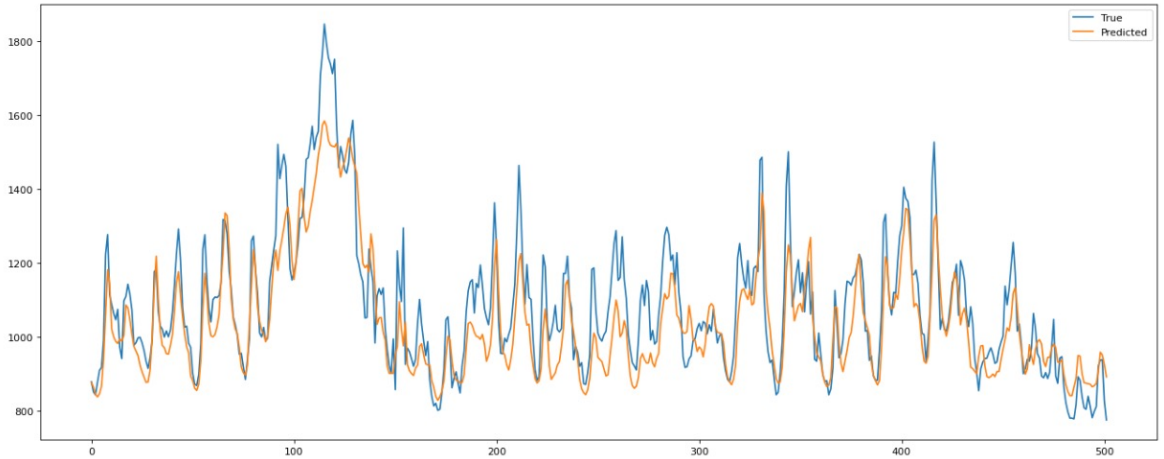
Operazioni di **preprocessing** necessarie:

- Normalizzazione dataset
- Rigenerazione train test set
- Reshape strutture dati

Previsioni sul test con il modello migliore con la seguente **architettura**:

- Layer LSTM da 10 neuroni
- LeakyReLU come funzione di attivazione
- Layer di Droput con dropout rate di 0.2
- Layer LSTM da 5 neuroni
- LeakyReLU come funzione di attivazione
- Layer di output di tipo Dense da 1 neurone

100 epoche, batchsize pari a 32, shuffle=False, Adam optimizer, EarlyStopping



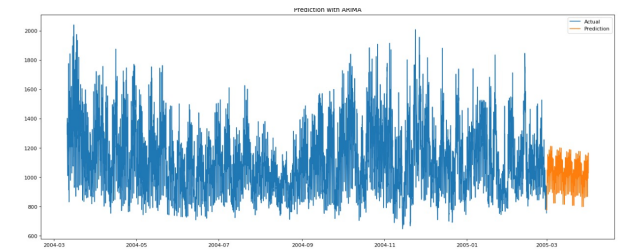
04

# Conclusioni e Previsioni

# Previsioni

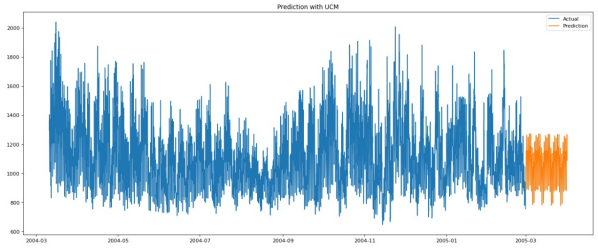
## ARIMA

MAPE: 10.62%



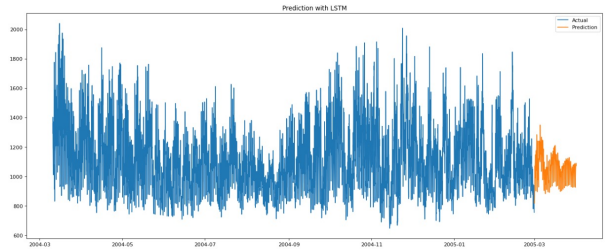

## UCM

MAPE: 11.02%




## LSTM

MAPE: 06.45%




Previsioni marzo 2005

# Conclusioni

## Conclusioni

Il miglior modello osservando le performance sul test set in termini di MAPE è stato ottenuto dalla rete **LSTM**, della famiglia machine learning. Per quanto riguarda invece i modelli ARIMA e UCM registrano performance molto simili.

## Sviluppi futuri

Ipotizzando alcuni sviluppi futuri si potrebbe porre una maggiore attenzione alla modellazione della **stagionalità** per cercare di avere dei risultati migliori e ai parametri delle reti neurali per evitare problemi di **overfitting**.

# Grazie per l'attenzione!