

UNIVERSITÀ DEGLI STUDI DI  
MILANO-BICOCCA

STREAMING DATA MANAGEMENT AND TIME  
SERIES ANALYSIS

FINAL PROJECT

---

## Predizione dei livelli di ossido di carbonio

---

*Authors:*

Nataliya Zayeva - 867981- n.zayeva@campus.unimib.it  
17 gennaio 2022



# 1 Introduzione

Il progetto di Streaming Data Management and time series analysis ha l'obiettivo di predire la serie storica, presa in considerazione, per il mese di marzo 2005. Per la previsione vengono sviluppati i seguenti modelli:

- Modelli **ARIMA**
- Modelli **UCM**
- Modelli **non lineari** (Machine Learning)

# 2 Dataset

Il Dataset utilizzato per la previsione dei livelli orari di ossido di carbonio va dal 10 Marzo 2004 alle 18 al 23 Febbraio 2005 alle 23, per un totale di 8526 ore. È suddiviso in 3 colonne:

- Date: data di riferimento in formato *yyyy-mm-dd*
- Hour: ora di riferimento, i cui valori vanno da 0 a 23, rappresentando gli intervalli di tempo. (Si va da 0 che rappresenta l'intervallo 00:00 - 00:59 fino a 23 che rappresenta l'intervallo 23:00 - 23:59)
- CO: valore di ossido di carbonio rilevato

Per l'analisi, il dataset è stato diviso in training e validation dataset, considerando per il primo il periodo che va dal 10 Marzo 2004 ore 18 al 1 Gennaio 2005 ore 0, e per il secondo i restanti due mesi. Tale decisione è stata presa per studiare una possibile stagionalità giornaliera, settimanale o mensile e valutare la performance del modello. Di seguito viene rappresentata la suddivisione del dataset (*figura 1*)

Inizialmente il dataset conteneva il 4% di valori mancanti che sono stati successivamente sostituiti con la predizione del modello migliore sul training dataset, ottenendo una serie completa

L'obiettivo dello studio è di prevedere l'andamento della serie storica, cioè l'andamento del livello dell'ossido di carbonio in varie fasce orarie per un periodo che va dal 1 Marzo 2005 ore 0 al 31 Marzo 2005 ore 23.

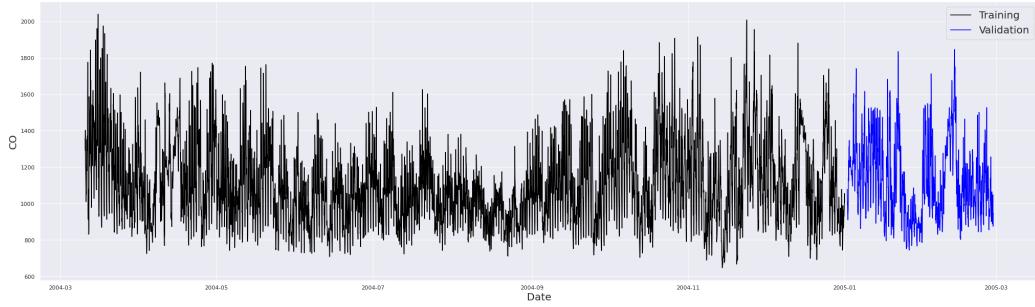


Figura 1: Suddivisione della serie storica in Training e Validation set

### 3 Approccio metodologico

Prima di procedere allo sviluppo dei modelli, per la previsione, sono state effettuate delle analisi qualitative dei dati per studiare meglio eventuali andamenti di tipo crescente/decrescente o di eventuali picchi positivi/negativi della serie.

Da una prima osservazione della *figura 1* si nota una possibile presenza di stagionalità mensile o settimanale, portando a supporre che la serie storica non è stazionaria. Per studiare un possibile pattern la serie è stata decomposta per trend, stagionalità e sono stati visualizzati i residui (*figura 2*).

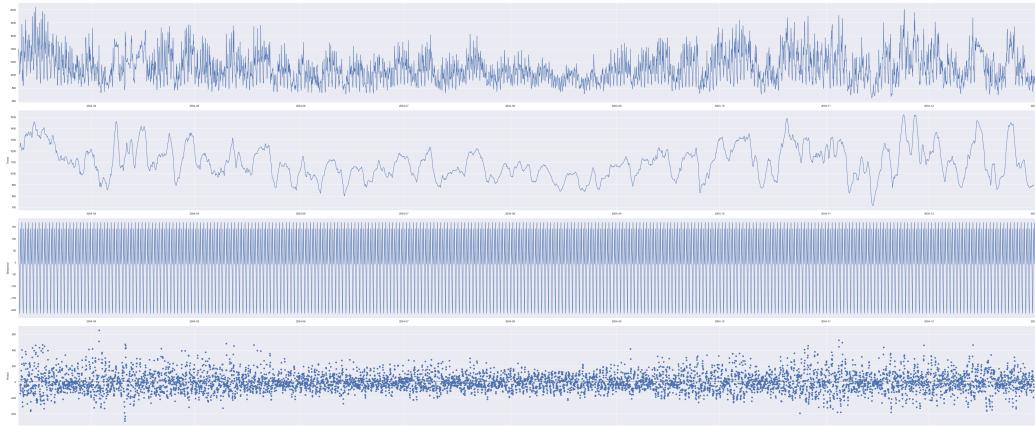


Figura 2: Decomposizione della serie storica

Dalla decomposizione si può osservare che, come predetto, non c'è un trend significativo (crescente/decrescente), ma la serie presenta una stagionalità,

che è stata maggiormente rappresentata in *figura 3*.

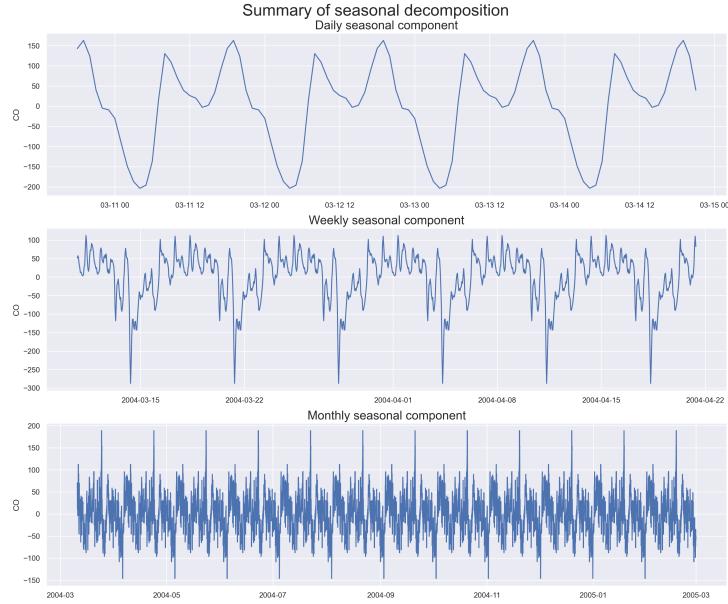


Figura 3: Decomposizione stagionale della serie per giorni, settimane e mesi

La serie è stata decomposta per stagionalità giornaliera, settimanale e mensile. Nel primo grafico in *figura 3* si può osservare un andamento sinusoidale, con dei picchi negativi nelle ore notturne dove l'ossido di carbonio rilevato è minore rispetto alle ore diurne. Nel secondo grafico sono presenti dei picchi settimanali negativi in corrispondenza dei weekend.

Nella *figura 4* sono rappresentati i boxplot della serie aggregata per le ore.

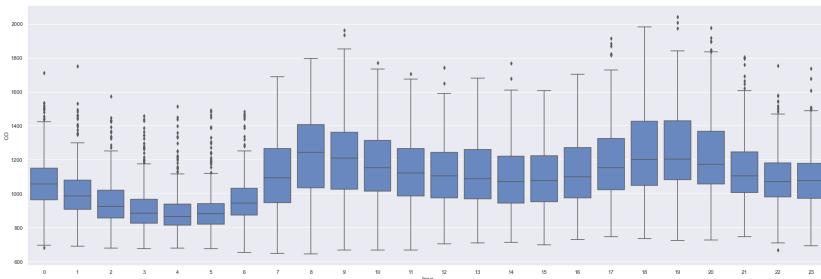


Figura 4: Boxplot per ore del giorno

Il grafico ci conferma che nelle ore notturne l'ossido di carbonio è minore rispetto alle ore diurne, si osservano due orari di picco (le 7 e le 18), forse dovuto a un maggior afflusso di persone che si spostano, per poi diminuire nelle ore notturne.

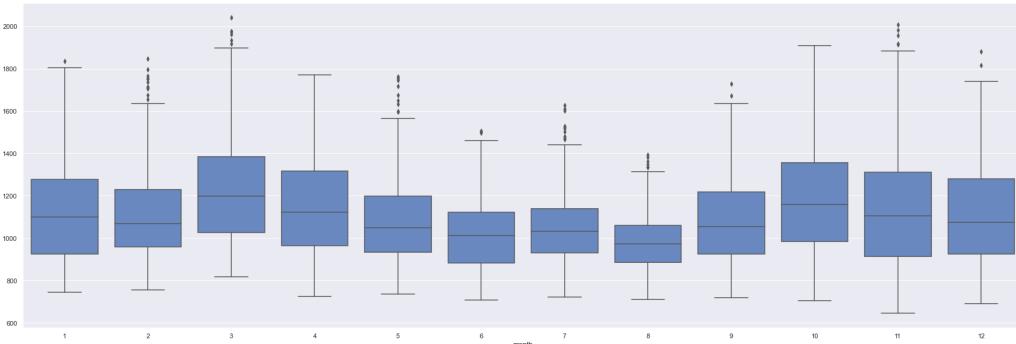


Figura 5: Boxplot per mese

In *figura 5* si possono osservare i boxplot per l'aggregazione mensile da cui emerge che sono presenti delle leggere diminuzioni del livello dell'ossido di carbonio nei mesi estivi che poi aumentano nei mesi invernali. La diminuzione nei mesi estivi può essere dovuta a una diminuzione dell' utilizzo del riscaldamento e una maggiore produzione di ossigeno da parte della vegetazione.

## 4 Modelli di previsione

### 4.1 ARIMA

Prima di procedere ai modelli di previsione bisogna seguire la procedura di Box e Jenkins che ci permette di costruire, a partire dall'osservazione dei dati, un modello ARMA per approssimare il processo generatore della serie storica. La prima analisi da eseguire è l'analisi dei correlogrammi della serie (*figura 6 e 7* )

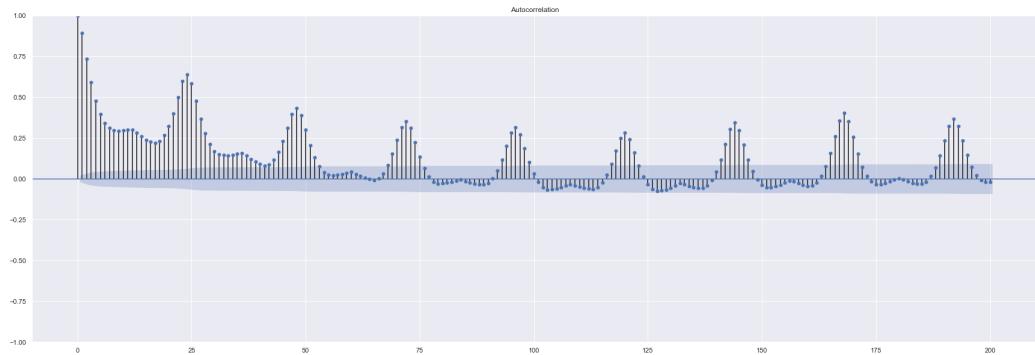


Figura 6: ACF

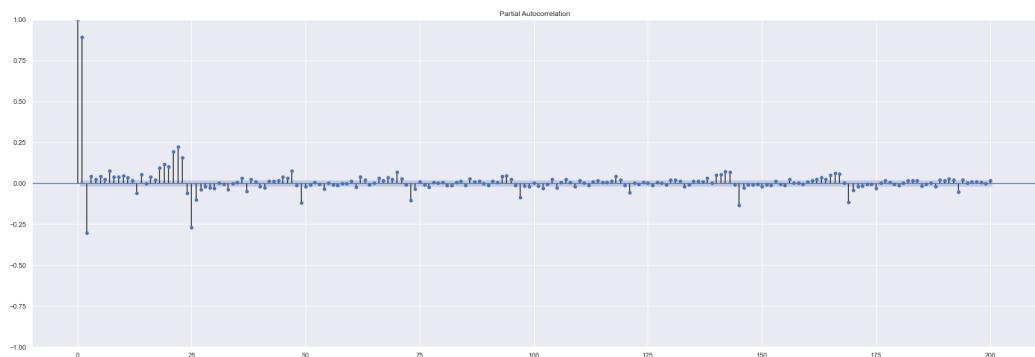


Figura 7: PACF

La funzione di autocorrelazione globale (ACF) ai ritardi stagionali tende a zero lentamente e ha un andamento sinousidale, questo conferma che la serie storica non è stazionaria. Nella PACF si può osservare un andamento stagionale ogni 24 lag che tende a diminuire e può essere usato nel SAR(24). Dai correlogrammi possiamo supporre che la serie storica ha sia un processo AR (autoregressivo) sia un processo MA (a media mobile). Dall'analisi delle funzioni di autocorrelazione e dopo aver provato diversi modelli, si è giunti ad utilizzare un ARIMA(4,0,1) testandolo con o senza stagionalità.

Sono stati studiati i seguenti 3 modelli:

- SARIMA(4,0,1)(1,1,1)<sub>24</sub>
- ARIMAX(4,0,1) con regressori sinusoidali e dummy stocastiche
- SARIMAX(4,0,1)(1,0,1)<sub>24</sub> con regressori sinusoidali e dummy stocastiche .

Per la valutazione e il confronto dei modelli statistici ARMA è stato utilizzato l'AIC(*Akaike's information criterion*). Il primo modello considerato è il SARIMAX(4,0,1)(1,1,1)<sub>24</sub> e l'AIC risulta essere 81915 ma dal grafico delle previsioni confrontate con il validation set emerge che il modello non riesce a modellare la stagionalità.

Per risolvere tale problema vengono aggiunte delle dummy stocastiche e i regressori sinousidali. Le dummy stocastiche considerato sono: le ore diurne (dalle 7 alle 20) , le ore di traffico (ore 6 e 18) e i mesi di primaverili/estivi. Invece per quanto riguarda i regressori sinusoidali è stato utilizzato Fourier per gli anni, mesi, giorni e settimane. Gli altri due modelli considerati sono stati: ARIMAX e SARIMAX con l'aggiunta dei regressori sinusoidali e le dummy stocastiche, che hanno portato ai seguenti risultati:

Modello	AIC
SARIMA(4,0,1)(1,1,1) <sub>24</sub>	81915
ARIMAX(4,0,1)regressori sinusoidali e dummy stocastiche	83336
SARIMAX(4,0,1)(1,0,1) <sub>24</sub> regressori sinusoidali e dummy stocastiche	82767

Il modello migliore risulta essere SARIMAX(4,0,1)(1,0,1)<sub>24</sub> con regressori sinusoidali e dummy stocastiche, la predizione sul validation set è in *figura 8*

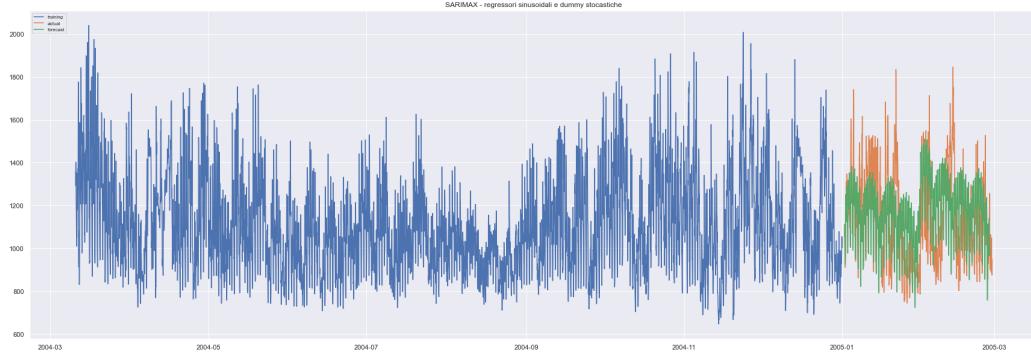


Figura 8: SARIMAX regressori sinusoidali e dummy stocastiche

Per concludere la procedura di Box-Jenkins è stata effettuata l’analisi dei residui sul modello migliore, testando le ipotesi che i residui sono distribuiti normalmente e incorrelati. Per controllare la normalità sono stati usati 3 test: Test di Shapiro-Wilk, Test di Kolmogorov-Smirnov e Test Anderson-Darling in tutti e tre i casi non possiamo accettare la normalità dei residui. In *figura 9 e 10* si osservano l’ACF e PACF dei residui del modello migliore, e si osserva una consistente diminuzione delle autocorrelazioni, quindi si può accettare l’ipotesi che i residui siano incorrelati.

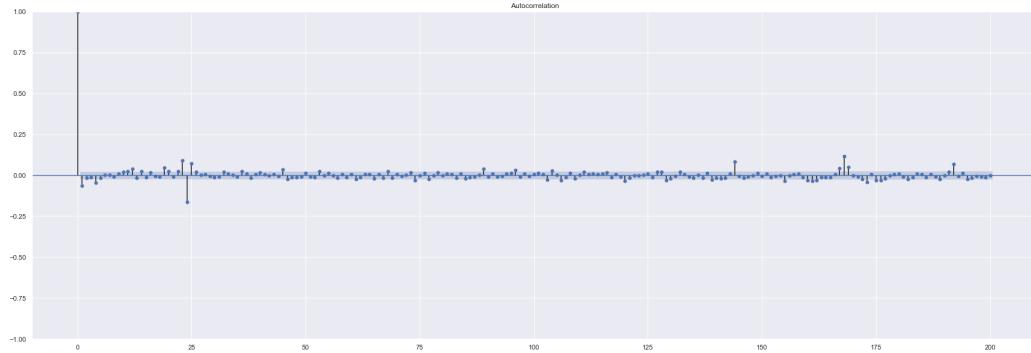


Figura 9: ACF SARIMAX con regressori sinusoidali e dummy stocastiche

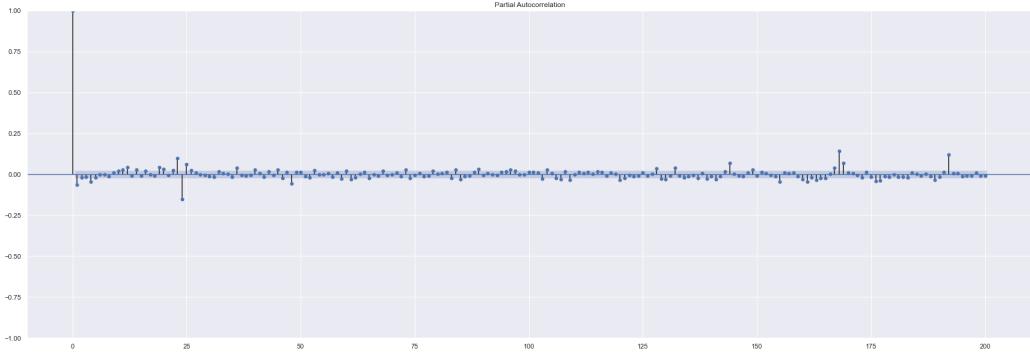


Figura 10: PACF SARIMAX con regressori sinusoidali e dummy stocastiche

## 4.2 UCM

I modelli UCM (*Unobserved Components Model*) a componenti non osservabili performano la decomposizione della serie storica in componenti come trend, ciclo e stagionalità. Per tutti i modelli sono stati utilizzati le sinusoidi stocastiche e le dummy stocastiche per i giorni della settimana, i mesi, i weekend, i mesi primaverili/estivi e le ore di traffico. Le sinusoidi stocastiche sono state modellate con l'utilizzo delle frequenze nella libreria *UnobservedComponents* di Python. Per la stima del trend sono stati testati 3 diversi metodi:

- Local Linear Trend
- Random Walk
- Deterministic Trend

Per ogni modello è stato calcolato il MAPE e il modello migliore risulta essere con il Local Linear Trend. La *figura 11* rappresenta la predizione del modello Local Linear Trend sul validation set. L'area rossa rappresenta l'intervallo di confidenza.

Facendo la previsione su tutta la serie storica si osserva che la curva ha un declino per i mesi predetti (marzo2005), quindi il modello non performa bene. Lo stesso succede anche per il Random Walk, quindi si considera come modello migliore il Deterministic Trend che ha un MAPE superiore agli altri due. In *figura 12* è rappresentato l'andamento della serie storica predetta sul validation set .

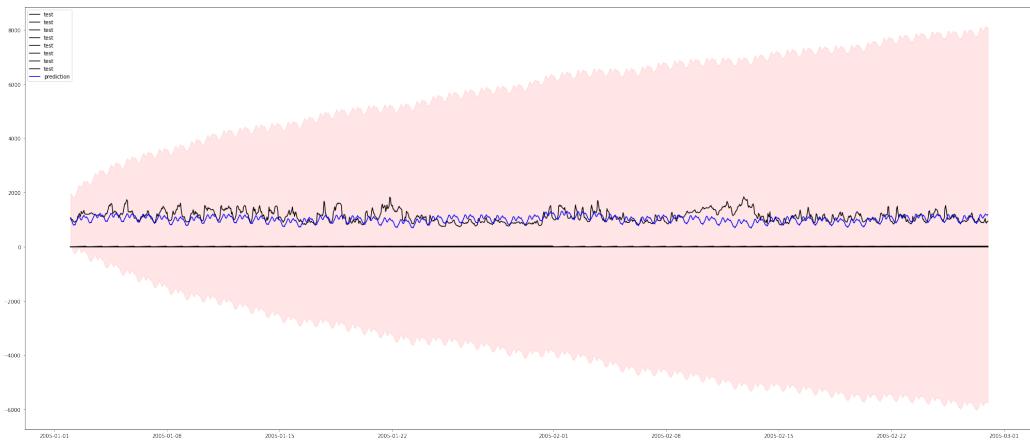


Figura 11: UCM-Local Linear Trend

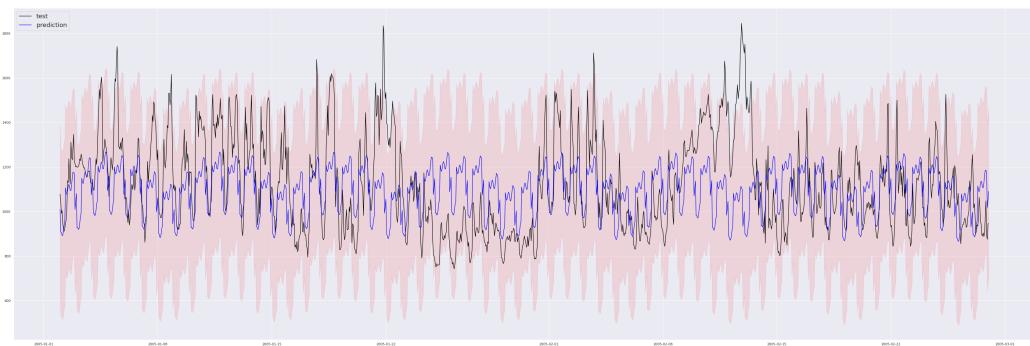


Figura 12: UCM-Local Linear Trend

### 4.3 Modelli non lineari

Per i modelli non lineari sono state utilizzate due architetture diverse dei Recurrent Neural Network:

- **LSTM** (Long ShortTerm Memory)
- **GRU** (Gated Recurrent Unit)

Sono stati scelti questi due modelli poiché permettono l'utilizzo di uno degli strati come memoria di stato, e consentono di modellare un comportamento dinamico temporale dipendente dalle informazioni ricevute agli istanti di tempo precedenti. Prima di iniziare, la serie storica deve essere normalizzata per poi applicare il modello di machine learning, questo serve per modificare i valori in una scala comune, senza distorcere le differenze negli intervalli di valore.

Successivamente, bisogna trasformare il training e validation set in un array di 3 dimensioni: la numerosità del campione, il numero di lags (quantità dei dati che deve passare attraverso le Reti Neurali Ricorrenti, che funziona come una memoria) e il numero di features. Il numero dei lags è stato posto a 24, poichè nei modelli ARIMA abbiamo osservato che i modelli hanno performato bene con tale stagionalità, invece il numero delle features è stato posto a 8 poichè sono state aggiunte delle dummy come: ore, settimane, giorni, mesi, weekend, traffico e mesi primaverili/estivi.

L'archittutura del modello **LSTM** è stata costruita da: un bidirectional layer che ha come argomento il LSTM layer di 128 neuroni, un layer Dense di 24 neuroni la cui funzione di attivazione un ReLu (*Rectified Linear Unit*), un layer di Dropout con tasso di 0.3 per prevenire il problema di overfitting, e come layer finale un Dense. Come ottimizzatore è stato utilizzato Adam con un learning rate di 0.01, e come funzione di perdita il mean absolute error.

Invece, l'archittutura del modello **GRU** è stata costruita da un layer GRU di 128 neuroni, un layer Dense di 24 neuroni la cui funzione di attivazione un ReLu (*Rectified Linear Unit*), un layer di Dropout con tasso di 0.3 per prevenire il problema di overfitting, e layer finale un Dense. Come ottimizzatore è stato utilizzato, come in precedenza, un Adam con un learning rate di 0.01, e come funzione di perdita il mean absolute error.

Tutti e due i modelli sono stati fittati con un batch\_size di 32 e 15 epoche. La scelta del numero di epoche è stata presa dopo aver effettuato un training su

100 epoche con il quale il modello è andato in overfitting, ragion per cui si è preferito lasciarne 15. Per il fitting è stata utilizzata una parte del training e 0.1 del training è utilizzato per la validazione. Il grafico della loss del training e validation in *figura 13* del modello LSTM.

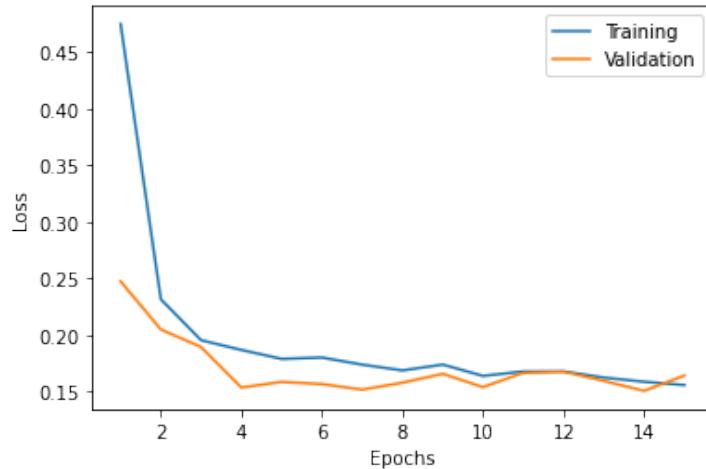


Figura 13: Loss training e validation - LSTM

I modelli sono stati validati con la metrica MAPE e il modello che ha performato meglio è stato il LSTM. In *figura 14* si può osservare il fitting tra la previsione e la validation.

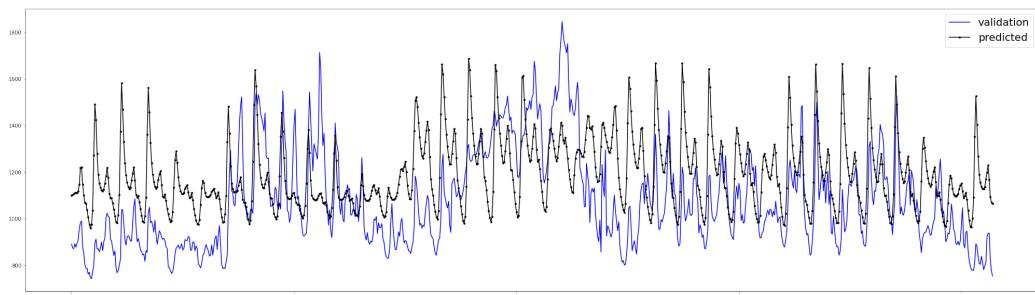


Figura 14: LSTM

## 5 Conclusioni

Per il calcolo della bontà dei modelli è stata utilizzata la metrica MAPE (*mean absolute percentage error*) sui validation set. Tale metrica è stata potuta applicare sui modelli poiché la serie storica non contiene valori nulli o prossimi a zero.

Nella seguente tabella, vengono mostrati i valori del MAPE per i modelli ARIMA sul validation set:

Modelli ARIMA	MAPE
<i>SARIMA(4,0,1)(1,1,1)<sub>24</sub></i>	13.61
<i>ARIMAX(4,0,1) regressori sinusoidali e dummy stocastiche</i>	14.45
<i>SARIMAX(4,0,1)(1,0,1)<sub>24</sub> regressori sinusoidali e dummy stocastiche</i>	14.28

Dal MAPE viene di nuovo confermato che il modello migliore per gli ARIMA è SARIMAX(4,0,1)(1,0,1)<sub>24</sub> regressori sinusoidali e dummy stocastiche.

Le performance dei modelli UCM, invece, hanno ottenuto i seguenti risultati:

Modelli UCM	MAPE
<i>Local Linear Trend</i>	14.32
<i>Random Walk</i>	14.41
<i>Random Walk with Drift</i>	14.33

Il migliore modello è risultato essere il Local Linear Trend. Ma come specificato precedentemente il modello utilizzato per la previsione è il Deterministic Trend.

Per quanto riguarda invece i modelli non lineari, si hanno avuto i seguenti risultati:

Modelli non lineari	MAPE
<i>LSTM</i>	18.80
<i>GRU</i>	32.55

Nel modelli non lineari la miglior performance è stata del LSTM. Selezionando i modelli migliori per ogni categoria, questi sono stati allenati su tutta la serie storica, per poi predire il periodo dal 1° Marzo 2005 alle ore 00 al 31° Marzo 2005 alle ore 23. Di seguito sono rappresentati le previsioni ottenute (*figura 15*).

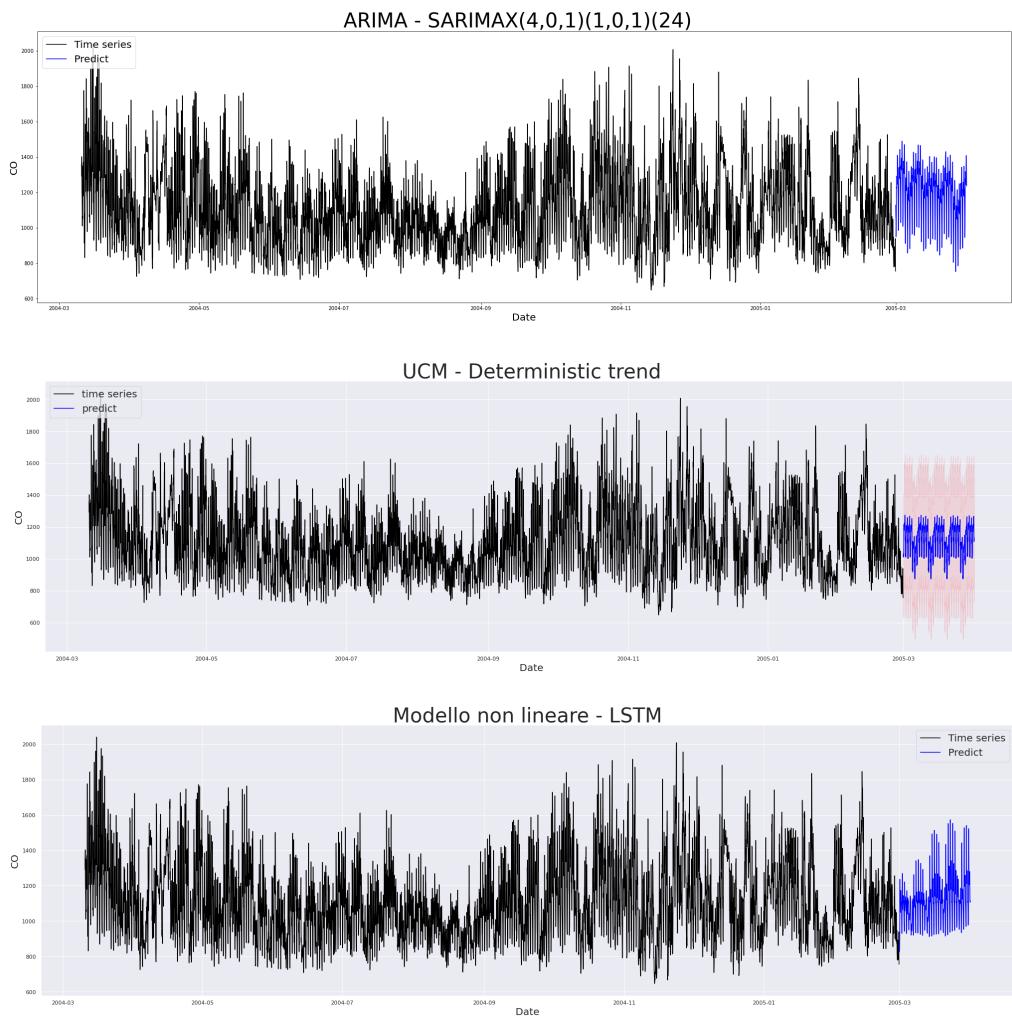


Figura 15: Predizione dei 3 migliori modelli