UNIVERSITÀ DEGLI STUDI DI
MILANO-BICOCCA

CORSO DI LAUREA MAGISTRALE IN
DATA SCIENCE

STREAMING DATA MANAGEMENT AND TIME SERIES ANALYSIS
PROGETTO FINALE

Previsioni di misurazioni orarie di ossido di carbonio (CO)

Autore
Claudio Fadda – 813499 – c.fadda1@campus.unimib.it

A.A. 2021/2022

Sommario

1. Abstract

L'obiettivo di questo report è andare a predire, attraverso l'uso di algoritmi della famiglia ARIMA (Autoregressive Integrated Moving Average), UCM (Unobserved Components Model) e ML (Machine Learning), i valori di emissioni di monossido di carbonio (CO). L'ossido di carbonio è un gas che viene prodotto dalla non totale combustione da parte degli idrocarburi presenti nei carburanti e combustibili in generale; la sua fonte primaria è collegata al traffico stradale, ma anche all'uso dei riscaldamenti durante il periodo invernale. Il tempo di permanenza nell'atmosfera (ma in generale anche nell'aria che respiriamo) può essere molto lungo, influiscono in questo senso gli agenti meteorologici (pioggia e vento in modo particolare) che permettono di ripulire l'aria e abbassare di conseguenza le concentrazioni.

---

2. Introduzione

Il dataset a disposizione è costituito da tre colonne "Date", "Hour", "CO" che rispettivamente vanno a identificare la data nel formato (*yyyy-mm-dd*), l'ora del giorno (da 0 a 23) e il valore di CO registrato; si tratta quindi di una serie storica oraria con le relative misurazioni di CO. I dati a disposizione partono dal 10 marzo 2004 ore 18, fino al 28 febbraio 2005 ore 23 e si vuole andare a prevedere un mese di dati: dalle ore 00 del 1° marzo 2005 fino alle ore 23 del 31 marzo dello stesso anno.

3. Preprocessing

Sono stati adottati degli accorgimenti per risolvere il problema dei dati mancanti (su 8526 osservazioni, 365 risultavano nulle). Essendo una serie storica molto variabile, dato che come abbiamo detto i valori di CO possono dipendere da tanti fattori in gioco si è deciso per ogni singolo valore nullo di sostituirlo facendo una media tra il valore sette giorni prima e sette giorni dopo alla medesima ora; nel caso invece uno dei due non fosse disponibile, il valore mancante veniva rimpiazzato attraverso la copia dello stesso valore sette giorni prima o sette giorni dopo della stessa ora.

Inoltre, il dataset è stato diviso in training e validation set, rispettivamente con 7110 (dal 10 marzo 2004 ore 18, al 31 dicembre 2004 ore 23) e 1416 osservazioni (dal primo gennaio 2005 ore 00, al 28 febbraio 2005 ore 23).

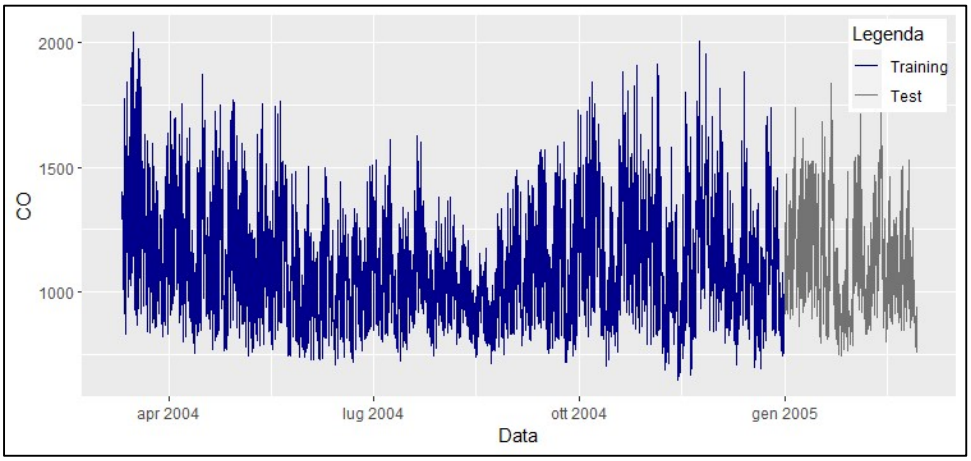

Fig. 1 - Serie storica con la divisione train e validation set.

Dalla *Fig.1* possiamo osservare notevoli variazioni nel corso di un anno, notiamo un abbassamento delle concentrazioni nei mesi estivi, soprattutto durante il mese di agosto, presumibilmente dovuto al calo del traffico automobilistico per via delle vacanze estive.

Se ci focalizziamo su due settimane centrali di marzo 2004 (*Fig.2*), ci accorgiamo della presenza di due tipi di stagionalità:

- Giornaliera, si ripete ogni 24 ore (24 osservazioni) e ha generalmente un aumento verso l'alto nelle prima ore della giornata, un momento di stasi nelle ore pomeridiane, per poi avere un andamento crescente la sera e un picco verso il basso nelle ore notturne.
- Settimanale, si ripete ogni 168 ore (24 osservazioni ogni giorno per 7 giorni). In questo caso vediamo valori alti per i primi giorni della settimana, con uno smorzamento negli ultimi giorni feriali fino al raggiungimento del week-end nel quale vengono registrati i valori più bassi.

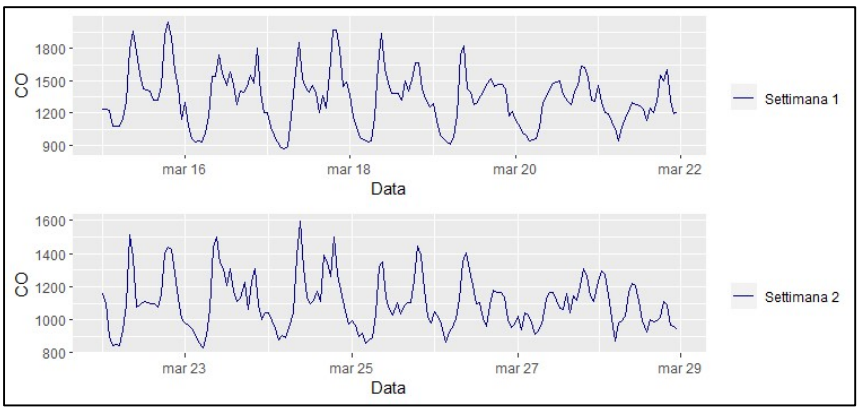

Fig. 2 - Andamento di due settimane centrali del mese di marzo 2004.

Da ricordare che questo è un andamento generale, che non è sempre costante: festività, ponti festivi e meteo influenzano in modo indiretto le emissioni di CO portando ad andamenti diversificati.

4. Metodologia usata

Le performance di predizione relative ai diversi modelli costruiti per questo studio sono state valutate attraverso il Mean Absolute Percentage Error (MAPE).

5. ARIMA

Non rilevando una particolare non stazionarietà in varianza ci si concentra direttamente sull'autocorrelogramma (*Fig.3*). L'ACF mostra una discesa graduale verso 0 e dei picchi. I picchi più pronunciati sono quelli che si ripetono ogni 24 osservazioni (24 ore), i picchi meno intensi si ripetono ogni 12 osservazioni. Possiamo trascurare i picchi più lievi, dovuti alle differenze tra mattina-pomeriggio-sera delle emissioni di CO; molto più importante è invece la periodicità ogni 24 ore. L'ACF quindi permette inizialmente di stabilire una non stazionarietà in media stagionale con lag stagionali multipli di 24. Il PACF mostra sempre dei lag ogni 24.

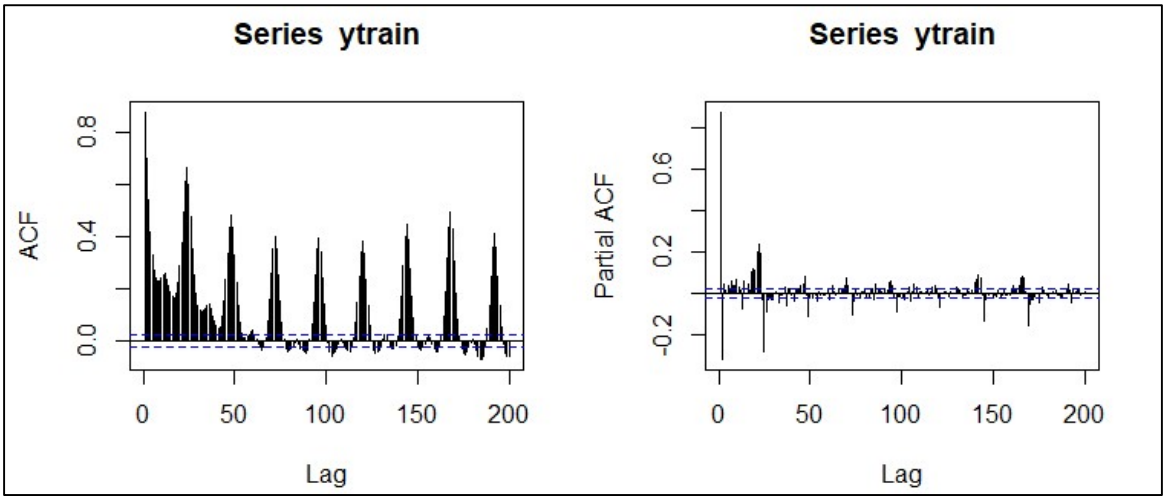

FIg. 3 - ACF e PACF della serie originale.

Sui lag successivi notiamo anche un picco più alto rispetto agli altri (al 168° lag), questo è sintomo di una stagionalità ogni 168 osservazioni, quindi settimanale.

Inizialmente ci si concentra sulla stagionalità ogni 24 ore che è quella più visibile.

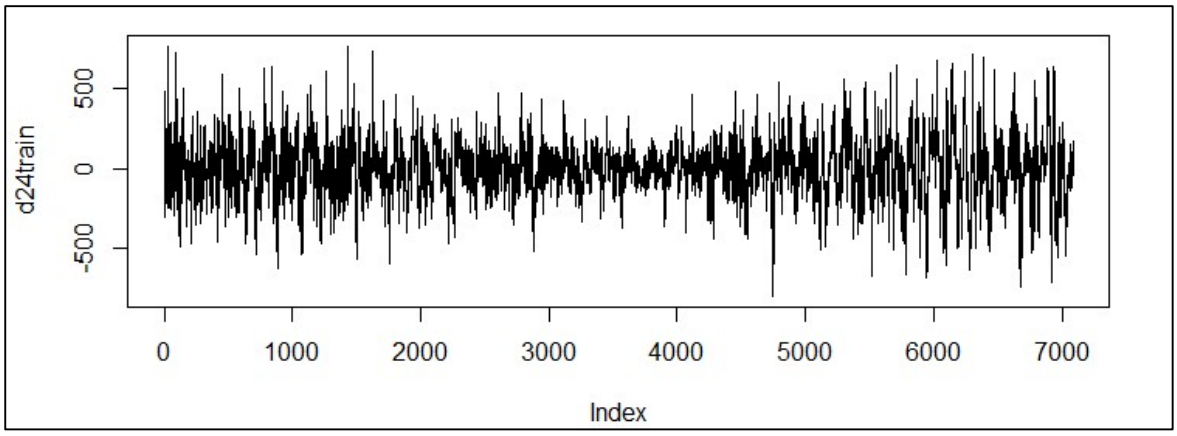

Fig. 4 - Differenziazione stagionale delle serie storica.

Dopo la differenziazione stagionale con un periodo di 24 la serie risulta stazionaria in media (*Fig.4*), si procede quindi con un primo modello $SARIMA(0,0,0)(1,1,1)_{24}$

Osservando il correlogramma di questo primo modello (*Fig.5*) si nota l'ACF scendere rapidamente a 0; nel PACF si notano dei lag significativi nel primo ritardo e ogni 24 osservazioni, sinonimo che la periodicità ogni 24 ore non è stata del tutto risolta attraverso la differenziazione.

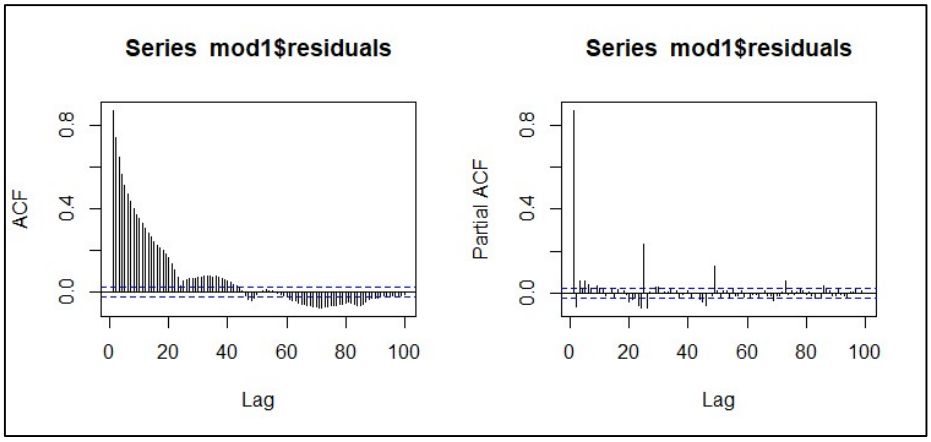

FIg. 5 - Correlogramma SARIMA(0,0,0)(1,1,1)$_{24}$.

Si cerca di modificare il modello andando a stimare un modello del tipo $SARIMA(1,0,1)(1,1,1)_{24}$ andando ad aumentare l'ordine della componente autoregressiva. Non ottenendo buoni risultati, si preferisce mantenere il modello parsimonioso e quindi confermare un modello $SARIMA(2,0,1)(1,1,1)_{24}$ che ottiene un valore di AICc più basso rispetto a tutti quelli precedentemente stimati.

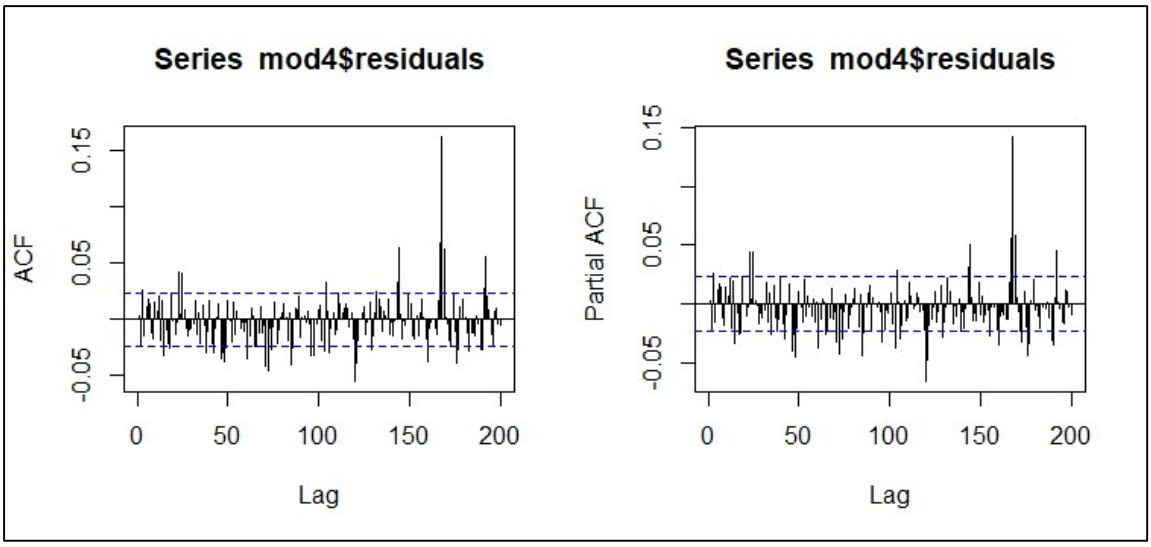

Fig. 6 - ACF e PACF relativo a SARIMA(2,0,1)(1,1,1)$_{24}$.

Analizzando il correlogramma di *Fig.6* vediamo come tutti i lag iniziali con quest'ultimo modello siano più bassi, anche se diversi lag escono fuori dalle bande. Notiamo un lag sia nell'ACF che nel PACF che spunta fuori a ritardo 168, questo ci fa pensare che non è stata del tutto risolta la stagionalità ogni 168 osservazioni (settimanale). Si procede quindi con l'aggiunta di regressori deterministici al nostro modello.

In particolare, vengono aggiunte delle sinusoidi deterministiche, nello specifico sei sinusoidi con frequenza pari a $\frac{2\pi}{168}$.

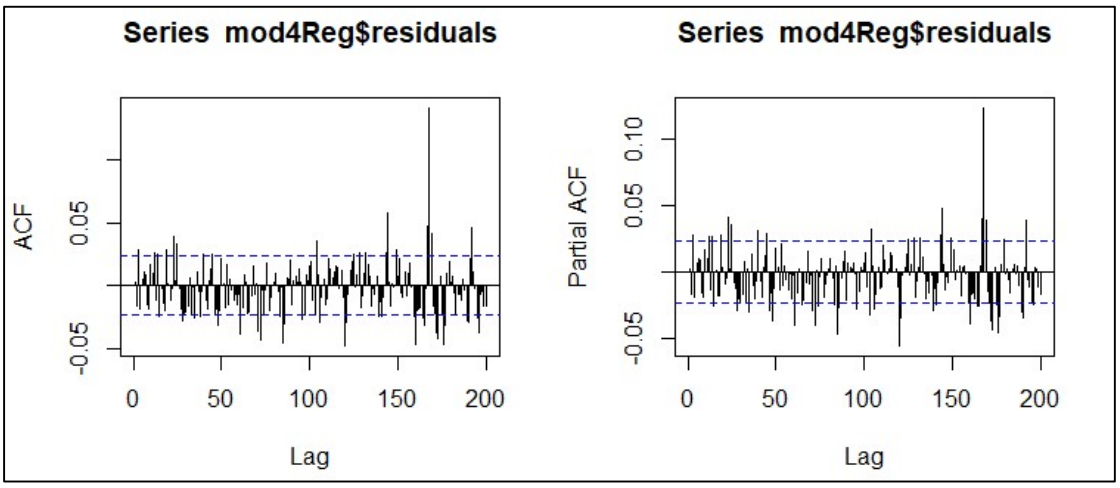

Fig. 7- ACF e PACF relativo a SARIMA(2,0,1)(1,1,1)$_{24}$ con l'aggiunta di regressori deterministici.

Il modello, nonostante nessun miglioramento apparente (*Fig. 7*), ha ottenuto un AICc migliore (passando da AICc=82167.74 (senza regressori) a AICc=82013.87 (con l'aggiunta di regressori)), inoltre il lag a 168 ha ora un valore più basso. Si decide quindi di confermare questo modello e testarlo sul validation set (*Fig.8*).

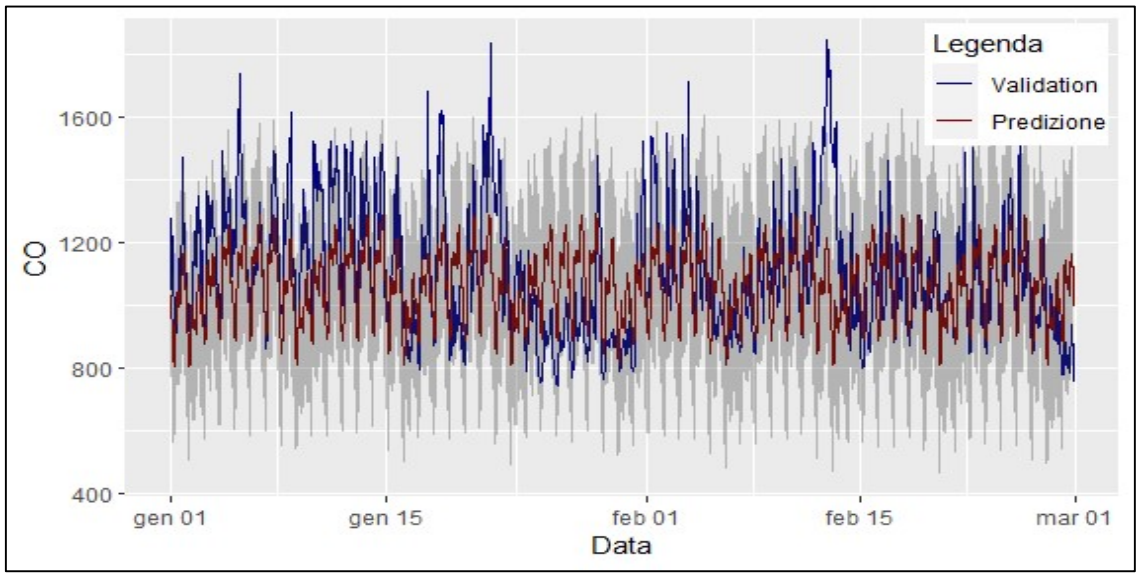

Fig. 8 - Previsione del modello SARIMA(2,0,1)(1,1,1)$_{24}$ sul validation set.

In termini di MAPE, la previsione sul validation set si attesta su un valore paria a 12.62.

Infine, andiamo a fare un plot dei residui visibile in (*Fig.9*).

Possiamo affermare che la media sia nulla, nonostante ci siano dei residui che assumono valori abbastanza alti. Per quanto riguarda il grafico dell'ACF, ci mostra che ancora è presente della memoria; mentre la normalità sembra essere verificata. Bisogna comunque aggiungere che essendo questi dei dati reali, è complesso riuscire ad ottenere dei residui che siano totalmente dei white noise.

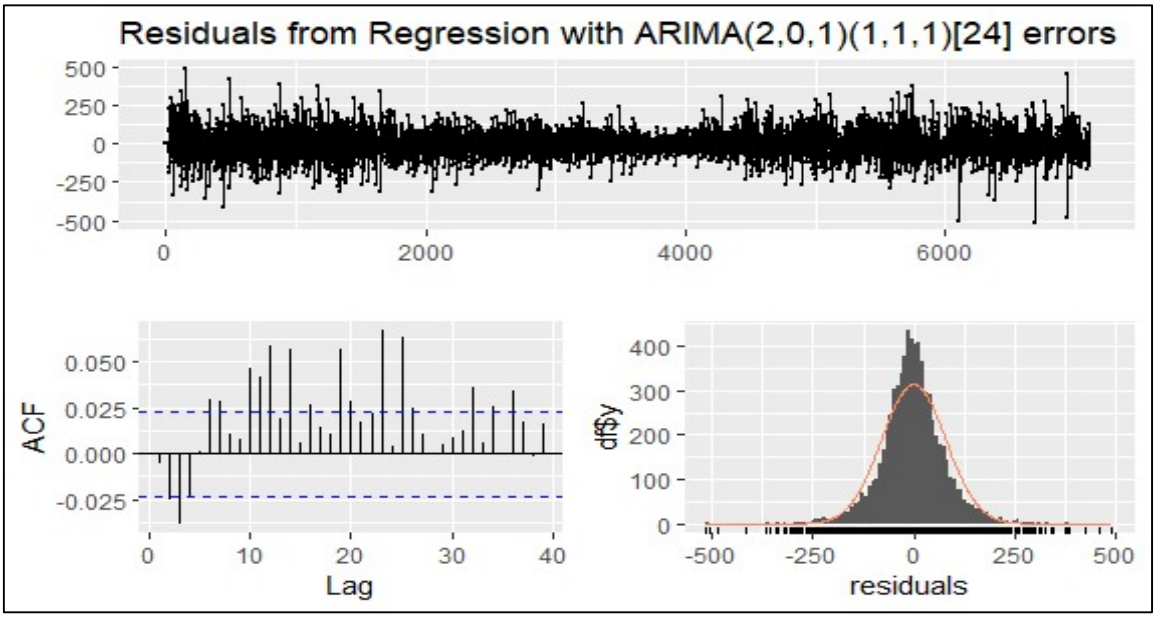

Fig. 9 - Residui del modello SARIMA(2,0,1)(1,1,1)$_{24}$.

6. UCM

I modelli UCM permettono di scomporre una serie storica additivamente nelle sue componenti come il trend, la stagionalità, il ciclo e una componente di errore, tutte queste componenti possono essere definite in modo stocastico.

La stima del trend è stata effettuata attraverso il local linear trend, random walk e random walk integrato. In questi modelli non sono stati usati dei regressori esterni, ma le stagionalità ogni 24 ore e quella settimanale, sono state modellate rispettivamente attraverso l'uso di dummy e sinusoidi stocastiche. In termini di MAPE le migliori performance sono state raggiunte dal modello con la componente random walk per un valore uguale a 13.09.

Il modello è stato poi testato sul validation set con la previsione visibile in *Fig.10*.

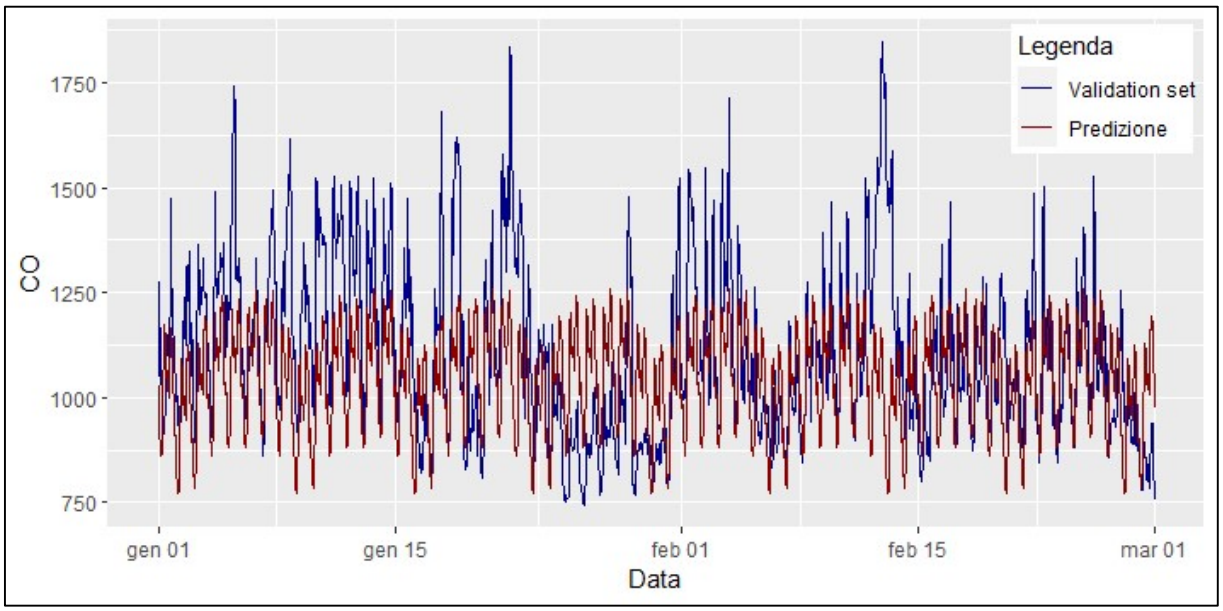

Fig. 10 - Previsione del modello UCM sul validation set.

7. Machine Learning

Per quanto riguarda l'area del machine learning, è stato scelto di predire i valori di CO tramite il k-Nearest Neighbours (k-NN) e tramite le Recurrent Neural Network (rNN).

k-NN

Il k-NN è un metodo model-free, l'idea si basa sul fatto che se il futuro è in qualche modo correlato con il passato allora si possono cercare delle k sotto-sequenze simili al presente e usarle per predire il futuro. Questo metodo fa uso di iper-parametri che devono essere impostati all'inizio: p, rappresenta il numero di osservazioni del passato che vogliamo tenere in considerazione; k, è il parametro di tuning, permette di cercare k sotto-sequenze più simili alla "query" data in input; h, è l'orizzonte previsionale e quindi specifica quante osservazioni prevedere nel futuro.

I parametri p e k sono strettamente legati tra loro: maggiore è p, minor è il numero k che possiamo trovare. In questo caso tenendo conto del fatto di avere una serie temporale minore di un anno, si è deciso di avere p pari a 744 osservazioni (un mese) in modo da non essere troppo legati alle ultime osservazioni che potrebbero essere fuorvianti per predire il futuro (una settimana, per esempio, potrebbe avere basse concentrazioni di CO per eventuali giorni di festa, ponti festivi e così via

quindi si vuole evitare di prenderla come riferimento). Inoltre, si decide di utilizzare la metodologia Multi-Input Multi-Output (MIMO), e di utilizzare la mediana come funzione per aggregare le k sotto-sequenze trovate dall'algoritmo.

Per quanto riguarda la scelta di k, considerando che p è stato preso pari a 744, si è deciso di scegliere l'iper-parametro iterando il modello diverse volte con gli iper-parametri facendo variare il parametro k e andando a calcolare per ogni iterazione il MAPE rispetto al validation set.

Dalla Fig.11 si può notare che il valore di MAPE più basso è associato ad un k pari a 8.

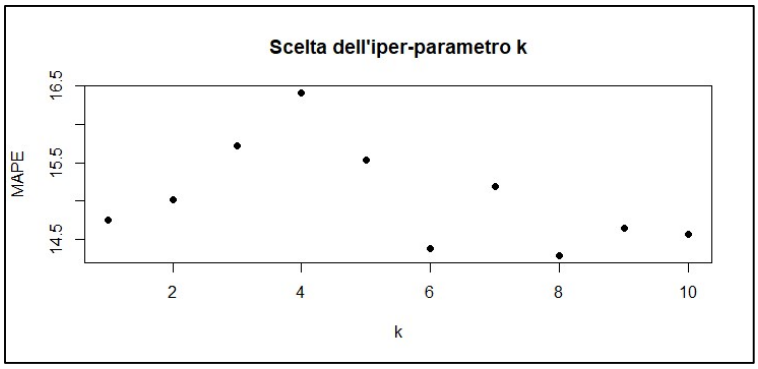

Fig.11 Scelta di k.

Si riportano nella Fig.12 le previsioni ottenute sul validation set.

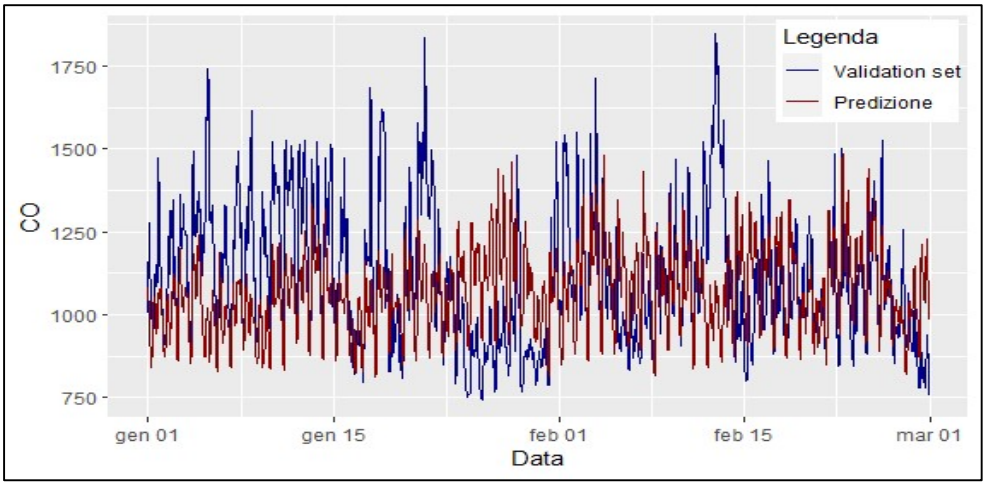

Fig. 12 – Previsione sul validation set.

Il modello k-NN ha ottenuto un MAPE paria a 14.28.

rNN

La rete neurale ricorrente mantiene memoria di ciò che ha visto fino all'ultimo dato e in base a ciò produrrà un output opportuno.

I dati, prima di essere usati dal modello, sono stati scalati e poi centrati; per quanto riguarda il modello invece è stata sviluppata una architettura LSTM (Long Short-Term Memory) con due layers LSTM da 100 e 50 neuroni, un numero di lags pari a 168 e un batch size sempre di 168, è stato inserito anche uno strato di dropout con un rate pari a 0.5. Inoltre, il modello è stato allenato per cento epoche, come ottimizzatore è stato utilizzato Adam e come funzione di perdita si è scelto di usare il mean absolute error.

In questo caso si è scelto di andare a prevedere un passo alla volta, di seguito la previsione sul validation set (*Fig.13*).

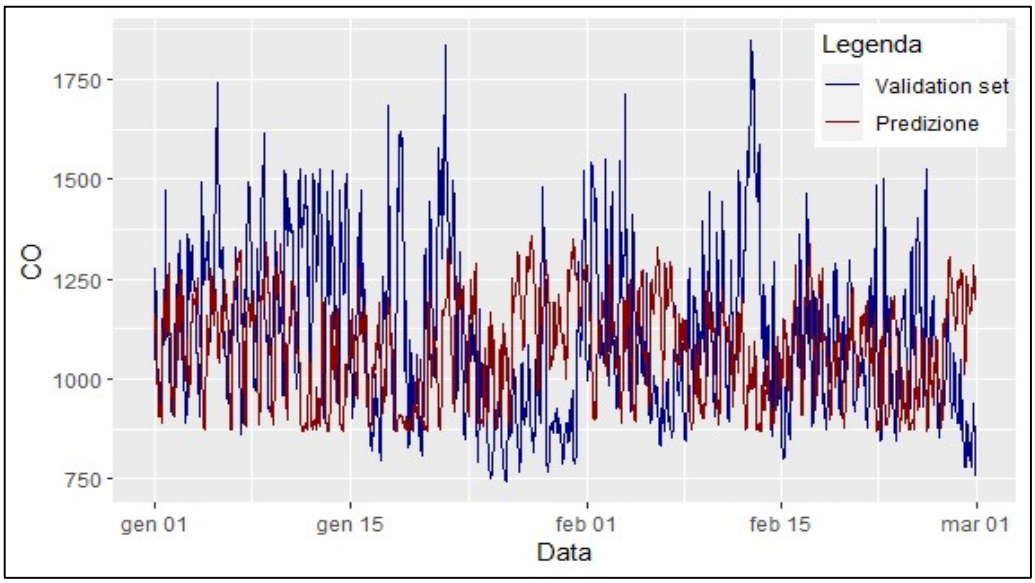

Fig. 13 – Previsione sul validation set.

In termini di MAPE il modello ha ottenuto un valore di 18.53.

8. Conclusioni

I modelli ARIMA, UCM e ML che sono stati validati sul validation e che hanno ottenuto il valore migliore di MAPE, sono stati infine allenati su tutto il dataset di 8526 osservazioni ed è stato così predetto il mese di marzo 2005. Di seguito le previsioni ottenute rispettivamente per il modello $SARIMA(2,0,1)(1,1,1)_{24}$ (*Fig.14*), il modello UCM (*Fig.15*) e il modello k-NN (*Fig.16*).

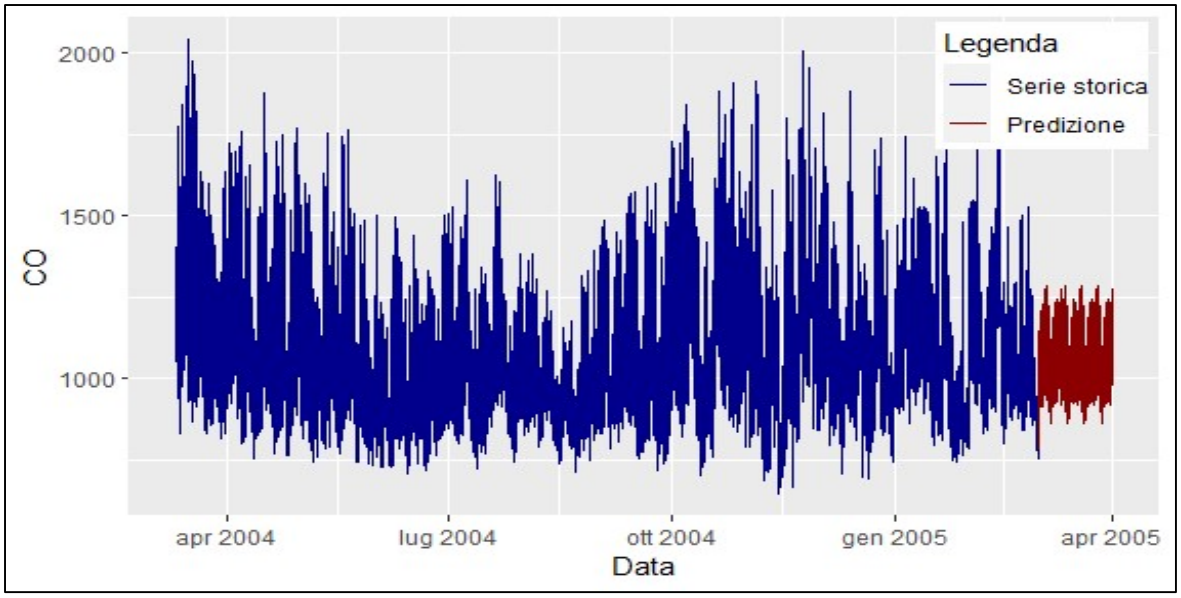

Fig. 14 – Previsione marzo 2005 modello SARIMA(2,0,1)(1,1,1)$_{24}$.

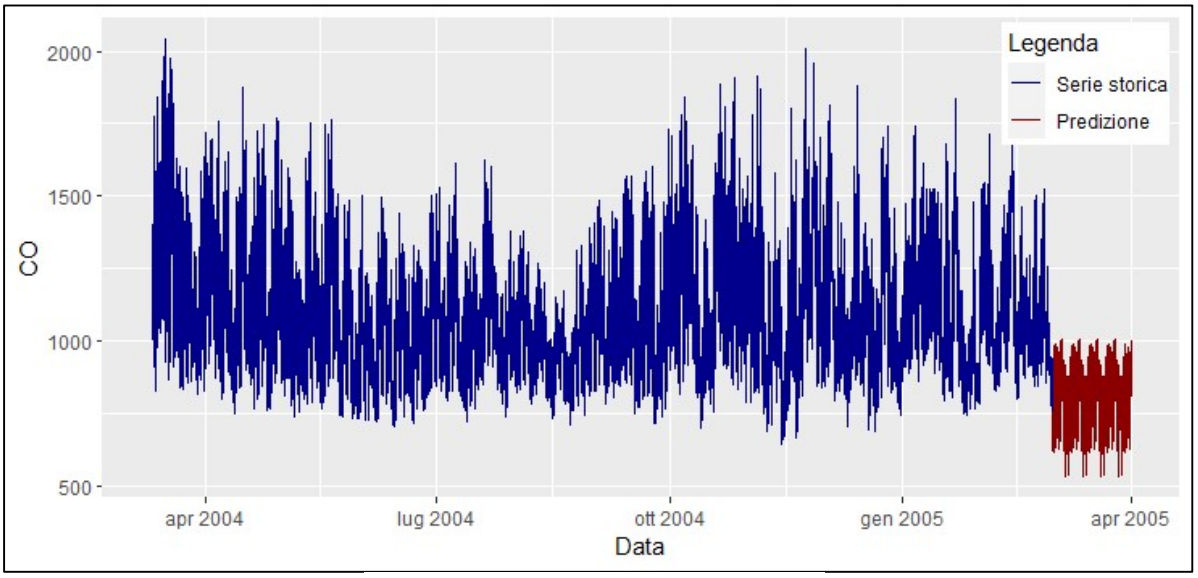

Fig. 15 – Previsione marzo 2005 modello UCM.

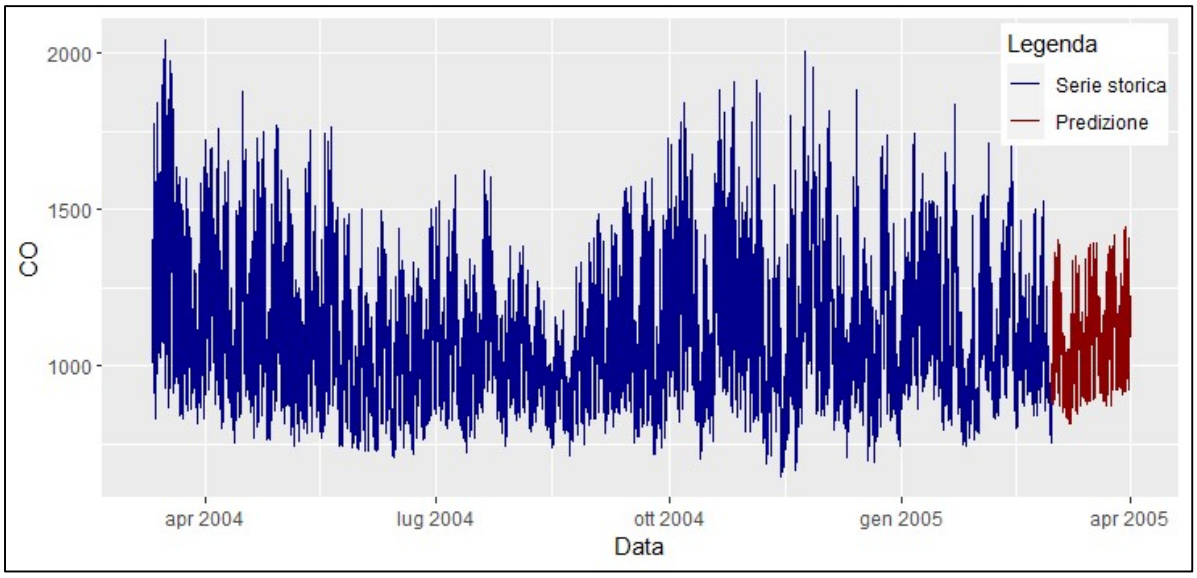

Fig.16 – Previsioni modello k-NN.