Laurea Magistrale in
DATASCIENCE

PROGETTO STREAMING DATA MANAGMENT & TIME SERIES ANALYSIS

Serie storica del CO

Francesco Fustini, 830697
Corso di Laurea Magistrale in Data Science - UniMiB

Data:
7 febbraio 2022

Abstract: L'elaborato espone diversi approcci utilizzati per fare previsioni su una serie storica rappresentate le rilevazioni di monossido di carbonio (CO). I dati coprono a livello orario il periodo che parte dalle ore 18:00 del 10 marzo 2004 alle ore 23:00 del 28 febbraio 2005. Le previsioni effettuate sono su tutto il mese successivo. Per lo scopo sono stati provati diversi modelli sia lineari (ARIMA, UCM) che di machine learning (k-NN, LSTM, GRU), per la scelta del modello la metrica utilizzata è il Mean Absolute Percentage Error.

Key-words: Serie storica, previsioni, modelli lineari, machine learning

1. Analisi Esplorativa

Il file analizzato si compone come una serie storica univariata con 8526 osservazioni, ogni osservazione è caratterizzata dalle variabili: *Date*, *Hour* e *CO*. I dati sono rilevati a livello orario a partire dal 10 marzo 2004 alle ore 18:00 fino al 28 febbraio 2005 alle ore 23:00. Non sono presenti cambi di ora legale/solare.

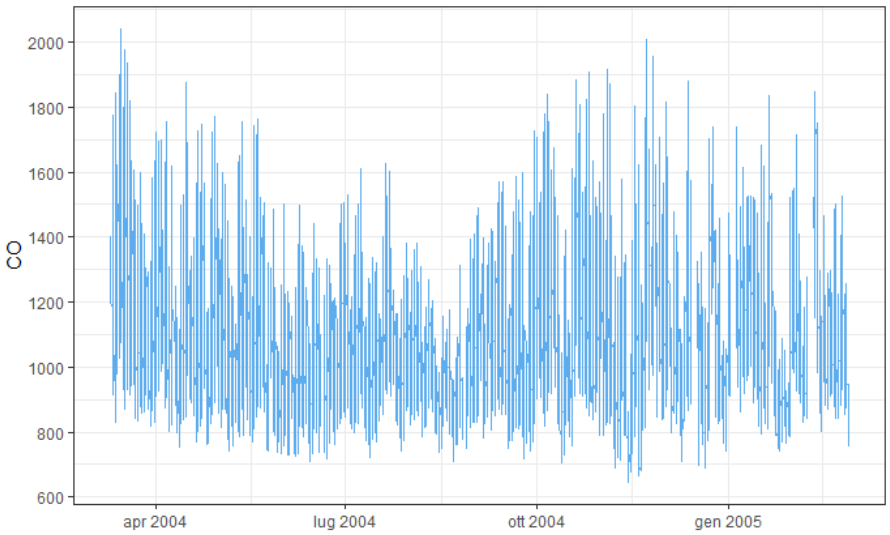
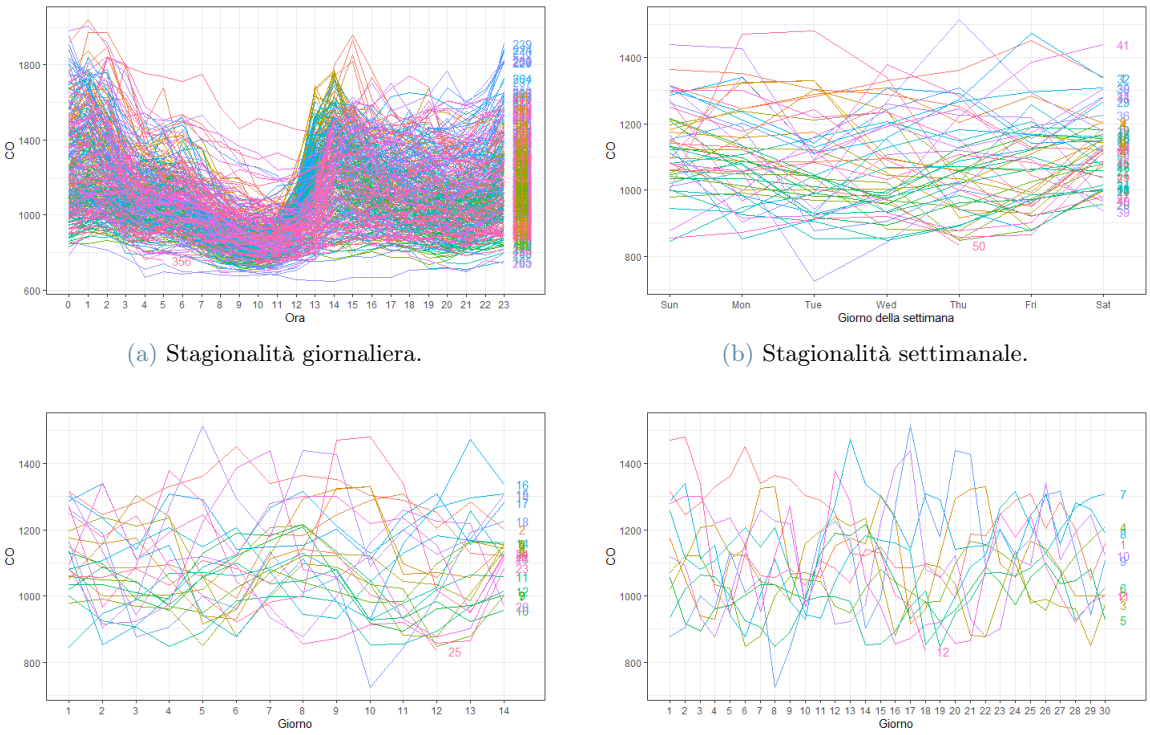


Figura 1: Grafico serie storica.

La variabile *CO* assume valori compresi tra 647 e 2040 con media 1097 e mediana 2059. Inoltre sono presenti 365 valori mancanti.

1.1. Stagionalità



(a) Stagionalità giornaliera.

(b) Stagionalità settimanale.

(c) Stagionalità bisettimanale.

(d) Stagionalità 4 settimane.

Figura 2: Grafici della stagionalità.

L'analisi grafica della stagionalità (Figura 2) diagnostica una stagionalità giornaliera mentre esclude quella settimanale, bisettimanale e di 4 settimane.

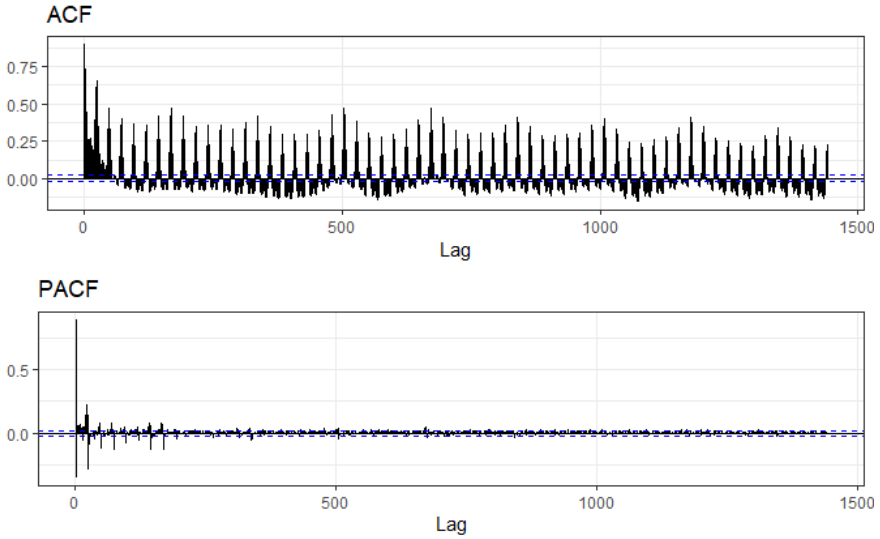


Figura 3: grafici ACF e PACF.

Il grafico dell'autocorrelazione parziale (Figura 3) invece sembra segnalare una presenza di stagionalità settimanale, verrà tenuta di conto nelle prossime analisi.

1.2. Stazionarietà

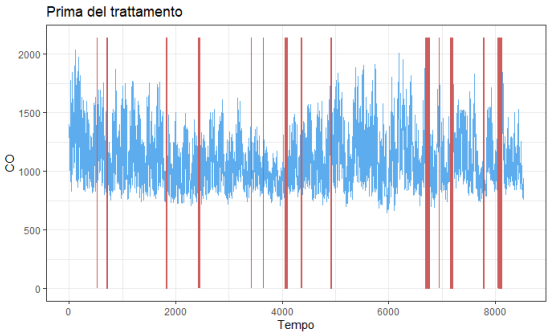
La stazionarietà in media è stata verificata tramite il test di **Dickey-Fuller aumentato** per le radici unitarie.

La diagnostica grafica (Figura 1) rileva presenza di non stazionarietà in varianza. Per risolvere potrebbe servire una trasformazione.

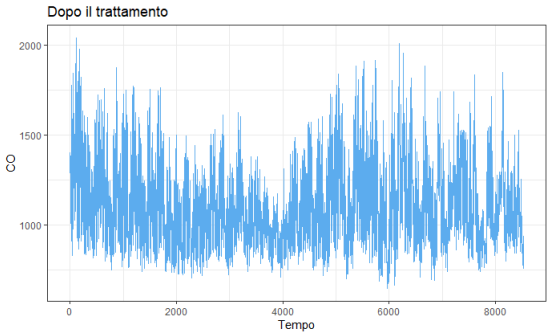
2. Pre-Processing

2.1. Valori Mancanti

Come già anticipato dall'analisi esplorativa sono presenti 365 valori mancanti sulla variabile *CO* pari al 4.28% delle osservazioni totali. La tecnica scelta per trattarli è l'imputazione della media settimanale a livello orario. Quindi ad esempio il valore imputato alla variabile *CO* del giorno 1 aprile 2004 alle ore 14 è la media dei valori alle ore 14 dei 3 giorni prima: 29 marzo, 30 marzo e 31 marzo e 3 giorni dopo: 2 aprile, 3 aprile e 4 aprile.
La media settimanale serve per catturare i trend locali, i 7 giorni sono anche condizionati dalla presenza (anche se non elevata) della stagionalità settimanale. Inoltre la scelta di non fare la media tra valori contigui ma tra osservazioni con stesso orario è stata fatta dopo la diagnostica di stagionalità giornaliera.



(a) Distribuzione NA pre trattamento.



(b) Serie storica post trattamento.

2.2. Trasformazioni

Per quanto riguarda i modelli lineari per risolvere la non stazionarietà in varianza viene applicata una trasformazione **Box-Cox** ai dati:

$$f(x, \lambda) = sign(x)\frac{|x|^{\lambda} - 1}{\lambda}$$

Il λ trovato è -0.89.

Alle le reti neurali alla serie viene applicata una normalizzazione:

$$\hat{x} = \frac{x - min(x)}{max(x) - min(x)}$$

3. Modelli

Al fine di verificare le bontà dei modelli è stato diviso il dataset in Train e Validation set con di percentuale 80% e 20%.
Il parametro su cui ottimizzare i modelli è il **Mean Absolute Percentage Error (MAPE)**, calcolato nel seguente modo:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}|\frac{x_i - \hat{x}_i}{x_i}|$$

dove:

- n è il numero di osservazioni
- x sono i valori reali

- $\hat{x}$ sono i valori previsti

Per ogni famiglia di modelli (**ARIMA**, **UCM** e **ML**) viene scelto quello con **MAPE** minore e viene addestrato su tutto il dataset per prevedere tutto il mese di marzo 2005.

3.1. ARIMA

La prima famiglia di modelli lineari usata è quella degli **ARIMA**. Nel particolare sono stati testati i seguenti modelli:

- ARIMA$(2,0,0)(0,1,0)_{24}$
- ARIMA$(5,0,1)(2,1,0)_{24}$
- ARIMA$(1,0,0)(4,1,0)_{24}$

I valori dei parametri per il primo modello sono ottenuti osservando il grafico ACF e PACF (Figura 3), invece il secondo e il terzo modello sono ottenuti tramite un test con su più parametri.

Modello	**MAPE**
ARIMA$(2,0,0)(0,1,0)_{24}$	15.91%
ARIMA$(5,0,1)(2,1,0)_{24}$	13.52%
ARIMA$(1,0,0)(4,1,0)_{24}$	13.61%

Tabella 1: Confronto MAPE per modelli ARIMA.

Il modello scelto della famiglia **ARIMA** per prevedere il mese di marzo 2005 è l'ARIMA$(5,0,1)(2,1,0)_{24}$
Inoltre possiamo affermare l'incorrelazione dei residui dopo aver effettuato i test di **Ljung-Box** e **Box-Pierce**.

3.2. UCM

L'altra famiglia di modelli lineari presa in considerazione è quella degli **UCM**. I modelli presi in considerazione sono i seguenti:

- Local Linear Trend
- Local Linear Trend con stagionalità trigonometrica
- Local Linear Trend con stagionalità dummy
- Local Linear Trend con stagionalità dummy e ciclo settimanale
- Local Linear Trend con stagionalità dummy e ciclo di 4 settimane
- Random Walk con stagionalità dummy

Le componenti degli **UCM** nel formato state space sono la componente Trend(μ), la componente ciclo(ψ) e la componente stagionale(γ) giornaliera.

Componente			
μ	ψ	γ	**MAPE**
LLT			15.96%
LLT		dummy	12.62%
LLT		trigonometrica	12.54%
LLT	settimanale	dummy	15.08%
LLT	4 settimane	dummy	13.53%
RW		dummy	12.67%

Tabella 2: Confronto MAPE per modelli UCM.

In questo caso il modello scelto è il Local Linear Trend con stagionalità trigonometrica.

3.3. Machine Learning

Sono state testate due tipologie di modelli di machine learning (**ML**): il k-Nearest Neighbors (**k-NN**) e la Recurrent Neural Network (**RNN**). Della famiglia **RNN** nel particolare sono stati testati il Long Short-Term Memory (**LSTM**) e il Gated Recurrent Unit (**GRU**).

L'algoritmo **k-NN** applicato alla previsione nelle serie storiche è parametrizzato da k e p. Il parametro p indica quante osservazioni passate usare per fare la previsione e k quanti gruppi di osservazioni uguali alle p passate considerare.
Sono stati provati diversi valori di k e p. Di seguito i migliori risultati.

p	k	MAPE
24	25	31.91%
168	2	13.44%
168	7	13.12%
168	8	17.43%
672	3	13.52%
672	6	13.58%

Tabella 3: Confronto MAPE per modelli kNN.

Il miglior risultato è dato dal modello con $p = 168$ (settimanale) e $k = 7$.

Gli algoritmi **RNN** invece richiedono una scelta sui *neuroni* e *look back*. Il numero di *neuroni* ottimale è quello di 512 mentre per il *look back* ne sono stati trovati in particolare due buoni: 24 e 168.

Modello	look back	MAPE
LSTM	24	14.36%
	168	12.02%
GRU	24	14.86%
	168	17.43%

Tabella 4: Confronto MAPE per modelli RNN.

Il modello con **MAPE** minore è l'**LSTM** con *look back* = 168

4. Limit

Le previsioni effettuate hanno un problema, infatti la presenza di quasi il 5% di valori mancanti potrebbe compromettere la stima dei modelli.
Inoltre una serie storica più lunga e la presenza di altre covariate significative migliorerebbero ulteriormente le performance.

5. Conclusione

I modelli scelti per effettuare la previsione a livello orario sul mese di marzo 2005 sono: ARIMA$(5,0,1)(2,1,0)_{24}$, Local Linear Trend con stagionalità trigonometrica e LSTM con 512 neuroni e 168 (24x7) di look back. Di questi sul validation set il migliore sembra essere il modello di **ML** con **MAPE** al 12.02% contro il 13.52% dell'**ARIMA** e il 12.54% dell'**UCM**.