UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Progetto di Streaming Data Mangement and Time Series

Andrea Cattaneo, 815585

Indice:

Abstract

Il progetto consiste in un task di previsione di una serie temporale univariata relativa alle rilevazioni orarie di monossido di carbonio (CO). La previsione è relativa al periodo dal 01/03/2005 00:00 al 31/03/2005 23:00. Per effettuarla verranno selezionati i 3 modelli migliori da tre diverse categorie di algoritmi, gli ARIMA, gli UCM e i modelli di Machine Learning. Il miglior algoritmo per ciascuna categoria verrà selezionato in base alle sue performance predittive misurate con la metrica MAPE.

Data exploration e preprocessing:

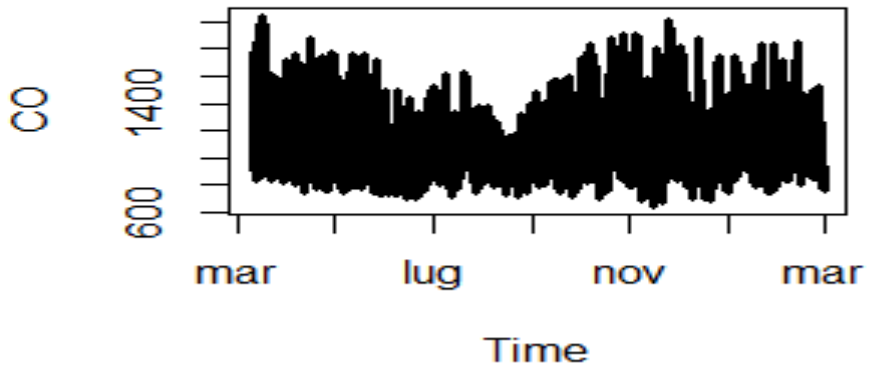
Il dataset si compone di 8526 osservazioni e 3 colonne:

1. “Date” campo di tipo stringa che contiene la data in formato “yyyy-mm-dd”
2. “Hour” campo di tipo stringa che contiene l’ora e assume valori nel range di interi [0-23]
3. “CO” campo di tipo numeric che contiene la quantità di monossido di carbonio rilevata nell’aria.

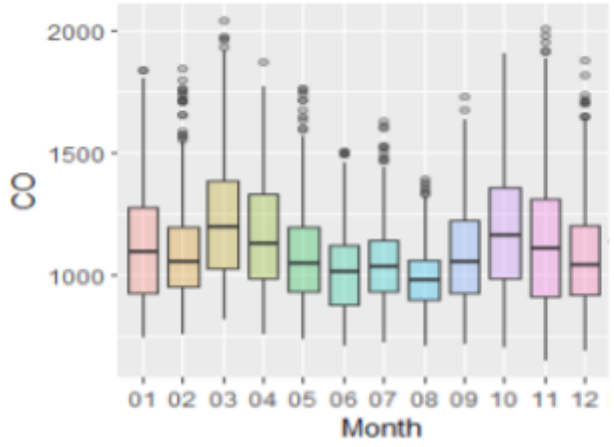
I dati a disposizione vanno dalle 18.00 del 10 marzo 2004 fino alle 23.00 del 28 febbraio 2005 e le rilevazioni hanno frequenza oraria. Non è considerato il cambio dell’ora (legale/solare).

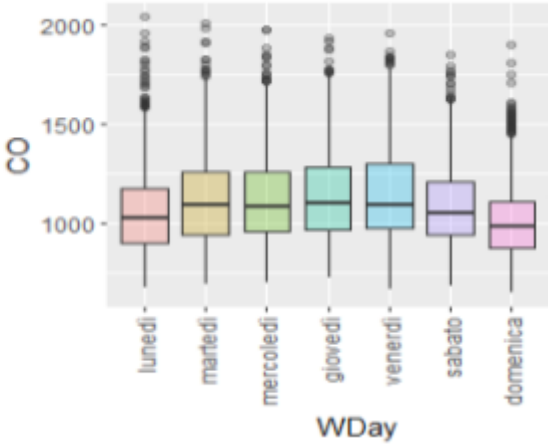
Per prima cosa si è creato un campo “Time” di tipo POSIXct che unisce “Date” e “Hour” in un unico campo di tipo datetime. Dopodiché si è controllata l’eventuale presenza di valori nulli nei dati, rilevandone solo all’interno del campo “CO” che ne presentava 365. A questo punto, prima di decidere come rimpiazzare questi valori nulli è stata effettuata una prima esplorazione degli andamenti di “CO” in base a giorno, mese e ora. Da questa prima esplorazione è emerso come il valore di “CO” fosse molto simile negli stessi giorni e alle stesse ore. Si osserva, inoltre, che i valori nulli non sono concentrati in uno stesso periodo ma abbastanza distribuiti nel dataset, per questo si è deciso di sostituire i nulli con la media dei valori di “CO” rilevati lo stesso giorno e alla stessa ora nella settimana precedente e in quella successiva. Qualora uno di quei due valori dovesse risultare anch’esso nullo si utilizzerebbe il valore non nullo dei due per il rimpiazzo.

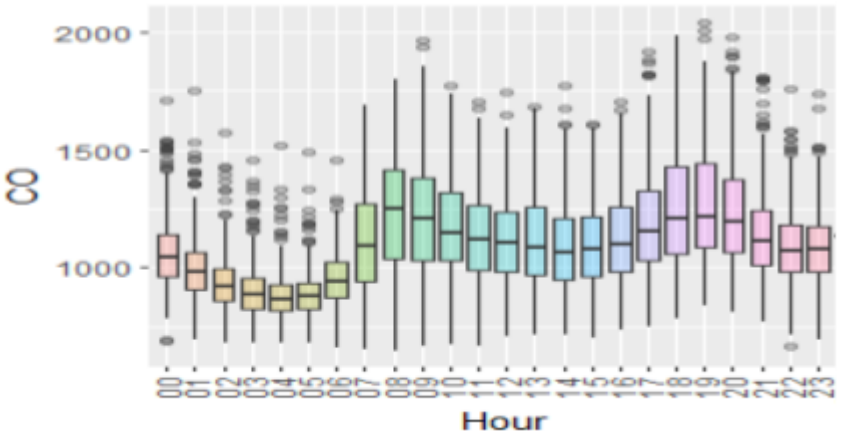
Risolto il problema dei nulli si è condotta un’analisi esplorativa del dataset completo, volta ad individuare eventuali ciclicità all’interno della serie storica.



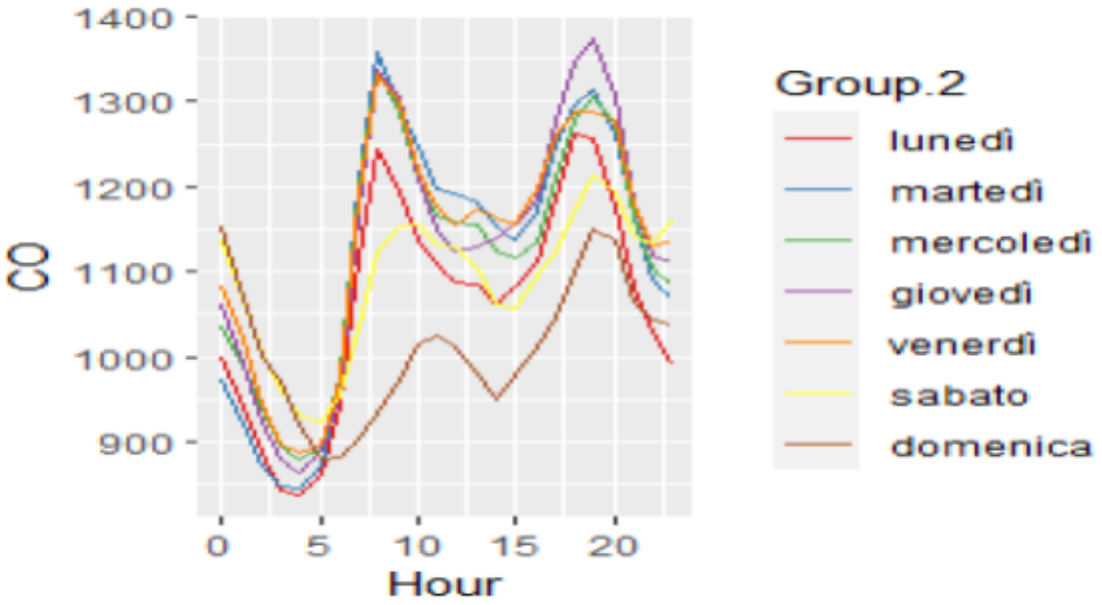
Da un primo sguardo si nota come la time series non sembri presentare alcun tipo di trend, ma presenti una sensibile riduzione della media e della varianza nei mesi estivi. Avendo solo un anno di dati però, non possiamo stabilire con certezza se il comportamento nei mesi estivi sia un evento sporadico o un comportamento stagionale, anche se è decisamente plausibile che si tratti del secondo. La presenza di molti picchi nella serie, inoltre, fa ritenere plausibile la presenza di stagionalità giornaliere e settimanali al suo interno. Per verificare ciò sono stati plottati box plot per ora, giorno della settimana e mese.







Si nota la presenza di stagionalità sia a livello giornaliero che a livello settimanale. I valori di “CO”, infatti, tendono ad assumere valori sempre più bassi nelle ore notturne, dalle 21 alle 5 per poi ricominciare a salire tra le 6 e le 8 e calando nuovamente dalle 9 alle 16 quando iniziano a risalire. Dai box plot emerge anche come nel weekend e il lunedì i valori di “CO” siano mediamente più bassi.



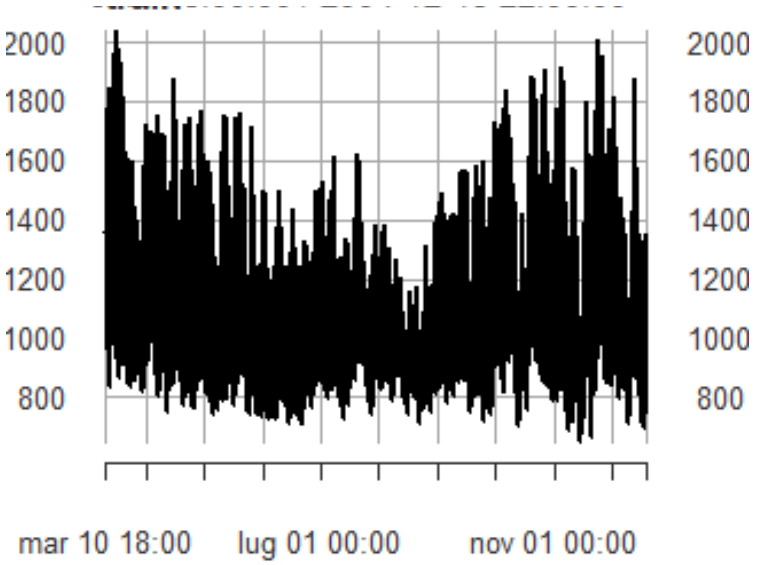
Il grafico sovrastante mostra l’andamento medio orario delle rilevazioni di “CO” alle diverse ore, si nota come i picchi siano all’incirca nelle stesse ore da lunedì a venerdì con qualche differenza di intensità, mentre il sabato e la domenica i picchi sembrano spostarsi in avanti di qualche ora oltre ad essere molto meno intensi.
Possiamo dunque concludere dicendo che vi è sicuramente una stagionalità ogni 24 ore e molto probabilmente anche una stagionalità settimanale.

Train-Test split:

Per poter addestrare e valutare le performance predittive dei modelli si è deciso di utilizzare il metodo holdout, splittando i dati originali in train set e test set. Nello specifico il train set si compone dell’80% dei valori della serie e il test set del 20%.

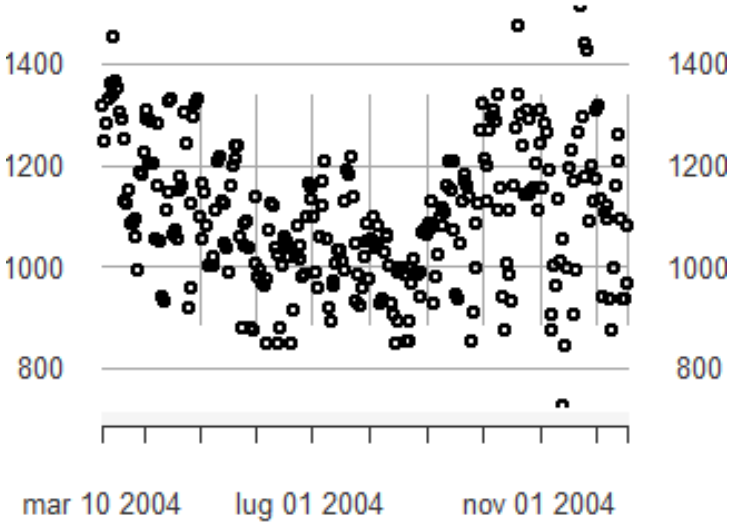
Modelli ARIMA

La prima famiglia di modelli utilizzata per la previsione sono gli ARIMA, Auto Regressive Integrated Moving Average, dei modelli additivi per la gestione di serie storiche. Requisito fondamentale per l’applicazione dei modelli ARIMA è che la serie storica fornitagli sia stazionaria sia in media che in varianza.

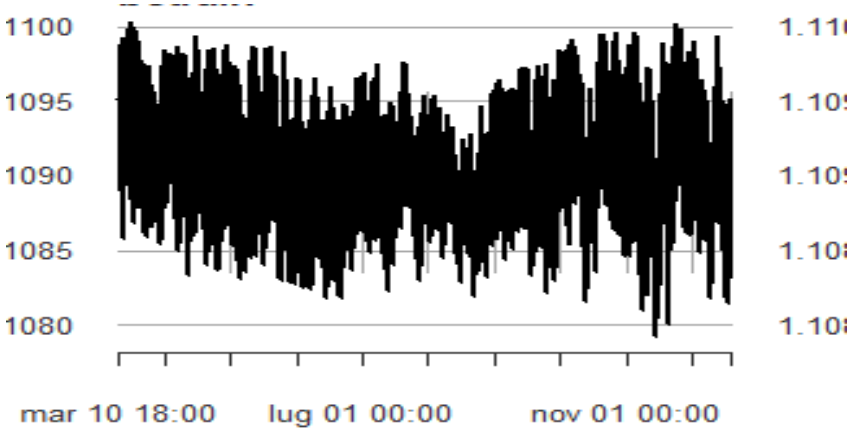


Dal plot della serie storica si nota come questa non sia stazionaria né in media, né in varianza. Si cerca dunque di renderla stazionaria partendo dalla stazionarietà in varianza.
Non avendo rilevato una relazione lineare tra il livello e la deviazione standard non si reputa sufficiente una trasformazione logaritmica per rendere la serie stazionaria in varianza. Di conseguenza verrà utilizzata una trasformazione di BoxCox con lambda = -0.8999268 determinato in maniera automatica dalla funzione BoxCox presente in R.

[PLOT MEDIE SU STD]

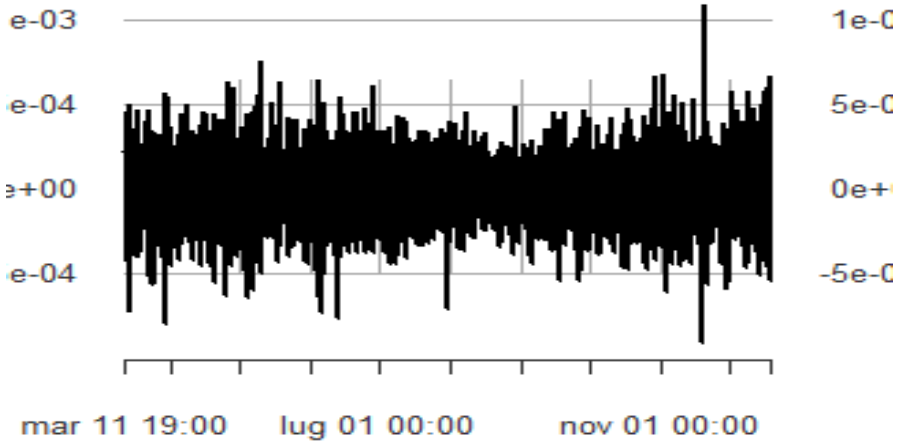


[PLOT TS TRASFORMATA BOX COX]:

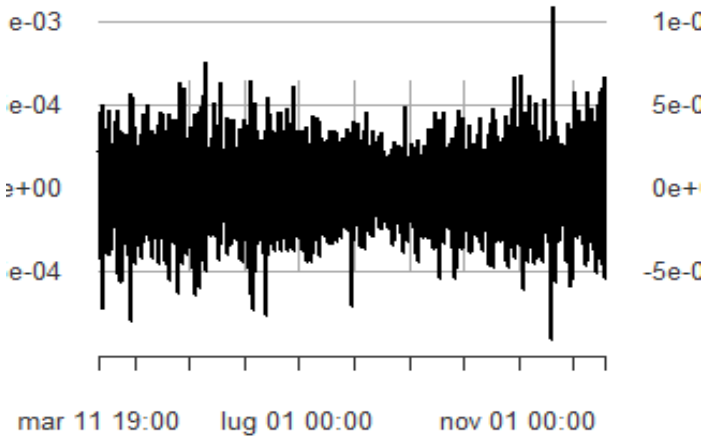


Risolta la non stazionarietà in varianza ci si occupa della non stazionarietà in media rilevata anche dal test di Dickey-Fuller. Sappiamo dall'esplorazione che esiste una stagionalità con periodo pari a 24, procediamo dunque ad una differenziazione stagionale di ordine 24. Dal grafico della serie dopo la trasformazione è possibile notare come sia stata ridotta la non stazionarietà in media. Già a questo punto dal test ADF risulta stazionaria la serie, dal grafico però si vede come non sia del tutto stazionaria in media e perciò si applica una differenziazione semplice di ordine 1. La serie così ottenuta risulta stazionaria.

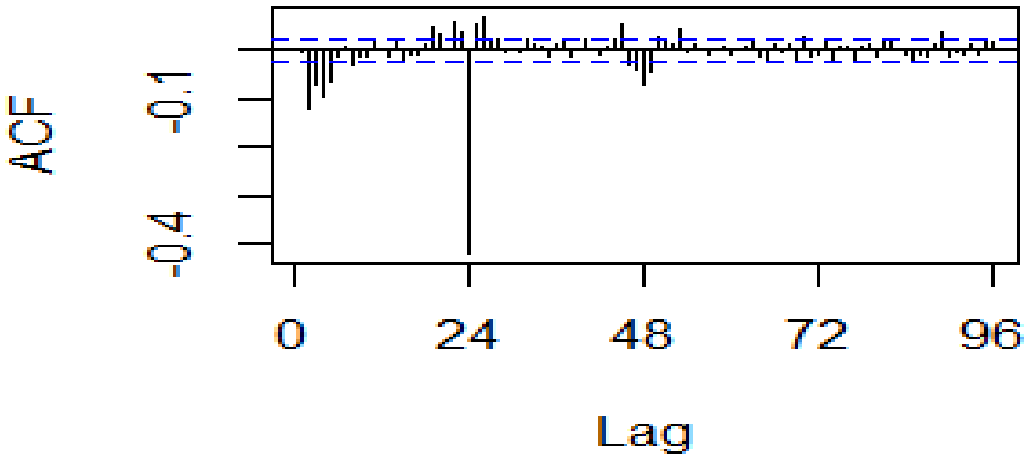
[PLOT TS DIFF 24]:

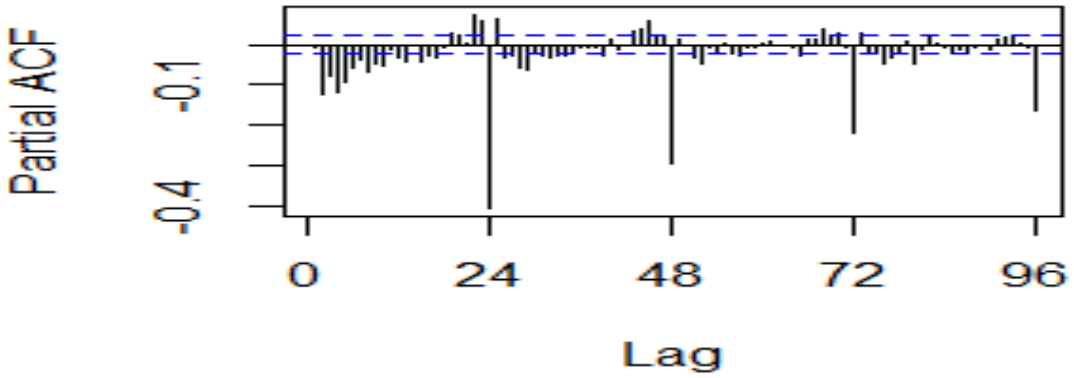


[PLOT TS DIFF 1]:



A questo punto è possibile iniziare l'analisi delle funzioni di autocorrelazione per determinare p,q e P,Q del modello ARIMA migliore (sappiamo già che d=1 e D=1 con S=24).



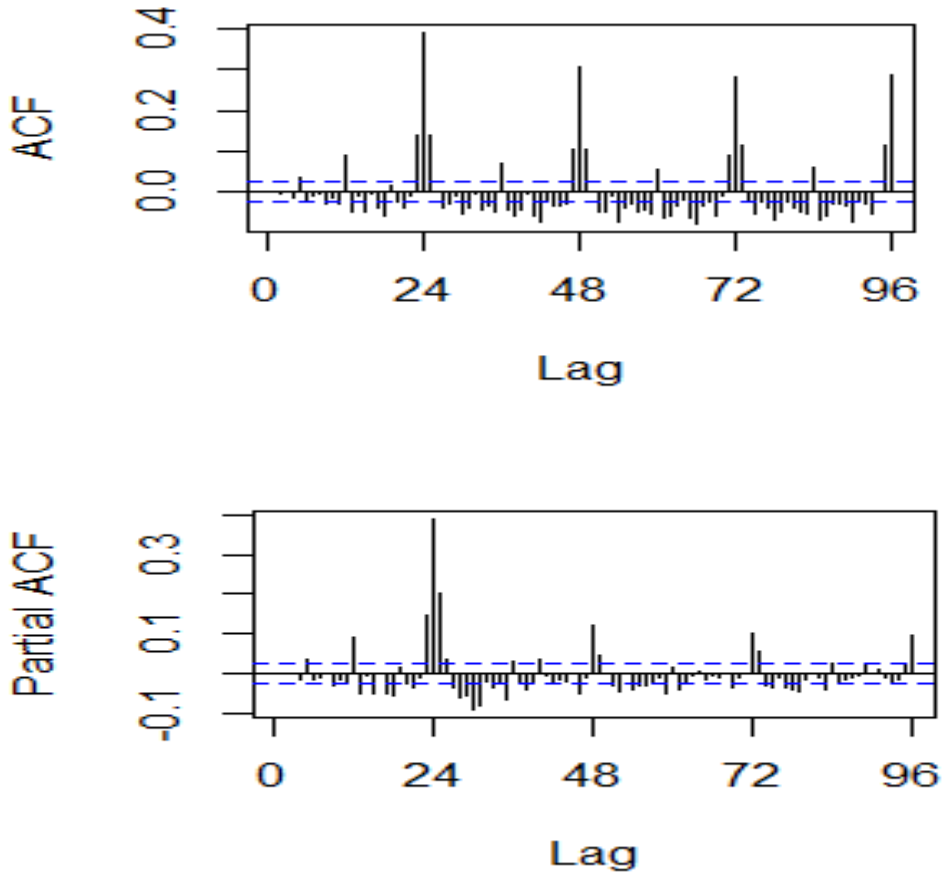


L'ACF presenta delle spikes ai lag 2, 3, 4 e 5 e mostra la presenza di memoria storica al lag 24. PACF, invece, presenta spikes ai lag 2, 3, 4 e 5 e presenza di memoria storica ai lag multipli di 24. Da questa prima analisi non è possibile stabilire immediatamente i parametri del modello più adatto. Sembra però emergere una leggera prevalenza della componente MA dato che ACF sembra rientrare a zero dopo il lag 4 e dopo il lag 24. Il comportamento di ACF e PACF, dunque, suggerirebbe di partire da un modello MA o SMA e, in base all'analisi dei residui arrivare a quel modello che presenta residui tra loro incorrelati. Per velocizzare la ricerca della combinazione migliore dei parametri si è scelto di effettuare un auto.arima sulla componente non stagionale in modo da trovare la combinazione di p,d,q tale da minimizzare

l'AICc. Si è poi utilizzato il modello così trovato come base di partenza per determinare la componente stagionale P,D,Q. auto.arima è stato eseguito dando come limiti massimi per p e q il valore 5, di solito, infatti, si preferisce non utilizzare valori di p,q e P,Q troppo elevati nei modelli. L'output restituito per la componente stagionale è un modello con p=5, d=1 e q=1.

Si è quindi addestrato il modello ARIMA(5, 1, 1) e si sono plottate ACF e PACF dei suoi residui.

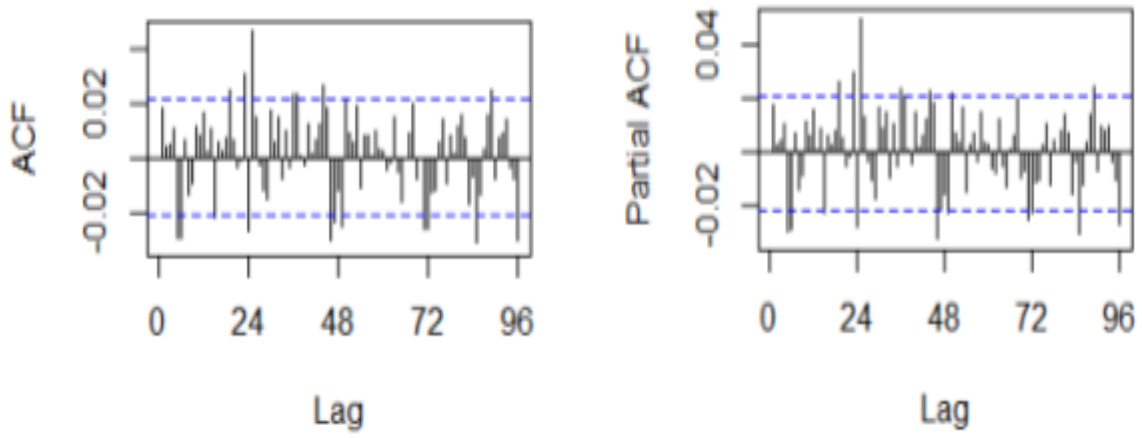
[PLOT ACF E PACF RESIDUI ARIMA(5, 1, 1)]



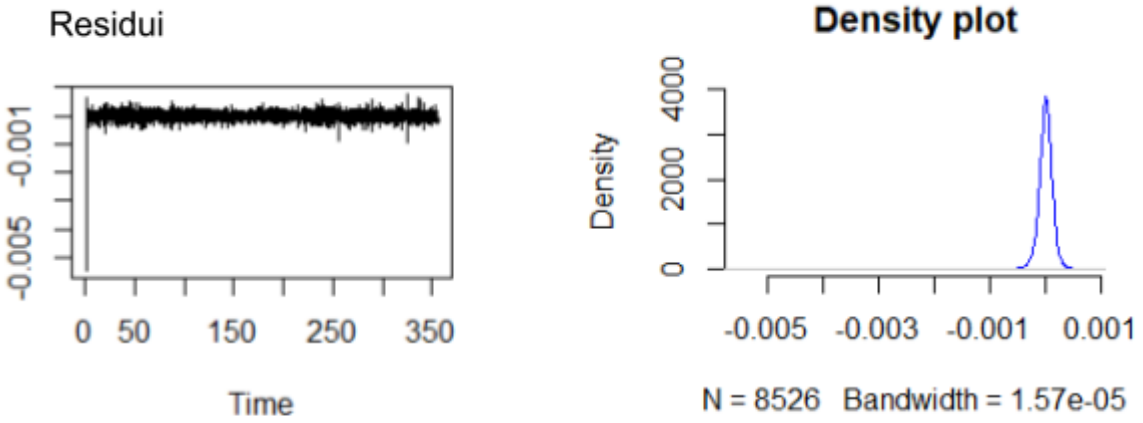
Dalle ACF e PACF dei residui si evince la presenza di memoria stagionale ai lag 24 e multipli. Nessuna delle due funzioni rientra nettamente a 0 dopo i primi 4 lag, ma la PACF sembra avere maggiore tendenza a farlo dopo il primo lag stagionale, per questo si è costruito un modello con p=5, d=1, q=1 e P=1, D=1, Q=0. Anche questo modello non portava a dei residui privi di memoria stagionale, per questo, partendo dalle funzioni di autocorrelazione di questi si è addestrato un nuovo modello, portando a 2 il parametro Q dato che l'ACF sembrava rientrare a zero dopo 2 lag mentre la PACF rientrava più lentamente. Questa procedura iterativa è stata seguita fino ad arrivare al modello con l'AICc più basso, vale a dire un modello SARIMA(5, 1, 1)(1, 1, 2). Anche in questo modello

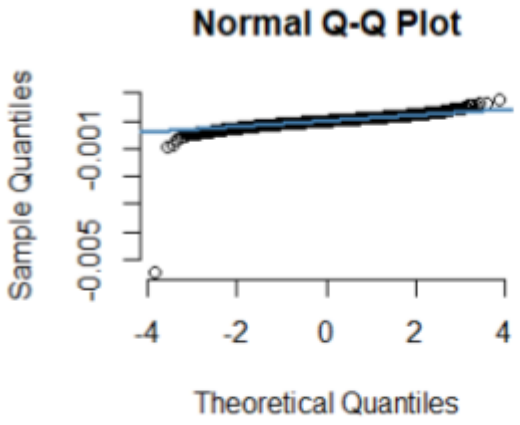
persisteva della memoria nei residui ai lag stagionali, ma si è scelto di fermarsi qui ritenendo di non poter migliorare i risultati ulteriormente.

[ACF E PACF DEI RESIDUI DI SARIMA(5, 1, 1)(1, 1, 2)[24]]



[RESIDUI DI SARIMA (5, 1, 1)(1, 1, 2)[24]]

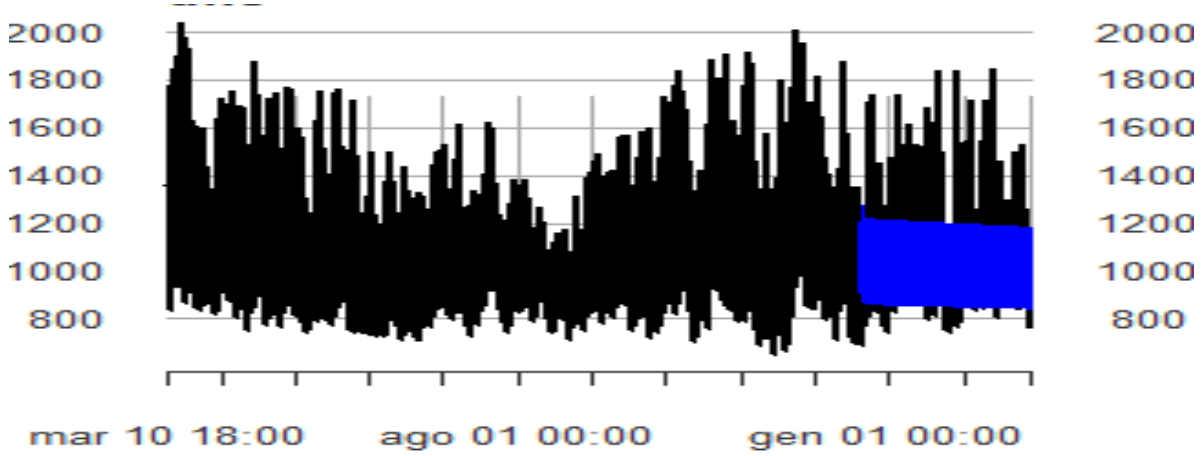




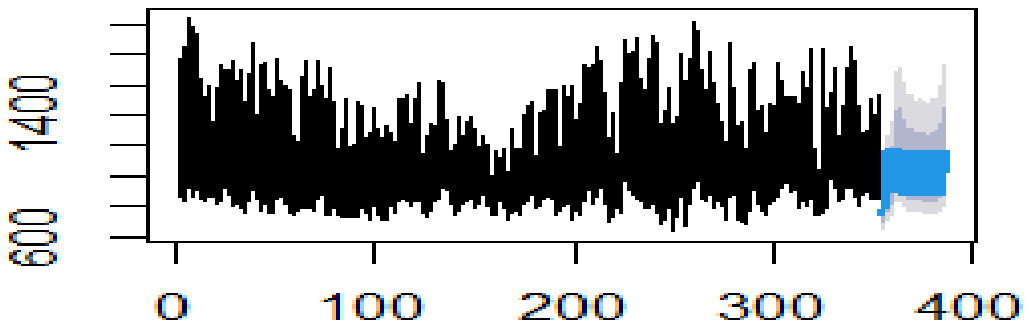
Partendo dal modello con l'AICc migliore si è provato a costruire un modello SARIMAX che presentasse un regressore aggiuntivo in grado di modellare la stagionalità a 168 ore (settimanale) emersa nell'esplorazione. La stagionalità settimanale è stata modellata con uno sviluppo in serie di Fourier di 8 sinusoidi.
Si evidenzia anche che per alcuni modelli non è risultata possibile la stima con il metodo di Maximum Likelihood e si è dovuto ricorrere al metodo CSS. Questi modelli non potendo essere valutati con l'AICc sono stati valutati direttamente in termini di performance (MAPE) sul test set.

Scelti i modelli più promettenti in termini di AICc si è deciso di selezionare il migliore in base alle performance predittive sul test set. In particolare si è scelta come metrica il MAPE. Da quest'ultima selezione è emerso come modello migliore fosse SARIMA(5, 1, 1)(1, 1, 2)[24], con un MAPE di 9.08, preferito al modello SARIMA(5, 1, 1)(3, 1, 3)[24] perchè a parità di prestazioni si è scelto il modello più parsimonioso. I modelli con il regressore sinusoidale hanno registrato performance leggermente peggiori di quelli che non lo presentavano.
Nessuno dei modelli addestrati è stato in grado di rimuovere completamente la memoria stagionale ogni 24 ore.

[PLOT PREVISIONI SU TEST SARIMA(5, 1, 1)(1, 1, 2)[24]]



[PREVISIONI MARZO SARIMA(5, 1, 1)(1, 1, 2)[24]]



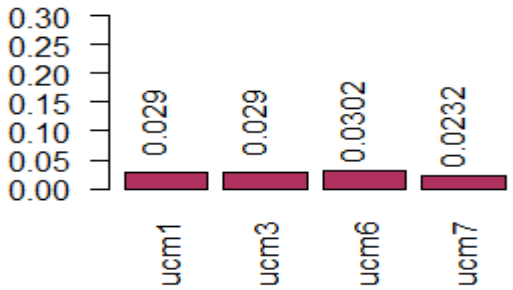
Modelli UCM

I modelli UCM (Unobservable Component Models) si basano sull'idea che una serie storica sia costituita da una somma di componenti stocastiche (trend, ciclo, stagionalità) e da eventuali regressori aggiuntivi. In questo caso non è stata considerata la componente ciclo in quanto si è ritenuto che la serie storica fosse troppo corta per presentare delle ciclicità di medio-lungo periodo e che bastasse la stagionalità per modellare i comportamenti ciclici all'interno della serie. Osservando i dati, inoltre, si è notato che non appare al loro interno un trend evidente o molto forte. Per questo nei modelli addestrati la componente di trend è sempre un random walk o un trend deterministico. Trattandosi di modelli additivi si è preferito fornire loro una serie già stazionaria in varianza applicandole la medesima trasformazione di BoxCox già applicata negli ARIMA.
Sono stati addestrati più modelli che combinavano un trend di tipo deterministico o di tipo random walk con stagionalità giornaliera a dummy stocastiche e stagionalità settimanale di tipo trigonometrico stocastica o deterministica. Sono state effettuate diverse prove, confrontando i modelli in base alle loro performance predittive sul test set, misurate in termini di MAPE. Di seguito vengono riportati i 4 modelli migliori tra quelli provati.
Modelli:

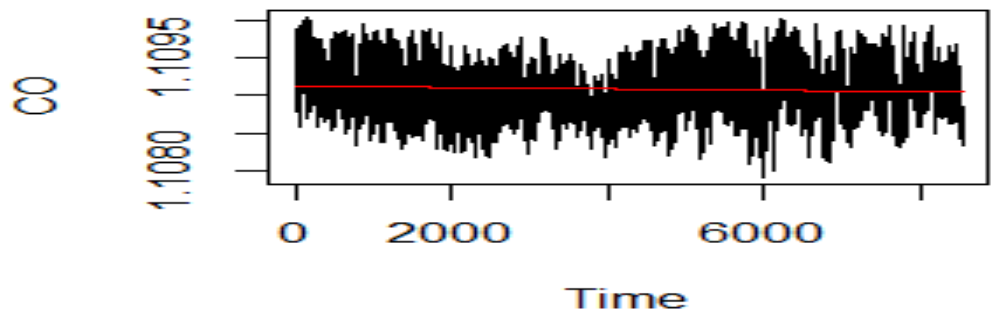
- mod1: trend RW, stagionalità giornaliera dummy stocastiche, stagionalità settimanale trigonometrica deterministica con 10 sinusoidi
- mod3: trend RW, stagionalità giornaliera dummy stocastiche, stagionalità settimanale trigonometrica stocastica con 10 sinusoidi
- mod6: LLT, stagionalità giornaliera dummy stocastiche, stagionalità settimanale trigonometrica deterministica con 10 sinusoidi
- mod7: trend deterministico, stagionalità giornaliera dummy stocastiche, stagionalità settimanale trigonometrica deterministica con 10 sinusoidi.

I modelli hanno performance abbastanza simili, ma il migliore è risultato essere quello con trend deterministico, stagionalità giornaliera a dummy stocastiche e stagionalità settimanale trigonometrica deterministica, che ha fatto registrare un MAPE sul test di circa 0.0232.

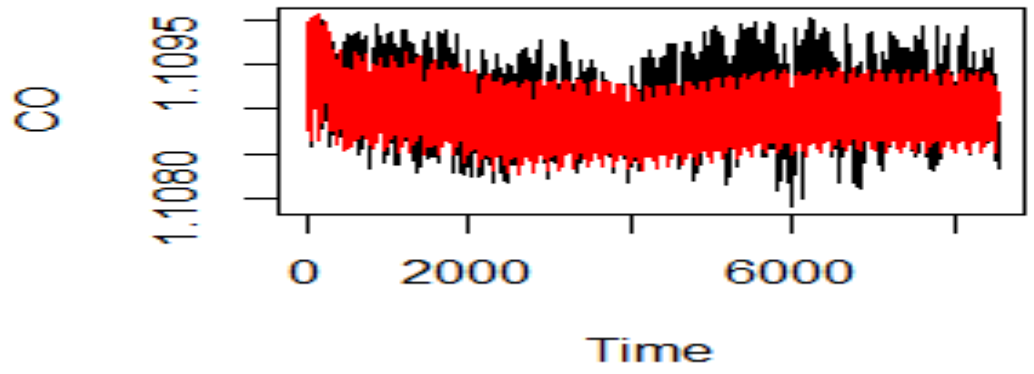
[MAPE UCM]



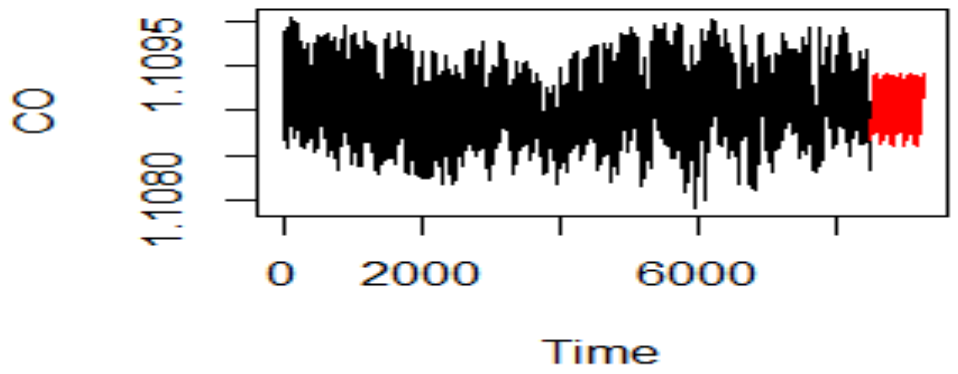
[PLOT TREND UCM]



[PLOT PREVISTI VS EFFETTIVI UCM]



[PLOT PREVISIONI MARZO UCM]



Modelli ML

L'ultimo tipo di modelli addestrati sono delle reti neurali di tipo LSTM e GRU. Le LSTM (Long Short Term Memory), così come le GRU (Gated Recurrent neural network) sono delle reti neurali appartenenti alla famiglia delle Recursive Neural Network. In particolare le LSTM presentano un layer LSTM che si compone di 3 diversi tipi di gate: Forget gate: decide se dimenticare il passato o meno in base all'output della sua funzione di attivazione sigmoidale, Input gate, Output gate. Le GRU utilizzano dei meccanismi di gate simili a quelli delle LSTM ma hanno un gate in meno, presentando solo il Rest gate che determina come combinare il nuovo input con i precedenti e l'Update gate che definisce quanta parte della memoria già presente trasmettere agli istanti successivi.

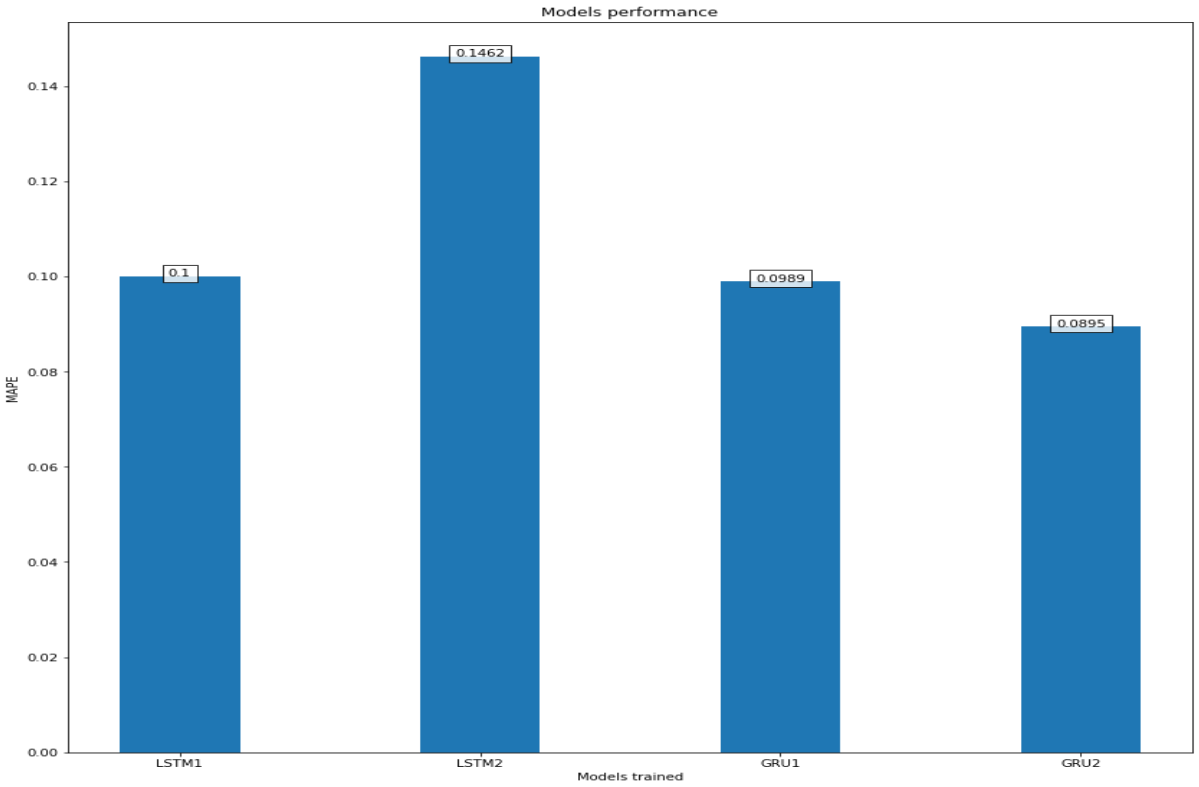
Entrambi i tipi di rete presi in considerazione sono piuttosto sensibili alla scala dei dati che, per questo, sono stati normalizzati portandoli nel range [0, 1].
Per confrontare i modelli sono state valutate le loro performance nella previsione del test, come metrica è stato utilizzato il MAPE.
Tra gli iperparametri da settare per questi modelli c'è il lookback, vale a dire il numero di osservazioni da prendere in considerazione per prevedere la successiva. In questo caso, essendo presenti una stagionalità giornaliera e una settimanale si è ritenuto adeguato un lookback pari a 168 osservazioni orarie (168=24*7). Gli altri iperparametri come il numero di layer, il numero di neuroni per layer, il numero di epoche e il batch size sono stati settati provando più combinazioni e scegliendo la migliore in termini di bontà di adattamento ai dati.

Di seguito sono riportate le caratteristiche dei modelli:

- Modello1:
  Layer LSTM da 128 neuroni
  Funzione di attivazione LeakyReLU
  Layer di dropout con rate 0.5
  Layer di output Dense con 1 solo neurone
- Modello2:
  Layer LSTM da 128 neuroni
  Funzione di attivazione LeakyReLU
  Layer di dropout con rate 0.3
  Layer LSTM da 64 neuroni
  Layer di dropout con rate 0.2
  Layer di output Dense con 1 solo neurone
- Modello 3:
  Layer GRU da 128 neuroni
  Layer di dropout con rate 0.5
  Layer GRU da 64 neuroni
  Layer di dropout con rate 0.3
  Layer GRU da 32 neuroni
  Layer di output Dense con 1 solo neurone
- Modello 4:
  Layer GRU da 128 neuroni
  Layer di dropout con rate 0.5
  Layer GRU da 64 neuroni
  Layer di dropout con rate 0.5
  Layer di output dense con 1 solo neurone

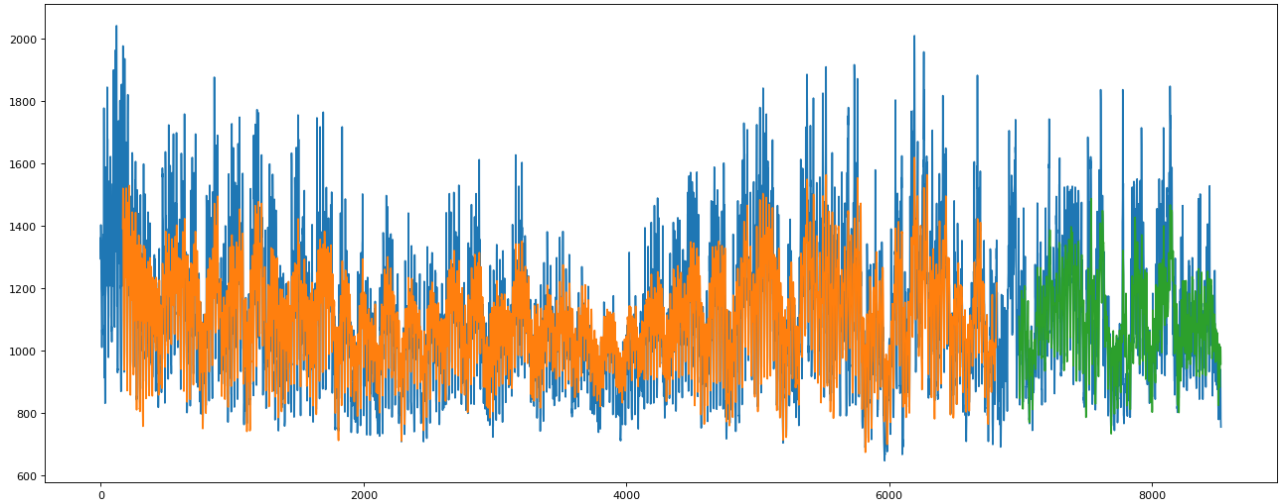
Tutti i modelli descritti sopra sono stati addestrati per 100 epoche utilizzando l'RMSE come loss function, un batch size di 32 e un ottimizzatore di tipo "Adagrad". Per cercare di contenere l'overfitting è anche stato settato un meccanismo di early stopping che interrompe il training quando la loss subisce peggioramenti per 5 epoche consecutive.

Di seguito è riportato un barplot con il MAPE relativo alle previsioni sul test dei 4 modelli:



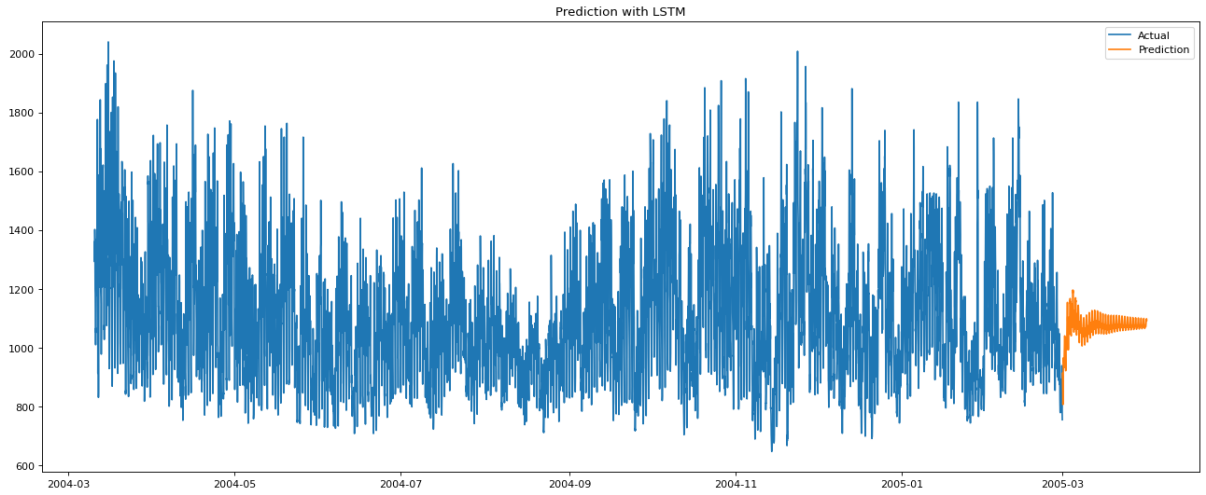
Il modello migliore risulta essere il Modello 4. Questa rete GRU verrà utilizzata per prevedere il mese di marzo.

Di seguito si riporta il grafico che mostra in blu la serie storica originale, in arancione i valori fittati dal modello sul train e in verde le previsioni del modello sulla parte di dati non osservati durante l'addestramento.



Il modello sembra avere un buon adattamento sia sul train che sul test e non sembra andare in overfitting.

L'immagine sottostante, invece, mostra la serie storica utilizzata per l'addestramento e, in arancione, la previsione per il mese di marzo.



Conclusioni

Per concludere, i modelli migliori relativi a ciascuna famiglia sono risultati:
SARIMA(5, 1, 1)(1, 1, 2)[24].

UCM con trend deterministico, stagionalità giornaliera a dummy stocastiche e stagionalità settimanale trigonometrica deterministica con 10 sinusoidi.

GRU con architettura:

- Layer GRU da 128 neuroni
- Layer di dropout con rate 0.5
- Layer GRU da 64 neuroni
- Layer di dropout con rate 0.5
- Layer di output dense con 1 solo neurone