Progetto analisi serie storica ossido di carbonio

Luca Maggi 866654

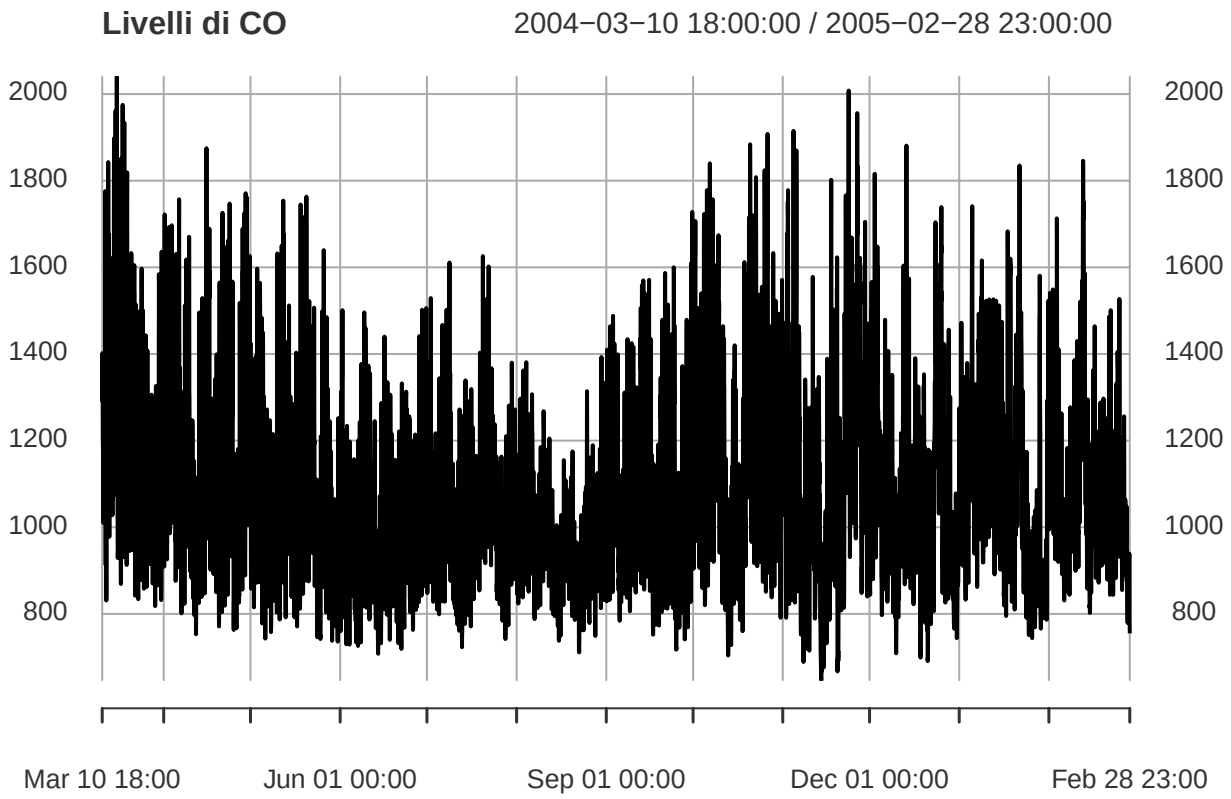
Contents

Introduzione e operazioni preliminari

La serie storica in esame comprende le osservazioni orarie delle emissioni di ossido di carbonio (CO) per il periodo che va dal 10 Marzo 2004, ore 18:00, al 28 Febbraio 2005, ore 23:00. Il totale delle osservazioni è di 8526, tra queste si contano 365 *missing values* da trattare prima di passare alla fase di analisi.

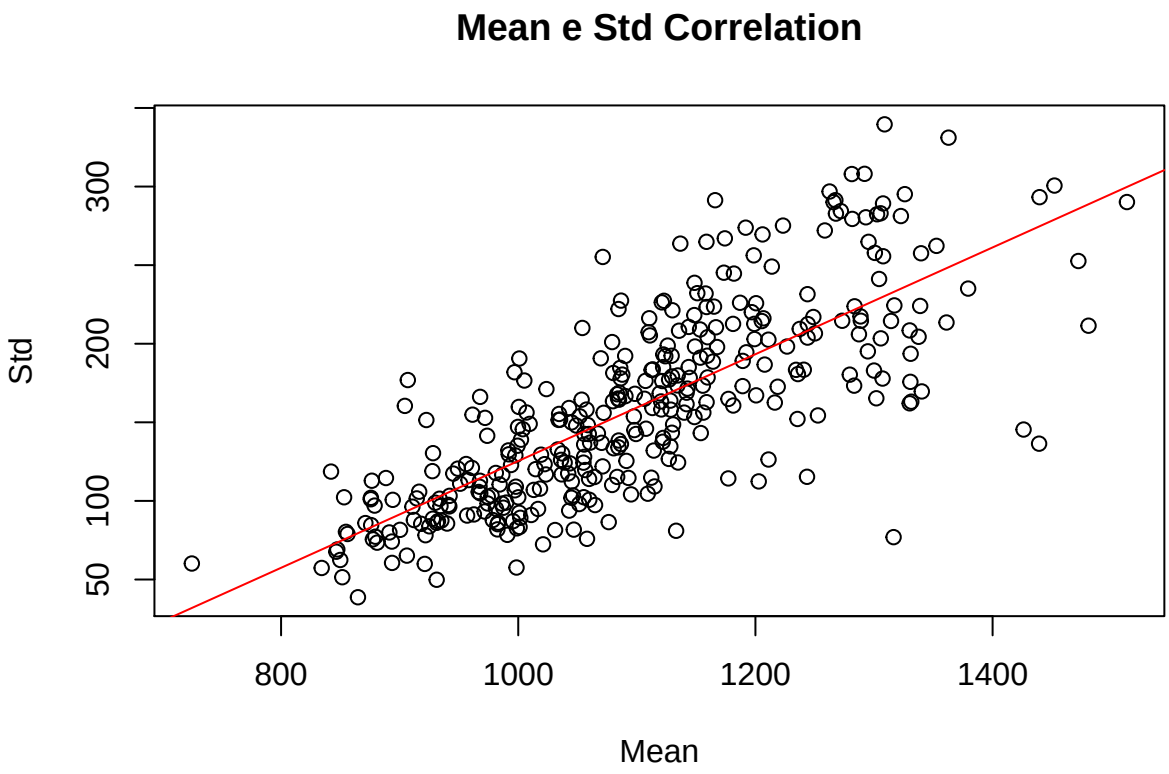
Data la natura della serie in esame nel trattare i valori mancanti è supposta una possibile correlazione tra il valore da sostituire, il mese ed il giorno della settimana in cui questo valore avrebbe dovuto manifestarsi; inoltre, ovviamente, è posta una particolare attenzione all'ora dell'osservazione. Partendo da questi presupposti i valori assenti sono sostituiti da una media delle osservazioni aventi questi tre elementi in comune. Per esempio, supponendo mancante l'osservazione di Venerdì 12 Marzo 2004, ore 17:00, il suo sostituto è trovato prendendo la media dei valori registrati alle 17:00 di ogni Venerdì di Marzo. Ovviamente, prima di procedere con questa imputazione, è necessario aggiungere una colonna nuova al dataset, contenente i giorni della settimana relativi ad ogni osservazione.

Imputati i valori mancanti la colonna della data, dell'ora e quella dei valori di CO sono trasformate in un unico oggetto *time series* di R per rendere il loro impiego più comodo. Alla fine la serie storica in esame risulta la seguente:

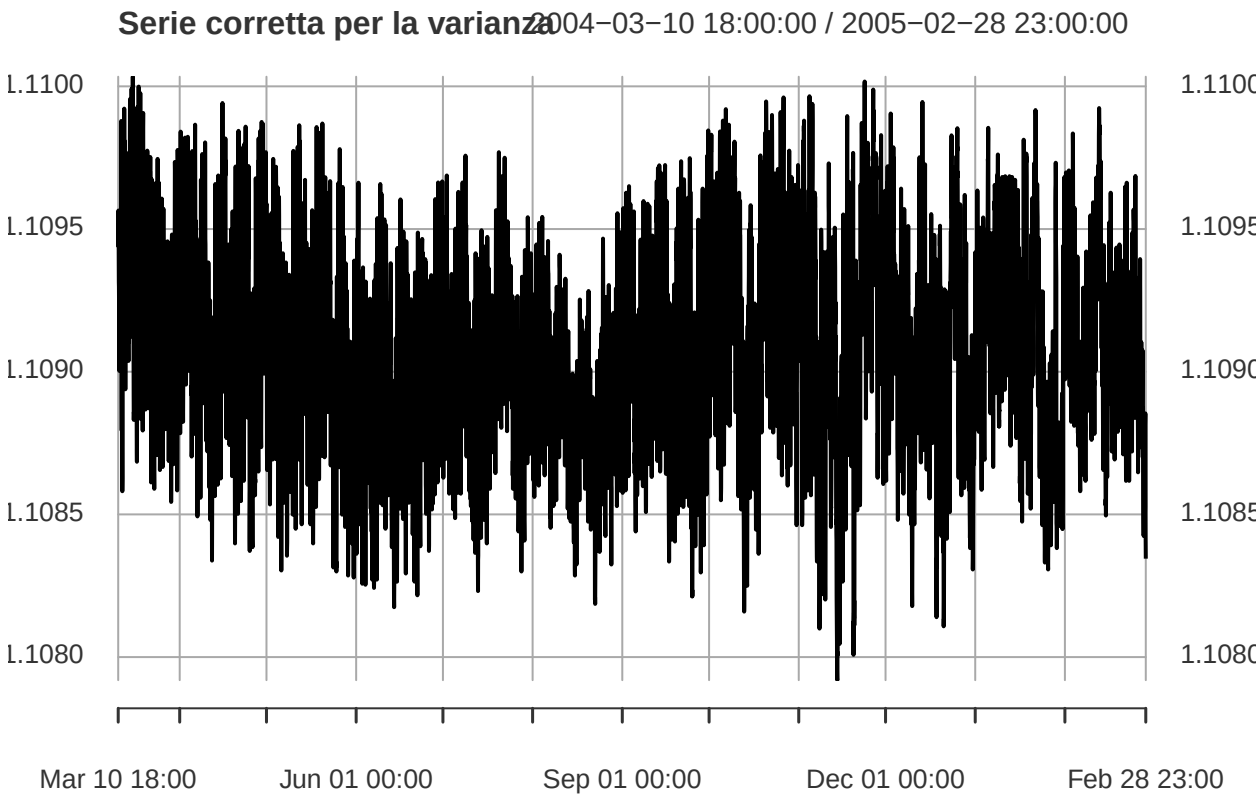


Modelli ARIMA

Il primo passo per la costruzione di un modello ARMA consiste nello studiare la varianza della serie, che, idealmente, dovrebbe essere costante nel tempo. Osservando il grafico precedente è possibile notare come i suoi valori tendano ad aumentare progressivamente. Questo effetto potrebbe essere semplicemente dovuto al passaggio dai mesi estivi a quelli invernali dell'anno; logicamente, infatti, avrebbe senso supporre un aumento delle emissioni di CO nei mesi freddi a causa, principalmente, del riscaldamento. Questa teoria è supportata anche dalla presenza di valori più alti nel primo mese della serie, corrispondente a Marzo 2004. Malgrado queste considerazioni è comunque indubbia la presenza di una varianza crescente che è opportuno trattare. Ad ulteriore supporto di ciò di seguito è mostrata la relazione tra la media della serie e la sua deviazione standard e, come si può notare, spicca una relazione lineare nel tempo.



Una trasformazione di Box-Cox è quindi applicata per cercare di riportare la varianza a valori costanti. In particolare, è utilizzato il pacchetto *forecast* per trovare il valore di lambda ideale da applicare, che risulta essere di circa -0.9. Questa famiglia di trasformazioni è ovviamente invertibile, condizione vitale per poter poi ottenere previsioni sensate rispetto ai valori originali della serie. Il grafico della serie corretta per la varianza è riportato di seguito



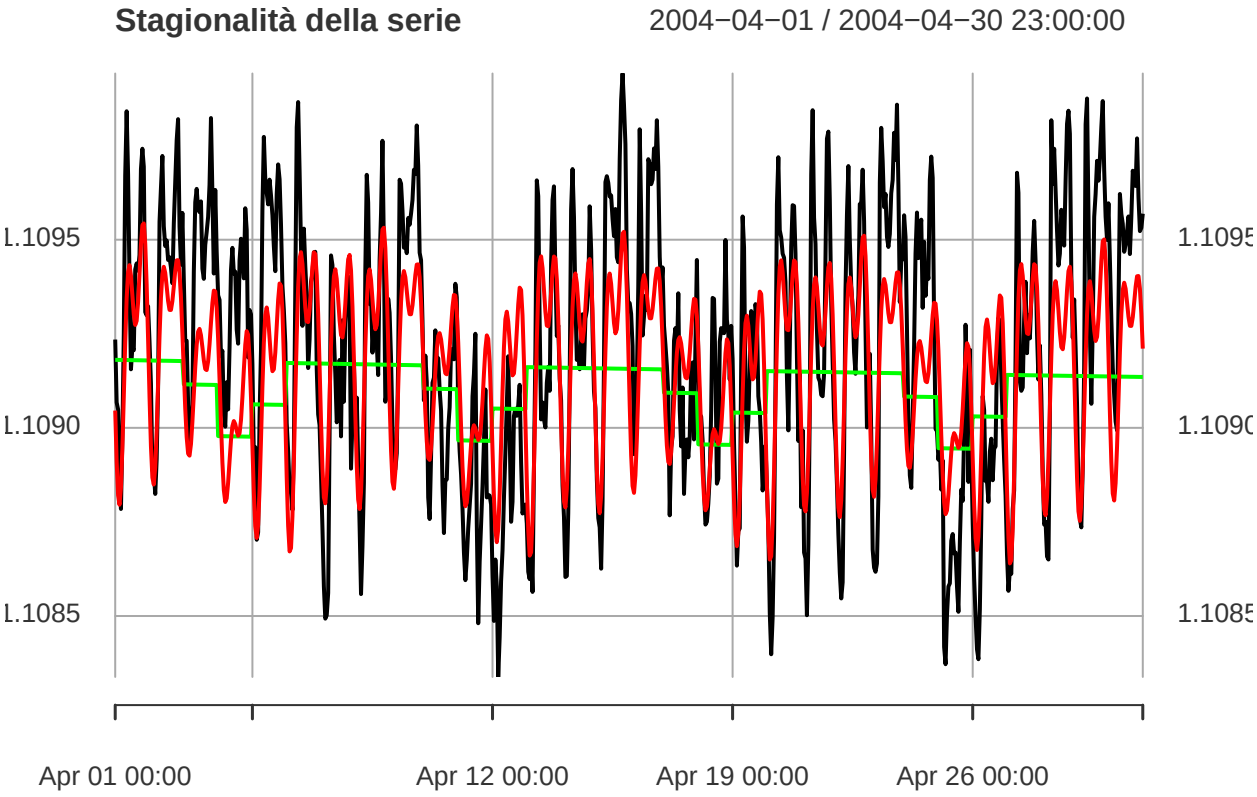
Un'analisi del grafico mostra come il problema sia stato alleviato, ma non risolto completamente. Ad ogni modo, dopo varie prove, è bene notare che, empiricamente, i modelli aggiustati per la varianza performano solo leggermente meglio rispetto a quelli che utilizzano la serie pura in questo caso specifico.

Il secondo aspetto da prendere in considerazione è la presenza di una stagionalità. Logicamente è facile intuire la presenza di una stagionalità giornaliera nei dati, questa intuizione è confermata anche dall'osservazione della serie. Un'altra possibile stagionalità è quella settimanale, molto probabilmente, però, questa è solo un multiplo di quella giornaliera e, quindi, risolvere la prima potrebbe portare ad una risoluzione anche della seconda e viceversa. Tenendo questo in mente diversi approci sono confrontati per modellare la stagionalità di questa serie; in particolare tre: una differenziazione stagionale, una modellazione mediante variabili *dummies* e una mediante l'uso di sinusoidi. In particolare la differenziazione è testata su un periodo di 24 ore, per cercare di eliminare la stagionalità giornaliera, le sinusoidi sono costruite sia con frequenza giornaliera che con frequenza settimanale (168 ore), mentre le dummies sono modellate su ogni giorno della settimana, meno il Giovedì. Nel caso delle *dummies* procedere in questo modo permette di ottenere le stime dell'impatto degli altri giorni della settimana confrontati a quello escluso.

Inizialmente, per comprendere l'influenza e l'entità dei regressori stagionali sulla serie viene sviluppata una semplice regressione lineare. Questo approcio soffre di due problemi principali: per prima cosa la presenza di correlazione nei residui inficia la precisione dei vari statistica test e relativi p-value, secondariamente l'utilizzo di sole componenti deterministiche nella modellazione della serie considera solo una parte della realtà. A causa di queste criticità il processo è quindi impiegato solo come mera base intuitiva per orientarsi nella selezione preliminare dei coefficienti. Nel performare queste regressioni è ipotizzato anche un trend nella serie, sia in forma normale che quadratica e, pur con le limitazioni di cui sopra, entrambi sono parsi significativi. La prima regressione performata comprende le *dummies* settimanali e indica che le uniche significative sono quelle di Lunedì, Sabato e Domenica. Gli altri giorni non hanno riscontrato un comportamento significativamente differente rispetto a quello del Giovedì e, in ottica di parsimoniosità del modello, sono scartate dalle successive analisi. In seguito una regressione con le sinusoidi a frequenza giornaliera ed una con le sinusoidi

a frequenza settimanale è performata. A livello giornaliero risultano significative solo sei sinusoidi, mentre a livello settimanale fino a 16 portano un certo grado di significatività.

Di seguito è riportata la modellazione della serie originale compiuta sia con variabili *dummies* settimanali, in verde, che con le sinusoidi giornaliere, in rosso; i dati mostrati coprono solo il mese di Aprile per maggiore chiarezza.



Come già accennato la serie sembra possedere un trend, questa supposizione è confermata anche da un *Augmented Dickey-Fuller unit root test* che, a tal proposito, restituisce un esito positivo.

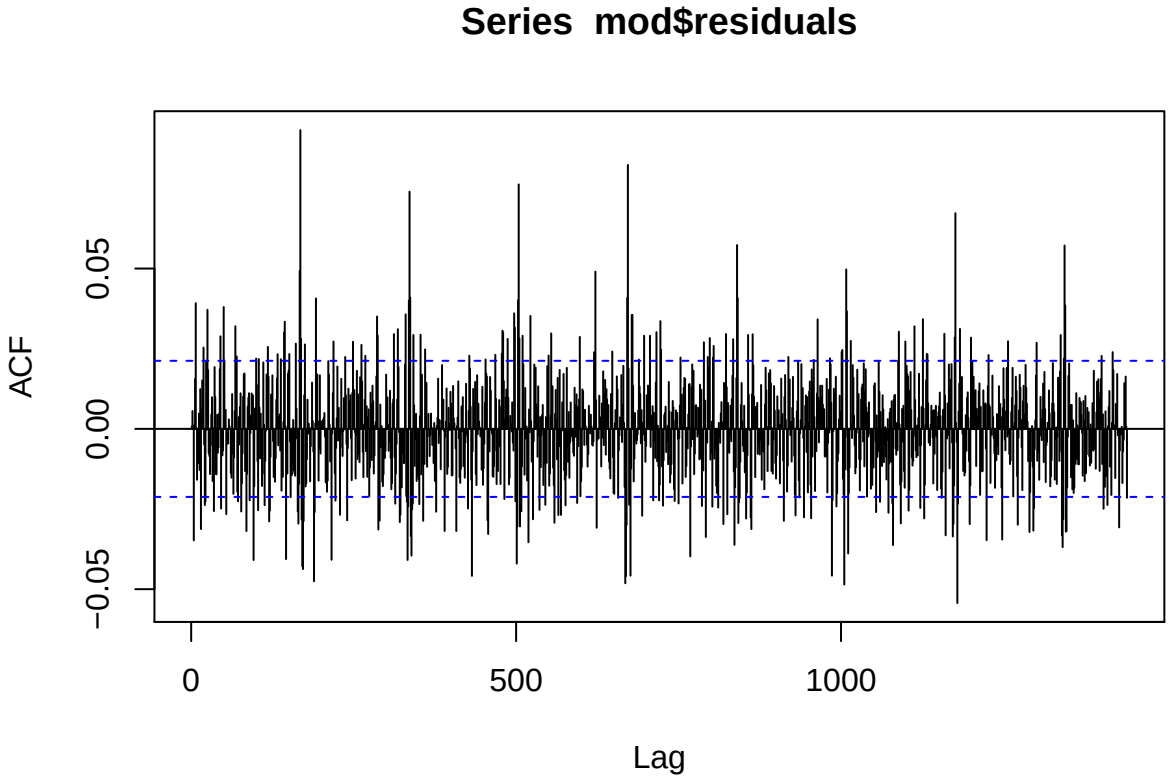
A questo punto diversi modelli ARIMA sono testati e messi in competizione tra loro; ogni modello proposto si basa sulla serie storica modificata dalla trasformazione di Box-Cox di cui sopra e aggiustata per il bias. Di volta in volta i parametri di ogni modello sono aggiustati in base ai diagrammi di correlazione dei residui ottenuti. Il primo di questi gestisce sia il trend che la stagionalità giornaliera mediante una differenziazione stagionale su 24 lags; dopo uno studio dei diagrammi di autocorrelazione e di autocorrelazione parziale il modello ottimale per questo approcio sembra essere un ARIMA(2,0,2)(1,1,1)[24]. Poi è utilizzato un modello che considera le sinusoidi per modellizzare la stagionalità giornaliera; in particolare esso è un ARIMA(3,0,3)(5,0,3)[24] a cui sono aggiunti i coefficienti delle sei sinusoidi e la media. Infine il focus è posto sulla stagionalità settimanale per la quale sono tentati due approci differenti. Il primo consiste in un ARIMA(1,0,1)(1,0,1)[24], con costante, a cui sono aggiunti i regressori delle *dummies* per i giorni della settimana. Il secondo è un ARIMA(2,0,2)(3,0,3)[24] con i regressori delle 16 sinusoidi settimanali e l'intercetta. In entrambi i casi la stagionalità settimanale è modellata mediante l'aggiunta di regressori, mentre il periodo del modello stesso è tenuto di 24 ore. Questo deriva tanto dalla volontà, confermata empiricamente, di cogliere il massimo da entrambe le stagionalità, quanto dalle limitazioni computazionali che non permettono di utilizzare adeguatamente i coefficienti AR(p) e MA(q) stagionali con periodi di 168 osservazioni.

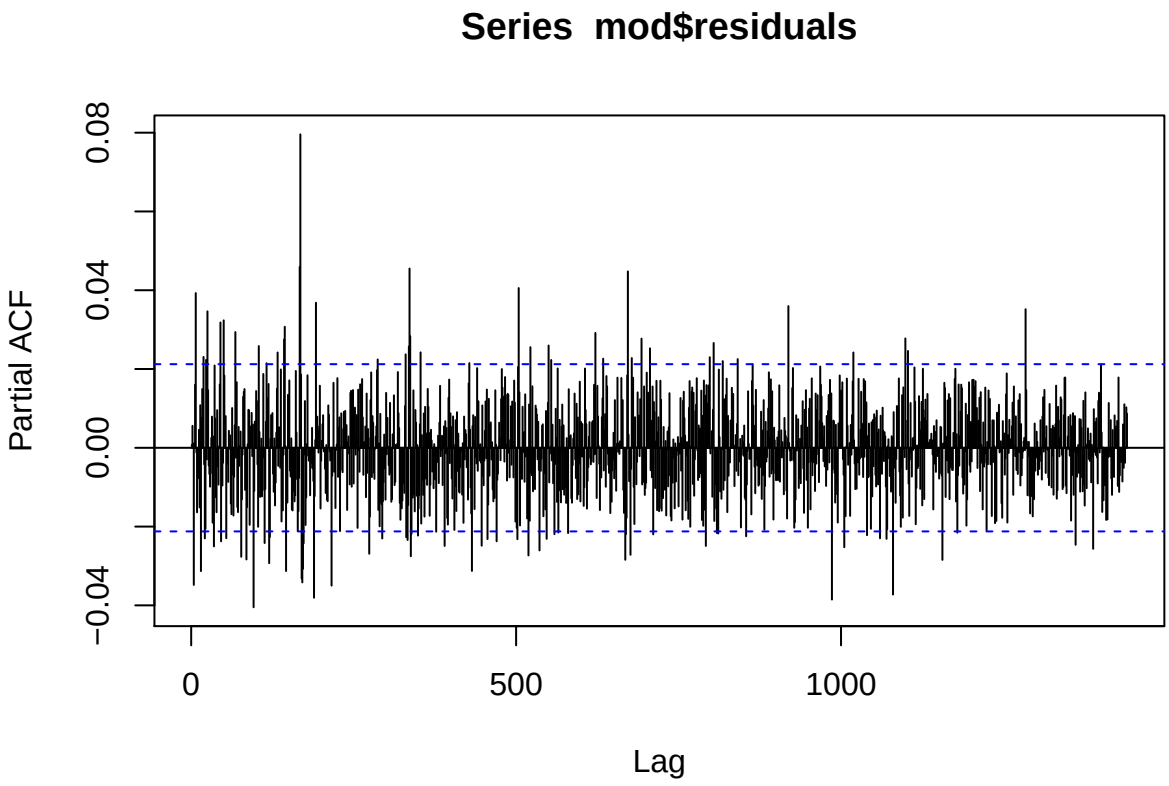
Come criterio di valutazione per trovare il modello migliore tra tutti quelli proposti è utilizzato l'AIC corretto. Poi i candidati più promettenti sono addestrati su tutta la serie meno l'ultimo mese (Febbraio 2005) ed utilizzati per prevedere la parte della serie mancante. Avendo le previsioni e i dati reali a cui esse si riferiscono

il MAPE (Mean Average Percentage Error) è calcolato; questo è utilizzato come stima della precisione del modello nel prevedere il continuo della serie. Dal confronto degli indici AIC i modelli a base di sinusoidi sono quelli che sembrano performare meglio, in particolare quello che impiega 16 sinusoidi per modellare la stagionalità settimanale. Calcolato il MAPE sulle previsioni per Febbraio 2005 per entrambi risulta che il modello con la differenza giornaliera raggiunge un valore di 11.6, mentre quello a base settimanale ottiene un errore di 10.9, quindi, alla fine, avendo raggiunto un errore percentuale più basso quest'ultimo è scelto come modello definitivo. E' empiricamente dimostrabile che aggiungere una differenziazione al modello scelto ne abbassa l'efficacia predittiva, quindi, discapito di quanto prima affermato, questo passaggio è scartato. Il modello finale risulta quindi essere un ARIMA(2,0,2)(3,0,3)[24] con 16 regressori delle sinusoidi e l'intercetta.

Nel corso della sperimentazione diverse altre variabili *dummies* sono state testate per cercare di migliorare le performance delle previsioni. In particolare è stata creata una variabile contenete le maggiori festività nazionali, come Natale, Capodanno o Pasqua, ma la sua implementazione non si è rivelata particolarmente fruttifera. Anche altri tentativi non hanno prodotto i risultati sperati, come la creazione di una variabile per differenziare i mesi caldi estivi da quelli più freddi invernali o una per cogliere la differenza nelle emissioni tra la notte ed il giorno.

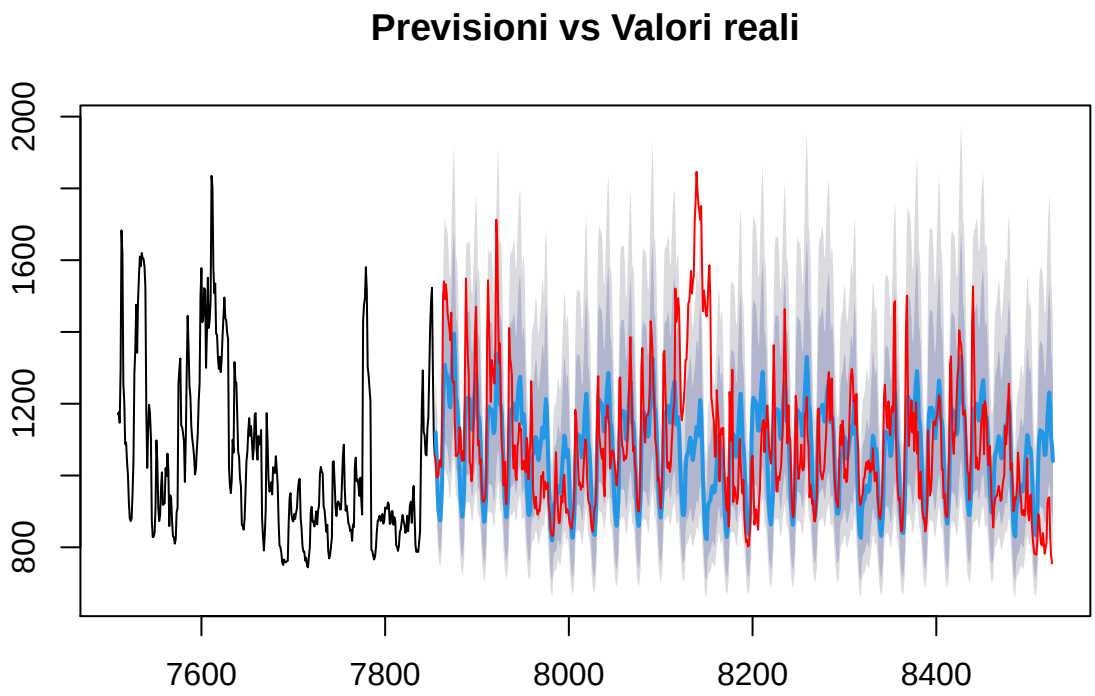
Di seguito sono riportati i diagrammi dell'autocorrelazione e della autocorrelazione parziale per i residui del modello scelto, i grafici si riferiscono ad un intervallo di circa due mesi (60 giorni).





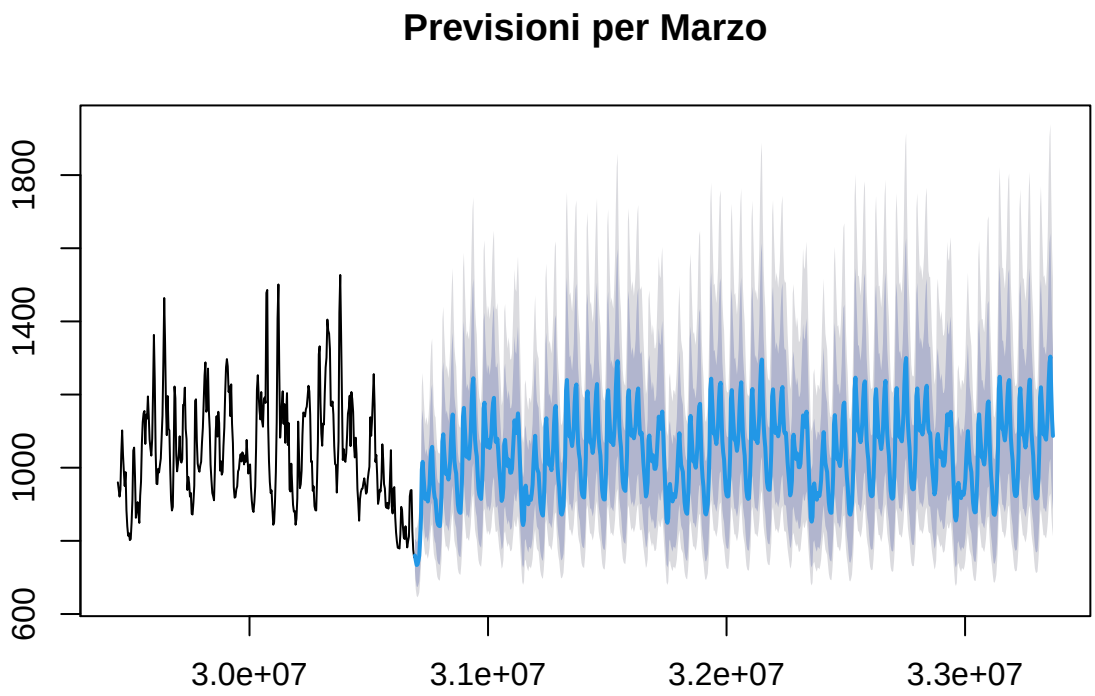
I valori del grafico ACF tendono progressivamente a zero, trend molto evidente se si osservano le correlazioni sulla serie per intero, mentre persistono dei lag significativi nella prima parte del PACF. Questo è dovuto in parte alla ristrettezza delle barre di confidenza, causata dall'alto numero di osservazioni disponibili, ed in parte dalla presenza di una correlazione residuale difficilmente eliminabile con questo tipo di modelli. In generale, comunque, i residui sono sufficientemente simili a quelli di un *White Noise* per poter ritenere il modello soddisfacente.

Infine si mostrano le previsioni ottenute sull'ultimo mese della serie storica, in blu, confrontate ai valori effettivi, in rosso.



Anche in questo caso si nota la difficoltà del modello ad adattarsi a repentini cambiamenti. In linea generale, però, l'andamento della serie è adeguatamente modellato, con tutte le limitazioni del caso.

A questo punto è possibile prevedere il mese successivo alla fine della serie ed ottenere così i valori richiesti.



Modelli UCM

Il modello UCM scelto per rappresentare la serie è costituito da due componenti: un trend e una stagionalità. Il trend risultato migliore è un local linear trend (LLT), attraverso questo, infatti, è permessa un'evoluzione nel tempo sia al coefficiente angolare sia all'intercetta del trend stesso. Per quanto riguarda la stagionalità la decisione è ricaduta su una rappresentazione trigonometrica a 16 armoniche, di periodo settimanale e con varianza evolutiva nel tempo.

Per garantire una migliore convergenza del modello non sono usate le matrici inizali di medie e varianze (a1 e P1) a distribuzione difusa, ma i valori al loro interno sono impostati manualmente. In particolare, per il valore iniziale dell'intercetta e delle sinusoidi sono utilizzati i coefficienti del modello ARIMA precedentemente addestrato. Per popolare la matrice della varianza iniziale, invece, si è calcolato il valore della varianza totale delle osservazioni.

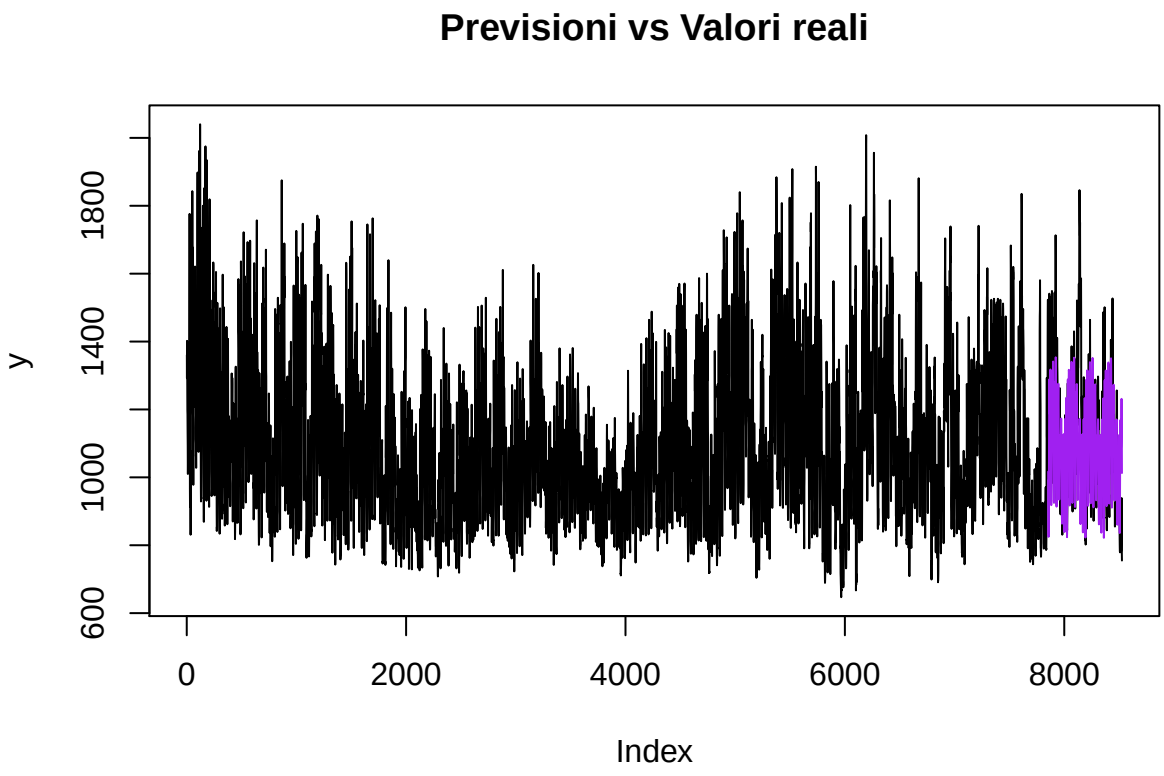
Infine è definita una funzione di *update* delle matrici Q e H la quale, partendo da una parametrizzazione iniziale, si occupa della loro evoluzione. Definire questa funzione risulta particolarmente importante nel modello in esame per motivi computazionali. Infatti, avendo impostato una componente stagionale con varianze diverse da zero in Q, il software tenderebbe a trattare l'evoluzione di ogni varianza di ogni sinusoide singolarmente, appesantendo di molto il calcolo. Con il procedimento svolto, invece, si spinge il programma a trattare l'evoluzione di tutte le sinusoidi ugualmente, ottenendo così risultati migliori. Per quanto riguarda la parametrizzazione iniziale si sono utilizzate diverse frazioni della variazna totale dei dati, di cui si è poi preso il logaritmo per imporre un segno positivo.

Per essere testato il modello è addestrato su tutta la serie meno l'ultimo mese di dati, poi sono calcolate le previsioni per il mese mancante e confrontate ai valori reali.

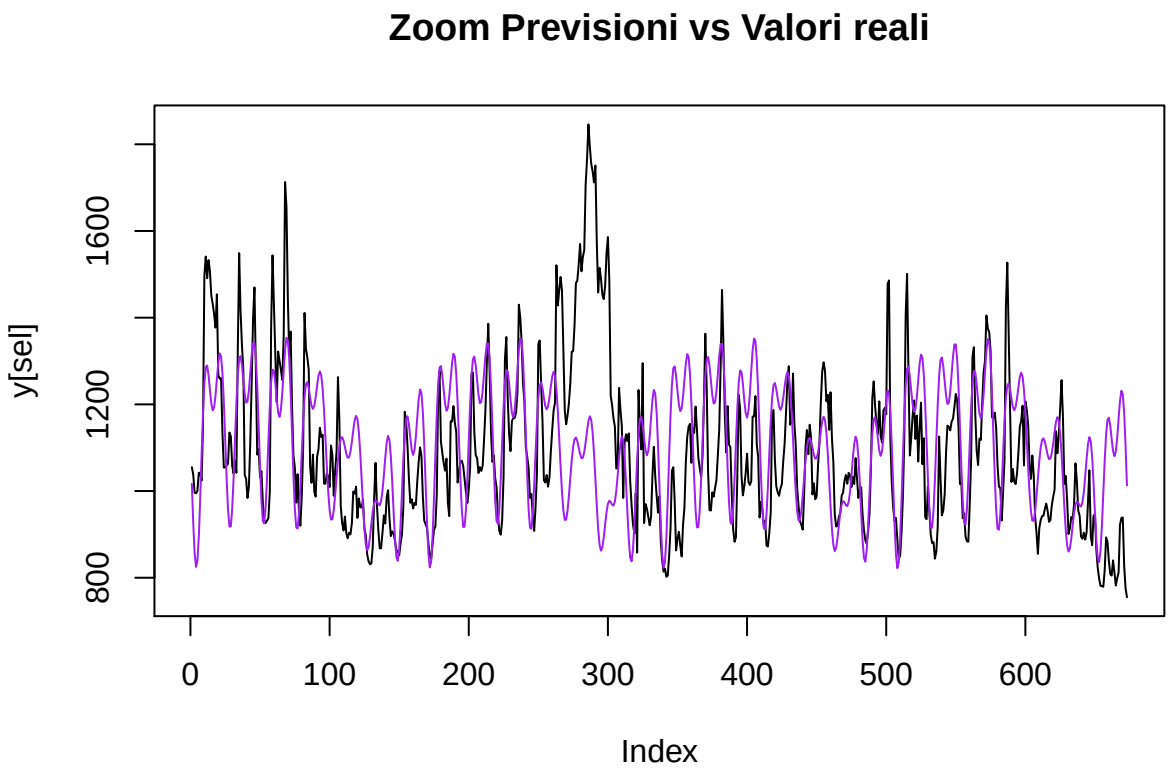
Di seguito è riportata la previsione per l'ultimo mese di dati, in viola, rispetto ai valori effettivi, in nero.

```
## Warning in `dim<-.zoo`(`*tmp*`, value = length(x)): setting this dimension may
## lead to an invalid zoo object
```

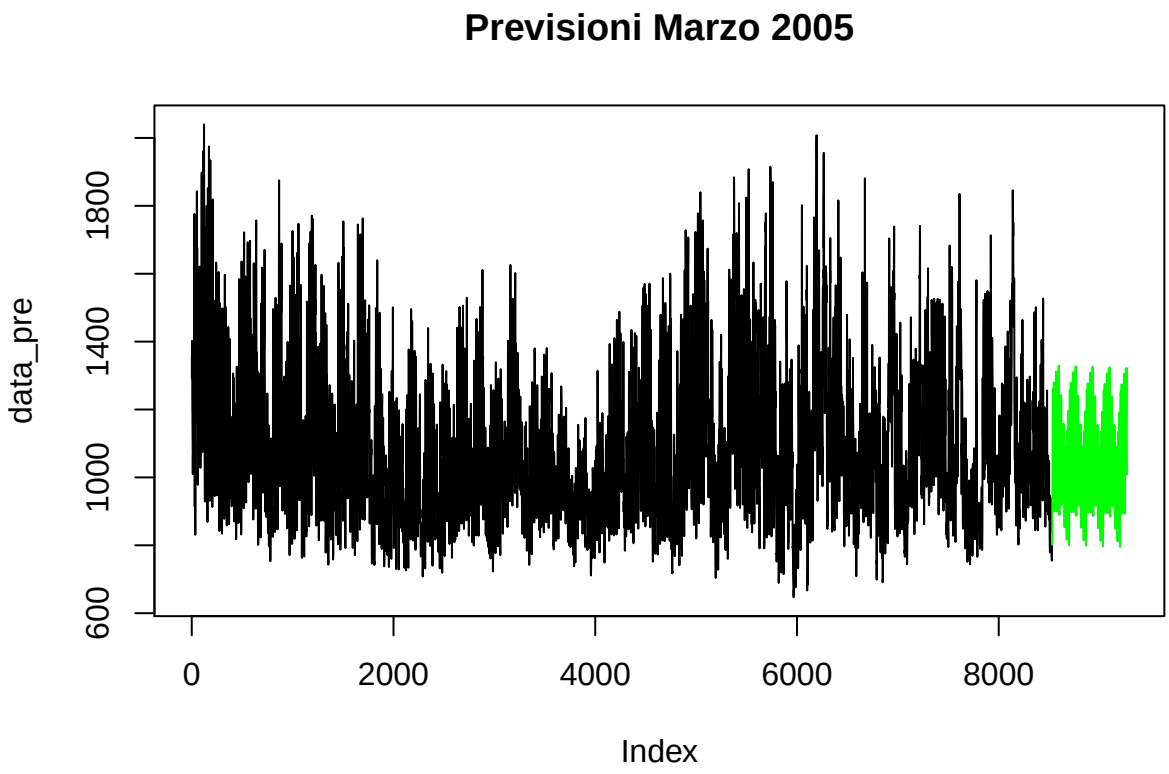
```
## Warning in `dim<-.zoo`(`*tmp*`, value = c(n, p)): setting this dimension may
## lead to an invalid zoo object
```



Uno zoom sui valori predetti consente di valutare meglio l'accuratezza del modello. In generale le previsioni sembrano seguire bene il modello di riferimento, anche se, a volte, sovrastimano leggermente i dati. Il calcolo del MAPE segnala che l'errore del modello è, in media, del 12%.



Infine il modello validato è riaddetsrato su tutta la serie disponibile ed è utilizzato per prevedere il mese di Marzo 2005. Le previsioni ottenute sono riportate nel grafico di seguito in verde.



Modelli Machine Leraning

L'ultima famiglia di modelli utilizzati è quella basata sul *Machine Learning*, in particolare è impiegata una rete neurale. Per addestrare questo modello il dataset è diviso in un subset di training, circa 80% della serie, ed uno di validazione, il rimanente 20%. La suddivisione prende in considerazione la natura sequenziale della serie storica, infatti il campionamento non avviene casualmente, ma la serie è tagliata ad uno specifico punto, preservandone così la sua natura. Prima della suddivisione le osservazioni corrispondenti all'ultimo mese disponibile, Febbraio 2005, sono estratte e conservate per servire da test set; saranno poi utilizzate per valutare l'accuratezza delle previsioni.

Diverse forme del modello, combinazioni di neuroni per ogni suo strato, numero di strati e ampiezza delle finestre di previsione sono provati. Alla fine la configurazione che risulta empiricamente migliore è quella costituita da una rete a 6 strati: uno di input, due LSTM, due di drop-out e uno di output. Gli strati LSTM hanno rispettivamente 122 e 66 neuroni ciascuno e, dopo ognuno di essi, è utilizzato un layer di regolarizzazione con una percentuale di drop-out del 50%. Come input il modello utilizza sequenze di 168 osservazioni successive (due settimane), senza sovrapposizione, e restituisce una sequenza di pari lunghezza di previsioni. Tenendo presente la natura dei dati utilizzati i layer LSTM sono impostati come "*stateful*" e, in nessun caso, sono utilizzate tecniche di "*shuffle*" dei dataset. Le funzioni di attivazione sono quelle classiche di un layer LSTM, ovvero una tangente iperbolica per l'attivazione ed un sigmoide per la parte ricorrente. Un algoritmo di *early stopping* conclude l'addestramento del modello alla decima epoca, la Figura 1 mostra il processo di *fitting*.

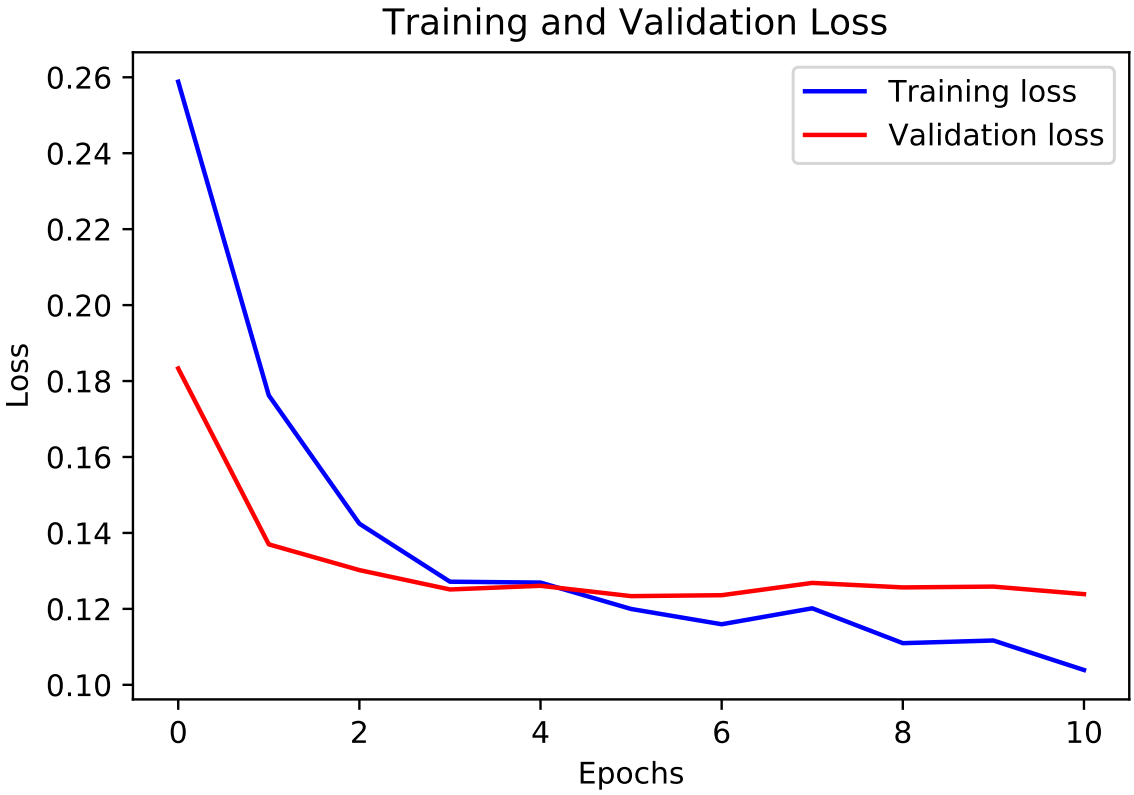


Figure 1: Addestramento del modello

Dopo l'addestramento, per valutare le capacità del modello, sono create le previsioni per il mese di Febbraio 2005 e, queste, sono confrontate ai valori reali. La prima settimana di Febbraio è predetta sull'ultima settimana di osservazioni, mentre, le previsioni per le tre settimane successive, sono basate sui valori a loro volta predetti della settimana subito precedente. Per questo motivo più si cerca di prevedere nel futuro, più il modello è soggetto ad un errore. Nella Figura 2 è mostrato il confronto tra i valori predetti (arancione) ed i valori reali del mese di Febbraio 2005 (blu), il MAPE ottenuto è circa del 15%.

Un focus sull'ultimo mese è utile per valutare meglio le previsioni(Figura 3).

Infine sono prodotte le previsioni per il mese di Marzo 2005, come da richiesta. Anche in questo caso la prima settimana risulta la più precisa e, man mano che ci si allontana dalla serie originale, i valori sono soggetti ad un errore sempre maggiore. Di seguito sono mostrati i grafici delle previsioni (Figura 4) e della serie completa (Figura 5), composta dai valori osservati e dalle previsioni dell'ultimo mese.

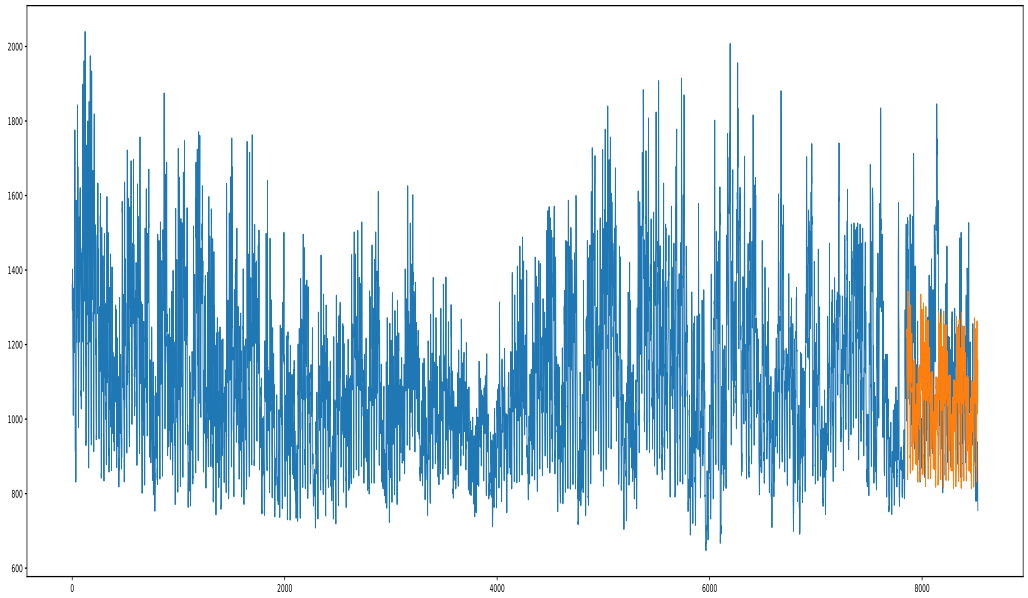


Figure 2: Previsioni vs Valori reali Febbraio 2005

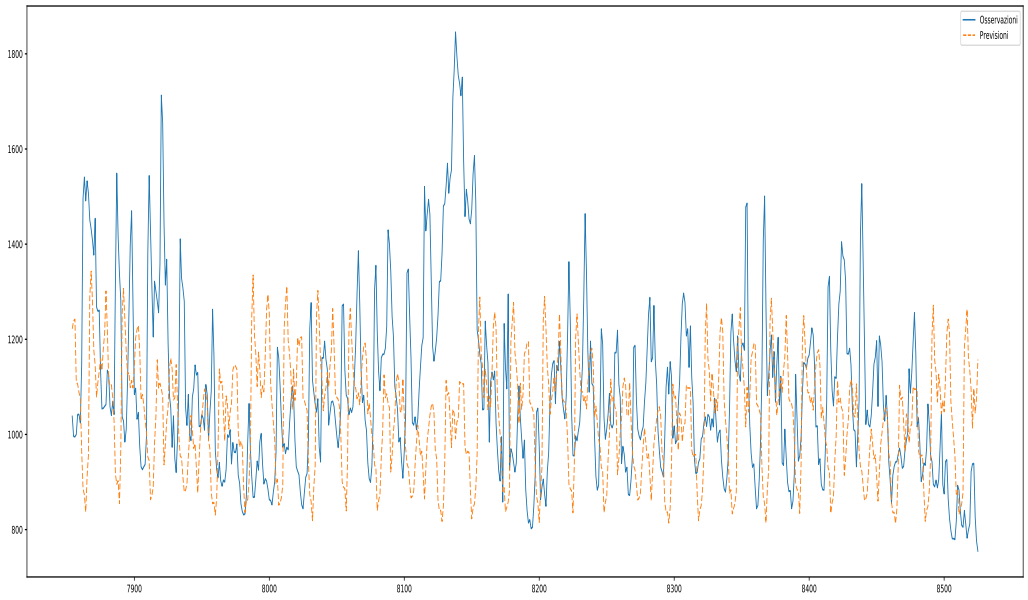


Figure 3: Zoom Previsioni vs Valori reali Febbraio 2005

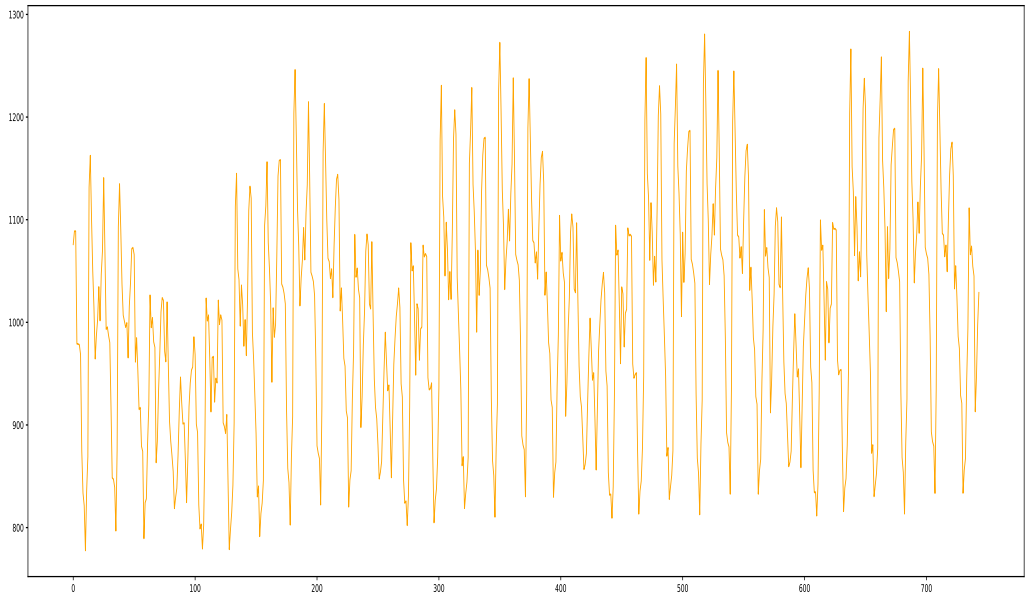


Figure 4: Previsioni Marzo 2005

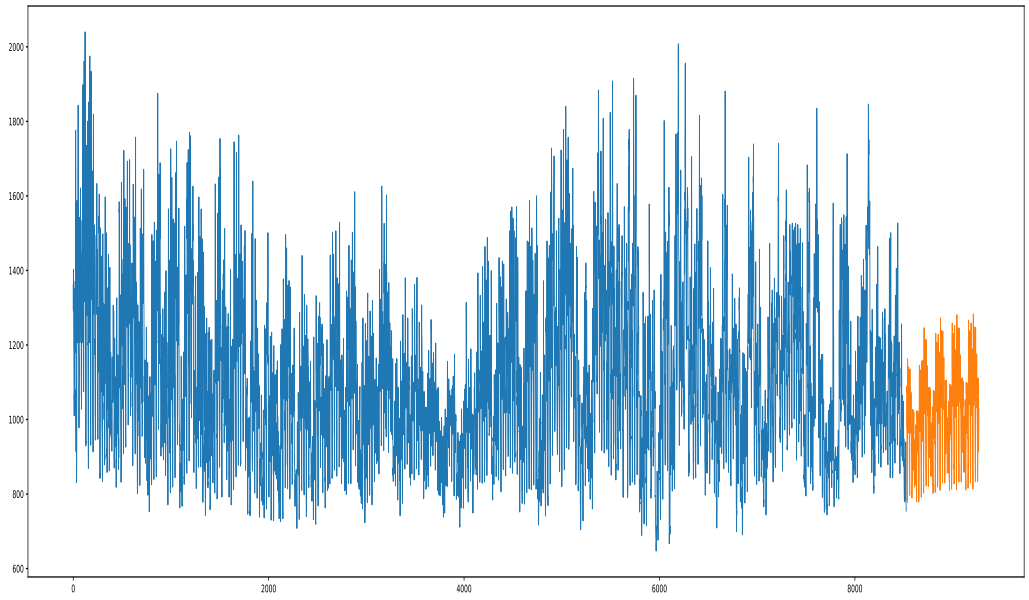


Figure 5: Serie completa