Progetto Streaming Data Management and Time Series Analysis

Alberto Raimondi

Il seguente elaborato riguarda il progetto svolto per l'esame di streaming data management e time series analysis: il dataset analizzato è Appliance Energy Prediction Dataset in cui il task consiste nel predire i consumi energetici futuri di elettrodomestici e luci domestiche.

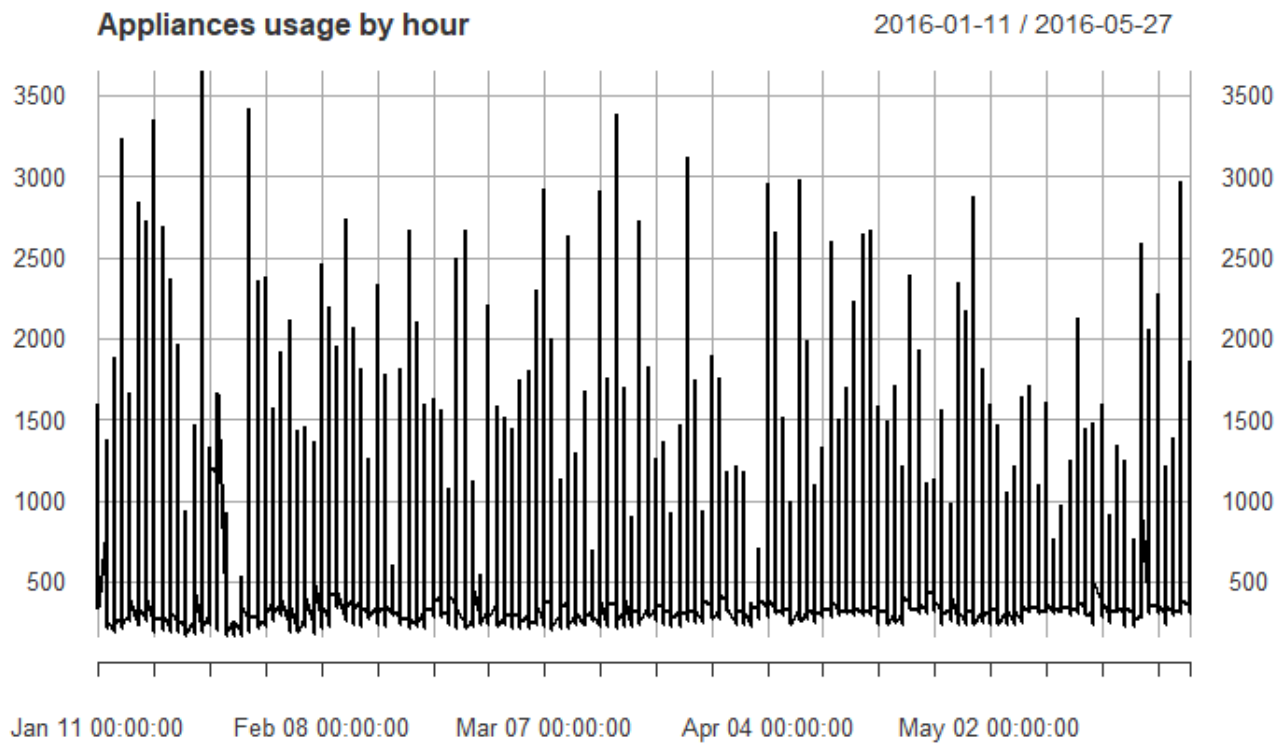
Per prima cosa si procede all'osservazione del dataset e si nota che abbiamo un dataset indicizzato per data con due target variables (appliances e lights) e 26 variabili riguardanti le condizioni metereologiche al tempo della misurazione.

Essendo due le serie storiche da predire procediamo modellandone una per volta, sarebbe possibile creare un modello unico per una regressione multivariata ma per facilitare il confronto con modelli di tipo RNN si è deciso di creare modelli univariati.

Si è deciso inoltre di considerare solamente le osservazioni sul consumo dei giorni per i modelli UCM e ARIMA scartando le altre variabili indipendenti.

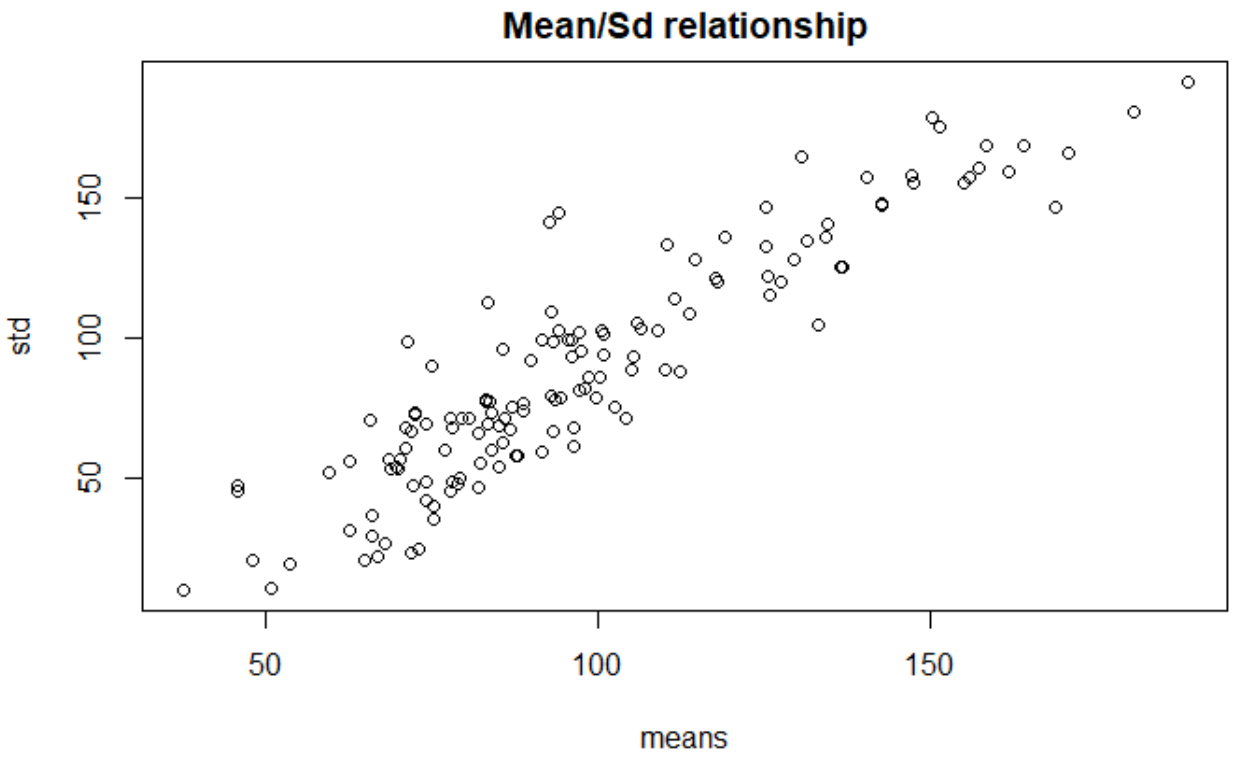
Appliances

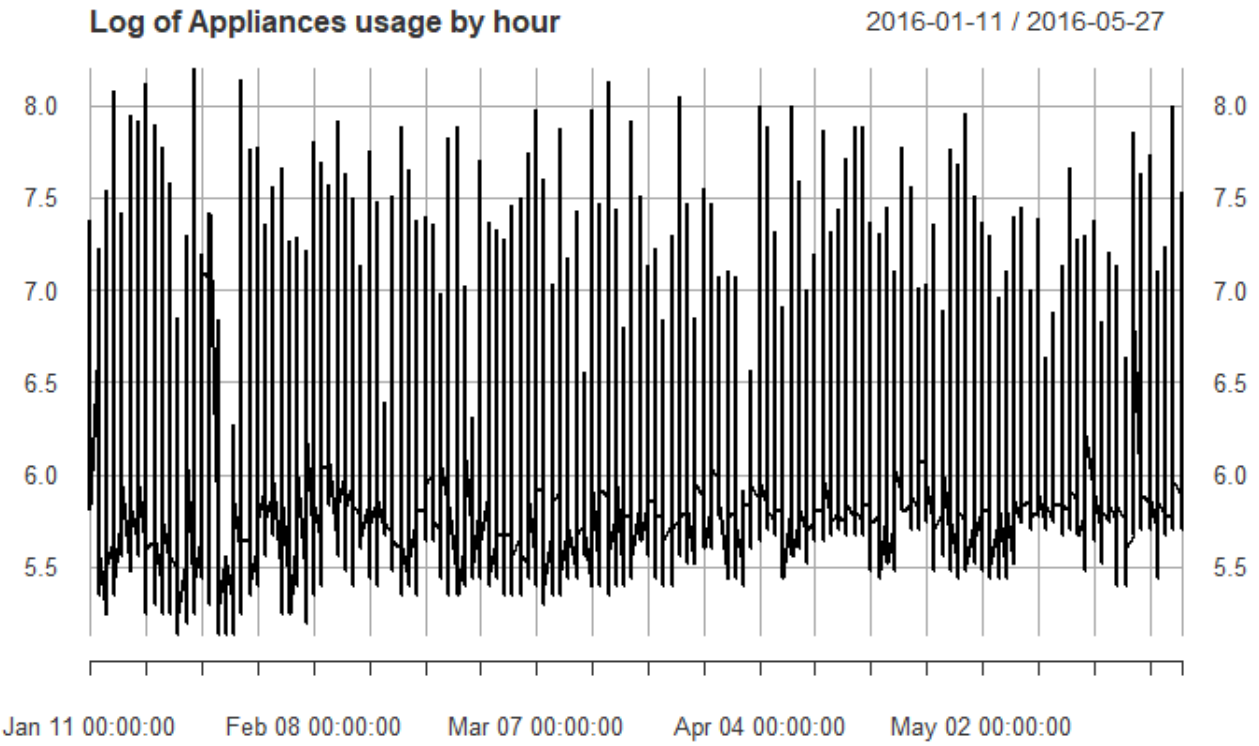
Iniziamo dall'analisi del consumo degli elettrodomestici, come consigliato nella consegna aggreghiamo i dati in maniera oraria con una semplice somma.



la serie sembra non presentare un trend lineare e si nota come prevedibile una stagionalità che probabilmente sarà di 24 ore dato il tipo di dati trattati.

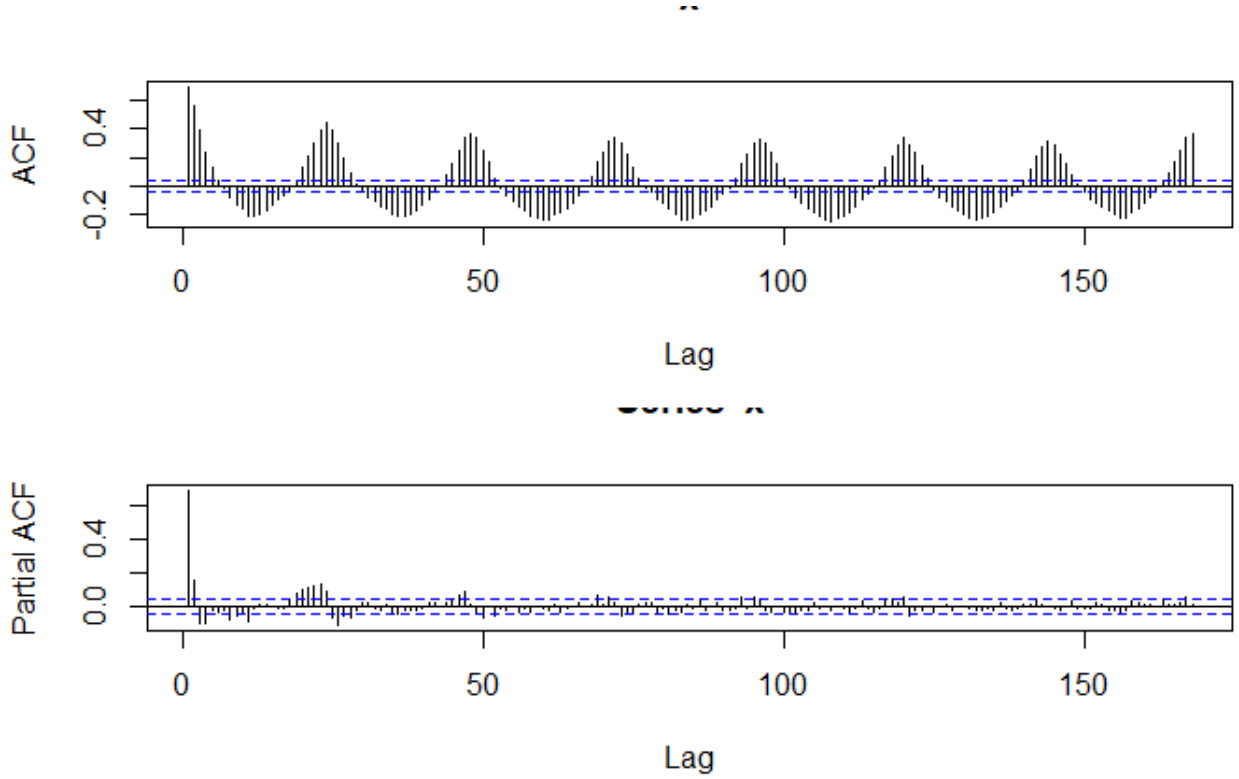
Se andiamo a visualizzare il rapporto tra le medie giornaliere e le varianze giornaliere possiamo notare una fortissima correlazione che ci porta ad utilizzare una trasformazione logaritmica prima di modellare la serie storica.





Come da consegna prima di provare a modellare la serie dividiamo i dati in train e test utilizzando l'11 Marzo 2016 come soglia per i dati di train allocando il resto a test set.

Osservando il grafico della cross correlation possiamo notare un evidente stagionalità di 24 ore, nel plot pacf possiamo invece vedere che c'è una lag di 2 osservazioni.



Modelli ARIMA

Modello 1 (ARIMA(2,0,0))

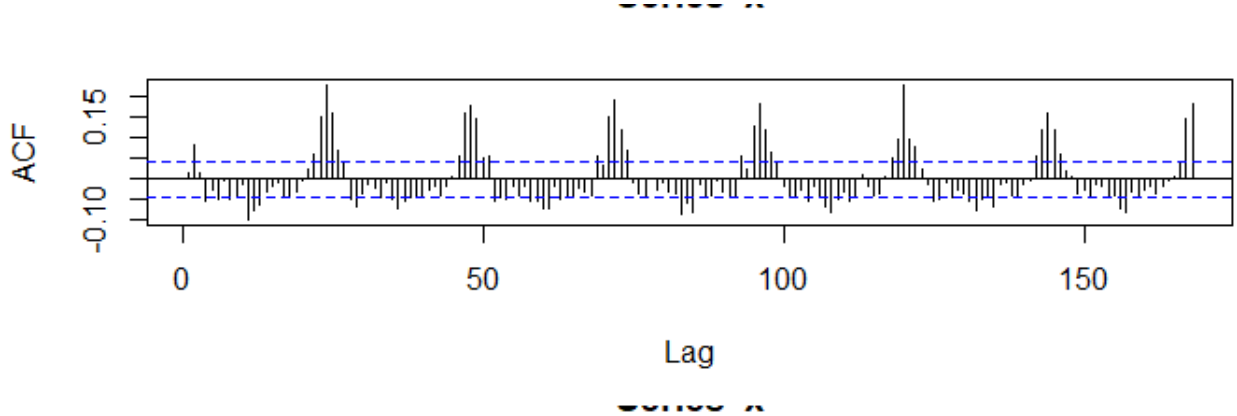
Proviamo a modellare la parte autocorrelata della serie attraverso un modello autoregressivo di ordine 2 e notiamo che i coefficienti sembrano essere significativi.

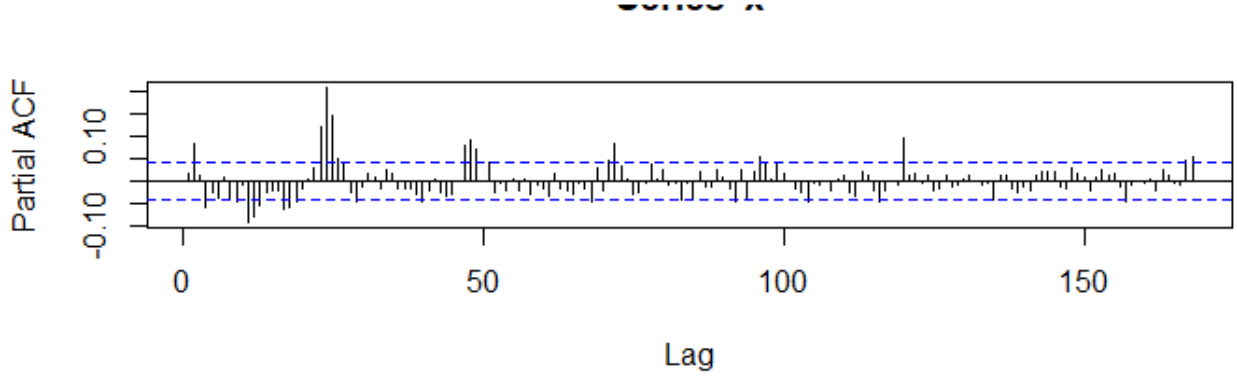
Osservando i grafici acf e pacf dei residui notiamo che la parte problematica del pacf non è più presente.

Coefficienti

coef	Estimate	Std. Error	z value	Pr(> |z|)
ar1	0.586455	0.021135	27.7475	< 2.2e-16
ar2	0.156734	0.021148	7.4114	1.25e-13
intercept	6.135876	0.037848	162.1209	< 2.2e-16

ACF/PACF Residui





Modello 2 (ARIMA(2,0,0), stagionalità (0,1,0))

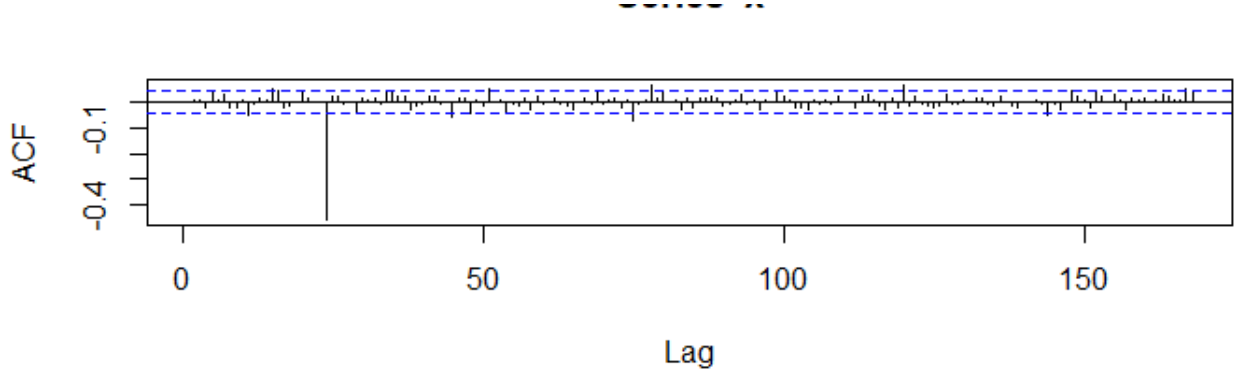
La serie presenta una ovvia stagionalità di 24 ore come prevedibile, tentiamo quindi di modellarla inserendo una componente stagionale integrata di ordine 1 nel nostro modello.

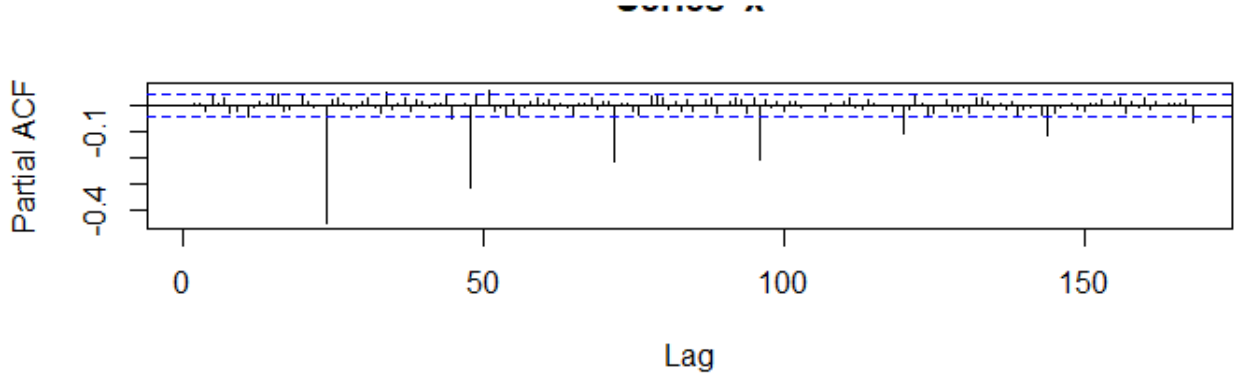
Vediamo che i coefficienti risultano significativi e i plot riguardanti i residui indicano una autocorrelazione con andamento discendente che fa presuppore una componente a media mobile.

Coefficienti

coef	Estimate	Std. Error	z value	Pr(> |z|)
ar1	0.407097	0.020885	19.492	< 2.2e-16
ar2	0.248592	0.020903	11.892	< 2.2e-16

ACF/PACF Residui





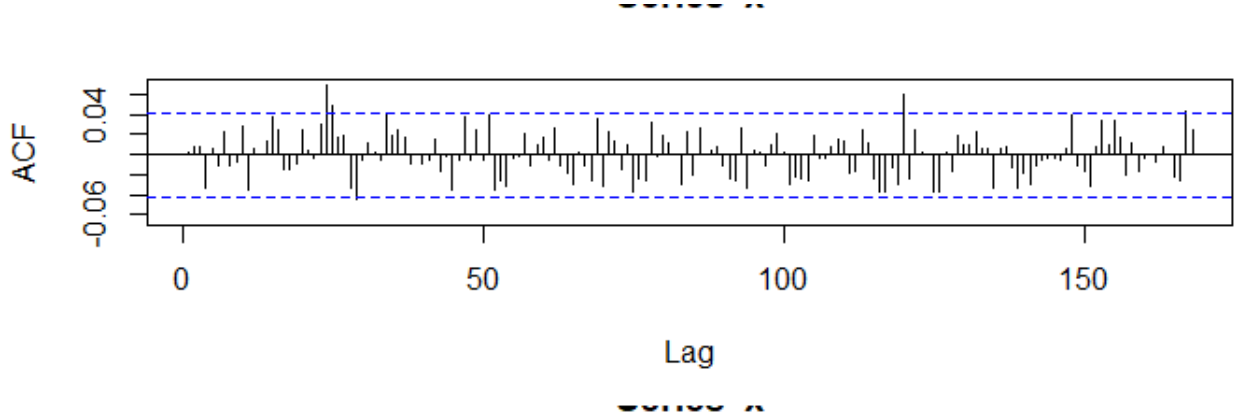
Modello 3 (ARIMA(2,0,0), stagionalità (0,1,1))

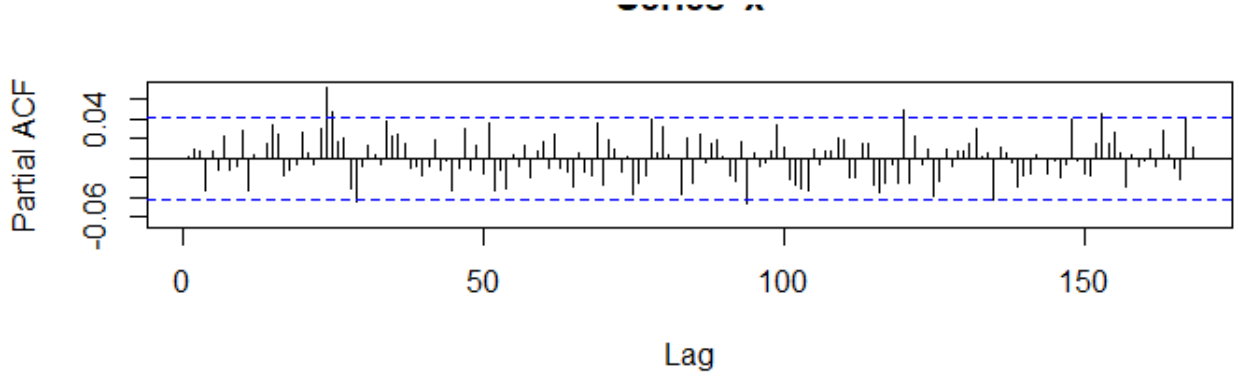
Il modello con la componente stagionale a media mobile è significativo tuttavia presenta una autocorrelazione all'osservazione 24 che proviamo a modellare aggiungendo una componente AR alla stagionalità.

Coefficienti

coef	Estimate	Std. Error	z value	Pr(> |z|)
ar1	0.446313	0.020962	21.292	< 2.2e-16
ar2	0.233309	0.020994	11.113	< 2.2e-16
sma1	-0.959166	0.011931	-80.391	< 2.2e-16

ACF/PACF Residui





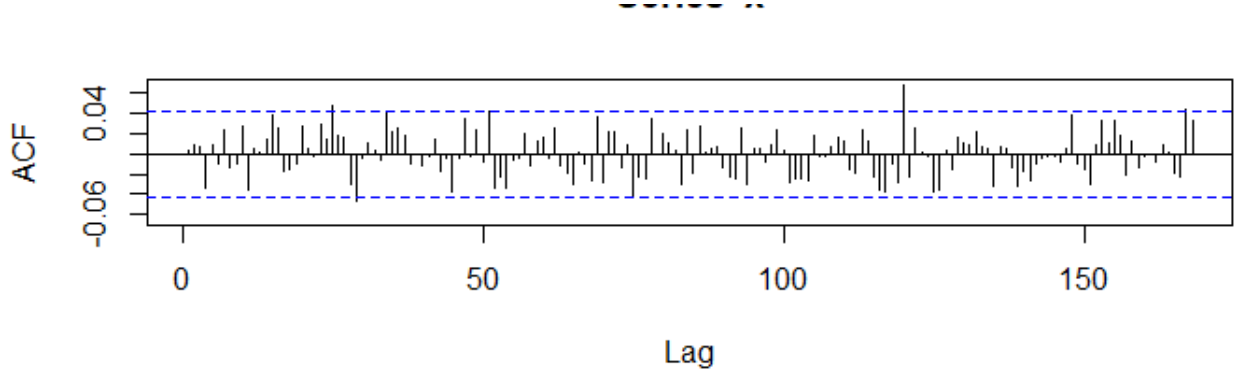
Modello 4 (ARIMA(2,0,0), stagionalità (1,1,1))

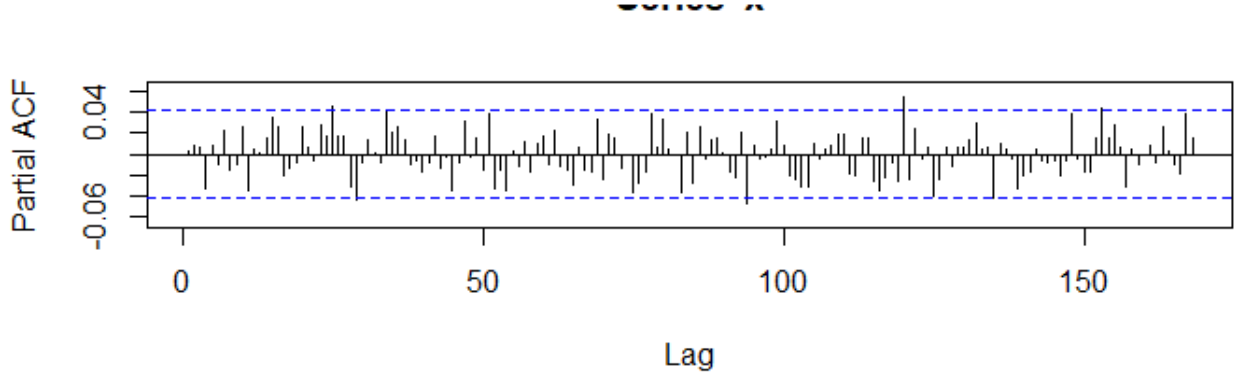
Il modello risulta significativo ed efficace secondo i plot di autocorrelazione, tuttavia è ancora presenta una anomalia allo step 24.

Coefficienti

coef	Estimate	Std. Error	z value	Pr(> |z|)
ar1	0.442165	0.021009	21.0469	< 2.2e-16
ar2	0.234046	0.020978	11.1567	< 2.2e-16
sma1	-0.959166	0.011931	-80.391	< 2.2e-16
sar1	-0.971885	0.012068	-80.5308	0.004444

ACF/PACF Residui





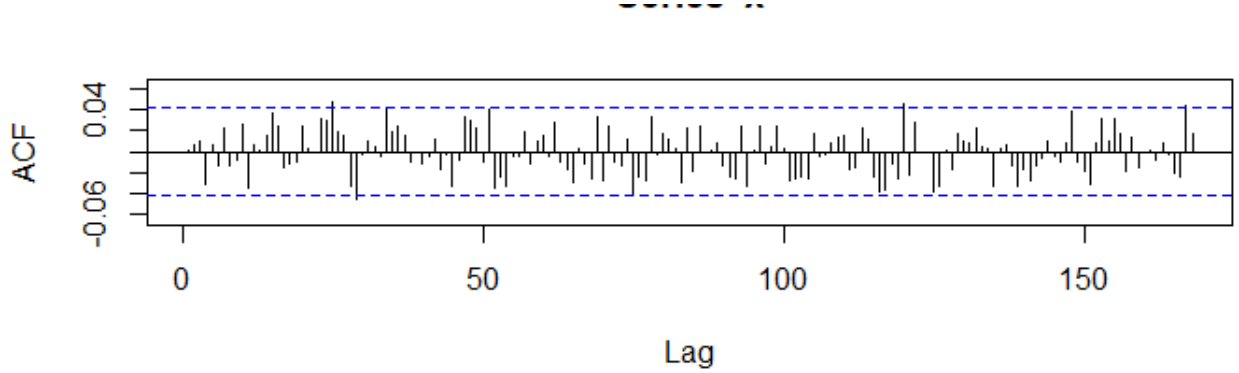
Modello 5(ARIMA(2,0,0), stagionalità (1,1,2))

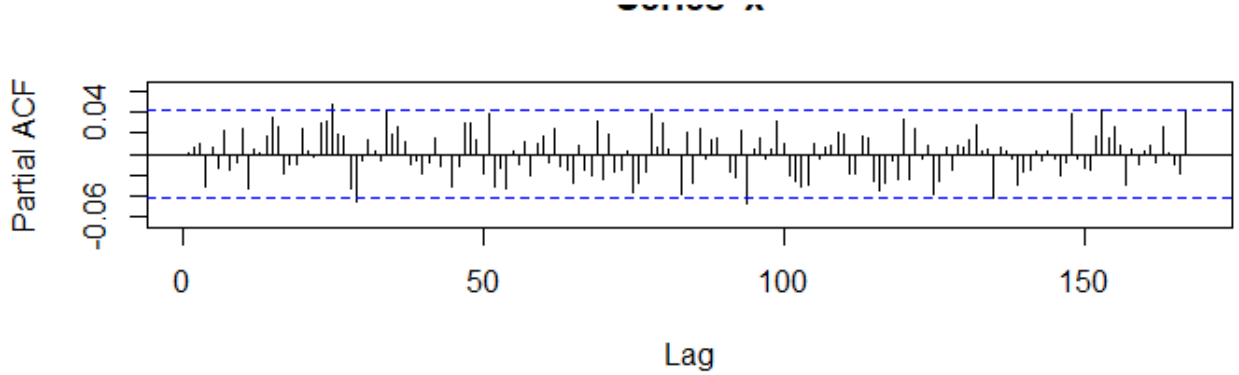
Provando un modello con una componente AR di ordine 2 vediamo che i grafici acf e pacf migliorano ma il modello perde di significatività, ci limitiamo quindi ad utilizzare il modello precedente.

Coefficienti

coef	Estimate	Std. Error	z value	Pr(> |z|)
ar1	0.444817	0.020978	21.2036	< 2.2e-16
ar2	0.232532	0.020999	11.0733	< 2.2e-16
sma1	-0.118671	0.062875	-1.8874	0.0591
sma2	-0.814952	0.058740	-13.8738	< 2.2e-16
sar1	-0.797361	0.069100	-11.5392	< 2.2e-16

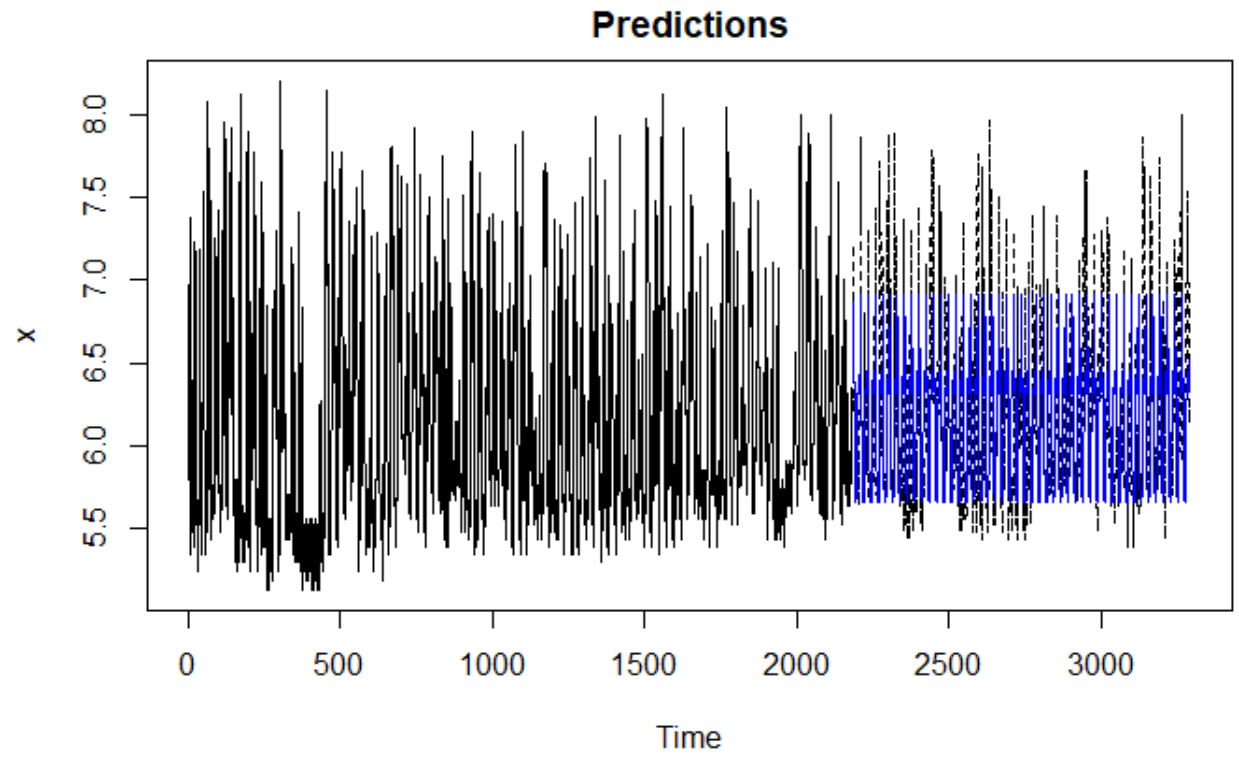
ACF/PACF Residui





Results

Come metrica per la valutazione è stato utilizzato il mean squared error in modo da penalizzare non linearmente l'errore delle previsioni, i risultati solo i seguenti.



Model MSE	Baseline MSE	Model MSE/Baseline
0.2083586	0.2965399	0.7026325

Notiamo un buon rapporto tra l'errore del nostro modello e quello baseline che va ad indicare che il nostro modello prevede la serie meglio che una modello baseline che preveda la media.

Modelli UCM

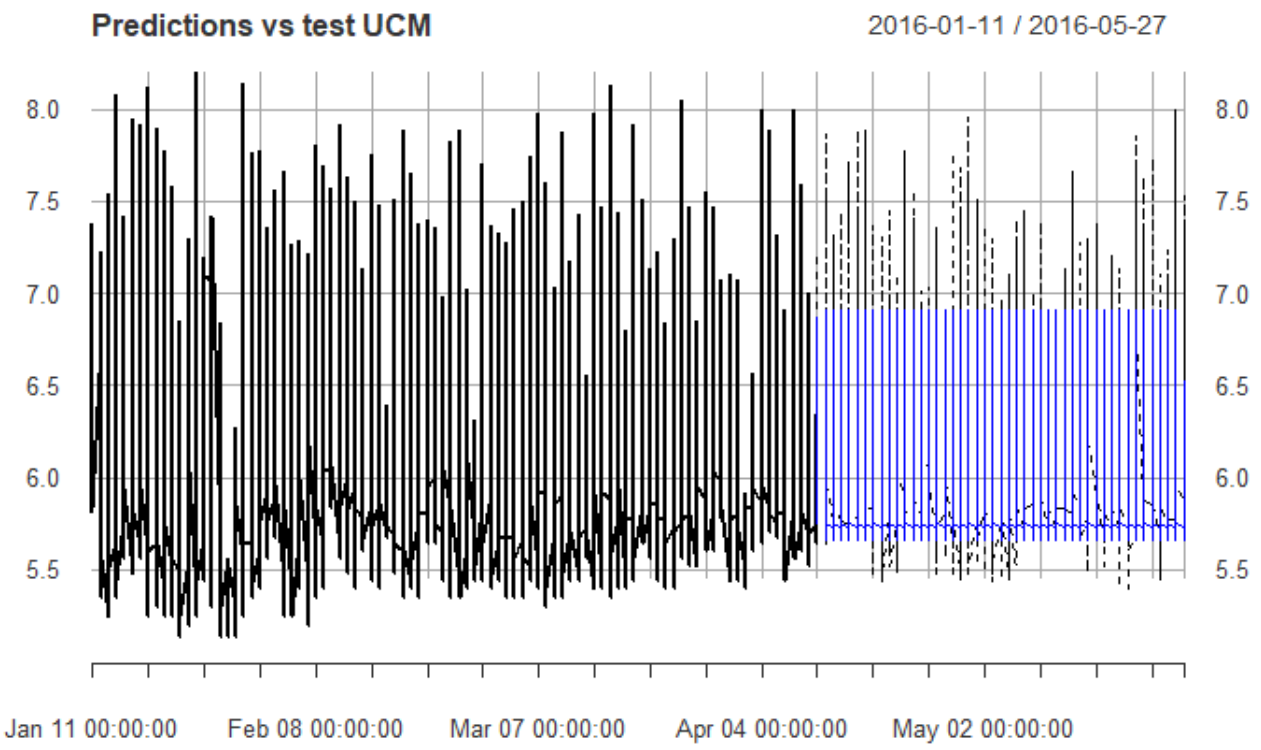
Proviamo ora ad utilizzare un modello a componenti inosservabili per modellare la serie storica. Con modelli UCM è necessario sostituire i dati nel train con tanti NA quanti sono i dati che vogliamo prevedere, perciò dopo aver svolto le operazioni di preprocessing andiamo a costruire un modello di previsione.

Modello UCM

Dato che abbiamo già osservato precedentemente la stagionalità della serie andiamo a costruire direttamente un modello con un trend a varianza nulla e una stagionalità di 24 ore modellata attraverso sinusoidi.

Dopo aver definito una funzione di update ed aver inizializzato le varianza a valori ritenuti opportuni(osservabili nel codice) proviamo a fittare il modello per trovare i parametri iniziali migliori, una volta raggiunta la convergenza passiamo il costrutto ad un filtro di Kalman che vada ad allenare il nostro modello per predire i dati.

Results



	Model MSE	Baseline MSE	Model MSE/Baseline
Train	0.06956837	0.4114526	0.1690799
Test	0.2152392	0.2965399	0.7258356

Il modello ovviamente è molto più performante sul train che sul test set, notiamo che comunque i risultati sul test set sono simili a quelli ottenuti utilizzando modelli ARIMA.

Modelli LSTM

Per i modelli di tipo LSTM l'approccio considerato è stato differente: questi modelli infatti hanno molta più difficolta ad analizzare rapporti tra dati a molti timesteps di distanza per via della loro instabilità e costo computazionale, si è deciso quindi di limitare il training di modelli ricorrenti a serie di 24 ore. Sono state inoltre aggiunte input anche le variabili indipendenti prima non utilizzate per provare ad ottenere un modello più performante.

Preprocessing

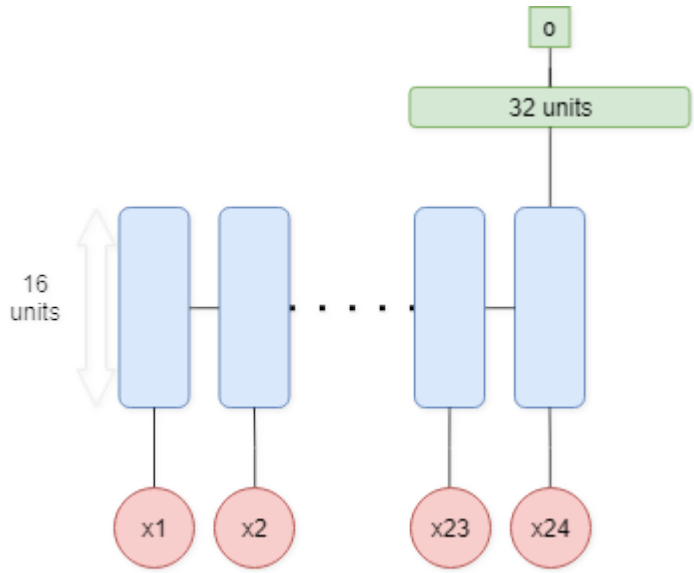
Il preprocessing per il modello ricorrente è stato simile a quello per gli altri modelli, è stata aggiunta tuttavia una variabile artificiale indicante il giorno della settimana in cui la misurazione ha luogo.

Prima di procedere all'allenamento del modello tramite LSTM tuttavia è stato necessario andare a trasformare la serie storica in un tensore 3D in cui ognuna delle precedenti righe va a corrispondere ad una matrice contenente le 24 osservazioni dei giorni precedenti con le relative variabili indipendenti.

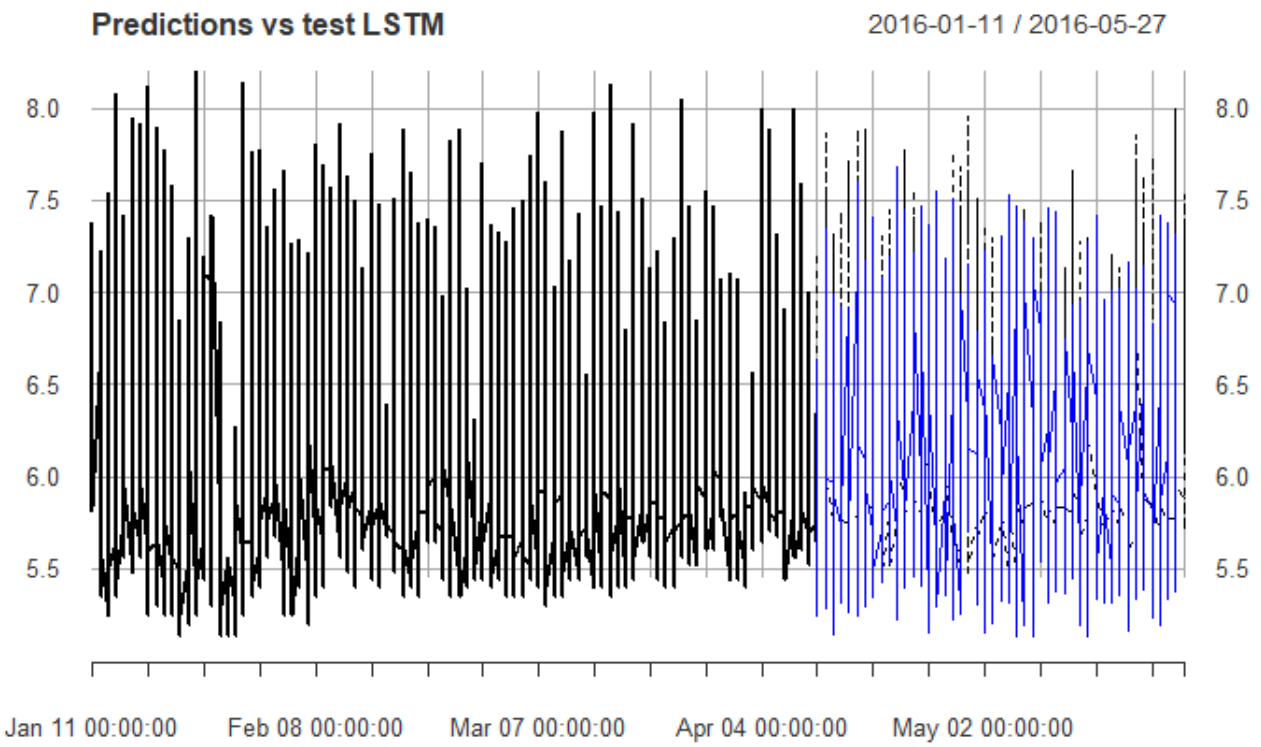
Modello

Il modello è una semplice rete neurale ricorrente che fa uso di celle LSTM per facilitare la propagazione del gradiente, la rete è di tipo *many to one* che significa che per ogni sequenza di input andiamo a prevedere un solo output (per ogni 24 ore prevediamo l'osservazione successiva). Oltre la parte ricorrente utilizziamo un livello fully connected con attivazione ReLU ed un singolo ouput semi lineare (utilizziamo un attivazione relu perchè i consumi non possono essere negativi).

La loss utilizzata è il mean squared error dato che è la metrica che vogliamo minimizzare ed il modello è allenato per 100 epoche con ottimizzatore adam.



Results



	Model MSE	Baseline MSE	Model MSE/Baseline
Train	0.1695	0.3675	0.4612
Test	0.2467	0.3812	0.6473

I risultati utilizzando un modello ricorrente sono significativamente migliori rispetto ai modelli precedenti, è interessante notare inoltre quanto le previsioni fatte da questo modello siano reattive agli input rispetto ai modelli ARIMA e UCM.

Lights

Essendo la variabile riguardante il consumo di elettricità per l'illuminazione molto simile a quella per il consumo degli elettrodomestici proviamo a modellarla con le stesse modalità. Notiamo che otteniamo ancora dei risultati soddisfacenti per i modelli ARIMA e LSTM ma non per modelli UCM.

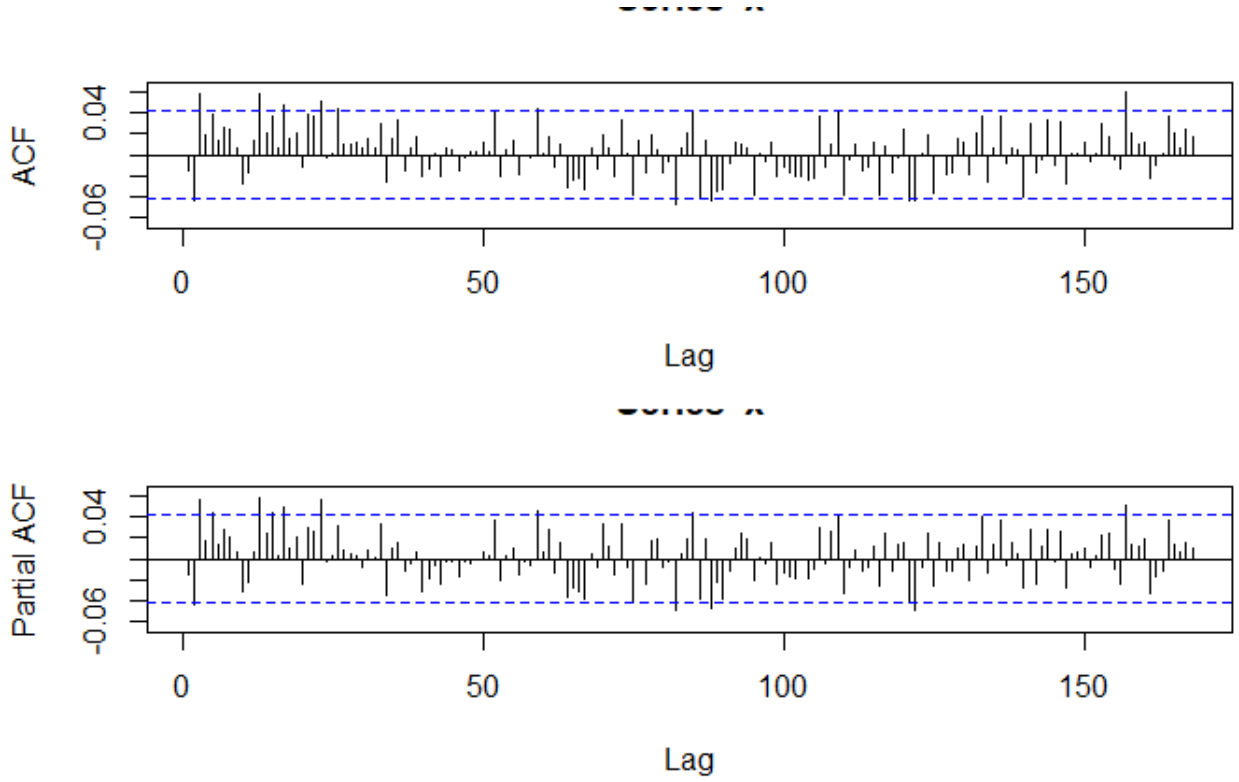
Si rimanda al codice per ulteriori dettagli.

ARIMA(2,0,0), stagionalità (1,1,1))

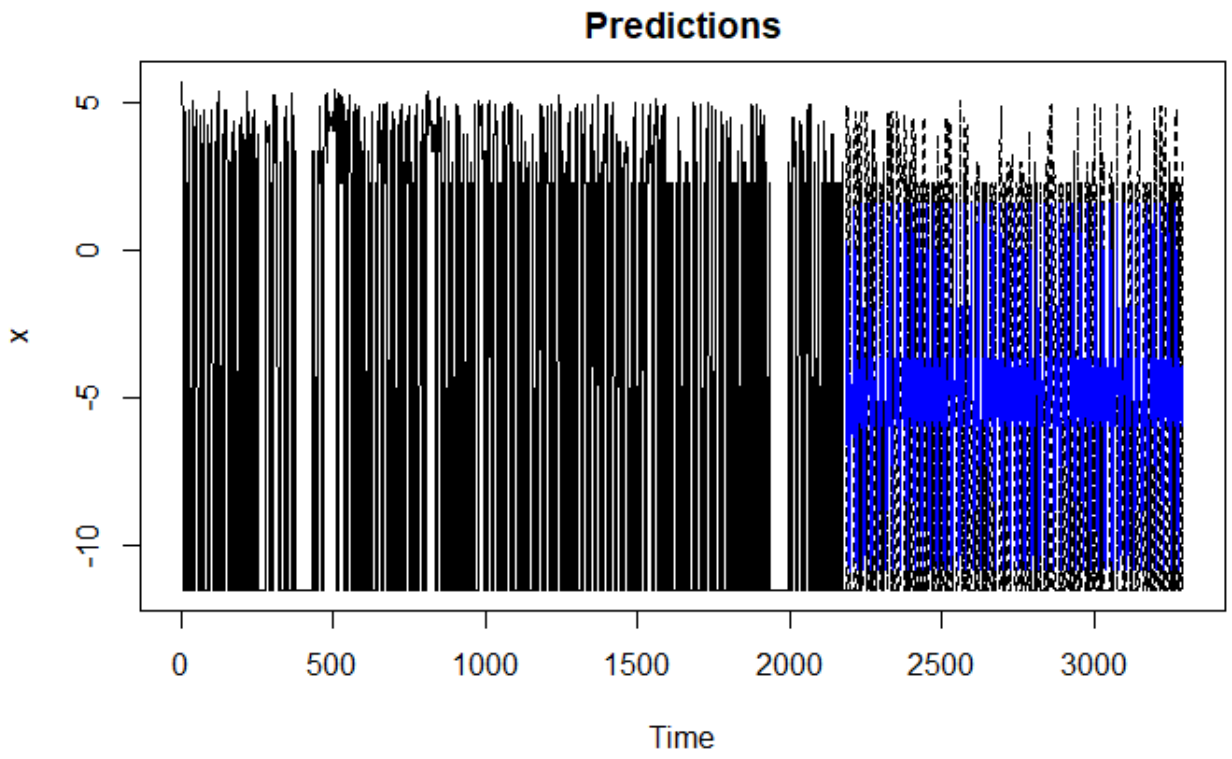
Coefficienti

coef	Estimate	Std. Error	z value	Pr(> |z|)
ar1	0.401685	0.021272	18.8837	< 2.2e-16
ar2	0.157652	0.021294	7.4037	1.325e-13
sma1	-0.966804	0.011629	-83.1378	< 2.2e-16
sar1	0.071891	0.023439	3.0672	0.002161

ACF/PACF Residui

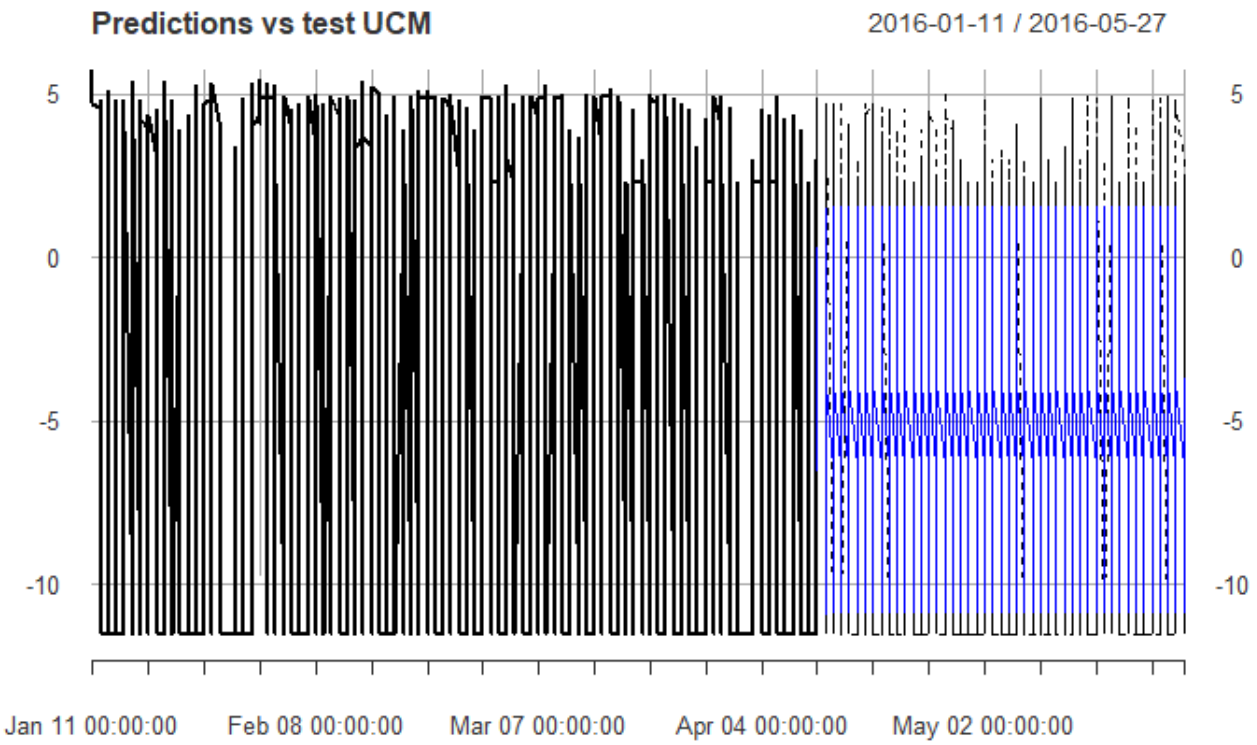


Results



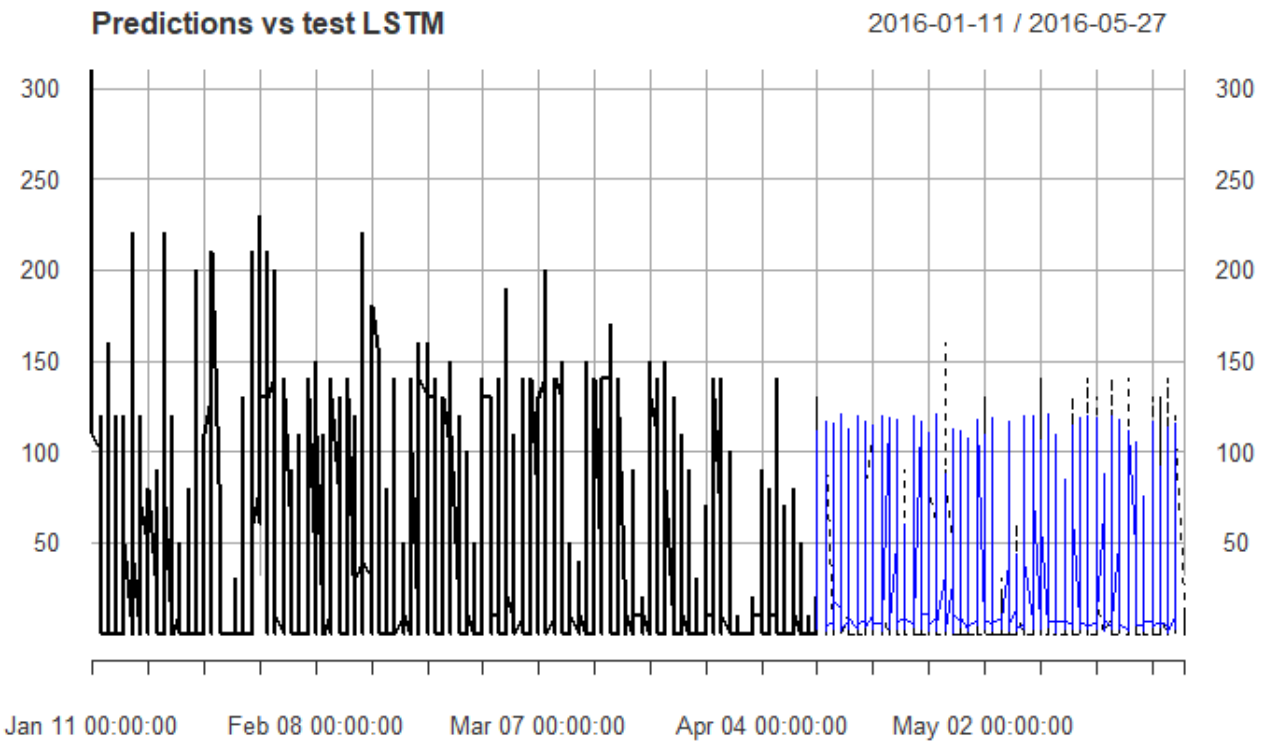
Model MSE	Baseline MSE	Model MSE/Baseline
47.725	54.006	0.884

UCM



	Model MSE	Baseline MSE	Model MSE/Baseline
Train	16.312	57.357	0.2844
Test	68.25413	54.00681	1.263806

Modello LSTM



	Model MSE	Baseline MSE	Model MSE/Baseline
Train	496.18	1713.13	0.2896
Test	1720.14	786.1	0.4570