Previsione serie storica consumo elettrico

Andrea Lucini Paioni - n° matricola 826578

Appello di settembre 2023

Indice

Introduzione

Le previsioni di serie temporali sono un processo di analisi di dati, relativi a serie temporali, utilizzando statistiche. metodologie e modellazione per ottenere previsioni col fine di informare il processo decisionale strategico di un'azienda/ente/servizio/privato. Sebbene non si tratti sempre di previsioni esatta al 100%, e infatti la probabilità di previsione tende a variare notevolmente, la previsione delle serie temporali può fornire spunti fondamentali per valutare la tendenza futura più probabile del fenomeno di interesse: si tratta di un approccio sempre più importante, utilizzato in svariati ambiti economici e industriali

Questo progetto, dunque, pone come obiettivo l'implementazione di una serie di algoritmi col fine di prevedere valori di una serie storica: si tratta di una serie in cui sono raccolte le rilevazioni di monossido di carbonio (CO), in un periodo compreso tra il 01/01/2017 alle ore 00:00:00, e il 30/12/2017 alle ore 23:50:00: tuttavia il dataset a nostra disposizione comprende i dati fino al 30/11/2017, dunque escludendo l'ultimo mese col fine di prevederne i valori. Per questo scopo sono stati considerati algoritmi di tre categorie differenti, ARIMA, UCM e machine learning, in modo da individuare i modelli dalle migliori performance predittive, considerando il Mean Absolute Error (MAE), che indica l'errore assoluto medio, come la metrica da considerare.

Il Mean Absolute Error viene calcolato nel seguente modo: $MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$

Import ed esplorazione

Il dataset fornito è formato da 48096 osservazioni con due colonne, rispettivamente:

- **date**: indica data e ora in cui è stato osservato il consumo di corrente, in formato *dd/mm/yyyy HH:MM:SS*;
- **power**: indica il consumo di corrente rilevato all'orario di interesse.

Le misurazioni sono comprese in un periodo che va dal 01/01/2017 alle ore 00:00:00 al 30/11/2017 alle ore 23:50:00, con una misurazione ogni 10 minuti per tutto l'intervallo temporale di interesse. Il numero di valori da prevedere è pari a 4464, ovvero un'osservazione ogni 10 minuti (6 all'ora), per tutto il giorno (24 ore), in tutto il mese di dicembre (31 giorni).

Di seguito sono riportate, in ordine: la serie temporale originale, l'andamento giornaliero della serie, l'andamento settimanale della serie e l'andamento mensile.

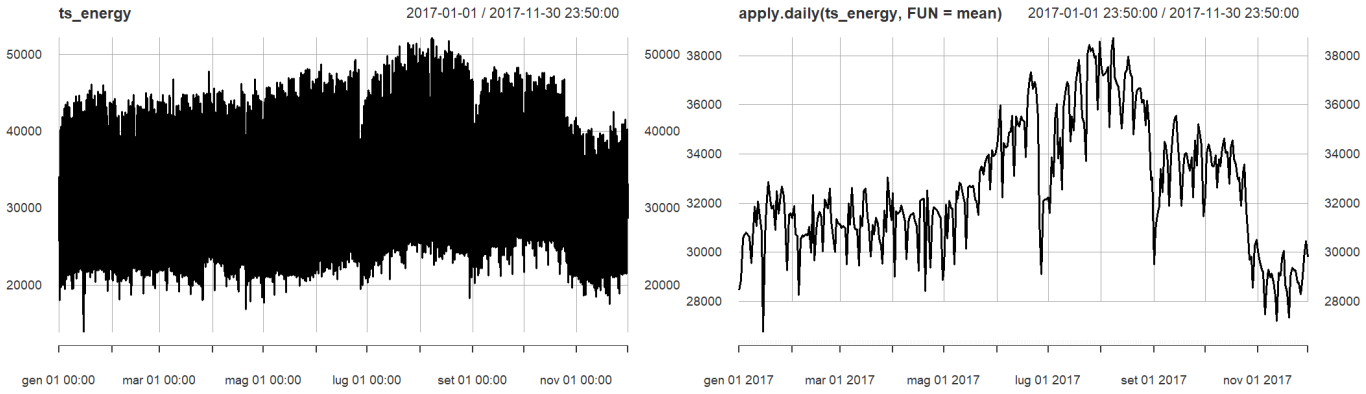

Figure 1: a sinistra, la serie temporale originale; a destra, l'andamento giornaliero.

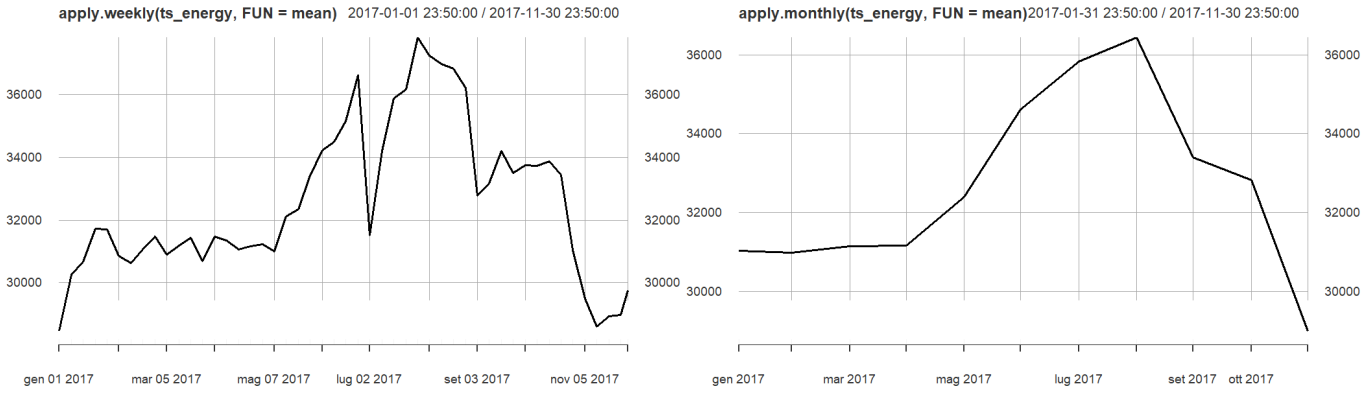

Figure 2: a sinistra, l'andamento settimanale; a destra, l'andamento mensile.

È facile notare, dall'andamento mensile e settimanale soprattutto, valori che tendono ad essere stabili nei primi mesi dell'anno per poi aumentare sensibilmente con l'inizio dell'estate (probabile conseguenza dell'aumento delle temperature e dunque dell'utilizzo massiccio di condizionatori); i valori raggiungono il picco intorno ad agosto, per poi calare con l'arrivo dell'autunno. In generale, si inizia ad osservare sia la presenza di trend stagionale, sia una componente mensile, da analizzare ulteriormente.

Osserviamo ora un boxplot e un istogramma della variabile power, in modo da capire com'è distribuita:

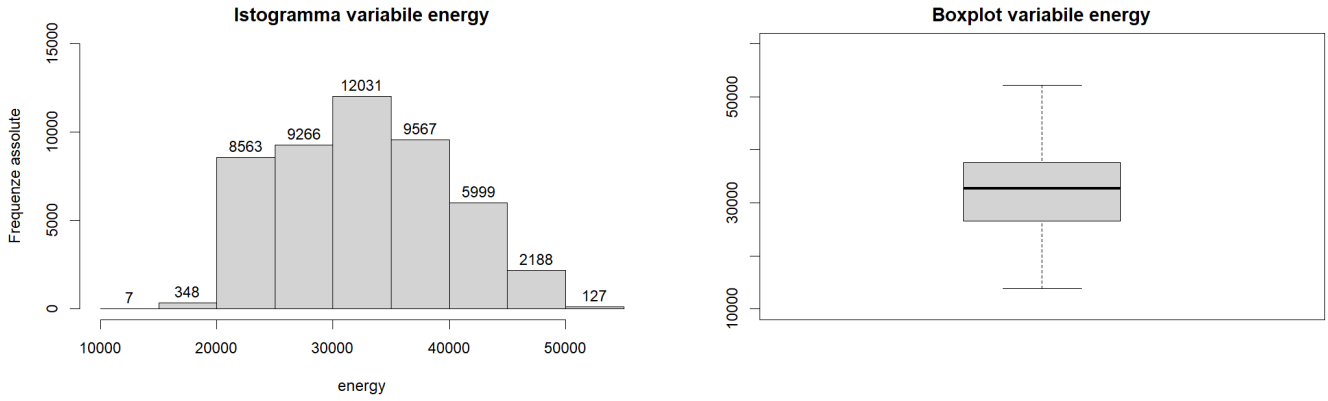

Figure 3: a sinistra, l'istogramma della variabile energy, a destra il boxplot della variabile energy

Il valore minimo è di 13896, il massimo di 52204, e i valori di media e mediana sono molto simili (rispettivamente 32643 e 32651). Notiamo una distribuzione dei valori, nell'istogramma, che può ricordare una normale, con una leggera asimmetria verso sinistra. Inoltre, si possono notare anche alcune situazioni interessanti osservando il consumo medio nei singoli giorni della settimana e nei mesi dell'anno, come dai seguenti boxplot:

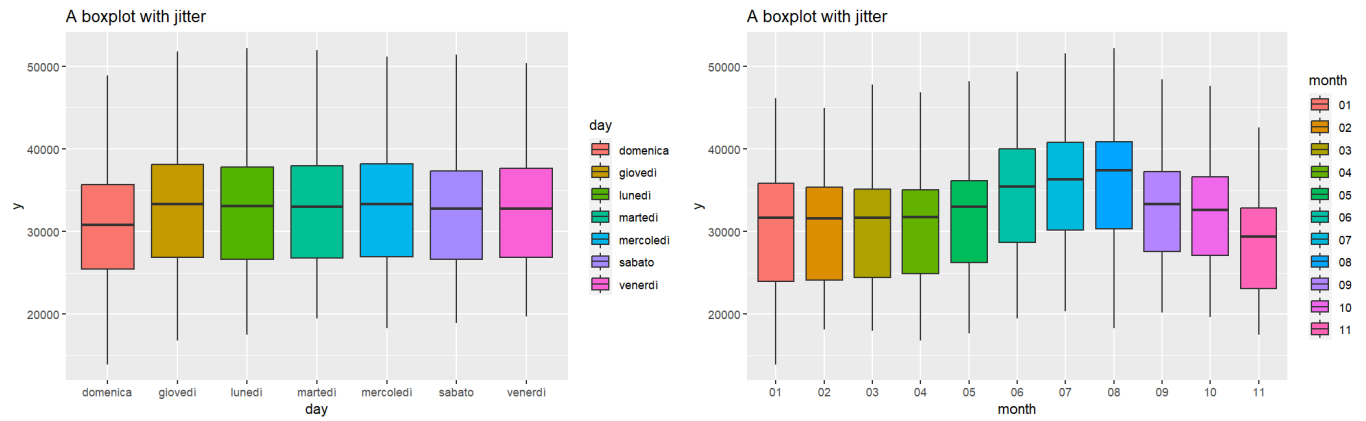

Figure 4: a sinistra, boxplot variabile energy in base al giorno; a destra, boxplot variabile energy in base al mese.

L'utilizzo di corrente sembra essere costante all'interno della settimana, tranne che la domenica, per cui si registrano valori leggermente più bassi (media attorno ai 30000 rispetto alle medie tra i 32000 e i 33000 per gli altri giorni). Viene confermato il discorso fatto con l'andamento mensile, ovvero con valori costanti più o meno per tutto l'anno fino all'inizio dell'estate, con giugno e soprattutto luglio e agosto con i valori più elevati.

Infine, osserviamo l'andamento di una singola settimana e di un singolo giorno, per verificare se sono presenti ulteriori trend:

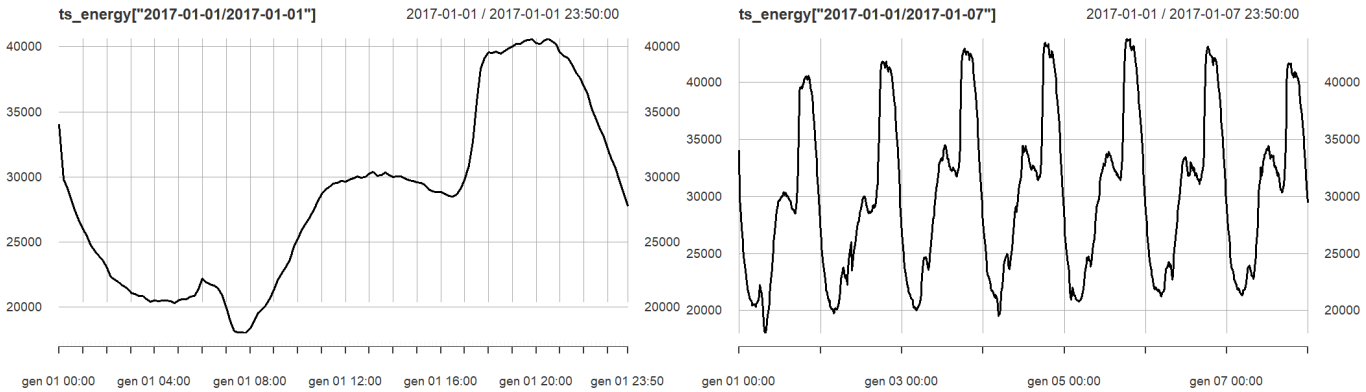

Figure 5: a sinistra, andamento giornaliero del 01/01/2017; a destra, andamento della prima settimana del 2017.

Subito si può notare un calo del consumo di notte, e un aumento durante il giorno, fino ad arrivare al picco della sera (momento in cui verosimilmente il maggior numero di persone è a casa e utilizza la corrente): il trend si ripete non solo per la prima settimana del 2017, ma per tutto l'anno.

Infine, creiamo una divisione del dataset tra training dataset (compreso tra il 01/01/2017 alle 00:00:00 e il 31/10/2017 alle 23:50:00) e validation dataset (dal 01/11/2017 alle 00:00:00 al 30/11/2017 alle 23:50:00); l'obiettivo della previsione sarà di individuare i valori compresi tra il 01/12/2017 alle 00:00:00 e il 30/12/2017 alle 23:50:00.

ARIMA

I modelli ARIMA sono un punto di riferimento nell'ambito dello studio di serie storiche: appartengono alla famiglia dei processi stocastici lineari non stazionari, e sono un'estensione dei modelli ARMA. L'acronimo ARIMA sta per Auto-Regressive Integrated Moving Average:

- AR sta per modellazione autoregressiva;
- I sta per l'integrazione della serie storica, usata in caso ci trovi di fronte ad una serie non stazionaria (la differenza che contraddistingue un modello ARMA, senza integrazioni, da un modello ARIMA);
- MA sta per Moving Average, o modellazione a media mobile.

Per ognuna delle 3 componenti, dunque, viene stabilito un parametro, che definisce rispettivamente:

- Il numero di lag della componente autoregressiva (definito con la lettera *d*);
- Il numero di integrazioni (definito con la lettera *d*);
- Il numero di lag della componente moving average (definito con la lettera *q*).

I modelli ARIMA richiedono che la serie sia stazionaria, dunque si procede in quest'ottica per rendere la serie, attualmente non stazionaria (data la presenza di diversi trend, come quello stagionale e giornaliero), come stazionaria. Verifichiamo l'andamento di media e varianza della serie, e l'eventuale presenza di correlazione tra essi:

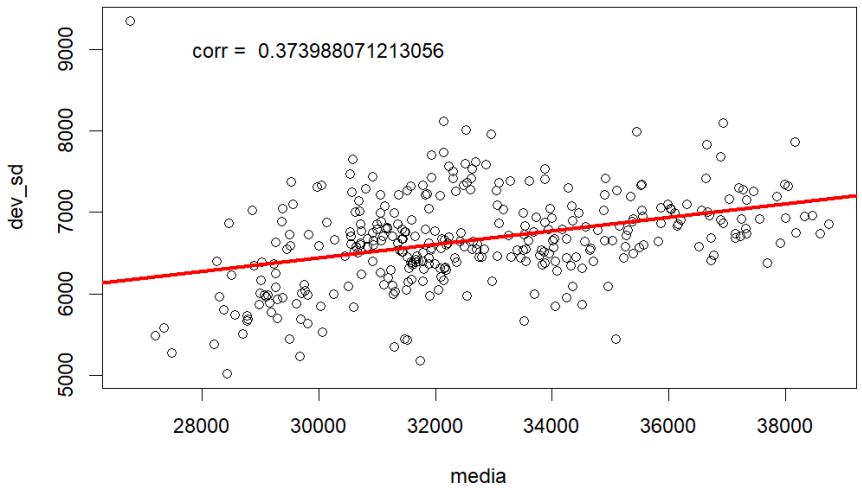

Figure 6: correlazione tra media e deviazione standard della serie temporale

Si può notare dalla Figura 6 una leggera correlazione (valore pari a 0.37) tra media e deviazione standard nella distribuzione (osservando le osservazioni raggruppate per ogni giorno). Inoltre, osserviamo un outlier con valore molto basso della media ma valore della deviazione standard elevatissimo rispetto al resto della distribuzione. Proviamo ad applicare sia una trasformazione di Box-Cox (il valore della lambda osservato col metodo è pari a 0.22), sia una trasformazione logaritmica per risolvere il problema della stazionarietà in varianza.

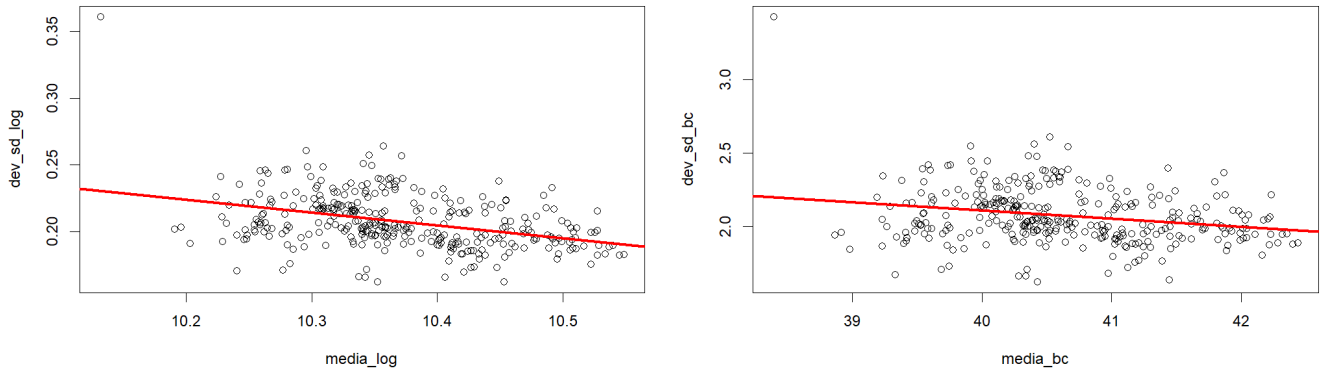

Figure 7: a sinistra, correlazione tra media e deviazione standard della serie temporale con trasformazione logaritmica; a destra, correlazione tra media e deviazione standard della serie temporale con trasformazione di Box-Cox.

Decidiamo di considerare la serie temporale trasformata per i modelli ARIMA.

Una serie di modelli ARIMA sono stati testati sulla serie temporale (trasformata con trasformazione logaritmica): valutando i risultati del MAE ottenuti, è stato considerato in particolare il modello ARIMA(3,0,0)(0,1,0)[144]: il modello raggiunge un valore del MAE pari a 1215.022, e risulta essere il migliore tra quelli testati, nonostante dai grafici di ACF e PACF sembra esserci ancora qualche potenziale problema di stagionalità:

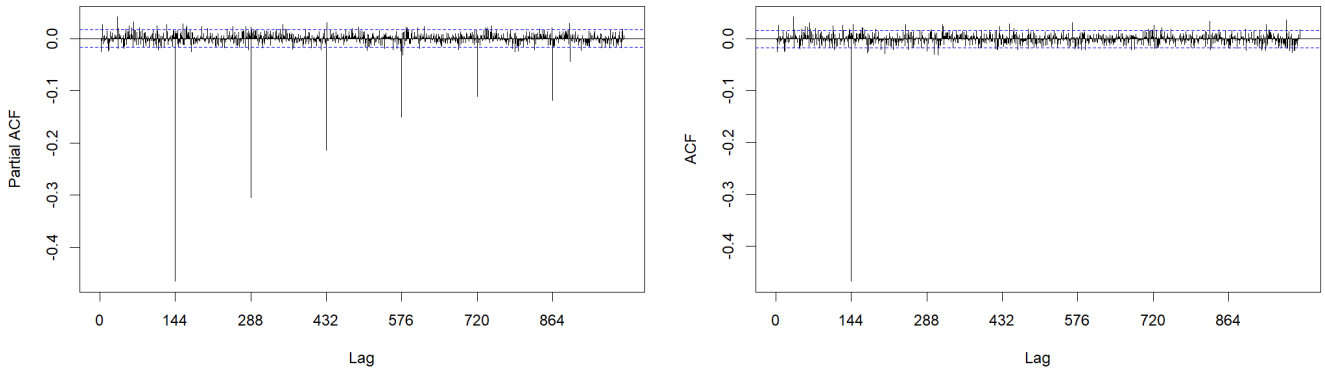

Figure 8: a sinistra Partial ACF del modello ARIMA(3,0,0)(0,1,0)[144]; a destra ACF del modello ARIMA(3,0,0)(0,1,0)[144].

Questi sono i risultati delle previsioni ottenute sul mese di novembre 2017:

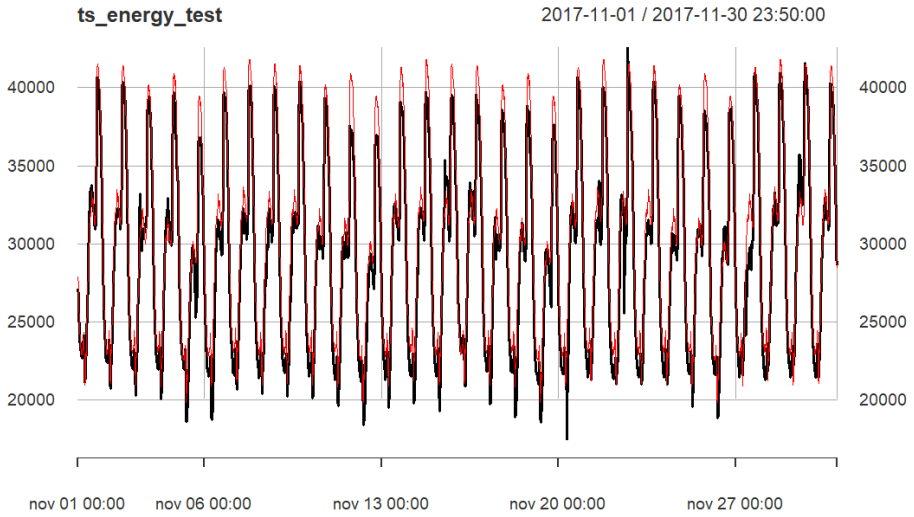

Figure 9: in rosso le previsioni ottenute dal modello ARIMA(3,0,0)(0,1,0)[144], a partire dalla serie temporale originale (in nero).

Tra i modelli ARIMA, è stata inoltre utilizzata anche la funzione auto.arima(), che permette di testare in successione i parametri in modo da ottenere il modello di serie temporale migliore: il modello restituito dalla funzione è il seguente ARIMA(4,0,0)(0,1,0)[144]. Il modello, tuttavia, sembra mostrare ancora dei pattern, come è possibile notare dai grafici per l'acf e il pacf seguenti:

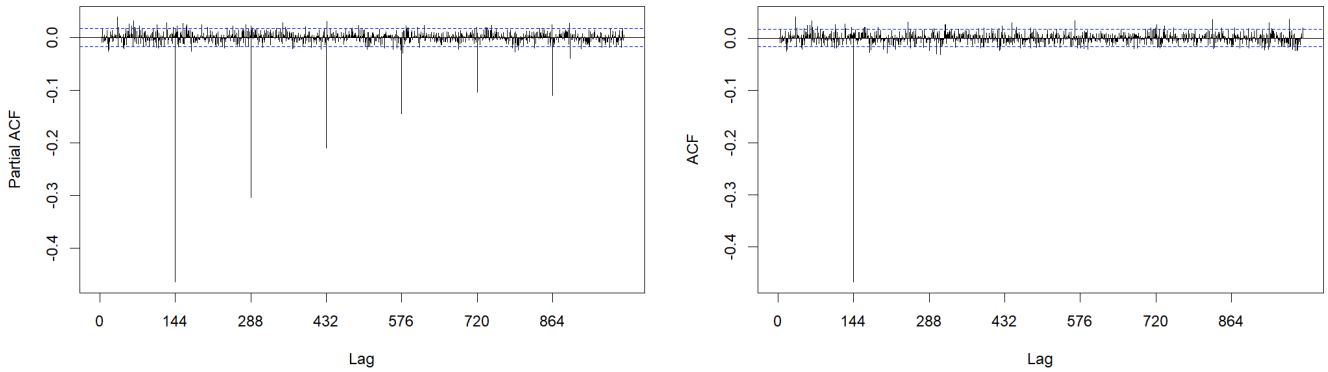

Figure 10: a sinistra Partial ACF del modello ARIMA(4,0,0)(0,1,0)[144]; a destra ACF del modello ARIMA(4,0,0)(0,1,0)[144].

Inoltre, il valore del MAE ottenuto è pari a 29976.77, dunque molto più elevato del valore del modello ottenuto col metodo ARIMA, dunque per le previsioni dei modelli ARIMA consideriamo il primo modello indicato, quello del tipo ARIMA(3,0,0)(0,1,0)[144].

UCM

Una possibile alternativa ai modelli ARIMA sono i modelli UCM, o Unobserved Component Models: si tratta di una classe di modelli che permette decomporre le serie temporali come somma di una serie di componenti non direttamente osservabili (come stagionalità, white noise, trend, ciclo…).

Tra i vari modelli testati, quello effettivamente considerato è formato da un trend LLT (Local Linear Trend), sotto forma di SSMtrend di ordine 2 e dalle due stagionalità trigonometriche (giornaliera e settimanale), ovvero rispettivamente SSMseasonal giornaliera con una sinusoide di 10 armoniche e SSMseasonal settimanale con una sinusoide di 10 armoniche.

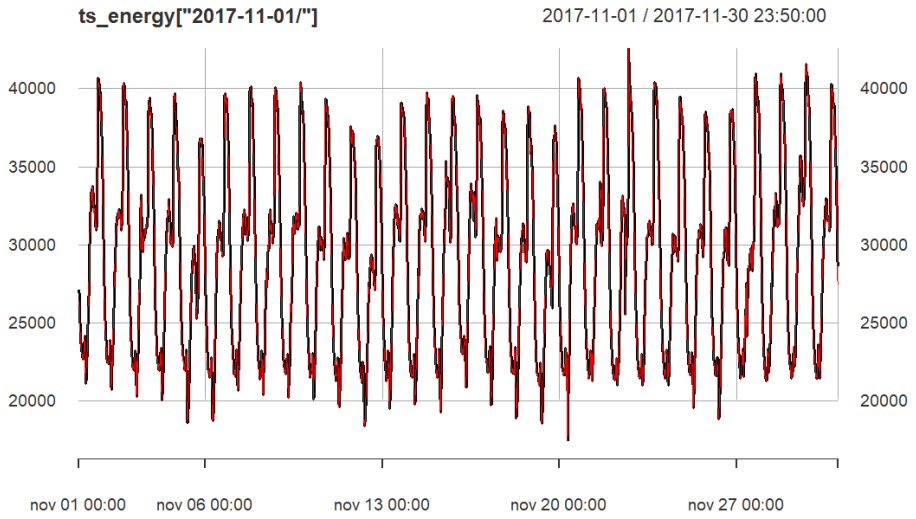

Figure 11: in rosso le previsioni ottenute dal modello UCM, a partire dalla serie temporale originale (in nero).

Il modello ha avuto prestazioni molto importanti, con valore del MAE pari a 688.3647.

Machine Learning

Una terza classe di modelli per le serie temporali è quella dei modelli di machine learning: si tratta di modelli in grado di apprendere direttamente dalla tendenza dei dati presenti per formulare previsioni attendibili, senza definire delle componenti o dei parametri in particolare. Per questo motivo, ad esempio, non è necessario definire a priori l'eventuale presenza di stagionalità nei dati, perché questa viene intercettata dall'algoritmo.

Tra i modelli considerati, quello delle KNN, o dei k-nearest neighbors, è il modello di machine learning che ha avuto prestazioni migliori:

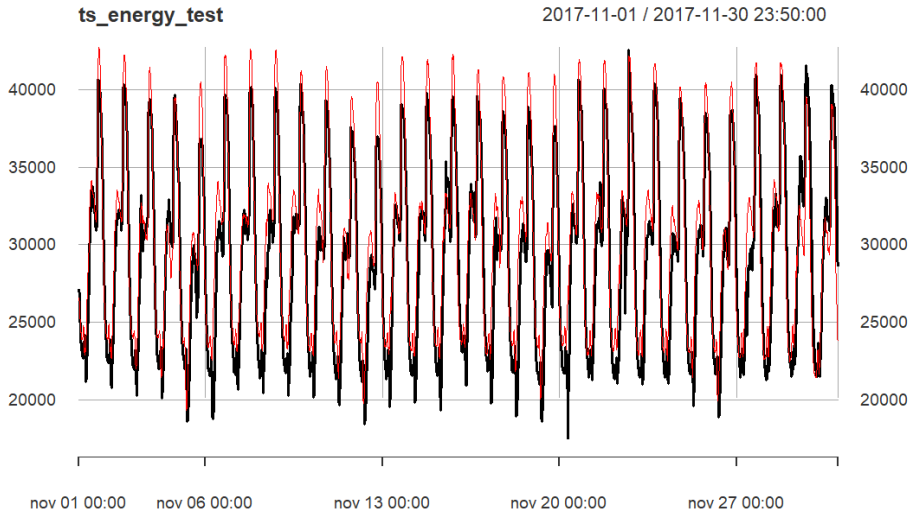

Figure 12: in rosso le previsioni ottenute dal modello di machine learning, a partire dalla serie temporale originale (in nero).

Tuttavia, il valore del MAE ottenuto risulta essere il peggiore tra quelli trovati, pari a 1870.728.

Conclusione

Tra i differenti modelli provati, il modello UCM è risultato il migliore dal punto di vista del MAE, con un valore di 688.3647, rispetto ai 1215.022 del modello ARIMA e ai 1870.728 del modello di machine learning dei KNN. Tra i vari modelli, quelli computazionalmente più pesanti sono stati gli UCM, mentre i modelli di machine learning sono stati i più veloci e performanti.

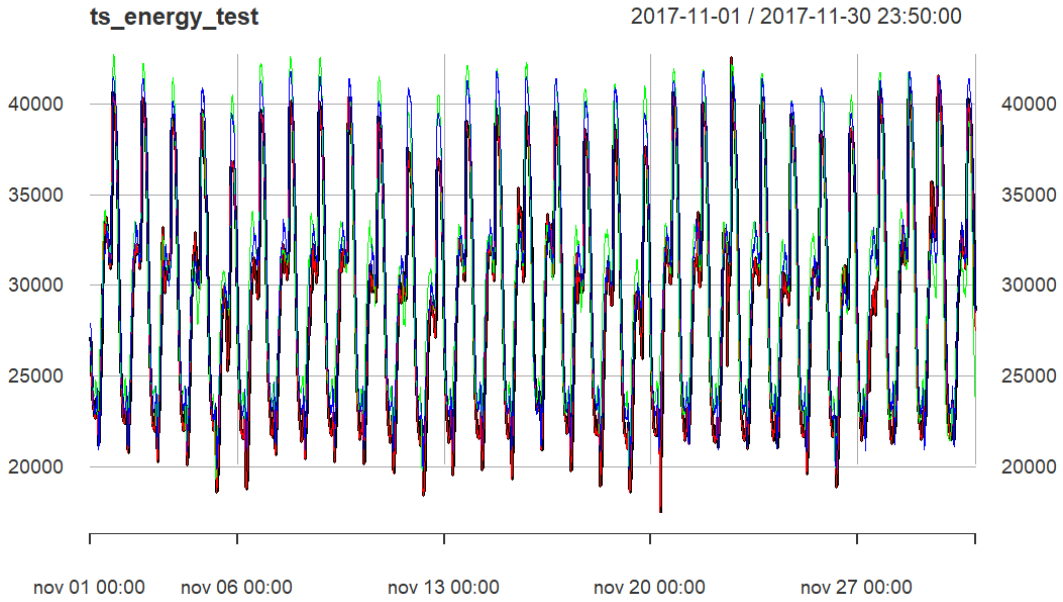

Figure 13: in blu le previsioni ottenute dal modello ARIMA, in verde le previsioni ottenute dal modello di machine learning, in rosso le previsioni ottenute dal modello UCM, a partire dalla serie temporale originale (in nero).

Utilizzando questi modelli sono state poi effettuate le previsioni per il mese di dicembre 2017. Come potenziali sviluppi futuri è possibile aggiungere delle covariate (come le festività, o informazioni del meteo quali tempo,

temperatura e precipitazioni), sviluppare ulteriori modelli (anche combinando le diverse modalità) e testando ancora più modelli con parametri differenti.

In generale, comunque, tutte le classi di modelli sembrano essere validi per l'implementazione di previsioni, e in tutti i casi ci sono margini per migliorare ulteriormente le previsioni.