Analisi e previsione del consumo di energia
Progetto di Streaming Data Management and Time Series Analysis

Puccinelli Niccolò, 881395

---

10 febbraio 2023

1 Introduzione

Questo lavoro è relativo alla modellazione di una serie storica ad alta frequenza (10 minuti), riguardante il consumo di energia, i cui dati fanno riferimento al 2017, più precisamente al periodo compreso tra il 01/01/2017 e il 30/11/2017. L'obiettivo è la previsione del consumo di energia durante il mese di dicembre, tramite 3 diverse famiglie di modelli: **ARIMA**, **UCM** e **Machine Learning**. Nello specifico, sono state provate diverse combinazioni di differenti modelli, scegliendo quello con la performance migliore per ogni tipo di approccio. L'obiettivo è la minimizzazione del *MAE* (*Mean Absolute Error*), definito nel seguente modo:

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

2 Metodologia

Le osservazioni a disposizione sono 48096, dal 01/01/2017 00:00:00 al 30/11/2017 23:50:00. Considerando quest'ultimo come training-set, il test-set è dato dai 4320 valori corrispondenti al mese di dicembre 2017, dal 01/12/2017 00:00:00 al 30/12/2017 23:50:00, a noi sconosciuti. La serie non contiene valori nulli.

Per ogni approccio è stato impiegato come validation set il mese di novembre, dal 01/11/2017 00:00:00 al 30/11/2017 23:50:00. Tuttavia, le valutazioni su tale insieme di dati sono state effettuate con un certo riserbo. Come si può notare dal grafico sottostante (Fig. 1), che indica l'andamento orario del consumo di energia dal 01/01/2017 00:00:00 al 30/11/2017, nel mese di novembre vi è una drastica flessione dell'andamento della curva. Questo è probabilmente dovuto al cambiamento dell'ora da legale a solare. Questo "salto" può portare a stime sfalsate (il cambiamento avviene a fine ottobre e il modello potrebbe non essere in grado di adattarsi in maniera soddisfacente) e ad un adattamento eccessivo dei modelli al validation set.

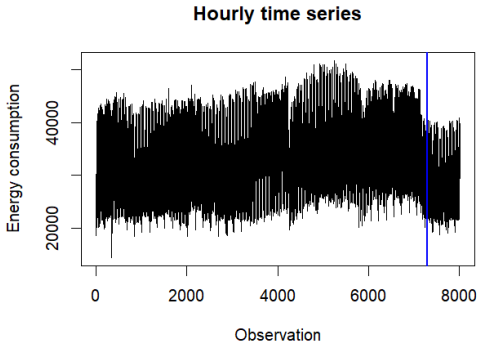

Figura 1: *Serie storica aggregata per ora: la linea blu indica l'inizio del mese di novembre.*

Pertanto, per la scelta finale del modello, è stato preso in considerazione anche questo fattore di adattamento al validation-set e si è cercato di massimizzare la capacità di generalizzazione.

Infine, i modelli sono stati ristimati su tutti i dati (i.e. training-set + validation-set).

3 ARIMA

Prima di procedere con le stime dei modelli, la serie è stata raggruppata in modo da avere 24 serie temporali giornaliere, una per ogni ora del giorno. In particolare, per ogni ora è stata presa la media delle 6 osservazioni relative (e.g. per l'ora 4 è stata effettuata la media tra 04:00:00, 04:10:00, 04:20:00, 04:30:00, 04:40:00 e 04:50:00). In questo modo viene meno la correlazione tra le ore, ma la stagionalità viene modellata in maniera più efficiente e, inoltre, viene ridotta la dimensionalità del dataset.

Per la valutazione della stazionarietà, sono stati creati i grafici di dispersione tra media e deviazione standard delle serie temporali, dai quali non si evince una relazione lineare crescente tale da far pensare alla presenza di non stazionarietà in varianza. Tuttavia, la serie è chiaramente integrata. La non stazionarietà in media risulta risolvibile tramite una differenziazione semplice e una differenziazione stagionale (una settimana).

Sono stati stimati vari modelli, prendendo inizialmente come riferimento le ore 17. Ogni valutazione è stata eseguita sequenzialmente, valutando l'andamento dei grafici di ACF e PACF.

Una volta differenziata la serie tramite una differenza semplice e una stagionale, ci si accorge subito della presenza di MA stagionale, pertanto il primo modello stimato è un **ARIMA(0,1,0)(0,1,1)[7]**. Successivamente, i grafici di ACF e PACF suggerivano l'inserimento di un fattore MA(1). Una volta ottenuto il modello **ARIMA(0,1,1)(0,1,1)[7]**, si sono valutate le performance sul validation-set. Dunque, ognuna delle 24 ore è stata modellata tramite modello Airline. Le previsioni di ogni ora, concatenate, sono state rese "compatibili" con il validation-set considerando 2 previsioni consecutive della serie ed estraendo 6 valori equidistanti all'interno del loro range (i.e. la serie è stata riportata nella metrica originale, con un valore ogni 10 minuti). Il MAE risultante sul validation-set è pari a 1466.737.

A questo punto si è cercato di modellare la serie in maniera diversa.

Modellando tutte le 24 ore, si è notato che il consumo di energia di giorno e il consumo di energia di notte sono molto diversi, pertanto potrebbe essere necessario modellare in modo diverso questi due diversi andamenti. A tal fine, sono stati stimati due modelli ARIMA diversi: uno per la fascia oraria compresa tra le 4 e le 20 (fascia giornaliera) e uno per la fascia oraria compresa tra le 21 e le 3 (fascia notturna). L'orario considerato è quello standard, ovvero UTC.

Dopo i cambiamenti iterativi dei modelli sulla base degli ACF e PACF plot, i due modelli scelti sono: **ARIMA(0,1,1)(0,1,1)[7]** per la fascia giornaliera e **ARIMA(0,1,1)** per la fascia notturna. Questo è anche molto sensato: di notte infatti difficilmente i consumi di energia presentano una stagionalità settimanale.

Analizzando i grafici dei residui si notano diversi picchi negativi. Andando ad esplorare queste osservazioni per ogni ora, ci si accorge presto che sono relativi a giorni di festa, precisamente in Marocco (questo spiega anche la teoria sul cambio orario a fine ottobre). Pertanto, sono state modellate diverse dummy deterministiche relative alle festività, analizzando le informazioni trovate su diversi siti internet.[1][2]

Per dimostrare la teoria sul cambio orario, molto influente per le stime sul validation set, è stata costruita una variabile di regressione, che vale 0 per l'ora solare e 1 per quella legale.

Dopo aver applicato questi cambiamenti alla modellistica, il MAE ottenuto è migliorato notevolmente, pari a 1287.33 (Fig. 2). Tutta-

[1] https://www.timeanddate.com/holidays/morocco/2017

[2] https://www.officeholidays.com/countries/morocco/20177

via, questa stima è influenzata dalla dummy aggiunta per il cambiamento orario. Rimuovendola, infatti, il MAE sale a 1367.583. Per la stima finale tale fattore non verrà dunque considerato.

Una volta ristimati i due modelli ARIMA su tutti i dati, analizzando gli ACF e PACF plot di diversi orari giornalieri, si è valutato l'inserimento di un fattore AR con p = 2. Dopo l'inserimento di tale fattore, i due grafici sembrano più puliti, pertanto il modello finale è composto da:

- **ARIMA(2,1,1)(0,1,1)[7]** per la fascia giornaliera (4-20).
- **ARIMA(0,1,1)** per la fascia notturna (21-3).
- Regressore dummy riguardante i giorni festivi in Marocco.

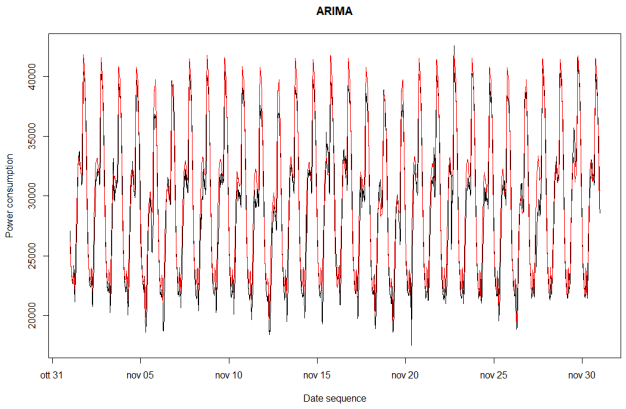

Figura 2: *Previsioni ARIMA, MAE = 1287.33*

4 UCM

Per questa tipologia di approccio si è deciso di raggruppare i dati per ore, al fine di diminuire l'altrimenti elevato tempo di computazione necessario. Pertanto, anche in questo caso è stata presa la media oraria e, in fase di predizione, i 6 punti di unione tra due valori consecutivi, al fine di ritornare alla metrica originale di 10 minuti.

Anche in questo caso sono stati testati diversi modelli:

- Il primo modello si compone di un Local Linear Trend con 24 dummy stocastiche. Il trend ottenuto cresce eccessivamente rispetto all'andamento della serie e il MAE è di conseguenza altissimo.
- Come secondo modello, si è provato ad aggiungere un ciclo ogni 168 ore (i.e. una settimana). La stima è migliorata notevolmente e il MAE si attesta a 1873.31. Ciò nonostante, il grafico delle previsioni suggerisce che il ciclo stocastico ha identificato un ciclo diverso da 168 ore (confermato anche dai parametri di ritorno).
- Il terzo modello è stato costruito come il precedente, ma introducendo una stagionalità intra-annua con 24 sinusoidi stocastiche al posto delle 24 dummy stocastiche. Le stime sono completamente sfalsate e il MAE è di nuovo altissimo.
- Si è inoltre provato ad aggiungere le dummy deterministiche concernenti le festività come nei modelli ARIMA, ma le performance non sono migliorate.
- Il quarto e ultimo modello testato contiene un Local Linear Trend, 24 dummy stocastiche e una componente stagionale trigonometrica con periodo di 168 ore (i.e. una settimana). E' stato inoltre ricercato il numero di sinusoidi ottime tra 1 e 16 (in grado di minimizzare il MAE sul validation-set). Alla fine ne sono state selezionate 6.

Il modello finale è dunque composto da un Local Linear Trend, una componente stagionale a dummy stocastiche e una componente stagionale a sinusoidi stocastiche. Il MAE ottenuto sul validation-set è pari a 1299.51 (Fig. 3). Tale modello è stato dunque ristimato su tutti i dati (training + validation), da cui sono state estratte le previsioni per il test-set.

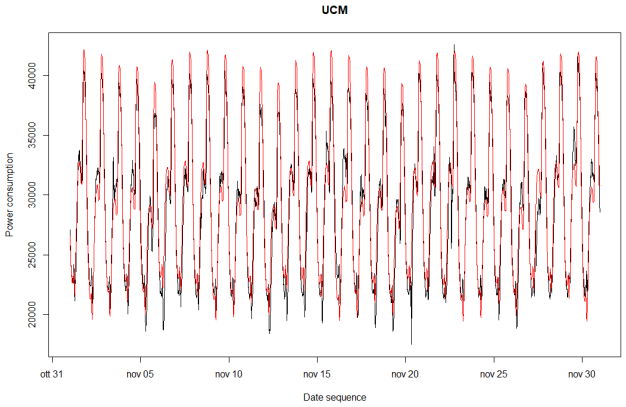

Figura 3: *Previsioni UCM, MAE = 1299.51*

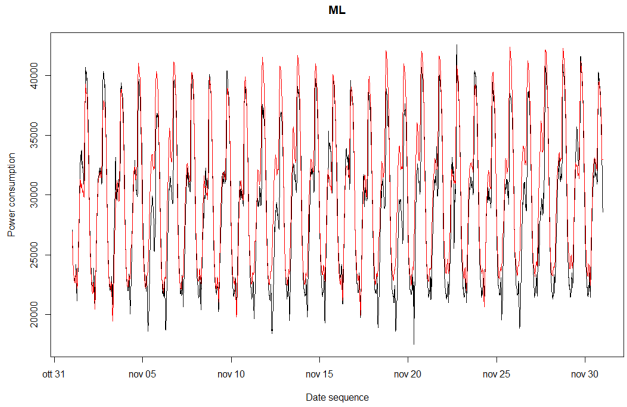

Figura 4: *Previsioni ML, MAE = 1716.4*

5 Machine Learning

Per i modelli di Machine Learning si è scelto di raggruppare la serie in maniera oraria (come per UCM) e tramite 24 serie giornaliere (come per ARIMA). Per entrambe le tipologie di raggruppamento sono stati testati 3 algoritmi differenti: **Random Forest**, **XGBoost** e **Support Vector Machines**.

Le previsioni sono state effettuate con metodo ricorsivo per ogni modello, così da essere comparabili.

Per le varie combinazioni sono stati testati diversi lag e ogni modello sembra prediligere un determinato tipo di raggruppamento + lag. Tuttavia, il raggrupamento mediamente più efficace è risultato essere il primo, per il quale sono stati esaminati diversi lag, in particolare 2, 8, 16, 24 (1 giorno), 48 (2 giorni) e 168 (una settimana).

L'aggiunta di regressori esterni, quali sinusoidi e dummy deterministiche relative ai giorni festivi, non ha portato alcun miglioramento.

Il metodo Random Forest è risultato quello con le performance peggiori, mentre SVM e XGBoost raggiungono performance comparabili, seppur con lag diversi come regressori: 24 per SVM e 168 per XGBoost.

In ultima analisi, il modello che ha restituito le performance migliori è un XGBoost con 168 lag (i.e. lag settimanale) e un numero di round pari a 1000, con un MAE sul validation-set di 1716.4 (Fig. 4).

6 Conclusioni

In conclusione, il modello migliore sul validation-set risulta essere il modello **ARIMA(2,1,0)(0,1,1)[7]**. Tuttavia, per tutti i 3 modelli selezionati, ci si potrebbe aspettare un significativo miglioramento delle performance, per le motivazioni precedenti riguardanti il validation-set e per l'alta capacità di generalizzazione che si è cercato di mantenere durante il processo di selezione dei modelli.

Ranking test-set Osservando il ranking sul test-set, si può confermare la teoria sul validation-set e la capacità di generalizzazione. Tutti i 3 modelli registrano infatti una significativa riduzione del MAE, dovuta sia al ri-addestramento con i dati di training insieme ai dati di validation (mese di novembre), sia all'elevata generalizzazione che si è cercato di mantenere durante lo svolgimento di questo progetto.

Viene riportato di seguito un breve sommario delle performance (Tabella 1).

Model	Val	Test	Δ_{V-T}
ARIMA	1287.33	1020.2	267.13
UCM	1299.51	1121.72	177.79
ML	1716.4	1526.51	189.89

Tabella 1: *Sommario delle performance (**MAE**) su validation-set e test-set.*