

# Previsione Consumi Elettrici

Tramite utilizzo di modelli ARIMA, UCM e  
LSTM

Gianmarco Russo mat.887277

Msc Data Science

Streaming Data Management and Time Series Analysis



January 14, 2023

## Contents

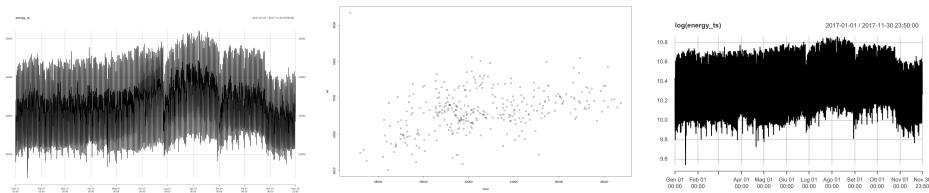
<b>1</b>	<b>Introduzione</b>	<b>1</b>
<b>2</b>	<b>Analisi Esplorative</b>	<b>1</b>
<b>3</b>	<b>ARIMA</b>	<b>2</b>
3.1	Sarima . . . . .	2
3.2	Primo modello ARIMA . . . . .	3
3.3	Auto arima . . . . .	4
3.4	Terzo modello ARIMA . . . . .	4
<b>4</b>	<b>UCM</b>	<b>5</b>
4.1	Modello proposto . . . . .	5
<b>5</b>	<b>Deep Learning</b>	<b>6</b>
5.1	LSTM . . . . .	7
<b>6</b>	<b>Conclusioni</b>	<b>7</b>

## 1 Introduzione

Le previsioni dei consumi energetici sono cruciali per diversi scopi, dall'acquisto dell'energia dal produttore, alla gestione di sovraccarichi. In questo elaborato viene analizzata una serie storica univariata di consumo energetico con frequenza di campionamento ogni 10 minuti. Il periodo fornito va dal 01/01/2017 al 30/11/2017 con l'obiettivo di stimare i consumi del mese di dicembre. Il dataset nel complesso è composto da 48096 osservazioni. Non sono presenti ulteriori informazioni come meteo, festività ed altre caratteristiche tipiche del luogo di interesse.

## 2 Analisi Esplorative

La serie storica completa si presenta così:



**Figure 1:** (a) Serie storica (b) Mean/Stdev plot (c) log transformed series

Tra le analisi esplorative valutiamo inizialmente la necessità o meno di trasformazioni nei dati, iniziando con un plot media su deviazione standard per capire se c'è correlazione e quindi bisogno di una log transformation: come si può notare vi è una discreta tendenza lineare(specie se non consideriamo l'outlier in alto a sinistra). Si lavorerà perciò con i

logaritmi della serie storica. A causa di limiti computazionali a gestire l'intera serie storica con R studio in locale(specialmente nella stime dei modelli ARIMA e UCM, in quanto la parte di deep learning è stata svolta su python in colab) in questo elaborato verranno discussi i risultati ottenuti utilizzando solamente i mesi di Settembre, Ottobre per voi validare sul mese di Novembre e ottenere un MAE(mean absolute error) di riferimento, in quanto si, il nostro obiettivo è stimare dicembre, ma non avendo i veri valori del suddetto mese non sarebbe poi possibile valutare le performance dei modelli.

### 3 ARIMA

Arima è un punto di riferimento della modellazione delle serie storiche. E' composto da 3 componenti:

- **AR:** Modellazione autoregressiva:  

$$X_t = c + \sum_1^p \phi_i X_{t-i} + \epsilon_t$$
- **I:** Integrazione della serie storica, utilizzata per serie non stazionarie;
- **MA:** Modellazione a media mobile(moving average):  

$$X_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q}$$

Il modello ARIMA deriva dal modello ARMA a cui sono state applicate le differenze di ordine  $d$  per renderlo stazionario. I parametri del modello sono 3:

- **p:** numero di lag della componente autoregressiva;
- **d:** numero di integrazioni;
- **q:** numero di lag della componente moving average.

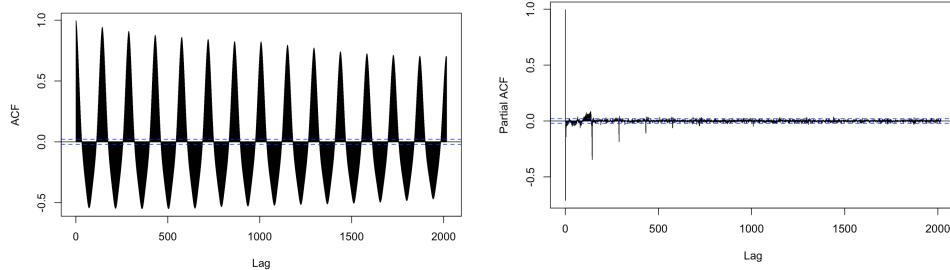
#### 3.1 Sarima

Seasonal ARIMA, necessario in quanto i modelli arima non considerano le stagionalità all'interno delle serie storiche. Una stagionalità è definita come la ricorrenza ciclica di pattern all'interno della serie storica. I modelli arima si aspettano dati senza stagionalità o con stagionalità rimossa con metodi come la differenziazione stagionale. Sarima è perciò una estensione di arima e oltre ai parametri già descritti si aggiungono anche parametri puramente stagionali:

- **P:** numero di stagioni della componente autoregressiva;
- **D:** numero di integrazioni nelle stagioni;
- **Q:** numero di stagioni della componente moving average;
- **m:** numero di punti componenti la stagione.

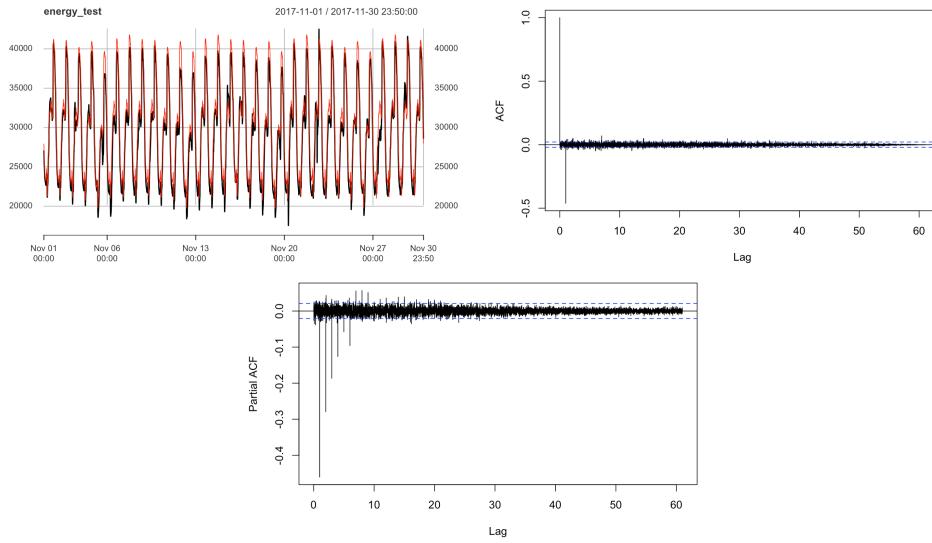
### 3.2 Primo modello ARIMA

Per capire con quali parametri iniziare a stimare il nostro modello, guardiamo i grafici di autocorrelazione ed autocorrelazione parziale:



**Figure 2:** (a) ACF (b) PACF

Notiamo come ci sia una stagionalità giornaliera, con picchi ogni 144 osservazioni, che sicuramente includeremo nel parte seasonal del modello arima. Per quanto riguarda i ritardi autoregressivi è stato provato con 3 ritardi. Il primo modello è stato quindi composto così:  $arima(3, 0, 0)(0, 1, 0)_{144}$ . Come regressori sono state create 10 sinusoidi per aiutare a modellare la serie storica. Di seguito le stime:

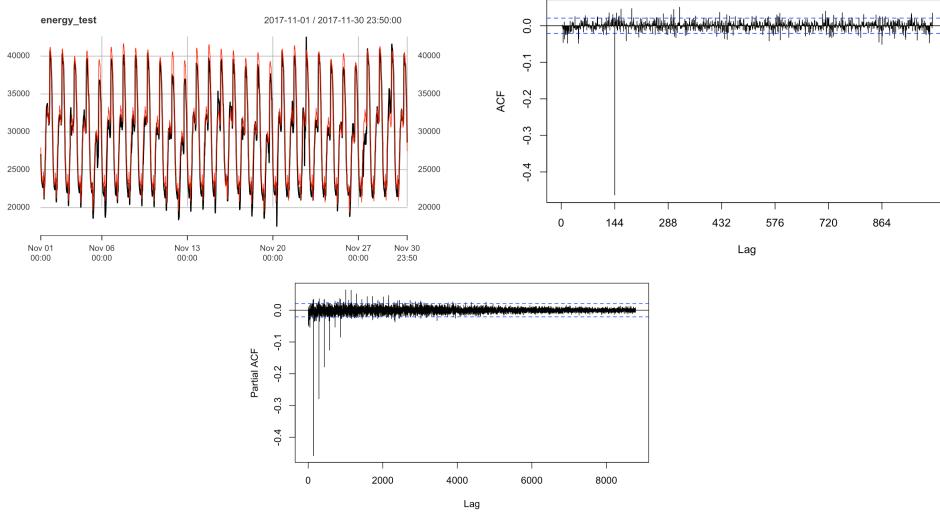


**Figure 3:** (a) previsioni modello (in rosso) (b) ACF residui (c) PACF residui

Il mean absolute error (**MAE**) su novembre è pari a 1206. Per avere contezza della bontà della stima ricordiamo che il valore medio della serie storica è di 32643, quindi un errore intorno al 3%. Analizzando i residui del modello possiamo notare come non vi sia un vero e proprio white noise, il che significa che qualcosa nel modello è ancora migliorabile.

### 3.3 Auto arima

La funzione auto arima prova iterativamente diversi parametri nei modelli arima ed identifica quello che performa meglio. In questo caso la funzione ha restituito un  $arima(2, 1, 0)(0, 1, 0)_{144}$  come modello migliore. Analizziamo ora le stime e i residui di questo modello: Notiamo come

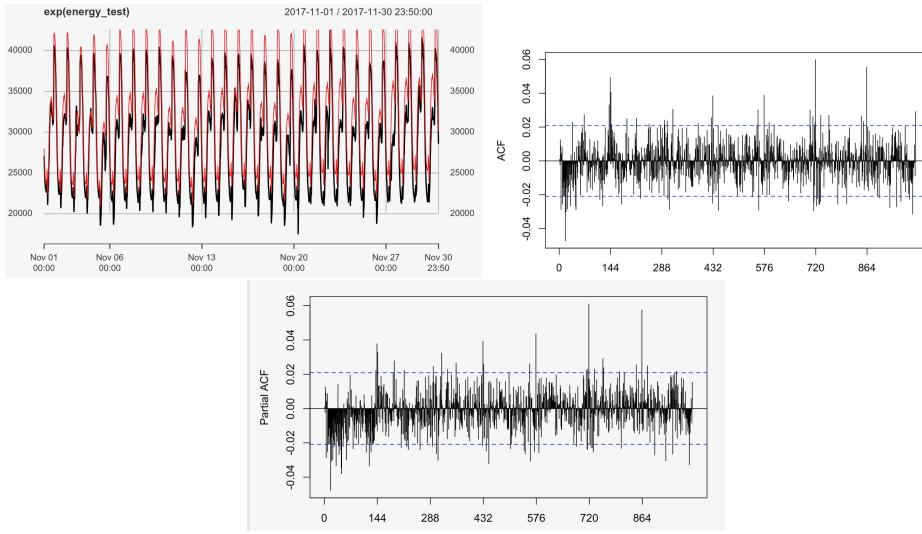


**Figure 4:** (a) previsioni modello (in rosso) (b) ACF residui (c) PACF residui

nei residui vi sia ancora un pattern. in particolare sembrerebbe uno **SMA(1)** con stagionalità 144. La funzione auto arima ottimizza in base alle performance e probabilmente non ha scelto di includere uno SMA per questo motivo. Il modello performa meglio del precedente, con un **MAE** su novembre di 1030.

### 3.4 Terzo modello ARIMA

Visti i residui del precedente modello si è voluto provare ad aggiungere uno SMA(1) per ottenere dei residui white noise, il modello quindi stimato è un  $ARIMA(2, 1, 0)(0, 1, 1)_{144}$ :



**Figure 5:** (a) previsioni modello (in rosso) (b) ACF residui (c) PACF residui

Come possiamo notare abbiamo i residui totalmente casuali, come anticipato nella sezione precedente. Purtroppo però il **MAE** risulta essere 1300, più alto dei modelli precedenti. Questo sembrerebbe essere il modello più **giusto**, ma purtroppo non il più performante.

## 4 UCM

Nei modelli a componenti non osservabili una serie storica è considerata come la somma di alcune componenti non direttamente osservabili. Nello specifico può essere composta da: trend, ciclo, stagionalità e rumore bianco:  $Y_t = \mu_t + \psi_t + \gamma_t + \epsilon_t$ .

Le singole componenti vengono derivate da funzioni deterministiche del tempo come la rette e le sinusoidi, rese stocastiche con l'aggiunta di shock casuali.

### 4.1 Modello proposto

Per il modello trattato in questo elaborato è stato creato un modello UCM con le seguenti componenti:

- **Local Linear Trend** (SSMtrend di ordine 2),
- **Stagionalità trigonometrica giornaliera** (SSMseasonal con 10 armoniche),
- **Stagionalità trigonometrica settimanale** (SSMseasonal con 10 armoniche).

Ricordiamo brevemente il compito di queste componenti:

- Il *trend* è responsabile per la variazione della media del processo nel lungo periodo, la sua equazione può essere scritta:  $\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t$  e  $\beta_t = \beta_{t-1} + \zeta_t$ . In questo modo in base al valore della varianza degli errori  $\sigma_\eta^2$  e  $\sigma_\zeta^2$  e al valore di  $\beta$ , possiamo modellare diversi comportamenti della media. In questo caso è stato scelto appunto un local linear trend, quindi non vi sono restrizioni sui valori di questi ultimi.

- La *stagionalità trigonometrica* si ottiene dalla rappresentazione di una funzione periodica a somma nulla, come somma di sinusoidi a frequenze di fourier. Teoricamente il numero di sinusoidi da utilizzare dovrebbe essere pari a  $s/2$  con  $s$  stagionalità. In questo caso risultava impossibile per ovvi motivi computazionali introdurre 72 (stagionalità giornaliera/2) o 516 (stagionalità settimanale/2) sinusoidi, perciò si è sperimentato con valori compresi tra 10 e 20 sinusoidi.

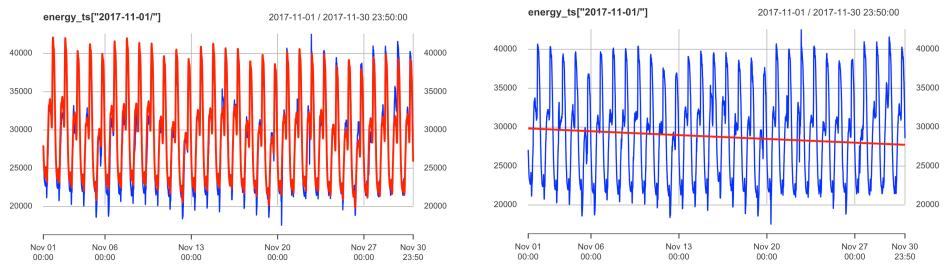
Altri parametri fondamentali sono l'inizializzazione delle varianze del modello. Sia  $vy$  la varianza della serie storica:

- $\log VarEta$ :  $vy/10$  varianza della retta,
- $\log VarZeta$ :  $vy/10000$  varianza dello slope,
- $\log VarOmega$ :  $vy/1000$  varianza stagionalità giornaliera,
- $\log VarOmega7$ :  $vy/10000$  varianza stagionalità settimanale,
- $\log VarEpsilon$ :  $vy/10$  varianza white noise.

Ed infine:

- $a1$ : valore atteso del vettore di stato al tempo 1, inizializzato come la media della serie storica,
- $P1$ : matrice di covarianza del vettore di stato al tempo 1, inizializzato come  $vy^*10$

Sono state effettuate diverse prove con valori iniziali delle varianze e sono stati riportati solo i valori del modello migliore. Di seguito i risultati:



**Figure 6:** (a) previsioni modello (in rosso) (b) Trend (in rosso)

Il seguente modello ha ottenuto un **MAE** di 1180, quindi in linea con i modelli ARIMA e precedentemente discussi.

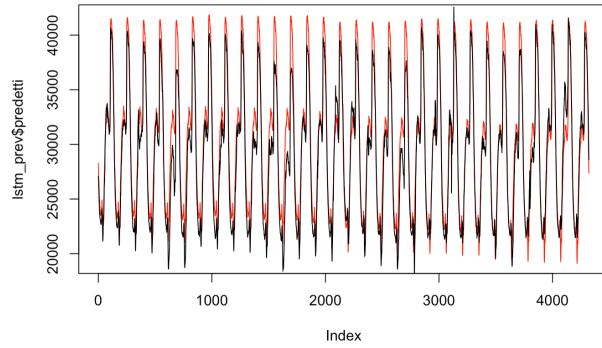
## 5 Deep Learning

La maggior parte dei modelli di machine learning performa previsioni one step ahead, quindi con orizzonte previsivo pari ad 1 e quindi non confrontabili con i modelli ARIMA e UCM che effettuano previsioni k-step ahead (con k pari ad un mese di previsioni in questo caso). Per questo motivo in questo elaborato ci si è concentrati sulla definizione di un modello LSTM(long short term memory), una recurrent neural network molto usata in ambito di natural language processing ed elaborazione di serie storiche.

## 5.1 LSTM

Rispetto ai modelli precedenti in questo caso è stata applicata manualmente una differenziazione a 144 osservazioni (giornaliera) per imporre una stagionalità alla rete. Sono stati provati diversi parametri della rete, variando sia la quantità di layer lstm nella rete, sia la grandezza di questi ultimi. Per questo modello i dati utilizzati sono stati da giugno in avanti, cercando di ottenere un tradeoff tra limiti computazionali e sufficienti dati per addestrare algoritmi di deep learning. Di seguito vengono riportati i parametri della rete che ha dato risultati migliori:

- a) **lags**: 4320,
- b) **layers**: 2 layer da 16 unità,
- c) **epochs**: 30,



**Figure 7:** Previsioni modello (in rosso)

Il modello ottiene un **MAE** di 1376, quindi di poco superiore ai modelli precedentemente discussi, il che è un ottimo risultato considerando che non sono stati inclusi regressori (le sinuosoidi).

## 6 Conclusioni

In questo elaborato sono state discusse le scelte e le performance di diverse famiglie di modelli per serie storiche. In particolare il modello più performante è stato l'ARIMA. La famiglia dei modelli lineari ha formato meglio delle LSTM, godendo inoltre di maggiore trasparenza ed explainability rispetto ai modelli deep learning che per natura sono più "black box". Nel suo piccolo si dimostra anche come i modelli lineari risultino ad oggi più che validi per la previsione di serie storiche. L'aggiunta di regressori migliorerebbe sicuramente le performance, come ad esempio dati meteo-geologici o festività specifiche della zona di riferimento.