Previsione dei consumi elettrici nella città di Tétouan, Marocco

Simone Farallo[1]

Abstract
La previsione di serie temporali riveste un ruolo essenziale in un mondo sempre più orientato ai dati, consentendo di anticipare eventi futuri basandosi su dati storici, prendere decisioni più accurate ed individuare tendenze e pattern nei dati. Nel contesto specifico del consumo di energia elettrica, la previsione svolge un ruolo cruciale fornendo informazioni ai fornitori di energia per migliorare le prestazioni dei propri sistemi in termini di produttività ed efficienza. L'obiettivo di questo progetto consiste nell'analizzare l'andamento del consumo di energia e confrontare tre differenti approcci di previsione: ARIMA, UCM e Machine Learning. I dati a disposizione sono una serie storica univariata relativa al consumo di energia elettrica nella città di Tètouan in Marocco nel 2017; i risultati dei tre metodi di previsione sono confrontati utilizzando il Mean Absolute Error come metrica di riferimento. Attraverso l'analisi dei risultati, sarà possibile determinare quale tra i tre approcci si adatta meglio alla previsione del consumo di energia elettrica per il periodo compreso tra l'1 dicembre 2017 e il 31 dicembre 2017.

Keywords
Forecasting - Time Series - ARIMA - UCM - Machine Learning - Electricity Consumption

[1,2,3] *Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli studi di Milano-Bicocca, Milano, Italia*

Contents

1. Introduzione

La modellazione e previsione di serie storiche per il consumo di energia elettrica rappresenta un'area di ricerca importante nell'ambito dell'analisi dei dati, in questo studio, l'obiettivo è stato modellare e prevedere la serie temporale relativa al consumo di energia elettrica nel Marocco nel corso del 2017, confrontando le prestazioni di tre metodi di previsione: ARIMA, UCM e Machine Learning.
In particolare, si vuole identificare il modello che offre la massima precisione nella previsione dei valori di consumo di energia il periodo compreso tra il 1° dicembre 2017 e il 31 dicembre 2017. Per confrontare e scegliere i modelli migliori, è stata presa come riferimento la Mean Absolute Error(MAE), l'obiettivo è minimizzare tale misura:

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

2. Esplorazione dei dati

Il dataset fornito per questo progetto consiste in una serie storica univariata regolare relativa al consumo di energia elettrica, osservata ogni 10 minuti. I dati sono suddivisi in due colonne: "date" e "power",la prima colonna rappresenta la data e l'ora della misurazione , mentre la seconda colonna contiene i valori di consumo rilevati.
Prima di iniziare le analisi ci si è accertati che la serie non contenesse valori nulli o mancanti; la serie temporale copre un periodo specifico, che va dal 1° gennaio 2017 alle 00:00:00 fino al 30 novembre 2017 alle 23:50:00, in totale sono presenti 48.096 osservazioni. Succesivamente sono stati realizzati dei grafici per comprendere l'andamento della serie storica.

Dal grafico [1] che indica l'andamento orario del consumo di energia dal 01/01/2017 al 30/11/2017, si nota che nel mese di novembre vi è una drastica flessione dell'andamento della curva, dovuta probabilmente al cambiamento dall'ora da legale a solare; questo può portare a stime sfalsate infatti 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 nel validation set.

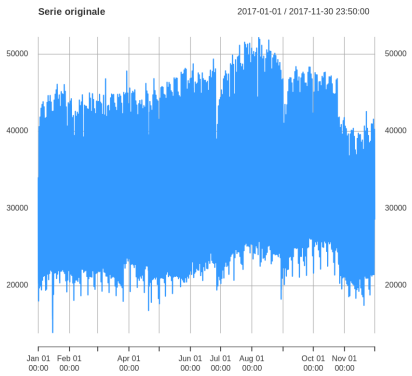

Figure 1. Serie storica originale.

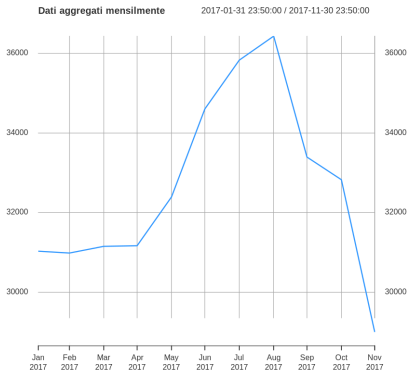

Figure 2. Andamento mensile

Osservando il grafico [2] si nota che nei mesi primaverili i consumi iniziano ad aumentare rispetto ai mesi precedenzi e i consumi massimi vengono raggiunti nei mesi estivi, con il picco nel mese di agosto.

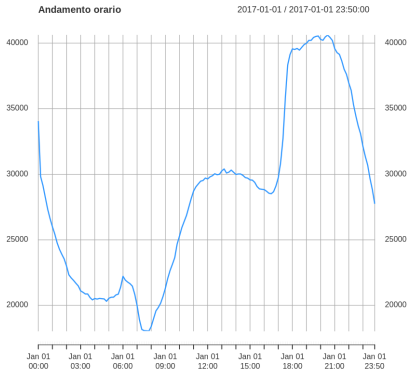

Figure 3. Andamento orario del primo giorno di Gennaio.

Osservando il grafico [3] si nota come i consumi sono minimi nelle ore notturne dalle 21 alle 6, crescono nelle ore diurne dalle 8 alle 18 e raggiungono il massimo attorno alle ore 20 e 21.

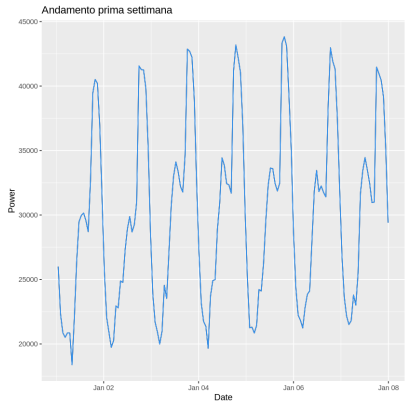

Figure 4. Andamento prima settimana di Gennaio.

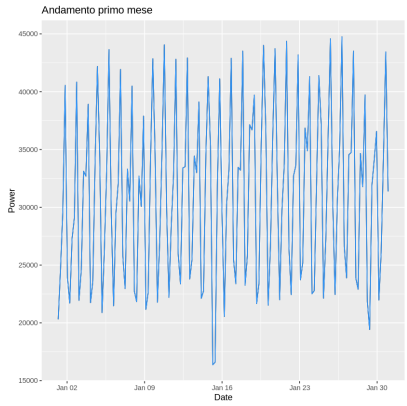

Figure 5. Andamento del mese di Gennaio.

Osservando i grafici [4] e [5], si possono già individuare una componente stagionale e una componente mensile. Successivamente sono stati controllati ed analizzati alcuni outliers, rappresentati nella figura [6].

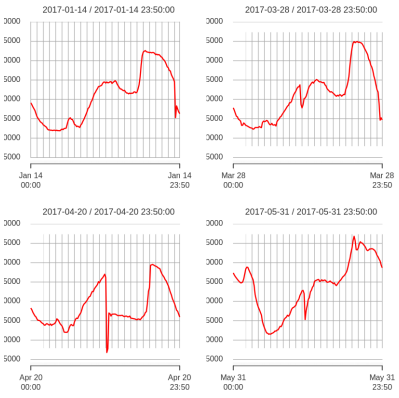

Figure 6. Outliers.

Per gestire gli outliers ci sono diverse tecniche che possono essere utilizzate, ma in questo caso gli outliers sono relativamente pochi e per evitare di creare distorsione nelle stime, non è stata intrapresa nessuna azione.
La serie poi è stata divisa in due parti:

- Train set: dal 1 Gennaio 2017 al 31 Ottobre 2017.
- Validation set: dal 1 Novembre 2017 al 30 Novembre 2017.

3. ARIMA

I modelli ARIMA appartengono alla famiglia dei processi stocastici lineari non stazionari e sono un'estensione dei modelli ARMA, quest'ultimi richiedono la stazionarietà del processo.

3.1 Stazionarietà

Per la valutazione della stazionarietà in varianza, 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; per risolvere il problema della stazionarietà in varianza, è stata applicata la trasformazione di Box-Cox.

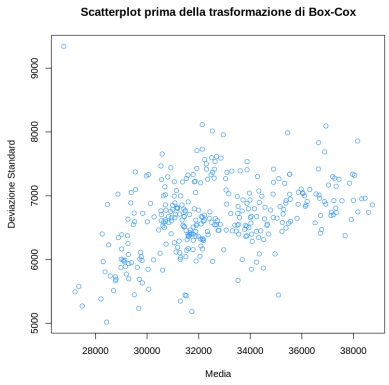

Figure 7. Scatterplots prima della trasformazione di Box-Cox.

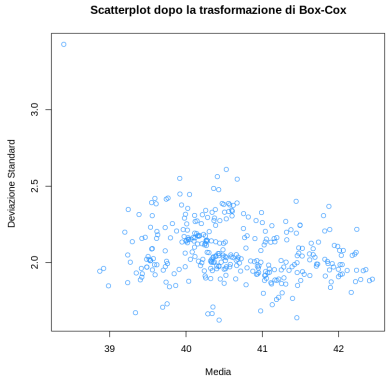

Figure 8. Scatterplots dopo la trasformazione di Box-Cox.

Per la stazionarietà in media sono stati utilizzati i test di Dickey-Fuller e KPSS i quali hanno confermato la non stazionarietà della serie, risolvibile tramite una differenziazione semplice e una differenziazione stagionale.

3.2 Modellazione

I modelli che seguono partono dunque dalla modellizzazione di questa non stazionarietà stagionale e proseguono con cambi iterativi dei parametri (p, d, q) (P, D, Q) basati sui grafici Acf[1] [9] e Pacf[2] [10].

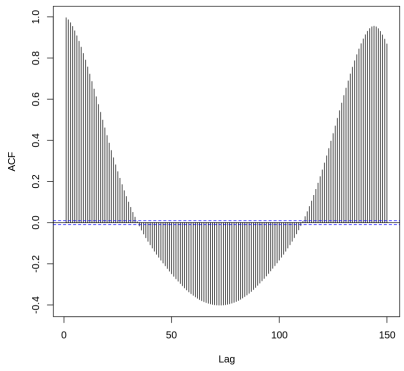

Figure 9. ACF

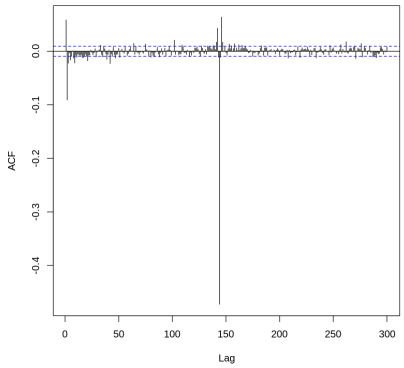

Figure 10. PACF

Il primo approccio è stato modellare solo la stagionalità giornaliera con la differenza stagionale, poiché è quella dominante; osservando i grafici ACF e PACF e provando diverse combinazioni, la migliore è risultata

$$ARIMA(0,0,0)(1,1,0)[144]$$

Il valore del MAE ottenuto validation set è stato 1152.399.

Il secondo approccio è stato modellare i giorni con differenza stagionale e la settimana con variabili dummy, nel dettaglio sono state create 6 variabili dummy per modellare i 7 giorni della settimana da utilizzare come regressori, anche in questo caso sono state provate diverse combinazioni, la migliore è risultata essere

$$ARIMA(0,0,0)(1,0,0)[144]$$

Il valore del MAE ottenuto sul validation set è stato **1050.714**.

[1] Funzione di autocorrelazione
[2] Funzione di autocorrelazione parziale

Nel terzo approccio è stato effettuato un raggruppamento della serie storica al fine di creare 24 serie temporali giornaliere, corrispondenti alle 24 ore del giorno. Questo raggruppamento è stato realizzato calcolando la media dei valori di consumo delle 6 osservazioni relative a ciascuna ora, questo consente una modellazione più efficiente della stagionalità presente nella serie storica, il risultato migliore si ottiene con

$$ARIMA(0,1,1)(0,1,1)[144]$$

Il valore del MAE ottenuto sul validation set è 1245.199.

4. UCM

Gli Unobservable Component Models (UCM) sono una classe di modelli statistici utilizzati per l'analisi delle serie temporali, questi modelli si basano sulla decomposizione della serie temporale in componenti osservabili e non osservabili. La componente osservabile corrisponde ai dati stessi, mentre le componenti non osservabili includono fattori come tendenza, ciclo, stagionalità e così via.
Gli UCM sono altamente flessibili e possono essere adattati a diverse situazioni e problemi specifici, rendendoli uno strumento potente per molte applicazioni, inoltre rispetto ai modelli ARIMA non richiedono nessuna assunzione di stazionarietà.
Per applicare questi modelli, è necessario innanzitutto assegnare il valore NA ai dati da prevedere per consentire al filtro di Kalman di prevedere i valori sulla base delle componenti non osservate, successivamente è necessario esaminare i risultati dell'analisi esplorativa delle serie temporali per valutare le componenti da includere.

Il primo approccio è stato modellare la serie storica con un trend lineare[3] e due componenti stagionali (144 e 1008 osservazioni) ottenute con due sinusoidi composte rispettivamente da 2 e 1 armoniche; il valore del MAE ottenuto sul validation set è di 2145.8.

Anche nel secondo approccio la serie storica è stata modellata con un trend lineare e due componenti stagionali (144 e 1008 osservazioni), ma in questo caso, ottenute con due sinusoidi composte rispettivamente da 10 e 1 armoniche; Il valore del MAE sul validation set si è ridotto notevolmente arrivando a 1366.786.

Prendendo in considerazione il secondo modello, la componente stagionale di periodo 1008 è stata sostituita da una componente ciclica con lo stesso periodo, questo significa che la stagionalità viene modellata come un ciclo senza una regolarità specifica; il MAE ottenuto sul validation set equivale a **1341.666**.

L'ultimo approccio è stato modellare la serie utilizzando un trend lineare, una componente stagionale ogni 144 osservazioni ed utilizzando variabili dummy, questo può fornire maggiore flessibilità nel modellare le componenti stagionali, consentendo di incorporare effetti stagionali più complessi o di gestire casi specifici, ma in questo caso il modello risulta essere meno performante rispetto ai precedenti, il valore di MAE sul validation set equivale ad 1920.99.

[3] Local Linear Trend

5. Machine Learning

I modelli di Machine Learning rappresentano la terza e ultima classe di modelli presa in considerazione, l'idea è quella di essere in grado di apprendere direttamente dai dati la tendenza della serie temporale, senza dover necessariamente manipolarla o definire componenti specifiche. Questi modelli sono utili perché non richiedono una modellizzazione preliminare della stagionalità presente nei dati, in quanto questa può essere acquisita dal modello durante la fase di addestramento. Di conseguenza, non è necessario definire in anticipo la presenza di stagionalità nei dati poiché il modello è in grado di rilevarla autonomamente durante il processo di addestramento.

Come primo modello è stato utilizzato la SVM, quest'ultimo è stato testato sia sulla serie originale che con dati orari, i risultati migliori sono stati ottenuti sulla serie originale, in particolare inserendo come lag 1008 corrispondente alla periodicità settimanale; il MAE ottenuto sul validation set corrisponde a **1748.029**.

Come secondo modello è stato utilizzato il Random Forest, sono stati utilizzati dati aggregati su base oraria, in modo da avere un numero sufficiente di lag, inoltre è stato aggiunto un regressore per indicare il giorno della settimana, sono stati utilizzati 350 alberi e un lag pari a 168, corrispondente alla periodicità settimanale; il MAE ottenuto sul validation set è corrisponde a 1840.538.

Il terzo ed ultimo modello utilizzato è stato l'XGBoost, sono stati utilizzati dati aggregati su base orarie con un lag pari a 168 corrispondente alla periodicità e un numero di round pari a 1000; il MAE ottenuto sul validation set di 1775.7.

6. Conclusioni

La modellizzazione e la previsione di tali serie storiche possono aiutare a identificare i fattori che influiscono sui consumi energetici, essendo però la serie ad alta frequenza è molto difficile effettuare delle previsioni accurate.
Dopo aver testato diversi approcci con diverse combinazioni, per ogni classe di modelli è stato scelto il più performante; i 3 modelli scelti sono stati riaddestrati sull'intera serie temporale disponibile (da gennaio a novembre) e le previsioni sono state fatte per il mese di dicembre[11].
Nella tabella [1] possiamo osservare i valori di MAE ottenuti sul validaion set con i modelli più performanti.

Modello	RMSE	MAPE	**MAE**
ARIMA	1342.476	3.742	1050.814
UCM	1703.600	4.721	1341.666
ML	2205.156	6.310	1748.029

Table 1. Risultati dei modelli migliori sul Validation set.

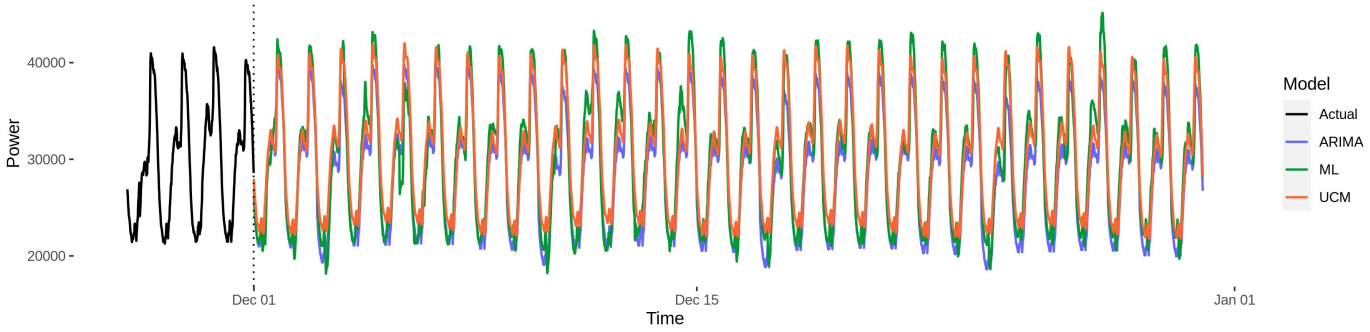

Figure 11. Previsioni di dicembre.

Nel confronto dei risultati ottenuti dai tre modelli, basati sui dati di validazione compresi tra l'1 novembre e il 30 novembre, si è osservato che i modelli statistici più tradizionali hanno fornito risultati più accurati a scapito dei lunghi tempi di calcolo dovuti all'elevato numero di osservazioni della serie temporale mentre i modelli di Machine Learning, invece, si sono dimostrati molto funzionali e veloci a scapito di previsioni meno accurate. Si può presumere che gli errori sulle previsioni saranno leggermente ridotti, tenendo conto della disponibilità di un mese aggiuntivo per l'addestramento, questo aspetto risulta particolarmente rilevante per i modelli di machine learning, i quali richiedono un ampio volume di dati per ottenere previsioni accurate.
Un approccio che si è rivelato potenzialmente molto utile per migliorare tutti i modelli è l'aggregazione dei dati per ora anche se non forniscono direttamente misure molto accurate. Come sviluppo futuro si possono combinare le previsioni a 10 minuti per ottenere stime migliori, il problema consiste nel trovare il peso o i pesi migliori da utilizzare nella combinazione lineare.
I risultati ottenuti in questo progetto dimostrano l'importanza di sperimentare diversi metodi per analizzare e prevedere serie storiche complesse, in quanto ognuno può fornire informazioni uniche e valide che possono essere utilizzate per prendere decisioni informate.

References

[1] Kfas: Exponential state space models in r. https://cran.r-project.org/web/packages/KFAS/vignettes/KFAS.pdf.

[2] Time series modelling with unobserved components. https://www.taylorfrancis.com/books/mono/10.1201/b18766/time-series-modelling-unobserved/-components-matteo-pelagatti.

[3] Forecasting: Principles and practice. https://otexts.com/fpp3/.