Streaming Data Management & Time Series Analysis Project **Alberto Gadda, Mat.824029**

1. INTRODUZIONE:

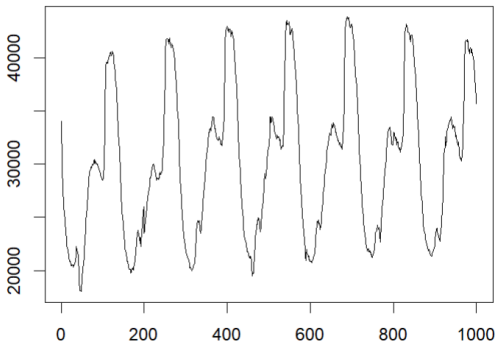
Il presente lavoro parte da una serie storica contenente dati relativi al consumo di energia elettrica, rilevati ogni 10 minuti per circa 10 mesi, e si pone l'obiettivo di fare le migliori previsioni sul mese successivo, sconosciuto al momento del lavoro.
Per lo svolgimento di questo progetto sono stati utilizzati tre diversi tipi di approccio al problema:

- Modelli ARIMA
- Modelli UCM
- Modelli ML-based

Le performance di questi tre modelli sono state confrontate su un validation-set al fine di eleggere il miglior previsore sviluppato, poiché le performance sull'effettivo test-set non sono ottenibili, essendo quest'ultimo ignoto. All'interno della stessa categoria di modelli si sono provate diverse soluzioni.

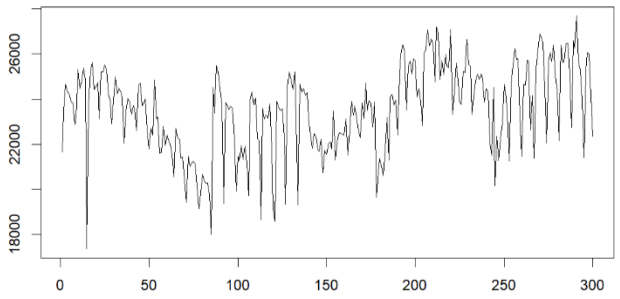
2. APPROCCIO UTILIZZATO:

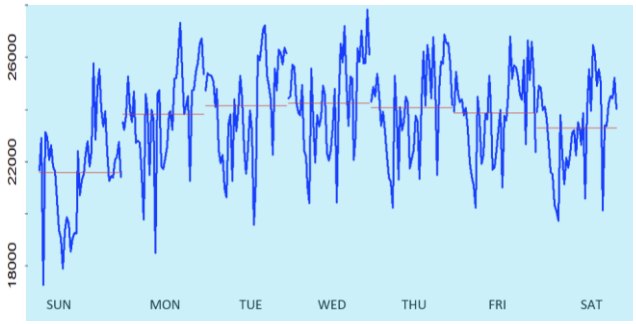
Dato che la serie è composta da osservazioni campionante ogni 10 minuti, il segnale è soggetto a molti tipi di stagionalità.
Emerge chiaramente una stagionalità prevalente sulle altre: quella ogni 144 osservazioni (1 giorno), ma ne sono presenti anche altre: settimanale e mensile.
In primo luogo, sono stati provati approcci che partivano dalla serie storica nella sua forma originale, ma a causa di questa sua complessa caratteristica i risultati sono stati scadenti.
Si è quindi deciso di scomporla in 144 serie storiche diverse, in maniera tale che ognuna di esse avesse frequenza giornaliera.
Si è poi svolta l'analisi su una serie di esempio al fine di selezionare il miglior modello per ogni categoria.
Per scegliere invece il miglior modello complessivo si sono analizzati i migliori 3 modelli su un validation-set (ultimi 30 giorni) della serie originale.



3. ANALISI PRELIMINARE

Per l'analisi dei modelli si è scelta, in maniera randomica, la quarantesima colonna del dataset, corrispondente alle h. 6:30 am.
La serie storica ottenuta è composta da 334 osservazioni: le ultime 34 sono state oscurate al fine di prevederle e valutare le performance.

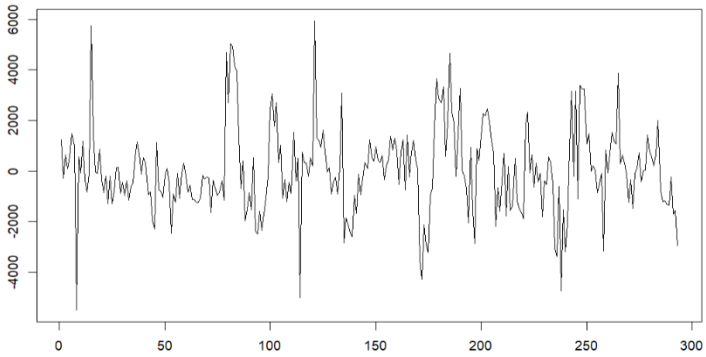


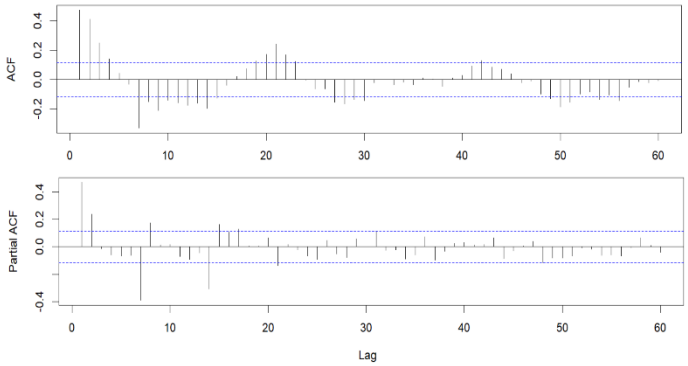


Dalle analisi svolte, l'unica stagionalità che emerge in maniera significativa è quella settimanale.

ARIMA

Per l'applicazione dei modelli ARIMA è innanzitutto stata valutata la stazionarietà della serie storica.
È risultato necessario eliminare la non stazionarietà in media e anche la stagionalità presente.
La serie risultava invece essere stazionaria in varianza, perciò non è stata applicata alcuna trasformazione di Box-Cox.
La differenza stagionale (s=7) sembra risolvere entrambi questi problemi.



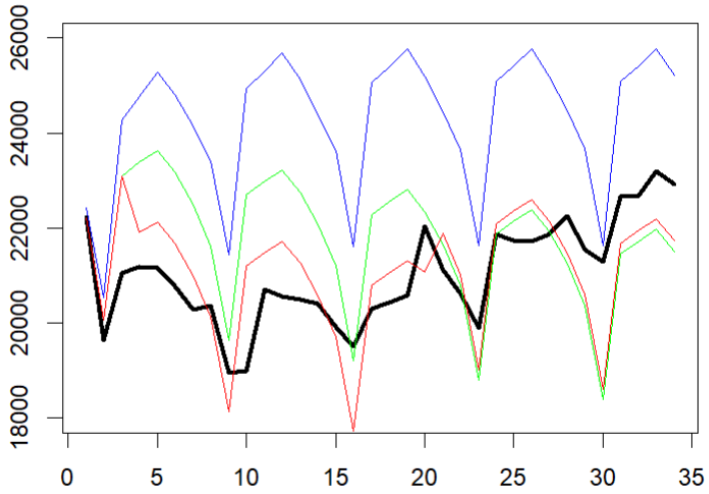


Dal correlogramma si notano chiaramente una componente stagionale e una componente AR di ordine 2.

Vengono provati i seguenti modelli:

MODEL	MAE	COLOR
ARIMA $(2,0,0)(0,1,1)_7$	3265	
ARIMA $(2,0,0)(0,1,1)_7$	1395	
ARIMA $(2,1,0)(0,1,1)_7$ + features esterne	859	

Di seguito si mostra il grafico delle previsioni dei modelli sopra elencati sui 34 valori finali della serie di esempio.
I colori utilizzati fanno riferimento alla tabella precedente.



Il modello migliore risulta essere quello che oltre alla differenza stagionale utilizza anche una differenza semplice, e in più usa per la previsione anche dei regressori costruiti a mano.
Le feature a cui si fa riferimento sono:

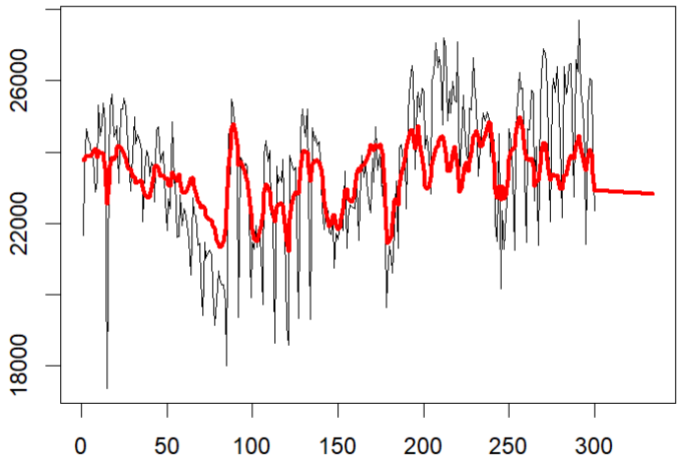
- Venerdì / Sabato / Domenica (3 colonne Booleane)
- Festività prese dal calendario del Marocco (Utilizzando diverse fonti, come ad esempio i giorni di chiusura delle banche in quell'anno reperita al seguente link: https://assiomforex.it/archivioFiles/Calendario2017.pdf)
- Festività intuite dai dati, stando attenti a non incombere nel fenomeno dell'overfitting.

UCM

Per l'applicazione dei modelli UCM non è stato necessario gestire la stazionarietà, e di conseguenza si è potuto lavorare sulla serie originale. Si è deciso di utilizzare le seguenti componenti:

- Local Linear Trend: *per modellare i diversi livelli*
- Stagionalità a Dummy Stocastiche (usando s=7): *per modellare la stagionalità settimanale*
- Stagionalità a Sinusoidi Stocastiche: *per modellare le altre stagionalità*

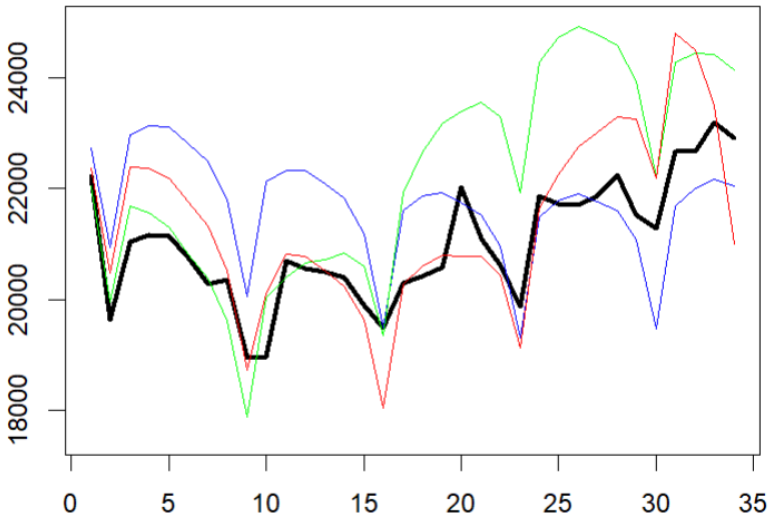
I risultati sono piuttosto soddisfacenti, di seguito si mostra ad esempio il livello estratto dalla serie:



Anche questa volta si è scelto di confrontare diverse variazioni del modello, provando due diverse quantità di sinusoidi stocastiche, e provando come prima ad utilizzare le features aggiuntive.

MODEL	MAE	COLOR
6 Sinusoidi	1130	
16 Sinusoidi	1316	
6 Sinusoidi + features esterne	789	

Di seguito si mostra il grafico delle previsioni dei modelli sopra elencati sui 34 valori finali della serie di esempio.
I colori utilizzati fanno riferimento alla tabella precedente.



Anche in questo caso le features aggiuntive sembrano essere importanti ai fini della previsione.

ML - based

Per l'applicazione dei metodi basati su algoritmi di Machine learning si sono sperimentati diversi approcci, anche se purtroppo non è stata trovata una soluzione che producesse risultati interessanti.

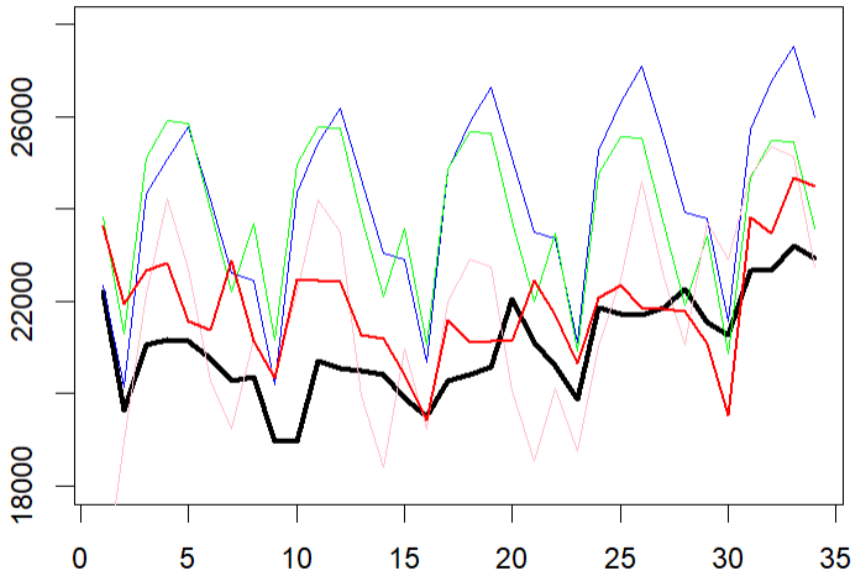
Si sono provate due famiglie di algoritmi, entrambe valutate usando sia RandomForest che XGBoost come classificatori:

- Algoritmi a previsioni ricorsive
- Algoritmi a previsioni dirette

Dato che nessuna di esse ha prodotto risultati accettabili, si è provato un approccio diverso utilizzando il KNN.
I parametri di quest'ultimo sono stati analizzati in una Grid-Search effettuata su una serie storica diversa da quella usata fino ad ora (h 10.00 am), in modo da poter poi valutare sul nostro campione (h 6.30 am) le performance del modello vincente.

MODEL	MAE	COLOR
recursive	3229	
Direct (RF)	2824	
Direct (XGBoost)	1732	
KNN	1103	

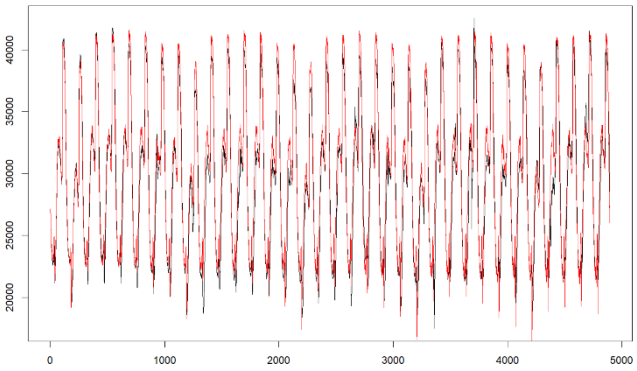
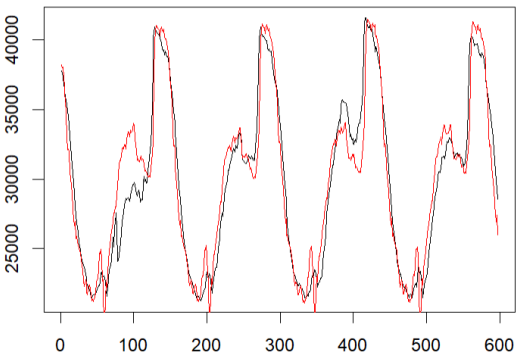
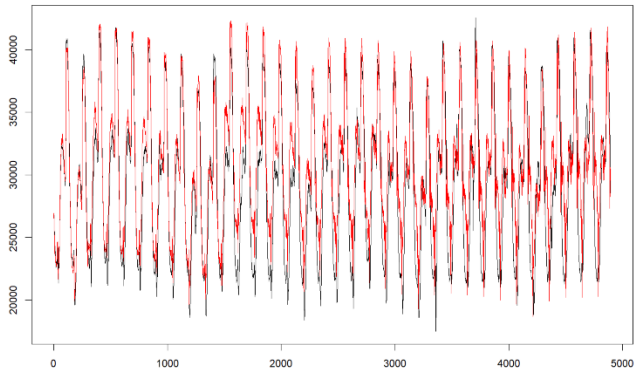
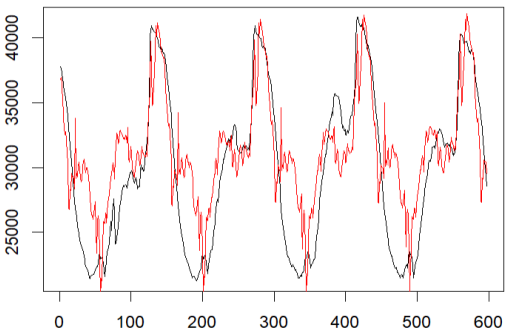
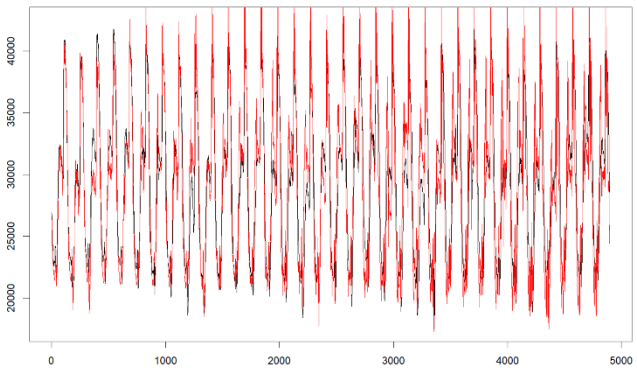
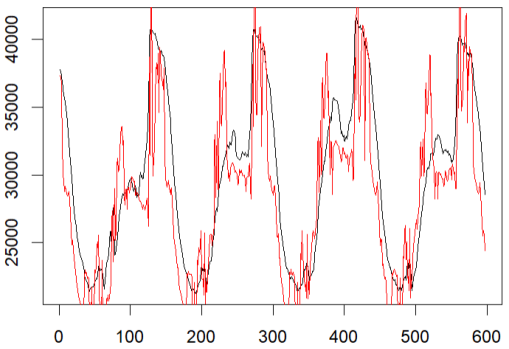
Di seguito si mostra il grafico delle previsioni dei modelli sopra elencati sui 34 valori finali della serie di esempio.
I colori utilizzati fanno riferimento alla tabella precedente.



Nonostante le previsioni non siano ottime, emerge comunque che il KNN è nettamente migliori degli altri approcci.
Si nota però che questa performance è dovuta solo al fatto che effettua previsioni su un livello simile a quello della serie storica, senza però aver colto particolarmente l'andamento.

VALUTAZIONE FINALE

Dopo aver provato i modelli descritti sulla serie storica di esempio, si ha un'idea di quali sono le opzioni che performano meglio all'interno di ogni famiglia, tuttavia per avere una stima più robusta della bontà delle previsioni, in quest'ultima fase, il modello migliore di ogni categoria viene valutato sull'ultimo mese del dataset originale.
Viene anche riportato un focus sugli ultimi quattro giorni (i più difficili da prevedere), in modo che sia possibile valutarlo visivamente.

	PREVISIONE TOTALE	ULTIMI 4 GIORNI
ARIMA *(MAE=987)*		
UCM *(MAE=1371)*		
KNN *(MAE=2138)*		

In conclusione, il miglior modello tra quelli utilizzati sembra essere:

- **ARIMA (2,1,0) $(0,1,1)_7$ con features aggiuntive**