Streaming data management and time series analysis
Forecast power time series

Andrea Maver - matricola 828725

Sommario

Con il seguente progetto ci si pone il problema di studiare una serie storica di consumi elettrici con il fine di prevederne gli andamenti un mese in avanti. Sono stati utilizzati tre metodi principali, cioè ARIMA, UCM e modelli di machine learning; per ognuno di essi sono state sviluppate molte versioni con configurazioni diverse, in modo tale da poter scegliere la più adatta. Le valutazioni sono state effettuate utilizzando la metrica MAE, oltre che la valutazione grafica.

1 Introduzione

L'obiettivo dello studio è quello di confrontare diversi metodi di studio e previsione delle serie storiche, per valutare quale sia il migliore a prevedere un mese di dati elettrici. Le metodologie scelte sono state ARIMA (*modello autoregressivo integrato a media mobile*), UCM (*modello a componenti non osservabili*) e modelli machine learning (in particolare *k-nearest neighbors* e una rete neurale ricorrente).

La serie fornita è composta da osservazioni ogni 10 minuti e va dal 2017-01-01 00:00:00 al 2017-11-30 23:50:00, cioè dal 1 gennaio al 30 novembre. L'obiettivo finale sarà quello di prevedere il mese di dicembre, cioè le osservazioni da 2017-12-01 00:00:00 a 2017-12-31 23:50:00. In figura 1 si può osservare la serie completa e delle sue sotto-sequenze, che rendono più efficace la visualizzazione.

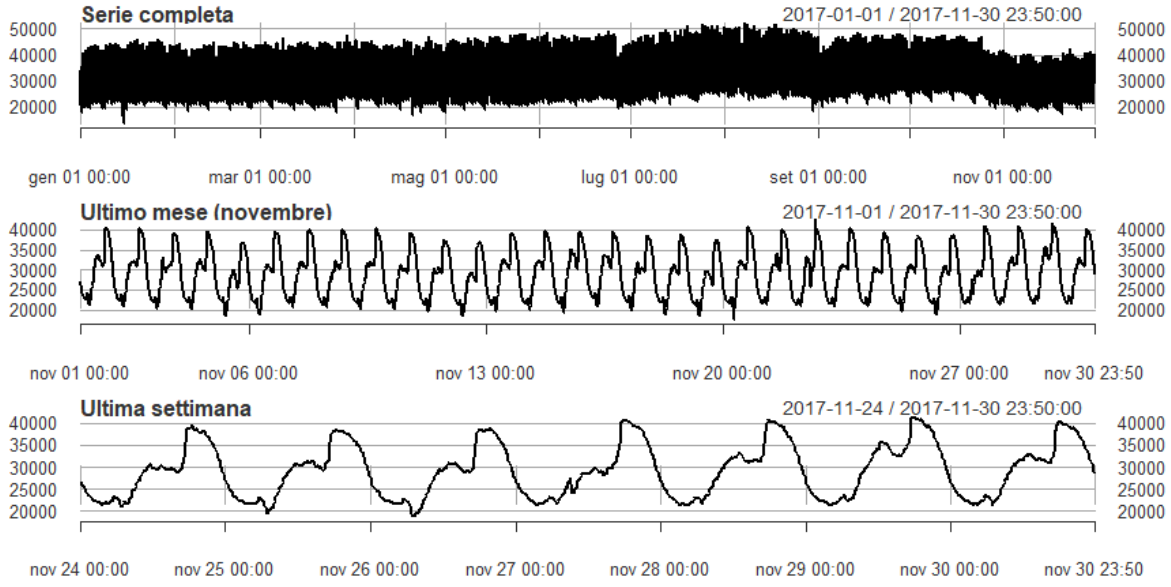

Figura 1: Serie totale, mese di novembre e ultima settimana

Dal primo grafico si osserva come la serie sia essere piuttosto regolare, senza particolari picchi o avvallamenti; sembra esserci un leggero trend crescente nella prima parte che si inverte

intorno ad agosto, dove inizia una discesa. Dalla seconda visualizzazione, relativa solo al mese di novembre, si nota una forte ripetitività giorno per giorno, ribadita anche nel terzo grafico.

Tramite il test di Dickey-Fuller si verifica la condizione di stazionarietà della serie, rifiutando l'ipotesi nulla. Utilizzando la funzione BoxCox si può cercare se una trasformazione si adatti al meglio per modellare la serie nelle fasi successive; il valore di 0.22 fornito dal test non dà chiare indicazioni, in quanto una trasformazione del genere potrebbe rendere più difficile l'interpretazione del modello. Alcuni test sono stati svolti con la più classica trasformazione logaritmica, ma non avendo portato a miglioramenti osservabili si è scelto di operare sulla serie originale.

Per poter testare e valutare i modelli è necessario dividere la serie in due sottoinsiemi, uno di training e uno di validation. Per fare ciò si sono scelti i primi 10 mesi come serie di addestramento e l'ultimo mese di novembre come serie di validazione dei risultati, che rappresentano il 91% del totale per il primo e il restante 9% per il secondo. In figura 2 si può osservare a che altezza della serie è stato effettuato lo split.

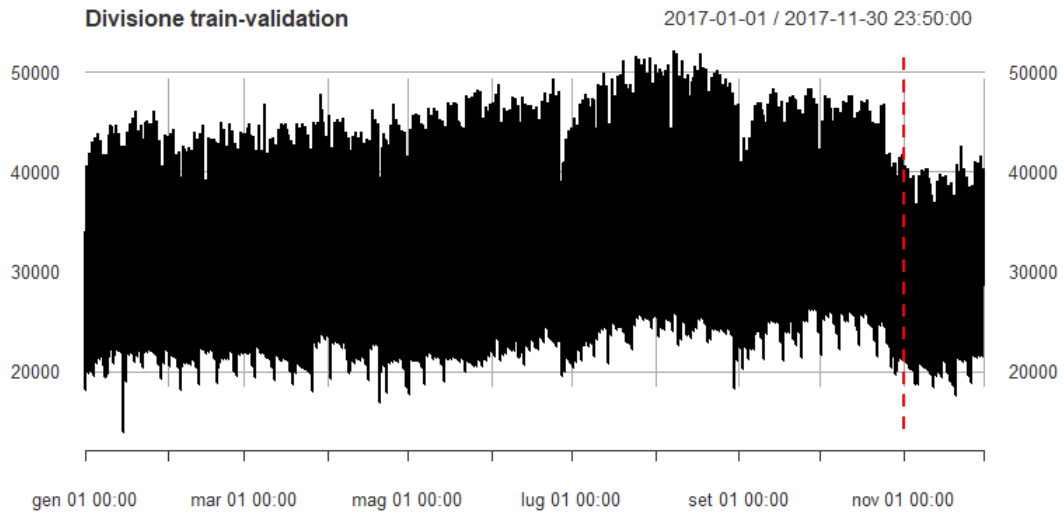

Figura 2: Split train-validation

Per valutare i modelli è stata scelta la metrica MAE (mean absolute error) che indica l'errore medio di ogni previsione rispetto all'osservazione reale; a questa è sempre stata aggiunta una valutazione grafica, necessaria in questo tipo di analisi per osservare che i risultati siano verosimili.

2 Modello baseline - Regressione lineare

Per iniziare l'analisi è stato deciso di applicare un modello molto semplice, cioè la regressione lineare, che non tiene conto della dipendenza temporale tra le osservazioni, in modo tale da analizzarne i residui per poter applicare le tecniche successive, sopratutto ARIMA, in modo più ragionevole.

Nella regressione la serie viene modellata utilizzando un trend quadratico, scelto dopo aver osservato l'andamento completo, delle variabili dummy relative al giorno della settimana e delle sinusoidi di frequenza 144. Le dummy sono variabili di comodo che assumono sempre valore zero tranne nei momenti che rappresentano, in questo caso i giorni della settimana; se ne utilizzano sei, in quanto i valori assunti dalle stime saranno in relazione con quella mancante. Per questa implementazione è stato scelto di rimuovere la variabile relativa al mercoledì. Le sinusoidi vengono utilizzate per modellare stagionalità morbide e sono state scelte dopo aver osservato la serie settimanale; la frequenza di 144 è stata applicata in quanto ogni giorno è composto da altrettante osservazioni, perciò si può modellare una stagionalità giornaliera.

Osservando i coefficienti dei vari regressori si osserva come tutte le dummy siano significative tranne quella relativa al giovedì; per quanto riguarda le sinusoidi, per il modello bastano otto coseni e sette seni. È necessario ricordare però che i p-value non sono particolarmente attendibili, in quanto i dati presentano relazioni temporali tra loro; detto questo, possono comunque essere un buon indicatore iniziale per fare delle prime considerazioni.

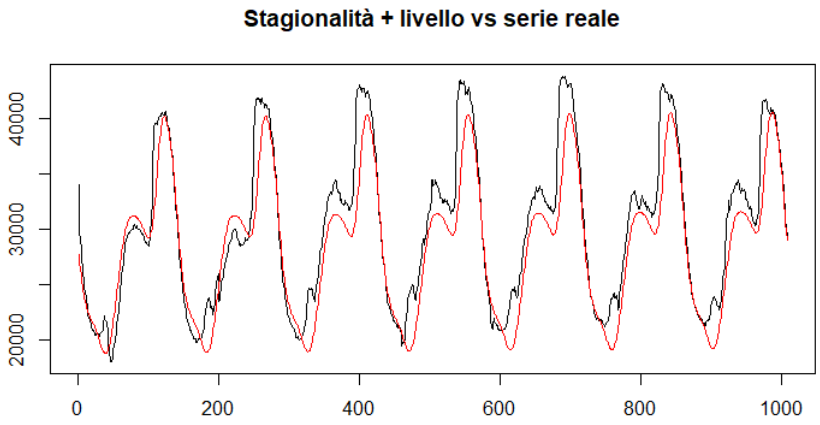

Figura 3: Stagionalità + livello e serie reale

La figura 3 confronta la prima settimana di dati, in nero, con la somma di stagionalità e livello, in rosso; si può osservare come i regressori sembrano essersi adattati bene alla ripetitività della serie, riuscendo a modellare la crescita compreso l'avvallamento centrale.

Per poter procedere con le analisi è necessario osservare i grafici Acf e Pacf dei residui della regressione.

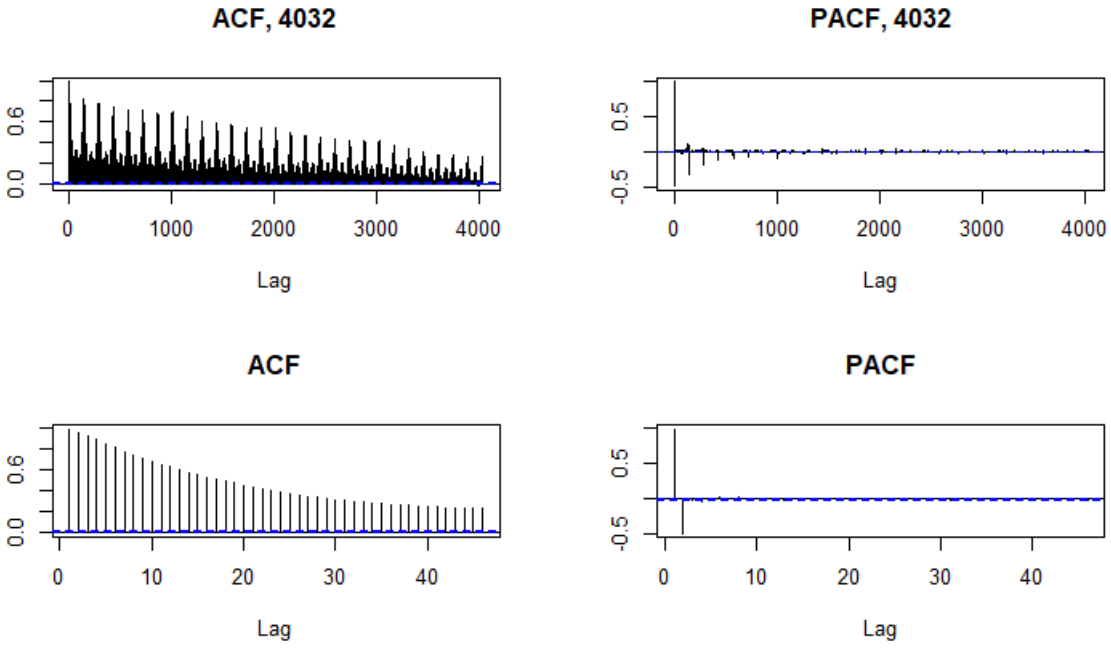

Figura 4: Acf e Pacf rispettivamente per 4032 lag e per 46 lag

Sulla prima riga della figura 4 si possono osservare i grafici per 4032 ritardi, cioè quattro settimane di dati; si nota come la Acf scenda molto lentamente verso zero con dei picchi a lag

144 e multipli; la Pacf presenta due picchi iniziali e poi si discosta dalle bande di confidenza solo in occasione degli stessi lag 144. Andando più nello specifico e osservando meno ritardi totali nella seconda riga si nota lo stesso pattern discendente per l'Acf e si notano chiaramente valori a 1 e 2 lag nella Pacf, oltre che per 3 e 4 ma con molta meno incidenza. Queste considerazioni saranno la base per lo sviluppo dei successivi modelli ARIMA.

3 ARIMA

In virtù delle considerazioni fatte nella sezione precedente è possibile iniziare a sviluppare i modelli ARIMA, o più nello specifico SARIMAX, in quanto verranno testate anche configurazioni con stagionalità e regressori esterni. I diversi tentativi non verranno trattati allo stesso modo per non rendere troppo ripetitiva la spiegazione, ma verranno comunque esposti tutti i motivi che hanno portato a particolari scelte e allo sviluppo di questo processo in un certo senso definibile *trial and error*. In Appendice sarà possibile anche visionare alcuni grafici di Acf e Pacf, oltre che le distribuzioni dei residui. Le performance di ogni modello saranno fornite alla fine di questa sezione e permetteranno di scegliere il modello migliore con cui eseguire le previsioni.

Il primo modello testato è un Arima(2,0,0) molto semplice e senza parte stagionale, che si è invece deciso di configurare tramite regressori, cioè le dummy e le sinusoidi descritte in precedenza, che dovrebbero permettere di modellare la stagionalità stagionale e quella giornaliera; a queste componenti viene anche aggiunto un drift. Nonostante i residui sembrino piuttosto normali, da Acf e Pacf si nota che è rimasta una forte dipendenza stagionale a 144 e anche al primo lag. Gli errori sul validation sono molto alti, perciò si decide di proseguire con i tentativi.

Il secondo modello è simile al precedente, cioè un (3,0,0) con regressori e drift ma senza componente stagionale. Nonostante gli errori siano molto minori dei precedenti, a dimostrare la preferibilità del modello, sono ancora presenti forti dipendenze a lag 144. Questo potrebbe significare che modellare la stagionalità utilizzando solamente i regressori non sia opportuno.

Per il terzo modello viene scelto di abbandonare l'uso dei regressori per creare un modello Sarima (3,0,0)(0,1,0)[144]. Acf e Pacf sono chiaramente i migliori tra quelli ottenuti, infatti il primo presenta solo una banda a lag 144 e alcune leggermente fuori soglia per 1008, mentre il secondo scende gradualmente verso zero con picchi solo rispettivamente ai multipli di 144. Anche gli errori sul validation sono i migliori riscontrati finora.

Tramite le considerazioni fatte si decide di aggiungere al modello tre un'ulteriore componente MA(1) stagionale, in modo da creare il quarto e finale modello (3,0,0)(0,1,1)[144]. Questo porta ai risultati migliori in termini di errori, sia sul train che sul validation set; Acf e Pacf sono mostrati in figura 5.

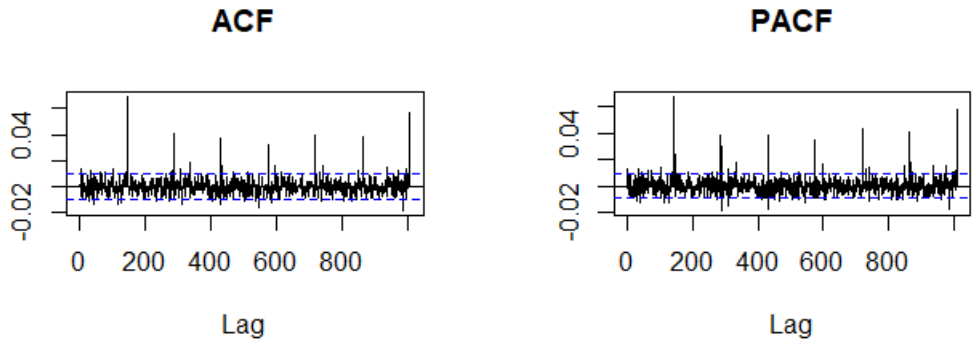

Figura 5: Acf e Pacf Sarima (3,0,0)(0,1,1)[144]

Si nota come, rispetto ai casi precedenti, i grafici siano molto simili tra loro, ancora con le solite dipendenze a 144, ma questa volta molto meno significative. I residui mostrati in figura

6 possono definirsi normali e i picchi che compaiono non seguono delle particolari distribuzioni; osservando il comportamento delle previsioni di validation in blu, rispetto alle osservazioni reali in nero (figura 7), si nota come si adatti bene alla ripetitività della serie, sovrastimando leggermente i picchi e sottostimando gli avvallamenti.

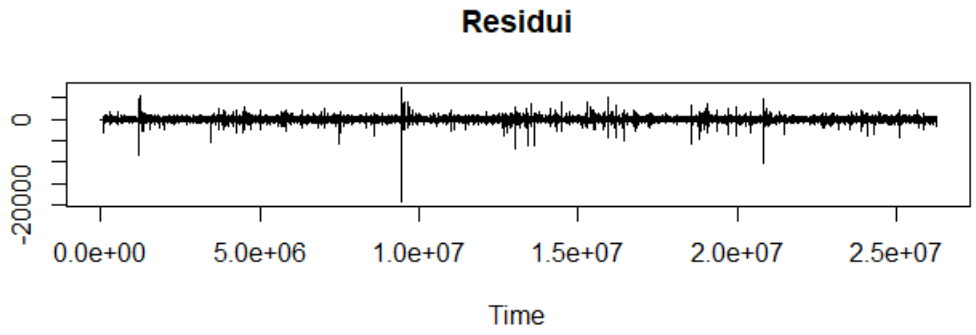

Figura 6: Residui Sarima (3,0,0)(0,1,1)[144]

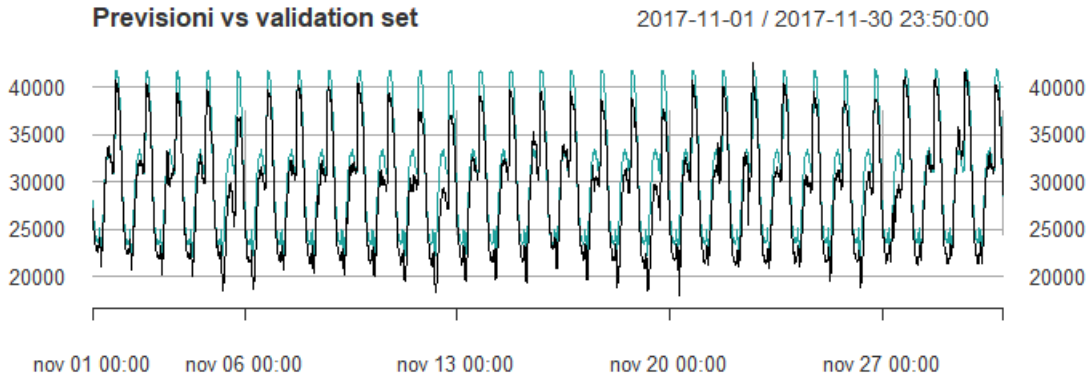

Figura 7: Previsioni validation (3,0,0)(0,1,1)[144]

Di seguito (tabella 1) vengono mostrati i risultati in termini di MAE e MAPE dei quattro modelli testati.

Modello	**MAE**	**MAPE**
Arima(2,0,0) + reg	6530.932	23.35308
Arima(3,0,0) + reg	1901.971	6.456155
Sarima(3,0,0)(0,1,0)[144]	1858.162	6.692717
Sarima(3,0,0)(0,1,1)[144]	1839.028	6.69136

Tabella 1: Risultati modelli ARIMA

Il quarto modello è il migliore in termini di MAE e il secondo migliore in termini di MAPE (*mean absolute percentage error*), e le sue previsioni sono errate in media solo del 6.69% rispetto alla realtà; per questo motivo, oltre che quelli esposti precedentemente, verrà utilizzato per le successive previsioni del mese di dicembre.

4 UCM

Lo sviluppo di questa sezione rispecchia quello della parte ARIMA, con una breve descrizione dei modelli testati, dei loro risultati e delle scelte successive che hanno portato allo sviluppo del migliore.

Per iniziare si è deciso di optare per un modello molto semplice, composto solamente da un random walk per la componente trend e 20 sinusoidi con frequenza 144 per la componente stagionale. Per inizializzare il fit del modello la varianza è stata ridistribuita su tutti gli errori e a seguire viene applicato uno smoother per ottenere le previsioni; il modello raggiunge convergenza. Osservando i risultati numerici e grafici sembra che questa semplice configurazione si adatti bene ai dati, di cui sottostima leggermente gli avvallamenti.

Il secondo modello è uno sviluppo del primo, al quale vengono aggiunte delle ulteriori sinusoidi ma di frequenza 1008, cioè sette giorni di osservazioni, a rappresentare una stagionalità settimanale. Lo sviluppo è identico al precedente, con l'unica differenza che come parametri iniziali vengono forniti quelli ottimali stimati dal primo modello, in modo tale che questo possa partire da una situazione più vantaggiosa. Anche questo raggiunge convergenza; le performance sono leggermente migliori del precedente, e dal punto di vista grafico (figura 8) il fit sembra ancora migliore.

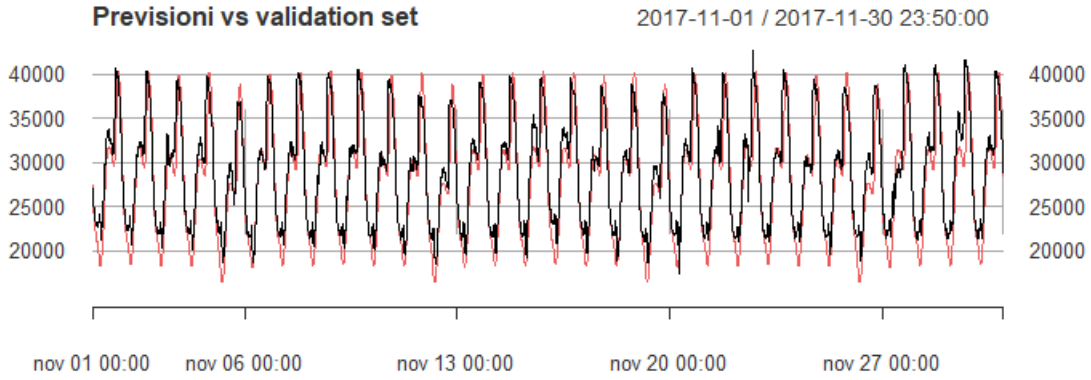

Figura 8: Previsioni validation UCM

Il terzo e ultimo modello utilizza dei regressori, cioè le sei dummy create in precedenza, per modellare la stagionalità settimanale al posto delle sinusoidi a frequenza 1008, mentre le altre vengono mantenute. Le performance sono migliori rispetto a quelle del primo modello, a dimostrazione che la definizione anche della stagionalità settimanale porta a dei miglioramenti, ma peggiori rispetto a quelle del secondo, che porta a concludere che le sinusoidi in questo caso pratico sono più adatte delle dummy stocastiche.

Nella tabella 2 sono mostrate le prestazioni dei tre modelli esposti; tramite queste e le valutazioni grafiche il secondo modello è stato scelto come migliore per la previsione dei nuovi dati.

Modello	**MAE**	**MAPE**
UCM stag 144	1990.408	6.818897
UCM stag 144 + stag 1008	1879.281	6.475925
UCM stag 144 + dummy	1983.751	6.794245

Tabella 2: Risultati modelli UCM

5 Machine learning

Per quanto riguarda la parte di machine learning sono state utlizzate due tecniche, cioè k-nearest neighbors e una rete neurale ricorrente costruita tramite layer LSTM.

5.1 KNN

Questa tecnica di previsione delle serie storiche è molto particolare perchè non si basa su un modello specifico sottostante, ma cerca le sotto-sequenze più simili rispetto a un target per poi osservare come si distribuiscono i valori successivi a queste ultime; è particolarmente indicata per serie molto ripetitive, come in questo caso pratico. Il pacchetto R 'tsfknn' permette di applicare la tecnica facilmente, cambiando i parametri base per testare diverse configurazioni. I più significativi sono *lags*, che indica la lunghezza massima delle sotto-sequenze da individuare, k cioè quante cercarne, *msas*, in questo caso è stato scelto MIMO e *cf*, settato come mediana. Nella tabella 3 sono presenti alcune delle configurazioni testate, con le rispettive performance; si specifica che altri tentativi sono stati effettuati con diversi valori di *msas* e *cf*, ma che non hanno portato a miglioramenti globali delle previsioni.

Modello	**Lags**	**K**	**MAE**	**MAPE**
KNN 1	1:1008	10	2346.035	8.034549
KNN 2	1:4032	10	3065.367	10.82601
KNN 3	1:1008	5	2365.022	7.99821
KNN 4	1:1008	20	2348.963	8.04137

Tabella 3: Risultati e configurazioni KNN

Osservando le performance sia a livello di metriche che graficamente si nota che la prima configurazione è la migliore tra quelle testate, perciò, dopo il confronto con la rete neurale, sarà quella che verrà eventualmente utilizzata per le previsioni future. In figura 9 si possono osservare le previsioni sul validation set rispetto alle osservazioni reali.

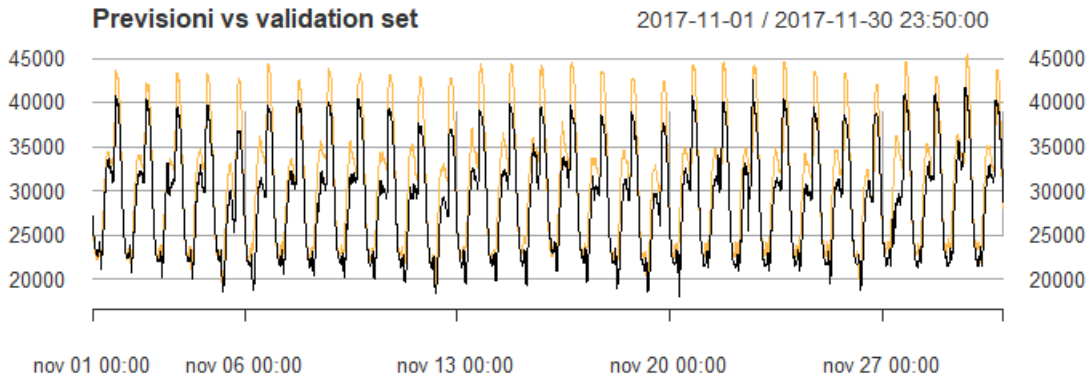

Figura 9: Previsioni validation KNN

5.2 Recurrent neural network

L'altra tecnica ML testata è una rete neurale ricorrente, specifica per dati in cui la disposizione ordinale è importante. La rete implementata è molto semplice ed è formata da due layer LSTM (*long short term memory*) composte da rispettivamente 25 e 20 unità e da un layer Dense con ottimizzatore *adam*. I dati vengono scalati tramite un MinMaxScaler che li porta in un range [0,1] e per creare il train viene scelto come periodo di look-back 4320, cioè 4 settimane di dati. Questa metrica influenza la lunghezza delle finestre che la rete utilizzerà per l'addestramento;

influenza fortemente i tempi di allenamento, perciò si è dovuto cercare un trade off tra un valore non troppo basso ma che non facesse crollare i tempi computazionali.

Il network viene addestrato per sole cinque epoche, sempre a causa dei citati problemi computazionali, ma nonostante questo raggiunge bassi errori sul set di validazione (MAE = 856.101; MAPE =2.9414); anche osservando la figura 10 si nota come le previsioni si adattino molto bene alle osservazioni reali.

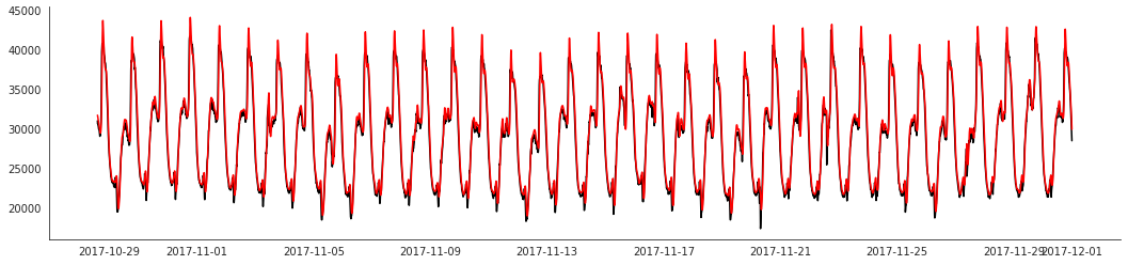

Figura 10: Previsioni validation LSTM

Nonostante questa rete neurale performi molto bene in termini di validation, quando testata per le nuove previsioni ha avuto performance molto peggiori di quelle degli altri modelli concorrenti; infatti, i nuovi valori oscillavano tra massimi e minimi molto più schiacciati rispetto a quelli della serie originale, poco verosimili considerando l'andamento generale e il tipo di dato analizzato. Per dare ulteriore contesto, nella seguente tabella 4 vengono mostrate le varianze della serie e delle previsioni con KNN e con la rete. Come si può notare quella della rete è minore di due ordini di grandezza rispetto alle altre, a dimostrare la poca bontà del network; anche graficamente (figura 11) si nota subito la differenza con le previsioni del KNN, molto più plausibili.

Anche se non si dovrebbe scegliere il modello migliore in base ai dati di test ma solo su quelli di validation, date le considerazioni fatte e osservando che in una materia come il forecast di serie storiche l'osservazione delle previsioni sia una parte imprescindibile del processo di sviluppo, per quanto riguarda la sezione di machine learning il modello scelto è stato il k-nearest neighbors.

	Var
Serie completa	50906683
KNN	45765264
LSTM	837826

Tabella 4: Confronto varianze

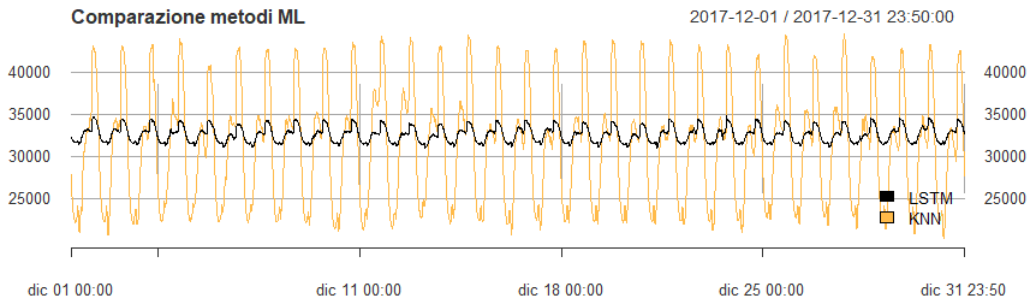

Figura 11: Confronto tra previsioni LSTM e KNN

6 Previsioni e conclusioni

Dopo aver confrontato tutti i modelli e scelto i migliori è necessario ora completare lo studio prevedendo il mese di dicembre: per farlo i tre modelli vengono addestrati nuovamente ma sull'intera serie invece che solamente sul train set. Si ricordano sotto i modelli scelti e le performance sul validation:

Modello	MAE	MAPE
ARIMA	1839.028	6.69136
UCM	1879.281	6.475925
KNN	2346.035	8.034549

Tabella 5: Modelli finali

Di seguito vengono mostrate le nuove serie predette, confrontandole l'una con l'altra (figura 12) e in relazione alla serie completa (figure 13, 14 e 15).

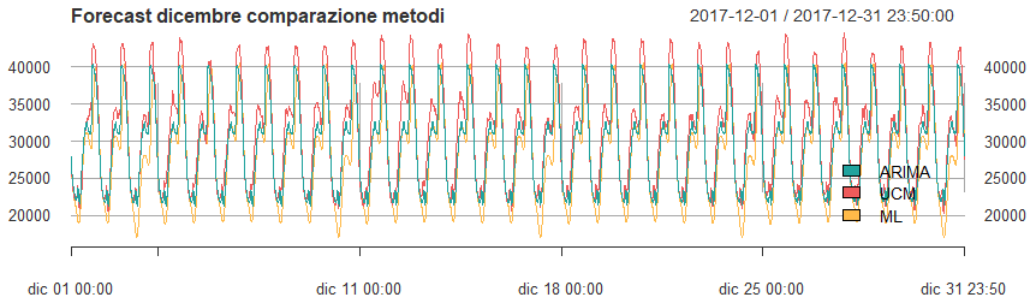

Figura 12: Confronto mese di dicembre previsto dai tre modelli

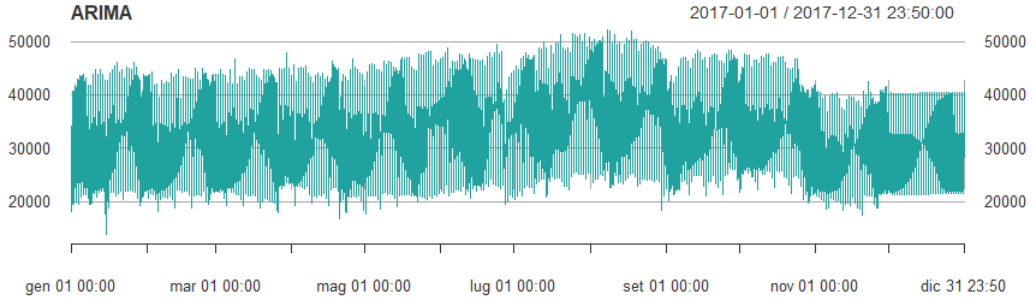

Figura 13: Serie totale con previsioni ARIMA

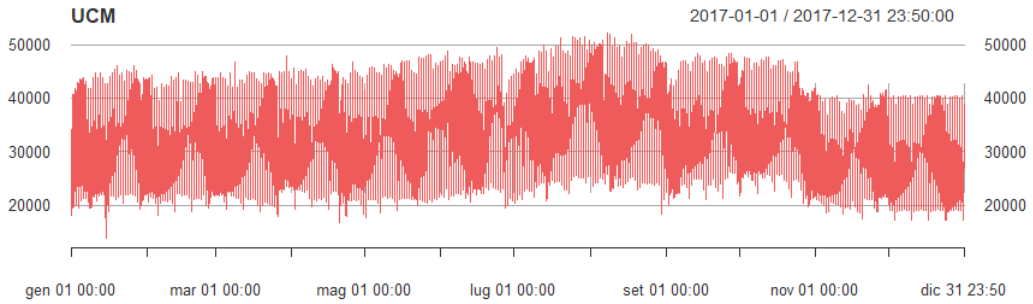

Figura 14: Serie totale con previsioni UCM

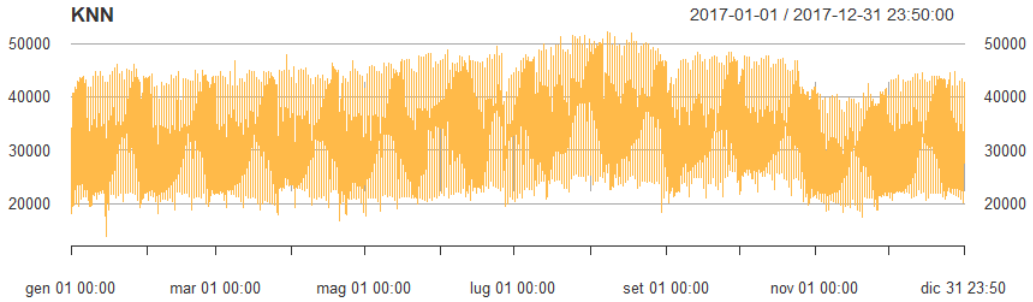

Figura 15: Serie totale con previsioni KNN

Dal punto di vista delle performance sul validation i tre modelli sono molto simili tra loro, con il KNN che sbaglia mediamente di due punti percentuali in più rispetto agli altri.

Osservando i grafici si può concludere che le previsioni ARIMA sono le più equilibrate, mentre UCM e KNN hanno rispettivamente i picchi più alti e gli avvallamenti più bassi delle tre. Per quanto riguarda il livello generale delle serie totali, ARIMA e UCM rimangono molto stabili, emntre KNN presenta delle oscillazioni leggermente più accentuate.

Appendice

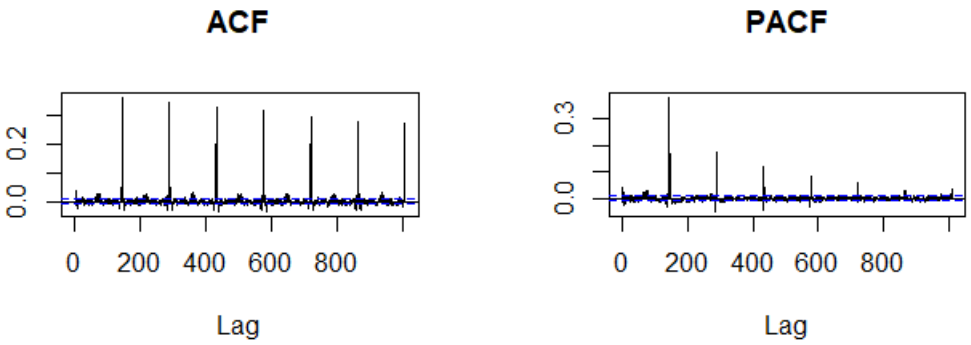

Figura 16: Acf e Pacf Arima 1

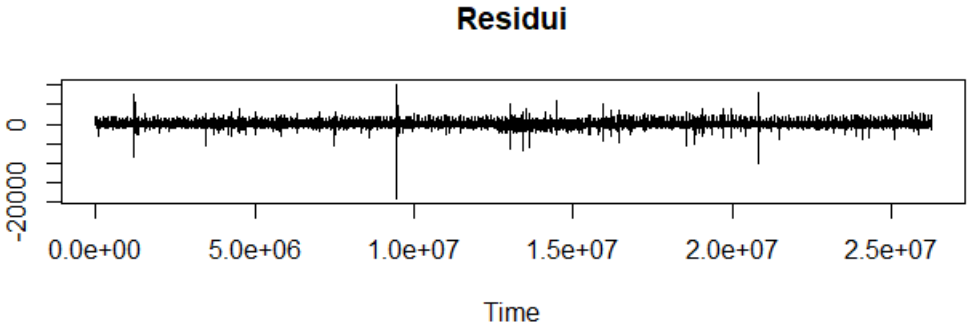

Figura 17: Residui Arima 1

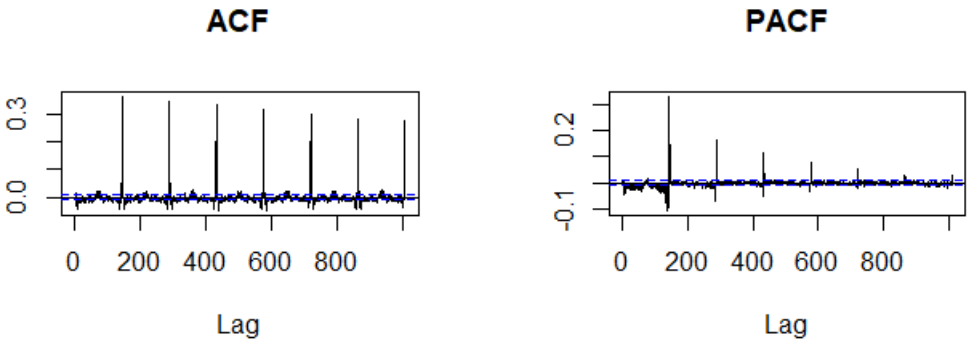

Figura 18: Acf e Pacf Arima 2

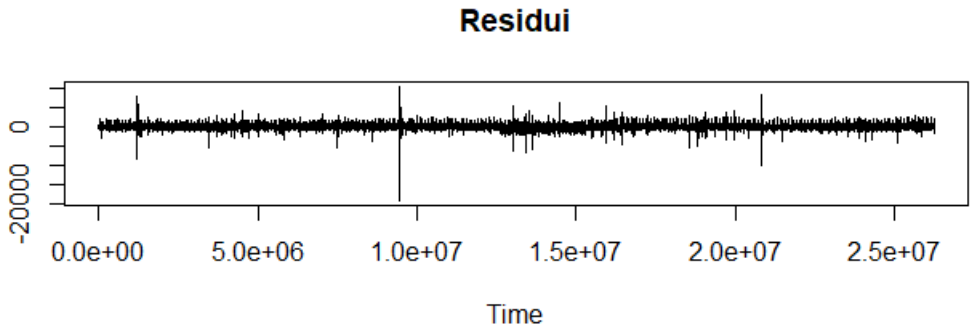

Figura 19: Residui Arima 2

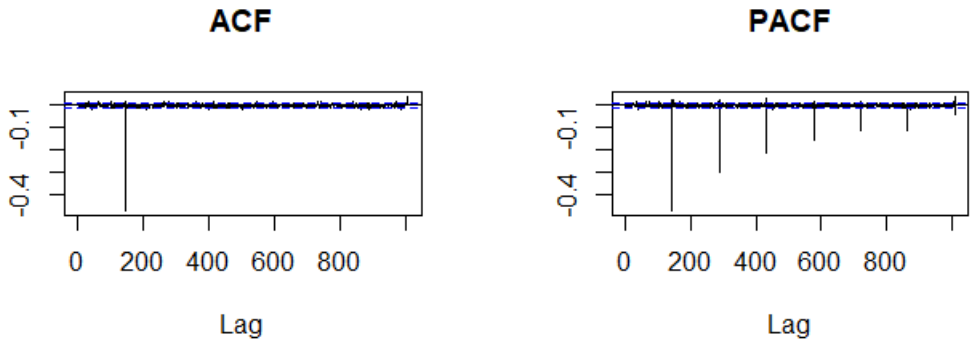

Figura 20: Acf e Pacf Arima 3

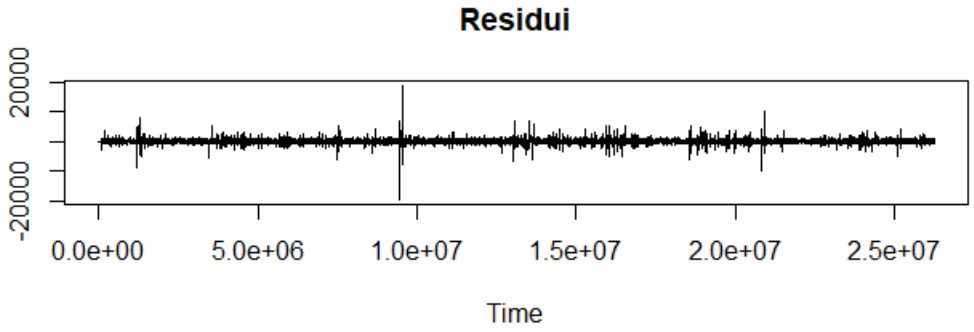

Figura 21: Residui Arima 3