Università degli studi di
Milano-Bicocca

Streaming Data Management and Time Series Analysis

Final Project

Previsioni su serie storiche

Authors:
Simone D'Amico - 850369
s.damico4@campus.unimib.it

1 Introduzione

Il progetto ha come obiettivo l'individuazione del modello predittivo migliore per serie temporali scelto tra vari modelli di tre differenti tipologie: modelli ARIMA, modelli UCM e modelli di Machine Learning.
Per ogni tipologia si individueranno e valideranno vari modelli con lo scopo di scegliere il miglior rappresentante da usare per ottenere le previsioni.

2 Preprocessing e visualizzazione dei dati

La serie temporale è una serie oraria con dati dal 01/09/2018 al 31/08/2020. Il dataset si compone di tre colonne: la colonna del giorno, la colonna dell'ora del giorno e la colonna del valore misurato e sul quale fare previsioni.
Nella prima fase si sono esplorati i dati cercando eventuali valori mancanti studiando le varie stagionalità in modo grafico e creando dei regressori per le successive analisi. Si è deciso di creare una colonna `datetime` concatenando i valori delle colonne data e ora in modo da identificare univocamente ogni riga. Nella serie sono presenti dei valori mancanti: la terza ora dei giorni

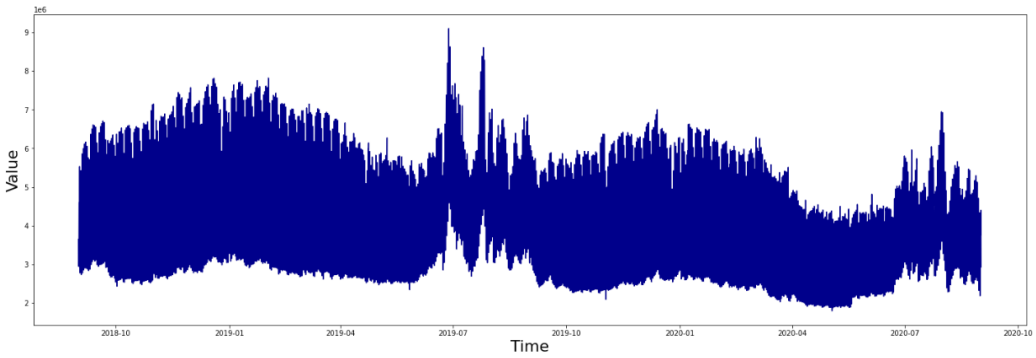

Figure 1: Plot della serie originale.

31/03/2019 e 29/03/2020, dovuto al passaggio all'ora legale e le osservazioni dell'intero giorno 31/05/2020. Nel primo caso, come suggerito dal fornitore dei dati, si è scelto di replicare il valore dell'ora precedente, mentre per il giorno mancate, una domenica, si è deciso di calcolare la media dei valori, per ogni ora, della domenica precedente e successiva. In figura 1 si mostra la serie completa.

2.1 Analisi della stagionalità

Per l'identificazione della stagionalità si è effettuata in prima battuta un'analisi grafica, visualizzando l'andamento della serie per gli anni, per i mesi, per le settimane e per i giorni; i valori sono stati aggregati con la media. In figura 2

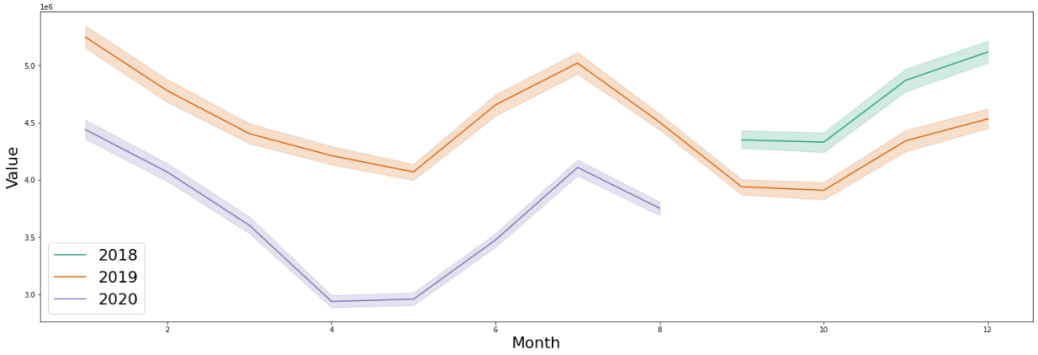

Figure 2: Stagionalità annua

si può vedere la stagionalità annuale e un andamento decrescente della serie con il picco negativo che si registra durante il periodo di lockdown. In

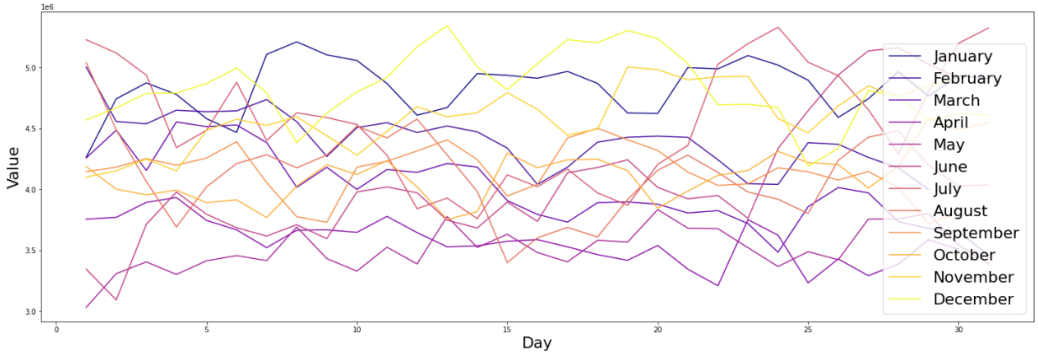

Figure 3: Stagionalità annua

figura 3 si può vedere l'assenza di una stagionalità mensile mentre in figura 4 e in figura 5 è presente una forte stagionalità, rispettivamente settimanale e giornaliera, che sarà poi considerata nella definizione dei modelli.

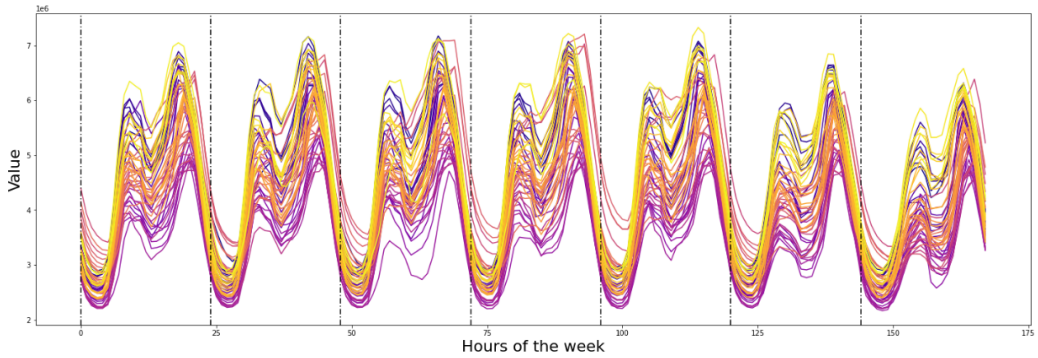

Figure 4: Stagionalità settimanale

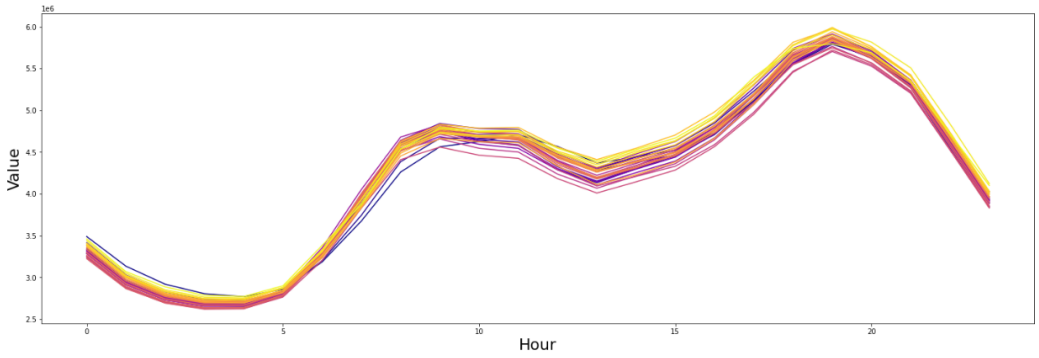

Figure 5: Stagionalità giornaliera

2.2 Aggiunta dei regressori e split del dataset

Infine si sono aggiunti dei regressori esterni, usati poi per l'analisi dei modelli lineari, come variabili binarie per ogni osservazione del dataset. Si sono costruiti tre tipi di regressori: se il giorno è in un weekend, se è una festa e un regressore chiamato `lockdown` che identifica il periodo di chiusura generale, dal 09/03/2020 al 18/05/2020, per l'emergenza Covid-19. Per determinare se il giorno è in un weekend o se è una festa si sono usate delle apposite librerie *Python*. Si è poi diviso il dataset in train set e validation set come mostrato in tabella 1. includendo nel train set anche i primi due mesi di lockdown per avere delle stime più precise.

Table 1: Divisione tra train e validation set.

	# osservazioni	% sul totale
Train set	14592	83.17%
Validation set	2952	16.83%

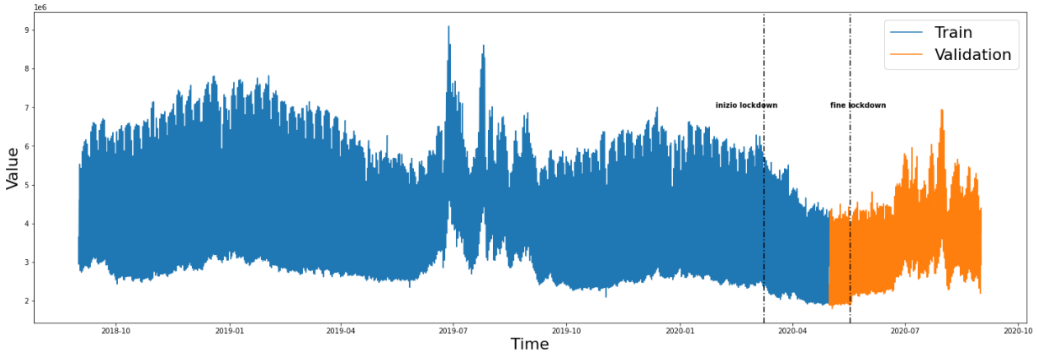

Figure 6: Serie divisa in train set e validation set.

3 Modello SARIMA

In questa sezione si descrive la scelta del modello SARIMA da usare, si parte con l'analisi della stazionarietà della serie per capire se sia opportuno o meno applicare delle trasformazioni o delle differenze alla serie. Successivamente si testano dei modelli per individuare il migliore.

3.1 Analisi stazionarietà

La condizione di stazionarietà di una serie è un requisito importante per utilizzare i modelli lineari. Si analizzano due forme di stazionarietà: la *stazionarietà in varianza* e la *stazionarietà in media*. Per la prima si studia la relazione tra media e varianza per capire se applicare delle differenze, per la seconda si utilizzano i correlogrammi e dei test statistici. Come si può vedere

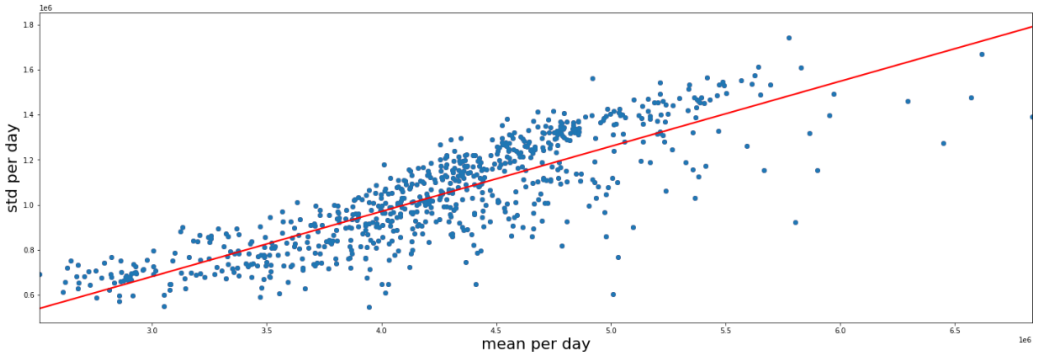

Figure 7: Relazione tra media e varianza per i giorni.

in figura 7 la relazione tra la media e la varianza per giorno sembrerebbe lineare, per questo si applica una trasformazione logaritmica ai dati.

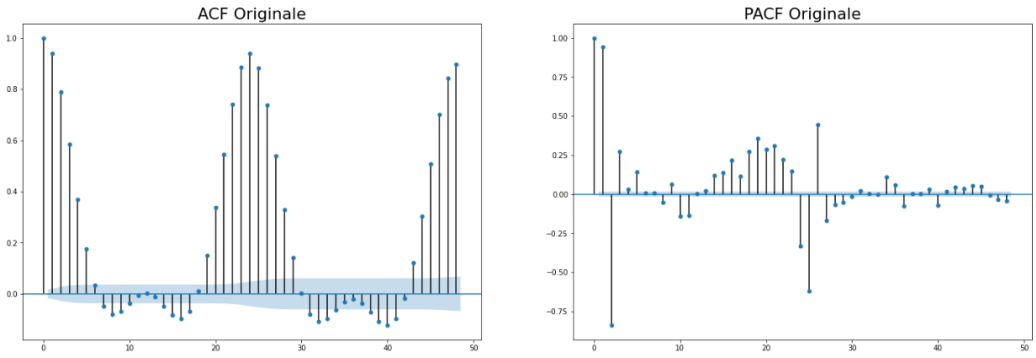

Figure 8: Correlogrammi della serie originale.

In figura 8 si mostrano i correlogrammi in cui si vede una stagionalità giornaliera ogni 24 ritardi. Per questo motivo si decide di differenziare la serie con un periodo di 24 lag, come mostrato in figura 9, la serie sembra stazionaria.

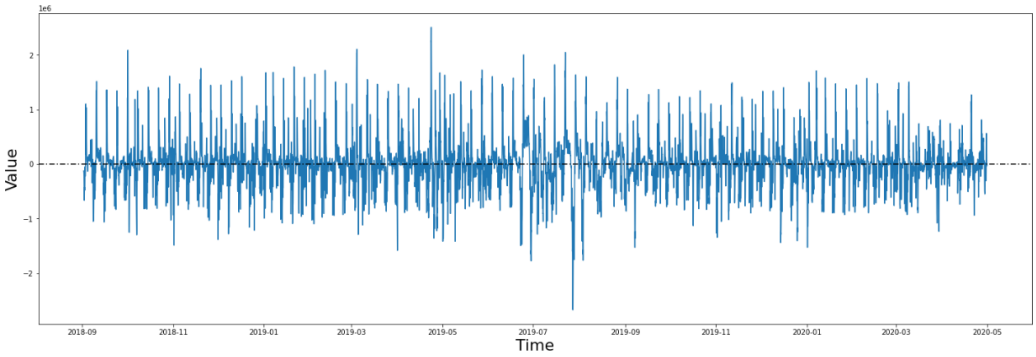

Figure 9: Serie differenziata di un periodo di 24 lag.

Per avere una maggiore conferma si esegue il test Augmented Dickey-Fuller (ADF) in cui l'ipotesi nulla è che sia presente una radice unitaria che rende la serie non stazionaria mentre l'ipotesi alternativa è l'assenza di una radice unitaria. Applicando il test alla serie differenziata si ottiene un *p-value* maggiore di 0.05 e un valore test di molto più piccolo del valore critico quindi si può rifiutare l'ipotesi di non stazionarietà della serie e si è deciso di non applicare ulteriori differenze.
In figura 10 si mostrano i grafici della ACF e PACF della serie differenziata. Il correlogramma della PACF suggerisce come coefficienti della parte regressiva $p = 2$ e $P = 1$, i coefficienti della parte MA non sono così evidenti nella ACF. Si utilizza una *grid search* per determinare i coefficienti migliori.

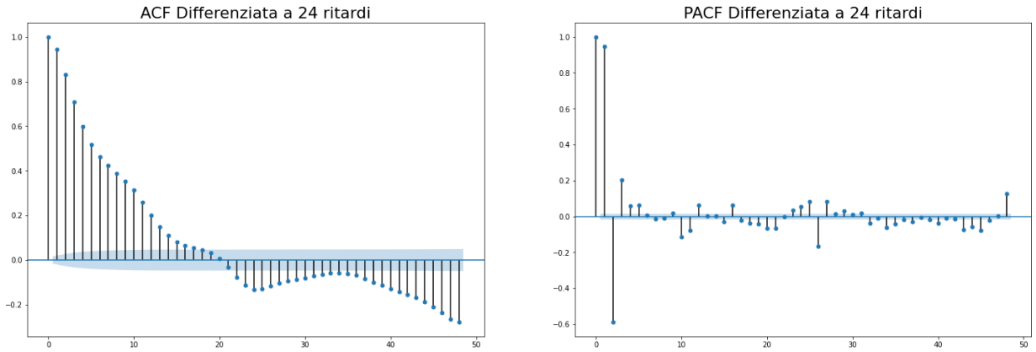

Figure 10: Serie differenziata di un periodo di 24 lag.

3.2 Scelta del modello

Si decide quindi di effettuare una *grid search* variando i coefficienti p e q da 0 a 2 e mantenendo fissi i valori di P e Q a 1 e la stagionalità a 24, il tutto viene applicato ai dati trasformati con il logaritmo. Nella tabella 2 sono riportati i valori di MAE ottenuti sia sul Test set che sul Validation set.

Table 2: Risultati dei modelli SARIMA.

Modello	Log-likelihood	MAE train	MAE Validation
SARIMA(0,0,0)$(1,1,1)_{24}$	14843.5	274785.7	726388.8
SARIMA(0,0,1)$(1,1,1)_{24}$	23309.3	152181.4	726320.8
SARIMA(0,0,2)$(1,1,1)_{24}$	28571.7	104576.2	726938.0
SARIMA(1,0,0)$(1,1,1)_{24}$	31651.2	82676.4	726202.6
SARIMA(1,0,1)$(1,1,1)_{24}$	34215.6	66682.4	727167.8
SARIMA(1,0,2)$(1,1,1)_{24}$	34650.2	63291.2	727247.9
SARIMA(2,0,0)$(1,1,1)_{24}$	34460.0	64429.8	726196.4
SARIMA(2,0,1)$(1,1,1)_{24}$	34669.0	62807.1	726737.0
SARIMA(2,0,2)$(1,1,1)_{24}$	**34696.6**	**62564.1**	**726840.6**

Il modello migliore secondo il valore di *Log-likelihood* è il SARIMA(2,0,2)$(1,1,1)_{24}$, il modello senza trasformazione logaritmica è migliore in termini di MAE sul valitadion (**711547.3**) anche se peggiore sul MAE sul train (**65584.2**), si decide quindi di mantenere il modello senza trasformazione.
Si trattano ora la stagionalità settimanale e annuale utilizzando delle serie di Fourier con periodo 168 (settimanale) e 8760 (annuale). Si applicano al modello precedentemente individuato una serie per ognuna delle due stagionalità, anche in questo caso si utilizza una grid search per trovare il numero di armoniche migliore tra 5, 10 per entrambe le stagionalità; i risultati sono

mostrati in tabella 3.

Table 3: Risultati del modello SARIMA(2,0,2)(1,1,1)$_{24}$ con serie di Fourier.

# Armoniche settimanali	# Armoniche annue	Log-likelihood	MAE train	MAE Validation
5	**5**	**-188346.3**	**65321.3**	**333732.0**
5	10	-188339.2	65286.8	384878.3
10	5	-200329.5	131894.7	335200.7
10	10	-125323722.7	334913.8	382500.9

Il miglior modello è quello che ha 5 armoniche per la settimana e 5 armoniche per l'anno in termini di MAE sul validation. A questo modello si vanno ad aggiungere i regressori per le feste e per il periodo di lockdown andando a provare tutte le combinazioni. I modelli con i regressori `holiday`

Table 4: Risultati del modello SARIMA(2,0,2)(1,1,1)$_{24}$ con regressori.

Regressori	Log-likelihood	MAE train	MAE Validation
holiday	**-188713.5**	**67206.0**	**328893.0**
weekend	-188440.0	66817.3	333931.3
lockdown	-188346.3	65321.3	333732.0
holiday+weekend	-188797.9	68581.7	329022.7
holiday+lockdown	-188713.5	67206.0	328893.0
weekend+lockdown	-188440.0	66817.3	333931.3
holiday+weekend+lockdown	-188797.9	68581.7	329022.7

e `holiday+lockdown` hanno le stesse performance sul MAE, si sceglie quindi il primo per la sua semplicità. In definitiva il modello migliore risulta essere SARIMA(2,0,2)(1,1,1)$_{24}$ con 5 armoniche per le stagionalità e l'uso del regressore `holiday`, nelle figure 11 e 12 si mostrano le sue predizioni del modello. In figura 13 si mostrano le predizioni del modello su tutta la serie, si vede che il modello riesce a cogliere meglio le stagionalità sul validation.

4 Modello UCM

Per la scelta del modello UCM si è eseguito un approccio simile al caso dell'ARIMA: tramite una grid search si individua la componente level-trend che ha il miglior valore per il MAE e successivamente si aggiungono stagionalità e ciclo. Per i modelli testati si è usata la stagionalità giornaliera *dummy*

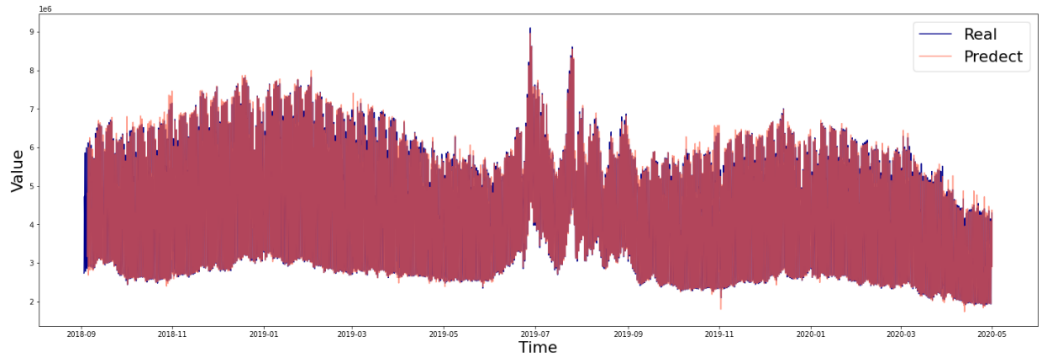

Figure 11: SARIMA: plot delle predizioni sul train.

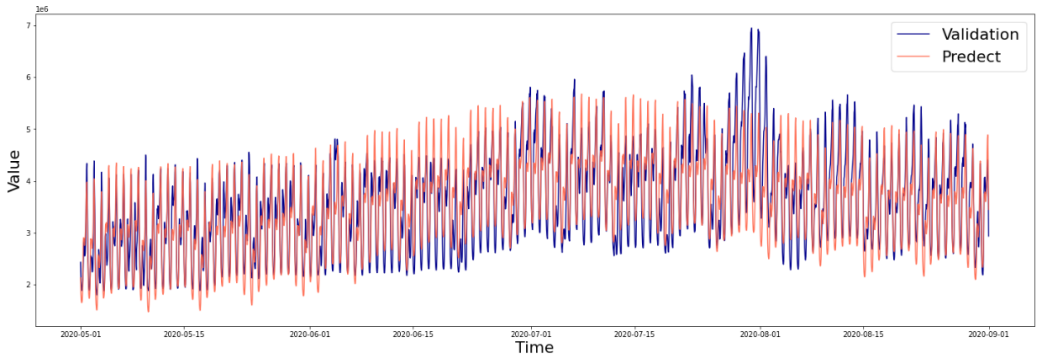

Figure 12: SARIMA: plot delle predizioni sul validation.

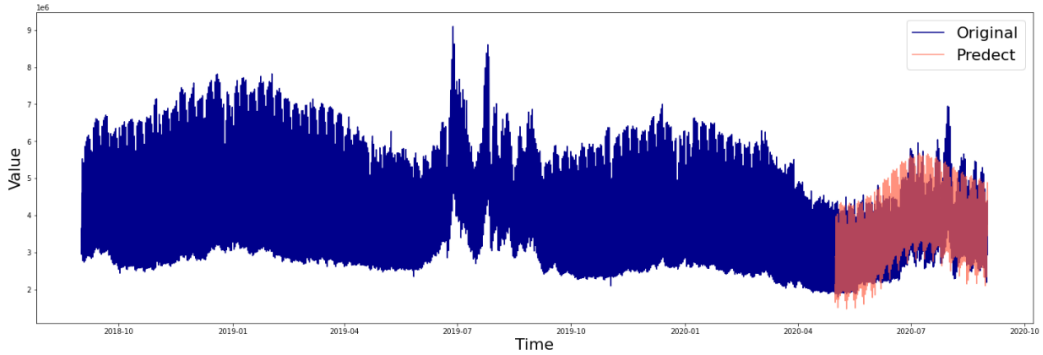

Figure 13: SARIMA: plot delle predizioni sull'intera serie.

e una stagionalità settimanale con 15 armoniche e una annuale con 15 armoniche. Durante l'analisi è emerso che la stagionalità annuale abbassa, in alcuni casi anche sensibilmente, le prestazioni dei modelli e quindi si è scelto di non inserirla. I risultati dei modelli in questa prima fase sono mostrati

nella tabella 5.

Table 5: Risultati dei modelli UCM per il trend.

Modello	MAE Train	MAE Validation
No trend	1023536.3	3645233.1
Deterministic constant	1023575.9	3645172.2
Local level	**103983.0**	**752009.1**
Random walk	**102648.6**	**752061.9**
Local linear det. trend	**103958.2**	**742714.1**
Random walk with drift	**102623.3**	**743458.6**
Local linear trend	105028.5	13418369.2
Smooth trend	105031.0	13418507.5
Random trend	103951.7	13096502.3

I modelli migliori per il MAE sul train e sul validation sono: Local level, Random walk, Local linear deterministic e Random walk with drift. Si continua l'analisi concentrandosi solo su questi trends inserendo la componente ciclo e i vari regressori.
L'aggiunta del ciclo non produce un aumento della precisione del modello e quindi si decide di non considerarlo, mentre per i regressori si provano tutte le combinazioni. Per questioni di leggibilità nella tabella 6, per ogni trend, si mostra la combinazione di regressori che ottiene i risultati migliori.

Table 6: Risultati dei modelli UCM per i regressori.

Modello	regressori	MAE Train	MAE Validation
Local level	holiday	103970.3	749240.0
Random walk	holiday	102636.1	749291.2
Local linear det. trend	**holiday**	**103945.6**	**739937.5**
Random walk with drift	holiday	102610.9	740681.1

Il miglior modello, intermini di MAE sul validation, ha il *local linear deterministic trend*, con stagionalità giornaliera e settimanale e con i regressori per le vacanze.

In figura 16 si mostrano le predizioni del modello su tutta la serie, nelle figure 14 e 15 ci si concentra rispettivamente sul train set e sul validation set. Il modello non riesce a cogliere bene i picchi che si verificano nell'ultima parte del validation set.

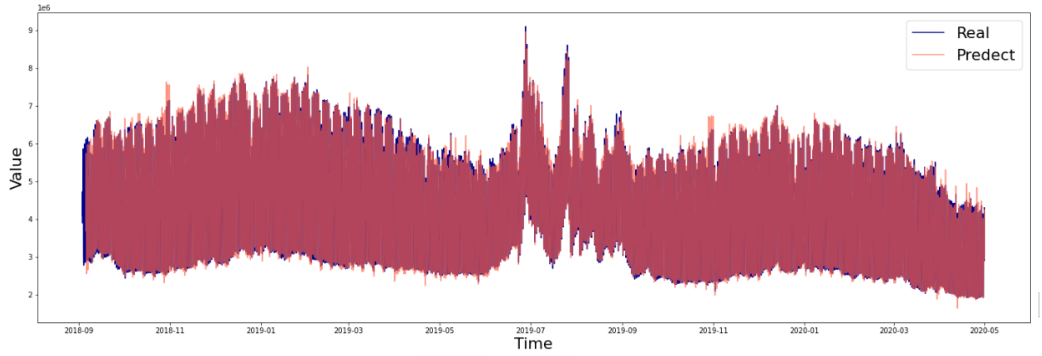

Figure 14: UCM: plot delle predizioni sul train.

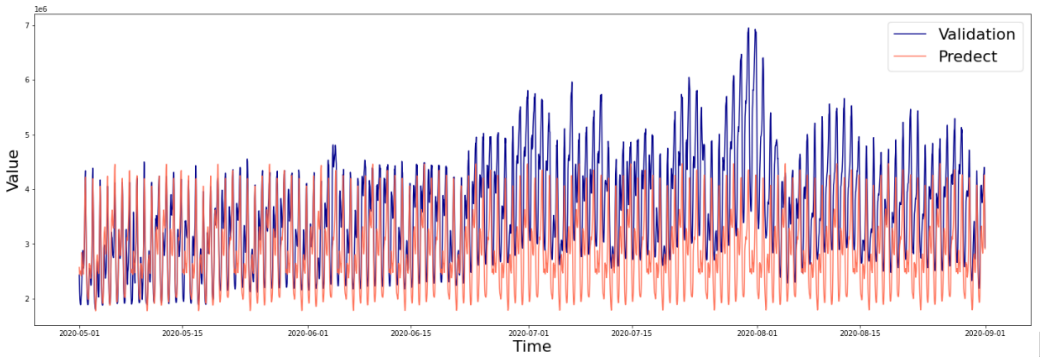

Figure 15: UCM: plot delle predizioni sul validation.

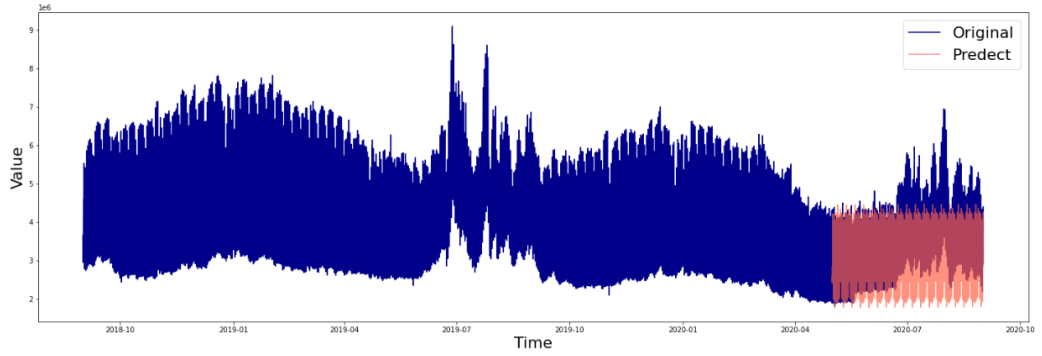

Figure 16: UCM: plot delle predizioni sull'intera serie.

5 Modello ML

Per i modelli di Machine Learning si sono testati due differenti reti neurali: una LSTM e una GRU. Una LSTM ha connessioni di feedback. Non solo può elaborare singoli punti dati ma anche intere sequenze di dati (come voce

o video) mentre una GRU è una versione più semplice di una LSTM con meno gate. Dopo una prima fase di preparazione dei dati secondo il formato adatto all'uso delle reti. Per rendere più equo il confronto, per entrambi i tipi di RNN si provano le stesse architetture:

- 1 layer da 128 neuroni con dropout a 0.33.
- 1 layer da 64 neuroni, 1 layer da 32 neuroni con dropout

Dopo ogni livello c'è un dropout impostato a 0.33 e il livello finale di ogni architettura è uno strato denso per l'output. Il training è stato effettuato su 5 epoche.

Table 7: Risultati dei modelli ML.

Modello	MAE Train	MAE Validation
GRU 1 layer	2979502.0	970576.5
LSTM 1 layer	3029389.6	928614.0
GRU 2 layer	**3149814.3**	**786970.3**
LSTM 2 layer	2964050.3	1071215.0

Il modello migliore risulta essere la rete GRU con 2 layers, anche se il MAE sul train è il più alto, ha l'errore sul validation sensibilmente più baso degli altri modelli.

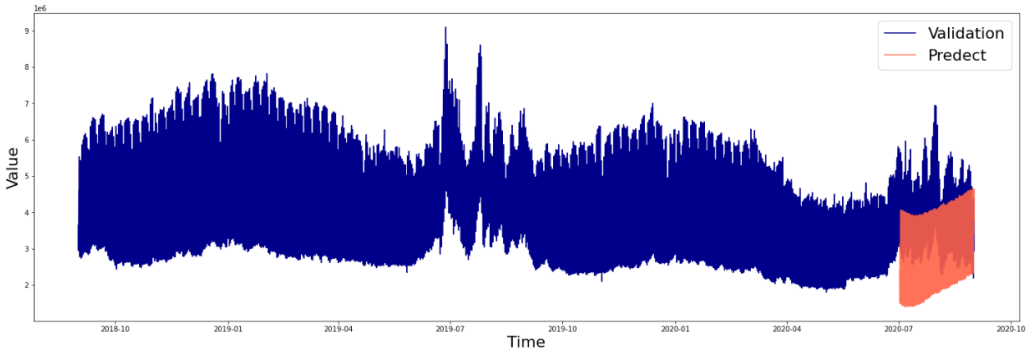

Figure 17: GRU: plot delle predizioni sull'intera serie.

6 Previsioni

In questa sezione si calcolano le previsioni dei tre migliori modelli sul periodo richiesto dal 01/09/2020 al 31/10/2020: modello SARIMA$(2,0,2)(1,1,1)_{24}$ con

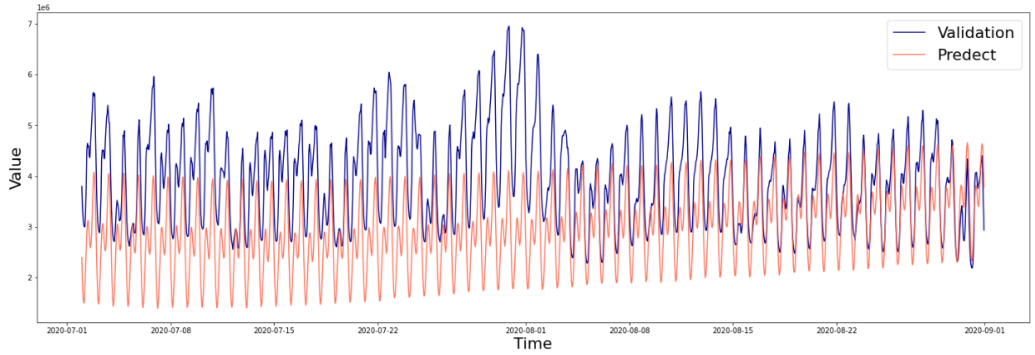

Figure 18: GRU: plot delle predizioni sul validation.

5 armoniche per le stagionalità e uso di regressori `holiday`, modello UCM con *Local linear deterministic trend* con stagionalità giornaliera e settimanale e regressore `holiday` e rete GRU con due layers. Nella tabella 8 si mostrano i valori del MAE sul train e sul validation.

Table 8: Risultati finali dei modelli.

Modello	MAE Train	MAE Validation
ARIMA	67206.0	328893.0
UCM	103945.6	739937.5
GRU	3149814.3	786970.3

Nelle figure seguenti si mostrano graficamente le predizioni al buio dei tre modelli.

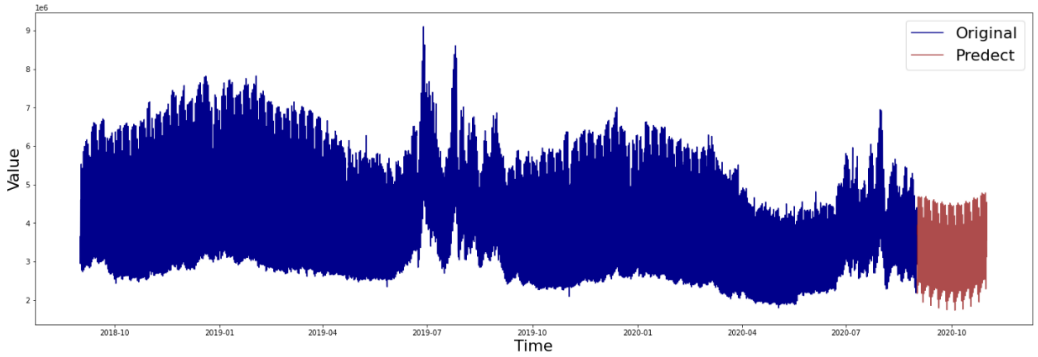

Figure 19: ARIMA: predizioni al buio.

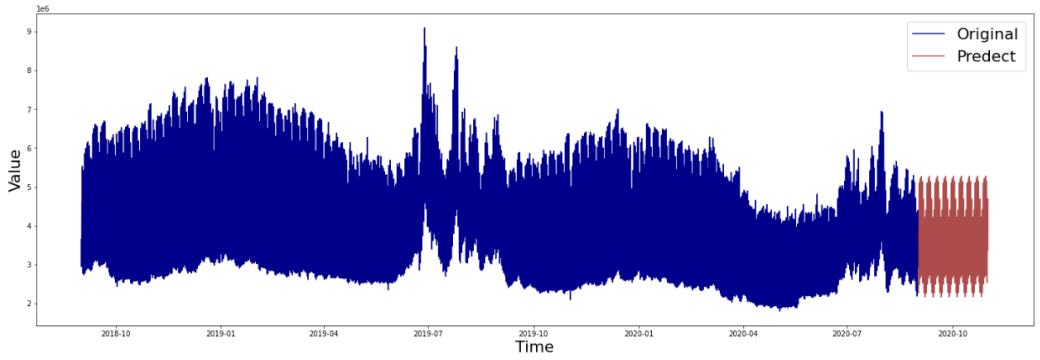

Figure 20: UCM: predizioni al buio.

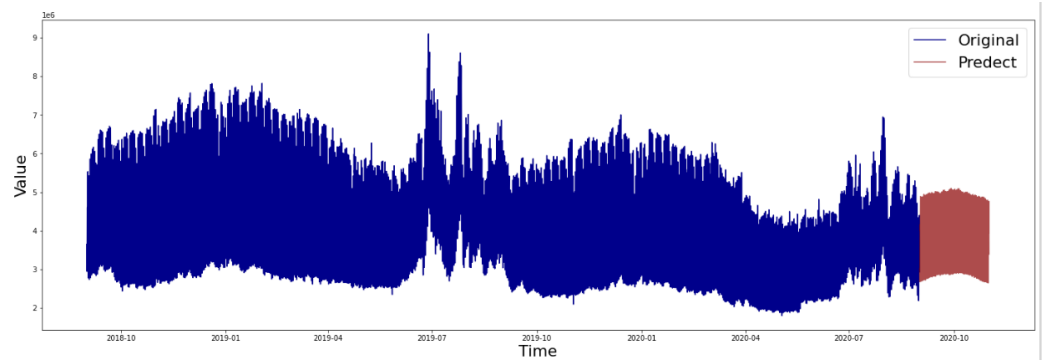

Figure 21: GRU: predizioni al buio.

7 Conclusioni

In questo report si analizzano le più usate tecniche per lo studio di serie temporali, si definiscono diversi modelli, sia lineari che appartenenti al mondo del machine learning, per individuare lo strumento migliore con cui fare previsioni sul periodo richiesto. Il miglior modello è risultato essere il modello ARIMA.
Una questione da considerare è che il contesto dei dati non è dato, se fosse stato noto si sarebbero potuti individuare altri regressori per migliorare le performance dei modelli lineari. Per quanto riguarda i modelli di Machine Learning, per questioni di tempo si sono limitate le epoche di addestramento e i livelli nascosti della rete, al cambio di questi parametri si sarebbero potuti ottenere risultati migliori.