

Progetto di Streaming Data Management and Time Series Analysis

Abstract

Il progetto ha come obiettivo la definizione, lo sviluppo e la validazione di un sistema predittivo per dati di tipo time series. Per questo sono stati implementati diversi tipi di modelli quali:

- **ARIMA**
- **UCM** (modelli a componenti non osservabili)
- **PROPHET**
- **LSTM**
- **GRU**

cercando di modellare i diversi tipi di stagionalità contenute nei dati.

1 PREPROCESSING DEI DATI

Il dataset a disposizione contiene dati orari che vanno dal 2018-09-01 al 2020-08-31. Una particolarità del dataset è che dovrebbe essere affetto dalla crisi Covid-19. Dal preprocessing dei dati, risultano mancanti i valori associati alle datetime 2019-03-31 03:00:00 e 2020-03-29 03:00:00. La causa di questo vuoto è che quella data corrisponde al passaggio ad ora solare-legale. Decido quindi di sostituire i due valori mancanti con il valore all'ora precedente. Inoltre, mancano anche tutti valori del 2020-05-31. Quella data risulta associata alla domenica di Pentecoste. Per sostituire i valori mancanti riferiti a quella data decido di calcolare la media dei valori di una settimana prima e dopo, sempre alla stessa ora.

2 VISUALIZZAZIONI INIZIALI DEI DATI

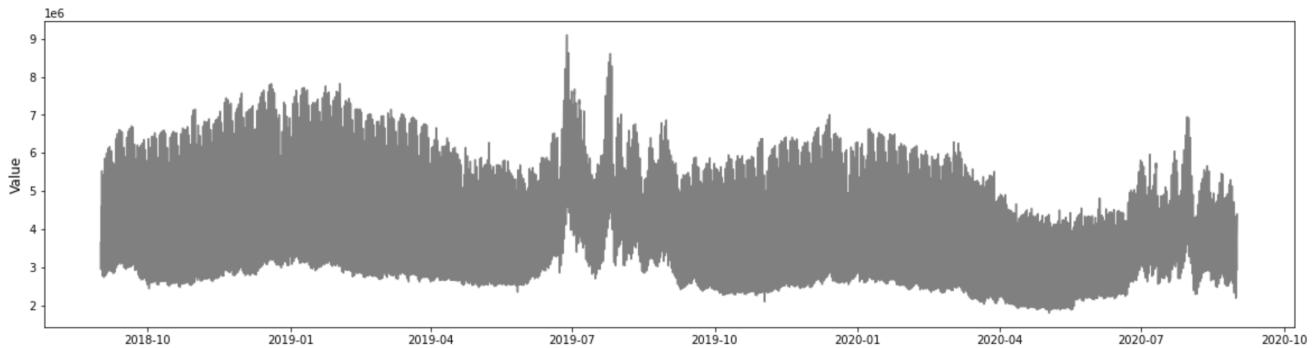


Figure 1: Visualizzazione dell'intera serie storica.

Dalla prima visualizzazione complessiva dell'intera serie storica Figura 1, si evince che sembrano presenti due picchi anomali verso fine giugno e a luglio.

2.1 Andamenti annuale e mensile

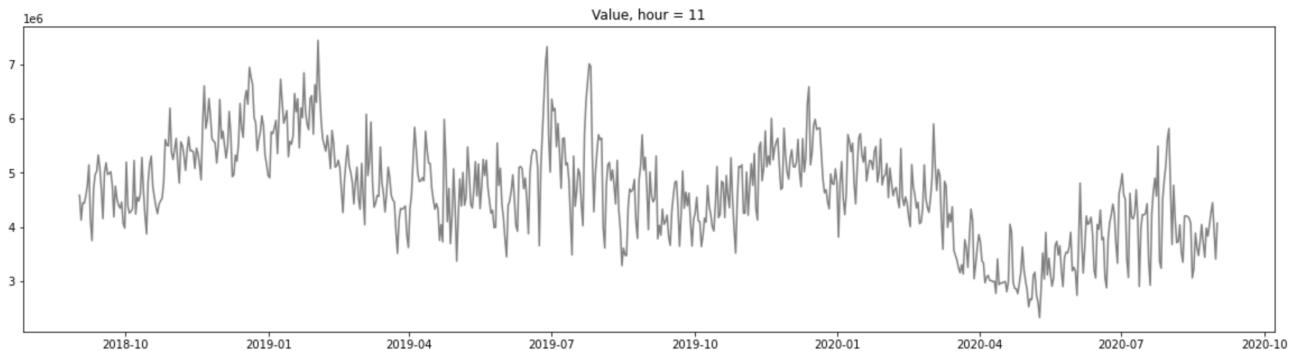


Figure 2: Visualizzazione dell’intera serie storica all’ora 11.

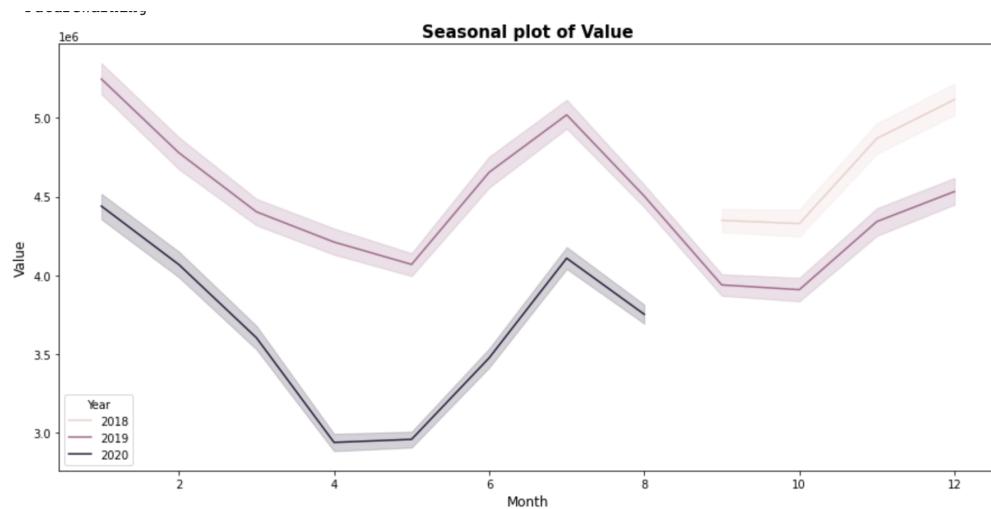


Figure 3: Visualizzazione della stagionalità annuale.

La visualizzazione 2 rappresenta l’intera serie storica all’ora 11. Questa permette di dare una rappresentazione più chiara dei dati mostrando le eventuali anomalie presenti nel dataset. Si vede dalla Figura 3 che esiste un andamento stagionale annuale con dei picchi nei mesi di dicembre-gennaio e di giugno-luglio. Si nota anche un trend decrescente al variare degli anni. Inoltre, nei mesi di marzo-aprile-maggio 2020 si notano chiaramente dei trend decrescenti dovuti molto probabilmente alla pandemia COVID-19.

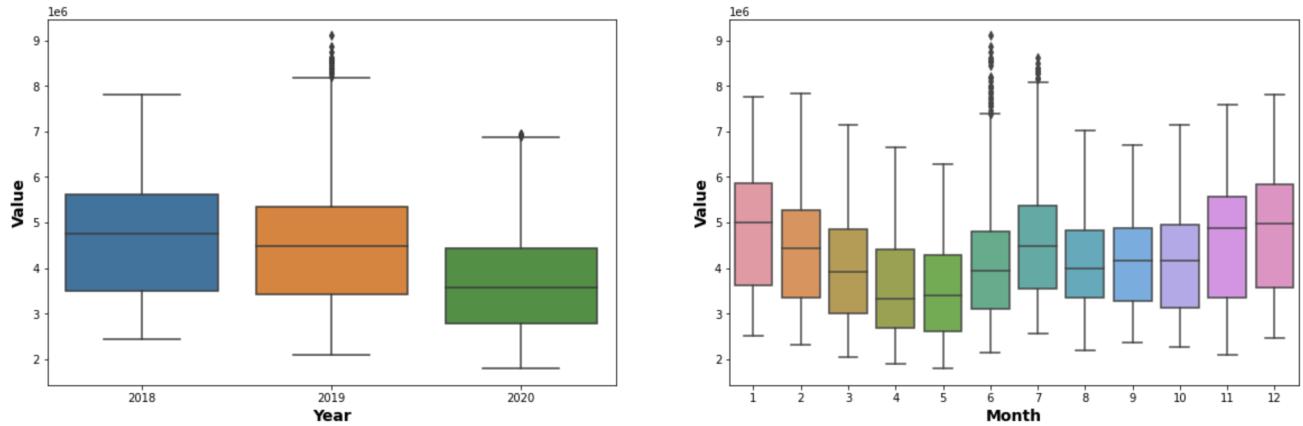


Figure 4: Box plot dei dati raggruppati per anni (sinistra) e per mese (destra).

Il box plot (Figura 4) di sinistra conferma che è presente un andamento decrescente durante gli anni e sembrano presenti numerosi outliers nell'anno 2019. Più in particolare, il box plot di destra mostra che sono effettivamente presenti dei valori anomali nei mesi di giugno e luglio. Sembra anche esserci un andamento intra annuo.

2.2 Andamenti settimanali e giornalieri

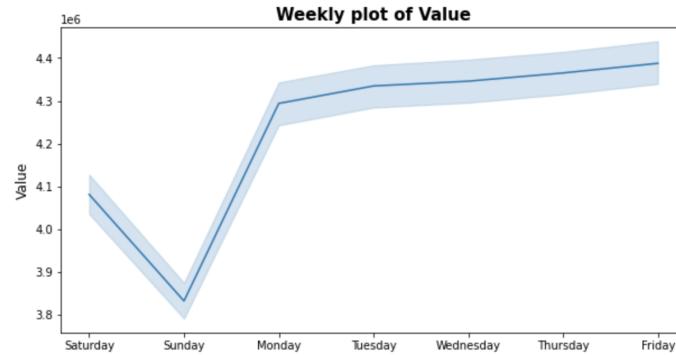


Figure 5: Visualizzazione dell'andamento settimanale.

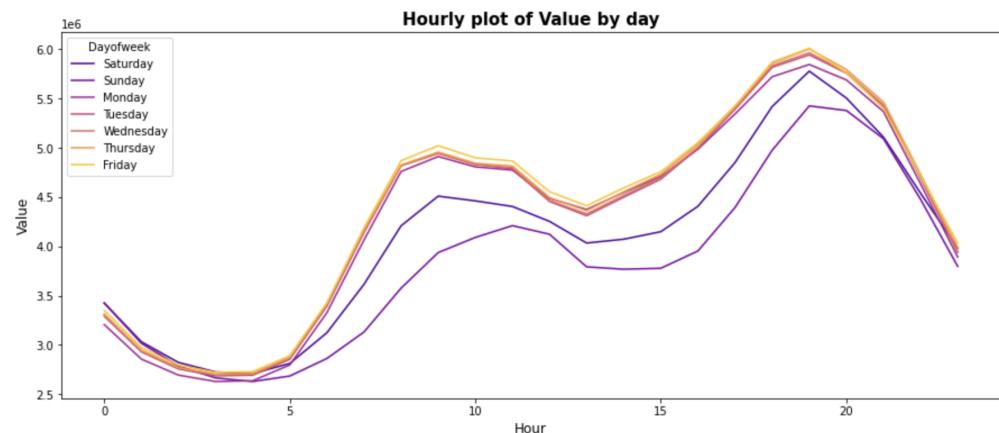


Figure 6: Visualizzazione dell'andamento giornaliero in funzione del giorno della settimana.

La Figura 5 mostra che è presente un andamento infra settimanale: i valori registrati del sabato e della domenica sono decisamente più bassi di quelli del resto della settimana. Andando a guardare più nello specifico di ciascun giorno della settimana 6, si nota un andamento infra giornaliero con picchi tra le 9 e le 11 e tra le 15 e le 21. Gli andamenti rimangono molto simili per i giorni lavorativi della settimana. Invece per il sabato e la domenica l'andamento è simile ma il trend è più basso.

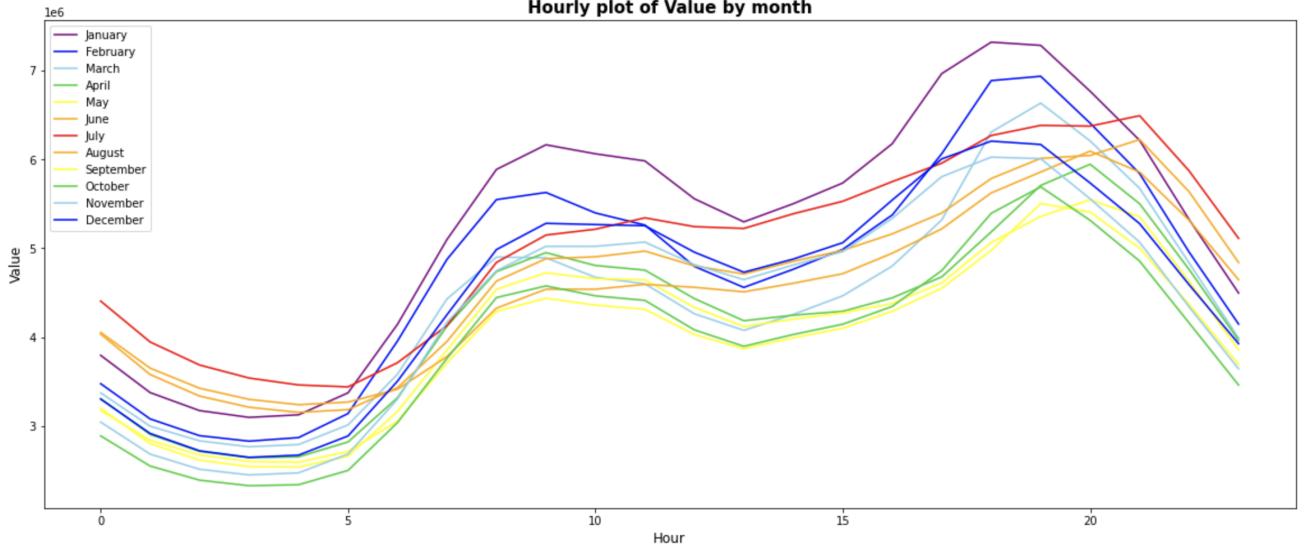


Figure 7: Visualizzazione dell'andamento giornaliero in funzione del mese.

Dalla Figura 6, si evince che la stagionalità giornaliera cambia al variare dei mesi dell'anno. Nei mesi invernali si vedono con più chiarezza due picchi ben distinti: uno alla mattina dalle 9 alle 12 e uno al pomeriggio dalle 15 alle 20. Nei mesi estivi invece si nota un unico picco giornaliero dalle ore che vanno dalle 10 alle 20. Sembra che i dati siano influenzati da luce solare o da temperatura.

3 DIVISIONE TRAINING VALIDATION

La divisione training validation è stata effettuata in questo modo:

- Observations Train: dal 2018-09-01 al 2020-04-30 con 14592 dati ossia 83.17%.
- Observations Validation: dal 2020-05-01 al 2020-08-31 con 2952 dati ossia 16.83%.

In questo modo il training test contiene i dati riferiti al primo lockdown dovuto alla pandemia COVID-19.

4 ARIMA

Per l'applicazione del modello ARIMA è necessario che la serie storica sia stazionaria, ovvero che le sue proprietà temporali non varino nel tempo. La trasformazione logaritmica può aiutare a stabilizzare la serie temporale in varianza. La differenza aiuta invece a stabilizzare la serie in media.

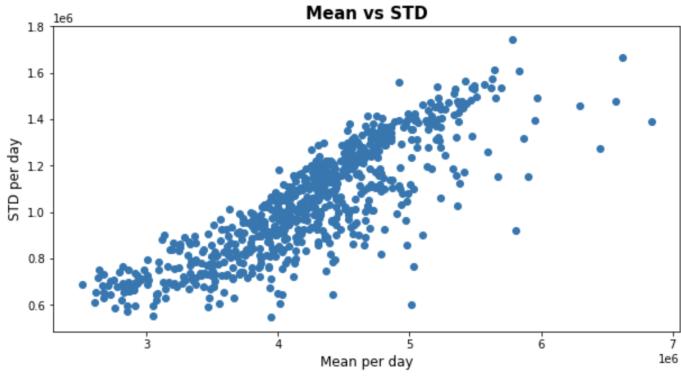


Figure 8: Visualizzazione della relazione tra le medie e le deviazioni standard giornaliere.

Da questo grafico si evince che la relazione tra media e deviazione standard per ciascun giorno è approssimativamente lineare. Le trasformazioni più comuni per rendere la serie stazionaria sono quelle di Box and COX di cui fa parte la trasformazione logaritmica. Quindi la trasformazione logaritmica potrebbe essere in grado di rendere la serie storica stazionaria in varianza. Un modo per capire se questa migliora il modello è quello di applicarla e confrontarla con il modello senza nessuna trasformazione. Vari test permettono di verificare se una serie storica è stazionaria o meno.

Results of Dickey-Fuller Test:		Results of KPSS Test:	
Test Statistic	-5.810651e+00	Test Statistic	9.070323
p-value	4.409576e-07	p-value	0.010000
Lags Used	4.200000e+01	Lags Used	42.000000
Number of Observations Used	1.454900e+04	Critical Value (10%)	0.347000
Critical Value (1%)	-3.430800e+00	Critical Value (5%)	0.463000
Critical Value (5%)	-2.861739e+00	Critical Value (2.5%)	0.574000
Critical Value (10%)	-2.566876e+00	Critical Value (1%)	0.739000
dtype: float64		dtype: float64	

Figure 9: Risultato dei tests Dickey Fuller (ADF) e KPSS.

Il Dickey-Fuller ha come ipotesi nulla che almeno una radice unitaria sia presente nel modello autoregressivo, il che implica che la serie non è stazionaria. È possibile vedere come il p-value risulti piccolo (significativa del 0.99), permettendo di rifiutare l'ipotesi nulla. Quindi, la serie è stazionaria secondo il test ADF.

Il test Kwiatkowski–Phillips–Schmidt–Shin (KPSS) ha come ipotesi nulla che la serie storica osservata sia stazionaria attorno ad un trend deterministico rispetto all'alternativa di una radice unitaria. È possibile in questo caso rifiutare l'ipotesi nulla che la serie storica sia stazionaria. Quindi la serie non è stazionaria secondo il test KPSS. Questi risultati mostrano che la serie storica è stazionaria per differenza. Si può quindi pensare di applicare una differenza ai dati.

Per capire l'ordine di differenziazione da applicare ai dati è necessario guardare il grafico *ACF* (Figura 10) che rappresenta i coefficienti di correlazione di una serie storica con i suoi ritardi. Il grafico dei dati originali mostra una stagionalità ogni 24 ore. Per questo motivo ho deciso di applicare una differenza sui dati di 24.

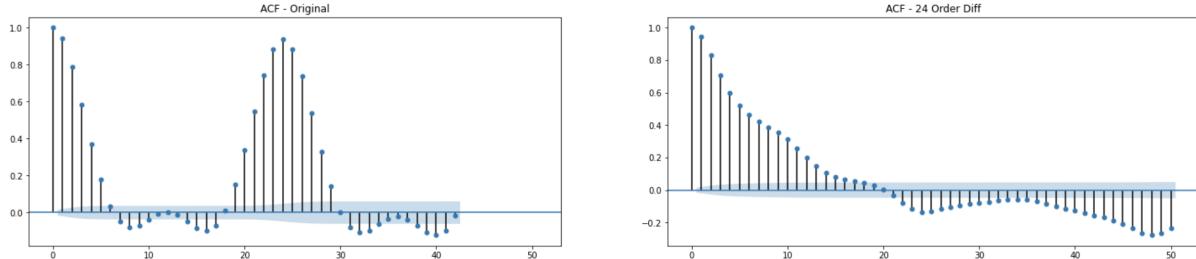


Figure 10: Grafico di autocorrelazione (ACF) della serie storica originale e differenziata di ordine 24.

```
Results of Dickey-Fuller Test:
Test Statistic           -22.449561
p-value                  0.000000
Lags Used                42.000000
Number of Observations Used 14525.000000
Critical Value (1%)      -3.430800
Critical Value (5%)      -2.861739
Critical Value (10%)     -2.566876
dtype: float64
```

```
Results of KPSS Test:
Test Statistic          0.018266
p-value                  0.100000
Lags Used                42.000000
Critical Value (10%)    0.347000
Critical Value (5%)     0.463000
Critical Value (2.5%)   0.574000
Critical Value (1%)     0.739000
dtype: float64
```

Figure 11: Risultato dei tests Dickey Fuller (ADF) e KPSS dopo avere applicato una differenza Δ^{24} .

Dai risultati della Figura 11, si nota che in questo caso, l'ipotesi nulla è rifiutata per il Dickey-Fuller test ed è accettata per il KPPS test. Quindi per entrambi i test, la serie temporale $\Delta^{24}Z_t$ è stazionaria.

Il grafico *PACF* mostra i coefficienti di autocorrelazione parziale tra una serie storica e i suoi ritardi. Un'autocorrelazione parziale è la quantità di correlazione tra una variabile e un ritardo che non risulta spiegata da correlazioni di tutti i ritardi di ordine inferiore. Il grafico *PACF - 24 Order Diff* mostra che potrebbe avere senso un ordine $p = 1$, $p = 2$ o anche addirittura $p = 3$. Per quanto riguarda la parte stagionale non si riesce bene a capire quanti picchi si potrebbero considerare significativi. Quindi inizio a considerare $P = 1$ e $Q = 1$.

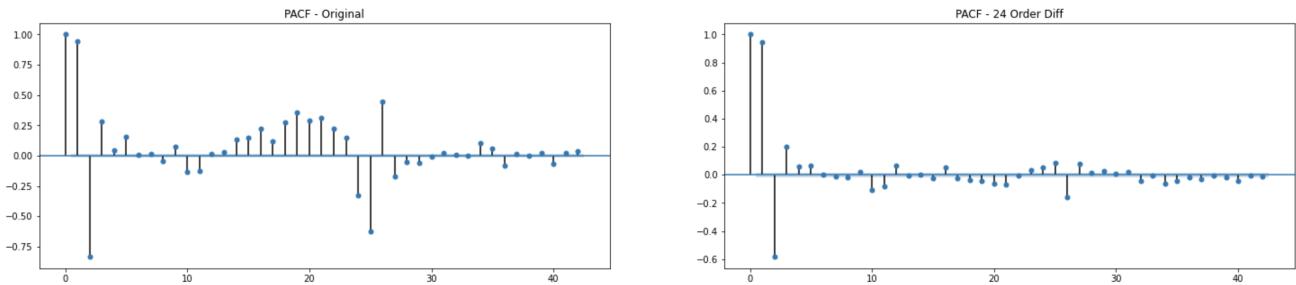


Figure 12: Grafico di autocorrelazione parziale (PACF) della serie storica originale e differenziata di ordine 24.

5 Modello SARIMA

Per trovare il modello migliore, applico il metodo grid search dei parametri non stagionali p e q sia usando i dati originali che quelli log-trasformati.

La trasformazione logaritmica migliora il MAE del training ma non del validation. Una prima osservazione importante che si può fare è che la maggior parte dei modelli sembrano overfittare i dati. Nonostante questo, il modello che sembra funzionare meglio in termini di $AIC = 376946.0$ è il SARIMA(3,0,3)(1,1,1)₂₄ con un MAE del training di 64858.7 e del validation di 709131.6.

Modelli	Senza log trasformazione			Con log trasformazione		
	MAE Train	MAE Validation	AIC	MAE Train	MAE Validation	AIC
SARIMA(0,0,0)(1,1,1) ₂₄	276458.0	716203.0	417218.0	274785.7	726388.8	-29681.0
SARIMA(0,0,1)(1,1,1) ₂₄	157702.5	710287.1	406377.0	152181.4	726320.8	-46611.0
SARIMA(0,0,2)(1,1,1) ₂₄	110341.7	707535.0	400669.0	104576.2	726938.0	-57133.0
SARIMA(0,0,3)(1,1,1) ₂₄	87141.9	707871.0	396015.0	82140.1	727884.9	-63102.0
SARIMA(1,0,0)(1,1,1) ₂₄	84912.7	718818.9	383424.0	82676.4	726202.6	-63294.0
SARIMA(1,0,1)(1,1,1) ₂₄	69836.8	711543.9	378389.0	66682.4	727167.8	-68421.0
SARIMA(1,0,2)(1,1,1) ₂₄	66466.6	710414.0	377501.0	63291.2	727247.9	-69288.0
SARIMA(1,0,3)(1,1,1) ₂₄	65730.6	709639.1	377347.0	62452.4	726817.8	-69391.0
SARIMA(2,0,0)(1,1,1) ₂₄	67167.6	713421.9	377707.0	64429.8	726196.4	-68910.0
SARIMA(2,0,1)(1,1,1) ₂₄	65770.2	711881.0	377324.0	62807.1	726737.0	-69326.0
SARIMA(2,0,2)(1,1,1) ₂₄	65584.2	711547.3	377291.0	62564.1	726840.6	-69379.0
SARIMA(2,0,3)(1,1,1) ₂₄	65545.7	711244.0	377267.0	62770.6	726879.8	-69333.0
SARIMA(3,0,0)(1,1,1) ₂₄	65279.9	712395.3	377141.0	62409.3	726862.3	-69419.0
SARIMA(3,0,1)(1,1,1) ₂₄	64907.0	708234.3	377009.0	63270.5	727131.3	-69204.0
SARIMA(3,0,2)(1,1,1) ₂₄	64872.9	707677.4	376990.0	61865.9	724401.9	-69544.0
SARIMA(3,0,3)(1,1,1) ₂₄	64858.7	709131.6	376946.0	61816.5	725938.2	-69551.0

Table 1: Performance del modello SARIMA al variare dei parametri non stagionali p e q.

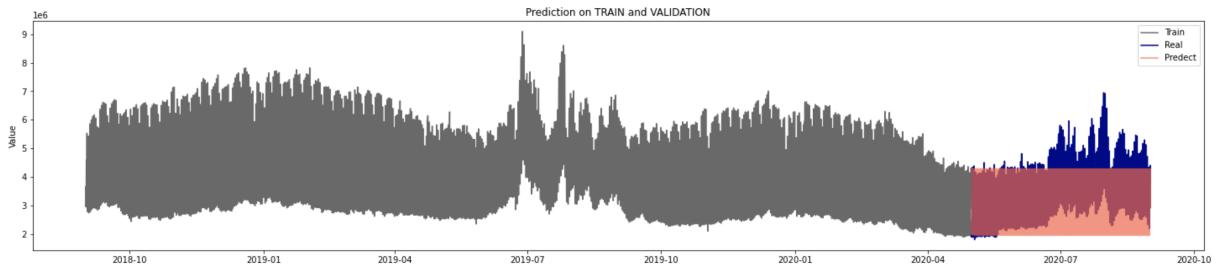


Figure 13: Previsioni con modello SARIMA(3,0,3)(1,1,1)₂₄ sull'intero validation set.

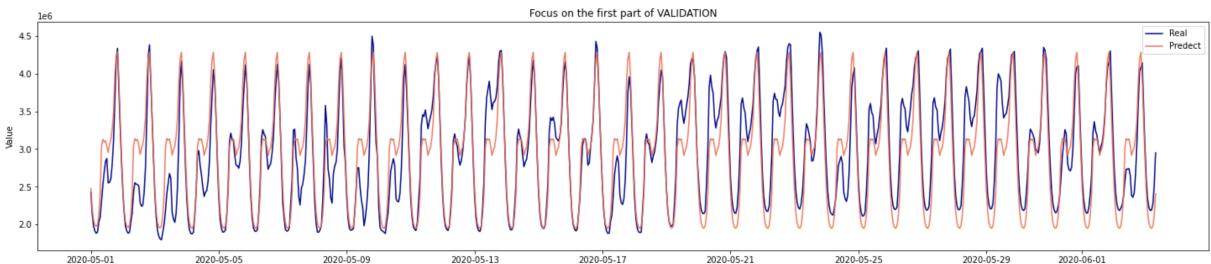


Figure 14: Focus delle previsioni con modello SARIMA(3,0,3)(1,1,1)₂₄ sulla prima parte del validation set.

Dai vari plot in Figura 13 e in Figura 14 si evince che il modello ha colto la stagionalità infragiornaliera ma non riesce a cogliere quella settimanale e infrannuale.

5.0.1 Modello SARIMAX con regressori di Fourier

Per riuscire a modellare gli altri tipi di stagionalità presenti nei dati è necessario aggiungere altri tipi di regressori quali le serie di Fourier. Per quanto riguarda la stagionalità settimanale inserisco una serie di Fourier con periodo

$24*7 = 168$. Per la stagionalità annuale inserisco una serie di Fourier con periodo $24*365,25 = 8766$. Anche in questo caso si applica un approccio grid search, facendo variare le armoniche settimanali da 6 a 12 e quelle annuali da 5 a 10 ogni 5.

Arm settimanali	Arm annuali	MAE Train	MAE Validation	AIC
6	5	64300.7	325296.3	376258.0
6	10	64287.3	379698.9	376271.0
12	5	545056.1	338712.5	1160898398.0
12	10	579614.4	386710.4	1308199833.0

Table 2: Performance del modello SARIMAX(3,0,3)(1,1,1)₂₄ al variare del numero di armoniche settimanali e annuali.

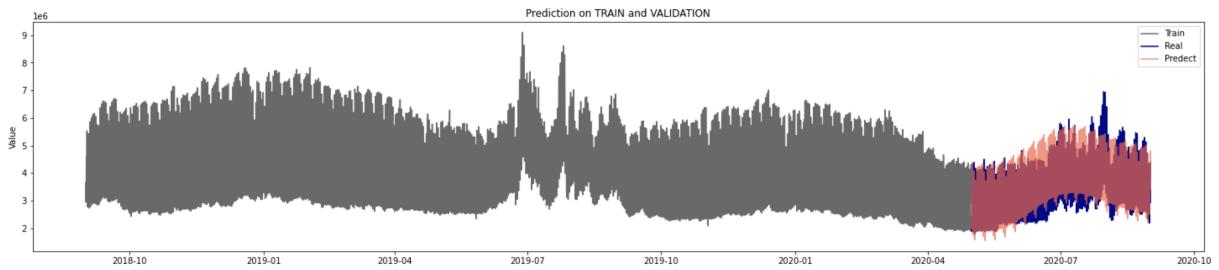


Figure 15: Previsioni con modello SARIMAX(3,0,3)(1,1,1)₂₄ con 6 armoniche settimanali e 5 armoniche annuali sull'intero validation set.

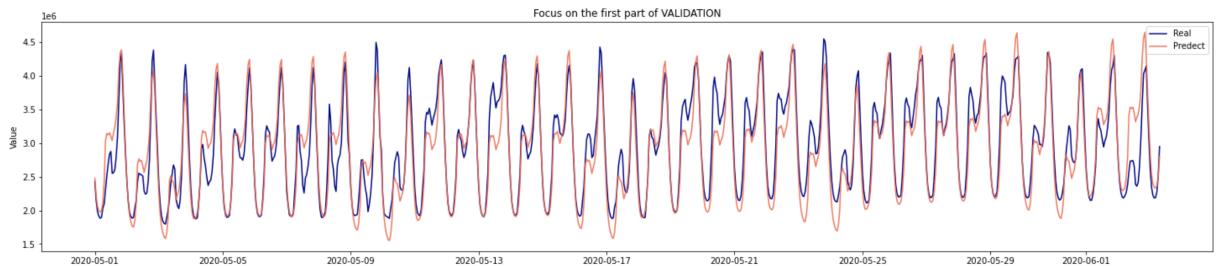


Figure 16: Focus delle previsioni con modello SARIMA(3,0,3)(1,1,1)₂₄ con 6 armoniche settimanali e 5 armoniche annuali sulla prima parte del validation set.

Dalle Figure 33 e 34 si nota che il modello è riuscito a cogliere le stagionalità settimanale e annuali. Effettuando le previsioni con questo modello, si ottengono: MAE train: 64300,7 e MAE validation: 325296,3.

SARIMAX Results						
Dep. Variable:		Value	No. Observations:	14592		
Model:	SARIMAX(3, 0, 3)x(1, 1, [1], 24)	Log Likelihood	-188098.078			
Date:	Sat, 04 Sep 2021	AIC	376258.156			
Time:	10:26:16	BIC	376493.280			
Sample:	09-01-2018 - 04-30-2020	HQIC	376336.286			
Covariance Type:	approx					
	coef	std err	z	P> z	[0.025	0.975]
sin(1,168)	-1.903e+05	1.72e+04	-11.068	0.000	-2.24e+05	-1.57e+05
cos(1,168)	-6.947e+04	1.72e+04	-4.033	0.000	-1.03e+05	-3.57e+04
sin(2,168)	-1.103e+05	1.07e+04	-10.268	0.000	-1.31e+05	-8.93e+04
cos(2,168)	9.909e+04	1.08e+04	9.208	0.000	7.8e+04	1.2e+05
sin(3,168)	3.19e+04	7803.404	4.088	0.000	1.66e+04	4.72e+04
cos(3,168)	5.439e+04	7806.021	6.968	0.000	3.91e+04	6.97e+04
sin(4,168)	1.623e+04	6699.805	2.422	0.015	3097.719	2.94e+04
cos(4,168)	-5.318e+04	6700.875	-7.936	0.000	-6.63e+04	-4e+04
sin(5,168)	-9.11e+04	6868.502	-13.263	0.000	-1.05e+05	-7.76e+04
cos(5,168)	-3.337e+04	6874.943	-4.854	0.000	-4.68e+04	-1.99e+04
sin(6,168)	-7.268e+04	8554.034	-8.496	0.000	-8.94e+04	-5.59e+04
cos(6,168)	6.954e+04	8555.105	8.128	0.000	5.28e+04	8.63e+04
sin(1,8766)	2.829e+05	2.010	1.41e+05	0.000	2.83e+05	2.83e+05
cos(1,8766)	1.981e+04	94.645	209.263	0.000	1.96e+04	2e+04
sin(2,8766)	-4.905e+05	3.648	-1.34e+05	0.000	-4.91e+05	-4.91e+05
cos(2,8766)	-6.345e+04	95.289	-665.879	0.000	-6.36e+04	-6.33e+04
sin(3,8766)	-1.102e+05	5.952	-1.85e+04	0.000	-1.1e+05	-1.1e+05
cos(3,8766)	-1.397e+05	97.343	-1434.766	0.000	-1.4e+05	-1.39e+05
sin(4,8766)	4.686e+04	7.645	6129.201	0.000	4.68e+04	4.69e+04
cos(4,8766)	-3.377e+04	98.338	-343.402	0.000	-3.4e+04	-3.36e+04
sin(5,8766)	8.38e+04	6.649	1.26e+04	0.000	8.38e+04	8.38e+04
cos(5,8766)	5.426e+04	100.785	538.375	0.000	5.41e+04	5.45e+04
ar.L1	2.1475	0.025	84.738	0.000	2.098	2.197
ar.L2	-1.6854	0.040	-41.649	0.000	-1.765	-1.606
ar.L3	0.5076	0.021	23.845	0.000	0.466	0.549
ma.L1	-0.6073	0.027	-22.827	0.000	-0.659	-0.555
ma.L2	-0.0314	0.024	-1.283	0.199	-0.079	0.017
ma.L3	0.0435	0.021	2.072	0.038	0.002	0.085
ar.S.L24	0.3105	0.011	27.070	0.000	0.288	0.333
ma.S.L24	-0.7151	0.007	-95.414	0.000	-0.730	-0.700
sigma2	1.173e+10	0.048	2.42e+11	0.000	1.17e+10	1.17e+10
Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	61197.96			
Prob(Q):	0.90	Prob(JB):	0.00			
Heteroskedasticity (H):	0.70	Skew:	-0.05			
Prob(H) (two-sided):	0.00	Kurtosis:	13.05			

Figure 17: Coefficienti del modello SARIMAX.

Dalla Figura 25, tutte le variabili risultano significative tranne il coefficiente MA con q = 2.

6 Modello UCM

La seconda tipologia di modelli utilizzati riguarda i modelli lineari Unobserved Components Model (UCM). UCM decomponete una serie storica univariata in diverse componenti trend, stagionali, cicliche, e irregolari. Ciascuna componente del modello cattura importanti caratteristiche della serie dinamica. UCM può modellare il trend in due modi: il primo consiste nell'usare il modello random walk che implica che la tendenza rimane approssimativamente costante nel periodo di tempo della serie, e il secondo consiste nell'usare un trend localmente lineare con una pendenza verso l'alto o verso il basso. Per determinare il migliore tipo di level-trend ho usato un approccio grid search riportato nella Tabella 3

La stagionalità giornaliera è rappresentata tramite variabili dummy, quella settimanale tramite serie di Fourier con 6 armoniche. Il migliore modello in termini di MAE sembra essere il random walk with drift con un MAE train = 136033.5 e un MAE validation = 775391.1.

Level-trend	MAE Train	MAE Validation
rwalk	136033.6	782625.4
dconstant	600530.6	3651114.1
ntrend	600509.3	3651178.7
llevel	141739.8	784019.9
lldtrend	141742.8	775439.0
rwdrift	136033.5	775391.1
lltrend	145298.4	27220429.0
strend	145302.6	27216492.4
rtrend	139492.4	30142522.5

Table 3: Performance del modello UCM al variare della tipologia di Level-trend.

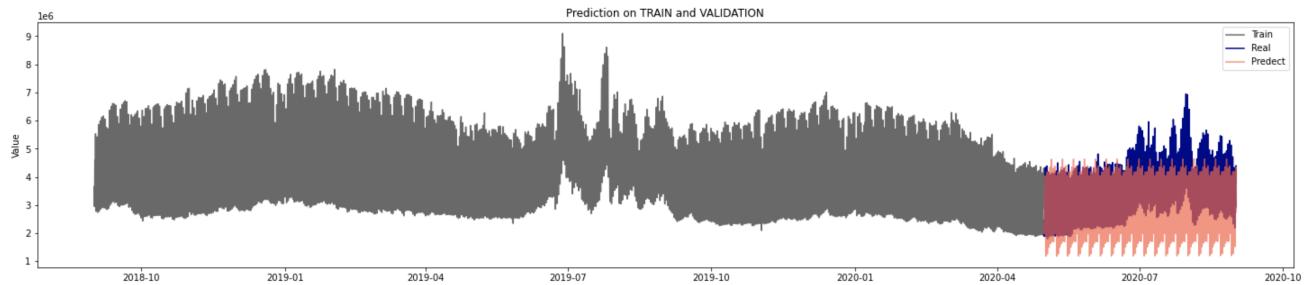


Figure 18: Previsioni con modello local linear deterministic trend sul validation set.

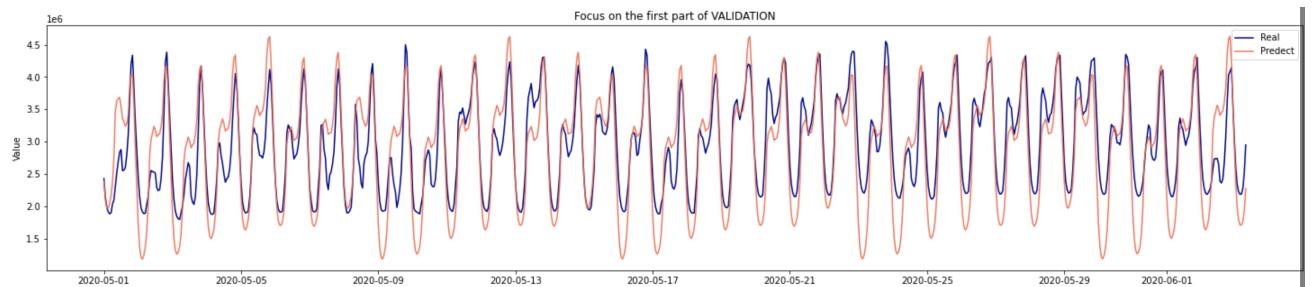


Figure 19: Focus delle previsioni con modello local linear deterministic trend sulla prima parte del validation set.

Le previsioni sul validation set suggeriscono che il modello riesce a cogliere la stagionalità settimanale e giornaliera ma non quella annuale. Sembrerebbe opportuno aggiungere le sinusoidi per cogliere tale stagionalità.

7 PROPHET

L'ultimo modello lineare testato è il Prophet che appartiene ad una libreria open source per previsioni di serie temporali univariate sviluppata da Facebook. Prophet implementa un modello additivo che supporta trend, stagionalità e festività ed è progettato per essere facile e completamente automatico. Il modello migliore trovato include una stagionalità giornaliera, settimanale e annuale. Utilizza 10 armoniche per rappresentare la stagionalità annuale, 3 per quella settimanale e 4 per quella giornaliera. Ottengo le seguenti performance: MAE train: 366467.4, MAE validation: 545685.0.

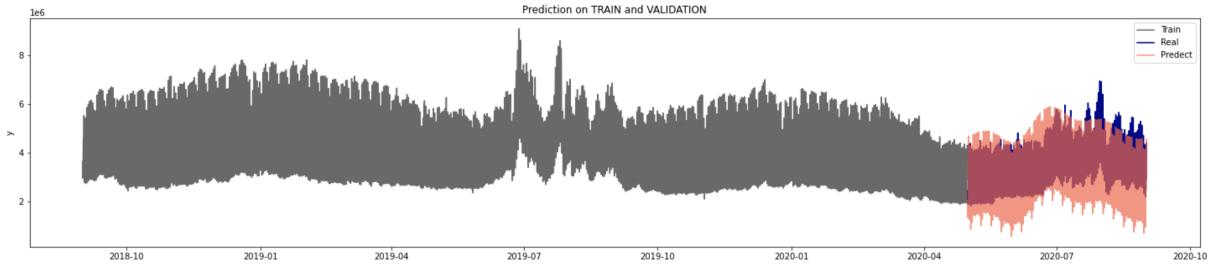


Figure 20: Previsioni con modello Prophet sul validation set.

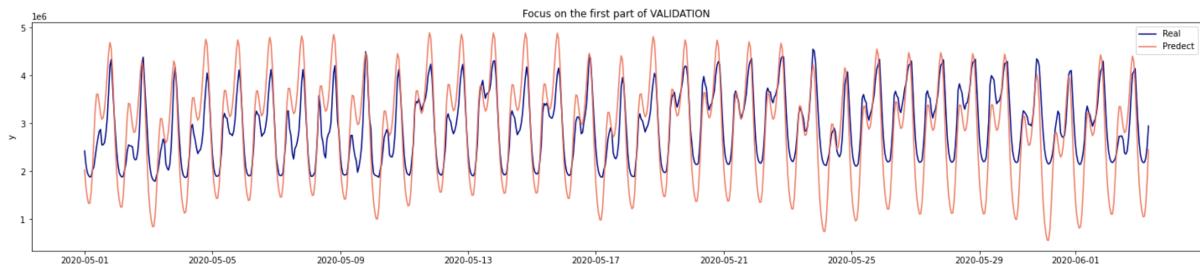


Figure 21: Focus delle previsioni con modello Prophet sulla prima parte del validation set.

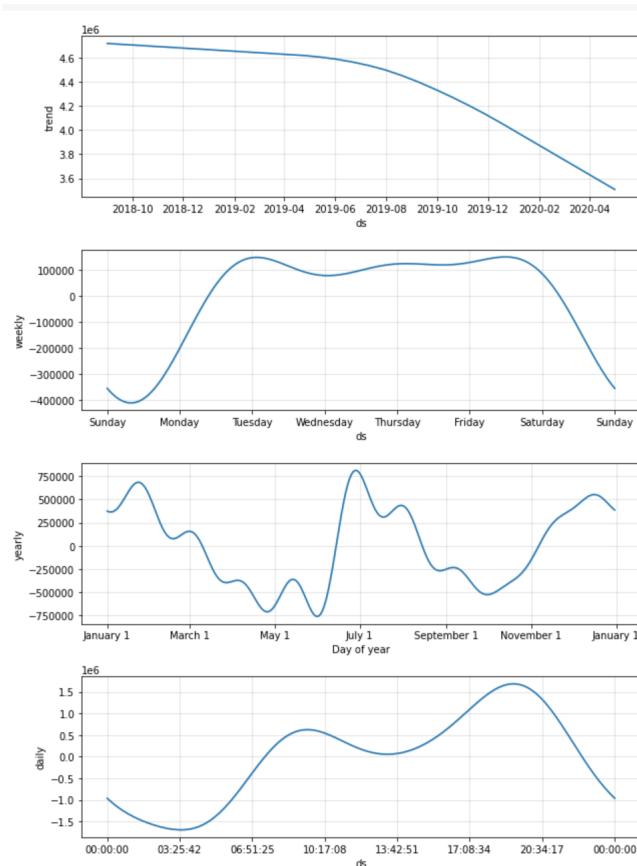


Figure 22: Grafico delle componenti del modello.

7.1 Tuning degli iperparametri

Il modello Prophet possiede una serie di iperparametri che si possono ottimizzare. Per questo, utilizzo la cross validation sull'intero dataset. I valori che portano al modello più performante sono: $changepoint_{prior,scale} = 0.01$ e $seasonality_{prior,scale} = 0.01$. Questi valori verranno usati per fare le previsioni.

change point prior scale	seasonality prior scale	MAE
0.001	0.01	633981
0.001	0.10	636224
0.001	1.00	637349
0.001	10.00	638559
0.01	0.01	612461
0.01	0.10	619521
0.01	1.00	625381
0.01	10.00	626003
0.1	0.01	659848
0.1	0.10	996928
0.1	1.00	852193
0.1	10.00	735807
0.5	0.01	1061018
0.5	0.10	2946970
0.5	1.00	3663978
0.5	10.00	3150483

Table 4: Performance del modello Prophet al variare degli iperparametri.

8 MACHINE LEARNING

L'ultima tipologia di modelli utilizzati riguardano i modelli non lineari di tipo machine learning. Per effettuare le previsioni due mesi avanti della serie storica, diversi possibili due approcci:

- Ricorsiva: le previsioni per gli step successivi al primo si ottengono in modo ricorsivo utilizzando le previsioni effettuate fino a quel punto per quelle future.
- Multi-output: il modello prevede direttamente tutto l'intervallo di previsione.
- Ibrida: il vettore di output ha più dimensioni ma il modello prevede l'intero intervallo di previsione in modo ricorsivo.

Per questioni di RAM ho affrontato la prima e terza tipologia di previsione. I due modelli usati per entrambe le tipologie, sono LSTM e GRU con implementazione stateful. Questa implementazione è utile per avere memoria dei dati tra i batch di un'epoca di addestramento. I modelli RNN sono sensibili alla dimensione dei dati di input. I dati sono stati riscalati in un intervallo tra 0 e 1. Dopodiché i dati sono stati divisi in sequenze e quindi posti nella dimensione: [samples, timesteps, features].

8.1 Metodo ricorsivo

Entrambi i modelli LSTM e GRU sono stati costruiti utilizzando un hidden layer LSTM o GRU con 200 neuroni seguito da un LeakyReLU e da un Dropout. Il LeakyReLU consente di restituire valori molto piccoli per valori

di input minori di zero. Il Dropout consente di evitare l'overfitting dei dati. È stato posto un tasso di neuroni da tagliare pari al 20%. Si osserva per entrambi i modelli che le previsioni sono buone solamente per il primo giorno di forecast.

8.1.1 LSTM

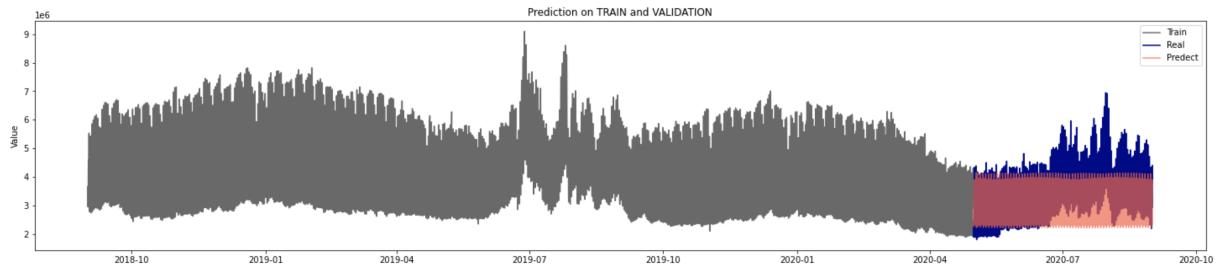


Figure 23: Previsioni con modello LSTM e implementazione stateful.

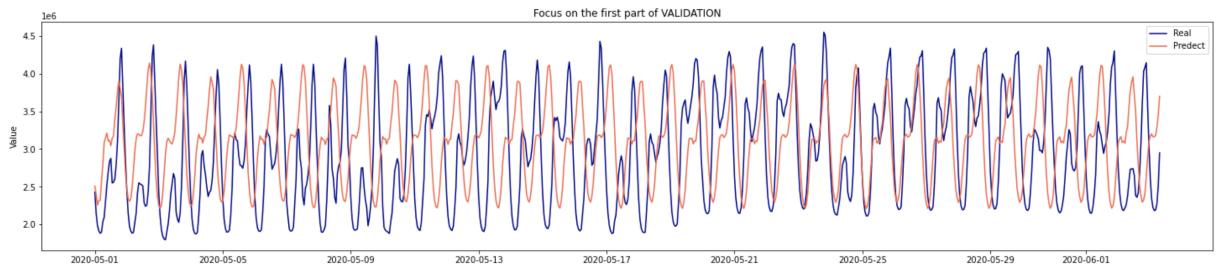


Figure 24: Previsioni con modello LSTM e implementazione stateful.

8.1.2 GRU

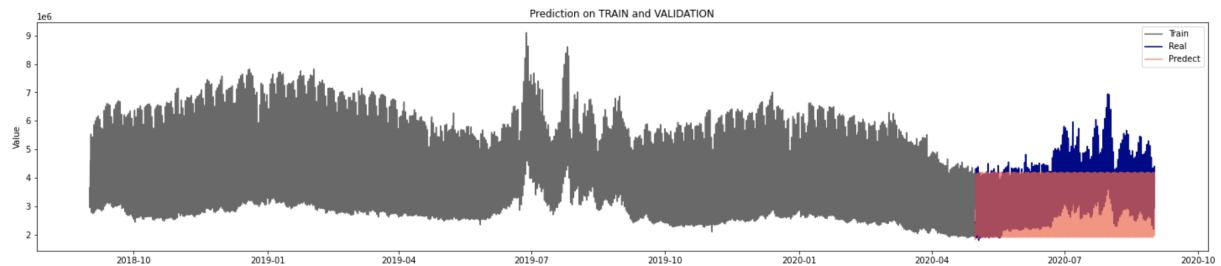


Figure 25: Previsioni con modello GRU e implementazione stateful.

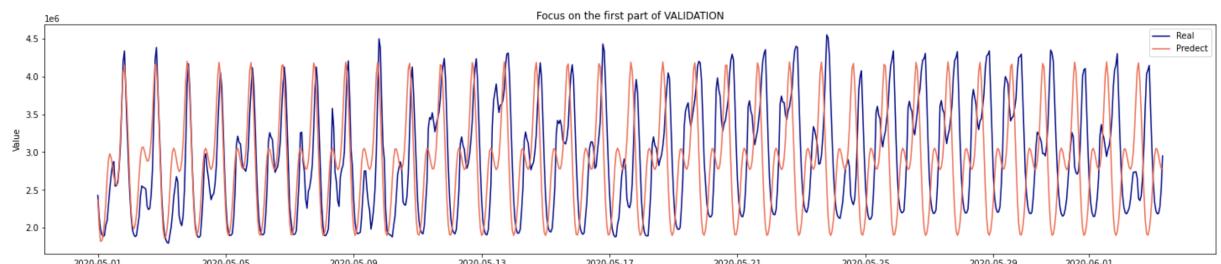


Figure 26: Previsioni con modello GRU e implementazione stateful.

8.2 Metodo ibrido

La previsione ibrida è stata costruita in diversi modi per predire diversi archi temporali: un giorno in avanti, 1 settimana in avanti. È stata testata soltanto l'architettura LSTM, con la stessa implementazione usata per il metodo ricorsivo, poiché l'architettura GRU non dava risultati soddisfacenti.

8.2.1 Previsione 1 giorno in avanti

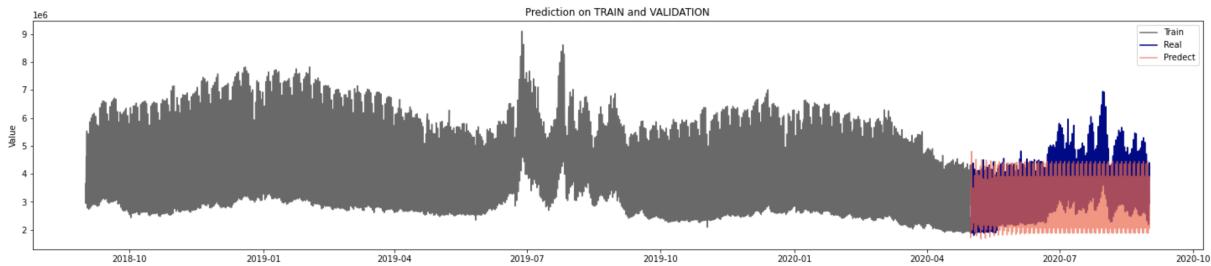


Figure 27: Previsioni ibride 1 giorno in avanti con modello LSTM e implementazione stateful.

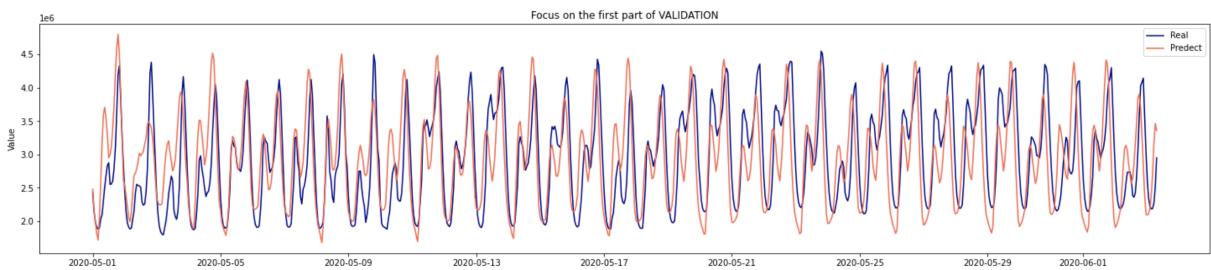


Figure 28: Previsioni ibride 1 giorno in avanti con modello LSTM e implementazione stateful.

8.2.2 Previsione 1 settimana in avanti

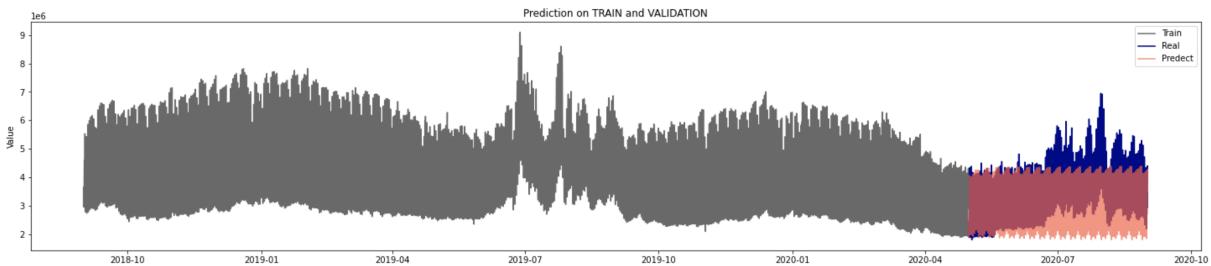


Figure 29: Previsioni ibride 1 settimana in avanti con modello LSTM e implementazione stateful.

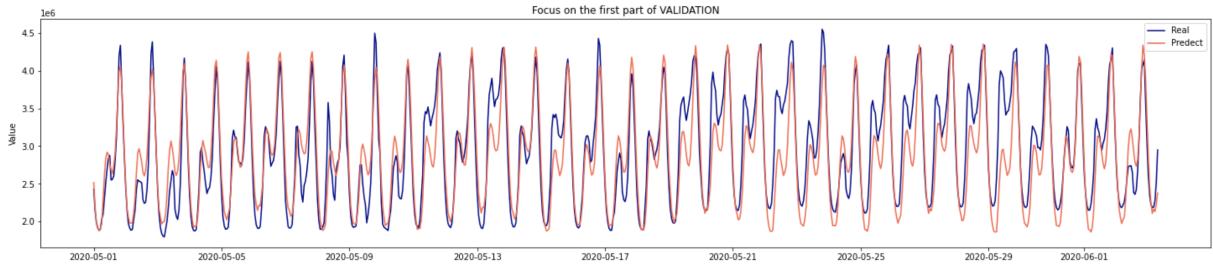


Figure 30: Previsioni ibride 1 settimana in avanti con modello LSTM e implementazione stateful.

Una considerazione che si può fare è che questi modelli non riescono a catturare la stagionalità annua. Il miglior modello risulta comunque quello ibrido con previsione una settimana in avanti. Ottengo un MAE train: 1405659,3 e un MAE validation: 740741,4. Data la multi-stagionalità dei dati, architetture quali CNN-LSTM o encoder-decoder potrebbe essere opportune.

9 Risultati

Per ciascuna tipologia di modello ho considerato quello che ha ottenuto performance migliori in termini di MAE e rieseguo il fit sull'intero dataset. Graficamente ottengo le seguenti previsioni:

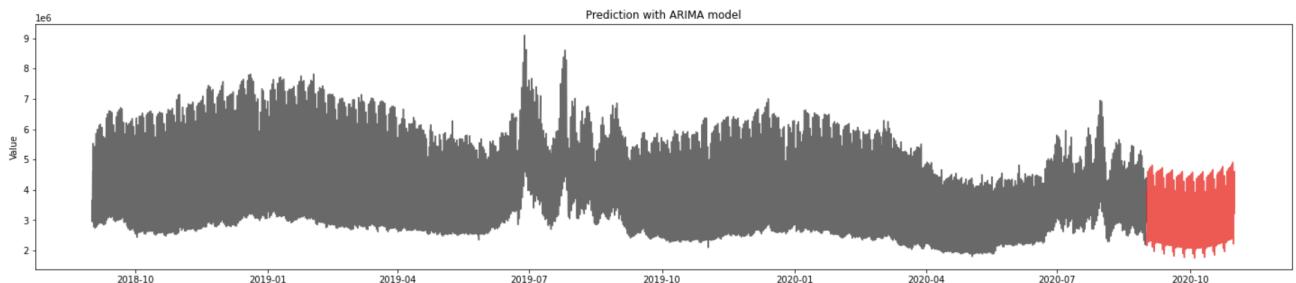


Figure 31: Previsioni con modello SARIMAX.

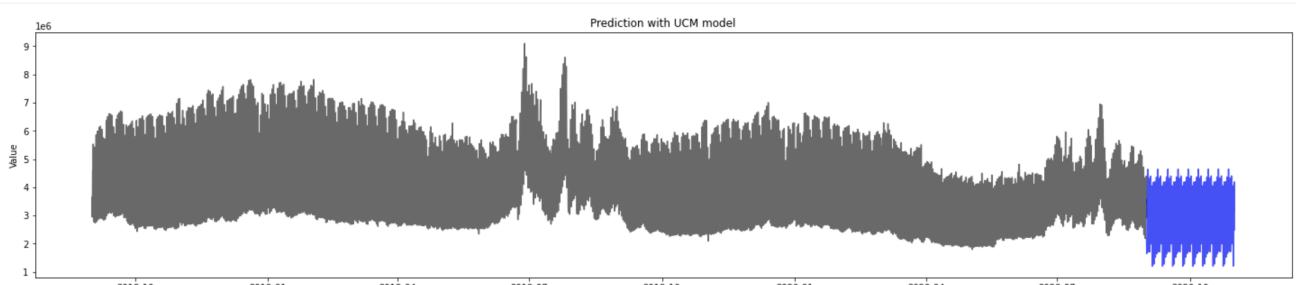


Figure 32: Previsioni con modello UCM.

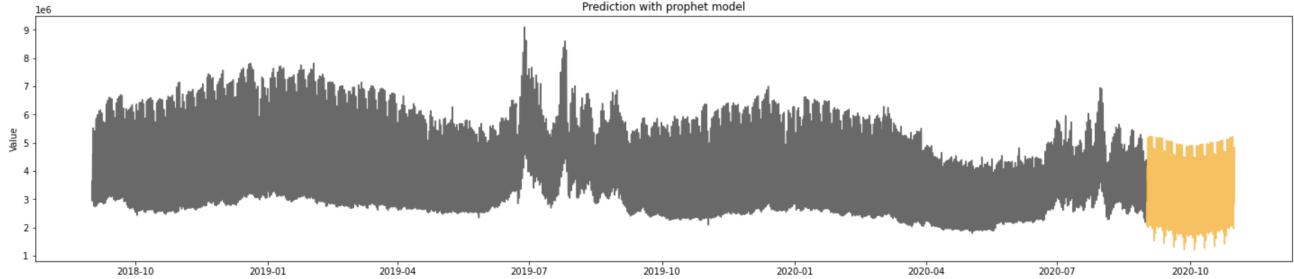


Figure 33: Previsioni con modello Prophet.

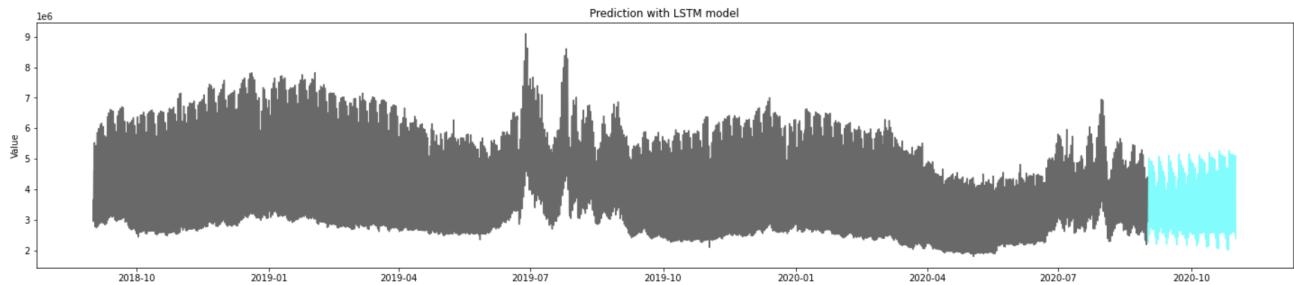


Figure 34: Previsioni con modello LSTM.

10 Conclusioni

Riassumendo le performance dei vari modelli:

Modello	MAE training	MAE validation
SARIMAX	64300.7	325296.3
UCM	136033.5	775391.1
PROPHET	367604.6	597051.0
LSTM	1405659.3	740741.4

Table 5: Performance dei vari modelli.

Il modello migliore risulta essere SARIMAX(3,0,3)(1,1,1)₂₄ con 6 armoniche settimanali e 5 armoniche annuali. Quest'ultimo riesce a cogliere la stagionalità giornaliera, settimanale e annuale.

Per poter ottenere risultati più soddisfacenti potrebbe essere utile conoscere la natura dei dati.