

Streaming Data Management and Time Series Analysis


Sviluppo di un sistema di previsione di dati temporali

800928

Chiaretti Giulia



Agenda

1. Introduzione e obiettivo
2. Dataset
3. Data Exploration
4. Modelli sviluppati 
 - Lineari
 - Non lineari
5. Conclusioni

1 - Introduzione e obiettivo



Previsione dei prezzi
giornalieri del
mercato energetico.

2010 - 2018



2019

Modelli ARIMA

Modelli UCM

Modelli ML: kNN, RNN



MAPE

2 - Dataset

1 gen 2010 - 31 dic 2018



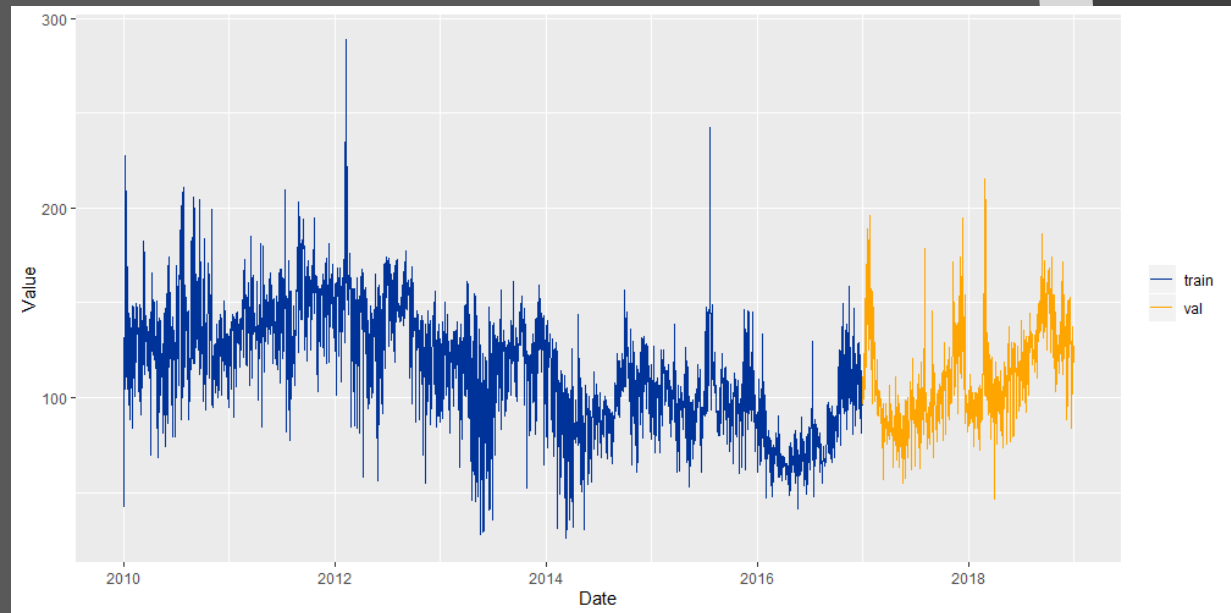
3287 osservazioni

Training set

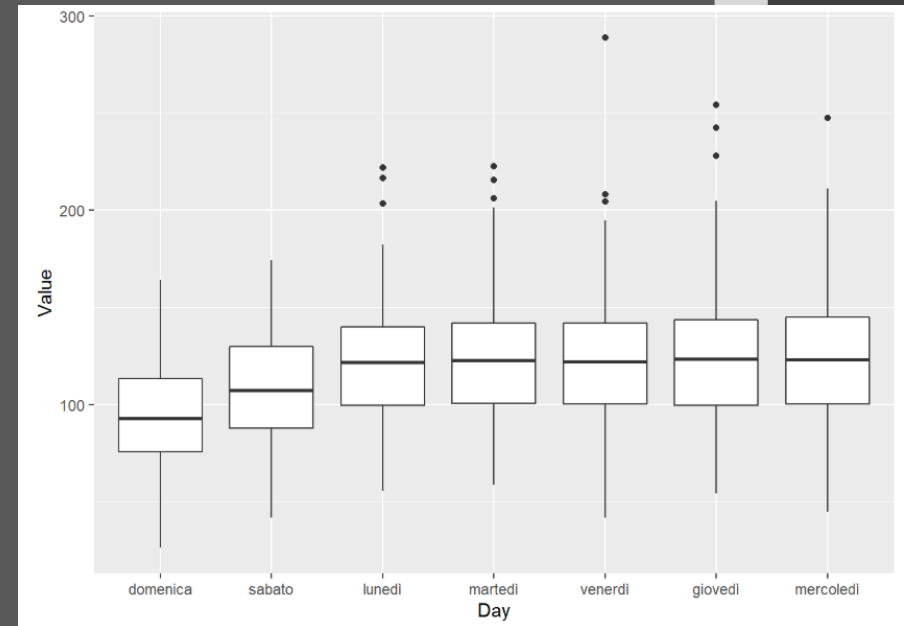
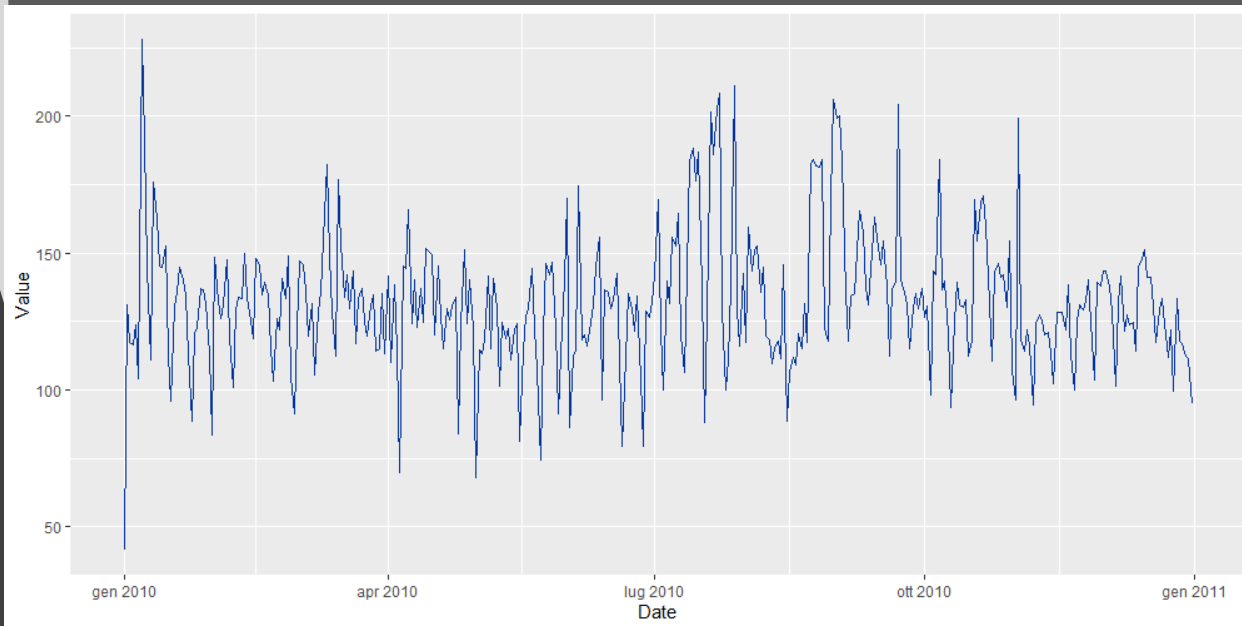
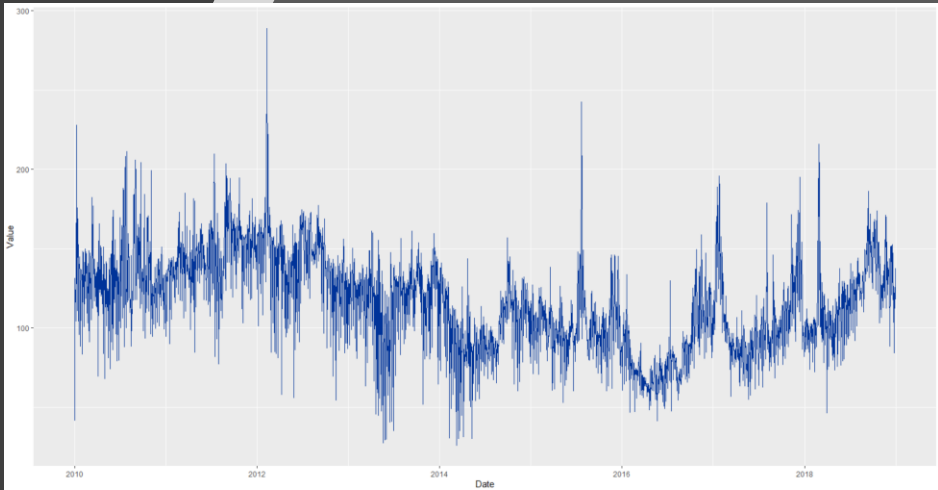
01-01-2010 – 31-12-2016

Validation set

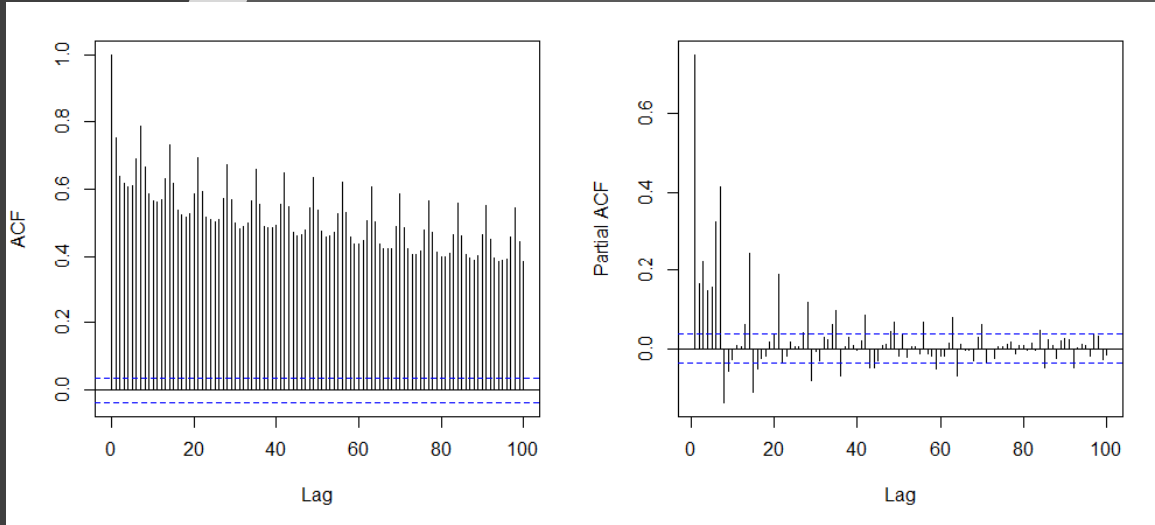
01-01-2017 – 31-12-2018



3 - Data Exploration

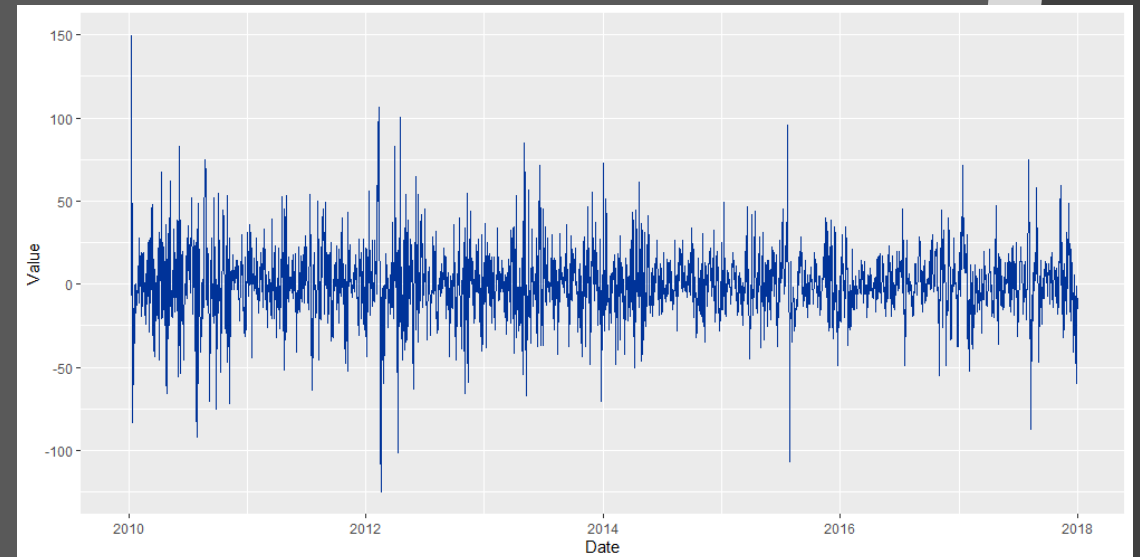


4 - Modelli Sviluppati - ARIMA

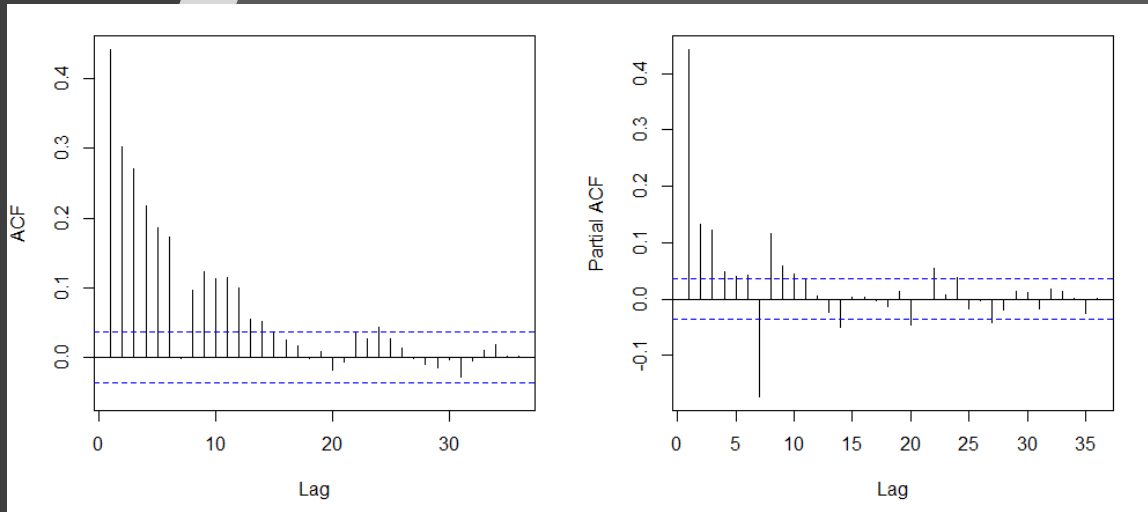


Correlogrammi della
serie originale

Serie differenziata
stagionalmente



4 - Modelli Sviluppati - ARIMA



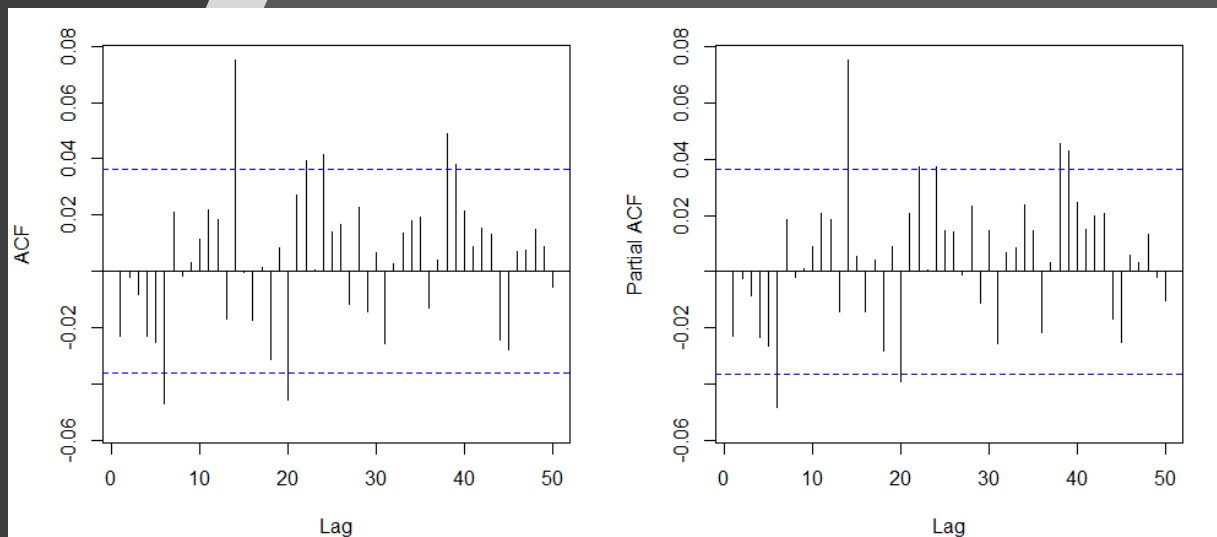
Correlogrammi dei residui del modello SARIMA(0, 0, 0)(1, 1, 1)₇

```
## [1] "AIC del modello ARIMA (2,0,0) (1,1,1): 20930.31"  
## [1] "AIC del modello ARIMA (3,0,0) (1,1,1): 20856.24"  
## [1] "AIC del modello ARIMA (4,0,0) (1,1,1): 20831.23"  
## [1] "AIC del modello ARIMA (5,0,0) (1,1,1): 20810.74"  
## [1] "AIC del modello ARIMA (6,0,0) (1,1,1): 20766.23"
```

Confronto tra i modelli SARIMA con diverso ordine della componente AR non stagionale.

Modello migliore in termini di AIC: SARIMA(6, 0, 0)(1, 1, 1)₇

4 - Modelli Sviluppati - ARIMA



Modulo delle radici della componente autoregressiva del modello.

```
## [1] 0.7026783 0.6879106 0.6879106 0.9323356 0.7026783 0.6464541
```

Correlogrammi dei residui del modello SARIMA(6, 0, 0)(1, 1, 1)₇

Si valuta la possibilità di una differenziazione di primo ordine ma il modello SARIMA(6, 1, 0)(1, 1, 1)₇ non genera miglioramenti né in termini di AIC né sui correlogrammi dei residui.

4 - Modelli Sviluppati - ARIMA



Necessità di maggiore
adattamento ai dati



Previsione sul validation set del
modello SARIMA(6, 0, 0)(1, 1, 1)₇
AIC: 20766.23

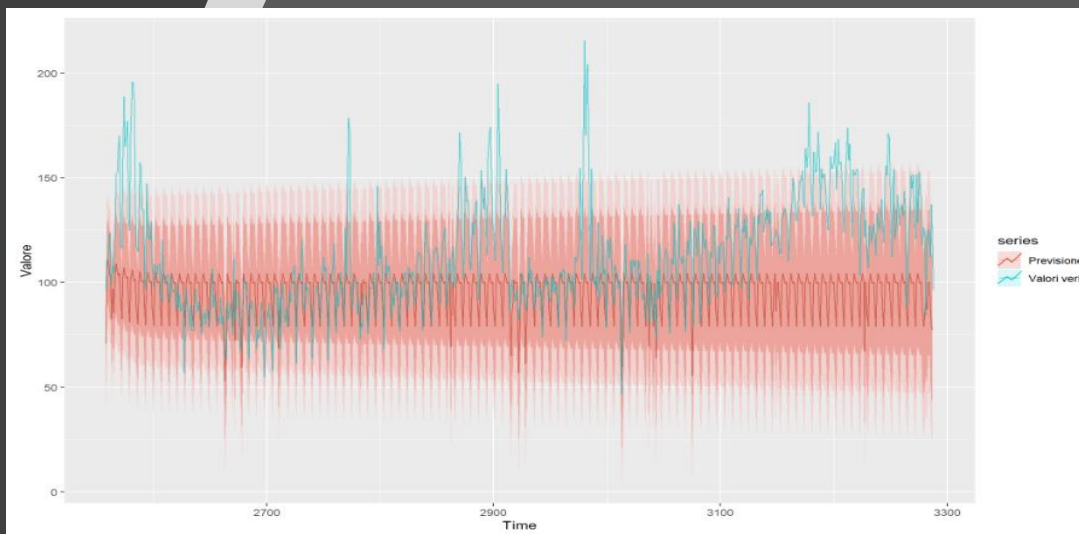
Regressori esterni:

- 18 sinusoidi con frequenza $\left(\frac{2\pi}{365.25}\right)$
- 9 variabili dummy con le principali festività del calendario italiano

4 - Modelli Sviluppati - ARIMA

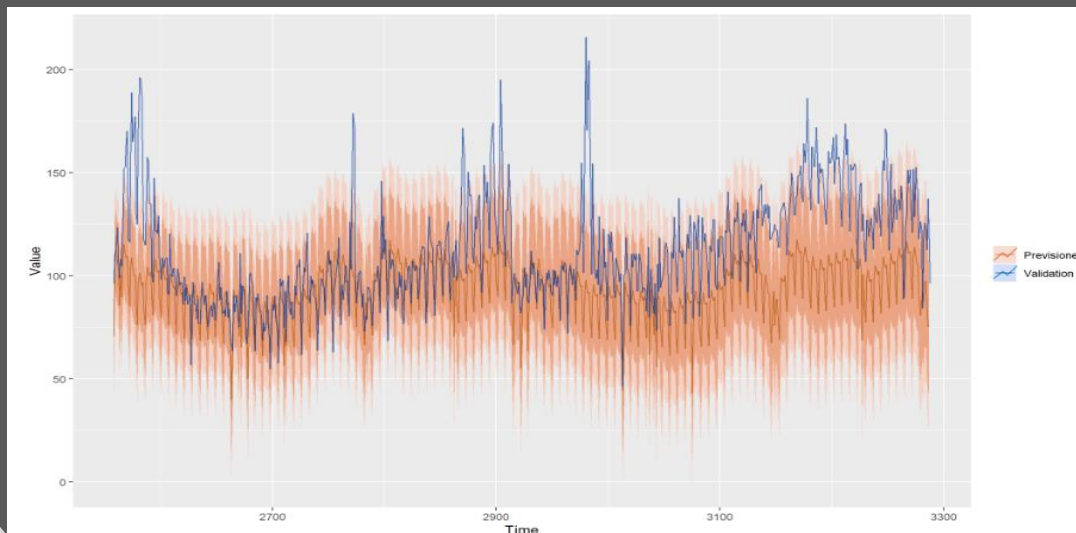
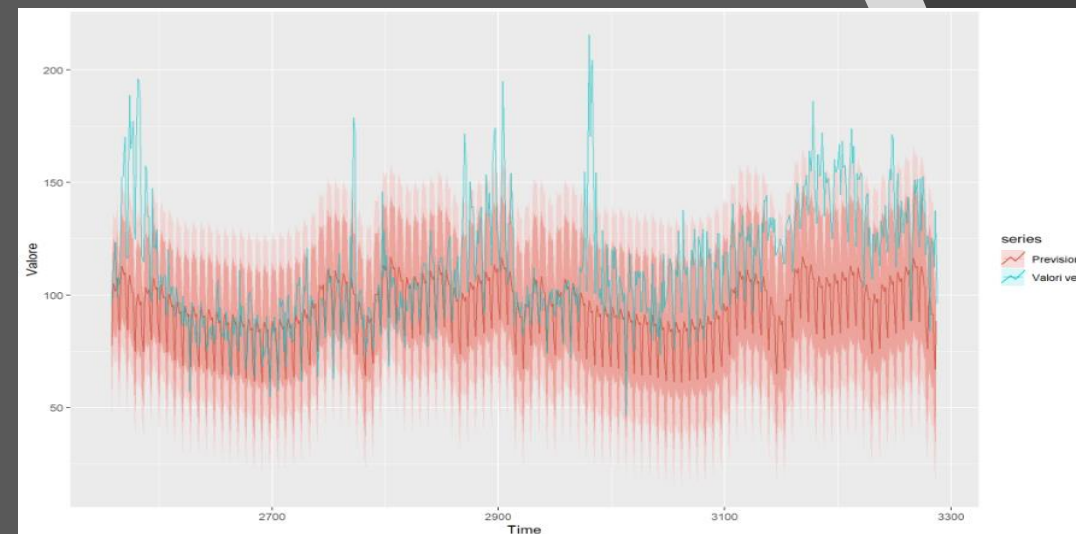
SARIMA(6, 0, 0)(1, 1, 1)₇ + dummy

AIC: 20482 MAPE(train): 9.39



SARIMA(6, 0, 0)(1, 1, 1)₇ + sinusoidi

AIC: 20769 MAPE(train): 9.69



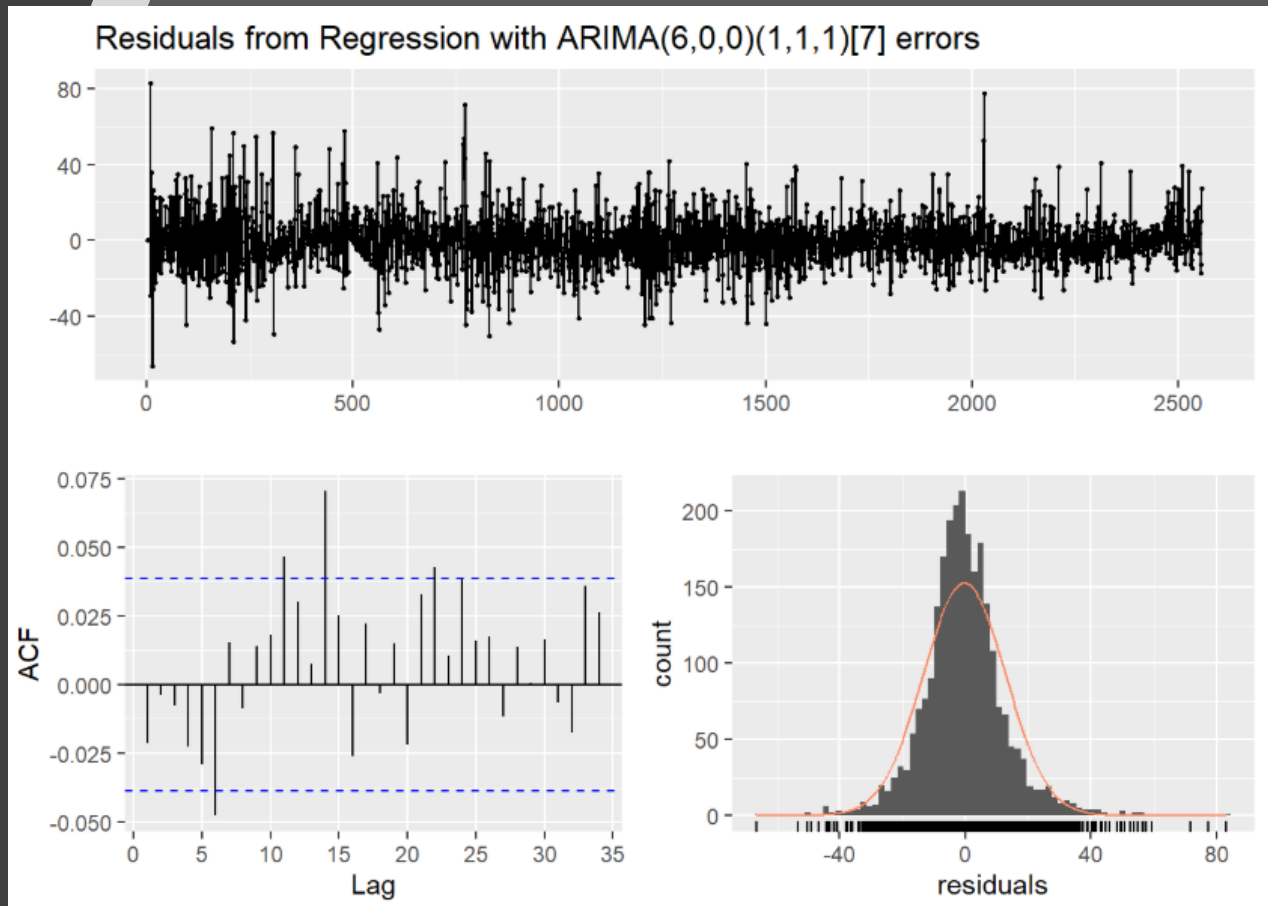
SARIMA(6, 0, 0)(1, 1, 1)₇
+ dummy + sinusoidi

AIC: 20495

MAPE(train): 9.27



4 - Modelli Sviluppati - ARIMA



Si verifica che i residui del modello siano generati da un processo white noise.

Test di Ljung-Box di autocorrelazione globale.

```
## Ljung-Box test
##
## data: Residuals from Regression with ARIMA(6,0,0)(1,1,1)[7] errors
## Q* = 93.663, df = 3, p-value < 2.2e-16
##
## Model df: 53. Total lags used: 56
```

4 - Modelli Sviluppati - UCM

STAGIONALITA'

TREND

Regressori Esterni

Stagionalità settimanale a dummy stocastiche

+

Stagionalità intra-annua a sinusoidi stocastiche

+

Local Linear
Trend

Local Linear
Trend

Random Walk

Random Walk
Integrato

+

Dummy
festività

+

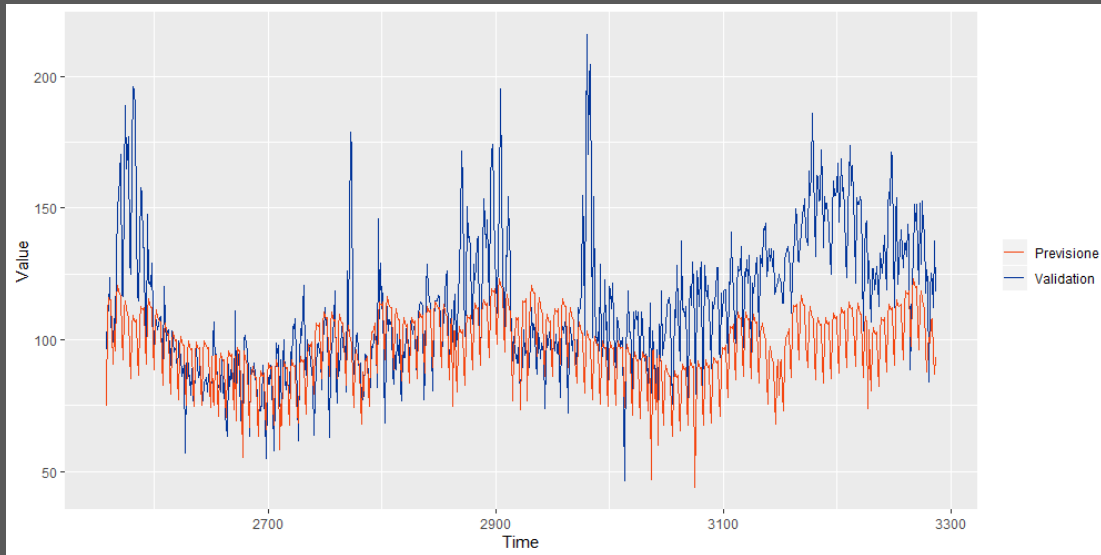
Dummy
festività

+

Dummy
festività

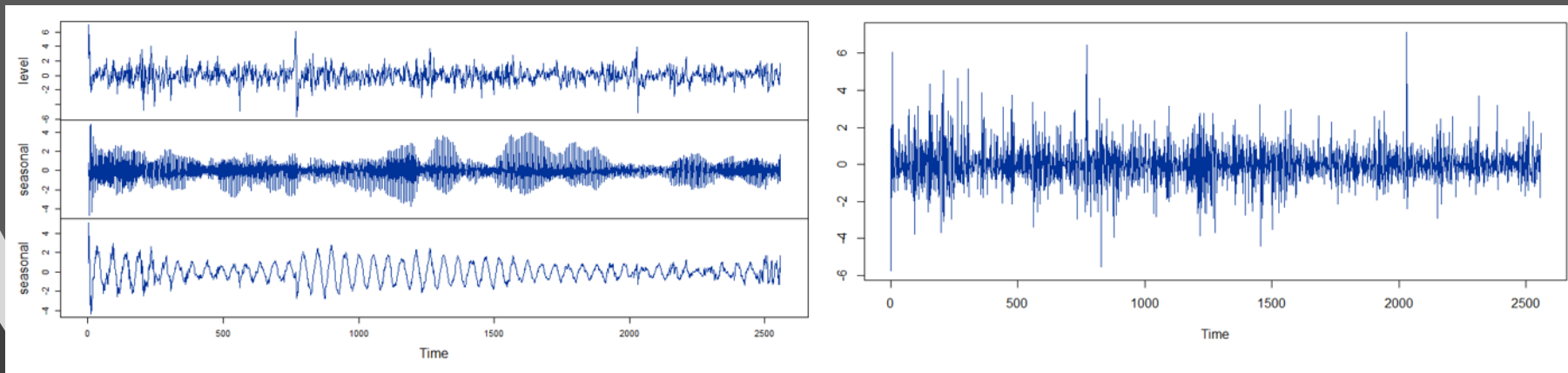
	LLT (no regr)	LLT	RW	IRW
Training	7.91	7.32	7.34	7.32
Validation	18.78	18.43	18.27	22.28

4 - Modelli Sviluppati - UCM



Previsione sul validation
set del miglior modello
UCM

Analisi del disturbance
smoother e degli errori di
osservazione



4 - Modelli Sviluppati – ML: kNN

L'algoritmo kNN prevede i valori futuri basandosi sui k Nearest Neighbors, ovvero le k serie più simili all'ultimo lag temporale che precede i valori da prevedere.

Metodo ricorsivo

Lags = 365

h = 730 (validation)

Media pesata

k = 50 (euristica)

Metodo ricorsivo

Lags = 365

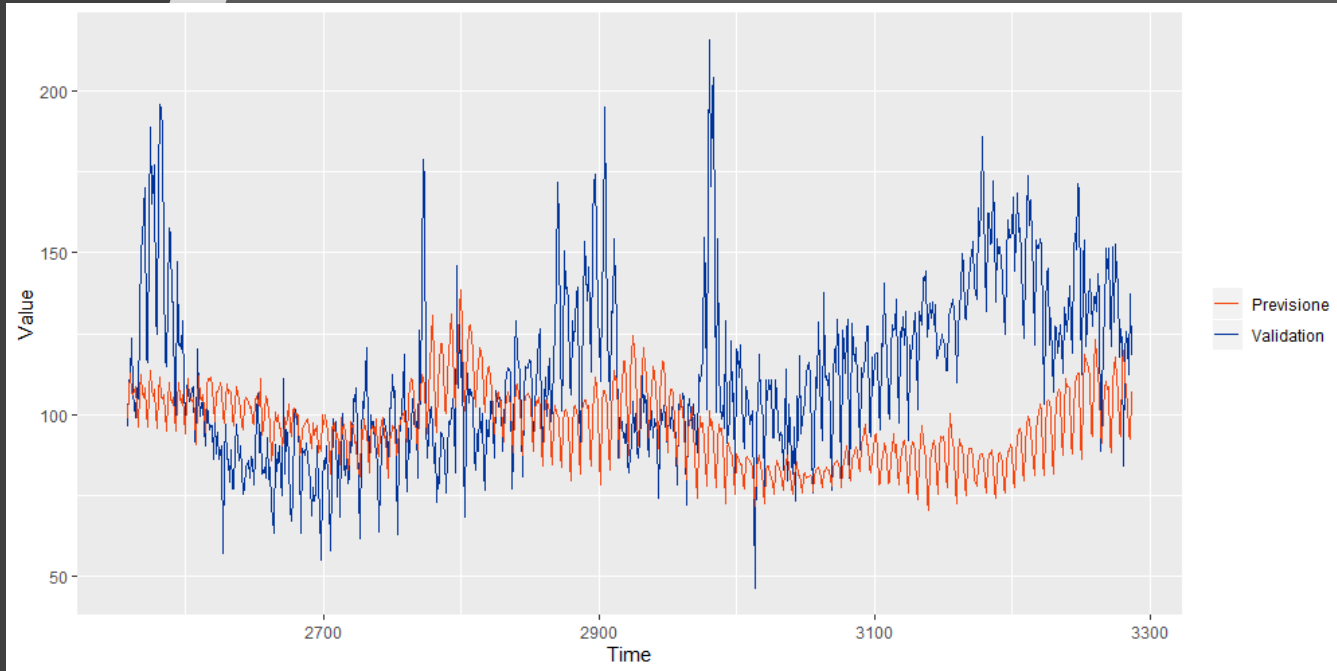
h = 730 (validation)

Media pesata

k = 30, 35, 40, 45, 50, 55,
60, 65, 70

	kNN (k=50)	kNN (multiple k)
Validation	17.64	17.51

4 - Modelli Sviluppati – ML: kNN



Previsione sul validation set
del modello kNN (k multiplo)

4 - Modelli Sviluppati – ML: RNN

Preprocessing dei dati:

- Dati centrati e scalati
- Posti sottoforma di array con 3 dimensioni:
 - Dimensione del campione
 - Numero di feature
 - Timesteps

Batch size = 365

Epoche = 200

Adam optimizer lr = 0.001

Loss = mae

LSTM

Layer LSTM - tanh (100 units)

Dropout 0.3

Layer LSTM - tanh (90 units)

Final dense layer - linear

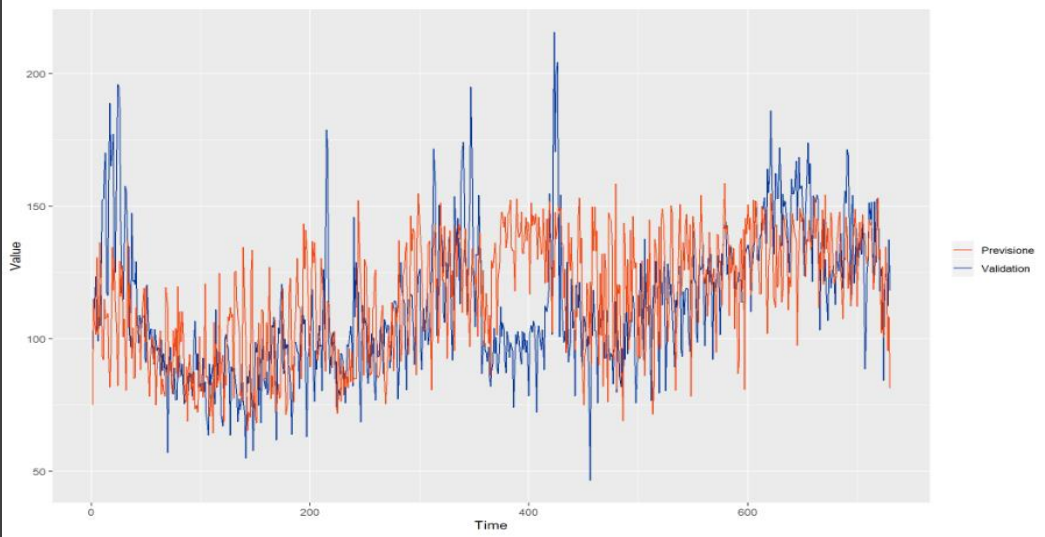
GRU

Layer GRU - tanh (90 units)

Dropout 0.3

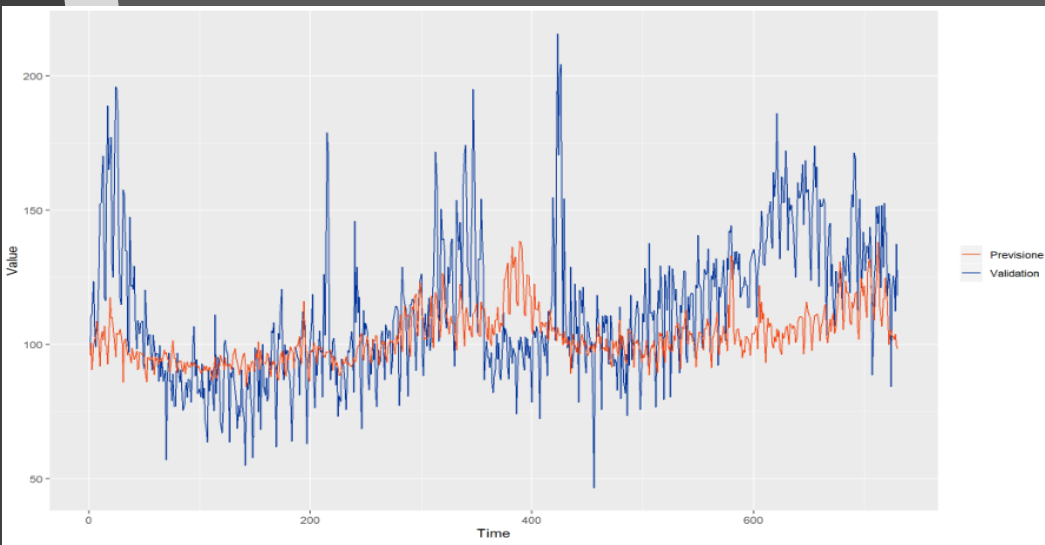
Final dense layer - linear

4 - Modelli Sviluppati – ML: RNN



Previsione sul validation set
della rete LSTM

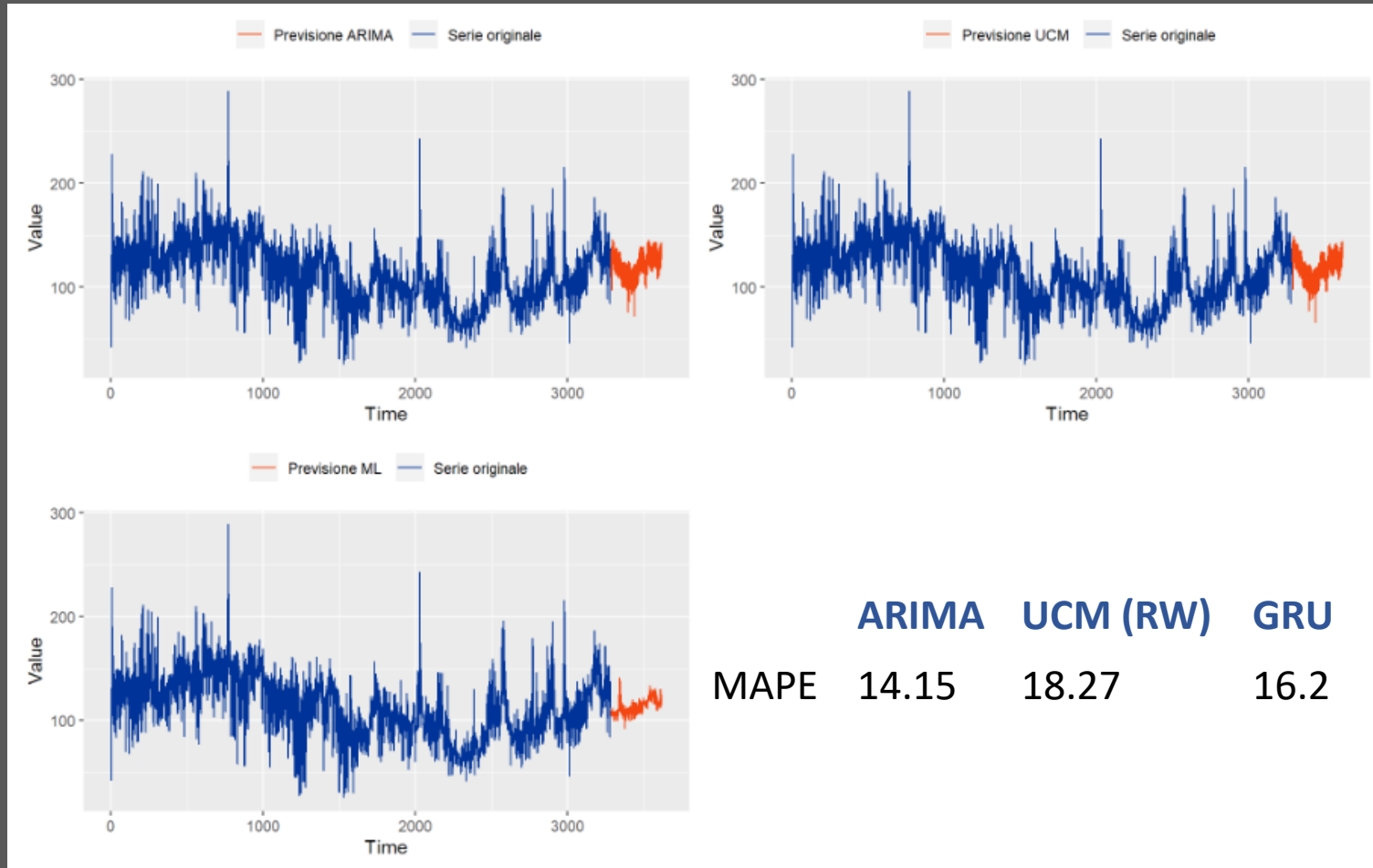
MAPE: 20.17



Previsione sul validation set
della rete GRU

MAPE: 16.2

5 - Conclusioni





Grazie per l'attenzione

Streaming Data Management and Time Series Analysis

800928

Chiaretti Giulia

