



UNIVERSITÀ DI PISA

CORSO DI LAUREA MAGISTRALE IN ARTIFICIAL  
INTELLIGENCE AND DATA ENGINEERING

---

## Statistica

---

Progetto 3

Giuseppe MARTINO

23 dicembre 2020

# Indice

<b>1</b>	<b>Introduzione e caratteristiche dei dati</b>	<b>2</b>
<b>2</b>	<b>Visualizzazione della serie</b>	<b>2</b>
<b>3</b>	<b>Decomposizione della serie</b>	<b>3</b>
3.1	Decomposizione additiva . . . . .	3
3.2	Decomposizione moltiplicativa . . . . .	4
3.3	Confronto tramite analisi dei residui . . . . .	5
<b>4</b>	<b>Analisi dei dati</b>	<b>6</b>
4.1	Analisi dei residui . . . . .	7
4.2	Previsioni . . . . .	9
4.3	Autovalutazione . . . . .	9

# 1 Introduzione e caratteristiche dei dati

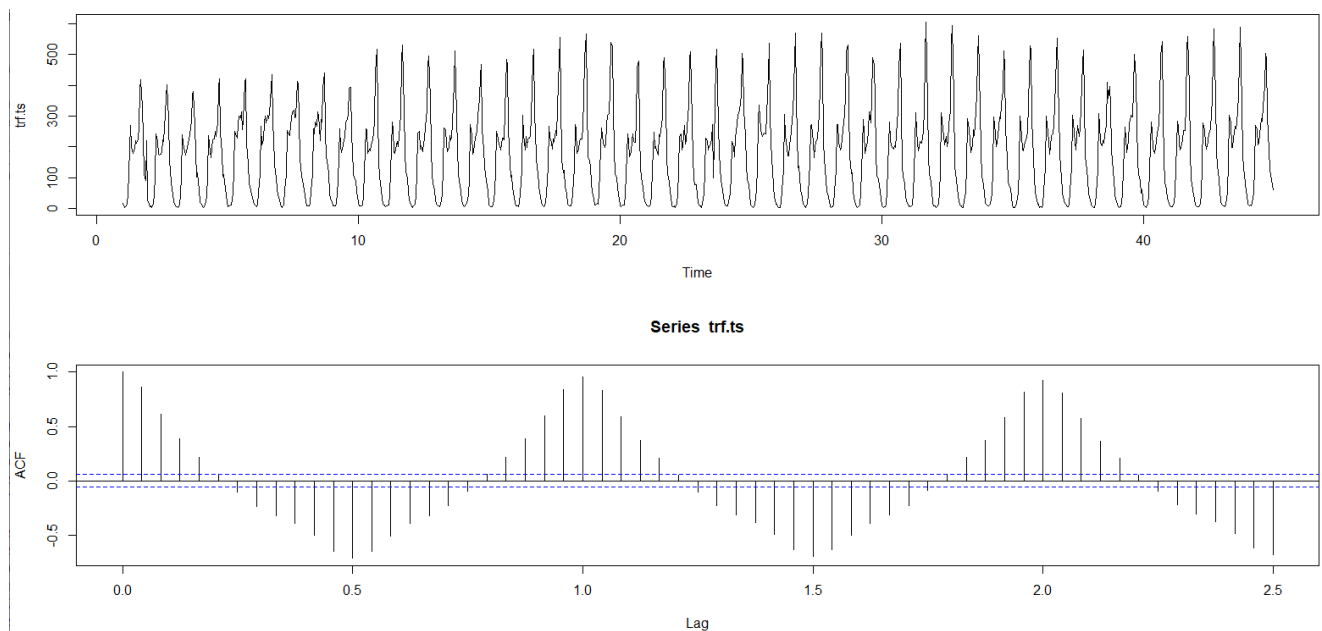
L'obiettivo dello studio consiste nell'analisi di una serie temporale riguardante il traffico stradale. Il dataset è stato costruito utilizzando i dati del 2019 disponibili al seguente link: <https://data.gov.uk/dataset/c7a7c892-16e9-4d07-a539-10390b45c14c/leeds-annual-traffic-g> e contiene il numero dei veicoli transitati in una strada per fascia oraria. In particolare sono stati considerati i dati per la strada A58 Cockersdale (ID:9913) in direzione sud ovest della città di Leeds, nel Regno Unito, dei mesi di Ottobre e Novembre in quanto senza valori mancanti.

Il modello ricavato dall'analisi può essere utile per i gestori di esercizi commerciali (si pensi a tipi di attività come i mcDrive e simili) per prevedere quali sono gli orari con più affluenza di automobili nella strada in cui si trova l'attività, in modo da ottimizzare gli orari di apertura e chiusura, per coordinare gli orari di lavoro dei dipendenti e per capire il numero stesso di dipendenti necessari a un certo orario del giorno. Per questo sono stati presi in considerazione i giorni lavorativi (dal lunedì al venerdì) per ogni settimana.

Alla fine del processo di pulizia dei dati, la serie strutturata da 1056 osservazioni, dove ognuna rappresenta il numero di veicoli transitati in ogni ora del giorno

## 2 Visualizzazione della serie

Per prima cosa la serie è stata visualizzata ed è stata tracciata la funzione di autocorrelazione allo scopo di riconoscere la natura della serie e distinguerne le componenti strutturali.



Dato che è stata riconosciuta una netta componente stagionale, è stata eseguita una decomposizione della serie.

## 3 Decomposizione della serie

### 3.1 Decomposizione additiva

Per decomporre la serie, è stato provato inizialmente un modello additivo.

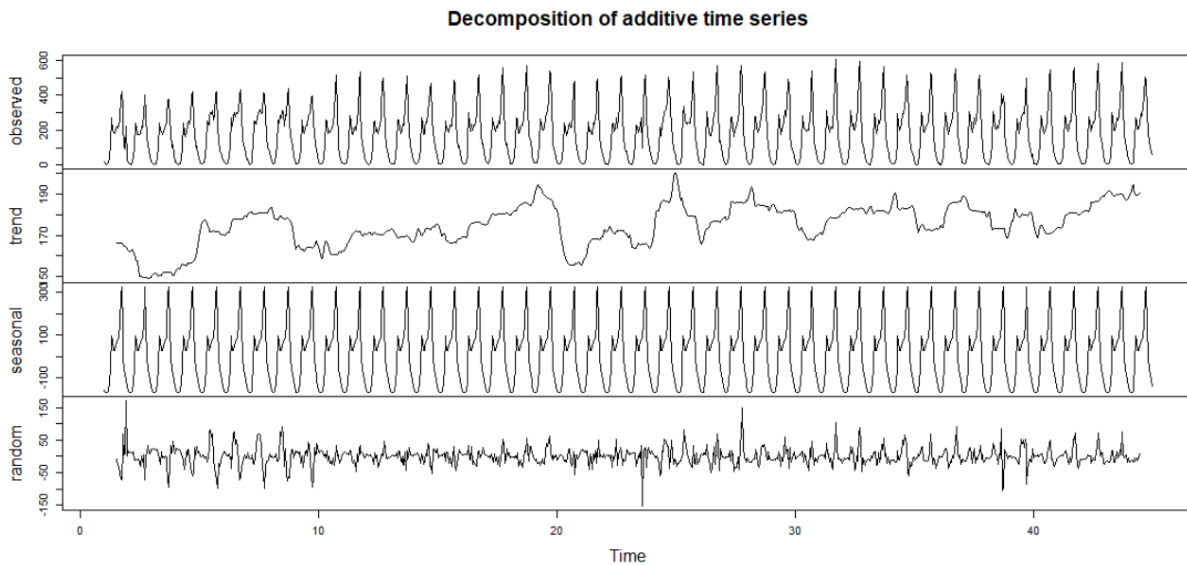


Figura 2: Decomposizione additiva

La decomposizione conferma l'analisi preliminare di una serie prevalentemente stagionale. Infatti, se confrontate le scale di stagionalità e trend, si nota che l'ampiezza del trend rilevato è minore rispetto a quella della stagionalità. I residui non sono molto soddisfacenti, infatti si notano zone con diversa variabilità, segno che i residui non hanno perso del tutto le caratteristiche temporali e che non sono rumore vero.

### 3.2 Decomposizione moltiplicativa

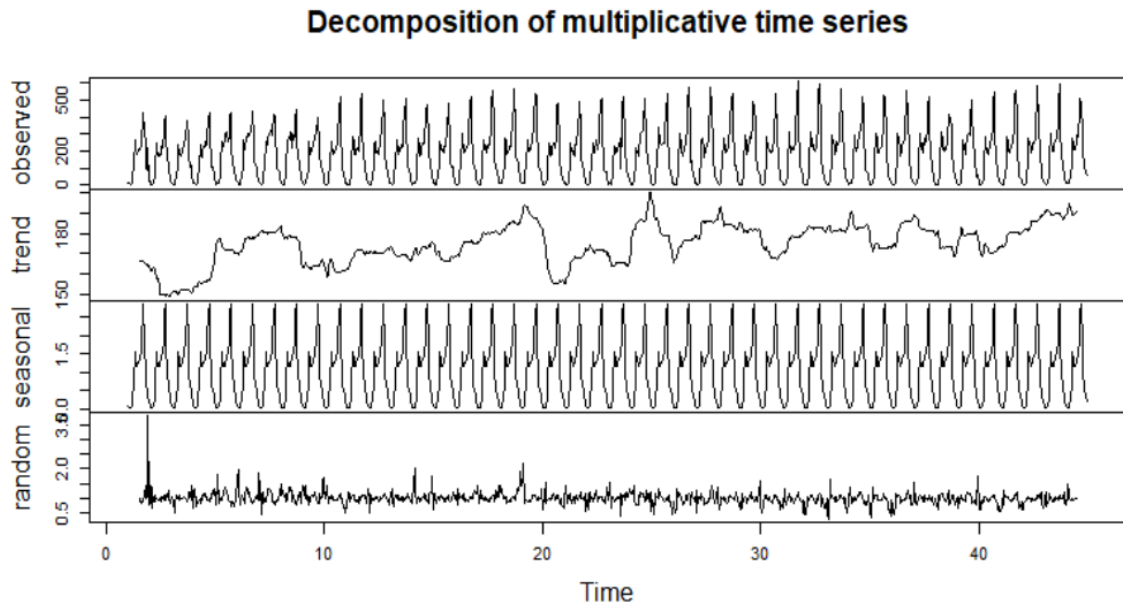


Figura 3: Decomposizione moltiplicativa

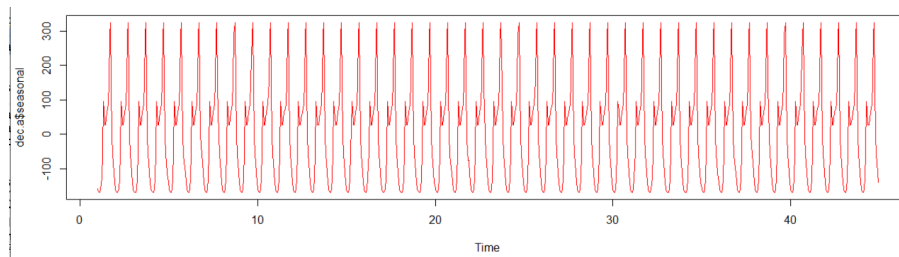


Figura 4: Confronto tra stagionalità moltiplicativa (in rosso) e additiva (nero)

La stagionalità delle due decomposizioni è praticamente la stessa. I residui sembrano essere in generale un po' migliori nella decomposizione moltiplicativa, ma per decidere quale decomposizione fosse la più appropriata, è stata fatta un'analisi dei residui

### 3.3 Confronto tramite analisi dei residui

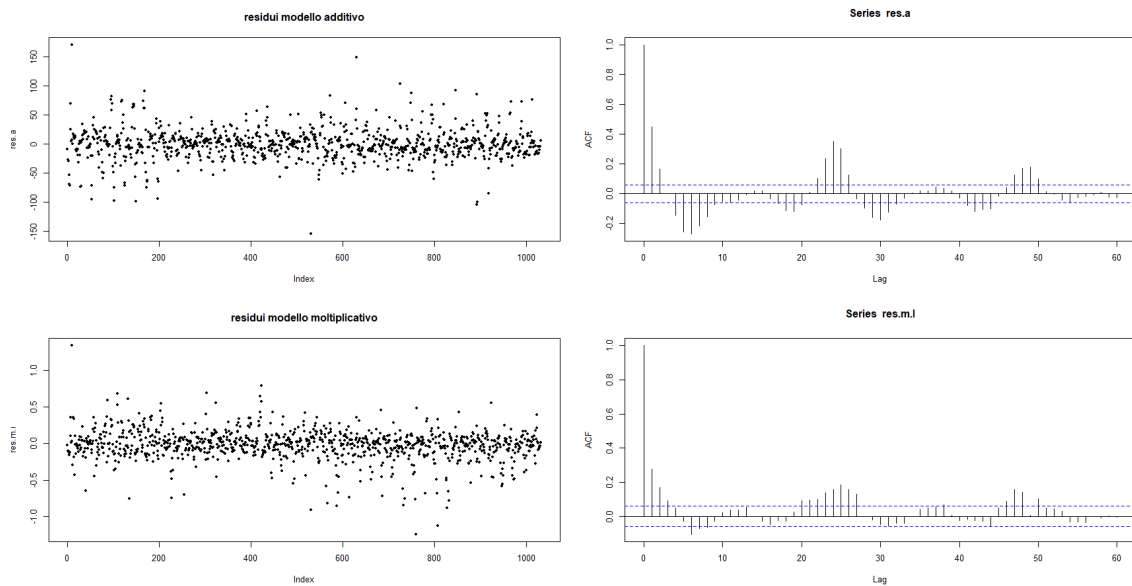


Figura 5: Residui del modello additivo (sopra) e di quello moltiplicativo (sotto)

```
> var(res.a)/var(window(trf.ts,c(1,13),c(44,12)))  
[1] 0.03580994  
> var(res.m)/var(window(log(trf.ts),c(1,13),c(44,12)))  
[1] 0.02330409
```

Figura 6: Proporzioni di varianza spiegata date ai residui

```
> sd(acf(res.a,plot = F)$acf)  
[1] 0.2492495  
> sd(acf(res.m,plot = F)$acf)  
[1] 0.1937032
```

Figura 7: Variabilità delle acf dei residui

Sono stati eseguiti ulteriori test (grafico quantile-quantile, confronto densità empiriche e gaussiane, test di Shapiro-Wilk) per controllare se i residui potessero essere assimilati a una distribuzione gaussiana ma i risultati sono stati deludenti. I residui di entrambi i modelli infatti hanno conservato struttura, ma combinando i risultati grafici e numerici, è evidente che c'è più struttura nel modello additivo. Già da un'analisi visiva dei grafici di autocorrelazione, il modello additivo mostra più struttura, ipotesi confermata da una maggiore proporzione di varianza spiegata e una maggiore variabilità della funzione di autocorrelazione calcolata numericamente.

È stata anche provata una decomposizione con una stagionalità non costante usando il comando `stl` ed effettivamente si è riusciti meglio a catturare la variabilità della stagionalità. La scala della stagionalità è anche aumentata rispetto a quella dei residui che però non perdono ancora del tutto struttura.

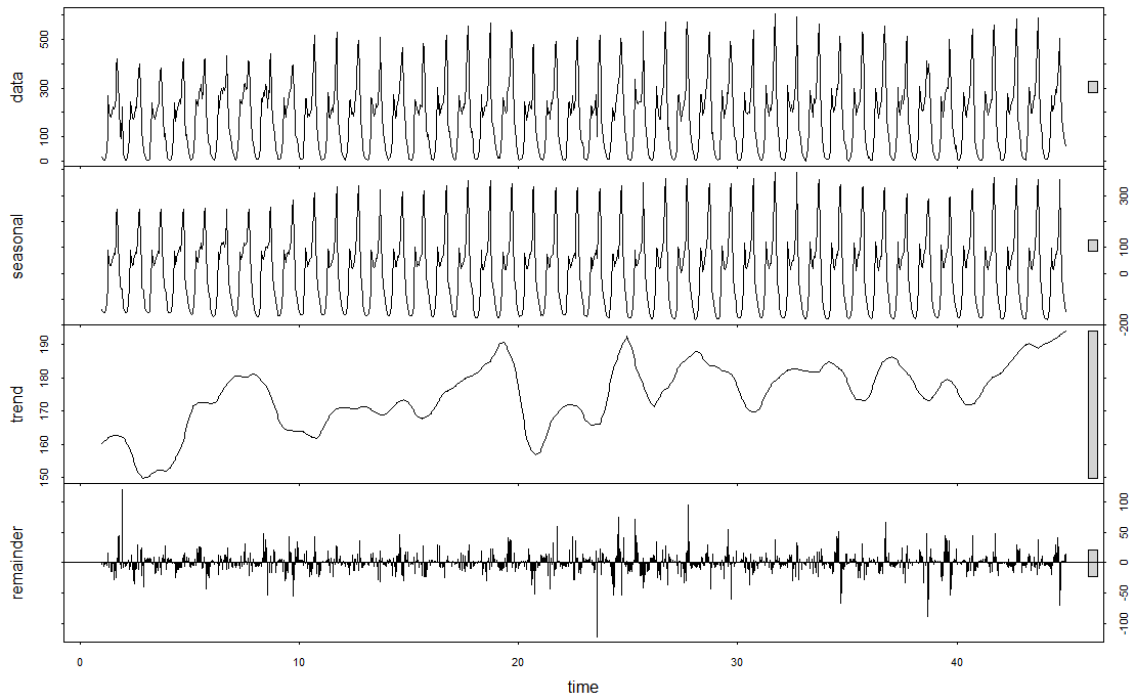
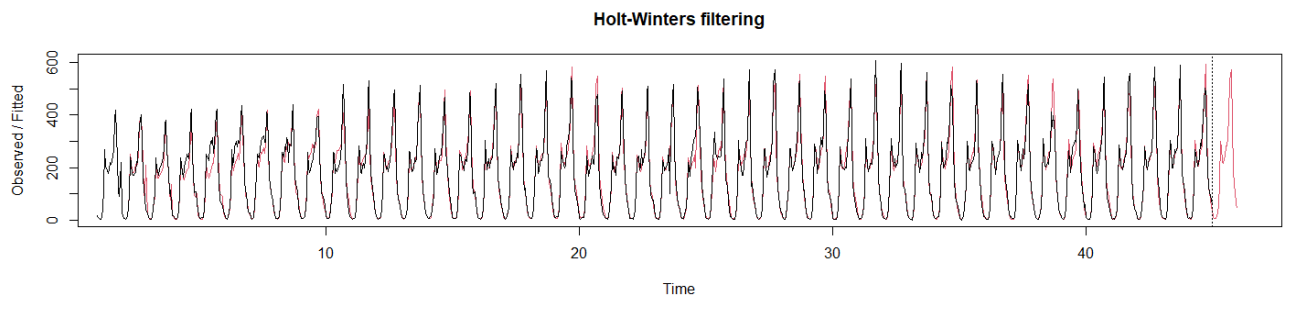


Figura 8: Decomposizione con stl

## 4 Analisi dei dati

Per l'analisi e la previsione della serie ho provato metodi con smorzamento esponenziale e metodi autoregressivi. Tutti i metodi hanno restituito risultati non troppo diversi, per cui di seguito vengono illustrati e confrontati due metodi: **Holt-Winters moltiplicativo** e **metodo autoregressivo dei minimi quadrati**.

Al fine di migliorare i risultati dati dal modello di Holt-Winters, ho modificato i parametri calcolati automaticamente dal comando minimizzando lo scarto quadratico dei residui. In particolare ho diminuito il valore di  $\alpha$  (settandolo a 0.01) che enfatizzava troppo gli ultimi valori più alti della serie, dando a sua volta una previsione troppo alta.



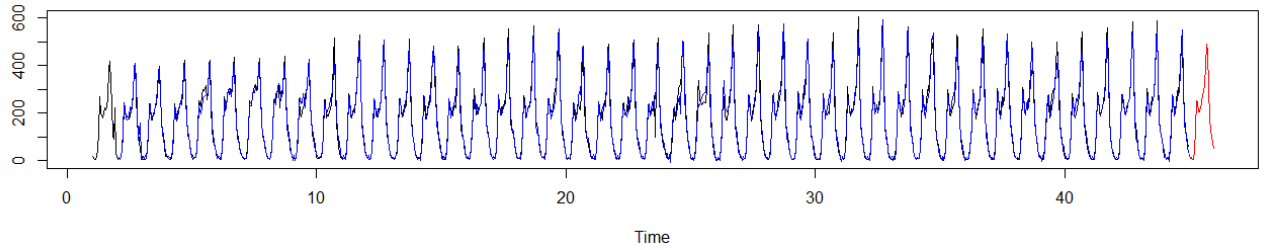


Figura 9: Analisi e previsione. Holt-Winters (sopra) e minimi quadrati (sotto)

Da una prima visualizzazione grafica i metodi sembrano fare abbastanza bene, data una buona sovrapposizione degli andamenti con la serie originale. Una differenza sembra essere presente in alcuni picchi, dove il metodo di Holt-Winters sovrastima rispetto al metodo autoregressivo.

#### 4.1 Analisi dei residui

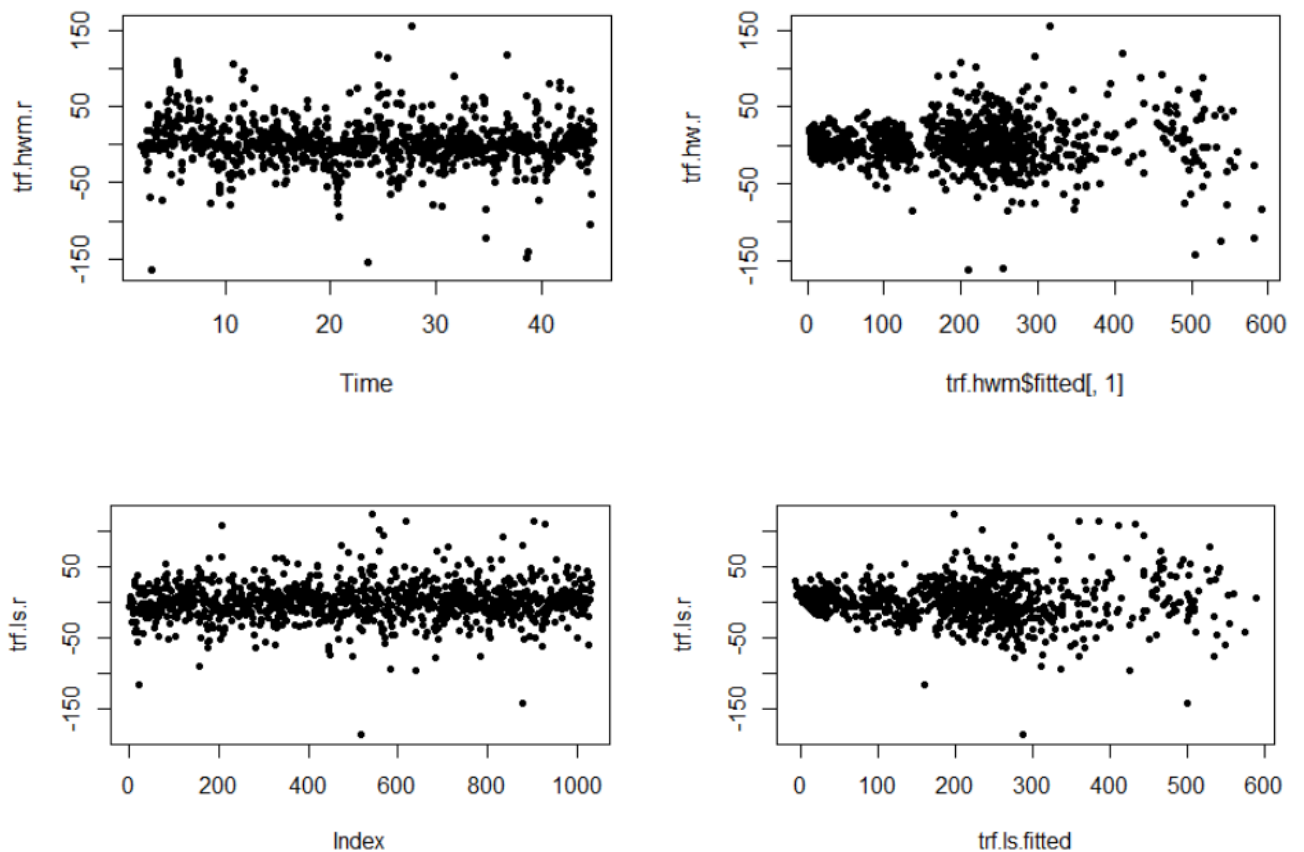


Figura 10: Distribuzione dei residui. Holt-Winters (sopra) e minimi quadrati (sotto)



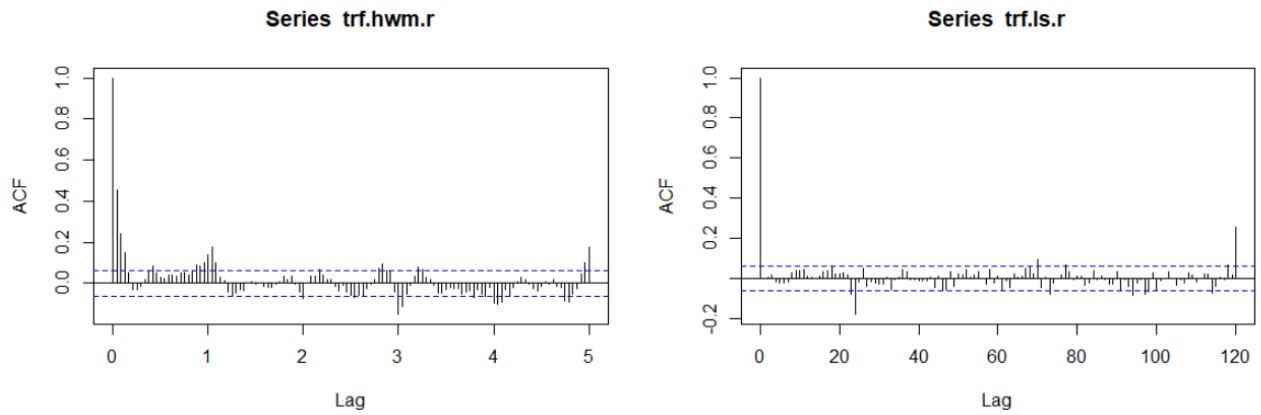


Figura 11: Funzioni di autocorrelazione dei residui. Modello Holt-Winters (sinistra) e minimi quadrati (destra)

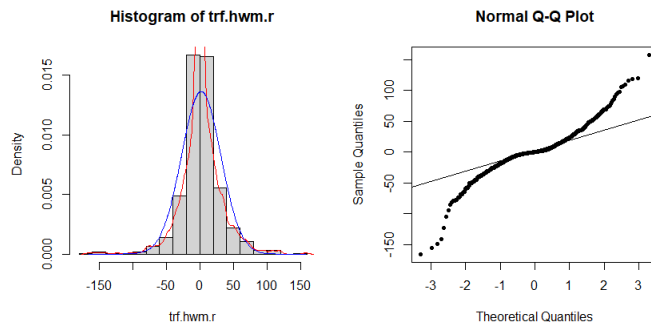


Figura 12: Densità empiriche e QQ-plot residui modello Holt-Winters

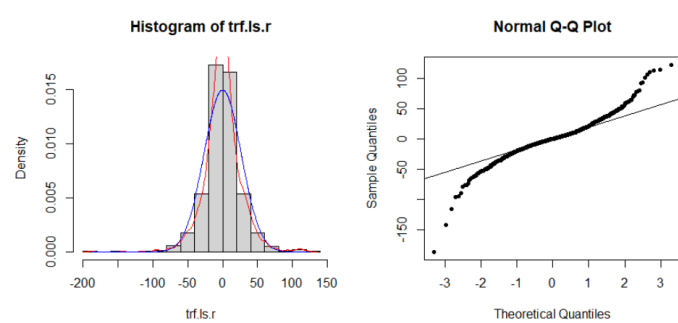


Figura 13: Densità empiriche e QQ-plot residui modello minimi quadrati

Inoltre, è stata calcolata la proporzione di varianza non spiegata, che è del 4,26% per Holt-Winters e del 3,51% per il modello autoregressivo.

Nonostante i residui del modello autoregressivo sono migliori, entrambi sembrano ancora mantenere struttura. Inoltre si può certamente escludere che abbiano una forma simile a quella di una distribuzione gaussiana.

## 4.2 Previsioni

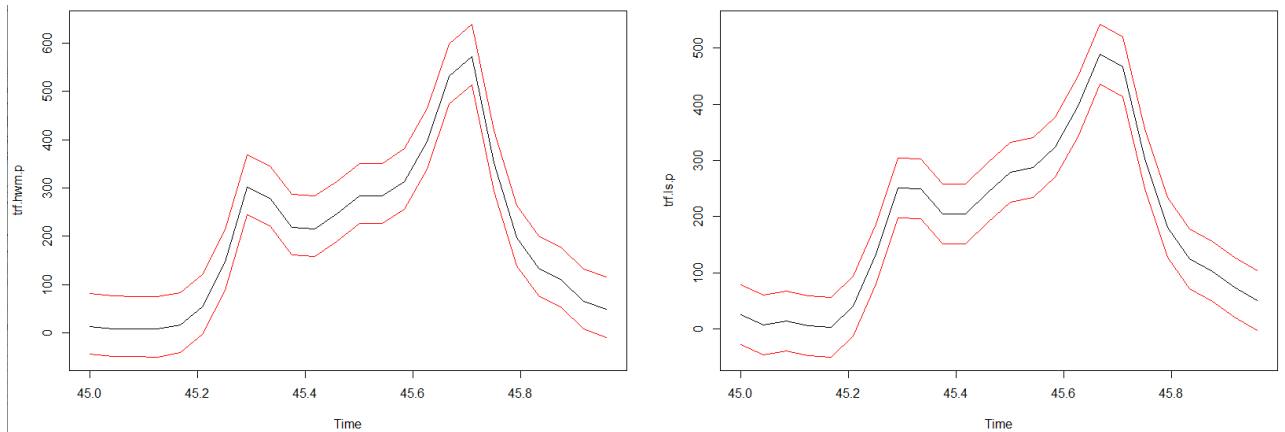
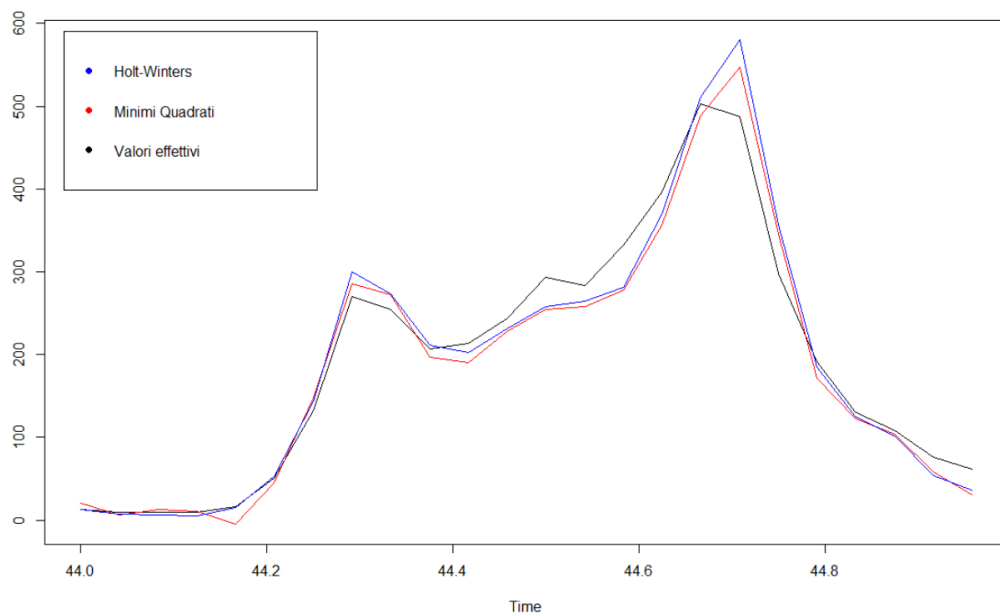


Figura 14: Previsioni e incertezze dei modelli. Holtr-Winters (sinistra) e Minimi Quadrati (destra)

Vediamo nel dettaglio le previsioni dei due modelli, corredate dalle rispettive incertezze calcolate in maniera non parametrica, ovvero in maniera totalmente empirica sulla base dei residui. L'analisi dei residui non ci permette infatti di fare una stima parametrica, data la non somiglianza dei residui con una distribuzione gaussiana.

## 4.3 Autovalutazione

Un'ulteriore prova di confronto tra i due modelli consiste nel valutare la loro capacità di predizione attraverso un'autovalutazione.



```
> sqrt(mean((trf.ls.p - test)^2))  
[1] 26.34998  
> sqrt(mean((trf.hwm.p - test)^2))  
[1] 28.81171
```

Figura 15: errori, il primo del modello autoregressivo, il secondo del modello di Holt-Winters

Guardando i risultati dell'autovalutazione, le capacità di predizione dei modelli differiscono di poco, con il modello autoregressivo che in questo caso sembra fare di poco meglio.

Tuttavia possiamo concludere che non ci sono differenze così nette nei modelli analizzati tali da far scegliere un metodo piuttosto che l'altro.