Analisi dei pernottamenti turistici mensili in Italia.

Mantovani Giacomo

23 dicembre 2019

Indice

1 Introduzione **3**

2 Dati **3**
2.1 Descrizione del set di dati 3
2.2 Preprocessing ed importazione dei dati 3

3 Visualizzazione dei dati **3**
3.1 Decomposizione della serie 4
3.2 Andamento annuale della serie 5

4 Analisi dei dati **5**
4.1 Smorzamento Esponenziale con e senza trend 5
4.2 Predizione 6
4.3 Analisi dei residui 7
4.4 Autovalidazione 8

1 Introduzione

Lo scopo di questo esperimento è quello di analizzare la serie temporale riguardante i pernottamenti dei turisti nei vari stabilimenti in Italia e di prevedere i valori degli anni successivi. A partire dal set di dati abbiamo effettuato il preprocessing utilizzando il software Eclipse (IDE per programmi in linguaggio Java) ed i dati ottenuti da questa fase sono stati analizzati utilizzando il software R confrontando tre modelli di analisi per le serie storiche (Smorzamento esponenziale senza trend, con trend e con trend e stagionalità).

2 Dati

Il dati utilizzati nell'esperimento è stato reperito dal sito https://ec.europa.eu/eurostat/data/database tramite il link: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=tour_occ_nim&lang=en

2.1 Descrizione del set di dati

Il set di dati iniziale presenta al suo interno 6 colonne ed è composto da 357 record, i quali non contengono valori mancanti. di questi 6 attributi abbiamo utilizzato solo l'attributo "Value" che indica il numero totale di pernottamenti nell'intera nazione nel mese considerato. I restanti 5 attributi non sono utili ai fini dell'esperimento pertanto non saranno descritti.

2.2 Preprocessing ed importazione dei dati

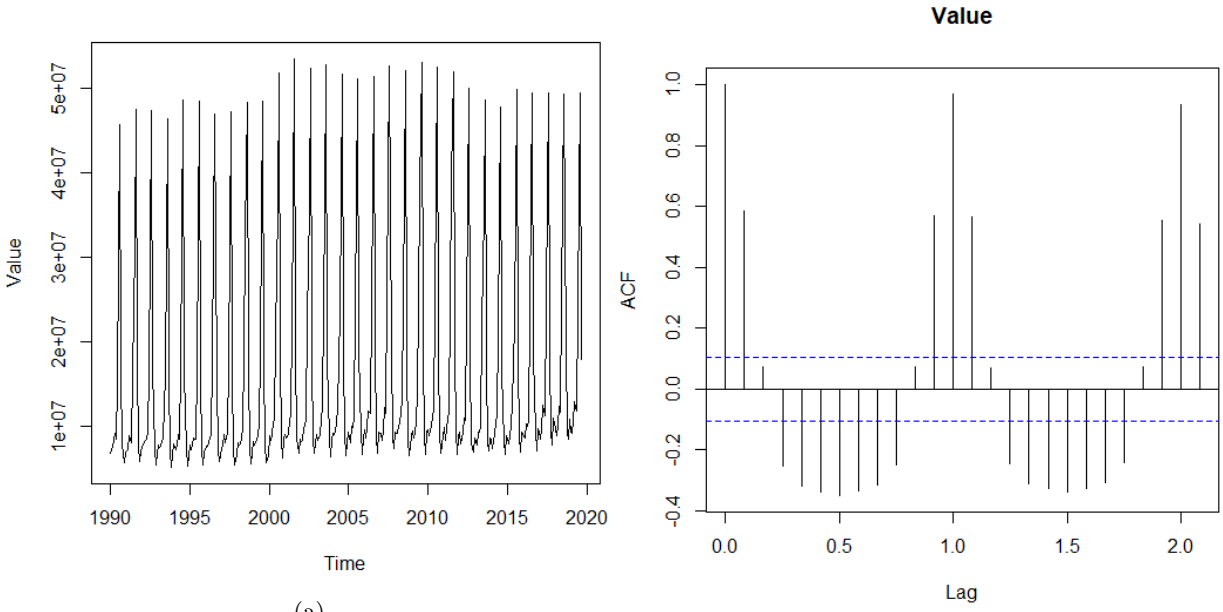
Tramite il software Eclipse abbiamo modificato il set di dati selezionando esclusivamente l'attributo "Value", inoltre da tale valore abbiamo eliminato i separatori delle migliaia (le virgole) ottenendo un numero senza punteggiatura aggiunta in modo da convertire tali valori nel formato previsto dal software R. Successivamente sono stati importati su R per iniziare ad effettuare l'analisi. Di seguito è stato riportato il codice java utilizzato per il preprocessing dei dati.

```
String line = null;
String str = null;
String[] splitted;
File f1 = new File("C:\\Users\\Geghi\\Desktop\\Italy.txt");
File f2 = new File("C:\\Users\\Geghi\\Desktop\\ItalyMonth.txt");

FileReader fr = new FileReader(f1);
BufferedReader br = new BufferedReader(fr);
FileWriter fw = new FileWriter(f2);
BufferedWriter out = new BufferedWriter(fw);
line = br.readLine();
while (line != null) {
        splitted = line.split("\",");
        //elimina tutte le virgolette e tutte le virgole.
        str = splitted[5].replaceAll("\"", "").replaceAll(",", "");
        System.out.println(str);
        out.write(str + "\n");
        line = br.readLine();
}
fr.close();
br.close();
out.flush();
out.close();
```

3 Visualizzazione dei dati

Il primo approccio esplorativo dei dati sarà di tipo grafico. Riportiamo di seguito il grafico dell'andamento della serie temporale (figura 1a) il grafico dell'autocorrelazione (figura 1b) ed il relativo codice. Il grafico dell'autocorrelazione ci suggerisce la presenza di stagionalità nella serie in quanto nei punti diversi dal periodo sono presenti valori negativi e nei punti che coincidono con il periodo abbiamo dei picchi.



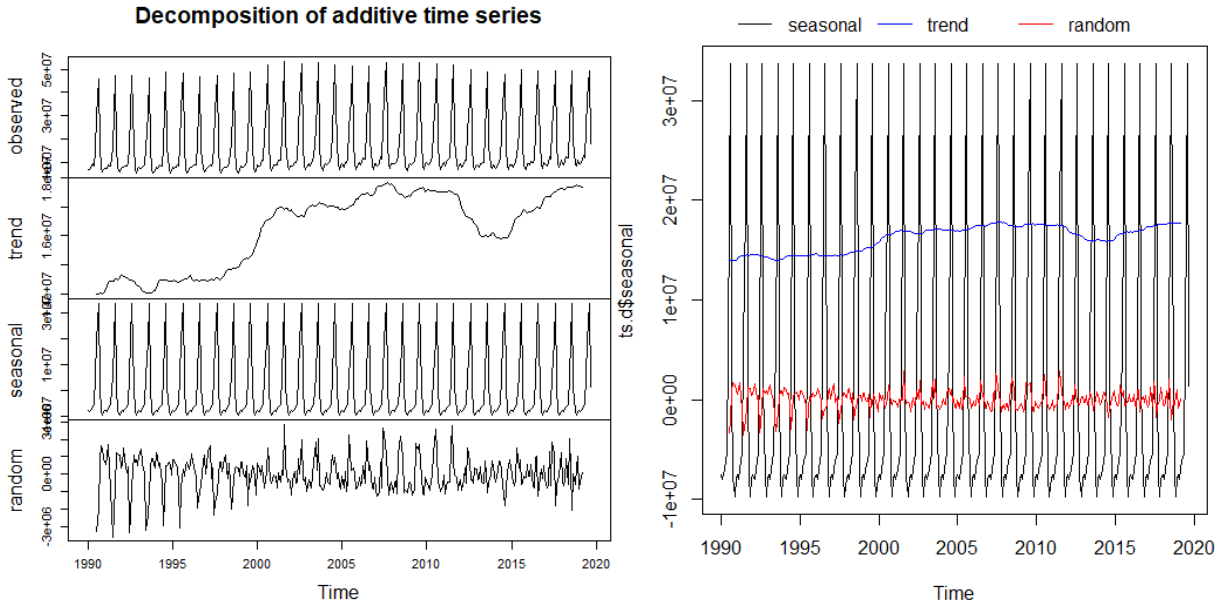
(a) (b)

Figura 1: Andamento della serie temporale (a) e visualizzazione dell'autocorrelazione (b).

```
data = read.csv("tabella.csv")
ts = ts(data, frequency = 12, start = 1990)
plot(ts)
acf(ts)
```

3.1 Decomposizione della serie

Procediamo con la decomposizione della serie, riportata nella (figura 2a) Per una migliore interpretazione visualizziamo i valori del trend , stagionalità ed errore sulla stessa scala (figura 2b). Le informazioni fornite dall'autocorrelazione ci vengono quindi confermate in quanto è presente una forte stagionalità ed un'errore limitato in confronto ad essa. Inoltre, sembra essere presente un trend lievemente crescente ma la serie risulta essere prevalentemente stagionale.

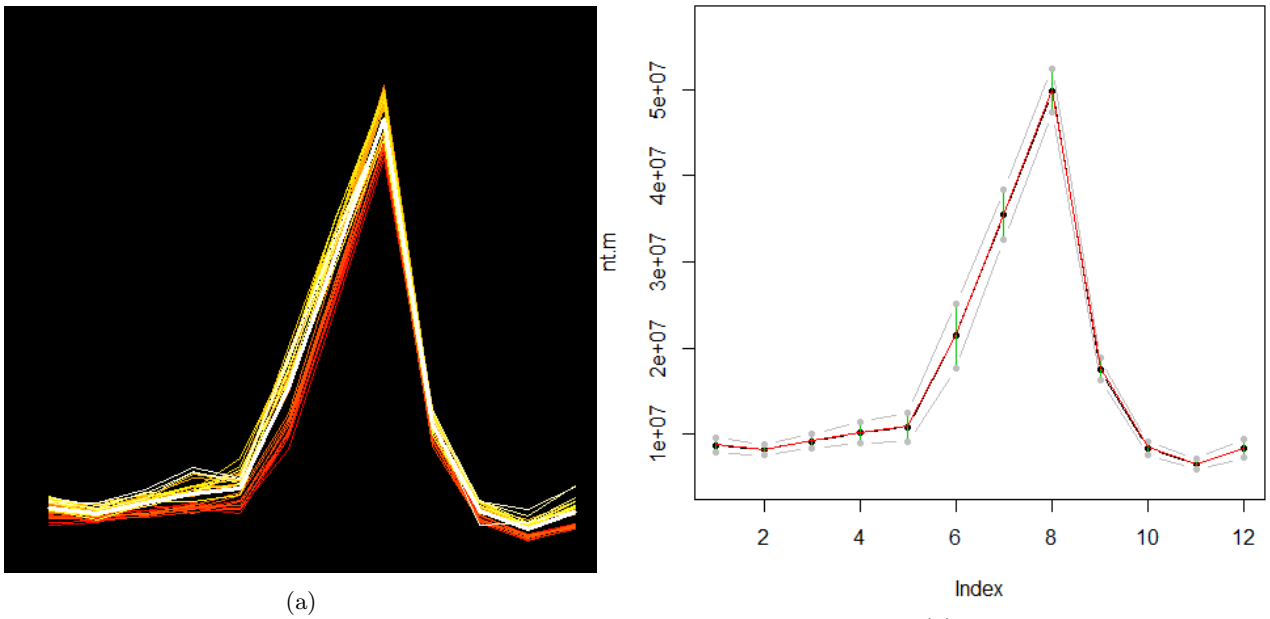


(a) (b)

Figura 2: Visualizzazione grafica della serie decomposta (a) e dei grafici di quest'ultima su stessa scala (b).

3.2 Andamento annuale della serie

Procediamo visualizzando la sovrapposizione dell'andamento della serie nei vari anni come mostrato in (figura 3). Nel corso degli anni la serie mantiene la sua struttura con piccole variazioni.



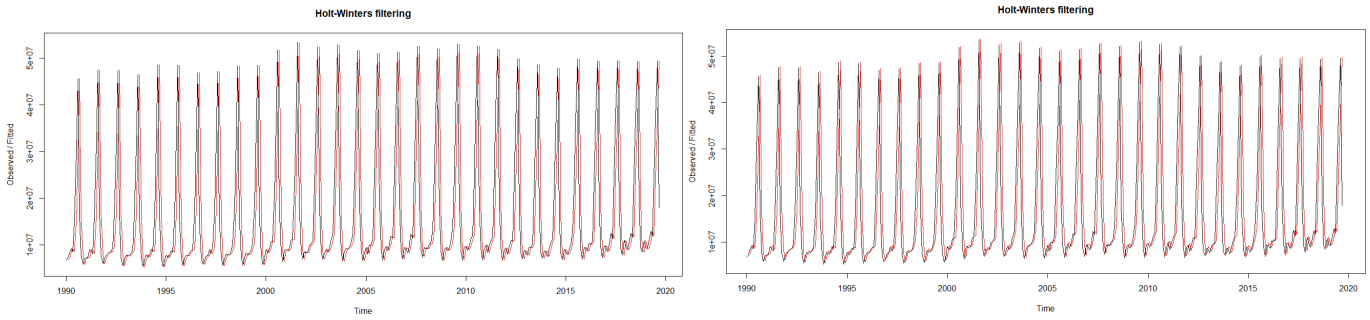
(a) (b)

Figura 3: Sovrapposizione dell'andamento annuale della serie ed andamento medio (a) e confronto di quest'ultimo con bande di confidenza empiriche (b).

4 Analisi dei dati

4.1 Smorzamento Esponenziale con e senza trend

Procediamo analizzando due metodi di analisi delle serie storiche: SE e SET. Entrambi i metodi risultano essere fedeli alla serie senza catturare il trend (figura 4).



(a) SE (b) SET

Figura 4: Smorzamento Esponenziale (a) e Smorzamento Esponenziale con trend (b).

```
ts.se=HoltWinters(ts,beta=F,gamma=F)
plot(ts.se)
ts.set=HoltWinters(ts,gamma=F)
plot(ts.set)
```

Proviamo a modificare le condizioni iniziali per catturare il trend (figura 5).

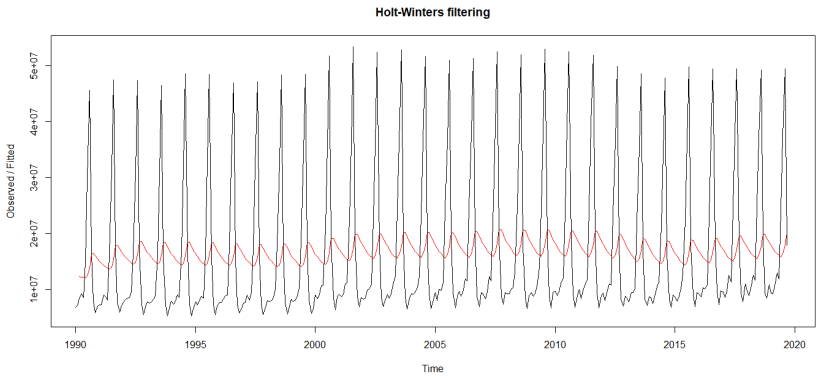
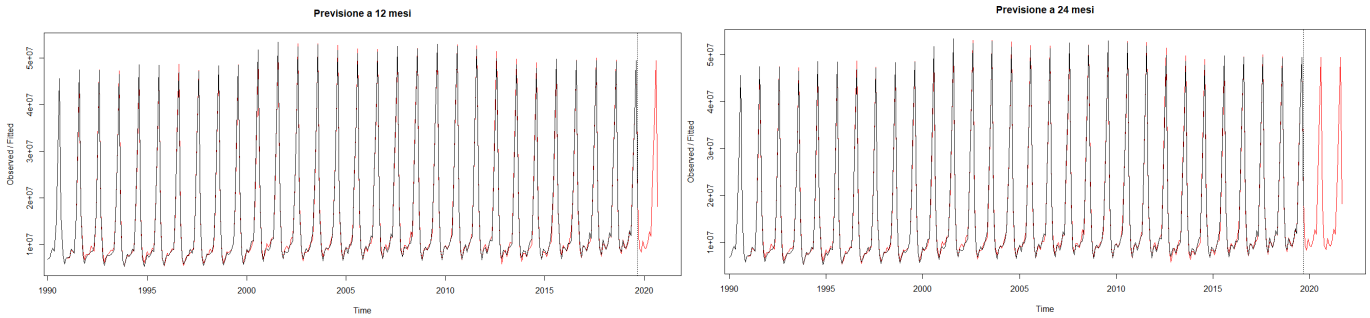


Figura 5: Cattura del trend tramite la funzione di HoltWinters.

```
x=1:24
coefficients(lm(ts[1:24]~x))
plot(HoltWinters(ts,alpha=0.07,gamma=F,l.start=12181677,b.start=161082))
```

4.2 Predizione

In questa fase andremo ad effettuare una predizione dei valori nei prossimi 12 mesi e 24 mesi (figura 6) utilizzando il modello di smorzamento esponenziale con stagionalità e trend.

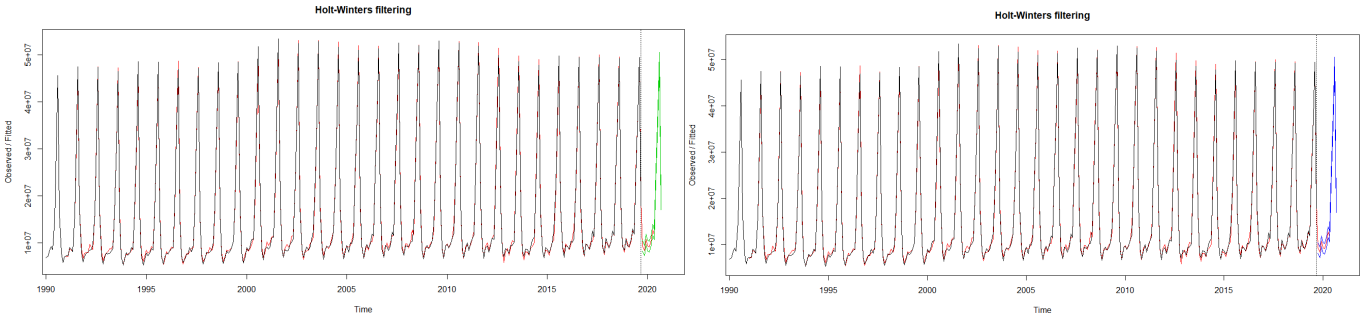


(a) (b)

Figura 6: Predizioni degli anni 2020 (a) e 2020-2021 (b).

```
plot(ts.hw,predict(ts.hw,12),main="Previsione a 12 mesi")
plot(ts.hw,predict(ts.hw,24),main="Previsione a 24 mesi")
```

Andiamo successivamente ad estrarre i residui ed utilizziamoli per visualizzare l'incertezza per via parametrica e non parametrica dei valori predetti come mostrato in (figura 7) ed andiamo a sovrapporli (figura 8). per calcolare l'incertezza per via parametrica il tipo di densità utilizzato è stato il tipo gaussiano, ed abbiamo selezionato la gaussiana con media e deviazione standard empiriche dei residui.



(a) (b)

Figura 7: Incertezza calcolata per via parametrica (a) e non parametrica (b).

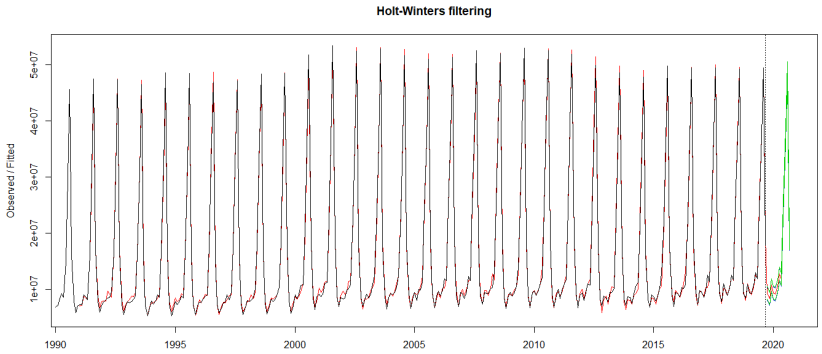


Figura 8: Incertezze sovrapposte.

```
ts.hw.r=residuals(ts.hw)
plot(ts.hw,predict(ts.hw,12))
lines(predict(ts.hw,12)+quantile(ts.hw.r, 0.05),col="green3")
lines(predict(ts.hw,12)+quantile(ts.hw.r, 0.95),col="green3")
```

4.3 Analisi dei residui

Procediamo analizzando i residui andando a verificare se questi mantengono una struttura dei dati. La loro visualizzazione grafica è riportata in figura 9.

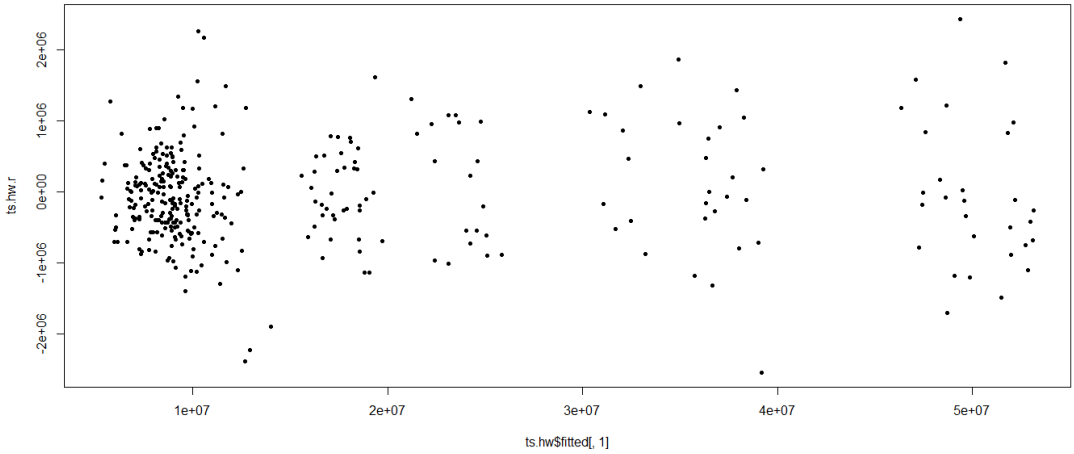


Figura 9: Plot dei residui.

Calcoliamo la varianza spiegata dei residui su finestra comune ottenendo un valore pari a 0.99 (notare che la varianza spiegata nel caso di serie storiche risulta essere un valore indicativo):

```
start(ts)
end(ts)
start(ts.hw.r)
end(ts.hw.r)
1-var(ts.hw.r)/var(window(ts,c(1991,1)))
```

Analizzando l'autocorrelazione dei residui (figura 10) possiamo notare che questi catturano una piccola parte di stagionalità, di conseguenza nelle successive analisi non ci aspetteremo valori completamente casuali, ma dipenderanno in piccola parte dalla struttura della serie.

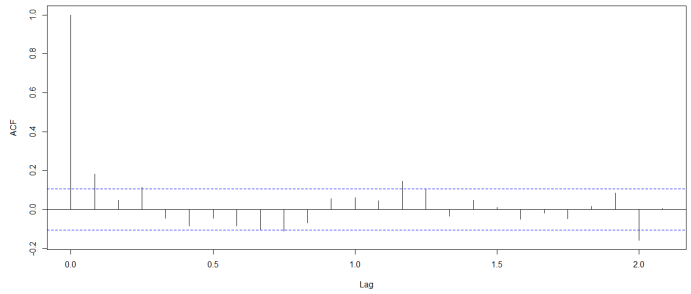
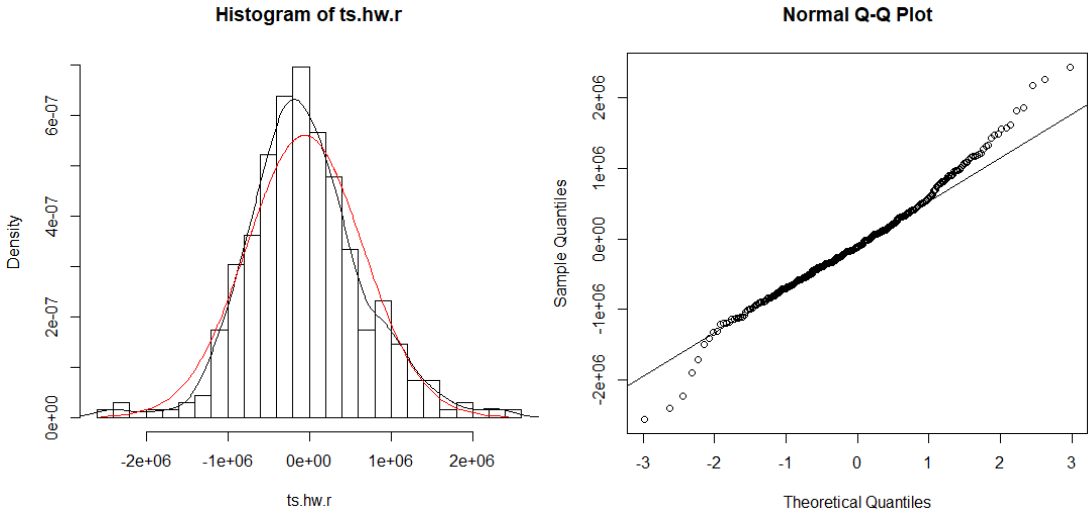


Figura 10: Autocorrelazione dei residui.

Tale ipotesi è confermata dall'analisi della densità e dal quantile-quantile plot dei residui dove possiamo notare che questi non sono completamente casuali ma vi si avvicinano, i valori ottenuti risultano, infatti, buoni nel complesso. Un ulteriore test effettuato è il test di Shapiro-Wilk che presenta un ottimo valore di W pari a 0.98.



(a) (b)

Figura 11: Desità dei residui (a) e Q-Q plot (b).

4.4 Autovalidazione

Nell'ultima fase di questo esperimento andiamo a testare le capacità predittive del metodo di smorzamento esponenziale con trend e stagionalità. A causa della mancanza dei valori futuri con i quali confrontare le previsioni verrà effettuata una procedura di autovalidazione.

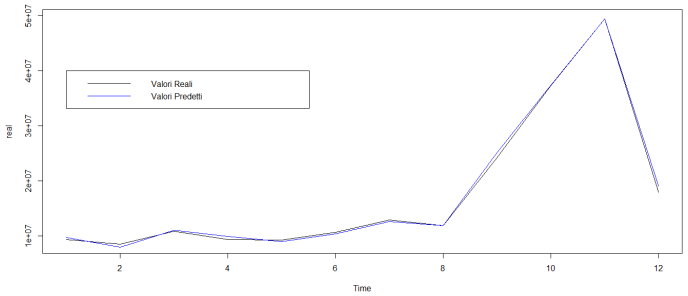


Figura 12: Differenze tra predizioni e valori reali.

L'errore nella previsione è molto limitato (pari a 0.04) e come possiamo vedere dalla figura 12 le previsioni ottenute risultano essere estremamente buone. Possiamo quindi affermare che il modello utilizzato è dotato di un'elevata capacità predittiva.