Università della Calabria

Dipartimento di Economia, Statistica e Finanza "Giovanni Anania"

Corso di

MODELLI E TECNICHE DI PREVISIONE

Elaborato sull'analisi di una serie storica a carattere economico

Studente: Micieli Ottavio

Matricola: 214209

Professore

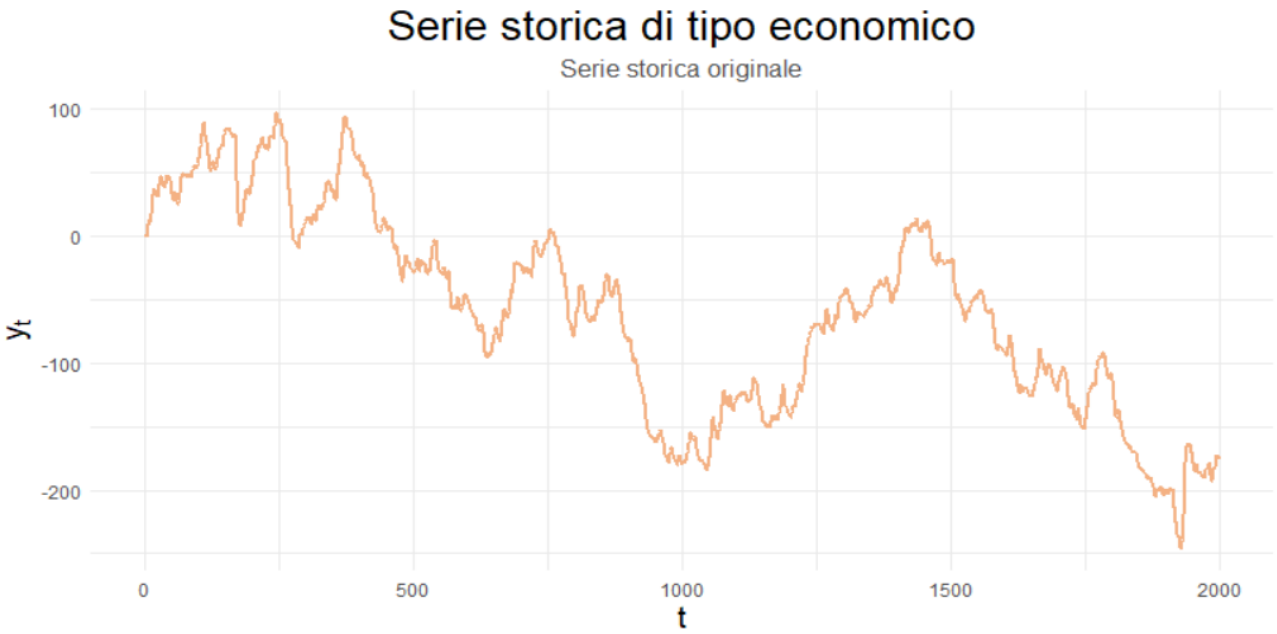
Perri Pier Francesco

Anno accademico 2020/2021

INDICE

1 ANALISI PRELIMINARE DELLA SERIE STORICA

L'obiettivo di tale elaborato è quello di analizzare una serie storica di carattere economico, utilizzando l'approccio di Box e Jenkins, che prevede l'utilizzo di un insieme di modelli, i cosiddetti modelli ARIMA, al fine di definire il processo stocastico che si presume possa aver generato i dati osservati.
Tale analisi condurrà a determinare quale sia il migliore modello in termini di adattamento ai valori osservati, che possa essere anche utilizzato per prevedere i valori futuri del fenomeno oggetto di studio, con un margine di errore il più contenuto possibile.
Prima di procedere con l'individuazione del l'ordine del modello che verosimilmente si presume possa aver generato i dati, si suole condurre un'analisi di tipo descrittivo al fine di individuare le principali caratteristiche distributive del fenomeno. Le misure statistiche comunemente utilizzate vengono corredate da un'analisi di tipo grafico, in cui, in prima battuta, si osserva quale potrebbe essere l'andamento di fondo del fenomeno e se vi è la presenza o meno di una componente di tipo stagionale.
Di seguito si riporta il grafico nel quale si può osservare l'evoluzione del fenomeno oggetto di studio e le stime delle principali statistiche descrittive del processo generatore dei dati, sulla base dei valori che sono stati osservati.

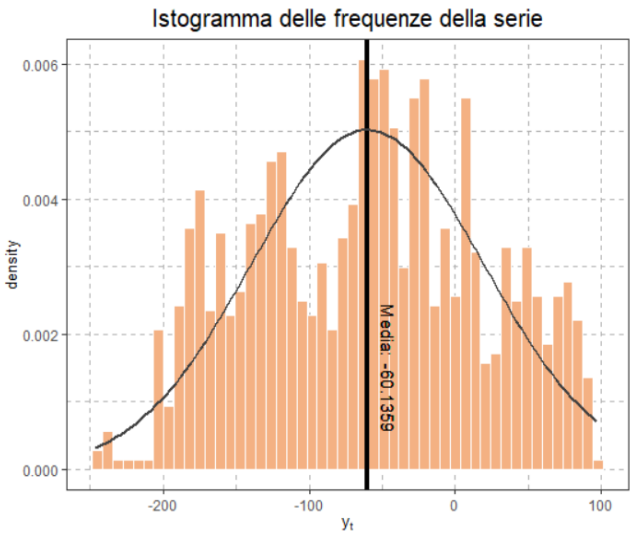


La serie storica è composta da 2001 osservazioni di cui non se ne conosce l'unità temporale di misura.
Il range di variazione dei valori registrati è compreso tra un minimo pari a -246.36 e un massimo pari a 96.71, con un valore medio che si attesta intorno a -60.14. Tale valore risulta anche essere una stima del momento primo del processo generatore dei dati, nel momento in cui si ipotizza che la serie possa essere almeno stazionaria in media. In caso di stazionarietà è difatti possibile sfruttare tutte le informazioni contenute all'interno della serie storica per giungere ad una stima delle statistiche descrittive principali relative al processo che si suppone possa aver generato i dati osservati.

Osservando l'andamento generale del fenomeno si può osservare un declino nel valor medio dello stesso; in particolare, mentre nelle prime 450 osservazioni circa il fenomeno presentava dei valori positivi, in seguito, fino al termine del periodo di osservazione, si sono registrati unicamente valori negativi, ad eccezione di qualche periodo di ridotte dimensioni. Tale aspetto ci porta a dedurre la presenza di una non stazionarietà in media, che dovrà essere trattata con appositi metodi di trasformazione dei dati, in particolar modo attraverso l'uso dell'operatore differenza.
Si ha anche un forte andamento oscillatorio intorno al valor medio, che può far pensare alla presenza di una componente stagionale all'interno della serie oppure alla presenza di eteroschedasticità su cui bisogna indagare.

Nobs	2001
NAs	0
Minimum	-246.361649
Maximum	96.7137736
1. Quartile	-124.598204
3. Quartile	0.766073
Mean	-60.135869
Median	-56.912275
Sum	-120331.874051
SE Mean	1.773451
LCL Mean	-63.613875
UCL Mean	-56.657863
Variance	6293.404950
Stdev	79.330984
Skewness	-0.020123
Kurtosis	-0.888419

Si analizzano, ora, le statistiche di forma, quali l'indice di asimmetria e l'indice di curtosi, incominciando dal primo. L'indice di simmetria è stimato molto prossimo allo 0, pertanto si può affermare che la distribuzione di tale processo sia simmetrica. Tale aspetto viene confermato anche dal test di D'Agostino sull'indice di asimmetria, che presenta un p-value molto al di sopra della soglia prefissata del 5%.
Per quanto riguarda l'indice di curtosi, per precisione la misura inerente all'eccesso di curtosi, essa viene stimata essere pari a -0.888419. Si accerta, quindi, la presenza di una distribuzione platicurtica, aspetto che viene confermato anche dal test di Anscombe-Glynn. Tali aspetti sulla distribuzione di probabilità del processo generatore dei dati possono anche osservati tramite l'istogramma delle densità di probabilità e il QQ plot riportati di seguito.



Tuttavia, il fatto che la distribuzione si stimi essere simmetrica non ci garantisce che il processo generatore dei dati abbia un andamento gaussiano; infatti, l'allontanamento dal un valore pari a 3 dell'indice di curtosi ci porta a dedurre di non avere di fronte una distribuzione normale. Tale considerazione è supportata da una serie di test statistici utili a valutare l'allontanamento (o avvicinamento) che è presente tra una distribuzione normale e la distribuzione di probabilità che è possibile dedurre dai dati osservati. I risultati di tali test statistici sono riportati qui di seguito.

```
Kolmogorov-Smirnov test
Statistic: 0.7419
P value two-sided: <2.2e-16

Shapiro-Wilk Normality Test
Statistic: 0.9787
P value:  <2.2e-16

Jarque-Bera Normality Test
Statistic: 65.6296
Asymptotic p value: 5.662e-15
```

L'allontanamento dalla normalità della distribuzione di probabilità dei dati si può anche osservare nel Q-Q plot, in maniera particolare per quanto riguarda le osservazioni presenti sulle code.
Tuttavia, il fatto che la distribuzione che si cela dietro le informazioni in nostro possesso non sia significativamente equiparabile ad una distribuzione gaussiana non comporta dei problemi in termini di individuazione del miglior modello e, soprattutto, di stima dei parametri incogniti. L'allontanamento dalla normalità ha effetto unicamente sulla precisione delle stime che si ottengono dei parametri, i cui stimatori presenteranno un livello di efficienza minore rispetto a quelli che si sarebbero potuti ottenere in presenza di normalità all'interno dei dati, oppure nel momento in cui si conosca la vera distribuzione di probabilità associata al processo stocastico generatore delle osservazioni presenti nella serie storica.

2 ANALISI SULLA STAZIONARIETÀ E SULLA COMPONENTE STAGIONALE

2.1 Stazionarietà del processo

Il passo successivo consiste nel valutare la presenza o meno di stazionarietà, almeno in senso debole, all'interno della serie storica osservata. Qualora risulti che la serie storica sia non stazionaria, vi è la necessità di manipolare opportunatamente i dati al fine di renderla tale. Una serie non stazionaria risulta non essere trattabile attraverso l'approccio classico di Box & Jenkins.
La valutazione della stazionarietà all'interno della serie storica avviene attraverso l'uso sia di strumenti grafici, quali l'osservazione del grafico relativo alla stima della funzione di autocorrelazione, ma anche di un insieme di test statistici che prendono il nome di test per radici unitarie. Tra i test statistici più utilizzati vi è l'Augmented Dickey-Fuller (ADF) test, che pone sotto l'ipotesi nulla la non stazionarietà della serie, individuata come presenza di radici unitarie all'interno delle soluzioni del polinomio caratteristico. L'ipotesi alternativa può essere scelta arbitrariamente, optando tra l'ipotesi di avere una serie stazionaria oppure l'ipotesi che si abbia esplosività all'interno della serie storica. I risultati che si ottengono dall'applicazione dell'ADF test sono riportati qui di seguito:

```
Augmented Dickey-Fuller Test
Alternative hypothesis: stationary
Dickey-Fuller statistic: -2.4189
p-value: 0.401
```

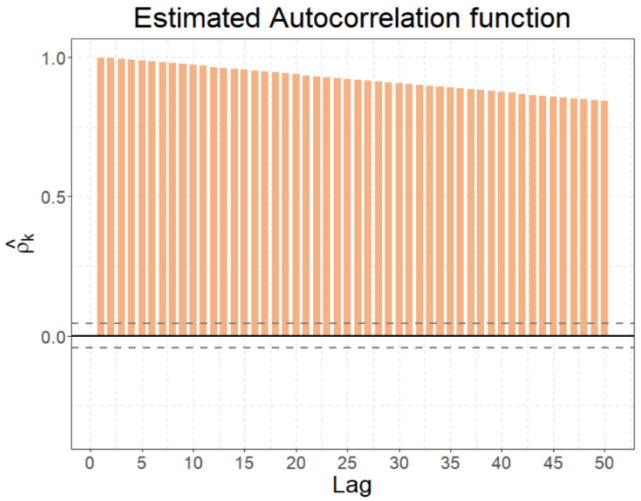
```
Augmented Dickey-Fuller Test
Alternative hypothesis: explosive
Dickey-Fuller statistic: -2.4189
p-value: 0.599
```

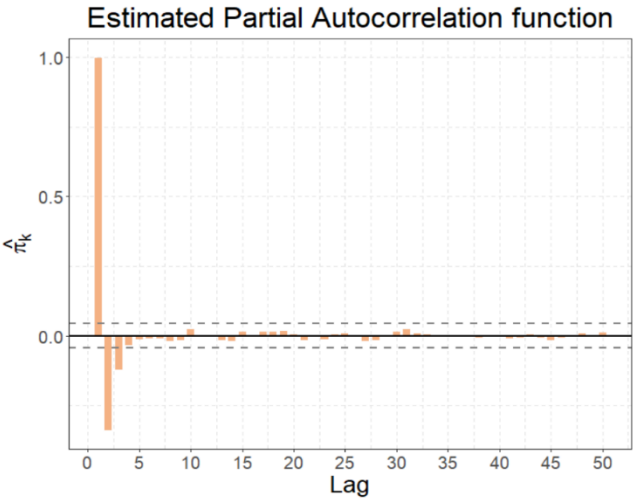
Sia per quanto riguarda il test con ipotesi alternativa la presenza di stazionarietà all'interno della serie, che per quello in cui l'ipotesi alternativa risulta essere l'esplosività della serie, il livello del p-value non ci consente di rifiutare l'ipotesi. Pertanto, la serie risulta essere non stazionaria, ma comunque non si ha un processo esplosivo, il quale non potrebbe essere trattato con la metodologia statistica relativa all'approccio di Box & Jenkins. Altri test statistici sono stati condotti per completezza di trattazione, quale il test di Phillips-Perron per radici unitarie, il quale prevede una formulazione in termini di sistema d'ipotesi identica al test di Dickey-Fuller, e il test di Kwiatkowski, Phillips, Schmidt e Shin (KPSS test), il quale a differenza dei due precedenti presenta sotto l'ipotesi nulla la stazionarietà della serie storica, mentre l' ipotesi alternativa prevede la non stazionarietà, senza

specificare se essa sia esplosiva o meno. Anche questi ultimi due test statistici confermano la non stazionarietà della serie storica.

```
KPSS Test for Level Stationarity
KPSS statistic: 12.818
p-value: 0.01

Phillips-Perron Unit Root Test
Alternative hypothesis: stationary
PP statistics: -9.669
p-value: 0.5704

Phillips-Perron Unit Root Test
Alternative hypothesis: explosive
PP statistics: -9.669
p-value: 0.4296
```

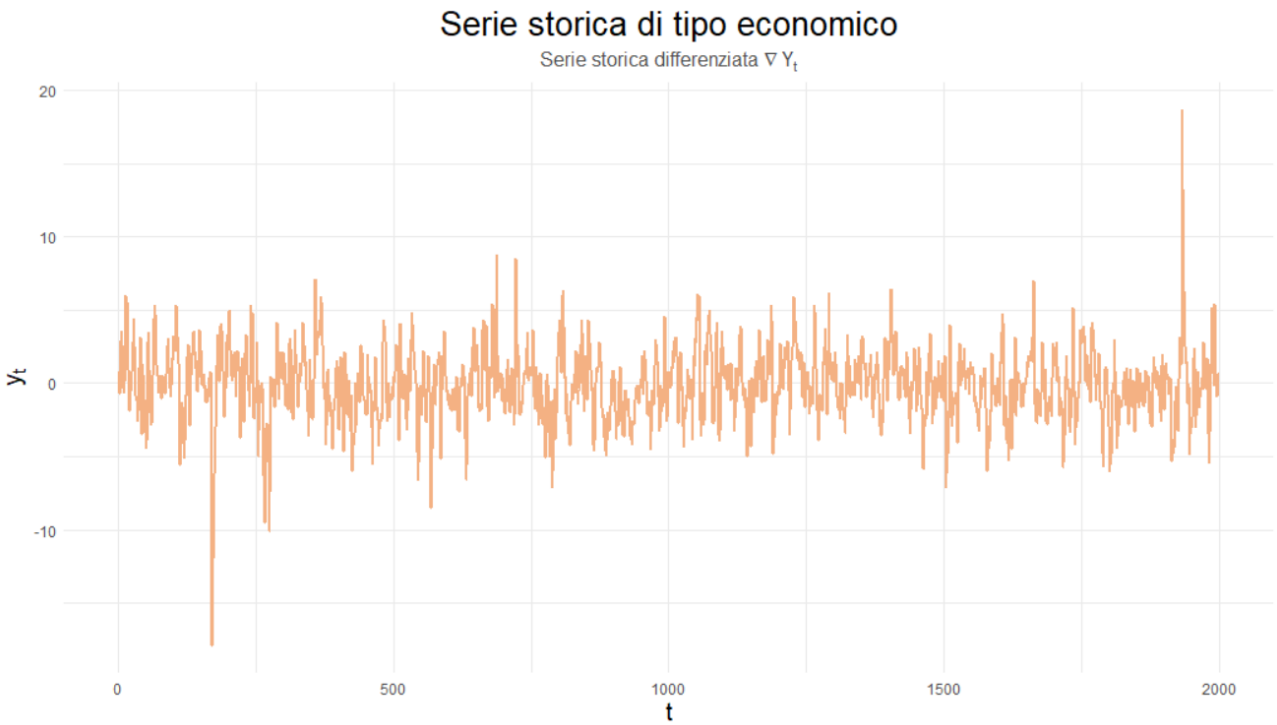
Come detto in precedenza, la non stazionarietà può essere osservata anche attraverso l'utilizzo del grafico riportante la funzione di autocorrelazione stimata per un numero prefissato di lag.





Come si può notare dal grafico soprastante, il valore assunto dalla funzione di autocorrelazione decade molto lentamente con l'aumentare del lag considerato. Questo è sintomo di presenza di non stazionarietà all'interno della serie storica.
Si procede, pertanto, ad applicare l'operatore differenza sulla serie storica originale al fine di cercare di raggiungere la stazionarietà all'interno del processo.
Sulla nuova serie storica che si è generata viene testata la presenza di radici unitarie sfruttando i medesimi test statistici visti in precedenza. Di seguito vengono riportati i risultati per tali test d'ipotesi, i quali p-value ci permettono di accertare che la nuova serie storica originata sia stazionaria.

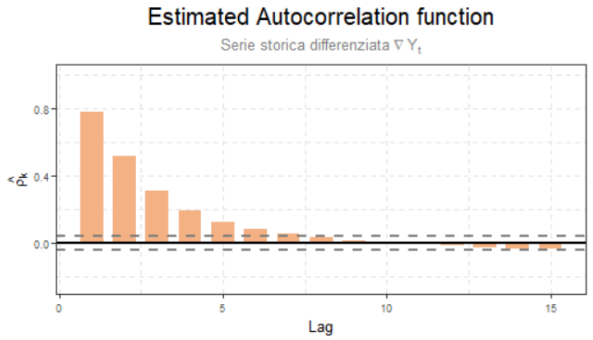
```
Augmented Dickey-Fuller Test
Alternative hypothesis: stationary
p-value: 0.01

Augmented Dickey-Fuller Test
Alternative hypothesis: explosive
p-value: 0.99

KPSS Test for Level Stationarity
p-value: 0.1

Phillips-Perron Unit Root Test
Alternative hypothesis: stationary
p-value: 0.01

Phillips-Perron Unit Root Test
Alternative hypothesis: explosive
p-value: 0.99
```



2.2 Analisi sulla componente stagionale

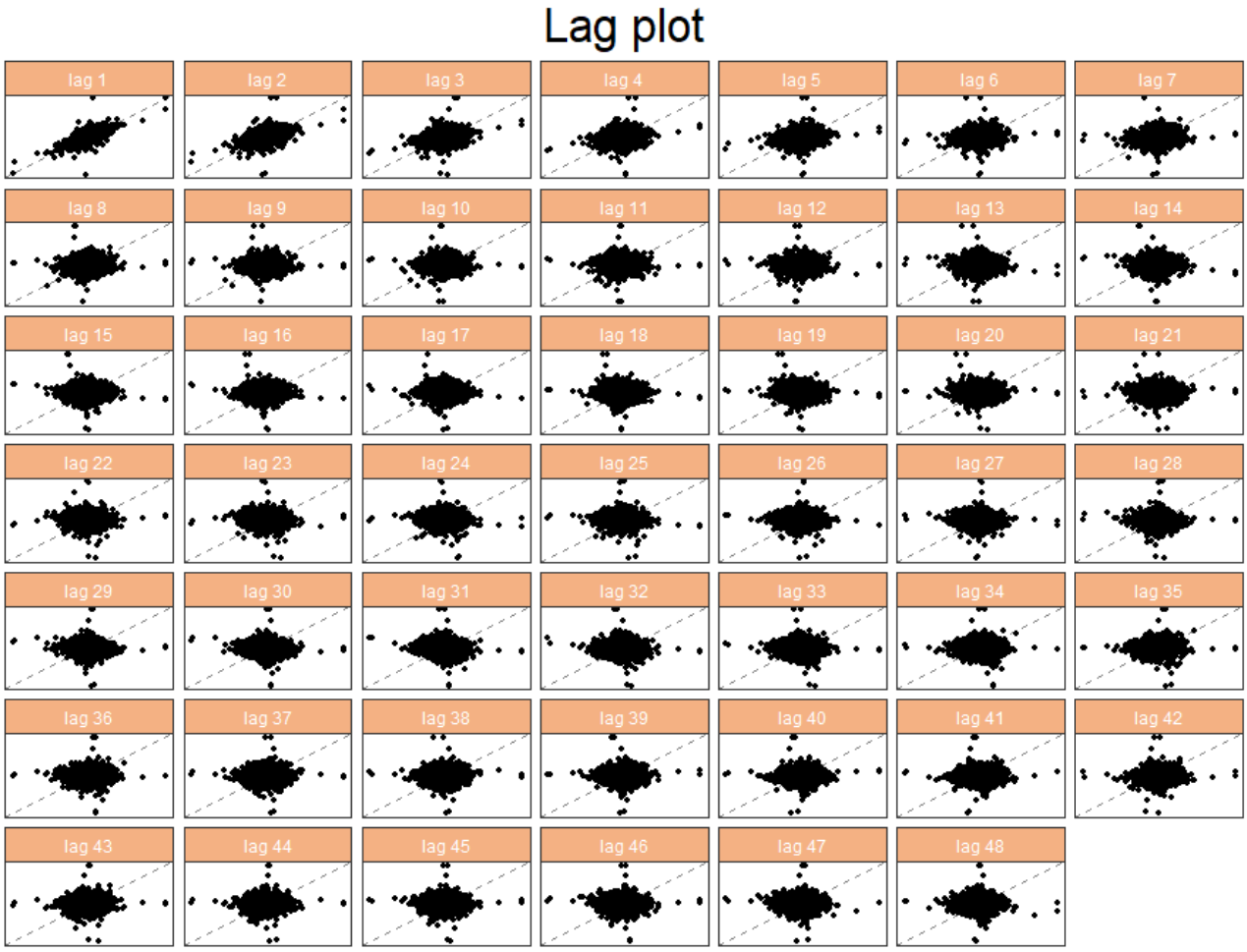
Affrontiamo, in questo paragrafo, il problema inerente alla presenza o meno di una componente stagionale all'interno della serie storica osservata.
Da un punto di vista pratico, tale problema può essere trattato attraverso l'utilizzo di alcuni strumenti grafici, quali il grafico in cui si riportano i valori assunti dalla serie storica differenziata in ordine temporale (riportato sopra), il grafico in cui vengono riporti i valori stimati della funzione di autocorrelazione e il lag plot.
Dal grafico riportante i valori della serie storica differenziata rispetto al tempo notiamo la presenza di movimenti oscillatori intorno al valor medio, che possono far pensare alla presenza di una componente stagionale. Bisogna notare la presenza di alcuni valori, in particolare due, che sembrano essere anomali rispetto agli altri. A tali valori bisogna prestare le dovute attenzioni affinché non compromettano l'individuazione del modello migliore per descrivere l'andamento della serie storica e su cui basarci per prevedere lo sviluppo futuro del fenomeno.

L'ipotesi di stagionalità all'interno della serie storica osservata non è supportata, tuttavia, dai valori stimati della funzione di autocorrelazione, che decadono piuttosto velocemente fino ad annullarsi significativamente per un lag pari a 8.



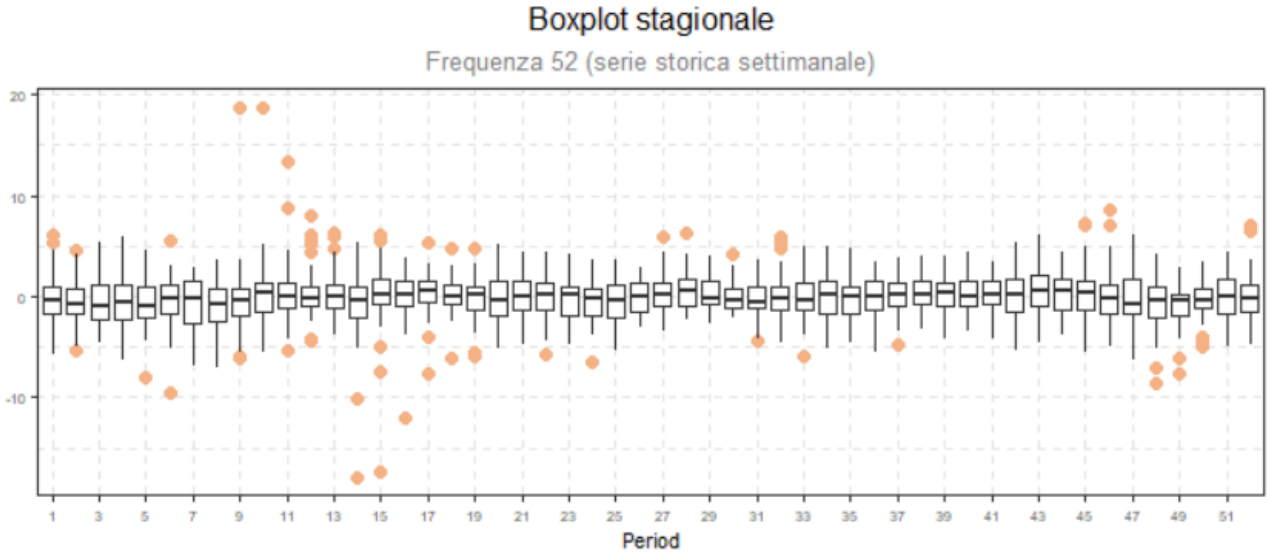
In presenza di stagionalità si osserverebbe, infatti, che i valori della funzione di autocorrelazione tendono a ripetersi, mentre in tale situazione ciò non accade. Anche per quanto riguarda lo strumento grafico del lag plot, esso non fornisce informazioni utili circa una possibile stagionalità presente all'interno della serie.

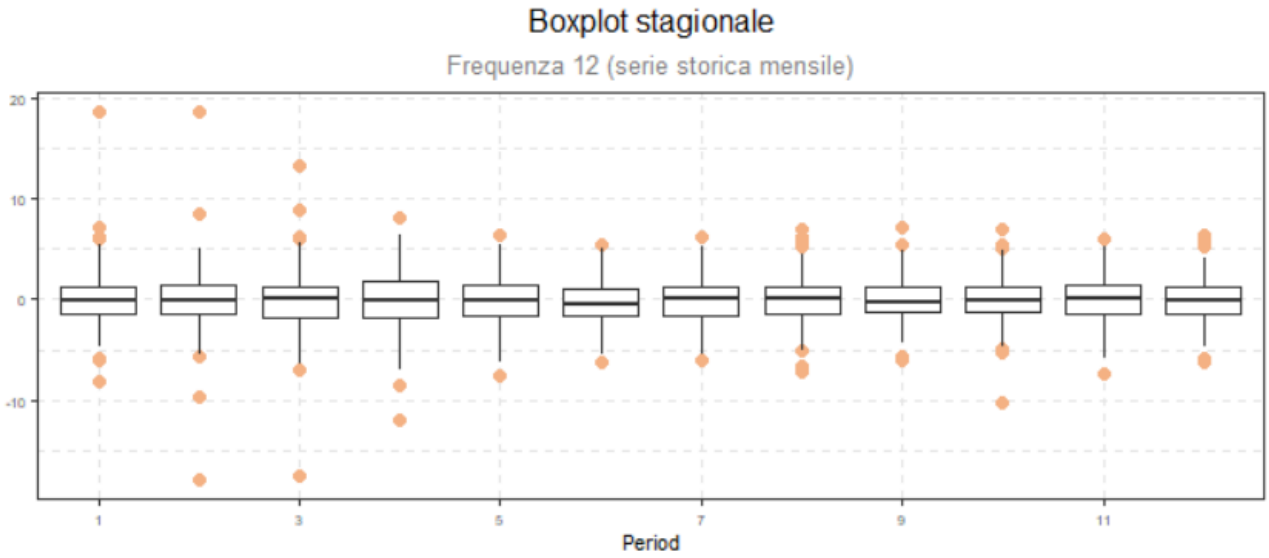
Tranne per quanto riguarda i primi due lag, i grafici che vengono riportati di seguito non riescono ad evidenziare qualche forma di dipendenza tra il valore al generico tempo t e il valore assunto dalla serie al tempo $t + k$, dove k e il lag che si sta considerando.

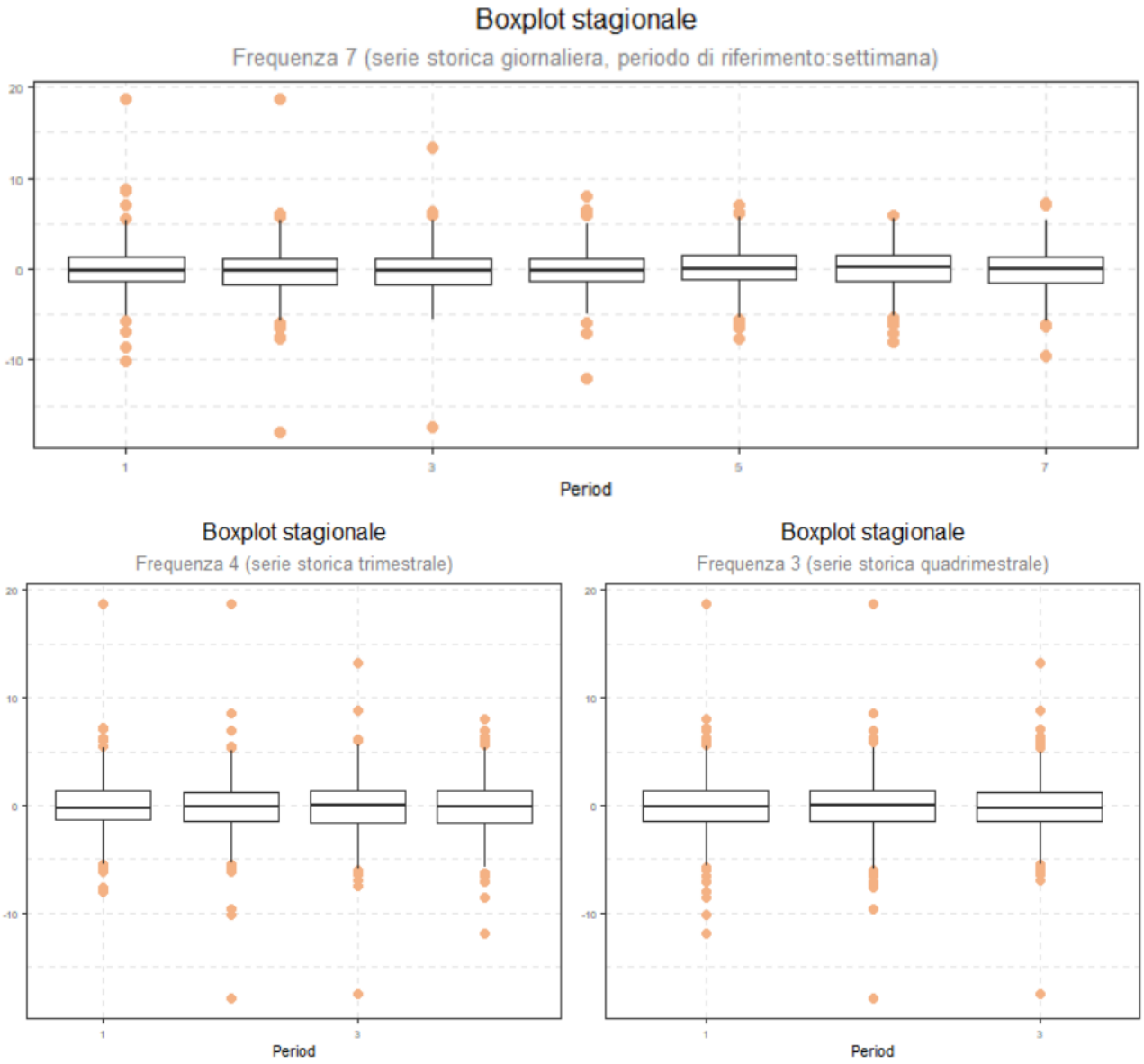


Per maggiore scrupolo, sono stati condotti anche alcuni test statistici, riportati all'interno del pacchetto del software R `seastests`, che hanno lo scopo proprio di testare la presenza di stagionalità all'interno della serie storica. In maniera particolare, sono stati adoperati il test di Webel-Ollech e il test di Kruskall-Wallis. Utilizzando direttamente il comando `isSeasonal` e specificando di volta in volta il tipo di test statistico da utilizzare e il lag da considerare ci viene riferito se la serie presenta una componente stagionale o meno. Sono stati considerate tutte le principali tipologie di stagionalità che possono essere rinvenute all'interno di una serie storica quali una stagionalità mensile (ponendo la frequenza pari a 12), una stagionalità trimestrale (ponendo la frequenza pari a 4) e una stagionalità quadrimestrale (ponendo la frequenza pari a 3). Sono, inoltre, state considerate stagionalità di tipo settimanale, considerando come unità di tempo principale l'anno e ipotizzando la presenza di 52 settimane all'interno di ogni unità, e stagionalità di tipo giornaliero, considerando come unità di tempo la settimana.
Tutti i test hanno comunque fornito un esito negativo, pertanto si può concludere l'assenza di stagionalità all'interno della serie storica osservata.

Per scrupolo è stato utilizzato un altro strumento grafico, che prevede l'utilizzo della tipologia di grafico che prende il nome di Boxplot. Quello che si è fatto è stato raggruppare tutte quelle osservazioni equidistanti tra di loro per un lag prefissato e considerarle come appartenenti ad un unico gruppo. Ad esempio, considerando una stagionalità con una frequenza pari a 12, ossia ipotizzando di avere in possesso una serie storica di tipo mensile, quello che si fa e raggruppare le osservazioni registrato al tempo $1, 1+k, 1+k\times 2, 1+k\times 3, \dots$ nel primo gruppo, le osservazioni $2, 2+k, 2+k\times 2, 2+k\times 3, \dots$ nel secondo gruppo e così via fino ad ottenere il dodicesimo gruppo con osservazioni $12, 12+k, 12+k\times 2, 12+k\times 3, \dots$. In tal modo, in presenza di stagionalità, le osservazioni che si ipotizzano essere correlate verticalmente apparteranno al medesimo gruppo. Fatto ciò, si procede a creare i boxplot per ciascun gruppo, in modo da individuarne il valor medio e la distribuzione degli stessi.
Da tale procedimento si ottiene ciò che è riportato qui di seguito.





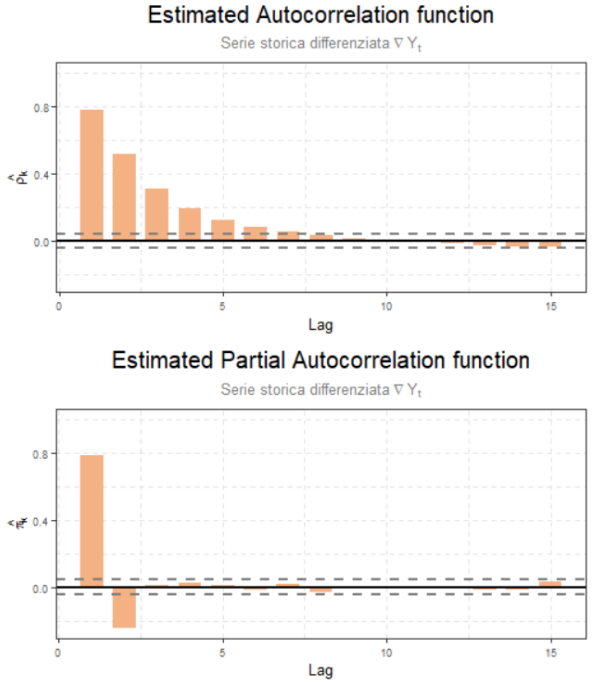


In tutti i grafici riportati non si evidenzia una qualche differenza significativa del valor medio in ciascun gruppo, che risulta in tutti i casi essere circa pari a 0, qualsiasi sia la frequenza considerata. Un qualche andamento può essere visibile se si considera una serie storica settimanale, ma tale andamento risulta essere abbastanza erratico per definire una qualche forma di stagionalità.
Forti del fatto che i test per l'individuazione della stagionalità diano esito negativo, possiamo concludere che non è presente una componente stagionale all'interno della serie storica osservato. Pertanto, non vi è la necessità di ricorrere ai cosiddetti modelli SARIMA, poiché non vi è nessuna correlazione di tipo verticale all'interno dei dati della serie storica.

3 INDIVIDUAZIONE DEL MIGLIOR MODELLO E STIMA DEI PARAMETRI

3.1 Considerazioni su ACF e PACF

Dopo aver reso stazionaria la serie storica che si sta analizzando e accerta l'assenza di stagionalità all'interno della stessa, quello che bisogna fare è individuare la famiglia di modelli che può essere sfruttata in base alle informazioni in nostro possesso. Come già detto nel capitolo precedente, non occorre utilizzare un modello SARIMA, poiché all'interno della serie storica non compare alcuna componente stagionale.
Pertanto, la serie storica potrà essere modellata attraverso l'utilizzo di un modello ARIMA, o meglio, si modella la serie storica differenziata di un periodo attraverso un modello ARMA.
Il passo successivo consiste, quindi, nell'individuazione dell'ordine del modello ARMA da utilizzare al fine di modellare la serie storica differenziata, indicata da ∇Y_t. Per agevolare tale compito si possono sfruttare le informazioni che si ottengono dalla stima della funzione di autocorrelazione sia totale che parziale. Di fianco viene riportato il correlogramma utilizzato per la visualizzazione grafica di ciascuna delle due funzioni.
Per quanto riguarda la stima della funzione di autocorrelazione globale, essa ci fornisce indicazioni circa l'ordine della parte a media mobile del modello ARMA. In particolare, possiamo notare come



essa si annulli in maniera significativa per un valore del lag pari a 8, per cui questo ci suggerisce che i parametri relativi alla componente a media mobile siano 7. In altri termini il valore del parametro q si ipotizza essere pari a 7. Tale valore risulta essere alquanto eccessivo, per cui si ipotizza che in realtà occorrono un numero inferiore di parametri.
Per quanto riguarda, invece, il numero di parametri della componente autoregressiva, dal grafico della funzione di autocorrelazione parziale possiamo osservare come essa si annulli per un lag pari a 3 quindi, verosimilmente, il numero di parametri per quando riguarda tale componente saranno pari a 2.
Ragioniamo ora su quest'ultimo aspetto: se risultasse vero che il valore assunto dalla serie in un determinato periodo

risultasse essere in relazioni con i due valori assunti in passato, allora tale relazione presente si ripercuoterebbe anche sui valori anche più passati della serie. In presenza di un modello autoregressivo di secondo ordine è ragionevole aspettarci che il valore della funzione di autocorrelazione decada dopo un certo numero di lag, ma che non si annulli del tutto. Allo stesso modo con cui Y_t è messo in relazione con Y_{t-1} e Y_{t-2}, allora si avrà anche che si conseguenza Y_{t-1} è messo in relazione con Y_{t-2} e Y_{t-3}, e così di seguito a ritroso. Pertanto, si avrà anche che, seppur in misura sempre minore, Y_t sia sempre collegato con la sua storia passata.
Per tale ragionamento, si ipotizza che il modello che possa verosimilmente adattarsi con maggiore aderenza ai dati osservati possa essere un modello autoregressivo del secondo ordine. Tuttavia, per il momento si accantona tale ipotesi e si ricerca, tra i vari modelli che si possono ottenere modificando l'ordine delle due componenti, quello che possa adattare meglio alla serie storica differenziata di un periodo.

3.2 Individuazione del miglior modello

All'interno del pacchetto `forecast` è possibile rinvenire la funzione `auto.arima`, che può essere utilizzata per avere una prima idea su quello che può essere l'ordine del modello e, in base a quest'ultimo, stimare il valore dei parametri incogniti del stesso. Tale funzione può essere applicata in maniera indifferente alla serie storica differenziata, e quindi stazionaria, che alla serie storica originale, in quanto in maniera automatica considera la presenza di non stazionarietà all'interno dei dati. I risultati dell'applicazione di tale funzione alla serie storica originale sono di seguito riportati. Si riportano anche gli intervalli di confidenza per un livello di fiducia pari al 95% dei parametri stimati insieme ai test statistici di significatività degli stessi.

```
ARIMA (1,1,2)
Coefficients:
      Estimate  Std. Error   Pr(>|z|)
ar1  0.610027  0.031751  <2.2e-16
ma1  0.363733  0.036872  <2.2e-16
ma2  0.109366  0.033044  0.000934

          2.5%      97.5%
ar1  0.547797  0.672257
ma1  0.291464  0.436002
ma2  0.044602  0.174131

Sigma^2 estimated as 2.256
log likelihood=-3650.48
AIC=7308.96     AICc=7308.98
BIC=7331.36
```

In termini dell'equazione che specifica un modello ARIMA, il modello stimato può essere scritto come

$$\nabla(1 - 0.61(B))Y_t = (1 + 0.36(B) + 0.11(B^2))\varepsilon_t$$

Ricordando che il modello specificato nella funzione `arima` presenta i parametri della componente a media mobile con segno positivo, mentre da un punto di vista teorico tali parametri sono stati presentati con segno negativo, per cui vi è la necessità di cambiare il segno dei valori stimati. Risolvendo rispetto a Y_t l'equazione precedente si ha

$$Y_t = 1.61Y_{t-1} - 0.61Y_{t-2} + \varepsilon_t + 0.36\varepsilon_{t-1} + 0.11\varepsilon_{t-2}$$

Tale funzione implementata su R, tuttavia, non è sempre precisa e si preferisce indagare in maniera approfondita, vagliando differenti modelli attraverso un'apposita funzione costruita per far ciò. Tale algoritmo iterativo, costruito appositamente per essere sfruttato in tale contesto, sfrutta, in maniera indipendente ma del tutto equivalente, la funzione `arima` presente all'interno del pacchetto base di R, oppure la funzione `Arima`, implementata invece nel pacchetto `forecast`. Tale funzione autoprodotta stima differenti modelli ARIMA, cambiando ad ogni iterazione l'ordine dei parametri, considerando le diverse combinazioni che si originano alternando differenti valori per il parametro p e il parametro q. Dalle differenti combinazioni si ottengono 36 differenti modelli da stimare e, quindi, da confrontare. La funzione iterativa fornisce come output finale una tabella in cui vengono ordinati i modelli stimati in base al valore dell'AIC, insieme all'ordine individuato per tale modello sia per quanto riguarda la componente autoregressiva, che per la componente a media mobile. Viene anche proposto l'indice BIC. I primi sei modelli in ordine crescente del livello di AIC sono riportati qui di seguito.

	p	d	q	AIC	BIC
1	2	1	0	7308.143	7324.946
2	1	1	2	7308.958	7331.362
3	2	1	3	7309.775	7343.380
4	3	1	0	7309.949	7332.353
5	2	1	1	7309.986	7332.389
6	4	1	0	7310.355	7338.359

Il modello che migliore sia in termini di AIC che di BIC risulta essere un processo ARIMA(2,1,0), di cui ne stimiamo i parametri incogniti, ottenendo le seguenti informazioni

```
ARIMA (2,1,0)
Coefficients:
      Estimate  Std. Error   Pr(>|z|)
ar1   0.97266   0.021674   <2.2e-16
ar2   -0.24425  0.021680   <2.2e-16

          2.5%      97.5%
ar1   0.93018   1.01514
ar2  -0.28675  -0.20176

Sigma^2 estimated as 2.256
log likelihood=-3651.07
AIC=7308.14     AICc=7308.16
BIC=7324.95
```

In termini dell'equazione che specifica il valore assunto da Y_t, il modello può essere espresso come

$$\nabla(1 - 0.97(B) + 0.29(B^2))Y_t = \varepsilon_t$$

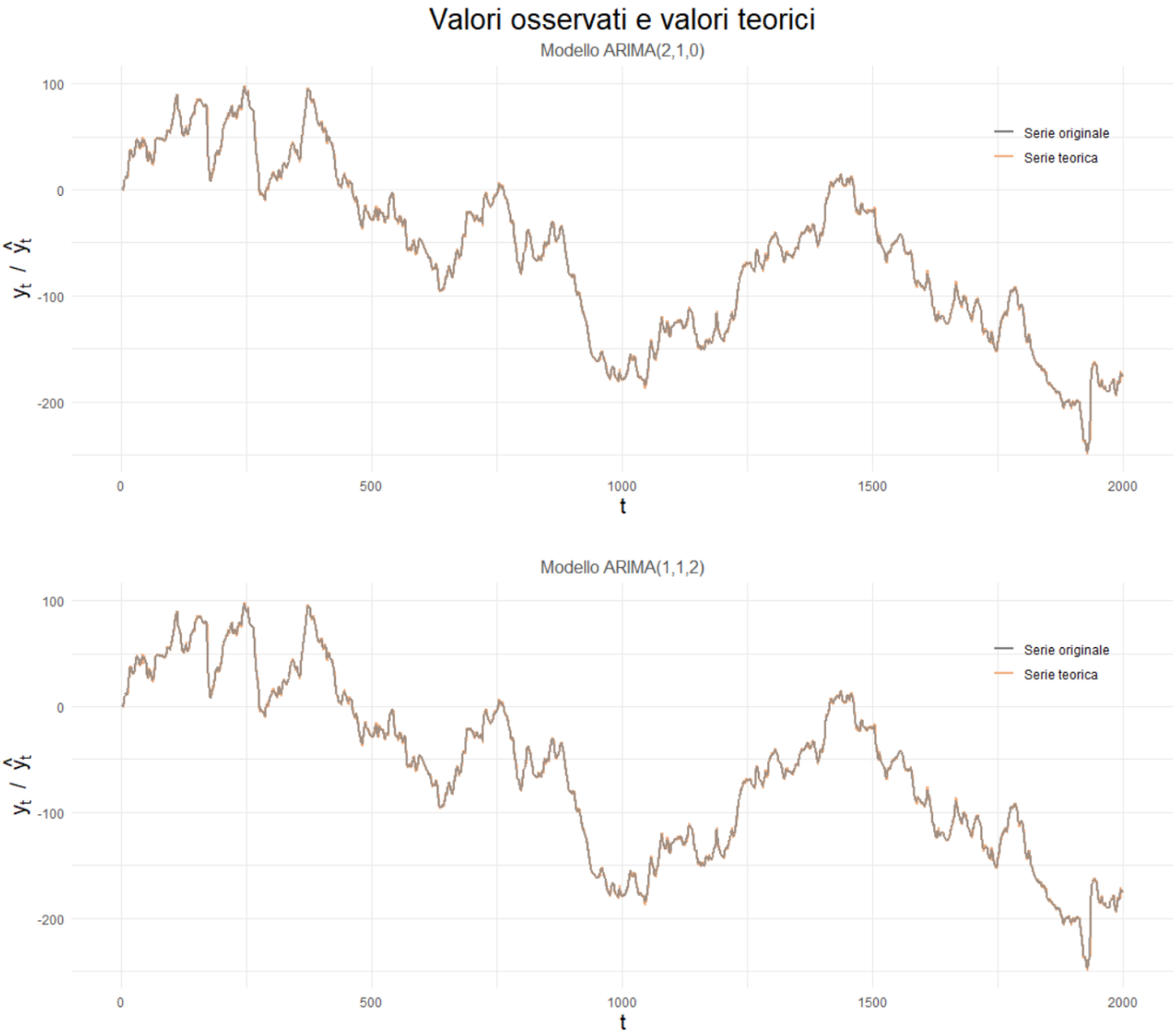
Risolvendo tale equazione rispetto a Y_t si ottiene

$$Y_t = 1.97Y_{t-1} - 1.26Y_{t-2} + 0.29Y_{t-3} + \varepsilon_t$$

Tuttavia, vi è un particolare da notare. L'intervallo di confidenza per il coefficiente autoregressivo ϕ_1 contiene il valore 1. Considerando tale aspetto, la serie differenziata, qualora provenga da un processo in cui uno dei parametri è maggiore o pari a uno in valore assoluto, dovrebbe essere una serie storica non stazionaria. Tuttavia, per i test statistici svolti in precedenza, si è potuto valutare come essa sia stazionaria.

Ovviamente, il parametro incognito potrebbe assumere un valore molto prossimo ad uno ma rimanere al di sotto di tale soglia, e in tal senso si avrebbe una serie stazionaria. Tale aspetto, tuttavia, volge a favore del modello stimato in precedenza.

Nel proseguivo della trattazione verranno analizzati entrambi i modelli stimati, con lo scopo di riuscire ad ottenere maggiori informazioni circa il miglior modello da adottare.

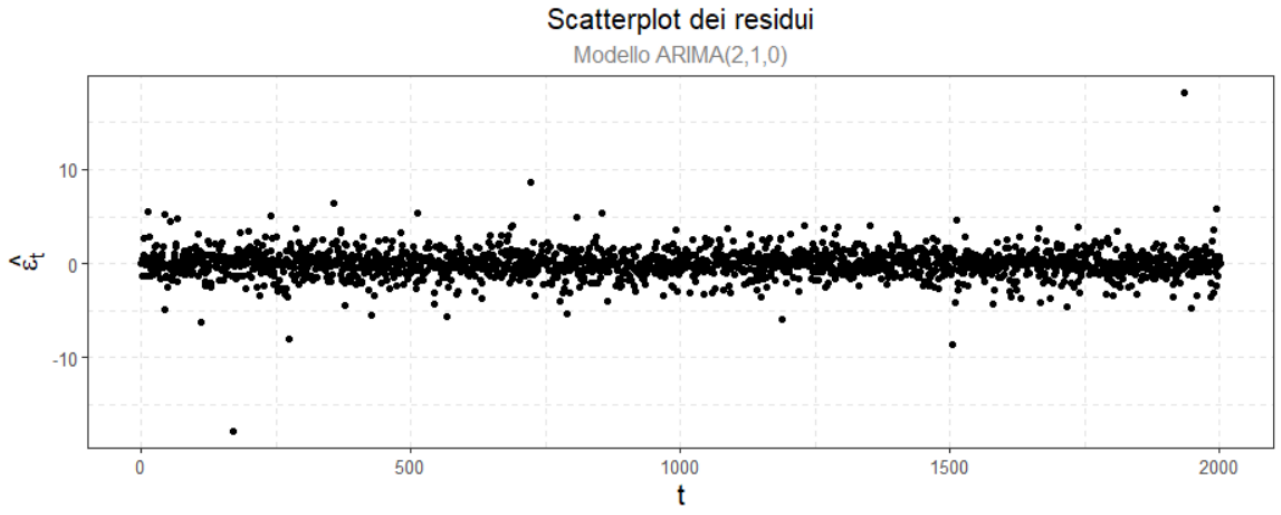


Prima di procedere all'analisi dei residui su entrambi i modelli, valutiamo l'adattamento degli stessi sui dati osservati. In particolare, si costruiscono due differenti grafici, uno per ciascun modello stimato, in cui vengono tracciati lo sviluppo della serie storica originale e, contemporaneamente, i valori teorici che si possono definire attraverso l'utilizzo del modello stimato. In entrambi i casi si può notare come la serie storica teorica ricalchi abbastanza fedelmente la serie storica originale, riuscendo anche a cogliere quei picchi e quelle decadute più repentine presenti nell'evoluzione del fenomeno. Tale aspetto si può analizzare in maniera approfondita sfruttando i residui.

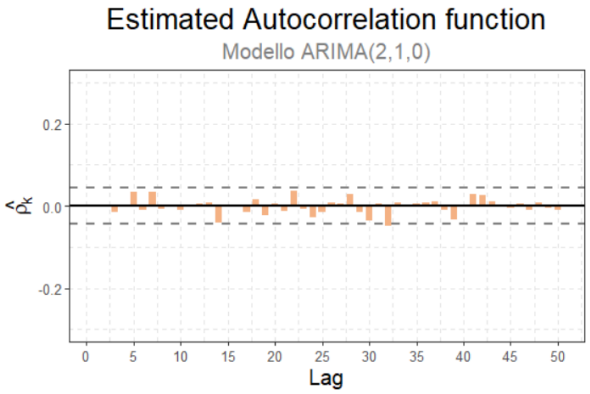
4 DIAGNOSTICA DEL MODELLO: ANALISI DEI RESIDUI

4.1 Analisi dei residui modello ARIMA(2,1,0)

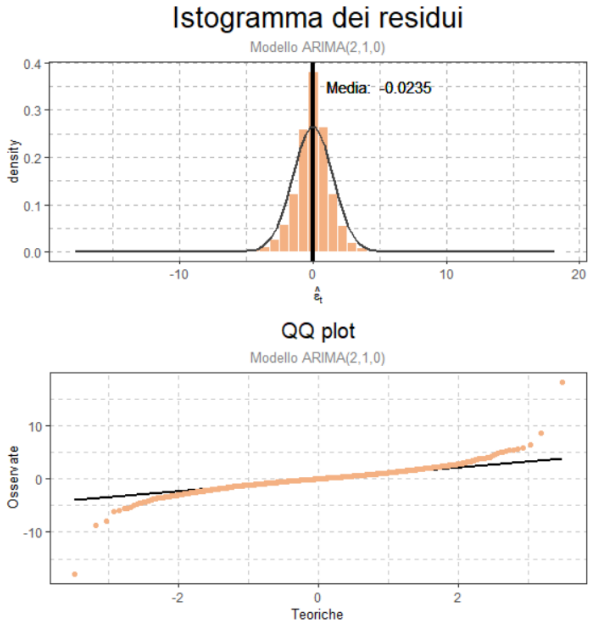
Verifichiamo, ora, in maniera separata l'adeguatezza dei due modelli stimati attraverso lo studio dei residui che è possibile ottenere dalla differenza tra il valore osservato e il valore teorico per ciascun tempo di osservazione.
In tale paragrafo vengo analizzati i residui che si originano nel momento in cui si stima un modello ARIMA(2,1,0).
Poiché i residui risultano equiparabili a stime della componente erratica presente all'interno del modello, essi dovranno presentare pressocché la medesima distribuzione che si è ipotizzata per tale componente, affinché si possa affermare che il modello riesca a cogliere in maniera ottimale le caratteristiche della serie. Ricordando che la componente erratica si ipotizza avere distribuzione secondo un processo white noise, con valore atteso nullo e varianza costante, dovremmo osservare che la nuvola di punti che si ottiene riportando su un grafico bidimensionali i residui ordinati rispetto al tempo abbia un andamento erratico, con valori attorno allo zero, e che non presenti aumenti o diminuzioni di variazione. Come si può osservare dal grafico sottostante, i residui mostrano un andamento abbastanza erratico; essi si dispongono quasi interamente all'interno di un intervallo intorno allo zero, con giusto qualche eccezione dovuta, verosimilmente, alla presenza di valori estremi rispetto agli altri registrati. Non si registrano, inoltre, sintomi di sistematicità, dovuti alla presenza di una qualche dipendenza tra i residui stessi.



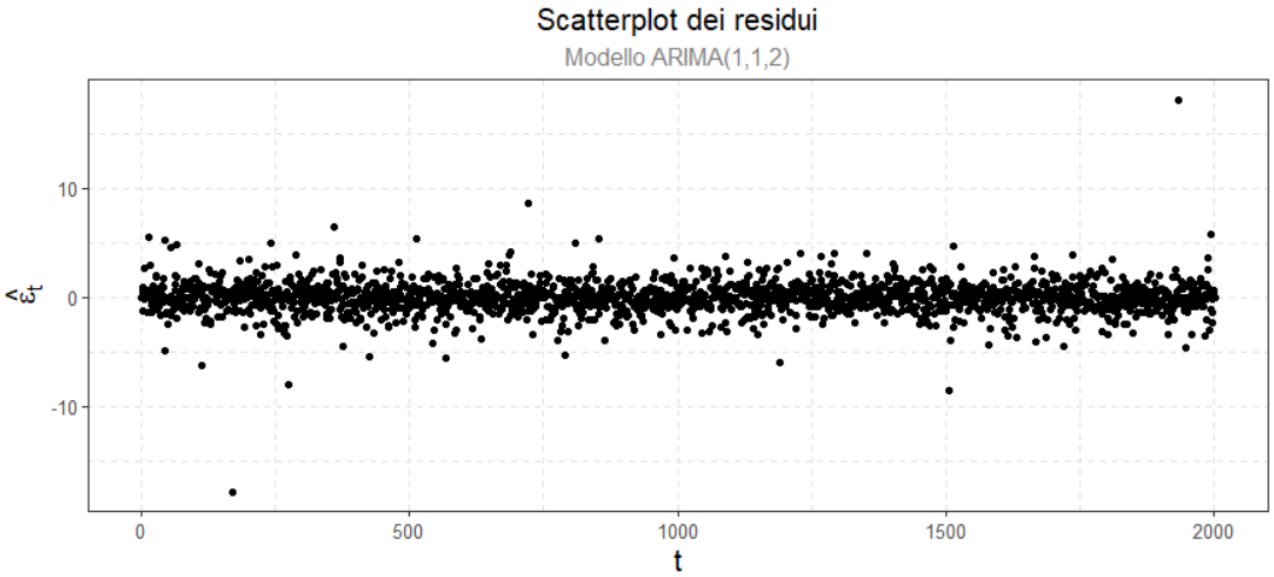
Un altro aspetto da considerare è proprio quella di assenza di dipendenza tra i residui. Tale aspetto può essere valutato osservando la funzione di autocorrelazione globale dei residui, a cui vengono affiancati alcuni test statistici al fine di verificare che gruppi di autocorrelazioni continue in senso temporale sia tutte statisticamente pari a 0 oppure no.
Sia il test statistico di Box-Pierce, che quello di Ljung-Box, presentano un valore del p-value di gran lunga superiore alla soglia di significatività prefissata, per cui si può affermare l'assenza di dipendenza lineare tra i residui. Affermazione che è anche supportata dal fatto che la funzione di autocorrelazione sia statisticamente pari a 0 per ciascuno dei lag considerati, come si può osservare dal correlogramma riportato.



Un altro aspetto che può essere valutato, ma che risulta in secondo piano rispetto all'erraticità dei residui, è quello per cui si verifica se i residui abbiano o meno una distribuzione gaussiana. Utilizzando gli strumenti grafici per testare la normalità di una serie, quali l'istogramma delle densità di probabilità e QQ plot, possiamo notare che da un punto di vista grafico i residui sembrano possedere un andamento normale, che si disperde nelle due code della distribuzione. Tuttavia, i test statistici condotti per testare la normalità di una distribuzione, test già applicati in precedenza sulla serie storica originale, rigettano l'ipotesi di normalità dei residui. In particolare, si ipotizza che i residui abbiano una distribuzione leptocurtica, in quanto è presente un eccesso di curtosi positivo. Tale aspetto si può anche scorgere dall'istogramma delle densità di probabilità di seguito riportato.

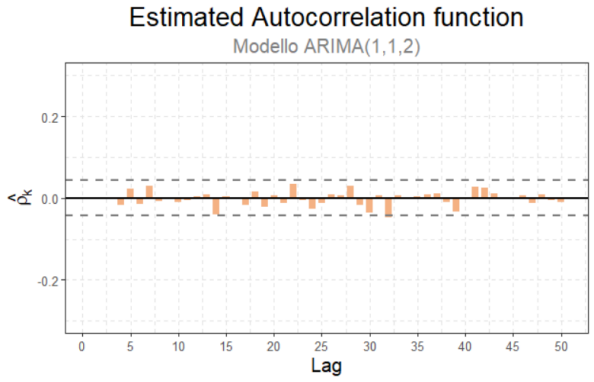


Tuttavia, l'assenza di normalità all'interno dei residui, e quindi di conseguenza nella serie storica originale, non inficia le stime dei parametri incogniti del modello, che rimangono comunque consistenti.
L'ultimo test statistico che è stato condotto sui residui riguarda il fatto di accertare la presenza o meno di eteroschedasticità all'interno di quest'ultimi. Il test condotto è il McLeod-Li test, il quale fornisce un valore del p-value superiore della soglia di significatività comunemente considerata per ogni lag che viene considerata, per cui non può essere rigettata l'ipotesi nulla di omogeneità all'interno dei dati. Tale test si fonda, comunque, sul test di Ljung-Box che viene applicato ai residui, considerandone il quadrato.



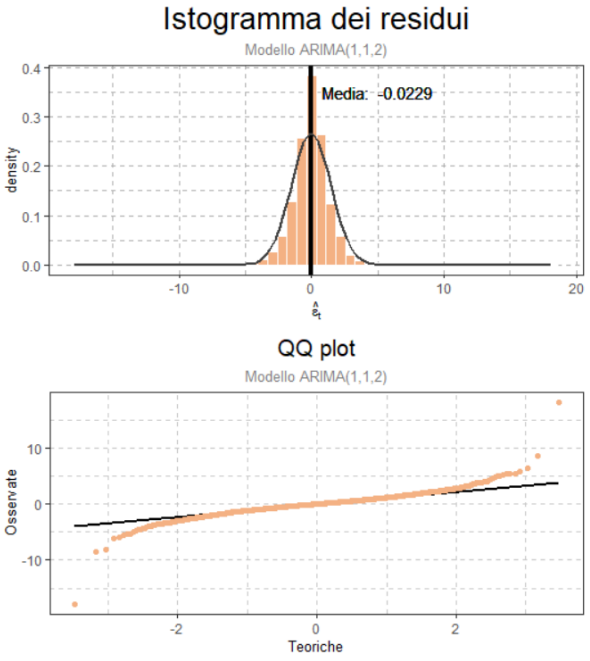
4.2 Analisi dei residui modello ARIMA(1,1,2)

Le medesime analisi che sono state condotte su i residui relativi al modello ARIMA(2,1,0) sono state condotte sui residui che possono essere calcolati sulla base delle osservazioni teoriche che si riescono ad ottenere dal modello ARIMA(1,1,2) che si è stimato.
Anche in questo caso, se si osserva il grafico in cui viene riportato il valore assunto dai residui rispetto al tempo, si può osservare come essi non presenti nessun pattern di sistematicità, nessuna anomalia tale da farci ipotizzare la presenza di una qualche forma di dipendenza non catturata dal modello stimato. I residui, così come nel modello precedente, si dispongono all'interno di un intervallo che si viene ad originare intorno al valor medio, che risulta essere 0 anche in questo caso. Vi sono sempre alcuni casi che si allontanano maggiormente dalla nuvola di punti, ma che comunque non destano preoccupazione.
Per quanto riguarda l'eventuale dipendenza lineare presente tra i residui, i test di Box-Pierce e di Ljung-Box escludono la presenza della stessa; tale aspetto viene anche confermato dalla stima della funzione di autocorrelazione per quanto riguarda sempre i residui, il cui correlogramma è riportato di seguito.



Si può infatti notare come la funzione di autocorrelazione globale, per ogni lag che si è considerato, sia significativamente pari 0, rimarcando l'assenza di correlazione tra i residui, il che conferma l'adeguatezza del modello stimato nel cogliere tutti gli aspetti relativi al processo stocastico generatore dei dati osservati presenti in quest'ultimi.
Anche in questo caso si è condotto il test di McLeod-Li, il quale ha confermato l'assenza di eteroschedasticità

all'interno dei residui provenienti dal modello ARIMA(1,1,2) stimato.
Si è proceduto, infine, a valutare la normalità dei residui attraverso l'utilizzo dei consueti test statistici, nonché degli strumenti grafici quali l'istogramma delle densità di probabilità e il QQ plot, riportati di seguito anche per tale modello.
In generale si può rilevare pressoché la medesima distribuzione che si osserva per i residui provenienti dal modello precedente stimato.
Anche i test statistici condotti ci portano ad accettare il fatto che i residui non presentino una distribuzione normale, in quanto si ha un eccesso positivo di curtosi. Nuovamente, tale aspetto non deve comunque preoccuparci, in quanto le stime dei parametri incogniti del modello risultano comunque essere consistenti.



Riassumendo l'analisi condotta finora sui residui di entrambi i modelli, si ha che, in generale, si riescono a trarre le medesime conclusioni da entrambi i modelli stimati, per cui risulta impossibile determinare quale sia il migliore, ossia quello che si adatta meglio ai dati osservati.

5 ANALISI SUL LIVELLO DI ACCURATEZZA DELLE PREVISIONI

5.1 Indici di errore e stima dell'accuratezza sul training set

Tale fase è rivolta a testare l'abilità del modello stimato nel predire i valori della serie, valori che non sono stati utilizzati nella fase di training del modello.
Le previsioni che si ottengono vengono valutate attraverso l'utilizzo di alcuni indici di errore quali la radice dell'errore quadratico medio (**RMSE**), l'errore assoluto medio (**MAE**) e il l'errore assoluto medio rapportato al valore osservato al tempo considerato (**MAPE**). Maggiore sarà il livello di tali indici, superiore sarà la distanza, la differenza presente tra il valore osservato e il valore teorico, e quindi previsto, attraverso il modello stimato.

La prima cosa che si può fare è utilizzare l'intera serie storica durante la fase di stima dei parametri incogniti del modello e riutilizzare i medesimi dati per valutare la capacità del modello di predire i valori futuri. Tale approccio risulta essere immediato da applicare e anche abbastanza banale, ma in tal modo si riesce ad ottenere una prima idea sul livello di accuratezza del modello stimato. Gli indici citati in precedenza possono essere ottenuti sfruttando la funzione `accuracy` presente all'interno della libreria `forecast`. I risultati che si ottengono considerando i modelli precedenti stimati sono riportati di seguito, evidenziando di volta in volta il modello che presenta un valore minore rispetto all'indice che si sta considerando.

	ARIMA(1,1,2)	ARIMA(2,1,0)
RMSE	1.500478	1.500923
MAE	1.018736	1.018936
MAPE	4.904975	4.88904

Si può notare come considerando l'indice RMSE e il MAE, il modello ARIMA(1,1,2) risulterebbe essere quello con un minore livello di errore. Tale tendenza si inverte se si considera l'indice relativizzato MAPE, per il quale risulta essere più accurato il modello ARIMA(2,1,0).
Tuttavia, basare l'analisi circa l'accuratezza del modello sulle previsioni future sulla medesima serie storica, intesa come medesima finestra temporale, che è stata utilizzata per la stima dei parametri incogniti del modello che si sta considerando può condurre a risultanti che possono essere forvianti; inoltre, vi è la possibilità di incorrere nel fenomeno dell'*overfitting*. Si ha, infatti, che il modello si adatti in maniera ottimale ai dati osservati ma, nel momento in cui viene sfruttato per previsioni future, le sue performance risultano non essere

all'altezza delle aspettative. Bisogna infatti considerare che la stima di un modello che possa descrivere il processo generato dei dati non è fine a sé stessa, ma che quest'ultimo dovrà essere utilizzato al fine di prevedere l'andamento futuro del fenomeno oggetto di studio. Pertanto, si preferisce stimare l'errore che si compierebbe utilizzando il modello individuato per prevede i valori futuri della serie confrontandoli con un insieme di informazioni che non sono state utilizzate durante la fase di stima dei parametri.
Il metodo più comune, e anche più semplice da implementare, consiste nel suddividere l'intera serie storica in due parti, ciascuna delle quali contenenti un determinato numero di osservazioni contigue tra di loro. Tali partizioni prendono, rispettivamente, il nome di partizione di training e partizione di testing. A differenza di quanto avviene nel momento in cui si analizza l'accuratezza di un modello di regressione lineare, oppure di un modello di regressione logistica, in cui le unità che andranno a formare l'insieme di testing vengo estratte dall'intero campione osservato in maniera casuale, nello studio delle serie storiche bisogna tenere conto della dipendenza temporale delle osservazioni. Pertanto, quello che si fa è destinare l'ultimo gruppo di osservazioni, in senso temporale, all'analisi dell'accuratezza del modello, mentre le altre rimanenti vengono utilizzate per la stima dei parametri incogniti.

Tuttavia, anche in questo caso si potrebbero avere dei difetti di stima del livello di errore, poiché questa si baserebbe unicamente su un campione di unità, anche abbastanza esiguo, che potrebbero presentare delle particolarità tali per cui non si otterrebbe una stima corretta dell'errore di previsione. Per risolvere tale problema di può pensare di implementare il metodo *sliding window*, per cui dalla serie storica vengono, in un certo senso, estrapolati un numero maggiore di campioni test su cui basare la stima degli indici RMSE, MAE e MAPE.

5.2 Stima dell'accuratezza con tecnica *sliding window*

In tale strategia di stima quello che si fa è far scivolare lungo l'asse temporale sia l'insieme di osservazioni che costituisce la partizione di training del modello, che l'insieme appartenente al campione test, su cui basare la stima dell'accuratezza del modello stimato.
Si è scelto di considerare come prima partizione per la stima dei parametri incogniti del modello le osservazioni dall'inizio della serie fino alla millesima, e considerare un campione test di ampiezza 5, che di conseguenza sono le osservazioni dalla numero 1001 alla numero 1005 comprese. Di volta in volta viene stimato un nuovo modello ARIMA dell'ordine prefissato sulla base della partizione di training, e ne viene valutata l'accuratezza basandoci su le 5 osservazioni future teoriche e quelle presenti all'interno del campione test. Infine, si fanno slittare entrambe le partizioni di un numero di osservazioni pari all'ampiezza della partizione test, che in tale contesto risulta essere pari a 5. Ad ogni passo vengono, pertanto, registrati i valori di RMSE, MAE e MAPE e di tali valori ne viene infine fatto la media, che viene utilizzata come stima del livello di errore. I risultati che si ottengono, considerando quindi un campione test composto dalle 5 osservazioni immediatamente successive a quelle presenti all'interno della partizione di training del modello

sono riportate di seguito, evidenziando sempre il modello per cui si ha un minor valore dell'indice di errore che si sta considerando.

	ARIMA(1,1,2)	ARIMA(2,1,0)
RMSE	3.804252	3.799415
MAE	3.21387	3.208775
MAPE	0.08583645	0.08571918

Basandoci sul valore assunto dai tre indici statistici si può notare come il modello ARIMA(2,1,0) risulti essere il migliore in termini di previsioni, se esse sono basate su osservazioni non presenti all'interno della partizione utilizzata nella fase di stima dei parametri incogniti.
Tuttavia, il livello di errore aumenta con l'espandersi della numerosità presente all'interno della partizione di test. Può, inoltre, succedere che per alcune particolari ampiezze del test set si registri una maggiore accuratezza del modello ARIMA(1,1,2) rispetto al modello ARIMA(2,1,0). In maniera particolare, se si considerano differenti ampiezze campionarie per il test set, da un minimo pari ad un'ampiezza unitaria, fino ad un massimo di 50 unità, aumentando ad ogni iterazione tale ampiezza di una unità, possiamo notare che per quanto riguarda l'indice RMSE il modello ARIMA(2,1,0) performa meglio del modello ARIMA(1,1,2) 26 volte su 50, per l'indice MAE 29 volte su 50, mentre per l'indice MAPE solo 20 volte su 50.
La stima di tali indici statistici è stata svolta sempre sfruttando la tecnica delle sliding window, variando di volta in volta l'ampiezza della partizione relativa al test dell'accuratezza.
In definitiva, anche valutando il livello di accuratezza delle previsioni non si riesce in maniera univoca a definire quale sia il modello più adatto alla serie storica osservata e che possa essere utilizzato per la stima dei valori futuri della serie.

6 PREVISIONI FUTURE SU MODELLO SELEZIONATO E CONSIDERAZIONI FINALI

6.1 Scelta del modello finale

Riconsideriamo, in maniera sintetica, le analisi che sono state condotte fino a questo momento, in maniera tale da individuare in maniera definitiva il modello da utilizzare al fine di prevedere i valori futuri del fenomeno oggetto di studio.

L'analisi sulla funzione di autocorrelazione globale e parziale mostra come quest'ultima di annulli per un lag pari a 3, il che ci porta a dedurre che la componente autoregressiva del modello sia di ordine pari a 2, così come accade nel modello ARIMA(2,1,0).

Nella ricerca del miglior modello effettuata basandoci sui valori dell'AIC in maniera particolare, ma lo stesso discorso vale per l'indice BIC, è risultato maggiormente adatto un processo ARIMA(2,1,0), mentre la funzione `auto.arima` fornisce come migliore modello il processo ARIMA(1,1,2). Tuttavia, proprio da un punto di vista dell'indice AIC e dell'indice BIC, risulta preferibile il modello ARIMA(2,1,0).

Nella stima dei parametri incogniti del modello, si può notare come il primo parametro inerente il modello ARIMA(2,1,0) sia molto prossimo ad 1, e il suo intervallo di confidenza contiene tale valore. Se fosse da tale processo che verosimilmente provengono i dati osservati, allora dovremmo osservare che anche differenziando la serie di un periodo, la traiettoria che si origina potrebbe essere ancora non stazionaria, ma questo non avviene, per quanto visto attraverso l'utilizzo dei test statistici relativi alle radici unitarie.

Per quanto riguarda l'analisi dei residui che si producono da entrambi i modelli, non si sono notati aspetti rilevanti tali per cui si preferisce l'utilizzo di un modello piuttosto dell'altro. In entrambi i casi si hanno residui non correlati tra di loro e che non presentano elementi di sistematicità tali per cui si sospetti la non correttezza dell'ordine del modello individuato. Inoltre, in entrambi i casi si hanno dei residui che non presentano una distribuzione propriamente normale.

Infine, considerando l'accuratezza delle previsioni, si ha che variando l'ampiezza su cui tali previsioni vengono effettuate, varia anche il modello che presenta valori minori rispetto agli indici di errore RMSE, MAE e MAPE. Pertanto, l'accuratezza non può essere utilizzata come discriminante per individuare il miglior modello che si adatta ai dati.

Un ultimo controllo che si può pensare di fare è confrontare le stime della funzione di autocorrelazione sia globale che parziale con i valori teorici di tali funzioni, che è possibile ricavare sfruttando le stime dei parametri incogniti dei due modelli.
Nella pratica, dopo aver calcolato i valori teorici della funzione di autocorrelazione globale e di quella parziale di entrambi i modelli, considerando un lag massimo pari a 50, si considera la media delle differenze in valore assoluto tra i valori teorici e i valori stimati. Notiamo che per quanto riguarda sia la funzione di autocorrelazione globale, che per la funzione di autocorrelazione parziale, si ha che il modello che ricalca meglio l'andamento delle due funzioni risulta essere il modello ARIMA(1,1,2). Si scopre, pertanto, che tale modello produce dei valori teorici delle funzioni di autocorrelazione globale e parziale più aderenti ai valori stimati di tali funzioni rispetto al modello ARIMA(2,1,0), che inizialmente sembrava essere il candidato ideale osservato il correlogramma sui dati osservati.

Per quanto osservato in precedenza si conclude che il modello migliore in termini di adattamento ai dati osservati risulta essere il processo ARIMA(1,1,2), anche se in termini di AIC e BIC risulterebbe essere il processo ARIMA(2,1,0). Si presume che tale aspetto si ottenga per via del fatto che l'ultimo modello presenta un parametro in meno rispetto al primo e, pertanto, si ha una minore penalizzazione del valore della funzione di verosimiglianza, che risulta essere simile per entrambi i modelli.

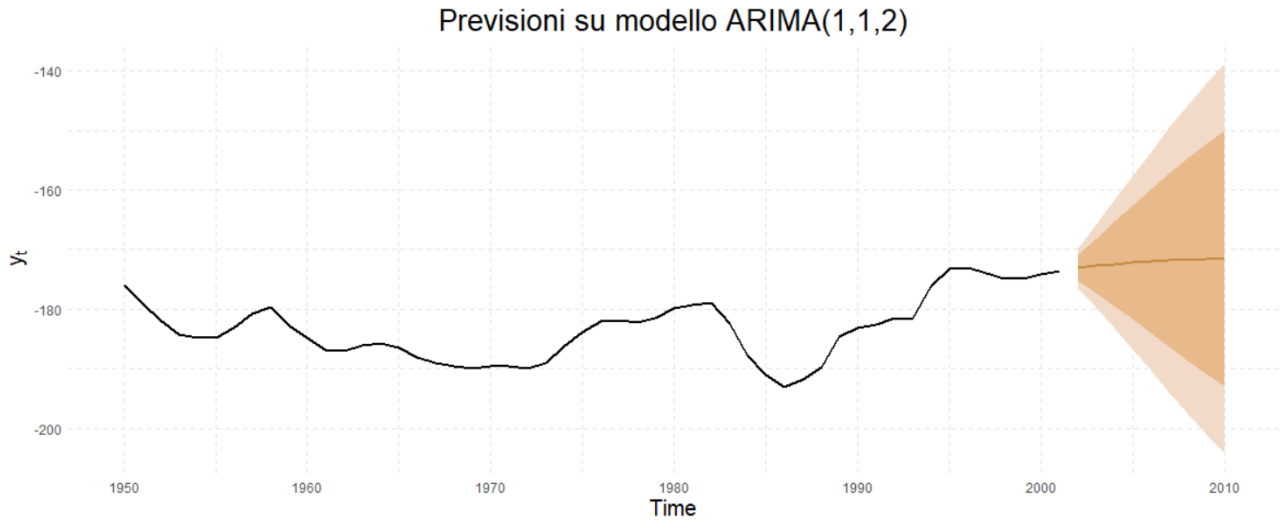
6.2 Previsioni future

Riportiamo le stime sul modello finale scelto da utilizzare per le previsioni future.

```
ARIMA (1,1,2)
Coefficients:
      Estimate  Std. Error   Pr(>|z|)
ar1  0.610027  0.031751  <2.2e-16
ma1  0.363733  0.036872  <2.2e-16
ma2  0.109366  0.033044  0.000934

         2.5%     97.5%
ar1  0.547797  0.672257
ma1  0.291464  0.436002
ma2  0.044602  0.174131

Sigma^2 estimated as 2.256
log likelihood=-3650.48
AIC=7308.96     AICc=7308.98
BIC=7331.36
```

L'ultimo passo consiste nell'effettuare alcune previsioni dei valori futuri del fenomeno, tenendo in considerazione la storia passata dello stesso. In particolare, si effettuano previsioni per 9 periodi di osservazione in avanti rispetto all'ultima osservazione registrata. Di tali previsioni se ne considera anche l'intervallo di confidenza per un livello di fiducia pari a 0.80 e 0.95. I 9 valori previsti vengono di seguito riportati, corredati di grafico.

```
1   -173.1683      6   -172.5698
2   -172.9134      7   -172.5483
3   -172.7578      8   -172.5351
4   -172.6630      9   -172.5271
5   -172.6051
```



Si può osservare come l'intervallo di confidenza delle previsioni aumenti con l'aumentare del lag considerato, come è ovvio aspettarsi. Si propone, anche, una previsione 50 passi in avanti, al fine di notare come da un certo punto in poi la previsione puntuale tende a stabilizzarsi attorno ad un valore medio.