Analisi Serie Storiche

Quando aprire un negozio sportivo

Giuseppe Pio Zito 583233

1 Introduzione

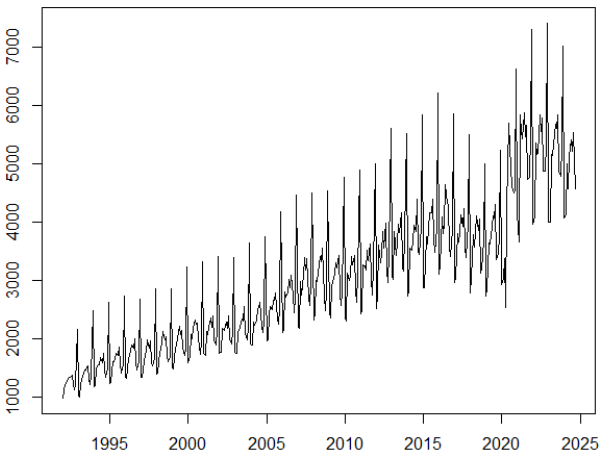
1.1 Scopo dello studio

Dopo le analisi sportive svolte per selezionare i partecipanti al *Six Kings Slam* e per suddividere i tennisti per tipologia di gioco, la federazione tennistica internazionale *ATP* contatta ancora una volta la nostra agenzia per effettuare uno studio sulle vendite di articoli sportivi negli USA durante l'anno, in modo da individuare il periodo migliore nel quale aprire il proprio store ufficiale dove vendere il proprio merchandising ufficiale e una vasta gamma di articoli tennistici (racchette, abbigliamento, scarpe, ecc.).

1.2 Caratteristiche della tabella

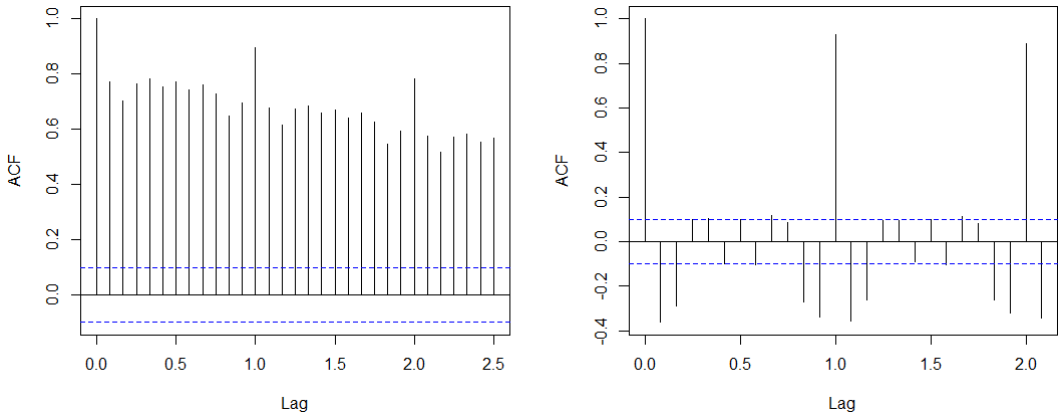
I dati usati (disponibili su `https://fred.stlouisfed.org/series/MRTSSM45111USN`) riguardano i ricavi mensili in milioni di dollari dalle vendite al dettaglio dei negozi di articoli sportivi negli USA nel periodo tra gennaio 1992 e settembre 2024. Il codice R completo usato è disponibile in appendice 4.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1992	972	1100	1214	1267	1293	1334	1340	1377	1206	1120	1207	2153
1993	1032	984	1213	1367	1387	1457	1457	1531	1313	1199	1377	2474
1994	1168	1193	1488	1551	1551	1677	1584	1737	1470	1324	1471	2611
1995	1224	1248	1599	1606	1645	1750	1705	1846	1569	1398	1548	2731
1996	1327	1309	1649	1702	1773	1895	1833	1992	1594	1453	1608	2675
1997	1337	1331	1609	1726	1788	1970	1838	1955	1595	1521	1651	2846
1998	1381	1407	1754	1820	1938	2120	1975	2038	1701	1607	1685	2858
1999	1485	1469	1827	1888	1993	2215	2062	2143	1847	1711	1832	3227
2000	1591	1681	2065	1991	2187	2319	2237	2257	2007	1725	1939	3309
2001	1739	1704	2111	2071	2185	2352	2199	2382	1984	1882	2151	3398
2002	1739	1763	2167	2157	2141	2276	2177	2384	1998	1897	2133	3387
2003	1761	1747	2121	2135	2182	2325	2293	2548	2075	1970	2200	3643
2004	1901	1888	2287	2221	2305	2514	2530	2621	2216	2103	2306	3748
2005	1956	1976	2408	2539	2493	2655	2652	2788	2338	2246	2492	4170
2006	2100	2154	2792	2713	2776	3012	2872	3085	2752	2447	2710	4456
2007	2200	2171	2969	2737	3087	3380	3165	3378	2712	2570	2939	4496
2008	2315	2415	2995	2984	3276	3408	3308	3543	2725	2482	2796	4530
2009	2442	2347	2894	2959	3119	3290	3209	3418	2779	2571	2788	4758
2010	2373	2303	3118	3001	3078	3413	3266	3415	2828	2611	3107	4894
2011	2414	2456	3266	3246	3169	3505	3334	3623	3011	2746	3227	4998
2012	2505	2723	3607	3304	3455	3829	3551	3969	3161	2959	3478	5601
2013	3071	3016	3831	3428	3657	3954	3819	4161	3230	3162	3750	5518
2014	2722	2803	3553	3520	3681	3934	3842	4386	3519	3435	3902	5840
2015	2874	2867	3739	3621	3882	4156	4161	4392	3702	3472	3883	6219
2016	3106	3237	4077	3794	3942	4634	4318	4307	3591	3401	3878	5849
2017	2966	3041	3795	3657	3813	4117	3921	4226	3566	3390	3978	5495
2018	2774	3023	3782	3564	3809	4095	3845	4048	3124	3189	3672	4988
2019	2730	2826	3636	3615	3762	4096	4073	4302	3346	3463	3758	5236
2020	2923	3029	3365	2526	4249	5696	5284	4948	4611	4491	4543	6617
2021	4144	3662	5830	5558	5430	5881	5468	5657	4733	4762	5509	7310
2022	3956	4079	5349	5152	5182	5843	5558	5781	4867	4874	5347	7416
2023	3991	3998	5161	5145	5322	5675	5611	5835	4879	4781	5274	7008
2024	4061	4126	4991	4573	5128	5399	5210	5539	4561			

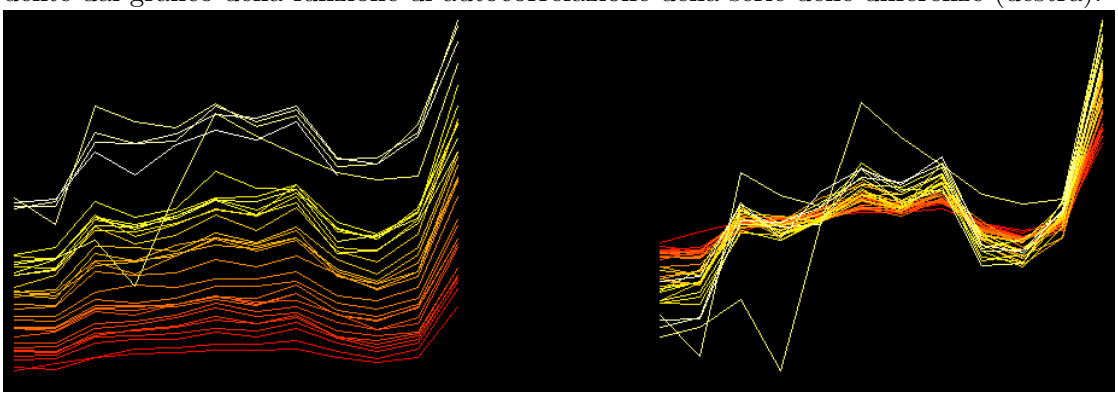


2 Scelta della decomposizione

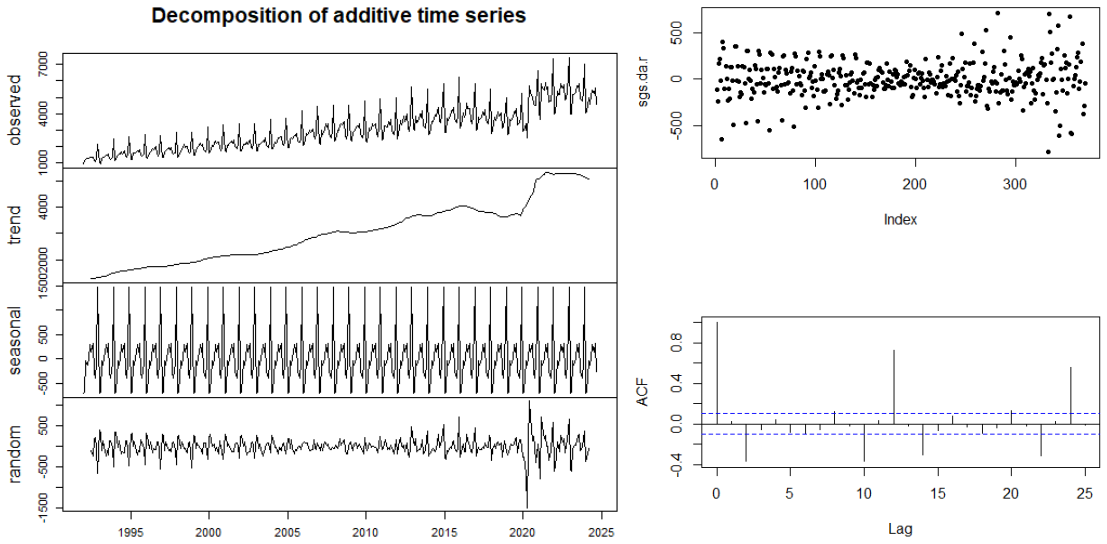
Dal grafico della serie sembrano presenti sia una componente di trend che una componente stagionale annuale, ma possiamo anche osservare un comportamento anomalo in corrispondenza dell'anno 2020, probabilmente a causa delle restrizioni dovute alla pandemia di Covid-19 scoppiata in quell'anno.



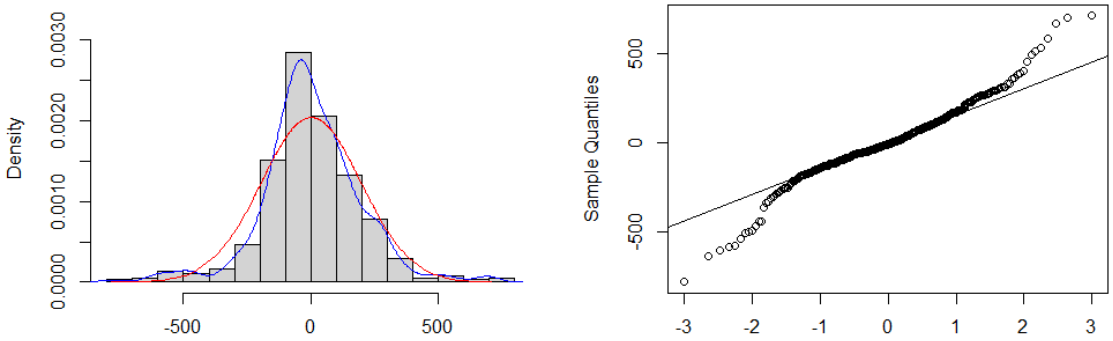
Il grafico della funzione di autocorrelazione della serie (sinistra) conferma l'impressione precedente riguardo la presenza di trend e stagionalità, quest'ultima maggiormente evidente dal grafico della funzione di autocorrelazione della serie delle differenze (destra).



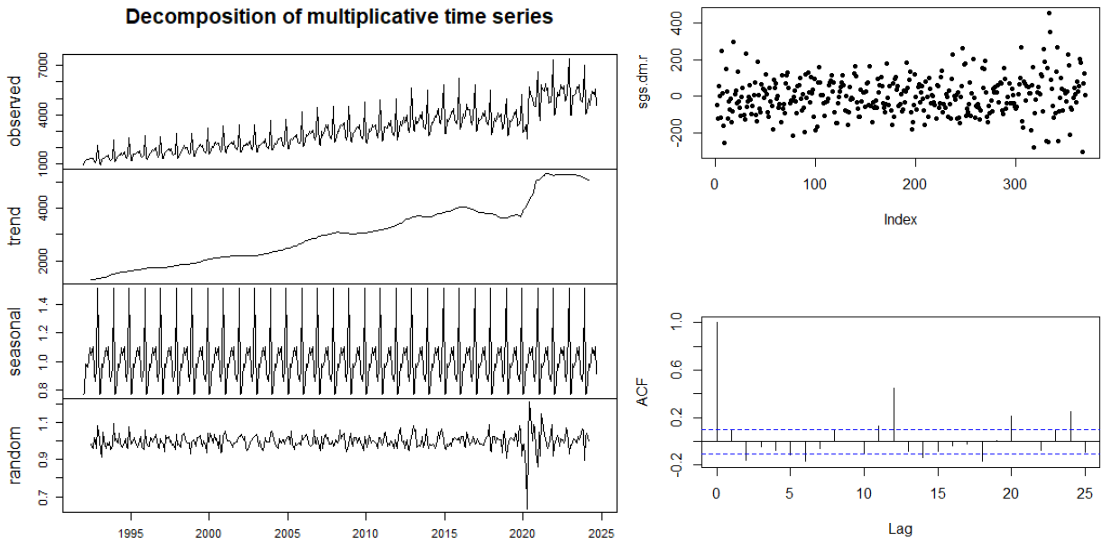
Confrontando l'andamento della serie di anno in anno (sinistra) dove dal rosso al bianco sono indicati gli anni dal più remoto al più recente rispettivamente, possiamo osservare che il trend sia tendenzialmente crescente. Andando ad annullare l'effetto del trend (destra), possiamo osservare che la serie di anno in anno ha un andamento piuttosto simile tranne in due casi che, dal colore giallo chiaro che denota la loro recentezza, possiamo intuire essere gli anni 2020 e 2021 maggiormente colpiti dalle restrizioni dovute alla pandemia di Covid-19. Inoltre, possiamo osservare che con l'avanzare degli anni i picchi minimi e massimi della serie si sono sempre più estremizzati. Questo fatto combinato con un trend tendenzialmente crescente sembra suggerire che in questo caso una decomposizione moltiplicativa può essere una scelta ragionevole. Per confermare questa ipotesi confrontiamo le decomposizioni additiva e moltiplicativa.



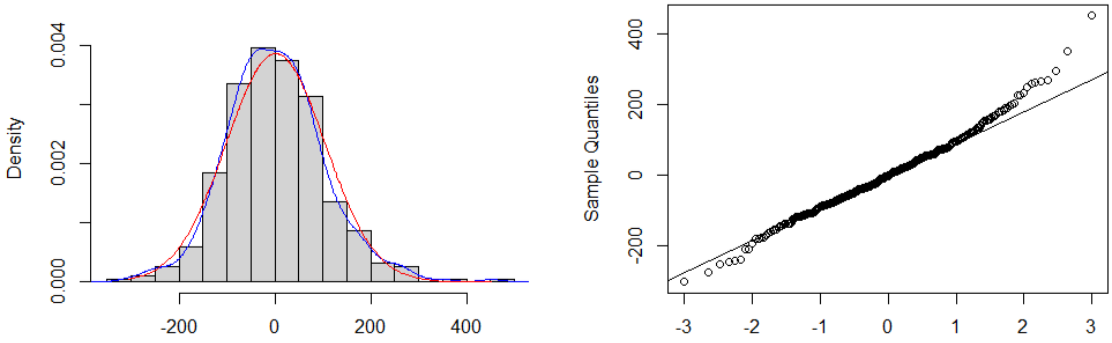
Per quanto riguarda la decomposizione additiva (sinistra) dagli ordini di grandezza delle componenti si può notare che il fenomeno principale della serie è il trend e che stagionaltà e rumore sono paragonabili. Inoltre, c'è un picco di rumore in corrispondenza del 2020 e sembra che una parte di stagionalità sia stata interpretata come rumore, ciò si vede abbastanza chiaramente soprattutto nella parte iniziale. Per verificare rigorosamente questa impressione svolgiamo una breve analisi del rumore; in particolare, per rendere più evidente l'eventuale presenza di struttura, escludiamo i dati del 2020 che già sappiamo avere un comportamento anomalo dovuto alla pandemia di Covid-19. Già dal grafico del rumore (destra in alto) si può notare della struttura nei primi e negli ultimi dati. Anche la funzione di autocorrelazione (destra in basso) conferma la presenza di una componente stagionale annuale, in particolare la deviazione standard della funzione di autocorrelazione è circa 0.3.



Possiamo osservare che (sinistra) la distribuzione empirica del rumore (in blu) è piuttosto simmetrica, come confermato anche da un valore di skewness pari a circa -0.05 molto vicino a 0, ma non è particolamente aderente alla distribuzione teorica gaussiana (in rosso), come confermato da un valore di kurtosi pari a circa 2.3 lontano da 0. Anche il grafico quantile-quantile (destra) sembra confermare le impressioni precedenti. Infine, il test di Shapiro-Wilk rigetta l'ipotesi di gaussianità del rumore restituendo valore molto basso del p-value (dell'ordine di 10^{-8}).



Per quanto riguarda la decomposizione moltiplicativa (sinistra) il trend sembra molto simile al caso additivo, ma non sembra esserci una struttura evidente nel rumore oltre al picco nel 2020. Per facilitare il confronto, analizzaremo il rumore escludendo il 2020 e riportandolo sulla stessa scala del caso additivo. Dall'ordine di grandezza del grafico del rumore (destra in alto) possiamo osservare che c'è meno rumore rispetto al caso additivo, mentre in generale si nota una leggera struttura nei primi e negli ultimi dati, come confermato dalla funzione di autocorrelazione (destra in basso). Questo fenomeno però è meno pronunciato rispetto al caso additivo, come confermato da un valore minore di deviazione standard della funzione di autocorrelazione pari a circa 0.24.

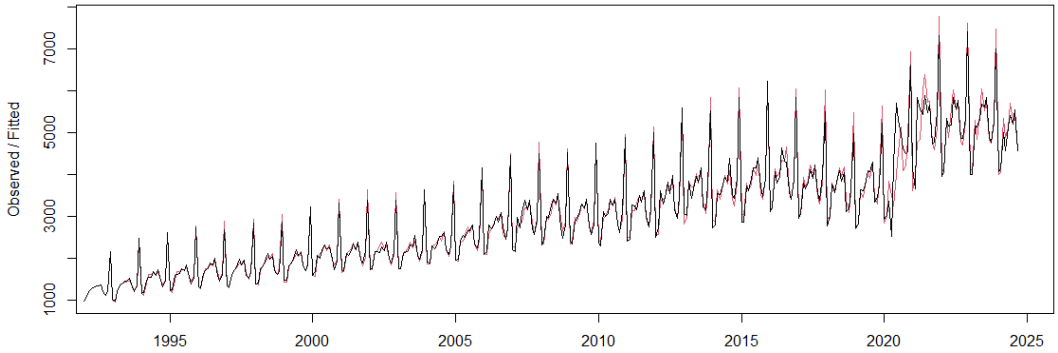


In questo caso (sinistra) la distribuzione empirica del rumore (in blu) è meno simmetrica con skewness pari a circa 0.39, ma in generale è più aderente alla distribuzione teorica gaussiana (in rosso), come confermato da un valore di kurtosi pari a circa 1.15, minore rispetto al caso additivo. Il grafico quantile-quantile (destra) sembra confermare le impressioni precedenti, in particolare l'asimmetria dovuta ad un leggero discostamento in alto a destra per alcuni valori. Infine, il test di Shapiro-Wilk, seppur restituendo un p-value (dell'ordine di 10^{-3}) maggiore del caso additivo, anche in questo caso rigetta l'ipotesi di gaussianità. Alla luce di quanto visto decidiamo di procedere con la decomposizione moltiplicativa che sembra catturare meglio le componenti della serie.

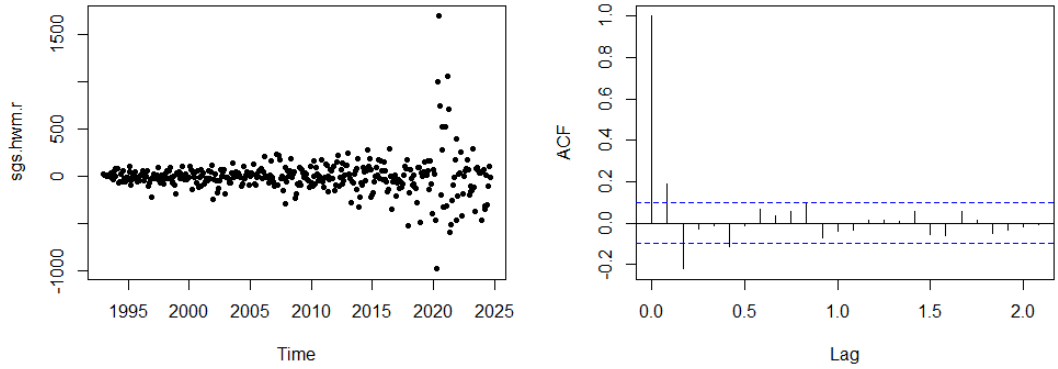
3 Analisi

3.1 Metodi di Smorzamento Esponenziale (Holt-Winters)

Poichè dallo studio precedente la serie risulta avere sia trend che stagionalità e la sua decomposizione è moltiplicativa, useremo lo smorzamento esponenziale con trend e stagionalità moltiplicativa. Iniziamo dal modello ottenuto con i parametri α, β e γ ottimizzati automaticamente dal software.

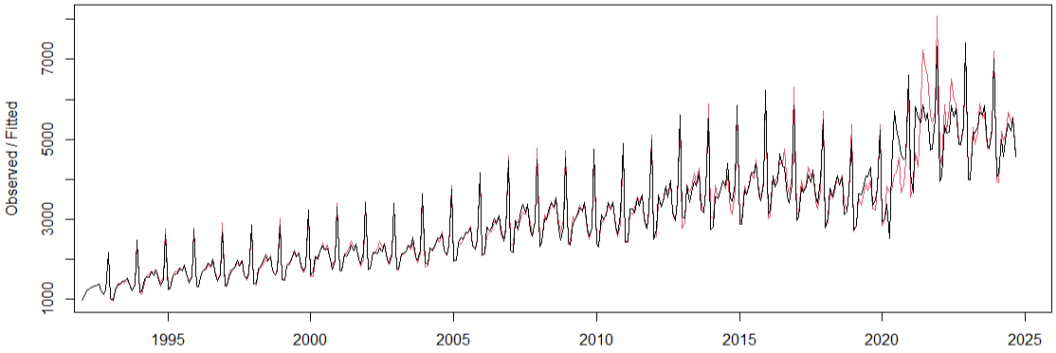


I parametri che si ottengono sono $\alpha = 0.422$, $\beta = 0$ e $\gamma = 0.323$ e possiamo osservare che il modello così ottenuto (in rosso) sembra aderire abbastanza bene alla serie tranne negli anni 2020 e 2021 che sappiamo essere anomali.

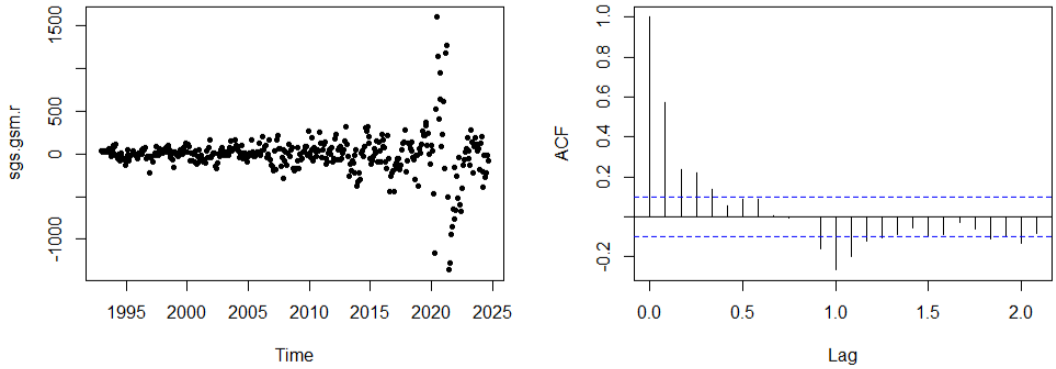


Dal grafico dei residui (sinistra) possiamo osservare che il modello sembra catturare abbastanza bene il comportamento della serie all'inizio per poi andare sempre più a perdere precisione, in particolare il picco tra il 2020 e il 2021 causato dalla pandemia di Covid-19 rigetta qualsiasi ipotesi di gaussianità dei residui. Comunque non sembra esserci una particolare struttura nei residui, come confermato anche dal grafico della funzione di autocorrelazione (destra). Infine, la varianza non spiegata dal modello è circa il 2.3%.

Adesso possiamo provare a cercare direttamente dei parametri che minimizzano le somme degli scarti quadratici medi nelle previsioni tramite grid-search, in particolare faremo variare i parametri $\alpha, \beta, \gamma = 0.1, 0.2, ..., 0.9$ effettuando 15 test per ogni scelta dei parametri con previsione di un anno (che è l'arco temporale che ci interessa). Aggiungendo anche i parametri ottimizzati automaticamente, possiamo condurre in parallelo un confronto per autovalidazione tra il modello precedente e quelli della grid-search.



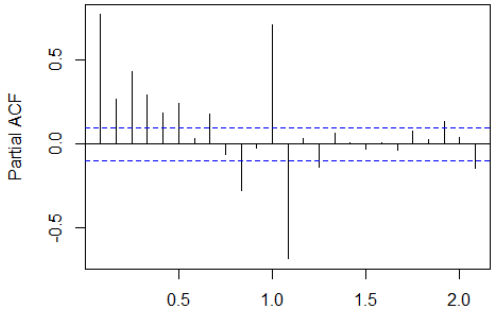
I coefficienti risultanti dalla grid-search sono $\alpha = 0.1$, $\beta = 0.2$ e $\gamma = 0.5$ con un errore (nel senso descritto in precedenza) pari a 425828, mentre il modello precedente ha commesso un errore pari a 929635. Il modello ottenuto dalla grid-search (in rosso) aderisce abbastanza bene alla serie all'inizio e alla fine, anche se sembra avere un comportamento peggiore rispetto al modello precedente negli anni anomali 2020 e 2021.



Dal grafico dei residui (sinistra) possiamo osservare che l'ordine di grandezza è paragonabile a quello del modello precedente, ma nella parte finale i residui sembrano meglio distribuiti. Inoltre, si può osservare la presenza di una leggera struttura nei residui, come confermato anche dal grafico della funzione di autocorrelazione (destra), in particolare si può affermare che i residui non sono gaussiani. Infine, la varianza non spiegata dal modello è circa il 3.8%, superiore a quella del modello precedente.

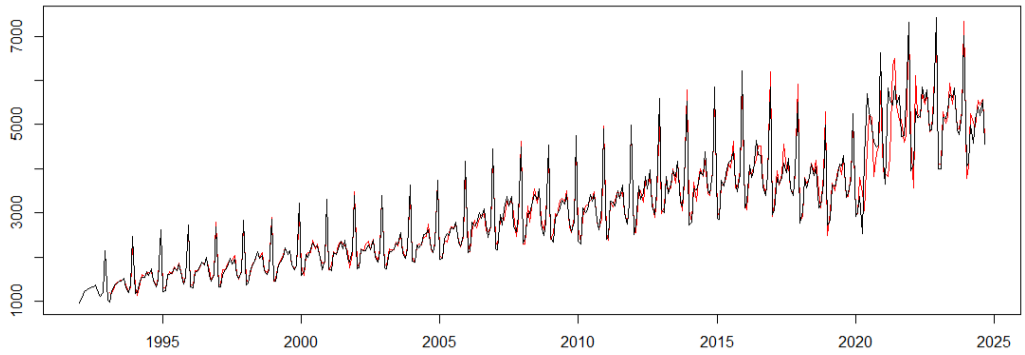
3.2 Metodi Autoregressivi

Innanzitutto proviamo a costruire un modello diretto.

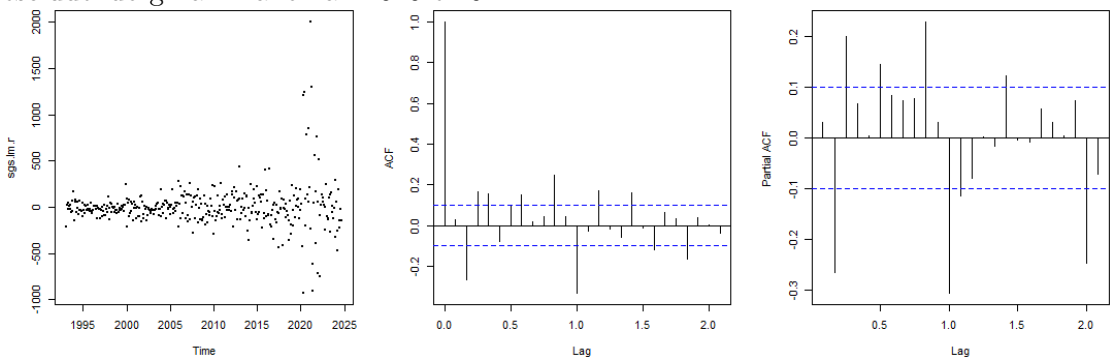


.

La funzione di autocorrelazione parziale sembra suggerire una dipendenza che arriva a arriva fino a 13 lag. Il modello lineare ottenuto con questa scelta ha una varianza spiegata pari al 96.91%, ma presenta molti fattori d'ingresso con p-value alto. Effettuando una riduzione del modello, otteniamo un modello lineare con soli 3 fattori d'ingresso corrispondenti ai lag 1, 12, 13 (che sono quelli con valore maggiore della funzione di autocorrelazione parziale) e varianza spiegata pari al 96.7%. Poichè la perdita di varianza spiegata è molto bassa, proseguiamo con l'analisi del solo modello ridotto.

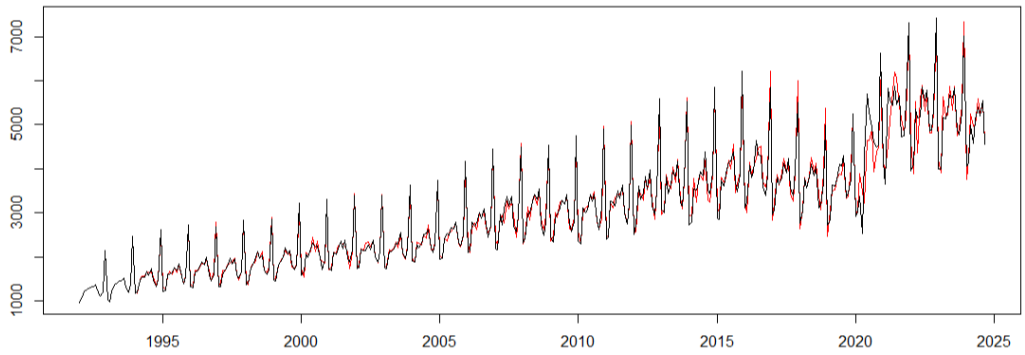


Anche in questo caso il modello (in rosso) sembra aderire abbastanza bene alla serie escludendo gli anni anomali 2020 e 2021.

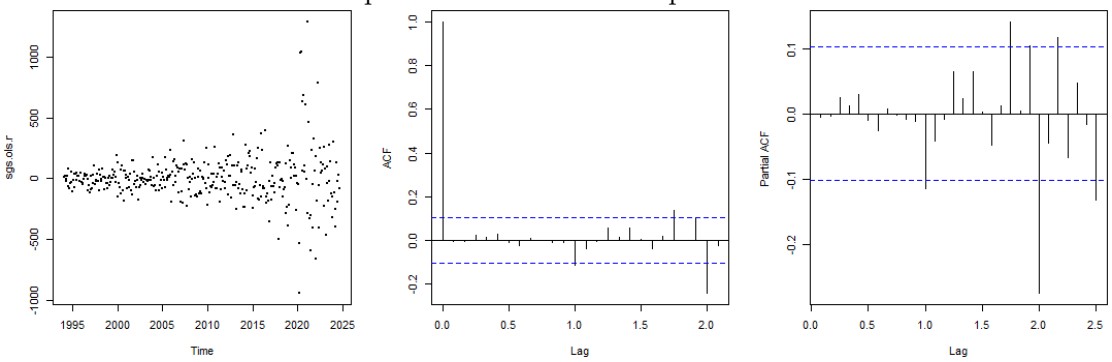


Dall'ordine di grandezza del grafico dei residui (sinistra) si può osservare che il picco negli anni anomali 2020 e 2021 è leggermente più pronunciato rispetto ai modelli precedenti. Inoltre, anche in questo caso si può osservare la presenza di una leggera struttura, come confermato dalla funzione di autocorrelazione (centro), in particolare i residui non sono gaussiani. La funzione di autocorrelazione parziale dei residui (destra) ci suggerisce che la presenza di struttura può essere causata dalla riduzione del modello, essendo che alcuni lag esclusi dal modello hanno valore della funzione di autocorrelazione parziale rilevante. Infine, la varianza non spiegata dal modello è circa il 3.3%.

Poichè la serie presenta un trend non banale, proviamo adesso ad utilizzare il metodo autoregressivo OLS, il quale sceglie il numero di lag in ingresso ideale come quello che minimizza lo scarto quadratico medio.



Il numero di lag in ingresso scelto dal metodo è pari a 25, più del doppio di quello identificato mediante la funzione di autocorrelazione parziale. Anche in questo caso il modello (in rosso) sembra aderire abbastanza bene alla serie e sembra approssimare meglio gli anni anomali 2020 e 2021 rispetto ai metodi visti in precedenza.



In effetti, anche l'ordine di grandezza del grafico dei residui (sinistra) sembra confermare tale impressione, mentre come nei casi precedenti si può affermare che i residui non sono gaussiani. Inoltre, non sembra esserci una particolare struttura nei residui, come confermato dalle funzioni di autocorrelazione (centro) e autocorrelazione parziale (destra), anche se entrambe presentano un valore rilevante per 24 lag. Infine, la varianza non spiegata dal modello è circa il 2.4%.

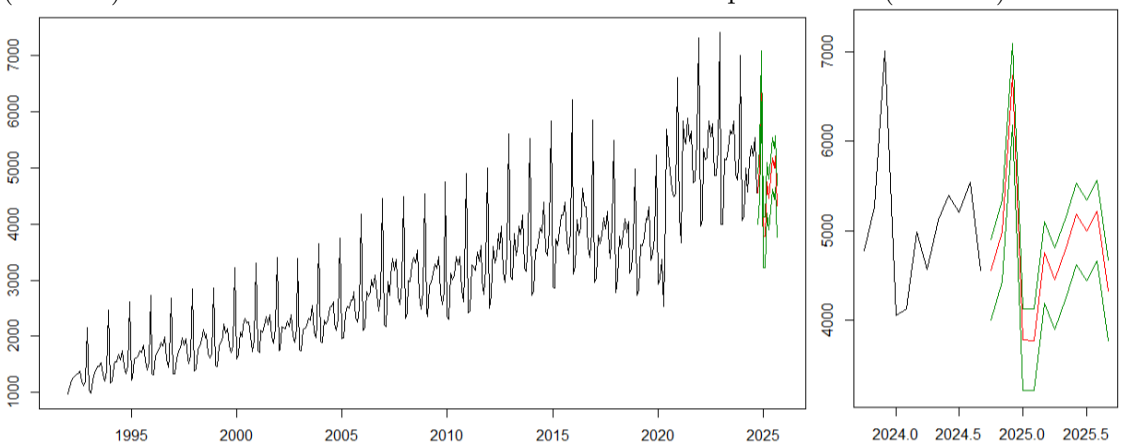
3.3 Scelta del metodo

Svolgendo un confronto per autovalidazione nelle stesse modalità usate per i modelli di smorzamento esponenziale tra i due modelli autoregressivi analizzati, otteniamo che il modello diretto ridotto commette un errore pari a 927669 mentre il modello OLS commette un errore pari a 1790848. Ricordando che il modello con coefficienti ottimizzati per grid-search ha commesso un errore pari a 425828 e il modello con coefficienti ottimizzati automaticamente ha commesso un errore pari a 929635, possiamo affermare che il modello che effettua le previsioni migliori nell'arco di un anno è il modello di smorzamento esponenziale con coefficienti ottimizzati per grid-search, che dunque sarà il modello con cui effettueremo la previsione finale.
Vale la pena osservare che il modello autoregressivo OLS, il quale sembrava meglio approssimare la serie e avere i residui migliori, risulta il modello che effettua le previsioni

peggiori. Potrebbe essere questo un caso di overfitting, nel quale il modello adattandosi troppo bene ai dati di addestramento perde sensibilmente capacità predittiva.

4 Previsione e conclusioni

Poichè abbiamo osservato che i residui non sono gaussiani, per completare la previsione (in rosso) useremo solamente intervalli di incertezza non parametrici (in verde).



Dunque il periodo migliore nei prossimi 12 mesi per aprire lo store dovrebbe essere dicembre 2024, ma grazie a questa previsione l'*ATP* potrà anche gestire al meglio gli articoli in magazzino in modo da non essere impreparata nei periodi in cui si prevedono maggiori vendite.

Appendice

Script R

```
   #funzione utile
#restituisce la data dopo l'unità successive a s nella forma (maggiore,minore)
ts_data <- function(s,t,l){
  #s: data di inizio come (maggiore,minore)
  #t: periodo
  #l: unità  successive
  e=s+c(floor(l/t),l%%t)
  e=c(e[1]+floor(e[2]/t),e[2]%%t)
  return(e)
}
tab=read.csv("tabella.csv") #carico la tabella
sgs=ts(tab[,2],frequency=12,start=c(1992,1))
sgs
plot(sgs)
n=length(sgs)
idt=start(sgs)
fdt=end(sgs)
pdt=frequency(sgs)
   #decomposizone serie storica
layout(t(1:2))
acf(sgs,30,main="")
acf(diff(sgs),main="")
#confronto anni
par(bg="black")
m_sgs=matrix(ts(c(tab[,2],rep(NA,3)),frequency=12,start=c(1992,1)),12,33)
ts.plot(m_sgs,col=heat.colors(33))
ts.plot(scale(m_sgs,scale=F),col=heat.colors(33))
par(bg="white")
layout(1)
  #decomposizone additiva
sgs.da=decompose(sgs)
plot(sgs.da)
layout(1:2) #analisi dei residui escluso 2020
sgs.da.r=c(na.omit(window(sgs.da$random,end=c(2019,12))),
           na.omit(window(sgs.da$random,c(2021,1))))
plot(sgs.da.r,pch=20)
acf(sgs.da.r,main="")
sd(acf(sgs.da.r,plot=F)$acf)
layout(t(1:2)) #confronto gaussiana
```

```
hist(sgs.da.r,20,freq=F,ylim=c(0,0.0033),main="")
lines(density(sgs.da.r),col="blue")
lines(sort(sgs.da.r),dnorm(sort(sgs.da.r),mean(sgs.da.r),sd(sgs.da.r)),
      col="red")
#skewness
mean((sgs.da.r-mean(sgs.da.r))^3)/(mean((sgs.da.r-mean(sgs.da.r))^2)^(3/2))
#kurtosi
mean((sgs.da.r-mean(sgs.da.r))^4)/(mean((sgs.da.r-mean(sgs.da.r))^2)^2)-3
qqnorm(sgs.da.r,main="")
qqline(sgs.da.r)
shapiro.test(sgs.da.r)
layout(1)
  #decomposizione moltiplicativa
sgs.dm=decompose(sgs,type="multiplicative")
plot(sgs.dm)
layout(1:2) #analisi dei residui escluso 2020
#rumore portato sulla stessa scala del caso additivo e centrato
sgs.dm.r=mean(sgs.dm$trend,na.rm=T)*
         (c(na.omit(window(sgs.dm$random,end=c(2019,12))),
            na.omit(window(sgs.dm$random,c(2021,1))))-1)
plot(sgs.dm.r,pch=20)
acf(sgs.dm.r,main="")
sd(acf(sgs.dm.r,plot=F)$acf)
layout(t(1:2)) #confronto gaussiana
hist(sgs.dm.r,20,freq=F,ylim=c(0,0.0045),main="")
lines(density(sgs.dm.r),col="blue")
lines(sort(sgs.dm.r),dnorm(sort(sgs.dm.r),mean(sgs.dm.r),sd(sgs.dm.r)),
      col="red")
#skewness
mean((sgs.dm.r-mean(sgs.dm.r))^3)/(mean((sgs.dm.r-mean(sgs.dm.r))^2)^(3/2))
#kurtosi
mean((sgs.dm.r-mean(sgs.dm.r))^4)/(mean((sgs.dm.r-mean(sgs.dm.r))^2)^2)-3
qqnorm(sgs.dm.r,main="")
qqline(sgs.dm.r)
shapiro.test(sgs.dm.r)
layout(1)
   #analisi
  #metodi di smorzamento esponenziale (Holt-Winters)
 #coefficienti ottimizzati automaticamente
sgs.hwm=HoltWinters(sgs,seasonal="multiplicative")
plot(sgs.hwm,main="")
print(paste("alpha =",round(sgs.hwm$alpha,3),"beta =",sgs.hwm$beta,
            "gamma =",round(sgs.hwm$gamma,3)))
```

```
layout(t(1:2)) #analisi residui
sgs.hwm.r=resid(sgs.hwm)
plot(sgs.hwm.r,type="p",pch=20)
acf(sgs.hwm.r,main="")
layout(1)
#varianza non spiegata
var(sgs.hwm.r)/var(window(sgs,start(sgs.hwm.r),end(sgs.hwm.r)))
 #scelta coefficienti con grid-search
nt=15 #numero di test set
ft=12 #unità di tempo nel futuro su cui valutare la previsione
min=c(0,0,0)
err_gsm=1e+10
err_hwm=0
#includiamo anche i coefficienti ottimizzati automaticamente
a=c(sgs.hwm$alpha,1:9/10)
b=c(sgs.hwm$beta,1:9/10)
c=c(sgs.hwm$gamma,1:9/10)
for(i in a){
  for(j in b){
    for(k in c){
      err=rep(0,nt)
      for(l in (n-nt-ft):(n-1-ft)){
        train=window(sgs,idt,ts_data(idt,pdt,l))
        test=window(sgs,ts_data(idt,pdt,l+1),ts_data(idt,pdt,l+ft))
        train.hw=HoltWinters(train,alpha=i,beta=j,gamma=k,
                             seasonal='multiplicative')
        err[l-(n-nt-ft)+1]=mean((as.numeric(test)-
                                 as.numeric(predict(train.hw,ft)))^2)
      }
      if(err_gsm>sum(err)){
        err_gsm=sum(err)
        min[1]=i
        min[2]=j
        min[3]=k
      }
      if(i==sgs.hwm$alpha && j==sgs.hwm$beta && k==sgs.hwm$gamma){
        err_hwm=sum(err)
      }
    }
  }
}
print(paste("alpha =",min[1],"beta =",min[2],"gamma =",min[3]))
sgs.gsm=HoltWinters(sgs,alpha=min[1],beta=min[2],gamma=min[3],
```

```
                    seasonal='multiplicative')
plot(sgs.gsm,main="")
print(paste("Modello Holt-Winters automatico - errore:",round(err_hwm)))
print(paste("Modello Holt-Winters grid-search - errore:",round(err_gsm)))
layout(t(1:2)) #analisi residui
sgs.gsm.r=resid(sgs.gsm)
plot(sgs.gsm.r,type="p",pch=20)
acf(sgs.gsm.r,main="")
layout(1)
#varianza non spiegata
var(sgs.gsm.r)/var(window(sgs,start(sgs.gsm.r),end(sgs.gsm.r)))
  #metodi autoregressivi
 #metodo diretto ridotto
pacf(sgs,main="") #l'ultimo valore di lag rilevante è 13
lg=13 #creazione modello
msgs=matrix(nrow=n-lg,ncol=lg+1)
for(i in 1:(lg+1)) {
  msgs[,i]=sgs[i:(n-lg-1+i)]
}
msgs<-data.frame(msgs)
msgs.lm<-lm(X14~.,data=msgs) #riduzione modello
summary(msgs.lm) #R^2=0.9691, X12 p-value=0.44569
msgs.lm<-lm(X14~.-X12,data=msgs)
summary(msgs.lm) #R^2=0.969, X6 p-value=0.57174
msgs.lm<-lm(X14~.-X12-X6,data=msgs)
summary(msgs.lm) #R^2=0.969, X5 p-value=0.19685
msgs.lm<-lm(X14~.-X12-X6-X5,data=msgs)
summary(msgs.lm) #R^2=0.9689, X7 p-value=0.09986
msgs.lm<-lm(X14~.-X12-X6-X5-X7,data=msgs)
summary(msgs.lm) #R^2=0.9686, X4 p-value=0.0581
msgs.lm<-lm(X14~.-X12-X6-X5-X7-X4,data=msgs)
summary(msgs.lm) #R^2=0.9683, X8 p-value=0.1016
msgs.lm<-lm(X14~.-X12-X6-X5-X7-X4-X8,data=msgs)
summary(msgs.lm) #R^2=0.9681, X10 p-value=0.1766
msgs.lm<-lm(X14~.-X12-X6-X5-X7-X4-X8-X10,data=msgs)
summary(msgs.lm) #R^2=0.9679, X3 p-value=0.05690
msgs.lm<-lm(X14~.-X12-X6-X5-X7-X4-X8-X10-X3,data=msgs)
summary(msgs.lm) #R^2=0.9676, X9 p-value=0.176
msgs.lm<-lm(X14~.-X12-X6-X5-X7-X4-X8-X10-X3-X9,data=msgs)
summary(msgs.lm) #R^2=0.9675, X11 p-value=0.161
msgs.lm<-lm(X14~.-X12-X6-X5-X7-X4-X8-X10-X3-X9-X11,data=msgs)
summary(msgs.lm) #R^2=0.967
sgs.lm=window(sgs,ts_data(idt,pdt,lg))-resid(msgs.lm)
```

```
ts.plot(sgs.lm,sgs,col=c("red","black"))
layout(t(1:3)) #analisi residui
sgs.lm.r=ts(resid(msgs.lm),frequency=12,start=ts_data(idt,pdt,lg))
plot(sgs.lm.r,type="p",pch=20)
acf(sgs.lm.r,main="")
pacf(sgs.lm.r,main="")
layout(1)
#varianza non spiegata
var(sgs.lm.r)/var(window(sgs,ts_data(idt,pdt,lg)))
 #OLS
ols=ar(sgs,method="ols")
ols$order #25 lag
sgs.ols=window(sgs,ts_data(idt,pdt,ols$order))-na.omit(ols$resid)
ts.plot(sgs.ols,sgs,col=c("red","black"))
layout(t(1:3)) #analisi residui
sgs.ols.r=na.omit(ols$resid)
plot(sgs.ols.r,type="p",pch=20)
acf(sgs.ols.r,main="")
pacf(sgs.ols.r,30,main="")
layout(1)
#varianza non spiegata
var(sgs.ols.r)/var(window(sgs,ts_data(idt,pdt,ols$order)))
 #confronto modelli autoregressivi per autovalidazione
err_lm=rep(0,nt)
err_ols=rep(0,nt)
for(l in (n-nt-ft):(n-1-ft)){
  train=window(sgs,idt,ts_data(idt,pdt,l))
  test=window(sgs,ts_data(idt,pdt,l+1),ts_data(idt,pdt,l+ft))
  #metodo diretto ridotto
  L=length(train)
  mtrain=matrix(nrow=L-lg,ncol=lg+1)
  for(i in 1:(lg+1)){
    mtrain[,i]=train[i:(L-lg-1+i)]
  }
  mtrain<-data.frame(mtrain)
  train.lm<-lm(X14~X1+X2+X13,data=mtrain)
  train.lm.p=rep(0,L+ft)
  train.lm.p[1:L]=train
  for(i in 1:ft){
    train.lm.p[L+i]=coef(train.lm)%*%c(1,train.lm.p[L+i-13],
                                       train.lm.p[L+i-12],train.lm.p[L+i-1])
  }
  err_lm[l-(n-nt-ft)+1]=mean((as.numeric(test)-train.lm.p[(L+1):(L+ft)])^2)
```

```
  #metodo OLS
  train.ols=ar(train,method="ols")
  err_ols[l-(n-nt-ft)+1]=mean((as.numeric(test)-
                               as.numeric(predict(train.ols,n.ahead=ft,
                                                  se.fit=F)))^2)
}
print(paste("Modello autoregressivo ridotto - errore:",round(sum(err_lm))))
print(paste("Modello autoregressivo OLS - errore:",round(sum(err_ols))))
   #previsione finale con HW con coefficienti grid-search
sgs.gsm.p=predict(sgs.gsm,12)
ts.plot(sgs,sgs.gsm.p,col=c("black","red"))
#intervalli non parametrici
lines(sgs.gsm.p+quantile(sgs.gsm.r,0.025),col="green4")
lines(sgs.gsm.p+quantile(sgs.gsm.r,0.975),col="green4")
  #zoom del periodo finale
wsgs=window(sgs,start=c(2023,10))
ts.plot(wsgs,sgs.gsm.p,ylim=c(3200,7300),col=c("black","red"))
#intervalli non parametrici
lines(sgs.gsm.p+quantile(sgs.gsm.r,0.025),col="green4")
lines(sgs.gsm.p+quantile(sgs.gsm.r,0.975),col="green4")
```