UNIVERSITÀ DI PISA

Terza Relazione di Statistica
Serie Storiche

Manni Matteo

PRESENTAZIONE DEL PROBLEMA

Lo scopo di queste analisi sarà quello di fornire una previsione sull'andamento del prezzo dell'elettricità. Immaginiamo che un'azienda americana necessiti della previsione affinché possa programmare la produzione per i mesi successivi in base anche alle spese sull'energia elettrica. Collochiamoci temporalmente ad inizio 2020. Per riuscire a fornire delle previsioni analizzeremo una serie storica riguardante il prezzo medio dell'elettricità elettrica per kilowattora (kWh). In prima battuta la studieremo da un punto di vista qualitativo, cercando di capirne la natura; successivamente proveremo a fornire delle previsioni mediante diverse metodologie.

DATASET

La serie storica analizzata è stata reperita dal database "Federal Reserve Economic Data", nello specifico al link: https://fred.stlouisfed.org/series/APU000072610. Sono stati estrapolati i dati che vanno dal gennaio del 1986 al dicembre del 2020. In particolare, si tratta di una serie storica con frequenza mensile, le cui informazioni sono raccolte dal Dipartimento dell'Energia tramite questionari. I dati riguardano 75 aree diverse degli Stati Uniti e l'unità di misura con il quale il prezzo dell'elettricità viene registrato è il dollaro statunitense.

ANALISI PRELIMINARE

fig.1

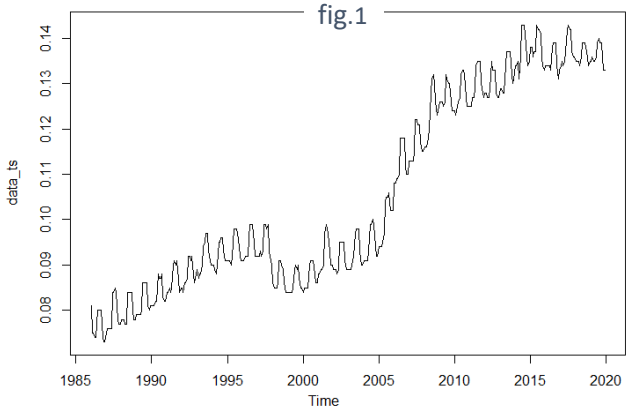
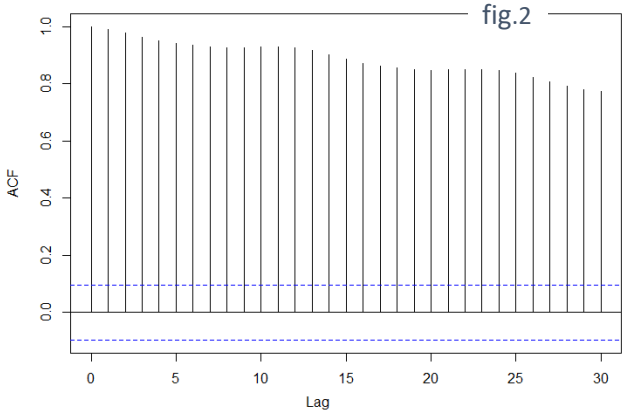


fig.2



Iniziamo l'analisi osservando l'andamento della serie storica nel tempo (fig.1). Notiamo subito che la struttura è composta da un fattore di trend al quale potrebbe aggiungersi anche una componente stagionale. Approfondiamo con una visualizzazione della funzione di autocorrelazione (fig.2), dalla quale osserviamo un andamento ondulato con rialzi attorno ai multipli di 12 *Lag*.

Ci convinciamo ulteriormente della presenza di stagionalità annuale se esaminiamo la funzione di autocorrelazione al netto del trend (fig.3), la quale presenta picchi proprio a multipli di 12 *Lag*.

fig.3

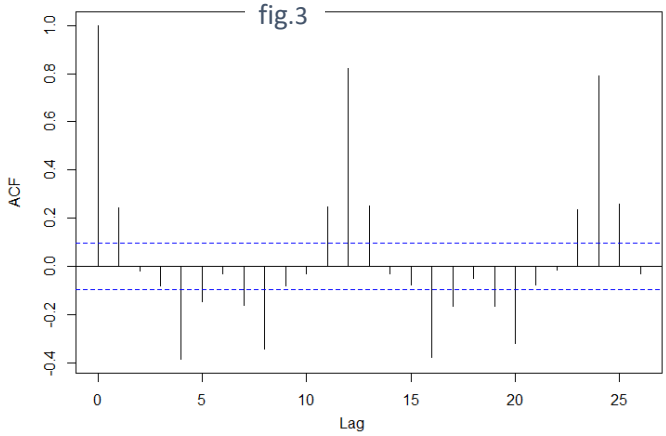
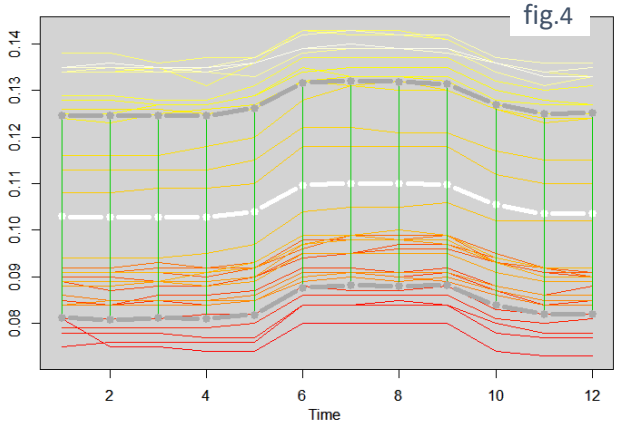


fig.4



Sovrapponendo anni diversi (fig.4) notiamo somiglianze evidenti negli andamenti, con valori che tendono ad aumentare a maggio per poi tornare a regime entro la fine di ottobre. Dalla colorazione degli anni di quest'ultimo grafico ribadiamo anche la ovvia componente di trend, della quale ci eravamo subito accorti. Dopo queste considerazioni possiamo passare alla decomposizione della serie.

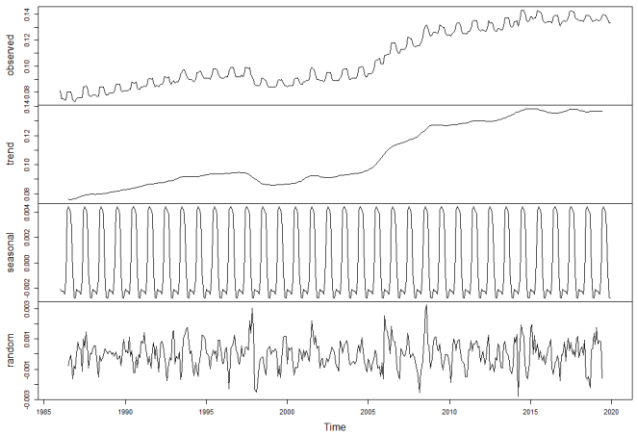
DEOCOMPOSIZIONE

L'analisi preliminare che abbiamo fatto potrebbe suggerire un modello di decomposizione additiva. In ogni caso andiamo ad esplorare anche quella moltiplicativa e successivamente i modelli STL additivo e moltiplicativo.

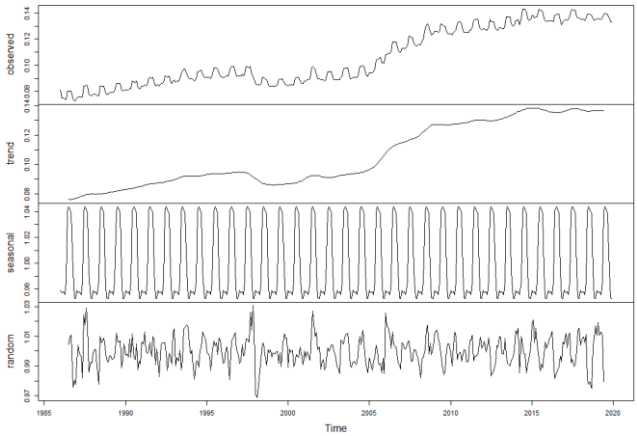
Decomposizione Additiva – Il modello di decomposizione additiva si comporta discretamente. Il rumore ha lo stesso ordine di grandezza della stagionalità, sembra però presentare lievi picchi in corrispondenza di variazioni sostanziali di trend. Quest'ultimo fenomeno, fortunatamente, non è molto marcato, se lo fosse stato avremmo pensato ad una stagionalità di tipo moltiplicativo.

Decomposizione Moltiplicativa – Tramite un primo confronto del rumore catturato da questo modello e quello catturato dal modello additivo non osserviamo sostanziali differenze. Anche sovrapponendo la stagionalità arriviamo ad un'analoga considerazione, le differenze sono irrilevanti.

Decomposizione Additiva:



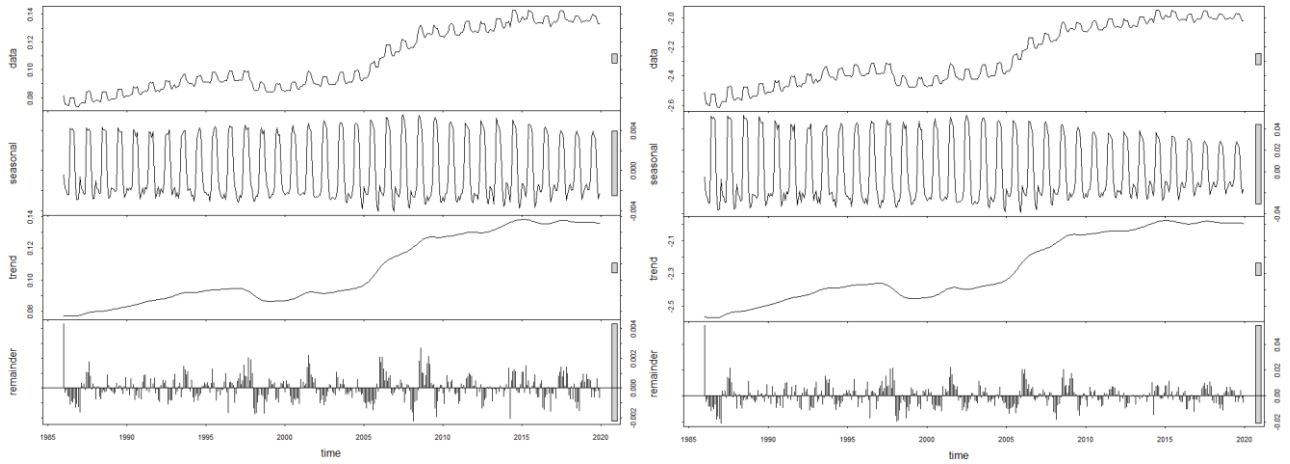
Decomposizione Moltiplicativa:



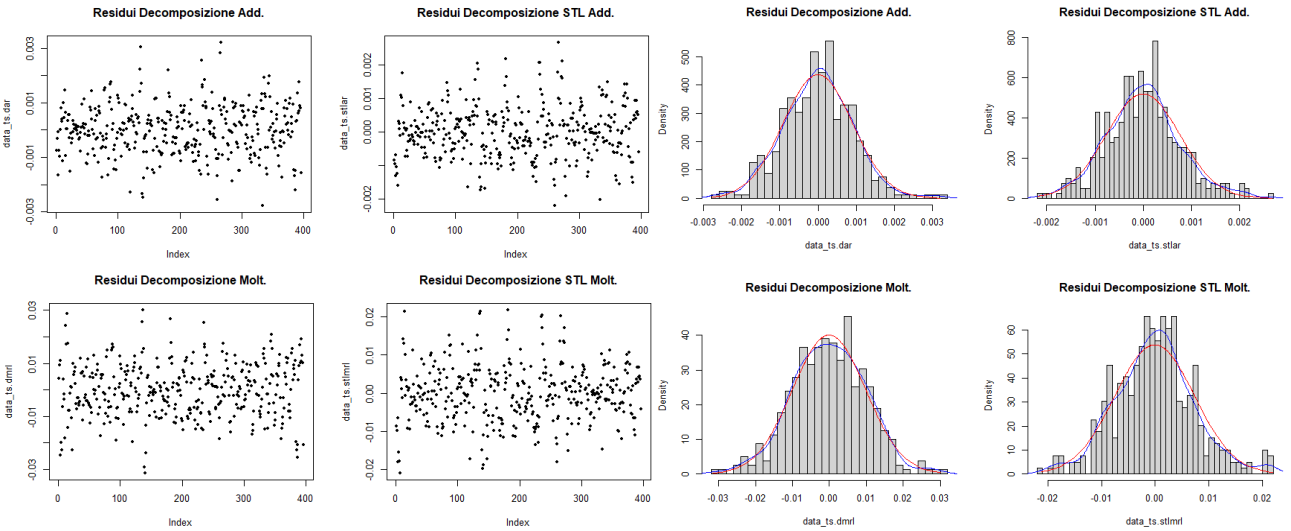
Decomposizione STL Additiva e Moltiplicativa – Tramite questi due modelli non otteniamo risultati particolarmente diversi, sia tra di loro che rispetto ai modelli precedenti. I trend sembrano quasi

perfettamente sovrapponibili, mentre per quanto riguarda la stagionalità abbiamo sì delle 'modulazioni' ma non particolarmente marcate.

Decomposizione STL Additiva: *Decomposizione STL Moltiplicativa:*



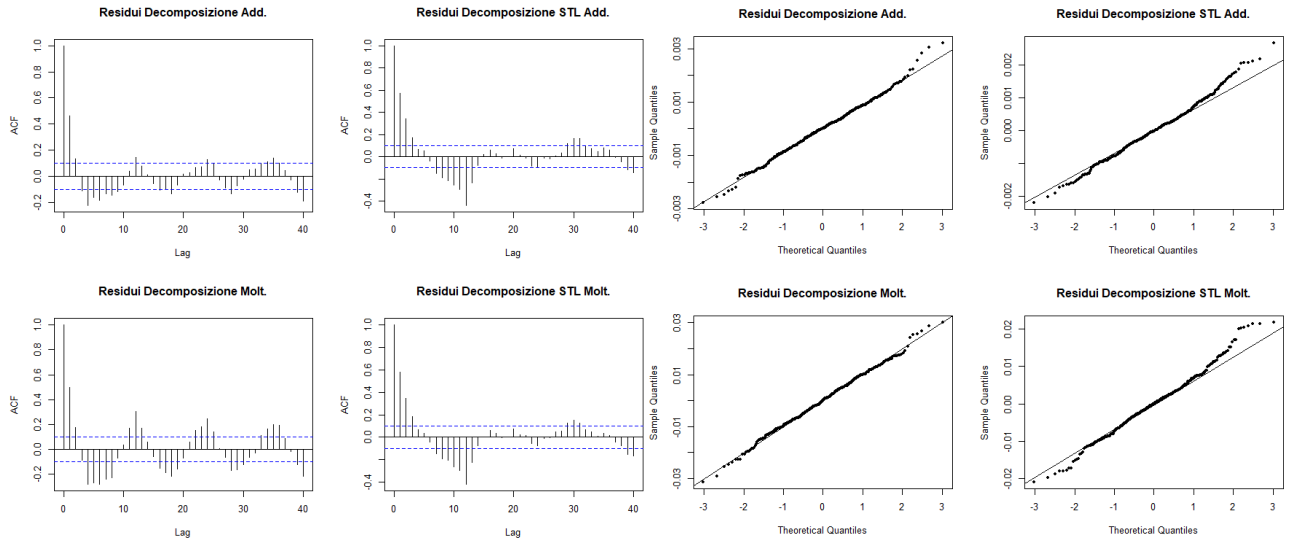
Le differenze delle decomposizioni non sono particolarmente apprezzabili, di conseguenza non riusciamo a preferire un modello rispetto agli altri. Considerando però le analisi preliminari, probabilmente la decomposizione da scegliere è quella additiva. Per prendere una decisione analizziamo anche i residui dei quattro metodi nel dettaglio.



Per come sono distribuiti i residui, dei diversi modelli, non riusciamo a dare immediatamente una preferenza. Quelli della decomposizione additiva sembrano presentare lievi anomalie intorno all'indice 200; in ogni caso, assieme a quelli del modello moltiplicativo, appaiono i migliori. Queste osservazioni si rafforzano se andiamo ad esaminare gli istogrammi e le distribuzioni a confronto con le distribuzioni gaussiane. Integriamo anche con i grafici quantile-quantile e con la funzione di autocorrelazione, che ci permette di apprezzare la struttura residua.

Eseguiamo anche il test di Shapiro-Wilk, che da esito negativo con il solo modello STL moltiplicativo. Mentre per quanto riguarda le varianze, i quattro modelli sono paragonabili.

Da queste analisi concludiamo che le decomposizioni con i residui migliori sono quelle di carattere additivo e moltiplicativo. In termini di struttura residua il modello additivo è migliore dell'altro; quindi, utilizzeremo quest'ultimo in supporto alle analisi di previsione.



PREVISIONE

HOLT WINTERS

Con Holt-Winters abbiamo due diverse possibilità di utilizzo, possiamo impostare manualmente i parametri oppure lasciare quelli calcolati automaticamente dall'algoritmo. Quest'ultimi sono scelti in modo da minimizzare lo scarto quadratico dei residui, nel nostro caso sono alpha = 0.78, beta = 0.02 e gamma = 0.81.

Cerchiamo di ottimizzare il metodo andando a cercare manualmente dei parametri migliori. Possiamo impostare anche l'intercetta iniziale e la pendenza iniziale, lo facciamo scegliendoli con una regressione lineare sui primi due anni. Utilizziamo a questo punto il metodo di validazione per confrontare modelli di Holt-Winters con parametri diversi. Per tentativi non si riescono a trovare dei parametri nettamente migliori di quelli scelti dal software.

Inoltre, possiamo anche esplorare Holt-Winters con stagionalità moltiplicativa anziché additiva, ma in seguito alle analisi precedenti ipotizziamo che l'opzione additiva sia la migliore. Infatti, confrontando i due modelli tramite validazione non si ottengono risultati significativamente favorevoli alla stagionalità moltiplicativa. Continuiamo quindi le analisi con il modello additivo e parametri scelti automaticamente.

Sovrapponendo la stagionalità di Holt-Winters e quella della decomposizione (fig.5) notiamo che le differenze sono minime, il che rafforza le nostre analisi dal punto di vista della coerenza.

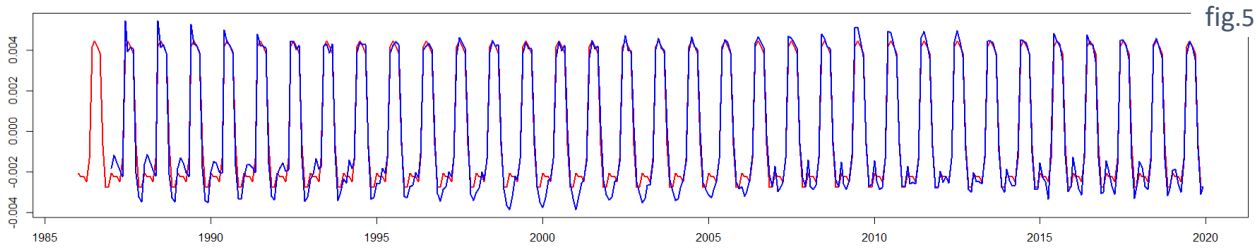


fig.5

Per capire quanto il modello sia incerto sulle previsioni possiamo fare un'analisi preliminare dei residui. Vediamo che non sembrano esserci criticità sostanziali nella distribuzione (fig.6), tranne qualche valore che si distacca particolarmente dalla media. Anche per quanto riguarda la rappresentazione tramite istogramma (fig.7) le considerazioni sono le medesime, osserviamo una buona aderenza alla distribuzione gaussiana. Nel grafico quantile-quantile (fig.8) abbiamo un addensamento sulla diagonale che si estende da -2 a 2, il che va bene. Possiamo anche esaminare i residui tramite la funzione di autocorrelazione (fig.9), sono presenti dei picchi fuori dalle bande ma comunque niente di eccessivo, quindi accettabile.

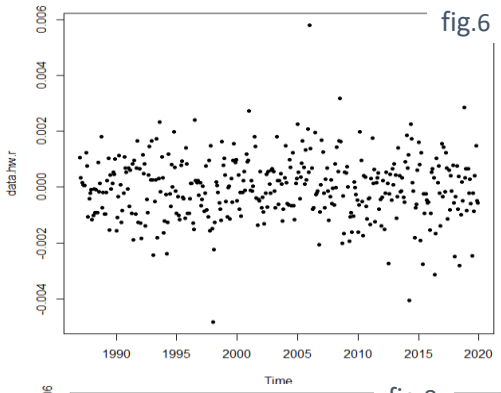


fig.6

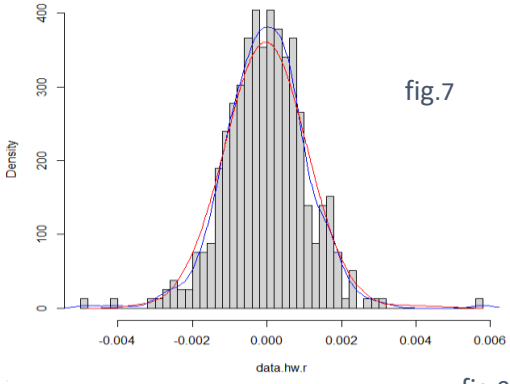


fig.7

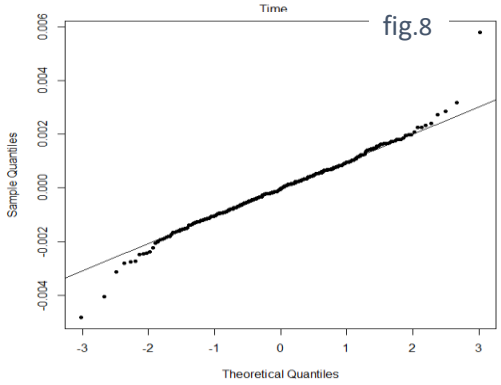


fig.8

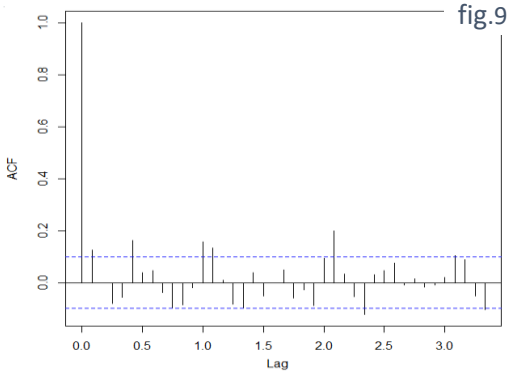


fig.9

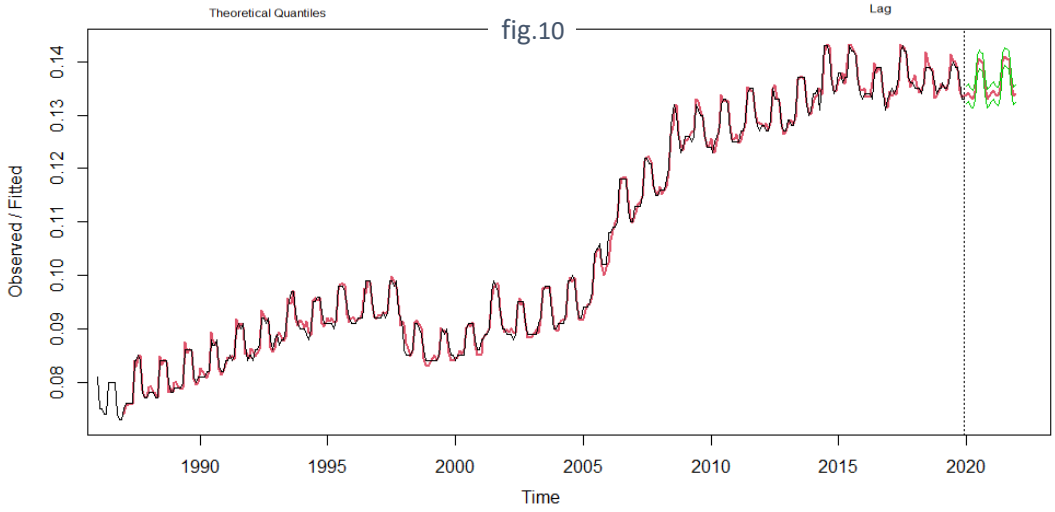


fig.10

Successivamente all'esito positivo dell'analisi sui residui, visualizziamo la predizione prodotta da Holt-Winters (fig.10) con una stima non parametrica al 95% delimitata dalle bande in verde.

AUTOREGRESSIONE

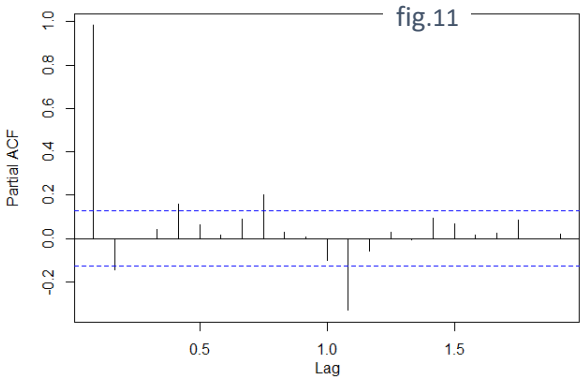
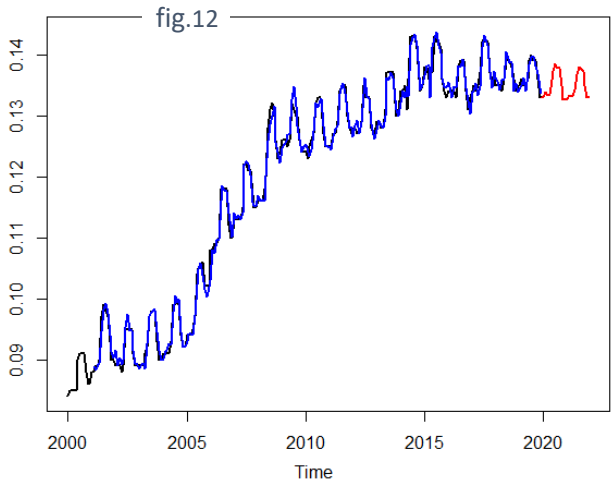
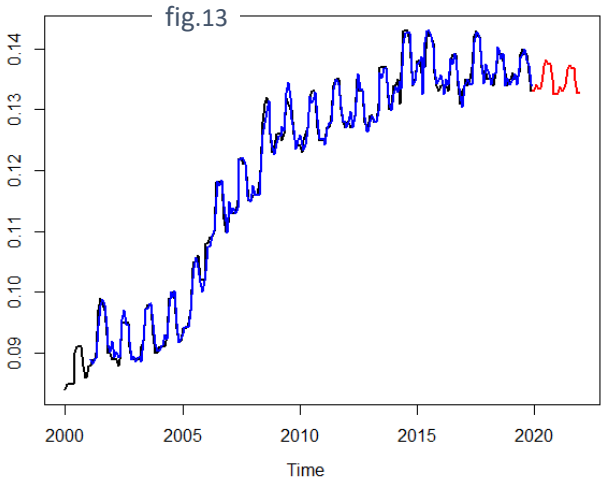


fig.11

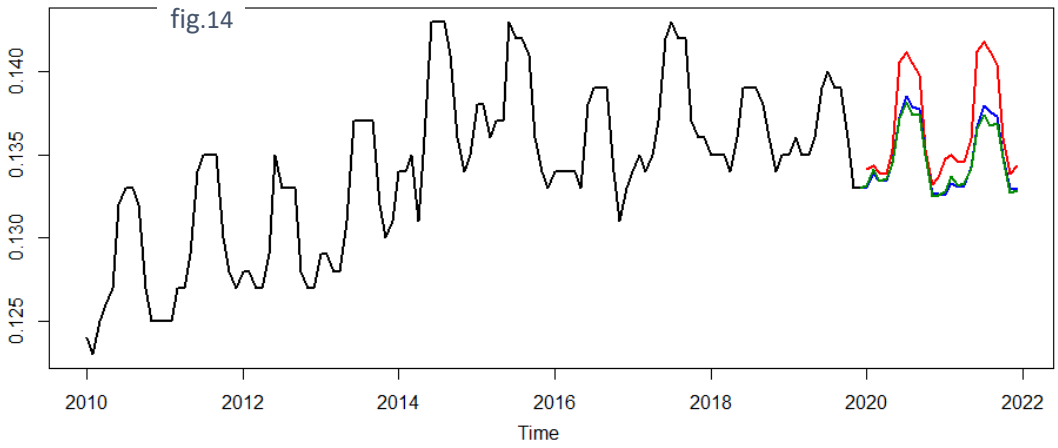
Iniziamo l'analisi dei modelli autoregressivi con il grafico della funzione di autocorrelazione parziale (fig.11). Da quest'ultimo notiamo una dipendenza fino a 13 Lag. Con questa informazione possiamo creare un modello regressivo con 13 fattori in ingresso. Tramite riduzione del modello scremiamo i fattori di ingresso preferendo quelli con i p-value più bassi: il primo, il secondo e il tredicesimo. La differenza tra la percentuale di varianza spiegata del modello completo e quella del ridotto è irrilevante; entrambe superano il 99%.



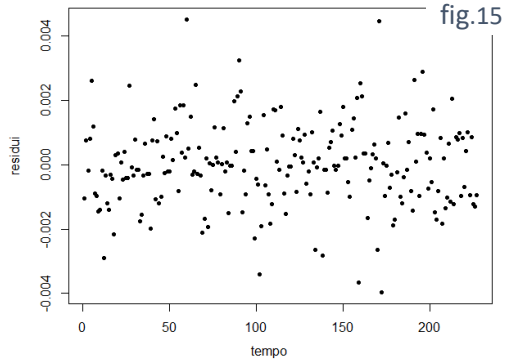


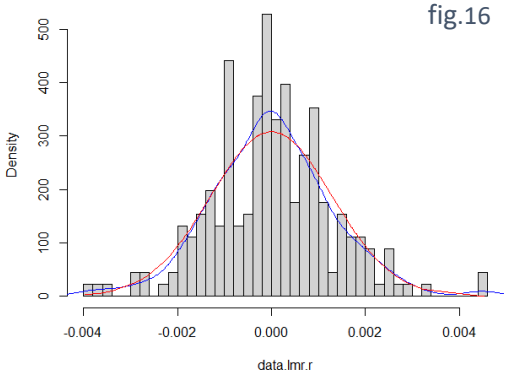
Visualizziamo le predizioni dei due modelli, prima quella del completo (fig.12) e poi quella del ridotto (fig.13).

Come potevamo aspettarci, le differenze tra le due sono minime; vediamo anche un paragone con la previsione del modello Holt-Winters (fig.14). Quello che osserviamo sono due andamenti leggermente diversi. Infatti, il modello Holt-Winters propone un trend lievemente positivo a contrasto con quelli che sembrerebbero lievemente negativi dei modelli regressivi. Per quanto riguarda la stagionalità notiamo che Holt-Winters sovrastima rispetto agli altri due modelli che si mantengono più bassi.



Dopo aver appurato che le differenze tra i due modelli regressivi sono minime, preferiamo il modello ridotto rispetto all'altro, questo per mantenere il numero di fattori d'ingresso contenuto. Eseguiamo l'analisi dei residui. Da una prima occhiata non sembrano esserci particolari anomalie (fig.15), purtroppo però la distribuzione non segue perfettamente la distribuzione gaussiana e l'istogramma non è ottimale (fig.16).





Per quanto riguarda il grafico quantile-quantile (fig.17) abbiamo un'aderenza discreta; mentre il grafico della funzione di autocorrelazione mette in evidenza una struttura residua con un picco negativo a 12 Lag. Se svolgiamo il test di Shapiro-Wilk otteniamo un p-value uguale a 0.053, appena accettabile.

fig.17

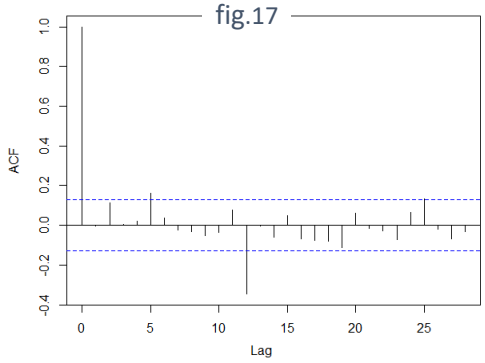
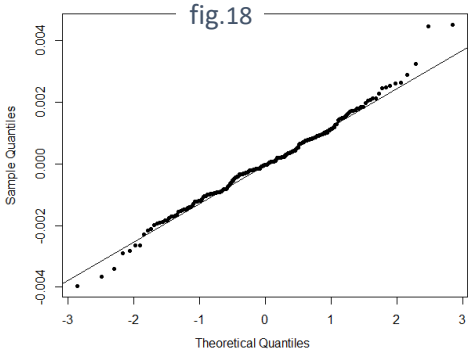


fig.18

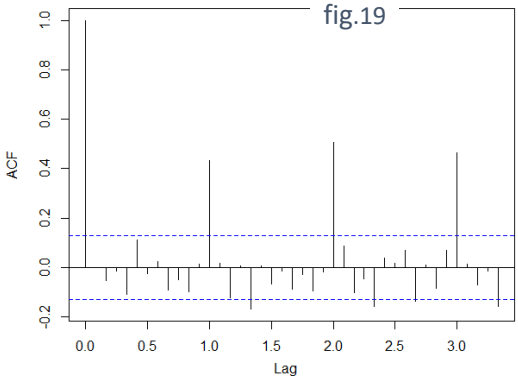


I residui non sono ottimi ma possiamo valutarli ammissibili.

MINIMI QUADRATI

Dall'algoritmo che implementa il modello autoregressivo con il metodo dei Minimi Quadrati emerge la dipendenza di 19 Lag. Diversa da quella che avevamo supposto con la funzione di autocorrelazione parziale.

fig.19



Esaminiamo questo metodo anziché Yule-Walker perché dopo una rapida analisi dei residui di quest'ultimo emergono criticità dalla funzione di autocorrelazione (fig.19), in particolare una forte struttura residua.

Tramite una prima rappresentazione (fig.20), i residui del modello con il metodo dei Minimi Quadrati non sembrano essere disastrosi.

L'aderenza della distribuzione a quella gaussiana non è ottimale (fig.21), così come anche il grafico quantile-quantile (fig.22). Confrontando i residui del metodo Yule-Walker con quelli dei Minimi Quadrati tramite la funzione di autocorrelazione notiamo un netto miglioramento. Consideriamo questi risultati accettabili.

fig.20

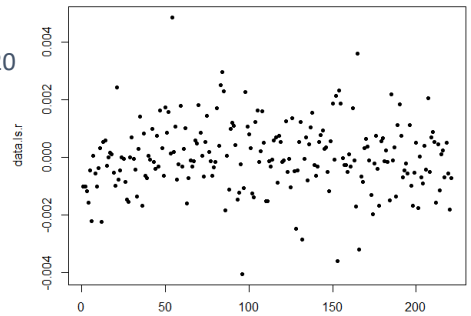


fig.21

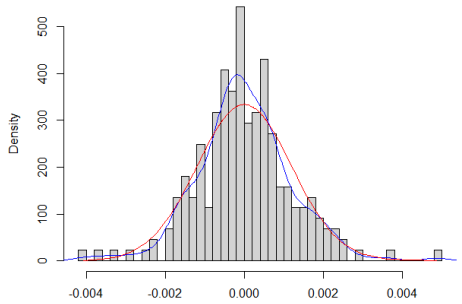


fig.22

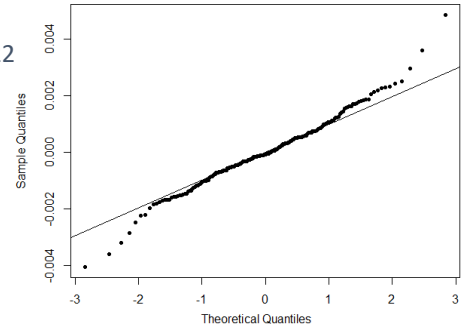
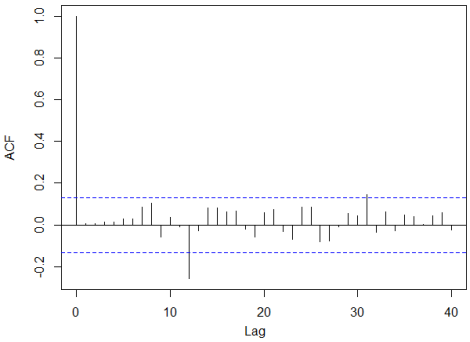
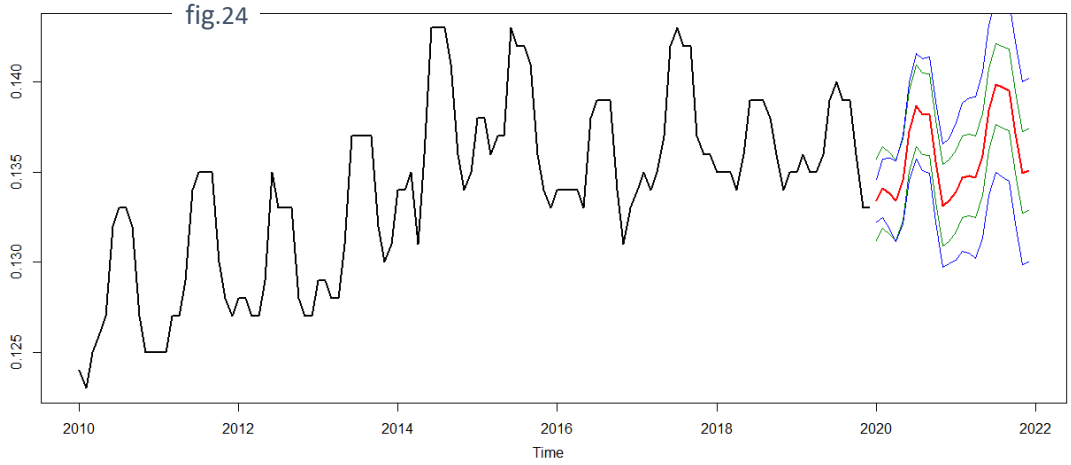


fig.23



Confrontando il modello dei Minimi Quadrati con Holt-Winters tramite autovalidazione notiamo che gli errori hanno lo stesso ordine di grandezza, possiamo procedere con la visualizzazione della previsione. Nella fig.24 andiamo a rappresentare, oltre che la previsione in rosso, anche la stima empirica dell'incertezza (in verde) e la stima parametrica dell'incertezza (in blu).

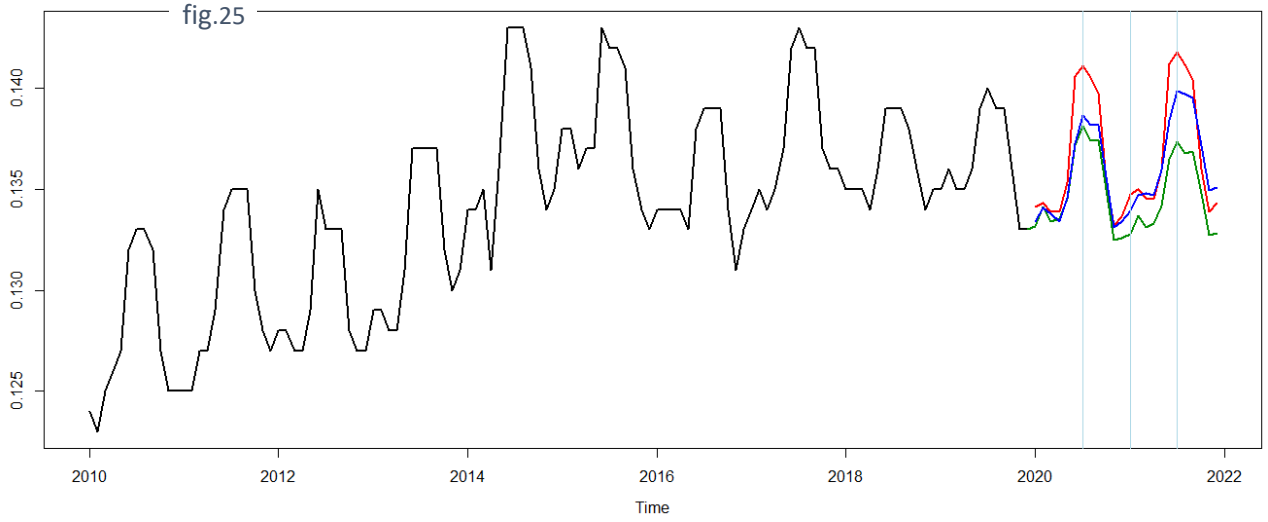
fig.24



CONCLUSIONI

Per concludere mettiamo a confronto graficamente le previsioni dei diversi modelli analizzati (fig.25). Vediamo in rosso la previsione di Holt-Winters, in verde quella prodotta tramite modello autoregressivo ridotto ed infine quella del modello autoregressivo dei Minimi Quadrati in blu. Le differenze tra le previsioni dei modelli non sono estremamente marcate. Notiamo che quella dei Minimi Quadrati sembra mantenersi intermedia tra le altre due. In realtà quest'ultima ha un trend lievemente maggiore della previsione di Holt-Winters; infatti, vediamo che a fine 2022 suggerisce dei valori più alti delle altre.

fig.25



Ricollegandoci al problema aziendale di partenza possiamo dire che si potrebbero minimizzare i costi concentrando la produzione durante i mesi di inizio e di fine anno. In particolare, il prezzo della corrente aumenta nei mesi di giugno, luglio, agosto e settembre, questi mesi sono quindi da evitare volendo limitare le spese relative all'energia elettrica. Queste considerazioni sono indipendenti dal modello predittivo che decidiamo di utilizzare.

Se volessimo portare avanti degli studi quantitativi sui costi potrebbe avere senso esaminare tutte e tre le predizioni e magari farne una media.

APPENDICE

```
data<-read.csv("C:/Users/matte/OneDrive/Desktop/STATISTICA/Terza_Relazione/AveragePriceElectricity.csv", stringsAsFactors=F)
data_num<-as.numeric(data[87:494,2]) #1986 al 2019
#CERCO STAGIONALITà e IL PERIODO
plot(data_num,type="l")
acf(data_num)
acf(diff(data_num))
#confrontiamo gli anni tra di loro
par(bg="lightgrey")
# andamenti in anni diversi
m_data=matrix(data_num,12,34)
ts.plot(m_data,col=heat.colors(34))
# andamento medio annuale
data.m=rowMeans(m_data)
lines(data.m,pch=20,type="b",lwd=5,col="white")
# bande empiriche
data.sd=vector("numeric",12)
for(i in 1:12){
  data.sd[i]=sd(m_data[i,])
}
arrows(1:12,data.m-data.sd,1:12,data.m+data.sd,length=0.02,angle=90,code=3,col="green3")
lines(data.m+data.sd,type="b",pch=20,col="darkgray",lwd=5)
lines(data.m-data.sd,type="b",pch=20,col="darkgray",lwd=5)
par(bg="white")
#Impostiamo periodo 12
data_ts<-ts(data_num,frequency=12,start=1986)
plot(data_ts)
#DECOMPOSIZIONE ------------------------------------------------------------------------
#decomposizione additiva
data_ts.da = decompose(data_ts)
plot(data_ts.da)
#decomposizione moltiplicativa
data_ts.dm = decompose(data_ts, type="multiplicative")
plot(data_ts.dm)
#ovviamnete il trend è lo stesso, confrontiamo prima le stagionalità e poi il rumore
plot(data_ts.da$seasonal)
lines(mean(data_ts.dm$trend,na.rm=T)*(data_ts.dm$seasonal-1),col="red")
plot(data_ts.da$random)
lines(mean(data_ts.dm$trend,na.rm=T)*(data_ts.dm$random-1),col="red")
#ESAMINIAMO I RESIDUI ##
#RESIDUI Modello Add
#per semplicità eliminiamo i termini NA
data_ts.dar = as.vector(window(data_ts.da$random,c(1986,7),c(2019,6)))
plot(data_ts.dar,pch=20)
#calcoliamo la varianza spiegata
var(data_ts.dar)/var(window(data_ts,c(1986,7),c(2019,6)))
#cerchiamo struttura con acf
acf(data_ts.dar,40, main="Residui Decomposizione Add.")
#confronto con la distribuzione normale
layout(t(1:2))
hist(data_ts.dar,40,freq=F, main="Residui Decomposizione Add.")
lines(density(data_ts.dar),col="blue")
lines(sort(data_ts.dar),dnorm(sort(data_ts.dar),mean(data_ts.dar),sd(data_ts.dar)),col="red")
qqnorm(data_ts.dar, pch=20, main="Residui Decomposizione Add.")
qqline(data_ts.dar)
layout(1)
shapiro.test(data_ts.dar)
#RESIDUI Modello Mol
data_ts.dmr=as.vector(window(data_ts.dm$random,c(1986,7),c(2019,6)))
data_ts.dmrl=log(data_ts.dmr)
plot(data_ts.dmrl,pch=20)
#la varianza spiegata
var(data_ts.dmrl)/var(window(log(data_ts),c(1986,7),c(2019,6)))
#cerchiamo struttura con acf
acf(data_ts.dmrl,40, main="Residui Decomposizione Molt.")
#confronto con la distribuzione normale
layout(t(1:2))
hist(data_ts.dmrl,40,freq=F, main="Residui Decomposizione Molt.")
lines(density(data_ts.dmrl),col="blue")
lines(sort(data_ts.dmrl),dnorm(sort(data_ts.dmrl),mean(data_ts.dmrl),sd(data_ts.dmrl)),col="red")
qqnorm(data_ts.dmrl, pch=20, main="Residui Decomposizione Molt.")
qqline(data_ts.dmrl)
layout(1)
shapiro.test(data_ts.dmrl)
#STL----
#decomposizione stl per stagionalità non uniforme additiva
data_ts.stla = stl(data_ts, s.window=7)
plot(data_ts.stla)
#decomposizione stl per stagionalità non uniforme moltiplicativa
data_ts.stlm = stl(log(data_ts), s.window=7)
plot(data_ts.stlm)
#RESIDUI Modello STL Add
#per confrontali con gli altri residui togliamo gli stessi valori
data_ts.stlar = as.vector(window(data_ts.stla$time.series[,3],c(1986,7),c(2019,6)))
plot(data_ts.stlar,pch=20)
#calcoliamo la varianza spiegata
var(data_ts.stlar)/var(window(data_ts,c(1986,7),c(2019,6)))
#cerchiamo struttura con acf
acf(data_ts.stlar,40, main="Residui Decomposizione STL Add.")
#confronto con la distribuzione normale
layout(t(1:2))
hist(data_ts.stlar,40,freq=F, main="Residui Decomposizione STL Add.")
```

```
lines(density(data_ts.stlar),col="blue")
lines(sort(data_ts.stlar),dnorm(sort(data_ts.stlar),mean(data_ts.stlar),sd(data_ts.stlar)),col="red")
qqnorm(data_ts.stlar, pch=20, main="Residui Decomposizione STL Add.")
qqline(data_ts.stlar)
layout(1)
shapiro.test(data_ts.stlar)
#RESIDUI Modello STL Mol
#per confrontali con gli altri residui togliamo gli stessi valori
data_ts.stlmr = as.vector(window(data_ts.stlm$time.series[,3],c(1986,7),c(2019,6)))
data_ts.stlmrl=log(data_ts.stlmr+1) #bisogna sommare +1 perchè stavolta i residui sono concentrati in 0
plot(data_ts.stlmrl,pch=20)
#calcoliamo la varianza spiegata
var(data_ts.stlmrl)/var(log(window(data_ts,c(1986,7),c(2019,6))))
#cerchiamo struttura con acf
acf(data_ts.stlmrl,40, main="Residui Decomposizione STL Molt.")
#confronto con la distribuzione normale
layout(t(1:2))
hist(data_ts.stlmrl,40,freq=F, main="Residui Decomposizione STL Molt.")
lines(density(data_ts.stlmrl),col="blue")
lines(sort(data_ts.stlmrl),dnorm(sort(data_ts.stlmrl),mean(data_ts.stlmrl),sd(data_ts.stlmrl)),col="red")
qqnorm(data_ts.stlmrl, pch=20, main="Residui Decomposizione STL Molt.")
qqline(data_ts.stlmrl)
layout(1)
shapiro.test(data_ts.stlmrl)

#PREVISIONE-----------------------------------------------------------------------------------------
#---------------------------------------------------------------------------------------------------
data_ts<-ts(data_num,frequency=12,start=1986)
inizio=2000
data_ts = window(data_ts, start = c(inizio, 1))
#Holt-Winters
data.hw=HoltWinters(data_ts)
plot(data.hw, lwd=2)
data.hw$alpha #0.7816148
data.hw$beta #0.02136604
data.hw$gamm #0.808237
#confrontro tra hw e dec.add
ts.plot(data_ts.da$seasonal,data.hw$fitted[,4],col=c('red','blue'),type="l", lwd=2)
plot(data.hw$fitted) #restituisce la sotto-struttura fitted, dove level sono le intercette e xhat sono i valori stimati
#con valori auto
plot(data.hw,type="l",lwd=2)
lines(data.hw$fitted[,1],col="blue")
#l.start è la intercetta iniziale, e b.start è la pendenza iniziale
#--- per cercare i parametri ----
x=1:24
coefficients(lm(data_ts[1:24]~x))
#con valaori ottimizzati manualmente
data.hwo = HoltWinters(data_ts, alpha=0.7, beta=0.1, gamma=0.5, l.start=0.075, b.start=0)
plot(data.hwo)
ts.plot(data_ts.da$seasonal,data.hwo$fitted[,4],col=c('red','blue'),type="l", lwd=2)
#validazione del modello tra quello automatico e manuale
nt=20 # numero di test set
ft=1 # unità di tempo nel futuro su cui valutare la previsione
n=length(data_ts) # numero totale di anni
idt=start(data_ts) # data di inizio della serie
fdt=end(data_ts)   # data di fine della serie
pdt=frequency(data_ts) # periodo della serie
err_data0=0
err_data1=0
(n-nt-ft):(n-ft-1)
for(j in (n-nt-ft):(n-ft-1)){
  # costruzione di train e test
  train=window(data_ts,idt,ts_data(idt,pdt,j))
  future=ts_data(idt,pdt,j+ft)
  test=window(data_ts,future,future)
  # HW standard
  train.data0=HoltWinters(train)
  err_data0=err_data0+sum((as.numeric(test)-as.numeric(predict(train.data0,ft)))^2)
  # HW parametri personalizzati
  train.data1=HoltWinters(train,alpha=0.7, beta=0.2, gamma=0.5, l.start=0.075, b.start=0)
  err_data1=err_data1+sum((as.numeric(test)-as.numeric(predict(train.data1,ft)))^2)
}
err_data0/nt
err_data1/nt
#HW Moltipllicativo
data.hwm = HoltWinters(data_ts, seasonal="multiplicative")
plot(data.hwm)
#confronto con HW moltiplicativo
nt=20 # numero di test set
ft=1 # unità di tempo nel futuro su cui valutare la previsione
n=length(data_ts) # numero totale di anni
idt=start(data_ts) # data di inizio della serie
fdt=end(data_ts)   # data di fine della serie
pdt=frequency(data_ts) # periodo della serie
err_data0=0
err_data1=0
for(j in (n-nt-ft):(n-ft-1)){
  # costruzione di train e test
  train=window(data_ts,idt,ts_data(idt,pdt,j))
  future=ts_data(idt,pdt,j+ft)
  test=window(data_ts,future,future)
  # HW standard
  train.data0=HoltWinters(train)
  err_data0=err_data0+sum((as.numeric(test)-as.numeric(predict(train.data0,ft)))^2)
```

```
  # HW parametri personalizzati
  train.data1=HoltWinters(train,seasonal="multiplicative")
  err_data1=err_data1+sum((as.numeric(test)-as.numeric(predict(train.data1,ft)))^2)
}
err_data0/nt
err_data1/nt
#Studiamo preliminarmente i residui---------------- HW standard
data.hw.r=resid(data.hw)
start(data_ts)
end(data_ts)
start(data.hw.r)
end(data.hw.r)
var(data.hw.r)/var(window(data_ts,c(inizio+1,1)))
layout(t(1:2))
plot(data.hw.r,type="p",pch=20)
plot(data.hw$fitted[,1],data.hw.r,pch=20)
layout(1)
acf(data.hw.r,40)
layout(t(1:2))
hist(data.hw.r,40,freq=F)
lines(density(data.hw.r), col="blue")
lines(sort(data.hw.r),dnorm(sort(data.hw.r),mean(data.hw.r),sd(data.hw.r)),col="red")
qqnorm(data.hw.r,pch=20)
qqline(data.hw.r)
layout(1)
shapiro.test(data.hw.r)
#previsione
plot(data.hw,predict(data.hw,24),main="Previsione a 24 mesi", lwd=2)
lines(predict(data.hw,24)+quantile(data.hw.r,0.05),col="green3", lwd=1.5)
lines(predict(data.hw,24)+quantile(data.hw.r,0.95),col="green3", lwd=1.5)
#AUTOREGRESSIONE------------------------------------------------------------
pacf(data_ts)
L = length(data_ts)
l = 13  # numero di lag in ingresso
mdata = matrix(nrow = L - l, ncol = l + 1)
for (i in 1:(l + 1)) {
  mdata[, i] = data_ts[i:(L - l - 1 + i)]
}
mdata <- data.frame(mdata)
data.lmc <- lm(X14 ~ ., data = mdata)  # X14 perché 13 lag in ingresso
summary(data.lmc)
data.lmr <- lm(X14 ~ X1 + X2 + X13, data = mdata)
summary(data.lmr)
#predizione modello ridotto
anni = 2
L = length(data_ts)
ptr = rep(0, L + 12 * anni)
ptr[1:L] = data_ts
for (i in 1:(12 * anni)) {
  ptr[L + i] = coef(data.lmr)%*%c(1, ptr[L + i - 13], ptr[L + i - 12], ptr[L + i - 1]) #fattori inclusi X1
}
data.lmr.pt = ts(ptr, frequency = 12, start = c(inizio, 1))
data.lmr.a = window(data_ts, c(inizio+1, 2)) - resid(data.lmr) #analisi del modello regressivo
ts.plot(data_ts, data.lmr.a, window(data.lmr.pt, c(2019, 12)), col = c("black","blue", "red"),lwd=2)

#predizione modello completo
anni = 2
ptc = rep(0, L + 12 * anni)
ptc[1:L] = data_ts
for (i in 1:(12 * anni)) {
  ptc[L + i] = coef(data.lmc) %*% c(1, rev(ptc[L + i - 1:l]))
}
data.lmc.pt = ts(ptc, frequency = 12, start = c(inizio, 1))
data.lmc.a = window(data_ts, c(inizio+1, 2)) - resid(data.lmc)
ts.plot(data_ts, data.lmc.a, window(data.lmc.pt, c(2019, 12)), col = c("black","blue", "red"),lwd=2)
#primo confronto grafico dei due modelli autoregressivi tra loro e con la previsione di Holt-Winters
data.hw = HoltWinters(data_ts) #<-------------- qui uso holt winters non ottimizzato
data.lm.ptc = window(data.lmc.pt, c(2019, 12))
data.lm.ptr = window(data.lmr.pt, c(2019, 12))
data.hw.pt = predict(data.hw, 24)
ts.plot(data_ts, data.hw.pt, data.lm.ptc, data.lm.ptr, col = c("black", "red", "blue","green4"),lwd=2)
#zoom del confronto dal 2010 in poi
ts.plot(window(data_ts, c(2010, 1)), data.hw.pt, data.lm.ptc, data.lm.ptr, col = c("black", "red", "blue","green4"),lwd=2)
# estrazione dei residui
data.hw.r = resid(data.hw)
data.lmc.r = resid(data.lmc)
data.lmr.r = resid(data.lmr)
# varianze non spiegate
var(data.hw.r)/var(window(data_ts, inizio+1))
var(data.lmc.r)/var(window(data_ts, inizio+1))
var(data.lmr.r)/var(window(data_ts, inizio+1))
# confronto grafico
layout(matrix(1:6, 2, 3, byrow = T))
plot(as.numeric(data.hw.r), pch = 20, main = "HW", xlab = "tempo", ylab = "residui")
plot(data.lmc.r, pch = 20, main = "AR completo", xlab = "tempo", ylab = "residui")
plot(data.lmr.r, pch = 20, main = "AR ridotto", xlab = "tempo", ylab = "residui")
plot(data.hw$fitted[, 1], data.hw.r, pch = 20, main = "HW", xlab = "stima", ylab = "residui")
plot(data.lmc.a, data.lmc.r, pch = 20, main = "AR completo", xlab = "stima",ylab = "residui")
plot(data.lmr.a, data.lmr.r, pch = 20, main = "AR ridotto", xlab = "stima", ylab = "residui")
layout(1)
# acf e pacf
layout(matrix(1:6, 2, 3, byrow = T))
acf(data.hw.r, 28)
acf(data.lmc.r, 28)
```

```
acf(data.lmr.r, 28)
pacf(data.hw.r, 28)
pacf(data.lmc.r, 28)
pacf(data.lmr.r, 28)
layout(1)
# frequenze
layout(t(1:3))
hist(data.hw.r, 20, freq = F, main = "HW")
lines(density(data.hw.r), col = "blue")
lines(sort(data.hw.r), dnorm(sort(data.hw.r), mean(data.hw.r), sd(data.hw.r)), col = "red")
hist(data.lmc.r, 20, freq = F, main = "AR completo")
lines(density(data.lmc.r), col = "blue")
lines(sort(data.lmc.r), dnorm(sort(data.lmc.r), mean(data.lmc.r), sd(data.lmc.r)),col = "red")
hist(data.lmr.r, 40, freq = F, main = "AR ridotto")
lines(density(data.lmr.r), col = "blue")
lines(sort(data.lmr.r), dnorm(sort(data.lmr.r), mean(data.lmr.r), sd(data.lmr.r)),col = "red")
layout(1)
# quantili
layout(t(1:3))
qqnorm(data.hw.r, pch = 20)
qqline(data.hw.r)
qqnorm(data.lmc.r, pch = 20)
qqline(data.lmc.r)
qqnorm(data.lmr.r, pch = 20)
qqline(data.lmr.r)
layout(1)
# test
shapiro.test(data.hw.r)
shapiro.test(data.lmc.r)
shapiro.test(data.lmr.r)
layout(1)

#Eseguiamo una autovalidazione dei tre modelli.
nt = 15  # numero di test set
ft = 1  # unità di tempo nel futuro su cui valutare la previsione
n = length(data_ts)  # numero totale di anni
idt = start(data_ts)  # data di inizio della serie
fdt = end(data_ts)  # data di fine della serie
pdt = frequency(data_ts)  # periodo della serie

err_hw = rep(0, nt)
err_lmc = rep(0, nt)
err_lmr = rep(0, nt)
for (j in (n - nt - ft):(n - ft - 1)) {
  # training e test set
  train = window(data_ts, idt, ts_data(idt, pdt, j))
  future = ts_data(idt, pdt, j + ft)
  test = window(data_ts, future, future)
  # HW
  train.hw = HoltWinters(train)
  err_hw[j - (n - nt - ft) + 1] = as.numeric(test) - as.numeric(predict(train.hw,ft)[ft])
  # AR
  L = length(train)
  l = 13  # numero di lag in ingresso
  mtrain = matrix(nrow = L - l, ncol = l + 1)
  for (i in 1:(l + 1)) {
    mtrain[, i] = train[i:(L - l - 1 + i)]
  }
  mtrain <- data.frame(mtrain)
  # AR completo
  train.lmc <- lm(X14 ~ ., data = mtrain)
  train.lmc.p = rep(0, L + ft)
  train.lmc.p[1:L] = train
  for (i in 1:ft) {
    train.lmc.p[L + i] = coef(train.lmc) %*% c(1, rev(train.lmc.p[L + i - 1:l]))
  }
  err_lmc[j - (n - nt - ft) + 1] = as.numeric(test) - train.lmc.p[L + ft]
  # AR ridotto
  train.lmr <- lm(X14 ~ X1 + X2 + X13, data = mtrain)
  train.lmr.p = rep(0, L + ft)
  train.lmr.p[1:L] = train
  for (i in 1:ft) {
    train.lmr.p[L + i] = coef(train.lmr) %*% c(1, train.lmr.p[L + i -13], train.lmr.p[L + i -12], train.lmr.p[L + i - 1])
  }
  err_lmr[j - (n - nt - ft) + 1] = as.numeric(test) - train.lmr.p[L + ft]
}
sum(err_hw^2)/nt
sum(err_lmc^2)/nt
sum(err_lmr^2)/nt
#Autoregressione con il metodo YULE-WALKER-----------------------------------
data.ar = ar(data_ts)
data.ar
ts.plot(data_ts, data_ts - data.ar$resid, col = c("black", "blue"),lwd = 2) #in blu le stime
#Predizione e incertezze
r = na.omit(data.ar$resid)
data.ar.pt = predict(data.ar, n.ahead = 24, se.fit = TRUE, level = 0.95)
data.ar.a = window(data_ts, start = c(2010, 1)) - r
ts.plot(window(data_ts, start = c(2010, 1)),data.ar.a, data.ar.pt$pred, col = c("black","blue", "red"), lwd = 2)
var(r)/var(window(data_ts, start = c(inizio+1, 3)))
layout(matrix(1:6, 2, 3, byrow = T))
plot(as.numeric(data.ar$resid), pch = 20)
plot(data.ar.a, data.ar$resid, pch = 20)
acf(r, 40)
pacf(r, 28)
```

```
hist(data.ar$resid, 40, freq = F)
lines(density(r), col = "blue")
lines(sort(r), dnorm(sort(r), mean(r), sd(r)), col = "red")
qqnorm(data.ar$resid, pch = 20)
qqline(data.ar$resid)
layout(1)
shapiro.test(data.ar$resid)
up = data.ar.pt$pred + quantile(r, 0.975)
lw = data.ar.pt$pred + quantile(r, 0.025)
ts.plot(data.ar.a, data.ar.pt$pred, col = c("black", "red"), lwd = 2)
lines(up, col = "blue", lwd = 2)
lines(lw, col = "blue", lwd = 2)
ts.plot(data.ar.a, data.ar.pt$pred, col = c("black", "red"), lwd = 2)
# non parametrico
lines(up, col = "blue", lwd = 2)
lines(lw, col = "blue", lwd = 2)
# parametrico
lines(data.ar.pt$pred - data.ar.pt$se, col = "green4", lwd = 2)
lines(data.ar.pt$pred + data.ar.pt$se, col = "green4", lwd = 2)
#confronto con holtwinters e modello ridotto
ts.plot(window(data_ts, c(2010, 1)), data.hw.pt, data.lm.ptr,data.ar.pt$pred, col = c("black", "red", "green4","blue"),lwd=2)
#Autoregressione con il metodo dei MINIMI QUADRATI
data.ls = ar(data_ts, method = "ols")
data.ls$order #da il lag
ts.plot(data_ts, data_ts - data.ls$resid, col = c("black", "blue"),lwd=2)
#residui
data.ls.r = as.double(na.omit(data.ls$resid))
data.ls.a = as.double(na.omit(data_ts - data.ls$resid))
#var(data.ls.r)/var(window(data_ts, start = c(1988, 3)))
layout(matrix(1:6, 2, 3, byrow = T))
plot(data.ls.r, pch = 20)
plot(data.ls.a, data.ls.r, pch = 20)
layout(1)
pacf(data.ls.r, 28)
hist(data.ls.r, 40, freq = F)
lines(density(data.ls.r), col = "blue")
lines(sort(data.ls.r), dnorm(sort(data.ls.r), mean(data.ls.r), sd(data.ls.r)), col = "red")
qqnorm(data.ls.r, pch = 20)
qqline(data.ls.r)
layout(1)
shapiro.test(data.ls.r)
data.ls.pt = predict(data.ls, n.ahead = 24, se.fit = TRUE, level = 0.95)
y.max = max(data.ls.pt$pred + quantile(data.ls.r, 0.975))
y.min = min(window(data_ts - data.ls$resid, 20107))
ts.plot(window(data_ts, 2007), data.ls.pt$pred,
        col = c("black", "red"), lwd = 2)
# stima empirica dell'incertezza
lines(data.ls.pt$pred + quantile(data.ls.r, 0.975), col = 'green4')
lines(data.ls.pt$pred + quantile(data.ls.r, 0.025), col = 'green4')
# stima parametrica dell'incertezza
lines(data.ls.pt$pred - data.ls.pt$se, col = "blue")
lines(data.ls.pt$pred + data.ls.pt$se, col = "blue")
#Confrontiamo il modello appena analizzato con Holt-Winters.
data.hw = HoltWinters(data_ts)
ts.plot(data_ts, data_ts - data.ls$resid, data.hw$fitted[, 1], col = c("black", "red", "blue"), lwd = 2)
# previsioni
data.hw.pt = predict(data.hw, 12)
ts.plot(window(data_ts, 2010), data.ls.pt$pred, data.hw.pt, col = c("black", "red", "blue"), lwd = 2)
#lines(data.ls.pt$pred - data.ls.pt$se, col = "green4", lwd = 2)
#lines(data.ls.pt$pred + data.ls.pt$se, col = "green4", lwd = 2)
#Confrontiamo i due metodi con l’autovalutazione.
nt = 20  # numero di test set
ft = 1  # unità di tempo nel futuro su cui valutare la previsione
n = length(data_ts)  # numero totale di anni
idt = start(data_ts)  # data di inizio della serie
fdt = end(data_ts)  # data di fine della serie
pdt = frequency(data_ts)  # periodo della seri
err_ls = rep(0, nt)
err_hw = rep(0, nt)
#confronto con tutti i modelli
ts.plot(window(data_ts, c(2010, 1)), data.hw.pt, data.lm.ptr, data.ls.pt$pred, col = c("black", "red", "green4","blue"),lwd=2)
abline(v=2020.5, col="lightblue")
abline(v=2021, col="lightblue")
abline(v=2021.5, col="lightblue")
```