

POLITECNICO DI TORINO

Master Degree course in Ingegneria Matematica 2022/2023

Apprendimento Statistico

Time Series

Racca Riccardo s315163 Filippo Scaramozzino s312856

Gennaio 2023

Abstract

Analisi di una serie temporale.

Contents

Ar	nalisi serie temporale
1.1	Introduzione
1.2	2 STL
1.3	ACF, PACF e Stazionarietà
	1.3.1 ACF
	1.3.2 PACF
1.4	AIC
1.5	6 ARIMA
	1.5.1 AR(1)
	1.5.2 AR(p)
	1.5.3 $MA(1)$
	1.5.4 $MA(q)$
	1.5.5 ARMA(p,q)
	1.5.6 Integrated Model
	1.5.7 ARI
	1.5.8 $IMA(1,8)$
	1.5.9 $ARIMA(p,d,q)$
	1.5.10 SARIMA
1.6	
	1.6.1 Analisi dei residui
	TEST

Chapter 1

Analisi serie temporale

1.1 Introduzione

Analizzeremo il data set 'elecequip' che descrive il numero di componenti elettronici prodotti in Europa dal Gennaio del 1996 a Marzo del 2012 normalizzato. Utilizzeremo il linguaggio R per procedere con le nostre analisi.

Esempio di una riga del dataset :

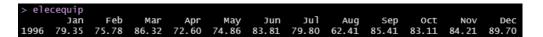


Figure 1.1. Una riga del dataset ottenuta tramite R

La serie temporale è descritta dal seguente grafico :

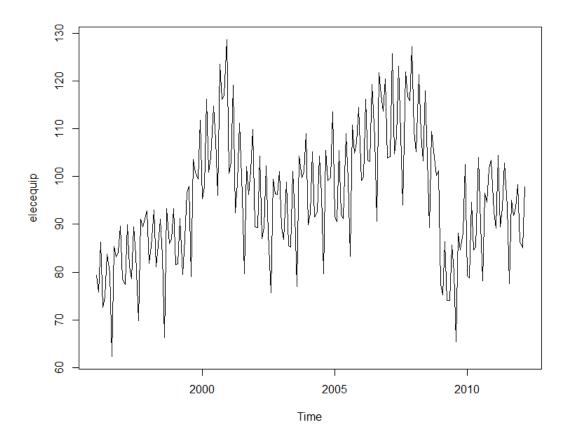


Figure 1.2. Grafico della serie

La nostra analisi si focalizzerà sul presentare le sue caratteristiche dal punto di vista teorico e trovare un modello in grado di descrivere il suo comportamento.

Procediamo con una analisi preliminare in cui andiamo a calcolare la media campionaria e la varianza campionaria calcolata con R.

 $\begin{array}{cc} Media & Varianza \\ 95.7 & 174.4 \end{array}$

1.2 STL

La prima analisi che siamo intenzionati a fare è la decomposizione STL.

Le serie temporali possono essere studiate attraverso l'analisi di alcune caratteristiche. Un metodo che può rivelarsi utile è quello di separare la serie temporale isolandola in diverse componenti.

Nello specifico andiamo ad analizzare: trend, seasonality, cycle e residual; definiamo brevemente questi termini.

- Il trend rappresenta l'andamento a lungo termine dei dati e non deve essere necessariamente lineare. Fornisce chiare spiegazioni sul fenomeno quando è monotono crescente o decrescente.
- La seasonality (stagionalità) rappresenta l'effetto che intervalli di tempo specifici e regolari hanno sulla serie temporale.
- Un cycle (ciclo) esibisce andamenti comuni durante il processo ma a frequenze non fissate.

I cicli e la stagionalità differiscono per frequenza (fissata per la stagionalità) e anche per la lunghezza dell'intervallo di interesse(maggiore per i cicli).

- Il residual sarà interpretato come l'errore della nostra decomposizione. Per la decomposizione della serie temporale abbiamo scelto il metodo STL.

Rappresentiamo la serie come la somma di tre componenti fondamentali :

$$y_t = S_t + T_t + R_t$$

dove:

- y_t è la serie
- S_t la componente stagionale
- T_t la componente del trend-cycle
- R_t la componente data dalla differenza tra la serie originale e la somma di S_t e T_t

Il risultato della decomposizione è mostrato in questo grafico:

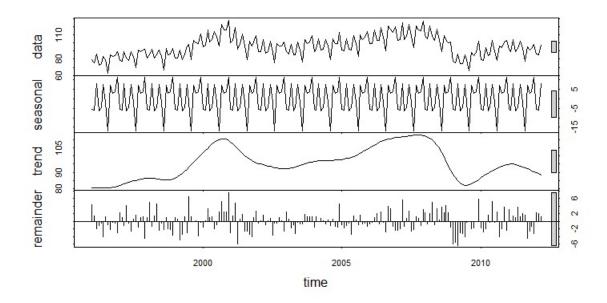


Figure 1.3. La decomposizione STL applicata alla serie

Analizzando la serie temporale notiamo come non ci sia particolare varianza nell'ampiezza dei dati ci aspettiamo quindi che un modello a stagionalità con effetto sulla media, e non sulla varianza, possa essere adeguato per questi dati.

Concentrandoci adesso sul trend osserviamo come non vi sia una monotonia costante, sembra infatti non esserci una tendenza dei dati nè ad aumentare nè al diminuire.

Non escludiamo che ci possa essere un comportamento ciclico, notiamo infatti come dal 2000 al 2001 vi sia un innalzamento rapido delle vendite che viene seguito da un periodo di diminuzione, questo comportamento sembra essere ripetuto a grandi linee dal periodo di tempo che va dal 2006 al 2008 dopo di cui si ha una sensibile diminuzione. Il trend descrive in maniera abbastanza accurata i dati analizzati, la stagionalità e il trend sembrano essere buoni parametri per ricostruire il modello, tuttavia il reminder non è trascurabile, è presente una leggera influenza ciclica su esso.

Queste analisi preliminari sui ruoli delle varie componenti saranno cruciali nell'applicazione e nella valutazione dei metodi futuri, perció a scopi successivi, evidenziamo come la stagionalitá non sia un fattore trascurabile e quindi sará necessario tenerne conto.

Concludiamo quindi che il modello scelto possa essere un buon modello descrittivo in cui è però presente un reminder non trascurabile.

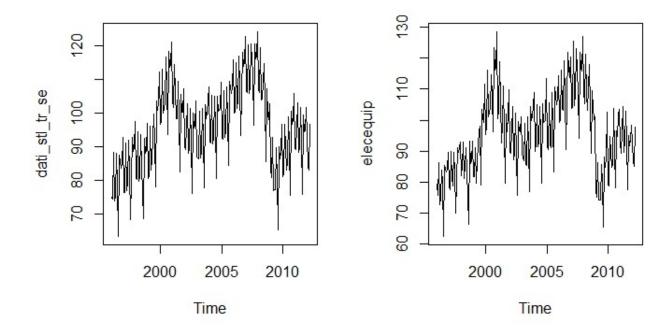


Figure 1.4. Confronto tra somma di trend e seasonal (a sinistra) e serie originale (a destra)

Per rafforzare quanto appena detto osserviamo come la decomposizione STL applicata al logaritmo della serie non conduca a sensibili miglioramenti, questo perchè l'azione del logaritmo di livellare i picchi della serie non viene sfruttata.

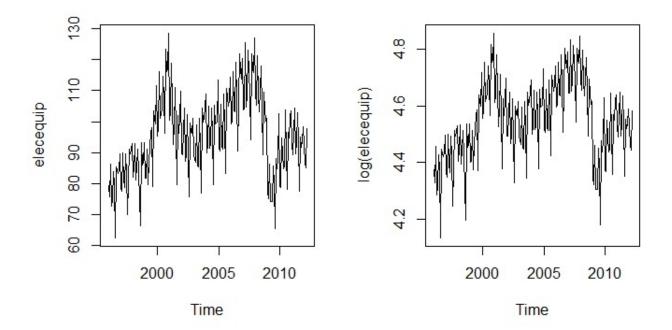


Figure 1.5. Confronto tra la serie originale (a sinistra) e il log della serie (a destra)

1.3 ACF, PACF e Stazionarietà

1.3.1 ACF

Quando un processo è stazionario debole di ordine due sappiamo che la funzione di autocovarianza dipende solamente dal lag temporale. come anche la sua funzione di autocorrelazione.

In questo caso è possibile effettuare delle stime tramite la covarianza campionaria e la correlazione campionaria. capire se un processo è stazionario o meno è difficile e quindi potrebbe non essere completamente giustificato l'utilizzo delle stime.

Quando rappresentiamo l'autocorrelazione in funzione del tempo questo viene chiamato correlogramma; in R può essere ottenuto con il comando Acf.

Ricordiamo che la autocorrelazione è definita tramite:

$$\rho(l) = \frac{Cov(X_{ti}, X_{ti+l})}{Var(X_{ti})} = \frac{\gamma(l)}{\gamma(0)}$$

dove l è il lag temporale, e γ è la funzione di autocorrelazione. Di seguito i risultati che otteniamo.

Series elecequip

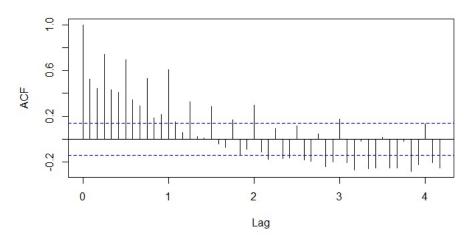


Figure 1.6. Grafico della Acf

Dal correlogramma sembra esserci una correlazione positiva durante lo stesso anno nelle vendite, inoltre osserviamo dei picchi positivi a distanza di due mensilità anche oltre l'anno, questo comportamento potrebbe essere dovuto al fattore stagionale. Invece riscontriamo correlazione negativa per alcuni mesi di anni successivi.

Dalla formula della stima della autocorrelazione sappiamo esserci al numeratore e al denominatore due sommatorie, per lag $\,l\,$ elevati lo stimatore peggiora nel valutare la stima perchè il numeratore ha una sommatoria con meno termini, infatti sappiamo essere consigliato per lag ridotti.

1.3.2 PACF

Questa quantitá misura un concetto chiave nelle serie temporali, che é anche molto intuitivo. In breve, indica quale correlazione sussiste tra due punti della serie temporale ad istanti diversi, avendo peró osservato gli eventi tra questi due istanti. Intuitivamente, ed in modo poco formale, possiamo prendere un esempio delle temperature, ovvero: se fossi interessato a conoscere la temperatura di domani, conoscendo solamente la temperatura rilevata due settimane fa, quest'ultima potrebbe darmi qualche informazione, potrebbe esserci qualche genere di correlazione tra queste due quantitá. Ma se io conosco la temperatura di ogni giorno delle ultime due settimane, allora il dato della temperatura rilevata quattordici giorni fa ha senso pensare che sia meno indicativo di quanto lo fosse prima. L'ultima considerazione importante prima di vedere la formula é anche questa, come l'ACF dipende solamente dal lag temporale quando si é davanti ad una serie stazionaria, cosí come fatto in precedenza mostriamo solamente la sua forma per serie temporali stazionarie.

Questo concetto intuitivo viene espresso dalla seguente formula:

$$\psi = \frac{Cov(X_{ti}, X_{ti+l} | X_{ti+1}, X_{ti+2}, ..., X_{ti+l-1})}{Var(X_{ti} | X_{ti+1}, X_{ti+2}, ..., X_{ti+l-1})}$$

Anche di questa quantitá esiste il corrispettivo campionario .

Series elecequip

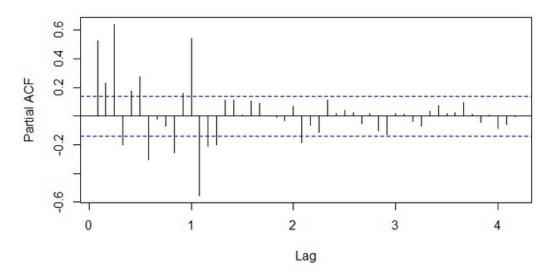


Figure 1.7. Grafico della Pacf

Analizzando il correlogramma parziale del nostro processo verrebbe naturale pensare che il processo non sia markoviano di ordine uno, viste le due correlazioni parziali molto elevate a lag maggiore di uno che non possono essere solamente errori di stima. Non escludiamo che possa godere della proprietà di Markov di ordine maggiore di uno. Visto il grafico qui sopra è ragionevole ipotizzare un comportamento stagionale della nostra serie. Sappiamo che si può introdurre la stagionalità tramite le espressioni $x_t = a_{tmods} + w_t$ e $x_t = x_{t-s} + w_t$. Come giá anticipato, il primo modello sembra più adeguato visto l'andamento della serie non essendoci chiari cambiamenti sulla varianza.

1.4 AIC

Prima di procedere nella vera e propria analisi della serie temporale in esame, é bene introdurre il parametro che valuterá la bontá dei metodi applicati. Questo parametro é l'AIC, acronimo di: Akaike information criterion.

L'AIC ha lo scopo di valutare quanto bene il metodo scelto riesca ad aderire (fit) alla serie temporale, aggiungendo talvolta una penalitá per i modelli con complessitá maggiore, vista la tendenza al fenomeno di overfitting.

L'obiettivo é di trovare il modello con il valore di AIC minore, indicante il miglior bilanciamento tra fit dei dati e fedeltá nel predire nuovi dati in input.

La formula é la seguente: $AIC = -2 \ln L + 2k$ dove, L indica una stima della verosimiglianza del modello e k indica il numero di parametri utilizzati dal modello.

1.5 ARIMA

$1.5.1 \quad AR(1)$

Come già accennato in precedenza, il processo non sembra essere markoviano di ordine uno e questo suggerisce che il modello auto regressivo di ordine uno non sia fedele al dataset.

Dopo aver applicato il modello AR(1) osserviamo quanto giá atteso, ovvero che la serie non sia in grado di aderire fedelmente alla serie temporale. Questo lo si puó osservare anche grazie al parametro AIC, che infatti assume un valore piuttosto elevato.

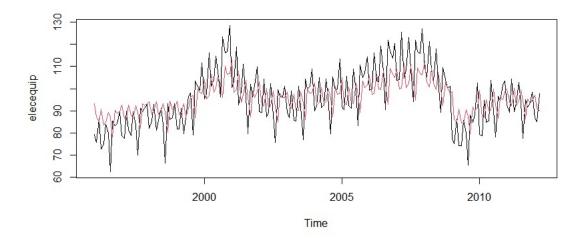


Figure 1.8. Grafico di un AR(1)

Di seguito i risultati di un AR(1), con il comando auto.arima():

$$\begin{array}{ccc} & ar_1 & intercept & AIC \\ AR(1) & 0.52 & 95.6 & 1501.7 \end{array}$$

1.5.2 AR(p)

Sappiamo dalla teoria che è difficile riconoscere un modello autoregressivo di ordine p solamente dall'Acf, per questo focalizzeremo le nostre osservazioni sulla Pacf infatti è nel correlogramma parziale che si evidenziano le intensità dei coefficienti autoregressivi ad esempio se notiamo un valore di Pacf prossimo a uno ad un lag l, siamo indotti ad assumere che il coefficiente α_l corrispondente sia elevato.

Vengono evidenziati chiari miglioramenti nel fit della serie temporale rispetto al modello AR(1), confermati anche dall'AIC. Notiamo comunque che l'AIC non migliora drasticamente e una giustificazione possibile potrebbe essere dovuta alla complessitá del modello. Tutto questo non ci stupisce, siccome il modello avendo a disposizione piú parametri riesce ad aderire meglio alla serie, ma allo stesso tempo viene anche penalizzato per il loro utilizzo.

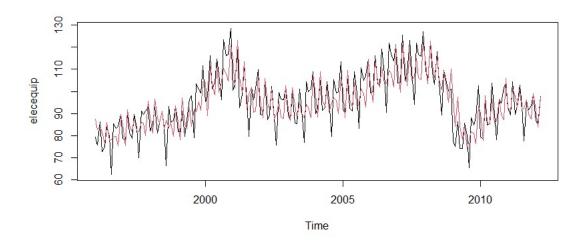


Figure 1.9. Grafico di un AR(8)

Di seguito i parametri di un AR(8) trovati con il comando auto.arima() di R:

$$ar_1$$
 ar_2 ar_3 ar_4 ar_5 ar_6 ar_7 ar_8 $intercept$ AIC $AR(8)$ 0.43 -0.054 0.42 -0.058 0.044 0.39 -0.29 -0.008 94.19 1348.5

$1.5.3 \quad MA(1)$

Valutiamo che il processo non venga descritto propriamente dal modello MA(1): questo perchè per valori alti di beta (circa 0.9) è presente una sorta di oscillazione nelle Pacf mentre per valori bassi beta (circa 0) il processo si comporta come un rumore bianco. Entrambi questi scenari sono molto distanti dal nostro processo così come per valori di beta <0 (molto inferiori a zero) siccome infatti la Pacf assume valori minori di zero.

Quanto detto trova supporto con i risultati ottenuti dal metodo applicato, ottenendo valori alti di AIC e scarsa abilita di aderenza alla seria.

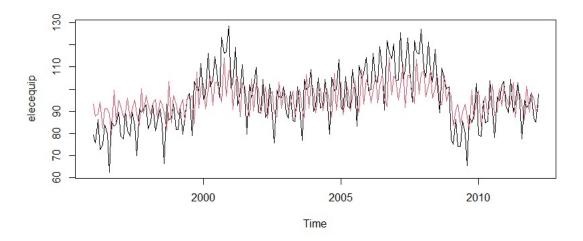


Figure 1.10. Grafico di un MA(1)

Di seguito i risultati di un MA(1) trovati con il comando auto.arima() di R:

$$\begin{array}{ccc} & ma_1 & intercept & AIC \\ \mathrm{MA}(1) & 0.63 & 95.7 & 1502.7 \end{array}$$

1.5.4 MA(q)

Così come accadeva con il modello AR(p) è difficile dedurre se il modello MA(q) sia appropriato o meno alla descrizione dei dati siccome vi sono molti parametri in gioco.

Si puó notare dalla figura sottostante che, malgrado il valore della AIC sia addirittura minore rispetto a quello ottenuto con AR(8), il modello MA(8) non sembra aderire fedelmente alla serie, questo é un indicatore che difficilmente riusciremo ad avere risulti ottimali solo con l'applicazione del "puro" MA(q).

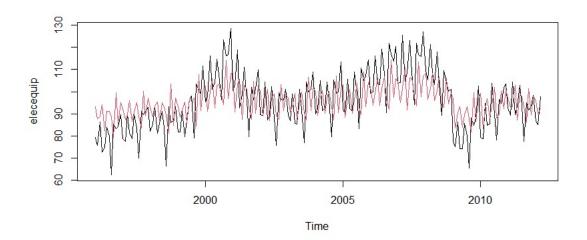


Figure 1.11. MA(8)

Di seguito i parametri di un MA(8) trovati con il comando auto.arima() di R:										
	ma_1	ma_2	ma_3	ma_4	ma_5	ma_6	ma_7	ma_8	intercept	AIC
MA(8)	0.38	0.37	0.82	0.15	0.53	0.69	-0.03	0.58	95.4	1338.9

1.5.5 ARMA(p,q)

I modelli AR(p) e MA(q) abbiamo visto essere in qualche modo collegati, infatti il modello MA(1) è un particolare modello $AR(\infty)$ cosí come il modello AR(1) è un particolare modello $MA(\infty)$; inoltre questi due metodi non vanno in conflitto e per queste ragioni possiamo unirli, semplicemente sommando il polinomio AR(p) e il polinomio MA(q), ricordando di contare una volta sola il rumore bianco ω_t .

Il modello appena definito viene chiamato ARMA(p,q) e sappiamo godere di maggiori vantaggi rispetto ai modelli presi singolarmente.

Questo viene rispecchiato totalmente sia dal fit della serie temporale che dal valore della AIC. Abbiamo forzato R ad utilizzare tutti e 14 i parametri, sapendo fosse non necessario, per mostrare che nonostante la penalizzazione da parte dell'AIC il valore di questo parametro fosse comunque inferiore agli altri due presi singolarmente.

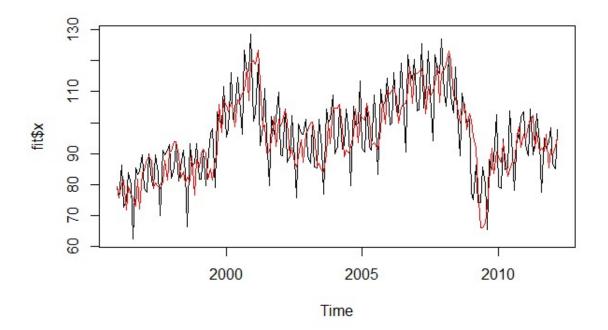


Figure 1.12. ARMA(7,7)

Di seguito i parametri di un ARMA(7,7) trovati con il comando Arima() di R:

ARMA(7,7)	$ar_1 \\ 0.91$	$ar_2 \\ 0.001$	ar_3 -0.012	ar_4 -0.014	ar_5 -0.001	$ar_6 \\ 0.99$	ar_7 -0.92	ma_1 -0.40	ma_2 -0.02	$ma_3 \\ 0.07$	ma_4 0.13	ma_5 -0.001	ma_6 -0.93
ma_7	intercept	AIC											

1.5.6 Integrated Model

Può capitare che si sia interessati a studiare dei processi che non sono stazionari come ad esempio il random walk, talvolta può essere utile per ovviare il problema, effettuare la differenza tra alcuni step del processo questo metodo è chiamato integrazione. Entrando nel dettaglio noi creeremo dei nuovi processi dati dalla integrazione del processo originale, il numero di volte sarà detto numero di integrazione e indicheremo questo modello come I(d) questo procedimento ci condurrà ad ottenere un processo stazionario su cui è possibile applicare i metodi AR(p) ed MA(q).

1.5.7 ARI

Cerchiamo anche un modello ARI(7,1) per vedere come approssima la nostra serie

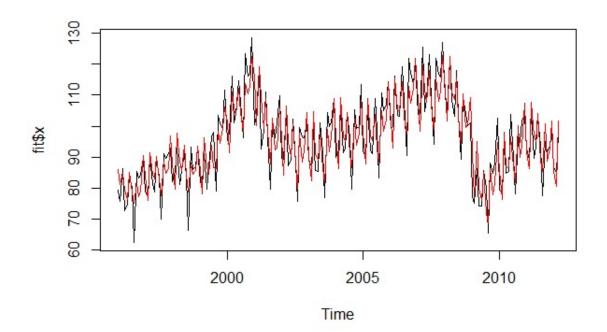


Figure 1.13. ARI(7,1)

Di seguito i parametri di un ARI(7,1) trovati con il comando auto.arima() di R:

1.5.8 IMA(1,8)

Cerchiamo anche un modello IMA(1,8) per vedere come approssima la nostra serie:

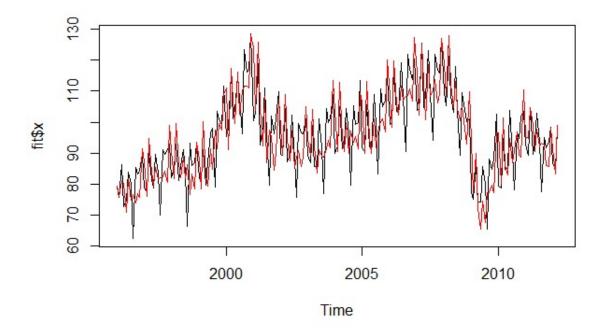


Figure 1.14. IMA(1,8)

Di seguito i parametri di un IMA(1,8) trovati con il comando auto.arima() di R:

1.5.9 ARIMA(p,d,q)

In linea con la trattazione seguita fino ad ora, introduciamo il penultimo modello che verrá analizzato, ovvero il modello: ARIMA (p,d,q). Esso consiste in un modello ARMA(p,q) preceduto dal metodo di integrazione. Come giá visto nel caso del modello ARMA(p,q), l'unione di piú metodi dovrebbe garantire un miglioramento delle prestazione del modello. A nostra sorpresa il comando auto.arima(), riconosce come migliore in assuluto il modello ARIMA(0,1,8), quindi senza l'utilizzo della componente autoregressiva, ricadendo difatto in un modello IMA(1,8).

Questo risultato stupisce, ma é giustificabile dal fatto che, innanzitutto il comando tiene in considerazione anche altri parametri oltre che all'AIC, e sopratutto non stiamo considerando la componente di stagionalitá che invece copriremo nella prossima sezione.

Di seguito i parametri di un ARIMA(p,d,q) trovati con il comando auto.arima() di R:

1.5.10 **SARIMA**

Ora possiamo finalmente giungere al modello che ci aspettiamo dia i risultati migliori, ovvero il modello SARIMA, ovvero con le componenti stagionali, denotato anche: ARIMA(p,d,q)(P,D,Q).

Al di lá del fatto che avendo a disposizione più strumenti diventa più facile creare un modello adeguato per la funzione auto.arima(), noi ci aspettiamo di ottenere risultati sensibilmente migliori di quelli ricevuti fin'ora, questo perché, come giá sottolineato nella sezione STL, la componente stagionale sembra giocare un ruolo importante nella descrizione della serie temporale in esame, e quindi é ragionevole pensare che quanto osservato venga rispecchiato anche a livello di risultati.

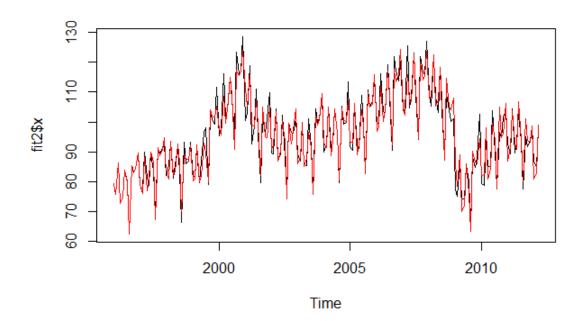


Figure 1.15. Grafico modello ARIMA(4,0,1)(0,1,1)

1.6 Forecasting

In questa sezione ci poniamo l'obiettivo di confrontare le 12 osservazioni future predette da tre dei metodi analizzati fino ad ora: ARMA, ARIMA, SARIMA.

Forecast con ARMA(1,5) delle prossime 12 osservazioni future:

Forecasts from ARIMA(1,0,5) with non-zero mean 120 - 2000 2005 Time

Figure 1.16. Forecasting del modello ARMA(1,5)

Forecast con $\operatorname{ARIMA}(0,1,8)$ delle prossime 12 osservazioni future:

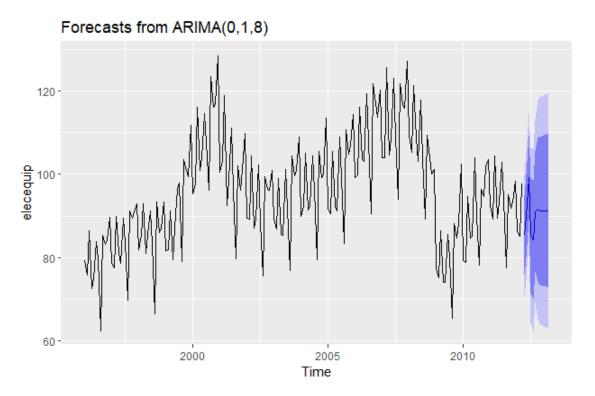


Figure 1.17. Forecasting del modello ARIMA(0,1,8)

Forecast con (SARIMA) ARIMA(4,0,1)(0,1,1) delle prossime 12 osservazioni future:

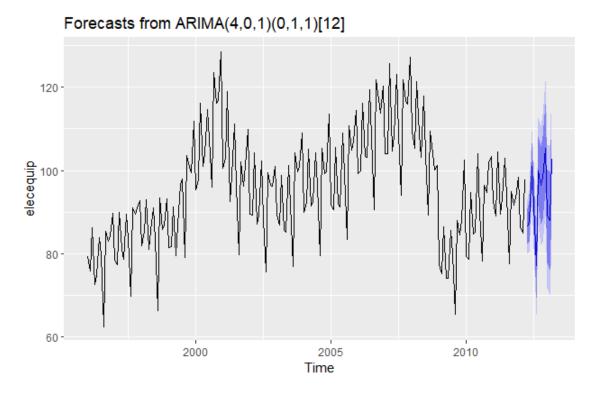


Figure 1.18. Forecasting del modello ARIMA(4,0,1)(0,1,1)

Commentiamo preliminarmente il risultato del modello ARMA(1,5) per poi man mano confrontare gli altri. Notiamo subito una varianza molto elevata nella previsione, indicata dalla regione blu, la quale si espande immediatamente dopo l'ultima osservazione, indice che forse questo modello non é particolarmente adatto per la previsione. Un altro indizio che qualcosa non vada viene dato dall'andamento della serie temporale fino all'istante noto e all'andamento previsto dal modello, sembra difficile immaginare che una serie che fino a quel punto sia stata caratterizzata da cambi repentini diventi improvvisamente quasi costante.

Per quanto riguarda il modello ARIMA(0,1,8) osserviamo un regione blu leggermente più piccola rispetto alla precedente con un andamento previsto meno "piatto" rispetto al modello ARMA(1,5), ma continuano a presentarsi alcune criticità riscontrate precedentemente, specialmente sulla poca certezza della previsione e sulla poca credibilità in merito all'andamento futuro, più credibile ma non a sufficienza.

Concludiamo invece con quella che sembrerebbe la miglior analisi predittiva tra i metodi visti, il forecasting effettuato dal modello SARIMA. E' possibile osservare come i problemi elencati di varianza e "credibilitá" siano stati risolti, risulta piú semplice credere a questa previsione, che semplicemente propone un andamento simile a quello visito fino all'ultima osservazione.

1.6.1 Analisi dei residui

I residui di una serie temporale sono quello che resta dopo aver fatto aderire il modello ai dati. Per molti modelli, ma non tutti, i residui sono la differenza tra le osservazioni e i corrispondenti valori stimati. Essi sono utili per controllare se un modello é in grado o meno di catturare le informazioni fornite dai dati.

Le caratteristiche fondamentali che devono soddisfare i residui di un modello per una serie temporale sono:

- I residui devono essere scorrelati, se vi é correlazione tra residui allora é rimasta dell'informazione non compresa dal modello, che bisognerebbe carpire.
- I residui devono avere media zero, se non accade allora il forecasting sará distorto.

Altre proprietá che il modello dovrebbe rispettare, ma non necessarie come le altre, sono:

- La varianza costante dei residui.
- I residui devono essere distribuiti secondo una normale.

Se un modello fornisce dei residui che rispettano queste proprieta, allora potrebbe essere un buon modello per fare previsione.

Di seguito vediamo due esempi di analisi dei residui, rispettivamente per il modello ARIMA senza componenti stagionali, e per il modello (S)ARIMA che comprende anche le componenti stagionali.

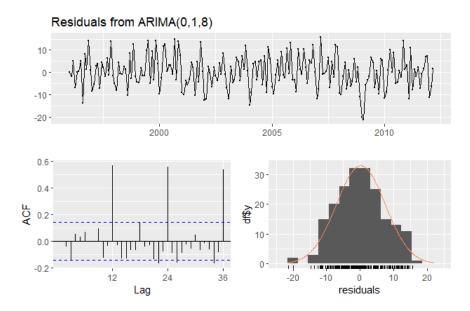


Figure 1.19. Residui dati dall'approssimazione con il modello ARIMA(0,1,8)

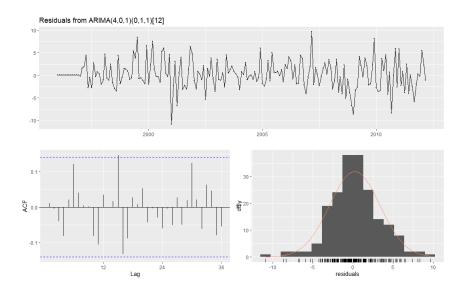


Figure 1.20. Residui dati dall'approssimazione con il modello ARIMA(4,0,1)(0,1,1)

La grande differenza tra i due residui ottenuti risiede nel fatto che nel caso dell'ARIMA senza stagionalitá siaa evidenziata una correlazione tra i residui, e questa rappresenta una violazione delle proprietá richieste in precedenza,

Quanto ottenuto sembra confermare le nostre osservazioni, mostrando una volta ancora che il modello più calzante per il forecasting sembri proprio essere dato dal ARIMA(4,0,1)(0,1,1).

1.7 TEST

Concludiamo il nostro elaborato valutando effettivamente il potere predittivo dei metodi visti fin'ora, dividendo la serie temporale in due porzioni: training set e test set. Il training set é stato selezionato rimuovendo le osservazioni successive al dicembre 2010, e chiaramente il test set sará costituito dalle osservazioni successive, ovvero dal gennaio 2011 fino al marzo 2012. L'idea sará quella di servirci del training set per valutare i modelli ed effettuare nuovamente il forecasting, questa volta effettuato conoscendo l'oggetto della previsione. E' giusto sottolineare che la serie temporale su cui si allena il nostro modello é diversa da quella precedente e quindi é ragionevole pensare che non necessariamente avremo gli stessi modelli di prima. Mostreremo una figura con i due grafici, predetto e reale, per vedere come i modelli si comportano in maniera facile da interpretare:

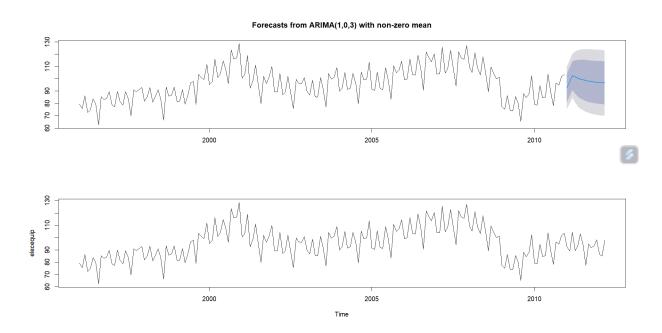


Figure 1.21. Forecasting comparison ARMA(1,3)

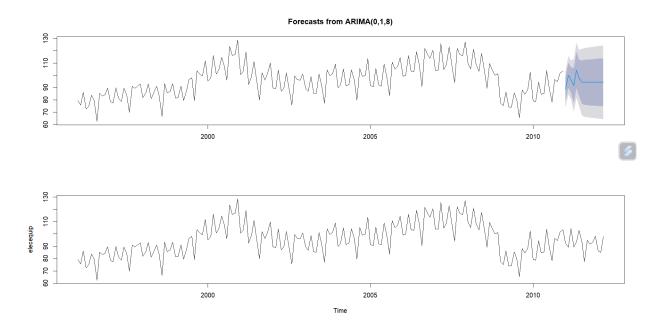


Figure 1.22. Forecasting comparison ARIMA(0,1,8)

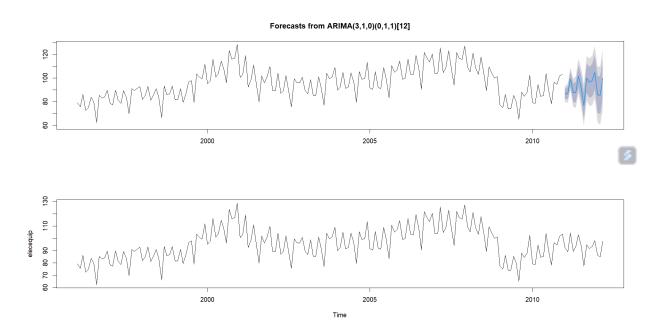


Figure 1.23. Forecasting comparison (S)ARIMA(1,0,3)(0,1,1)

Al termine di queste comparazioni tra effettivo andamento della serie temporale e previsione dei metodi, possiamo affermare che il modello ARIMA completo delle componenti stagionali é fedele alla serie temporale in esame e questo conclude il nostro elaborato.

Bibliography

Forecasting: Principles and Practice (2nd ed), Rob J Hyndman and George Athanasopoulos, Monash University, Australia, 2018