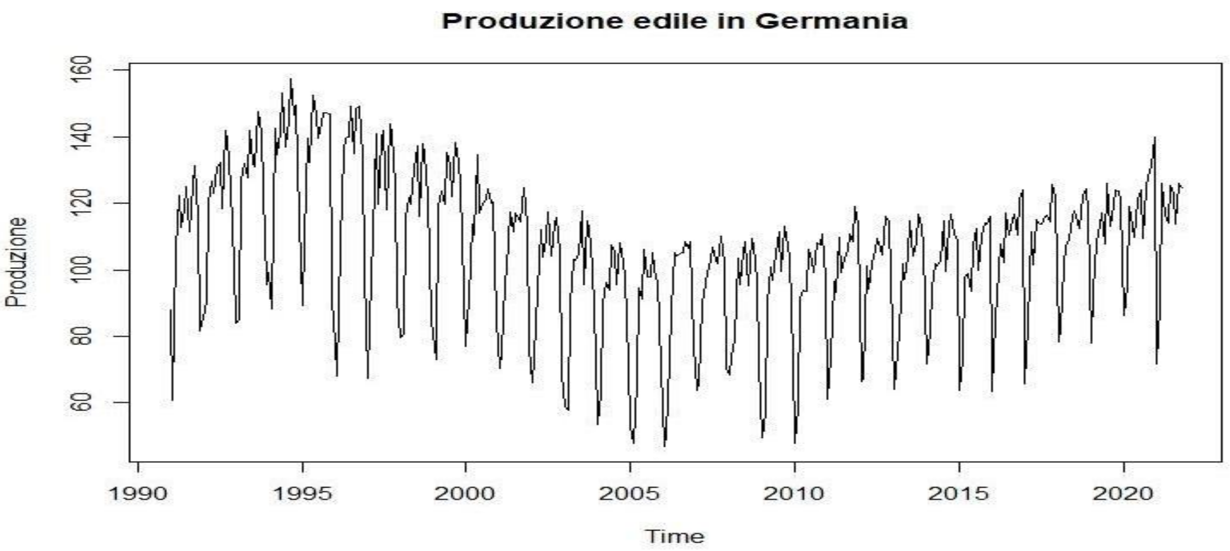
Analisi time series economica: “Produzione totale mensile del settore edile in Germania”

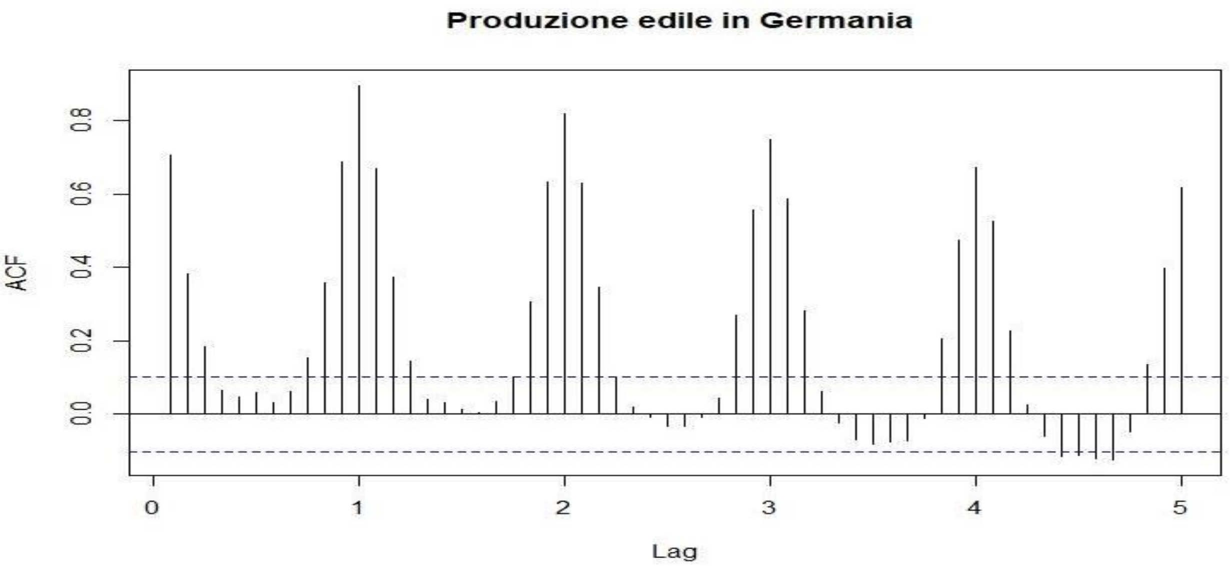
L’analisi della time series riguarda l’indice della produzione edile che misura le variazioni del volume della produzione mensile in Germania dal Gennaio 1991 fino a Ottobre 2021 (370 osservazioni). La fonte dei dati è il sito “Eurostat” (https://ec.europa.eu/eurostat/databrowser/view/STS_COPR_M__custom_1842339/default/table?lang=en). Fornisce una misura dell'andamento del volume del valore aggiunto in un determinato periodo di riferimento. L’obbiettivo principale dell’analisi è quello di prevedere l’indice di produzione edile nei successivi 12 mesi.

Analisi preliminari

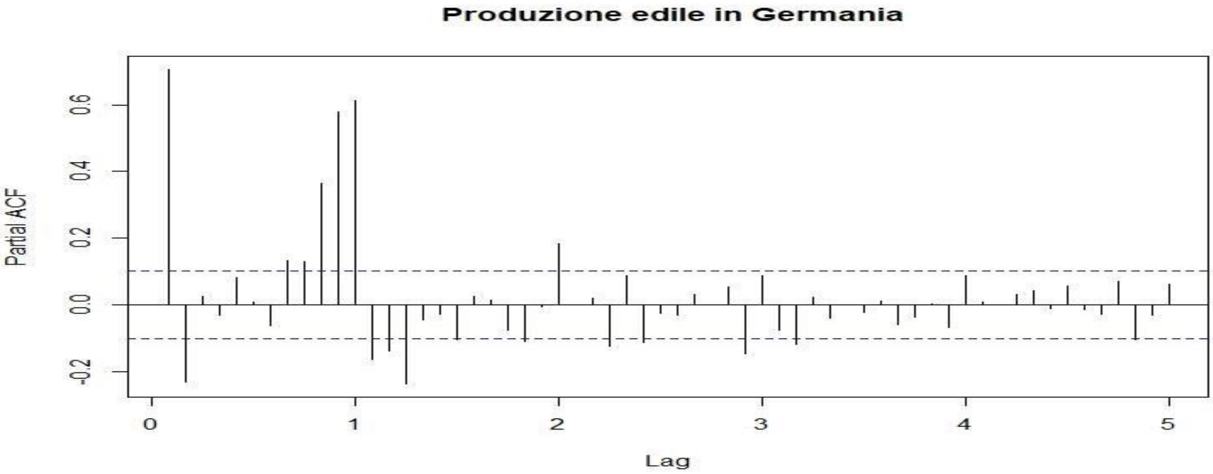


(1) Time Series Plot

L’immagine (figura 1) mostra il grafico della time series. Possiamo notare un trend di produzione costante (oscillante) e una stagionalità annuale. Si può notare che nel periodo 1996-2005 c’è stato un calo della produzione che dal 2010 in poi ha di nuovo iniziato a crescere linearmente.



(2) ACF



(3) PACF

Per l'analisi della time series utilizziamo il modello ARIMA stagionale (ARIMA (p,d,q)x(P,D,Q)S) con S=12 perché i dati sono a cadenza mensile. Il modello è composto da due parti auto-regressive (stagionale e non stagionale) e due parti Moving Average (stagionale e non stagionale).
I due grafici (figura 2 e 3) rappresentano l'autocorrelazione (Acf) e l'autocorrelazione parziale (Pacf) ai diversi lag.
Dal grafico (figura 2) si nota un decadimento lineare dell'Acf dovuto a lag multipli di S, mentre dal grafico (figura 3) che $\phi(1)$ è prossimo a 1: la stazionarietà sembra non esserci e ciò è dovuto dalla parte stagionale (D > 0).
Si procede con l'analisi tramite il test ADF (Augmented Dickey-Fuller) che testa la presenza o l'assenza di radici unitarie nel modello $\Delta Y_t=\omega+\xi t+\gamma Y_{t-1}+\delta_1 \Delta Y_{t-1}+\ldots+\delta_p \Delta Y_{t-p}+u_t$.

	tau3	phi2	phi3
statistic	-0.5198759	0.8428123	1.260848

	1pct	5pct	10pct
tau3	-3.98	-3.42	-3.13
phi2	6.15	4.71	4.05
phi3	8.34	6.30	5.36

H0(1): $\gamma=0$ Stat.test - > tau3
H0(2): $\gamma=\xi=0$ Stat.test - > phi3
H0(3): $\gamma=\xi=\omega=0$ Stat.test- > phi2

Le statistiche test tau3,phi3,phi2 non sono significative e portano ad accettare H0(1),H0(2) e H0(3) superando i primi due step della procedura. Si prosegue con lo step tre, utilizzando come DGP un RW senza drift e un modello senza trend lineare:

	tau2	phi1
statistic	-1.17637	0.695292

	1pct	5pct	10pct
tau2	-3.44	-2.87	-2.57
phi1	6.47	4.61	3.79

H0(5): $\gamma=0$ Stat.test - > tau2
H0(6): : $\gamma=\omega=0$ Stat.test - > phi1

Allo step tre accettiamo le due ipotesi H0(5) e H0(6) , che includono la presenza di unit roots. Le domande che sorgono sono: "Ma quante radici sono?"; "Dipendono dalla parte stagionale o non stagionale?". L'unica unit root trovata, guardando il time series plot e il grafico Acf e Pacf, pare sia dovuta dalla parte stagionale quindi si procede testando le differenze dodicesime. Ricalcolando il test ADF i risultati sono i seguenti:

	tau3	phi2	phi3
statistic	-3.527882	4.316884	6.449797

	1pct	5pct	10pct
tau3	-3.98	-3.42	-3.13
phi2	6.15	4.71	4.05
phi3	8.34	6.30	5.36

Questo è un caso borderline, ed essendo questi test poco potenti , i quali tendono ad accettare H0, si rifiuta anche se in dubbio e non abbiamo la presenza di altre unit root dalla parte non stagionale.

<u>Stima modello</u>

In base alle osservazioni e alle analisi precedentemente fatte passiamo alla stima dei parametri del modello ARIMA. Partiamo con d=0,D=1 ma notiamo subito che i risultati in termini di Information Criteria e in termini di significatività dei parametri sono miglliori settando d=D=1. Dopo varie combinazioni dei parametri p,q,P,Q il modello migliore risulta:

ARIMA(0,1,1)(1,1,2)[12]
Coefficents:

	ma1	sar1	sma1	sma2
	-0.7996	-0.6139	0.1396	-0.5820
s.e.	0.0336	0.1084	0.0984	0.0597

sigma^2 estimated as 51.45: log likelihood=-1212.63
AIC=2435.27 AICc=2435.44 BIC=2454.66

Inseriamo nel modello anche i regressori esterni: variabili di calendario e anomalie. Eliminando le varie variabili di calendario non significative e rimodellando l'ARIMA cercando di abbassare i valori dell'Information Criteria. Il modello stimato risulta:

ARIMA(0,1,1)(1,1,1)[12] errors
Coefficients:

	ma1	sar1	sma1	Tue	Wed	Thu	Fri
	-0.7148	0.1414	-0.6654	3.7644	3.8843	1.1694	3.1106
s.e.	0.0458	0.0784	0.0548	0.5632	0.5769	0.5664	0.5568

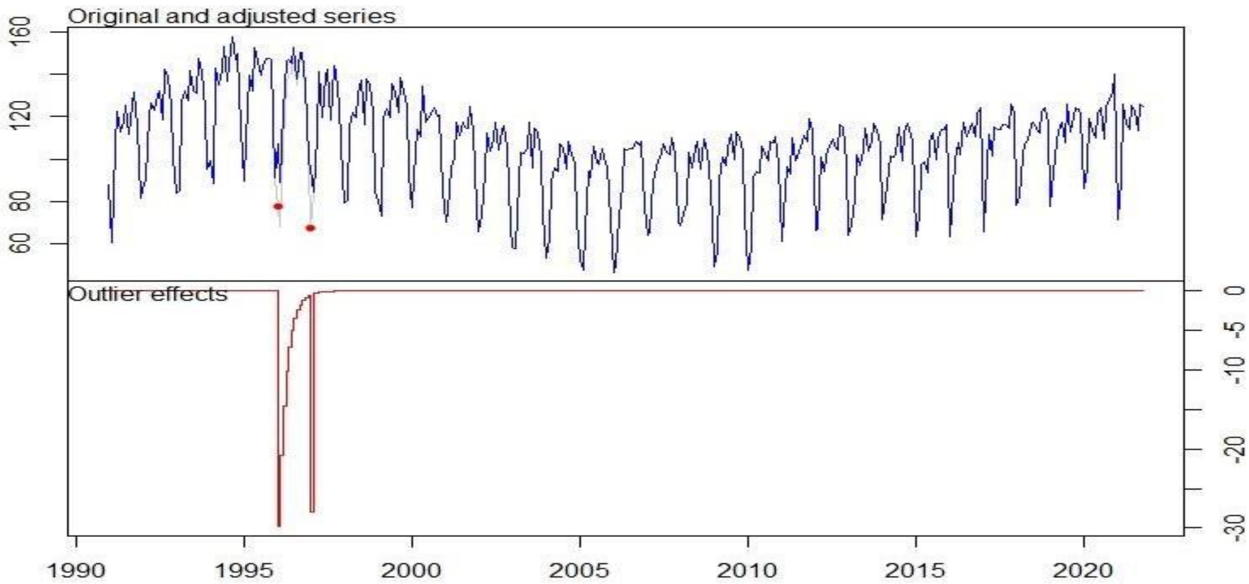
sigma^2 estimated as 40.05: log likelihood=-1164.53
AIC=2345.07 AICc=2345.48 p BIC=2376.09

A questo punto ricerchiamo la presenza di outliers tramite la funzione tso(), che cerca outliers di tipo AO (Additive Outlier), LS (Level Shift),TC(Transient Change). Con il modelo ARIMA(0,1,1)(1,1,2)[12] la funzione non trova outliers, se invece usiamo il modello con le variabili di calendario ARIMA(0,1,1)(1,1,1)[12] i risultati che otteniamo sono i seguenti:

Coefficients:

	ma1	sar1	sma1	Tue	Wed	Thu	Fri	TC61	AO73
	-0.7381	0.0508	-0.6098	4.1312	4.0388	1.4248	3.4398	-29.7985	-27.5374
s.e.	0.0373	0.0797	0.0574	0.5265	0.5410	0.5350	0.5238	4.4348	4.9260

sigma^2 estimated as 34.27: log likelihood=-1135.72
AIC=2291.44 AICc=2292.08 BIC=2330.22

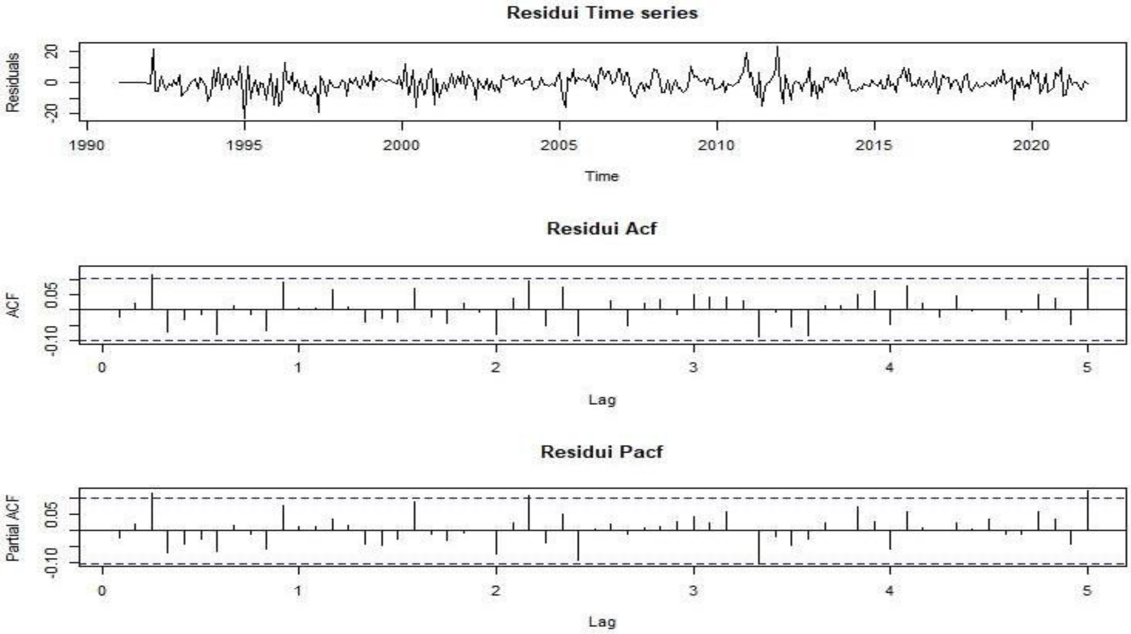


(4) Outliers

Il grafico (figura 4) mostra la presenza di due outliers. Il primo è un TC nel 1996, il secondo un AO nel 1997. Il modello ARIMA con variabili di calendario e anomalie sembra essere il migliore dal punto di vista di Information Criteria e dal punto di vista della significatività dei parametri.

Diagnostiche

Per le diagnostiche si considera il modello ARIMA con variabili di calendario e outliers, che abbiamo valutato essere il migliore.



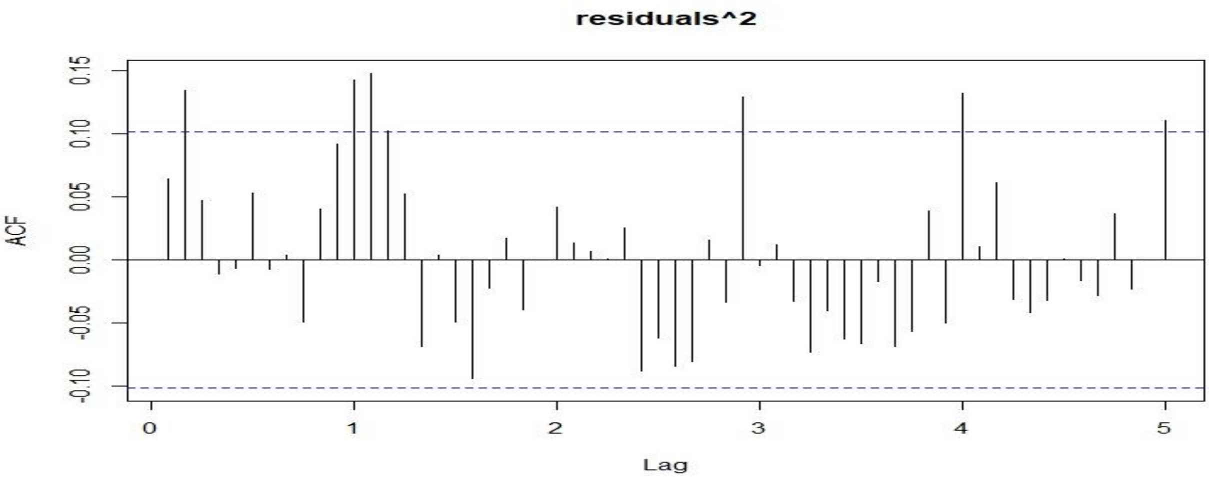
(5) Plot, Acf, Pacf sui residui

Dal grafico (figura 5) si nota che i residui del modello sembrano ben specificati.

La tabella riporta il test di Ljung-Box delle prime H autocorrelazioni:

lag	10	11	14	19	24	29
statistic	12.22848	15.37062	17.0528	20.68859	24.31073	34.73214
parameter	1	2	5	10	15	20
p.value	0.000470654	0.0004595285	0.004400733	0.02337281	0.06000111	0.02157815

Il test è altamente significativo per le prime tre H autocorrelazioni mentre significativo per le rimanenti.

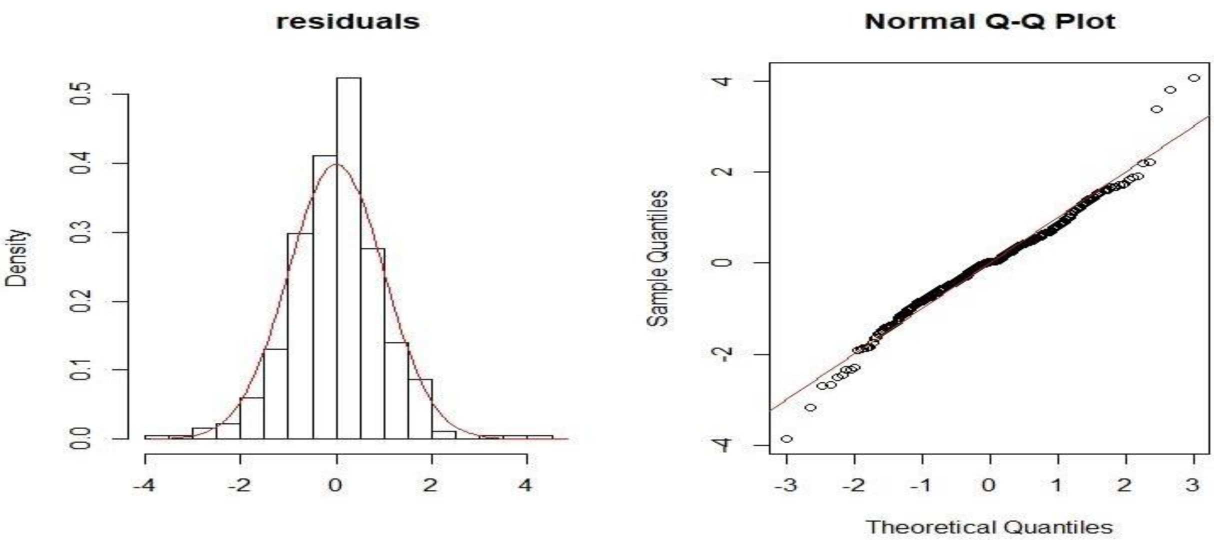


(6) Plot residui al quadrato

Facendo il test Arch, con i residui al quadrato (figura 6) degli M valori passati, viene testata H0 con l'uguaglianza a 0 dei coefficenti del modello.

statistic	0.0457943	0.2672806	0.3941379	1.111993	9.623928	14.9498
parameter	1	2	3	6	12	24
p.value	0.8305501	0.8749047	0.9414508	0.9810085	0.6489152	0.9222188

In questo caso accettiamo H0 , quindi non abbiamo bisogno di trasformare i dati.



(7) Controllo normalità

A questo punto procediamo con il test di Shapiro-Wilk per la normalità dei residui, i quali si dovrebbero distribuire come un NWN. Purtroppo però rifiutiamo l'ipotesi H0 di normalità :

W = 0.97512, p-value = 5.499e-06

Come vediamo dal grafico sulla distribuzione dei residui (figura7), la distribuzione gaussiana non si adatta bene ai dati, lascia qualcosina sulle code e soprattutto sulla punta. Questo confermato anche dal Q-Q plot (figura 7).

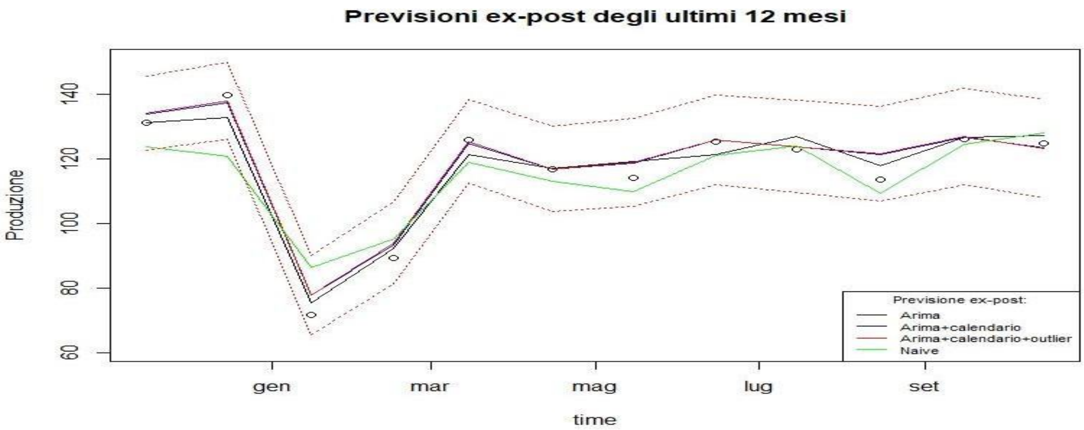
<u>Previsioni ex-post</u>

Nella previsione ex-post si mettono a confronto errori di misura in scala assoluta(ME,MAE,RMSE), in scala percentuale(MPE,MAPE,RMSPE), ed errori di misura scalati con l'errore del modello naive(ScMAE,ScRMSE).

	model	h	ME	MAE	RMSE
1	Arima	1	-0.4913138	5.484034	7.933890
2	Arima + Calendar	1	-0.3357008	3.980318	5.302499
3	Arima + Calendar + Outliers	1	-0.3254476	3.807281	5.006363
4	Naive	1	2.2333333	6.333333	8.144732

	MPE	MAPE	RMSPE	ScMAE	ScRMSE
1	-0.015998348	0.05361509	0.08746547	0.8659001	0.9741131
2	-0.010665569	0.03798225	0.05441793	0.6284713	0.6510341
3	-0.009916306	0.03606391	0.05034042	0.6011497	0.6146749
4	0.008183381	0.05851271	0.07971908	1.0000000	1.0000000

Concludiamo notando che i valori migliori appartengono al modello Arima + Calendar + Outliers e quindi lo utilizziamo per fare previsioni in modo più accurato, fatto confermato dal grafico (figura8) dove si nota che le osservazioni cadono tutte nelle bande di previsione. I 4 modelli vengono confrontati con un quinto modello chiamato Naive che nel nostro caso si comporta come un RW-stagionale.

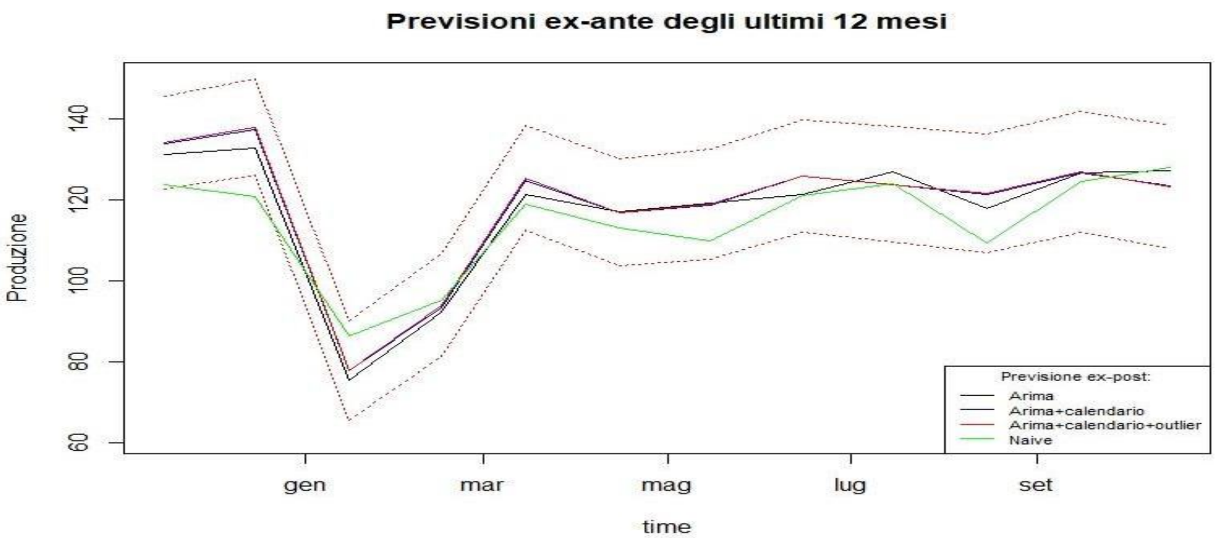


(8) previsioni ex-post

Previsioni ex-ante

Nelle previsioni ex-ante andiamo a prevedere l'Indice di produzione edile tra Novembre 2021 e Dicembre 2022. Utilizziamo i 4 modelli visti finora con l'aggiunta del modello naive per prevedere i dati.
Otteniamo i seguenti risultati:

t	pred1	pred2	pred4	naive
371	131.24700	133.97178	134.25536	123.8
372	132.95304	137.46165	138.08797	120.8
373	75.55263	77.79748	77.87553	86.3
374	92.21310	93.33513	93.99517	95.1
375	121.35826	124.89811	125.53238	118.9
376	117.06704	116.90521	116.94647	113.2
377	119.15909	118.72192	119.08422	109.9
378	121.34398	125.84031	125.88511	121.2
379	127.04136	123.91202	123.83693	124.0
380	117.86217	121.27849	121.65104	109.3
381	126.72111	126.63971	126.95434	124.6
382	127.19386	123.53514	123.32913	128.0



(9) Previsioni ex-ante

Le previsioni del grafico (figura 9) sembrano seguire quelle del grafico delle previsioni ex-post (figura 8), a parte per il modello naive un po' discostato in qualche punto.