Elaborato per Time Series, MD2SL 23/24
Serie aciclica

Dario Comanducci, 19 marzo 2025

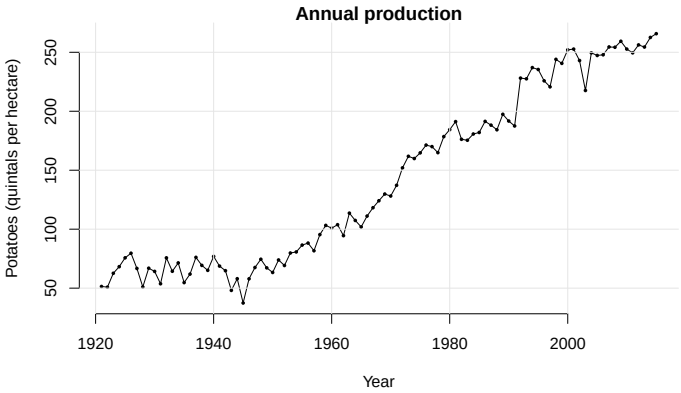
1

Describe the data, explaining which phenomenon they represent and providing details on the measurements (time span, frequency, scale). Display the time series and discuss its appearance in reference to stationarity and to the presence of outliers.

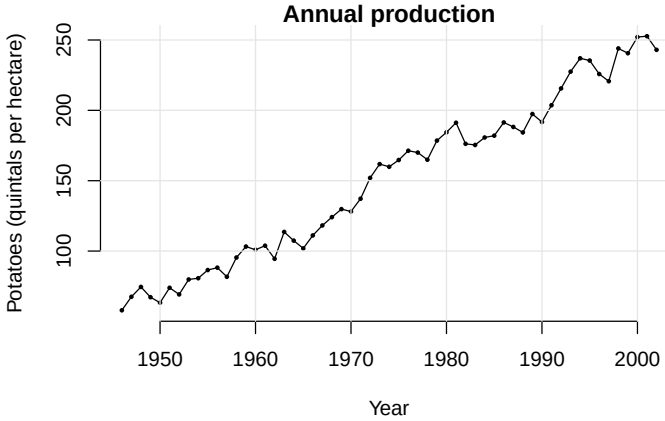
La serie in questione riguarda la produzione media (espressa in quintali per ettaro) su scala annuale delle patate in Italia, isolata all'interno del foglio excel scaricabile alla voce "Produzione media delle principali coltivazioni agricole - Anni 1921-2015"[1] all'interno della pagina web dell'Istat sulle serie storiche riguardanti agricoltura, zootecnia e pesca.[2]

Osservando l'andamento della serie nella sua interezza (Fig. 1a), in vista della successiva elaborazione notiamo che:

- prima della 2ª guerra mondiale la serie sembra piuttosto stazionaria, come se fosse soggetta ad un altro processo, per cui conviene rimuovere i dati prima del 1946;
- poiché nel foglio excel è indicato che dal 2006 la produzione media è "calcolata sulla base della produzione raccolta sulla superficie totale", merita scartare pure i dati dopo il 2005 nel dubbio che sia cambiato il loro calcolo a partire dal 2006;
- inoltre la presenza di due outlier (nel 1992 e nel 2003) induce a tagliare ulteriormente la serie, limitandosi agli anni 1946-2002;[3]
- per concludere, merita imputare l'outlier del 1992 interpolando i valori della serie per gli anni 1991 e 1992 secondo la retta che collega la produzione di patate nel 1990 con quella nel 1993.[4]



(a) serie completa



(b) serie tagliata ed imputata (1991-1992)

Fig. 1. Produzione media annuale di patate in Italia.

In definitiva la serie che verrà analizzata è quella in Fig. 1b, esibendo un chiaro trend che rende il processo non stazionario; applicando

[1] https://seriestoriche.istat.it/fileadmin/documenti/Tavola_13.16.xls

[2] https://seriestoriche.istat.it/index.php?id=1&no_cache=1&tx_usercento_centofe%5Bcategoria%5D=13&tx_usercento_centofe%5Baction%5D=show&tx_usercento_centofe%5Bcontroller%5D=Categoria&cHash=e3503d8195dd4231ff53ba078ad5c124

[3] L'outlier del 1992 non può essere tagliato, altrimenti non sarebbe rispettato il vincolo di almeno 50 osservazioni nella serie.

[4] Per una giustificazione della necessità di tale imputazione, si veda l'appendice A.

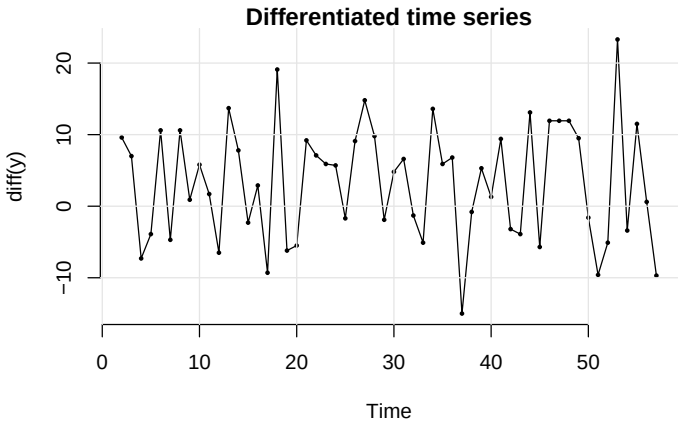


Fig. 2. Serie di Fig. 1b differenziata 1 volta ($\Delta y_t = y_t - y_{t-1}$).

i test ADF e KPSS per valutare presenza di radici unitarie, abbiamo che:

- il p-value di 0.217 per ADF è troppo elevato per rifiutare l'ipotesi nulla di non stazionarietà;
- il p-value < 0.01 per KPSS è abbastanza basso da rifiutare l'ipotesi nulla di stazionarietà.

Entrambi i test suggeriscono quindi la presenza di radici unitarie: differenziando solo una volta la serie, i dati sembrano poi in effetti assumere un comportamento stazionario (Fig. 2), confermato dai test ADF (p-value < 0.01) e PKSS (p-value > 0.1) su $\Delta y_t = y_t - y_{t-1}$.

2

Find an ARIMA model and an ARMA model with linear deterministic trend that fit adequately to the time series. To do this, make use of a combination of graphical checks, significance tests and automated search.

2.1 Modello ARIMA

Dai precedenti risultati dei test ADF e KPSS, lavoriamo direttamente sulla serie $\Delta y_t = y_t - y_{t-1}$ di Fig. 2 considerando in Fig. 3 sia ACF che PACF per Δy_t: in questo caso tutti i coefficienti di ACF e PACF sono statisticamente non diversi da 0 (eccetto la PACF a lag 11, ma considerabile come valore spurio), per cui il

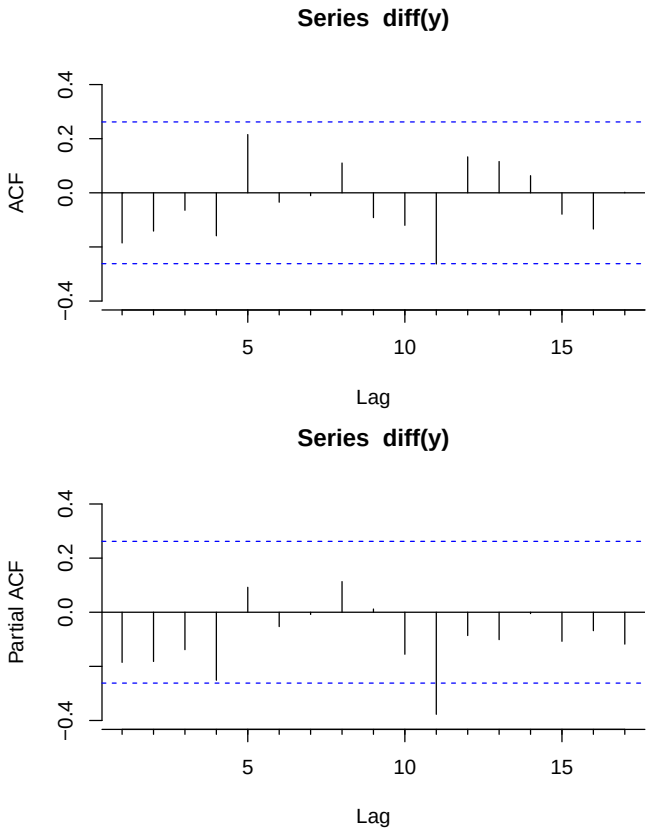


Fig. 3. ACF e PACF per Δy_t in Fig. 2.

modello candidato a descrivere y_t potrebbe essere un random walk (ossia ARIMA(0,1,0)), da valutare se con drift o senza.

Dalla PACF di Fig. 3 si nota anche che al lag 4 siamo al limite della banda di non significatività, per cui merita valutare anche un modello AR(4), con o senza drift.

Volendo considerare anche modelli alternativi, possiamo ricorrere alla funzione `auto.arima` (anche in questo caso con o senza drift):

- nel caso senza drift, il modello suggerito da `auto.arima` è il random walk;
- nel caso con drift, il modello trovato è un ARIMA(0,1,2);

Dal criterio informativo AICc in Tab. 1 per i tre modelli in esame vediamo che il modello migliore è l'ARIMA(0,1,2) con drift, sebbene possa essere preso in considerazione come alternativa anche il random walk con drift (applicando il rasoio di Occam, in quanto è un modello più semplice a fronte di un AICc simi-

Tab. 1. Modelli ARIMA a confronto

Modello	drift	AICc
ARIMA(4,1,0)	no	412.43
ARIMA(0,1,0)	no	403.81
ARIMA(4,1,0)	sì	397.61
ARIMA(0,1,0)	sì	397.31
ARIMA(0,1,2)	sì	395.33

le).[5] A tal proposito si vedano anche le analisi sui residui riportate in § 4.1.

Il test di significatività per ARIMA(0,1,2) con drift è superato in virtù dei valori ottenuti sulla statistica z in Tab. 2 (deve valere $|z| > 1.96$); inoltre per scrupolo determiniamo le radici per polinomio MA

$$1 + \psi_1 L + \psi_2 L^2$$

che valgono $L_1 \approx 1.48$, $L_2 \approx -2.69$, a garanzia dell'invertibilità del modello, in virtù dei coefficienti $\psi_1 \approx -0.304$ e $\psi_2 \approx -0.251$.[6]

Anche il coefficiente di drift $\omega \approx 3.3054$ per il random walk supera il test di significativià sulla statistica z, valendo $z \approx 3.057$.

Anticipando le analisi sui residui condotte in § 4.1, non è necessario effettuare trasformazioni nei dati: i due modelli trovati sono adeguati nel descrivere il processo in esame, producendo entrambi dei residui riconducibili a rumore bianco.

Tab. 2. Statistica z per i parametri del modello ARIMA(0,1,2) con drift.

	ψ_1	ψ_2	ω
z	−2.246997	−1.602865	7.224292

2.2 ARMA con trend deterministico

Sebbene i test ADF e KPSS abbiano indicato la presenza di un trend stocastico, proviamo a stimare sulla serie y_t di Fig. 1b un modello ARMA sui residui rispetto ad un trend lineare deterministico:

[5] Condividendo i tre modelli l'ordine di differenziazione ($d = 1$), sono confrontabili con AICc in quanto modelli annidati.

[6] L'invertibilità ci assicura che ACF e PACF siano determinate univocamente dal modello considerato.

Tab. 3. Statistica z per i parametri del modello ARMA(1,0) con trend deterministico lineare, ossia $y_t = \eta t + \xi_t$ dove $\xi_t = \mu + \varphi_1 \xi_{t-1} + \varepsilon_t$ (con ε_t rumore bianco).

	φ_1	μ	η
z	5.299659	11.891981	27.834752

- la funzione `auto.arima` identifica un modello AR(1) ≡ ARMA(1,0) con trend lineare, i cui coefficienti superano tutti il test di significatività sulla statistica z (Tab. 3);
- il coefficiente autoregressivo vale $\varphi_1 \approx 0.5695$, per cui il polinomio AR
$$\phi_1(L) = 1 - \varphi_1 L$$
ha come radice $L_1 = 1/\varphi_1 \approx 1.756$ (assicurando, al netto del trend lineare, la stazionarietà del modello).

3

Provide a formal definition (with formulas) of the two models, emphasizing the different theoretical properties, then report a summary of parameter estimates for each of them.

3.1 Modelli Arima

3.1.1 ARIMA(0,1,0) con drift

Il random walk con drift è definito da

$$Y_t = \omega + Y_{t-1} + \varepsilon_t$$
$$\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$$

con $\omega \approx 3.3054$ e $\sigma^2 \approx 66.66$ per la serie qui considerata.

Il fatto che sia $\omega \neq 0$ introduce un trend lineare nel processo, in quanto esso può essere riscritto come

$$Y_{t+1} = Y_0 + \omega t + \sum_{s=1}^{t} \varepsilon_s$$

Proprietà teoriche. Essendo il suo valor medio e varianza rispettivamente dati da

$$\mathbb{E}[Y_t] = \mathbb{E}[Y_0] + \omega t$$
$$\mathbb{V}[Y_t] = \mathbb{V}[Y_0] + \sigma^2 t$$

il random walk con drift di per sé non è stazionario, non avendo media e varianza costanti nel tempo; viceversa

$$\Delta Y_t \equiv Y_t - Y_{t-1} \sim \mathsf{WN}(\omega, \sigma^2)$$

Inoltre per ΔY_t il valore atteso condizionato e la varianza condizionata valgono

$$\begin{aligned}\mathbb{E}[\Delta Y_t | \mathscr{I}_{t-1}] &= \omega \\ \mathbb{V}[\Delta Y_t | \mathscr{I}_{t-1}] &= \sigma^2\end{aligned}$$

3.1.2 ARIMA(0,1,2) con drift

L'equazione del modello selezionato in § 2.1 è

$$Y_t = Y_{t-1} + \omega + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} \quad (1a)$$
$$\varepsilon_t \sim \mathsf{WN}(0, \sigma^2) \quad (1b)$$

dove, per la serie in questione,

$$\begin{aligned}\psi_1 &\approx -0.3039 \\ \psi_2 &\approx -0.2510 \\ \omega &\approx 3.3954 \\ \sigma^2 &\approx 61.2\end{aligned}$$

Proprietà teoriche. In questo caso il processo ΔY_t è descritto da un processo MA(2), non riscontrando contributi auto-regressivi nella stima prodotta da `auto.arima`. In generale per un processo MA(q) valgono le seguenti considerazioni:

- il processo è sempre ergodico e quindi anche stazionario;
- la ACF vale 0 per lag $L > q$;
- la PACF va a zero esponenzialmente.

Nel nostro caso per il modello MA stimato su ΔY_t possiamo aggiungere che

- il processo da esso prodotto è pure invertibile, poiché il polinomio $\psi_2(L) = 1 + \psi_1 L + \psi_2 L^2$ ha radici esterne al cerchio unitario (cfr. § 2.1), per cui il processo potrebbe essere espresso come AR(∞);
- la costante $\omega \neq 0$ fa sì che, ritornando da ΔY_t sui dati originali Y_t, si introduca un trend deterministico ωt, ossia[7]

$$Y_t = Y_0 + \omega t + \sum_{s=1}^{t} (\varepsilon_s + \psi_1 \varepsilon_{s-1} + \psi_2 \varepsilon_{s-2}) \quad (2)$$

Poiché, per un generico MA(q)

$$U_t = \mu + \sum_{j=0}^{q} \psi_i \varepsilon_{t-j} \quad (\text{con } \psi_0 = 1)$$

valore atteso e varianza valgono rispettivamente $\mathbb{E}[U_t] = \omega$ e $\mathbb{V}[U_t] = \sigma^2 \sum_{j=0}^{q} \psi_j^2$, nel nostro caso abbiamo

$$\mathbb{E}[\Delta Y_t] = \omega \Rightarrow \mathbb{E}[Y_t] = \mathbb{E}[Y_0] + \omega t$$

$$\mathbb{V}[\Delta Y_t] = \sigma^2 \sum_{j=0}^{2} \psi_j^2 \Rightarrow \mathbb{V}[Y_t] = \mathbb{V}[Y_0] + t\sigma^2 \sum_{j=0}^{2} \psi_j^2$$

mentre per valore atteso condizionato e varianza condizionata abbiamo

$$\begin{aligned}\mathbb{E}[\Delta Y_t | \mathscr{I}_{t-1}] &= \omega + \sum_{j=1}^{2} \psi_j \varepsilon_j \\ \mathbb{V}[\Delta Y_t | \mathscr{I}_{t-1}] &= \sigma^2\end{aligned}$$

3.1.3 Riepilogo

I due modelli proposti sono accomunati dalla presenza di un trend lineare dato dal drift ω: tali modelli sono stati comunque ricavati avendo applicato a monte la rimozione della radice unitaria, per poi stimare i modelli su ΔY_t.

Ciò che invece cambia è la modellazione della componente aleatoria ε_t, in quanto:

$$\text{ARIMA(0,1,0)} \quad Y_t = Y_0 + \omega t + \sum_{s=1}^{t} \varepsilon_s$$

$$\text{ARIMA(0,1,2)} \quad Y_t = Y_0 + \omega t + \sum_{s=1}^{t} \sum_{j=0}^{2} \psi_j \varepsilon_{s-j}$$

[7] Per induzione, da Eq. (2) ritorniamo a Eq. (1a):

$$Y_1 = Y_0 + \omega + \varepsilon_1 + \psi_1 \varepsilon_0 \qquad \text{(passo iniziale soddisfatto)}$$

$$\begin{aligned} Y_t &= Y_0 + \omega t + \sum_{s=1}^{t} (\varepsilon_s + \psi_1 \varepsilon_{s-1} + \psi_2 \varepsilon_{s-2}) \\ &= \underbrace{Y_0 + \omega(t-1) + \sum_{s=1}^{t-1} (\varepsilon_s + \psi_1 \varepsilon_{s-1} + \psi_2 \varepsilon_{s-2})}_{Y_{t-1}} + \omega + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} \end{aligned}$$

3.2 ARMA(1,0) con trend lineare

A differenza dei modelli ARIMA appena trovati, in questo caso il processo è modellato come ARMA lavorando sui residui lasciati dal trend lineare. Il modello selezionato in § 2.2 è descritto dall'equazione (con $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$)

$$
\begin{aligned}
y_t &= \varphi_1 y_{t-1} + \omega + \delta t + \varepsilon_t \\
\varphi_1 &\approx 0.5695 \\
\omega &\approx 52.150 \quad (\mu \approx 11.8920, \eta \approx 27.8348) \\
\delta &\approx 1.501 \\
\sigma^2 &\approx 53.85
\end{aligned}
$$

I valori di ω e δ tengono conto che il modello con trend lineare ha errori ξ_t AR(1):

$$
\begin{aligned}
Y_t &= \eta t + \xi_t \\
\xi_t &= \mu + \varphi_1 \xi_{t-1} + \varepsilon_t \\
\Rightarrow Y_t &= \eta t + \mu + \varphi_1 \xi_{t-1} + \varepsilon_t \\
&= \eta t + \mu + \varphi_1 (Y_{t-1} - \eta (t-1)) + \varepsilon_t \\
&= \underbrace{\eta (1 - \varphi_1)}_{\delta} t + \underbrace{(\mu + \varphi_1 \eta)}_{\omega} + \varphi_1 Y_{t-1} + \varepsilon_t
\end{aligned}
$$

Nel precedente svolgimento si è passati da una formulazione con trend lineare ηt ed errori ξ_t AR(1) alla forma equivalente AR(1) su Y_t con trend lineare $g(t) = \omega + \delta t$ e $\varepsilon \sim \mathrm{WN}(0, \sigma^2)$.

Proprietà teoriche. Adesso il modello è ottenuto da processo stazionario sommato ad una funzione deterministica (in questo caso lineare): stavolta il processo stazionario è descritto da un modello AR(1), il cui polinomio $\phi_1(L) = 1 - \varphi_1 L$ ha radice con modulo superiore a 1 (cfr. § 2.2) e pertanto il processo prodotto dal modello, una volta sottratto il trend lineare, gode delle proprietà di stazionarietà ed ergodicità, oltre a quella d'invertibilità (sempre vera per un processo AR).

Il valore atteso vale $g(t) = \omega + \delta t$, mentre la varianza è la stessa di ξ_t (la funzione di covarianza di Y_t coincide con quella di ξ_t).

4

For each of the two models, report and discuss some opportune diagnostics to check their adequacy.

4.1 Modelli Arima

4.1.1 Analisi visuale dei residui

Fig. 4 mostra le caratteristiche dei residui r_n per ARIMA(0,1,0) (ossia il random walk) e per ARIMA(0,1,2), entrambi con drift[8], allo scopo di valutare se siano riconducibili o meno a quelle del rumore bianco:

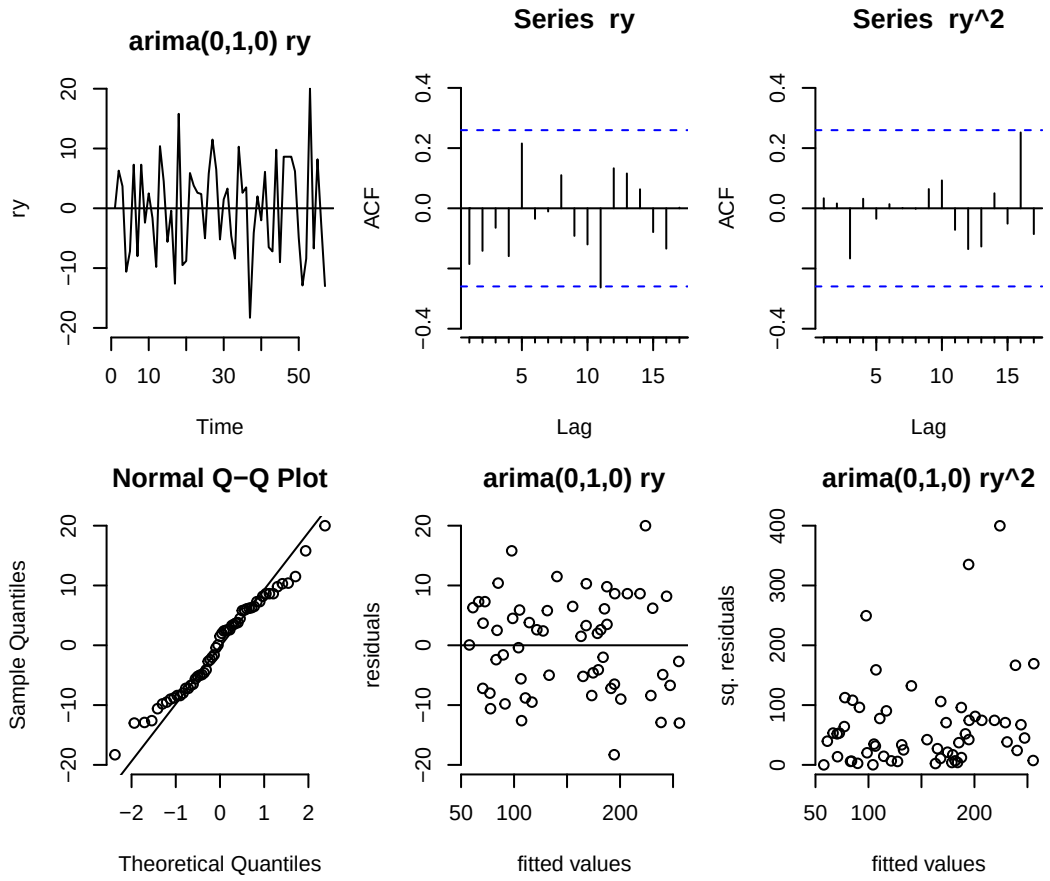
- i valori di ACF per r_n sono statisticamente non diversi da 0, a parte lag 11[9] su ARIMA(0,1,0) e ARIMA(0,1,2), per cui i residui possono considerarsi scorrelati;
- inoltre per essere compatibili con un processo di rumore bianco, anche i residui devono avere la stessa varianza, e questo è grossomodo verificato per tutti i modelli dall'assenza di autocorrelazione nell'ACF dei residui al quadrato r_n^2 (per lo spike a lag 10 su ARIMA(0,1,2) si veda § 4.1.2);
- i residui fluttuano in modo regolare rispetto alla baseline, senza mostrare pattern dovuti a eteroschedasticità condizionata;
- a supporto di ciò non si nota, al crescere dei valori stimati, un aumento nelle ampiezze dei residui (al più è presente qualche outlier, specie per il randon walk);
- le code del QQ-plot lasciano a desiderare per il modello ARIMA(0,1,2) mentre per ARIMA(0,1,0) i residui si distribuiscono in modo più Normale;

Ricapitolando, entrambi modelli analizzati producono residui caratterizzabili come rumore bianco; tuttavia:

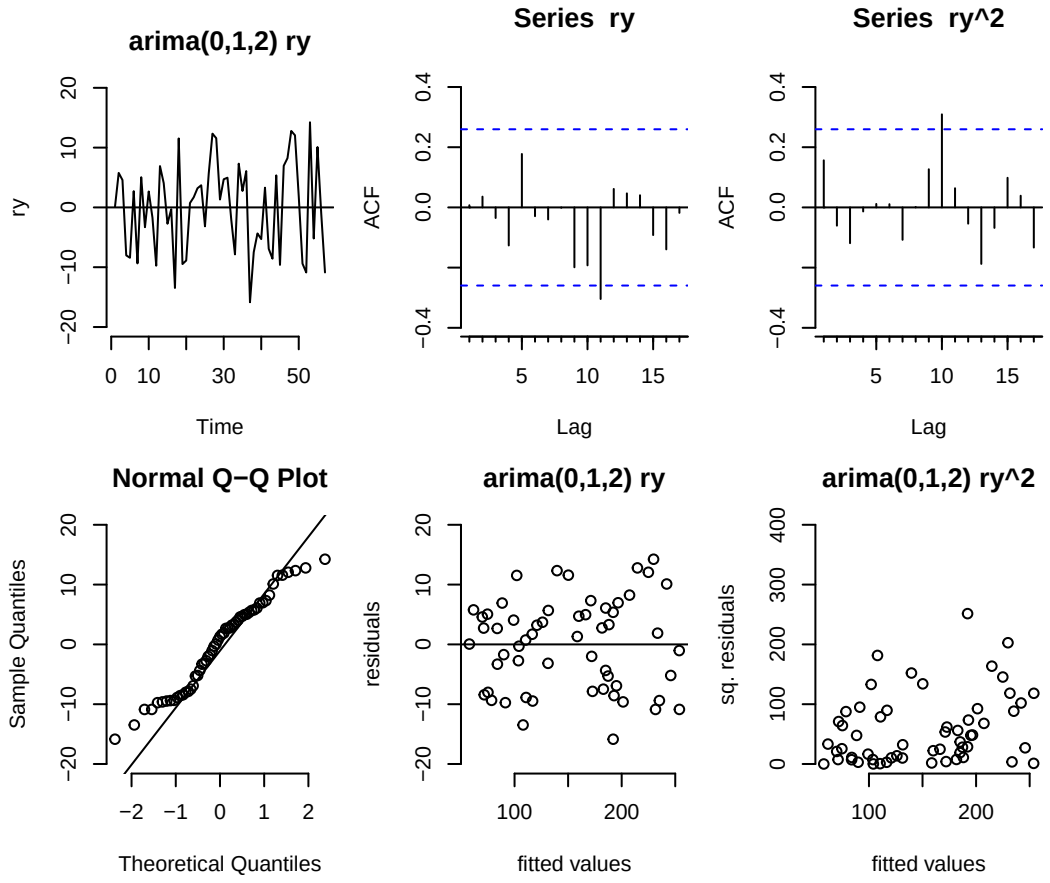
- ARIMA(0,1,2) mostra residui più contenuti del random walk, però nelle code il loro comportamento si discosta maggiormente dalla distribuzione Normale (e ciò impatta sugli intervalli di previsione, dato che sono ricavati ipotizzando dei residui gaussiani);

[8] Poiché i modelli ARIMA adesso considerati hanno tutti drift, d'ora in poi tale precisazione sarà sottintesa per brevità di esposizione.

[9] Tali occorrenze a lag così elevati sono ritenute spurie, in quanto le bande di accettazione sono riferite a un test di significatività per un lag alla volta (anche per ACF su r_n^2). Si veda anche il successivo test statistico in § 4.1.2.



(a) ARIMA(0,1,0) con drift



(b) ARIMA(0,1,2) con drift

Fig. 4. Residui a confronto per ARIMA(0,1,0), ARIMA(0,1,2).

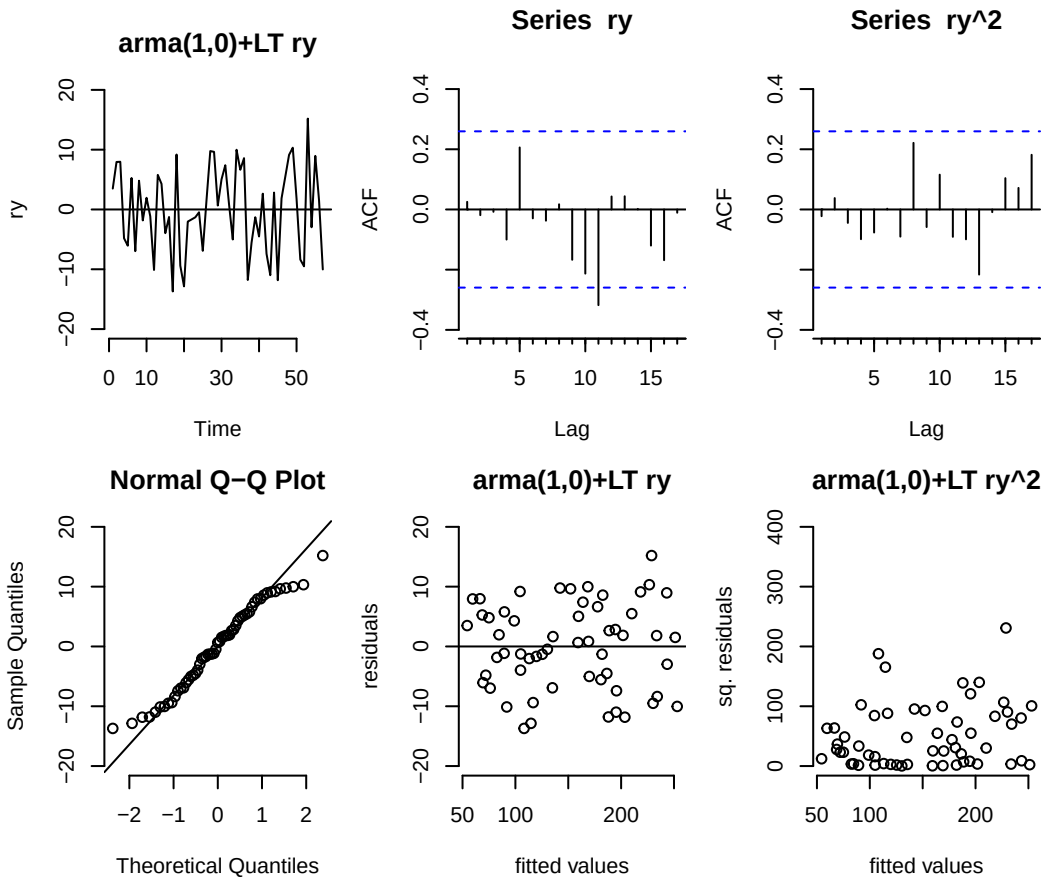


Fig. 5. Analisi dei residui per ARMA(1,0) con trend lineare.

- viceversa, il random walk esibisce il QQ-plot con punti meglio disposti rispetto alla diagonale teorica di Normalità.

4.1.2 Test statistici

Il test di Box-Pierce ci consente di verificare che i residui siano effettivamente scorrelati fino ad un certo lag: Tab. 4 riporta i test di Box-Pierce a vari lag sui residui $\{r_n\}$ e $\{r_n^2\}$ (per scrupolo il test è stato applicato anche a lag $= 16$ su $\{r_n^2\}$ nel caso del random walk).

Per p-value ben superiori a 0.05 siamo confidenti che i residui non mostrino correlazione significativa fino al lag testato: pertanto i test vengono superati senza difficoltà.

Tab. 4. Test di Box-Pierce per i modelli ARIMA.

	ARIMA(0,1,0)		ARIMA(0,1,2)	
residui	lag	p-value	lag	p-value
r_n	11	0.2665	11	0.3178
r_n^2	16	0.9260	10	0.4891

4.2 ARMA con trend lineare

4.2.1 Analisi visuale dei residui

Fig. 5 mostra l'analisi per i residui ottenuti dal modello AR(1) con trend lineare detrministico: i risultati sono paragonabili a quelli di ARIMA(0,1,2), e valgono le stesse considerazioni già fatte per Fig. 4, ma con un QQ-plot migliore nella parte centrale con punti più vicini alla diagonale (resta lo scarso fit nelle code).

4.2.2 Test statistici

Anche nel caso del modello ARMA(1,0) con trend lineare, merita investigare con Box-Pierce fino a lag 11 per i residui ottenendo 0.2846 come p-value; sui residui al quadrato abbiamo poi per lag 13 che p-value $= 0.7713$: il test è quindi superato.

5

Estimate the forecast accuracy of each of the two models at some horizons in the future th-

rough rolling window cross-validation, then provide a graphical comparison and discuss the differences as a function of the horizon.

Fig. 6 riporta, per i tre modelli proposti, le seguenti metriche per orizzonti $h = 1 \dots 10$:

$$\text{RMSE}_h = \sqrt{\frac{1}{T-h}\sum_{t=1}^{T-h}\underbrace{(y_{t+h} - \mathbb{E}(Y_{t+h}|\mathscr{I}_t))}_{\hat{e}_t(h)}^2}$$

$$\text{MAE}_h = \frac{1}{T-h}\sum_{t=1}^{T-h}|\hat{e}_t(h)|$$

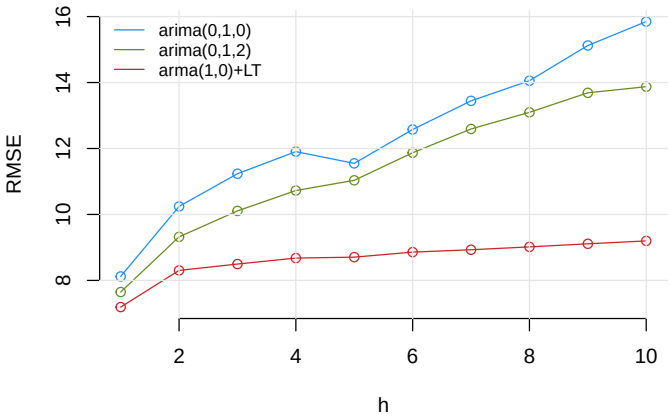
$$\text{MAPE}_h = \frac{100}{T-h}\sum_{t=1}^{T-h}\left|\frac{\hat{e}_t(h)}{y_{t+h}}\right|$$

$$\text{scal. MAPE} = \frac{\text{MAPE}_h}{(T-h)^{-1}\sum_{t=1}^{T-h}|y_{t+h} - y_t|}$$

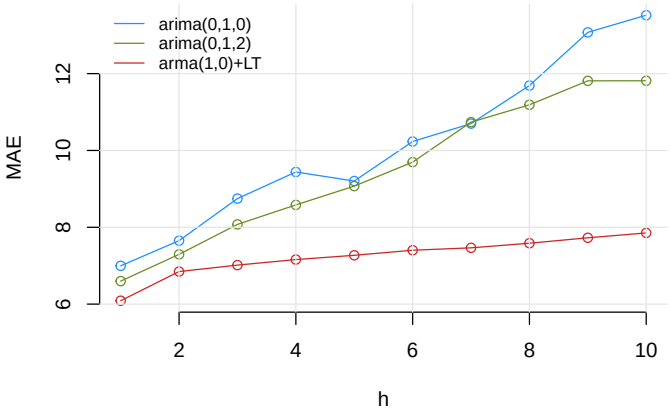
L'orizzonte massimo a 10 anni è stato fissato per estremizzare il comportamento dell'errore di predizione; nella pratica è ragionevole non superare i 4-5 anni: tale regola è confermata dal comportamento "ballerino" dell'errore per $h > 4$, in cui l'andamento talvolta decresce (questo fenomeno è imputabile alla mancanza di un numero adeguato di dati rispetto all'orizzonte di previsione). L'errore relativo fornito dallo scaled MAPE va letto diversamente: dal grafico di Fig. 6d ha un trend decrescente e risulta sempre inferiore a 1; significa che il modello lavora meglio del predittore naïve il cui costo è dato da $(T-h)^{-1}\sum_{t=1}^{T-h}|y_{t+h} - y_t|$, come era lecito aspettarsi.

A prima vista, il modello ARMA con trend deterministico lineare appare essere il migliore: ha un comportamento costante e risulta inferiore agli errori dei due modelli ARIMA, che invece tendenzialmente crescono all'aumentare di h. Tuttavia, constatando che la "scala" della serie fornita dalla sua deviazione standard vale `sd(y)` ≈ 58.96 e che gli errori in termini di RMSE per $h \leq 4$ differiscono al più di 3 (quintali per ettaro), così come gli errori percentuali del MAPE stanno tra 4% e 6%, possiamo ritenere i tre modelli equivalenti.

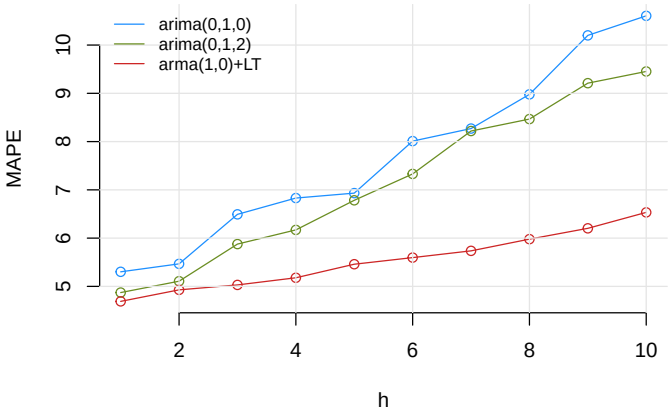
Pertanto il random walk ARIMA(0,1,0) con drift può essere preso come modello finale, in virtù della sua maggiore semplicità.



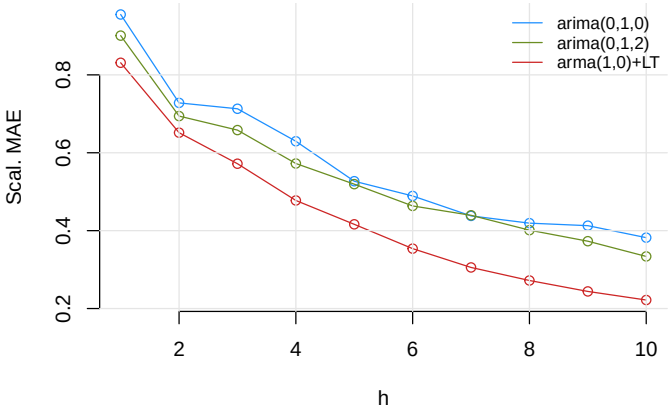
(a) Root Mean Squared Error



(b) Mean Absolute Error



(c) Mean Absolute Percentage Error

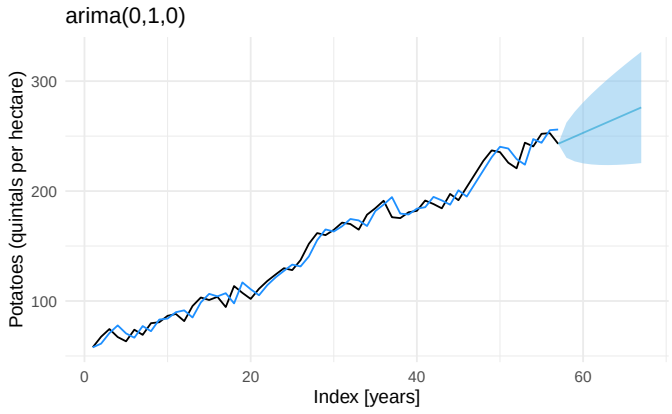


(d) Scaled Mean Absolute Error

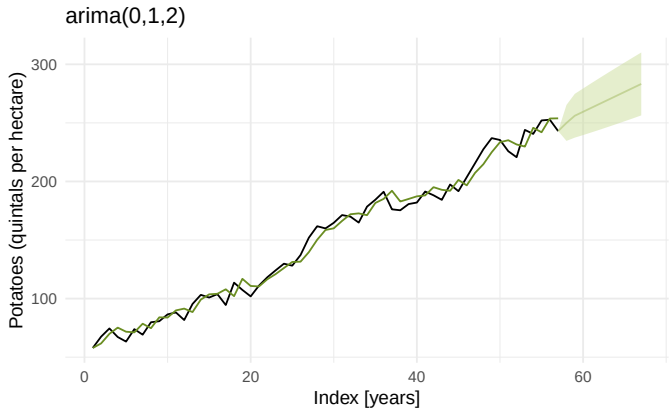
Fig. 6. Errori di predizione secondo varie metriche.

6

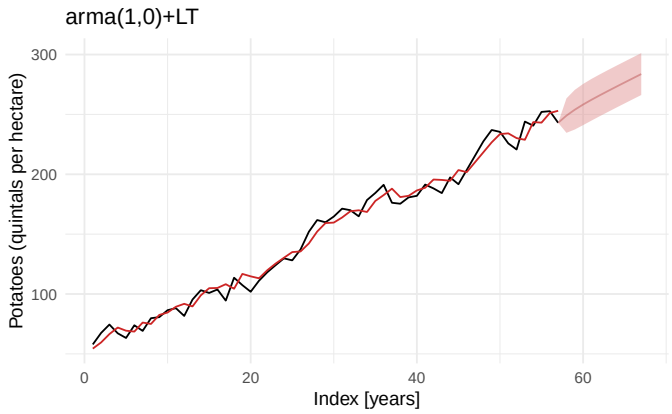
Display forecasts of future values of the time series based on each of the two models, then discuss the difference between them in light of their theoretical properties (e.g., presence of deterministic and/or stochastic trend)



(a) ARIMA(0,1,0)



(b) ARIMA(0,1,2)



(c) ARMA(1,0) + lin. trend

Fig. 7. Predizioni dei modelli.

Fig. 7 mostra le previsioni future fornite dai tre modelli proposti: la presenza di un trend lineare in tutti e tre fa sì che le previsioni crescano indefinitamente in modo lineare, una volta terminata la dipendenza dai valori a disposizione; viceversa, il trend stocastico nei modelli ARIMA non contribuisce all'andamento (in questo caso crescente) della successione.

Per quanto riguarda le bande di previsione, per il modello ARMA(1,0) con trend deterministico esse hanno ampiezza costante, mentre quelle per i modelli ARIMA continuano a crescere per la presenza anche del trend stocastico la cui varianza cresce nel tempo. Pertanto è più prudente effettuare previsioni, specialmente se a lungo termine, utilizzando i trend stocastici in quanto consentono una maggiore incertezza nella crescita futura.

A Serie senza imputazione

All'inizio dell'elaborato la serie è stata imputata per precauzione, ma senza mostrare che fosse necessario: riportiamo brevemente qui lo svolgimento senza la modifica apportata alla serie in corrispondenza del 1991 e 1992.

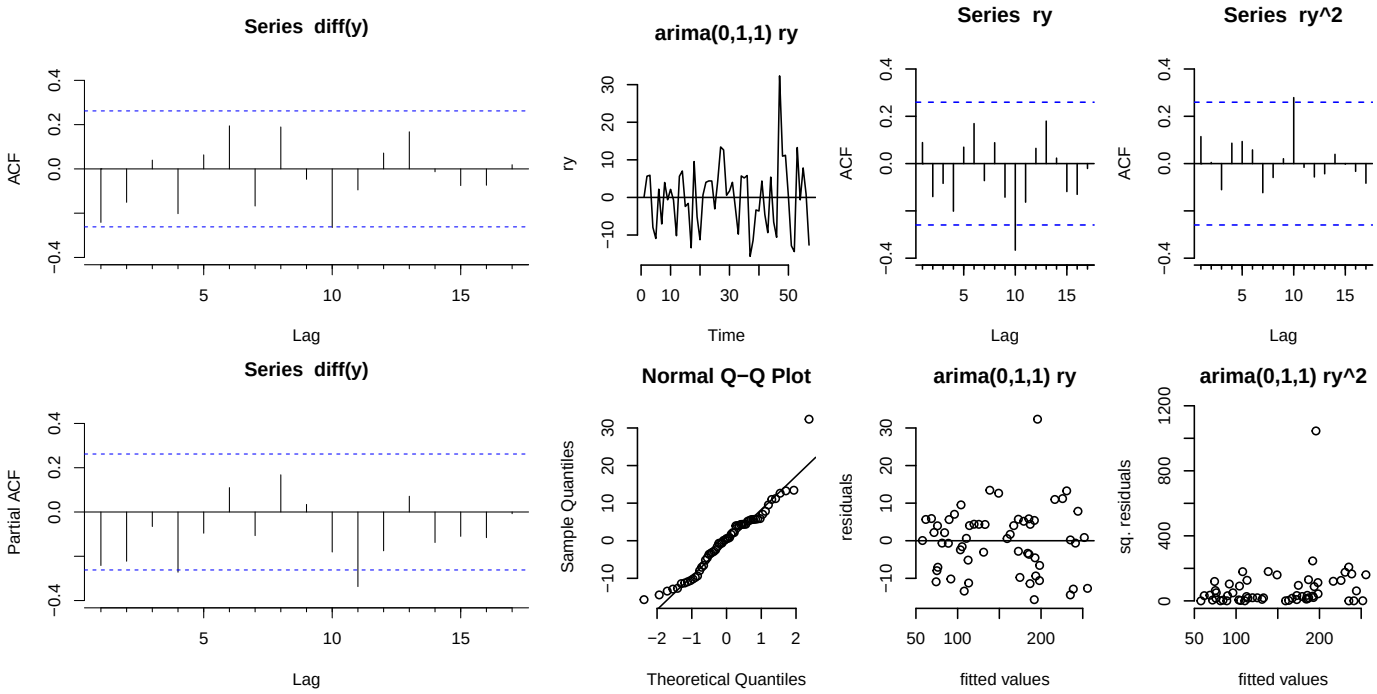
A.1 Selezione del modello

Fig. 8a riporta ACF e PACF sulla serie dopo aver rimosso la radice unitaria, molto simile a quello in Fig. 3: ripetendo le considerazioni già fatte in § 2.1, stavolta il modello con AICc più basso è lo ARIMA(4,1,0) con drift (AICc=411.89), mentre per il random walk ARIMA(0,1,0) con drift vale AICc=413.62.

La funzione `auto.arima` fornisce ancora un ARIMA(0,1,2) con drift (AICc=408.99): scartiamo perciò lo ARIMA(4,1,0) con drift sia a causa del suo AICc peggiore che per la sua maggior complessità. Tab. 5 riporta i valori dei coefficienti per ARIMA(0,1,2), insieme alla statistica z.

Tab. 5. Coefficienti e statistica z per ARIMA(0,1,2) con drift.

coeff.	valore	stat. z
ψ_1	-0.3797	-2.77
ψ_2	-0.2656	-1.54
ω	3.416	7.94



(a) ACF e PACF sulla serie differenziata.

(b) Analisi dei residui per ARIMA(0,1,1) con drift.

Fig. 8

Poiché ψ_2 non è statisticamente diverso da 0, riduciamo il modello a un ARIMA(0,1,1) con drift (AICc=409.47), e dalla funzione `Arima` otteniamo stavolta $\psi_1 = -0.4622$ ($z \approx -2.86$) e $\omega = 3.3676$ ($z \approx 5.22$):

$$Y_t = \omega + \varepsilon_t + \psi_1 \varepsilon_{t-1}$$

Fig.8b riporta l'analisi grafica sui residui del modello ARIMA(0,1,1) con drift, da cui emerge che i residui non siano del tutto riconducibili a rumore bianco:

- in particolare si nota uno spike esterno alla banda di significatività al lag 10 nel grafico ACF dei residui;
- anche nel grafico ACF sui residui al quadrato si ha una leggera fuoriuscita dalla banda di significatività sempre al lag 10;
- si nota infine un chiaro outlier nei residui per il campione #47 (ossia proprio per l'anno 1992).

D'altra parte il Box-test al lag 10 riporta un p-value pari a 0.1 (idem per lag 11), suggerendo assenza di correlazione fino al lag considerato; per quanto riguarda i Box-test sui residui al quadrato si ottengono rispettivamente dei p-value pari a 0.62 e 0.71, per lag 10 e 11.

Conclusione. Vista la situazione non del tutto chiara, conviene per semplicità di svolgimento provare ad imputare i valori della serie come descritto all'inizio dell'elaborato.