Elaborato per Time Series, MD2SL 23/24
Serie stagionale

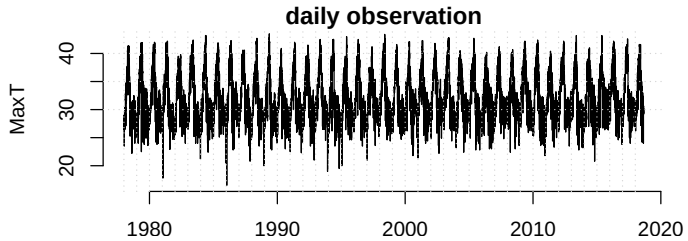
Dario Comanducci, 19 marzo 2025

1

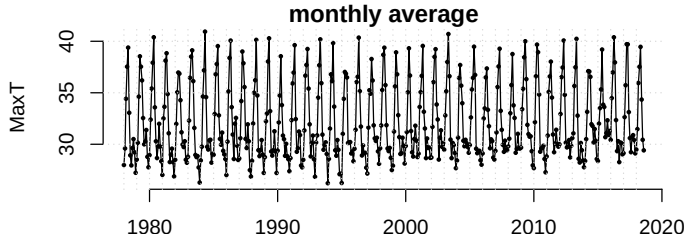
Describe the data, explaining which phenomenon they represent and providing details on the measurements (time span, frequency, scale). Display the time series and discuss its appearance in reference to stationarity and to the presence of seasonal patterns and outliers.

Per la serie con stagionalità i dati provengono dalle rilevazioni meteorologiche giornaliere per una località dell'India (Hyderabad) raccolti su un periodo di 40 anni, dal 1978 al 2018, disponibili su Kaggle[1] scaricando il file "ICRISAT Weather 1978 to 2018.xlsx".

Dal file excel è stata isolata la serie della temperatura massima `maxT` espressa in gradi Celsius (priva di valori mancanti), che da giornaliera è stata ridotta a mensile, calcolando per ogni mese la media dei valori rilevati (Fig. 1): viene qui studiata quest'ultima serie (Fig. 1b).



(a) osservazioni giornaliere

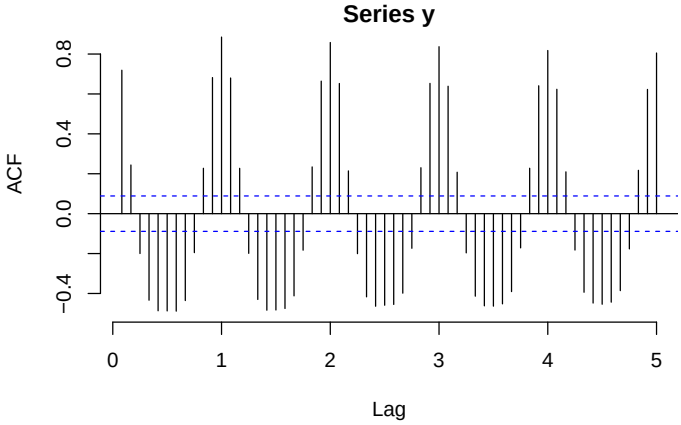


(b) medie mensili

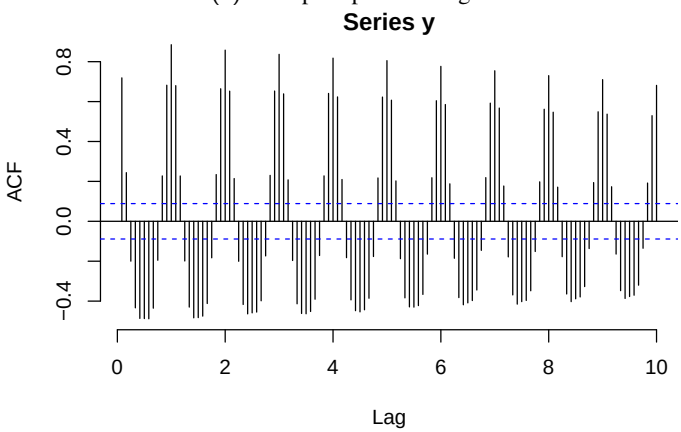
Fig. 1. Rilevazioni della temperatura massima a Hyderabad.

[1] https://www.kaggle.com/datasets/sulphatet/daily-weather-data-40-years

La serie esibisce un chiaro andamento periodico dovuto all'alternarsi delle stagioni, con pattern che si ripetono ogni 12 mesi: questo è confermato anche dal grafico della funzione ACF in Fig. 2, dove si nota uno schema regolare con massimi positivi ogni 12 lag, e similmente dei minimi negativi (di ampiezza minore) traslati di 6 (mesi) rispetto a quelli positivi, dovuti alla contrapposizione tra temperature estive ed invernali; inoltre tali picchi decrescono molto lentamente, restando rilevanti anche a lag molto elevati (Fig. 2b).



(a) ACF per i primi 60 lag



(b) ACF fino a 120 lag

Fig. 2. funzione ACF della serie in Fig.1b

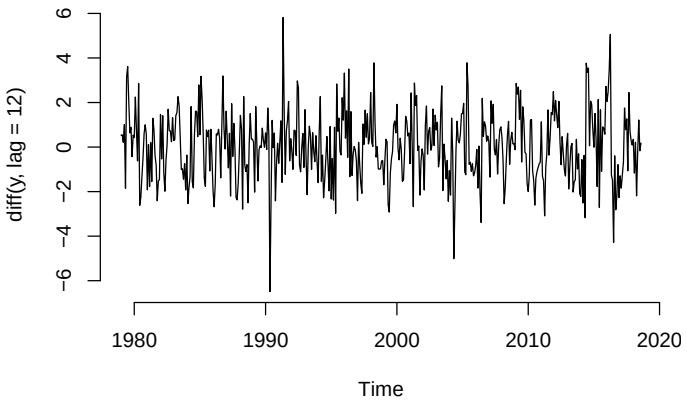


Fig. 3. Serie differenziata stagionalmente.

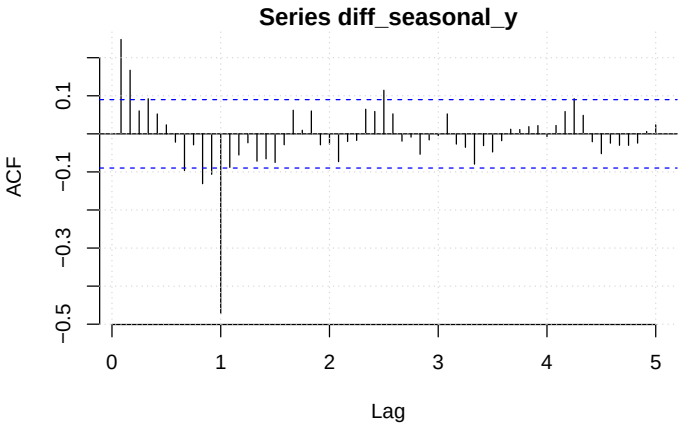
Tuttavia dai test statistici di stazionarietà abbiamo che:

- il p-value relativo al test ADF è inferiore a 0.01, per cui la serie dovrebbe essere stazionaria (viene rifiutata la presenza di radici unitarie);
- per il test KPSS invece il p-value è maggiore di 0.1, portando a non rifiutare l'ipotesi nulla di stazionarietà.

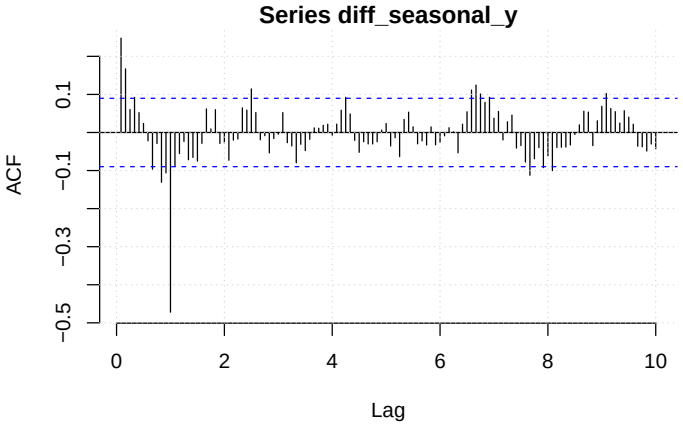
Il lento decrescere dei picchi nel grafico ACF di Fig. 2b però indica in maniera palese la presenza almeno di una radice unitaria stagionale[2] con periodo 12: si ha conferma di ciò considerando la serie ottenuta da `diff(y,lag=12)` (Fig. 3), la cui ACF in Fig. 4 mostra stavolta un solo spike elevato a lag 12, restando poi per lag maggiori di 2 sempre all'interno o poco sopra le bande di significatività; non pare quindi necessario applicare ulteriori differenziazioni non stagionali, anche grazie al supporto dei test statistici che stavolta sono in accordo all'ACF (avendo di nuovo un p-value inferiore a 0.01 per il test ADF e superiore a 0.1 per il KPSS).

Nella serie differenziata stagionalmente di Fig. 3 si notano un paio di spikes isolati in corrispondenza del maggio 1990 e 1991, outlier dovuti alla temperatura "anomala" per maggio 1990 (33.8° C), mentre per maggio 1989 e 1991 si ha rispettivamente 40.3°C e 39.6° C. Rimandiamo la gestione di tali spikes in § 2.2.

[2] Un errore di falso negativo dei test in caso di radici unitarie stagionali è abbastanza frequente.



(a) ACF su 60 lag



(b) ACF su 120 lag

Fig. 4. ACF per `diff(y,lag=12)`.

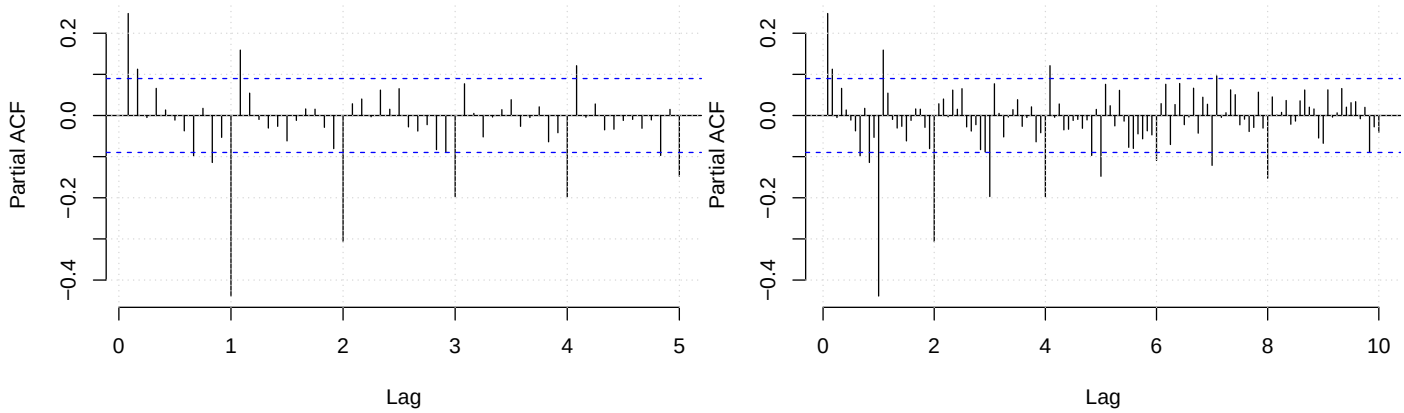
2

Find a seasonal ARIMA model 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 Analisi con autoarima

Poiché non sembrano necessarie ulteriori differenziazioni (né stagionali, né non), con il supporto anche della funzione PACF (Fig. 5) si può ipotizzare che la componente AR stagionale non superi ordine 8: la funzione `autoarima`, posto `max.P=8`, produce il modello ARIMA$(0,0,1)(2,1,2)_{12}$ con drift.[3]

[3] Da un'ispezione visuale di ACF e APCF, altri ordini papabili per un ARIMA$(p,d,q)(P,D,Q)_{12}$ sarebbero stati, fissati $Q=1$ e $q=2$, $P \in \{2,5,8\}$ e $p \in \{1,2\}$: al loro interno il modello migliore risulta essere lo ARIMA$(1,0,2)(2,1,1)_{12}$, con AICc $= 1452.27$, ma per il modello ottenuto con auto.arima vale AICc $= 1451.23$ (quindi preferibile). Anche con ARIMA$(1,0,2)(2,1,1)_{12}$ l'analisi dei residui risulta simile a quella di Fig. 6.



(a) PACF su 60 lag (b) PACF su 120 lag

Fig. 5. PACF per la serie di Fig. 3.

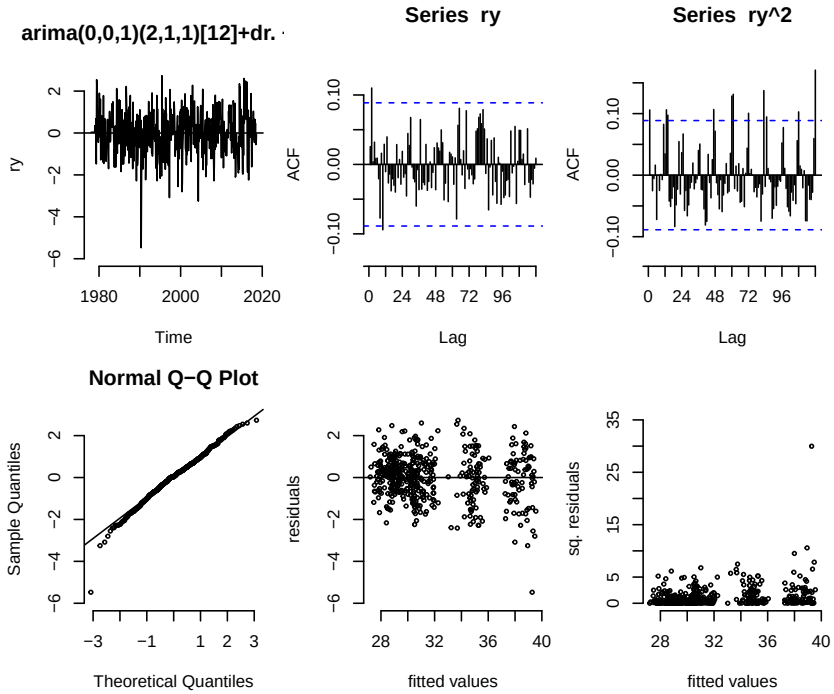


Fig. 6. Analisi grafica dei residui per ARIMA$(0,0,1)(2,1,2)_{12}$ con drift.

Tuttavia da Fig. 6 notiamo che i residui di tale modello potrebbero non essere riconducibili a rumore bianco:

- i coefficienti ACF sui residui al quadrato non sono tutti all'interno delle bande;
- inoltre i residui sembrano aumentare in ampiezza al crescere dei valori stimati, indizio di eteroschedasticità condizionata, seppur non sia uno schema così chiaro.
- anche nei Box-test in Fig. 7, dai p-value oltre 0.05, si ha assenza di correlazione significativa sui residui al quadrato perlopiù solo fino circa lag 60 (ossia 5 anni).

Sembra infine che i residui abbiano tre diversi comportamenti a seconda del valore predetto dal modello, producendo tre cluster distinti: residui più bassi per temperature predette inferiori a 34° C, un po' maggiori per valori stimati tra 34° C e 36° C e residui più alti per temperature sopra i 36° C.

2.2 Trasformazione della serie

Per avere una serie più facilmente modellabile con un SARIMA, innanzi tutto imputiamo

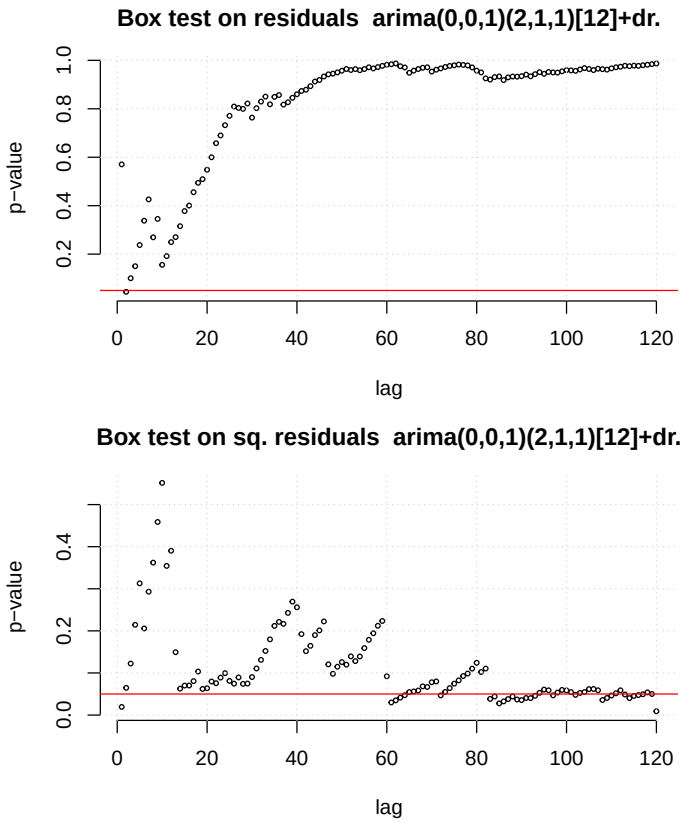


Fig. 7. Andamento dei p-value relativi ai test di Box sui residui al crescere del lag.

l'outlier rilevato a maggio[4] 1990, sostituendo al posto della temperatura anomala di 33.8° C il valore interpolato linearmente tra maggio 1989 e maggio 1991 (ossia 39.96° C), così da migliorare la leggibilità dei grafici sui residui.

La modifica più importante sui dati è comunque costituita dalla trasformazione Box-Cox con $\lambda = -0.9$ (valore ricavato tramite `BoxCox.ar(y)$mle`): Fig. 8 mostra la nuova serie già differenziata stagionalmente, dal momento che l'imputazione e la trasformazione non possono rimuovere la radice stagionale.

Stavolta il modello ottenuto sulla serie trasformata da `auto.arima` consiste in un $\mathrm{ARIMA}(0,0,1)(3,1,1)_{12}$ senza drift, con AICc $= -4848.31$. Osservando però Fig. 9 vale la pena di verificare anche un $\mathrm{ARIMA}(2,0,2)(5,1,1)_{12}$, a cui corrisponde AICc $= -4853.99$ in presenza di drift e AICc $= -4852.37$ in sua assenza: essendo lo AICc esibito dallo $\mathrm{ARIMA}(0,0,1)(3,1,1)_{12}$ superiore a quello

[4] Per aprile e giugno invece i valori dal 1989 al 1991 sono tutti simili.

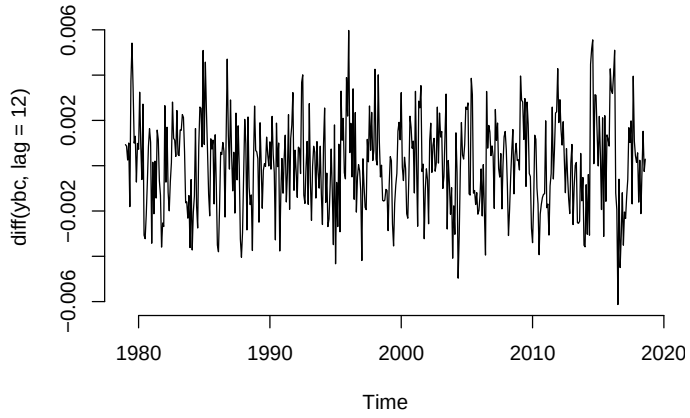


Fig. 8. Serie differenziata stagionalmente dopo imputazione e trasformazione Box-Cox.

dell'$\mathrm{ARIMA}(2,0,2)(5,1,1)_{12}$, ne segue che andrebbe impiegato quest'ultimo.

Dal confronto sui residui in § 4.1 nessuno dei due modelli è esente da difetti: anticipando il risultato dell'analisi condotta in § 4.1, il modello migliore pare essere lo $\mathrm{ARIMA}(2,0,2)(5,1,1)_{12}$ con drift.

3

Provide a formal definition (with formulas) and a description of the model, emphasizing its theoretical properties, then report a summary of parameter estimates.

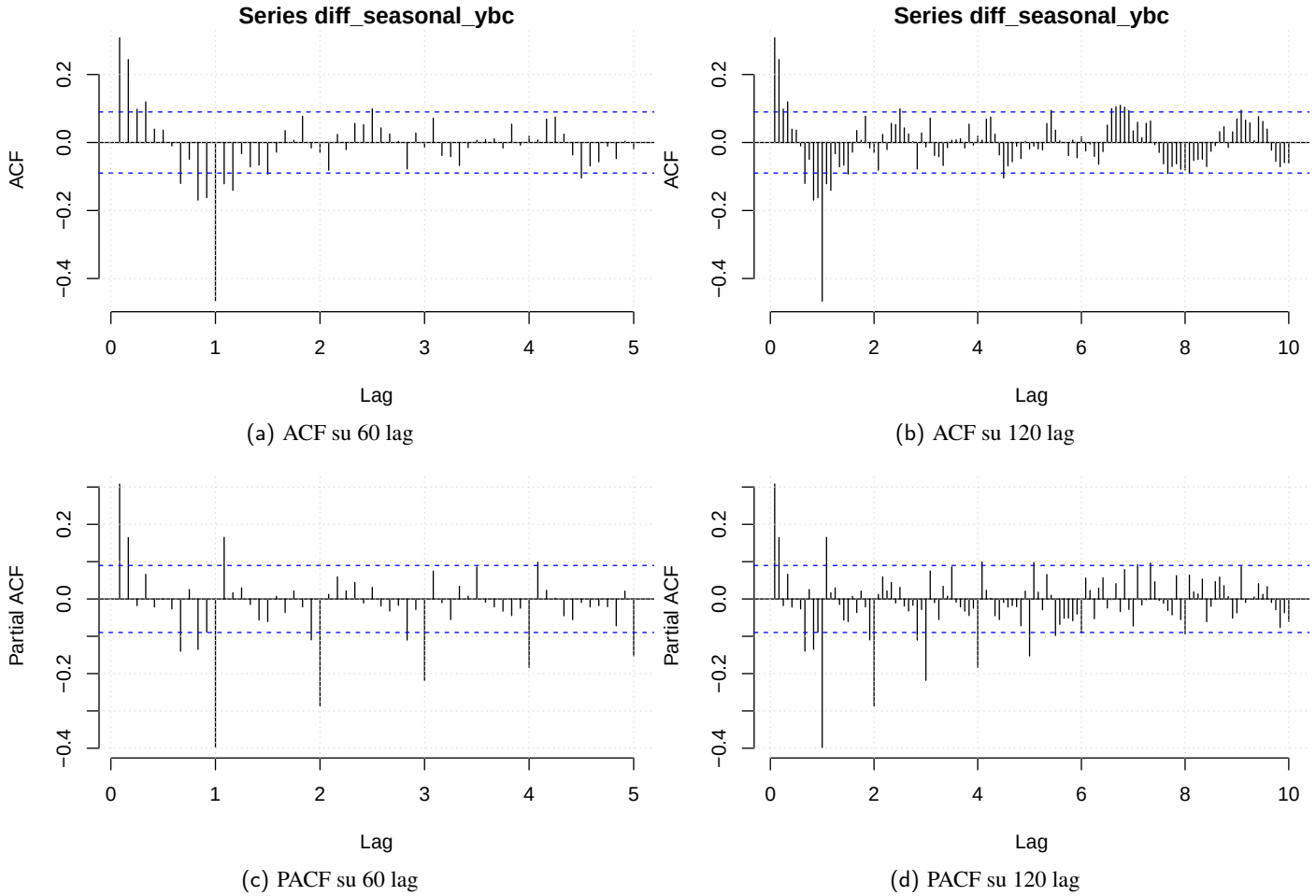
Il modello $\mathrm{ARIMA}(2,0,2)(5,1,1)_{12}$ con drift sulla serie trasformata con Box-Cox è in definitiva descritto dalla seguente equazione:

$$\Phi_p(\mathbb{L})\Phi_P^{(S)}(\mathbb{L}^{12})\Delta_S(\mathbb{L}^{12})Y_y = \omega + \Psi_q(\mathbb{L})\Psi_Q(\mathbb{L}^{12})\varepsilon_t$$

dove $\varepsilon_t \sim \mathrm{WN}(0,\sigma^2)$, con $\sigma^2 = 2.043 \cdot 10^{-6}$, ed essendo

$$\Phi_p(\mathbb{L}) = 1 - \phi_1\mathbb{L} - \phi_2\mathbb{L}^2$$
$$\Phi_P^{(S)}(\mathbb{L}^{12}) = 1 - \sum_{n=1}^{5}\phi_n^{(S)}\mathbb{L}^{12n}$$
$$\Delta_S(\mathbb{L}^{12}) = 1 - \mathbb{L}^{12}$$
$$\Psi_q(\mathbb{L}) = 1 + \psi_1\mathbb{L} + \psi_2\mathbb{L}^2$$
$$\Psi_Q^{(S)}(\mathbb{L}^{12}) = 1 + \psi_1^{(S)}\mathbb{L}$$

i cui coefficienti hanno i valori elencati in Tab. 1, assieme alla corrispondente statistica



(a) ACF su 60 lag

(b) ACF su 120 lag

(c) PACF su 60 lag

(d) PACF su 120 lag

Fig. 9. ACF in (a) e (b), e PACF in (c), (d) per la serie di Fig. 8.

z indicante la loro significativa differenza da 0. Notiamo che in realtà alcuni polinomi potrebbero essere semplificati, in quanto alcuni coefficienti mostrano valori per la statistica z inferiori in valore assoluto a 1.96:

$$\Phi_p(\mathbb{L}) = 1 - \phi_1\mathbb{L}$$
$$\Phi_P^{(S)}(\mathbb{L}^{12}) = 1 - \phi_5^{(S)}\mathbb{L}^{60}$$
$$\Psi_q(\mathbb{L}) = 1 + \psi_1\mathbb{L}$$

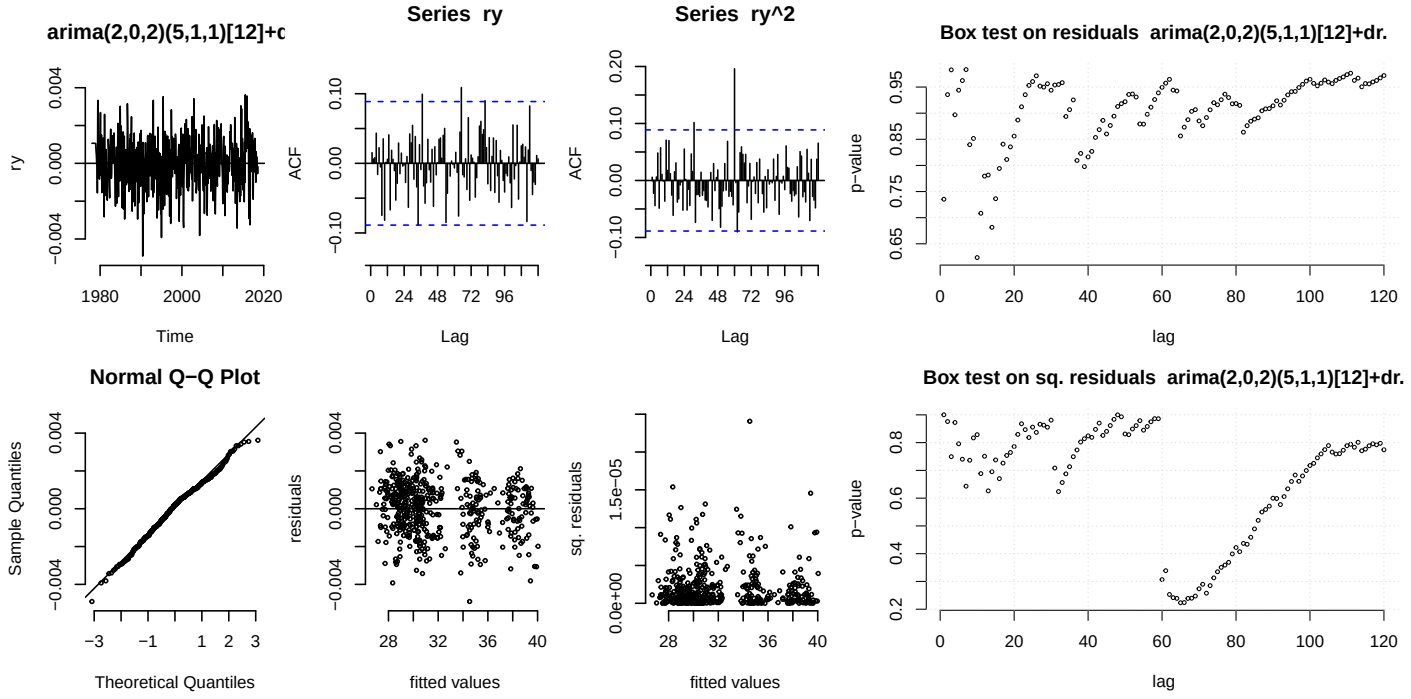
Discorso a parte merita il drift: R fornisce come valore $1.95 \cdot 10^{-6}$ che, combinato col la relativa statistica z pari a 0.03, può essere ritenuto non diverso da 0; evitiamo quindi la conversione in ω. Trascurare il contributo di ω consente inoltre di non introdurre trend deterministici (lineari) nel modello.

Un modello equivalente, ma più parsimonioso, sarebbe un ARIMA$(1,0,1)(5,1,1)_{12}$ senza drift. Per brevità di svolgimento, però ne omettiamo la stima.

Tab. 1. Coefficienti e statistica z per i polinomi nel modello ARIMA$(2,0,2)(5,1,1)_{12}$ con drift.

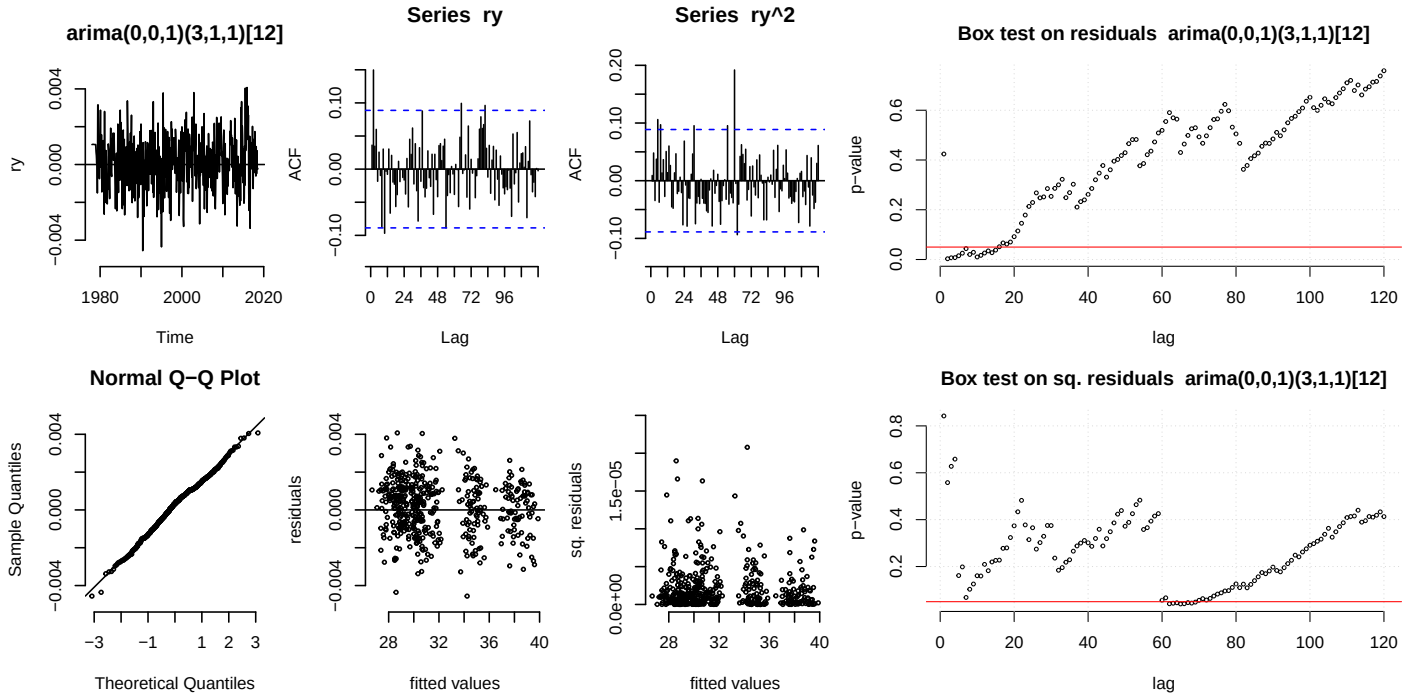
polinomio	coeff.	valore	stat. z
$\Phi_p(\mathbb{L})$	ϕ_1	0.0138	5.26
	ϕ_2	0.2296	1.29
$\Phi_P^{(S)}(\mathbb{L}^{12})$	$\phi_1^{(S)}$	−0.0200	−0.32
	$\phi_2^{(S)}$	−0.0702	−1.31
	$\phi_3^{(S)}$	−0.1030	−1.89
	$\phi_4^{(S)}$	−0.0574	−0.92
	$\phi_5^{(S)}$	−0.0348	−27.96
$\Psi_q(\mathbb{L})$	ψ_1	0.0138	5.61
	ψ_2	0.2296	−0.46
$\Psi_Q^{(S)}(\mathbb{L}^{12})$	$\psi_1^{(S)}$	−0.8941	−24.45

Riprendendo il discorso sul modello ARIMA$(2,0,2)(5,1,1)_{12}$, la presenza di una radice unitaria stagionale in $\Delta_S(\mathbb{L}^{12})$ è la causa dell'iniziale non stazionarietà della serie. La componente AR(5) della parte stagionale lega



(a) Analisi dei residui per ARIMA$(2,0,2)(5,1,1)_{12}$ con drift

(b) Box test per ARIMA$(2,0,2)(5,1,1)_{12}$ con drift



(c) Analisi dei residui per ARIMA$(0,0,1)(3,1,1)_{12}$

(d) Box test per ARIMA$(0,0,1)(3,1,1)_{12}$

Fig. 10. Analisi dei residui e Box-test sulla serie trasformata con Box-Cox.

per ciascun mese i valori del modello sui mesi corrispondenti dei 5 anni precedenti, mentre la componente MA(1) tiene conto dell'errore di previsione stagionale del periodo precedente. Infine il modello non stagionale mostra una dipendenza del valore corrente con gli ultimi 2 mesi (per AR(2)) e tiene conto anche degli errori negli ultimi due mesi (per MA(2)).

4

Report and discuss some opportune diagnostics to check the adequacy of the model. Estimate and discuss the forecast accuracy of the model at some horizons in the future through rolling window cross-validation.

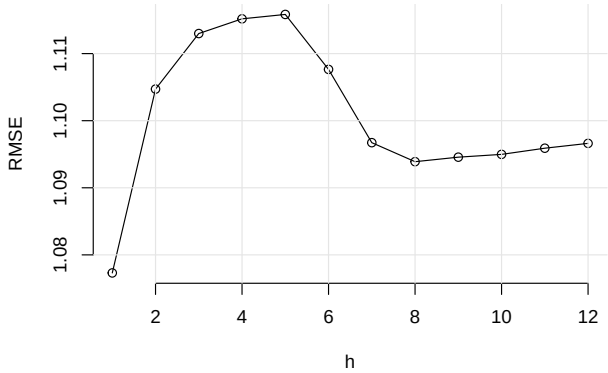
4.1 Diagnostiche sui residui

Fig. 10 mette a confronto l'analisi visuale dei residui ottenuti con i due modelli ARIMA$(2,0,2)(5,1,1)_{12}$ con drift (in alto) e ARIMA$(0,0,1)(3,1,1)_{12}$ (in basso), su cui facciamo le seguenti osservazioni:

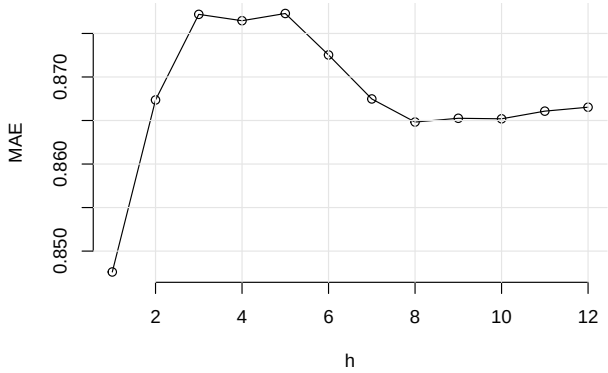
- non si nota più l'andamento crescente dei residui all'aumentare dei valori stimati presente in Fig. 6, ma la distribuzione dei residui sembra comunque piuttosto eterogenea lungo l'asse dei valori forniti da due modelli (permangono anche i tre cluster, stavolta con errori maggiori alle basse temperature);
- curiosamente è apparso un singolo picco che svetta a 60 lag (ossia 5 anni) nel grafico ACF sui residui al quadrato per entrambi i modelli;
- per quanto riguarda il Box test sui residui, per il modello ARIMA$(0,0,1)(3,1,1)_{12}$ si hanno p-value inferiori a 0.05 nei primi 15 lag (indicando quindi in quei casi correlazione fino a ciascun lag considerato), ed altri che lambiscono (oltre i 60 lag) la soglia 0.05 per i residui al quadrato.

Tra i due modelli, lo ARIMA$(2,0,2)(5,1,1)_{12}$ con drift è quello che esbisce meno criticità, sebbene il picco per l'ACF al lag 60 sui residui al quadrato lasci adito ad una situazione di eteroschedasticità condizionata.

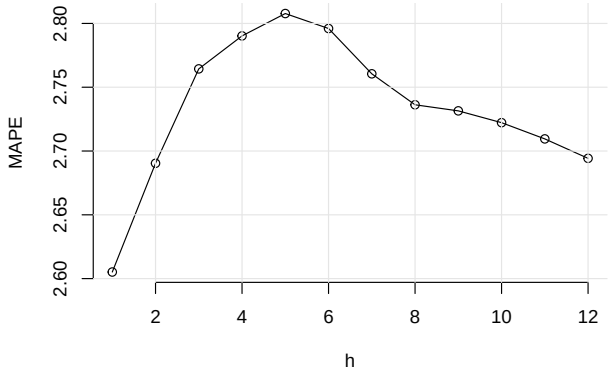
Visti tutti i tentativi fatti, non è chiaro se la serie delle temperature in esame possa essere adeguatamente modellata con un SARIMA.



(a) Root Mean Squared Error



(b) Mean Absolute Error



(c) Mean Absolute Percentage Error

Fig. 11. ARIMA$(2,0,2)(5,1,1)_{12}$ con drift: errori di predizione.

4.2 Previsione del modello

Continuiamo l'elaborato con il solo modello ARIMA$(2,0,2)(5,1,1)_{12}$ con drift descritto nelle precedenti sezioni.

Fig. 11 riporta le metriche RMSE, MAE e MAPE sugli errori di predizione per orizzonti

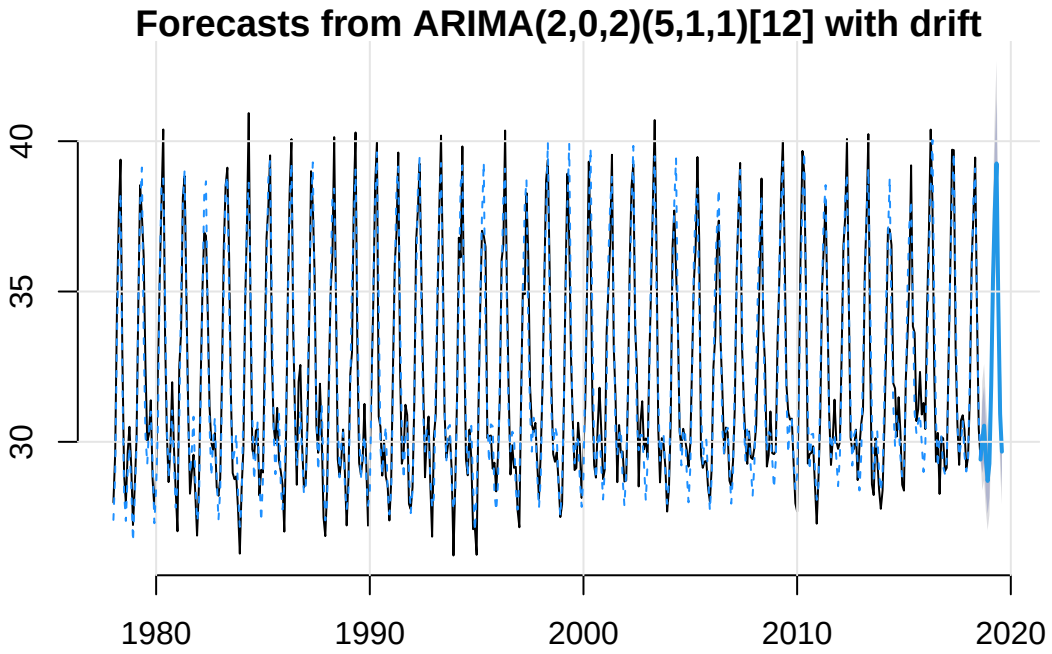


Fig. 12. Predizione fornita da ARIMA$(2,0,2)(5,1,1)_{12}$ con drift.

$h = 1 \dots 12$. Dagli andamenti di RMSE, MAE e MAPE è evidente che i modelli hanno senso solo su previsioni inferiori a 3-5 mesi.

Infine in Fig. 12 viene riportata la stima del modello (curva blu tratteggiata), compresa la previsione su 1 anno (curva blu piena) assieme ai margini di confidenza, confrontata con la serie originale, mostrando un buon accordo sebbene talvolta i valori del modello siano un po' più estremi di quelli presenti nella serie.