Relazione di serie storiche

D'alto Jacopo, Mattazzi Anna Chiara, Quercini Luca, Spinelli Sonia

Introduzione

Il dataset analizzato contiene le rilevazioni climatiche nella città di Nuova Delhi a partire dall'1 gennaio 2013 al 24 aprile 2017. Per ogni giorno sono registrati i valori della temperatura (gradi Celsius), dell'umidità (1 grammo di vapore acqueo per metro cubo), della velocità del vento (chilometri orari) e della pressione (ettoPascal).

	u.m.	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
`meantemp`	°C	6.00	18.86	27.71	25.50	31.31	38.71
`meanpressure`	hPa	-3.042	1001.580	1008.563	1011.105	1014.945	7679.333
`humidity`	g/mˆ3	13.43	50.38	62.62	60.77	72.22	100.00
`wind_speed`	km/h	0.000	3.475	6.222	6.802	9.238	42.220

Tabella 1: summary delle variabili

Una breve analisi descrittiva dei dati evidenzia la presenza di diversi outliers nei valori relativi alla pressione: il minimo risulta essere -3 hPa e il massimo 7679 hPa, rispetto a una media di 1011 hPa. Tali dati sono da considerarsi anomali: confrontando l'attendibilità dell'informazioni con una fonte esterna, si è giunti alla conclusione di escludere valori fuori dal range di 948-1083 hPa. Si sostituiscono, così, con la media.

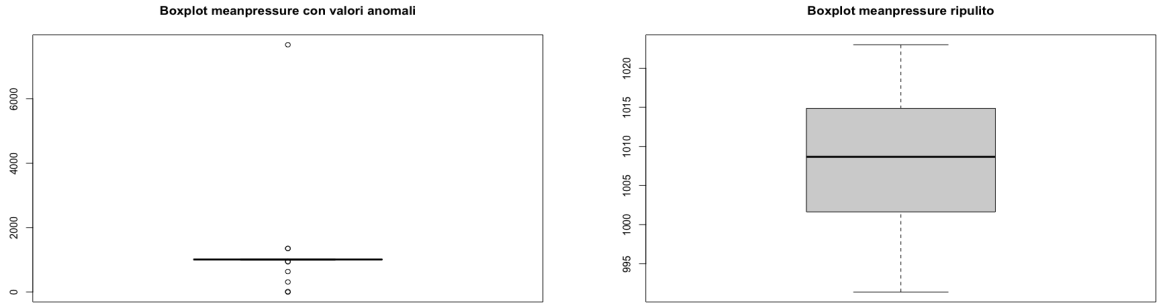

Figura 1: Boxplot di `meanpressure` prima e dopo aver ripulito i dati

Effettuando un'analisi delle correlazioni si evidenzia una forte correlazione negativa, pari a -0,87 , tra le variabili `meanpressure` e `meantemp`, mentre le altre correlazioni risultano essere poco significative.

	`meantemp`	`meanpressure`	`wind_speed`	`humidity`
`meantemp`	1.000	-0.876	0.306	-0.572
`meanpressure`	-0.876	1.000	-0.293	0.332
`wind_speed`	0.306	-0.293	1.000	-0.374
`humidity`	-0.572	0.332	-0.374	1.000

Tabella 2: matrice di correlazione tra le variabili

Graficamente tale relazione di dipendenza può essere rappresentata mediante un modello di regressione lineare:

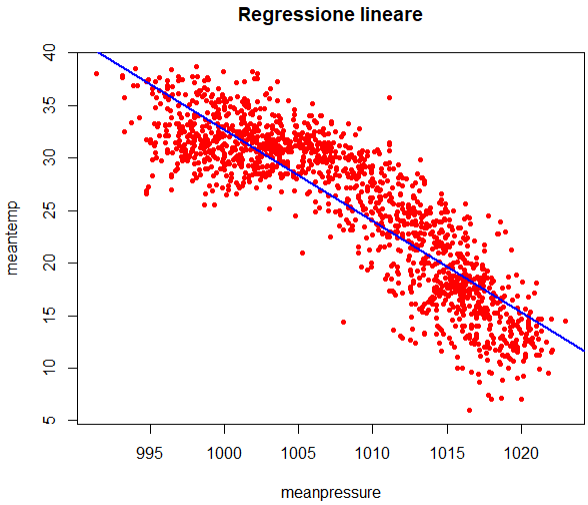

Figura 2: Regressione lineare

Osservando la *Figura 2*, si può affermare che la pressione atmosferica e la temperatura sono legate da una dipendenza lineare. Tale fenomeno è verificato dal punto di vista fisico: quando l'aria si riscalda, si espande e di conseguenza la densità dell'atmosfera diminuisce. Al contrario, quando l'aria si raffredda, la sua densità e la pressione atmosferica aumentano. La regressione lineare sui dati osservati è abbastanza buona con un indice R^2 pari a 0.77.

Analisi delle Serie Storiche

Alla luce dei risultati ottenuti dalle precedenti analisi si decide di considerare le variabili `meantemp` e `meanpressure` per l'analisi temporale.

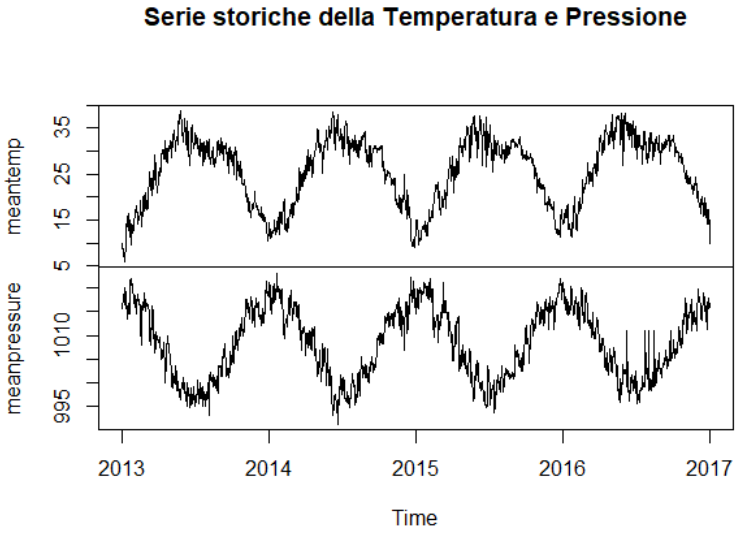

Figura 3: Plot delle serie storiche per le variabili `meanpressure` e `meantemp` con osservazioni giornaliere

Le due serie storiche mostrano un andamento stagionale annuale. Si osserva che i picchi positivi per `meantemp` si trovano in corrispondenza di quelli negativi per `meanpressure` e viceversa, aspetto che conferma il valore della correlazione tra le due variabili.

Trasformazioni:

Per semplificare la lettura dei grafici, si decide di ridurre il numero di osservazioni annuali: viene calcolata la media settimanale, in modo tale da avere 52 rilevazioni per ogni anno.

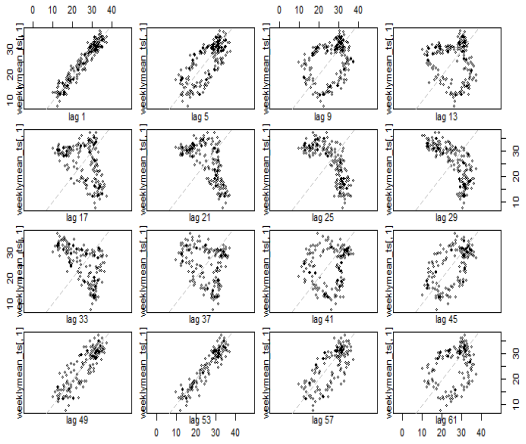

Figura 4.1: lag plot di `meantemp`

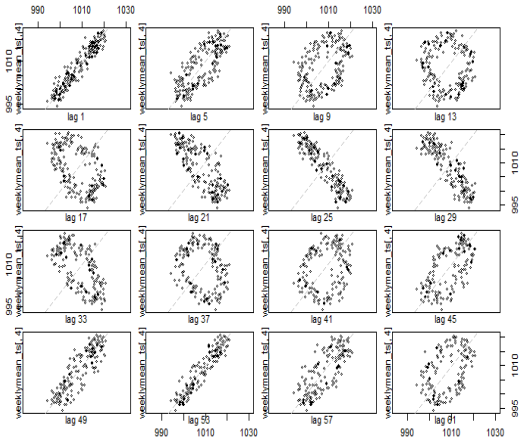

Figura 4.2: lag plot di `meanpressure`

Osservando i lag plot si conferma l'ipotesi di stagionalità annuale. Perciò, si calcolano per entrambe le differenze 52esime per eliminare la stagionalità annuale e successivamente le differenze prime per rimuovere il trend presente. Poichè i lag plot non presentano la tipica forma a imbuto, si può affermare la stazionarietà in varianza per entrambe le serie.

- Temperatura

Dalla decomposizione della serie storica della temperatura si evidenzia una stagionalità annuale, un trend positivo molto piccolo, in quanto il range di valori varia da 24.5°C a 27°C, e un errore che sembra aleatorio: è centrato in 0 e oscilla tra -6 e 6, assumendo valori piuttosto bassi.

Dalla *figura 5.2* si vede che le temperature non sono soggette a grandi cambiamenti da un anno all'altro: le linee risultano quasi sovrapponibili.

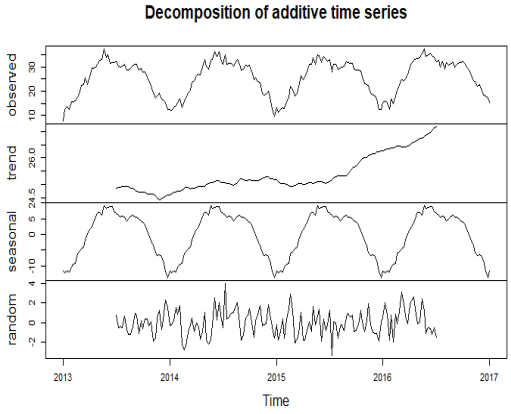

Figura 5.1: Decomposizione di `meantemp`

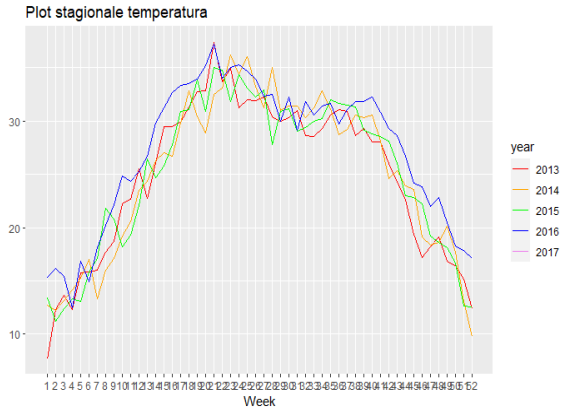

Figura 5.2: grafico stagionale di `meantemp`

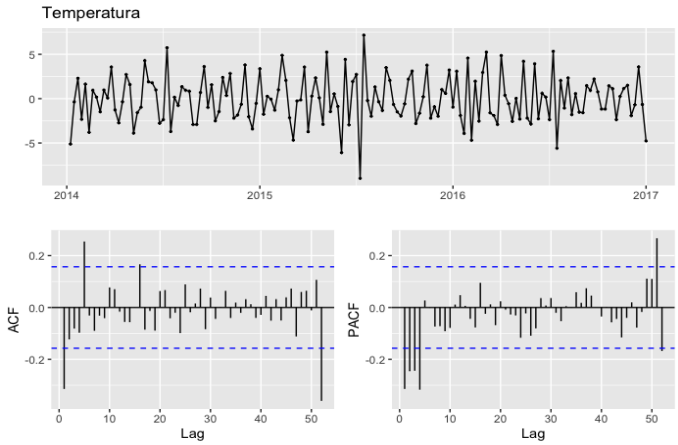

Figura 6: ACF e PACF di `meantemp`

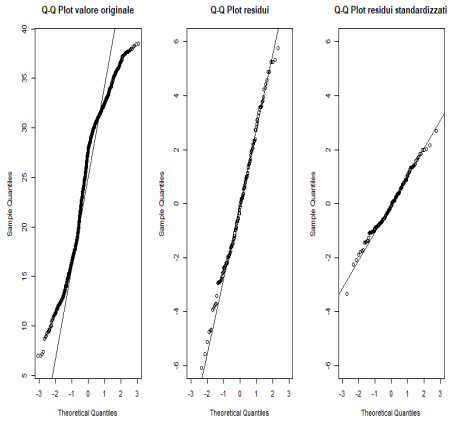

Figura 7: grafico stagionale di `meantemp`

Dopo aver trasformato la serie, si verifica l'aleatorietà dei residui. Il confronto con i quantili di una normale standard e i residui standardizzati in *figura 7* conferma l'assunzione.

Pertanto, si può formulare un ipotetico modello generativo dei dati a partire dai grafici in *figura 6*. La ciclicità annuale suggerisce un periodo pari a 52. Siccome sono state effettuate le differenze prime e quelle 52esime per lisciare la serie, sono scelti d=1 e D=1.

Dal grafico di ACF, si osserva che, in corrispondenza del lag 1, è presente una linea fuori dalle bande, dunque q=1. Per quanto riguarda la componente stagionale, il lag 52 esce fuori dal tratteggio, ovvero Q=1.

Dal grafico di PACF, si contano quattro picchi, che implicano p=4, mentre si deduce P=0, poichè il periodo non sembra avere un impatto così significativo.

Una volta definiti i parametri, viene applicato sui dati il modello scelto: SARIMA(4,1,1)(0,1,1)[52].

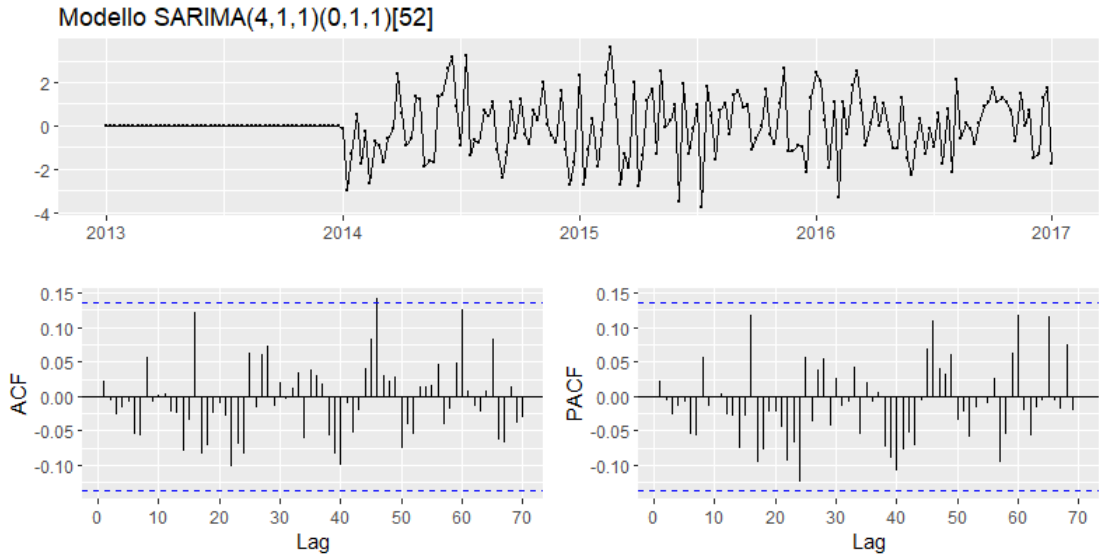

Figura 8: modello SARIMA(4,1,1)(0,1,1)[52] applicato su `meantemp`

Il grafico dei residui risultanti mostra un andamento aleatorio e i grafici di ACF e PACF non escono dalle bande di confidenza. Per il modello ipotizzato l'AIC è pari a 658.78.

D'altra parte, il modello proposto dal software è un SARIMA(3,1,1)(1,1,0)[52]. Tuttavia, i grafici di ACF e PACF mostrano linee uscenti dalle bande e l'AIC è pari a 672.15, più alto del modello precedente.

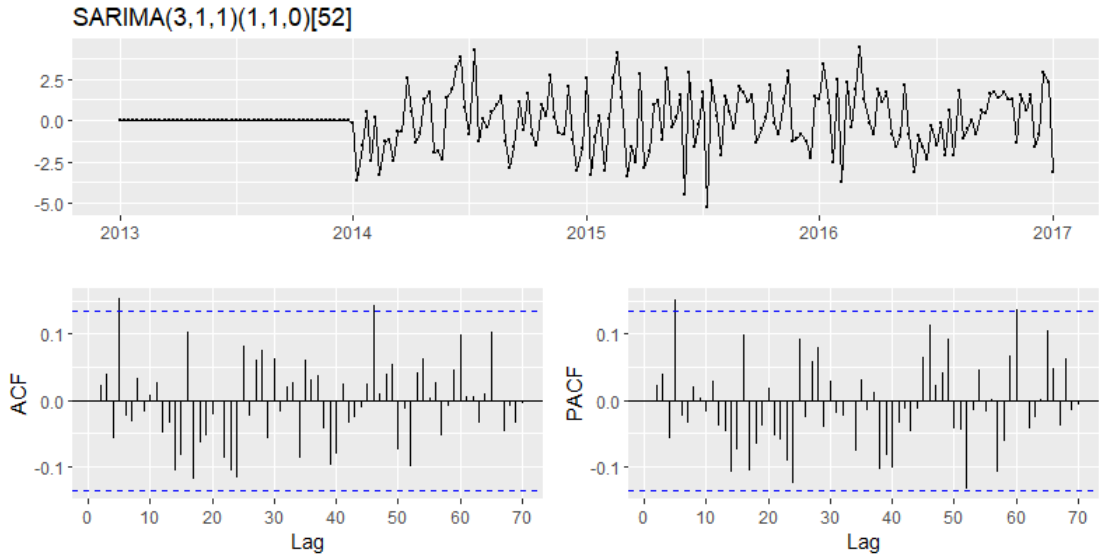

Figura 9: modello SARIMA(3,1,1)(1,1,0)[52] applicato su `meantemp`

Si opta per il primo modello ipotizzato e si fa una predizione sui dati relativi al 2017. La linea spessa rappresenta la stima puntuale del valore della variabile temperatura nell'istante di tempo, mentre le ombre intorno rappresentano gli intervalli di confidenza all'80%, colorata di rosso intenso, e al 95%, di rosso più trasparente. La predizione rappresentata nel grafico in *figura 10* rispetta la stagionalità e il trend crescente.

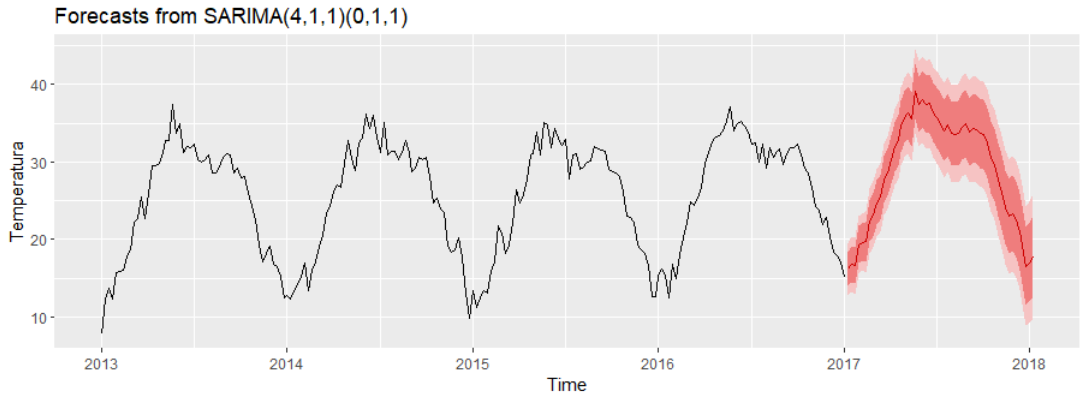

Figura 10: predizione 2017 rispetto al modello SARIMA(4,1,1)(0,1,1)[52]

- Pressione

Dalla decomposizione della serie storica della pressione risulta esserci una stagionalità annuale, come già evidenziato, e un trend leggermente positivo, con valori compresi nell'intervallo (1007.6 hPa, 1008.4 hPa). L'errore non sembra seguire un andamento particolare: è centrato in zero e oscilla tra -3 e 3. La serie presenta inoltre un andamento simile di anno in anno (*figura 11.2*).

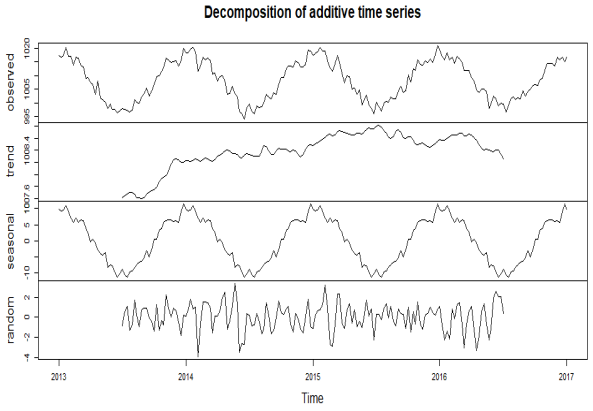

Figura 11.1: Decomposizione di `meanpressure`

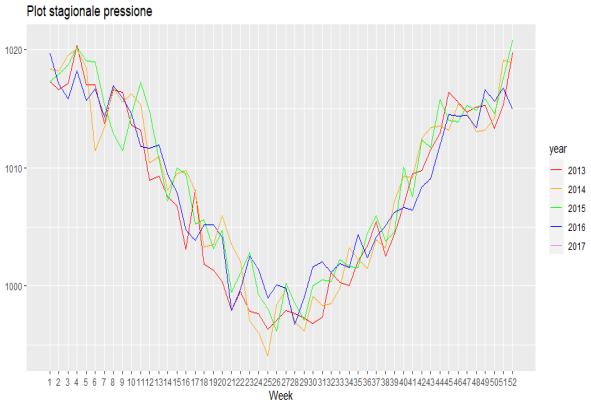

Figura 11.2: grafico stagionale di `meanpressure`

Dopo aver applicato le differenze 52esime per destagionalizzare e le differenze prime per detrendizzare, si studiano nei grafici in *figura 13* come si distribuiscono i residui della serie trasformata confrontati con i quantili di una normale standard. Essi sembrano discostarsi leggermente nelle code.

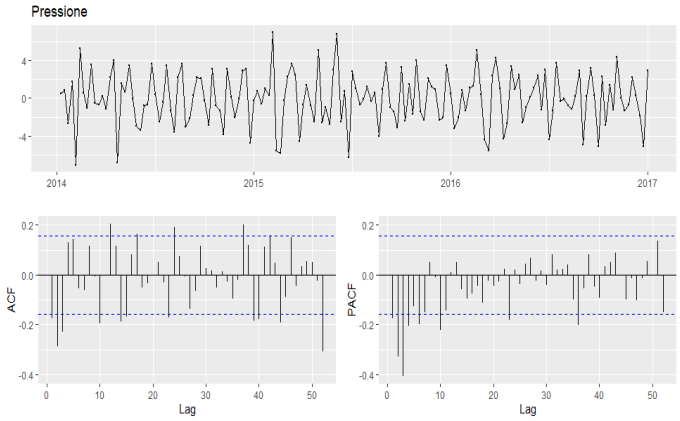

Figura 12: ACF e PACF di `meanpressure`

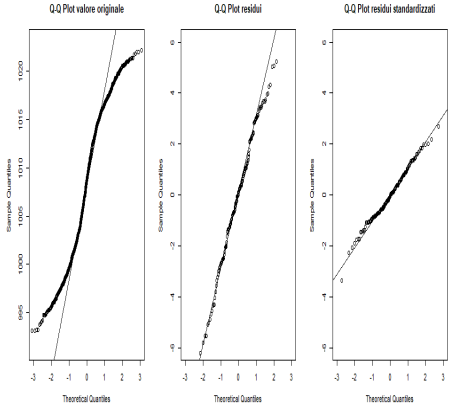

Figura 13: grafico stagionale di `meanpressure`

Si stimano i parametri del modello generatore dei dati, analizzando i grafici di ACF e PACF in *figura 12*. Nel grafico di autocorrelazione si nota un picco in corrispondenza del lag 52, suggerendo che il parametro Q sia 1. Inoltre, i primi tre lag escono dalle bande, il che suggerirebbe un q=3, ma per un problema del software si preferisce q=2.

Dal grafico di autocorrelazione parziale non si notano picchi intorno al periodo, mentre il primo lag interno alle bande è il quinto, quindi p = 5 e P = 0. Infine, analogamente a prima i parametri d e D sono entrambi 1.

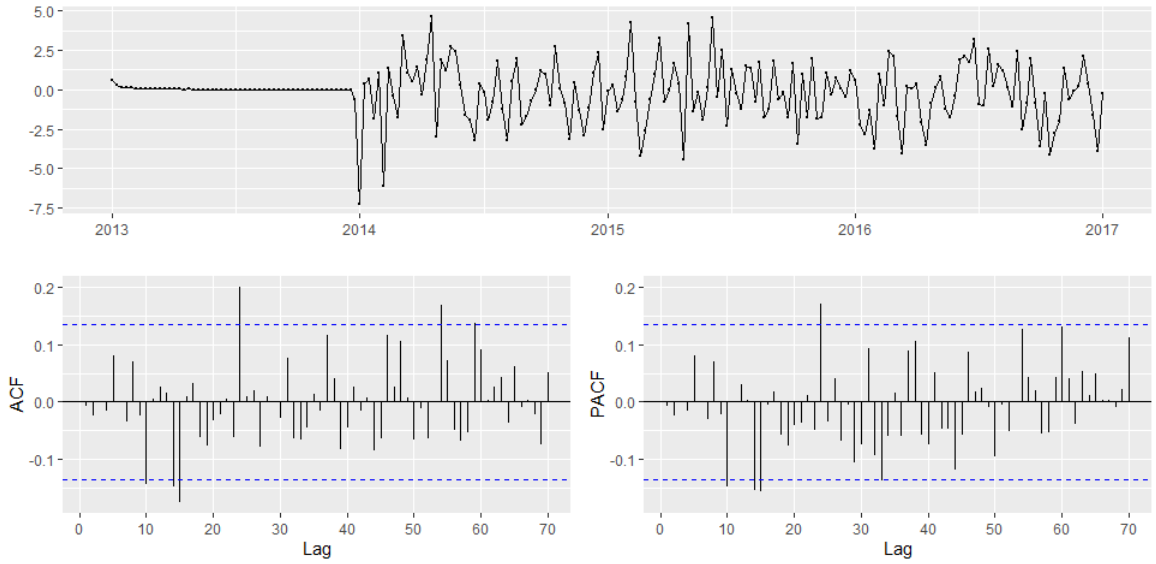

Figura 14: modello SARIMA(5,1,2)(1,1,0)[52] applicato su `meanpressure`

Si applica ai dati il modello scelto, il quale presenta un AIC di 683.25. Dall'analisi dei residui si può suppore che essi non siano puro rumore bianco, per via dei grafici di ACF e PACF, che presentano diversi picchi fuori dalle bande di confidenza.

Si considera il modello fornito dal software, che è un SARIMA(4,1,1)(1,1,0)[52], con un AIC leggermente più basso, di 679.48. I residui risultano correlati tra loro. Dato che il coefficiente AIC è più basso, si esegue la predizione con il modello suggerito dal software.

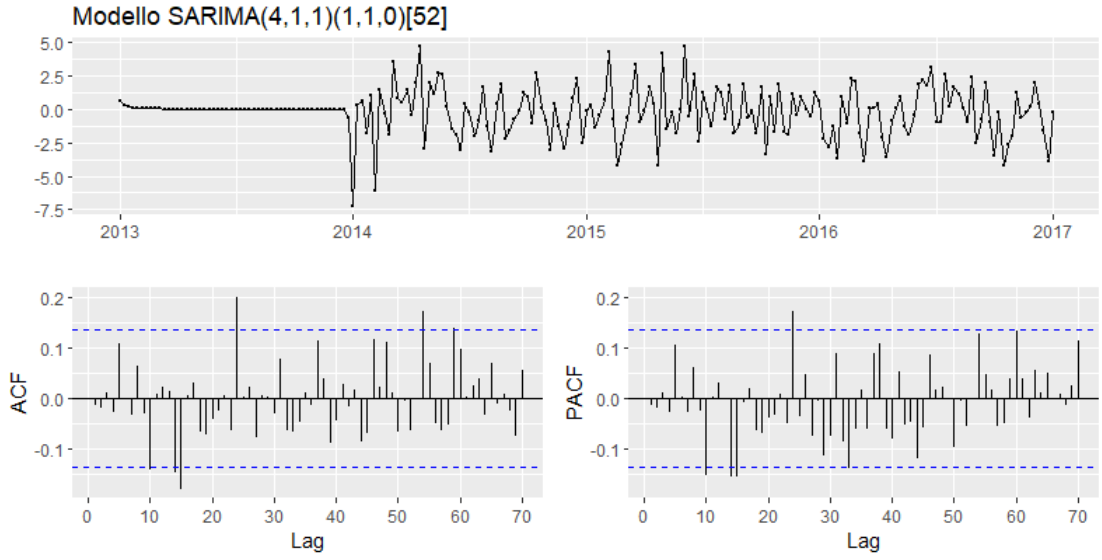

Figura 15: modello SARIMA(4,1,1)(1,1,0)[52] applicato su `meanpressure`

Grazie all'analisi dei dati nel tempo, si deduce l'andamento della serie dal 2017 al 2018. La predizione, che viene svolta sulla base della stagionalità e la ciclicità delle osservazioni precedenti, sembrerebbe essere una stima verosimile per il futuro.

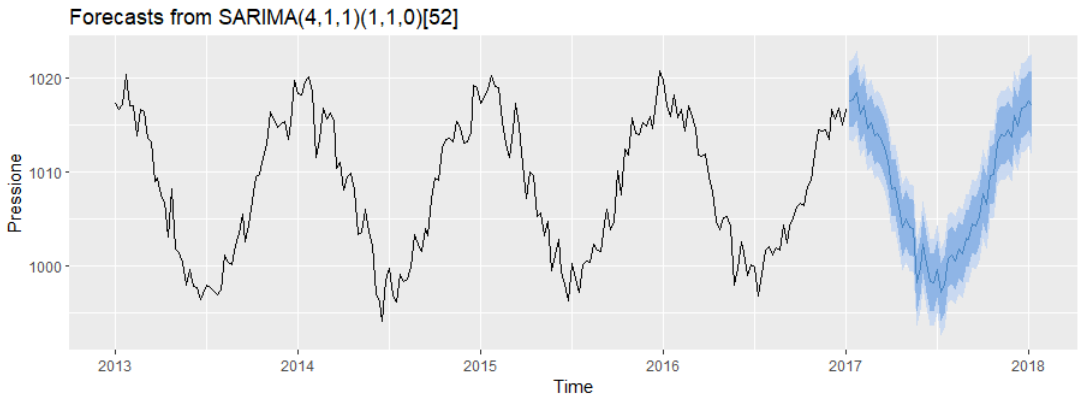

Figura 16: predizione 2017 rispetto al modello SARIMA(4,1,1)(1,1,0)[52]

Autocorrelazioni incrociate

Dal grafico di autocorrelazione incrociata tra le due variabili in *figura 17*, si nota che in corrispondenza del lag 0 la temperatura e la pressione sono correlate negativamente. All'aumentare del lag diventano sempre meno correlate fino ad evidenziare una correlazione positiva a lag 26, per poi tornare negativa in corrispondenza della 52esima settimana. All'aumentare degli anni, la correlazione oscilla con minor ampiezza.

Si deduce che le due serie hanno un andamento speculare l'una rispetto all'altra, fenomeno che è spiegato dal punto di vista fisico dal rapporto tra la temperatura e la pressione. Si può concludere che essere a conoscenza dell'evoluzione di una delle due serie implica poter stimare puntualmente anche quella dell'altra.

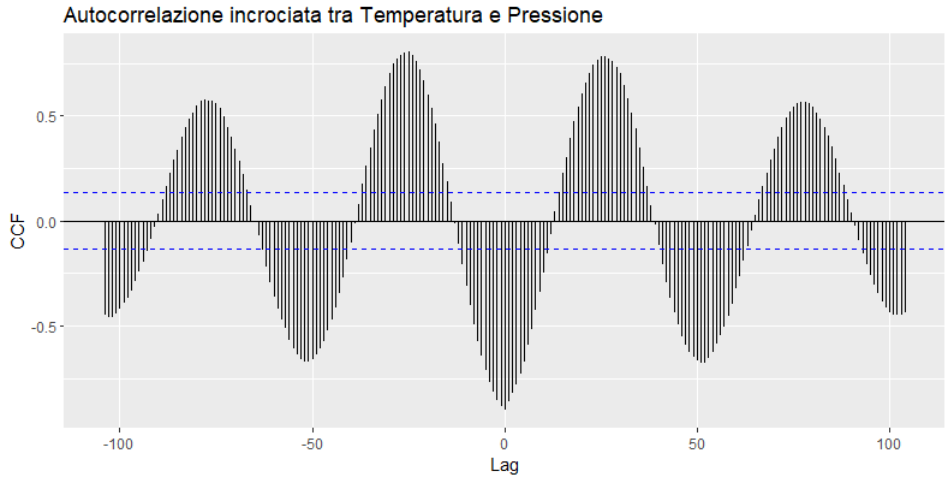

Figura 17: Plot dell'autocorrelazione incrociata tra `meantemp` e `meanpressure`

Fonti:

- https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data
- https://weatherspark.com/h/y/109174/2016/Historical-Weather-during-2016-in-New-Delhi-India#Figures-Pressure