DAILY CLIMATE DELHI (INDIA) – Time Series Analysis

Davide Mascolo, 01/06/2021

1. Abstract

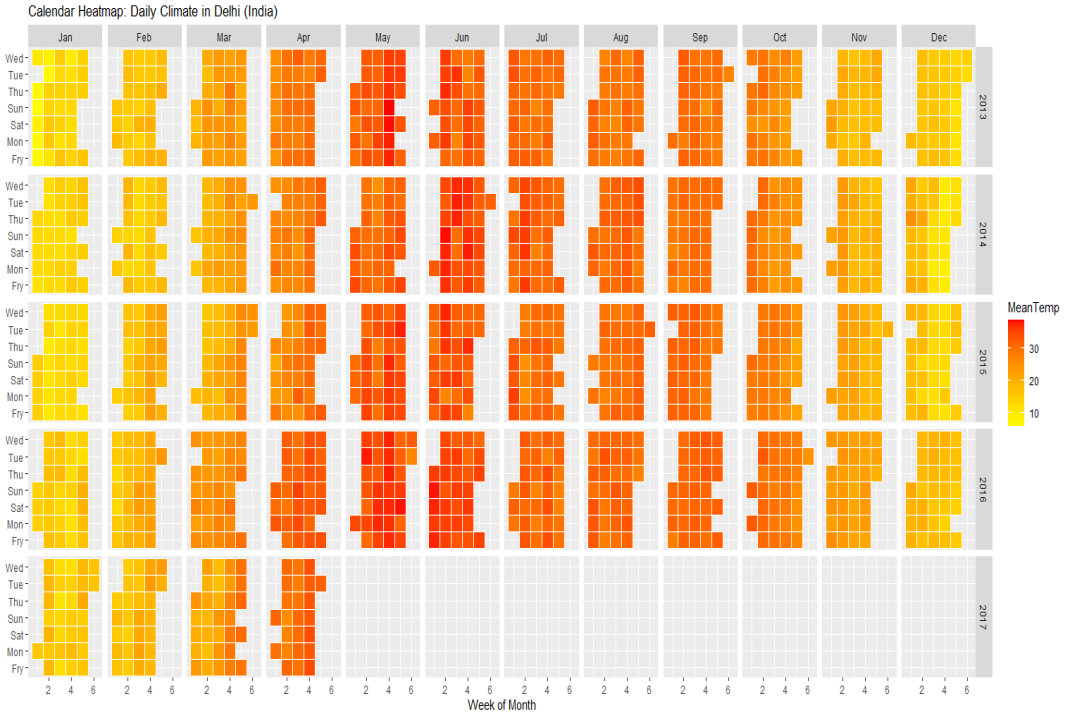
- In questo studio si vuole analizzare la serie storica relativa alle *temperature* registrate nella città di *Delhi* (India).
- I dati presi in considerazione vanno dal *1 gennaio 2013* al *24 aprile 2017*.

2. Data Wrangling

- In questa fase preliminare di lavorazione del dato, si parte da una serie storica splittata in *train* e *test*; di conseguenza, il primo passo è stato *aggregare* i dati per osservazione, facendo attenzione all'ordine delle osservazioni stesse.
- Non sono presenti *valori mancanti*.
- Successivamente, si è passati a *formattare* adeguatamente alcune variabili come la data, per poi concentrarsi sulla creazione di nuove variabili che saranno necessarie per l'analisi esplorativa.

3. Analisi Descrittiva

- Il *grafico1* proposto è un *Calendar Heatmap*.

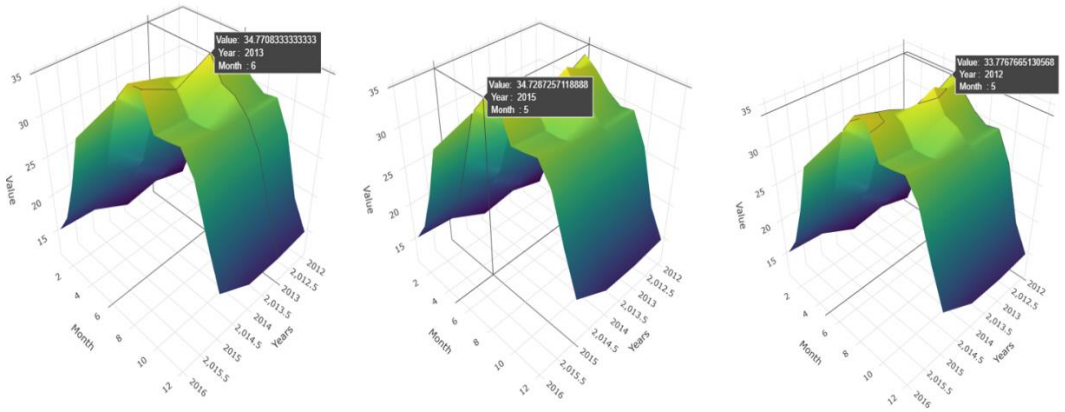


- Una mappa di calore rappresenta i valori per una variabile di interesse principale su due variabili dell'asse, come una griglia di quadrati colorati. Le variabili degli assi sono suddivise in intervalli come un grafico a barre o un istogramma ed il colore di ciascuna cella indica il valore della variabile principale nell'intervallo di celle corrispondente.

- Le mappe termiche del calendario sono utili quando i valori sono giornalieri o settimanali. Sono molto utili, in casi come questi, quando si vuole analizzare il valore giornaliero per l'intera serie storica.
- Tuttavia, sono meno utili se si vogliono vedere componenti come la stagionalità, la forma della serie, la stazionarietà e così via. La mappa termica in questione, mostra la distribuzione giornaliera della temperatura media nella città di Delhi, in India, raggruppata per mese e registrata dal gennaio 2014 al 24 aprile 2017.
- Ogni cella riporta un conteggio numerico, come in una tabella dati standard, ma il conteggio è accompagnato da un colore, con conteggi maggiori associati a colorazioni più scure. (La colorazione della cella è personalizzabile)
- Dai colori più scuri si vede come le temperature medie sono più elevate nei mesi estivi, come ci si aspetta e vanno al di sotto dei venti gradi nei mesi di Dicembre, Gennaio e Febbraio.
  Tuttavia, essendo l'India un paese molto caldo, si può notare come già dalla fine di Marzo le temperature, mediamente, iniziano a salire in maniera importante, fino ad arrivare nei mesi di Maggio, Giugno e nelle prime settimane di Luglio a superare mediamente i 30 gradi.
  Inoltre, i mesi di Luglio ed Agosto sono anche quelli in cui si registrano le maggiori precipitazioni.

Surface Plot

- Il *grafico2* racchiude una serie di rappresentazioni in più dimensioni che hanno lo scopo di cogliere i picchi della serie.



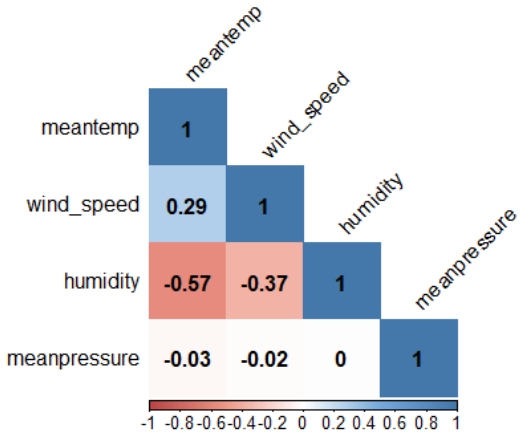
- Le temperature più alte, in tutta la serie, vengono registrate nel mese di *giugno 2013* con *34.77 gradi*. Altro picco si ha nel mese di *maggio 2015* con *34.72 gradi*. Infine, ultimo picco rilevante si ha nel mese di *maggio 2012* con *33.77 gradi*. Questa visualizzazione, in aggiunta alla mappa di calore, ci permette non solo di visualizzare in maniera interattiva i dati, ma anche di vedere effettivamente il valore numerico associato alla temperatura e non la semplice colorazione. Lo svantaggio è che non si ha il giorno e la settimana precisa della rilevazione, cosa che invece si ha nella mappa di calore.

Correlazione

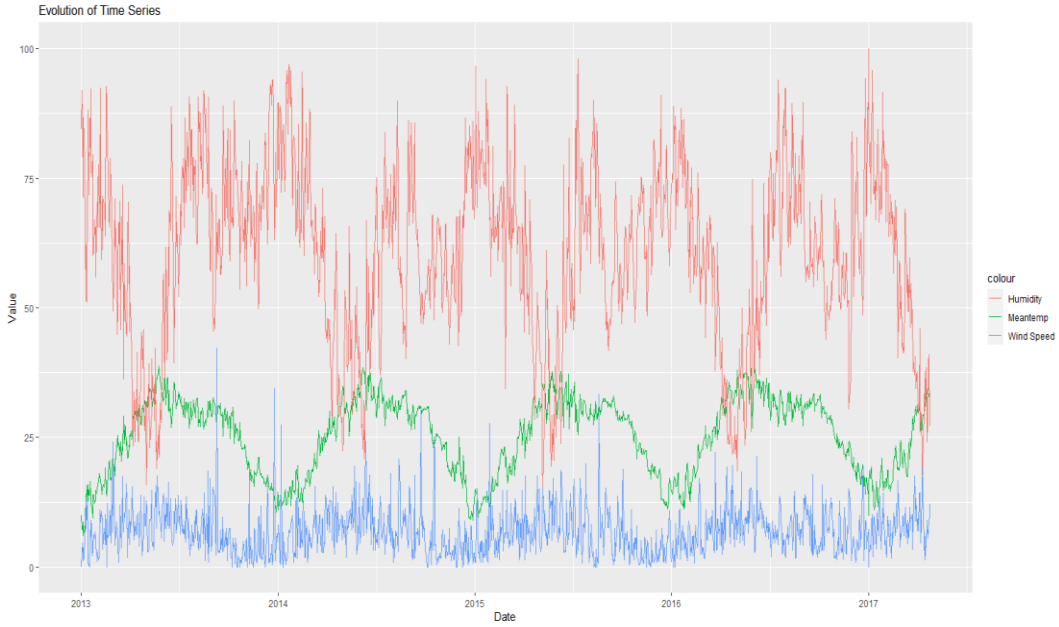
- Dallo studio della *correlazione* tra le variabili, è emerso ciò.

	MeanTemp	***WindSpeed***	***Humidity***	***MeanPressure***
MeanTemp	*1.000*			
WindSpeed	*0.288*	*1.000*		
Humidity	*- 0.575*	*- 0.374*	*1.000*	
MeanPressure	*- 0.035*	*- 0.017*	*- 0.002*	*1.000*

- Assodato il concetto teorico di correlazione, vediamo come le variabili siano tra di loro debolmente correlate. L'unica relazione lineare degna di nota è tra *MeanTemp* ed *Humidity*, pari a *-0.578*. Significa che se l'umidità aumenta di una singola unità, in media, ci aspettiamo che la temperatura media cali di *0.58* gradi circa.



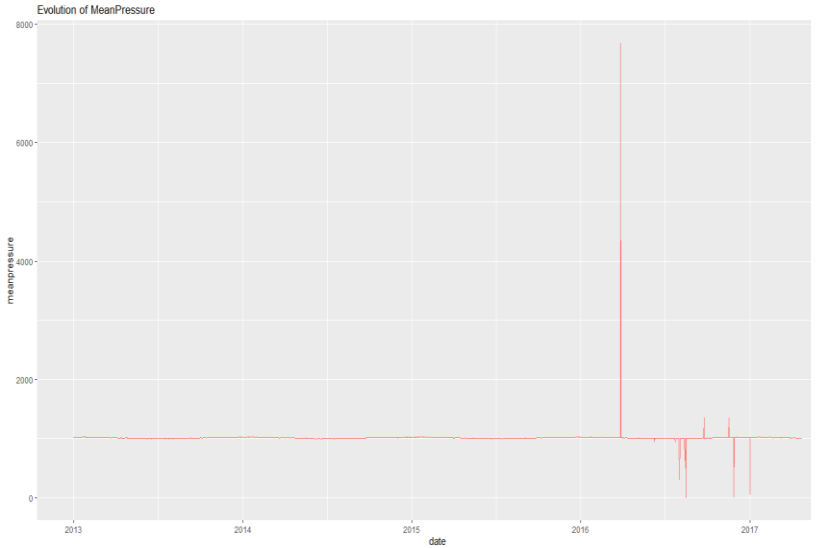
Andamento della serie



- Il *grafico3* mostra l'andamento dell'umidità, temperatura media e velocità del vento in relazione al tempo.
  Si nota come mediamente i valori associati all'umidità siano sempre maggiori rispetto a quelli della temperatura media. Risulta interessante osservare come la correlazione notata prima abbia un significato anche grafico. Infatti, per valori alti dell'umidità la temperatura media tende ad assumere valori bassi. Man mano che l'umidità cala la temperatura subisce un rialzo, fino ad arrivare in giorni in cui i valori di temperatura media ed umidità si "*incontrano*" e ciò avviene quando la temperatura raggiunge il picco mentre l'umidità raggiunge il valore minimo. Graficamente vediamo quest'*intersezione* tra le

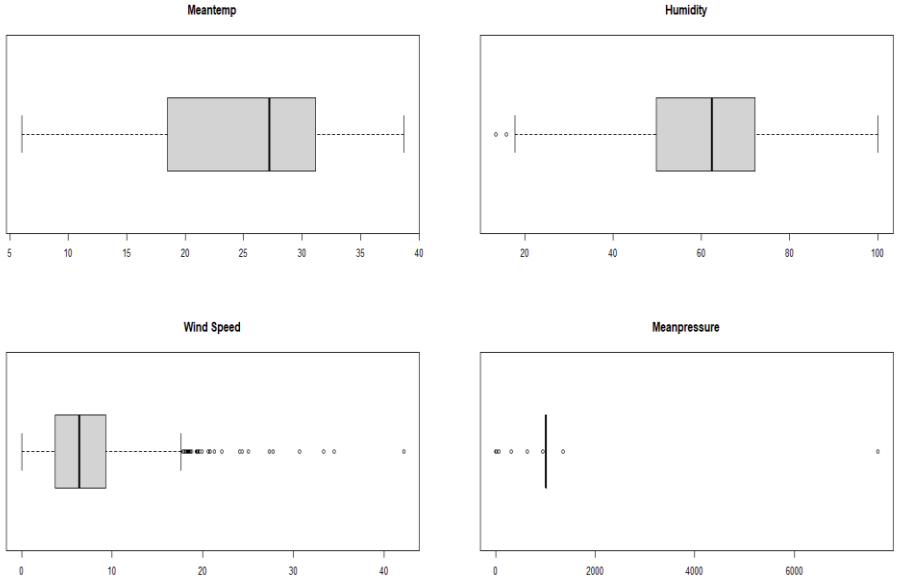
due curve in corrispondenza di ogni *picco* per la *temperatura* ed ogni *picco(negativo*) per *l'umidità*.

grafico4



- Da questa visualizzazione, invece, studiamo l'evoluzione della *pressione media* negli anni. I valori, per tutta la serie osservata, sono abbastanza *discretizzati* e non subiscono notevoli variazioni. Tuttavia, qualcosa di anomalo accade nella *prima metà del 2016* con un *incremento* importante che poi ha delle ripercussioni nella *seconda metà* dello stesso anno, causando variazioni al rialzo ed al ribasso. Data la correlazione molto bassa con la variabile di riferimento, però, è da escludere che ciò abbia avuto degli effetti sulla temperatura media e lo si evince anche dal *grafico1*. Infatti, confrontando il periodo *2016-2017(periodo della variazione legata alla pressione)* con gli altri anni, la temperatura media dimostra avere sempre lo stesso comportamento.

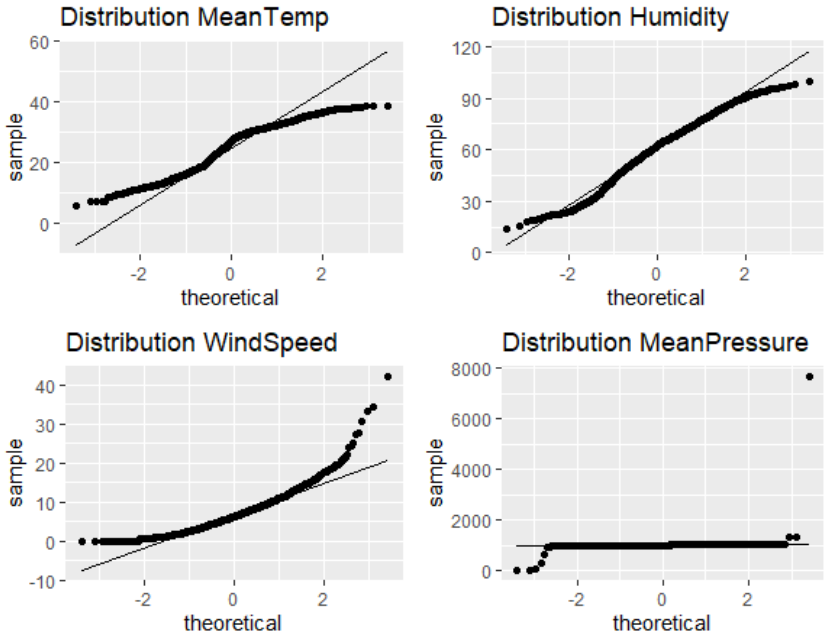
Distribuzione



- I boxplot del *grafico5* hanno come obiettivo quello di analizzare la *distribuzione* di ogni singola variabile.
- ***MeanTemp:*** guardando al range interquartile, vediamo come la varianza sia abbastanza grande. Se l'IQR è piccolo, significa che la metà delle osservazioni si trova fortemente *concentrata* intorno alla mediana; all'aumentare della distanza interquartilica aumenta la *dispersione* del 50% delle osservazioni centrali intorno alla *mediana*.
  Il box di sinistra è più lungo rispetto al box di destra è ciò indica un'*asimmetria negativa*. Le code della distribuzione sono abbastanza lunghe e lo si vede dai "*baffi*".

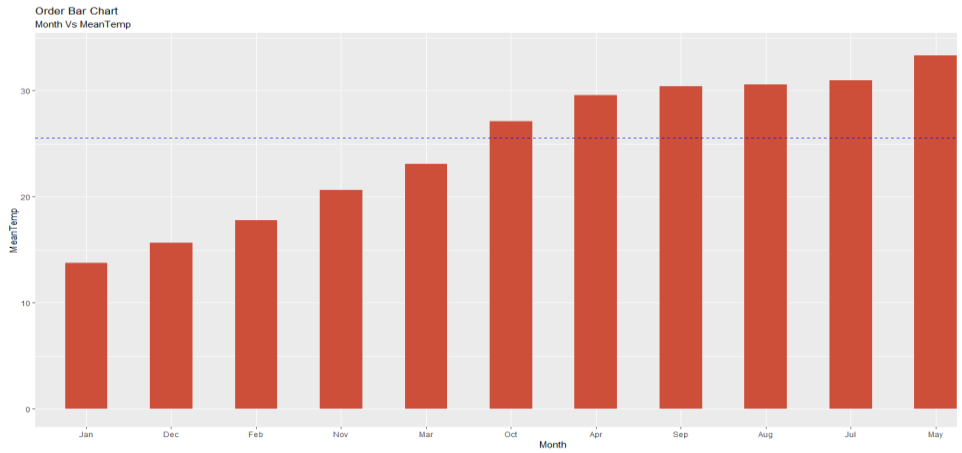
- ***Humidity***: distribuzione leggermente *asimmetrica negativa*. Varianza minore rispetto alla variabile *MeanTemp* e presenza di due valori eccezionalmente piccoli.
- ***WindSpeed***: distribuzione pressochè *simmetrica* con una coda destra molto lunga.
- ***MeanPressure***: distribuzione fortemente *discretizzata*

- Nel *grafico6*(che segue), si fa un approfondimento sulle distribuzioni delle variabili, andandole a confrontare con i quantili teorici della distribuzione Gaussiana. Questa visualizzazione permette anche di discutere in maniera più approfondita il *comportamento delle code* di ogni distribuzione, valutandone la *pesantezza* e la *lunghezza*.



- La distribuzione di ***MeanTemp*** risulta essere *platicurtica*, infatti per la coda di sinistra i valori sono *meno estremi* della Normale(quantili MeanTemp > quantili Normale) ed anche per la coda di destra i valori sono *meno estremi* della Normale(quantili MeanTemp < quantili Normale). Essendo platicurtica, la distribuzione della variabile di riferimento avrà *code meno spesse/pesanti/lunghe*, e quindi meno dati, rispetto alla Normale.
- Discorso simile per la variabile ***Humidity***, anche se in questo caso il comportamento devia da quello della distribuzione Normale *solo nelle code*, mentre nella zona centrale è pressochè identico. Tuttavia, nonostante nelle code si osservi qualche deviazione questa è meno marcata rispetto alla variabile studiata in precedenza.
- Per la variabile ***WindSpeed***, come detto anche dal boxplot, abbiamo una distribuzione abbastanza Normale nella zona *centrale* dei dati, ma spostandoci nelle *code non* viene mantenuto questo presupposto. Interessante notare come la *coda di destra* sia molto *più pesante* della Normale e ciò lo si poteva intuire anche dal *grafico5*.

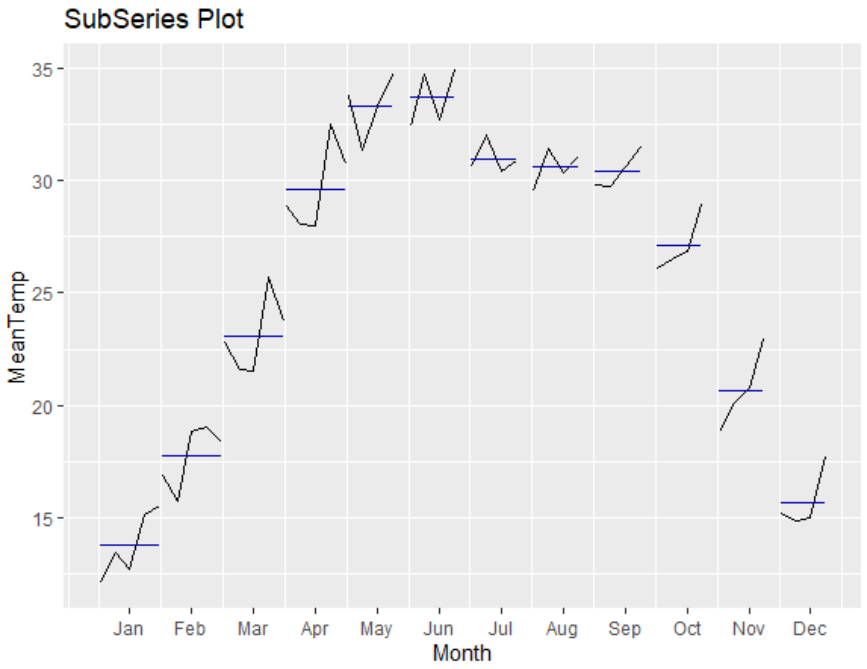
Ranking



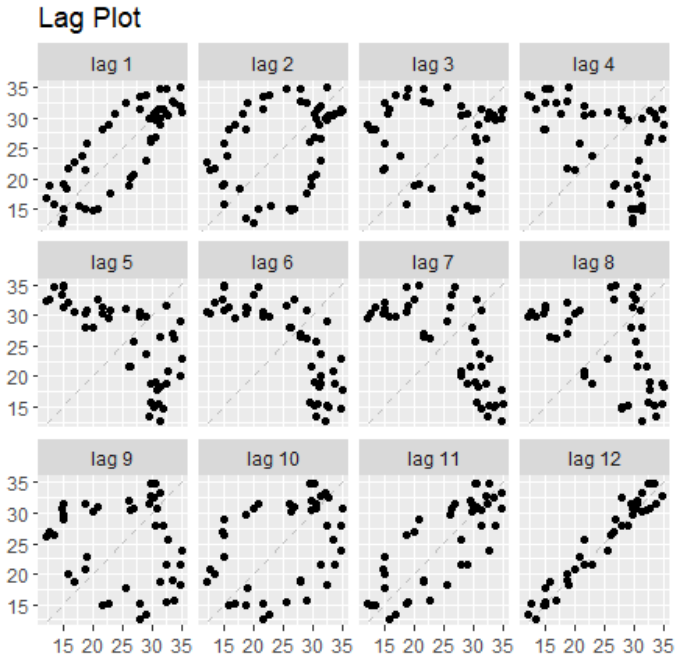
- Dal *grafico7* si ottiene un *ordinamento* dei mesi sulla base della *temperatura media*, con la linea orizzontale(blu) che rappresenta la media di tutti i mesi. Si ha una conferma di quanto detto al *grafico1* ovvero che i mesi invernali hanno breve durata rispetto a quelli estivi e che mediamente *sette* mesi su *dodici* fanno registrare temperature medie al di sopra dei *26 gradi*. Inoltre, anche per i mesi che si trovano al di sotto della media, solo *Dicembre*, *Gennaio* e *Febbraio* fanno registrare temperature medie al di *sotto dei 20 gradi*; già a partire dal mese di *Novembre* e successivamente *Marzo* le temperature medie *superano i 20 gradi*.

Forza del Trend e della Stagionalità

- Per iniziare a studiare le componenti della serie storica osservata, ci si concentra in questa fase preliminare sul trend e sulla stagionalità andandone a misurare la loro forza per capire quanto impatto abbiano effettivamente nella serie.
- **Fs = 0,982** e **Ft = 0,446.**
- Da questi risultati emerge che la componente stagionale nella serie storica osservata incide tantissimo, quasi il massimo ricordando che l'indice varia nell'intervallo[0;1]; mentre il trend incide in maniera importante ma sicuramente meno significativa.
- Si può notare, dal *grafico8*, come sia effettivamente marcata la presenza della componente stagionale.

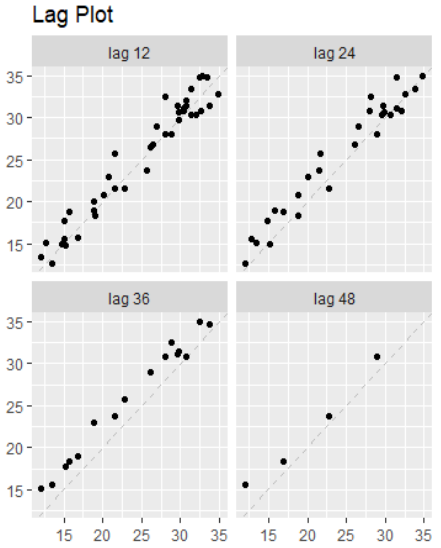


Analisi della stagionalità tramite i ritardi



- Il *grafico9* fornisce una dispersione bivariata, per ogni livello di ritardo(1-12lags). Una relazione si nota al ritardo 12 che presenta un forte legame lineare positivo con la serie. Per questo motivo si può approfondire sul ritardo della serie.

grafico10

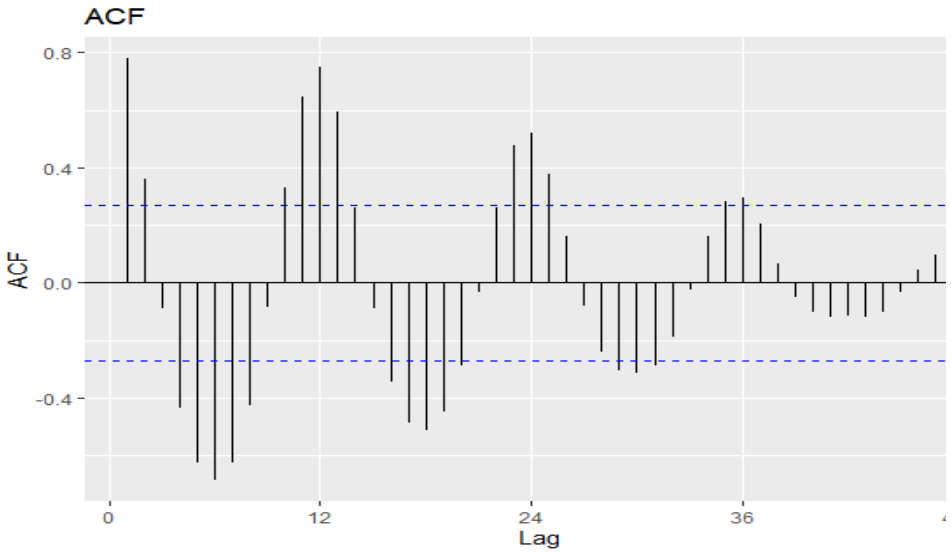


- Anche al ritardo *24* e *36* si hanno delle relazioni lineari e positive molto forti. Per il ritardo *48* le osservazioni sono poche perché ricordiamo che la serie termina nel mese di *Aprile 2017*.

Valori ACF

	ACF
LAG 12	0.748
LAG 24	0.519
LAG 36	0.296
LAG 48	0.083

grafico11

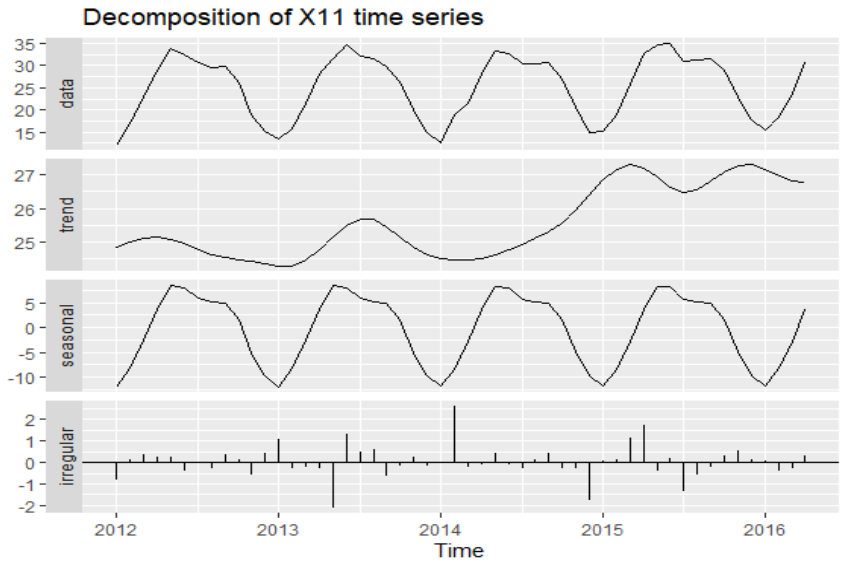


- Anche da questa visualizzazione, si nota come l'autocorrelazione al ritardo *12* è più alta rispetto alle autocorrelazioni degli altri ritardi. Ciò è dovuto all'andamento stagionale dei dati e quindi i picchi tendono ad essere a distanza *12* tra di loro.
  Anche i minimi sono a distanza *12* tra loro e data la componente stagionale, vengono osservati proprio a partire dal suddetto ritardo.
- Inoltre, i valori delle autocorrelazioni tendono velocemente a zero, il che fa escludere la presenza di un trend, almeno marcato(come detto in precedenza), all'interno della serie.
- Infine, i dati non sono stazionari e ciò lo si vede sia dal ritardo uno che è il più grande e positivo, ma anche per la presenza di stagionalità che caratterizza la serie.

Decomposizione

- Si effettua una decomposizione per capire non solo il comportamento della serie, ma anche quello delle sue componenti.
- In questo caso, si utilizzerà solamente la decomposizione ***X11***.

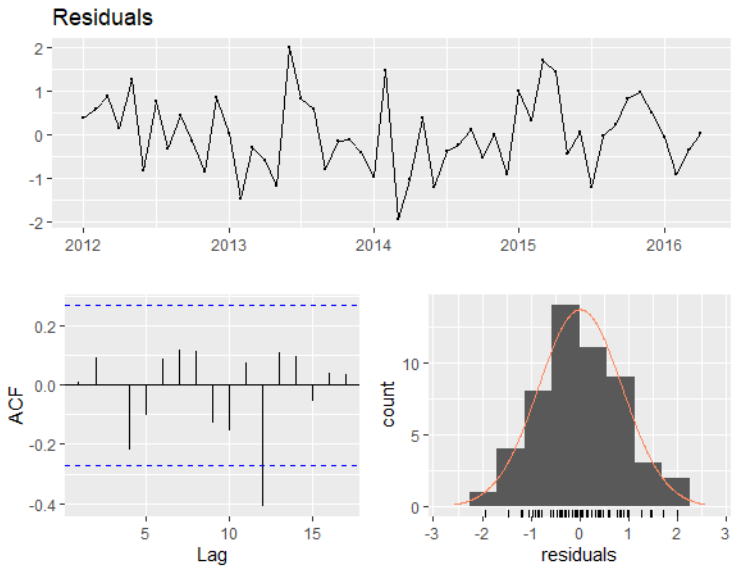
grafico12



Considerazioni:

1. **Trend**: In media il trend è crescente, ma subisce dei cambi di livello nel tempo. Si notano, infatti, periodi di grande crescita come nella *prima parte del 2013 e dal 2014 alla prima metà del 2015*, ma anche periodi al ribasso come il periodo *2012-2013*.
2. **Stagionalità**: Sembra essere della stessa intensità e struttura nel tempo e spiega gran parte della varianza della serie.

Analisi dei Residui



- I residui sembrano avere lo stesso range di variazione intorno allo zero e non presentano una struttura ben precisa.
- I valori delle autocorrelazioni sono abbastanza bassi, ma tuttavia il test di Shapiro-Wilks porta al rifiuto dell'ipotesi Nulla, quindi i residui non hanno distribuzione Normale.

Specificazione del Modello e Previsioni

- Nonostante la decomposizione abbia colto la significatività della componente stagionale, è possibile migliorare ancora la performance andando a specificare dei modelli.

- La prima specificazione presa in esame è una *regressione lineare con stagionalità e trend fisso*.

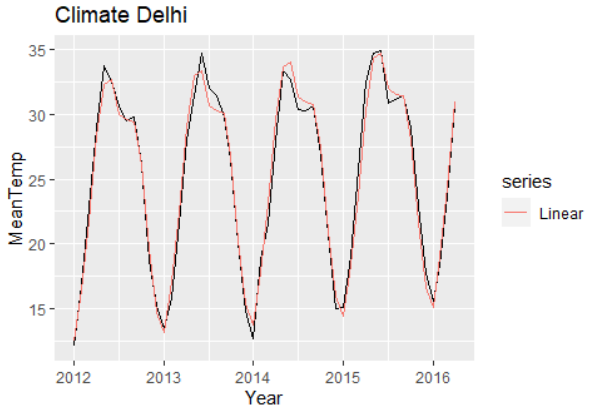
$$Meantemp \sim \beta o + \beta 1 * Season + \beta 2 * Trend + \varepsilon i$$

- La seconda specificazione è una regressione lineare con stagionalità fissa e l'aggiunta dei predittori *humidity* e *windspeed*.

$$Meantemp \sim \beta o + \beta 1 * Season + \beta 2 * Humidity + \beta 3 * WindSpeed + \varepsilon i$$

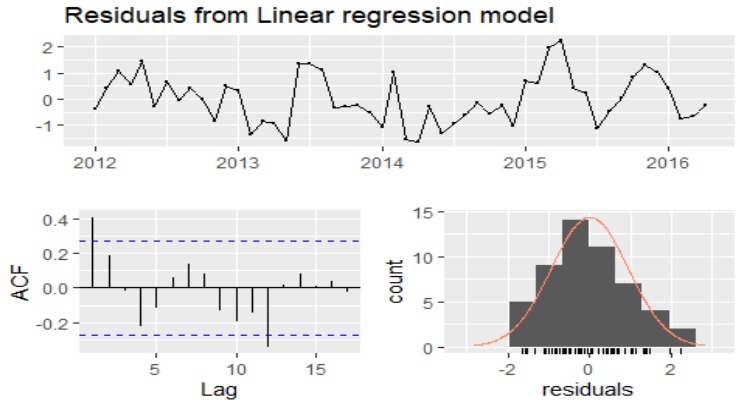
- La terza ed ultima specificazione è una regressione Cubic Spline con il predittore *humidity* e tenendo fisse stagionalità e trend. Per la scelta dei nodi, sono stati calcolati su Q1, Q2 e Q3 della distribuzione di *humidity*.

Prima Specificazione



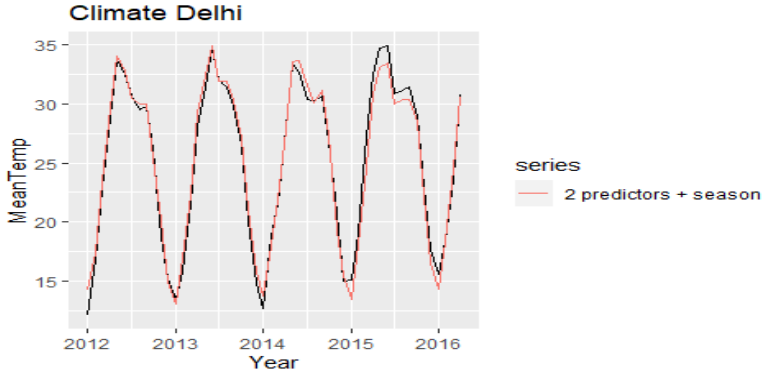
- Per un livello di alpha pari al 5%, si rifiuta l'ipotesi nulla che i parametri associati alla stagionalità ed al trend siano pari a zero.

Analisi dei Residui



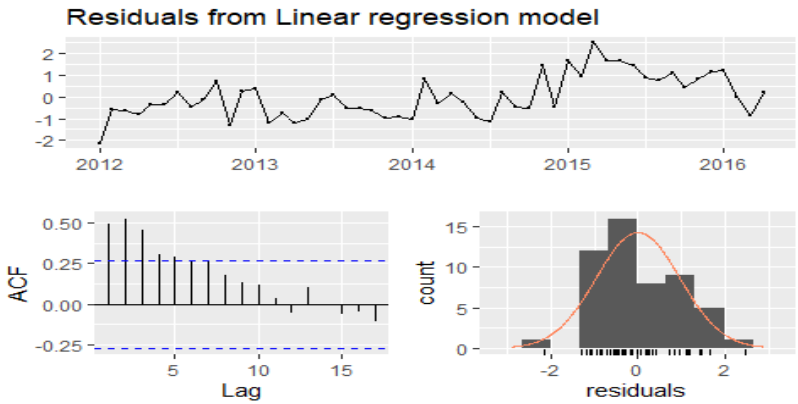
- I valori delle autocorrelazioni non superano mai la soglia, ad esclusione del valore registrato al ritardo *12*.
- Il test di Breusch-Godfrey suggerisce che non si può rifiutare l'ipotesi nulla, quindi vi è *autocorrelazione tra i residui*.

Seconda Specificazione



- Anche per questo modello, per un livello di alpha pari al 5%, si rifiuta l'ipotesi nulla che i parametri associati alle variabili ed alla stagionalità siano pari a zero.

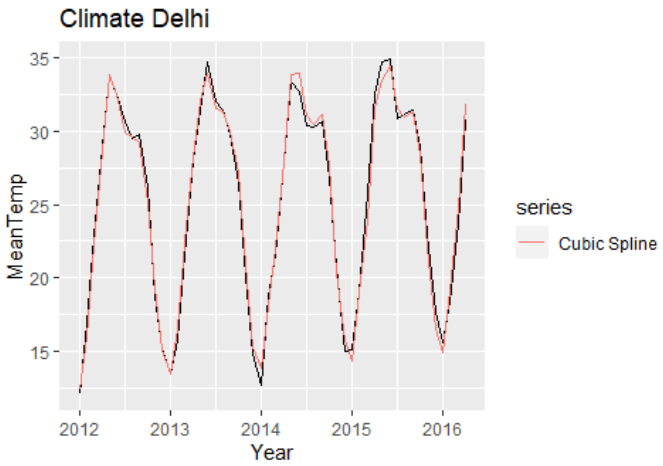
Analisi dei Residui



- I valori delle autocorrelazioni superano più frequentemente la soglia rispetto a prima e mostrano un *trend*; ciò sottolinea l'importanza della suddetta componente nella specificazione.

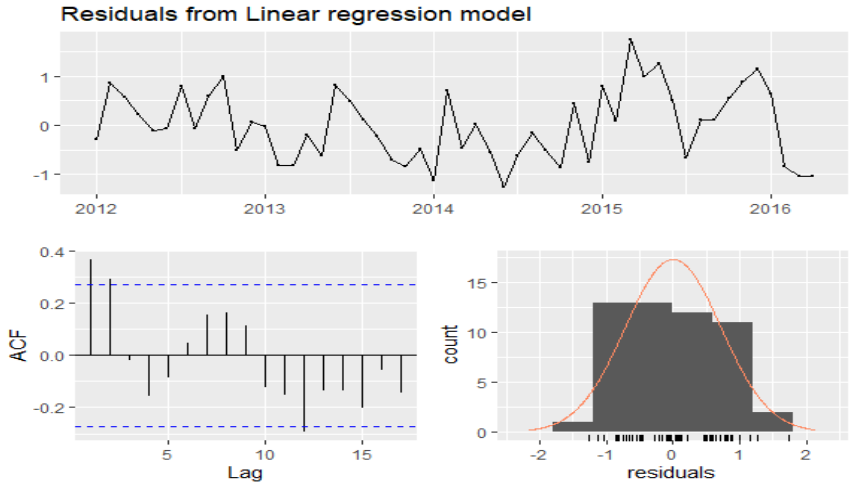
- Il test di Breusch-Godfrey suggerisce di rifiutare l'ipotesi nulla, quindi *non vi è autocorrelazione tra i residui*.

Terza Specificazione



- Anche in questo caso, per un livello di alpha pari al 5%, si rifiuta l'ipotesi nulla che i parametri associati alla stagionalità, al trend ed alla variabile *humidity* siano pari a zero.

Analisi dei Residui



- Si rifiuta l'ipotesi d'incorrelazione tra i residui che hanno anche una *distribuzione Normale* verificata con il test di Shapiro-Wilks e Jarque-Bera.

Normality Test	**Shapiro-Wilks**	**Jarque-Bera**
W	**0.970**	
X-Squared		**1.953**
P-value	**0.216**	**0.376**

Criteri di Scelta per i modelli

EVALUATION METRICS					
	CV	AIC	AICc	BIC	AjdR2
Model 1	1.489	20.341	31.666	47.632	0.976
Model 2	1.650	24.902	38.235	54.171	0.974
Model 3	1.218	4.556	31.652	43.581	0.983