2° PARTE

PROGETTO DI STATISTICA E MODELLI STOCASTICI (II MODULO)

Università degli studi di Bergamo 2020/2021, 13 giugno 2021

Nome del gruppo:
G25_MONACO

Componenti del gruppo:
Kenna Miriam Fatima (1058218), Dandis Iana (1065350),
Luca Lorenzi (1068520), Hurtado Beltran Ariana (1065528)

Docenti del corso:
Francesco Finazi
Frank Yannick MASSODA TCHOUSSI

Contenuto 2° parte:

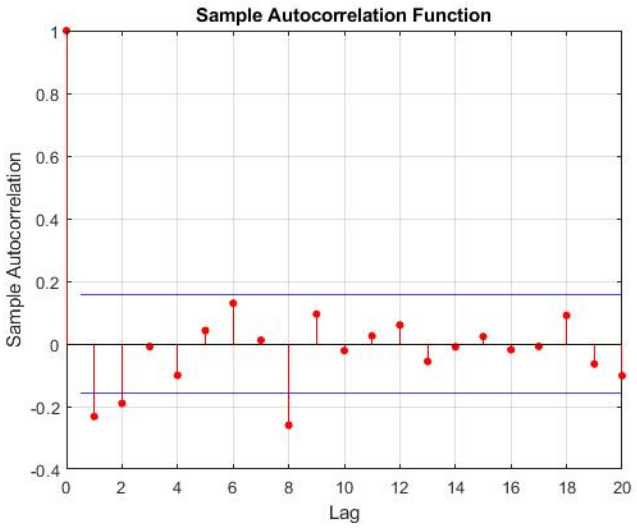
1. Introduzione e descrizione degli obiettivi
2. Autocorrelazione
3. Scelta del modello idoneo
4. Analisi della stazionarietà del modello
5. Diagnostica dei residui
    - 5.1 Autocorrelazione dei residui
    - 5.2 Varianza dei residui
    - 5.3 Residui IID
    - 5.4 Outliers
6. Bootstrap semi-parametrico per la stima di β1
7. Forecasting
8. Variazione [PM10] in relazione al cambiamento meteo

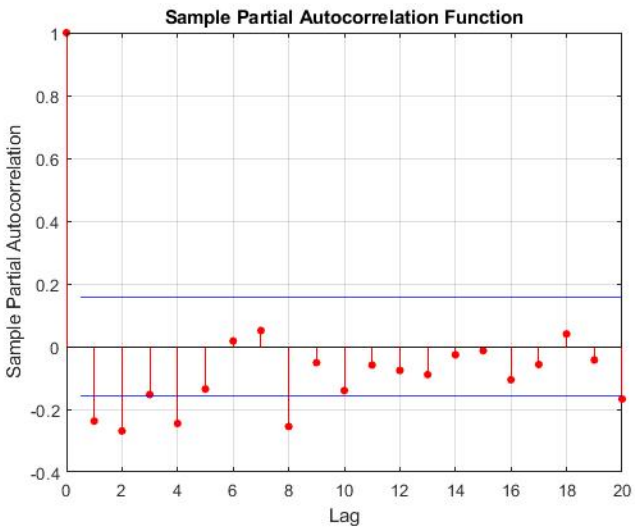
1. Introduzione e descrizione degli obiettivi

Nella prima parte del progetto ci siamo focalizzati sull'analisi del dataset utilizzando la statistica descrittiva, in particolare tecniche di regressione multipla, alcune tecniche di regressione avanzata e non parametrica.
Abbiamo però sempre ignorato il fatto che i dati del nostro database fossero stati raccolti nel tempo; non abbiamo quindi preso in considerazione la correlazione sulla componente stocastica. Utilizzare un modello di regressione classico, ma ignorare questa correlazione ci ha posto davanti il rischio di fare delle stime distorte (anche per $n \to \infty$). Precedentemente abbiamo gestito questa correlazione fra gli elementi tramite GLS, tuttavia senza tener conto che questo problema nasceva proprio perché i dati rappresentavano fenomeni che evolvono nel tempo.
L'obiettivo di questa 2° parte del progetto è introdurre nuovi strumenti che ci permetteranno di migliorare la qualità delle stime fatte in precedenza e di fare anche delle previsioni. Oltre ai Toolbox standard di Matlab in questa parte del progetto sono stati utilizzati *Financial Toolbox* e *System Identification Toolbox*.

2. Autocorrelazione

Studiamo l'autocorrelazione dei nostri dati destagionalizzati attraverso la funzione autocorr e parcorr, i quali tracciano la funzione di autocorrelazione totale e parziale campionaria della serie storica stocastica. I grafici ottenuti risultano essere:

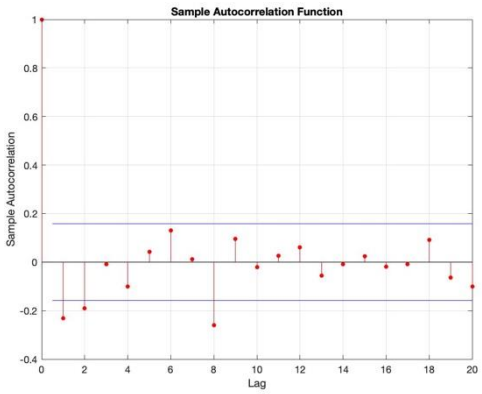


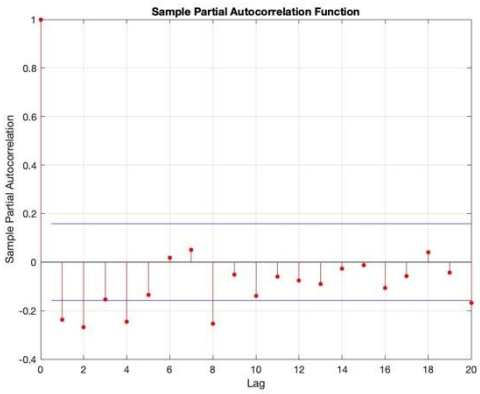


Possiamo notare come la maggior parte dei dati siano entro le bande di confidenza e da ciò possiamo affermare che i nostri dati non siano correlati tra di loro. Se vogliamo un'ulteriore conferma possiamo effettuare un test di Lijung-box il quale permette di verificare l'esistenza o meno di autocorrelazione entro una serie temporale. Il risultato è 1 quindi anche questo test conferma la nostra interpretazione dei grafici.

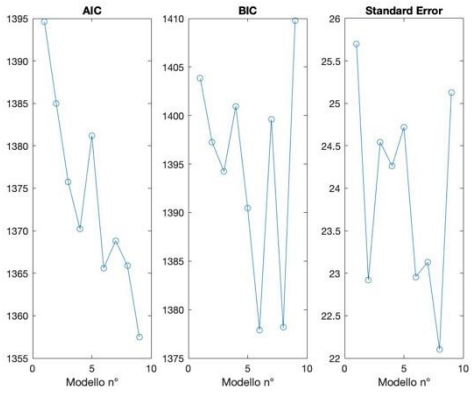
3. Scelta del modello idoneo

Successivamente allo studio dell'autocorrelazione e dell'autocorrelazione parziale è possibile determinare dei valori particolari di p e q, su cui poter costruire dei modelli ARMA. Indicati con: ☆





Tramite un algoritmo iterativo differenziato andiamo a valutare tutte le combinazioni di *p* e di *q*, con valori da 1 a 8. Delle 64 combinazioni, andiamo ad estrarre i due modelli con AIC e BIC migliore [*arma(7,0,8)* ed *arma(1,0,1)*], per poterli introdurre nella valutazione generale seguente:

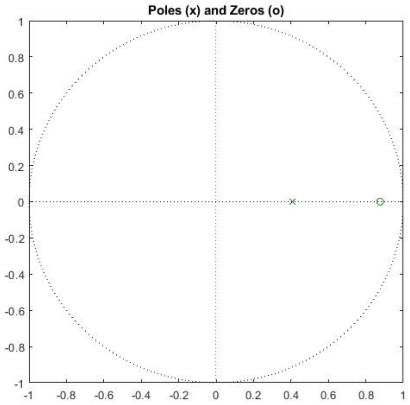


Sui modelli trovati applichiamo una particolare tecnica di validazione (criterio AIC + criterio BIC) per poter determinare qual è il migliore.

Utilizziamo anche la deviazione standard per poter avere un ulteriore elemento di conferma.

4. Analisi della stazionarietà del modello

Il modello ottimale risulta essere ***arma(1, 0, 1)***. Di esso studiamo la stazionarietà, la domanda che ci poniamo è:
le sue radici sono tutte interne al cerchio unitario?



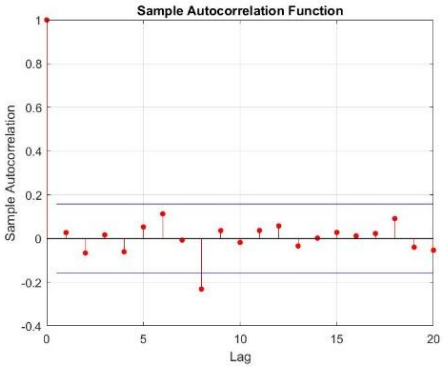
La risposta è affermativa, calcolando con il comando roots di matlab la radice otteniamo in valore assoluto un valore pari a 0,4210 < 1, come ci aspettavamo questo viene anche validato dall'utilizzo del *System Identification Toolbox*. Possiamo quindi concludere che il nostro modello è stazionario.

5. Diagnostica dei residui

Dopo aver scelto il *modello arma(1,0,1)* andiamo a diagnosticare i residui per poter confermare che il modello scelto da noi descriva correttamente il processo stocastico che stiamo analizzando.

5.1 Autocorrelazione dei residui

È necessario che i residui siano incorrelati nel tempo, altrimenti significa che non abbiamo utilizzato tutte le informazioni che era possibile estrarre dai residui.
Per verificare la presenza o meno dell'autocorrelazione abbiamo utilizzato Ljung-Box Q-test; l'ipotesi nulla del test è che i residui siano incorrelati. Nel nostro caso l'output del comando *lbqtest(…)* dei residui restituisce: *logical 0* quindi accettando l'ipotesi nulla, i nostri residui risultano essere incorrelati.



Riportiamo nella figura a destra una conferma grafica di quanto detto prima. Con il comando *autocorr (…)* applicato ai residui otteniamo un valore per ogni lag; nel nostro caso osserviamo che abbiamo un solo valore esterno alle bande blu(=intervallo di confidenza), il quale verrà ripreso successivamente.
Essendo che il resto dell'autocorrelazione campionaria è all'interno di queste bande, possiamo affermare che i nostri residui non presentano una correlazione temporale.

5.2 Varianza dei residui

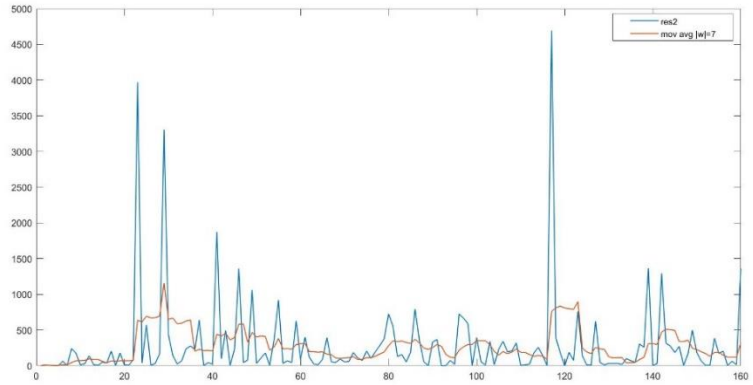
Essendo che il modello ARMA utilizzato per descrivere il processo è stazionario, ci aspettiamo che la varianza dei residui non cambi nel tempo altrimenti il processo sarebbe non stazionario (in contraddizione con il modello scelto).
Per verificare ciò utilizziamo "Engle test" per valutare l'eteroschedasticità dei residui. Nel nostro caso il comando *archtest()* ci consegna un *logical pari a 0*. Possiamo quindi concludere che la varianza dei nostri residui non ha variazioni significative e quindi il modello stimato è corretto per modellizzare i dati osservati.

5.3 Residui IID

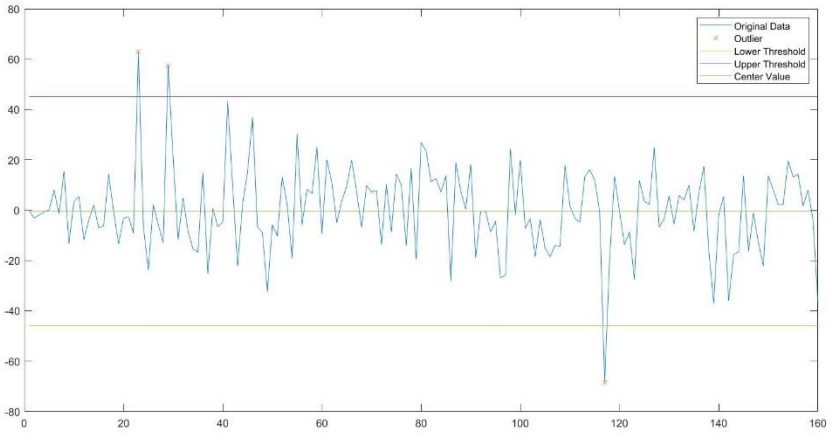
I residui dovrebbero essere incorrelati ed identicamente distribuiti (quindi con una distribuzione casuale senza nessun pattern temporale), altrimenti significa che seguono una struttura che è ancora possibile estrarre e utilizzare nel modello.
Controlliamo la normalità dei residui con il test di Jarque-Bera: l'output del test è *logical pari a 1* quindi i residui non sono normali. Per normalizzare i residui abbiamo aumentato il numero del campione (inizialmente uguale a 160) utilizzando il metodo del ricampionamento estraendo con riemmissione un numero casuale compreso fra il minimo e il massimo del vettore dei nostri residui. Otteniamo cosi un nuovo campione di 320 elementi costituito dai residui osservati e quelli campionati. Riproviamo a eseguire il test di Jarque-Bera per la normalità dei residui e otteniamo un output *logical pari a 0* e quindi essi risultano normali.

5.4 Outliers

Per poter osservare meglio il comportamento dei residui eliminiamo l'effetto del segno elevandoli al quadrato: sul grafico riportato osserviamo degli spike *[in particolare i valori anomali sono: X=23, 29, 117]* che potrebbero compromettere la varianza costante dei residui. Questi spike sono però dei punti isolati, quindi possono essere considerati degli *outliers*. Per capire se questi spike sono dovuti ad una variabilità sul vero processo che ci genera i dati oppure sono casuali, andiamo a costruire la media mobile con il comando "*movavg"* sui quadrati dei residui: in questo modo il modello risulta allisciato, infatti si può osservare che i picchi non sono più così alti.
Dato che la curva della media mobile non presenta alcun trend possiamo dire che i residui si comportano correttamente e quindi non siamo costretti ad utilizzare più di un modello per descrivere la traiettoria che abbiamo osservato.



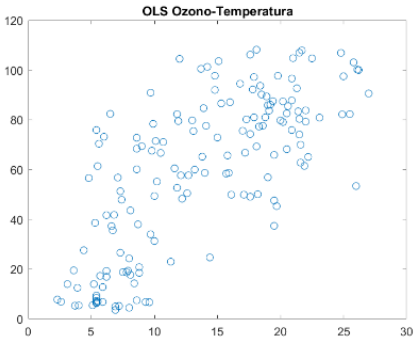
Per essere più precisi nell'identificazione degli outliers utilizziamo il commando **isoutlier**, che come possiamo vedere dal grafico riportato sotto definisce delle bande a 2-3 σ calcolate dai residui. Le X dei residui fuori dalle due bande sono residui anomali quindi ci aspettiamo di trovare sulla Y degli outliers; essi potrebbero essere dei valori dovuti ad una caratteristica del dataset quindi ad un evento accaduto nella realtà che ha influenzato l'andamento del dataset in modo anomalo. Per riuscire a capire meglio le cause prendiamo ad esempio l'outliers 117 che corrisponde al giorno 07/05/2020 e osserviamo che in quel giorno la quantità di biossido di azoto è più bassa (4,6) rispetto alla media (25,60).



A questo punto identificati gli outliers decidiamo di eliminarli dal dataset e osservare come cambiano le caratteristiche. Eliminando i valori di Y che corrispondono alle X degli outliers non osserviamo cambiamenti molto significativi né per quanto riguarda la media né la varianza.

6. Bootstrap semi-parametrico per la stima di β1

Usiamo i dati osservati per stimare β1. Il vantaggio di usare bootstrap per farlo è che possiamo rilassare l'ipotesi della distribuzione normale dei residui poichè andremo semplicemente a ricampionare i residui per creare nuove Y e stimare β1 sui dati ottenuti.



Prima quindi andremo ad usare OLS e scegliamo come variabile risposta Y ancora l'Ozono mentre come variabile esplicativa X scegliamo la Temperatura. Deduciamo dal grafico a destra (che rappresenta la distribuzione di ogni coppia di X e Y) che è possibile spiegarla con un modello lineare. Quindi l'equazione della retta che mi spiega i dati è:

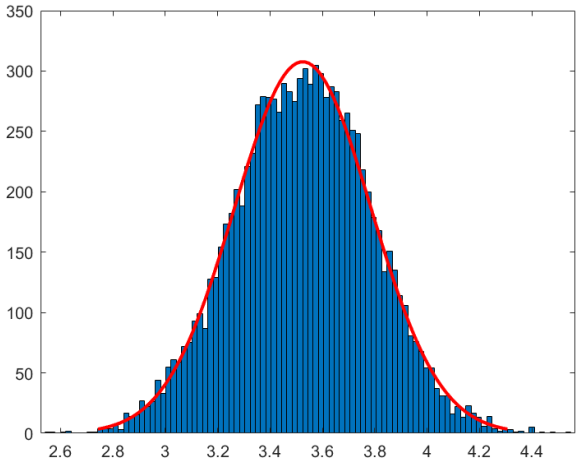
$\boldsymbol{Y = \beta_0 + \beta_1 X_1}$ ovvero **Ozono ~ 1 + Temperatura**

Avendo l'equazione della retta, come primo passo, ricostruiamo il vettore degli uni per l'intercetta; dopo di che useremo il comando *regress* [al quale passiamo i parametri: Y, X e α=0.05] il quale restituisce in output i residui che poi andremo a usare per costruire nuove Y

Facciamo una stima del coefficiente angolare estratta dal comando regress:

coefficiente angolare + IC 95%
3.0104 3.5281 4.0458

Ora per verificare che effettivamente l'intervallo di confidenza del comando *regress* sia corretto andremo ad effettuare il bootstrap. Creiamo quindi il vettore della previsione per le X deterministiche (= dati osservati). Poi usando un ciclo iterativo (abbiamo deciso di eseguire 10000 interazioni) campioniamo dei residui del modello e gli usiamo per generare un nuovo campione ad ogni interazione sommando il residuo all'equazione determinata prima.
Salviamo i risultati infine in un vettore dal quale poter poi estrarre β1. Nel grafico a destra riportiamo il risultato delle 10000 interazioni che riguardano l'inferenza su β1.



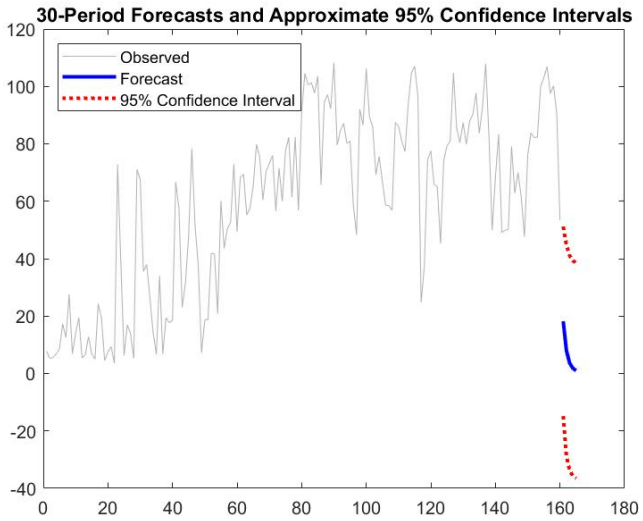
A questo punto ricalcoliamo il coefficiente angolare con un IC al 95% e lo confrontiamo con i valori di prima forniti dal comando *regress*: vediamo che i valori sono molto vicini e quindi deduciamo che la stima era corretta

coefficiente angolare + IC 95%
3.0130 3.5249 4.0511

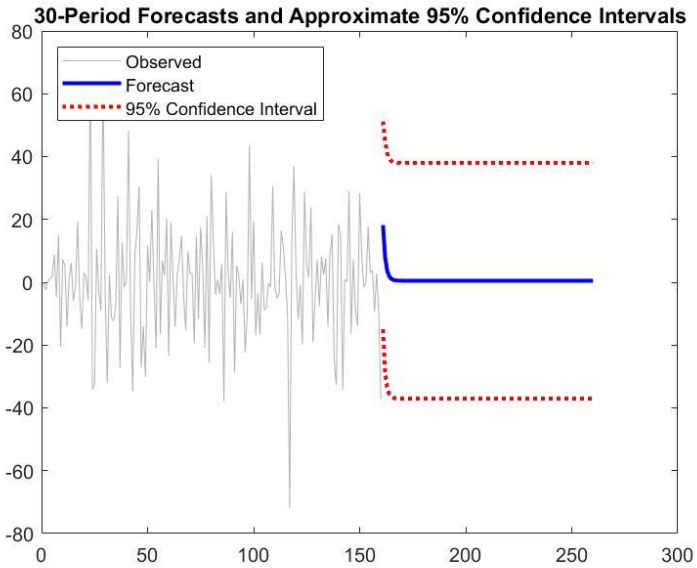
7. Capacità previsiva del modello: Forecasting

Dato che le premesse fondamentali per poter effettuare una previsione nel nostro caso sono rispettate, cioè il modello è stazionario mentre i residui sono normali, incorrelati e omoschedastici usiamo il comando *forecast* per poter osservare l'andamento del nostro modello in un futuro di lag da noi scelto. La previsione viene accompagnata da un intervallo di confidenza al 95%, in modo tale da tenere conto dell'eventuale incertezza della previsione.

I parametri passati al comando *forecast* sono: il miglior modello individuato precedentemente ARMA (1, 0, 1), il numero di step per i quali vogliamo stimare la previsione ed i dati appartenenti al dataset su cui lavoriamo, in quanto la previsione è una funzione di tutta la storia precedente. Guardando la figura possiamo notare che forecast mi restituisce i valori previsti a 5 step con la loro varianza associata.



Osservando il grafico concludiamo inoltre che il modello scelto è in grado di fare una previsione che segue l'andamento reale dei dati. Questo si verifica per un numero di step da prevedere in avanti relativamente piccolo, ad esempio prendendo in considerazione invece una previsione a cento passi il modello tende a zero, ossia la media marginale del nostro processo, in quanto effettuare la previsione con un lag temporale così ampio porta un'incertezza molto ampia. Concludiamo quindi che fare previsioni con un lag temporale cosi ampio non porta alcun vantaggio al fine di effettuare previsioni come si vedrà nell'ultimo paragrafo del nostro elaborato. Possiamo trovare conferma di quest'ultima conclusione anche nelle previsioni meteo esse risultano essere veritiere in un range dei lag da 1 a 14, ma già a 14 notiamo che l'incertezza è ampia



8. Variazione [PM10] in relazione al cambiamento meteo

Per la realizzazione di quest'ultima parte abbiamo introdotto un nuovo dataset "Meteo" in qui abbiamo raccolto variabili metereologiche di nostro interesse per poter proseguire con la nostra indagine e rispondere agli quisiti che ci siamo posti e che verranno ampiamente sviluppati successivamente.

Uno degli inquinanti atmosferici più interessanti che troviamo nell'area di nostro interesse è il PM10, il quale sappiamo dipendere particolarmente dalle caratteristiche metereologiche.

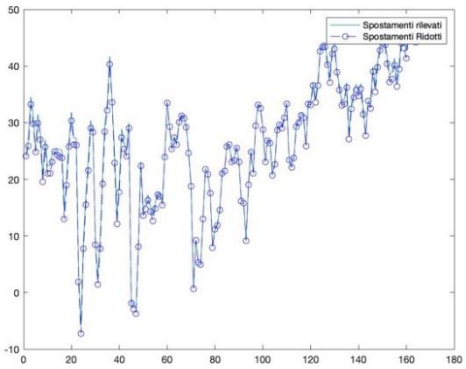
Per la 1° ipotesi però ipotizziamo che le condizioni meteo siano costanti, a tal punto da non inserirle nel modello. Questa ipotesi risulta essere errata in quanto le variazioni della concentrazione di PM10 non possono dipendere esclusivamente dalle variazioni degli Spostamenti in Macchina. Ciò è visibile anche eseguendo una fitlm tra il PM10 e gli spostamenti in macchina, la quale riporta un *R-squared di 0.0753*.

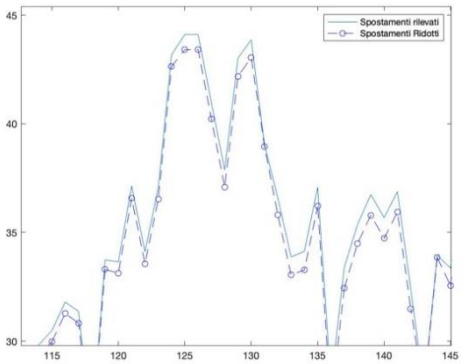
Difatti, la prima ipotesi presentata, di condizioni meteo costanti, risulta essere astratta e quindi irrilevante.

Il nostro obbiettivo è quindi individuare un modello che descriva la dipendenza del PM10 dalle variabili metereologiche che andiamo a considerare. Una volta individuato il modello migliore (tramite tecniche di validazione) andiamo a rispondere ad alcuni quesiti:

Se ipotizzassimo di diminuire gli spostamenti in macchina fino ad 1/3 del valore osservato, che cosa accadrebbe alla concentrazione di PM10?

Diminuendo idealmente gli spostamenti in macchina nel comune di Milano, si arriverebbe ad una diminuzione della concentrazione di PM10 non molto significativa: $\Delta = -2{,}24\ \%$.

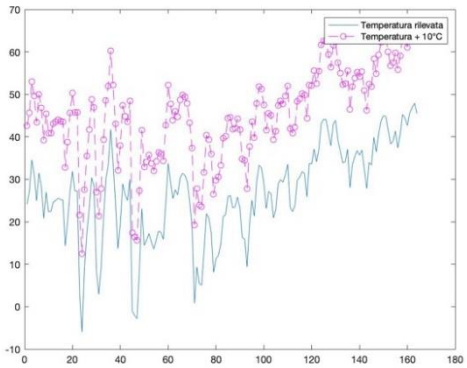


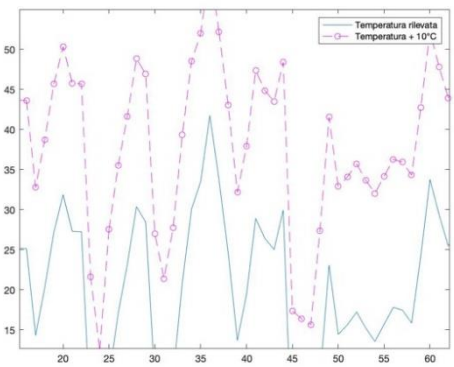


Ciò lascia intendere che certamente l'utilizzo dell'auto a Milano risulta influenzare la concentrazione di PM10, ma non è di sicuro una delle cause principali su cui poter andare ad agire con successo nella sua diminuzione.

Se ipotizzassimo che la temperatura media aumentasse di 10°C , a causa del surriscaldamento globale, cosa accadrebbe alla quantità di PM10 nell'aria?

Aumentando la temperatura di 10°C, si osserva che la concentrazione dell'inquinante di nostro interesse aumenterebbe significativamente: $\Delta = +75{,}59\ \%$.

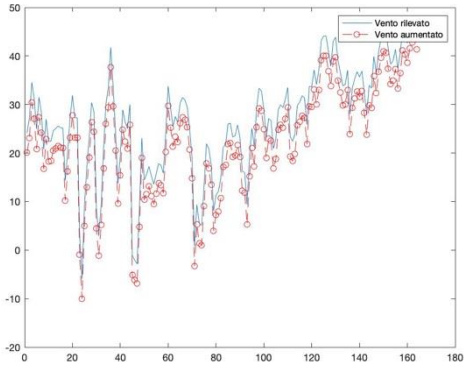


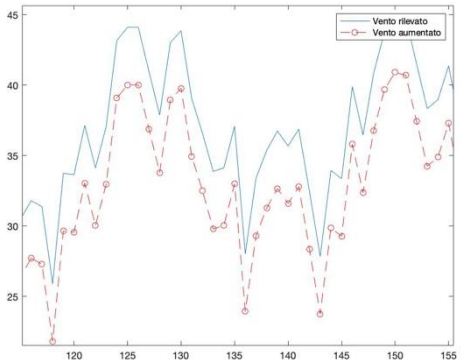


Ne consegue che un aumento drastico della temperatura, dovuto al surriscaldamento globale purtroppo sempre più in aumento, porterebbe ad un drastico aumento della concentrazione degli inquinanti.

Sappiamo che una delle cause del surriscaldamento globale, risulta essere l'albedo (=coefficiente di riflessione delle radiazioni). Quest'ultimo dipende anche dalla quantità di inquinanti che si trovano nell'aria, i quali vengono trasportati attraverso l'atmosfera dal vento. Di conseguenza, cosa accadrebbe all'aumento del vento rilevato?

Aumentando la presenza del vento di +5 unità, avremmo una discreta diminuzione del PM10, pari ad : $\Delta = -16{,}74\ \%$.





Ne consegue che un aumento del vento, porterebbe ad una diminuzione della concentrazione degli inquinanti con conseguente aumento dell'albedo e quindi ad un minore riscaldamento.