G30 - PODGORICA

Francesco Foresti - 1060073

Alberto Noris - 1054752

Davide Salvetti – 1057596

Descrizione Dataset

Il dataset scelto contiene dati sulla qualità dell'aria nell'area di Bergamo partendo dal 01/01/2017 al 31/12/2020, con dati a frequenza oraria. In particolare, i parametri che abbiamo considerato per il nostro studio sono stati: precipitazioni, temperatura, umidità relativa, radiazione globale, velocità del vento, direzione del vento, ossido di azoto, biossido di azoto, monossido di carbonio e ozono.

Quesiti

1. Vista l'importanza ambientale dell'azoto decidiamo di focalizzare la nostra analisi su di lui, ci chiediamo quindi quale sia il miglior modello che riesca a descrivere l'O3.
2. La temperatura nel tempo ha un andamento periodico, le Splines e le Basi di Fuorier sono tecniche che si adattano bene a comportamenti di questo tipo, decidiamo quindi di trovare il modello che riesca a stimare meglio i valori di temperatura nel tempo.

Quesito 1: Ricerca del miglior modello per l'O3

1.1) Introduzione

Dopo aver caricato il dataset, ed aver tolto i dati non validi, abbiamo svolto una prima analisi attraverso una matrice di correlazione e un plotmatrix delle variabili in studio.

	Precipitazione	Temperatura	Umidità	Radiazione	Vel Vento	Dir Vento	NO	NO2	CO	O3
Precipitazione	1	-0.017103	0.12761	-0.070073	0.15824	-0.049944	-0.038585	-0.052583	-0.0509	0.017016
Temperatura	-0.017103	1	-0.50651	0.49786	0.25021	-0.096438	-0.46667	-0.52022	-0.47249	0.83294
Umidità	0.12761	-0.50651	1	-0.579	-0.32641	0.038321	0.21829	0.18915	0.2478	-0.62865
Radiazione	-0.070073	0.49786	-0.579	1	0.28943	-0.02697	-0.19364	-0.2477	-0.19456	0.52652
Vel Vento	0.15824	0.25021	-0.32641	0.28943	1	-0.36199	-0.22902	-0.27456	-0.26423	0.36336
Dir Vento	-0.049944	-0.096438	0.038321	-0.02697	-0.36199	1	0.022721	0.015642	0.036285	-0.073063
NO	-0.038585	-0.46667	0.21829	-0.19364	-0.22902	0.022721	1	0.83485	0.84641	-0.52208
NO2	-0.052583	-0.52022	0.18915	-0.2477	-0.27456	0.015642	0.83485	1	0.71951	-0.604
CO	-0.0509	-0.47249	0.2478	-0.19456	-0.26423	0.036285	0.84641	0.71951	1	-0.49632
O3	0.017016	0.83294	-0.62865	0.52652	0.36336	-0.073063	-0.52208	-0.604	-0.49632	1

Come si può notare dalla matrice di correlazione, l'O3 è fortemente correlato con temperatura ($\rho = 0.83$), umidità ($\rho = -0.628$) e NO2 ($\rho = -0.604$), ma allo stesso tempo notiamo che NO e NO2 sono molto correlati, come anche CO ed NO. Dovremo porre attenzione nello scegliere quali tra questi regressori inserire nel modello, così da non incorrere in overfitting.

Procediamo nelle analisi con dei modelli di regressione lineare semplici tra O3 e i regressori con un ρ elevato e notiamo che i modelli con temperatura e umidità sono quelli che presentano un R^2 più alto.

1.2) Modello a 2 Regressori: modello di regressione lineare multipla a 2 regressori

Una volta individuati i regressori che si correlano meglio con l'O3 decidiamo di creare un modello di regressione lineare multipla. In questo modello abbiamo implementato la temperatura e l'umidità in quanto erano le due variabili che restituivano un R^2 più alto. Per la creazione del modello utilizziamo la tecnica dei Minimi Quadrati e per fare degli intervalli di confidenza e dei test di ipotesi sui coefficienti beta stimati controlliamo che gli errori si distribuiscano normalmente. Per controllare l'ipotesi nulla che i nostri residui provengano da una normale effettuiamo il test Jarque-Bera, il quale però rifiuta l'ipotesi portandoci quindi alla conclusione che i nostri errori non si distribuiscono normalmente. La normalità dei residui però non è una condizione necessaria per la normalità di beta cappello e quindi per dimostrare la normalità dei Beta proviamo a sfruttare il teorema del limite centrale. Perché valga il TLC gli errori devono essere IID e per controllare questa condizione iniziamo col verificare se i residui siano o meno omoschedastici.

Facciamo quindi un plot dei residui in relazione ai valori di y fittati e notiamo come l'andamento del grafico evidenzi una chiara eteroschedasticità (*Figura 1*). Escludiamo quindi che la distribuzione di beta cappello sia normale. A questo punto visto che gli errori sono eteroschedastici proviamo a controllare se siano correlati tra di loro: per farlo utilizziamo la funzione autocorr la quale però ci restituisce un grafico (*Figura 2)* che evidenzia una chiara autocorrelazione nei residui, non potremo quindi utilizzare la tecnica WLS.

Dopo aver dimostrato che gli errori non sono IID, decidiamo di verificare se la stima effettuata con OLS sia asintoticamente non distorta. Per farlo calcoliamo la matrice δ la quale ha tutti i componenti ≃ 0. Questo ci porta alla conclusione che, nonostante gli errori siano non IID la stima effettuata tramite OLS sia non distorta ma subottimale in quanto il miglior stimatore (BLUE) diviene GLS.

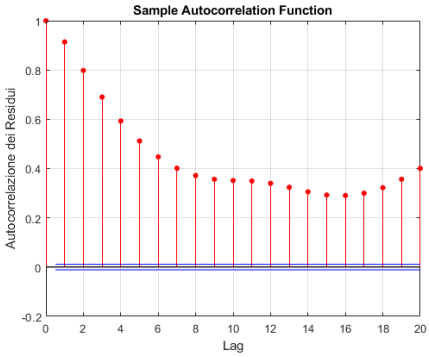

Figura 2

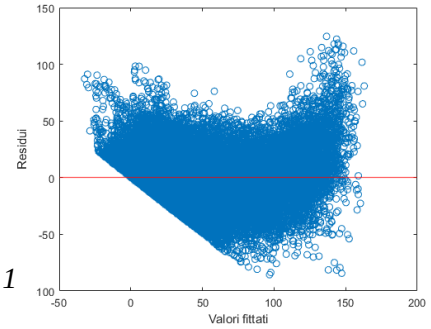

Figura 1

1.3) Analisi di 'Aumento numero Regressori' tramite Crossvalidazione

Arrivati a questo punto vogliamo verificare se abbia senso aggiungere altri regressori al modello di regressione lineare multipla analizzato in precedenza. Per fare ciò implementiamo un ciclo di cross-validazione che funziona in questo modo:

- partendo dal modello analizzato nel punto precedente O3 ~ Temperatura + Umidita creiamo altri 7 modelli dove in ognuno aggiungiamo un regressore al modello.
- Per ogni modello tramite la tecnica della cross-validazione con 10 fold calcoliamo l'MSE.
- Ripetiamo questo procedimento 10 volte e plottiamo i risultati ottenuti.

Analizzando i grafici ottenuti possiamo notare come all'aggiunta del 3° regressore vi sia una drastica diminuzione dell'EQM, accompagnata successivamente, con l'aumento del numero di regressori, da una lieve diminuzione di esso. Contemporaneamente calcoliamo per ogni modello l'R2 e possiamo notare (*Figura 3*) come al 3° regressore vi sia un aumento significativo accompagnato poi da un lieve incremento con l'aumentare del numero di regressori (d'altronde sappiamo che all'aumentare del numero dei regressori, R2 non può far altro che aumentare). Entrambe le analisi ci portano quindi alla conclusione di scegliere un modello con 3 regressori (evitando così overfitting). La scelta tiene conto anche della scarsa correlazione tra O3 e regressori come Precipitazioni, Vento e Direzione Vento, e tiene conto del fatto che regressori come NO e CO sono correlati con regressori come NO2 che fa parte del modello scelto.

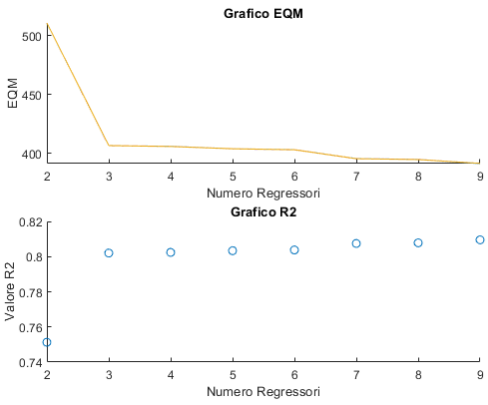

Figura 3

1.4) Modello a 3 Regressori: analisi del modello a 3 regressori ottenuto dalla crossvalidazione

Una volta capito che il modello che restituisce il miglior equilibrio tra MSE e R2, così da non andare in contro ad under o overfitting, è O3 ~ Temperatura + Umidita + NO2 procediamo con l'analisi così da verificare che la stima dei coefficienti sia accettabile.

Per farlo procediamo con la stessa tecnica analizzata al punto 1.2 arrivando anche qui alla conclusione che gli errori non sono Normali e nemmeno IID. Tuttavia, la matrice δ ha tutti i componenti ≃ 0 e quindi la stima è non distorta asintoticamente.

1.5) Modello Polinomiale: modello di regressione polinomiale

Infinte dal plotmatrix notiamo che l'O3 e la temperatura sembrino seguire un andamento non rettilineo e quindi decidiamo di provare a creare un modello dove l'O3 viene spiegato tramite ordini crescenti di temperatura. Per fare ciò creiamo 7 modelli diversi dove in ognuno aumentiamo il grado del polinomio e successivamente testiamo quale sia il migliore tramite la tecnica di crossvalidazione vista in precedenza. Comparando i risultati della crossvalidazione con

gli R2 di ogni modello possiamo notare come anche qui dopo il 2° grado del polinomio l'MSE e l'R2 rimangano quasi invariati portandoci quindi alla conclusione che il miglior modello polinomiale che descrivere l'O3 tramite la temperatura è quello di secondo grado. Terminiamo l'analisi verificando che gli errori sono non IID ma che anche qui δ ha tutti i componenti $\simeq 0$. La stima ottenuta per i beta quindi è non distorta.

1.6) Conclusioni Quesito 1:

Una volta ottenuti i 3 modelli di regressione multipla che descrivono l'azoto facciamo un confronto dei risultati ottenuti così da scegliere quello che lo descrive meglio. In particolare, confrontiamo l'MSE e gli R2 ottenuti nei 3 modelli analizzati in precedenza e possiamo notare come il modello a 3 regressori sembri essere quello che spiega meglio l'andamento dell'azoto preso in analisi.

	MSE	**R2**
Modello a 2 regressori	510	0.7513
Modello a 3 regressori	406	0.8020
Modello Polinomiale	505	0.7538

Quesito 2: Ricerca del miglior modello per la temperatura

2.1) Confronto B-Splines con Basi di Fourier tramite Crossvalidazione

Abbiamo deciso di modellizzare la temperatura rispetto al tempo. La temperatura è una variabile periodica e quindi le basi di Fourier sono ottime per modellizzarla. Tuttavia, abbiamo deciso di fare un confronto tra basi di Fourier e B-Spline di ordine 4 per vedere quale delle due si adattasse meglio a parità di numero di basi. Tramite crossvalidazione abbiamo estratto i GCV di entrambi i modelli per numero di basi crescente e ne abbiamo poi mostrato l'andamento.

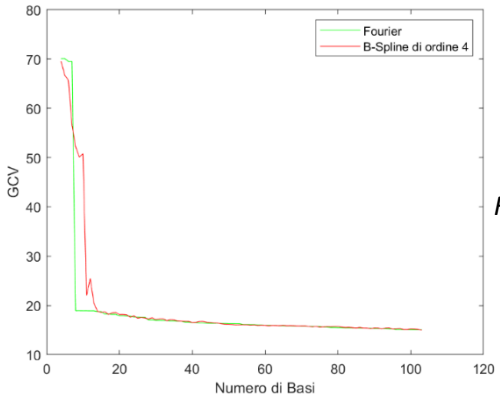

Figura 4

Come si può vedere dal grafico di fianco (*Figura 4*), all'aumentare del numero di basi il GCV tende continuamente a diminuire. Inoltre, non c'è una grossa differenza in termini di GCV tra BSpline e basi dsi fourier, se non per numero di basi minore di 20, dove le basi di Fourier si comportano meglio delle BSpline. Quindi abbiamo scelto di usare un modello con basi di Fourier dal momento che sono ottime per la modellazione di trend periodici.

Per scegliere il numero di basi adatto a descrivere correttamente il modello ci siamo affidati ai plot delle funzioni stimate: abbiamo subito notato come per un numero di basi alto (sopra i 100) il modello non è per niente smooth e porta a overfitting (*Figura 5*). Tuttavia, un numero di basi inferiore a 30 sembra essere troppo poco informativo (underfitting). Abbiamo quindi deciso di prendere come numero di basi un valore di 50 (*Figura 6*), nonostante questo abbia un GCV più alto rispetto a modelli con un maggior numero di basi. In aggiunta, abbiamo calcolato e plottato sul grafico anche gli intervalli di confidenza dei singoli punti del modello con 50 basi.

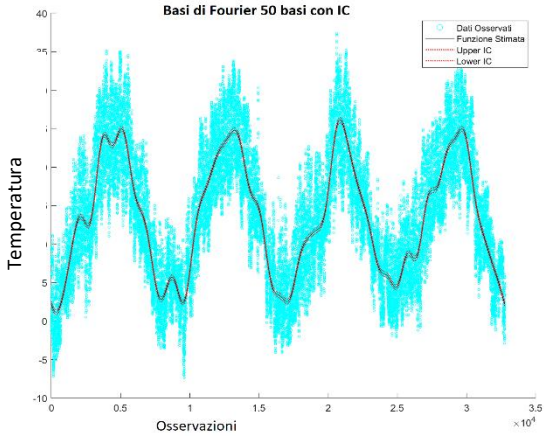

Figura 6

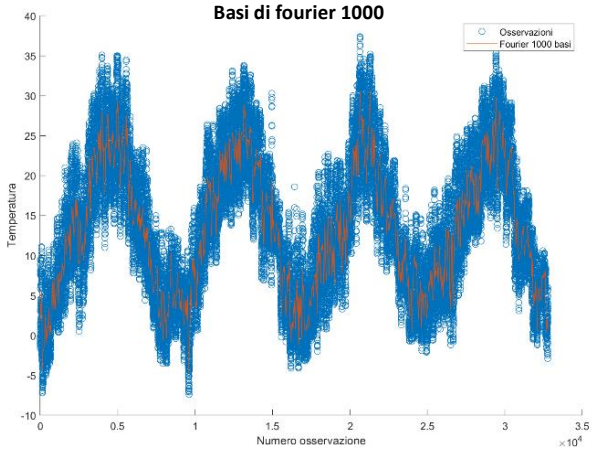

Figura 5

2.2) Conclusioni Quesito 2:

Dall'analisi effettuata arriviamo alla conclusione che entrambe le tecniche di B-Splines e Basi di Fuorier seguano bene l'andamento della temperatura ma dalla letteratura sappiamo come quest'ultime rappresentino meglio gli andamenti periodici e quindi decidiamo di scegliere loro. Otteniamo quindi un modello finale con Basi di Fourier a 50 basi. Questa soluzione sembra restituire un giusto equilibrio tra Overfitting e inseguimento del modello.