LAVORO DI GRUPPO DI STATISTICA

Influenza di fenomeni atmosferici sull'altezza della neve nella località di Livigno

Relatori:

- Francesco Corrini, matricola 1067709
- Federico Imberti, matricola 1066358
- Andrea Ingoglia, matricola 1068751
- Enrico Perani, matricola 1066174

Introduzione

I fenomeni atmosferici e naturali possono influire significativamente sulle attività commerciali ed i loro ricavi. Possiamo prendere come esempio gli stabilimenti balneari, che generano la maggior parte dei ricavi nella stagione estiva, oppure le attività sciistiche durante l'inverno.

Poter prevedere, con una certa precisione, il periodo migliore in cui aprire e chiudere la stagione, può essere determinante al fine di ottimizzare gli investimenti. Il nostro gruppo ha dunque deciso di porsi come obiettivo quello provare a stimare l'altezza della neve nella località di Livigno (SO) conoscendo i principali fenomeni atmosferici.

I dati utilizzati sono stati forniti dal sito di Arpa Lombardia: questi sono stati misurati dalla stazione meteorologica di Livigno posta ad un'altitudine di 2660 metri. I dati sono stati raccolti ogni 10 minuti dal giorno 22-03-2020 al 22-03-2021. I fenomeni atmosferici presenti nel nostro dataset sono temperatura, precipitazioni, radiazione globale, umidità relativa e velocità del vento, oltre all'altezza della neve.

La nostra domanda è quindi: possiamo costruire un modello per prevedere l'altezza della neve con questi dati?

Descrizione dei metodi utilizzati per l'analisi del dataset

Con il dataset a nostra disposizione abbiamo deciso costruire un modello di regressione lineare. Il modello mette in relazione gli eventi atmosferici (che fungeranno da covariate) e l'altezza della neve (che sarà la nostra variabile risposta).

Siccome l'andamento dei dati non è periodico abbiamo scartato l'ipotesi di utilizzare delle basi di fourier per stimare il modello. Anche le spline si sono dimostrate poco efficaci per i nostri scopi: oltre ad avere un'elevata complessità computazionale, non danno molta informazione su ciò che stiamo cercando, infatti, tentare di stimare la neve dando in input il tempo (l'istante in cui si vuole avere la previsione) non sembra molto utile.

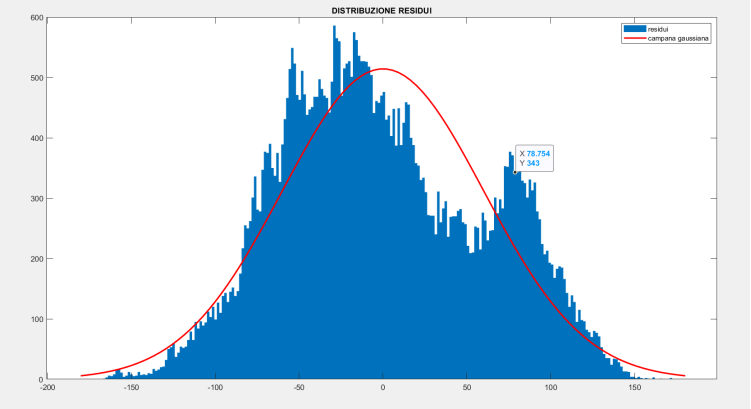
Abbiamo dunque optato per un metodo di regressione lineare. La stima che abbiamo scelto è quella OLS (minimi quadrati ordinari), che ha però diversi problemi: essendo il nostro dataset formato da serie storiche, è possibile che gli stimatori del modello ottenuto siano distorti. Tuttavia, un modello GLS (minimi quadrati generalizzati) avrebbe una complessità computazionale elevata, in quanto andrebbe calcolata l'inversa di una matrice quadrata con

più di 50000 dati. Il metodo dei minimi quadrati ponderati, oltre a presentare quest'ultima problematica, non sarebbe comunque efficace perché richiede che gli errori siano incorrelati, esattamente come le stime OLS. Abbiamo quindi scelto queste ultime per la loro semplicità.

Una volta ottenuto il modello abbiamo calcolato delle informazioni necessarie per valutare la qualità delle nostre stime: il coefficiente di determinazione ed il mean square error. Abbiamo poi fatto un *test t* sui coefficienti del modello per verificare che siano significativi. Infine, abbiamo verificato che il modello non soffrisse di overfitting tramite un algoritmo di cross-validazione. Inizialmente avevamo optato per un algoritmo *leave one out*, abbandonato però perché computazionalmente oneroso (ricordando sempre l'ampiezza del dataset). Abbiamo quindi implementato un algoritmo *k-fold* (con k = 10), meno complesso del precedente ma comunque molto efficace.

Analisi dei risultati del modello ottenuto

Utilizzeremo la notazione *X* per definire la matrice delle covariate e con *y* indichiamo il vettore della variabile risposta.
In un modello di regressione è necessario che i regressori siano tra loro linearmente indipendenti. Questa condizione si riflette in particolare nella condizione di esistenza delle soluzioni per i coefficienti stimati, la quale prevede che $det\left(X^{'}X\right) \gg 0$. Se così non fosse, due o più regressori porterebbero lo stesso contenuto informativo e risulterebbero quindi ridondanti.
Nel nostro caso, $det\left(X^{'}X\right) = 9,46 * 10^{35}$, quindi la condizione è ampiamente rispettata.
Il modello ha coefficiente di determinazione $R^2 = 0.52$. Ciò significa che solo il 52% della varianza della y è spiegata dal modello.
Analizzando i residui, si nota che essi non si distribuiscono secondo una normale. Questo è anche confermato dalla statistica di Jarque-Bera che ha valore 1413 (più il valore è alto, meno i residui sono normali). La media dei residui però è nulla e per questo possiamo dire che i coefficienti stimati non sono distorti (non asintoticamente). Tuttavia, l'assunzione di omoschedasticità degli errori non viene rispettata. I residui di questo modello sono eteroschedastici, ovvero presentano varianze ciascuna diversa dalla altre.



I coefficienti trovati sono raccolti nella seguente tabella. Viene inoltre calcolata la statistica test per effettuare un test d'ipotesi e verificarne la significatività. Il test è così costruito:
H0: $\beta_i = 0$**, H1:** $\beta_i \neq 0$
Fissiamo **α** = 1%, con il quale calcoliamo la tCritica = 2,576. Se la tStat del coefficiente si troverà nell'intervallo di confidenza [-tCritica, +tCritica] allora dovremo accettare il test

d’ipotesi, e quindi dedurre che il coefficiente non è significativo. Al contrario, dimostreremo che il coefficiente è significativo e deve rimanere nel modello.

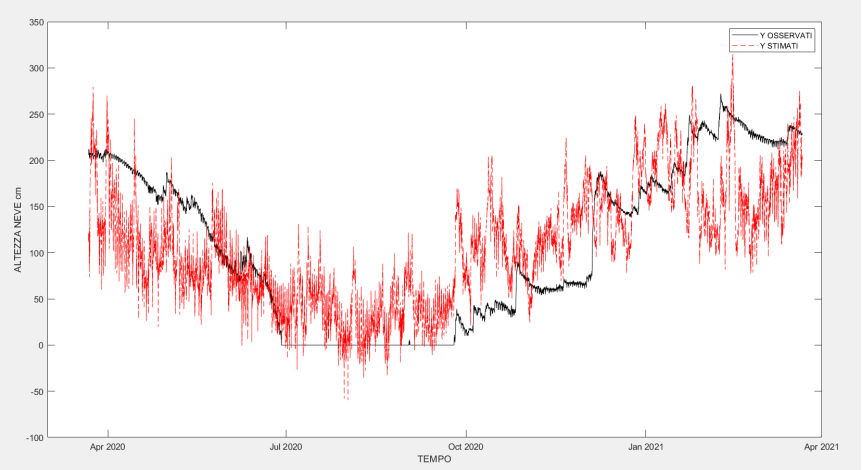
Nome del regressore	Valore del coefficiente	Valore della tStat	E’ significativo?
Intercetta	153,7	156,6	Si
Temperatura	-8,702	-226,1	Si
Precipitazioni	-14,45	-6,274	Si
Umidità relativa	-0,8201	-65,49	Si
Radiazione globale	0,0565	48,47	Si
Velocità del vento	1,521	8,591	Si

Come visto dalla tabella, l’ipotesi H0 deve essere rifiutata per tutti i regressori, e quindi risultano tutti significativi.

Abbiamo poi implementato un algoritmo di cross-validazione per capire se il modello soffre di overfitting. L’algoritmo dà come risultato un MSE_test ≃ 3611, molto simile all’MSE calcolato all’inizio. Possiamo dire che il modello stimato non soffre di overfitting.

Conclusione

L’obiettivo principale del progetto è quello di stimare l’altezza della neve sulla base degli effetti atmosferici. Dalle analisi effettuate, si arriva alla conclusione che il modello trovato si dimostra poco utile, soprattutto per l’elevata varianza: infatti, avendo ottenuto una deviazione standard di 60 (confermata in fase di validazione), si ottengono delle stime troppo distanti dai valori misurati, come mostrato dal seguente grafico.



Questo risultato è confermato anche dal coefficiente di determinazione basso, il quale ci dice che solo il 52% della varianza dell’altezza della neve viene spiegata dal nostro modello.
Per questo motivo, per stimare l’altezza della neve servono modelli più complessi, che tengano in considerazione la correlazione dei dati nel tempo.