Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Strategie di Imputazione in Ambiti Complessi: Studio Comparativo tra Modelli MICE e modelli di Machine Learning

Relatore Prof. Omar Paccagnella
Dipartimento di Scienze Statistiche

Laureando Luigi Venuto
Matricola 2037878

Anno Accademico 2023/2024

Indice

Introduzione

Nel vasto panorama dell'analisi statistica dei dati, uno degli ostacoli più comuni è rappresentato dalla presenza di dati mancanti per alcune variabili, noti come NA (*not available*). Questa carenza può derivare da diverse cause legate alla natura stessa dei dati. In particolare per i dataset di grandi dimensioni, la probabilità di avere un numero considerevole di dati mancanti aumenta notevolmente.

Per affrontare questa problematica, sono state sviluppate negli anni varie tecniche di analisi, ognuna con efficacia diversa a seconda della natura dei dati mancanti. Un approccio comune è l'eliminazione di intere unità statistiche (listwise deletion - LD) contenenti almeno un dato mancante. Tuttavia, questo metodo può causare una perdita significativa di informazioni e distorsioni nei risultati, maggiormente quando i dati a disposizione per l'analisi sono limitati, dove l'eliminazione delle unità statistiche che presentano almeno un dato mancante porterebbero ad un'ulteriore perdita di informazioni.

L'approccio utilizzato in questa tesi si distingue per l'utilizzo di una vasta gamma di tecniche di imputazione, mirate a compensare le lacune informative nel dataset. Tra queste, sono state esplorate e applicate con attenzione metodologie avanzate, compresi alcuni metodi per l'imputazione multipla e l'integrazione di modelli di machine learning per l'imputazione singola.

Attraverso l'esame di diverse tecniche di trattamento dei dati mancanti, questa tesi non solo si propone di colmare le lacune informative, ma anche di contribuire alla crescente comprensione delle sfide e delle soluzioni legate a questo importante aspetto dell'analisi statistica. L'indagine comprende anche una fase di simulazione per valutare le differenze tra i metodi di imputazione adottati, offrendo così una prospettiva critica sui risultati ottenuti. Queste soluzioni verranno applicate ad un dataset comprendente dati di rilevazioni di elementi chimici di depositi solfuri massivi vulcanogenici.

In sintesi, questa tesi offre una panoramica approfondita sulle principali metodologie di trattamento dei dati mancanti, evidenziando il ruolo cruciale che svolgono nell'analisi statistica e contribuendo allo sviluppo di approcci più sofisticati e adattabili per affrontare questa sfida in ambiti specifici.

Questo lavoro è strutturato nel seguente modo. Nel primo capitolo verranno introdotti i concetti fondamentali riguardanti i dati mancanti, mentre nel secondo capitolo si parlerà di alcuni metodi per il loro trattamento. Il terzo e il quarto capitolo riguarderanno rispettivamente metodi per l'imputazione semplice e multipla, mentre il capitolo cinque vedrà applicati questi metodi al dataset in esame. Infine verranno presentati i rusultati delle analisi e le conclusioni.

Capitolo 1

I dati mancanti

Prima di esplorare le strategie per gestire i dati mancanti, è fondamentale acquisire una comprensione dei concetti fondamentali come la non risposta, la struttura delle risposte mancanti e il meccanismo generatore dei dati mancanti. Quest'ultimo aspetto è cruciale poiché stabilisce la relazione tra dati mancanti e valori osservati nelle variabili.

1.1 Concetti fondamentali di dati mancanti

Le metodologie statistiche convenzionali sono generalmente costruite per esaminare dataset strutturati in modo rettangolare, in cui le righe rappresentano le osservazioni e le colonne delineano le variabili misurate su ciascuna di esse. Quando si fa riferimento alla non risposta (Barcaroli et al., 1999), si fa riferimento a situazioni in cui il valore di una variabile per un'unità specifica è:

- **Mancante**: se non è stato possibile acquisirlo;
- **Errato**: se non corrisponde al valore effettivamente associato all'unità considerata.

Quest'ultimo può essere classificato ulteriormente in tre categorie:

1. **Fuori dominio**: se si colloca al di fuori dell'intervallo accettabile di valori;
2. **Anomalo (*outlier*)**: se la risposta fornita da un'unità si discosta significativamente dalle risposte date da tutte le altre unità;

3. **Incompatibile**: se sussiste una contraddizione con i valori delle altre variabili rilevate sulla stessa unità.

La non risposta contribuisce ad aumentare la variabilità degli stimatori, a causa della riduzione della base campionaria di analisi e/o dell'applicazione di metodi di trattamento specifici. Inoltre, può condurre a stimatori distorti nel caso in cui i rispondenti differiscano in modo sistematico dai non rispondenti rispetto a particolari caratteristiche di interesse.

Le due principali forme di non risposta sono:

1. **Non risposta totale (*unit non response*)**: si verifica quando non si dispone di alcuna informazione rilevata per alcune unità campionarie, a causa di motivi quali impossibilità di contatto, non reperibilità, incapacità di rispondere, rifiuto o mancata restituzione del questionario, e così via.
2. **Non risposta parziale (*item non response*)**: si verifica quando manca la risposta a uno o più quesiti di un questionario, con motivazioni che possono essere molteplici, come la non comprensione del quesito, la percezione di domande troppo personali, il rifiuto a rispondere, ecc.

La non risposta parziale rappresenta la situazione più agevole da gestire, poiché si dispone ancora di una serie di informazioni sull'individuo in questione, che è presente nel dataset, sebbene con alcuni campi vuoti. La presenza di una non risposta porta dunque ad un dataset non rettangolare, rendendo le tradizionali analisi statistiche non direttamente applicabili.

1.2 Rilevanza della gestione dei dati mancanti

La rilevanza della gestione dei dati mancanti è cruciale nell'ambito dell'analisi statistica e della ricerca scientifica. La presenza di informazioni mancanti all'interno di un dataset può compromettere la validità e l'affidabilità delle analisi condotte, influenzando direttamente le proprietà degli stimatori e, di conseguenza, i risultati inferenziali ottenuti. Ignorare o trattare in modo inadequato i dati mancanti può portare a distorsioni significative e condurre a conclusioni errate, minando l'integrità delle indagini condotte.

La scelta di una strategia efficace per affrontare i dati mancanti diventa pertanto un aspetto chiave della gestione e dell'analisi dei dati. L'adozione di metodologie appropriate, come le tecniche di imputazione o altri approcci avanzati, diventa fondamentale per mantenere l'integrità e la rappresentatività del dataset. Inoltre, considerando la natura sempre più complessa e multidimensionale dei dati disponibili, la corretta gestione dei dati mancanti diventa ancor più critica nell'analisi di dataset di grandi dimensioni.
In questa ricerca è stata affrontata questa sfida fornendo un approfondimento sulle diverse tipologie di dati mancanti e presentando metodologie avanzate di trattamento.

In conclusione, la rilevanza della gestione dei dati mancanti in questo lavoro di tesi va oltre la semplice correzione di vuoti nel dataset; essa rappresenta un elemento fondamentale per garantire la solidità e la validità delle analisi condotte, contribuendo così alla costruzione di risultati accurati e generalizzabili nell'ambito della statistica e della ricerca scientifica.

1.3 Concetto MAR e MNAR

Risulta doveroso fornire una breve introduzione ai termini MCAR, MAR e MNAR. Rubin (1976) ha classificato i problemi di dati mancanti in tre categorie. Nella sua teoria, ogni punto dati ha una certa probabilità di essere mancante. Il processo che governa queste probabilità è chiamato meccanismo dei dati mancanti o meccanismo di risposta. Il modello per il processo è chiamato modello dei dati mancanti o modello di risposta.

Se la probabilità di essere mancanti è la stessa per tutti i casi, allora i dati sono detti mancanti completamente a caso (MCAR). Questo implica effettivamente che le cause dei dati mancanti non sono correlate ai dati. Possiamo di conseguenza ignorare molte delle complessità che sorgono perché i dati sono mancanti, oltre alla perdita ovvia di informazioni. Un esempio si ha quando viene preso un campione casuale da una popolazione dove ogni membro ha la stessa possibilità di essere incluso nel campione. I dati (non osservati) dei membri della popolazione che non sono stati inclusi nel campione sono MCAR.

Se la probabilità di essere mancanti è la stessa solo all'interno dei gruppi definiti dai dati osservati, allora i dati sono mancanti in modo casuale (MAR). MAR è una classe molto più ampia rispetto a MCAR. Un esempio di MAR si ha quando viene considerato un campione da una popolazione dove la probabilità di essere incluso dipende da una proprietà nota. MAR è più generale e più realistico di MCAR. I metodi moderni per i dati mancanti partono generalmente dall'assunzione di MAR.

Se né MCAR né MAR sono validi, allora si parla di mancanti non casualmente (MNAR). Nella letteratura si può trovare anche il termine NMAR (non mancanti casualmente) per lo stesso concetto. MNAR significa che la probabilità di essere mancanti varia per ragioni che sono anche sconosciute. MNAR è il caso più complesso. Le strategie per gestire MNAR sono di trovare più dati sulle cause della mancanza o di eseguire analisi "what-if" per vedere quanto siano sensibili i risultati in vari scenari.

La distinzione di Rubin è importante per capire perché alcuni metodi non funzioneranno ed altri si nella gestione di dati mancanti. La sua teoria stabilisce le condizioni sotto le quali un metodo per i dati mancanti può fornire inferenze statistiche valide. La maggior parte delle soluzioni semplici funziona solo sotto l'assunzione restrittiva e spesso irrealistica di MCAR. Se MCAR è improbabile, tali metodi possono fornire stime distorte.

La matrice R memorizza le posizioni dei dati mancanti in Y. La distribuzione di R può dipendere da $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, e questa relazione è descritta dal modello dei dati mancanti. Definita con Θ l'insieme contenente i parametri del modello dei dati mancanti, allora l'espressione generale del modello dei dati mancanti è $\Pr(R|Y_{\text{obs}}, Y_{\text{mis}}, \Theta)$.

I dati sono detti MCAR se

$$\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}, \Theta) = \Pr(R = 0|\Theta) \tag{1.1}$$

quindi la probabilità di essere mancanti dipende solo da alcuni parametri Θ.

I dati sono detti MAR se

$$\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}, \Theta) = \Pr(R = 0|Y_{\text{obs}}, \Theta) \tag{1.2}$$

quindi la probabilità di mancanza può dipendere dalle informazioni osservate, compresi eventuali fattori di design.

Infine, i dati sono detti MNAR se

$$\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}, \Theta) \tag{1.3}$$

non si semplifica, quindi qui la probabilità di essere mancanti dipende anche dalle informazioni non osservate, compreso Y_{mis} stesso.

Come anticipato le tecniche semplici di solito funzionano solo con MCAR, ma questa assunzione è molto restrittiva e spesso irrealistica. L'imputazione multipla può gestire sia MAR che MNAR.

Sono stati proposti diversi test per testare MCAR rispetto a MAR. Questi test non sono ampiamente utilizzati e il loro valore pratico non è chiaro. Si veda Enders (2010, pp. 17-21) per una valutazione di due procedure. Non è possibile testare MAR rispetto a MNAR poiché le informazioni necessarie per tale test mancano.

Capitolo 2

Metodi per il trattamento dei dati mancanti

La trattazione scientifica dei dataset che includono dati mancanti è una tematica di ricerca relativamente recente, caratterizzata dalla diversità di tecniche e metodologie proposte per gestire i dati mancanti. Le principali tecniche per il trattamento di dataset incompleti sono:

- Metodi di eliminazione
- Metodi di imputazione singola
- Metodi di imputazione multipla

2.1 Metodi di eliminazione

Un metodo utilizzato, soprattutto nei casi in cui si ha a disposizione un elevato numero di osservazioni, l'ammontare dei dati mancanti è limitato ed il meccanismo che li ha generati è di tipo MCAR, è cancellare le osservazioni mancanti (case deletion). I metodi per far questo sono due: *listwise deletion* e *pairwise deletion.*

2.1.1 Listwise deletion

L'approccio noto come listwise deletion, o analisi dei casi completi, rappresenta la pratica più diffusa. Essenzialmente, comporta l'eliminazione di tutte le righe, cioè le osservazioni, che presentano almeno un valore mancante tra le variabili considerate. Le sue qualità positive includono la facilità di esecuzione e la possibilità di confrontare le statistiche univariate, poiché sono calcolate sul medesimo insieme di casi. Tuttavia, questo metodo presenta svantaggi significativi derivanti dalla potenziale perdita di informazioni, manifestata attraverso le osservazioni scartate. Questa perdita può tradursi in una diminuzione di precisione e persino in distorsioni, specialmente se il meccanismo alla base dei dati mancanti non segue una distribuzione completamente casuale (MCAR), ma piuttosto una distribuzione associata al meccanismo (MAR). Pertanto, l'impiego di questa tecnica risulta giustificato solamente in situazioni in cui l'imprecisione e la distorsione sono limitate, e non possono essere attribuite esclusivamente alla proporzione di casi eliminati rispetto all'intero campione.

2.1.2 Pairwise deletion

Un'alternativa valida al metodo della listwise deletion, che comporta la perdita di informazioni anche per le variabili in cui i dati sono completi, è rappresentata dalla pairwise deletion, o analisi dei casi disponibili. Questa strategia coinvolge tutte le unità statistiche per le quali la variabile di interesse è stata registrata. Si procede alla creazione di diversi dataset, ciascuno destinato a uno specifico studio, e in ognuno vengono considerate solo le variabili rilevanti per l'analisi, eliminando successivamente i valori mancanti. Tuttavia, questa procedura presenta uno svantaggio significativo: la composizione del campione varia in base ai diversi dataset creati e alle variabili coinvolte. La variabilità nella struttura del campione introduce complessità, rendendo difficile l'utilizzo di strumenti diretti per verificare la corretta costruzione dei dataset. D'altro canto, il vantaggio della pairwise deletion risiede nella riduzione della distorsione delle stime rispetto alla listwise deletion. Questo risultato positivo, tuttavia, è ottenuto a costo di un aumento dei requisiti computazionali necessari per gestire la varietà di dataset generati.

2.2 Metodi di imputazione singola

Esplorando il complesso scenario dei dati mancanti, risulta cruciale adottare strategie appropriate per affrontare situazioni in cui la mancanza di informazioni segue un modello non completamente casuale (MAR).Diventa opportuno ricorrere a metodi di imputazione per sostituirli con funzioni adeguate dei dati effettivamente osservati. Si esaminano ora i diversi approcci per imputare i valori mancanti, che possono essere implementati per sostituire un singolo valore per ogni dato mancante (imputazione singola) o, in alcuni scenari, per imputare più di un valore al fine di valutare in modo appropriato l'incertezza dell'imputazione (imputazione multipla).

L'imputazione singola, una pratica statistica finalizzata a eliminare i valori mancanti all'interno di un dataset, si basa sulla sostituzione di tali valori con altri validi per la variabile considerata, al fine di ripristinare la completezza della matrice dei dati. Questo metodo risulta interessante in quanto contribuisce a ridurre la perdita di informazioni; tuttavia, è necessario affrontare con cautela, poiché, come indicato da Dempster e Rubin (1983), presenta rischi significativi. Sebbene, una volta sostituiti i valori mancanti, consenta di trattare il dataset come completo, semplificando l'analisi e la presentazione dei risultati, il pericolo principale risiede nella tendenza a considerare, nelle successive analisi, i dati imputati come osservazioni reali. Tale approccio non tiene conto dell'incertezza derivante dall'ignoranza circa il vero valore delle variabili in cui manca l'informazione, con conseguente riduzione della variabilità complessiva.

I principali metodi includono:

- **Imputazione con la media:** Sostituzione del dato mancante con la media della variabile corrispondente, calcolata sull'intero set di risposte osservate. Questo metodo, sebbene semplice, può introdurre distorsioni nella distribuzione della variabile, generando un picco artificiale e sottostimando la variabilità.
- **Campionamento aleatorio:** Sostituzione del dato mancante con un valore estratto casualmente tra quelli disponibili per la variabile di riferimento.
- **Imputazione con regressione:** Questo approccio si basa sulle informazioni disponibili per altre variabili, con l'uso di modelli di regressione per variabili quantitative o modelli log-lineari o logistici per variabili qualitative. L'imputazione avviene prevedendo i valori mancanti sulla base della stima dell'equazione di regressione. Se da una parte questo metodo può ridurre le distorsioni causate dalle

mancate risposte, dall'altra parte può introdurre alcune problematiche, come la distorsione nelle relazioni tra variabili non utilizzate nel modello e il rischio di imputare valori non plausibili.

A questo punto, è fondamentale sottolineare l'importanza di selezionare attentamente il metodo di imputazione in base al contesto specifico del dataset e alla tipologia di dati mancanti. Questa decisione richiede una comprensione approfondita delle relazioni tra le variabili coinvolte e delle possibili fonti di distorsione nei risultati. Allo stesso tempo, occorre considerare che l'imputazione è solo una delle molte strategie per gestire dati mancanti, e la scelta tra diverse metodologie dipende dall'obiettivo dell'analisi e dalle caratteristiche specifiche del dataset in esame. In tal senso, un approccio ponderato e informato alla gestione dei dati mancanti può contribuire in modo significativo alla validità e all'affidabilità delle analisi statistiche condotte.

2.3 Metodi di imputazione multipla

L'imputazione multipla è una strategia avanzata per gestire i dati mancanti, che va oltre l'approccio dell'imputazione singola. La sua peculiarità risiede nel fatto che, oltre a sostituire i dati mancanti con valori plausibili, tiene anche conto dell'incertezza associata alla generazione di tali imputazioni. Questo aspetto rende i risultati inferenziali non solo validi in termini di stima puntuale dei parametri di interesse, ma offre anche la possibilità di costruire intervalli di confidenza che riflettono l'incertezza introdotta dalla presenza di dati mancanti.

L'idea di generare più di un valore ($m > 2$) per ogni dato mancante è stata introdotta da Rubin nel 1978. Questo approccio consente di creare m dataset completi, su ciascuno dei quali vengono eseguite le analisi di interesse. La diversità tra gli m dataset riflette l'incertezza intrinseca alla natura mancante dei dati.

Nella fase successiva, i risultati derivanti dalle analisi condotte separatamente su ciascuno degli m dataset vengono combinati attraverso regole predeterminate. Questo processo di combinazione mira a produrre un risultato inferenziale finale che tiene conto dell'incertezza causata dalla mancanza di dati, valutata attraverso la variabilità tra le diverse uscite indipendenti.

Va sottolineato, tuttavia, che l'imputazione multipla comporta un costo computazionale significativo. Dopo l'imputazione dei dati, è necessario infatti ripetere le analisi

m volte e combinare i risultati per ottenere stime valide e affidabili. Nonostante questo svantaggio, l'imputazione multipla si configura attualmente come la metodologia più potente nell'affrontare la sfida dei dati mancanti in ambito statistico, offrendo una prospettiva più completa e robusta rispetto a molte altre tecniche.

Il processo di imputazione multipla può essere sintetizzato in tre fasi principali:

- Fase di imputazione: creazione di m dataset completi sostituendo i valori mancanti
- Fase di analisi dei m dataset creati
- Fase di pooling: unificazione delle diverse stime ottenute per ottenere un unico valore di stima per ciascun parametro di interesse.

2.3.1 Fase di imputazione

In questa fase si procede con l'imputazione dei dati mancanti, utilizzando le varie tecniche impiegate per l'imputazione singola. Questo processo viene ripetuto m volte al fine di generare m dataset. Tale approccio consentirà poi di condurre analisi sullo stesso insieme di dati con lievi variazioni, ottenendo stime diverse dei parametri di interesse.

La scelta dei metodi dipende dalla natura della variabile da imputare. Ad esempio, per variabili binarie, la tecnica più comune coinvolge l'utilizzo della regressione logistica, mentre per le variabili continue si ricorrono spesso ai metodi precedentemente descritti per l'imputazione singola: imputazione con la media, campionamento aleatorio, imputazione con regressione lineare, il metodo PMM, lasso o metodi che utilizzino campioni bootstrap.

2.3.2 Fase di analisi e fase di pooling

Una volta ottenuti gli m dataset è possibile applicare su ciascuno di essi le opportune analisi statistiche. Supponendo che θ sia un parametro incognito di interesse, al termine della procedura di inferenza sugli m dataset sono disponibili m coppie di valori composte dalla stima puntuale del parametro di interesse $\hat{\theta}_i$ e dalla stima della varianza dello stimatore, $\hat{U}_i$ (con $i = 1, \ldots, m$).

La stima puntuale di θ è data dalla media delle singole stime calcolate sui m dataset completi:

$$\bar{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i \tag{2.1}$$

La varianza della stima puntuale è invece data dalla somma di una componente di variabilità entro l'imputazione ($\bar{U}$) e da una componente di variabilità tra imputazioni (B):

$$\bar{U} = \frac{1}{m} \sum_{i=1}^{m} \hat{U}_i \tag{2.2}$$

La varianza tra imputazioni B è calcolata come segue:

$$B = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}_i - \bar{\theta})^2 \tag{2.3}$$

La varianza totale associata a $\bar{\theta}$ è quindi ottenuta combinando le due componenti (Rubin, 1987):

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B = \bar{U} + B + \frac{B}{m} \tag{2.4}$$

dove il terzo termine tiene conto del fatto che il numero di imputazioni è finito: può essere considerato come una sorta di errore di simulazione e dunque non è presente se il numero di simulazioni è molto grande.

Capitolo 3

Metodi di imputazione singola

L'imputazione singola è un metodo statistico per eliminare i valori mancanti all'interno di un dataset, sostituendoli con valori ammissibili per la variabile considerata, al fine di ripristinare la completezza della matrice dei dati. È un'opzione interessante poiché riduce la perdita di informazioni, consentendo al contempo un'analisi dei dati facilitata e una presentazione immediata dei risultati. Allo stesso tempo però, presenta il rischio di considerare i dati imputati come osservati, ignorando l'incertezza associata al vero valore delle variabili in cui mancano dati, soprattutto se questi dati sono MCAR

Per sostituire i valori mancanti, è necessario definire criteri di imputazione, e ci sono diverse tecniche disponibili, ciascuna con risultati differenti. Le principali sono:

- **Imputazione con la media**: sostituisce il dato mancante con la media della variabile, calcolata sull'intero campione. Questo metodo è limitato alle variabili quantitative e può introdurre distorsioni nella distribuzione della variabile, oltre a sottostimare la variabilità.
- **Campionamento casuale**: sostituisce il dato mancante con un valore estratto casualmente dai dati disponibili per la stessa variabile.
- **Imputazione con regressione**: si basa sulle informazioni disponibili dalle altre variabili per predire i valori mancanti dopo la stima di un modello di regressione. Questo approccio preserva meglio le relazioni tra le variabili rispetto ad altri metodi, ma può introdurre distorsioni nella distribuzione della variabile e nelle relazioni tra variabili non utilizzate nel modello.

- **Nearest Neighbor Imputation**: sostituisce il valore mancante con il valore di un'unità simile e completa del campione, utilizzando una funzione di distanza appropriata per valutare la "vicinanza" delle unità. Le varianti includono, tra le altre, distanza Euclidea, distanza ponderata, distanza di Mahalanobis e distanza Minmax.

La presente tesi si propone di esaminare metodologie e approcci che, sebbene non siano esclusivamente destinati all'imputazione di dati mancanti, possono trovare applicazione in questo ambito. Attraverso un'analisi approfondita, verranno esplorate tecniche che, pur avendo originariamente differenti finalità, si sono dimostrate utili e efficaci nel trattamento dei dati mancanti. Questo studio si concentrerà sull'adattamento e l'applicazione di tali metodologie al fine di migliorare la qualità e l'affidabilità delle analisi condotte sui dati contenenti valori mancanti. La ricerca si propone di esaminare come queste tecniche possano essere integrate nei processi di trattamento dei dati mancanti, contribuendo così a una maggiore completezza e accuratezza delle analisi statistiche e dei modelli predittivi.

3.1 Modello lineare

Per studiare la relazione tra due variabili, un primo punto di partenza può essrere rappresentato da un tradizionale modello del tipo

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{3.1}$$

dove y rappresenta la variabile risposta, x una variabile dipendente, e ε è un termine di "errore" casuale non osservabile, di media zero e varianza costante σ^2. Ipotizziamo inoltre l'assenza di correlazione tra i termini di errore.

Si vuole ottenere una stima dei parametri di regressione sconosciuti β_0 e β_1 usando n coppie di osservazioni, indicate da (x_i, y_i), per $i = 1, \ldots, n$. L'equazione (3.1) è il caso più semplice per una formulazione più generale del tipo

$$y = f(x; \beta) + \varepsilon \tag{3.2}$$

che diventa (3.1) quando f è l'espressione della retta e $\beta = (\beta_0, \beta_1)$.

Per stimare β, il criterio dei minimi quadrati ci porta a identificare i valori per cui otteniamo il minimo, rispetto a β, della funzione obiettivo

$$D(\beta) = \sum_{i=1}^{n} (y_i - f(x_i; \beta))^2 = \|y - f(x; \beta)\|^2 \tag{3.3}$$

dove l'ultima espressione utilizza la notazione matriciale per rappresentare il vettore $y = (y_1, \ldots, y_n)$, $f(x; \beta) = (f(x_1; \beta), \ldots, f(x_n; \beta))$, e $\| \cdot \|$ indica la norma euclidea del vettore, cioè la radice quadrata della somma dei quadrati degli elementi.

La soluzione a questo problema di minimizzazione è indicata da $\hat{\beta}$, e indichiamo i corrispondenti valori adattati

$$\hat{y}_i = f(x_i; \hat{\beta}), \quad i = 1, \ldots, n \tag{3.4}$$

che, nel caso lineare (3.1), sono del tipo

$$haty_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \tilde{x}_i \hat{\beta} \tag{3.5}$$

dove $\tilde{x}_i = (1, x_i)$.

Dalla stessa formula, possiamo anche scrivere l'espressione del valore predetto

$$haty_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{3.6}$$

per un valore x_0 della variabile esplicativa, che non corrisponde necessariamente a nessuna osservazione.

Tuttavia, quando il trend della relaione tra variabili non si presta a essere espresso da una retta, possiamo muoverci in diverse direzioni alternative. La più immediata è probabilmente considerare una forma più elaborata di funzione $f(x; \beta)$, ad esempio una forma polinomiale

$$f(x; \beta) = \beta_0 + \beta_1 x + \ldots + \beta_{p-1} x^{p-1} \tag{3.7}$$

dove β è ora un vettore con p componenti, $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$. Utilizzare una funzione polinomiale ha il doppio vantaggio di essere concettualmente e matematicamente semplice e di offrire un trattamento semplice per quanto riguarda l'uso del criterio dei minimi quadrati.

Poiché (3.4) è lineare nei parametri, può essere riscritta come

$$f(x;\beta) = X\beta \tag{3.8}$$

dove X è una matrice $n \times p$, chiamata matrice di progetto, definita da

$$X = \begin{bmatrix} 1 & x & \ldots & x^{p-1} \end{bmatrix}$$

dove x è il vettore delle osservazioni della variabile esplicativa, e le varie colonne di X contengono potenze di ordine da 0 a $p-1$ degli elementi di x.

In questa formulazione, la soluzione esplicita al problema di minimizzazione di (3.3) è

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3.9}$$

con il quale il vettore dei valori adattati è

$$haty = X\hat{\beta} = Py \tag{3.10}$$

dove

$$P = X(X^T X)^{-1} X^T \tag{3.11}$$

è una matrice $n \times n$, chiamata matrice di proiezione. Le proprietà $P^T = P$, $P^2 = P$ e $\mathrm{tr}(P) = \mathrm{rk}(P) = p$ sono verificate.

Il valore minimo di (3.3) può essere scritto in varie forme equivalenti

$$D(\hat{\beta}) = \|y - \hat{y}\|^2 = y^T(I_n - P)y = \|y\|^2 - \|\hat{y}\|^2 \tag{3.12}$$

dove I_n denota la matrice identità di ordine n. La quantità $D = D(\hat{\beta})$ è chiamata devianza, in quanto è una quantificazione della discrepanza tra i valori adattati e osservati.

Da qui, otteniamo anche la stima di σ^2, di solito data da

$$s^2 = \frac{D(\hat{\beta})}{n-p} \tag{3.13}$$

e questo ci permette di valutare la varianza delle stime di β attraverso

$$\text{var}(\hat{\beta}) = s^2(X^T X)^{-1} \tag{3.14}$$

La radice quadrata degli elementi diagonali fornisce gli errori standard delle componenti di $\hat{\beta}$ - essenziali per i procedimenti inferenziali.

Per valutare la bontà dell'adattamento, è necessario calcolare il coefficiente di determinazione

$$R^2 = 1 - \frac{\text{devianza residua}}{\text{devianza totale}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{3.15}$$

dove $D(\hat{\beta})$ è calcolato usando X, la matrice corrispondente al modello; e $\bar{y} = \sum_i y_i/n$ indica la media aritmetica o media di y. Tuttavia, non possiamo ridurre la valutazione dell'adeguatezza di un modello alla considerazione di un singolo indicatore. Altre indicazioni sono fornite dalla diagnostica grafica. Ce ne sono diverse, e tutte ci riportano più o meno esplicitamente all'esame del comportamento dei residui

$$hat\varepsilon_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n \tag{3.16}$$

che servono come surrogati degli errori ε_i, che non sono osservabili. I residui hanno vari aspetti che dobbiamo valutare in base a varie ipotesi.

3.2 Metodi di regolazione

Quando un gran numero di covariate è disponibile, le stime dei minimi quadrati di un modello lineare spesso presentano una bassa deviazione ma una alta varianza rispetto ai modelli con un minor numero di variabili. I metodi di selezione delle variabili e riduzione delle dimensioni possono aiutare a migliorare l'accuratezza della previsione consentendo una deviazione maggiore ma una varianza minore. Tuttavia, questi metodi possono essere poco attraenti per motivi di onere computazionale (selezione delle variabili) o interpretazione (riduzione delle dimensioni). Un approccio diverso è modificare il metodo di stima abbandonando il requisito di uno stimatore non distorto dei parametri e invece considerare la possibilità di utilizzare uno stimatore distorto, che potrebbe avere una varianza più piccola. Esistono diversi stimatori di questo tipo, per lo più basati sulla regolarizzazione: tutte le variabili vengono mantenute nel modello, ma quando il modello viene adattato, i loro coefficienti si restringono. L'idea è ottenere un restringimento verso la media, in modo che di solito l'intercetta non venga penalizzata. È possibile quindi operare in due fasi: prima si ottiene la media di y come stima per l'intercetta; poi viene sostituito ogni valore y_i con $y_i - \bar{y}$, e gli x_{ij} con le variabili centrate $x_{ij} - \bar{x}_j$ (per $j = 1, \ldots, p-1$).

La regressione ridge è probabilmente il metodo di restringimento più comune. Consideriamo un modello lineare $y = X\beta + \varepsilon$, per il quale i coefficienti di regressione ridge minimizzano una forma vincolata del tipo:

$$\sum_{i=1}^{n} (y_i - x_i\beta)^2 \quad \text{con vincolo} \quad \sum_{j=1}^{p-1} \beta_j^2 \leq s \tag{3.17}$$

Un'altra formulazione equivalente può essere ottenuta con la forma di Lagrange, in modo che i coefficienti di regressione ridge minimizzino la somma residua penalizzata dei quadrati

$$D_{\text{ridge}}(\beta, \lambda) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2 = \|y - X\beta\|^2 + \lambda\beta^T\beta \tag{3.18}$$

dove λ è determinato univocamente da s che rappresenta un iperparametro che regola il livello di regolarizzazione applicato al modello. La soluzione è $\hat{\beta}_\lambda = (X^TX + \lambda I)^{-1}X^Ty$, dove I è la matrice identità. Lo stimatore $\hat{\beta}_\lambda$ è distorto, ma per alcuni valori di $\lambda > 0$ può avere un errore quadratico medio più piccolo rispetto all'estimatore dei minimi quadrati.

Si noti che $\lambda = 0$ dà lo stimatore dei minimi quadrati e, se $\lambda \to \infty$, allora $\hat{\beta} \to 0$. La regressione ridge è particolarmente utile quando le variabili esplicative sono collineari, poiché anche un piccolo $\lambda > 0$ rende la soluzione $\hat{\beta}_\lambda$ numericamente e statisticamente stabile. Il parametro λ dovrebbe essere scelto in modo adattivo, ad esempio, tramite la cross-validazione.

La regressione ridge ha una semplice interpretazione geometrica secondo la PCA, poiché proietta la variabile di risposta y sui componenti principali e poi restringe i coefficienti dei componenti a bassa varianza più di quelli ad alta varianza. È, infatti, spesso (sebbene non sempre) ragionevole aspettarsi che la variabile risposta varierà di più nella direzione delle variabili esplicative con alta varianza. Pertanto, rispetto alla trasformazione delle componenti principali delle variabili esplicative, la regressione ridge restringe i coefficienti delle componenti principali, applicando relativamente più restringimento alle componenti più piccole che a quelle più grandi, mentre la regressione delle componenti principali scarta le componenti con autovalori più piccoli (Hastie et al., 2009).

La scelta di una penalità alternativa da aggiungere alla somma dei quadrati può fornire un metodo di restringimento che, oltre alla restrizione dei parametri, richiede che alcuni coefficienti siano nulli. Quando il vincolo quadratico è sostituito dal vincolo del valore assoluto $\sum_{j=1}^{p-1} |\beta_j| \leq s$ e viene scelto un s sufficientemente piccolo, la minimizzazione vincolata della somma dei quadrati imposta alcuni coefficienti esattamente a zero, eseguendo una sorta di selezione del modello continua. Questo metodo di restringimento è chiamato lasso e minimizza

$$\sum_{i=1}^{n} (y_i - x_i\beta)^2 \quad \text{con vincolo} \quad \sum_{j=1}^{p-1} |\beta_j| \leq s \tag{3.19}$$

o, in forma di Lagrange:

$$D_{\text{lasso}}(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p-1} |\beta_j| = \|y - X\beta\|^2 + \lambda \sum_{j=1}^{p-1} |\beta_j|. \tag{3.20}$$

Le soluzioni sono non lineari in y a causa della natura del vincolo. Come per la regressione ridge, il parametro di regolarizzazione s (o λ) dovrebbe essere scelto in modo adattivo

Quando si confronta le stime dei coefficienti ottenute mediante regressione ridge e lasso, è possibile osservare che i coefficienti della regressione ridge sono ottenuti dalla moltiplicazione dei coefficienti dei minimi quadrati per una costante tra 0 e 1, mentre il lasso li trasla verso 0 per una costante, come mostra la Figura 3.1.

Le caratteristiche interessanti del lasso sono controbilanciate dal complicato algoritmo di stima necessario per stimare i coefficienti. Negli ultimi anni sono stati proposti diversi algoritmi più veloci. L'algoritmo più utilizzato ed efficiente si basa sulla regressione ad angolo minimo (LAR), una modifica dell'algoritmo di Gram-Schmidt per stimare i coefficienti dei minimi quadrati mediante ortogonalizzazione successiva

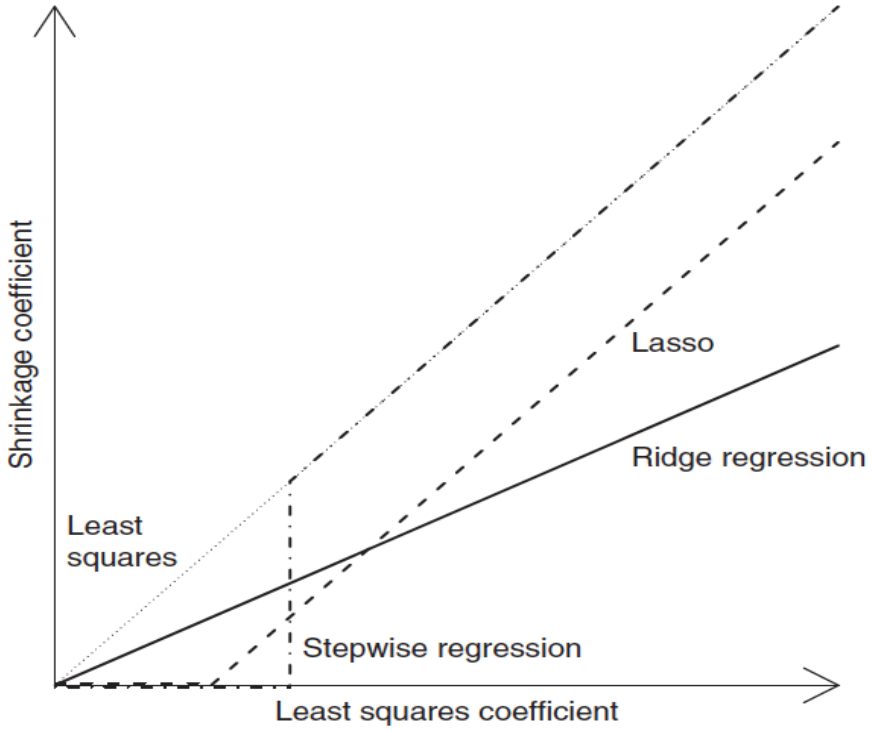

FIGURA 3.1: Coefficiente trasformato rispetto al coefficiente della regressione lineare per la regressione ridge, lasso e regressione stepwise nel caso ortogonale

La regressione stepwise in avanti aggiunge una variabile alla volta al modello identificando la variabile da includere in quel modello ad ogni passo. LAR utilizza una strategia simile, ma aggiunge al modello solo quella parte di informazione inclusa in una variabile che è necessaria. LAR inizia aggiungendo al modello la variabile più correlata con la risposta e, anziché adattare questa variabile mediante minimi quadrati, sceglie il coefficiente muovendo il suo valore continuamente tra 0 e il valore dei minimi quadrati. Man mano che il coefficiente stimato si sposta tra di essi, la correlazione tra la variabile

e i residui diminuisce in valore assoluto. A un certo punto in questa evoluzione del primo coefficiente, la correlazione tra la variabile e i residui diventa uguale alla correlazione tra un'altra variabile e gli stessi residui. Questa seconda variabile viene quindi inclusa nel modello, e il suo coefficiente viene scelto insieme al primo, spostandoli nella direzione del loro coefficiente dei minimi quadrati, fino a quando un'altra variabile avrà la stessa correlazione con i residui correnti. Il processo continua fino a quando tutte le variabili sono incluse nel modello e otteniamo i coefficienti dei minimi quadrati.

L'aspetto interessante del LAR è la sua semplice relazione con il lasso: una modifica dell'algoritmo può generare l'intero percorso della sequenza di stima. Infatti, è sufficiente aggiungere un nuovo passaggio all'algoritmo indicando che se un coefficiente diverso da zero incrocia lo 0, la variabile corrispondente deve essere rimossa dal modello. La migliore direzione congiunta dei minimi quadrati è quindi ricalcolata, richiedendo all'algoritmo di ripartire da questa nuova migliore direzione. Chiaramente, il numero di passaggi nell'algoritmo LAR modificato dal lasso (chiamato LARS) potrebbe essere maggiore rispetto a quello dell'algoritmo LAR stesso, ma l'ordine di grandezza dei calcoli rimane lo stesso.

3.3 MARS

Quando il numero di covariate è elevato è importante utilizzare un metodo che, partendo dalle informazioni presenti nei dati, ci consenta di selezionare variabili in modo ragionevole e fornisca criteri per la scelta del numero di nodi necessari per ciascuna variabile.

Le regressioni spline adattive multivariate (MARS) rappresentano una specifica iterativa particolare delle spline di regressione, il cui obiettivo è modellare problemi con molte variabili esplicative. Le funzioni di base utilizzate sono coppie di funzioni lineari a tratti, del tipo $(x-\xi)_+$ e $(\xi-x)_+$, con un solo nodo al punto ξ

L'obiettivo è trovare la relazione tra una variabile dipendente y e le p covariate $x=(x_1,\ldots,x_p)^T$. Per ogni variabile esplicativa x_j, viene determinato una coppia di funzioni di base con il nodo in ogni valore osservato x_{ij}, per $i=1,\ldots,n$ oltre a quella lineare. Ciò fornisce l'insieme di funzioni di base che sono considerate come funzioni sull'intero spazio $\mathbb{R}^p$:

$$C=\{x_j,(x_j-\xi)_+,(\xi-x_j)_+ : \xi\in\{x_{i1},x_{i2},\ldots,x_{ip}\},i=1,2,\ldots,n,j=1,\ldots,p\}.$$

Verrà quindi selezionato un sottoinsieme di funzioni di base in C da combinare in un modello appropriato per adattarsi ai dati. Le funzioni di base a tratti sono incluse nel modello a coppie del tipo $\{(x_j - \xi)_+, (\xi - x_j)_+\}$. Il modello MARS è quindi del tipo:

$$f(x) = \beta_0 + \sum_{k=1}^{2K} \beta_k h_k(x) \tag{3.21}$$

dove $h_k(x)$ sono funzioni appartenenti a C o prodotti di due o più di tali funzioni, e K è il numero di coppie di funzioni di base da includere nel modello.

Per selezionare le funzioni h_k e stimare i parametri β, seguiamo un processo ricorsivo.

- Si parte con $K = 0$. E viene innanzitutto introdotta prima la funzione costante $h_0(x) = 1$.
- Passo generico K. Si ipotizza che il modello abbia già $2(K-1)$ termini. Consideriamo, come nuova coppia di funzioni di base, ciascuna delle possibili coppie di prodotti di una funzione h_k, $k \in \{1, \ldots, K\}$, già inclusa nel modello, con un'altra coppia di funzioni in C. Verrà quindi scelta la coppia di funzioni di base che aggiunge i termini

  $$\hat{\beta}_{2K-1} h_m(x)(x_j - \xi)_+ + \hat{\beta}_{2K} h_m(x)(\xi - x_j)_+$$

  che minimizzano il criterio dei minimi quadrati. Qui, h_m indica una funzione già inclusa nel modello, e $\hat{\beta}_{2K-1}$ e $\hat{\beta}_{2K}$ sono parametri stimati per i minimi quadrati insieme a tutti gli altri parametri β del modello.
- Il processo continua fino a quando si raggiunge un K massimo predefinito.

Questo modello è generalmente molto grande e può sovradattare i dati. Può essere opportuno formulare una procedura inversa in cui selezionare e rimuovere iterativamente i termini dal modello uno per uno, eliminando ad ogni passo i termini che contribuiscono in modo marginale alla somma dei quadrati dei residui. In questa procedura inversa, di solito vengono eliminati singoli termini, quindi il modello finale non è necessariamente caratterizzato da una coppia di funzioni di base per ogni nodo.

I sottoinsiemi del modello vengono quindi confrontati mediante alcuni criteri di adattamento. Quando sono disponibili molti dati, verrà scelto il miglior sottoinsieme di

modelli utilizzando un diverso set di dati come test. In alternativa, è possibile utilizzare la cross-validazione, che tuttavia richiede un notevole carico computazionale.

Un'altra alternativa è utilizzare la cross-validazione generalizzata (GCV). Per ciascun modello da confrontare, il GCV è definito come

$$GCV = \sum_{i=1}^{n} \frac{(y_i - \hat{f}(x_i))^2}{(1 - d/n)^2} \tag{3.22}$$

dove d è un indicatore del numero effettivo di parametri nel modello. Per il contesto di MARS, d è la somma del numero di termini nel modello e il numero di nodi definiti nel processo di selezione delle basi pesato da una penalità che, dopo alcuni risultati teorici e di simulazione, di solito viene fissata a 2 o 3. Un'altra approssimazione frequentemente utilizzata sceglie d proporzionale al numero di termini nel modello. Si noti che la formula utilizzata da GCV approssima l'errore che sarebbe determinato dalla cross-validazione con un solo dato escluso per un modello lineare: ecco perché è chiamata cross-validazione generalizzata.

Le coppie di funzioni lineari scelte come funzioni di base per MARS hanno il vantaggio di operare localmente. Quando queste funzioni di base vengono moltiplicate insieme, sono diverse da zero solo nella parte dello spazio in cui tutte le funzioni univariate sono positive, e ciò consente al modello di adattarsi ai dati con un numero relativamente piccolo di parametri. Queste funzioni hanno anche il vantaggio che possono essere moltiplicate insieme in modo semplice, con una complessità computazionale notevolmente ridotta.

La logica costruttiva del modello è chiaramente gerarchica, nel senso che è possibile moltiplicare nuove funzioni di base che coinvolgono nuove variabili solo alle funzioni di base già nel modello; quindi, un'interazione di un ordine superiore può essere introdotta solo quando sono presenti interazioni di un ordine inferiore. Questo vincolo, introdotto per motivi computazionali, non riflette necessariamente il comportamento reale dei dati, ma spesso aiuta nell'interpretazione dei risultati. Tuttavia, per una interpretazione più facile, spesso viene vincolato il modello ad avere solo interazioni di primo o al massimo secondo ordine.

3.4 GAM

Le soluzioni fino ad ora descritte consentono di esaminare la relazione tra una variabile di risposta y e un certo numero p di variabili esplicative. Tutte queste tecniche sono valide per lo scopo, ma si scontrano anche con gli stessi problemi quando p è alto: la maledizione della dimensionalità, che si presenta quando ci si trova ad affrontare un aumento delle dimensioni delle variabili esplicative in un contesto di regressione non parametrica. In pratica, ciò comporta una rapida dispersione dei dati nello spazio, rendendo difficile ottenere stime accurate della funzione desiderata a causa della scarsità di punti osservati rispetto alla vastità dello spazio delle covariate. Per superare questa difficoltà, si può ricorrere a strategie di riduzione delle dimensioni, come l'estrazione delle componenti principali, che mirano a conservare il massimo contenuto informativo possibile riducendo il numero di variabili coinvolte. Questo approccio non solo contribuisce a ridurre la complessità computazionale, ma anche a migliorare la comprensione e l'interpretazione dei risultati ottenuti.

Per superare questo ostacolo, da un lato è possibile introdurre una qualche forma di "struttura", cioè un modello della forma della funzione di regressione
$f(x)$, $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$. Dall'altro lato, si possono evitare delle strutture rigide che tuttavia mantengono un'ampia flessibilità.

Una opzione che è stata molto apprezzata per la sua utilità pratica e la sua semplicità logica è la seguente. Sia data una rappresentazione del tipo

$$f(x) = f(x_1, \ldots, x_p) = \beta_0 + \sum_{j=1}^{p} f_j(x_j) \tag{3.23}$$

valida per $f(x)$, dove $f_1, \ldots, f_p$ sono funzioni di una variabile, ognuna con un comportamento liscio, e β_0 è una costante.

Si noti che per evitare ciò che è essenzialmente un problema di identificabilità del modello, è necessario che le varie f_j siano centrate attorno a 0, cioè

$$\frac{1}{n} \sum_{i=1}^{n} f_j(x_{ij}) = 0, \quad (j = 1, \ldots, p), \tag{3.24}$$

dove x_{ij} è la j-esima variabile per l'unità i.

Per adattare il modello ai dati, esiste un processo iterativo basato su un metodo di stima non parametrico di funzioni univariate per stimare f_j. Questa procedura, è chiamata backfitting ed è essenzialmente una variazione dell'algoritmo di Gauss-Seidel.

Algoritmo 3.1 Backfitting

1. Inizio: $\hat{\beta}_0 \leftarrow \frac{1}{n}\sum_i y_i/n$, $\hat{f}_j \leftarrow 0$ per tutti j.
2. Ciclo per $j = 1, 2, \ldots, p, 1, 2, \ldots, p, 1, 2, \ldots$:
   (a) $\hat{f}_j \leftarrow S\left[\frac{1}{n}\sum_i \left(y_i - \hat{\beta}_0 - \sum_{k=j} \hat{f}_k(x_{ik})\right)\right]$,
   (b) $\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{n-1}\sum_{i=1}^{n} \hat{f}_j(x_{ij})$, fino a quando le funzioni $\hat{f}_j$ si stabilizzano.

Il metodo specifico per la stima non parametrica non è cruciale, si possono anche scegliere diversi metodi per diversi f_j, ma di solito solo uno viene applicato (indicato genericamente con S nell'algoritmo (3.1), nel senso che $S(y)$ costituisce la stima non parametrica, calcolata sui valori osservati $y = (y_1, \ldots, y_n)$, di una funzione scalare). In molti casi, S è uno stimatore lineare, del tipo Sy, dove S è una matrice di smoothing appropriata.

Una generalizzazione del modello è del tipo

$$f(x_1, \ldots, x_p) = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \sum_{j=1}^{p}\sum_{k<j} f_{kj}(x_k, x_j) + \sum_{j=1}^{p}\sum_{k<j}\sum_{h<k<j} f_{hkj}(x_h, x_k, x_j) + \ldots \tag{3.25}$$

che consente di tenere presente l'effetto di interazione tra coppie di variabili, triplette o altre interazioni di ordine superiore.

Un'altra direzione in cui il modello (3.17) è frequentemente generalizzato è del tipo

$$g\left(\mathbb{E}\left[Y \mid x_1, \ldots, x_p\right]\right) = \beta_0 + \sum_{j=1}^{p} f_j(x_j) \tag{3.26}$$

e viene chiamato modello additivo generalizzato (GAM). Come nel GLM standard, la funzione di collegamento g deve essere specificata. Ad esempio, nel caso di Y binomiale, si presume comunemente che g sia la funzione logit. Invece, il termine sul lato destro

è ora espresso da una forma additiva, e di conseguenza il contributo della variabile generale x_j non è più lineare $\beta_j x_j$ ma è del tipo più generale $f_j(x_j)$.

Per stimare le funzioni per un modello di tipo GAM, è possibile utilizzare una combinazione adeguata dell'algoritmo (3.1) con quello delle iterazioni dei minimi quadrati pesati, applicato nel caso del GLM.

3.5 Alberi decisionali e mistura di modelli

Negli ultimi decenni, gli algoritmi di machine learning hanno rivoluzionato il modo in cui affrontiamo i problemi di previsione e classificazione. Uno di questi algoritmi, noto come albero di regressione, si è dimostrato particolarmente efficace nel modellare relazioni complesse tra variabili indipendenti e dipendenti.

In questa tesi, verrà esplorato il concetto di alberi di regressione, il loro funzionamento, le loro applicazioni pratiche e le loro limitazioni. Inoltre, si esamineranno le tecniche avanzate di regolarizzazione e ottimizzazione per migliorare le prestazioni degli alberi di regressione e si discuterà delle migliori pratiche per l'uso di questi modelli in contesti reali.

Gli alberi di regressione sono una forma di modello di apprendimento supervisionato che, a differenza di altri approcci, suddivide iterativamente lo spazio delle caratteristiche in sottoinsiemi più piccoli e più semplici. Questo processo di suddivisione si basa su regole decisionali che mirano a ridurre l'errore di previsione al minimo. Una volta costruito, l'albero di regressione può essere utilizzato per stimare il valore di una variabile dipendente per nuove osservazioni, in base alle loro caratteristiche.

L'appeal degli alberi di regressione risiede nella loro capacità di gestire dati complessi e non lineari senza richiedere molte ipotesi sulla struttura dei dati. Inoltre, sono facili da interpretare, consentendo agli utenti di comprendere il processo decisionale sottostante e di identificare le variabili importanti per la previsione.

3.5.1 Alberi di regressione semplici

In un certo senso, il modo più semplice per approssimare una funzione generica $y = f(x)$, con $x \in \mathbb{R}$, è utilizzare una funzione a gradini, ovvero una funzione costante per tratti (Figura 3.2).

Tuttavia, ci sono varie scelte da fare: quanti sottoinsiemi dell'asse x devono essere considerati? Dove devono essere posizionati i punti di suddivisione? E quale valore di y deve essere assegnato ad ogni intervallo?

Di queste domande, la più facile da rispondere è l'ultima, perché è completamente naturale scegliere il valore

$$\frac{\int_{R_j} f(x)\,dx}{|R_j|} \tag{3.27}$$

per ogni intervallo R_j, avendo indicato la lunghezza di tale intervallo con $|R_j|$. Per quanto riguarda il posizionamento dei punti di suddivisione di R, e quindi la definizione degli intervalli, è meglio scegliere piccoli intervalli dove $f(x)$ è più ripida. La scelta del numero di suddivisioni è il punto più soggettivo dei tre: intuitivamente, qualsiasi aumento del numero di passaggi aumenta la qualità dell'approssimazione, e quindi, in un certo senso, siamo portati a pensare a suddivisioni infinite. Tuttavia, ciò contrasta con il requisito di utilizzare una rappresentazione approssimativa "parsimoniosa", e quindi di adottare un numero finito di suddivisioni.

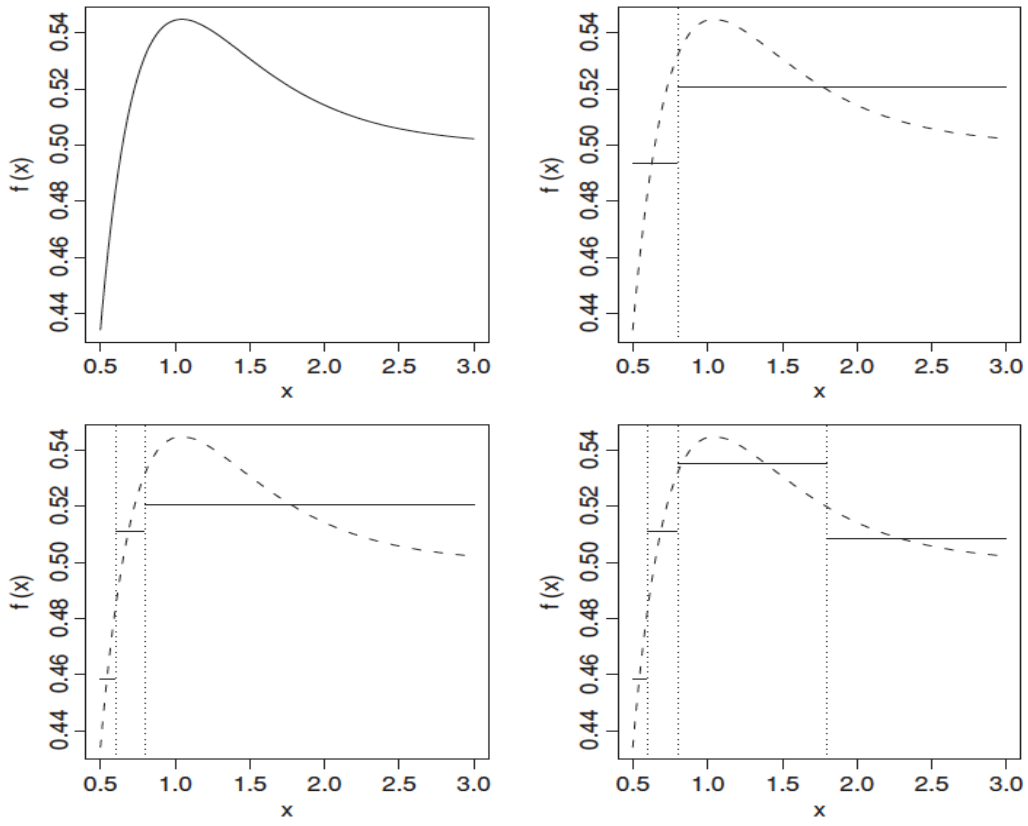

FIGURA 3.2: Una funzione continua e alcune approssimazione tramite funzioni a gradini

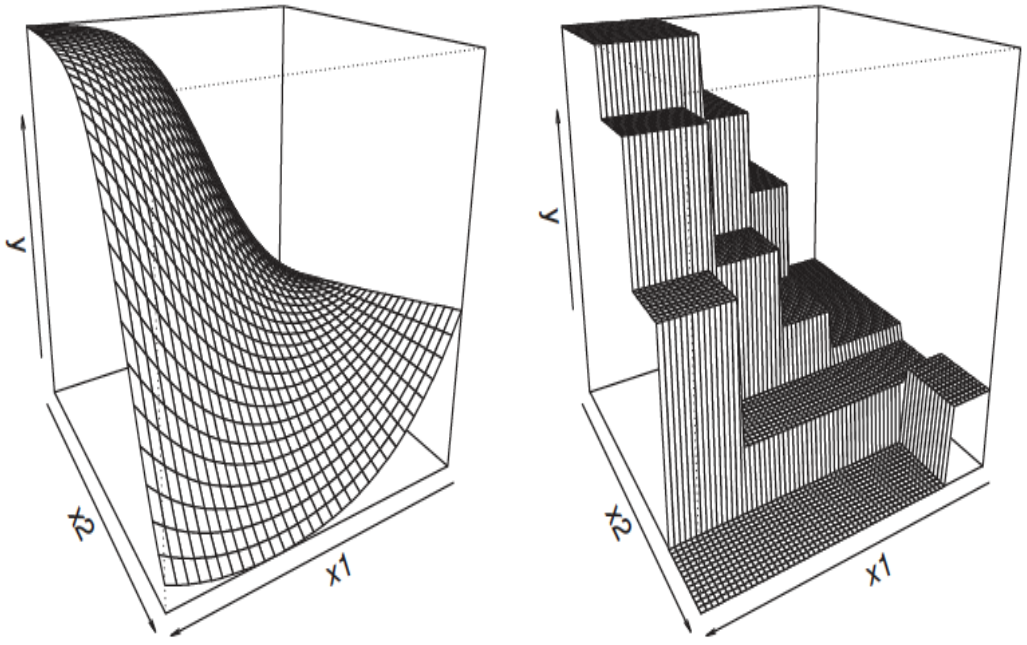

FIGURA 3.3: Una funzione continua in $\mathbb{R}^2$ e un'approssimazione tramite una funzione a gradini.

Lo schema può essere esteso al caso di funzioni di p variabili: sia quindi $y = f(x)$ dove $x \in \mathbb{R}^p$. Ci sono molti modi di estendere l'idea dal caso $p = 1$ al caso generale p. La Figura 3.3 mostra una funzione in $\mathbb{R}^2$ e la sua approssimazione mediante una funzione a gradini: le regioni con valori costanti sono quindi rettangoli, i lati dei quali sono paralleli agli assi di coordinate.

Queste caratteristiche di una funzione approssimata, con alcune specifiche aggiuntive da descrivere in seguito, consentono di rappresentarla come un albero binario, mostrato nel pannello superiore della Figura 3.4; il pannello inferiore mostra la corrispondente suddivisione del dominio della funzione $f(x)$ e i valori della funzione approssimante in ciascun rettangolo.

I componenti dell'albero sono disuguaglianze, chiamate nodi, relative a qualsiasi componente x ad esempio $x_2 < 1.725$. È possibile iniziare esaminando la disuguaglianza della radice dell'albero, che si trova in cima. Seguiamo il ramo sinistro se la disuguaglianza è vera e il ramo destro se non lo è. Procediamo nello stesso modo, esaminando sequenzialmente tutte le disuguaglianze fino a raggiungere i nodi terminali, chiamate foglie, che forniscono i valori della funzione approssimata.

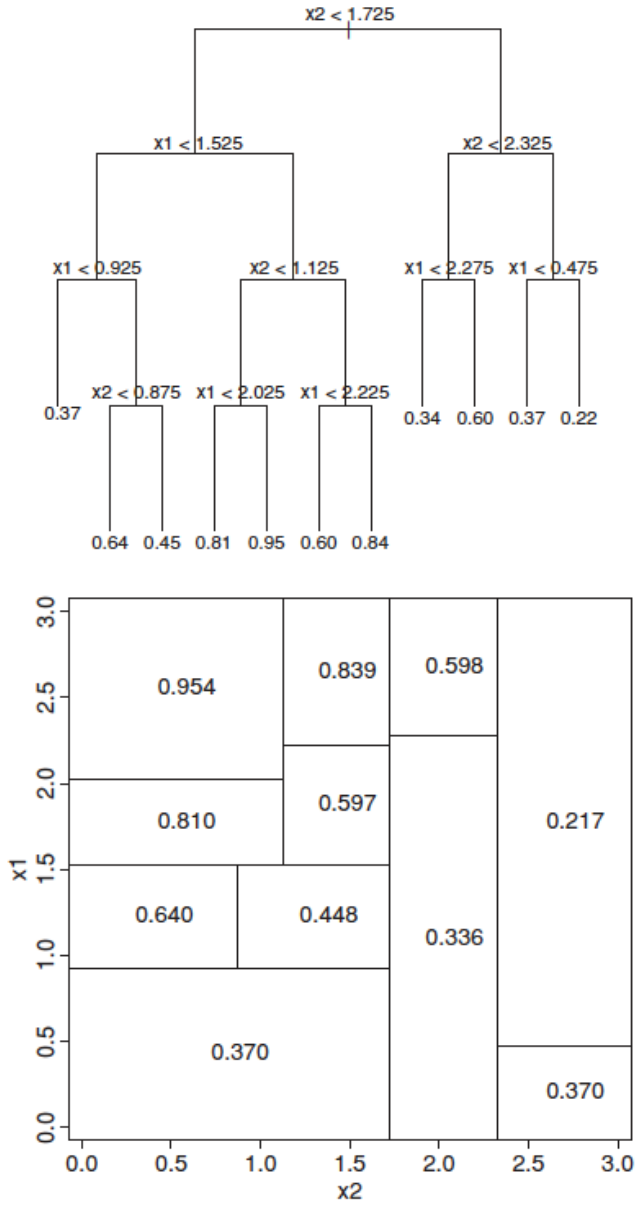

FIGURA 3.4: Albero corrispondente all'approssimazione del pannello inferiore (in alto), e partizione del dominio indotta dall'albero (in basso)

La rappresentazione grafica come un albero ha importanti vantaggi: poiché l'albero è identificato da pochi elementi numerici, può essere facilmente memorizzato. Un secondo importante vantaggio è che è possibile passare da un'approssimazione a una più accurata suddividendo una delle componenti in due sotto-rettangoli con le stesse caratteristiche dell'originale. Ciò corrisponde all'estensione di un ramo dell'albero a un ulteriore livello di ramo. Questa caratteristica permette immediatamente di costruire ricorsivamente una sequenza di approssimazioni che sono sempre più accurate, ciascuna ottenuta rifinendo quella precedente.

3.5.1.1 Albero di regressione: crescita

Si utilizzi l'idea di approssimazione con una funzione a gradino per approssimare le funzioni di interesse, che sono le funzioni di regressione. Nel contesto di questo lavoro di tesi, la funzione di regressione $f(x)$ non sarà nota, ma potrà essere osservarla indirettamente attraverso n osservazioni campionarie.

Per semplicità, si consideri il caso in cui $p = 1$ e i dati suddivisi in due gruppi: 'ieri' e 'domani'. È possibile stimare la curva di regressione $f(x)$ sottostante i dati con una funzione a gradino del tipo appena descritto, cioè

$$\hat{f}(x) = \sum_{h=1}^{J} c_h \mathbb{I}(x \in R_h) \tag{3.28}$$

dove $c_1, \ldots, c_J$ sono costanti e $\mathbb{I}(z)$ è la funzione indicatrice 0-1 del predicato logico z. In generale, gli insiemi $R_1, \ldots, R_J$ sono rettangoli, nel senso p-dimensionale, con i loro lati paralleli agli assi delle coordinate. Nel caso specifico in cui $p = 1$, ovviamente R_h si riducono a segmenti di retta.

È necessaria una funzione obiettivo per scegliere R_h e c_h. Il criterio di riferimento è la devianza, ma la sua minimizzazione, anche se il numero di passi J viene fissato, comporta calcoli molto complessi. Pertanto, operativamente viene seguito un approccio subottimale di ottimizzazione passo dopo passo, nel senso che costruiamo una sequenza di approssimazioni gradualmente più raffinate e per ciascuna di queste minimizziamo la devianza relativa al passaggio dall'approssimazione corrente a quella precedente.

L'algoritmo inizia suddividendo la retta reale associata a una delle variabili, ad esempio, x_j, in due parti; quale variabile considerare viene discusso successivamente. Ad ogni sottointervallo viene assegnato un valore, c_h, dato dalla media aritmetica dei valori osservati y_i aventi componente x_j che cade in questo sottointervallo, indipendentemente dalle altre covariate. Notare che questo passo divide lo spazio $\mathbb{R}^p$ in due regioni tramite un iperpiano parallelo all'asse delle coordinate j-esimo. I passaggi successivi dell'algoritmo procedono allo stesso modo, dividendo ogni volta una delle regioni esistenti di $\mathbb{R}^p$ in due regioni ulteriori, sempre con una divisione parallela a uno degli assi delle coordinate.

La Figura 3.5 mostra tre istanze di porzioni non compatibili con il processo precedente; la quarta è ammissibile.

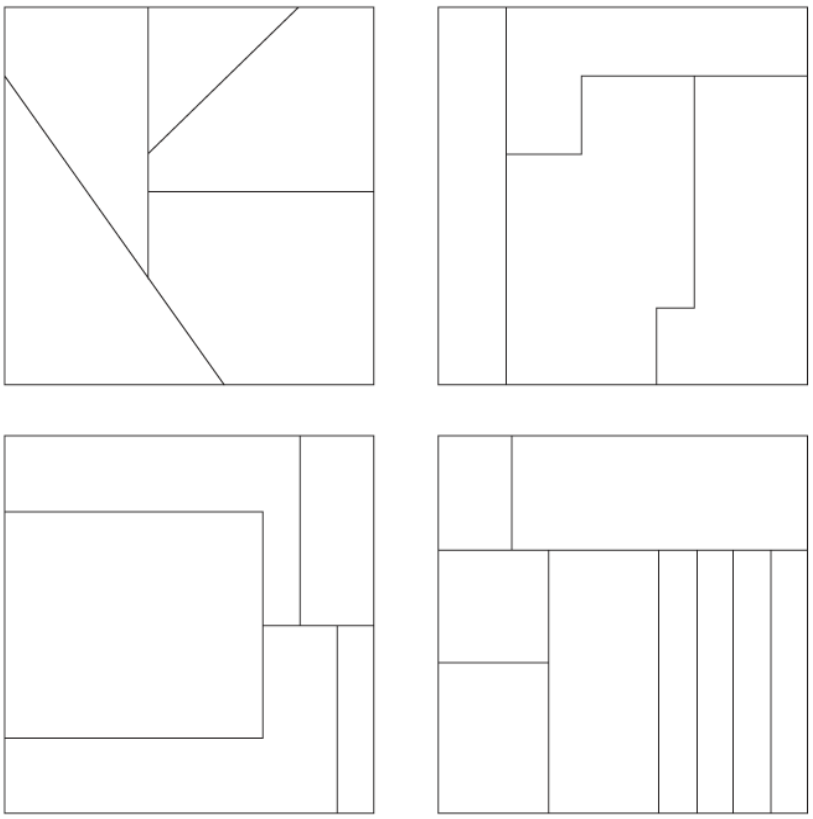

FIGURA 3.5: Suddivisioni tramite iperpiano dello spazio

Un aspetto cruciale è il fatto che, ad ogni passo, uno dei rettangoli già costruiti viene diviso in due, così come la porzione di dati ad esso appartenente; viene ottimizzata la devianza rispetto a questa operazione. Pertanto, si tratta di una procedura di ottimizzazione miope: anche se non garantisce la minimizzazione globale della devianza, fornisce soluzioni accettabili, mantenendo una complessità computazionale limitata. Almeno in linea di principio, questa procedura può essere applicata iterativamente attraverso successive suddivisioni di $\mathbb{R}^p$ fino a quando non si è più in grado di distinguere insiemi contenenti una singola osservazione campionaria e quindi ottenere un albero con n foglie. Per essere utile, il numero di foglie deve essere inferiore a n, preferibilmente molto inferiore. Pertanto, dopo la fase di crescita dell'albero, con lo sviluppo completo o quasi completo di tutte le foglie, si passa ad una fase di potatura dell'albero. Viene descritto ora l'algoritmo di crescita, mentre la fase di potatura verrà ripresa successivamente.

Per sviluppare l'algoritmo di crescita, si noti innanzitutto che qualunque sia la suddivisione di $\mathbb{R}^p$ in iper-rettangoli, è possibile scomporre la devianza come segue

$$D = \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2 = \sum_{h=1}^{J} \left\{ \sum_{i \in R_h} (y_i - \hat{c}_h)^2 \right\} = \sum_h D_h. \tag{3.29}$$

Si tenga presente anche la proprietà generale che il minimo di $\sum_{i=1}^{n}(z_i - a)^2$ rispetto a a è ottenuto per $a = M(z)$, dove $M(\cdot)$ è l'operatore di media del vettore.

Il processo di crescita inizia con $J = 1$, $R_J = \mathbb{R}^p$, $D = \sum_i (y_i - M(y))^2$, e procede iterativamente per un numero di cicli, secondo lo schema seguente:

- una volta scelto un rettangolo R_h, il valore appropriato $\hat{c}_h$ è la media dei valori corrispondenti $\hat{c}_h = M(y_i : x_i \in R_h)$
- se la regione R_h viene divisa in due parti, R_h e $R_{h'}$ (si passa quindi a $J+1$ foglie), il sommando D_h di D viene sostituito da

$$D_h^* = \sum_{i \in R_h} (y_i - \hat{c}_h)^2 + \sum_{i \in R_{h'}} (y_i - \hat{c}_{h'})^2 \tag{3.30}$$

  con un "guadagno" di $g_h = D_h - D_h^*$
- è possibile ispezionare tutte le p variabili esplicative e, per ciascuna di esse, tutti i possibili punti di suddivisione, selezionando la variabile e il suo punto di suddivisione che massimizzano g_h.

La procedura si ferma quando $J = n$, almeno concettualmente. Principalmente, se n è molto grande, ci si ferma prima, ad esempio, quando tutte le foglie contengono un numero di elementi campionari inferiore a un valore predefinito, o quando il calo relativo della devianza è inferiore a una soglia prefissata.

3.5.1.2 Albero di regressione: potatura

Un grande albero con n foglie è concettualmente equivalente all'interpolazione tramite un polinomio di grado $n - 1$ che passa esattamente per tutti i punti, quindi non è molto utile. È quindi necessario potare l'albero rimuovendo rami di scarsa o nessuna utilità. Viene introdotto quindi una funzione obiettivo che tiene conto sia del costo della complessità dell'albero sia della devianza. Questa funzione obiettivo è data da

$$C_\alpha(J) = \sum_{h=1}^{J} D_h + \alpha J, \tag{3.31}$$

dove α è un parametro di penalità non negativo. Breiman et al. (1984) hanno dimostrato che l'insieme di sottoalberi che minimizzano la misura di costo-complessità è nidificato, il che significa che aumentando α possiamo trovare gli alberi ottimali tramite una sequenza di operazioni di potatura sull'albero corrente. Per minimizzare (3.18), si procede eliminando sequenzialmente una foglia alla volta e scegliendo ad ogni passo la foglia che causa il minor aumento nella devianza. La scelta del parametro α può essere fatta utilizzando uno dei metodi descritti in precedenza. Il processo di previsione per nuovi dati segue un approccio simile, consentendo all'osservazione di scendere dall'albero disponibile fino a raggiungere una foglia con il valore predetto. Questo processo viene ripetuto per ogni osservazione nel set di test, calcolando infine la devianza osservata. L'albero può essere potato utilizzando diverse tecniche, con l'obiettivo di ottenere un albero più piccolo ma comunque accurato nella previsione.

3.5.2 Bagging

Sia $Z = \{(\tilde{x}_1, y_1), (\tilde{x}_2, y_2), ..., (\tilde{x}_n, y_n)\}$ il set di addestramento e $C(x)$ un classificatore ottenuto con uno dei metodi presentati in precedenza. In seguito, il modello associato a $C(x)$ viene chiamato modello di base. Per semplicità, si consideri il caso con $K = 2$. Adottando una procedura bootstrap, viene esaminato il campione Z_1^* ottenuto estraendo n elementi dal set di addestramento Z con rimpiazzo. Si ottiene un nuovo classificatore $C_1^*(x)$ adattando a Z_1^* uno dei modelli presentati in precedenza in questo capitolo, ad esempio, un albero decisionale. In generale, per un x fisso, il nuovo modello adattato è diverso dall'originale. L'applicazione ripetuta di questo passaggio, ad esempio B volte, produce un insieme di campioni Z_b^* $(b = 1, ..., B)$, ognuno di dimensione n, che producono a loro volta B nuovi classificatori $C_b^*(x)$, $b = 1, ..., B$. Un nuovo classificatore che è una media dei risultati di ciascun $C_b^*(x)$ per il dato x può essere introdotto. La forma più naturale di media è la media aritmetica

$$C_{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^{B} C_b^*(x) \tag{3.32}$$

che assegna l'unità con le variabili esplicative x a $y = 1$ se $C_{\text{bag}}(x) > \frac{1}{2}$ e a $y = 0$ altrimenti. Questa procedura di classificazione è chiamata bootstrap aggregating, da cui deriva il termine abbreviato "bagging". L'errore di classificazione della nuova procedura è spesso inferiore a quello dei modelli di base. Molti procedure di classificazione producono anche una funzione $\hat{p}(x)$, che fornisce la probabilità che un'unità con variabili esplicative x appartenga a ciascuna classe. Una variante del bagging funziona mediando i $\hat{p}_b^*(x)$, che stimano le probabilità di classe per il modello adattato a ciascuno dei B campioni bootstrap Z_b^* e utilizzando questa nuova $\hat{p}_{\text{bag}}(x) = \frac{\sum_b \hat{p}_b^*(x)}{B}$ come indicatore di probabilità di appartenenza a una classe. La strategia del bagging può essere facilmente adattata al contesto della regressione, dove al posto dei classificatori $C(x)$ si usano le previsioni derivate dai modelli discussi precedentemente. In questo caso, non è necessario tornare al criterio del voto maggioritario, poiché si può usare direttamente la media dei predittori ottenuti mediante campionamento bootstrap come nuovo predittore. La nuova previsione può avere una varianza inferiore a quella del modello originale. Le procedure di bagging spesso migliorano notevolmente la capacità predittiva, soprattutto quando i classificatori utilizzati sono molto instabili, ad esempio, alberi o reti neurali. Tuttavia, con procedure più stabili, il bagging può peggiorare leggermente la qualità della previsione. È anche ovvio che l'operazione di combinare i risultati dei singoli modelli attraverso la media aritmetica comporta la perdita di qualsiasi struttura semplice esistente nel modello di base, portando a una maggiore difficoltà nell'interpretare i risultati.

L'uso di campioni casuali di osservazioni consente l'uso di una tecnica chiamata out-of-bag per una facile stima degli errori di previsione. Infatti, in ogni campione bootstrap, alcuni dei dati del set di addestramento originale sono esclusi. Di conseguenza, per ciascun classificatore $C_b^*(x)$, i dati del set di addestramento Z che non sono nel campione Z_b^* possono essere utilizzati come set di test. Possiamo quindi stimare, ad esempio, l'errore di classificazione su questi dati al di fuori del campione utilizzato per l'adattamento (out-of-bag), senza richiedere un set di test o dover scegliere soluzioni computazionalmente intensive, come la cross-validation.

3.5.3 Random Forest

Il bagging costruisce modelli diversi che poi vengono combinati cambiando ad ogni iterazione l'insieme di unità o il peso assegnato a ciascuna unità su cui adattare il modello, utilizzando tutte le p variabili esplicative disponibili ad ogni iterazione. Un altro modo per ottenere combinazioni di modelli consiste nel considerare diversi sottoinsiemi delle variabili esplicative, invece di considerare sottoinsiemi delle unità. Una strategia di questo tipo è stata proposta con alberi come classificatori di base, scegliendo le variabili da inserire in ogni modello per selezione casuale: questa procedura è chiamata random forest. Si noti che questo termine viene talvolta utilizzato con un significato più generale, riferendosi a qualsiasi classificatore ottenuto come combinazione di un insieme di alberi di classificazione. Ad esempio, in questa interpretazione del termine, il bagging e il boosting appartengono anche alle random forests quando applicati agli alberi. La procedura consiste nel selezionare casualmente, ad ogni nodo dell'albero, un piccolo gruppo di covariate, che vengono esaminate per trovare il loro miglior punto di suddivisione. Pertanto, anziché esplorare tutte le variabili possibili in ciascun nodo, solo q ($q \leq p$) variabili scelte casualmente vengono esaminate. L'albero cresce fino alla dimensione massima ma non viene potato. In effetti, la combinazione risultante di vari alberi evita l'overfitting. Il numero q di variabili da selezionare in ciascun nodo è un parametro di regolazione da determinare e viene generalmente mantenuto costante su tutti i nodi. Il numero è spesso scelto considerando foreste costruite con diversi valori di q e determinando il valore che minimizza l'errore su un set di test. L'altro parametro di regolazione è il numero di alberi, (ad esempio B), che compongono la foresta. Si può dimostrare che l'errore globale converge a un limite inferiore quando B aumenta e che non causa problemi di overfitting quando vengono aggiunti alberi aggiuntivi. Se, quindi, si sceglie un valore sufficientemente grande per B, si può essere certi che l'errore di previsione ottenuto non sarà molto lontano dal suo minimo. Nella costruzione di una foresta, di solito è associata anche una procedura di bagging con selezione casuale delle variabili. Ogni albero è fatto crescere su un diverso campione bootstrap con un numero q di variabili selezionate casualmente per ogni nodo. Qui, il bagging, il cui obiettivo principale è migliorare l'accuratezza delle previsioni, consente anche di utilizzare la tecnica out-of-bag per scegliere il parametro di regolazione q e ottenere le misure di importanza delle covariate. È possibile utilizzare l'errore di previsione ottenuto dai dati out-of-bag quando viene determinato q, invece dell'errore su un set di test. Per ottenere una misura dell'importanza di ciascuna variabile esplicativa nella previsione della risposta, si può procedere utilizzando i dati out-of-bag nel seguente modo. Per ogni albero, si ottiene

l'errore di classificazione sulla parte out-of-bag dei dati. Lo stesso viene fatto dopo la permutazione casuale dei valori di ogni variabile esplicativa. Si calcola la media della differenza tra i due errori di classificazione e si divide per la deviazione standard delle differenze, fornendo un indicatore di come quella variabile influenzi le previsioni. Un altro indicatore della rilevanza delle variabili si basa sulla misura di importanza per un singolo albero. Questo viene ottenuto come media su tutti gli alberi nella foresta di quella misura di importanza, calcolata separatamente per ciascuna variabile. Rispetto ad altri metodi di combinazione di modelli, le random forests hanno alcuni vantaggi interessanti. L'accuratezza delle loro previsioni è paragonabile a quella del boosting e in alcuni casi è migliore, ma sono molto più veloci perché ogni singolo albero si basa su un numero inferiore di variabili e il carico computazionale è quindi inferiore. È anche relativamente semplice costruire un algoritmo che, sfruttando il calcolo parallelo, può accelerare ulteriormente la procedura random forest.

3.6 Rete neurale

Il termine rete neurale comprende una vasta famiglia di tecniche sviluppate nell'apprendimento automatico. Viene qui brevemente descritto solo la versione più semplice.

La Figura 3.6 mostra p variabili esplicative (input) in relazione con q variabili di risposta, o output. L'aspetto più caratteristico è il livello di r variabili latenti, che non è osservabile (nascosto) e si trova tra i due gruppi precedenti nel senso che le covariate influenzano le variabili latenti; queste a loro volta influenzano le variabili di risposta. Il numero di variabili di input e di output è determinato dal problema, ma il numero r di variabili latenti è qualcosa che possiamo scegliere, poiché sono solo entità concettuali. Nella Figura 3.6, si ha $p = 4$, $r = 3$ e $q = 2$, e sono mostrate anche alcune "variabili costanti", identiche a 1.

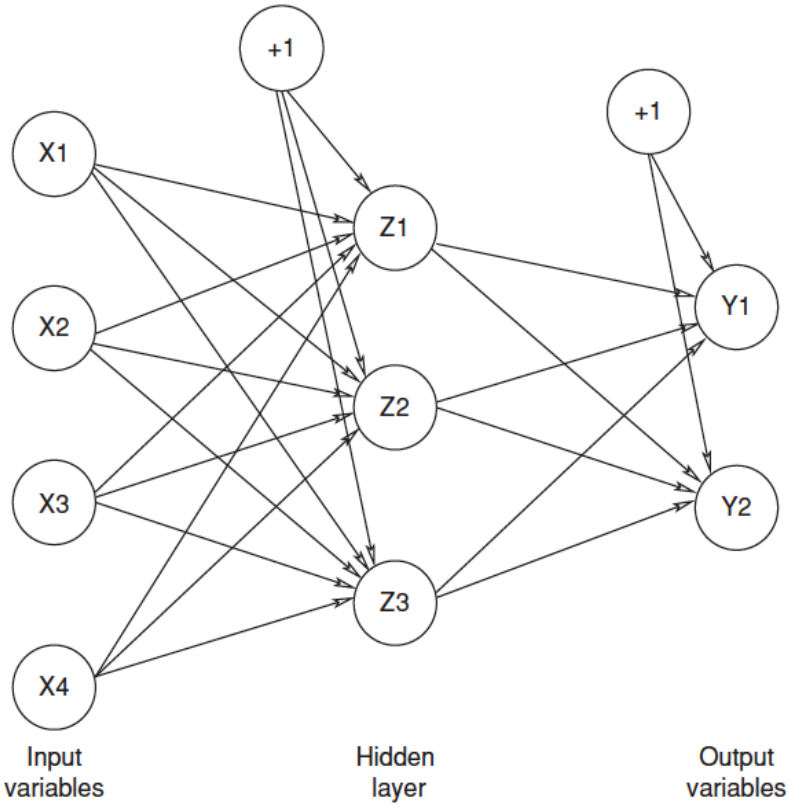

FIGURA 3.6: Rete neurale semplice

Il termine rete neurale ha avuto origine come modello matematico che in passato si credeva fosse il meccanismo che controllava il funzionamento del cervello animale: ogni nodo del grafico rappresentava un neurone e gli archi rappresentavano le sinapsi.

Ora è noto che il cervello animale sia molto più complesso, ma il termine rete neurale sopravvive. Una rete neurale è essenzialmente uno schema di regressione a due fasi, generalmente di tipo non lineare o almeno parzialmente non lineare. Indichiamo le variabili di input, latenti e di output generiche con x_h, z_j e y_k, rispettivamente, e si aggiungano variabili costanti x_0 e z_0 uguali a 1. Lo schema precedente può ora essere espresso come

$$z_j = f_0\left(\sum_{h\to j} \alpha_{hj} x_h\right) \tag{3.33}$$

$$y_k = f_1\left(\sum_{j\to k} \beta_{jk} z_j\right) \tag{3.34}$$

dove α_{hj} e β_{jk} sono parametri da stimare e le somme sono sugli indici delle variabili per le quali è prevista una relazione di dipendenza. La Figura (3.6) mostra queste dipendenze con frecce e coinvolge tutte le variabili compatibili, anche se questo non è necessariamente il caso. Si passa quindi a vedere che la struttura risultante sia un grafo aciclico con archi diretti e pesi associati ai coefficienti α e β. Per completare il quadro, devono essere specificate le funzioni di attivazione f_0 e f_1. Nei problemi di regressione, dove i y_k sono generalmente non limitati, si ipotizza

$$f_0(u) = \frac{e^u}{1+e^u} \tag{3.35}$$

$$f_1(u) = u \tag{3.36}$$

dove la scelta di f_0 è la funzione logistica. Almeno una delle due funzioni f_0 e f_1 deve essere non lineare per evitare di ridurre l'intera rete a un insieme di relazioni lineari, eliminando efficacemente il livello latente. Esistono risultati matematici che danno origine a proprietà interessanti per il framework. In particolare, si può dimostrare che una rete neurale con unità di output lineari può approssimare uniformemente qualsiasi funzione continua f su insiemi compatti, aumentando opportunamente il numero di unità del livello latente (Ripley 1996). Sono possibili estensioni in varie direzioni. Una delle più comuni è considerare diversi livelli di variabili latenti. Un'altra è introdurre archi che saltano un livello: nel caso del singolo livello latente considerato qui, ciò significa inserire un arco direttamente tra alcune variabili del livello di input e alcune dell'output. Dopo

aver fissato r, si stimano i coefficienti α e β secondo le osservazioni campionarie. Questo viene fatto minimizzando la solita funzione obiettivo

$$D = \sum_i \|y_i - f(x_i)\|^2, \tag{3.37}$$

dove y_i indica il vettore q-dimensionale delle variabili di risposta dell'i-esima osservazione. Analogamente, x_i è il vettore p-dimensionale delle covariate corrispondenti e $f(x)$ è il vettore, il cui componente k-esimo è

$$f(x)_k = f_1\left(\sum_{j\to k} \beta_{jk} f_0\left(\sum_{h\to j} \alpha_{hj} x_h\right)\right) \tag{3.38}$$

$(k = 1, \ldots, q)$. Versioni più elaborate di questa funzione obiettivo possono essere ottenute includendo un termine di penalizzazione per evitare problemi di sovradattamento, ad esempio, funzioni del tipo

$$D_0 = D + \lambda J(\alpha, \beta), \tag{3.39}$$

dove λ è un parametro di regolazione positivo e $J(\alpha, \beta)$ è una funzione di penalizzazione. Tra le forme di penalizzazione più comuni ci sono

$$J(\alpha, \beta) = \frac{\#\,(h,k)}{n} \sum_i \frac{\partial^2 y_{ki}}{\partial x_{hi} \partial x_{hi}} \tag{3.40}$$

$$J(\alpha, \beta) = \sum_h \alpha_h^2 + \sum_k \beta_k^2 \tag{3.41}$$

delle quali la prima forma penalizza l'ampiezza della seconda derivata e la seconda tende a restringere i parametri a 0; quest'ultima è chiamata riduzione del peso. Qui y_{ki} indica il k-esimo componente di y_i. Queste formulazioni, sia D che la funzione di penalità J, hanno senso se le variabili sono misurate sulla stessa scala. Come operazione preliminare, è quindi meglio normalizzarle, ad esempio, ridimensionando tutte le variabili tra 0 e 1 (almeno approssimativamente). Per il parametro di regolazione λ, Venables & Ripley (2002) consigliano di scegliere un valore tra 10^{-4} e 10^{-2}. Chiaramente, la minimizzazione di D_0 richiede un processo di ottimizzazione numerica. Molto sforzo

è stato investito nello sviluppo di tali algoritmi. Il metodo più comune è chiamato back-propagation, che ha proprietà interessanti. Uno degli aspetti più importanti in questo contesto è che esiste una variante dell'algoritmo di back-propagation, che consente di aggiornare successivamente le stime dei parametri man mano che diventano disponibili nuovi dati. Deve essere sottolineato che l'esperienza pratica ha fornito ampie evidenze che la funzione obiettivo D_0 ha spesso molti punti di minimi locali, ed è quindi saggio avviare l'algoritmo di ottimizzazione da diversi punti iniziali. Questa difficoltà a sua volta influenza qualcos'altro: nella scelta di λ è difficile sfruttare tecniche come la cross-validazione, poiché l'algoritmo varia ampiamente nel individuare il minimo.

3.7 Valutazione della performance dei modelli

Nel vasto territorio della statistica e dell'apprendimento automatico, la tecnica della Validazione Incrociata con Esclusione di Uno (LOOCV) svolge un ruolo cruciale nell'analisi dei modelli.

Prima di entrare in dettagli della LOOCV, è importante comprendere il concetto più ampio di validazione incrociata (CV). La CV è una tecnica utilizzata per valutare l'accuratezza di un modello predittivo utilizzando una strategia di partizione dei dati. La sua importanza risiede nel fornire una stima attendibile delle prestazioni del modello su dati non osservati, contribuendo così a evitare il sovradattamento e valutare la generalizzazione del modello. La CV è ampiamente utilizzata nell'ambito della modellazione statistica e dell'apprendimento automatico per ottimizzare la selezione del modello e valutarne l'efficacia.

Tornando alla LOOCV, viene suddiviso il dataset in un training set e un testing set, ma con un twist: il testing set contiene solo un'osservazione alla volta. Successivamente, viene addestrato il modello utilizzando solo i dati del training set, escludendo un'osservazione alla volta. Il modello addestrato viene quindi utilizzato per prevedere il valore di risposta dell'osservazione esclusa, calcolando l'errore quadratico medio (MSE) tra la previsione e il valore reale. Questo processo viene ripetuto per ogni osservazione nel dataset, e alla fine viene calcolata la media degli MSE ottenuti da tutte le iterazioni.

La formula per il calcolo dell'errore quadratico medio (MSE) con LOOCV è:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.42}$$

dove:

- n è il numero totale di osservazioni nel dataset.
- y_i rappresenta il valore di risposta dell'osservazione i.
- $\hat{y}_i$ è il valore di risposta previsto per l'osservazione i utilizzando il modello addestrato su tutti gli altri dati tranne l'osservazione i.

Come ogni strumento, la LOOCV ha i suoi pro e contro. Da una parte, fornisce una misura di errore meno distorta rispetto all'utilizzo di un singolo testing set e tende a non sovrastimare l'errore di previsione. Dall'altra, richiede un notevole sforzo computazionale e può essere sensibile a outlier o dati particolarmente influenti.

In sintesi, la LOOCV si rivela uno strumento prezioso per valutare l'accuratezza dei modelli e stimare l'errore di previsione su dati non visti, offrendo un approccio robusto e affidabile nell'analisi statistica dei dati.

Capitolo 4

Metodi di imputazione multipla

La Multiple Imputation (MI) è emersa negli anni '70 come soluzione innovativa al problema dei dati mancanti, sviluppata da Donald B. Rubin. Inizialmente, l'attenzione si concentrò sulle lacune nei dati di reddito nel March Income Supplement to the Current Population Survey (CPS), dove l'approccio tradizionale di "hot deck imputation" si rivelava inadeguato per il calcolo accurato della varianza. Rubin (1977) propose di utilizzare multiple versioni complete del dataset, ciascuna riflettente l'incertezza associata ai dati mancanti. Questo concetto, rivoluzionario per l'epoca, portò alla creazione delle "regole di Rubin", che forniscono le formule necessarie per combinare le stime derivate da ciascuna imputazione.

L'idea di creare più versioni dei dati sembrava rivoluzionaria, poiché si discostava radicalmente dall'approccio tradizionale di stimare un unico valore per i dati mancanti. Rubin osservò che l'imputazione singola di un valore per i dati mancanti potrebbe non essere corretta in generale, in quanto richiedeva un modello per relazionare i dati non osservati a quelli osservati e poteva generare valori imputati con incertezza. La soluzione proposta da Rubin fu semplice e brillante: creare più imputazioni che riflettessero l'incertezza dei dati mancanti.

Rubin (1987a) fornì le basi metodologiche e statistiche per il metodo, delineando le condizioni in cui l'inferenza statistica sotto imputazione multipla sarebbe stata valida. Da allora, sono stati proposti diversi miglioramenti e soluzioni a problematiche specifiche. Ad esempio, sono stati sviluppati test per le combinazioni di parametri (Li et al., 1991a; Li et al., 1991b; Meng and Rubin, 1992) e algoritmi iterativi per dati mancanti multivariati (Rubin, 1987; Schafer, 1997; Van Buuren et al., 1999; Raghunathan et al., 2001; King et al., 2001).

Tuttavia, negli anni '90, la Multiple Imputation è stata oggetto di critiche, soprattutto da parte di Fay (1992), che sollevò dubbi sulla validità del metodo in determinate situazioni. La critica principale riguardava la dipendenza della validità della Multiple Imputation dalla forma dell'analisi successiva. Tuttavia, Meng (1994) sottolineò che i modelli di imputazione di Fay omettevano importanti relazioni necessarie nel modello di analisi, una situazione indesiderabile denominata "uncongeniality". Inoltre, si evidenziarono problemi legati alla stima della varianza, con alcune ricerche che indicavano un bias nelle stime di Rubin (Wang and Robins, 1998; Robins and Wang, 2000; Nielsen, 2003; Kim et al., 2006). Rubin (2003) tuttavia argomentò che l'obiettivo principale della stima della varianza era la costruzione di intervalli di confidenza, e che il bias osservato non sembrava influenzare significativamente la copertura di tali intervalli in una vasta gamma di casi pratici.

Negli ultimi anni, la MI ha ricevuto crescente accettazione come metodo standard per il trattamento dei dati mancanti, con molte criticità affrontate attraverso il continuo sviluppo e miglioramento del metodo. La sua adozione diffusa ha consolidato la MI come punto di riferimento contro cui valutare nuovi metodi sviluppati per gestire dati mancanti.

4.1 Metodo MICE

Ci sono diversi modi per implementare l'imputazione sotto modelli specificati condizionalmente. L'Algoritmo che segue descrive un'istanza particolare: l'algoritmo MICE (Multiple Imputation by Chained Equations) (Van Buuren e Groothuis-Oudshoorn, 2000, 2011). L'algoritmo inizia con un'estrazione casuale dai dati osservati e imputa i dati incompleti per ogni variabile. Un'iterazione consiste in un ciclo attraverso tutti Y_j. Il numero di iterazioni T può spesso essere basso, ad esempio 5 o 10. L'algoritmo MICE genera MI eseguendolo in parallelo m volte.

L'algoritmo MICE è un metodo di catena di Markov Monte Carlo (MCMC), dove lo spazio degli stati è la collezione di tutti i valori imputati. Più specificamente, l'algoritmo MICE è un campionatore di Gibbs, una tecnica di simulazione bayesiana che campiona dalle distribuzioni condizionali al fine di ottenere campioni dalla distribuzione congiunta (Gelfand e Smith, 1990; Casella e George, 1992). Nelle applicazioni convenzionali del campionatore di Gibbs, le distribuzioni condizionali complete sono derivate dalla distribuzione congiunta di probabilità (Gilks, 1996). Nell'algoritmo MICE, le distribuzioni

condizionali sono sotto il controllo diretto dell'utente, e quindi la distribuzione congiunta è conosciuta solo implicitamente e potrebbe non esistere effettivamente. Anche se quest'ultima è chiaramente indesiderabile dal punto di vista teorico (poiché non si conosce la distribuzione congiunta a cui l'algoritmo converge), in pratica non sembra ostacolare le applicazioni utili del metodo.

Per convergere a una distribuzione stazionaria, una catena di Markov deve soddisfare tre proprietà importanti (Roberts, 1996; Tierney, 1996):

1. **Irreversibilità o Ergodicità**: La catena di Markov deve essere irreversibile, il che significa che deve essere possibile passare da uno stato a qualsiasi altro stato della catena con una certa probabilità non nulla. In altre parole, la catena deve essere ergodica, garantendo che il processo possa esplorare completamente lo spazio degli stati nel lungo termine.

2. **Aperiodicità**: La catena di Markov deve essere aperiodica, il che significa che il periodo di ritorno di uno stato a se stesso deve essere finito e non dipendere dal numero di passaggi. Questo assicura che non vi siano cicli regolari di transizione tra gli stati, consentendo alla catena di esplorare in modo efficace l'intero spazio degli stati.

3. **Assorbimento**: La catena di Markov deve essere assorbente, il che significa che deve esistere almeno uno stato (o più stati) che una volta raggiunto non può essere lasciato. Questi stati sono chiamati stati assorbenti e rappresentano i "punti di non ritorno" della catena, poiché una volta raggiunti non si può più lasciare il sistema.

Queste proprietà sono fondamentali affinché una catena di Markov converga a una distribuzione stazionaria e rappresentano criteri essenziali per garantire l'efficienza e l'affidabilità degli algoritmi basati su catene di Markov, come ad esempio l'algoritmo di Metropolis-Hastings utilizzato per il campionamento da distribuzioni di probabilità complesse.

Sintetizzando quindi i vari passi:

1. Specificare un modello di imputazione $P(Y_{\text{mis}j}|Y_{\text{obs}j}, Y_{\text{-}j}, R)$ per la variabile Y_j con $j = 1, \ldots, p$.
2. Per ciascun j, riempire le imputazioni iniziali $\hat{Y}_{0\text{j}}$ con estrazioni casuali da $Y_{\text{obs}j}$.
3. Ripetere per $t = 1, \ldots, T$:
4. Ripetere per $j = 1, \ldots, p$:
5. Definire $\hat{Y}_{\text{t-j}} = (\hat{Y}_{\text{t-1,1}}, \ldots, \hat{Y}_{\text{t-1,j-1}}, \hat{Y}_{\text{t-1,j+1}}, \ldots, \hat{Y}_{\text{t-1,p}})$ come i dati completi attualmente tranne Y_j.
6. Estrarre $\hat{\theta}_{\text{tj}} \sim P(\theta_{\text{tj}}|Y_{\text{obs}j}, \hat{Y}_{\text{t-j}}, R)$.
7. Estrarre le imputazioni $\hat{Y}_{\text{tj}} \sim P(Y_{\text{mis}j}|Y_{\text{obs}j}, \hat{Y}_{\text{t-j}}, R, \hat{\theta}_{tj})$.
8. Fine ripetizione per j.
9. Fine ripetizione per t.

La periodicità è un potenziale problema e può sorgere nella situazione in cui i modelli di imputazione sono chiaramente inconsistenti. Un esempio piuttosto artificiale di comportamento oscillatorio si verifica quando Y_1 viene imputato da $Y_2 + 1$ e Y_2 viene imputato da $-Y_1 + 2$ per qualche α fisso e non nullo. Il campione oscillerà tra due stati qualitativamente diversi, quindi la correlazione tra Y_1 e Y_2 dopo l'imputazione di Y_1 sarà diversa da quella dopo l'imputazione di Y_2. In generale, si vuole che le inferenze siano indipendenti dal punto di sospensione. Un modo per diagnosticare il problema è fermare la catena in diversi punti. Il punto di sospensione non dovrebbe influenzare le inferenze statistiche. L'aggiunta di rumore per creare imputazioni è un salvagente contro la periodicità e consente al campionatore di "uscire" più facilmente.

La non ricorrenza potrebbe anche essere un potenziale problema, manifestandosi come comportamento esplosivo o non stazionario. Ad esempio, se le imputazioni vengono effettuate attraverso funzioni deterministiche, la catena di Markov potrebbe bloccarsi. Tali casi possono talvolta essere diagnosticati dalle linee traccia del campionatore. Le proprietà richieste del metodo MCMC possono essere tradotte in condizioni sugli autovalori della matrice delle probabilità di transizione (MacKay, 2003).

4.1.1 Predict mean matching

Il Predictive Mean Matching (PMM) sottocampiona dai dati osservati. Il metodo calcola il valore predetto della variabile target Y secondo il modello di imputazione specificato. Per ogni voce mancante, il metodo forma un piccolo insieme di potenziali donatori (tipicamente con 1, 3 o 10 membri) da tutti i casi completi che hanno valori predetti vicini al valore predetto per l'osservazione mancante. Viene effettuata un'estrazione casuale tra i potenziali donatori, e il valore osservato dei donatori viene preso per sostituire il valore mancante. L'assunzione è che all'interno di ogni insieme, i dati dei riceventi seguano la stessa distribuzione dei dati dei potenziali donatori.

Il PMM è un metodo facile da usare e versatile. È abbastanza robusto alle trasformazioni della variabile target, quindi imputare $\log(Y)$ spesso produce risultati simili all'imputazione di $\exp(Y)$. Il metodo permette anche variabili target discrete. Le imputazioni si basano su valori osservati altrove, quindi sono realistiche e ammissibili. Le imputazioni al di fuori dell'intervallo dei dati osservati non si verificheranno, evitando così problemi con imputazioni prive di significato. Il modello è implicito, il che significa che non c'è bisogno di definire un modello esplicito per la distribuzione dei valori mancanti. Per questo motivo, il PMM è meno vulnerabile alla specificazione errata del modello rispetto ad altri metodi.

Il PMM è un esempio di metodo di tipo hot deck, dove i valori vengono imputati utilizzando valori dai casi completi abbinati rispetto a qualche metrica. L'espressione "hot deck" si riferisce letteralmente a un pacchetto di schede di controllo del computer contenenti i dati dei casi che sono in qualche modo vicini.

Sono possibili varie metriche per definire la distanza tra i casi. Le metriche del PMM sono particolarmente utili per le applicazioni con dati mancanti perché è ottimizzata per ciascuna variabile target separatamente. Il valore predetto è generalmente un comodo riepilogo numerico delle informazioni importanti che riguardano il target. Il calcolo è semplice, ed è facile includere variabili nominali e ordinali.

Una volta definita la metrica, ci sono vari modi per selezionare il donatore. Sia $\hat{y}_i$ il valore predetto delle righe con un y_i osservato dove $i = 1, \ldots, n_1$. Allo stesso modo, sia $\hat{y}_j$ il valore predetto delle righe con y_j mancante dove $j = 1, \ldots, n_0$. distinguono quattro metodi:

1. Viene scelta una soglia t, e si prendono tutti i i per cui $|\hat{y}_i - \hat{y}_j| < t$ come donatori potenziali per l'imputazione di j. Viene eseguito un campionamento casuale di un donatore tra i candidati e prendi il suo y_i come valore di sostituzione.

2. Si prende il candidato più vicino, cioè il caso i per cui $|\hat{y}_i - \hat{y}_j|$ è minimo, come donatore. Questo è noto come "hot deck del vicino più prossimo", "hot deck deterministico" o "predittore più vicino".

3. Si trova i d candidati per cui $|\hat{y}_i - \hat{y}_j|$ è minimo e campiona uno di loro. I valori usuali per d sono 3, 5 e 10. Esiste anche un metodo adattivo per specificare il numero di donatori.

4. Si campiona un donatore con una probabilità che dipende da $|\hat{y}_i - \hat{y}_j|$.

Il pericolo evidente del PMM è la duplicazione dello stesso valore del donatore molte volte. Questo problema è più probabile che si verifichi se il campione è piccolo o se ci sono molti più dati mancanti che dati osservati in una particolare regione del valore predetto. Tali regioni sbilanciate sono più probabili se la proporzione di casi incompleti è alta o se il modello di imputazione contiene variabili molto fortemente correlate alla mancanza di dati.

È disponibile un certo lavoro di simulazione sui diversi metodi per definire l'insieme di potenziali donatori. Impostare $d = 1$ è generalmente considerato troppo basso, poiché potrebbe selezionare nuovamente lo stesso donatore più volte. Il PMM funziona molto male quando d è piccolo e ci sono molte corrispondenze per i predittori tra gli individui da imputare. La ragione è che gli individui legati ottengono tutti lo stesso valore imputato in ciascun set di dati imputati quando $d = 1$. Impostare d su un valore alto (ad esempio $n = 10$) allevia il problema della duplicazione, ma potrebbe introdurre un bias poiché la probabilità di corrispondenze sbagliate aumenta. Sono stati valutati $d = 3$, $d = 10$ e uno schema adattivo. Il metodo adattivo è stato leggermente migliore rispetto all'uso di un numero fisso di candidati, ma le differenze erano piccole. Viene notato che potrebbero anche esserci situazioni in cui l'estimazione adattiva potrebbe essere più vantaggiosa, ma ulteriori lavori su questo problema sono ancora mancanti.

È utile distinguere tre tipi di abbinamento:

1. Tipo 0: $\hat{y} = X_{\text{obs}}\hat{\theta}$ è abbinato a $\hat{y}j = X_{\text{mis}}\hat{\theta}$;
2. Tipo 1: $y = X_{\text{obs}}\hat{\theta}$ è abbinato a $y_j = X_{\text{mis}}\tilde{\theta}$;
3. Tipo 2: $y = X_{\text{obs}}\tilde{\theta}$ è abbinato a $y_j = X_{\text{mis}}\tilde{\theta}$.

Qui $\hat{\theta}$ è la stima di θ, mentre $\tilde{\theta}$ è un valore estratto casualmente dalla distribuzione a posteriori di θ. L'abbinamento di Tipo 0 ignora la variabilità del campionamento in $\hat{\theta}$, portando a imputazioni improprie. L'abbinamento di Tipo 2 sembra risolvere questo problema. Tuttavia, è insensibile al processo di estrazione casuale di θ se ci sono solo poche variabili. Nel caso estremo, con un singolo X, l'insieme di potenziali donatori basato su $|\tilde{y}_i - \tilde{y}j|$ rimane invariato sotto valori diversi di $\tilde{\theta}$, quindi gli stessi donatori vengono selezionati troppo spesso. L'abbinamento di Tipo 1 è un piccolo ma astuto adattamento della distanza di abbinamento che sembra alleviare il problema. La differenza con l'abbinamento di Tipo 0 e Tipo 2 è che nell'abbinamento di Tipo 1 solo $X_{\text{mis}}\tilde{\theta}$ varia stocasticamente e non si annulla più. Di conseguenza, $\tilde{\theta}$ incorpora la variazione tra imputazioni.

4.1.2 CART

I metodi CART hanno proprietà che li rendono attraenti per l'imputazione: sono robusti contro i valori anomali, possono gestire la multicollinearità e distribuzioni asimmetriche, e sono sufficientemente flessibili da adattarsi alle interazioni e alle relazioni non lineari. Inoltre, molti aspetti dell'adattamento del modello sono stati automatizzati, quindi c'è "poca necessità di regolazioni da parte dell'imputatore" (Burgette and Reiter, 2010). L'idea di utilizzare i metodi CART per l'imputazione è stata suggerita da una vasta varietà di autori in vari modi. Si veda Saar-Tsechansky e Provost (2007) per una panoramica introduttiva. Alcuni ricercatori (He, 2006; Vateekul e Sarinnapakorn, 2009) semplicemente inseriscono la media o la moda. La maggior parte dei metodi di imputazione basati su alberi utilizzano una forma di imputazione singola basata sulla previsione (Barcena and Tusell, 2000; Conversano and Cappelli, 2003; Siciliano et al., 2006; Creel and Krotki, 2006; Ishwaran et al., 2008; Conversano and Siciliano, 2009). Sono stati sviluppati metodi di imputazione multipla da Harrell (2001), che lo ha combinato con la scalatura ottimale delle variabili di input, da Reiter (2005b) e da Burgette e Reiter (2010). Wallace et al. (2010) presentano un metodo di imputazione multipla che fa la media delle imputazioni per produrre un singolo albero e che non raggruppa le varianze. Parker (2010) indaga sui metodi di imputazione multipla per vari algoritmi di apprendimento non supervisionato e supervisionato.

Algoritmo 4.1

1. Viene estratto un campione bootstrap $(\bar{y}_{\text{obs}}, \bar{X}_{\text{obs}})$ di dimensione n_1 da $(y_{\text{obs}}, X_{\text{obs}})$.
2. Si adatta $\bar{y}_{\text{obs}}$ a $\bar{X}_{\text{obs}}$ con un modello ad albero $f(X)$.
3. Si prevedono i nodi terminali g_j da $f(X_{\text{mis}})$.
4. Si costruiscono n_0 insiemi Z_j di tutti i casi nel nodo g_j, ognuno contenente d_j donatori candidati.
5. Si estraggono casualmente un donatore i_j da Z_j per $j = 1, \ldots, n_0$.
6. Si calcola le imputazioni $y_j = y_{i_j}$ per $j = 1, \ldots, n_0$.

L'idea è identica al PMM, dove il "valore predittivo" è ora calcolato da un modello ad albero invece che da un modello di regressione. Come prima, l'incertezza dei parametri può essere incorporata adattando l'albero su un campione bootstrap. L'Algoritmo 4.1

descrive i principali passaggi di un algoritmo per creare imputazioni utilizzando un albero di classificazione o regressione. C'è una notevole libertà al passo 2, dove il modello ad albero viene adattato ai dati di addestramento $(\bar{y}_{\text{obs}}, \bar{X}_{\text{obs}})$. Potrebbe essere utile adattare l'albero in modo che il numero di casi in ogni nodo sia uguale a un certo numero predefinito, ad esempio 5 o 10. La composizione dei gruppi di donatori varierà su diverse ripetizioni del bootstrap, il che incorpora l'incertezza del campionamento sull'albero. Gli studi finora condotti si sono concentrati sulla precisione predittiva, che non è un criterio utile nel contesto dell'imputazione. Nessuno degli studi ha riportato statistiche di copertura. Il potenziale dei metodi basati su alberi e di altre tecniche di apprendimento automatico (Hastie et al., 2009) per creare imputazioni multiple adeguate deve ancora essere esplorato.

4.1.3 Approccio Joint Modeling

Il concetto di *joint modeling* nell'ambito dei metodi MICE si erge come una sofisticata procedura analitica che si prefigge di integrare il processo di imputazione dei dati mancanti con l'analisi statistica dei dati completi. Tale approccio avanzato si articola attorno alla creazione di una serie di modelli dedicati per ciascuna variabile contenente dati mancanti, dove tali modelli sono attentamente disegnati per catturare le intricate relazioni intercorrenti tra le variabili presenti nel dataset. Più in dettaglio, il processo di *joint modeling* implica la costruzione di modelli statistici multivariati in grado di tenere conto della complessa struttura dei dati e delle dipendenze esistenti tra le variabili. Questi modelli possono essere molteplici, e vengono impiegati per predire in modo ottimale i valori mancanti, sfruttando appieno le informazioni disponibili nel dataset.

L'obiettivo primario del *joint modeling* è quello di massimizzare l'accuratezza delle stime imputate, adottando un approccio integrato che incorpora le caratteristiche peculiari e le intricate interazioni presenti nei dati. Ciò consente di affrontare in maniera robusta e completa il problema dei dati mancanti, evitando distorsioni o perdite di informazioni rilevanti durante il processo di imputazione. Inoltre, il *joint modeling* permette di esplorare e comprendere in modo più approfondito la struttura e le dinamiche sottostanti dei dati, consentendo di ottenere insight più ricchi e informativi sul fenomeno oggetto di studio. Questa complessa metodologia rappresenta quindi un notevole balzo in avanti nell'analisi dei dati mancanti, garantendo risultati più precisi e affidabili nelle analisi statistiche condotte.

Capitolo 5

Caso in esame

5.1 Il dataset in esame

Il dataset utilizzato per l'analisi riguardante l'imputazione di dati mancanti in questa tesi comprende dati di rilevazione di elementi chimici di quattro depositi di solfuri massivi vulcanogenici (VMS), localizzati nelle Liguridi Interne delle ofioliti dell'Appennino settentrionale, precisamente nella regione della Liguria (C.Benedetti, 2023). I depositi considerati comprendono varie tipologie di elementi chimici ospitati in rocce mafiche e ultramafiche che si sono formati in un ambiente oceanico. Le rocce mafiche e ultramafiche sono due tipi di rocce ignee, cioè rocce che si sono formate dal raffreddamento e dalla solidificazione del magma. Le rocce mafiche sono caratterizzate da un'elevata quantità di minerali ferromagnesiani, come ad esempio il pirosseno e l'olivina. Esse sono di solito di colore scuro e presentano una densità elevata. Le rocce ultramafiche, invece, contengono una quantità ancora maggiore di minerali ferromagnesiani, con prevalenza di olivina e/o piroxeni. Sono le rocce più ricche in magnesio e contengono spesso minerali come serpentino e talco. Entrambe le tipologie di rocce si formano generalmente in ambienti geologici particolari, come ad esempio nelle zone di rift oceanico o all'interno dei mantelli terrestri e sono associate a processi geologici come l'attività vulcanica e la solidificazione del magma nelle profondità della crosta terrestre.

I depositi di solfuri massivi vulcanogenici sono caratterizzati principalmente dalla presenza di minerali come pirite, calcopirite, sfalerite e galena. Questi depositi si trovano in ambienti tettonici divergenti e convergenti e sono di fondamentale importanza per

l'associazione di metalli Fe-Cu-Zn, offrendo preziose informazioni sui processi geologici che hanno portato alla loro formazione.

Ogni rilevazione contiene informazioni su diversi elementi chimici, ma non tutte le variabili sono sempre state rilevate, il che ha portato alla presenza di dati mancanti. È importante notare che i dati mancanti sono distribuiti in modo sparso nel dataset, il che significa che in ogni rilevazione potrebbero mancare uno o più elementi chimici e questi elementi mancanti possono variare da rilevazione a rilevazione.

In dettaglio, il dataset è composto da 577 osservazioni. Sono presenti otto variabili categoriali, dove oltre alla variabile identificativa del campione sono disponibili anche variabili come la location in cui è stato raccolto il campione o il tipo del substrato del terreno. Le variabili qualitative non presentano dati mancanti. Le variabili categoriali forniscono informazioni aggiuntive sui contesti in cui sono stati effettuati i rilevamenti e possono includere, ad esempio, informazioni sulla posizione geografica, sul periodo di tempo in cui è stato effettuato il rilevamento, sul metodo di estrazione del campione o la sua profondità. È importante notare che una eventuale presenza di dati mancanti in queste variabili può influenzare direttamente sull'accuratezza delle analisi chimiche e delle valutazioni scientifiche relative ai campioni in esame. Gli elementi chimici coinvolti in queste rilevazioni sono tredici: oro (Au), argento (Ag), arsenico (As), cobalto (Co), molibdeno (Mo), nichel (Ni), antimonio (Sb), selenio (Se), stagno (Sn), rame (Cu), ferro (Fe), zinco (Zn), piombo (Pb). Due elementi non presentano dati mancanti, vale a dire cobalto (Co) e zinco (Zn). È inoltre presente un'altra variabile numerica senza dati mancanti cioè la profondità a cui è stato estratto il campione (Depth)

Il trattamento dei dati mancanti in questo contesto è cruciale per garantire l'affidabilità e la validità delle analisi condotte all'interno dataset. Pertanto, questa ricerca si concentra sull'applicazione di metodi di imputazione che tengano conto della struttura e della distribuzione dei dati nel dataset, nonché delle caratteristiche specifiche delle variabili coinvolte, al fine di produrre stime accurate e attendibili dei valori mancanti.

Per comprendere meglio le caratteristiche del dataset in esame, verranno inizialmente forniti alcune statistiche descrittive che riassumono le principali informazioni riguardanti le variabili presenti nel dataset. La Tabella 5.1 riassume i principali risultati delle analisi descrittive.

	Depth (m)	*Au (ppb)*	*Ag (ppm)*	*As (ppm)*	*Co (ppm)*	*Mo (ppm)*	*Ni (ppm)*
Media	2730.652	3005.308	81.549	1431.317	539.142	73.825	175.393
Dev. Standard	806.255	8145.685	228.352	14550.689	979.857	96.732	544.843
Minimo	1443.000	0.020	0.220	0.100	0.500	0.000	0.000
Massimo	4960.000	93600.000	1920.000	261000.000	9780.000	1040.000	4500.000

	Sb (ppm)	*Se (ppm)*	*Sn (ppm)*	*Cu (wt%)*	*Fe (wt%)*	*Zn (wt%)*	*Pb (wt%)*
Media	25.116	139.749	83.753	7.372	29.486	6.320	0.110
Dev. Standard	59.473	212.155	260.149	10.696	13.020	12.791	0.639
Minimo	0.064	0.050	0.200	0.020	0.390	0.000	0.000
Massimo	521.000	1505.000	1720.000	59.540	64.400	57.900	7.100

Tabella 5.1: Statistiche descrittive delle principali caratteristiche del campione

Nel contesto della stima dei dati mancanti, è fondamentale considerare i valori massimi delle variabili coinvolte nel dataset. I valori massimi rappresentano i massimi teorici o pratici che una variabile può assumere, e superarli potrebbe essere indicativo di anomalie o errori nei dati.

Ad esempio, in questo dataset, le variabili di interesse hanno dei valori massimi, che corrispondono ai limiti superiori delle loro gamme di misura. Ottenere una stima oltre questi valori massimi sarebbe anomalo e potrebbe indicare problemi di precisione o di validità nei metodi di imputazione utilizzati.

Per comprendere meglio i valori massimi delle variabili nel nostro dataset, si riportano in Tabella 5.2 che mostra i valori massimi per ciascuna variabile.

Au	*Ag*	*As*	*Co*	*Mo*	*Ni*	*Sb*	*Se*	*Sn*	*Cu*	*Fe*	*Zn*	*Pb*
(ppb)	*(ppm)*	*(ppm)*	*(ppm)*	*(ppm)*	*(ppm)*	*(ppm)*	*(ppm)*	*(ppm)*	*(wt%)*	*(wt%)*	*(wt%)*	*(wt%)*
200000	2000	5000	20000	1000	2000	1000	2000	2000	60	60	60	8

Tabella 5.2: Valori massimi ammissibili

Prima di procedere con l'analisi dei dati mancanti, è importante valutare la distribuzione e la frequenza delle osservazioni mancanti per ciascuna variabile nel dataset. Questo permette di comprendere meglio la natura e l'estensione degli stessi e di pianificare strategie appropriate per gestirli.

Nella Tabella 5.3, verrà presentato il numero totale di osservazioni mancanti per ogni variabile, insieme alla corrispondente frequenza relativa rispetto al numero totale di osservazioni nel dataset.

Variabile	*Frequenza assoluta*	*Frequenza relativa*
Cu	1	0.0017
Fe	9	0.0156
Pb	9	0.0156
Ag	13	0.0225
Ni	15	0.0260
Au	16	0.0277
As	39	0.0676
Sb	42	0.0728
Se	48	0.0832
Mo	60	0.1040
Sn	211	0.3657

Tabella 5.3: Frequenza assoluta e relativa dei dati mancanti per gli elementi chimici

Si può notare che il numero di dati mancanti varia notevolmente tra le variabili, con percentuali che vanno dallo 0.2% al 37%. Questo ampio range di valori suggerisce la presenza di differenze importanti nella completezza delle informazioni raccolte per ciascuna variabile.

Nel contesto dell'analisi dei dati, comprendere le relazioni esistenti tra le variabili numeriche è fondamentale per cogliere appieno le dinamiche sottostanti il dataset. Analizzare le correlazioni tra le variabili può fornire preziose informazioni sulle interazioni tra di esse, aprendo la strada a nuove prospettive di analisi e interpretazione.

Dopo aver esplorato le caratteristiche delle variabili del dataset, è giunto il momento di investigare più approfonditamente le relazioni che possono esistere tra di loro. Per fare ciò, verrà presentato un grafico che illustra le correlazioni tra tutte le variabili numeriche in questo dataset.

Il grafico delle correlazioni offre una rappresentazione visiva delle associazioni lineari tra le variabili, permettendo di individuare pattern e tendenze significative nei dati. Questo strumento aiuta a comprendere meglio come le diverse variabili si influenzino reciprocamente e quale grado di relazione esista tra di esse.

Nella Figura 5.1 è riportato il grafico delle correlazioni tra variabili numeriche. Questo grafico rappresenta i coefficienti di correlazione di Pearson tra tutte le coppie di variabili nel nostro dataset.

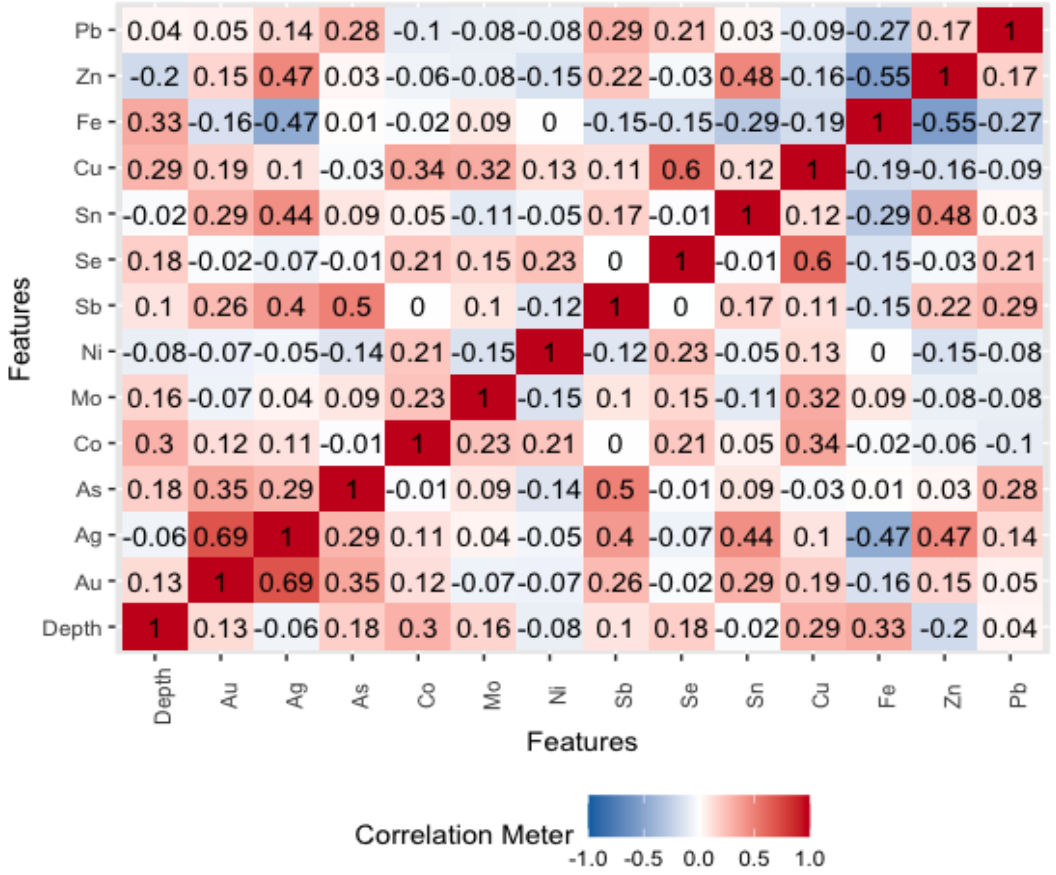

FIGURA 5.1: Correlazione tra le variabili numeriche presenti nel dataset.

L'analisi delle correlazioni permetterà di individuare le relazioni più significative tra le variabili numeriche e di identificare le variabili più rilevanti per le successive analisi. Questo fornirà una base solida per approfondire ulteriormente la comprensione del dataset e per sviluppare modelli analitici più efficaci.

5.2 Metodo MICE

Nei seguenti paragrafi verranno presentati due scenari differenti, quello in cui il processo generatore dei dati mancanti risulta essere MAR e quello in cui risulta essere NMAR. Verrà quindi, per ciascun caso, simulato un dataset utilizzando le distribuzioni marginali delle variabili numeriche originarie.

La simulazione del dataset e l'implementazione dei due metodi sono stati condotti con l'obiettivo di riprodurre un approccio accurato per il calcolo del Mean Squared Error (MSE) dei modelli. Dal momento che non era possibile valutare le performance del modello sul dataset originale, dato che non erano disponibili le vere rilevazioni riferite ai dati mancanti, si è optato per un'alternativa strategica. Questa procedura è stata adottata al fine di confrontare le prestazioni dei diversi modelli di stima del metodo MICE, nonché per confrontare il metodo MICE con gli approcci di imputazione singola. L'implementazione di questa metodologia consente di ottenere una misura quantitativa dell'errore, fornendo così una base solida per il confronto e la valutazione dei modelli di imputazione dati.

5.2.1 Caso *Missing At Random*

Per simulare il caso MAR, è stato generato un nuovo dataset utilizzando le distribuzioni marginali delle variabili numeriche presenti nel dataset originale. Questo nuovo dataset è stato creato con 1000 osservazioni attraverso il processo di reinserimento, che consiste nel campionare casualmente le osservazioni dal dataset originale con sostituzione. In questo modo, è stato possibile mantenere le stesse caratteristiche delle variabili nel dataset originale, consentendo una simulazione realistica delle situazioni in cui i dati mancanti sono osservati in modo casuale rispetto alle variabili note.

Il processo di reinserimento è stato eseguito in modo tale da mantenere la struttura delle variabili numeriche, preservando le loro distribuzioni marginali. Questo approccio ha permesso di simulare un dataset che riflettesse fedelmente le relazioni tra le variabili presenti nel dataset originale. Utilizzando questo nuovo dataset simulato, si è in grado di esaminare il comportamento dei metodi di imputazione dei dati mancanti in condizioni controllate, consentendo una valutazione accurata delle prestazioni di tali metodi sotto il presupposto di MAR.

Una volta ottenuto il dataset simulato, è stato necessario creare la stessa proporzione di dati mancanti del dataset originale per ogni variabile. A tale scopo, sono state selezionate casualmente le osservazioni da eliminare, garantendo che la proporzione di dati mancanti fosse coerente con quella del dataset originale. Per aumentare ulteriormente l'aleatorietà del processo, questa operazione è stata ripetuta tre volte, eliminando ogni volta osservazioni diverse. In questo modo, sono stati ottenuti tre dataset distinti, ciascuno con una proporzione di dati mancanti uguale a quella del dataset originale, ma con osservazioni mancanti disposte in modo diverso. Utilizzando tutti e tre i dataset per le analisi, è stata effettuata la media dei risultati ottenuti per gli errori dei modelli. Questo approccio ha permesso di ottenere una stima più robusta delle prestazioni dei metodi di imputazione dei dati mancanti, considerando la variabilità introdotta dalle diverse configurazioni di dati mancanti.

5.2.2 Caso *Missing Not At Random*

Dopo aver esaminato il caso NMAR e compreso l'importanza di simulare questa situazione, è stato necessario sviluppare un approccio accurato per generare un dataset simulato che rispecchiasse le caratteristiche di questa condizione. Per raggiungere questo obiettivo, è stato adottato un approccio che si basa sull'utilizzo di un modello binomiale per valutare la probabilità che ciascuna osservazione sia mancante, consentendo così di selezionare in modo selettivo le osservazioni da eliminare dal dataset simulato.

Il primo passo consisteva nell'allenare il modello binomiale utilizzando il dataset originale. Questo processo ha permesso di creare una variabile dicotomica per ciascuna variabile numerica nel dataset, che indicasse se un dato fosse mancante o meno. Allenando il modello su dati reali, si è in grado di catturare le relazioni complesse tra le variabili e la mancanza di dati, rendendo il modello più adatto per fare previsioni sui dati simulati.

Successivamente, è stato utilizzato il modello allenato per effettuare previsioni sul dataset simulato, valutando la probabilità che ciascuna osservazione fosse mancante. Questo ha permesso di identificare le osservazioni con una maggiore probabilità di essere mancanti e quindi di selezionare le osservazioni da eliminare dal dataset simulato.

È importante sottolineare che il processo di selezione delle osservazioni da eliminare è stato guidato dalle previsioni del modello binomiale, il quale ha considerato le relazioni

complesse tra le variabili nel dataset originale. In questo modo, è stato possibile simulare una situazione NMAR realistica, in cui la probabilità che i dati fossero mancanti dipendeva dai valori non osservati delle variabili stesse.

In conclusione, il processo di simulazione per il caso NMAR ha richiesto un'attenta pianificazione e un'accurata implementazione per garantire la creazione di un dataset simulato che rispecchiasse fedelmente le caratteristiche di questa condizione.

5.2.3 Valutazione della performance metodo MICE

Per il caso MAR, gli MSE sono stati calcolati utilizzando le osservazioni simulate eliminate per creare i dati mancanti. In pratica, dopo aver generato il dataset simulato, è stato selezionato casualmente un certo numero di osservazioni da eliminare per ciascuna variabile, mantenendo proporzioni simili di dati mancanti rispetto al dataset reale. Questo ha permesso di ottenere tre dataset MAR simulati, ognuno con una diversa configurazione di dati mancanti.

Successivamente, i modelli MICE sono stati stimati utilizzando i tre dataset simulati e calcolato gli MSE per ciascuna variabile e per ciascun metodo di stima. In particolare, sono state eseguite tre iterazioni per ogni variabile, ognuna utilizzando un diverso metodo di stima per i modelli MICE (PMM, Cart e Random Forest). Una volta ottenuti gli MSE per i tre dataset del caso MAR, è stata calcolata la media per presentare una singola tabella degli errori di previsione per ogni variabile, rappresentando così l'efficacia dei modelli MICE nel gestire i dati mancanti nel contesto di una situazione MAR.

L'obiettivo finale di questa analisi è stato determinare quale dei tre metodi di stima si dimostrasse migliore per quel tipo di dataset, presupponendo di trovarsi nel caso MAR. Questo ha permesso di individuare il metodo più adatto per affrontare i dati mancanti e di fornire raccomandazioni valide per situazioni simili in futuro.

La Tabella 5.4 evidenzia, attraverso gli errori MSE, quali siano i metodi di stima migliori per ogni singolo elemento chimico, nel caso in cui ci trovassimo in una situazione MAR

	PMM	CART	RF
Au	**112909058,380**	140193154,257	171038163,330
Ag	99479,103	**6158,190**	40214,743
As	192000,860	159702,563	**112742,647**
Co	1821610,957	1768889,477	**1533750,007**
Mo	**10724,640**	11442,590	11399,440
Ni	**96622,870**	103468,033	106821,940
Sb	**6889,780**	7668,063	7497,797
Se	**83204,113**	95697,527	92366,713
Sn	62640,833	60702,220	**59575,183**
Cu	**4,207**	12,027	11,703
Fe	**23,667**	24,890	29,340
Zn	24,673	42,947	**20,603**
Pb	0,130	0,123	**0,110**
Media	**8867868,016**	10954381,762	13307891,812

Tabella 5.4: MSE per ogni variabile stimata per ogni metodo di stima nel caso MAR

Durante la fase di valutazione della performance del modello MICE, è stata estesa l'analisi al caso NMAR. Questa volta, invece di simulare più dataset con diverse configurazioni di dati mancanti, ci si è concentrati su un unico dataset simulato, utilizzando le distribuzioni marginali delle variabili del dataset originale.

Per creare il dataset simulato nel contesto NMAR, è stato adottato un modello binomiale per stimare la probabilità che ciascuna singola osservazione fosse mancante. Questo modello è stato addestrato sul dataset originale, generando una variabile binaria per ogni variabile numerica che indica se il dato è mancante o meno. Una volta addestrato il modello, è stato applicato al dataset simulato per prevedere quali osservazioni avrebbero avuto una maggiore probabilità di essere eliminate.

Successivamente, sono state eliminate le osservazioni con le probabilità di mancanza più elevate per ogni variabile, mantenendo proporzioni simili di dati mancanti rispetto al dataset reale. Questo ha permesso di ottenere un dataset simulato NMAR, prontamente utilizzato per valutare i modelli MICE.

I modelli ottenuti con il metodo MICE sono stati stimati utilizzando il dataset simulato NMAR e calcolato gli MSE per ogni variabile e per ciascun metodo di stima.

In questo caso è stata ottenuta la tabella degli errori di previsione per ogni variabile, rappresentando l'efficacia dei modelli MICE nel contesto NMAR, non più come media di tre tabelle degli errori derivati da tre dataset differenti.

L'obiettivo finale di questa fase di valutazione è stato determinare quale dei tre metodi di stima si fosse dimostrato più efficace per il tipo di dataset che presumibilmente si trova nel caso NMAR. Questo ha permesso di identificare il metodo più adatto per affrontare i dati mancanti in situazioni non casuali, fornendo così indicazioni preziose per le future analisi di dati simili.

La Tabella 5.5 evidenzia, attraverso gli errori MSE, quali siano i metodi di stima migliori per ogni singolo elemento chimico, nel caso in cui ci trovassimo in una situazione NMAR

	PMM	**CART**	**RF**
Au	**74172521,18**	252824690,23	193493554,11
Ag	**216041,86**	219638,31	353573,98
As	320637,33	270242,69	**83635,58**
Co	3390589,49	**2708902,17**	3006830,14
Mo	**8856,90**	10113,53	10998,67
Ni	298188,06	134280,02	**103619,55**
Sb	**2213,85**	4525,80	3333,33
Se	94943,08	82232,78	**66188,90**
Sn	98623,91	**45117,80**	69006,08
Cu	**0,96**	1,28	3,14
Fe	**32,25**	38,66	40,60
Zn	28,05	31,04	**25,12**
Pb	0,08	**0,01**	**0,01**
Media	**6046359,77**	19715370,33	15168523,79

Tabella 5.5: MSE per ogni variabile stimata per ogni metodo di stima nel caso NMAR

Durante la fase di valutazione della performance dei modelli ottenuti con il metodo MICE, è stato considerato il calcolo di un singolo errore numerico rappresentativo per ciascun metodo di stima. Questo errore è stato ottenuto come la media degli Errori Quadratici Medi (MSE) di ogni singolo elemento chimico per quel metodo.

Dopo aver stimato i modelli utilizzando il dataset simulato NMAR, sono stati calcolati gli MSE per ogni variabile e per ciascun metodo di stima. Successivamente, è stato calcolato un valore medio di MSE per ogni metodo di stima, calcolato come la media degli MSE di tutti gli elementi chimici considerati.

5.2.3.1 Mistura di metodi di stima

Nella valutazione dei modelli MICE, potrebbe non sempre emergere un unico metodo di stima come il più efficace per gestire i dati mancanti. Nel caso in esame, dai risultati delle Tabelle 5.4 e 5.5 emerge chiaramente come non esista un metodo di stima più efficace rispetto ad altri. A seconda dell'elemento chimico considerato, la scelta cadrebbe su metodi diversi, soprattutto nell'ipotesi NMAR. Inoltre, per uno stesso elemento, la scelta cambierebbe a seconda della situazione MAR o NMAR. Per questo motivo invece di limitarci a una singola strategia, si potrebbe considerare un approccio più sofisticato, fondato sull'idea di combinare più metodi di stima al fine di ottenere risultati più robusti e affidabili.

Questa prospettiva introduce un elemento di flessibilità e adattabilità nel processo di gestione dei dati mancanti. Piuttosto che vincolarci rigidamente a una singola tecnica, si potrebbe sfruttare la diversità dei metodi disponibili per cogliere al meglio la complessità dei dati e ottenere stime più accurate. Questo approccio può essere particolarmente utile in contesti in cui le variabili presentino eterogeneità o complessità intrinseche, rendendo difficile individuare un unico metodo di stima che si adatti a tutte le situazioni.

La combinazione di metodi di stima offre anche un'opportunità per mitigare i potenziali limiti o debolezze di ciascun metodo singolo. Ad esempio, un metodo potrebbe eccellere nella gestione di determinati tipi di dati mancanti, mentre un altro potrebbe essere più adatto per altre situazioni. Sfruttando la complementarietà tra i diversi metodi, si potrebbe superare le limitazioni individuali e ottenere una soluzione più completa e bilanciata.

Tuttavia, la scelta dei metodi da combinare non è un compito banale e richiede un'attenta valutazione delle caratteristiche del dataset e delle specifiche esigenze dell'analisi. È importante considerare fattori come la natura dei dati mancanti, la distribuzione delle variabili, e le proprietà del modello di previsione. Inoltre, è necessario sviluppare criteri chiari e trasparenti per la selezione dei metodi, al fine di garantire la coerenza e l'affidabilità dei risultati.

In conclusione, l'approccio della combinazione di metodi di stima offre un modo promettente per affrontare la complessità dei dati mancanti nei modelli utilizzati per il metodo MICE. Integrando diverse tecniche e sfruttando la diversità dei metodi disponibili, è possibile migliorare la qualità delle previsioni e ottenere una rappresentazione più accurata e completa dei dati.

5.3 Modelli di imputazione singola

Nella seconda fase di questa analisi, ci si immerge nel complesso mondo dell'imputazione dei dati mediante l'utilizzo di modelli di machine learning. Questo approccio sofisticato abbraccia la complessità intrinseca dei dati e sfrutta la potenza dei modelli predittivi per affrontare la sfida dei dati mancanti in modo più avanzato e accurato. Tale processo è strutturato in diverse fasi, ciascuna delle quali mira a integrare e ottimizzare l'uso delle informazioni disponibili per stimare con precisione i valori mancanti.

La prima fase di questa esplorazione prevede un'attenta analisi della variabile con il minor numero di dati mancanti. Questa variabile diventa il fulcro delle nostre indagini e viene utilizzata come target primario da predire nei modelli di machine learning. Per costruire tali modelli predittivi, vengono sfruttate tutte le variabili che non presentano dati mancanti come covariate. Questo approccio consente di massimizzare l'utilizzo delle informazioni disponibili e di catturare al meglio le relazioni complesse tra le variabili nel nostro dataset.

L'allenamento dei modelli avviene su un dataset costituito dalle covariate selezionate e dalla variabile target, escludendo le righe contenenti dati mancanti. Questo permette di garantire un'analisi accurata e consistente, concentrando l'attenzione sui dati completi e preservando l'integrità delle informazioni durante il processo di predizione.

Per valutare le prestazioni dei modelli e determinare i parametri di regolazione ottimali, viene adottata una strategia di convalida incrociata leave-one-out (LOOCV). Questo approccio consente di valutare l'efficacia dei modelli su dati non visti, riducendo il rischio di sovrapprendimento e garantendo una stima affidabile degli errori di predizione.

Durante il processo di LOOCV, vengono calcolati gli MSE per ciascun modello e se ne confrontano le prestazioni. Il modello che mostra il minor MSE viene selezionato

come il più adatto per stimare l'osservazione mancante, fornendo una stima accurata e robusta che riflette al meglio le relazioni presenti nei dati.

Una volta stimati i valori mancanti per la variabile in esame, si aggiorna dataset ridotto includendo la variabile appena stimata. Questo processo viene iterato per ogni variabile con dati mancanti, aggiungendo una variabile alla volta e stimando i valori mancanti utilizzando il modello ottimale identificato nella fase precedente.

Questa metodologia consente di affrontare in modo graduale e sistematico la sfida dei dati mancanti, integrando progressivamente le informazioni disponibili e sfruttando la potenza dei modelli di machine learning per ottenere stime accurate e affidabili. Alla fine di questo processo, il dataset sarà completo, privo di valori mancanti e pronto per ulteriori analisi e interpretazioni.

Il completamento del primo ciclo di imputazioni segna l'inizio di una nuova fase del processo di analisi. Si consideri quindi l'idea di un secondo giro di imputazioni, che si distingue per una maggiore complessità e ricchezza concettuale. Ancora una volta il punto di partenza è la selezione della variabile con il minor numero di dati mancanti. Tuttavia, questa volta si adatta un approccio più sofisticato, in cui sfrutta appieno le conoscenze acquisite nel ciclo precedente.

Al centro di questo secondo ciclo di imputazioni c'è una strategia di ottimizzazione dell'informazione, volta a massimizzare l'utilizzo delle stime già ottenute. Si rimuovano le osservazioni con dati mancanti e utilizzano le variabili senza dati mancanti come covariate nei modelli. Tuttavia: invece di considerare solo le variabili con dati completi, si includono ora anche quelle che presentano dati mancanti, sostituendo tali valori mancanti con le imputazioni generate nel primo ciclo.

Questa metodologia, sebbene complessa, si dimostra incredibilmente potente nel capitalizzare l'informazione già disponibile. Integrando le stime dei dati mancanti ottenute precedentemente, si arricchisce ulteriormente il processo di imputazione, consentendo una migliore modellazione e una maggiore precisione nelle stime.

Il concetto chiave qui è la ricorsività: ogni ciclo di imputazioni si basa sulle stime del ciclo precedente, creando un feedback loop dinamico che guida il processo verso una convergenza verso stime sempre più accurate e affidabili. Questa iterazione continua e la costruzione di modelli sempre più sofisticati consentono di affrontare con successo la complessità dei dati mancanti e di ottenere risultati di alta qualità.

Con questo secondo ciclo di imputazioni, ci si immerge ancora più a fondo nel cuore del problema e si adatta un approccio avanzato e multifase per gestire la sfida dei dati mancanti.

Quest'operazione viene ripetuta per un totale di quattro volte, con l'obiettivo è di massimizzare l'utilizzo delle informazioni disponibili, allenando i modelli su un numero crescente di osservazioni e cercando di ottenere stime sempre più precise dei dati mancanti. Con ogni iterazione, il processo di imputazione si evolve e si affina, consentendo di ottenere una comprensione sempre più dettagliata e accurata del dataset sottostante.

Prima di procedere con ulteriori analisi, viene presentata una tabella che riporta gli errori per ogni metodo di stima utilizzato nel primo giro di imputazioni. Questa tabella fornisce una panoramica dei risultati ottenuti per ciascuna variabile, consentendoci di valutare le prestazioni dei diversi metodi di imputazione in relazione a ciascuna caratteristica del dataset. (Tabella 5.6)

	Cu	**Fe**	**Pb**	**Ag**	**Ni**	**Au**	**As**	**Sb**	**Se**	**Mo**	**Sn**
Lineare Stepwise	74.372	92.625	0,343	21431,88	103098,9	50578232	189783950	1.390,489	23497,97	6.739,28	27754.19
Ridge	59.518	72.993	0,336	18486,19	37719,28	33974440	182347325	1.372,166	18120,44	5.223,91	15432.66
Lasso	58.683	73.074	0,337	18512,45	37746,06	33825760	185695979	1.386,380	18199,41	5.245,63	14416.77
GAM Stepwise	50.132	68.160	0,282	16028,06	28306,74	36698079	143261421	582,262	17458,42	5.097,32	11546.84
Mars	59.149	81.389	0,326	211269,67	52456,43	39050394	440526931	1.792,030	25450,40	8.678,97	38116,12
Tree	69.128	79.621	0,191	**17164,82**	74138,24	41057273	249883401	1.573,127	24230,67	7.965,24	19859.23
Bagging	**55.425**	**66.636**	0,246	21344,31	**40608**	**30529435**	202092658	**1.298,302**	**17029,82**	**4.796,35**	**14436.82**
Random Forest	88.563	112.641	**0,228**	25393,76	85737,15	39240802	**172430344**	1.601,348	26605,75	6.622,00	20254.63
Rete Neurale	58.103	105.537	0,344	20829,22	107025,16	57561076	224752457	1.862,882	29870,10	6.679,64	42399.71

TABELLA 5.6: MSE per ogni variabile e ogni metodo relativi al primo giro di imputazioni

È interessante notare come, durante il processo di imputazione, il metodo di stima scelto per ciascuna variabile possa variare in base alla natura dei dati e all'adattamento dei modelli alla singola caratteristica. Tuttavia, analizzando i risultati, è possibile osservare che il metodo più frequentemente selezionato è il bagging, seguito dal random forest. Questo fenomeno potrebbe essere attribuibile al fatto che entrambi i metodi utilizzano campioni bootstrap, che contribuiscono ad aumentare la numerosità campionaria. In un contesto in cui si dispone di pochi dati, è comprensibile preferire modelli che fanno uso di tecniche che aumentino la dimensione del campione e migliorare la robustezza delle stime.

Viene ora presentata la tabella degli errori per ogni metodo di stima, relativa al secondo giro di imputazioni (Tabella 5.7). Come ci si poteva aspettare, si osserva un abbassamento degli errori per ogni variabile. Questo risultato è probabilmente derivato dal fatto che nel secondo giro di imputazioni vengono utilizzate più osservazioni e più variabili, consentendo ai modelli di apprendere da un maggior numero di informazioni e di generare stime più accurate.

	Cu	**Fe**	**Pb**	**Ag**	**Ni**	**Au**	**As**	**Sb**	**Se**	**Mo**	**Sn**
Lineare Stepwise	47.946	88.520	0.207	22123.90	105918.33	48494549	190974749	1.289,238	23748,15	6.699,34	24657,24
Ridge	40.921	70.695	0.170	19343.41	104408.39	33665003	178544989	1.336,972	18412,98	5.214,74	15604,16
Lasso	40.508	70.828	0.170	19537.20	104246.28	33598712	181863412	1.355,209	18431,16	5.233,63	14389,05
GAM Stepwise	27.860	55.181	0.072	13738.55	62737.86	30040406	110917871	847,677	17754,32	4.988,22	11639,09
Mars	34.667	60.874	0.057	62191.52	60045.50	50874861	11689773	2.930,450	25673,65	7.794,08	36499.35
Tree	61.495	80.062	0.103	28890.86	77026.11	39981456	250227386	1.452,757	23736,54	5.628,05	23768,03
Bagging	**33.534**	**60.747**	0.247	**21702.92**	**63432.80**	**31392275**	201097634	**1.315,772**	**17040,26**	**4.808,31**	16904,34
Random Forest	63.593	112.773	**0.129**	24953.05	92761.54	39836723	**146217042**	1.629,817	27210,63	6.640,48	**17361,28**
Rete Neurale	51.696	94.200	0.176	32725.69	134251.55	52459899	250533489	1.364,781	31458,67	7.427,32	28901,30

TABELLA 5.7: MSE per ogni variabile e ogni metodo relativi al secondo giro di imputazioni

È cruciale sottolineare che per alcune variabili, la selezione del metodo di stima definitivo per l'imputazione non coincide necessariamente con quello che registra il minor MSE. Tale fenomeno può essere attribuito a due fattori chiave. In alcuni scenari, i modelli adottati possono generare stime che eccedono i limiti superiori delle variabili, come precedentemente discusso. Ciò può comportare una selezione alternativa di metodi di imputazione, al fine di evitare l'inserimento di valori non plausibili o anormali. In altri casi, invece, le stime prodotte dai modelli possono presentare una ridotta varianza all'interno del campione di imputazioni. Ad esempio, potrebbero essere generate solo un limitato numero di imputazioni numeriche distribuite in modo sparso all'interno del gruppo di osservazioni mancanti, determinando una preferenza per metodi di stima più conservativi o meno sensibili alla variabilità dei dati.

Questo fenomeno si spiega in parte con il raggruppamento delle osservazioni mancanti per la stessa variabile, in cui i dati mancavano in modo contiguo o "a blocchi". Per esempio, in una specifica località o contesto, tutte le osservazioni relative a una variabile potevano presentare dati mancanti. In tal caso, l'utilizzo di variabili che mostravano valori costanti per quel blocco di osservazioni per le variabili categoriali e valori molto simili per le variabili numeriche avrebbe condotto alla produzione di stime identiche o

molto simili tra loro. Questo può riflettere una limitata variabilità dei dati nell'ambito di quel blocco, inducendo modelli di imputazione a generare stime omogenee o poco variabili per le osservazioni mancanti.

Riguardo i successivi due giri di imputazione, non verranno presentate le tabelle degli errori poiché si è osservato un fenomeno di stabilizzazione sia del metodo di stima per la stessa variabile, sia degli errori. Questo significa che, a partire dal terzo giro di imputazioni, i modelli di imputazione hanno raggiunto una certa stabilità nel selezionare il metodo di stima più appropriato per ciascuna variabile e nell'ottenere stime consistenti e affidabili. Di conseguenza, non si è ritenuto necessario riportare ulteriori tabelle degli errori, poiché i risultati erano sostanzialmente simili a quelli ottenuti nel secondo giro di imputazioni.

5.4 Risultati

Le analisi condotte nel corso di questa ricerca hanno fornito importanti informazioni sul processo di imputazione dei dati mancanti in dataset contenenti misurazioni di elementi chimici. Attraverso l'utilizzo di tecniche di modellazione avanzate e approcci sofisticati, siamo stati in grado di ottenere quattro dataset con variabili imputate, ciascuno adottando un approccio diverso.

Il primo dataset è stato ottenuto utilizzando l'algoritmo MICE con il metodo di imputazione Predictive Mean Matching il quale, considerando che per la maggioranza degli elementi chimici risulta essere il metodo di stima che genera l'errore più basso, ha dimostrato di essere il miglior metodo di imputazione sia per i dati mancanti MAR che per quelli NMAR. Questo approccio ha garantito un'elevata precisione nella stima dei valori mancanti, contribuendo a mantenere l'integrità e la coerenza del dataset.

Il secondo dataset è stato generato utilizzando il miglior metodo di stima del metodo MICE a seconda dell'elemento chimico analizzato nel caso NMAR, mentre il terzo dataset ha impiegato il miglior metodo di stima del metodo MICE a seconda dell'elemento chimico analizzato nel caso MAR. Entrambi questi dataset hanno consentito una buona imputazione dei dati mancanti, sebbene in presenza di differenti meccanismi di mancanza dei dati.

Infine, il quarto dataset è stato ottenuto utilizzando metodi di imputazione singola riconducibili al machine learning, selezionando per ogni variabile il metodo di imputazione ritenuto più appropriato. Questo approccio ha fornito una panoramica della varietà di metodi disponibili per l'imputazione dei dati mancanti, mostrando come diverse strategie possano essere adottate in base alla natura dei dati e al contesto specifico dell'analisi.

Sono state esaminate diverse strategie di imputazione. Ogni approccio ha i suoi vantaggi e svantaggi, e si ha avuto l'opportunità di sperimentare diverse tecniche per valutarne l'efficacia.

In particolare, il confronto dei risultati è basato sulle correlazioni come metrica di valutazione. Le correlazioni forniscono una misura della relazione tra le variabili nel dataset, e ci si aspettava che il metodo di imputazione preservasse queste relazioni nel modo più fedele possibile, in accordo con lo spirito del processo di imputazione

Questo approccio ha permesso di valutare non solo la precisione delle singole imputazioni, ma anche l'impatto complessivo sulle relazioni tra le variabili nel dataset. Alla luce di queste considerazioni, è possibile ora esaminare i risultati e trarre conclusioni significative sul metodo di imputazione più adatto allo specifico contesto di studio.

Per prima cosa, sono state calcolate le matrici delle correlazioni per ciascun dataset, sia quello originale che quelli stimati. Queste matrici sono strumenti fondamentali per valutare come le relazioni tra le variabili variano tra i diversi dataset. Successivamente, sono stati confrontati i dataset stimati con quello originale, analizzando la variazione delle correlazioni tra le variabili. Questo ha fornito una visione dettagliata di quanto i dataset stimati si discostassero dal dataset originale in termini di correlazioni tra variabili.

I risultati di questo confronto sono stati riassunti in una tabella (Tabella 5.8), che mostra la variazione media delle correlazioni per ogni variabile in ciascun dataset stimato rispetto al dataset originale. Questo riassunto fornisce un'indicazione chiara e concisa di come le correlazioni tra variabili siano cambiate con l'imputazione dei dati.

Per ottenere un'indicazione più comprensiva della variazione delle correlazioni, sono state calcolate la media delle differenze assolute tra le correlazioni di ciascun dataset stimato e quelle del dataset originale, come:

$$\text{Media delle differenze assolute} = \frac{1}{n}\sum_{i=1}^{n} |\rho_{\text{dataset stimato},i} - \rho_{\text{dataset originale},i}| \tag{5.1}$$

dove:

- n è il numero totale di coppie di variabili.
- $\rho_{\text{dataset stimato},i}$ è la correlazione tra le variabili nel dataset stimato per la i-esima coppia di variabili.
- $\rho_{\text{dataset originale},i}$ è la correlazione tra le variabili nel dataset originale per la i-esima coppia di variabili.

La tabella riassuntiva (Tabella 5.8) consente una rapida valutazione delle differenze nelle correlazioni tra i dataset, offrendo così un'importante prospettiva sull'efficacia delle diverse tecniche di imputazione nei dati.

	NMAR Mistura Stimatori	**MAR Mistura Stimatori**	**PMM**	**Imputazione singola**
Au	0,80692	0,29114	**0,22645**	0,23997
Ag	0,80734	0,30140	**0,27794**	0,28613
As	0,80898	0,30942	**0,21351**	0,22946
Co	0,92680	0,36673	0,21818	**0,16751**
Mo	0,91014	0,35007	0,18139	**0,14121**
Ni	0,93642	0,37635	0,21826	**0,17036**
Sb	0,74087	**0,29833**	0,30597	0,31440
Se	0,88419	0,35692	0,20657	**0,16952**
Sn	0,77364	0,27351	**0,26216**	0,27099
Cu	0,90489	0,37762	0,23680	**0,21438**
Fe	1,02022	0,46014	0,29147	**0,23464**
Zn	0,87799	0,34834	0,28250	**0,27972**
Pb	0,83182	0,28838	**0,21334**	0,21560
Media	0,86386	0,33833	0,24112	**0,22568**

TABELLA 5.8: Variazione media delle correlazioni per ogni variabile in ciascun dataset stimato rispetto al dataset originale

D'ora in avanti, per tutte le considerazioni e le analisi successive, il riferimento sarà esclusivamente il dataset creato utilizzando il metodo di imputazione PMM all'interno dell'algoritmo MICE. Questa decisione è motivata dal fatto che questo metodo si è dimostrato il migliore sia in situazioni in cui i dati mancanti seguono un meccanismo NMAR che MAR, come confermato dai risultati delle analisi condotte finora. Pertanto, i dataset ottenuti utilizzando le misture di stimatori saranno tralasciati nelle considerazioni successive.

Nella valutazione delle tecniche di imputazione, non ci si limiterà tuttavia solo all'analisi delle correlazioni tra variabili. Si ritiene fondamentale esaminare anche come cambia la struttura dei dati, inclusa la variazione di elementi chiave come la media e la deviazione standard.

Oltre all'analisi delle correlazioni, sono stati condotti degli studi approfonditi sui cambiamenti nella distribuzione dei dati nei vari dataset stimati rispetto al dataset

originale. Questo approccio ha permesso di esaminare come le tecniche di imputazione influenzino la tendenza centrale e la dispersione dei dati.

I risultati di questa analisi sono riassunti in Tabella 5.9 la quale illustra la variazione della media per ciascuna variabile in ciascun dataset stimato rispetto al dataset originale. Questa panoramica ha fornito un'ulteriore comprensione di come le tecniche di imputazione abbiano modellato la struttura complessiva dei dati.

L'analisi dei cambiamenti nella media, insieme all'analisi delle correlazioni, ha fornito una visione più completa dell'efficacia delle tecniche di imputazione e del loro impatto sulla struttura dei dati.

	Media originale	Media imputazione singola	Media PMM
Au	2915,30	3537,00	3318,21
Ag	80,56	101,17	82,69
As	1410,19	1365,23	1378,42
Co	512,11	735,41	542,64
Mo	72,41	67,49	68,33
Ni	154,04	93,18	156,47
Sb	23,68	45,06	27,48
Se	128,29	204,84	149,54
Sn	75,17	82,84	69,73
Cu	7,37	7,38	7,39
Fe	29,49	29,44	29,65
Zn	6,20	6,54	5,76
Pb	0,10	0,20	0,11

Tabella 5.9: Media delle variabili nei vari dataset

I risultati presentati nella Tabella 5.9 offrono interessanti spunti per valutare l'efficacia dei metodi di imputazione singola e del metodo PMM nel contesto dell'analisi dei dati. Emergono similitudini tra i due approcci, poiché entrambi tendono a fornire medie dei valori degli elementi chimici nei dataset imputati che sono confrontabili tra loro. Tuttavia, va notato che la variazione dalla media originale dei valori imputati varia considerevolmente a seconda dell'elemento chimico considerato. Ad esempio, per

l'oro (Au), la media ottenuta con il metodo di imputazione singola è più lontana alla media originale rispetto al metodo PMM, mentre per lo zinco (Zn), la situazione è inversa. Questo suggerisce che la scelta del metodo di imputazione può influenzare in modo significativo i risultati ottenuti.

È importante considerare che la variazione nella media dei valori imputati può essere influenzata da diversi fattori, tra cui la scala di valori che ciascuna variabile può assumere e il numero di dati mancanti per ciascuna variabile. Inoltre, altre caratteristiche non osservabili dei dati possono giocare un ruolo determinante nel processo di imputazione. Pertanto, questi risultati mettono in evidenza l'importanza di valutare attentamente i metodi di imputazione utilizzati e di considerare i fattori che possono influenzare la precisione delle stime ottenute.

Capitolo 6

Conclusioni

6.1 Conclusioni

Alla luce dei risultati emersi dall'analisi comparativa tra l'imputazione singola e i modelli utilizzati con il metodo MICE, emergono considerazioni importanti sulle prestazioni e sull'efficacia di entrambi, nonché sulle implicazioni delle scelte metodologiche nelle situazioni di dati mancanti.

In primo luogo, l'imputazione singola si distingue per la sua semplicità concettuale e per la facilità di implementazione, consentendo una rapida gestione dei dati mancanti attraverso la generazione di un unico valore sostitutivo per ciascuna osservazione mancante. Tuttavia, la sua efficacia può essere influenzata dalla scelta del modello di imputazione singola. In particolare, l'adozione di modelli che fanno uso del bootstrap si è rivelata cruciale in scenari in cui la numerosità ridotta del dataset ha costituito una sfida significativa. Infatti, l'incorporazione del bootstrap nei modelli di imputazione singola ha contribuito a mitigare gli effetti negativi della ridotta numerosità del dataset, consentendo di ottenere stime più robuste e affidabili. Tuttavia proprio l'utilizzo di questo metodo di stima, per la sua natura, nel caso in esame ha generato imputazioni che presentano poca variabilità all'interno del gruppo di valori ammissibili per ogni elemento, probabilmente dovuto al fatto che in alcuni casi i valori mancanti per un singolo elemento chimico risultavano mancanti a blocchi per una determinatà modalità di una delle variabili categoriali. È importante sottolineare che l'applicazione di questa tecnica a dataset di dimensioni maggiori potrebbe potenzialmente portare a risultati ancora più soddisfacenti, evidenziando così il ruolo cruciale delle strategie metodologiche nella gestione dei dati mancanti.

D'altra parte, il metodo MICE si distingue per la sua complessità e la sua capacità di catturare le relazioni complesse tra le variabili nei dati attraverso l'uso di più stime per ciascuna osservazione mancante e l'iterazione attraverso una serie di modelli. È proprio la generazione di più stime che in contesti come questo risulta fondamentale, al fine di garantire una maggiore variabilità nei valori imputati per ogni elemento chimico. Tuttavia, la sua efficacia può essere influenzata dalla scelta dei modelli utilizzati per le stime imputate, nonché dalle risorse computazionali e dal tempo di esecuzione necessari per l'analisi.

In conclusione, la scelta tra l'imputazione singola e il metodo MICE dipende dalle specifiche esigenze dell'analisi, dalle caratteristiche dei dati e dalle risorse disponibili. Mentre l'imputazione singola può essere preferibile per dataset con pochi dati mancanti e una struttura semplice, il metodo MICE può essere più adatto per dataset complessi con relazioni intricate tra le variabili, come ad esempio il caso in esame. Tuttavia, è importante considerare attentamente le opzioni disponibili e adottare un approccio che sia adeguato al contesto specifico dell'analisi dei dati, tenendo conto anche delle strategie metodologiche che possono migliorare le prestazioni dei modelli di imputazione.

Appendice

.0.1 Grafici delle correlazioni

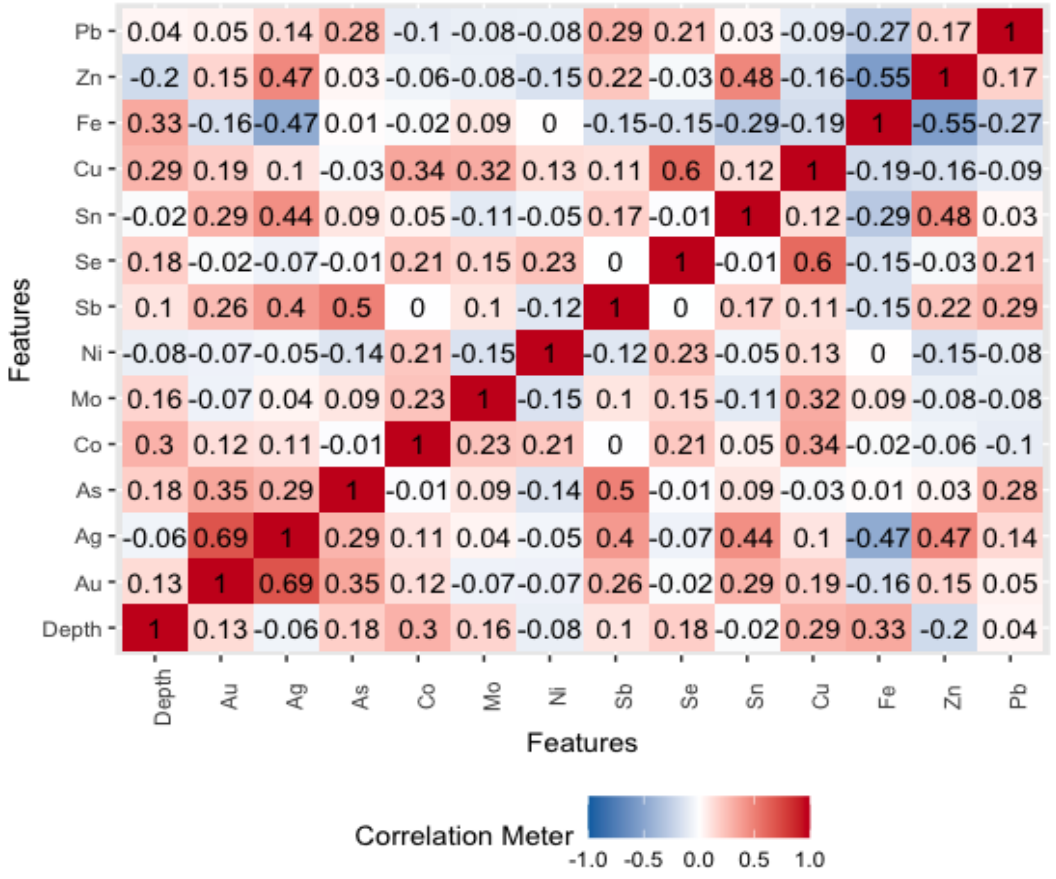

FIGURA .1: Correlazione tra le variabili numeriche presenti nel dataset.

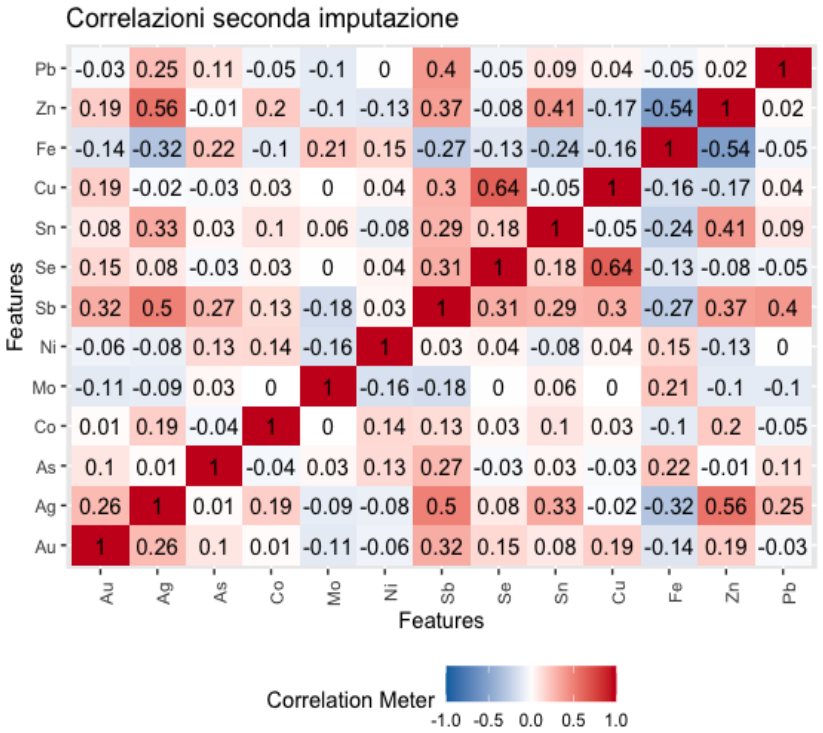

FIGURA .2: Correlazione tra le variabili numeriche presenti nel dataset ottenuco con imputazione semplice.

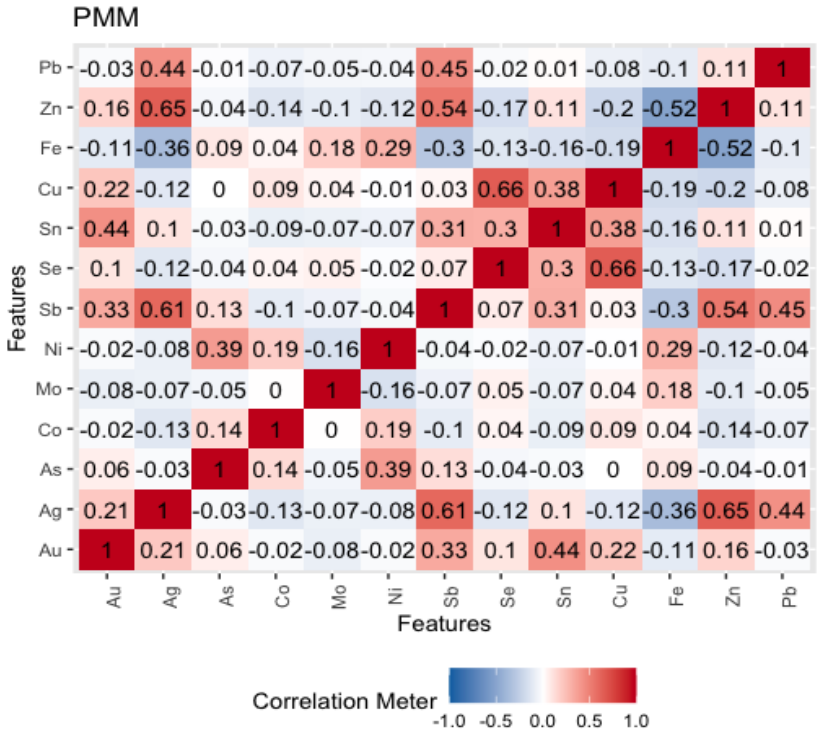

FIGURA .3: Correlazione tra le variabili numeriche presenti nel dataset ottenuto con stima PMM.

.0.2 Codice funzione LOOCV (Senza Selezione Del Parametro di Regolazione)

```
LOOCV <- function(y, x, pred) {
  set.seed(1234)
  k <- nrow(x)  # Numero di campioni, corrisponde al numero di fold (LOOCV)
  eqm <- rep(NA, k)

  for (i in 1:k) {
    tic()
    oof <- rep(FALSE, k)
    oof[i] <- TRUE
    prev <- pmax(pred(y[!oof], x[!oof,], x[oof,]),0)
    eqm[i] <- sum((y[oof] - prev)^2) / length(y[oof])
    cat("Iterazione", i, "di", k, "Tempo:")
    toc()
  }

  return(eqm)
}
```

.0.3 Codice funzione LOOCV (Selezione Del Parametro di Regolazione)

```
Codice funzione LOOCV  (Selezione Del Parametro di Regolazione)
LOOCV.reg <- function(y, x, pred, valori) {
  set.seed(1234)
  k <- nrow(x)  # Numero di fold corrisponde al numero di campioni

  eqmm <- rep(NA, length(valori))
  sd <- rep(NA, length(valori))
  cat("Valori i-esimi di", length(valori), ": ")

  for (q in 1:length(valori)) {
    eqm <- rep(NA, k)

    for (i in 1:k) {
      oof <- i  # Lascia fuori un singolo campione alla volta
      prev <- pmax(pred(y[-oof], x[-oof, ], x[oof, ], valori[q]),0)
      eqm[i] <- (y[oof] - prev)^2
    }

    eqmm[q] <- mean(eqm)
    sd[q] <- sd(eqm)
    cat("[", q, "]")
  }

  return(list(EQM = eqmm, SD = sd))
}
```

Bibliografia

Little, R.J.A., Rubin, D.B. *Statistical Analysis with Missing Data*. Wiley, New York, 2002.

Ford, D.J. *Hot Deck Methods*. Journal of the American Statistical Association, 78(382), 97–102, 1983.

Brick, J.M., Kalton, G. *Handling Missing Data in Survey Research*. Statistical Methods in Medical Research, 5(3), 215–238, 1996.

Koller-Meinfelder, F. *Methods of Imputation for Missing Data - Hot Deck Methods*. University of Mannheim, 2009.

Andridge, R.R., Little, R.J.A. *A Review of Hot Deck Imputation for Survey Non-response*. International Statistical Review, 78(1), 40–64, 2010.

De Waal, T., Pannekoek, J., Scholtus, S. *Flexible Imputation of Missing Data: Multiple Imputation for Repeated Measurement Data*. Chapman and Hall/CRC, 2011.

Rubin, D.B. *Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations*. Journal of Business Economic Statistics, 4(1), 87–94, 1986.

Little, R.J.A. *Missing-Data Adjustments in Large Surveys*. Journal of Business Economic Statistics, 6(3), 287–296, 1988.

Schenker, N., Taylor, J.M.G. *Partially Parametric Techniques for Multiple Imputation*. Computational Statistics Data Analysis, 22(4), 425–446, 1996.

Siddique, J., Belin, T.R. *Multiple Imputation Using an Incomplete Multinomial Outcome Measure*. Statistics in Medicine, 27(26), 5159–5175, 2008.

Heitjan, D.F., Little, R.J.A. *Multiple Imputation for the Fatal Accident Reporting System.* Journal of the American Statistical Association, 86(415), 91–99, 1991.

Rubin, D.B. *Matched Sampling for Causal Effects.* Cambridge University Press, 2006.

Adelchi Azzalini, Bruno Scarpa *Data Analysis and Data Mining.*

D'Orazio, V., Marco, M.D., Mignone, F., Salmaso, L. *File Matching Methods, Non-ignorable Non-response and Weighting to Adjust for Missing Data in the Second Survey on the Use of Information and Communication Technologies (ICT) in Enterprises.* Survey Methodology, 32(1), 77–88, 2006.

White, I.R., Royston, P., Wood, A.M. *Multiple Imputation Using Chained Equations: Issues and Guidance for Practice.* Statistics in Medicine, 30(4), 377–399, 2011.