Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica per l'Economia e l'Impresa

Analisi Statistica della Qualità Biologica del Suolo in Veneto

Relatore: Prof. Antonio Canale
Dipartimento di Scienze Statistiche

Laureando: Giulia Poidomani
Matricola: 1227361

Anno Accademico 2022/2023

Indice

Introduzione

L'elaborato nasce dall'esperienza di stage effettuata presso ARPAV: Agenzia Regionale per la Prevenzione e Protezione Ambientale del Veneto. In particolare presso l'unità organizzativa del suolo, che ha sviluppato una rete di monitoraggio della qualità biologica del suolo condotta dal 2012.

Il principale obiettivo che ci si è posti è determinare da cosa dipende e quali sono i fattori che influenzano maggiormente la qualità biologica del suolo in Veneto, misurata attraverso l'indice QBS-ar. A tale scopo verranno utilizzati modelli statistici di regressione lineare multipla normale con effetti casuali.

Nel *primo capitolo* si descrive cos'è la qualità biologica del suolo e la sua importanza. Si definiscono gli obiettivi dell'analisi e si presenta l'indice QBS-ar, usato per misurare la qualità biologica del suolo. Infine si spiega come quest'ultimo viene calcolato.

Nel *secondo capitolo* si introducono i metodi di campionamento e si illustra la disposizione dei punti di rilevazione nel territorio Veneto. Vengono presentati i dati e si descrive la fase di pulizia, dove si riscontra il problema di dati mancanti. Infine si presenta l'analisi esplorativa, osservando la distribuzione di tutte le singole variabili e poi quella della variabile risposta, ovvero il QBS-ar, condizionata alle esplicative.

Nel *terzo capitolo* si descrivono a livello teorico i metodi e i modelli statistici usati. In particolare si presenta il modello di regressione lineare multipla normale, descrivendo come esso è strutturato, quali sono le sue assunzioni e come avviene la stima dei parametri. Si introduce poi il modello a effetti casuali, un particolare modello per l'analisi di dati correlati. Infine si descrivono le procedure di selezione del modello forward e backward.

Nel *quarto capitolo* si presentano i modelli stimati e si spiega come essi vengono ottenuti. Si tratta poi il problema dei dati mancanti, risolto con la stima di un modello di regressione lineare multipla normale.

Capitolo 1

Definizione obiettivi

1.1 Qualità biologica del suolo

Il suolo è il punto nodale per gli equilibri ambientali che assicurano la continuità della vita sulla Terra e la salute del territorio. I servizi che esso fornisce all'ecosistema sono dovuti principalmente agli organismi viventi che lo popolano. Questi ultimi, infatti, svolgono un ruolo primario nei processi di formazione del suolo, nella decomposizione e trasformazione della sostanza organica, nel rilascio di elementi disponibili per piante e altri organismi, nel controllo del regime delle acque, nell'attenuazione della contaminazione chimica e biologica. Per questi motivi è importante tutelarne la biodiversità.
Il suolo è anche una risorsa non rinnovabile nel breve periodo ed estremamente fragile, che può essere soggetta a intensi processi degradativi essenzialmente legati all'antropizzazione. Poiché generalmente il degrado del suolo è un processo lento, che raramente comporta effetti drammatici immediati, i problemi sono evidenziati solo quando sono in uno stato avanzato o ad un grado tale da renderne estremamente oneroso e economicamente poco proponibile il ripristino. A questi livelli anche la capacità di adattamento e la fondamentale azione di mitigazione degli agenti inquinanti svolta dagli organismi che vivono nel suolo è notevolmente ridotta o, nei casi estremi, annullata. Dunque risulta fondamnetale il monitoraggio della qualità biologica del suolo.

1.2 Obiettivi dell'analisi

Il principale obiettivo dell'analisi è determinare da cosa dipende e quali sono i fattori che influenzano maggiormente la qualità biologica del suolo in Veneto. Determinare quest'ultima conoscendo le caratteristiche del suolo e i fattori esterni che lo condizionano è un ulteriore obiettivo che ci si pone.
Per indicare il livello di qualità biologica del suolo è stato usato l'indice QBS-ar, di cui se ne spiegheranno i dettagli nella Sezione 1.3, che si basa sulla presenza di microartropodi nel terreno e il loro adattamento alla vita in questo ambiente. I microartropodi sono molto sensibili alle alterazioni naturali o causate dall'uomo e agli equilibri chimico-fisici di questo ambiente; maggiore è il loro grado di adattamento al suolo, minore sarà la loro capacità di abbandonarlo quando si trova in condizioni sfavorevoli, per cui si possono considerare dei buoni indicatori del livello di disturbo del suolo.
Dunque ci si pone l'obiettivo di definire la relazione tra l'indice QBS-ar e le caratterstiche del suolo.

1.3 Che cos'è il QBS-ar?

Il QBS-ar (Qualità Biologica del Suolo, attraverso microartropodi) è un indice sintetico per la valutazione della qualità biologica del suolo. Si basa sul grado di adattamento anatomico alla vita nel suolo di microartropodi presenti nel terreno (insetti, aracnidi, miriapodi, crostacei), che vengono utilizzati come bioindicatori. Questi organismi presentano una serie complessa di adattamenti alla vita nell'ambiente edafico e si dimostrano sensibili allo stato di sofferenza di un suolo. Infatti se l'ecosistema suolo è indisturbato prevarranno i gruppi particolarmente adattati a questo ambiente, ossia di piccole dimensioni, depigmentati o con l'eventuale pigmentazione criptica per confondersi con le particelle di terra, privi di occhi e ali; se il suolo subisce impatti disturbanti, i gruppi più adattati tenderanno a scomparire mentre prevarranno quelli meno adattati.

Per ottenere l'indice QBS-ar si distinguono 5 fasi: prelievo del campione, estrazione e conservazione dei microartropodi, determinazione delle forme biologiche contenute e infine calcolo dell'indice QBS-ar.

Il metodo prevede, per ogni osservazione, la raccolta di tre zolle di terreno di dimensioni pari a 10 cm^3 (lettiera o copertura erbacea esclusa). Entro le 24 ore seguenti le tre zolle vengono posizionate nel selettore di Berlese-Tullgren (Figura 1.1) che consiste in un imbuto in cui viene posto un setaccio, sotto il quale vi è un recipiente di raccolta contenente una soluzione composta da due parti di alcol a 90 gradi e una parte di glicerina. A 20 cm al di sopra di esso è posizionata una moderata sorgente di calore, generalmente una lampada da 40 watt, che provoca lo spostamento progressivo della pedofauna attiva verso il basso per sfuggire all'essiccamento, fino a cadere nel recipiente. Nell'arco di 15-20 giorni viene estratto oltre il 95% degli animali (definiti in seguito "selettura").

Figura 1.1: Estrazione della selettura con il selettore di Berlese-Tullgren.

Successivamente viene svolta l'analisi al microscopio della selettura per il riconoscimento delle forme biologiche.

Il grado di adattamento delle forme biologiche alla vita nel suolo varia in base alla presenza e alla combinazione di alcuni caratteri e per quantificarlo si utilizza una scala di riferimento di punteggi chiamata EMI (Eco-Morphological Index): per ogni carattere che evidenzia l'adattamento al suolo si attribuisce un punteggio, da un minimo di 1 ad un massimo di 20, a seconda che la forma considerata sia pochissimo o decisamente adattata al suolo (Figura 1.2). Quando in un campione di pedofauna prelevato dal suolo sono presenti diverse forme biologiche appartenenti allo stesso gruppo, si tiene conto solamente del valore di EMI più alto riscontrato.

L'Indice di Qualità Biologica del Suolo (QBS-ar) è un punteggio totale attribuito a un campione di terreno, dato dalla somma di tutti i valori dei singoli EMI.

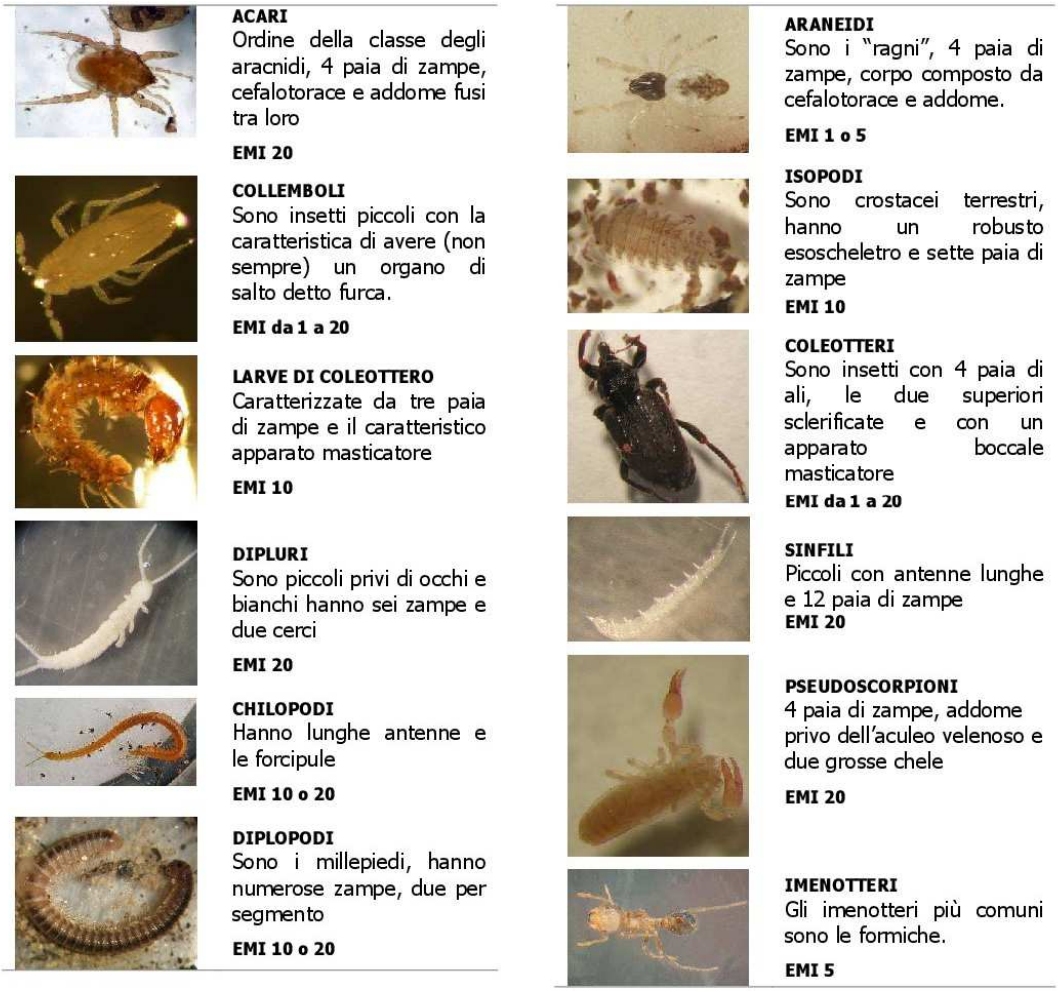

Figura 1.2: Microartropodi più comuni e loro indice EMI

Capitolo 2

I dati

2.1 Il campionamento

I dati analizzati sono stati raccolti da ARPAV, che ha sviluppato una rete di monitoraggio condotta dal 2012. Essi sono rappresentativi dell'ambiente regionale veneto per: uso del suolo, caratteristiche del terreno e condizioni climatiche. Sono stati rilevati tramite la raccolta di zolle di terreno, trivellate e profili, effettuati in diverse località del Veneto, la cui disposizione nel territorio si può vedere in Figura 2.3.
Il profilo consiste nello scavo di una trincea della profondità di 150 cm, che permette di vedere una sezione verticale di suolo, di cui è possbile vedere un esempio in Figura 2.1. In questo modo è possibile effettuare il campionamento dei diversi strati, detti orizzonti. La trivellata è l'estrazione di una carota di terreno, realizzata per mezzo di una trivella a mano fino a 120 cm di profondità (vedi Figura 2.2), che viene usata per la descrizione di alcune caratteristiche del suolo.

Figura 2.1: Esempio di un profilo di suolo.

Figura 2.2: Esempio di una trivellata.

Figura 2.3: Località e punti in cui è stato effettuato il campionamento.

2.2 Il dataset

Il campione è costituito da 243 osservazioni, per ognuna delle quali è stata registrata la località di rilevamento, la coltura impiegata su quel suolo in 2 livelli di specificità e il tipo di lavorazione, l'esposizione del suolo a quale punto cardinale (quando non risulta essere pianeggiante), l'indice QBS-ar, i valori di pH, salinità (EC1:2) in classi, carbonio organico, carbonati (CaCo3), umidità in percentuale sul peso, densità apparente, livello di quota e quantità di sabbia e argilla in percentuale. Come già detto nel Capitolo 1, ciò che si vuole indagare è il valore dell'indice QBS-ar al variare delle caratteristiche del suolo, dunque il QBS-ar è la variabile risposta, mentre tutte le altre sono le covariate.

Inizialmente è stata svolta una fase di pulizia del dataset con la correzione di qualche dato e la creazione di alcune variabili, quali la coltura aggregata e specifica. Quest'ultima è una variabile qualitativa con delle modalità in più rispetto all'altra, che specificano quello che nella coltura aggregata è indicato in modo generale come seminativo; quindi entrambe indicano la stessa cosa ma con un diverso livello di specificità.
Inoltre alcune unità statistiche erano caratterizzate dalla presenza di valori mancanti di densità apparente, umidità, pH o quanità di argilla e sabbia. Sia per il pH che per la quantità di argilla e sabbia i dati mancanti erano pari a 2, dunque si è preferito sostituirli con dei valori, in modo da non perdere le relative osservazioni. Per capire quale fosse il metodo migliore per l'imputazione di questi dati mancanti sono state visualizzate le osservazioni nello spazio. Si è notato che entrambe le unità statistiche a cui mancava il pH appartenevano alla stessa località, ovvero Asiago (Figura 2.4), dunque i valori mancanti sono stati sostituiti con la media effettuata solo sulle osservazioni situate ad Asiago. Per quanto riguarda argilla e sabbia, entrambe avevano valori mancanti per le stesse osservazioni, che erano situate a Sista Bassa. Come è possibile vedere in Figura 2.5, entrambe si trovano in prossimità di altre unità statistiche e dato che la quantità di argilla e sabbia non cambia nè tra punti molto vicini tra loro nè nel tempo, si è deciso di sostituire i valori mancanti con i valori registrati per le relative osservazioni vicine. In particolare all'osservazione QBS1O0021 sono stati imputati i dati di argilla e sabbia registrati per la QBS1O0064, mentre alla QBS1O0020 quelli di QBS1O0065.
Per quanto riguarda l'umidità e la densità apparente i valori mancanti sono rispettivamente 40 e 57, dunque sostituirli con una media non è la soluzione migliore. Si vedrà nel Capitolo 4

una possibile soluzione per l'umidità.

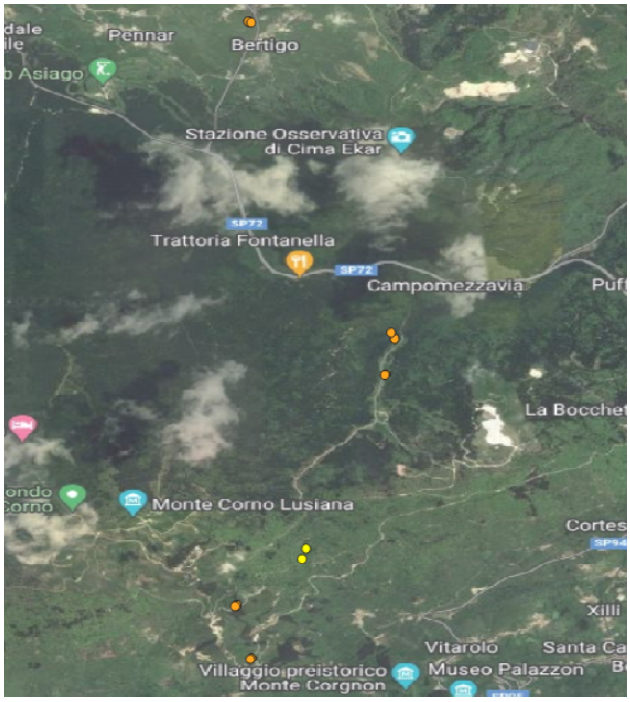

Figura 2.4: Unità statistiche situate ad Asiago e in giallo le osservazioni con dati di pH mancanti.

Figura 2.5: In blu le osservazioni complete, in giallo quelle con dati di sabbia e argilla mancanti.

2.3 Analisi esplorativa

Si effettua una prima analisi esplorativa dei dati in modo da osservare la distribuzione della variabile d'interesse QBS-ar e la relazione che quest'ultima ha con ogni altra covariata. I valori di minimo e massimo che assume l'indice QBS-ar sono rispettivamente 61 e 253, mentre la media assume valore pari a 163.2. Nella Figura 2.6 è possibile osservare la distribuzione del QBS-ar tramite boxplot, in cui si nota che non sono presenti valori anomali (outliers), la mediana è molto vicina alla media e i due baffi hanno approssimativamente la stessa lunghezza, dunque la distribuzione sembra essere simmetrica. In Figura 2.7 è possibile osservare invece la distribuzione dell'indice QBS-ar attraverso un istogramma, dove si nota che essa non si concentra sullo 0 e che non risulta essere molto simmetrica come ci si aspettava.

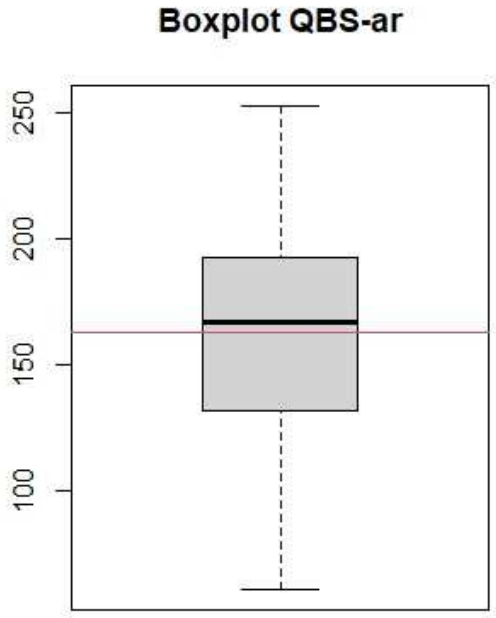

Figura 2.6: Boxplot dell'indice QBS-ar, la linea rossa indica la media.

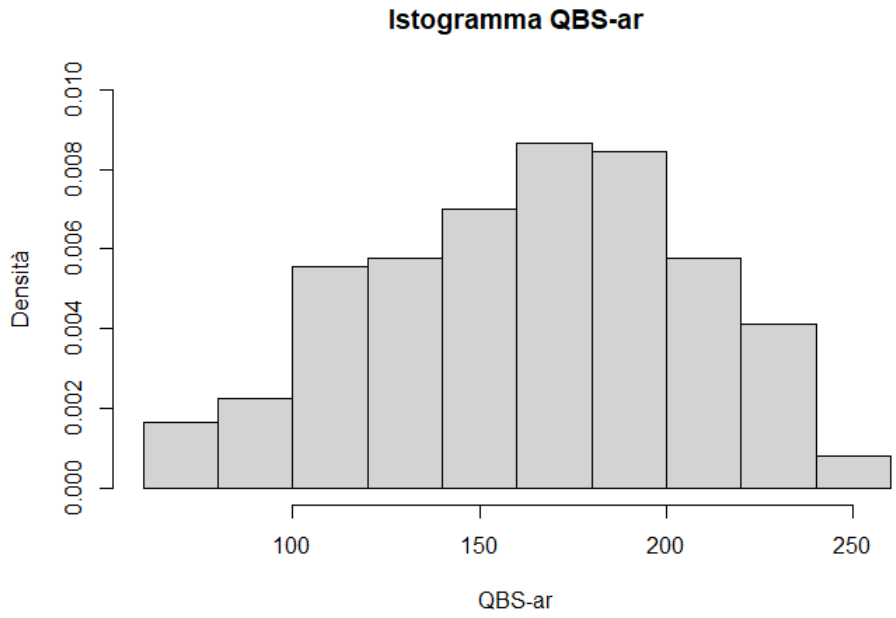

Figura 2.7: Istogramma dell'indice QBS-ar.

In seguito si valuta la distribuzione delle variabili quantitative tramite boxplot, che è possibile vedere in Figura 2.9. Le distribuzioni di pH, carbonio organico, carbonati e sabbia sembrano essere asimmetriche, contrariamente a quelle di argilla e densità apparente. Inoltre sono presenti molti outliers per quanto riguarda carbonio, sabbia e umidità, alcuni dei quali anche molto estremi. Nella distribuzione della quota sembrano essere presenti molti valori anomali, in realtà questi si riferiscono alle osservazioni di montagna, dunque non sono da considerare outliers.

Si valuta poi la relazione tra l'indice QBS-ar e le variabili concomitanti, tramite grafici di dispersione nel caso di variabili quantitative (Figura 2.8) e boxplot della distribuzione del QBS-ar condizionata alle variabili qualitative (da Figura 2.10 a 2.16).
Osservando i grafici di dispersione si nota che i carbonati non assumono valori tra circa 20 e 45 e che per valori alti il QBS-ar non assume valori elevati (maggiori di 200 circa). Con grandi quantità di sabbia (superiori al 60%) non si hanno valori alti di QBS-ar (maggiori di circa 180). Inoltre con scarse quantità di argilla (inferiori al 10%) il QBS-ar tende a non assumere valori alti (maggiori di 200), mentre per una quantità di argilla maggiore del

35% sembra che il QBS-ar non assuma valori nè estremamente bassi nè estremamente alti, ma compresi tra circa 100 e 200. Per quanto riguarda l'umidità, all'aumentare di essa il QBS-ar tende a non assumere valori bassi.
Analizzando la distribuzione del QBS-ar condizionata alle variabili qualitative si nota come l'indice assuma valori molto diversi in base alla località, per esempio a Cornuda, Gosaldo, Le Poscole e Maser tende ad assumere valori più alti rispetto a quelli assunti a Sista Bassa (Figura 2.10). Anche la coltura condiziona la distribuzione del QBS-ar, inoltre specificare il tipo di seminativo con la coltura specifica sembra essere rilevante (Figure 2.11 e 2.12). Invece considerare anche il tipo di lavorazione sembra non modificare molto la distribuzione dell'indice (Figure 2.13 e 2.14). Per quanto riguarda l'esposizione il QBS-ar assume valori più bassi a nord-ovest e più alti a nord, nord-est, sud e sud-ovest. Si nota poi la presenza della modalità "x", questa si riferisce nella maggior parte dei casi ai suoli di pianura, perchè non essendo collocati su un rilievo non si ha un'esposizione verso un determinato punto; ciò si può osservare nella Tabella 2.1. Infine in Figura 2.16 si può notare che all'aumentare della salinità il QBS-ar tende a diminuire.

Un elemento molto importante da valutare è la correlazione tra le variabili esplicative, questo per evitare il problema della multicollinearità, che si verifica in caso di alta correlazione. Per farlo si osservano i grafici di dispersione per ogni coppia di variabili, in Figura 2.17. Si nota la presenza di forte correlzione tra sabbia e argilla e tra umidità e densità apparente, di cui si terrà conto nelle successive analisi.

	Esposizione								
Quota	NO	N	NE	E	SE	S	SO	O	x
pianura	0	0	0	0	0	0	0	0	175
collina	0	2	3	3	6	3	0	2	4
montagna	3	10	6	3	14	0	6	0	3

Tabella 2.1: Distribuzione dell'esposizione condizionata alla quota raggruppata in classi.

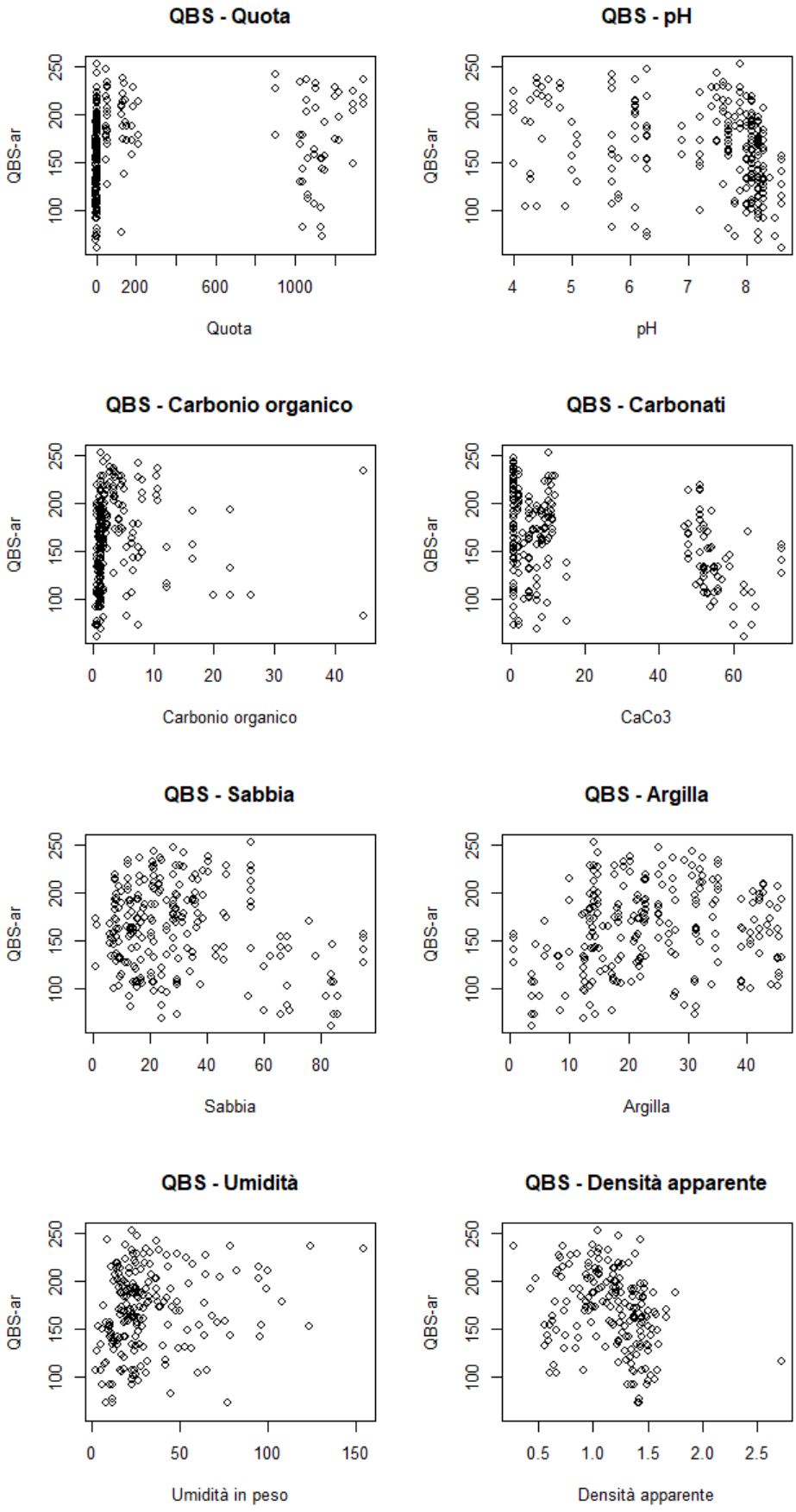

Figura 2.8: Grafici di dispersione tra QBS-ar e le variabili esplicative quantitative.

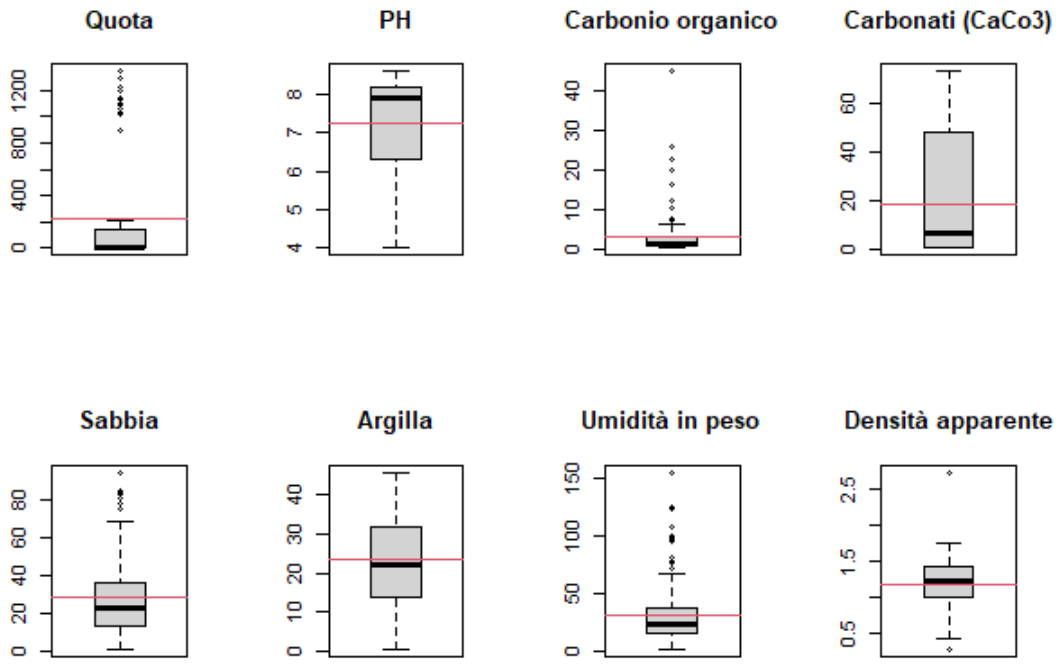

Figura 2.9: Boxplot delle variabili esplicative quantitative.

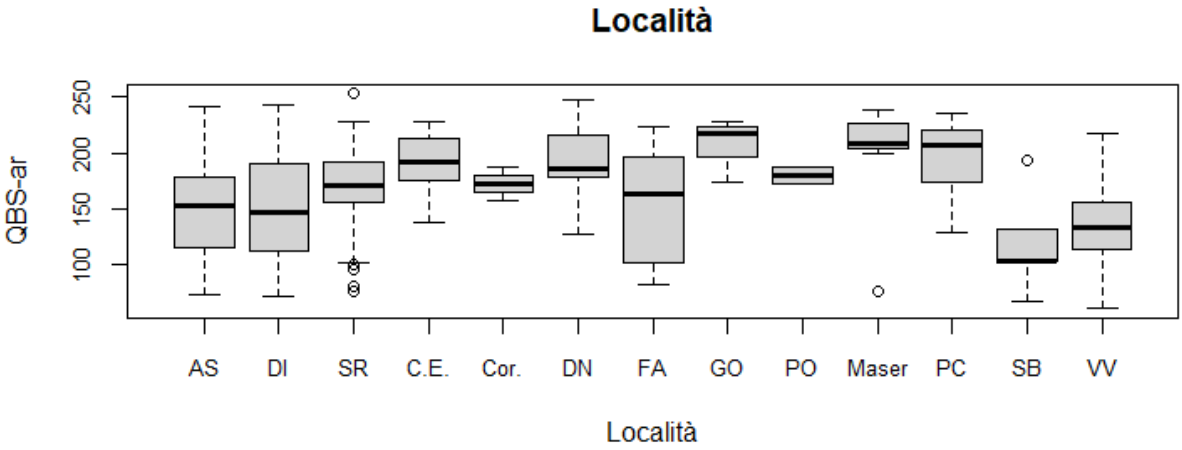

Figura 2.10: Boxplot della distribuzione del QBS-ar condizionata alla località.

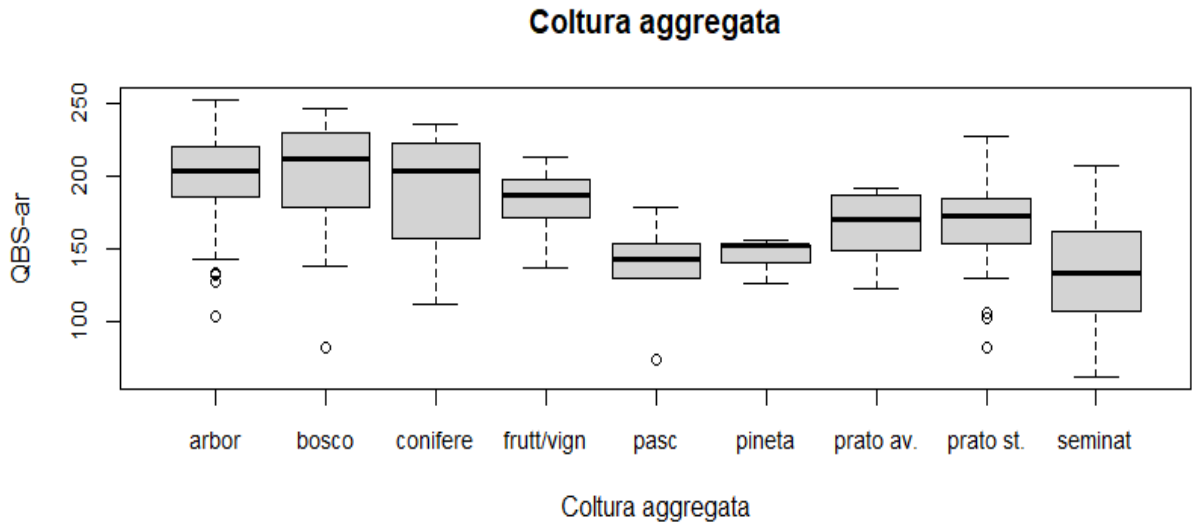

Figura 2.11: Boxplot della distribuzione del QBS-ar condizionata alla coltura aggrergata.

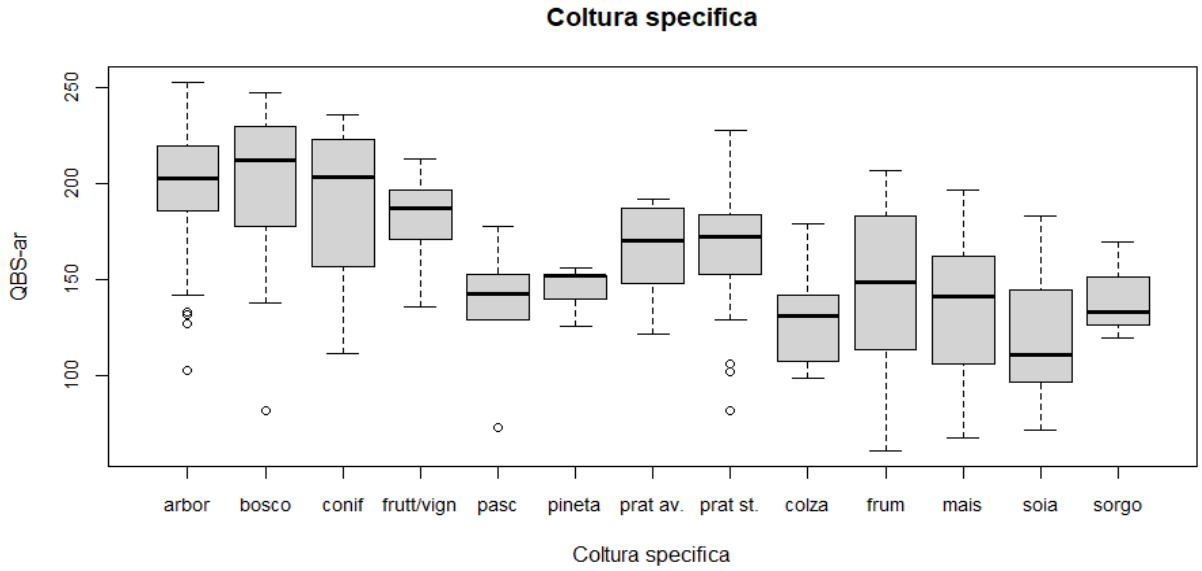

Figura 2.12: Boxplot della distribuzione del QBS-ar condizionata alla coltura specifica.

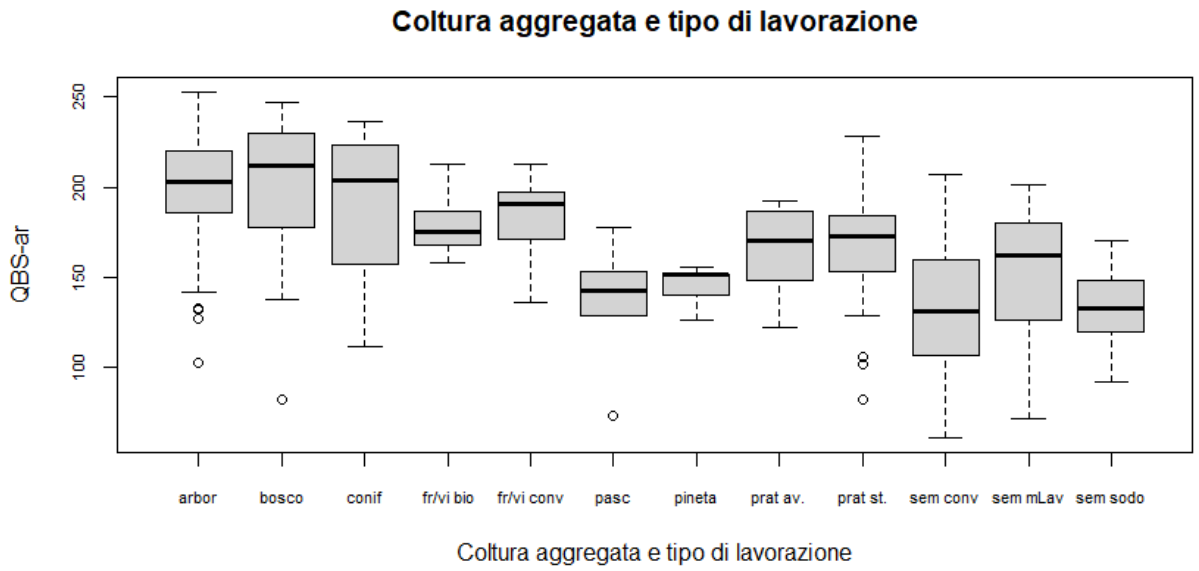

Figura 2.13: Boxplot della distribuzione del QBS-ar condizionata alla coltura aggregata e al tipo di lavorazione.

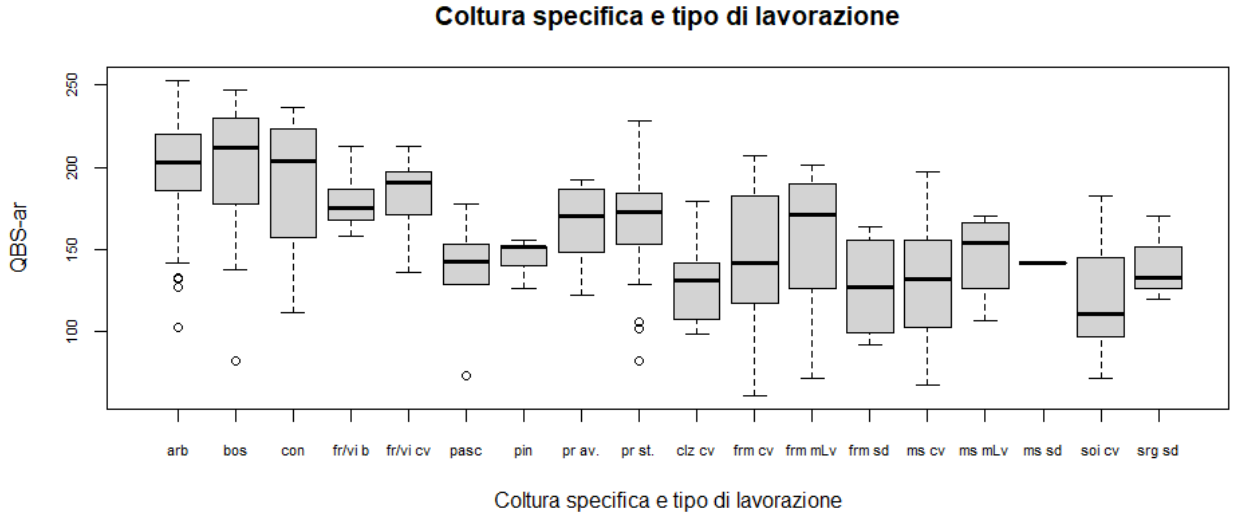

Figura 2.14: Boxplot della distribuzione del QBS-ar condizionata alla coltura specifica e al tipo di lavorazione.

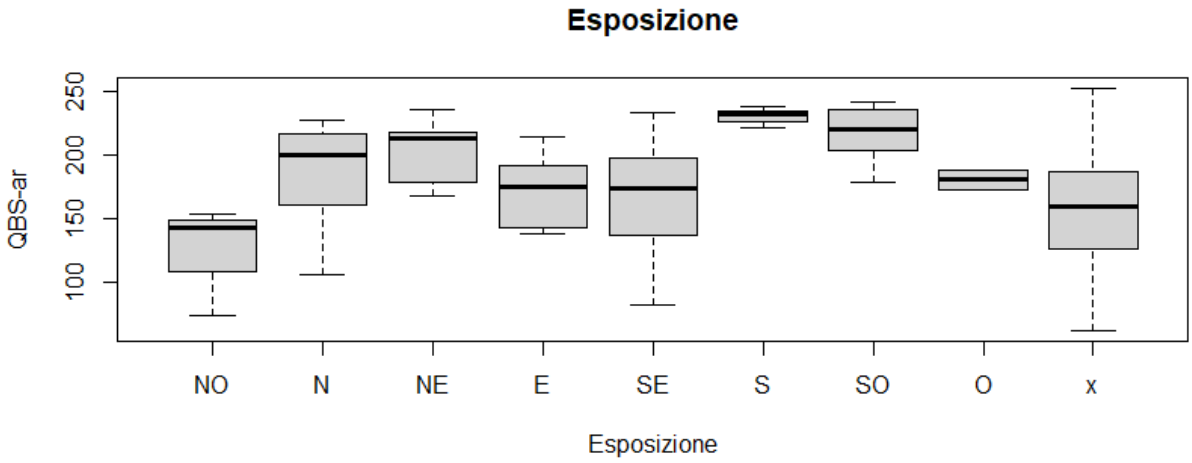

Figura 2.15: Boxplot della distribuzione del QBS-ar condizionata all'esposizione del suolo.

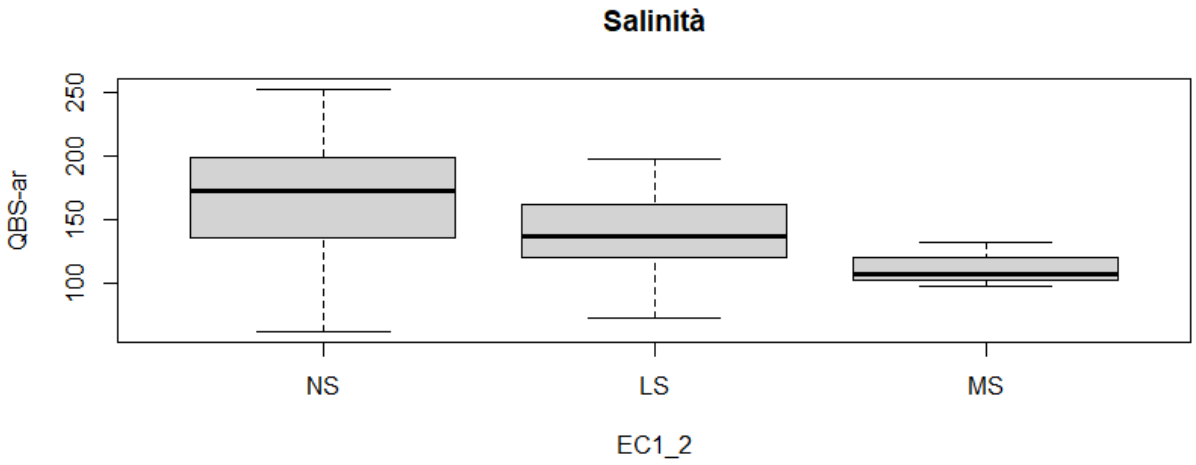

Figura 2.16: Boxplot della distribuzione del QBS-ar condizionata alla salinità del suolo.

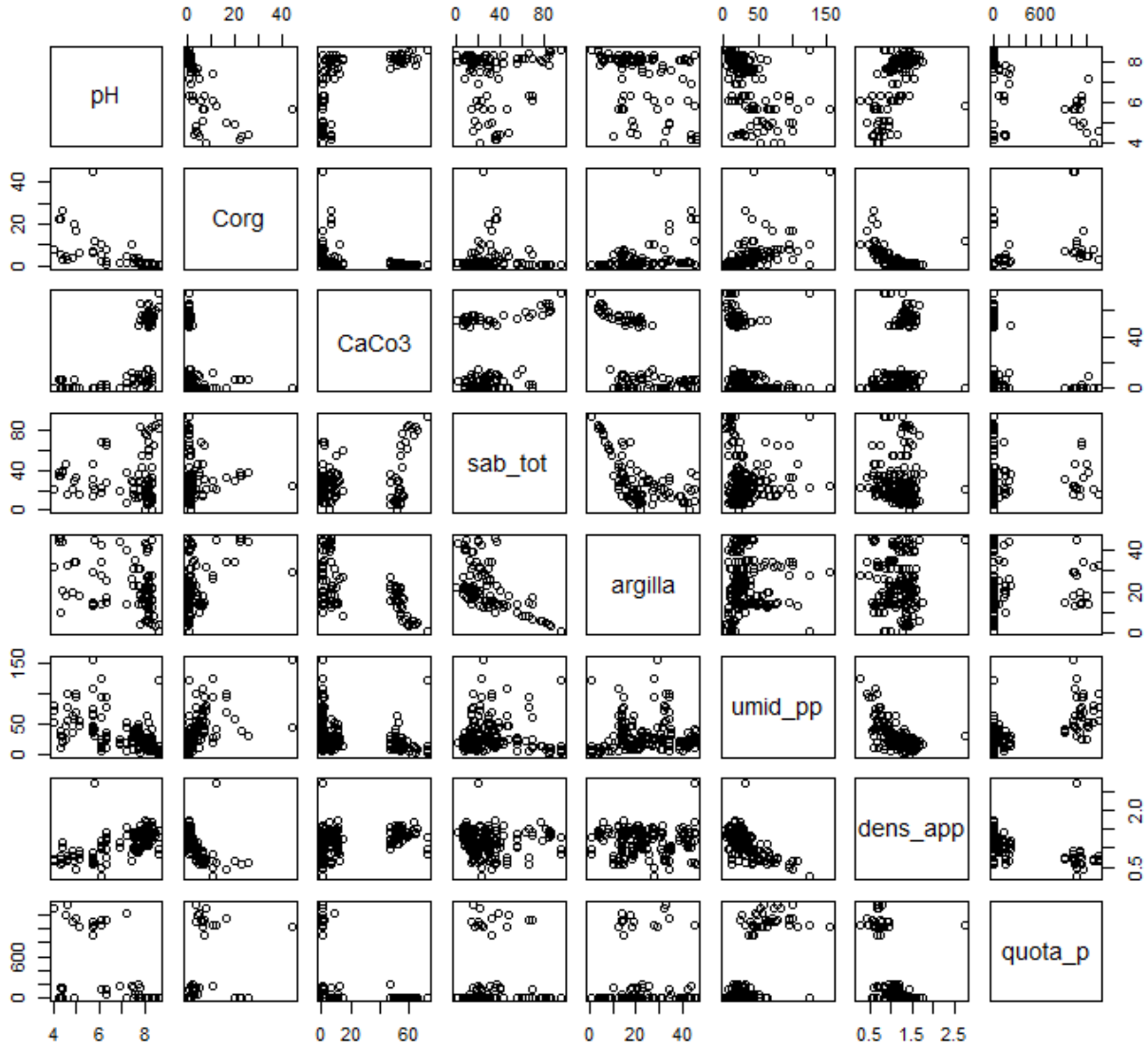

Figura 2.17: Grafici di dispersione per coppie di variabili esplicative.

Capitolo 3

Metodi e Modelli Statistici

Un problema centrale della statistica è studiare la relazione tra una variabile di interesse, detta risposta o dipendente, e altre variabili, dette in genere variabili esplicative o concomitanti. I modelli di regressione servono per capire come il comportamento della risposta può essere spiegato dalle variabili esplicative. La modellazione statistica descrive la variabilità della risposta assumendo che le osservazioni siano realizzazioni di variabili casuali. Interessa studiare se e come la legge di probabilità della risposta è influenzata dai valori delle variabili esplicative. Questi modelli sono adatti a trattare sia variabili esplicative quantitative sia qualitative, dette fattori, opportunamente codificati per mezzo di variabili indicatrici. L'eventuale presenza di fattori non rilevabili o imprevedibili viene inclusa nel modello tramite un termine, il cui comportamento viene completamente attribuito al caso. Il modello si compone quindi di una parte strutturale, che spiega la relazione tra la variabile di interesse e le variabili che ne determinano il comportamento, e di un termine casuale.

3.1 Modello di regressione lineare multipla normale

Date n osservazioni per una variabile risposta e per p variabili esplicative, il modello di regressione lineare multipla normale su una risposta quantitativa $y = (y_1, \ldots, y_n)^T$ con variabili esplicative $x_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, assume che y sia realizzazione di una variabile casuale n-dimensionale Y che soddisfa le seguenti ipotesi:

- Linearità:
$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

- Media nulla, omoschedasticità, normalità e indipendenza degli errori:
$$\varepsilon_i \sim N(0, \sigma^2) \qquad \text{indipendenti}, \qquad i = 1, \ldots, n.$$

- Indipendenza lineare tra le variabili esplicative:
$$\text{i vettori } x_j \text{ sono linearmente indipendenti}, \qquad j = 1, \ldots, p.$$

Dunque si assume che le Y_i sono indipendenti, $i = 1, \ldots, n$, e che
$$Y_i \sim N(\mu_i, \sigma^2)$$
con
$$E[Y_i] = \mu_i = \eta_i = x_i\beta = \beta_1 x_{i1} + \cdots + \beta_p x_{ip},$$
dove $\beta = (\beta_1, \ldots, \beta_p)^T$ e η_i è il predittore lineare.
Se nel modello si vuole includere l'intercetta vi è una variabile che assume valore 1 in corrispondenza di tutte le osservazioni, ossia x_{i1} per $i = 1, \ldots, n$.
Per ogni fissato $r \in \{2, \ldots, p\}$, il modello assume che $E[Y_i]$ aumenti di β_r unità se x_{ir} viene incrementato di una unità, rimanendo inalterati i livelli delle altre variabili esplicative.

Per stimare il modello bisogna effettuare la stima dei parametri β. La stima di massima verosimiglianza e quella ai minimi quadrati coincidono e lo stimatore risulta
$$\hat{\beta} = (X^T X)^{-1} X^T y$$
dove X è la matrice delle variabili esplicative di dimensione $n \times p$.
Inoltre la distribuzione di $\hat{\beta}$ è la seguente
$$\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1}).$$
Uno stimatore non distorto di $Var(\hat{\beta}_r)$ è $w_r^2 S^2$, dove $S^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p}$ e w_r^2 è l'elemento di posto (r, r) della matrice $(X^T X)^{-1}$. Dunque lo standard error, ovvero la stima della deviazione standard di $\hat{\beta}_r$, risulta essere $se(\hat{\beta}_r) = s\sqrt{w_r^2}$.

3.2 Modello a effetti casuali per risposte normali

Il modello a effetti casuali è adatto a situazioni in cui la variabile risposta è multivariata e le osservazioni sono dipendenti tra loro. In questo caso la risposta viene analizzata come realizzazione di un vettore casuale con componenti dipendenti. Siano y_{ij} le osservazioni sulla risposta per l'i-esima unità statistica, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$. Il numero totale di osservazioni è $N = \sum_{i=1}^{n} m_i$ e per ogni soggetto ne sono disponibili m_i. Il vettore delle osservazioni sulla risposta per l'i-esima unità è indicato con $y_i = (y_{i1}, \ldots, y_{im_i})^T$. Il vettore riga p-dimensionale delle variabili esplicative per l'osservazione j-esima sull'unità i-esima è indicato con x_{ij}. Si assume che y_{ij} sia realizzazione di una variabile casuale Y_{ij}, dove Y_{ij} e Y_{ih}, $j \neq h$, che descrivono osservazioni relative all'unità i-esima, sono correlate.

L'assunzione principale del modello a effetti casuali è la presenza di alcune caratteristiche non osservabili comuni a tutte le osservazioni relative ad una stessa unità, descritte come realizzazione di una variabile casuale, detta effetto casuale. Gli effetti casuali relativi a unità diverse sono assunti indipendenti. Il modello risulta essere il seguente

$$Y_{ij} = x_{ij}\beta + z_{ij}u_i + \varepsilon_{ij},$$

con β vettore p-dimensionale di effetti fissi, $u_i \sim N_q(0, \Sigma_u)$ vettore q-dimensionale di effetti casuali, mentre marginalmente $\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$, indipendente da u_i. Il modello prevede che $E[Y_{ij}] = \mu_{ij} = x_{ij}\beta$. Il termine $z_{ij}u_i$ descrive la variabilità tra unità (o cluster), mentre ε_{ij} descrive la variabilità interna alle unità. Le componenti del vettore degli effetti fissi β possono essere associate a variabili esplicative che dipendono unicamente dall'unità i-esima, dette effetti fissi fra unità (between subject) o a variabili esplicative che dipendono anche da j, dette effetti fissi entro le unità (within-subject). Si assume inoltre indipendenza tra effetti casuali e errori casuali.

L'inferenza e la stima dei parametri si basa su una verosimiglianza marginale ottenuta a partire dal modello statistico per una trasformazione AY di Y con densità non dipendente da β. La trasformazione AY fornisce i residui linearmente indipendenti della regressione lineare di Y su X. Per questo motivo, il metodo è detto della massima verosimiglianza residua o ristretta, REML (restricted maximum likelihood estimation).

3.3 Metodi di selezione del modello: *forward* e *backward*

I metodi di selezione permettono di scegliere quante e quali variabili includere in un modello, evitando di dover stimare i modelli con tutti i possibili sottoinsiemi di variabili, che con p potenziali esplicative sarebbero $2^p - 1$. I diversi metodi possono portare a modelli anche molto differenti e ciò richiede una valutazione non automatica di quali variabili è meglio considerare come esplicative. Di seguito si illustrano la procedura di selezione *forward* e *backward*. Inoltre in questo contesto si fa riferimento a una selezione definita *manuale* e una *automatica*, ciò che le differenzia è il criterio usato per la selezione delle variabili.

La selezione **forward manuale** considera inizialmente il modello con solo una variabile esplicativa, quella che fornisce l'indice R^2 più alto e che risulta maggiormente significativa, ovvero con il minor p-value per il test di nullità del coefficiente di regressione corrispondente. L'R^2 è un indice di bontà di adattamento del modello:

$$R^2 = \frac{SQ_{reg}}{SQ_{tot}} = 1 - \frac{SQ_{res}}{SQ_{tot}},$$

dove $SQ_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $SQ_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, $SQ_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ e $SQ_{tot} = SQ_{reg} + SQ_{res}$[1].
Successivamente si stimano tutti i modelli che includono una nuova variabile oltre quelle precedentemente selezionate. Si sceglie di aggiungere al modello ridotto la variabile che risulta maggiormente significativa e a cui corrisponde il modello con R^2 *corretto* maggiore. Si usa l'R^2 *corretto*

$$R^2_{adj} = 1 - \frac{SQ_{res}/(n-p)}{SQ_{tot}/(n-1)},$$

in modo da penalizzare la complessità del modello, ovvero il numero di variabili in esso incluse. Per le variabili qualitative si usa il test di confronto tra modelli annidati e si sceglie eventualmente quella il cui test fornisce p-value minore (almeno <0.05), indicando un alta significatività della variabile. Si ripete la procedura finchè l'aggiunta di un'ulteriore variabile non apporta miglioramenti al modello, ovvero quando tutte le possibili nuove variabili risultano non significative; oppure finchè tutte le variabili vengono incluse nel modello, se queste risultano tutte significative.

[1] $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ indica la media campionaria e $\hat{y}_i$ indica la stima di y_i

La selezione **backward manuale** procede in senso opposto, quindi considera come modello iniziale quello completo, che include tutte le potenziali variabili esplicative. Successivamente si elimina dal modello la variabile meno significativa, ovvero quella il cui test di nullità del coefficiente fornisce p-value maggiore. La significatività di variabili qualitative si valuta osservando il p-value fornito dal test di confronto tra il modello corrente, che la include, e quello ridotto, che la esclude. I passi si ripetono finchè non è possibile eliminare ulteriori variabili, ovvero quando tutti i test per verificare la significatività dei singoli coefficienti forniscono p-value minore di 0.05.

Per quanto riguarda le procedure di selezione **backward** e **forward automatiche** ciò che cambia è il criterio di selezione delle variabili. La scelta dei modelli avviene minimizzando il criterio d'informazione AIC (Akaike Information Criterion)

$$AIC = \frac{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}{\sigma^2} + 2p,$$

che permette di confrontare modelli anche non annidati e penalizza quelli con un alto numero di variabili.

Capitolo 4

Stima dei modelli

Per analizzare la relazione tra il QBS-ar e le altre variabili esplicative è stato stimato un modello di regressione lineare multipla normale e un modello a effetti casuali. Nel primo tutte le unità statistiche sono considerate indipendenti, mentre nel secondo si tiene conto della dipendenza tra le osservazioni relative alla stessa località.
Per la stima di entrambi i modelli sono state usate la procedura di selezione *forward* e *backward* manuale e *backward* automatica. Si noti che, per quanto riguarda i metodi di selezione manuale per i modelli a effetti casuali, non si valutano gli indici R^2 e R^2 *corretto*, ma solo la significatività dei coefficienti e i test di confronto per modelli annidati.

4.1 Stima del modello di regressione lineare multipla normale

Inizialmente è stato stimato un modello di regressione lineare multipla normale tramite procedura di selezione forward manuale. Si noti che, in questo caso, una delle assunzioni del modello non viene rispettata, poichè alcune osservazioni sono dipendenti tra loro, essendo state rilevate nella stessa località. Tuttavia per semplicità e per una valutazione iniziale si procede comunque con la stima del modello. Al primo passo i modelli con la variabile esplicativa relativa alla coltura presentano un R^2 maggiore rispetto agli altri, dunque sembra che essa spieghi maggiormente il QBS-ar. In particolare si ottiene un R^2 maggiore utilizzando la coltura specifica e il tipo di lavorazione, ma questo livello di specificità porta a un R^2 *corretto* minore rispetto a quello del modello con la sola coltura specifica, dunque

quest'ultimo risulta preferibile agli altri. Nei 2 step successivi vengono aggiunte località e sabbia. Al passo seguente il test di nullità del singolo coeffciente che fornisce p-value minore è quello relativo all'umidità, essa però è caratterizzata da 40 dati mancanti, che si riferiscono solamente ai dati campionati nel 2013 e 2014, dunque includerla nel modello significherebbe non considerare i campionamenti effettuati in questi 2 anni.
Data questa situazione si decide di stimare i modelli su 3 diversi insiemi di dati: uno in cui sono escluse le unità statistiche relative al 2013 e 2014 ma sono presenti tutte le variabili, uno con tutte le osservazioni ma in cui viene esclusa l'umidità, e uno in cui i valori mancanti di umidità sono sostituiti dai valori stimati tramite un modello di regressione lineare multipla (l'imputazione dei dati mancanti è trattata nella Sezione 4.3). In quest'ultimo caso è stata usata solamente la procedura di selezione backward, sia manuale che automatica.

Per la stima dei modelli tramite procedura di selezione backward, il modello completo iniziale è stato stimato comprendendo non tutte le variabili relative alla coltura, ma solamente la coltura specifica, dato che essa risulta quella maggiormente significativa.

I modelli finali stimati su tutte le osservazioni, esclusa l'umidità, tramite procedura forward e backward manuale e backward automatica risultano uguali e includono le seguenti variabili: coltura specifica, località, sabbia e esposizione (Modello $A1$). Per quanto riguarda i modelli stimati sui dati raccolti dal 2015, le procedure di selezione backward e forward manuale portano allo stesso modello finale, che include: coltura specifica, località, sabbia, umidità e carbonio organico; mentre quello stimato tramite selezione backward automatica, differisce dal precedente solo poichè include una variabile in più: il pH. Essendo 2 modelli annidati è possibile valutare la significatività della variabile in più, per poter determinare quale sia il modello preferibile tra i 2, che risulta essere quello ottenuto dalla selezione manuale (Modello $A2$). Il modello stimato tramite selezione backward manuale, che coincide con quello ottenuto dalla modalità automatica, sui dati in cui i valori di umidità mancanti sono stati sostituiti dalle stime, comprende le seguenti variabili: coltura specifica, località, sabbia, umidità e esposizione (Modello $A3$).
I modelli $A1$, $A2$ e $A3$ sono presentati in Tabella 4.1

Per verificare se le assunzioni sono rispettate (eccetto l'indipendenza di tutte le osservazioni) e valutare la bontà di adattamento dei modelli si esegue l'**analisi dei residui** studentizzati. Essa viene svolta graficamente anche per verificare che i residui non siano caratterizzati da un qualche comportamento che non viene colto dal modello.

In Figura 4.2 è possibile osservare il grafico di dispersione tra i valori stimati e i residui del Modello $A1$, dove la linea rossa indica la media di quest'ultimi. Sembra che i residui non abbiano un andamento sistematico e che ci sia omoschedasticità. Inoltre si nota che essi hanno media nulla. Dall'istogramma dei residui, con sopra la curva della distribuzione normale, è possibile vedere come la loro distribuzione segua abbastanza bene quella della normale. Il Q-Q Plot (in Figura 4.1(a)) mostra un adattamento abbastanza buono della distribuzione dei residui a quella normale, a parte un lieve discostamento sulle code, in particolar modo su quella sinistra. Dunque sembra che il Modello $A1$ rispetti tutto sommato le assunzioni.
In Figura 4.3 è possibile osservare l'analisi dei residui del Modello $A2$. Il grafico di dispersione, dove la linea rossa indica la media, mostra che i residui non hanno un andamento sistematico e che vi è omoschedasticità. Inoltre si nota che essi hanno media nulla. Dall'istogramma dei residui, a cui è sovrapposta la curva della distribuzione normale, si nota che la distribuzione dei residui non segue quella normale abbastanza bene. Il Q-Q Plot, in Figura 4.1(b) conferma quanto detto, infatti la distribuzione si discosta dalla normale sulle code.
Per quanto riguarda l'analisi dei residui del Modello $A3$, in Figura 4.4, dal grafico di dispersione è possibile notare le stesse caratteristiche dei Modelli $A1$ e $A2$. Mentre l'istogramma dei residui mostra un buon adattamento alla distribuzione normale. Dal Q-Q Plot in Figura 4.1(c) è possibile osservare che la distribuzione dei residui si adatta abbastanza bene a quella normale.

Confrontando i 3 modelli $A1$, $A2$ e $A3$, si nota che i valori di R^2 e R^2 *corretto* sono molto simili. il modello $A2$ ha un R^2 *corretto* leggermente maggiore rispetto agli altri, dunque sembrerebbe che esso riesca a spiegare il fenomeno un po' meglio, tuttavia la differenza è molto piccola, quindi i modelli hanno capacità di spiegazione del fenomeno molto simile. Dal confronto dei Q-Q Plot dei 3 modelli (in Figura 4.1) emerge che i residui di $A1$ e $A3$ si adattano molto meglio alla distribuzione normale rispetto a quelli del modello $A2$. In conclusione l'assunzione di normalità dei residui è rispettata maggiormente nei modelli $A1$ e $A3$, che risultano dunque preferibili. Inoltre il modello $A1$ risulta preferibile al modello $A3$ perchè, a parità di capacità di spiegazione del fenomeno, quest'ultimo è stato stimato su dati a loro volta stimati. Dunque in conclusione il modello $A1$ risulta preferibile agli altri.

	Modello A1	Modello A2	Modello A3
Intercetta	188.09 (29.11)***	178.94 (15.82)***	164.33 (30.66)***
localita Az. Diana	88.22 (30.55)**	28.77 (15.11)$^{\cdot}$	101.39 (30.80)**
localita Ceregnano - Sasse Rami	105.33 (30.05)***	46.24 (14.24)**	117.29 (30.22)***
localita Colli Euganei	48.78 (15.19)**	41.27 (14.61)**	63.72 (16.41)***
localita Cornuda	44.03 (23.09)$^{\cdot}$	30.45 (21.31)	55.69 (23.43)*
localita Dueville, Novoledo	91.61 (29.30)**	29.78 (12.47)*	99.53 (29.22)***
localita Falcade	44.86 (22.06)*	12.46 (16.45)	49.50 (21.94)*
localita Gosaldo	42.39 (26.65)	68.44 (21.90)**	52.98 (26.80)*
localita Le poscole	21.58 (34.44)	36.39 (22.27)	34.96 (34.60)
localita Maser	112.23 (33.73)**	54.30 (16.29)**	126.70 (34.00)***
localita Pian Cansiglio	35.10 (12.19)**	25.32 (10.17)*	32.99 (12.11)**
localita Sista Bassa	46.66 (32.14)	−19.03 (20.82)	52.56 (31.93)
localita Vallevecchia	80.93 (30.14)**	12.17 (14.63)	93.51 (30.35)**
coltura specifica bosco latifoglie	−2.75 (11.31)	3.59 (10.35)	−1.53 (11.21)
coltura specifica conifere	−3.69 (16.53)	−13.12 (13.20)	−3.73 (16.37)
coltura specifica frutteto/vigneto	−30.74 (9.85)**	−34.31 (9.24)***	−31.29 (9.76)**
coltura specifica pascolo	−31.21 (24.14)	−54.39 (17.02)**	−35.20 (23.97)
coltura specifica pineta litoranea	−8.03 (15.65)	−13.83 (15.32)	−11.71 (15.58)
coltura specifica prato avvicendato	−34.82 (12.05)**	−37.93 (11.95)**	−35.45 (11.94)**
coltura specifica prato stabile	−19.69 (11.06)$^{\cdot}$	−35.87 (10.44)***	−19.34 (10.95)$^{\cdot}$
coltura specifica colza	−67.08 (11.11)***	−47.56 (15.11)**	−66.58 (11.00)***
coltura specifica frumento	−54.44 (7.74)***	−53.99 (7.66)***	−53.19 (7.69)***
coltura specifica mais	−63.85 (8.23)***	−55.30 (8.46)***	−63.31 (8.15)***
coltura specifica soia	−81.68 (8.53)***	−78.08 (9.16)***	−81.70 (8.44)***
coltura specifica sorgo	−34.48 (16.44)*	−33.25 (15.50)*	−34.15 (16.28)*
esposizione Nord	−11.86 (29.65)		−9.43 (29.38)
esposizione Nord-Est	3.93 (32.76)		4.54 (32.44)
esposizione Est	−19.41 (29.56)		−21.10 (29.28)
esposizione Sud-Est	−18.38 (26.85)		−12.60 (26.71)
esposizione Sud	−46.62 (47.71)		−46.40 (47.25)
esposizione Sud-Ovest	28.46 (30.54)		29.55 (30.24)
esposizione x	−65.87 (41.43)		−61.97 (41.06)
sabbia	−0.51 (0.11)***	−0.44 (0.12)***	−0.49 (0.11)***
carbonio organico		−1.43 (0.57)*	
umidità		0.43 (0.13)**	0.31 (0.14)*
R^2	0.61	0.63	0.62
Adj. R^2	0.55	0.57	0.55
AIC	2363.29	1942.13	2359.3
Num. obs.	243	203	243

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Tra parentesi lo standard error

Tabella 4.1: Modelli di regressione lineare multipla normale.

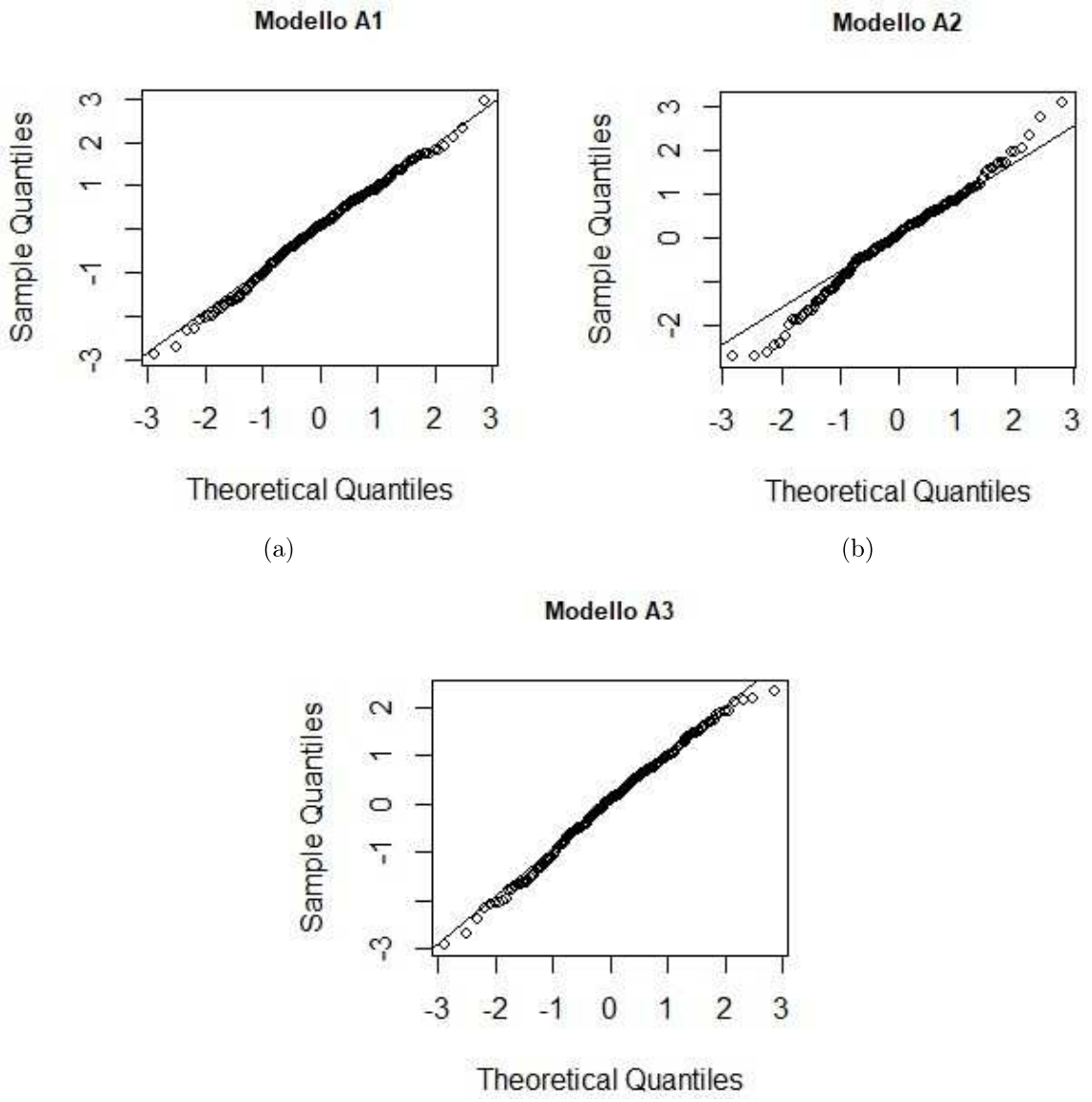

Figura 4.1: Q-Q Plot dei residui dei modelli di regressione lineare multipla normale.

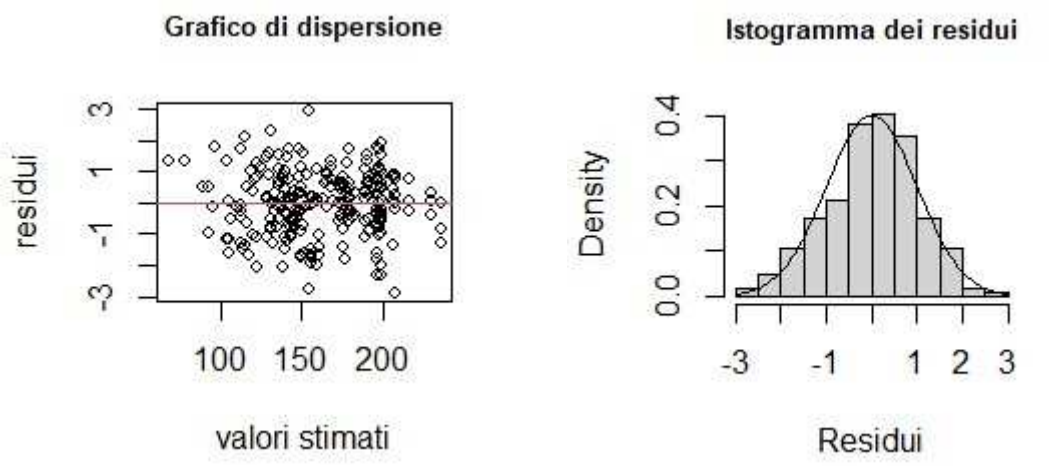

Figura 4.2: Analisi dei residui del Modello $A1$

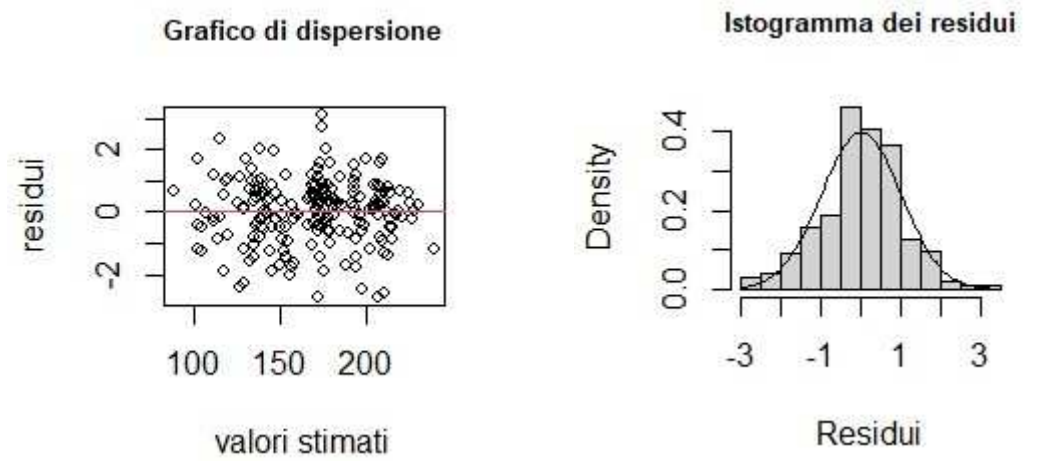

Figura 4.3: Analisi dei residui del Modello $A2$

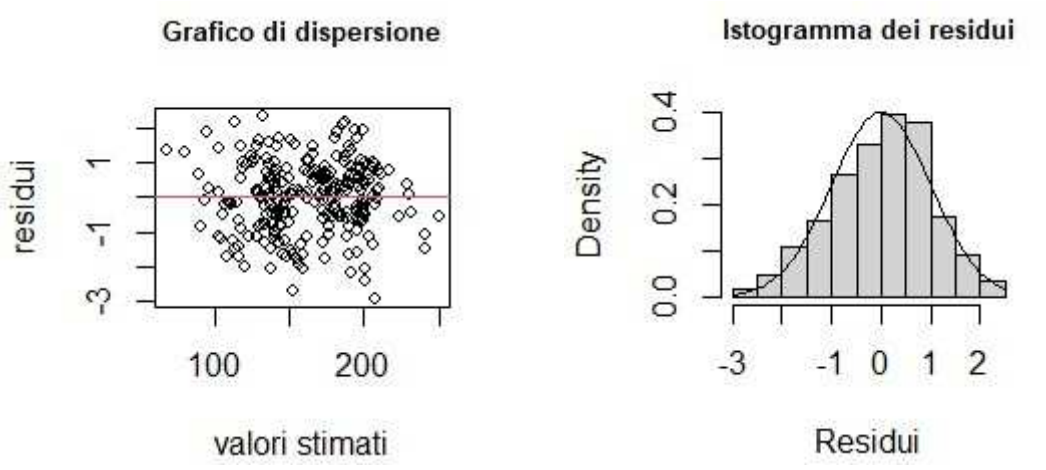

Figura 4.4: Analisi dei residui del Modello $A3$

4.2 Stima del modello a effetti casuali

Per tenere conto della dipendenza tra le osservazioni relative alla stessa località e includerla nel modello, è stato stimato un modello a effetti casuali. Anche in questo caso sono stati stimati i modelli sui 3 insiemi di dati precedentemente descritti. Per la stima viene usato il metodo "REML", ovvero della massima verosimiglianza ristretta, tuttavia per poter effettuare il test di confronto tra modelli annidati è necessario utilizzare il metodo "ML", dove viene massimizzata la log-verosimiglianza.

Il modello stimato su tutte le osservazioni, escludendo a priori l'umidità, tramite procedura di selezione forward manuale include solo 3 variabili: coltura specifica, sabbia e carbonio organico (Modello $B1.1$). I modelli stimati tramite selezione backward manuale e automatica risultano essere annidati. Il primo include coltura specifica, esposizione, sabbia e quota; il secondo include carbonio organico e carbonati, oltre le variabili precedentemente elencate. Essendo 2 modelli annidati è possibile effettuare il test di confronto, che indica il modello ridotto come preferibile, dunque quello ottenuto dalla selezione backward manuale (Modello $B1.2$).

Il modello stimato sui dati raccolti dal 2015 tramite selezione forward manuale risulta essere lo stesso di quello ottenuto dalla selezione backward manuale, e include le seguenti

	Modello B1.1	Modello B1.2	Modello B2	Modello B3
Intercetta	214.92 (7.91)***	234.95 (31.68)***	208.32 (9.11)***	218.11 (30.57)***
coltura specifica bosco latifoglie	10.20 (9.26)	0.35 (11.02)	9.31 (9.41)	5.13 (10.88)
coltura specifica conifere	2.60 (10.41)	8.10 (15.19)	−8.70 (11.02)	10.04 (14.77)
coltura specifica frutteto/vigneto	−25.41 (8.87)**	−27.79 (9.28)**	−29.65 (8.72)***	−28.54 (9.15)**
coltura specifica pascolo	−36.59 (15.51)*	−21.36 (22.72)	−54.88 (15.75)***	−28.65 (22.63)
coltura specifica pineta litoranea	−9.80 (15.76)	−6.71 (15.54)	−14.72 (15.23)	−13.38 (15.59)
coltura specifica prato avvicendato	−34.71 (12.24)**	−33.25 (12.00)**	−36.89 (11.94)**	−35.15 (11.96)**
coltura specifica prato stabile	−26.69 (9.14)**	−18.95 (10.57)$^{\cdot}$	−34.84 (9.33)***	−20.25 (10.44)$^{\cdot}$
coltura specifica colza	−67.85 (11.22)***	−66.19 (11.03)***	−48.47 (15.11)**	−66.70 (10.97)***
coltura specifica frumento	−54.69 (7.73)***	−53.21 (7.63)***	−53.28 (7.60)***	−52.82 (7.60)***
coltura specifica mais	−66.02 (8.27)***	−63.91 (8.15)***	−55.70 (8.44)***	−65.07 (8.11)***
coltura specifica soia	−81.56 (8.51)***	−80.46 (8.42)***	−76.22 (9.12)***	−80.95 (8.35)***
coltura specifica sorgo	−36.44 (16.71)*	−33.68 (16.39)*	−34.33 (15.49)*	−35.33 (16.30)*
sabbia	−0.52 (0.11)***	−0.53 (0.11)***	−0.43 (0.12)***	−0.48 (0.11)***
carbonio organico	−1.27 (0.47)**		−2.10 (0.50)***	−1.25 (0.53)*
esposizione Nord		0.25 (28.69)		9.01 (28.26)
esposizione Nord-Est		20.27 (30.36)		27.15 (29.58)
esposizione Est		−17.05 (29.40)		−12.02 (29.14)
esposizione Sud-Est		−14.74 (26.23)		0.96 (26.08)
esposizione Sud		5.14 (37.80)		15.88 (35.97)
esposizione Sud-Ovest		28.82 (29.92)		31.36 (29.47)
esposizione Ovest		−20.73 (41.26)		−9.87 (38.68)
esposizione x		−26.31 (31.07)		−13.54 (29.65)
quota		−0.03 (0.02)$^{\cdot}$		−0.04 (0.02)*
umidità			0.39 (0.12)**	0.36 (0.14)*
AIC	2297.22	2247.73	1878.15	2245.73
Log Likelihood	−1131.61	−1098.86	−921.07	−1095.86
Num. obs.	243	243	203	243

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Tra parentesi lo standard error

Tabella 4.2: Modelli a effetti casuali

variabili: coltura specifica, sabbia, carbonio organico e umidità. Il modello stimato tramite selezione backward automatica include, oltre le variabili precedentemente elencate, pH e salinità. Dunque si tratta di 2 modelli annidati, per cui si effettua il test di confronto, che porta a preferire il modello ridotto, ovvero quello ottenuto sia dalla selezione forward che backward manuale (Modello $B2$).

Per quanto riguarda le osservazioni con imputazione dei dati mancanti, anche in questo caso i modelli stimati tramite procedura bacward manuale e automatica risultano essere annidati. In particolare il secondo comprende i carbonati, oltre coltura specifica, esposizione, carbonio organico, sabbia, quota e umidità, che sono incluse nel primo. Dato che l'AIC di entrambi è molto simile e i carbonati risultano non significativi, è preferibile il modello stimato tramite procedura di selezione backward manuale (Modello $B3$).

Confrontando i 4 modelli si nota che il $B1.1$ e il $B2$ includono le stesse variabili, eccetto l'umidità che viene inclusa solo dal $B2$, anche perchè il $B1.1$ la esclude a prescindere. Dunque sembra che l'umidità sia significativa. Il modello $B1.2$ presenta un AIC minore rispetto al $B1.1$, dunque sembra essere preferibile a quest'ultimo. Si noti che l'AIC del modello $B2$ non è confrontabile con gli altri, poichè il modello è stato stimato su un numero minore di osservazioni, di conseguenza anche l'AIC risulta minore. Il modello $B3$ ha AIC minore, ma essendo stimato su valori a loro volta stimati, è meno preferibile rispetto agli altri. Dunque i modelli preferibili sono il $B1.2$ e $B2$, sembra infatti che essi abbiano capacità di spiegazione del fenomeno molto simile, dunque possono essere usati entrambi.

4.3 Imputazione dei dati mancanti

Per poter effettuare la stima dei modelli su tutte le osservazioni e considerare tutte le variabili come potenziali esplicative è stata effettuata l'imputazione dei dati mancanti. Per far ciò è stato utilizzato un modello di regressione lineare multipla, dove l'umidità è la variabile risposta e le covariate sono tutte le altre variabili, eccetto il QBS-ar e la densità apparente. Quest'ultima è stata esclusa in quanto presenta valori mancanti per le stesse osservazioni in cui mancano i dati di umidità. Per la stima del modello è stata usata la procedura di selezione backward, sia manuale che automatica. I 2 modelli finali risultano essere annidati. In particolare il modello ottenuto dalla selezione automatica (Modello $C2$) include carbonati e argilla oltre località, pH, carbonio organico e quota, incluse nel Modello

$C1$ (ottenuto dalla selezione manuale). Entrambi i modelli stimati sono presentati nella Tabella 4.3. Dato che i modelli sono annidati è possibile svolgere il test di confronto. Esso fornisce p-value maggiore di 0.1, che porta a non rifiutare l'ipotesi nulla, per cui il modello ridotto risulta preferibile. Dunque il Modello $C1$ è stato utilizzato per stimare i valori mancanti di umidità.

	Modello C1	Modello C2
Intercetta	−46.74 (30.32)	−78.94 (35.37)*
localita Az. Diana	26.40 (25.28)	46.42 (27.46)$^{\cdot}$
localita Ceregnano - Sasse Rami	29.26 (25.19)	49.78 (27.46)$^{\cdot}$
localita Colli Euganei	19.78 (21.86)	42.78 (25.15)$^{\cdot}$
localita Cornuda	20.22 (22.89)	40.69 (25.43)
localita Dueville, Novoledo	39.67 (24.11)	59.12 (26.23)*
localita Falcade	−22.83 (9.54)*	−28.08 (9.92)**
localita Gosaldo	−23.34 (12.30)$^{\cdot}$	−24.26 (12.30)$^{\cdot}$
localita Le poscole	27.17 (24.11)	42.99 (25.50)$^{\cdot}$
localita Maser	28.22 (23.40)	50.69 (26.29)$^{\cdot}$
localita Pian Cansiglio	9.71 (5.99)	9.87 (5.97)
localita Sista Bassa	41.09 (27.13)	71.87 (31.79)*
localita Vallevecchia	26.48 (25.22)	61.36 (31.92)$^{\cdot}$
pH	4.78 (2.13)*	7.35 (2.64)**
carbonio organico	1.22 (0.30)***	1.29 (0.30)***
quota	0.07 (0.02)**	0.09 (0.03)***
carbonati		−0.37 (0.23)
argilla		−0.22 (0.14)
R^2	0.64	0.65
Adj. R^2	0.61	0.61
Num. obs.	203	203

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Tabella 4.3: Modelli di regressione lineare multipla normale per la stima dell'umidità.

Conclusioni

Questo lavoro di tesi è incentrato sullo studio della relazione tra la qualità biologica del suolo e i potenziali fattori che la influenzano. Si sono proposti dei modelli di regressione lineare multipla normale e a effetti casuali per spiegare il comportamento dell'indice QBS-ar in relazione alle covariate che descrivono le caratterisitche del suolo. Per la stima dei modelli sono state usate le procedure di selezione backward e forward, sia in modo manuale che automatico. In particolare, la modellazione proposta attraverso un modello a effetti casuali considera la dipendenza tra le osservazioni relative alla stessa località, includendo nel modello un termine, detto effetto casuale, che descrive la presenza di caratteristiche non osservabili comuni a queste osservazioni. Per questo motivo il modello a effetti casuali risulta preferibile rispetto alla regressione lineare multipla normale. Infatti in quest'ultimo le stime puntuali dei coefficienti sono corrette, ma non quelle degli standard error, quindi la loro variabilità.

Durante l'analisi è stato affrontato anche il problema di dati mancanti, che si è cercato di risolvere attraverso la stima di un modello di regressione lineare multipla. Tuttavia sono stati stimati diversi modelli su diversi insiemi di dati.

I modelli risultati preferbili nelle precedenti analisi sono $B1.2$ e $B2$ (in Tabella 4.2). Inoltre, a parità di performance, il modello $B2$ sembra preferibile al $B1.2$ poichè ha quasi tutti i coefficienti significativi e include l'umidità, che sembra essere molto significativa, ma che viene esclusa a prescindere dal $B1.2$.

Ad ogni modo si nota che tutti i modelli includono la sabbia e la coltura specifica, dunque si può affermare che il tipo di coltivazione e la quantità di sabbia presente nel suolo, oltre naturalmente la località, sono i fattori che maggiormente influenzano il comportamento del QBS-ar, quindi la qualità biologica del suolo.

Le stime dei parametri del Modello $B2$ indicano che mediamente il valore dell'indice

QBS-ar, a parità dei valori delle altre variabili, tende ad aumentare nei suoli su cui vi sono boschi di latifoglie, mentre tende a diminuire in presenza di altre coltivazioni. Per quanto riguarda la sabbia, all'aumentare di essa il QBS-ar tende in media a diminuire, a parità dei valori delle altre variabili. Inoltre all'aumentare del carbonio organico, l'indice tende mediamente a diminuire, mentre tende ad aumentare all'aumentare dell'umidità, rimanendo inalterati i valori delle altre variabili.

Appendice A

Codice R

Librerie

```
library(MASS)
library(tidyverse)
library(nlme)
```

Pulizia del dataset

```
# eliminazione quota in coltura
coltura_lista <- strsplit(prova$coltura, split = ">|<")
coltura <- rep(0, length(prova$coltura))
for (i in 1:length(prova$coltura)) {
  coltura_tipoLav[i] <- coltura_tipoLav_lista[[i]][1]
}

# creazione variabile coltura specifica
dati.all$coltura_specifica <- dati.all$coltura_aggr
dati.all[which(dati.all$coltura_specifica=="seminativo"),]$coltura_specifica <- dati.all[which(
    dati.all$coltura_specifica=="seminativo"),]$coltura_noQuota

# definizione fattori
data$localita <- factor(data$localita)
data$coltura_aggr <- factor(data$coltura_aggr)
data$coltura_specifica <- factor(data$coltura_specifica, levels = c("arboricoltura", "bosco
    latifoglie", "conifere", "frutteto/vigneto", "pascolo", "pineta litoranea", "prato
    avvicendato", "prato stabile", "colza", "frumento", "mais", "soia", "sorgo"))
```

```
data$coltura_specifica_tipoLav <- factor(data$coltura_specifica_tipoLav, levels = c("
    arboricoltura X", "bosco latifoglie X", "conifere X", "frutteto/vigneto biologico", "
    frutteto/vigneto convenzionale", "pascolo X", "pineta litoranea X", "prato avvicendato X",
    "prato stabile X", "colza convenzionale", "frumento convenzionale", "frumento minime
    lavorazioni", "frumento sodo", "mais convenzionale","mais minime lavorazioni", "mais sodo",
     "soia convenzionale", "sorgo sodo"))
data$coltura_aggr_tipoLav <- factor(data$coltura_aggr_tipoLav, levels = c("arboricoltura X", "
    bosco latifoglie X", "conifere X", "frutteto/vigneto biologico", "frutteto/vigneto
    convenzionale", "pascolo X", "pineta litoranea X", "prato avvicendato X", "prato stabile X"
    , "seminativo convenzionale", "seminativo minime lavorazioni", "seminativo sodo"))
data$EC1_2_CLASSI <- factor(data$EC1_2_CLASSI, levels = c("NS", "LS", "MS"))
data$esposizione <- factor(data$esposizione, levels = c("NO", "N", "NE", "E", "SE", "S", "SO", "
    O"))

# Sostituzione dati mancanti di sabbia, argilla e pH.
## sabbia ####
# sostituzione con osservazione vicina perche' la sabbia non cambia negli anni
data[data$ID_OSS=="QBS1O0020",]$sab_tot <- data[data$ID_OSS=="QBS1O0065",]$sab_tot
data[data$ID_OSS=="QBS1O0021",]$sab_tot <- data[data$ID_OSS=="QBS1O0064",]$sab_tot

## argilla ####
# sostituzione con osservazione vicina perche' l'argilla non cambia negli anni
data[data$ID_OSS=="QBS1O0020",]$argilla <- data[data$ID_OSS=="QBS1O0065",]$argilla
data[data$ID_OSS=="QBS1O0021",]$argilla <- data[data$ID_OSS=="QBS1O0064",]$argilla

## ph ####
# sostituzione con media di asiago
summary(data$pH)
summary(data$pH[data$localita=="Asiago"])
data[data$ID_OSS=="QBS1O0174",]$pH <- mean(data$pH[data$localita=="Asiago"], na.rm = T)
data[data$ID_OSS=="QBS1O0222",]$pH <- mean(data$pH[data$localita=="Asiago"], na.rm = T)
```

Analisi esplorativa

```
# Statistiche descrittive QBS
summary(data$QBSar)

# Boxplot QBS
boxplot(data$QBSar, main = "Distribuzione QBS")
abline(h=mean(data$QBSar), col=2)
```

```r
# Istorgamma QBS
hist(data$QBSar, main = "Distribuzione QBS", probability = T, ylim = c(0,0.01))

# Statistiche descrittive delle variabili qualitative
table(data$localita, useNA = "ifany")
table(data$coltura_aggr, useNA = "ifany")
table(data$coltura_aggr_tipoLav, useNA = "ifany")
table(data$coltura_specifica, useNA = "ifany")
table(data$coltura_specifica_tipoLav, useNA = "ifany")
table(data$esposizione, useNA = "ifany")
table(data$EC1_2_CLASSI, useNA = "ifany")

# Statistiche desrittive e distribuzione delle variabili quantitative
summary(data$quota)
summary(data$pH)
summary(data$Corg)
summary(data$CaCo3)
summary(data$sab_tot)
summary(data$argilla)
summary(data$umid_pp)
summary(data$dens_app)

# Boxplot
par(mfrow=c(2,4))
# quota
boxplot(data$quota, main = "Quota")
abline(h=mean(data$quota, na.rm = T), col=2)
# pH
boxplot(data$pH, main = "PH")
abline(h=mean(data$pH, na.rm = T), col=2)
# Corg
boxplot(data$Corg, main = "Carbonio organico")
abline(h=mean(data$Corg, na.rm = T), col=2)
# CaCo3
boxplot(data$CaCo3, main = "Carbonati (CaCo3)")
abline(h=mean(data$CaCo3, na.rm = T), col=2)
```

```r
# sab_tot
boxplot(data$sab_tot, main = "Sabbia")
abline(h=mean(data$sab_tot, na.rm = T), col=2)
# argilla
boxplot(data$argilla, main = "Argilla")
abline(h=mean(data$argilla, na.rm = T), col=2)
# umid_pp
boxplot(data$umid_pp, main = "Umidita' in peso")
abline(h=mean(data$umid_pp, na.rm = T), col=2)
# dens_app
boxplot(data$dens_app, main = "Densita' apparente")
abline(h=mean(data$dens_app, na.rm = T), col=2)

# Grafici di dispersione
par(mfrow=c(2,4))
plot(data$quota, data$QBSar, main = "QBS - Quota", xlab = "Quota", ylab = "QBS")
plot(data$pH, data$QBSar, main = "QBS - pH", xlab = "pH", ylab = "QBS")
plot(data$Corg, data$QBSar, main = "QBS - Carbonio organico", xlab = "Carbonio organico", ylab =
     "QBS")
plot(data$CaCo3, data$QBSar, main = "QBS - Carbonati", xlab = "CaCo3", ylab = "QBS")
plot(data$sab_tot, data$QBSar, main = "QBS - Sabbia", xlab = "Sabbia", ylab = "QBS")
plot(data$argilla, data$QBSar, main = "QBS - Argilla", xlab = "Argilla", ylab = "QBS")
plot(data$umid_pp, data$QBSar, main = "QBS - Umidita'", xlab = "Umidita' in peso", ylab = "QBS")
plot(data$dens_app, data$QBSar, main = "QBS - Densita' apparente", xlab = "Densita' apparente",
    ylab = "QBS")

# Distribuzione della risposta condizionata a variabili qualitative
boxplot(QBSar ~ localita , data = data, xlab = "Localita'", ylab = "QBS", main = "Localita'",
    cex.axis = 0.6, names = c("Asiago", "Az.Diana", "Baone-C.E.", "Breganze-P.M.", "Ceregnano-S.
    R.", "Cornuda", "Dueville, N.", "Falcade", "Gosaldo", "Le poscole", "Maser", "Pian
    Cansiglio", "Pirio-C.E.", "Sista Bassa", "Vallesana-C.E.", "Vallevecchia"))
boxplot(QBSar ~ coltura_aggr , data = data, xlab = "Coltura aggregata", ylab = "QBS", main = "
    Coltura aggregata")
boxplot(QBSar ~ coltura_aggr_tipoLav , data = data, xlab = "Coltura aggregata e tipo di
    lavorazione", ylab = "QBS", main = "Coltura aggregata e tipo di lavorazione")
boxplot(QBSar ~ coltura_specifica , data = data, xlab = "Coltura specifica", ylab = "QBS", main
    = "Coltura specifica")
boxplot(QBSar ~ coltura_specifica_tipoLav , data = data, xlab = "Coltura specifica e tipo di
    lavorazione", ylab = "QBS", main = "Coltura specifica e tipo di lavorazione")
```

```
boxplot(QBSar ~ esposizione , data = data, xlab = "Esposizione", ylab = "QBS", main = "
    Esposizione")
boxplot(QBSar ~ EC1_2_CLASSI , data = data, xlab = "EC1_2", ylab = "QBS", main = "Salinita'")

# Correlazione tra covariate
cor(data[,c(10,12:18)], use = "complete.obs")
pairs(data[,c(10,12:18)])
```

Modelli di regressione lineare multipla normale

```
## Procedura stepwise forward manuale
# Passo 1
fit.coltSp <- lm(QBSar ~ coltura_specifica, data = data)
summary(fit.coltSp)
# Passo 2
fit.loc2 <- update(fit.coltSp, .~.+localita)
summary(fit.loc2)
anova(fit.coltSp, fit.loc2)
# Passo 3
fit.sab3 <- update(fit.loc2, .~.+sab_tot)
summary(fit.sab3)
# Passo 4 - Modello finale su tutte le osservazioni, senza umidita'
fit.esp4 <- update(fit.sab3, .~.+esposizione)
summary(fit.esp4)
anova(fit.sab3, fit.esp4)

# Passo 5 - Modello finale sui dati raccolti dal 2015
fit.corg5 <- update(fit.umid4, .~.+Corg)
summary(fit.corg5)

## Procedura stepwise backward manuale
# Modello completo stimato sui dati dal 2015
fit.completo <- lm(QBSar ~ localita + coltura_specifica + esposizione + pH + EC1_2_CLASSI + Corg
     + CaCo3 + sab_tot + argilla + umid_pp + quota_p, data = data.noNAumid)
summary(fit.completo)
# Passo 1
fit1 <- update(fit.completo, .~.-CaCo3)
summary(fit1)
# Passo 2
fit2 <- update(fit1, .~.-quota_p)
```

```r
summary(fit2)
# Passo 3
fit3 <- update(fit2, .~.-pH)
summary(fit3)
# Passo 4
fit4 <- update(fit3, .~.-argilla)
summary(fit4)
# Passo 5
fit5 <- update(fit4, .~.-esposizione)
summary(fit5)
# Passo 6 - Modello finale
fit6 <- update(fit5, .~.-EC1_2_CLASSI)
summary(fit6)

# Modello completo stimato su tutti i dati
fit.completoU <- lm(QBSar ~ localita + coltura_specifica + esposizione + pH + EC1_2_CLASSI +
    Corg + CaCo3 + sab_tot + argilla + quota_p, data = data)
summary(fit.completoU)
# Passo 1
fit1U <- update(fit.completoU, .~.-quota_p)
summary(fit1U)
# Passo 2
fit2U <- update(fit1U, .~.-Corg)
summary(fit2U)
# Passo 3
fit3U <- update(fit2U, .~.-CaCo3)
summary(fit3U)
# Passo 4
fit4U <- update(fit3U, .~.-argilla)
summary(fit4U)
# Passo 5
fit5U <- update(fit4U, .~.-pH)
summary(fit5U)
# Passo 6 - Modello finale
fit6U <- update(fit5U, .~.-EC1_2_CLASSI)
summary(fit6U)
```

```r
# Modello completo stimato sui dati con umidita' stimata
fit.completoS <- lm(QBSar ~ localita + coltura_specifica + esposizione + pH + EC1_2_CLASSI +
    Corg + CaCo3 + sab_tot + argilla + umid_pp + quota_p, data = data.umidStim)
summary(fit.completoS)
# Passo 1
fit1S <- update(fit.completoS, .~.-CaCo3)
summary(fit1S)
# Passo 2
fit2S <- update(fit1S, .~.-quota_p)
summary(fit2S)
# Passo 3
fit3S <- update(fit2S, .~.-argilla)
summary(fit3S)
# Passo 4
fit4S <- update(fit3S, .~.-Corg)
summary(fit4S)
# Passo 5
fit5S <- update(fit4S, .~.-EC1_2_CLASSI)
summary(fit5S)
# Passo 6 - Modello finale
fit6S <- update(fit5S, .~.-pH)
summary(fit6S)

## Procedura stepwise backward automatica
# su tutti i dati
fit.completoU <- lm(QBSar ~ localita + coltura_specifica + esposizione + pH + EC1_2_CLASSI +
    Corg + CaCo3 + sab_tot + argilla + quota_p, data = data)
fit.step.U <- step(fit.completoU, direction = "backward")
summary(fit.step.U)
# sui dati dal 2015
fit.completo <- lm(QBSar ~ localita + coltura_specifica + esposizione + pH + EC1_2_CLASSI + Corg
     + CaCo3 + sab_tot + argilla + umid_pp + quota_p, data = data.noNAumid)
fit.step <- step(fit.completo, direction = "backward")
summary(fit.step)
# sui dati con umidita' stimata
fit.completoS <- lm(QBSar ~ localita + coltura_specifica + esposizione + pH + EC1_2_CLASSI +
    Corg + CaCo3 + sab_tot + argilla + umid_pp + quota_p, data = data.umidStim)
fit.stepS <- step(fit.completoS, direction = "backward")
summary(fit.stepS)
```

Analisi dei residui di uno dei modelli

```
# Grafico di dispersione tra i valori stimati e i residui
residui.corg5 <- rstandard(fit.corg5)
plot(fit.corg5$fitted.values, residui.corg5, xlab = "valori stimati", ylab = "residui")
abline(h = mean(residui.corg5, na.rm = T), col=2)
# Istogramma dei residui
hist(residui.corg5, probability = T, main = "Istogramma dei residui", xlab = "Residui")
curve(dnorm(x), add=T)
# Q-Q plot
qqnorm(residui.corg5)
qqline(residui.corg5)
```

Modelli a effetti casuali

```
# Procedura forward manuale
# Su tutti i dati
# Passo 1
fit.eColtSp <- lme(QBSar ~ coltura_specifica, random = ~ 1 | localita, data = data)
# Passo 2
fit.eSab2 <- update(fit.eColtSp, .~.+sab_tot)
# Passo 3 - Modello finale
fit.eCorg3 <- update(fit.eSab2, .~.+Corg)

# Sui dati dal 2015
# Passo 1
fit.eColtSpU <- lme(QBSar ~ coltura_specifica, random = ~ 1 | localita, data = data.noNAumid)
# Passo 2
fit.eSab2U <- update(fit.eColtSpU, .~.+sab_tot)
# Passo 3
fit.eCorg3U <- update(fit.eSab2U, .~.+Corg)
# Passo 4 - Modello finale
fit.eUmid4 <- update(fit.eCorg3U, .~.+umid_pp)

# Procedura backward manuale
# Modello completo stimato su tutti i dati
fit.eCompleto <- lme(QBSar ~ coltura_specifica + esposizione + pH + EC1_2_CLASSI + Corg + CaCo3
    + sab_tot + argilla + quota_p, random = ~1 | localita, data = data)
# Passo 1
fit.e1.b <- update(fit.eCompleto, .~.-argilla)
```

```r
# Passo 2
fit.e2.b <- update(fit.e1.b, .~.-pH)
# Passo 3
fit.e3.b <- update(fit.e2.b, .~.-EC1_2_CLASSI)
# Passo 4
fit.e4.b <- update(fit.e3.b, .~.-CaCo3)
# Passo 5 - Modello finale
fit.e5.b <- update(fit.e4.b, .~.-Corg)

# Modello completo sui dati dal 2015
fit.eCompletoU <- lme(QBSar ~ coltura_specifica + esposizione + pH + EC1_2_CLASSI + Corg + CaCo3
     + sab_tot + argilla + quota_p + umid_pp, random = ~1 | localita, data = data.noNAumid)
# Passo 1
fit.e1.bU <- update(fit.eCompletoU, .~.-argilla)
# Passo 2
fit.e2.bU <- update(fit.e1.bU, .~.-pH)
# Passo 3
fit.e3.bU <- update(fit.e2.bU, .~.-CaCo3)
# Passo 4
fit.e4.bU <- update(fit.e3.bU, .~.-esposizione)
# Passo 5
fit.e5.bU <- update(fit.e4.bU, .~.-quota_p)
# Passo 6 - Modello finale
fit.e6.bU <- update(fit.e5.bU, .~.-EC1_2_CLASSI)

# Modello completo sui dati con umidita' stimata
fit.eCompletoUs <- lme(QBSar ~ coltura_specifica + esposizione + pH + EC1_2_CLASSI + Corg +
    CaCo3 + sab_tot + argilla + quota_p + umid_pp, random = ~1 | localita, data = data.umidStim
    )
# Passo 1
fit.e1.bUs <- update(fit.eCompletoUs, .~.-pH)
# Passo 2
fit.e2.bUs <- update(fit.e1.bUs, .~.-argilla)
# Passo 3
fit.e3.bUs <- update(fit.e2.bUs, .~.-CaCo3)
# Passo 4 - Modello finale
fit.e4.bUs <- update(fit.e3.bUs, .~.-EC1_2_CLASSI)
```

```r
# Procedura backward automatica
# su tutti i dati
fit.eCompleto.ml <- lme(QBSar ~ coltura_specifica + esposizione + pH + EC1_2_CLASSI + Corg +
    CaCo3 + sab_tot + argilla + quota_p, random = ~1 | localita, data = data, method = "ML")
fit.eStep.ml <- stepAIC(fit.eCompleto.ml, direction = "backward")
summary(fit.eStep.ml)
# sui dati dal 2015
fit.eCompletoU.ml <- lme(QBSar ~ coltura_specifica + esposizione + pH + EC1_2_CLASSI + Corg +
    CaCo3 + sab_tot + argilla + quota_p + umid_pp, random = ~1 | localita, data = data.noNAumid
    , method = "ML")
fit.eStepU.ml <- stepAIC(fit.eCompletoU.ml, direction = "backward")
summary(fit.eStepU.ml)
# sui dati con umidita' stimata
fit.eCompletoUs.ml <- lme(QBSar ~ coltura_specifica + esposizione + pH + EC1_2_CLASSI + Corg +
    CaCo3 + sab_tot + argilla + quota_p + umid_pp, random = ~1 | localita, data = data.umidStim
    , method = "ML")
fit.eStepUs.ml <- stepAIC(fit.eCompletoUs.ml, direction = "backward")
summary(fit.eStepUs.ml)
```

Stima dei dati mancanti di umidità

```r
# da procedura backward manuale
fit.Uarg <- lm(umid_pp ~ localita + pH + Corg + quota_p, data = data.noNAumid)
summary(fit.Uarg)
# da procedura backward automatica
fit.Ustep <- lm(umid_pp ~ localita + pH + Corg + CaCo3 + argilla + quota_p, data = data.noNAumid
    )
summary(fit.Ustep)
# test di confronto
anova(fit.Ustep, fit.Uarg)

# Stima dei dati mancanti
umidNA <- data[is.na(data$umid_pp),]
umid.stim <- predict(fit.Uarg, newdata = umidNA)

# Creazione Dataset
data.umidStim <- data
data.umidStim[is.na(data.umidStim$umid_pp),]$umid_pp <- umid.stim
```

Bibliografia

ARPAV (2005). «Carta dei Suoli del Veneto».

ARPAV, Suolo (2018). «Risultati del monitoraggio biologico dei suoli del Veneto».

Grigoletto, M., F. Pauli e L. Ventura (2017). *Modello Lineare – Teoria e Applicazioni con R*. Torino: Giappichelli.

Salvan, A., N. Sartori e L. Pace (2020). *Modelli Lineari Generalizzati*. Springer.