Università degli Studi di Napoli
Federico II

Strutture ad albero per l'analisi di regressione in presenza di effetti moderanti

Contributi metodologici ed applicativi

Gianfranco Giordano

Tesi di Dottorato di Ricerca in
Statistica Computazionale

XXII Ciclo

Dipartimento
di Matematica e Statistica
Università degli Studi di Napoli "Federico II"
via Cintia, Monte Sant'Angelo – 80126 Napoli

Strutture ad albero per l'analisi di regressione in presenza di effetti moderanti

Contributi metodologici ed applicativi

Napoli, 30 novembre 2009

Indice

Elenco delle tabelle

Elenco delle figure

Introduzione

Sempre più spesso e in vari ambiti disciplinari (come ad esempio nelle ricerche sociologiche, economiche, demografiche, epidemiologiche, etc.), si analizzano fenomeni con una struttura informativa gerarchica, in cui i dati si presentano a più livelli: individuale, familiare, territoriale, sociale e così via. In particolare, lo studio delle relazioni tra l'individuo e il contesto che lo circonda, può essere ricondotto all'analisi di fenomeni a struttura gerarchica. I modelli sviluppati negli ultimi tempi, risultati più idonei al trattamento di dati con struttura di varianza complessa, sono denominati *Multilevel Model*.

Questa classe è composta da modelli caratterizzati da una certa flessibilità che permette di includere più dimensioni di analisi: una dimensione micro, relativa all'individuo, e una dimensione macro, riferita al contesto in cui l'individuo vive, formalizzando l'interazione individuo/ambiente attraverso lo studio dell'effetto di variabili macro su scelte e comportamenti individuali.

L'effetto delle variabili a livello macro su quelle a livello micro può essere definito *moderante*, poiché l'influenza che esso rappresenta, modera la relazione di tipo causale tra le variabili di risposta e quelle esplicative (esse sono componenti essenziali di qualsiasi analisi di regressione).

Tra i principali limiti di questa classe di modelli è possibile individuare le restrittive assunzioni sulla parte erratica e l'elevata com-

plessità in presenza sia di un numero elevato di livelli che di variabili esplicative.

Per fronteggiare le problematiche evidenziate, si propone una metodologia di regressione ad albero per l'analisi di una struttura gerarchica dei dati, non soggetta ad ipotesi distribuzionali e senza vincoli sulla parte erratica.

I metodi di segmentazione classici, pur presentandosi come strumenti di analisi di forte validità applicativa, in alcuni contesti possono risultare inadeguati al raggiungimento degli scopi esplorativi o predittivi prefissati. Quindi, pur seguendo la filosofia *divide et impera*, tipica degli approcci ad albero, si definisce un algoritmo innovativo di partizionamento ricorsivo che individua la migliore partizione finale, considerando l'*effetto condizionamento* dovuto alla presenza di una o più variabili moderatrici, espressione della gerarchia di stratificazione della popolazione analizzata.

L'obiettivo di questo lavoro è duplice: metodologico ed applicativo. In un contesto metodologico, si intende effettuare una disamina dei più recenti sviluppi in tema di alberi esplorativi, partendo dalla metodologia CART (*Classification and Regression Trees*, Breiman et al., 1984) [15], pietra miliare di questa classe di modelli, per poi trattare gli algoritmi accelerati a due stadi che fanno uso di misure alternative di impurità (*Two-Stage* [72] e *FAST* [74]). Successivamente si proporranno alcune metodologie di analisi di dati a struttura gerarchica in problemi di regressione ad albero, concentrandosi in particolar modo sulla definizione di due criteri di partizione alternativi e nella ricerca di una misura di *goodness of fit* che consenta di valutare l'efficacia della metodologia proposta.

Tali contributi sono stati sviluppati interamente in ambiente MatLab, sfruttando la versatilità e la modularità del software *Tree Harvest* [94].

In un'ottica applicativa, lo studio delle metodologie considerate sarà completato attraverso la presentazione di diverse applicazioni compa-

rative che hanno ad oggetto sia dati reali sia dataset simulati; ciò al fine di mostrarne i punti di forza e le differenze sostanziali rispetto alle metodologie classiche di analisi multilivello che seguono un approccio di tipo parametrico. Le analisi proposte saranno effettuate con l'ausilio dei software *Tree Harvest* e *MlWin* e per la loro comparazione si farà uso di numerose rappresentazioni tabellari.

Il lavoro di tesi è strutturato in quattro capitoli.

I dati con una particolare struttura di tipo gerarchico, le relazioni in essi presenti e gli strumenti "classici" per il trattamento degli stessi, sono elementi chiave dei temi affrontati nel **primo capitolo**. In particolare si farà riferimento ai concetti di *livello gerarchico* e *tipologia di relazione*, che sintetizzano la particolare struttura, spesso latente, dei dati.

Dopo aver affrontato in maniera dettagliata le tipologie di variabili e le relazioni presenti nei dati, si effettuerà una disamina degli approcci classici, ed in maniera più approfondita dei modelli multilevel che rappresentano, oggigiorno, il punto di riferimento tra le tipologie deputate ad affrontare questi tipi di problematiche. Si cercherà di inquadrare il ruolo svolto da queste metodologie cercando di evidenziare i vantaggi, ma soprattutto i limiti che gli approcci di tipo parametrico necessariamente subiscono a causa dei forti vincoli sulla parte erratica.

Il **secondo capitolo** tratta i metodi di segmentazione binaria. Le metodologie considerate sono il CART e le strutture ad albero a due stadi, soffermandosi sulle proposte *Two-Stage* e FAST per la riduzione del costo computazionale dell'analisi. Inoltre, si effettua una disamina sia dell'aspetto esplorativo sia di quello confermativo, per cui verranno presentate le tecniche di *pruning* e di validazione dell'albero.

Per quanto riguarda l'analisi di dataset con una particolare struttura gerarchica si è notato, da applicazioni ed analisi empiriche, che i vantaggi degli approcci non parametrici, quali i metodi di classificazione e regressione ad albero, sono notevoli, ma allo stesso tempo la metodologia in esame ha il grosso limite di non far emergere la struttu-

ra latente presente nei dati, a causa della procedura di partizionamento che caratterizza i metodi di segmentazione. Questo e ulteriori limiti sono trattati alla fine del capitolo.

Nel **terzo capitolo** si propongono due metodologie di segmentazione binaria denominate *Regression Trees with Moderating Effects*, per consentire l'esplorazione delle relazioni e la struttura gerarchica presente nei dati, attraverso le tecniche di segmentazione binaria. L'idea è quella di seguire un approccio non parametrico, utilizzando perciò una tecnica di regressione che sfrutti tutti i vantaggi dei metodi di segmentazione, superando in tal modo i limiti che tali metodi presentano quando si affrontano strutture di dati multilivello. Tale approccio è basato sulla ricerca di un criterio di partizione che impiega al suo interno una misura che tenga conto dell'influenza delle variabili relative a differenti livelli gerarchici per la determinazione del taglio ottimale ad ogni nodo.

Più dettagliatamente, la prima proposta metodologica (*criterio di partizionamento moltiplicativo*) è basata sul coefficiente di correlazione intraclasse (ICC come acronimo di *intra-class correlation*) che considera il ruolo giocato da una generica variabile moderatrice nella spiegazione della variabile di risposta. Questa misura verrà implementata nell'algoritmo di partizione classico CART e saranno riportate in dettaglio le sua proprietà.

La seconda proposta metodologica (*criterio di partizionamento additivo*), che affronta il problema degli effetti moderanti, è rappresentata dalla definizione di una misura di impurità che considera un effetto additivo della variabile moderatrice sul legame causale tra la variabile di risposta e i predittori.

La fase successiva per entrambi i criteri ha riguardato la definizione di una idonea misura di *goodness of fit* che permette la valutazione dei metodi proposti e, la comparazione sia con la metodologia classica di segmentazione sia con i modelli multilevel.

Infine, nel **quarto capitolo**, verrà presentato uno studio dettaglia-

to delle applicazioni effettuate. Per quanto riguarda lo studio comparativo con tecniche parametriche, quali i modelli multilevel, è doveroso sottolineare che quest'ultimi rappresentano il *benchmarking* verso cui misurare e incrementare le performance delle nuove metodologie proposte, poiché un confronto (*strictu sensu*) sarebbe non corretto dal punto di vista metodologico.

In un primo momento sono stati affrontati degli studi su dataset reali, i quali sono noti in letteratura per essere stati applicati in contesti di analisi multilivello. Successivamente è stata compiuta un'analisi comparativa che misurasse in maniera più dettagliata la capacità di interpretare le relazioni esistenti nella gerarchia dei dati.

La seconda parte di questo capitolo affronta lo studio, in maniera più dettagliata, dei vantaggi evidenziati in ogni metodo, in particolare dell'importanza del numero di osservazioni, della dimensione dei gruppi, dell'effetto moderante definito dalla misura di correlazione intraclasse e, infine, della regola di arresto della procedura.

In ultimo, nell'appendice, si riporta il codice sorgente in linguaggio MatLab delle principali routine sviluppate per la definizione dei nuovi criteri di spilt e i risultati delle applicazioni col software *MlWin*.

Capitolo 1

Dati a struttura gerarchica

1.1 Introduzione

Intorno alla metà degli anni 80 un numero sempre crescente di ricercatori iniziarono a studiare come introdurre, secondo un approccio sistematico, i modelli statistici nell'analisi di strutture di dati gerarchici. In vari ambiti disciplinari (sociologico, economico, demografico, sanitrio etc.), si ha a che fare spesso con fenomeni a struttura gerarchica, in cui i dati si presentano a più livelli: individuale, familiare, territoriale, sociale. In queste circostanze bisogna procedere all'analisi di una relazione tra gli individui e la società.

Gli individui interagiscono col contesto sociale cui appartengono, cioè i soggetti sono influenzati dalle caratteristiche dei gruppi di cui fanno parte e, a loro volta, le proprietà di questi gruppi risentono dell'influenza dei singoli individui.

In simili circostanze, individui (unità) e gruppi (macrounità) sono presi in considerazione come un sistema gerarchico osservabile a differenti livelli; ciò conduce ad un'analisi dell'interazione tra le variabili che caratterizzano gli individui, con quelle che caratterizzano i gruppi.

In molti casi il ricercatore si trova ad analizzare dati che presentano una struttura gerarchica o multilevello: si pensi ai pazienti raggruppati nelle strutture ospedaliere, agli individui appartenenti alla stessa regione, agli impiegati nelle aziende, o ancora ai bambini nelle unità familiari (da non dimenticare poi il caso delle misure ripetute).

Gli approcci classici per la risoluzione di questi problemi si rifanno all'uso di metodologie di analisi di regressione OLS (*Ordinary Least Squares*)[21] [79], ANOVA (Analysis of Variance), ed in particolare ai modelli multilevel, i quali, nascono principalmente con lo scopo di far emergere dai dati le caratteristiche, espressione della gerarchia di stratificazione della popolazione analizzata[1]

Prima di esaminare le tecniche maggiormente utilizzate per la risoluzione di tali problematiche, è utile effettuare un'analisi delle tipologie di variabili e del tipo di relazioni esistenti tra i differenti livelli.

1.2 La struttura dei dati

La maggior parte delle applicazioni, per le quali le tecniche sopra citate sono impiegate, considerano le unità rilevate come appartenenti ad un unico insieme, e quindi, provenienti da una stessa popolazione.

La presenza di una gerarchia nei dati è rilevabile, sovente, nelle applicazioni industriali, dove occorre studiare le relazioni tra le variabili in presenza di una struttura di gruppo, come ad esempio stratificazione di prodotti, di consumatori, segmentazione del mercato, che comporta l'organizzazione dei dati in matrici totalmente o parzialmente appaiate.

La specificazione di un modello statistico che consideri le relazioni sopra descritte, consiste nell'esplicitare un legame tra i fenomeni di

[1]Le problematiche in esame sono discusse in maniera dettagliata in J. Hox. *Multilevel Analysis*[55].

interesse nel modo seguente:

$$Y = f(X_1, X_2, ...X_p) \tag{1.1}$$

dove Y è la variabile da spiegare mentre $X_1, X_2, ...X_p$ sono variabili prescelte per spiegare Y grazie alla funzione $f(.)$. Tale relazione deriva dalla interazione tra conoscenze a priori e risultati sperimentali. Inoltre, è importante definire la natura di queste variabili e il loro ruolo (cosa spiega Y, come si misurano le variabili, cosa è maggiormente rilevante per spiegare Y, etc.), perchè ciò aiuta a formulare più correttamente il legame funzionale.

Alcune di queste variabili X verranno distinte nel prosieguo di questa trattazione come appartenenti ad un livello gerarchico superiore, per cui saranno definite con la lettera Z.

1.2.1 *Tipologia di variabili definite nei differenti livelli gerarchici*

Le variabili presenti in tali strutture di dati possono essere definite per tipologie [66] ad ogni livello gerarchico e si distinguono in:

- globali;
- relazionali;
- analitiche;
- strutturali;
- contestuali.

Le variabili **globali** sono tutte quelle variabili che si riferiscono ad un solo livello di cui fanno parte senza considerare gli altri; esse descrivono le caratteristiche di ogni singola unità (es. genere, dimensione della scuola di appartenenza).

Le variabili **relazionali** sono quelle che, pur appartenendo sempre ad un solo tipo di livello esprimono al contempo la relazione tra le varie unità in esso contenute (es. indici sociometrici quali quello di popolarità o reciprocità di relazioni).

Le variabili **analitiche** e quelle **strutturali** differiscono dalle precedenti in quanto si costruiscono a partire dalle proprie sub unità: quelle analitiche, ad esempio, possono riferirsi all'intelligenza media di un gruppo di studenti, costruito sulla base delle singole rilevazioni in una classe; quelle strutturali, invece, si riferiscono alla distribuzione di variabili relazionali al livello più basso (es. indici di reti sociali). In ogni caso, sia le analitiche sia le strutturali sono costruite per aggregazione (processo *bottom-up*).

Le variabili **contestuali**, infine, vengono costruite per disaggregazione, ovvero seguono un processo di tipo *top-down*. Un tipico esempio è il processo attraverso il quale si assegna ad ogni alunno di una scuola, la dimensione di quest'ultima a cui appartiene.

1.2.2 *Tipologia di relazioni tra variabili definite tra i differenti livelli gerarchici*

Matrici di dati che presentano una struttura gerarchica sono caratterizzate da relazioni tra variabili ai differenti livelli. In situazioni simili, si è in presenza quasi sempre, di osservazioni dipendenti e, di conseguenza, si verifica una situazione di eteroschedasticità della variabile casuale errore. In tali casi sono necessari modelli che permettono di evidenziare l'effetto delle macro unità sulle unità al livello più basso e viceversa.

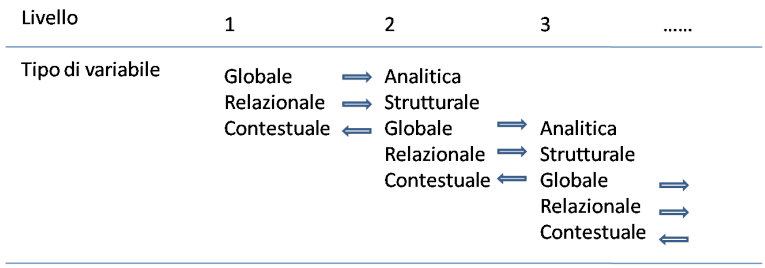

Figura 1.1: Uno schema semplificato della tipologia di variabili ai differenti livelli

Nelle analisi *multilevel* la struttura dei dati della popolazione è gerarchica e i dati campionari sono visti come risultato di un campionamento a più stadi. In molti casi, il campionamento casuale semplice non risulta essere efficiente da un punto di vista dei costi, soprattutto quando si hanno a disposizione una serie di informazioni sulla popolazione, per cui applicare un campionamento a stadi potrebbe risultare più conveniente. In tale situazione bisogna sempre tenere conto del problema della dipendenza delle osservazioni in fase di analisi.

In molte altre situazioni, l'uso di un disegno campionario a stadi è giustificato proprio dall'interesse per le relazioni tra le variabili ai differenti livelli gerarchici. In questo caso la dipendenza tra le osservazioni per l'effetto dei gruppi rappresenta il focus dell'analisi stessa [40].

In altri contesti la struttura gerarchica può riflettere l'annidamento di micro unità in macro unità, può anche essere rappresentato da una serie di misure ripetute all'interno di singoli individui (analisi di dati longitudinali) o da un collettivo di soggetti all'interno di differenti studi scientifici (meta-analisi).

Di conseguenza, nei casi in cui i dati presentano una struttura ge-

rarchica, non può essere ritenuto efficiente il campionamento casuale, ma è opportuno prendere in considerazione un campionamento a stadi, poiché interessati alle relazioni fra le variabili ai differenti livelli.

Nel campionamento a stadi si estraggono le macro-unità, e successivamente da esse, si estraggono le unità all'interno del gruppo; ciò implica la dipendenza fra le unità appartenenti allo stesso gruppo. In questi casi le probabilità di scelta sono note, ma non costanti.

Un errore che frequentemente si commette è quello di ignorare la struttura dei dati e pretendere che le unità al livello più basso siano selezionate indipendentemente da quelle di livello superiore. In realtà, una volta selezionata l'unità primaria, aumentano le probabilità di scelta di un'unità secondaria appartenente a quel gruppo.

Un disegno campionario a stadi può essere descritto graficamente come nella figura seguente:

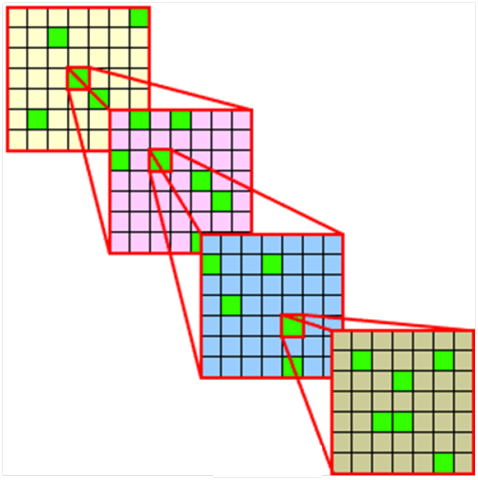

Figura 1.2: Il campionamento a stadi

Le unità in verde sono quelle selezionate ad ogni livello, a partire dalla matrice delle macro-unità in alto e, seguendo un percorso di tipo *top-down*, si scende a cascata fino al arrivare al livello 1 delle unità "elementari".

E' preferibile utilizzare il campionamento a stadi in quanto i costi, per la fase di intervista o testing, sono fortemente ridotti se i soggetti da intervistare sono riconducibili a raggruppamenti geografici di riferimento o ad altri tipi di organizzazione in gruppi.

Dopo la specificazione del disegno campionario, per lo studio di gerarchie, o in generale dei sistemi multilivello, è necessario distinguere tra le relazioni presenti tra le micro-unità, le macro unità e le macro-micro unità. Inoltre, è bene sottolineare che i modelli statistici multilevel, per essere correttamente impiegati, necessitano di un disegno campionario a più stadi (*multi-stage*).

L'uso di tale disegno campionario può sembrare ovvio se si è interessati alle relazioni macro-micro, ma lo è meno nelle restanti situazioni. Seguendo l'impostazione grafica di Tacq (1986) [107], ben nota in letteratura e adottata anche da Snijders e Bosker [104], le figure seguono le seguenti convenzioni: la linea tratteggiata indica la presenza di due livelli, al di sotto c'è il micro livello, al di sopra il macro livello; le lettere maiuscole servono ad indicare variabili misurate al livello macro, mentre quelle minuscole al livello micro; infine la freccia indica la presunta relazione causale.

Il caso classico che si affronta nei sistemi gerarchici multilivello può essere descritto come nella figura 1.3:

L'obiettivo è rilevare l'influenza che la variabile Z al livello macro ha sulla variabile y al livello micro, in cui vi è la presenza congiunta anche dell'effetto della variabile x, legata ad essa da un nesso di causalità.

Lo schema in figura 1.4, invece, rappresenta l'ipotesi in cui non vi è un'influenza delle variabili al macro livello su quello inferiore, per

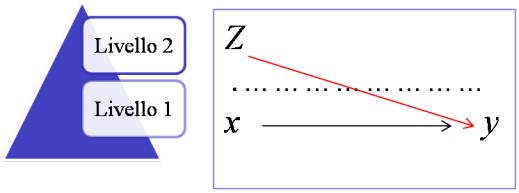

Figura 1.3: Struttura di una proposizione multilivello

cui un disegno di campionamento a stadi è usato solo allo scopo di estrarre le unità al livello più basso.

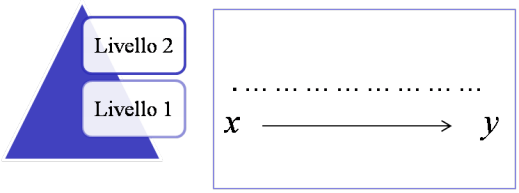

Figura 1.4: Struttura di una proposizione al micro livello

Caso speculare a quello appena visto, si verifica quando si è interessati soltanto alla relazione tra le variabili al livello macro (fig. 1.5).

La situazione più comune nei vari ambiti della ricerca sociale si verifica quando si suppone una interazione delle variabili tra i differenti livelli. Nella figura 1.6 sono illustrati i tre, più ovvi, casi di relazioni macro-vs-micro.

Nel primo caso c'è un effetto netto delle unità al secondo livello su quelle del primo. La seconda situazione è relativa al caso in cui c'è una

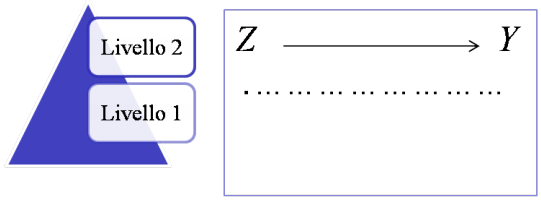

Figura 1.5: Struttura di una proposizione al macro livello

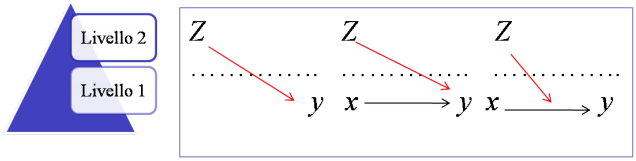

Figura 1.6: Struttura di una proposizione macro-vs-micro

relazione tra Z e y, dato che l'effetto di x su y è preso in considerazione a priori. L'ultima ipotesi rappresenta la macro-micro-interazione, conosciuta anche come la *cross-level-interaction*; in questa circostanza la relazione tra x e y dipende dall'influenza di Z.

Accanto alle precedenti proposizioni, ci sono altri due casi da prendere in esame:

- la relazione micro-macro;
- la catena causale macro-micro-micro-macro.

Nella figura 1.7 una variabile di micro livello influenza la variabile macro (ad esempio il comportamento degli alunni in una classe può incidere su quello degli insegnanti). Nella 1.8 si ha, invece, una catena

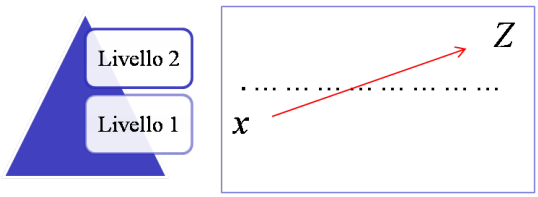

Figura 1.7: Struttura di una proposizione micro-vs-macro

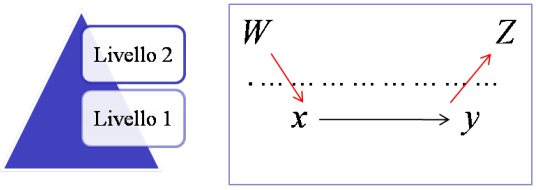

Figura 1.8: Struttura di una catena cusale macro-micro-micro-macro

che spiega attraverso quali variabili a livello micro, si raggiunge la relazione fra le macro variabili W e Z.

1.3 Approcci classici per l'analisi di dataset strutturati

La classificazione delle tipologie di variabili effettuata nei paragrafi precedenti non deve essere considerata come un rigido schema nel quale ogni variabile occupa un proprio ruolo prefissato; l'obiettivo è di fornire un ausilio affinchè questi schemi concettuali facciano chiarezza sul tipo di relazione presente all'interno dei dati e al livello cui essa appartiene.

I modelli di regressione possono essere impiegati per lo studio di "dati dipendenti", circostanza che si verifica quando le osservazioni individuali sono correlate, come accade, per esempio, con una certa frequenza, nelle ricerche in ambito della valutazione dell'istruzione, della medicina e della biologia ecc. Come più volte ricordato, tale dipendenza può derivare dall'esistenza di una struttura gerarchica nei dati (*nested data*) che spesso caratterizza il contesto in cui avviene la rilevazione (dati raggruppati, quali ad esempio gli studenti appartenenti a diverse classi) oppure la stessa operazione di rilevazione (dati longitudinali, si pensi ad esempio alle misurazioni effettuate da diversi soggetti sullo stesso paziente). In entrambi i casi, è opportuno considerare il legame esistente tra le singole osservazioni all'interno di ciascun gruppo di unità statistiche, ricorrendo all'impiego di modelli statistici e stimatori adeguati per l'analisi di queste strutture di dati.

Il presupposto logico di tali tecniche deriva dalla considerazione che il risultato individuale Y dipende da fattori riferibili all'unità statistica oggetto di studio (unità di primo livello) e da fattori riferibili al gruppo cui essa appartiene (unità di secondo livello). Ciò che è possibile

osservare (fattori osservabili) è rappresentato da una o più variabili x riferite all'unità di primo livello e da una o più variabili Z riferite all'unità di secondo livello.

Storicamente, nei problemi di tipo multilevel venivano considerati approcci di analisi che portavano all'aggregazione o disaggregazione di tutte le variabili ad un singolo livello di interesse, seguendo un modello di regressione OLS, ANOVA, oppure altri metodi "standard" di analisi[2]. In questi casi, analizzare le variabili provenienti dai vari strati della gerarchia, come se appartenessero ad un unico livello, determina il sorgere di due principali problemi [55]:

- Il primo è di natura statistica: se i dati subiscono un processo di aggregazione, i differenti valori provenienti dalle varie sub-unità sono combinati in poche unità a livello superiore, comportando inevitabilmente una perdita di informazione che attribuisce di conseguenza, un basso potere esplicativo dell'analisi statistica. Al contrario, se i dati sono disaggregati si rischia di assegnare, per un tipo di variabile, molti più valori (tra l'altro omogenei) alle unità del primo livello, comportando un aumento della dimensione del campione. In tale situazione, i test statistici ordinari considerano i valori disaggregati come informazioni indipendenti provenienti dal vasto insieme di unità di basso livello. Eppure, la reale numerosità campionaria per le variabili disaggregate è costituita dal più basso numero di unità di livello superiore. Pertanto, l'utilizzo dell'ampio insieme di casi disaggregati come numerosità campionaria porta ad ottenere risultati statisticamente significativi che sono totalmente spuri, dando perciò, risultati che sono falsati da questo processo di disaggregazione.
  Questo errore è associato alla presenza di una correlazione intraclasse e alla numerosità degli individui in ogni gruppo. Secondo

[2]Per una disamina più approfondita, si veda Raudenbush and Bryk [82], e Snijders and Bosker, [104]

Barcikowski (1981) [8] la probabilità di commettere l'errore di prima specie cresce nel caso in cui ci si trova in presenza di una alta correlazione intraclasse e nell'ipotesi di un elevato numero di individui in ogni gruppo.

- Il secondo tipo di errore è di natura concettuale: si rischia di analizzare i risultati di un livello e formulare false considerazioni su di esso, poiché le variabili sono inerenti ad un livello differente. Questo tipo di errore è conosciuto in letteratura col nome di *ecological fallacy*.[3]
  Simile all'*ecological fallacy* è l'*atomistic fallacy* [10] [3], che si verifica quando si compiono deduzioni riguardanti la variabilità tra unità di livello superiore inter-gruppo (o la relazione tra variabili a livello di gruppo) sulla base di dati relativi alle unità di livello inferiore. Più in generale, l'*atomistic fallacy* nasce dal fatto che le associazioni tra due variabili a livello individuale possono differire da associazioni tra le variabili analoghe misurate a livello di gruppo. I fattori che spiegano la variabilità tra individui all'interno dei gruppi non sono necessariamente gli stessi che spiegano la variabilità tra i gruppi.

1.3.1 *La correlazione intraclasse*

La correlazione intraclasse è una misura del grado di dipendenza degli individui. Più gli individui condividono le esperienze comuni dovute alla vicinanza nel tempo e nello spazio, più sono simili. Il più alto livello di dipendenza può essere trovato, ad esempio, tra due osserva-

[3] Questo tipo di problema fu affrontato per la prima volta da Robbinson nel 1950 [84], nella descrizione di dati aggregati relativi alla relazione tra la percentuale di persone di colore e il livello di analfabetismo in nove regioni statunitensi nel 1930. Nel suo lavoro Robinson conclude che una *ecological correlation* è quasi sicuramente non uguale alla corrispondente correlazione al livello degli individui.

zioni di gemelli monozigoti, oppure bambini nati e cresciuti nella stessa famiglia. Un altro ben conosciuto esempio di osservazioni dipendenti riguarda le "misure ripetute" sulla stessa persona.

La caratteristica principale dell'analisi multilevel è costituita dal fatto che in genere, trattandosi di dati gerarchicamente organizzati, le osservazioni individuali non sono del tutto indipendenti. Ne consegue che la correlazione media tra individui appartenenti allo stesso gruppo risulterà più elevata di quella tra individui che afferiscono a gruppi differenti.

Riconoscere l'esistenza della correlazione intraclasse è importante perchè cambia l'errore relativo alla stima della varianza nei modelli lineari di regressione tradizionali. Questo errore di stima della varianza rappresenta l'effetto di tutte le variabili omesse e le misure degli errori, assunto che questi siano non correlati.

Nei modelli tradizionali lineari si assume che le variabili omesse hanno un effetto casuale non strutturale; questo aspetto è discutibile nei dati che contengono osservazioni strutturate in maniera gerarchica.

L'*intra-class correlation*, generalmente indicato da ρ, può essere spiegata in diversi modi, ad esempio, può anche essere definita come misura di omogeneità di un gruppo. In modo più formale, con i dati organizzati in una struttura gerarchica a due livelli, l'*intra-class correlation* è definita come la proporzione della varianza della variabile di risposta, che si trova tra le unità al livello superiore. Se siamo in presenza di correlazione intraclasse, come potrebbe succedere con questo tipo di dati, il presupposto delle osservazioni indipendenti dei modelli lineari tradizionali non è rispettato. Un esempio lampante dell'effetto di tale violazione è l'incremento di probabilità di commettere l'errore di prima specie (livello α), in letteratura associato proprio alla presenza della correlazione intraclasse [64].

I test statistici tradizionali sono fortemente basati sull'assunto di indipendenza tra le osservazioni. Se questa ipotesi risulta violata, le

stime degli errori standard prodotte dai test statistici convenzionali sono troppo piccole e, di conseguenza, i risultati che si ottengono appaiono "impropriamente" significativi. La dipendenza tra le osservazioni individuali può essere considerata come un fattore che "riduce" la numerosità campionaria effettiva. Considerando un campionamento a due stadi in cui tutti i gruppi sono costituiti dallo stesso numero di unità elementari, la numerosità campionaria effettiva n_{eff} può essere calcolata come segue (Kish, 1965) [62] [63]:

$$n_{eff} = \frac{n}{1+(n_{clus}-1)\rho}$$

dove n è la numerosità campionaria totale, n_{clus} è la dimensione di ciascun gruppo e ρ è una opportuna misura della correlazione intraclasse [4]. Le correzioni per gli effetti da disegno, come quella proposta da Kish, presentano due pesanti limiti. In primo luogo, la correlazione intraclasse varia al variare della variabile di interesse. In secondo luogo, i problemi relativi all'analisi di strutture gerarchiche sono in genere resi più complessi dalla presenza di variabili misurate su tutti i livelli della gerarchia. Emerge, quindi, la necessità di utilizzare un modello statistico che tenga conto della non indipendenza delle osservazioni e che consenta, allo stesso tempo, di analizzare simultaneamente variabili che "provengono" da diversi livelli della gerarchia [5].

[4] In letteratura sono stati proposti diversi coefficienti per la misura della correlazione intraclasse. Tra i più importanti si segnalano quelli di Donner (1986) [37], e Searle, Casella e McCulloch (1992) [90]

[5] Per maggiori dettagli sull'argomento si faccia riferimento a Barcikowski, (1981) [8] e Cochran (1977) [20]

1.4 Modelli parametrici per l'analisi di problemi multilivello

Per la risoluzione dei problemi affrontati nel precedente paragrafo, sono stati sviluppati nel corso degli anni delle metodologie statistiche di natura parametrica che considerano la presenza di gerarchie strutturate all'interno di matrici complesse.

I **Modelli multilevel**, quali metodologie statistiche sicuramente più adatte ad estrapolare al meglio le informazioni presenti all'interno delle strutture gerarchiche, tengono conto, in maniera esaustiva, sia della presenza di relazioni tra le variabili appartenenti ad ogni livello, sia delle relazioni tra i livelli differenti, considerando in tal modo l'effetto netto sulle unità e le interazioni in esse presenti.

In letteratura sono stati proposti diversi modelli di regressione multilevel: *random coefficient model*, *variance component model* e *hierarchical linear model*. Questi modelli, essendo basati su un approccio comune, formano la classe dei *multilevel regression models*. Essi partono dall'assunto che ci sia un dataset strutturato in maniera gerarchica, una sola variabile esplicativa misurata al livello più basso e più variabili esplicative ad ogni livello presente nella struttura. Concettualmente si è soliti immaginare i modelli di regressione multilevel, come sistemi gerarchici di equazioni di regressioni. Nel prosieguo di questa trattazione si considereranno solitamente due livelli, senza trascurare le possibili generalizzazioni.

La prima grande differenzazione delle tecniche multilevel riguarda il caso in cui si considerano i modelli ad intercetta variabile (*random intercept model*), oppure i modelli a coefficienti casuali (*random slopes*). Per entrambi si useranno le notazioni di seguito indicate.

1.4.1 Notazioni e definizioni

Data un popolazione strutturata in J gruppi, si definisce n_j con $j = 1....J$ la numerosità del $j - esimo$ gruppo. Sia Y la variabile risposta misurata al livello più basso della gerarchia (rilevata, quindi, per ciascun individuo), sia X una variabile esplicativa misurata sul livello degli individui e sia Z una variabile esplicativa misurata sul livello dei gruppi.

Il livello degli individui costituisce il primo livello, quello dei gruppi il secondo livello. Sarà utilizzato l'indice j per i gruppi ($j = 1,.., J$) e l'indice i per gli individui ($i = 1,.., n_j$). I pedici i e j indicano rispettivamente, che la Y assume valori rispetto ai diversi individui (pedice i) presenti nei diversi gruppi (pedice j). La realizzazione della Y per un dato individuo ad un dato gruppo sarà indicato con y_{ij}.

1.4.2 *Il modello ad intercetta casuale*

Questo modello rappresenta un caso semplice del cosiddetto modello gerarchico lineare, conosciuto anche col nome di *Random Intercept Model* (Snijders e Bosker, 1999) [104]. Come nel classico modello di regressione lineare, si è in presenza di una variabile dipendente Y e di un set di predittori X, entrambi misurati al livello degli individui. In particolare, la formalizzazione del modello avviene nel seguente modo:

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + e_{ij} \tag{1.2}$$

Y_{ij} rappresenta la variabile di risposta, con i relativo agli individui e j relativo alle unità di secondo livello[6]. L'obiettivo del multilevel è quello di stimare il valore atteso di Y_{ij}, considerando l'effetto del predittore X sia a livello individuale sia a livello di gruppo. Si ipotizza

[6]In questo modello non compaiono variabili esplicative di secondo livello; l'effetto su di esso sarà specificato nei modelli *random slopes*

infatti che la variabile indipendente, sia caratterizzata da livelli medi differenti, quindi diversi in ogni gruppo.

Tale modello considera l'effetto "gruppo" del predittore attraverso le variazioni dell'intercetta. In altre parole, si stima un modello in cui il coefficiente di regressione è costante nei gruppi e ciò che distingue gli stessi rispetto al predittore è la diversa intercetta. Gli e_{ij} sono invece gli errori al livello degli individui.

L'intercetta variabile a livello di gruppo può essere scomposta in due parti:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{1.3}$$

dove γ_{00} rappresenta l'intercetta media tra tutti i gruppi, mentre u_{0j} rappresenta la parte aleatoria. In altre parole l'intercetta è la somma della media generale e dell'effetto casuale a livello di gruppo, ovvero la misura della sua deviazione intorno alla media.

Sostituendo l'equazione 1.3 nella 1.2 si può ottenere il modello completo:

$$Y_{ij} = \gamma_{00} + \beta_1 X_{ij} + u_{0j} + e_{ij} \tag{1.4}$$

Nel modello così ottenuto, gli u_{0j} possono essere considerati sia come parametri fissi, che come variabili casuali indipendenti ed identicamente distribuite. Il primo caso ha senso quando i gruppi hanno un'interpretazione distinta, riconducendosi quindi all'analisi della covarianza in cui la variabile di raggruppamento è un fattore; nel secondo caso gli u_{0j} sono gli effetti di gruppo non spiegati dalla regressione multilivello. Tale interpretazione porta alla definizione del *Random Intercept Model* in cui l'intercetta varia tra i gruppi in maniera casuale, poiché i gruppi sono considerati un campione estratto casualmente da una popolazione.

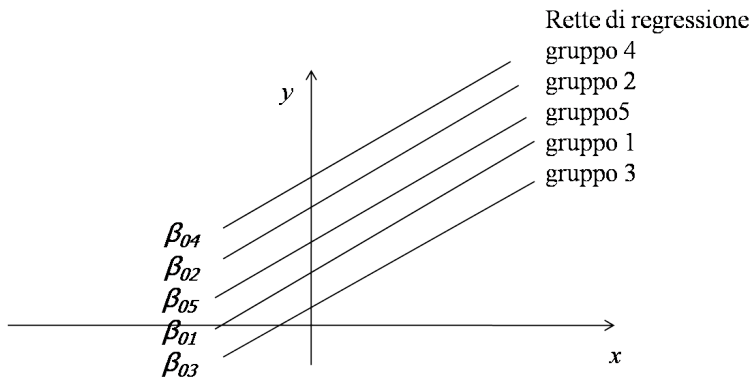

Figura 1.9: Random intercept model

Per comprendere come si giunge a questo modello, bisogna in realtà partire dal considerare il modello ANOVA ad effetti casuali, in cui le variabili esplicative (X e Z) ai diversi livelli non compaiono (questo modello contiene solo i gruppi casuali e le sue variazioni casuali interne). Questo modello è definito *Empty Model*.

Esso può essere espresso come un modello in cui la variabile dipendente è uguale alla somma della media generale γ_{00}, dell'effetto casuale a livello di gruppo u_{0j} e dell'effetto casuale a livello individuale e_{ij}.

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \tag{1.5}$$

I gruppi con elevato u_{0j} avranno in media Y elevato, mentre i gruppi con basso u_{0j} avranno in media Y basso.

Si può assumere che le variabili casuali u_{0j} e e_{ij} abbiano media 0, siano mutuamente indipendenti. Tale modello permette, in questo modo, la partizione base della variabilità dei dati tra i due livelli.

Nel modello 1.4 infatti la varianza totale di Y può essere scomposta come la somma delle varianze a livello 1 e a livello 2 nel seguente modo:

$$\text{var}\,(Y_{ij}) = \text{var}\,(u_{0j}) + \text{var}\,(e_{ij}) = \tau_0^2 + \sigma^2 \tag{1.6}$$

La covarianza tra due individui(i e i' con $i \neq i'$) appartenenti allo stesso gruppo j è uguale alla varianza u_{0j} condivisa dagli stessi:

$$\text{cov}\,(Y_{ij}, Y_{i'j}) = \text{var}\,(u_{0j}) = \tau_0^2 \tag{1.7}$$

e la loro correlazione è

$$\rho\,(Y_{ij}, Y_{i'j}) = \frac{\tau_0^2}{(\tau_0^2 + \sigma^2)} \tag{1.8}$$

Il parametro ρ è un coefficiente di correlazione intraclasse, e indica la correlazione tra due individui dello stesso gruppo o anche la quota di variabilità totale a livello di gruppo. Si può affermare, nell'ipotesi in cui il coefficiente di correlazione é significativamente alto, che ha senso effettuare un'analisi multilevel in quanto buona parte della variabilità è attribuibile ai gruppi, e quindi, il macro livello influenza il micro.

A questo punto il successivo step è l'inclusione nel modello delle variabili esplicative. Come nel classico modello di regressione lineare esse sono usate per spiegare parte della variabilità della Y; nel caso specifico si riferisce alla variabilità sia del primo che del secondo livello. Se si considera una sola variabile indipendente X si ritrova nuovamente il modello 1.4:

$$Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + u_{0j} + e_{ij} \tag{1.9}$$

Le assunzioni fondamentali sono che tutti gli errori u_{0j} e e_{ij} siano mutuamente indipendenti e abbiano medie pari a 0, date dal valore x_{ij} della variabile esplicativa. Per u_{0j} e e_{ij} si assume che siano estratti da popolazioni distribuite normalmente e le loro varianze siano ancora τ_0^2 e σ^2. Le variabili casuali u_{0j} possono essere viste come errori a livello

di gruppo, cioè come effetti di gruppo non spiegati da X. Dal momento che gli errori casuali, contengono quella parte di variabilità della variabile dipendente che non è considerata come funzione di variabili esplicative, si può affermare che questo modello contiene variabilità non spiegata a due livelli annidati. La partizione della variabilità non spiegata sui vari livelli è l'essenza dei modelli gerarchici ad effetti casuali.

All'interno del modello, γ_{00} è sempre l'intercetta media dei gruppi e γ_{10} può essere visto come un coefficiente di regressione non standardizzato come nel modo usuale (infatti in tale equazione $\gamma_{10} = \beta_1$); cioè l'aumento unitario nel valore di X è associato con un aumento medio in Y di "β_1" unità. La varianza residua condizionata al valore di X è:

$$\text{var}\,(Y_{ij}|x_{ij}) = \text{var}\,(u_{ij}) + \text{var}\,(e_{ij}) = \tau_0^2 + \sigma^2 \tag{1.10}$$

mentre la covarianza tra due differenti individui (i e i' con $i \neq i'$) nello stesso gruppo è:

$$\text{cov}\,(Y_{ij}, Y_{i'j}|x_{ij}, x_{i'j}) = \text{var}\,(u_{ij}) = \tau_0^2 \tag{1.11}$$

La frazione di variabilità residua ascrivibile al livello 1 è data da $\sigma^2/(\sigma^2 + \tau_0^2)$ e per il livello 2 questa frazione è :$\tau_0^2/(\sigma^2 + \tau_0^2)$.

Della covarianza o correlazione tra due individui dello stesso gruppo, una parte può essere spiegata dai rispettivi valori di x, mentre l'altra parte non è spiegata. Questa è il coefficiente di correlazione intraclasse residuo:

$$\rho_I\,(Y|X) = \frac{\tau_0^2}{\sigma^2 + \tau_0^2} \tag{1.12}$$

Questo parametro è analogo all'usuale coefficiente di correlazione intraclasse, ma ora i parametri r_0^2 e σ^2 sono riferiti alle varianze del modello 1.4, che include gli effetti della variabile x, mentre prima erano riferiti alle varianze nell'*Empty Model*.

Quando il coefficiente di correlazione intraclasse è 0, ad esempio, u_{0j} è uguale a 0 per tutti i gruppi J, allora il raggruppamento è irrilevante per la variabile Y che condiziona X, e si può usare il normale modello di regressione lineare. Se il coefficiente di correlazione intraclasse residuo, o equivalentemente τ_0^2 è significativo, allora il modello lineare gerarchico è un metodo di analisi migliore di quella di regressione *Ordinary Least Squares* (OLS).

In conclusione nel *Random Intercept Model*, i parametri da stimare sono quattro:

- i coefficienti di regressione γ_{00} e γ_{10} o (β_1);
- le componenti della varianza τ_0^2 e σ^2.

Ovviamente è possibile generalizzare il modello *Random Intercept Model* a più di due livelli.

1.4.3 *Il modello "completo" a coefficienti casuali: Random slopes*

Nei modelli ad intercetta casuale, i gruppi differiscono rispetto al valore medio della variabile dipendente in cui l'unico effetto casuale è attribuibile all'intercetta. La relazione fra variabile dipendente e variabile esplicativa può tuttavia differire tra i gruppi in più modi: è possibile, ad esempio, che gli effetti dello stato socio-economico degli studenti di una scuola sul loro rendimento, sia più forte in alcune classi rispetto ad altre. Questo fenomeno, nell'analisi della covarianza, è conosciuto come eterogeneità della regressione fra i gruppi, nei modelli gerarchici è noto come *random slopes* [104].

Nella situazione appena descritta, la stima dei parametri di un modello multilevel può essere concettualmente distinta in due fasi successive. Nella prima fase a livello degli individui, vengono realizzati,

all'interno di ciascun gruppo, modelli di regressione separati, al fine di predire la variabile risposta Y in funzione della variabile esplicativa X; nella seconda fase si introducono le variabili esplicative misurate a livello di gruppo che comportano la variazione dei coefficienti di regressione. Il modello in esame può essere specificato come segue:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \tag{1.13}$$

Ritornando al modello con una variabile esplicativa, si ha che in questa equazione di regressione, β_{0j} è la classica intercetta, β_{1j} è l'usuale coefficiente di regressione per la variabile esplicativa X misurata sul livello degli individui, mentre e_{ij} rappresenta il solito termine d'errore.

Come nel *random intercept model*, anche in questo caso la differenza rispetto al modello di regressione non gerarchico consiste nel fatto che ogni gruppo possiede una diversa intercetta, β_{0j} e un differente coefficiente di regressione β_{1j}. Inoltre, si assume che, all'interno di ciascun gruppo, gli errori al livello individuale siano indipendenti e normalmente distribuiti con media nulla e varianza σ^2, $e_{ij} \sim \mathrm{N}(0,\sigma^2)$. A causa della variazione tra le unità di livello superiore, i coefficienti in esame prendono il nome di coefficienti casuali. Le macro-unità sono perciò viste come un campione proveniente da una più vasta popolazione di gruppi.

A questo punto i coefficienti β_{0j} e β_{1j} del modello di regressione gerarchico, possono essere suddivisi in un coefficiente medio e una parte che risente della dipendenza delle unità a livello superiore, come nel seguente modo:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \tag{1.14}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j} \tag{1.15}$$

Anche in questo caso si assume che i termini di errore nelle equazioni di regressione a livello di gruppo u_{0j} e u_{1j}, spesso denominati macro-errori, siano normalmente distribuiti con media nulla e varianze τ_0^2 e τ_1^2, rispettivamente. Inoltre, si assume che i macro-errori siano indipendenti tra i gruppi e dagli errori di livello individuale e_{ij}, mentre σ_{u01}^2 rappresenta la covarianza tra i macro-errori u_{0j} e u_{1j}.

$$u_{0j} \sim \mathrm{N}\,(0,\tau_0^2);\ u_{1j} \sim \mathrm{N}\,(0,\tau_1^2);\ \mathrm{cov}\,(u_{0j};u_{1j}) = \sigma_{u01}^2$$

I coefficienti γ_{00}, γ_{10}, γ_{01}, γ_{11}, poiché non hanno la caratteristica di variare tra le unità di appartenenti al livello macro, sono detti coefficienti fissi.

Sostituendo le equazioni 1.14 e 1.15 nella equazione 1.13, il modello di regressione multilevel può essere illustrato in un'unica equazione di regressione:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}X_{ij}Z_j + u_{1j}X_{ij} + u_{0j} + e_{ij} \quad (1.16)$$

Il termine $X_{ij}Z_j$ è denominato *cross-level interaction* poiché risente dell'effetto moderante delle variabili esplicative misurate su differenti livelli della gerarchia come mostrato in figura 1.10. La parte $[\gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}X_{ij}Z_j]$ nell'equazione 1.16 viene denominata parte fissa del modello, poiché contiene i coefficienti fissi, la parte $[u_{1j}X_{ij} + u_{0j} + e_{ij}]$ al contrario, contiene i termini casuali di errore, ragion per cui viene denominata parte casuale del modello. Questo segmento dell'equazione costituisce una struttura complessa di errore e, come si può notare dalla formula, gli errori per le osservazioni all'interno delle macro unità sono correlati poiché u_{0j} e u_{1j} risultano comuni per le osservazioni che appartengono al medesimo gruppo.

Il modello 1.16 implica non solo che gli individui all'interno dello

stesso gruppo hanno valori di Y correlati, ma anche che questa correlazione, così come la varianza di Y è dipendente dal valore di X, (il termine d'errore u_{1j} è connesso con X_{ij}). Da ciò deriva che l'errore totale sarà differente per differenti valori di X_{ij}, situazione questa, che nei modelli di regressione ordinari, prende il nome di "eteroschedasticità". Risultano, pertanto, violate le assunzioni di indipendenza e di omoschedasticità degli errori su cui si basano i modelli di regressione ordinari.

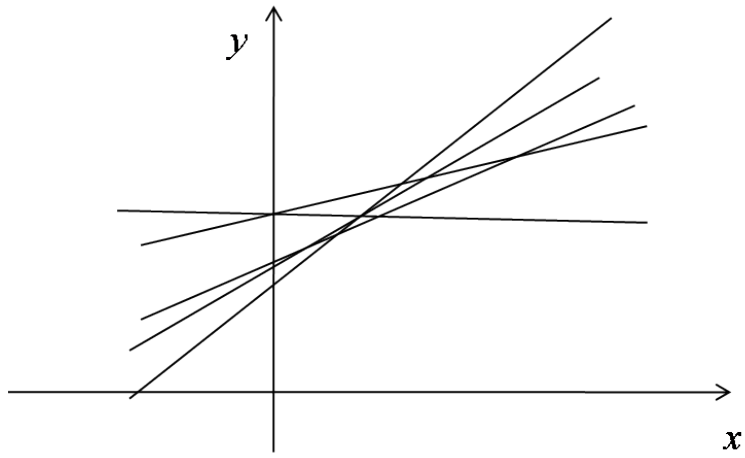

Figura 1.10: Random slope model

Attraverso l'equazione di regressione 1.16 è, dunque, possibile stimare, attraverso i coefficienti ad effetti fissi, gli effetti indipendenti delle variabili di secondo livello, di quelle di primo livello e la loro interazione. Il modello multilevel, inoltre, permette di quantificare la variabilità nei diversi livelli della gerarchia:

- variabilità entro il gruppo, espressa dalla varianza σ^2;

- variabilità tra i gruppi, espressa dalle varianze degli effetti casuali τ_0^2 e τ_1^2.

Gli effetti stimati dal modello possono essere suddivisi in due segmenti:

il primo riguarda la **parte fissa**

- γ_{00} è l'intercetta; rappresenta il valore di Y qualora sia X che Z presentano valore zero;
- γ_{01} è l'effetto del predittore del livello 2 (variabile esplicativa Z);
- γ_{10} è l'effetto del predittore del livello 1 (effetto di X su Y quando Z assume valore zero);
- γ_{11} è l'effetto dell'interazione tra i predittori del livello 1 e del livello 2.

il secondo riguarda la **parte casuale**

- σ^2 varianza intra-classe (tra le unità di livello inferiore) controllando per l'effetto di X;
- τ_0^2 varianza condizionata dell'intercetta rispetto a Z, (esprime la variabilità tra le macro unità per la parte relativa alla sola intercetta);
- τ_1^2 varianza condizionata del coefficiente di regressione ripetto a Z, (esprime la variabilità tra le macro unità per la parte legata all'effetto interazione);
- σ_{u01}^2 covarianza condizionata tra intercetta e coefficiente di regressione di primo livello.

Quando nel modello in esame si verifica che la variabilità residua tra le unità di secondo livello relativa alle intercette e ai coefficienti di regressione risulta trascurabile, la parte casuale a livello macro nelle equazioni 1.14 e 1.15 tende allo zero; di conseguenza tendendo a zero anche le stime delle varianze ad esse collegate τ_0^2 e τ_1^2.

In una simile circostanza, il coefficiente di correlazione intraclasse è prossimo allo zero ed il modello di regressione multilevel 1.16 si riduce ad un classico modello di regressione multipla che include variabili indipendenti misurate indistintamente sia nel primo che nel secondo livello, poiché è inesistente la struttura gerarchica. In questa situazione, gli individui all'interno dei gruppi possono essere considerati indipendenti. Al contrario, l'esistenza di una variabilità significativa tra le intercette o tra i coefficienti di regressione, comporta la presenza di una elevata correlazione intraclasse e giustifica l'adozione del modello di equazione 1.13.

1.4.4 *Stima dei parametri nei modelli multilevel*

Il metodo maggiormente utilizzato per la stima dei parametri nei modelli multilevel (intercetta, coefficienti di regressione e componenti della varianza) è quello della Massima Verosimiglianza (Maximum Likelihood - ML). Questo metodo rappresenta una generale procedura di stima: produce stime asintoticamente efficienti e consistenti e, inoltre, in presenza di grandi campioni, le stime di Massima Verosimiglianza sono generalmente robuste rispetto a leggere violazioni dell'assunzione di normalità distributiva degli errori [55].

Nonostante presenti molteplici punti di forza, in alcuni contesti di analisi, il metodo della Massima Verosimiglianza è risultato non ottimale, ragion per cui sono stati sviluppati metodi di stima alternativi. Tra i principali si ricordano il metodo dei minimi quadrati generalizzati, (*Generalized Least Squares*), l'approccio bayesiano, le equazioni

di stima generalizzate (*Generalized Estimating Equations*).

Maximum Likelihood e Restricted or Residual Maximum Likelihood

Il metodo della Massima Verosimiglianza rappresenta il metodo di stima più utilizzato nei modelli multilevel e consiste nel massimizzare la Funzione di Verosimiglianza, definita in base alla probabilità di osservare una data realizzazione campionaria condizionatamente ai valori assunti dai parametri oggetto di stima.

Nell'ambito dei modelli di regressione multilevel esistono due differenti Funzioni di Verosimiglianza:

1. la Funzione di Verosimiglianza Completa (*Full Maximum Likelihood* - FML)

2. le Funzione di Verosimiglianza Ristretta (*Restricted Maximum Likelihood* - RML)

Nella prima sono inclusi sia i coefficienti di regressione che le componenti della varianza, mentre nella funzione del secondo tipo sono incluse le sole componenti della varianza. Per quanto riguarda i coefficienti di regressione, questi vengono stimati in una fase successiva. Entrambi i metodi producono le stime puntuali dei parametri, gli errori standard e la devianza complessiva del modello.

In particolare, il metodo FML, nel momento in cui vengono stimate le componenti della varianza, tratta i coefficienti di regressione come quantità fisse e incognite, senza prendere in considerazione i gradi di libertà persi stimando gli effetti fissi. Il metodo RML, al contrario, stima le componenti della varianza dopo la rimozione degli effetti fissi dal modello. Ne consegue che le stime FML delle componenti della varianza, rispetto a quelle ottenute con il metodo RML, sono distorte e risultano generalmente troppo piccole.

Le stime RML sono, dunque, meno distorte e presentano inoltre la proprietà che nelle situazioni in cui i gruppi siano perfettamente bilanciati, coincidono con le stime ottimali, ottenute mediante l'analisi della varianza.

Pertanto, sotto il punto di vista teorico, il metodo RML produce stime migliori rispetto al metodo FML, soprattutto nei casi in cui il numero dei gruppi è piccolo. Nella pratica, tuttavia, le differenze tra i due metodi di stima sono piuttosto contenute e il metodo FML continua ad essere frequentemente utilizzato; esso presenta sia il vantaggio di avere un costo computazionale generalmente ridotto rispetto al metodo RML, sia la caratteristica di poter utilizzare un test chi-quadrato basato sulle verosimiglianze per confrontare modelli annidati che differiscono sia nella parte fissa che nella parte casuale, in quanto i coefficienti di regressione sono inclusi nella funzione di verosimiglianza. Col metodo RML, al contrario, il test chi-quadrato basato sulle verosimiglianze, può essere utilizzato per confrontare modelli che differiscono esclusivamente nella parte casuale.[7]

Il calcolo delle stime di massima verosimiglianza richiede una procedura iterativa. Al primo *step* della procedura vengono generati valori iniziali per i diversi parametri, generalmente ottenuti stimando i parametri incogniti come se la regressione riguardasse un singolo livello gerarchico. Nei passaggi successivi, le stime iniziali vengono iterativamente migliorate fino al momento in cui la procedura di stima converge. Tuttavia, una ridotta numerosità campionaria o una non corretta specificazione del modello potrebbero causare problemi di convergenza impedendo l'arresto della procedura di stima. Nelle analisi multilivello, inoltre, la mancata convergenza dell'algoritmo è spesso attribuibile alla presenza, nel modello, di molte componenti della varianza prossime allo zero.

[7]Per una accurata trattazione dei metodi FML e RML si rimanda a Searle, Casella, McCulloch, (1992)[90] e Longford, (1993)[69].

Il metodo di stima della Massima Verosimiglianza risulta perciò non adatto in situazioni in cui la numerosità campionaria è esigua e quando non risulta soddisfatta l'ipotesi di normalità distributiva degli errori.

Metodi Bayesiani

In molti casi è utile stimare anche gli effetti di gruppo u_{0j}, pur essendo questi ultimi variabili latenti e non parametri[104]. Il criterio che si adotta per la stima è *"l'empirical Bayes estimation"*, che conduce alla definizione delle cosiddette "medie a posteriori"[104].

Per giungere alla stima di u_{0j} è necessario prendere in considerazione due insiemi di informazioni:

1. quelle relative ai dati inerenti ciascun gruppo j;
2. l'assunzione che i τ_j^2 siano variabili casuali normali con media nulla e varianza costante pari a τ^2 .

In pratica si combinano le notizie relative ai dati con quelle relative alla popolazione. Si può partire considerando per semplicità un Empty Model:

$$Y_{ij} = \beta_{0j} + e_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij} \tag{1.17}$$

Nel momento in cui la media generale γ_{00} è stata stimata, stimare u_{0j} equivale a stimare β_{0j}. Per fare ciò bisogna prendere in esame la media di gruppo e la media generale, cioè β_{0j} è stimato in base alla seguente combinazione:

$$\hat{\beta}_{01}^{EB} = \lambda_j \beta_{0j} + (1 - \lambda_j)\, \gamma_{00} \tag{1.18}$$

Questa è la stima *"Empirical Bayes"*, dove i pesi sono funzione delle componenti della varianza. La formulazione dello stimatore può essere vista come una stima OLS per i gruppi j, corretta in base alla media generale. La varianza stimata dello stimatore è:

$$\text{var}\left(\hat{\beta}_{01}^{EB} - \beta_{0j}\right) = (1 - \lambda_j)\,\tau_0^2 \tag{1.19}$$

Generalized Least Squares

Le stime dei minimi quadrati generalizzati (Generalized Least Squares - GLS) possono essere ottenute da una procedura di massima verosimiglianza restringendo il numero di iterazioni a uno. Una caratteristica che contraddistingue tale metodo è la velocità impiegata, particolarmente ridotta, per la stima dei parametri del modello.

Le stime ottenute col metodo dei minimi quadrati generalizzati si possono approssimare alle stime di massima verosimiglianza poiché sono asintoticamente equivalenti, pertanto, su grandi campioni, le due procedure producono stime praticamente indistinguibili. Dal momento che le stime GLS sono molto più rapide da calcolare rispetto alle stime FML, queste possono essere utilizzate come stime provvisorie nel caso in cui l'algoritmo di massima verosimiglianza risulti particolarmente lento, come nel caso di data set estremamente grandi. Inoltre, il metodo dei minimi quadrati generalizzati risulta particolarmente utile nei casi in cui le procedure di massima verosimiglianza non convergono.

L'analisi dei risultati GLS può essere molto utile al fine di individuare il problema che determina la non convergenza dell'algoritmo di massima verosimiglianza.

Tuttavia, ricerche basate su simulazioni, hanno evidenziato come gli stimatori GLS siano meno efficienti rispetto agli stimatori ML; ne segue che qualora non sussistano problemi di rapidità computaziona-

le o di convergenza, il metodo della massima verosimiglianza risulta certamente preferibile [119].

1.4.5 *Principali test di ipotesi nei modelli multilevel*

Test di Wald

Uno dei test più utilizzati per la verifica di ipotesi nei modelli di regressione multilevel è il test di Wald in cui la statistica test Z viene calcolata rapportando la stima puntuale del parametro di interesse all'errore standard della stima stessa. La distribuzione di riferimento per la statistica Z è la normale standardizzata. Il test di Wald si basa sull'assunto che i parametri sottoposti a verifica di ipotesi abbiano una distribuzione campionaria normale, con una varianza campionaria che può essere stimata a partire dalla matrice di informazione. Come discusso da Fears et al. (1996) [42], in situazioni particolari, la statistica di Wald non risulta adatta a testare le componenti della varianza, soprattutto nei casi in cui queste siano prossime allo zero o nei casi in cui la numerosità campionaria sia molto ridotta. Si precisa, inoltre, che gli errori standard utilizzati per la costruzione del test sono di natura asintotica, pertanto sono validi per grandi campioni [8].

E' opportuno precisare che nelle regressioni multilivello, la numerosità campionaria, rilevante per i coefficienti di regressione e le componenti della varianza di secondo livello, è costituita dal numero dei gruppi che, generalmente, non è molto elevato.

[8] Non si conosce con precisione quale sia la numerosità campionaria sufficiente affinché gli errori standard possano essere considerati accurati; per gli approfondimenti su questo tema si rimanda agli studi di simulazione condotti da Van der Leeden et al. (1997) [120].

Deviance test

Il *deviance test* o anche *likelihood ratio test* si basa sul principio che, quando i parametri di un modello statistico sono stimati attraverso il metodo *maximum likelihood* (ML), la stima fornisce la *likelihood*, che può essere trasformata nella devianza. Infatti questa è definita come meno due volte il logaritmo naturale della *likelihood*. In genere non si considera direttamente il valore della devianza, ma le differenze nelle devianze di diversi modelli applicati agli stessi dati, ad esempio:

- M_0 è il modello con m_0 parametri e devianze D_0;
- M_1 è il modello con m_1 parametri e devianze D_0.

Il test considerato sarà :

$$D_0 - D_1 = -2 \ln L_0 + 2 \ln L_1 \tag{1.20}$$

L'ipotesi nulla sarà:

$H_0 : D_0 - D_1 \sim \chi^2$ con $m_1 - m_0$ gradi di libertà.

Questo test può essere applicato sia alla parte fissa che alla parte random del modello. Se la devianza è stata calcolata in base al criterio di stima Residual ML, si possono effettuare confronti solo tra modelli che presentano stessa parte fissa e differiscono solo nella parte random.

1.5 Vantaggi e limiti dei modelli gerarchici lineari

Nei precedenti paragrafi sono state analizzate le caratteristiche principali dei modelli multilevel, tralasciando le possibili generalizzazioni

a più livelli gerarchici e la loro presentazione in forma matriciale; sono state evidenziate le ragioni del loro utilizzo e i limiti principali dei modelli di regressione classici.

Volendo riassumere quanto fin qui esposto, allo scopo di sottolineare le differenze principali tra i modelli di regressione lineare ed i modelli multilevel, si riportano in tabella 1.1 le assunzioni di base su cui essi sono fondati.

Modelli di regressione lineari	Modelli multilevel
- Linearità della relazione funzionale	- Linearità della relazione funzionale
- Normalità distribuzionale	- Normalità distribuzionale
- Omoschedasticità	- Eteroschedasticità
- Indipendenza delle osservazioni	- Dipendenza delle osservazioni

Tabella 1.1: Confronto tra modelli parametrici

Per dati a struttura gerarchica l'applicazione dell'analisi multilevel comporta i seguenti vantaggi:

- Interdipendenza: l'idea generale che spiega questa caratteristica è che individui appartenenti alla medesima rete di relazioni siano più vicini o abbiano dei comportamenti tra loro più simili di quanto non accada con individui appartenenti a reti di relazioni diverse (correlazione intra-classe)

- Scomposizione struttura dell'errore (varianza) in una o più fonti di variabilità (una o più componenti) corrispondenti alle diverse unità di analisi (es. primo e secondo livello) riuscendo così ad esprimere anche la variabilità tra i gruppi.

- Le fonti di variabilità possono essere generate da variabili esplicative relative a ciascun livello.

Inoltre, i modelli multilevel consentono di:

- Eliminare la distorsione nella stima degli errori standard dei parametri;

- Stimare l'effetto del gruppo (*group effects*) scomponendo la variabilità in due componenti: quota interna (*within*) ai gruppi e tra gruppi (*between*);

- Introdurre variabili esplicative a livello di gruppo (*group-level predictors*) cercando così di dare una descrizione della variabilità tra gruppi (*random effects model*);

- Modellare gli effetti di interazione o *cross-level*.

Tra i principali limiti dell'analisi multilevel va sottolineato che:

- Nonostante il rigore metodologico di tali modelli c'è la necessità di sviluppare teorie che specifichino a livello di gruppo ed a livello individuale quali fattori possano congiuntamente configurare un determinato *outcome*, ad es. il supporto sociale;

- Come tutti i modelli statistici, anche i modelli multilevel necessariamente semplificano processi complessi. Un limite intrinseco che l'analisi multilevel condivide con gli alti metodi di regressione è il fatto di verificare separatamente gli effetti indipendenti di variabili.

- L'analisi multilevel non consente infine di abbracciare la complessa fenomenologia delle possibili relazioni tra variabili, poiché implica una struttura di regressione in cui una singola variabile dipende da un insieme di altre variabili.

Capitolo 2

Analisi esplorativa attraverso metodi di segmentazione binaria

2.1 Introduzione

Nel presente capitolo, sono trattate le tecniche di segmentazione binaria, a partire dalla metodologia CART (*Classification and Regression Trees*, Breiman et al., 1984 [15]) che ha rappresentato la pietra miliare per questa classe di modelli. Successivamente si è passati ad una disamina dei recenti sviluppi in tema di alberi esplorativi.

Tale percorso metodologico è stato affiancato, in ogni sua fase, da un supporto computazionale rappresentato dal software *Tree Harvest* [94] che ha il compito di implementare le diverse proposte metodologiche, sia quelle note, sia quelle proposte nel successivo capitolo, in un approccio interattivo [4]. Tali proposte consistono nell'applicazione di un metodo di regressione ad albero che tenga conto in maniera esaustiva della presenza di relazioni tra le variabili appartenenti a differenti

livelli gerarchici di una matrice di dati, considerando sia l'effetto netto sulle unità sia le interazioni in esse presenti.

Nel capitolo successivo, allo scopo di risolvere le problematiche tipiche dell'analisi multilevel e superare i limiti dei metodi di segmentazione, si proporrà una metodologia di regressione ad albero innovativa, non soggetta ad ipotesi distribuzionali, e senza vincoli sulla parte erratica. Seguendo la filosofia *divide et impera*, si definisce un algoritmo di partizionamento ricorsivo che individua la migliore partizione finale, condizionata dalla presenza di una o più variabili moderatrici, espressione della gerarchia di stratificazione della popolazione analizzata.

La proposta metodologica si basa sulla generalizzazione del criterio di partizionamento del CART, attraverso la definizione di due algoritmi di split che tengano in considerazione il principale effetto moderante della variabile Z rispetto alla predizione di X su Y secondo due criteri alternativi:

- **Criterio additivo** attraverso l'impiego di una misura d'impurità che tiene conto in maniera additiva dell'effetto della variabile Z sulla relazione diretta tra la variabile di risposta Y e i predittori X;
- **Criterio Moltiplicativo** attraverso l'impiego di una misura di impurità che si fonda sulle proprietà del coefficiente di correlazione intraclasse.

2.2 Le strutture ad albero

La segmentazione è un'analisi asimmetrica che presuppone la presenza di una variabile dipendente o di risposta che deve essere spiegata da un insieme di predittori. In generale, i metodi di segmentazione seguono un approccio "supervisionato" di tipo non parametrico per l'analisi dei dati caratterizzati da alta dimensionalità, sia nel numero

di variabili sia nel numero di unità e dalla non linearità nel legame di dipendenza tra le variabili.

L'approccio supervisionato, si differenzia da un approccio non supervisionato, tipico dei metodi di *cluster analysis*, per la presenza di una variabile di risposta che guida il processo iterativo di apprendimento fino al raggiungimento dell'obiettivo di classificazione o regressione. Tale approccio è non parametrico o *distribution free* perché non sono richieste assunzioni probabilistiche, ed è volto, quindi, alla definizione di modelli intrinsecamente più flessibili, capaci di gestire interazioni non lineari tra le variabili, i cui risultati sono di facile interpretazione. I modelli che ne derivano sono delle strutture ad albero, che, in quanto tali, non sono esprimibili attraverso una semplice forma funzionale dipendente da parametri.

Obiettivo dei metodi di segmentazione è, quindi, la costruzione di una struttura ad albero per descrivere la dipendenza di una variabile di risposta da un insieme di variabili esplicative in problemi di classificazione e regressione. In particolare se la variabile di risposta è qualitativa, si perviene ad una classificazione ad albero, se la variabile dipendente è numerica si perviene ad una regressione ad albero.

Per albero si intende un modello grafico costituito da un insieme finito di elementi, detti *nodi*, che si ripartiscono a partire da un nodo iniziale, che è la cosiddetta *radice* della struttura. La numerazione del generico nodo t è tale che al nodo figlio di sinistra sarà attribuito il numero $2t$ mentre al nodo figlio di destra $2t + 1$[1].
Si tratta di un grafo aciclico diretto, in cui l'orientamento dei segmenti che uniscono i nodi, i cosiddetti *archi*, indica la direzione dei legami esistenti tra i nodi stessi. Si distinguono poi i *nodi interni*, usualmente rappresentati da cerchi dai *nodi terminali* (foglie), rappresentati da

[1]Tale approccio fu proposto dagli autori del software statistico SPAD (Cisia Institute, France). In questo modo è possibile risalire alla posizione di ogni nodo dal suo numero, potendo risalire la struttura dal nodo padre e viceversa.

quadrati, a seconda che siano ulteriormente bipartiti in due nodi discendenti o meno (si veda fig. 2.1).

Il punto di partenza è rappresentato da un insieme di N individui sui quali sono osservate K variabili esplicative ed una variabile dipendente. Come è possibile notare dalla figura 2.1, si ottiene un modello grafico in cui è definita una partizione finale dell'insieme dei dati presente al nodo radice in sottogruppi disgiunti ed esaustivi rappresentati dai nodi terminali dell'albero ai quali sarà assegnata una classe o un valore di risposta. Per definizione, i nodi terminali presenteranno un grado di omogeneità interna maggiore rispetto al gruppo di partenza, omogeneità che è valutata in riferimento alla distribuzione della variabile di risposta. I predittori quindi, giocano il ruolo di generatori delle possibili partizioni (*Split*) in modo da caratterizzare il passaggio delle unità dal nodo radice ai suoi discendenti.

Il nome di segmentazione binaria deriva dal fatto che in ciascuna partizione, il numero di sottogruppi è costante e pari a due, pervenendo in questo modo ad una struttura elementare ad albero binario.

Una possibile generalizzazione definisce ad ogni passo del percorso di suddivisione, la corrispondenza nel diagramma di due (segmentazione binaria, es. AID, CART) o più (segmentazione multipla, es. CHAID) nodi che rappresentano i gruppi di unità che si formano a quello stadio del processo. La segmentazione migliore (a questo e ad ogni passo successivo) è individuata sulla base di una regola di ottimalità che tiene conto della omogeneità *entro* e della eterogeneità *tra* i sottoinsiemi per la variabile criterio. Ciascun gruppo formato ad uno stadio può essere ulteriormente suddiviso in uno stadio successivo fino a quando il termine è raggiunto all'interno di una regola di arresto.

Una volta ottenuta la struttura ad albero, sarà poi possibile individuare quali interazioni tra i diversi split caratterizzeranno l'appartenenza ad un determinato nodo terminale piuttosto che ad un altro.

I metodi di segmentazione hanno una duplice valenza, sia esplorati-

va sia decisionale, in base all'impiego dell'albero per scopi descrittivi, oppure per scopi decisionali, quali strumento induttivo di previsione o di classificazione di nuovi individui che impiega "algoritmi supervisionati", ossia basati sull'"apprendimento" del fenomeno, attraverso la conoscenza a-priori delle osservazioni di una variabile criterio distinta dalle altre variabili esplicative.

È infatti possibile condurre un'analisi esplorativa mediante la segmentazione al fine di descrivere la struttura di dipendenza tra le variabili e utilizzarla come strumento di stratificazione. Ogni procedura di segmentazione è caratterizzata da un criterio di partizione, da una regola di arresto della procedura e, infine, da una regola di assegnazione di una classe o di un valore, alle unità di un nodo terminale.

Al fine poi di impiegare una struttura ad albero per l'analisi confermativa o decisionale, occorre definire una procedura induttiva per il passaggio dal campione osservato ad un ipotetico nuovo campione di cui si vuole prevedere la variabile di risposta.

I vantaggi derivanti dall'uso di questa tecnica sono da ricercarsi in due motivazioni principali: l'intuitività con la quale i risultati sono esposti, derivante dall'output grafico che l'algoritmo produce, dalla caratteristica forma ad albero (grafo aciclico orientato) e l'agevole interpretazione delle regole che discriminano l'appartenenza all'una o all'altra categoria.

In relazione alla natura qualitativa o quantitativa della variabile di risposta si parlerà di classificazione o regressione ad albero. Nei problemi di classificazione, dove le modalità della variabile di risposta sono definite a priori, il nodo terminale sarà caratterizzato da quella modalità che presenta il più alto numero di osservazioni. Nei problemi di regressione ad albero le unità di un nodo terminale saranno caratterizzate dal valore medio ad esse associate. In entrambe i casi l'assegnazione è probabilistica nel senso che ad essa è associata una misura di errore.

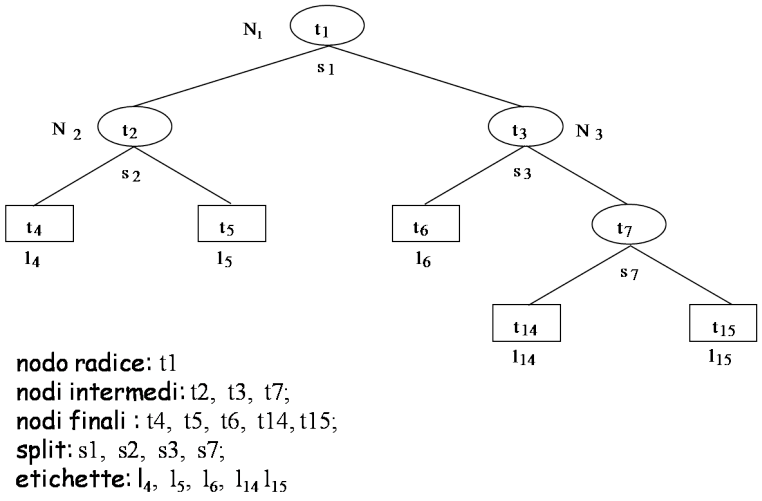

Figura 2.1: Un esempio di struttura ad albero

2.2.1 La costruzione dell'albero esplorativo

Ogni procedura di segmentazione è caratterizzata da un certo numero di fasi che guidano la costruzione dell'albero esplorativo. Tali fasi possono essere sintetizzate nel seguente modo:

- *Il criterio di partizione.* E' il passaggio chiave che caratterizza le diverse metodologie impiegate (CART, FAST, Two-Stage, ecc.) e consiste in un algoritmo di partizione ricorsivo che genera, partendo dal nodo radice, i gruppi sempre più omogenei internamente ed eterogenei esternamente;

- *La regola di arresto della procedura.* Si compone di una serie

di vincoli definiti al fine di controllare la dimensione dell'albero finale;

- *L'assegnazione della risposta.* Consiste in una regola di assegnazione di una classe, o di un valore, alle unità di un nodo terminale.

2.3 Il criterio di partizione ricorsiva

Come già più volte ricordato, il principale obiettivo della procedura è quello di definire una regola di classificazione/predizione sulla base di un campione di apprendimento (denominato anche *training set*) i cui i valori sono relativi ad una variabile di risposta Y, e un set di K variabili esplicative $(X_1, \ldots, X_k, \ldots, X_K)$.

Data $(Y, \mathbf{X})$ una variabile multivariata, il primo problema da affrontare nella costruzione dell'albero è come determinare lo split binario degli N oggetti, in modo da ottenere delle partizioni sempre più piccole e maggiormente omogenee al loro interno rispetto alla distribuzione della variabile di risposta.

L'ammontare massimo dei possibili split è un numero finito ed è dato dal numero di predittori e dalla loro natura. Ad esempio, se si considera una variabile binaria il numero dei possibili split è paria a uno, se invece la stessa è categorica il numero è pari a $2^{m-1} - 1$, se è ordinale il numero è $m-1$ e infine se è numerica il numero è $N-1$. La tabella 2.1 riporta il numero di tutte le possibili variabili di split che possono essere generate da ogni variabile esplicativa secondo le loro scale di misura.

Ad ogni nodo dell'albero, l'algoritmo genera una ingente quantità di split, poiché prende in considerazione un predittore alla volta e, di esso, ne calcola tutti i possibili split.

Natura del predittore	Numero di modalità	Numero di split
Variabile binaria	2	1
Variabile nominale	m	$2^{m-1}-1$
Variabile ordinale	m	$m-1$
Variabile continua	N	$N-1$

Tabella 2.1: Origine delle variabili di split

Al fine di scegliere la partizione $p \in P$, tra tutte quelle generate dal set di predittori X, che deve essere considerata per suddividere le unità contenute in un generico nodo t, occorre fare riferimento ad una misura che consenta una valutazione comparativa di tutti i possibili tagli, in modo tale da poterne poi selezionare il "migliore".

La definizione di una tale misura consiste nella traduzione in termini formali dell'obiettivo principale della segmentazione ad albero, e cioè, la suddivisione di un nodo, ovvero di un insieme di osservazioni, in più sottonodi (sottoinsiemi di osservazioni), che siano quanto più possibile omogenei al loro interno ed eterogenei esternamente.

A tal proposito, in letteratura si parla di *impurità* quale concetto sintesi di quanto detto fino ad ora. Con il termine impurità si intende l'inverso del grado di omogeneità di un collettivo di unità statistiche sulle quali sono osservate una serie di caratteristiche.

Dato $p\,(j|t) \geq 0$ il numero di unità del nodo t che appartiene alla classe j della variabile Y, con $\sum_{j=1}^{J} p\,(j|t) = 1$, si definisce *indice di impurità* una funzione non negativa ϕ tale che:

1. ϕ è massimo solo al punto $(1/J, 1/J, ..., 1/J)$

2. ϕ raggiunge il minimo solo nei punti $(1, 0, ..., 0)$, $(0, 1, ..., 0)$, $(0, 0, ..., 1)$

3. ϕ è una funzione simmetrica di $p(j|t)$.

dove J rappresenta il numero di distinte modalità (o valori) della variabile considerata.

Tale funzione è massima quando le osservazioni cadute in un nodo sono uniformemente distribuite tra le diverse modalità (o valori) della Y mentre è minima quando tutte le osservazioni assumono la stessa modalità (o valore).

Ci sono diverse funzioni di impurità che soddisfano queste tre proprietà. Nei problemi di classificazione, le misure di impurità maggiormente utilizzate sono le seguenti:

1. *tasso di errata classificazione*

$$i(t) = 1 - \max_j p(j|t) \tag{2.1}$$

2. *indice di eterogeneità del Gini*

$$i(t) = 1 - \sum_j p(j|t)^2 \tag{2.2}$$

3. *indice di entropia*

$$i(t) = -\sum_j p(j|t) log p(j|t) \tag{2.3}$$

Nei problemi di regressione, il criterio di split è basato sulla ricerca di uno split che riduce l'impurità e si traduce in una misura di varianza o di devianza di Y, dove quest'ultima sarà riferita alle sole unità del nodo:

$$i(t) = \sum_{\mathbf{x}_n \in t} (y_n - \bar{y}_t)^2 \tag{2.4}$$

dove $\bar{y}_t$ è la media dei valori di risposta nelle unità del nodo t, $\bar{y}_t = \frac{1}{N_t} \sum_{x_n \in t} y_n$, N_t è il numero totale di osservazioni nel nodo t dove la somma riguarda tutti gli y_n, tale che $\mathbf{x}_n \in t$.

A questo punto è possibile calcolare l'impurità totale dell'albero. Si definisce *impurità totale* dell'albero T la somma delle impurità nei nodi terminali appartenenti all'insieme $\tilde{T}$:

$$I(T) = \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t)p(t) \tag{2.5}$$

dove $I(t)$ è l'impurità nel nodo t pesata dal numero di unità che dal nodo radice cadono nel nodo t, con $p(t) = N_t/N$, mentre $\tilde{T}$ rappresenta il set di tutti i nodi terminali dell'albero T.

Se s è uno split candidato a dividere un generico nodo t in due nodi figli t_l e t_r e, p_l e p_r sono proporzioni di oggetti al nodo t splittati da s in t_l e t_r rispettivamente, allora una misura della riduzione dell'impurità nel passaggio dal nodo padre ai nodi figli si definisce:

$$\Delta i(t, s) = \{i(t) - [i(t_l)p_{t_l} + i(t_r)p_{t_r}] \tag{2.6}$$

dove Δi è chiamato **decremento di impurità**. Quest'ultimo può essere usato come criterio di split: un alto valore di esso significa che lo split generato ha un forte potere discriminante. La procedura, dunque, troverà il migliore split s^* che massimizza la precedente equazione. $p_{t_{l/r}}$ è la proporzione di unità del nodo t che cade nel discendente di sinistra o di destra rispettivamente.

Nelle procedure classiche di segmentazione la ricerca di una struttura ad albero, che descriva al meglio i dati analizzati, si concretizza nella identificazione di una partizione finale che minimizzi la somma ponderata dell'impurità contenuta nei nodi, cioè la (2.5), che equivale

a massimizzare la (2.6) ricorsivamente ad ogni nodo t.

2.3.1 La metodologia *CART*

Le partizioni sono determinate sulla base dei predittori: occorre definire il numero di modi possibili in cui partizionare in G gruppi le modalità di ciascun predittore. Nella *segmentazione binaria*, si costruiscono alberi binari (per $G = 2$) suddividendo in due soli gruppi, nel nodo figlio di destra t_r e di sinistra t_l rispettivamente, le unità di ciascun nodo interno t.

Il generico split è indicato con s mentre S rappresenta l'insieme di tutti i possibili split. In generale, l'insieme S include tutte le partizioni binarie possibili generate dall'insieme $\mathbf{X}$ dei predittori al nodo t. La migliore partizione s^* è determinata tra le migliori partizioni di ciascun predittore, ciascuna delle quali è ottenuta massimizzando il decremento di impurità al nodo t:

$$\max_{s \in S_k} \Delta i(t, s) = \max_{s \in S_k} \{i(t) - (i(t_l)p(t_l) + i(t_r)p(t_r))\} \qquad (2.7)$$

dove $S_k \in S$ è l'insieme delle partizioni generate dal generico predittore X_k con $(k = 1, ..., K)$ e $p(t_l)$ e $p(t_r)$ pari alle distribuzioni condizionate dei nodi di sinistra e di destra dovute allo split s.

In sostanza, l'**algoritmo CART** si compone dei seguenti passi:

- *Passo 1.* si genera l'insieme S di tutte le partizioni possibili ottenute dal set di predittori $\mathbf{X}$;

- *Passo 2.* per ogni partizione s dell'insieme S si calcola il decremento di impurità secondo la (2.7);

- *Passo 3.* si determina la miglior partizione s^* a cui è associato il massimo decremento di impurità.

L'algoritmo è applicato ad ogni nodo fino a che la costruzione dell'albero non si arresta.

Il costo computazionale di questa metodologia è molto elevato. Basti pensare, infatti, al caso in cui i predittori impiegati sono di numerosità elevata ed inoltre parte di essi sono in scala numerica o nominale. In questo caso il numero di split s che deve essere generato ad ogni nodo è considerevole soprattutto se si pensa che ogni volta, per ognuno di essi, l'algoritmo deve calcolare il decremento di impurità per poi selezionare la migliore partizione.

2.3.2 Il criterio di split (*Two-Stage*)

Sebbene la ricerca scientifica e tecnologica abbia permesso di poter disporre di calcolatori elettronici che consentono elevatissime prestazioni, facendo si che il metodo CART possa essere ampiamente utilizzato anche con dataset di "grandi" dimensioni, oggi, riferendosi ad un contesto di data mining, l'esplorazione di *huge database* richiede necessariamente una riduzione dei costi computazionali che garantisca tempi accettabili per l'esecuzione delle procedure.

Mola e Siciliano [72, 73] hanno proposto un criterio di partizione denominato *Two-Stage* per scegliere il migliore split. Tale approccio si basa sull'assunzione che un predittore X_k non è meramente usato come un generatore di partizioni, ma gioca anche un ruolo "globale" nell'analisi.

In particolare, la metodologia two-stage si pone in un'ottica diversa rispetto al ruolo classico giocato dai predittori nella determinazione della partizione finale dell'albero. In base a tale metodologia la determinazione della partizione avviene attraverso un algoritmo a due stadi che dapprima determina un sottoinsieme di predittori che meglio spiegano la variabile di risposta Y, considerando in questo modo l'aspetto "globale" nella predizione, e successivamente identifica la migliore partizione binaria scelta tra quelle generate dal sottoinsieme di

variabili precedentemente selezionate, delle quali si considera ora l'aspetto "locale" di generatori di split.

L'algoritmo di partizione prende in considerazione il costo computazionale, il quale dipende dalla natura ricorsiva della procedura e dal numero di possibili partizioni ad ogni nodo dell'albero. Gli sviluppi successivi di tale procedura affrontano il problema dell'efficienza computazionale; infatti, da questo punto di vista, diventa cruciale la crescita esponenziale della procedura, soprattutto quando si opera su dataset molto grandi oppure quando si ricorre all'utilizzo di tecniche *ensemble*.

Ad ogni nodo t il *two stages* può essere definito come:

- **selezione globale**; uno o più predittori sono scelti in quanto maggiormente predittivi nei confronti della variabile di risposta secondo un dato criterio; il predittore selezionato è usato per generare un set di partizioni o di split. In questo stadio occorre definire un indice al fine di valutare la Riduzione Globale Proporzionale dell'Impurità ***Global IPR*** (ottenendo in questo modo una gerarchia dei predittori sulla base del potere predittivo degli stessi) della variabile di risposta Y al nodo t, dato dal predittore X;

- **selezione locale**; la migliore partizione è selezionata come la più predittiva e col maggior potere discriminante per i sottogruppi secondo la regola data. In questo stadio occorre definire un indice al fine di valutare la Riduzione Locale Proporzionale dell'Impurità (**Local IPR**) della variabile dipendente Y dato dalla partizione p generata dal predittore X.

Nell'ambito delle tecniche di classificazione ad albero il Global IPR è definito come l'indice τ di Goodman e Kruskal [50]:

$$\tau_t(Y|X) = \frac{\sum_i \sum_j p_t^2(j|i)p_t(i) - \sum_j p_t^2(j)}{1 - \sum_j p_t^2(j)} \tag{2.8}$$

dove $p_t(i)$, per $i = 1, \ldots, I$, è la proporzione del numero di unità appartenenti al nodo t che hanno la i-esima modalità[2] i di X, e $P_t(j|i)$, per $j = 1, \ldots, J$, è la proporzione del numero di unità al nodo t appartenenti alla classe j di Y data la i-esima modalità i di X. Da notare che il denominatore nell'equazione 2.8 è l'indice di eterogeneità di Gini.

Nei problemi di regressione, la misura di impurità utilizzata si basa sulla devianza, per cui in questo caso, il Global IPR può essere definito come *rapporto di correlazione* η^2 *di Pearson*:

$$\eta^2_{Y|X}(t) = \frac{BSS_{Y|X}(t)}{TSS_Y(t)} \tag{2.9}$$

dove SST (*Total Sum of Squares*) è la devianza totale relativa alla variabile di risposta numerica Y e BSS è la devianza tra i gruppi (*between group sum of squares*) dato il predittore X.

In modo simile, la Local IPR sia per la classificazione che per la regressione ad albero sono definite dalle equazioni 2.8 e 2.9, con la differenza che in questi casi gli indici sono calcolati tra la variabile di risposta Y e il set di split s generati dalle funzioni Global IPR.
Più precisamente, per la classificazione ad albero, ad ogni nodo t della procedura di splitting, uno split s della $I - esima$ modalità di X in due sottogruppi (ad esempio $i \in l$ oppure $i \in r$), porta alla definizione di una variabile di split X_s con due categorie denotate da l e r. Il Local IPR è definito come:

$$\tau_t(Y|s) = \frac{\sum_j p^2_{tl}(j|l)p_{tl} + \sum_j p^2_{tr}(j|r)p_{tr} - \sum_j p^2_t(j)}{1 - \sum_j p^2_t(j)} \tag{2.10}$$

[2]In questo caso, il termine modalità è utilizzato indifferentemente per indicare le I categorie di una variabile qualitativa o gli I distinti valori assunti da un predittore quantitativo.

allo stesso modo per la regressione ad albero risulta:

$$\eta^2_{Y|s}(t) = \frac{BSS_{Y|s}(t)}{TSS_Y(t)} \tag{2.11}$$

In sintesi, le fasi dell'algoritmo Two-Stage sono le seguenti:

- *Passo 1.* **Selezione delle variabili (via Global IPR)**
  - *fase 1.* Si calcola per ogni predittore $X_k \in \mathbf{X}$ la riduzione globale proporzionale dell'impurità ottenendo in questo modo una gerarchia dei predittori sulla base del potere predittivo degli stessi;
  - *fase 2.* Si seleziona il migliore o un sottogruppo di migliori predittori sulla base della gerarchia definita alla fase 1;
- *Passo 2.* **Selezione dello split (via Local IPR)**
  - *fase 3.* Si genera l'insieme dei possibili split dai predittori selezionati e si calcola per ognuno di essi la riduzione proporzionale locale dell'impurità;
  - *fase 4.* Si seleziona la migliore partizione s^* come quella che massimizza l'indice di riduzione proporzionale locale dell'impurità.

2.3.3 L'algoritmo di partizione accelerato *FAST*

La metodologia FAST (*Fast Algorithm for Splitting Tree*) introdotta da Mola e Siciliano (1997) [74] rappresenta un contributo di notevole importanza nella riduzione del costo computazionale delle procedure ad albero garantendo allo stesso tempo, la generazione di un albero la cui partizione è proprio quella del CART. L'idea chiave del FAST è quella di selezionare la migliore partizione ad ogni nodo attraverso un

algoritmo accelerato che perviene alla soluzione ottima senza necessariamente esplorare tutti i possibili split contenuti in S.

Come discusso nel precedente paragrafo, quando si applica il criterio *two stage* il miglior predittore potrebbe essere trovato minimizzando il Fattore di Riduzione Globale Proporzionale dell'Impurità di ogni variabile esplicativa X, così come il Fattore di Riduzione Locale Proporzionale dell'Impurità determina lo split ottimo rispetto a tutte le partizioni derivanti dal miglior predittore selezionato precedentemente.

Il principale risultato che deriva dal FAST è che la misura del Global IPR soddisfa la seguente proprietà:

$$\gamma(Y|X) \geq \gamma(Y|s) \tag{2.12}$$

in cui γ è la generica misura Global IPR, e s è il set di split generati dalla variabile X.
L'algoritmo FAST consiste in due step:

- si calcola la misura di Global IPR come nell'equazione 2.8 o 2.9 per tutte le variabili appartenenti alla matrice dei predittori X e ordina in maniera decrescente queste misure (in questo modo si ottiene una gerarchia dei predittori sulla base del potere predittivo degli stessi);

- si calcola la misura di Local IPR come nell'equazione 2.10 o 2.11 secondo la gerarchia definita nella fase precedente cioè col massimo Global IPR. Se il Local IPR di questo predittore è più alto del Global IPR della variabile X successiva, la procedura si ferma, altrimenti continua fino a quando la disuguaglianza è soddisfatta.

In altre parole, l'algoritmo aggiorna la migliore partizione fino a quando il predittore selezionato presenta un indice di riduzione globale

inferiore all'indice di riduzione locale della soluzione corrente. Ciò significa che tale predittore genererà quale migliore partizione una soluzione certamente peggiore (comunque non migliore) di quella corrente; inoltre, la soluzione corrente sarà quella ottimale poiché la selezione dei predittori avviene in senso non decrescente rispetto al potere esplicativo o potere di riduzione dell'impurità e quindi ogni altro futuro predittore sicuramente genererà una riduzione globale inferiore e quindi una partizione non migliore di quella corrente.

Questo algoritmo accelerato permette di trovare la soluzione ottimale che si avrebbe massimizzando la (2.6) con un notevole risparmio del costo computazionale richiesto dalle procedure classiche in ciascun nodo (valutabile anche in base al numero di partizioni da provare prima di determinare la soluzione ottimale). Si dimostra teoricamente e mediante studi di simulazione che in media la riduzione relativa nel numero di split provati dal FAST rispetto all'approccio standard cresce al crescere del numero di modalità distinte del predittore ed al crescere del numero di unità presenti nel nodo (Mola e Siciliano, 1998). Questi risultati sono maggiormente evidenti in presenza di predittori fortemente esplicativi della variabile dipendente, come spesso si riscontra in applicazioni su dati reali.

2.4 L'arresto della procedura e l'assegnazione della risposta ai nodi terminali

Uno dei vantaggi dei metodi di segmentazione consiste nella semplicità interpretativa del diagramma ad albero, purché questo non sia di dimensioni elevate. Per assicurare che tale condizione sia soddisfatta si rende necessaria la definizione dei criteri di arresto che limitino la crescita dell'albero:

a) *Decremento minimo di impurità.* Un nodo è dichiarato "ter-

minale" se la riduzione dell'impurità conseguibile mediante la suddivisione del nodo stesso risulta inferiore ad una soglia prefissata. In questo modo si pone un freno alla crescita di branche il cui contributo alla purezza dell'albero è praticamente nullo;

b) *Numerosità minima del nodo.* Si fissa una soglia minima per il numero di osservazioni contenute in un nodo padre o eventualmente nei nodi figli generati da questo. La regola serve ad ottenere alberi i cui nodi non siano espressione di singole o pochissime unità fornendo così percorsi poco informativi;

c) *Taglia massima dell'albero.* Un'ulteriore regola d'arresto definisce la dimensione massima della taglia dell'albero al fine di limitarne l'espansione. La taglia può essere definita in termini di numero di nodi terminali che è anche pari al numero di suddivisioni (nodi interni) più uno, ma anche in riferimento al numero di livelli dell'albero che danno una misura della profondità della struttura.

I criteri d'arresto possono essere impiegati simultaneamente in modo da creare una struttura la cui espansione rispetti i diversi propositi finalizzati all'ottenimento di un albero con una limitata complessità e una facile interpretazione.

Le regole appena descritte rappresentano un sistema empirico per la scelta della dimensione di un albero di tipo esplorativo, dove l'interesse del ricercatore è quello di evidenziare la struttura di dipendenza esistente tra i dati. Tale sistema prescinde dal problema decisionale e non tiene conto di logiche statistiche. Quando l'obiettivo è invece quello di ottenere un *albero delle decisioni*, allora l'attenzione si sposta su un problema diverso. La scelta della taglia ottimale dell'albero non consisterà nella semplice definizione di alcune regole d'arresto, ma si concretizza in una procedura che attraverso la valutazione

dell'*accuratezza* della regola opererà semplificando una struttura sovradattata tagliando ricorsivamente i legami definiti 'deboli'[3].

Con i metodi di segmentazione si perviene ad una struttura ad albero i cui nodi terminali costituiscono una partizione del campione iniziale in gruppi "puri" al loro interno. Nell'interpretazione dell'albero esplorativo, si seguiranno i diversi percorsi della struttura gerarchica individuando le diverse interazioni tra predittori che conducono le unità a cadere in un nodo terminale piuttosto che in un altro. Data C la classe a-priori $C = (1, ..., j, ..., J)$ delle modalità della variabile si risposta Y, ciascun nodo terminale sarà etichettato attribuendo la classe modale di risposta (in problemi di classificazione):

$$p(j^*|t) = \max_{j \in C} [p(j|t)]$$

o il valore medio (in problemi di regressione):

$$\bar{y}_t = \frac{1}{N_t} \sum_{\mathbf{x}_n \in t} y_n$$

In tal modo, si definiranno ad esempio i diversi percorsi che conducono alla stessa classe di risposta, oppure che spiegano le variazioni in media della variabile di risposta al variare delle diverse interazioni tra predittori.

2.5 L'obiettivo confermativo e il *Pruning* selettivo

Gli alberi esplorativi possono essere usati per investigare circa la struttura dei dati o come strumento di stratificazione di un collettivo, ma

[3]Ne sono un esempio le tecniche di *pruning* basate su metodi di ricampionamento (Breiman, 1996 [13], Efron e Tbishirani, 1993 [38]) o su test statistici che valutano la significatività del taglio delle branche [17] [17]

non possono essere usati, senza trascurarne l'accuratezza, per poter fare induzione. L'accuratezza va valutata in termini di tasso di errata classificazione o previsione; c'è da considerare, inoltre, anche la dimensione dell'albero, poiché tale tasso dipende dalla taglia di quest'ultimo. Quanto più il numero di nodi terminali è alto, tanto più il tasso di errata classificazione sarà basso. Questo comporta un sovradimensionamento della struttura per cui ci si imbatte in quello che viene definito il fenomeno dell'*overfitting*, che conduce ad alti errori di classificazione o di previsione per nuove unità.

Per ovviare a questo deficit di interpretabilità e di predittività, e per ridurre il *trade-off* tra accuratezza e complessità della struttura, l'albero decisionale avrà un numero di nodi terminali contenuto rispetto a quello esplorativo e sarà costruito a partire da un altro campione (normalmente complementare a quello di apprendimento) detto campione test.

Per scegliere l'albero cosiddetto *"onesto"*in termini di dimensioni, Breiman *et al.* [15] definiscono il minimo costo-complessità del *pruning*. Prima di procedere con la descrizione della procedura di *pruning*, occorre definire una misura di errore della struttura ad albero.

- Per la classificazione ad albero, l'errore al generico nodo t è definito come:

$$r(t) = \frac{1}{N_t} \sum_{i=1}^{N_t} (\hat{Y}_t \neq Y_i) \tag{2.13}$$

  dove n_t è la dimensione al $t - esimo$ node, $\hat{Y}_t$ è la classificazione prodotta dall'albero alla stesso nodo. Il tasso di errore totale dell'albero è definito come

$$R(T) = \sum_{h \in H_T} r(t)p(t) \tag{2.14}$$

  dove H_T è il set di tutti i nodi terminali dell'albero T, e $p(t)$ è

la proporzione di malclassificati al $t-esimo$ nodo terminale.

- Per quanto riguarda la regressione ad albero il tasso di errore è definito esattamente come nell'equazione 2.4, che altro non è che la devianza totale (TSS) nel $t-esimo$ nodo diviso la dimensione totale del campione, dove l'errore di previsione totale dell'albero è definita come:
$$RR(T) = \frac{R(T)}{R(t_1)} \tag{2.15}$$
dove $R(t_1)$ rappresenta l'errore al nodo radice.

La procedura di *pruning* si esplicita nel seguente modo: considerando T_{max} come l'albero massimamente espanso, e indicando con $\left|\tilde{T}\right|$ il numero di nodi terminali dell'albero T_{max}. La misura costo-complessità è definita come:
$$R_\alpha(T) = R(T) + \alpha \left|\tilde{T}\right| \tag{2.16}$$
dove α è un parametro di complessità non negativo il quale "governa il *trade-off* tra dimensione dell'albero e il suo *goodness of fit* dei dati"[54].

L'idea è: per ogni α, trovare il sottoalbero $T^*_\alpha \supseteq T_{max}$ che minimizza $R_\alpha(T)$. Quando $\alpha = 0$ la soluzione è rappresentata dall'albero massimamente espanso T_{max}, mentre quando α cresce tanto più la dimensione dell'albero diminuisce.

La procedura di *pruning* risulta essere la stessa per problemi sia di classificazione sia di regressione ad albero, tale che è possibile focalizzare l'attenzione sul problema di classificazione senza perdersi in generalizzazione. La misura di costo complessità definita per ogni nodo interno t ed alla branca T_t è:
$$R_\alpha(t) = r(t)p(t) + \alpha$$
$$R_\alpha(T_t) = \sum_{h \in H_t} r(h)p(h) + \alpha \left|\tilde{T}_t\right|$$

dove $R_{(t)}$ è l'errore di risostituzione al nodo t, $p(t) = \frac{N_t}{N}$ è il peso del nodo t dato dalla proporzione di malclassificati in esso presenti H_t rappresenta l'insieme di nodi terminali della branca di cardinalità $\left|\tilde{T}\right|$. La branca T_t potrebbe essere trattenuta nel modello se:

$$R_\alpha(t) > R_\alpha(T_t) \tag{2.17}$$

l'errore di complessità del nodo t comporta un alto errore di complessità alla sua branca. Così α cresce quando le due misure tendono ad essere uguali; il valore critico di alfa α si ottiene risolvendo la seguente disuguaglianza:

$$\alpha = \frac{R(t) - R(T_t)}{\left|\tilde{T}_t\right| - 1} \tag{2.18}$$

così che α rappresenta per ogni nodo interno t l'aumento di costo per nodo terminale quando si pota la branca che si diparte dal nodo t.
La procedura di semplificazione dell'albero produce una sequenza di sottoalberi $\Omega = T_1 \subset T_2 \subset \ldots \subset T_{max}$, corrispondente ad una sequenza crescente di valori di α, dove T_1 è un albero costituito solo dal nodo radice. Si può dimostrare [15] che il minimo costo-complessità della procedura di *pruning* produce il sottoalbero col tasso minimo di errore dato dal numero dei suoi nodi terminali. In altre parole, se T_α ha cinque nodi terminali, non ci saranno altri sottoalberi $T_s \subseteq T_{max}$ che avranno cinque nodi terminali con un errore più piccolo ([15]).

Per la validazione della struttura ad albero si considera la sua accuratezza: il tasso di errata classificazione ovvero il tasso di errata previsione. In entrambi i casi non potendo determinare il "vero valore" del tasso di errore, occorre ricorrere ad una sua stima. Ci sono tre possibili modi per poter compiere tale stima:

- *Stima di risostituzione*
  La stima di risostituzione è calcolata usando la stesso dataset utilizzato per costruire l'albero. Può essere considerata una stima ottimistica e per questo motivo è poco utilizzata.

- *Stima test set*
  Tale metodo si concretizza con la divisione del dataset in due parti: la prima, che generalmente è costituita dal 70% dei dati a disposizione, prende il nome di campione di apprendimento e viene utilizzata per la costruzione dell'albero. La seconda, detta campione test, viene utilizzata per validare l'albero precedentemente ottenuto al fine di valutare la bontà della struttura stessa in termini di classificazione di nuove unità.
  .

- *Stima cross validation*
  Questa procedura di stima è solitamente utilizzata quando le osservazioni a disposizione sono di scarsa numerosità affinché si possa dividere in due il dataset e ricavarne due campioni di dimensioni soddisfacenti. In questi casi si procede dividendo i dati di partenza in V sottocampioni della stessa dimensione e costruendo tanti alberi a partire dai V gruppi escludendo di volta in volta il $V - esimo$ sottocampione, utilizzandolo successivamente per validare la struttura. Quando si sono esaurite le combinazioni e sono stati creati tutti gli alberi, la media delle stime test set così ottenute fornirà la stima cross validation.

2.6 Vantaggi e limiti dei metodi di segmentazione per l'analisi di dataset strutturati

In contrasto con i metodi classici, le metodologie ad albero presentano dei notevoli vantaggi che si possono riassumere principalmente nei seguenti punti:

- sono tecniche non parametriche che non abbisognano della specificazione di un modello;
- offrono la possibilità di utilizzare predittori di diversa natura;
- danno luogo ad una rappresentazione grafica di facile interpretazione che consente di visualizzare con immediatezza le relazioni esistenti tra la variabile di risposta e i predittori.

Si potrebbe dire che tali metodologie rispondono ad un problema classico della statistica senza presentare molti degli inconvenienti dei metodi classici impiegati al medesimo scopo.

Per quanto riguarda l'analisi di dataset con una particolare struttura gerarchica si è notato, da applicazioni ed analisi empiriche, che i vantaggi sopra menzionati permangono, ma allo stesso tempo la metodologia in esame ha il grosso limite di non far emergere la struttura latente presente nei dati.

Come visto precedentemente gli split generati dalla procedura di segmentazione sono effettuati ricorsivamente *step by step* seguendo una strategia cosiddetta *divide et impera*. Ciò comporta strutture ad albero di taglia considerevole e dunque complesse, ma soprattutto caratterizzate dal concatenarsi di numerose condizioni, dettate dalle

risposte che determinano la partizione dei nodi, derivanti dal criterio di split che considera singolarmente le modalità dei predittori.

In questo caso durante lo split, non viene considerata la presenza delle più complesse relazioni gerarchiche presenti nei dati. Inoltre, soprattutto se l'obiettivo è esplorativo, si rischia di ottenere un albero eccessivamente espanso per cui risulta difficile capire ed interpretare le relazioni in esso esplicitate.

Il risultato è che stratificazioni, gerarchie e relazioni trasversali presenti all'interno dei vari livelli della struttura vengono ignorate.

Capitolo 3

Regressione ad albero con effetti moderanti

3.1 Introduzione

Pur presentandosi come strumenti di analisi di forte validità applicativa, esistono alcuni contesti in cui le strutture ad albero "classiche" possono risultare inadeguate al raggiungimento degli scopi esplorativi o predittivi prefissati.

Si fa riferimento, in particolare, a problemi di regressione in cui, come noto, la variabile di risposta è di tipo numerico, mentre ci sono una serie di predittori ad essa legata che presentano relazioni di tipo gerarchico. In questi casi, la scelta di un metodo basato su un albero binario costruito ricorsivamente attraverso variabili *dummy*, quali gli split, può risultare inefficace nella spiegazione delle relazioni di dipendenza che legano i predittori alla Y.

Come visto nella prima parte di questa trattazione l'approccio più comune per affrontare questo tipo di problematiche risulta essere quello di tipo parametrico denominato *multilevel analysis*. Lo sviluppo di

questi metodi si è notevolmente incrementato negli ultimi anni, poiché l'attenzione verso strutture di dati molto complesse, in particolar modo quando in esse sono presenti relazioni gerarchiche, ha spinto soprattutto in ambito sociologico e medico ad avviare un filone di ricerca che ha prodotto una vasta letteratura a riguardo.

L'obiettivo che ci si prefigge è quello cercare di affrontare le problematiche evidenziate attraverso una metodologia che sia svincolata da una serie di assunzioni teoriche tipiche degli approcci parametrici classici e che sia di facile comprensione circa l'interpretazione dei risultati.

L'idea è quella di seguire un approccio non parametrico, utilizzando perciò una tecnica di regressione che sfrutti tutti i vantaggi dei metodi di segmentazione, e che superi i limiti di tali metodi quando si affrontano strutture di dati multilivello.

3.2 Approccio non parametrico per dati a struttura gerarchica

In questo lavoro di tesi si introduce un approccio ad albero per l'analisi di problemi caratterizzati da una variabile di risposta numerica e da una serie di predittori, sia di natura quantitativa che qualitativa, in cui sia presente un'organizzazione dei dati logico-strutturale su più livelli. Tale approccio è basato sulla ricerca ed applicazione di un criterio di partizione che impiega al suo interno una misura che tenga conto dell'influenza delle variabili relative a differenti livelli gerarchici (Giordano e Aria, 2009)[44] per la determinazione del taglio ottimale ad ogni nodo.

Questa metodologia prende spunto dal *partial predictability trees* (Tutore e al., 2007)[114], in cui si è utilizzato l'indice τ di Goodman

and Kruskal (1979) [50] e si riferisce a contesti di analisi di regressione.

La proposta metodologica si basa sulla generalizzazione del criterio di partizionamento del CART attraverso la definizione di due algoritmi di split, che tengano in considerazione i possibili effetti moderanti della variabile Z rispetto alla predizione di X su Y. In particolare si focalizzerà l'attenzione sulle relazioni ed interazioni di tipo *cross-level* come visto nella figura 1.6.

Con l'acronimo RTME (*Regression Trees with Moderating Effects*) si propone dunque, un innovativo metodo di regressione ad albero che ha il principale obiettivo di impiegare un algoritmo di partizionamento ricorsivo che identifichi la migliore partizione finale condizionata da una o più variabili moderatrici espressione della gerarchia di stratificazione della popolazione analizzata.

L'effetto delle due misure identificate sarà combinato con il classico indice di impurità del CART: una in maniera moltiplicativa, l'altra in maniera additiva rispetto al contributo della variabile Z.

3.2.1 Criterio di partizionamento moltiplicativo

La prima proposta metodologica è basata sul **coefficiente di correlazione intraclasse** ICC (*intra-class correlation*) che considera il ruolo giocato dalla variabile moderatrice Z nella spiegazione della variabile di risposta Y.

In particolare, si è considerata una misura di ICC che possiede le seguenti proprietà:

- l'indice è nullo quando l'effetto moderante è assente;
- l'indice cresce al crescere dell'effetto moderante;
- l'indice è uguale a uno quando l'effetto moderante è massimo.

Coefficiente di Correlazione Intra-Classe

Al fine di esplicitare meglio il coefficiente adottato nel criterio moltiplicativo è utile specificare in dettaglio le sue caratteristiche. Tale misura parte dalla scomposizione della varianza, in *within* e *between* come mostrato nell'equazione 1.6:

$$\text{var}\,(Y_{ij}) = \tau^2 + \sigma^2 \tag{3.1}$$

Per giungere alla sua formulazione si considerano:

- La media delle macro unità j: $Y_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$
- La media generale: $\bar{Y}_{..} = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} Y_{ij} = \frac{1}{N} \sum_{j=1}^{J} n_j \bar{Y}_{.j}$
- La varianza *within* al gruppo j è data da: $S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (Y_{ij} - Y_{.j})^2$

Quest'ultima può variare da gruppo a gruppo. La notazione N rappresenta il numero totale di individui appartenenti all'intero campione, mentre con J si intende il numero totale di individui appartenenti ad una macro unità.

La media pesata delle varianze *within* delle varie macro unità è pertanto:

$$S^2_{within} = \frac{1}{N - J} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(Y_{ij} - \bar{Y}_{.j}\right)^2 = \frac{1}{N - J} \sum_{j=1}^{J} (n_j - 1) S_j^2 \tag{3.2}$$

Per cui la varianza *within* attesa sarà: $E\,(S^2_{whithin}) = \sigma^2$. La varianza *within* deve essere interpretata come effetto residuo delle singole osservazioni all'interno delle macro unità.

Per quanto riguarda la varianza *between*, il discorso si complica in

riferimento alla dimensione delle macro unità. Per gruppi di uguale numerosità essa è definita come:

$$S^2_{between} = \frac{1}{(J-1)} \sum_{j=1}^{J} \left(\bar{Y}_{.j} - \bar{Y}_{..}\right)^2 \tag{3.3}$$

$$S^2_{between} = \frac{1}{\tilde{n}\,(J-1)} \sum_{j=1}^{J} n_j \left(\bar{Y}_{.j} - \bar{Y}_{..}\right)^2 \tag{3.4}$$

Nella formula 3.4 $\tilde{n}$ è definito come: $\tilde{n} = \frac{1}{J-1}\left\{N - \frac{\sum_j n_j^2}{N}\right\} = \bar{n} - \frac{s^2(n_j)}{J\bar{n}}$ dove $\bar{n} = N/J$ è la media della dimensione delle macro unità e $s^2\left(n_j\right) = \frac{1}{J-1}\sum_{j=1}^{J}\left(n_j - \bar{n}\right)^2$ è la sua varianza.

La varianza totale può, quindi, essere mostrata come una combinazione della varianza *within* e *between*:

$$\text{var}\left(Y_{ij}\right) = \frac{1}{(N-1)} \sum_{j=1}^{J}\sum_{i=1}^{n_j} \left(Y_{ij} - \bar{Y}_{..}\right)^2 = \frac{N-J}{N-1} S^2_{within} + \frac{\tilde{n}\,(J-1)}{N-1} S^2_{between} \tag{3.5}$$

Le complicazioni relative alla varianza *between* derivano dal fatto che i residui rispetto alle singole osservazioni, contribuiscono, benché in misura minore, alla sua costruzione. Il valore atteso della varianza *between* sarà perciò:

$$E\left(s^2_{between}\right) = \tau^2 + \frac{\sigma^2}{\tilde{n}} \tag{3.6}$$

In pratica non si conoscono la varianza *between* e *within* della popolazione ma è possibile stimarle attraverso i dati osservati: $\hat{\sigma}^2 = s^2_{within}$ e $\hat{\tau^2} = s^2_{between} - \frac{s^2_{within}}{\tilde{n}}$

Supponendo, ad esempio, di avere un set di dati strutturato su due livelli, dove le micro unità rappresentano l'insieme delle n_i. osservazioni (livello 1) e, le macro unità $n_{.j}$ i rispettivi gruppi di appartenenza (livello 2), si può dividere la variabilità totale in quella *within*, ossia all'interno dello stesso gruppo, e quella *between*, ovvero tra vari gruppi (vedi equazione 3.1).

In tale situazione il coefficiente di correlazione intraclasse ρ si può definire:

$$\rho\left(Y_{ij}, Y_{i'j}\right) = \frac{\text{varianza popolazione tra macro unità}}{\text{varianza totale}} = \frac{\tau^2}{\tau^2 + \sigma^2} \quad (3.7)$$

Il parametro ρ è il coefficiente di correlazione intraclasse e indica la correlazione tra due individui dello stesso gruppo o anche la quota di variabilità totale a livello di gruppo. Nel caso in cui il coefficiente di correlazione é significativamente alto si può affermare che buona parte della variabilità è attribuibile ai gruppi, e che, quindi, il macro livello influenza il micro.

La misura ICC adottata

Un indice che gode delle suddette proprietà è il ben noto *Intraclass Correlation Coefficient* analizzato da Donner nel 1986 [37].

$$\rho_{ICC} = \frac{\text{var (tra le classi)}}{\text{var (tra le classi)} + \text{var}\,(residua)} = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{residual}} \quad (3.8)$$

Dove per classe si intende un livello o uno strato della popolazione risultato della presenza di una variabile moderatrice, espressione della gerarchia informativa presente nei dati [1].

Il coefficiente di correlazione intraclasse è usato per stimare la correlazione di una variabile relativa a due unità dello stesso gruppo di

[1] Si noti come la 3.8 rappresenta una formulazione equivalente alla 3.7

appartenenza, per esempio due studenti della stessa classe. La correlazione intraclasse ha la caratteristica che media e varianza sono comuni relativamente a tutti i membri appartenenti al medesimo gruppo. Questo perché, la correlazione intraclasse fornisce la proporzione della varianza attribuibile alla differenza tra i gruppi.

Dato t un generico nodo dell'albero, si definisce una nuova misura di impurità con un effetto di tipo moltiplicativo come:

$$i_m(t) = \left[\frac{TSS(Y_t)}{v(t)}\right] \cdot [1 - \rho_{Y|Z}(t)] \tag{3.9}$$

dove

- $TSS\,(Y_t)$ devianza totale (Total Sum of Squares) della variabile di risposta al nodo t
- $v\,(t)$ sono i gradi di libertà
- $\rho_{Y|Z}$ è il coefficiente di correlazione intraclasse di Y dato la stratificazione definita da Z al nodo t.

Inoltre si definisce il decremento di impurità come segue:

$$\Delta i(t\,|s\,) = i\,(t) - [i\,(t_l) * p_l + i\,(t_r) * p_r] \tag{3.10}$$

e si identifica il migliore split $s^* \in S$ come

$$\Delta i(t\,|s*\,) = \max! \tag{3.11}$$

3.2.2 Criterio di partizionamento additivo

La seconda proposta metodologica che affronta il problema degli effetti moderanti è rappresentata dalla definizione di una misura di impurità che considera un effetto additivo di Z sul legame causale tra la risposta ed i predittori.

Predizione della Y = effetto della X + effetto moderante della Z

L'impurità al nodo t è:

$$i\,(t) = \left[\frac{TSS\,(Y_t)}{v\,(t)} + \sum_{j=1}^{J} \frac{WSS\,(Y_t\,|Z_j\,)}{g_j\,(t)}\right] \tag{3.12}$$

dove

- $TSS\,(Y_t)$ è la devianza totale (*Total Sum of Squares*) della variabile di risposta al nodo t
- $WSS\,(Y_t\,|Z_j\,)$ è la somma dei quadrati *within* della Y al nodo t condizionata al gruppo j della Z
- $v(t)$ e $g(t)$ sono i gradi di libertà.

L'equazione 3.12 definisce la misura di impurità al nodo t come la combinazione additiva del contributo di X nella predizione di Y, prendendo in considerazione, allo stesso tempo, l'effetto dello split sulla distribuzione condizionata di Y rispetto alla Z.

Seguendo il precedente approccio, si definisce il decremento di impurità al nodo t:

$$\Delta i(t\,|s\,) = i\,(t) - [i\,(t_l) * p_l + i\,(t_r) * p_l] \tag{3.13}$$

e si massimizza

$$\Delta i(t\,|s*) = \max! \tag{3.14}$$

E' possibile dimostrare, per entrambi i criteri proposti, che in assenza di effetto moderante della Z, il processo di apprendimento dell'albero coincide con la soluzione classica della metodologia CART:

$$i\,(t) = \left[\frac{TSS\,(Y_t)}{v\,(t)}\right] * \left[1 - \rho_{Y|Z}\,(t)\right] \equiv i\,(t)_{CART} \tag{3.15}$$

e

$$i(t) = \left[\frac{TSS(Y_t)}{v(t)} + \sum_{j=1}^{J} \frac{WSS(Y_t \,|\, Z_j)}{g_j(t)} \right] \equiv 2 \cdot i(t)_{CART} \qquad (3.16)$$

e gli alberi risultanti producono le stesse partizioni.

In entrambi gli approcci, la ricerca del migliore split s^* consiste nell'identificazione della migliore partizione binaria come compromesso tra la predizione di X e la capacità di esprimere gli effetti moderanti di Z. Questo compromesso può essere considerato in modo additivo o moltiplicativo ed è dipendente dalla tipologia del legame e della relazione gerarchica.

Nel prosieguo della trattazione, verranno effettuate prove empiriche su dataset reali e simulati e sarà inoltre mostrata la validità delle due proposte metodologiche.

3.3 Un'idonea misura di *Goodness of Fit*

La fase successiva alla definizione di questo nuovo criterio di partizionamento, capace di includere l'influenza delle variabili a livello di gruppo, riguarda la valutazione del metodo. A tal fine sarà condotta, nel quarto capitolo, un'analisi comparativa.

Il primo problema affrontato durante il processo di validazione e comparazione dei nuovi criteri è stato quello di misurare il potere esplicativo delle strutture ad albero che essi esprimono. Tali strutture hanno mostrato, fin dall'inizio, un diverso partizionamento degli N oggetti, ma soprattutto si è riscontrato che il numero di nodi terminali è spesso sensibilmente diverso sia rispetto alla classica procedura CART, sia rispetto alle procedure stesse (criterio additivo o moltiplicativo).

Come è noto, strutture ad albero che abbiano un differente numero di nodi terminali non possono essere comparate, se non senza commettere un errore metodologico. Infatti, come descritto nel paragrafo 2.5 ad ogni partizione finale è associata una misura di errore in termini di individui/oggetti malclassificati o predetti. Inoltre, non bisogna trascurare il principio di parsimonia del modello, poiché un albero eccessivamente espanso rischia di rendere incomprensibile la struttura.

Un ulteriore e rilevante problema, è quello di confrontare l'analisi di un fenomeno utilizzando due approcci completamente differenti. Come si è potuto osservare nel primo capitolo, i dati con una struttura gerarchica multilivello sono solitamente analizzati attraverso l'utilizzo di metodi parametrici, in primis i modelli multilevel, i quali si fondano su una serie di assunzioni e ipotesi distribuzionali. Per questa ragione il confronto con una tecnica non parametrica e *distribution free*, non trova la sua ragion d'essere se non per l'accuratezza della previsione/predizione di nuove unità.

In relazione alle considerazioni appena effettuate si è pensato di utilizzare una misura che ricorda, considerato lo schema concettuale, il ben noto Errore Quadratico Medio (MSE *Mean Squares Errors*). Quest'ultimo infatti è una misura di *badness of fit*, e misura, dunque, l'errore quadratico medio dello scostamento del valore osservato da quello predetto.

$$GoF_{overall} = 1 - \left[\frac{\sum_{i=1}^{N} (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2} \right] \tag{3.17}$$

Moderating Goodness of Fit:
Guadagno di accuratezza, in termini di *Within Sum of Squares* di $Y|Z$.

$$GoF_{\text{mod } moderating} = 1 - \left[\frac{\sum_{j=1}^{J} \sum_{i=1}^{N} (\hat{Y}_{ij} - Y_{ij})^2}{\sum_{j=1}^{J} \sum_{i=1}^{N} (Y_{ij} - \bar{Y}_{.j})^2} \right] \tag{3.18}$$

dove:

- i sono gli individui;
- j sono i gruppi.

Gli alberi caratterizzati da un differente numero di nodi terminali sono difficilmente comparabili, pertanto si è pensato di utilizzare una misura di efficienza dell'albero in termini di predizione della risposta, che tenga conto anche della sua complessità. Quindi, per entrambe le misure di *Goodness of fit* si ha rispettivamente:

$$\text{Accuracyratio(ACratio)} = \frac{\text{GoF}_{\text{overall}}}{\text{size of tree}} \tag{3.19}$$

in termini di *Within Sum of Squares*:

$$\text{Accuracyratio(ACratio)} = \frac{\text{GoF}_{\text{moderating}}}{\text{size of tree}} \tag{3.20}$$

Un alto valore dell'*AC ratio* significa che l'albero ha una maggiore capacità di esplicitare le relazioni e l'informazione in esso contenuto con una piccola struttura.

Capitolo 4

Applicazioni e confronto tra approcci differenti

4.1 Introduzione

Al giorno d'oggi, sempre più ricercatori si trovano a dover analizzare dataset con strutture di tipo gerarchico, in diversi ambiti disciplinari. Come mostrato all'interno di questo lavoro di tesi, il metodo più comune per affrontare tali strutture di dati è chiamato *multilevel analysis* (MLA). Tale metodo rappresenta attualmente, il punto di riferimento per questa classe di modelli, mentre, per quanto riguarda lo studio comparativo sviluppato in questa sezione, esso rappresenterà il *benchmarking* per misurare e incrementare le performance delle nuove metodologie proposte.

Questo innovativo metodo di regressione non parametrico è stato applicato per analizzare dataset in cui fossero presenti effetti moderanti (RTME). I nuovi criteri di split sono stati sviluppati per dati strutturati su due livelli, ma ovviamente è possibile una loro generalizzazione considerando più di due strati.

In un primo momento sono stati affrontati degli studi su dataset reali, i quali sono noti in letteratura per essere stati applicati in contesti di analisi multilivello, per verificare subito l'efficacia dei metodi proposti. Successivamente è stata compiuta un'analisi comparativa che misurasse in maniera più dettagliata la capacità di interpretare le relazioni esistenti nella gerarchia dei dati.

Verrà mostrato, in seguito, lo studio relativo all'influenza subita dall'indice di bontà di adattamento al variare della forza dell'effetto moderante, attraverso il coefficiente di correlazione intraclasse. Infine, lo studio proseguirà con l'analisi rispetto alla dimensione del gruppo, al numero totale di osservazioni presente nel dataset ed in base al criterio di stop della procedura.

4.2 I dati e i software utilizzati

La fase successiva alla definizione dei problemi evidenziati e delle soluzioni proposte, si traduce nella specificazione dei dati da utilizzare ai fini dell'analisi statistica; questo momento è strettamente legato a quello della scelta dell'approccio e del metodo statistico che si vuole adottare. In generale, si distinguono dati sperimentali, che sono costruiti ad hoc dal ricercatore, e dati di osservazione, che sono rilevati dal ricercatore ai fini della descrizione di una realtà già esistente.

La ricerca che si avvale del supporto statistico è rivolta principalmente alla generalizzazione, classificazione e spiegazione di una molteplicità di osservazioni, condotte in un ambito definito e delimitato, sulla base di una serie di ipotesi formulate in precedenza. Le conclusioni di una ricerca possono essere estese a un ambito spaziotemporale più ampio se vi persistono le stesse condizioni di base (nel rispetto del principio secondo il quale le condizioni uguali producono risultati simili). A seconda del tipo di indagine, i principali obiettivi dell'analisi statistica possono essere: la descrizione delle caratteristiche dei fe-

nomeni osservati (analisi esplorativa), l'individuazione di modelli per spiegare tali variazioni o la previsione di eventuali variazioni future (analisi confermativa). Per fare ciò sono stati utilizzati diversi metodi di analisi capaci di perseguire obiettivi descrittivi, esplorativi e induttivi.

Per una giusta comprensione del problema, l'analisi va condotta secondo livelli progressivi di complessità. In un primo momento essa riguarda principalmente l'esame dettagliato di ciascuna variabile separatamente (analisi univariata) con il proposito di individuare i dati anomali, le distorsioni, le asimmetrie. Un secondo livello prende in considerazione le relazioni tra variabili o il confronto tra casi. Un terzo livello coinvolge gruppi di variabili (analisi multivariata) tentando di far emergere le componenti strutturali comuni, ricorrendo successivamente a tecniche di aggregazione e disaggregazione, in modo che le componenti residuali non interferiscano con l'interpretazione delle regolarità riscontrate. A tutti gli approcci è possibile affiancare i metodi che consentono di effettuare l'analisi grafica[1].

La base di dati a cui si fa generalmente riferimento in questo tipo di analisi è una classica matrice *individui per variabile* in cui ogni riga rappresenta un'osservazione relativa ad un individuo o ad un qualsiasi fenomeno oggetto di studio. Sulle colonne della matrice sono rappresentate le variabili che misurano vari aspetti del fenomeno. La loro influenza sulla variabile di risposta può essere vista sotto un duplice aspetto, quello di semplice predittore oppure di variabile moderatrice, in modo tale da rendere la struttura un caso particolare delle matrici a tre vie.

[1]Nel presente lavoro, non è stato riportato lo studio descrittivo univariato, effettuato preliminarmente per l'applicazione dei metodi statistici su cui compiere lo studio comparativo, al fine di non appesantire la trattazione e perché esula dagli obiettivi dell'analisi

4.2.1 L'ambiente di lavoro Matlab

Il software *Tree Harvest* è stato sviluppato in ambiente MatLab R2008a e si compone di numerose routine di analisi integrate in un'unica interfaccia utente interattiva[2].

La scelta dell'ambiente di lavoro è dovuta all'elevata diffusione del linguaggio MatLab in ambito scientifico e alla grande potenza di calcolo che lo stesso assicura. Inoltre, l'enorme disponibilità sulla rete internet (attraverso siti specialisitici, newsgroup scientifici e il sito web MathWorks.com) di tool d'analisi per le metodologie più svariate, assicura la possibilità di affiancare al software *Tree Harvest* [94] gli strumenti necessari e complementari alle diverse esigenze dell'analista.

E' proprio grazie a queste proprietà, che si è avuto modo di implementare i due nuovi criteri di partizionamento, sfruttando così, in maniera modulare, il codice sorgente del software.

4.2.2 La procedura Multilevel

Per compiere l'analisi attraverso i modelli multilevel, è stato utilizzato il pacchetto software MlWin. Ogni modello è stato determinato *step by step* per creare quello con la migliore bontà di adattamento, che per quanto riguarda le tecniche in questione, è rappresentato dal valore più piccolo ***-2 log likelihood*** e dal maggior numero di gradi di libertà.

In generale, si parte dal modello cosiddetto banale a sola intercetta (un modello senza variabili esplicative). Questo sarà considerato la soglia di riferimento per il valore di log-verosimiglianza (*log-likelihood*), e

[2]L'interfaccia del software *Tree Harvest* [94], così come gli output grafici e testuali risultanti dall'analisi, sono scritti interamente in lingua inglese per assicurarne le più ampie possibilità di impiego anche da parte di ricercatori stranieri

la base di partenza dalla quale inserire le intercette casuali, producendo nella maggior parte dei casi la soluzione di adattamento migliore.

Le variabili esplicative sono inserite nel modello in ordine alla maggior correlazione con la variabile di risposta. Se la variabile esplicativa non apporta miglioramenti al modello viene cancellata dalla procedura di stima, se invece contribuisce apportando informazione significativa viene stimato anche il suo coefficiente angolare (*random slope*). Quando il valore di log-verosimiglianza migliora, la procedura di inserimento di nuovi predittori continua se, al contrario, il valore decresce in maniera significativa il modello viene costruito con le variabili esplicative, ma senza trattenere in esso i coefficienti angolari randomizzati. Questa procedura viene eseguita iterativamente fino all'introduzione di tutte le variabili esplicative.

Il modello che "interpola" la maggior parte della varianza è usato per calcolare gli *score* predetti. Grazie a tali valori teorici viene calcolata successivamente la misura di bontà dell'adattamento *Goodness of Fit overall e moderating* al fine di comparare i risultati del metodo parametrico con quelli dell'approccio non parametrico RTME [46].

Alcuni problemi emersi durante l'analisi col software MlWin, sono stati, ad esempio, fattori quali il tempo, particolarmente lungo, per applicare i *multilevel model* su dataset di grandi dimensioni, a causa soprattutto della lunghezza del processo iterativo. Inoltre, difficoltà si sono avute anche a causa di modelli contenenti un numero consistente di variabili, o ancora, a causa di variabili con molte modalità, provocando in tal modo una difficile convergenza verso una soluzione, ottimale[3].

[3]In appendice sono mostrati i risultati delle elaborazioni effettuate col software MlWin.

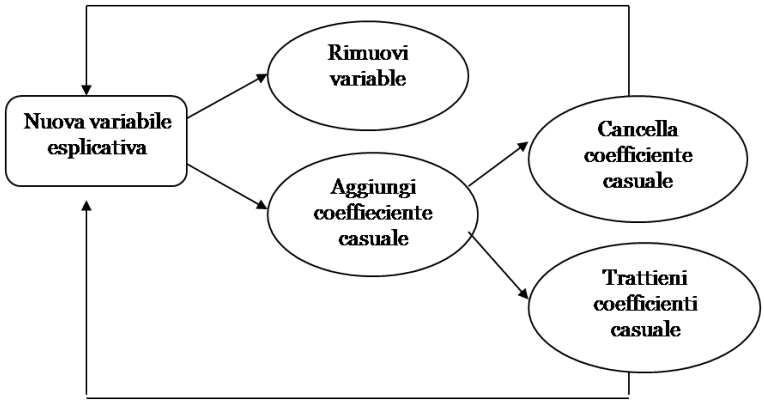

Figura 4.1: Metodo di costruzione del modello multilevel

4.3 Alcuni studi empirici

Nel prosieguo verranno presentate alcune applicazioni dei metodi trattati nel precedente capitolo attraverso l'impiego del software *Tree Harvest*. Gli obiettivi che si vogliono perseguire mediante questi studi empirici sono principalmente due: il primo è quello di mostrare, con diversi esempi applicativi, come le metodologie *RTME con criterio di split additivo e moltiplicativo* rispondono in maniera soddisfacente alle problematiche che sono chiamate ad affrontare; il secondo obiettivo è quello di illustrare le funzionalità e le potenzialità dell'approccio non parametrico. Inoltre, saranno presentati i risultati in tabelle riassuntive che permettono di visualizzare facilmente gli output prodotti e le differenze maggiormente significative tra i vari metodi impiegati.

4.3.1 Dataset reali utilizzati

In questo paragrafo si vuole fornire un esempio concreto dell'applicazione su dataset noti in letteratura, degli algoritmi implementati (criterio additivo e moltiplicativo RTME) attraverso l'impiego del software *Tree Harvest* [94], evidenziando come tali procedure consentono di interpretare l'effetto moderante, espressione della gerarchia presente nei dati, rispetto alla metodologia tradizionale CART e all'approccio parametrico dell'analisi multilivello.

Nella tabella seguente 4.1 sono riportati in maniera riassuntiva le principali caratteristiche dei dataset analizzati:

Nome Dataset	**Sugar Cane**	**Pulse Rate**	**ILE Authority**
Descrizione	*Questi dati riguardano la resa dello zucchero di canna per ogni piantagione nel Nord Queensland, nei periodi di raccolta del 1997*	*Battiti cardiaci prima e dopo l'esercizio. I dati riguardano, oltre alla frequenza, anche aspetti psicologici e stili di vita*	*I dati riguardano gli esami di studenti frequentanti 140 diverse scuole secondarie in differenti anni scolastici*
Fonte	*Denman, N., and Gregory, D. (1998)*	*R. J. Wilson, Univ. of Queensland (1998)*	*ILEA Research and Statistics (1987)*
Variabile di risposta	*Cane Quality*	*Relative Pulse Difference*	*Student's Score*
Variabile moderatrice	*Districts*	*Excercise Type*	*Student's Country*
N. di Predittori	*24*	*8*	*8*

Tabella 4.1: Descrizione sintetica dei dataset reali.

Il dataset Sugar Cane

La descrizione di questi dati riguarda la resa dello zucchero di canna per ogni piantagione nell'area Mulgrave del Queensland settentrionale, nella stagione del 1997[4] [35].

La canna da zucchero è l'industria principale della zona Mulgrave, situata intorno al Cairns nel Nord Queensland. Tutte la produzione di canna da zucchero in questo settore viene trasformata dalla *Mulgrave Central Mill*. I dati sono stati forniti dal *Bureau of Sugar Experiment Stations* (BSES) a nome della *Mulgrave Central Mill*[35].

La descrizione delle variabili:

Variabile di risposta. La variabile di risposta riguarda la produzione di canna in tonnellate per ettaro, appartenente a ciascuna piantagione. Rispetto allo studio originario, [35] non vengono considerate tali il contenuto della fibra per "pannocchia" e il contenuto di zucchero vendibile per pannocchia prodotta. C'è un *payoff* tra quantità (tonnellaggio) e qualità (contenuto di zucchero). Alcune varietà di canna da zucchero sono state sviluppate per avere un contenuto di zuccheri superiore rispetto a determinate condizioni del suolo, mentre altre varietà sono state sviluppate con l'intento di produrre un quantitativo maggiore.

La variabile moderatrice: *Distretto*. L'area del Mulgrave è stata suddivisa dal BSES in quindici distretti. Il BSES ha ulteriormente diviso i distretti in cinque macroaree basate sulla posizione fisica e la piovosità media.

I predittori *Tipo di terreno*: Il suolo è un fattore importante che determinerà le prestazioni del raccolto. Ci sono molti tipi di ter-

[4]Questo dataset è stato ottenuto da David Gregory e Nick Denman per il loro progetto di dati MS305 nel 1998.

reni, ciascuno avente caratteristiche particolari, come il contenuto di nutrienti, acidità e di drenaggio. In questa variabile viene fornito il nome del tipo di suolo di ogni piantagione.

Area: Piantagioni di vasta dimensione avranno più file, ciò significa che può essere prodotta una maggiore quantità per ettaro.

Varietà: Alcune specie di canna sono state modificate per essere in grado di sopravvivere in terreni asciutti, senza tanta pioggia, mentre altre sono state "progettate" per sfruttare i nutrienti di cui sono pieni i terreni vulcanici.

Età: La canna da zucchero è una pianta che, se tagliata, cresce nuovamente. L'agricoltore può scegliere di "arare" una piantagione di canna, una volta prelevato il raccolto. Ciò richiede all'agricoltore una nuova semina. La canna piantata l'anno precedente può essere considerata come se avesse età zero, la canna lasciata crescere fino all'anno successivo al taglio, può essere considerata di età pari ad un anno. Ci si può attendere che il contenuto di zucchero e il tonnellaggio di questa canna sia inferiore a quello delle piante più giovani. La variabile età rappresenta, dunque, il numero di anni di ricrescita prima della raccolta della canna da zucchero.

Mese di raccolta: La stagione di coltivazione della canna da zucchero inizia di solito nel mese di giugno e si conclude a metà novembre, tuttavia la data di conclusione può variare in base alle precipitazioni di stagione. E' possibile che possa esserci qualche interazione tra i mesi di raccolta e la varietà della canna da zucchero, così come è probabile che alcune varietà diano un maggiore contenuto di zucchero ad un certo livello di maturità, prima di altre.

Precipitazioni: Sono indicati i mesi delle precipitazioni totali, per

ciascun distretto a partire dal luglio 1996 fino al dicembre 1997. Denman e Gregory nel loro studio del (1998)[35], hanno raggruppato le precipitazioni in tre gruppi. Nel gruppo I il taglio per la stagione 1996 (da luglio a ottobre 1996); gruppo II stagione delle piogge 1996/1997 (novembre 1996 e febbraio 1997); gruppo III precipitazioni "fuori stagione" 1997 (da marzo a giugno 1997; taglio stagione 1997: e da luglio a ottobre 1997). Da notare che per novembre e dicembre 1997 le precipitazioni non sono state registrate, poiché la maggior parte della canna è già stata tagliata.

Variabile	Descrizione
Distretto	Nome del distretto
Gruppo di Distretto	Raggruppamento di BSES in 5 macro zone geografiche
Posizione del Distretto	Suddivisione in Nord, Sud, Est, Ovest e Centrale (N, S, E, W, C)
Identificativo del terreno	Tipo di terreno: il numero ID dettagliati
Tipo di suolo	Tipo di terreno: nome generico
Area	Area di pascolo (in ettari)
Varietà	Varietà di Zucchero
Ratoon	Germoglio o età ricrescita di canna da zucchero.
Età	Numero di anni di ricrescita dalla prima raccolta
Mesi di raccolta	Mese in cui il raccolto è stato avviato
Durata della raccolta	Durata della raccolta in giorni
Tonnellate	Tonnellate per ettaro di canna da zucchero raccolta
Fibre	Contenuto di fibre per pannocchia
Zucchero	Quantità di zucchero commercializzabile per pannocchia
lug-96	Precipitazioni per distretto a luglio 1996
ago-96	Precipitazioni per distretto ad agosto 1996
: :	: :
: :	: :
dic-87	Precipitazioni per distretto del dicembre 1997

Tabella 4.2: Descrizione del dataset *Sugar Cane*.

Per quanto riguarda l'analisi e l'elaborazione dei risultati non sono state considerate le variabili *Fibre*, *Sugar* e *Posizione del Distretto*. Le prime due sono state escluse perché nel set di dati sono considerate

variabili di risposta alternative, mentre la terza è stata eliminata per la sua natura di variabile moderatrice ad un livello superiore a quella scelta nell'analisi e non ancora implementata nel software.

Sulla variabile dipendente *Quality* è stata compiuta una trasformazione moltiplicando per cento i valori osservati; la variabile moderatrice individuata *Gruppo di Distretto* è caratterizzata dalla suddivisione in cinque "classi" diseguali.

Il dataset *Pulse rate*

Questo dataset riguarda i risultati ottenuti in un semplice esperimento, compiuto su un gruppo di studenti appartenenti ad una classe, in cui hanno insegnato i fautori della ricerca, il prof. John Eccleston e il dottor Richard Wilson [123] dell'Università del Queensland.

Si è misurata sugli studenti la frequenza cardiaca di ognuno. Successivamente sono stati invitati a lanciare una moneta e allorquando la moneta dava testa, dovevano correre sul posto per un minuto. In caso contrario, dovevano rimanere seduti. Trascorso il suddetto periodo di tempo su ognuno di essi si è rimisurata la frequenza dei battiti cardiaci. I tassi di impulso e di altri dati fisiologici sono indicati nei dati in tabella 4.3.

Inoltre, le rilevazioni sono state effettuate su cinque classi diverse tra il 1993 e il 1998. Il docente, Richard Wilson, era preoccupato del fatto che alcuni studenti avrebbero scelto l'opzione meno faticosa, restare suduti piuttosto che correre, anche se dalla loro moneta risultava testa; così negli anni 1995-1998 è stato utilizzato un diverso metodo di assegnazione casuale.

Per analizzare il suddetto dataset si è deciso di considerare, quale variabile dipendente, la variazione percentuale dei battiti cardiaci a seguito dell'esito dell'esperimento su ogni alunno. In questo modo i

Variabile	Descrizione
Altezza	Altezza (cm)
Peso	Peso (kg)
Età	Età (anni)
Genere	Genere (1 = maschio, 2 = femmina)
Fumo	Regolare fumatore? (1 = sì, 2 = no)
Alcol	Bevitore regolare? (1 = sì, 2 = no)
Esercizio	Frequenza di esercizio (1 = alto, 2 = moderata, 3 = bassa)
Ran	Misurazioni tra prima e seconda frequenza (1 = se ha corso, 2 = se seduto)
Pulse1	Misurazione delle pulsazioni First (tariffa al minuto)
Pulse2	Seconda misurazione delle pulsazioni (tariffa al minuto)
Anno	Anno di classe (93 - 98)

Tabella 4.3: Descrizione del dataset *Sugar Cane*.

predittori sono diventati otto, poiché non avrebbe avuto senso includere le variabili *Pulse1* e *Pulse2*. La variabile moderatrice, ovvero quella individuata al secondo livello, è stata individuata nella *frequenza di esercizio* la quale si caratterizza per la presenza di tre gruppi.

Il dataset *Ilea*

Il dataset proviene dall'*Inner London Education Authority* (ILEA) [56] [57] [58], in esso vengono descritti, per ogni record, gli esami sostenuti dagli studenti in 140 scuole secondarie negli anni 1985, 1986 e 1987. Tali dati sono un campione casuale composto da 15362 studenti, molto noto in letteratura perché utilizzato da Goldstein [47] [48] [76] nei sui lavori sui modelli multilevel. Questi dati sono stati usati principalmente per studiare l'efficacia della didattica nelle scuole. Di seguito viene riportata la rappresentazione tabellare delle variabili:

Si è ritenuto molto importante considerare un dataset che facesse parte della libreria di esempio[5], connessa all'uso di software che

[5]Tali librerie sono progettate per fini di insegnamento e formazione. Que-

Variabile	Descrizione
Anno	1985=1; 1986=2; 1987=3
Scuola	Codici 1-139
Punteggio esami	Punteggio
% FSM	Percentuale di studenti che beneficiano di pasti gratis
% VR1 band	Percentuale di studenti nella fascia degli esami orali
Genere	Maschio=0; Femmina=1
Fascia esami orali	VR1=alto; VR2=medio; VR3=basso
Etnia dello studente	Britannici=1; Africani=2; Arabi=3; Bengalesi=4; Caraibici=5; Greci=6; Indiani=7; Pakistani=8; Sud Est Asiatici=9; Turchi=10; Altri=11
Tipo di scuola (gen.)	Mista=1; Maschile=2; Femminile=3
Tipo di scuola (rel.)	Conservatrice=1; Chiesa Anglicana=2; Chiesa Cattolica=3

Tabella 4.4: Descrizione del dataset *Ilea*.

implementano i modelli multilevel, poiché è sicuramente presente un sistema di dati stratificato, sintesi dell'informazione gerarchica insita in esso.

Come è possibile osservare dalla tabella 4.1, è stata scelta come variabile dipendente *Punteggio (Student's Score)* che rappresenta i punteggi numerici ottenuti agli esami da ogni studente. Tra i vari predittori è stata scelta come variabile moderatrice *VR1 band* che, come la variabile *Scuola*, è espressione di una gerarchia informativa presente nei dati. Quest'ultima non è stata inclusa nell'analisi poiché, non potendo trattare più livelli nel criterio di split, avrebbe avuto un effetto distorsivo nei risultati. La variabile al secondo livello ha tre tipi di modalità, relative ai risultati di un test di ragionamento verbale, codificate in: 1=top 25%, 2=middle 50%, 3=bottom 25%.

sta appartiene al *Centre for Multilevel Modelling* dell'Università di Bristol, che ha sviluppato il software MlWin: http://www.cmm.bristol.ac.uk/learning-training/multilevel-m-support/datasets.shtml

4.3.2 Il piano di simulazione

Nel capitolo terzo, si è introdotta la metodologia *Regression Trees with Moderating Effects* illustrando come l'adozione di un criterio di partizione basato sul coefficiente di correlazione intraclasse consente di trattare tipologie di dati che per loro natura possiedono delle relazioni gerarchiche su più livelli. In particolare si fa riferimento a situazioni di tipo *cross level interaction*, in cui l'appartenenza di un soggetto ad una classe condiziona il suo comportamento e, viceversa, il comportamento di gruppo è la risultante dell'apporto individuale di ogni singolo individuo.

In questo paragrafo si mostra un applicazione della metodologia RTME su due dataset simulati, caratterizzati da diversi livelli di correlazione intraclasse e da diverse relazioni di dipendenza tra le variabili generate.

Nella tabella 4.5 sono indicate le principali caratteristiche delle variabili generate e delle relative distribuzioni.

4.3.3 Analisi dei risultati

Di seguito si illustrano i risultati dello studio comparativo in forma tabellare ed analitica[6]. Per quanto riguarda i modelli multilevel, i risultati del *deviance test* sono disponibili in appendice.

E' opportuno precisare, inoltre, che sia per la metodologia CART classica, sia per il metodo RTME, il criterio di arresto della procedura è stato definito in base alla numerosità minima all'interno di un generico nodo terminale, pari al 10% del totale delle osservazioni.

[6]In relazione alle tecniche di regressione ad albero si è ritenuto non attinente allo scopo del presente lavoro, la rappresentazione in forma grafica, mentre per lo stesso motivo non si è riportata la specificazione funzionale relativa all'analisi *Multilevel*.

Nome Dataset	**Simulazione 1**	**Simulazione 2**
Predittori	*I predittori sono stati generati da differenti distribuzioni casuali (Uniforme discreta, uniforme continua, multinomiale, normale)* *cinque predittori*	*I predittori sono stati generati da differenti distribuzioni casuali (Uniforme discreta, uniforme continua, multinomiale, normale)* *otto predittori*
Variabile di Risposta	*Relazioni di tipo lineare con i predittori*	*Relazioni di tipo non lineare con i predittori*
Effetto moderante	*Quattro gruppi* *Due livelli di influenza*	*Cinque gruppi* *Cinque livelli di influenza*
N. di osservazioni	*1000*	*5000*

Tabella 4.5: Sintesi delle caratteristiche dei dataset simulati.

Come si evince dalla tabella 4.6, l'analisi effettuata sui dataset reali, porta all'immediata considerazione che in presenza di strutture di dati gerarchiche il metodo RTME è sicuramente più adatto a descrivere questo tipo di relazioni. Infatti, in ogni dataset analizzato i risultati in termini sia di *Overall Goodness of Fit* che di *Moderating Goodness of Fit* sono migliori in entrambi i metodi (additivo e moltiplicativo) rispetto alla metodologia CART classica.

Inoltre, quando l'effetto moderante, esprimibile alternativamente attraverso il coefficiente di correlazione intraclasse, cresce, i risultati in termini di GoF sono migliori e più performanti. Tali affermazioni sono confermate sia dall'*Overall Accuracy Ratio* che dal *Moderating Accuracy Ratio*.

Dal punto di vista dell'accuratezza e semplicità del modello, tra il criterio Moltiplicativo e quello Additivo sembra essere più performante il primo, avendo un numero di nodi terminali inferiore o al massimo

Dataset	Metodo applicato	Numero di nodi	Overall GoF	Moderating GoF	Overall AC Ratio	Moderating AC Ratio
Pulse	*Multilevel Analysis*	–	0,7380	0,7306	–	–
	CART	23	0,9157	0,9542	39,8130	41,4870
mod.effect	*RTME Additive*	21	0,9221	0,9304	43,9095	44,3048
0,0110	*RTME Multiplicative*	21	0,8943	0,9467	42,5857	45,0810
Sugar Cane	*Multilevel Analysis*	–	0,2785	0,1575	–	–
	CART	96	0,4593	0,4219	4,7844	4,3948
mod.effect	*RTME Additive*	79	0,4464	0,4302	5,6506	5,4456
0,1192	*RTME Multiplicative*	69	0,4123	0,4672	5,9754	6,7710
ILE	*Multilevel Analysis*	–	0,3395	0,0925	–	–
	CART	105	0,1378	0,1578	1,3124	1,5029
mod.effect	*RTME Additive*	109	0,2655	0,1712	2,4358	1,5706
0,2705	*RTME Multiplicative*	104	0,2702	0,1907	2,5981	1,8337

Tabella 4.6: Sintesi dei risultati sui dataset reali.

uguale, ed una misura *AC Ratio* sia *overall* che *moderating* più elevata.

Per quanto riguarda il confronto con l'analisi multilevel, i risultati in termini di *Overall GoF* sono migliori nell'analisi col metodo RTME (sia additivo che moltiplicativo) quando l'influenza della variabile al secondo livello nella spiegazione della Y è bassa (con un basso effetto moderante). Nel caso in cui l'influenza della Z cresce, quindi con un coefficiente di correlazione intraclasse più elevato (come nel dataset ILE pari a 0,2705), il modello migliore in termini di *Goodness of Fit* è quello multilevel. D'altro canto, come già ricordato in precedenza, se l'effetto moderante è basso non ha senso effettuare un'analisi multilevel; in questi casi potrebbe essere sufficiente, ad esempio, applicare un modello di regressione *Ordinary Least Squares*.

Per poter effettuare lo stesso confronto valutando, però, i soli casi estremi dal punto di vista teorico (ovvero quando l'effetto moderante si avvicina molto allo zero o, al contrario, al suo massimo) è stato compiuto uno studio su un dataset simulato, descritto nella prima colonna della tabella 4.5.

I risultati mostrati nella tabella 4.7 confermano quanto descritto sopra:

- in termini di *Overall GoF* l'analisi multilevel è sempre preferibile in presenza di un rilevante effetto moderante;
- in termini di accuratezza e parsimonia del modello è preferibile (anche se la differenza è minima) tra i due criteri RTME quello moltiplicativo.
- entrambi i criteri RTME sono sempre preferibili alla regressione al albero classica CART.

Da entrambi gli studi effettuati, si può notare che i metodi Additivo e Moltiplicativo sono da preferire quando si vuole valutare una tecnica in termini di *Moderating Goodness of Fit*, ovvero di guadagno di accuratezza in termini di WSS *Within Sum of Squares* di $Y|Z$.

Simulaz. 1	Metodo applicato	Numero di nodi	Overall GoF	Moderating GoF	Overall AC Ratio	Moderating AC Ratio
Effetto moderante 0,001	*Multilevel Analysis*	–	0,3492	0,6759	–	–
	CART	26	0,4414	0,5037	0,01698	0,01937
	RTME Additive	24	0,4690	0,5420	0,01954	0,02258
	RTME Multiplicative	20	0,4414	0,5203	0,02207	0,02602
Effetto moderante 0,9541	*Multilevel Analysis*	–	0,9553	0,5161	–	–
	CART	26	0,4414	0,1366	0,01698	0,00525
	RTME Additive	20	0,7009	0,1410	0,03505	0,00705
	RTME Multiplicative	20	0,6114	0,3525	0,03057	0,01763

Tabella 4.7: Risultati della prima simulazione.

La tabella 4.8 mostra lo studio effettuato sulla seconda simulazione, dove a cambiare sono stati il tipo di relazione tra la variabile di risposta e i predittori, ed in particolare è cambiata la numerosità sia

Simulaz. 2	Metodo applicato	Numero di nodi	Overall GoF	Moderating GoF	Overall AC Ratio	Moderating AC Ratio
Effetto moderante 0,0001	*Multilevel Analysis*	–	0,1217	0,1193	–	–
	CART	31	0,1713	0,1879	0,00553	0,00606
	RTME Additive	25	0,1755	0,1900	0,00702	0,00760
	RTME Multiplicative	22	0,1679	0,1900	0,00763	0,00864
Effetto moderante 0,369	*Multilevel Analysis*	–	0,4189	0,0747	–	–
	CART	31	0,1713	0,1459	0,00553	0,00471
	RTME Additive	19	0,3277	0,1864	0,01725	0,00981
	RTME Multiplicative	20	0,3035	0,1523	0,01518	0,00762
Effetto moderante 0,4956	*Multilevel Analysis*	–	0,5311	0,0653	–	–
	CART	31	0,1713	0,1506	0,00553	0,00486
	RTME Additive	28	0,3774	0,1868	0,01348	0,00667
	RTME Multiplicative	24	0,3624	0,1749	0,01510	0,00729
Effetto moderante 0,7024	*Multilevel Analysis*	–	0,7258	0,0706	–	–
	CART	31	0,1713	0,1439	0,00553	0,00464
	RTME Additive	21	0,4543	0,1901	0,02163	0,00905
	RTME Multiplicative	19	0,4493	0,1965	0,02365	0,01034
Effetto moderante 0,9414	*Multilevel Analysis*	–	0,9420	0,0079	–	–
	CART	31	0,1713	0,0333	0,00553	0,00107
	RTME Additive	33	0,5572	0,0381	0,01688	0,00115
	RTME Multiplicative	34	0,5187	0,1187	0,01526	0,00349

Tabella 4.8: Risultati della seconda simulazione.

dei gruppi che la dimensione del dataset.

I risultati derivanti dal secondo studio di simulazione hanno confermato, in pratica, quanto osservato nelle applicazioni precedenti. In particolare si è notato come al crescere del coefficiente di correlazione intraclasse, i modelli multilevel sono sempre più performanti rispetto ai metodi RTME e come, al contrario, questi ultimi sia preferibili in riferimento al guadagno di accuratezza espressa in termini di WSS.

La spiegazione logica che si è tentata di dare in relazione a questo tipo di risultati è che, quando si valuta un modello dal punto di vista della migliore predizione che esso è capace di esprimere (Overall GoF), risultano più idonei i modelli Multilevel, mentre se si valutano tali tec-

niche in termini di *Within Sum of Squares*, si privilegia quale obiettivo principale dell'analisi il partizionare il gruppo iniziale di individui, in sottogruppi che siano espressione della gerarchia informativa presente nei dati e che abbiano caratteristiche simili al loro interno.

4.3.4 Influenza del criterio di arresto della procedura

Riguardo ai risultati summenzionati, relativi alle metodologie RTME e CART classica, è stato utilizzato un criterio di arresto della procedura di segmentazione, che porta alla costruzione di una albero che, in ogni nodo, contiene un numero minimo di osservazioni, pari a 10% della numerosità totale del campione. In questo paragrafo viene mostrato come la regola d'arresto influenza la bontà dell'adattamento. In particolare viene considerato il primo studio di simulazione e su di esso è stato calcolato sia il l'*overall GoF* che il *Moderating GoF* per tutte le procedure di segmentazione adottate. Nelle figure 4.2 e 4.3 sull'asse delle X viene riportato il numero di osservazioni presenti in un nodo (criterio di stop), mentre sull'asse delle Y viene riportato la misura di bontà *overall* / *moderating* GoF.

Dalle figure si evince come i metodi RTME siano migliori del CART classico, soprattutto se il criterio di stop consente di ottenere una numerosità inferiore all'interno di ogni nodo. In generale il metodo moltiplicativo risulta essere, a parità di condizioni, migliore.

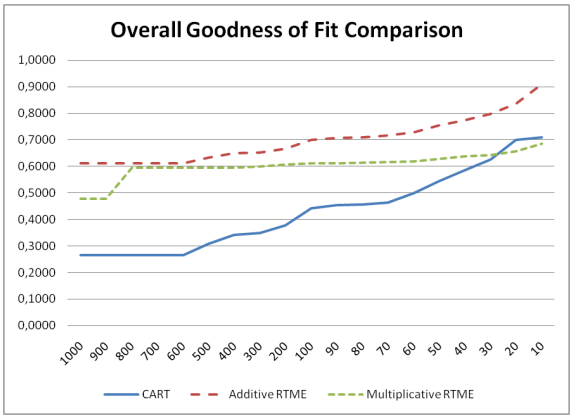

Figura 4.2: Comparazione dell'accuratezza globale tra le diverse metodologie di regressione ad albero

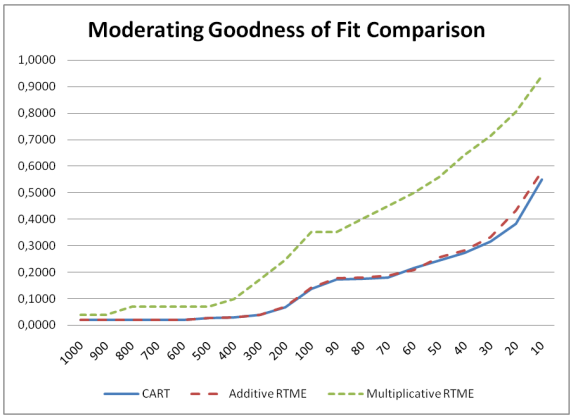

Figura 4.3: Comparazione del *Moderating GoF* tra le diverse metodologie di regressione ad albero

4.4 Uno studio di simulazione per la verifica dell'efficienza delle metodologie adottate

La seconda parte di questo capitolo riguarda lo studio, in maniera più approfondita, dei vantaggi evidenziati in ogni metodo, in particolare dell'importanza della dimensione del campione o in generale del numero di osservazioni, della dimensione dei gruppi, espressione della stratificazione ai livelli superiore al primo, e dell'effetto moderante definito dalla misura di correlazione intraclasse.

Lo studio è stato incentrato sulle assunzioni dell'analisi multilevel. E' stata considerata una variabile casuale multivariata per ottenere i dataset in cui fossero presenti stratificazioni gerarchiche. In questo modo tutte le variabili sono normalmente distribuite e, inoltre, nella simulazione si è tenuto conto, in particolare, della matrice di covarianza, di seguito mostrata, dove si evidenzia che nessuna ipotesi sarà violata. La correlazione tra le variabili esplicative è stata posta pari a zero, mentre tra queste e la variabile di risposta è stata posta pari a a 0,30. In tal modo nel corso dell'analisi non sarà violata l'ipotesi di multicollinearità. Nella simulazione in oggetto si hanno tre variabili esplicative e una variabile di risposta, la cui matrice di varianza-covarianza è la seguente:

$$\mathbf{Cov} = \begin{bmatrix} \sigma^2_{x1} & \rho\sigma_{x1}\sigma_{x2} & \rho\sigma_{x1}\sigma_{x3} & \rho\sigma_{x1}\sigma_y \\ \rho\sigma_{x2}\sigma_{x1} & \sigma^2_{x2} & \rho\sigma_{x2}\sigma_{x3} & \rho\sigma_{x2}\sigma_y \\ \rho\sigma_{x3}\sigma_{x1} & \rho\sigma_{x3}\sigma_{x2} & \sigma^2_{x3} & \rho\sigma_{x3}\sigma_y \\ \rho\sigma_y\sigma_{x1} & \rho\sigma_y\sigma_{x2} & \rho\sigma_y\sigma_{x3} & \sigma^2_y \end{bmatrix}$$

da cui

$$\mathbf{Cov} = \begin{bmatrix} \sigma_{x1}^2 & 0 & 0 & 0,30 \cdot \sigma_{x1}\sigma_y \\ 0 & \sigma_{x2}^2 & 0 & 0,30 \cdot \sigma_{x2}\sigma_y \\ 0 & 0 & \sigma_{x3}^2 & 0,30 \cdot \sigma_{x3}\sigma_y \\ 0,30 \cdot \sigma_y\sigma_{x1} & 0,30 \cdot \sigma_y\sigma_{x2} & 0,30 \cdot \sigma_y\sigma_{x3} & \sigma_y^2 \end{bmatrix}$$

Per ogni gruppo è stato creato un differente dataset. Il valore atteso della variabile di risposta varia per ogni gruppo, in modo tale da farlo corrispondere con il coefficiente di correlazione intraclasse richiesto. La deviazione standard dei valori medi della variabile di risposta deriva dalla trasformazione dell'equazione dell'ICC 3.8.

$$\sigma_{residual}^2 = \sigma_y^2 \tag{4.1}$$

La varianza della variabile di risposta corrisponde con la varianza dei residui. Il coefficiente di correlazione intraclasse differisce per ognuno dei dataset. Con le seguenti equazioni la somma dei quadrati *"between"* per il campione, può essere calcolata come:

$$\sigma_{between}^2 = \frac{-\sigma_y^2 \cdot \rho_{ICC}}{\rho_{ICC} - 1} \tag{4.2}$$

$$SS_{between} = \sigma_{between}^2 + \frac{\sigma_y^2}{\bar{n}} \tag{4.3}$$

dove n rappresenta la dimensione del gruppo.

In questo modo sono stati analizzati 44 dataset. Il coefficiente di

correlazione intraclasse ha un range che va da 0.01 a 0.4, poiché questi valori possono essere facilmente riscontrati in dataset reali. La dimensione del gruppo J varia tra 10 e 100, il numero totale delle osservazioni (N) tra 250 e 1000. Poiché i dataset sono stati costruiti grazie alla generazione di variabili casuali, su di essi è stato successivamente verificato l'ICC.

La stessa analisi mostrata nei paragrafi precedenti, relativamente allo studio comparativo, è stata inoltre utilizzata per la determinazione di tali dataset. Nel prosieguo si terrà conto solamente del valore dell'*overall GoF*.

4.4.1 Influenza dell'ICC, del numero di osservazioni e della dimensione dei gruppi

I dati mostrati nelle tabelle a tre vie di seguito riportate, la 4.9, la 4.10 e la 4.11, seppur di non facile interpretazione, esprimono in maniera globale tutti gli aspetti che caratterizzano una struttura gerarchica di dati. Tali aspetti sono misurati in termini di *overall GOF*, poiché rappresentano l'unica misura su cui poter effettuare una comparazione tra tutti i modelli in esame. In questo modo, sono riportati simultaneamente, gli effetti del coefficiente di correlazione intraclasse, del numero di osservazioni e la dimensione dei gruppi per le tre metodologie, (RTME con un criterio di impurità additivo e moltiplicativo e Multilevel Analysis (MLA)).

Prima di cominciare l'analisi dei risultati è utile sottolineare che si è tralasciato lo studio delle differenze degli *overall GoF* sui metodi RTME che, come visto nei paragrafi precedenti, si effettua attraverso l'*Accuracy Ratio*, poiché tale analisi esula dagli attuali obiettivi.

Se si osservano gli *score* riportati nelle tabelle 4.9 e 4.10 si può affermare che l'ICC influenza nella stessa misura i metodi RTME, ovvero

l'*overall GOF* decresce leggermente all'aumentare dell'ICC, mentre in generale, per l'analisi *multilevel* si nota una tendenza opposta.

RTME Additive Impurity					
		ICC			
N	*J*	0,01	0,1	0,2	0,4
250	10	0,862	0,837	0,849	0,827
	25	0,798	0,801	0,769	0,802
	50	0,814	0,813	0,808	0,742
	100	-	-	-	-
500	10	0,865	0,859	0,842	0,840
	25	0,838	0,842	0,850	0,815
	50	0,823	0,823	0,818	0,797
	100	0,771	0,755	0,760	0,778
1000	10	0,862	0,861	0,860	0,861
	25	0,852	0,867	0,857	0,844
	50	0,848	0,830	0,833	0,840
	100	0,813	0,808	0,806	0,810

Tabella 4.9: RTME Additive Impurity

Sia i metodi RTME che MLA sono influenzati dalla dimensione dei gruppi di cui è composta la variabile moderatrice. In particolare per quanto riguarda il metodo RTME additivo, si nota che l'*overall GoF* decresce all'aumentare della grandezza del gruppo, mentre tale effetto non si verifica per il metodo RTME moltiplicativo dove, seppur di poco, la bontà di adattamento risulta nella maggior parte dei casi migliore. Un comportamento simile all'RTME additivo si verifica per il metodo MLA.

Il confronto rispetto al numero delle osservazioni mostra che per entrambe i metodi RTME è vero che: più grande è il dataset, maggiore è l'*overall GoF*. In ogni caso tale effetto è meno apprezzabile nel metodo RTME moltiplicativo. Rispetto al metodo MLA si nota che l'*overall GoF* migliora all'aumentare della dimensione del campione, soprattutto se ciò accade in concomitanza di un ICC elevato.

RTME Multiplicative Impurity					
		ICC			
N	*J*	0,01	0,1	0,2	0,4
250	10	0,676	0,663	0,65	0,611
	25	0,695	0,643	0,654	0,667
	50	0,716	0,679	0,595	0,626
	100	-	-	-	-
500	10	0,652	0,624	0,613	0,609
	25	0,673	0,665	0,66	0,626
	50	0,706	0,66	0,665	0,622
	100	0,695	0,653	0,635	0,652
1000	10	0,631	0,644	0,624	0,591
	25	0,663	0,662	0,637	0,601
	50	0,678	0,667	0,636	0,638
	100	0,677	0,671	0,665	0,626

Tabella 4.10: RTME Multiplicative Impurity

Multilevel Analysis					
		ICC			
N	*J*	0,01	0,1	0,2	0,4
250	10	0,328	0,297	0,502	0,533
	25	0,243	0,282	0,413	0,571
	50	0,412	0,456	0,374	0,543
	100	-	-	-	-
500	10	0,284	0,388	0,464	0,608
	25	0,319	0,359	0,403	0,611
	50	0,263	0,372	0,426	0,536
	100	0,278	0,354	0,401	0,512
1000	10	0,342	0,378	0,489	0,609
	25	0,295	0,363	0,453	0,587
	50	0,289	0,344	0,428	0,565
	100	0,271	0,355	0,447	0,601

Tabella 4.11: Multilevel Analysis

Conclusioni

Molte tipologie di dati, soprattutto nel campo delle scienze economiche, sociali e biologiche, hanno una struttura gerarchica o sono caratterizzati da *cluster*. Oggi i modelli multilevel rappresentano una soluzione tra le più avanzate per la risoluzione delle problematiche affrontate nel presente lavoro.

Fino ad oggi si è assunto che la ricerca verificasse teorie interessanti sui sistemi a struttura gerarchica, sulla base di dati disponibili, oppure, partendo da conoscenze a priori dell'esistenza di dati multilivello e dalla volontà di compiere un'analisi esplorativa su questi ultimi. Ulteriore modo di agire è stato anche quello di raccogliere informazioni attraverso la costruzione di un disegno campionario che tenesse conto delle possibili interazioni tra soggetti che risultano influenzati dal contesto sociale cui appartengono.

L'approccio appena descritto si fonda su una modellizzazione di tipo parametrica classica, ragion per cui è necessario il rispetto di una serie di vincoli ed assunzioni in modo tale da garantire una stima accurata dei parametri. Tuttavia tali assunzioni teoriche non sempre sono verificabili, a causa di fattori esterni di disturbo che sfuggono alla percezione del ricercatore. In tali circostanze potrebbe risultare opportuno utilizzare tecniche capaci di superare tali limiti e in grado di adattare il modello ai dati, seguendo l'impostazione tipica del *Data Mining*.

Già da diversi anni i metodi di classificazione e regressione ad albero, fondati su un approccio non parametrico, si sono rivelati un utile strumento per il *data mining* e l'apprendimento supervisionato dai dati, in particolare in presenza di strutture complesse di essi e in assenza di ipotesi distribuzionali sulle variabili.

In particolare, nelle problematiche analizzate, si è posto maggiore attenzione alla regressione ad albero, poiché si è cercato di adattare tale tecnica di segmentazione binaria allo studio di fenomeni tipici dei modelli gerarchici lineari.

Lo studio effettuato può avere una duplice valenza: *esplorativa*, se si intende indagare sulla gerarchia di importanza delle relazioni di dipendenza tra i predittori e la variabile di risposta, e *decisionale*, se si intende costruire un modello predittivo per nuovi casi. Ciò avviene attraverso una procedura di partizione ricorsiva in gruppi internamente omogenei che consente di costruire un albero esplorativo, dalla cui semplificazione, mediante una procedura induttiva discende l'albero delle decisioni.

Il presente lavoro ha inteso proporre, in primo luogo, uno strumento utile alla costruzione di alberi esplorativi capaci di far emergere la struttura gerarchica latente, seguendo la filosofia metodologica della segmentazione binaria.

L'applicazione diretta del CART [15] non consente, però, di estrarre tale informazione, espressione della stratificazione presente nei dati. Una possibile soluzione a questo limite è stata individuata nel criterio di partizionamento, proponendo a tal riguardo due criteri di *split* alternativi. Grazie all'evidenza empirica su dataset reali e simulati, si è notato come entrambe le misure adottate superino gli inconvenienti derivanti dall'applicazione del CART, fornendo risultati più accurati. Questo nuovo tipo di regressione ad albero è stata denominata *Regression Trees with Moderating Effects* RTME [44], con la specificazione *Additive* o *Multiplicative*, a seconda del criterio di impurità adottato.

In secondo luogo si è voluto mostrare i risultati degli approfondi-

menti circa la valenza applicativa della nuova metodologia proposta, evidenziando perciò gli aspetti ritenuti maggiormente critici [46].

Poiché questo lavoro rappresenta il punto di partenza relativo all'utilizzo di metodi supervisionati non parametrici, come la regressione ad albero, si sono considerati quali *benchmarking* i modelli gerarchici lineari come i Multilevel.

Lo studio comparativo ha seguito due principali linee guida: il confronto tra le tecniche di regressione ad albero e il confronto tra queste e i modelli Multilevel. Tra i due metodi RTME, quello moltiplicativo è risultato preferibile, mentre la comparazione in termini di *goodness of fit* verso i Multilevel ha mostrato ancora ampi margini di miglioramento, legati soprattutto allo scopo confermativo.

La ricerca relativa all'aspetto decisionale degli alberi, nell'individuare soluzioni che siano stabili ed affidabili per la definizione del criterio di ottimalità da soddisfare, necessita di ulteriori approfondimenti. D'altronde, questo lavoro può rappresentare il primo passo di un'attività di ricerca da estendersi agli aspetti decisionali delle metodologie ad albero, soffermandosi in particolare, sull'accuratezza dei risultati attraverso tecniche *ensemble* in grado di migliorare la bontà di adattamento, soprattutto quando l'effetto moderante cresce sensibilmente e sulla definizione di criteri alternativi di semplificazione della struttura.

Questo elaborato può essere, quindi, considerato come la conclusione del percorso di studio e ricerca condotto durante il corso di dottorato in tema di apprendimento dai dati mediante le strutture ad albero, in particolare quando in essi è presente una gerarchia informativa.

Tra gli obiettivi ed i futuri sviluppi degli argomenti trattati resta da approfondire il miglioramento del criterio di split, la generalizzazione dei risultati ottenuti da due a più livelli gerarchici, ed infine, il miglioramento dell'interpretabilità della struttura ad albero quando la sua "taglia" è considerevole.

Appendice A

Il codice sorgente in linguaggio MatLab

Di seguito si riporta il codice sorgente di alcuni degli algoritmi di calcolo scritti per integrare il software *Tree Harvest*. La scelta di non riportare per intero il codice Matlab, ma di limitarsi solo ad alcune interessanti routine, è dettata dalla volontà di non appesantire questo lavoro di tesi con una appendice corposa ma di scarsa utilità pratica.

A.1 La generazione dell'albero RTME

```
function [tree sintchildren Imptree imp sintesi2]=intracart(X,Y,Z,num,decrmin,E)
%[tree matrix fitm sintfather sintchildren Imptree imp sintesi2]=intracart(X,Y,Z,num,decrmin)
%Z=instrumental variable
%E=1 Indirect Moderating effects
%trees Matrix
%Regression Tree Using RTME method
%tree = struct array sintesi dell'albero
%sintfather = sinesi relativi al nodo padre
%sintchildren = sintesi nodi figli
%Imptree = impurità dell'albero
%sintesi2 = [numero nodo terminale, numerosità, classe di appartenenza, impurità al nodo]

cicciput=X;
N=size(Y,1);
id=(1:1:N)';
it=0;it2=0;
tree.num=num;
tree.decrmin=decrmin;
tree.nodo(1).X=X;
tree.nodo(1).Y=Y;
tree.nodo(1).Z=Z;
tree.nodo(1).term=0;
tree.nodo(1).father=0;
tree.nodo(1).n=size(X,1);
%%%%%%%%%%%%%%%%%poi si vede
tree.nodo(1).overaclass=mean(Y);
tree.nodo(1).error=intraimpurity(Y,Z,E);
tree.nodo(1).impur=intraimpurity(Y,Z,E);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%tree.nodo(1).rho=intracorr(Y,Z);
L=1;
cont=0;
imp.decimpurita=0;
imp.nodo=0;
memnodo=[0,1];
lung=length(memnodo);
nodo1(1:size(Y,1),1)=1;
matrice=[nodo1(:,1) Y X Z]; %ho aggiunto la Z

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
while memnodo(lung) ~= 0
    it=it+1;
    while size(X,1)>=num
        %[XL XR YL YR indpred valsplit Impadre decr synt misclass ImpL ImpR]=split(X,Y,N);
        [XL XR YL YR ZL ZR indpred valsplit vsplit decr syntL syntR errornode
        ImpL ImpR measL measR]=splitintra(Y,X,Z,N,E);
```

```
if decr <= decrmin
    tree.nodo(L).decrimp=0;
    tree.nodo(L).term=1;
    break
end
cont=cont+1;

sintfather.numnode(cont)=L; %node nmber
sintfather.varsplit(cont)={['X' num2str(indpred) '<=' num2str(vsplit)]};
%variable which generates the split and rule of the split
sintfather.sizenode(cont)=length(X(:,1)); %size at node number

%%%%%%%%%%%
indiceL=find((matrice(:,indpred+2)<=vsplit)& (matrice(:,1)==L));
matrice(indiceL,1)=L*2;
indiceR=find(matrice(:,1)==L);
matrice(indiceR,1)=L*2+1;
%%%%%%%%%%%%%%%%%%%%%%%%

L=L*2
R=L+1;
tree.nodo(L).X=XL;
tree.nodo(R).X=XR;
tree.nodo(L).Y=YL;
tree.nodo(R).Y=YR;
tree.nodo(L).Z=ZL;
tree.nodo(R).Z=ZR;
tree.nodo(L/2).term=0;
tree.nodo(L).father=L/2;
tree.nodo(R).father=L/2;
tree.nodo(L/2).split={[num2str(vsplit) ' at X' num2str(indpred)]};
tree.nodo(L/2).col=indpred;
tree.nodo(L/2).valsplit=valsplit;

tree.nodo(L).impur=ImpL;  %impurità condizionata
tree.nodo(R).impur=ImpR;  %impurità condizionata

tree.nodo(L).class=syntL(2:end-1);%(1,2:end-1);  %medie generali
tree.nodo(R).class=syntR(2:end-1);%(2,2:end-1);  %medie generali

tree.nodo(L).n=measL(3,:);  %numerosità generale al nodo più numerorità parziali
tree.nodo(R).n=measR(3,:);  %numerosità generale al nodo più numerosità parziali

tree.nodo(L).intraclass=measL;  %misure intraclassi (medie-prima rige,
varianze-seconda riga, numerosità-terza riga)
tree.nodo(R).intraclass=measR;  %misure intraclassi (medie-prima rige,
varianze-seconda riga, numerosità-terza riga)
```

```
        tree.nodo(L).number=L;
        tree.nodo(R).number=R;

        tree.nodo(L).error=errornode.L;%*((size(XL,1)/N));
        tree.nodo(R).error=errornode.R;%*((size(XR,1)/N));

        sintchildren.R(cont,:)=[R syntR(1)]; %numero nodo, %# al nodo, media generale,
        medie parziali, impurità al nodo
        sintchildren.L(cont,:)=[L syntL(1)];

        memnodo=[memnodo,R,L];
        X=tree.nodo(L).X;
        Y=tree.nodo(L).Y;
        Z=tree.nodo(L).Z;
        lung=length(memnodo);
        memnodo(lung)=[];



        it2=it2+1;
        imp.decimpurita(it2)=decr;
        imp.nodo(it2)=(L/2);
    end
    if size(X(:,1)) < num
        tree.nodo(L).term=1;
    end
    lung=length(memnodo);
    L=memnodo(lung);

    if size(X(:,1)) < num | decr <decrmin
        memnodo(lung)=[];
        lung=length(memnodo);
    end


    if L > 1
       X=tree.nodo(L).X;
       Y=tree.nodo(L).Y;
       Z=tree.nodo(L).Z;
    end

end
if tree.nodo(1).term==1
    noditot=0;
    noditot2=0;
    sintchldren=0;
    sintfather=0;
```

```
    sintesi=0;
    sintesi2=0;
    decr=0;
    matrix=0;
else
% calcolo delle misure di bonta dell'albero
    noditot=[sintchildren.R(:,1:2) ; sintchildren.L(:,1:2)];  %numero nodo e
    (a destra e sinistra)
    noditot2=[sintchildren.R; sintchildren.L]; %numero nodo,%# al nodo,
    media generale, medie parziali, impurità al nodo
    noterm=sintfather.numnode';
    noterm(1)=[];
    noterm;
    n=length(noterm);
    m=length(noditot(:,1));
    cont=0;

%% definizione della lista dei nodi terminali
%% per il calcolo della funzione goodness
    for j=1:m
        term=1;
        for i=1:n
            if noterm(i) == noditot(j,1)
                term=0;
                i=n;
            end
        end
        if term==1
            cont=cont+1;
            sintesi(cont,1:2)=noditot(j,1:2); % elenco nodi terminali e
            numerosità degli stessi
            sintesi2(cont,:)=noditot2(j,:); %numero nodo terminale,frequenza,
            assegnazione e impurità
        end
    end
% %    matrix=matrice;
end
%%%%verifica%%%%
%[aa bb]=size(tree.nodo);
elenconodi=sort([sintfather.numnode';sintesi2(:,1)]);
bb=length(elenconodi);
for f=1:bb;
%     if decr <= decrmin
%         treeimp(p,1)=Impadre;
    p=elenconodi(f);
    if tree.nodo(p).term==1
       treeimp(p,1)=tree.nodo(p).impur;
       treemisc(p,1)=tree.nodo(p).error;
```

```
       %treenormimp(p,1)=tree.nodo(p).impur/tree.nodo(1).impur;
    else
       treeimp(p,1)=0;
       treemisc(p,1)=0;
       %bb
    end
end
Imptree.impurity=sum(treeimp);
Imptree.error=sum(treemisc);
Imptree.normerror=sum(treeimp)/tree.nodo(1).impur;
Imptree.gain=(tree.nodo(1).impur-Imptree.error)/tree.nodo(1).impur;

%
```

A.2 Calcolo della misura di impurità con effetto additivo

```
function R = intracartimpurity(node,Z,Y);
%%%%%node viene fuori dalla funzione [yfit node]=treeval(X,T)

%K=[yfit node Z];
tabnode=tabulate(node);
for j=1:size(tabnode,1)
    impuri(j)=intraimpurity(Y(node==tabnode(j,1)),Z(node==tabnode(j,1)));
    impur(j)=impuri(j)*tabnode(j,2)/size(Y,1);
end

R.impurity=sum(impur);
varimpadre=intraimpurity(Y,Z);
R.normimpurity=R.impurity/varimpadre;
R.gain=(varimpadre-R.impurity)/varimpadre;
```

A.3 Calcolo della misura di impurità con effetto moltiplicativo

```
function rho=intracorr(Y,Z);
%%inraclass correlation coefficient
%%Y = response variable
%%Z = instrumental variable


r=size(Y,1);
J=tabulate(Z);
J=J(:,1);
N=size(J,1);
GM=mean(Y);   %%mean of Y variable

for k=1:N
    index=find(Z==J(k));  %index of the Jth category of Z variable
    y=Y;
    y=y(index);
    y=y-GM;  %centred y variable
    if length(y)>1
        comb=combntns(1:size(y,1),2);  %index of all pairwise combinations
        mult=[y(comb(:,1)) y(comb(:,2))];    %all pairwise combinations
        (of the elements of the y variable)
        Num(k)=2*sum(mult(:,1).*mult(:,2));  %Jth sum of all pairwise
%     if size(y,1)>1   %what appen if we have just only one individual
in a group? CHECK THIS SITUATION
        ngroup(k)=size(y,1)-1;
%     else
%         ngroup(k)=1;
%     end
        Dengroup(k)=sum(y.^2)*ngroup(k);  %%kth denominator
    %Dengroup(k)=sum((Y(index,1)-GM).^2)*ngroup(k);
    %couple(k)=size(mult,1)*2;
    else
        Dengroup(k)=0;
        Num(k)=1;
    end
end
Num=sum(Num); %%Numerator of the intraclass correlation coefficient
Den=sum(Dengroup);
%Den=var(Y,1)*sum(couple) %%Denominator of the intraclass correlation coefficient
rho=Num/Den;  %intraclass correlation coefficient
```

A.4 Calcolo del coefficiente di correlazione intraclasse

```
function out = ICC(cse,sng,dat)
%function to work out ICCs according to shrout & fleiss' schema (Shrout PE,
%Fleiss JL. Intraclass correlations: uses in assessing rater reliability.
%Psychol Bull. 1979;86:420-428).'dat' is data whose columns represent
%different ratings/raters & whose rows represent different cases or targets
%being measured. Each target is assumed too be a random sample from a
%population of targets. 'cse' is either 1,2,3 & 'typ' is either string
%'single' or 'k'. 'typ' denotes whether the ICC is based on a single
%measurement or on an average of k measurements, where k = the number of
%ratings/raters. 'cse' is: 1 if each target is measured by a different set
%of raters from a population of raters, 2 if each target is measured by the
%same raters, but that these raters are sampled from a population of
%raters, 3 if each target is measured by the same raters and these raters
%are the only raters of interest. This has been tested using the example
%data in the paper by shrout & fleiss
%Kevin Brownhill, Imaging Sciences, KCL, London kevin.brownhill@kcl.ac.uk
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
%number of raters/ratings
k = size(dat,2);
%number of targets
n = size(dat,1);
%mean per target
mpt = mean(dat,2);
%mean per rater/rating
mpr = mean(dat);
%get total mean
tm = mean(mpt);
%within target sum sqrs
WSS = sum(sum(bsxfun(@minus,dat,mpt).^2));
%within target mean sqrs
WMS = WSS / (n * (k - 1));
%between rater sum sqrs
RSS = sum((mpr - tm).^2) * n;
%between rater mean sqrs
RMS = RSS / (k - 1);
% %get total sum sqrs
% TSS = sum(sum((dat - tm).^2));
%between target sum sqrs
BSS = sum((mpt - tm).^2) * k;
%between targets mean squares
BMS = BSS / (n - 1);
%residual sum of squares
ESS = WSS - RSS;
```

```
%residual mean sqrs
EMS = ESS / ((k - 1) * (n - 1));
switch cse
    case 1
        switch sng
            case 'single'
                out = (BMS - WMS) / (BMS + (k - 1) * WMS);
            case 'k'
                out = (BMS - WMS) / BMS;
            otherwise
               error('Wrong value for input sng')
        end
    case 2
        switch sng
            case 'single'
                out = (BMS - EMS) / (BMS + (k - 1) * EMS + k * (RMS - EMS) / n);
            case 'k'
                out = (BMS - EMS) / (BMS + (RMS - EMS) / n);
            otherwise
               error('Wrong value for input sng')
        end
    case 3
        switch sng
            case 'single'
                out = (BMS - EMS) / (BMS + (k - 1) * EMS);
            case 'k'
                out = (BMS - EMS) / BMS;
            otherwise
               error('Wrong value for input sng')
        end
    otherwise
        error('Wrong value for input cse')
end
```

A.5 Algoritmo per il confronto delle metodologie CART e RTME

```
function [R2int, R2glob,n]=CARTvsRTME(X,Y,Z)

j=0;

for i=90:-10:10
    i
    j=j+1;
    [tree sintchildren Imptree imp sintesi2]=intracart(X,Y,Z,i,eps,0);
    T=treefit(X,Y,'splitmin',i);
    [yfit node]=treeval(T,X);
    R2cart(j,1)=1-mean((Y-yfit).^2)/var(Y,1);
    [ris nodi ris2 nodi2 wss R2int R2glob Tcart]=confrontoCART(Y,node,Z,sintesi2,tree);

    R2cart(j,2)=R2int(1);
    R2g(j,1)=R2glob(2);
    R2i(j,1)=R2int(2);
    n(j,1)=size(nodi,1);
    n(j,2)=size(nodi2,1);

    [tree sintchildren Imptree imp sintesi2]=intracart(X,Y,Z,i,eps,1);
    [ris nodi ris2 nodi2 wss R2int R2glob Tcart]=confrontoCART(Y,node,Z,sintesi2,tree);
    R2g(j,2)=R2glob(2);
    R2i(j,2)=R2int(2);

    n(j,3)=size(nodi2,1);

    R2cart
    R2g
    R2i
end

R2int=[R2cart(:,2) R2i];
R2glob=[R2cart(:,1) R2g];
```

Appendice B

Risultati del *deviance test* dell'analisi Multilevel

Di seguito si riportano i risultati del *deviance test* relativi ai dataset reali utilizzati, ottenuti col software MLwiN, per la scelta del modello multilevel migliore (best fit).

ILEA Authority

Variables (RANDOM SLOPES)	2LL
empty model	121613,4
EMPTY MODEL	116813,6
EMPTY MODEL + fsm	116627,4
EMPTY MODEL + fsm + vr1	116498,2
EMPTY MODEL + fsm + vr1 + year	116393,8
EMPTY MODEL + fsm + vr1 + YEAR	116389,8
EMPTY MODEL + fsm + vr1 + YEAR + gender	115999,2
EMPTY MODEL + fsm + vr1 + YEAR + gender + etnic	115379,0
EMPTY MODEL + fsm + vr1 + YEAR + gender + etnic + schgender	115328,8
EMPTY MODEL + fsm + vr1 + YEAR + gender + etnic + schgender + religion	115252,3

26 parameters

Tabella B.1: ILEA Authority *deviance test*

Sugar Cane

variables (RANDOM SLOPE)	2LL
empty model	27949,51
EMPTY MODEL	27511,95
EMPTY MODEL + age	27258,63
EMPTY MODEL + age + area	27254,37
EMPTY MODEL + age + area + harvestmonth	26821,50
EMPTY MODEL + age + area + harvestmonth + harvestduration	26809,98
EMPTY MODEL + age + area + harvestmonth + harvestduration + tonn/hect	26801,19
EMPTY MODEL + age + area + harvestmonth + harvestduration + TONN/HECT	26791,77
EMPTY MODEL + age + AREA + harvestmonth + harvestduration + TONN/HECT	26771,38
EMPTY MODEL + age + AREA + harvestmonth + harvestduration + TONN/HECT + + jul96	26765,32
EMPTY MODEL + age + AREA + harvestmonth + harvestduration + TONN/HECT + + jul96 + aug96	26765,24
EMPTY MODEL + age + AREA + harvestmonth + harvestduration + TONN/HECT + + jul96 + sep96	26763,89
EMPTY MODEL + age + AREA + harvestmonth + harvestduration + TONN/HECT + + jul96 + sep96 + okt96	26762,91
EMPTY MODEL + age + AREA + harvestmonth + harvestduration + TONN/HECT + + jul96 + sep96 + okt96 + nov96	26747,66
EMPTY MODEL + age + AREA + harvestmonth + harvestduration + TONN/HECT + + jul96 + sep96 + nov96	26747,68
EMPTY MODEL + age + AREA + harvestmonth + harvestduration + TONN/HECT + + jul96 + sep96 + nov96 + mrt97	26743,98
EMPTY MODEL + age + AREA + harvestmonth + harvestduration + TONN/HECT + + jul96 + sep96 + nov96 + mrt97 + sep 97	26740,74
EMPTY MODEL + age + AREA + harvestmonth + harvestduration + TONN/HECT + + jul96 + sep96 + nov96 + mrt97 + sep 97	26740,70
parameters: 28	

Tabella B.2: Sugar Cane *deviance test*

Pulse Rate

Variables (RANDOM SLOPES)	2LL
empty model	1119,579
EMPTY MODEL	1119,579
ACCORDING TO MLWiN NO INTRACLASS CORRELATION	
EMPTY MODEL + height	1119,138
EMPTY MODEL + HEIGHT	1119,138
EMPTY MODEL + height + weight	1118,784
EMPTY MODEL + height + WEIGHT	1118,784
EMPTY MODEL + height + weight + age	1117,913
EMPTY MODEL + height + weight + AGE	1117,913
EMPTY MODEL + height + weight + age + gender	1117,408
EMPTY MODEL + height + weight + age + GENDER	1117,408
EMPTY MODEL + height + weight + age + gender + smokes	1116,337
EMPTY MODEL + height + weight + age + gender + smokes+ alcohol + ran	1116,25
EMPTY MODEL + height + weight + age + gender + smokes+ alcohol	967,427
EMPTY MODEL + ran	972,587

2 parameters

Tabella B.3: Pulse Rate *deviance test*

Secondo il software MlwiN l'ultimo modello si adatta meglio ai dati. Sebbene la misura -2ll cresce leggermente con l'inserimento delle variabili, questo effetto non è significativo poichè vi è la contestuale diminuzione dei gradi di libertà.

Bibliografia

[1] Agresti A. (2002). *Categorical Data Ananlysis*. J. Wiley.

[2] Aitkin, M., Longford, N., (1986). Statistical modelling in school effectiveness. *Journal of the Royal Statistical Society* A, 149,1-43.

[3] Alker HR., A typology of ecological fallacies. In: *Dogan M., Rokkam S., eds. Social ecology*. Boston: The MIT Press, 1969: 69-86.

[4] Aria, M. (2004): Un software fruibile ed interattivo per l'apprendimento statistico dei dati attraverso alberi esplorativi. Contributi metodologici ed applicativi. Phd thesis, University of Naples Federico II.

[5] Aria M., D'Ambrosio A., Siciliano R. (2007) Robust Incremental Trees for Missing Data Imputation and Data Fusion. *Classification and Data Analisys 2007, Book of short papers (Macerata, September 12-14, 2007)*, EUM Macerata, 287-290.

[6] Aria, M., Siciliano, R. (2003). Learning from Trees: Two-Stage Enhancements. *In Proceedings of Classification and Data Analysis Group (CLADAG 2003)*, 22-24 Settembre, Bologna.

[7] Barcena, M.J., Tusell, F. (1999). Enlace de encuestas: una propuesta metodológica y aplicación a la Encuesta de Presupuestos de Tempo. *Qüestiio*, vol. 23, núm. 2, pp. 297–320.

[8] Barcikowski, R.S., (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics*, 6(3), 267-285.

[9] Benzecri, J.P. (1973). *L'Analyse des Données*, 2 Vols. Dunod, Paris, France.

[10] Blakaly TA., Woodward AJ., Ecological effects in multi-level studies. *J Epidemiol Community Health* 2000; 54: 367-374.

[11] Bolasco, S. (1997). *Analisi Multidimensionale dei Dati, Metodi, Strategie e Criteri di Interpretazione*. Carocci.

[12] Breiman, L. (1996). Bagging Predictors, *Machine Learning*, 26, 46-59.

[13] Breiman L. (1996). Bias, Variance and Arcing Classifiers. *Dept. Of Statistics, University of California. Technical Report.*

[14] Brieman L.(1998). Arcing classifiers. *The Annals of statistics*, 26(3).

[15] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, California.

[16] Cappelli, C., Mola, F., Siciliano, R. (2000). Selecting Regression Tree Models: A Statistical Testing Procedure, in S. Borra, R. Rocci, M. Vichi, M. Schader (Eds.): *Advances in Classification and Data Analysis*, Berlin (D), Springer-Verlag, 249-256.

[17] Cappelli, C., Mola, F., Siciliano, R. (2002). A statistical approach to growing an honest reliable tree. Computational Statistics and Data Analysis, 38, 285-299, Elsevier Science.

[18] Cherkassky V., Mulier F. (1998). *Learning from Data: concepts, theory, and methods*. John Wiley & Sons., New York, USA.

[19] CISIA - CERESIA (2001), SPAD version 5.0, Manuel de Prise en Main, CISIA-CESTA, Montreuil, France.

[20] Cochran, W.G., (1977). *Sampling techniques*, 3rd edition. Wiley & Sons, New York.

[21] Cohen, J., Cohen, P. (1977, 1983). *Applied Multiple Regression Analysis for the Behavioral Sciences*. Hillslade NJ: Lawrence Erlbaum Associates.

[22] Conversano, C., Mola, F., Siciliano, R. (2000). Generalized Additive Multi-Model for Classification and Prediction, in H.A.L. Kiers, J.P. Rasson, P.J.F. Groen, M. Shader (Eds.): *Data Analysis, Classification and Related Methods*, Springer Verlag, Berlin (D), 205-210.

[23] Conversano, C., Mola, F., Siciliano, R. (2001). Partitioning and Combined Model Integration for Data Mining, presented at the Symposium on Data Mining and Statistics (Augsburg, November 2000), *Journal of Computational Statistics*, 16, 323-339, Physica Verlag, Heidelberg (D).

[24] Conversano, C., Siciliano, R. (2008). Statistical Data Editing, in Wang J. (eds.), *Encyclopedia of Data Warehousing and Data Mining*, IDEA Group. Inc., Hershey, USA, volume 2, 2nd edition.

[25] Conversano, C., Siciliano R. (2009). Incremental Tree-Based Imputation with lexicographic ordering, *Journal of Classification*, forthcoming.

[26] Conversano, C., Siciliano, R., Mola, F., (2000). Supervised Classifier Combination through Generalized Additive Multi-Model, in F. Roli, J. Kittler (Eds.): *Proceedings of the First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science*, Physica Verlag, Heidelberg (D), 167-176.

[27] Cover, T., Thomas, J. (1991). *Elements of Information Theory*. Wiley, New York.

[28] D'Ambra L., Lauro N.C. (1982). Analisi in componenti principali in rapporto ad un sottospazio di riferimento. *Rivista di Statistica Applicata*, 15, 1-25.

[29] D'Ambrosio, A., (2008). *Tree based methods for data editing and preference rankings*. Phd thesis, University of Naples Federico II.

[30] D'Ambrosio A., Aria M., Siciliano R. (2007). Robust Tree-based Incremental Imputation Method for Data Fusion. *Advances in Intelligent Data Analysis*, Springer-Verlag, pp 174-183.

[31] D'Ambrosio, A., Aria, M., Siciliano, R. (2007). Robust Incremental Trees for Missing Data Imputation and Data Fusion, in *Proceedings of the 6th Scientific Meeting of the Classification and Data Analysis Group* (Macerata 12-14 september, 2007).

[32] D'Ambrosio A., Aria M., Siciliano R. (2009). Robust Incremental Tree-Based Methodology for Missing Data Imputation and Data Fusion. *Journal of Classification*, to appear.

[33] D'Ambrosio A., Tutore V.A. (2009). Kemeny's axiomatic approach to find consensus ranking in tourist satisfaction, *Statistica*

Applicata (Italian Journal of Applied Statistics), vol 20(1), pp. 21-32.

[34] De'ath G. (2002). Multivariate regression trees: a new technique for modeling species-environment relationships. *Echology* 83

[35] Denman, N., and Gregory, D. (1998). Analysis of Sugar Cane Yields in the Mulgrave Area, for the 1997 Sugar Cane Season. MS305 *Data Analysis Project*, Department of Mathematics, University of Queensland.

[36] Di Martino, M. *Valutazione della persistenza in trattamento antiipertensivo: un'analisi multilivello paziente-medico attraverso modelli con effetti casuali*. Phd thesis, University of Bologna.

[37] Donner, A. (1986). A Review of Inference Procedures for the Intraclass Correlation Coefficient in the One-Way Random Effects Model. *International Statistical Review*. Vol. 54, No. 1, pp. 67-82.

[38] Efron, B., Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. London: Chapman and Hall.

[39] Efron, B., Tibshirani, R.J. (1993). Statistical analysis in the computer age, *Science*, 253: 390-395.

[40] Fabbris, L., (1989). *L'indagine campionaria. Metodi, disegni e tecniche di campionamento*. Carocci.

[41] Fabbris, L. (1997). *Statistica Multivariata*. McGraw-Hill.

[42] Fears, T.R., Benichou, J., Gail, M.H. (1996). A reminder of the fallibility of the Wald statistic. *American Statistician*, 50(3), 226-227.

[43] Freund Y., Schapire R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1).

[44] Giordano, G., Aria M., (2009). *Regression Trees with Moderating Effects*. Paper submitted.

[45] Giordano, G., D'Ambrosio, A., (2008). *Multi-Class Budget Tree as weak learner for ensemble procedures* - XLIV Riunione Scientifica SIS. Arcavacata di Rende (CS).

[46] Giordano, G., Remmerswaal, R., (2009). Non-parametric regression model for a hierarchical data-structure: a comparison with the classical approaches. *Seventh Scientific Meeting of the CLAssification and Data Analysis Group of the Italian Statistical Society*. Book of short papers (Catania, September 09-11, 2009), p. 259-262.

[47] Goldstein, H. (1987). *Multilevel models in educational and social research*. London, Griffin: New York, Oxford University Press.

[48] Goldstein, H. (1991). Multilevel modelling of survey data. *The Statistician*, 40 235-244.

[49] Goodman, L.A., Kruskal, W.H. (1954). Measures of association for cross-classification. *Journal of American Statistical Association*, 48, 732-762.

[50] Goodman, L.A., Kruskal, W.H. (1979). *Measures of association for cross classifications*. Springer.

[51] Gordon, A. (1999). *Classification (Second Edition)*, Chapman and Hall/CRC Press, London.

[52] Hand, D., (1998). Data Mining,: Statistics or more?. *Am.Statist.*, 52, 112-118.

[53] Hand.D., Mannila H., Smyth P. (2001). *Principles of Data Mining*. A Bradford Book, The MIT Press, Cambridge, Massachusetts, London, England.

[54] Hastie, T.J., Tibshirani, R.J., Friedman, J.H. (2001). *The Elements of Statistical Learning*. Springer Verlag.

[55] Hox J.J.; (2002). *Multilevel Analysis, techniques and applications*, Mahwah, NJ: Lawrence Erlbaum Associates.

[56] ILEA Research and Statistics (1986). *Looking at school performance*, (RS 1058/86). London: ILEA Research and Statistics.

[57] ILEA Research and Statistics (1987). *Actual and predicted examination scores in schools*, (RS 1129/87). London: ILEA Research and Statistics.

[58] ILEA Research and Statistics (1987). *Ethnic background and examination results-1985 and 1986*, (RS 1120/87). London: ILEA Research and Statistics.

[59] Jobson, J.D.,(1992). *Applied Multivariate Data Analysis Volume I: Regression and Experimental Design*. Springer-Verlag, New York.

[60] Jobson, J.D.,(1992). *Applied Multivariate Data Analysis Volume II: Categorical and Multivariate Methods*. Springer-Verlag, New York.

[61] Kim, H., Loh, W.Y. (2001). Classification Trees with Unbiased Multiway Splits, *Journal of the American Statistical Association*, 96, 454, 589-604.

[62] Kish L., (1965). Survey sampling, J. Wiley and Sons, New York, Cap. 13,1 - 13,5

[63] Kish L., (1987). Statistical design for research, J. Wiley and Sons, New York.Cap 1, 2.1,

[64] Kreft, I., de Leeuw, J., (1998). *Introducing Multilevel Modeling*. Sage Publications.

[65] Kreft, G., de Leeuw E. D., (1987). The See-Saw Effect: a multilevel problem? A reanalysis of some findings of Hox and de Leeuw. *Quality and Quantity*, 22, 127 - 137.

[66] Lazarsfeld, P.F., Menzel, H., (1961). On the Relation between Individual and Collective Properties, in Etzioni (ed.), *Complex Organizations*. New York: Holt, Rinehart and Winston.

[67] Lauro, N.C., Siciliano, R., (1989). Exploratory methods and modelling for contigency tables analisys: an integrated approach. *Statistica Applicata*, 1.

[68] Leeuw, J. de, Meijer, E., (2008). *Handbook of Multilevel Analysis*. Springer, New York.

[69] Longford, N.T., (1993). *Random Coefficient Models*. New York: Oxford University Press.

[70] Martinez W.L., Martinez, A.R., (2002). *Computational Statistics Handbook with MatLab*. Chapman & Hall/CRC, Boca Raton, Florida.

[71] Mola, F., (1993). *Aspetti metodologici e computazionali delle tecniche di segmentazione binaria: Un contributo basato su una funzione di predizione*. Unpublished Ph.D. thesis, Dipartimento di Matematica e Statistica, Universitá degli Studi di Napoli Federico II.

[72] Mola, F., Siciliano, R., (1992). A two-stage predictive splitting algorithm in binary segmentation, in Y. Dodge, J. Whittaker. (Eds.): *Computational Statistics: COMPSTAT 92*, 1, Physica Verlag, Heidelberg (D), 179-184.

[73] Mola, F., Siciliano, R., (1994). Alternative strategies and CATANOVA testing in two-stage binary segmentation, in E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, B. Burtschy (Eds.): *New Approaches in Classification and Data Analysis: Proceedings of IFCS 93*, Springer Verlag, Heidelberg (D), 316-323.

[74] Mola, F., Siciliano, R., (1997). A Fast Splitting Procedure for Classification and Regression Trees, *Statistics and Computing*, 7, Chapman Hall, 208-216.

[75] Mola, F., Siciliano, R., (1998). A general splitting criterion for classification trees, *Metron*, 56, 3-4.

[76] Nuttall,D.L., Goldstein,H., Prosser,R and Rasbash,J. (1990). Differential School Effectiveness. *International Journal of Educational Research*, 13, 769-776.

[77] Pecoraro, M. (2008) *Web Usage e Web Structure Mining: contributi per l'integrazione e la visualizzazione*. Phd thesis, University of Naples Federico II.

[78] Pecoraro, M., Siciliano, R. (2008). Statistical Methods for User Profiling in Web Usage Mining, in *Handbook of Research on Text*

and Web Mining Technologies, edited by Min Song and Yi-Fang Brook Wu, chapter XXII, IDEA Group. Inc., Hershey, USA.

[79] Pedhazur, E.J., (1997) *Multiple Regression in behavioral research: Explanation and Prediction*. Forth Worth, TA: Harcourt.

[80] Petrakos, G., Conversano, C., Farmakis, G., Mola, F., Siciliano, R., Stavropoulos, P. (2004) New ways to specify data edits. *Journal of Royal Statistical Society*, Series A, volume 167, part 2, 249-274.

[81] Piccolo D. (2000) *Statistica*. Il Mulino.

[82] Raudenbush, S. W., Bryk, A. S. (2002). *Hierarchical linear models*. Thousand Oaks, CA: Sage Publications.

[83] Rizzi, A. (1985). *Analisi dei Dati*. La Nuova Italia Scientifica.

[84] Robinson W. S. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, Vol. 15, N° 3 pp. 351-357.

[85] Sarle, W.S. (1998). *Prediction with Missing Inputs* Technical Report, SAS Institute.

[86] Schafer, J. L., (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

[87] Schapire R. E. (1990). The strength of weak learnability. *Machine learning* 5(2).

[88] Schapire R. E. (1999). A brief introduction to Boosting. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*.

[89] Schapire R.E., Singer Y., (1999). Improved boosting algorithms using confidence-rated predictions. *Machine learning* 37(3).

[90] Searle, S.R., Casella, G., e McCulloch, C.E.,(1992). *Variance Components*. New York: Wiley.

[91] Siciliano, R., (1998). Exploratory versus Decision Trees, invited lecture to COMPSTAT '98 (Bristol, August 24-28), in R. Payne, P. Green (Eds.): *Proceedings in Computational statistics: 13th Symposium of COMPSTAT*, Physica Verlag, Heidelberg (D).

[92] Siciliano, R., (1999). Latent budget trees for multiple classification, in M. Vichi, P. Optitz (Eds.): *Classification and Data Analysis: Theory and Application*, Springer Verlag, Heidelberg (D).

[93] Siciliano, R., Aria, M., (2009). TWO-CLASS Trees for Non-Parametric Regression Analysis. In *Studies in Classification, Data Analysis, and Knowledge Organization*, a cura di Fichet B., Piccolo D., Verde R. e Vichi M., to appear.

[94] Siciliano, R., Aria, M., Conversano, C., (2004). Harvesting trees: methods, software and applications. *In Proceedings in Computational Statistics: 16th Symposium of IASC Held In Prague*, (COMPSTAT2004), Eletronical Edition (CD)Physica-Verlag, Heidelberg.

[95] Siciliano, R., Aria., D'Ambrosio, A., (2005). Boosted stump algorithm for missing data incremental imputation. *CLADAG 2005, Book of Short Papers (Parma, June 6-8, 2005)*, MUP, Parma, 161-164.

[96] Siciliano, R., Aria., D'Ambrosio, A., (2006). Boosted incremental tree-based imputation of missing data. *Data Analysis, Classification and the Forward Search. Springer series in Studies*

in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, pp. 271-278.

[97] Siciliano, R., Aria, M., D'Ambrosio, A., (2008). Posterior Prediction Modelling of Optimal Trees, *Proceedings of COMPSTAT 2008*, 323-331 Physica Verlag.

[98] Siciliano, R., Conversano C., (2002). Tree-based Classifiers for Conditional Missing Data Incremental Imputation. *Proceedings of the International Conference on Data Clean* (Jyväskylä, May 29-31, 2002), University of Jyväskylä.

[99] Siciliano, R., Conversano, C., (2008). Decision Tree Induction, in Wang J. (eds.), *Encyclopedia of Data Warehousing and Data Mining*, IDEA Group. Inc., Hershey, USA, volume 2, 2nd edition.

[100] Siciliano, R., Mola F., (2000). Multivariate data analysis and modelling through classification and regression trees. *Computational Statistics & Data Analysis*.

[101] Siciliano, R., Mola, F., (2000). Multivariate Data Analysis through Classification and Regression Trees, *Computational Statistics and Data Analysis*, 32, 285-301, Elsevier Science.

[102] Siciliano, R., Mola, F., (2002). Discriminant Analysis and Factorial Multiple Splits in Recursive Partitioning for Data Mining, in Roli, F., Kittler, J. (eds.): *Proceedings of International Conference on Multiple Classifier Systems* (Chia, June 24-26, 2002), 118-126, Lecture Notes in Computer Science, Springer, Heidelberg.

[103] Siciliano, R., Mooijaart, A., (1999). Unconditional Latent Budget Analysis: a Neural Network Approach, in S. Borra, R. Rocci,

M. Vichi, M. Schader (Eds.): *Advances in Classification and Data Analysis*, Springer-Verlag, Berlin, 127-136.

[104] Snijders T., Bosker R., (1999) *Multilevel Analysis. An Introduction to basic and advanced mutilevel modeling*, SAGE Publications, London.

[105] Steverink M.H.M., Heiser W.J., van der Kloot W.A. (2002). Avoiding degenerate solutions in multidimensional unfolding by using additional distance information. *Technical report of University of Leiden*.

[106] Stone, M., (1974). Cross-validatory choice and assessment of statistical predictions, *Journal of the Rojal Statistical Society*, Series B, Vol. 36, pp. 111-133.

[107] Tacq, J., (1986) *Van muliniveau problem naar multiveau analyse*, Department of Research Methods and Techniques, Erasmus University, Rotterdam.

[108] Takeuchi, K., Yanai, H., Mukherjee, B. (1982). *The Foundations of Multivariate Analysis*, Wiley Eastern, New Dehli.

[109] Thisted, R.A., (1988). *Elements of Statistical Computing: Numerical Computation*. London, Chapman and Hall.

[110] Tibshirani R. (1996). Bias, variance and prediction error for classification rules. *Technical report*, University of Toronto.

[111] Tutore, V. A. (2008). *3Way classification and regression trees: methods, computations and applications*. Phd thesis, University of Naples Federico II.

[112] Tutore V.A., D'Ambrosio A. (2009).Three-Way Data Analysis by Tree-Based Partitioning. *Classification and Data Analisys 2009*, Book of short papers (Catania, September 9-11, 2009), CLEUP Padova, 641-644.

[113] Tutore, V.A., Siciliano, R., Aria, M., (2006). Three Way Segmentation in *Proceedings of Knowledge Extraction and Modelling (KNEMO06)* IASC INTERFACE IFCS Workshop, Capri, September 4th-6th 2006.

[114] Tutore, V.A., Siciliano, R., Aria, M., (2007). Conditional Classification Trees using Instrumental Variables. *Advances in Intelligent Data Analysis*, Springer-Verlag, pp 163-173.

[115] Tutore, V.A., Siciliano, R., Aria, M., (2007). 3-Way Trees, in *Proceedings of the 6th Scientific Meeting of the Classification and Data Analysis Group* (Macerata 12-14 september, 2007).

[116] Urbanek, S., (2002). Different ways to see a tree - KLIMT, in *Proc. of the 14th Conference on Computational Statistics*, (Compstat 2002), p303-308, Physica, Heidelberg.

[117] Valiant, L.G., (1999). A theory of the learnable. *Communication of the ACM* 27/11.

[118] van der Ark, L.A., (1999). *Contributions to Latent Budget Analysis. A Tool for the Analysis of Compositional Data*. DSWO Press, Leiden University.

[119] van der Leeden, R., Busing, F. M. T. A., (1994). *First iteration versus final IGLS/RIGLS estimates in two-level models: A Monte Carlo study with ML3*. PRM 94-02. Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

[120] van der Leeden, R., Busing, F., Meijer, E., (1997). *Applications of bootstrap methods for two-level models*. Paper, Multilevel Conference, Amsterdam, April 1-2, 1997.

[121] Vapnik, V., (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin.

[122] Vapnik, V., (1998). *Statistical Learning Theory*. Chichester, John Wiley & Sons, United Kingdom.

[123] Wilson, R.J., *Pulse Rates before and after Exercise* The data was supplied by Dr Richard J. Wilson, Department of Mathematics, University of Queensland.

[124] Zani, S., (1998) *Analisi dei dati statistici, vol. I, Osservazioni in una e due dimensioni*, Giuffré ed., Milano.

[125] Zani, S., (2000) *Analisi dei dati statistici, vol. II, Osservazioni multidimensionali*, Giuffré ed., Milano.