UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA SPECIALISTICA IN
SCIENZE STATISTICHE, ECONOMICHE, FINANZIARIE ED AZIENDALI

Modelli a Classi Latenti Multilivello
e
Criteri Alternativi di Valutazione di Adattamento

RELATORE: Ch.Ma Prof.Ssa Francesca Bassi

LAUREANDA: Casagrande Valentina

MATRICOLA: 601705-SEA

ANNO ACCADEMICO: 2010/2011

Alla mia
famiglia.

Indice

CAPITOLO 3

CAPITOLO 0

Introduzione

Nel mondo d'oggi il mercato è sempre più competitivo in tutti gli ambiti, le aziende hanno quindi l'esigenza di attuare strategie che permettano un vantaggio competitivo. Nel mercato, infatti, la competizione è sempre più intensa e diventa quindi importante conoscere in modo approfondito il proprio mercato sul piano competitivo e la propria struttura di domanda. La segmentazione del mercato è lo strumento che rende possibile alle aziende approfondire la conoscenza della propria domanda.

La segmentazione del mercato ha come scopo quello di aumentare l'effetto delle politiche di marketing dell'azienda, combinando gli strumenti del marketing mix alle esigenze manifestate dagli specifici segmenti e alle loro peculiarità. La segmentazione suddivide il mercato in gruppi omogenei e distinti di consumatori in base a caratteristiche comuni. I gruppi che si formano sono omogenei al loro interno ed eterogenei tra loro. Riconoscere l'esistenza di elementi di eterogeneità del mercato ha reso indispensabile segmentare il mercato stesso, per poter attuare le adeguate strategie di marketing. La base del target marketing risulta essere proprio la segmentazione, che permette l'individuazione del segmento, o dei segmenti, che l'azienda può raggiungere in modo efficiente ed efficace, in base alle sue risorse e competenze specifiche, per potersi poi posizionare nel segmento che rispetta le caratteristiche richieste.

I segmenti devono avere determinate caratteristiche:

- consitenza/dimensione;
- stabilità;

- identificabilità;
- accessibilità;
- capacità di risposta;
- propositività.

Le tecniche statistiche di segmentazione si dividono in due classi, la segmentazione a priori e la segmentazione a posteriori. La segmentazione a priori consiste nel suddividere il mercato in gruppi con modalità determinate a priori, mentre in quella a posteriori le modalità per il raggruppamento vengono determinate a posteriori.

La segmentazione si divede anche in due tecniche, per omogeneità e per obiettivi. Nelle tecniche per omogeneità le unità di interesse vengono raggruppate a seconda della loro similarità rispetto a determinate variabili e i gruppi che si formano sono caratterizzati da un'elevata omogeneità interna e un'elevata variabilità esterna. Tra queste tecniche ci sono quelle classiche in cui i gruppi che si formano non sono specificati a priori, la *Cluster Analysis,* e le tecniche flessibili in cui i segmenti che si formano dipendono dalle caratteristiche percepite dai consumatori, la *Conjoint Analysis.* Le tecniche per obiettivi suddividono le unità in base ad una variabile dipendente nota a priori, in seguito vengono individuate delle variabili esplicative che influiscono in modo rilevante sulla variabile d'interesse. Tra queste tecniche individuiamo l'AID, che suddivide in due parti il mercato, il CHAID, che lo suddivide in più gruppi, e la regressione logistica, che suddivide il mercato per obiettivi.

In questo lavoro si utilizzerà la tecnica statistica dei modelli a classi latenti multilivello utilizzando le procedure classica e una innovativa, proposta da Lukočienė, Varriale e Vermut (2010), per determinare il numero adatto di classi che permettono di descrivere il mercato in analisi.

L’obiettivo principale consiste nel confrontare i due procedimenti alternativi di valutazione dell’adattamento ottenendo un modello adeguato per il mercato in analisi.

Nel primo capitolo verrà presentata la segmentazione in modo statistico con le implicazioni nel marketing, descrivendo in modo approfondito le tecniche di classificazione.

Nel secondo capitolo si presenterà la tecnica a classi latenti (CL).

Il terzo capitolo riporterà i risultati principali delle analisi in cui si è utilizzato il modello CL multilivello, valutando l’adattamento nel modo classico e con il criterio che prevede di utilizzare la procedura a tre passi.

Nelle conclusioni si confronteranno i due procedimenti di valutazione dell’adattamento e i criteri d’informazione BIC e AIC3.

CAPITOLO 1

Segmentazione

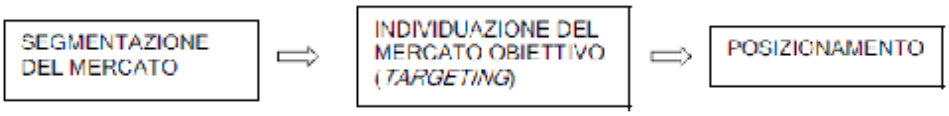
La segmentazione del mercato è un elemento essenziale nel marketing, in quanto un'azienda ha necessità di suddividere il mercato d'interesse per poter poi offrire i propri prodotti ad una specifica sezione, non potendo infatti raggiungere l'intero mercato, per problemi legati soprattutto all'eterogeneità delle preferenze e alla distanza geografica dai clienti. I segmenti vengono scelti in base alla loro posizione nel mercato e al fatto di essere o meno attraenti, cercando quindi il segmento più vantaggioso in base alle risorse a disposizione.

I gruppi che si formano dalla segmentazione vengono detti segmenti, essi sono dei gruppi di consumatori di una determinata tipologia di prodotti o servizi che richiedono all'azienda specifici e differenziati strumenti del marketing mix.

Dai segmenti che emergono dall'analisi verranno scelti i più adatti all'azienda e i risultati dell'indagine serviranno a scegliere le adeguate strategie di marketing.

Il posizionamento e la differenziazione del prodotto hanno come punto fondamentale l'utilizzo della tecnica di segmentazione. L'azienda, con l'utilizzo di questa tecnica, può quindi scegliere dei target, ossia degli specifici segmenti a cui proporre la propria gamma di offerte e, all'interno degli stessi, può variare la propria offerta (*Figura 1.1*).

Figura 1.1 Fasi del target marketing.



L'individuazione della funzione di domanda individuale, propria per ogni consumatore, è la base della segmentazione. I segmenti di consumatori sono accomunati da una stessa funzione di domanda che permette di adattare i prodotti dell'azienda a specifici bisogni.
Non essendo possibile osservarla direttamente, l'eterogeneità della domanda viene percepita in modo diverso da azienda a azienda, essa viene quindi rilevata con criteri differenti. L'interpretazione del mercato di interesse e della domanda a cui si è rivolti rimane comunque un punto molto importante.

1.1 Fasi della Segmentazione

La procedura di segmentazione è formata da quattro fasi:
- definizione del problema;
- indagine;
- scelta del metodo di analisi;
- elaborazione e interpretazione dei dati.

Nella prima fase si dovrà decidere che tipo di indagine si vuole fare, se un'unica indagine o se si vuole un'indagine ripetuta e andrà scelto un adeguato metodo per la segmentazione.

Le aziende che scelgono di fare un'unica indagine, hanno come obiettivo la conoscenza approfondita del mercato d'interesse e l'indagine risulta per esse molto dispendiosa.
Anche l'analisi che si svolge per periodi ripetuti, panel, risulta essere per l'azienda molto costosa ma permette di vedere l'evoluzione nel

tempo del mercato e di adeguare così le proprie strategie ai cambiamenti. Questo però non è sempre possibile da attuare.

Per implementare il piano di segmentazione è necessario definire le variabili di interesse, per poter classificare i consumatori e per descrivere i segmenti (descrittori). In seguito sarà necessario scegliere tra la segmentazione a priori o a posteriori.

Definito il modello da adottare, si dovrà passare al secondo passo, la raccolta dei dati. La raccolta dei dati primari avviene tramite opportuni metodi d'indagine, con l'osservazione del comportamento del consumatore, con metodi qualitativi, ad esempio focus group, interviste in profondità e tecniche proiettive, e quelli più usati sono i metodi qualitativi, che prevedono l'utilizzo del questionario. Il passo successivo sarà definire l'unità d'indagine. La scelta delle variabili è comunque un passo molto importante perché essa si ripercuote sulle procedure analitiche, sulla grandezza e sulla composizione dei segmenti che verranno definiti. Per poter avere un campione rappresentativo dell'intera popolazione di riferimento sarà necessario ricorre all'estrazione di un campione probabilistico, progettando opportune strategie di rilevamento. Non è comunque necessario sempre condurre un indagine con dati primari in quanto le aziende hanno a disposizione dati secondari, reperibili a basso costo e velocemente, che si possono utilizzare per la segmentazione. La scelta tra le due tipologie di dati verrà fatta tenendo conto dei costi, dei benefici e delle esigenze dell'azienda.

Gli aspetti che vengono considerati per la scelta della più adatta tecnica statistica per l'analisi sono molteplici, la tipologia dei dati, la qualità, gli obiettivi finali, le risorse disponibili e la tipologia del modello di segmentazione scelto. Il criterio di segmentazione adottato impone, per le sue caratteristiche, l'uso di determinate tecniche statistiche che meglio si adattano al mercato d'interesse.

L’importanza fondamentale nel corso dell’indagine è di tener sempre in considerazione gli obiettivi strategici dell’azienda, l’analista e il management dovranno quindi definire in modo dettagliato il problema in modo tale da garantire il successo dell’indagine.

La fase di interpretazione è la fase più delicata, in quanto i segmenti devono avere determinati requisiti per poter portare all’azienda il vantaggio competitivo.

1.2 Vantaggi della Segmentazione

I vantaggi che si ricavano dalla procedura di segmentazione sono molteplici, ma il più importante è la possibilità di migliorare e approfondire il mercato di riferimento permettendo di trovare l’adatto posizionamento. La conoscenza del mercato permette di valutare i punti di forza e di debolezza dell’azienda, ma anche della concorrenza, potendo così definire con più precisione ed efficacia gli obiettivi e consente di modificare il portafoglio prodotti in modo proficuo.

La segmentazione permette di definire i bisogni del cliente, possiamo percepire i cambiamenti che avvengono nel mercato, così da poterli affrontare al meglio.

Un buon posizionamento all’interno del mercato di riferimento permette di costruire una barriera d’entrata ai nuovi concorrenti.

Un altro aspetto dovuto alla segmentazione è la capacità dell’azienda di misurare gli effetti sulle vendite delle strategie di marketing attuate.

1.3 Criteri di Segmentazione

Esistono cinque tipologie di segmentazione del mercato. Esse si differenziano per le *basi* scelte ossia se si tratta di criteri di

segmentazione geografica, demografica, psicografica, comportamentale e per benefici attesi.

Il primo criterio è la segmentazione geografica che permette di dividere il mercato in aree territoriali, in quanto si pensa che le preferenze dei consumatori dipendano della caratteristiche del territorio in cui si trovano, alcune variabili base che potremmo usare sono città, densità del territorio e caratteristiche climatiche.

La segmentazione demografica ha come basi, ad esempio, età, sesso e numero dei componenti del nucleo famigliare.

Questi due soli criteri permettono all'azienda di conoscere in modo approfondito il profilo dei loro consumatori e mostra come poterli raggiungere. Non vi è data però considerazione ai desideri del cliente e alle sue aspettative, non riuscendo ad ottenere informazioni sul loro processo decisionale. Vi è quindi bisogno di utilizzare anche altri criteri per poter rimediare a questo problema.

Il criterio della segmentazione psicografica ha lo scopo di individuare dei segmenti con stili di vita simili[1] e per farlo si avvale dell'aiuto della psicologia, della sociologia, dell'antropologia e del behaviorismo[2]. Le basi scelte riguardano gli interessi, le opinioni, le attività e la convinzioni dei clienti.

Il quarto criterio, la segmentazione comportamentale, è centrato sul comportamento del consumatore, focalizzandosi sugli obiettivi e sulle caratteristiche richieste dal consumatore al prodotto. Questo criterio permette di raccogliere informazioni utili per quanto riguarda il posizionamento dell'azienda nel mercato. Alcune basi possibili sono le occasioni d'uso, vantaggi richiesti e il tipo di

[1] Giampaolo Fabris definisce gli stili di vita come"*insiemi di persone che per loro libera scelta adottano modi di comportarsi (in tutti i campi della loro vita sociale ed individuale) simili, condividono gli stessi valori ed esprimono opinioni ed atteggiamenti omogenei*" , (Fabris, 1992).

[2] Il behaviorismo è una disciplina che ricostruisce e cerca di spiegare i modelli di consumo emergenti in un determinato contesto, (Prandelli, Verona, 2006).

consumatore, ad esempio se si tratta di un consumatore occasionale o meno, e la propensione all'acquisto.

L'ultimo criterio è la benefit segmentation[3], essa raggruppa i consumatori in segmenti omogenei sulla base di benefici e vantaggi che sono richiesti al prodotto o al servizio.

Segmentando il mercato con i criteri di segmentazione psicografica, comportamentale e benefit segmentation si ha bisogno di dati più difficili da ottenere e soprattutto più costosi, essendo relativi a aspetti personali dei consumatori. Anche se comportano degli svantaggi, i dati che si ricavano permettono di ottenere informazioni strategicamente molto importanti.

1.4 Criteri dei Segmenti

Non potendo ottenere una segmentazione perfetta del mercato, anche utilizzando tutti i criteri a nostra disposizione, diviene utile individuare delle caratteristiche per i segmenti, per poter ottenere un'immagine completa del mercato e della domanda di riferimento che sia strategicamente funzionale.

Come già anticipato, i segmenti devono essere *identificabili*, omogenei al loro interno ed eterogenei tra di loro. I consumatori dovranno essere raggruppati in base a domande individuali simili ma ben distinte da quelle di consumatori appartenenti ad altri segmenti. I segmenti dovranno essere *consistenti* e *profittabili*, ossia essere di una certa ampiezza e avere una capacità di assorbimento tali da poter garantire il profitto all'azienda. Queste due caratteristiche garantiscono la *meaningful segmentation,* ossia l'efficacia delle strategie implementate. I segmenti dovranno essere *stabili* nel tempo. Dovranno essere *accessibili*, in quanto deve essere possibile

[3] Per un approfondimento si vedano Grandinetti (2002) e Haley (1968).

raggiungere questi segmenti con opportuni strumenti del marketing mix.

I segmenti dovranno avere *capacità di risposta*, quindi essere reattivi, e *propositivi*, per attuare opportune strategie di marketing e per capire aspettative e bisogni. Quest'ultime caratteristiche garantiscono l'*actionable segmentation*, l'efficienza delle strategie, ottenendo dei risultati che superino le risorse impiegate. Dato che ogni segmento ha gradi diversi per le varie caratteristiche, non è sempre possibile ottenere sia la segmentazione meaningful che actionable.

1.5 Modelli e Tecinche Statistiche di Segmentazione

Dopo aver definito le caratteristiche dei segmenti, si dovrà passare a scegliere lo schema di segmentazione. I modelli di segmentazione si dividono in a priori e in a posteriori. Nei modelli a priori la popolazione di interesse viene divisa in base a modalità scelte, appunto, a priori; verranno fissati prima dell'analisi la dimensione e la tipologia dei segmenti, ad esempio si useranno come basi l'area geografica. Nei modelli a posteriori, invece, le basi, la tipologia e il numero di segmenti vengono definiti durante l'indagine, sulla base di criteri di dissomiglianza. Questo tipo di modelli viene usato nel caso della segmentazione psicografica, per benefici attesi e comportamentale. Altre tecniche di segmentazione sono per omogeneità, per obiettivi e flessibili, la cui scelta avverrà nella fase di selezione della metodologia d'analisi.

Nella tecnica per omogeneità i consumatori vengono divisi in gruppi che dovranno avere un'elevata omogeneità interna ed eterogeneità esterna, in base alla similarità di determinate variabili (ad esempio la *Cluster Analysis*). Nelle tecniche flessibili, le unità vengono suddivise in base alla similarità dei profili, in termini di

preferenze per i prodotti (*Conjoint Analysis*). Le tecniche per obiettivi raggruppano le unità statistiche in base a una o più variabili dipendenti, definite a priori, da cui le variabili esplicative, che descrivono le caratteristiche dei segmenti, sono influenzate. Tra le tecniche per obiettivi ritroviamo Automatic Interaction Detection (AID), Chi Squared Automatic Interaction Detection (CHAID) e la regressione logistica.

La classificazione delle unità statistiche in segmenti e la determinazione dei profili sono gli scopi delle tecniche statistiche a priori. La suddivisione degli individui avviene secondo le modalità di una variabile scelta a priori mentre, nel caso fossero presenti più basi, la suddivisione seguirebbe modalità di classificazione incrociata delle unità. Nel caso fossimo in presenza di basi di tipo continuo è necessario convertirle a variabili categoriali, limitando le modalità e quindi i diversi impatti che sono provocati dall'uso di metriche diverse. Questa modifica essendo soggettiva comporta, però, la difficoltà di individuare in modo chiaro le relazioni tra le variabili che potrebbero essere di significative. La segmentazione per obiettivi racchiude le tecniche a priori, che hanno infatti l'obiettivo di legare alla variabile dipendente le variabili esplicative rilevanti, con lo scopo di descrivere i segmenti e spiegare le cause del comportamento del consumatore. Questo tipo di segmentazione precede diverse fasi: la selezione delle basi; poi le variabili esplicative o concomitanti che dovranno descrivere le caratteristiche delle unità; in seguito le unità verranno divise in due gruppi, esaustivi e mutuamente esclusivi, sulla base delle variabili esplicative. La regola di ottimalità permette di ricavare la migliore segmentazione, in base all'omogeneità interna e l'eterogeneità esterna secondo la variabile criterio. La migliore segmentazione viene determinata ad ogni passo dell'analisi. La suddivisione delle unità in gruppi si arresta al momento in cui si raggiungono le condizioni prefissate.

Le tecniche a posteriori usano un procedimento empirico di classificazione, non fissando anticipatamente le modalità. Tra queste tecniche troviamo la segmentazione per omogeneità, che lascia la ricerca della massima omogeneità interna e minima omogeneità esterna all'analisi statistica, e la segmentazione flessibile.

1.5.1 Tecniche per Obiettivi

Tra le tecniche per obiettivi troviamo principalmente AID e CHAID.

L'AID è una tecnica gerarchica di segmentazione binaria in cui vengono considerate, ad ogni passo dell'analisi, tutte le possibili suddivisioni dicotomiche in gruppi disgiunti, sulla base di una specifica variabile esplicativa. La scelta tra i vari gruppi viene fatta considerando la variabile che meglio scinde il mercato in due sottogruppi, con massima omogeneità entro e minima eterogeneità fra i gruppi, rispetto la variabile base. Questa scelta viene ripetuta ad ogni passo, essendo questa una procedura iter attiva, e termina quando si giunge ad un gruppo piccolo di unità, rispetto ad una soglia prefissata. Ci si arresta nel caso in cui il gruppo di origine non rende necessaria un'ulteriore suddivisione, avendo raggiunto la soglia minima di devianza tra i gruppi, se la bipartizione non provoca un incremento sufficiente della devianza tra i gruppi ed, infine, ci si arresta se si raggiunge il numero massimo di passi prefissati per il processo di analisi.

La CHAID è anch'essa una procedura gerarchica, ma attua una segmentazione multipla, non vi è quindi una bipartizione dei gruppi, come avviene per il metodo AID. La valutazione del criterio di ottimalità, avviene utilizzando il test Chi-quadro, χ^2. Viene testata l'ipotesi nulla di indipendenza delle caratteristiche, mettendo a confronto le frequenze osservate con quelle teoriche, quest'ultime

calcolate sotto l'ipotesi nulla di indipendenza; un elevato valore del test evidenzia una forte dipendenza tra le variabili, e quindi omogeneità interna. In questo modo la CHAID permette di individuare le variabili che sono più legate alla base e le più adatte a descrivere il profilo del segmento.

Un'ulteriore tecnica da inserire nell'analisi a priori è l'*analisi discriminante multipla* (Brasini, Tassinari F., Tassinari G., 1993). Questa tecnica analizza la relazione tra le variabili base, categoriche, e le predittive, che descrivono gli individui. L'analisi ha l'obiettivo di determinare una regola che predìca quale modalità della variabile criterio presenta un individuo, in base ad una funzione lineare che massimizza il rapporto di devianza tra ed entro i segmenti per la variabile criterio. Quest'analisi, quindi, rende possibile la classificazione delle unità anche se si conosce solamente il loro profilo, permette di verificare l'esistenza di eventuali differenze significative tra i valori medi delle variabili esplicative all'interno dei gruppi e consente di individuare le variabili specifiche che determinano le differenze tra i profili medi.

1.5.2 Tecniche per Omogeneita'

Tra le tecniche per omogeneità ritroviamo la Cluster Analysis e la Conjoint Analysis.

La Cluster Analysis risulta essere la più diffusa. Si tratta di una tecnica multivariata ed esplorativa, essa mira a scomporre la realtà, con molteplici osservazioni, in specifiche tipologie; permette di suddividere un insieme eterogeneo in sottoinsiemi mutuamente esclusivi, ed omogenei al loro interno. La Cluster Analysis tende a migliorare la comprensione dei comportamento d'acquisto degli individui, a valutare i legami tra i propri prodotti e quelli della concorrenza, a valutare l'opportunità di sviluppare nuovi prodotti e a

selezionare mercati prova per eventuali test di mercato. L'obiettivo di questa tecnica è di individuare i gruppi omogenei, non esiste però nessuna regola generale che guidi l'analista nella scelta del numero di gruppi ottimale. Un punto fondamentale risiede nella disponibilità del campione (n unità), esso viene rappresentato da un numero specifico di variabili (p) e i dati raccolti vengono inseriti in una matrice nxp, dove ogni riga rappresenta il profilo di un'unità statistica, che verrà sottoposta a elaborazioni successive. L'indagine inizia con la selezione degli elementi da analizzare e delle variabili da utilizzare per la segmentazione, successivamente emergeranno i raggruppamenti in base all'utilizzo di opportune procedure di clustering, tra cui i più diffusi sono l'algoritmo gerarchico e non gerarchico.

L'algoritmo gerarchico prevede di partire da n gruppi formati da un'unità e aggregare di volta in volta due gruppi per il criterio di similarità, per arrivare a n partizioni concatenate. Ogni gruppo è legato al gruppo del passo precedente, in quanto, quando avviene la fusione dei gruppi, essi non possono più essere separati. Questa concatenazione porta l'emergere dello svantaggio di questo metodo, ogni gruppo è legato al gruppo sorto al passo precedente. La Cluster Analysis utilizza cinque algoritmi gerarchici per valutare le distanze dai gruppi: il metodo del legame singolo, del legame completo, del legame medio, del centroide e di Ward.

Nell'algoritmo non gerarchico si parte da una partizione delle unità in g gruppi e successivamente si spostano le unità da un gruppo all'altro fino ad ottenere gruppi con massima omogeneità interna e minima omogeneità tra i gruppi. Tra questi algoritmi troviamo quello di McQueen e Foggy, che allocano le unità al centroide più vicino minimizzando la devianza entro relativamente alle p variabili.

Per analizzare se i cluster possono essere rappresentativi della realtà, avendo quindi dei valori medi di gruppo significativamente diversi tra loro, se confrontati con i valori medi usati nell'analisi del raggruppamento, si utilizza il test di Arnold e del coefficiente di correlazione cofenetico (solo per algoritmi gerarchici).

La Conjoint Analysis, fa parte delle segmentazione flessibile che prevede di chiedere agli intervistati di ordinare profili di uno stesso prodotto, che differiscono per almeno una modalità di un attributo, potendo così stimare l'*utility parthwoth*, l'utilità associata a ciascuna caratteristica. Viene ipotizzato che il consumatore agisca in modo razionale durante l'acquisto di un prodotto, scegliendo il bene che massimizza la propria utilità. L'utilità del bene deriva dagli attributi che lo compongono e la sua utilità complessiva deriva dalla somma delle utility parthworth. Per queste considerazione i mercati di riferimento sono quelli di prodotti a forte coinvolgimento psicologico. L'analisi Conjoint permette di comprendere e misurare i compromessi che i consumatori accettano, scegliendo un determinato prodotto.

L'obiettivo principale della Conjoint Analysis è di identificare la combinazione ottima delle caratteristiche del prodotto, per prevedere le preferenze dei consumatore e riconoscere i segmenti di mercato potenziali.

L'analisi Conjoint seleziona un campione di consumatori, in seguito individua gli attributi[4], incorrelati tra di loro, che potrebbero essere rilevanti per i consumatori; il passo successivo consiste nella definizione dei profili dei prodotti, *stimoli*, da sottoporre al giudizio

[4] Molto spesso si utilizza come attributo il prezzo, considerato dai consumatori come indice di qualità; va posta molta attenzione nella scelta di questo attributo in quanto se si sceglie sia prezzo che qualità nello schema di rilevazione, si rischia che esso venga sopravvalutato, nel caso di beni dal costo unitario elevato, o sottovalutato, nel caso di beni dal basso costo unitario.

tramite un piano degli esperimenti. Si dovrà scegliere tra sottoporre tutti gli stimoli (disegno fattoriale) o una loro parte (disegno fattoriale frazionato), garantendo l'affidabilità somministrando z profili, pari al numero totale dei livelli – il numero degli attributi + 1. Il consumatore dovrà ordinare gli attributi in base alla preferenza accordata ai profili, rilevando i punteggi assegnati ad ogni attributo. L'analisi continua stimando le utilità parziali utilizzando un modello di utilità scelto in precedenza (tra modelli vettore, punto ideale, parthworth), garantendo la massima corrispondenza tra punteggi di preferenza rilevati e punteggi previsti[5]. I possibili dati per l'analisi, oltre a quelli individuali, sono i dati aggregati. In questo modo diviene possibile calcolare l'importanza relativa associata ad ogni attributo e valutare l'utilità totale di ogni alternativa di prodotto. Anche se le alternative non sono valutate direttamente, è comunque possibile ricostruire le preferenze del consumatore utilizzando la simulazione del comportamento di scelta dello stesso con, ad esempio i criteri *first choice.*

Un ulteriore aspetto che viene valutato da questa tecnica è la disponibilità dei consumatori di combinare tra loro diverse modalità o livelli di attributi di prodotto, utilizzando semplici procedure di rilevazione e di stima. L'efficacia di questa tecnica emerge nel lancio di un nuovo prodotto o di un rilancio di prodotto e nel riconoscimento di nicchie di mercato.

La Conjoint Analysis presenta anche degli svantaggi, infatti il processo d'acquisto viene rappresentato solo in modo approssimato. I risultati dell'analisi sono condizionati dallo schema che collega le preferenze con l'utilità totale e dall'assenza di interazione tra gli attributi. Infine, sono gli attributi utilizzati che determinano la

[5] Tramite la correlazione tra le scelte previste e quelle scelte osservate, R di Pearson e τ di Kendall, è possibile valutare l'adattamento.

combinazione ideale, un cambiamento degli attributi può modificare i risultati, soprattutto se non sono stati inseriti attributi chiave.

CAPITOLO 2

Modelli a Classi Latenti

Molto spesso è necessario scindere il problema della segmentazione in due livelli per diminuirne la complessità, uno riguardante la definizione dei segmenti in base a criteri opportuni e un altro riguardante la qualità delle informazioni che dovranno essere utili per le strategie di marketing.

Già dagli anni ’80, per la segmentazione, si sono adoperati modelli a mistura finita, tra cui i più famosi modelli CL (Vermunt e Madigson, 2002), essi permettono di individuare la similarità in una popolazione omogenea. Questi modelli aiutano a segmentare il mercato in più livelli, permettendo di inserire delle variabili descrittori, *covariate*, nell’analisi e rendendo possibile suddividere il mercato in strutture gerarchiche multilivello. La procedura usata per la segmentazione è di tipo probabilistico e, quindi, più flessibile delle tecniche classiche che uniscono in gruppi le unità statistiche sulla base della loro similarità, o distanza, da una variabile di interesse, come accade per la Cluster Analysis.

In questo approccio si vuole riassumere l’eterogeneità osservata sulle unità statistiche con basi delle modalità di una variabile latente di tipo discreto, per poter definire delle classi omogenee. In quest’approccio si ipotizza che si possano riassumere in caratteristiche latenti quelle osservate nelle variabili a disposizione.

I segmenti si formano in base alle unità a disposizione, esse sono assegnate a cluster latenti individuali, determinati dalle probabilità condizionate di avere una determinata caratteristica, data l’appartenenza a una classe latente.

L'ipotesi fondamentale è l'*indipendenza locale* delle osservazioni, ossia la loro indipendenza data l'appartenenza ad una classe latente (in altri termini le variabili manifeste sono indipendenti, all'interno di ogni classe latente, e l'associazione fra di esse è spiegata dalle classi della variabile latente).

L'analisi multilivello permettere di tener in considerazione la struttura della popolazione, gerarchica o annidata, in più livelli, ossia raggruppa al primo livello le unità statistiche e individua nel secondo le classi latenti che raggruppano le unità statistiche. I modelli CL sono stati introdotti per la misurazione di variabili latenti attitudinali a partire da item dicotomici, da Lazarsfeld e Henry (1968). L'opportunità di utilizzare dati dicotomici permise di allargare il raggio d'azione di questi modelli, al contrario dell'analisi fattoriale che utilizza solo variabili continue[6]. La diffusione dell'approccio CL avvenne però solo grazie ai lavori di Goodman (1974a, 1974b) che formalizzò la metodologia dei modelli e ne estese l'applicazione anche a variabili nominali, elaborandone poi un algoritmo di stima di massima verosimiglianza. In seguito i modelli furono estesi anche alle variabili ordinali (Heinen, 1996), a indicatori continui, a variabili su scala nominale, ordinale e continua (Vermunt e Madigson, 2001) e a covariate. La possibilità di utilizzare le covariate permette di descrivere i gruppi in modo efficace, senza dover utilizzare in un secondo momento l'analisi discriminante, che stabilisce un criterio per assegnare correttamente ulteriori unità ai segmenti, precedentemente individuati.

Questi modelli permettono di relazionare una serie di variabili osservate discrete categoriali multivariate con un insieme di variabili latenti discrete categoriali. Le *classi* rappresentano le modalità delle

[6] L'analisi a CL applica alle variabili categoriali lo stesso principio che l'analisi fattoriale applica alle variabili cardinali, non viene quindi violata la natura delle variabili (Lazarsfeld, 1951).

variabili e ogni classe è caratterizzata da un insieme di probabilità condizionate che indicano la probabilità che le variabili assumano un determinato valore. I segmenti di un mercato sono rappresentati dalle classi di una variabile latente. Le unità statistiche vengono invece definite *casi*, e vengono assegnate alle classi in base alla loro probabilità di appartenere a una determinata classe, formando dei gruppi mutuamente indipendenti. In questo modo la popolazione, eterogenea, viene ridotta in sottogruppi di unità omogenei internamente ed eterogenei fra di loro, non rendendo più indispensabile porre delle assunzioni sui dati (linearità della relazione, normalità, omogeneità), riducendo eventuali distorsioni.

I modelli CL si fondano sull'ipotesi di poter riassumere le caratteristiche osservate nelle variabili in altre caratteristiche latenti.
Costruire gruppi internamente omogenei e mutuamente indipendenti, non imporre assunzioni restrittive ai dati e permettere l'uso delle covariate sono i vantaggi derivanti dall'utilizzo dei modelli CL. Un altro vantaggio di questo metodo è la possibilità di utilizzare software ad hoc per la stima, che permettono la sua diffusione nell'ambito di indagini di marketing, nonché in quelle sociali e biomediche.
I segmenti che vengono individuati sono il risultato dell'intera analisi svolta e non prefissati ad inizio analisi, questa tecnica è infatti predittiva a posteriori.

L'analisi CL permette di stimare la scelta dell'individuo e l'appartenenza ad un segmento latente simultaneamente, presentando dei risultati informativi e di facile interpretazione.
I modelli principali sono quelli tradizionali, fattoriali e multilivello (multigruppo).

2.1 Modelli Tradizionali

L'obiettivo dell'analisi tradizionale è di individuare il minor numero di classi latenti, che possono spiegare le associazioni osservate tra le variabili manifeste, partendo da una tabella di contingenza ad entrata multipla. È assunto che ogni osservazione possa appartenere ad una sola classe latente e che sussista l'ipotesi di indipendenza locale tra le variabili osservate, che implica la mutua indipendenza tra loro, condizionatamente all'appartenenza alla classe latente.

I modelli tradizionale vengono definiti da Madigson e Vermunt (2001) e Vermunt e Madigson (2002) *Latent Class Cluster Model*, sottolineando l'obiettivo dell'analisi, classificare le unità in T gruppi omogenei, comune alla Cluster Analysis.

Il modello tradizionale si può esprimere usando come parametri la probabilità (non condizionata) di appartenere ad ogni classe latente e le probabilità condizionate di risposta. Supponiamo di disporre di K variabili manifeste ($k=1,...,K$) per n soggetti di un'indagine (indicizzati con $i=1,...,n$) e una variabile latente X, con T $(t=1,...,T)$ classi; allora il modello CL può essere formulato come segue:

(1)

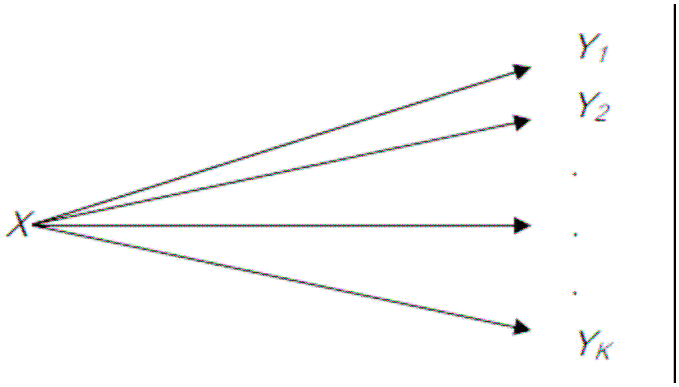
$$P(Y_i)=\sum_{t=1}^{T}P(X_i=t)P(Y_i\mid X_i=t)=\sum_{t=1}^{T}P(X_i=t)\prod_{k=1}^{K}P(y_{ik}\mid X_i=t)=\sum_{t=1}^{T}P(X_i=t)\prod_{k=1}^{K}P(y_{ik},\vartheta_{k_t})$$

In questo modello la probabilità che l'individuo i dia il vettore di risposte Y è indicata da $P(Y_i)$; la probabilità che l'individuo i appartenga alla specifica classe t è indicata con $P(X_i=t)$; la probabilità che l'individuo i dia la risposta y alla variabile manifesta k, dato che appartiene alla *classe* t, è $P(y_{ik}\mid Xi = t)$. Per definire la

distribuzione della variabile k, appartenente alla classe latente t, si stimano i parametri ϑ_{kt}.

Una rappresentazione grafica può essere data dalla *Figura 2.1*, in cui le variabili manifesto sono connesse tra loro grazie la variabile latente X, la quale spiega tutte le relazioni tra le variabili osservate.

Figura 2.1: path diagram; modello CL a K variabili osservate e una latente



2.1.1 Stima dei Modelli Tradizionali

Il primo passo dell'analisi è la stima del modello con variabili mutuamente indipendenti, ossia si fissa il numero di classi latenti (T) a uno.

$$P(Y_i) = \prod_{k=1}^{K} P(y_{ik})$$

Il secondo passo sarà aumentare le classi latenti, T=2, e l'analisi continua aumentando di volta in volta di uno la dimensione T, fino ad arrivare al modello che meglio descrive i dati.

2.1.2 Specificazione delle Distribuzioni

La scala delle variabili osservate incide sulla forma distributiva delle osservazioni. Per le variabili conteggio viene solitamente usata una distribuzione *Binomiale* o *Poisson*, per quelle continue si usa la distribuzione *Normale* e per le variabili categoriali si usa la distribuzione Multinomiale (nello specifico, logistico multinomiale per le variabili nominali e logistico ordinale per categorie adiacenti).

Ad esempio, per le variabili nominali avremmo la seguente distribuzione multinomiale:

$$P(y_{ik} = s | X_i = t) = \frac{\exp\left\{\eta^k_{s|t}\right\}}{\sum_{s=1} \exp\left\{\eta^k_{s|t}\right\}} \quad (2)$$

In questa distribuzione, $s=1,...,S^k$ indica una particolare categoria dell'osservazione y_{ik} e S^k rappresenta il numero di categorie di ogni y_k; la probabilità di rispondere una particolare categoria, s, data la variabile latente X, da parte dell'individuo i, è indicata con $P(y_{ik}=s \mid X_i=t)$; $\eta^k_{s|t}$ è il termine lineare che permette di distinguere i modelli logit multinomiali e i logit ordinali per categorie adiacenti.
Il predittore lineare è dato da:

$$\eta^k_{s|t} = \beta^k_{s0} + \beta^k_{st0}$$

con β^k_{s0} è indicata l'intercetta mentre β^k_{st0} è specifico per ogni classe.

La funzione di probabilità che corrisponde alle risposte del soggetto *i-esimo*, è formata da due probabilità, una relativa alla variabile latente e una per la variabile dipendente, come si può vedere dall'equazione (1). Per questo motivo, diviene necessario definire, successivamente alla determinazione della distribuzione della probabilità condizionata, anche la distribuzione della variabile

latente X, che è legata alla sua natura, nominale o ordinale. Essa ha forma multinomiale e viene parametrizzata nel seguente modo:

$$P(X=t)=\frac{\exp\{\eta_t\}}{\sum_{x'=1}^{T}\exp\{\eta_{t'}\}} \quad (3)$$

Il modello Logit Multinomiale standard viene utilizzato nel caso si disponga di una variabile latente. Il termine lineare per le t classi latenti è $\eta_t=\gamma_{t0}$ e il rispettivo vincolo per i parametri dell'intercetta γ_{t0} è:

$$\sum_{t=1}^{T}\gamma_{t0}=t$$

2.1.3 Le Covariate

Le covariate possono essere inserite nel modello sia per le variabili risposta che per le classi latenti. La distribuzione adatta a modelli con covariate è la Regressione Logistica Multinomiale, se si è in presenza di covariate per la variabile latente, e una regressione della famiglia dei Modelli Lineari Generalizzati, se è per la variabile risposta.

Il modello che contiene le covariate sia per le variabili latenti che per la risposta è definito come segue:

$$P(Y_i \mid Z_i)=\sum_{t=1}^{T}P(X_i=t \mid Z_i)\prod_{k=1}^{K}P(y_{ik} \mid X_i, Z_i)$$

dove Z_i è il vettore che contiene R covariate per l'individuo i.

I predittori lineari per la distribuzione condizionata sono (nel caso in cui gli indicatori sono nominali):

$$\eta_{s|t} = \beta_{s0}^{k} + \beta_{st0}^{k} + \sum_{r=1}^{R} \beta_{sr}^{k} z_{ir}$$

Essendo z_{ir} la *r-esima* covariata per l'individuo i ($r=1,...,R$).

Mentre i predittori lineari per la probabilità della variabile latente sono:

$$\eta_{t} = \gamma_{t0} + \sum_{r=1}^{R} \gamma_{tr} z_{ir}$$

In entrambi i casi, verranno rispettati i vincoli sui parametri (Vermunt e Madigson, 2005).

A differenza degli indicatori, le associazioni tra le covariate non sono spiegate dalla variabile latente.

2.1.4 Valutazione dell'Adattamento

Molteplici sono gli approcci per giudicare l'adeguatezza del modello, il più usato è L^2, rapporto statistico di verosimiglianza *Chi-quadro*, che permette di misurare la differenza tra le stime di massima verosimiglianza per le frequenze attese, $\hat{F}_i$, e le corrispondenti frequenze osservate f_i.

$$L^2 = 2\sum_{i=1}^{N} f_i \log \frac{\hat{F}_i}{f_i}$$

Un modello si adatta bene ai dati se il valore L^2 è sufficientemente basso (solitamente 0.5) da potersi attribuire al caso.

Il processo per ottenere le stime di massima verosimiglianza, per le frequenze $\hat{F}_i$, si divide in due fasi. Inizialmente si ottengono le stime di massima verosimiglianza delle probabilità di appartenenza alla classe latente, che sostituiscono quelle per i parametri del modello sulla parte destra del modello (1). Sommando le probabilità stimate per ogni classe latente si ottengono le stime di probabilità per ogni cella, se si moltiplicano per la numerosità, n, si ottengono le stime di massima verosimiglianza per le frequenze attese.

L'analisi a CL non è necessaria se il modello sotto ipotesi nulla fosse adeguato ai dati, ma dato che nella maggior parte dei casi questo non avviene viene utilizzato il valore L^2 come indice di adattamento. L^2 uguale a 0 indica che il modello spiega in modo perfetto i dati, ed avviene nel caso in cui vi sia una perfetta corrispondenza tra le frequenze stimate e quelle osservate; se L^2 è maggiore a 0, possiamo misurare la mancata adeguatezza del modello ai dati, potendo quantificare l'associazione non spiegata. Per determinare la percentuale di associazione spiegata introducendo le classi latenti possiamo confrontare il valore L^2 del modello sotto ipotesi nulla, in cui l'associazione tra le variabili è perfettamente spiegata, con $T>1$.

L^2 ha distribuzione asintotica X^2 sotto condizioni di regolarità con gradi di libertà pari al numero di celle nella tabella a più vie, a cui si sottrae il numero dei parametri distinti del modello, M, meno 1[7].

[7] Il numero di celle nella tabella multientrata è dato dal prodotto di tutte le modalità di ognuna delle k variabili manifeste, se ogni variabile osservata ha n_k modalità, il numero di celle è dato da $N_k = \prod_{k=1}^{K} n_k$. Il numero dei parametri, M, è dato da $M=T-1+\sum_{k=1}^{K} (n_k - 1)$. Il numero dei gradi di libertà è dato da $gl = N_k - M - 1$.

Se ci troviamo nel caso di dati sparsi[8], non è possibile utilizzare L^2 in quanto potrebbe non avere distribuzione asintotica *Chi-quadro*, può invece essere utilizzato *bootstrap approach* (Vermunt e Madigson, 2002).

L'alternativa all'utilizzo del rapporto di massima verosimiglianza, è l'utilizzo di criteri informativi, quali il criterio informativo Bayesiano (BIC, Schwarz, 1978) e il criterio informativo di Akaike (AIC, Akaike, 1974), che tengono conto sia della bontà delle stime che della parsimonia del modello.

Il criterio BIC è definito nel modo seguente (con gl i gradi di libertà della statistica L^2):

$$BIC_{L^2} = L^2 - \log(n)gl$$

La formulazione di BIC, sulla base della log-verosimiglianza è definita:

$$BIC_{LL} = -2LL + \log(n)M$$

con M il numero dei parametri.

Il valore BIC preferito è quello più basso.

L'indice AIC è invece definito :

$$AIC_{LL} = -2LL + 2M$$

Studi di simulazione hanno però mostrato che BIC tende a sottostimare il numero delle classi (ad esempio Dias, 2004) e AIC le sovrastima (ad esempio McLachlan e Peel, 2000). Sempre più

[8] Siamo in presenza di dati sparsi nel caso in cui il numero di variabili osservate o il numero delle loro categorie è molto alto, oppure quando il modello è esteso a variabili continue.

spesso nelle analisi a CL si utilizzano due versione aggiustate di AIC, AIC3 (Bozdogan, 1993) e CAIC, ossia l'indice AIC consistente (Bozdogan,1987).

$$AIC3_{LL} = -2LL + 3M$$

$$CAIC_{LL} = -2LL + (1 + \log(n))M$$

Per tutti i criteri viene preferito il modello con valore dell'indice minore.

2.1.5 Test di Significatività degli Effetti

Dopo la scelta del modello che meglio si adatta ai dati è necessario verificare la significatività delle stime eliminando le variabili che non sono significative.

La significatività dell'effetto della variabile viene esaminata testando l'ipotesi nulla di identica distribuzione delle categorie tra ogni classe:

$$H_0 : P(Y_{k1}|X = t) = P(Y_{k2}|X = t) = \ldots = P(Y_{kS}|X = t)$$

con k si denota la variabile di interesse; s è il numero delle categorie; t è il numero delle classi.

Il test può essere implementato usando la relazione tra probabilità di risposta condizionata e i parametri log-lineari. Si utilizza, a tal fine, la differenza tra L^2, ΔL^2, del modello senza la variabile in esame (*modello ristretto*) e quello con la variabile (*modello completo*, sotto ipotesi nulla H0). Sotto H0, l'ipotesi si distribuisce come un X^2, i gradi di libertà sono pari al prodotto tra la

differenza delle categorie della variabile sotto esame, n_k, meno 1, e il numero di classi meno 1 ($gl=(n_k-1)(T-1)$).

Per testare se i coefficienti di regressione sono uguali tra le classi, si utilizza il test di Wald. Questo test risulta meno potente del test ΔL^2. Il test si distribuisce come un X^2.

2.1.6 Classificazione

La classificazione degli individui nelle rispettive classi latenti è l'ultimo passo dell'analisi.

Per poter classificare gli individui è necessario calcolare la probabilità a posteriori, utilizzando il teorema di Bayes, che un individuo appartenga alla classe t, dato il suo pattern di risposte; il numeratore e il denominatore derivano dalla sostituzione delle stime al modello (1):

$$\hat{P}(X_i = t \mid Y_i) = \frac{\hat{P}(X_i = t)\hat{P}(Y_i \mid X_i = t)}{\hat{P}(Y_i)}$$

Le classi conterranno i casi che presentano probabilità a posteriore maggiore, in quanto una classe latente dovrà contenere unità con la stessa distribuzione di probabilità.

Utilizzare la probabilità per definire l'omogeneità dei gruppi, distingue l'analisi a CL dalla Cluster Analysis, che invece utilizza misure ad hoc.

2.2 Modelli a Classi Latenti Non Tradizionali

Con l'analisi CL classica possiamo incorrere nel problema di mancata adeguatezza del modello, dovuta al fatto che l'ipotesi di indipendenza locale non sussiste. Il problema può essere risolto aggiungendo un'ulteriore classe latente al modello ma, non essendo sempre accettabile questa aggiunta, esistono delle alternative più parsimoniose e congruenti con le ipotesi iniziali.

Un'alternativa consiste nell'aggiungere uno o più effetti diretti, che siano in grado di spiegare le associazioni residue tra le variabili osservate responsabili della dipendenza locale. Questa opzione si utilizza nel caso in cui dei fattori esterni, non correlati con la variabile latente, portino però alla creazione di associazione tra due variabili.

Un'altra opzione è l'eliminazione di item (y_{ik}), utilizzata quando la dipendenza locale è da imputarsi a due variabili e l'eliminazione di una delle due toglie anche l'effetto di dipendenza esistente, strategia da utilizzare soprattutto nel caso di item ridondanti.

L'ultima alternativa, per risolvere il problema di inadeguatezza del modello, consiste nell'aumentare il numero delle variabili latenti, da utilizzarsi soprattutto nel caso in cui la dipendenza locale sia data da un gruppo di variabili. Il modello CL fattoriale è il modello che meglio si adatta ai dati, secondo Madigson e Vermunt (2001), infatti questo modello prevede l'aumento del numero di variabili latenti, non di classi latenti.

Una statistica che aiuta nella scelta della strategia più appropriata è il *Bivariate Residual* (BVR), che misura quanta associazione osservata tra due variabili è spiegata dal modello, definendo le relazioni bivariate che il modello non riesce a spiegare. Questa statistica viene calcolata dividendo la statistica Chi-quadro di Pearson (usata per il test di indipendenza) per i gradi di libertà e

confrontando le frequenze osservate con le frequenze attese del modello CL utilizzato, partendo da una tabella a due entrate per le due variabili in esame.

2.3 Modelli a Classi Latenti Fattoriale

Goodman (1974a) introduce i modelli CL fattoriali nell'analisi a classe latente confermativa e in seguito riproposti come alternativa nell'analisi esplorativa CL tradizionale da Madigson e Vermunt (2001). Questi modelli permettono di includere più variabili latenti nel modello, incrementando così il numero di fattori, interpretati da Goodman (1974b) come una variabile congiunta. L'analisi permette di utilizzare variabile latenti multiple per modellare le relazioni multidimensionali che esistono tra le variabili manifeste.

Un esempio renderà più chiara quest'analisi. Supponendo di avere una variabile latente X a 4 classi ($T=4$), essa può essere espressa con due variabili latenti dicotomiche $V = \{1, 2\}$ e $W = \{1, 2\}$, nel seguente modo (*Figura 2.3*):

Figura 2.3

	W=1	W=2
V=1	X=1	X=2
V=2	X=3	X=4

Le restrizione imposte ai parametri, distinguono i modelli CL in varie tipologie e sono definite da Madigson e Vermunt (2001). Il modello fattoriale base è un modello con più fattori dicotomici mutuamente indipendenti, che esclude interazioni di ordine superiore nelle probabilità condizionate di risposta. Se il modello presenta R fattori avrà esattamente lo stesso numero di parametri distinti di un modello CL tradizionale con $R+1$ classi, anche specificando un modello con 2^R classi avremmo ancora lo stesso numero di parametri

del modello tradizionale con $R+1$ classi latenti. Questo modello è quindi più parsimonioso e di facile interpretazione, si adatta, inoltre a molteplici situazioni.

In questo modello possono essere introdotte covariate nonché usare dati di tipo categoriale, conteggio e una loro combinazione.

Si può parametrizzare un modello a quattro classi con quattro variabili nominali (A,B,C,D) tramite un modello fattoriale a CL con due variabili latenti dicotomiche (V, W) come segue:

$$\pi_{ijklrs} = \pi_{rs}^{VW} \pi_{ijklrs}^{ABCD|VW} = \pi_{rs}^{VW} \pi_{irs}^{A|VW} \pi_{jrs}^{B|VW} \pi_{krs}^{C|VW} \pi_{lrs}^{D|VW}$$

In questo modello π^{VW}_{rs} è la probabilità che V e W assumono rispettivamente i valori r ($r=1,\ldots,R$) e s ($s=1,\ldots,S$) e $\pi^{A|VW}_{irs}$ è la probabilità condizionata che l'item A presenti la risposta i dato $V=r$ e $W=s$, $\pi^{B|VW}_{jrs,}$ $\pi^{C|VW}_{krs}$ e $\pi^{D|VW}_{lrs}$ sono rispettivamente le probabilità condizionate degli item B, C e D.

2.4 Modelli a Classi Latenti Multilivello

Un'utile modello CL nel caso di osservazioni dipendenti e che presentano una struttura gerarchica, in cui si possono raggruppare i casi in gruppi (per esempio medici raggruppati per ospedali), è il modello multilivello. La struttura gerarchica può essere considerata utilizzando modelli CL e l'analisi multilivello standard in combinazione. A differenza dei modelli CL standard, questo modello permette ai parametri di variare introducendo una variabile latente discreta.

Un semplice modello CL per dati multilivello a J gruppi ($j=1,\ldots,J$) consiste in:

(4)

$$P(Y_{ij}=s)=\sum_{t=1}^{T}P(X_{ij}=t)P(Y_{ij}=s\mid X_{ij}=t)=\sum_{t=1}^{T}P(X_{ij}=t)\prod_{k=1}^{K}P(y_{ijk}=s_k\mid X_{ij}=t)$$

Il vettore di risposte dell'individuo i, che appartiene al *j-esimo* gruppo, è indicato da Y_{ij}; la risposta dell'individuo i, del gruppo j, all'item k è indicata con y_{ijk}; la variabile latente, riferita all'individuo i del gruppo j, è X_{ij}; s_k è un particolare livello dell'item k ($s_k=1,\ldots, S_k$).

In questo modello la media ponderata delle probabilità specifiche di classe, per la probabilità che l'unità i appartenente al gruppo j appartenga alla *t-esima* classe latente, rappresenta la probabilità di osservare un particolare pattern di risposte. Il modello presenta l'ipotesi di indipendenza locale, le osservazioni y_{ijk} vengono ipotizzate indipendenti, condizionatamente all'appartenenza ad una determinata classe latente.

Le assunzioni sulla distribuzione dei parametri rendono questo modello multilivello:

$$P(X_{ij} = t) = \frac{\exp\{\gamma_{tj}\}}{\sum_{r=1}^{T} \exp\{\gamma_{rj}\}} \quad (5)$$

$$P(y_{ijk} = s_k \mid X_{ij} = t) = \frac{\exp\{\gamma^k_{s_k}\}}{\sum_{r=1}^{S_k} \exp\{\gamma^k_{rtj}\}} \quad (6)$$

La variabilità dei parametri del modello emerge dalla comparsa dell'indice *j* nelle specificazioni (5) e (6).

Questi modelli assumo la possibilità che alcuni parametri varino tra gruppi, diversamente dai modelli tradizionali CL che invece ipotizzano la stabilità degli individui, e quindi dei parametri. La variabilità viene considerata assumendo o un approccio *fixed-effects* (Clogg e Goodman, 1984), introducendo variabili dummy di gruppo nel modello oppure usando l'approccio *random-effects*, che prevede che i coefficienti specifici di gruppo seguano una particolare distribuzione, i cui parametri devono essere stimati.

2.4.1 Fixed-Effect Approach

Questo approccio introduce una specificazione al modello (4), per mostrare il legame del modello stesso con la variabile di secondo livello:

$$P(Y_{ij} \mid G = j) = \sum_{t=1}^{T} P(X_{ij} = t \mid G = j) \prod_{k=1}^{K} P(y_{ijk} \mid X_{ij} = t, G = j) \quad (7)$$

G identifica la variabile di secondo livello ed identifica il gruppo di appartenenza delle unità statistiche.

È possibile porre delle restrizioni al modello fissando i parametri della probabilità condizionata nei gruppi.

Nel modello il numero dei parametri da stimare è pari al numero dei gruppi, questo comporta la complessità del modello nel caso di un elevato numero di gruppi ($J \geq 50$) e un esiguo numero di individui per gruppo ($n_j \leq 30$), complessità dovuta soprattutto all'emergere di stime instabili. Un altro svantaggio è dovuto all'impossibilità di determinare gli effetti delle covariate dei gruppi nella probabilità di appartenenza alla classe latente, in quanto le differenze tra i gruppi sono spiegate da dummy di gruppo.

2.4.2 Random-Effect Approach

Un modo per ovviare ai problemi dei modelli a effetti fissi è ipotizzare che gli effetti dei gruppi seguano una specifica distribuzione, questo è l'approccio ad effetti casuali. In questi modelli l'ipotesi base è la correlazione tra le osservazioni dei gruppi, i membri di uno stesso gruppo tendono ad appartenere alla stessa classe latente.

Nel modello si introduce una variabile latente continua al secondo livello, con uno o più effetti casuali a livello di gruppo, o discreta, i cui parametri variano tra le diverse classi latenti dei gruppi. Nel caso di aggiunta di una variabile continua il modello è di tipo parametrico, la variabile latente è assunta con distribuzione Normale, mentre con una variabile discreta siamo di fronte ad un modello non parametrico, la variabile avrà distribuzione Multinomiale.

Nell'approccio parametrico, il predittore, che emerge dalla probabilità condizionata (5) ha la seguente forma:

$$\gamma_{tj} = \gamma_t + \tau_t u_j \qquad \text{con } u_j \sim N(0,1)$$

Questo predittore ha due restrizioni a zero, una per τ_t e una per y_t.

L'assunzione base è la perfetta correlazione tra i componenti casuali y_{tj}, in specifico, il medesimo effetto casuale u_j viene riscalato in modo diverso per ogni t grazie al parametro ignoto τ_t. In questo modo ogni categoria nominale si suppone correlata ad una tendenza di risposta latente sottostante.

Nell'approccio parametrico le assunzioni sono molto forti sul modello, per questo una possibile alternativa è l'uso di distribuzioni discrete non specifiche, sia a livello uno che due. Definendo ora M come il numero delle classi latenti della variabile latente di secondo livello D, l'approccio non parametrico prevede che ad ognuna di esse appartenga un determinato gruppo. Indichiamo con D_j la classe di appartenenza del gruppo j e m una particolare classe latente, $1 \le D_j = m \le M$ (Vermunt, 2003b). Data una combinazione di componenti di primo e secondo livello otteniamo il modello a effetti casuali discreti:

(8)

$$P(Y_{ij} = s) = \sum_{m=1}^{M} \left(P(D_j = m) \prod_{i=1}^{n_j} \left(\sum_{t=1}^{T} P(X_{ij} = t \mid D_j = m) \prod_{k=1}^{K} P(y_{ijk} = s_k \mid X_{ij} = t) \right) \right)$$

In cui a livello di individui abbiamo:

$$P(Y_{ij} \mid D_j = m) = \sum_{t=1}^{T} P(X_{ij} = t \mid D_j = m) \prod_{k=1}^{K} P(y_{ijk} \mid X_{ij} = t, D_j = m) =$$

$$= \sum_{t=1}^{T} P(X_{ij} = t \mid D_j = m) \prod_{k=1}^{K} P(y_{ijk}, \vartheta_{ktm})$$

Essa descrive la probabilità di risposta Y_{ij}, condizionata all'appartenenza del gruppo j alla classe latente m. I diversi gruppi vengono distinti in base alla probabilità dei loro componenti di appartenere alla *t-esima* classe latente e in base ai parametri che definiscono le probabilità di risposta condizionate.

A livello di gruppo invece abbiamo:

$$P(Y_j) = \sum_{m=1}^{M} P(D_j = m) \prod_{i=1}^{n_j} P(y_{ijk} \mid D_j = m)$$

Le n_j risposte degli individui sono mutuamente indipendenti, condizionatamente al fatto di appartenere alla classe latente m del gruppo j.

Nel modello (8) troviamo la probabilità che un gruppo j appartenga a una classe latente m, la probabilità che un individuo i appartenga alla classe latente t, data l'appartenenza del gruppo alla classe m, e la probabilità che un consumatore di una determinata risposta y_{ijk}, data la sua appartenenza alla classe latente t. In questo modo la probabilità di osservare una risposta non è altro che una media pesata, in cui l'appartenenza ad una classe latente dei gruppi e degli individui sono i pesi usati.

I predittori sono descritti nel modo seguente:

$$P(X_{ij} = t \mid D_j = m) = \frac{\exp\{\gamma_{tm}\}}{\sum_{r=1}^{T} \exp\{\gamma_m\}}$$

in cui è possibile indicare $\gamma_{tm} = \gamma_t + u_{tm}$, e u_{tm} è distribuita in modo non specifico.

La struttura gerarchica che ne emerge è a tre livelli: individui, casi e risposte multiple; a differenza dell'analisi tradizionale che prevede due livelli (risposte multiple e individui).

Sia i modelli parametrici che non prevedono l'aggiunta di covariate. Nel modello parametrico con covariate vi è sia la presenza di effetti fissi (covariate nel primo livello) che casuali (secondo livello).

CAPITOLO 3

3.1 Presentazione del Lavoro

In questo lavoro si segmenta il mercato farmaceutico italiano utilizzando il modello gerarchico di data set a tre vie (Vermut, 2007) e verranno messi a confronto due diversi criteri di adattamento. Le due procedure di valutazione che verranno comparate sono quella utilizzata da Bassi (2007) e quella di Lukočienė, Varriale e Vermut (2010).

Il mercato in analisi, quarto in Europa, è molto competitivo, le attività di promozione e vendita sono molto costose e i budget che le aziende hanno a disposizione sono limitati, quindi è molto importante suddividere il mercato in gruppi con caratteristiche simili in modo da poter raggiungere tutti i segmenti, con le loro specifiche caratteristiche, in modo adeguato e con il minor dispendio di risorse. Diviene, quindi, importante individuare i fattori che influenzano i medici al momento della prescrizione dei farmaci, per poter così adottare appropriate strategie.

Le industrie farmaceutiche intendono capire le aspettative dei medici nei confronti dei loro prodotti e dei loro rappresentanti, per acquisire quote di mercato e dirigere gli investimenti per non sprecare risorse.

In questo lavoro è stato utilizzato un adattamento del modello CL standard, il modello gerarchico, in quanto viene violato l'assunto di indipendenza delle osservazioni (ogni industria farmaceutica può essere valutata da più medici). Il modello prevede di raggruppare i casi, medici, in base alla probabilità di appartenere ad una determinata classe latente. I medici sono stati aggregati in base ad atteggiamenti simili verso il lavoro dei rappresentanti farmaceutici e

verificando se vi sono differenze tra le unità per quanto riguarda l'importanza data ai servizi offerti dalle aziende.

In entrambi i tipi di analisi, sia nel caso dell'utilizzo del classico metodo di valutazione dell'adattamento che nel caso di utilizzo del metodo a tre passi, si considerano per la scelta del modello sia il criterio informativo Bayesiano, BIC (usato da Bassi, 2007) che il criterio informativo aggiustato di Akaike, AIC3 (Bozdogan, 1993). Questa scelta è dovuta al fatto che nell'articolo di Lukočienė, Varriale e Vermut (2010), il criterio AIC3 viene indicato come il miglior criterio per la scelta del numero di classi latenti in modelli CL multilivello, in base a studi di simulazione di Andrei e Currim (2003) e Dias (2004). Verranno utilizzati entrambi i criteri, BIC e AIC3, verificando la tesi degli autori sul nostro insieme di dati.

Il lavoro di Bassi incentra l'analisi nel modello in cui non sono presenti covariate, in quanto sono risultate non significative, per questo motivo anche in questo lavoro non saranno introdotte.

3.1.1 Dati

L'analisi sarà eseguita partendo dal data set usato da Bassi (Bassi, 2007). I dati riguardano un sondaggio di medicina generale italiana, i medici sono invitati a giudicare alcuni aspetti delle strategie di promozione delle industrie farmaceutiche.

I dati che si utilizzano nell'analisi sono stati raccolti presso 489 medici italiani di medicina generale. Gli intervistati dovevano rispondere ad un questionario su una scala a 7 punti, riguardante l'importanza di determinati aspetti per portarli a scegliere un farmaco proposto da un'azienda farmaceutica, al momento della prescrizione dello stesso.

Gli elementi giudicati dai medici sono:

a) ATT: attenzione dell'industria all'aggiornamento dei medici;

b) FRE: frequenza e regolarità delle visite dei rappresentanti delle case farmaceutiche;
c) ASS: assistenza sui problemi diagnostici e terapeutici;
d) EXP: considerazioni dei medici sull'esperienza e sui suggerimenti dati;
e) QUA: qualità della formazione dei rappresentanti farmaceutici;
f) INF: informazione sulle attività dell'industria;
g) PRO: qualità complessiva delle attività d'informazione e di promozione.

Il numero totale di giudizi a disposizione è 2537 poiché ogni medico giudica più di una casa farmaceutica. I dati sono stati raggruppati in un data set a tre vie.

3.1.2 Analisi

Questo lavoro si può dividere i due parti. La prima analisi ripropone lo studio effettuato da Bassi (2007), estendendo la procedura fino ad un massimo di 7 classi, sia per le unità di primo che di secondo livello. Nella seconda parte verrà utilizzata la procedura a tre passi introdotta da di Lukočienė, Varriale e Vermut (2010).
Entrambe le metodologie utilizzate per la valutazione dell'adattamento, prevedono l'utilizzo del modello multilivello.

Nello studio di Lukočienė, Varriale e Vermut (2010) si indaga la performance dei criteri d'informazione, in particolare, BIC e AIC3. Il modello scelto per entrambi gli indici sarà quello che presenterà il valore più basso. In entrambe le analisi, la procedura si arresta quando il valore dell'indice riprende a salire.

Per le analisi si è utilizzato il programma Latent Gold 4.0 (Vermut e Magidson, 2005).

3.2 Modello a Classi Latenti Multilivello: Metodo Classico

In questa analisi verranno individuati cluster di medici sulla base delle risposte che sono state fornite in merito a varie marche.
La prima parte prevede la scelta del modello secondo il criterio BIC, nella seconda parte verrà invece utilizzato il criterio AIC3.

Il numero di classi latenti che meglio si adatta ai dati viene valutato simultaneamente per le unità di primo e secondo livello.

3.2.1 Criterio d'Informazione BIC

Iniziamo l'analisi dei dati cercando il modello migliore in grado di spiegarli, utilizzando il criterio d'informazione BIC. Per ricavare il modello ottimo si è partiti fissando il primo livello del modello multilivello e aumentando gradualmente il numero di classi di livello superiore fino ad un massimo di 7. L'analisi continua aumentando in modo graduale anche le classi di primo livello. Il modello ottimo è il primo che ha indice BIC minore.

La *Tabella 3.2.1* riporta i principali valori che si sono ottenuti in modo schematico.

Tabella 3.2.1: stime del modello per varie classi e cluster

CLUSTER	CLASSI	BIC	Log-likelihood	Number of parameters
1	1	57060,7573	-28365,7651	42
1	2	57068,5960	-28365,7651	43
1	3	57076,4347	-28365,7651	44
1	4	57084,2735	-28365,7651	45
1	5	57092,1122	-28365,7651	46
1	6	57099,9510	-28365,7651	47
1	7	57107,7897	-28365,7651	48
2	1	52821,2453	-26214,6542	50
2	2	52685,2059	-26138,7958	52
2	3	52690,6082	-26133,6582	54
2	4	52706,2346	-26133,6327	56
2	5	52721,8399	-26133,5966	58
2	6	52737,4720	-26133,5739	60
2	7	52753,1174	-26133,5578	62
3	1	51542,0400	-25543,6966	58
3	2	51391,4869	-25456,6619	61
3	3	51321,9671	-25410,1440	64
3	4	51324,4686	-25399,6366	67
4	1	51125,6330	-25304,1381	66
4	2	50987,3526	-25219,3205	70
4	3	50901,9109	-25160,9222	74
4	4	50886,0778	-25137,3282	78
4	5	50901,6386	-25129,4311	82
5	1	51059,5201	-25239,7268	74
5	2	50821,0704	-25100,9051	79
5	3	50764,7154	-25053,1307	84
5	4	50695,3987	-24998,8755	89
5	5	50696,4324	-24979,7955	94
6	1	50996,2855	-25176,7545	82
6	2	50764,4592	-25037,3252	88
6	3	50585,3200	-24924,2393	94
6	4	50626,4347	-24921,2805	100
7	1	50963,3565	-25128,9351	90
7	2	50709,2844	-24974,4634	97
7	3	50555,8111	-24870,2912	104
7	4	50503,9934	-24816,9468	111
7	***5***	***50435,5526***	***-24755,2908***	***118***
7	6	50452,1694	-24736,1636	125

Dalla *Tabella 3.2.1* il modello scelto in base al BIC è quello con 7 cluster di giudizi sul lavoro dei rappresentanti e 5 gruppi di medici. Questo modello ha valore BIC uguale a *50435,5526* mentre nel modello successivo vi è un aumento dell'indice, infatti esso è pari a *50452,1694*.

Per poter accettare il modello, dobbiamo però osservare se esso può rappresentare in modo realistico la situazione considerata

nell'analisi, ossia se rappresenta il mercato farmaceutico italiano. La composizione di questo modello non è adatta alla situazione considerata, infatti, presenta dei cluster molto piccoli (Cluster 7, 4%) e ciò non verifica una delle proprietà dei segmenti di mercato: la dimensione.

Profile 7 Clusters- 5 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7
0,2533	0,2169	0,1998	0,1285	0,0947	0,0685	0,0382

Il successivo modello che, in base all'indice BIC, può ben rappresentare i dati a nostra disposizione, è il modello con 6 cluster e 3 classi (BIC=50585,3200); anch'esso dà però dei cluster di dimensione trascurabili (Cluster 6, 5%).

Profile 6 Clusters- 3 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
0,3704	0,2481	0,1439	0,1000	0,0889	0,0487

Il modello migliore, determinato tramite l'indice BIC, è dato dalla combinazione 5-4 (BIC=50695,3987), ossia 5 classi di livello inferiore e 4 di livello superiore. Le stime di questo modello mostrano che la dimensione dei cluster e delle classi formatasi è accettabile, è maggiore del 10% in tutti i casi.

Profile 5 Clusters- 4 Class

CLUSTER:

Cluster1(s.e.)	Cluster2(s.e.)	Cluster3(s.e.)	Cluster4(s.e.)	Cluster5(s.e.)
0,3261(0,0171)	0,2851(0,0185)	0,1477(0,0173)	0,1254(0,0112)	0,1157(0,0111)

CLASS:

Class1(s.e.)	Class2(s.e.)	Class3(s.e.)	Class4(s.e.)
0,3368(0,0421)	0,2621(0,0396)	0,2194(0,0296)	0,1817(0,0290)

- Analisi del Modello

Il modello multilivello scelto nell'analisi, che è formato da 5 classi di livello inferiore e 4 di livello superiore, presenta un mercato segmentato.

Tabella 3.2.2: modello multilivello LC

	Class1(s.e.)	**Class2(s.e.)**	**Class3(s.e.)**	**Class4(s.e.)**
SIZE	0,3368(0,0421)	0,2621(0,0396)	0,2194(0,0296)	0,1817(0,0290)
Clusters:				
Cluster1	0,6176(0,0361)	0,1168(0,0308)	0,2229(0,0356)	0,2123(0,0335)
Cluster2	0,2550(0,0392)	0,6078(0,0319)	0,0006(0,0026)	0,2191(0,0336)
Cluster3	0,0427(0,0209)	0,0003(0,0014)	0,5881(0,0348)	0,0229(0,0201)
Cluster4	0,0833(0,0157)	0,2575(0,0326)	0,0606(0,0141)	0,0910(0,0207)
Cluster5	0,0013(0,0056)	0,0175(0,0101)	0,1278(0,0218)	0,4547(0,0376)

Dalla *Tabella 3.2.2*, è possibile associare i cluster ai gruppi di medici. Il primo gruppo di medici, Class 1, è legato al primo cluster, Cluster 1 (*0,6176*), Class 2 è legato a Cluster 2 (*0,6078*), Class 3 è in relazione con Cluster 3 (*0,5881*) e infine Class 4 è legato a Cluster 5 (*0,4547*).

La *Tabella 3.2.3* presenta le caratteristiche di ogni gruppo, ogni colonna mostra i giudizi medi dati ad ogni caratteristica analizzata, per un singolo gruppo. La Class 3 (22%) raggruppa i medici che sono molto interessati a tutti gli aspetti tranne per quanto riguarda l'elemento informazione sulle aziende, infatti presenta valori molto vicini a 7 (valore massimo) ma assume valore *3,3607* nella variabile *INF*. I medici meno interessati sono quelli che sono presenti nella Class 4 (18%), infatti in questo gruppo si trovano i medici che hanno dato i giudizi più bassi. Alla Class 2 (26%) appartengono i medici che hanno dato giudizi medio alti e nella Class 1 (34%) ci sono quelli che

hanno dato giudizi medio bassi dato che il punteggio medio è attorno al valore 4.

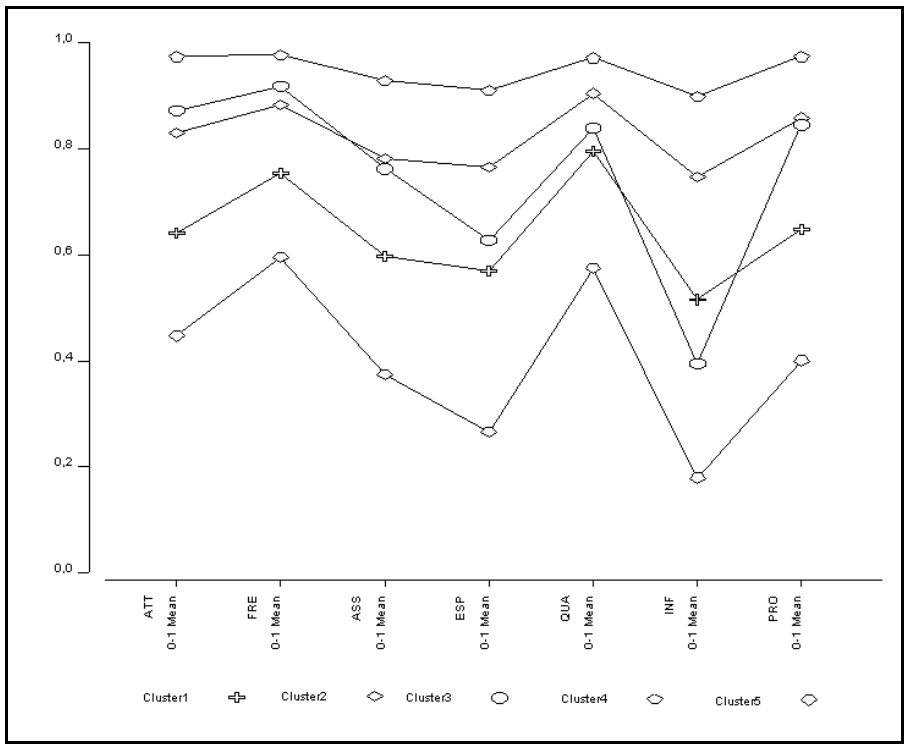
Da notare che il cluster 4 presenta giudizi ottimi per tutti gli aspetti (valori tutti superiori a 6) ma non è messo in relazione con nessuna classe di medici.

Tabella 3.2.3: stime dei risultati e standard error

Cluster1(s.e.)	**Cluster2**(s.e.)	**Cluster3**(s.e.)	**Cluster4**(s.e.)	**Cluster5**(s.e.)
SIZE				
0,3261(0,0171)	0,2851(0,0185)	0,1477(0,0173)	0,1254(0,0112)	0,1157(0,0111)
Indicators:				
ATT				
4,8336(0,0498)	5,9724(0,0491)	6,2229(0,0575)	6,8293(0,0318)	3,6691(0,0914)
FRE				
5,5131(0,0476)	6,2877(0,0428)	6,4976(0,0468)	6,8536(0,0276)	4,5562(0,1014)
ASS				
4,5675(0,0520)	5,6749(0,0506)	5,5678(0,0694)	6,5701(0,0486)	3,2374(0,0924)
ESP				
4,4130(0,0611)	5,5819(0,0578)	4,7533(0,0940)	6,4532(0,0571)	2,5907(0,1008)
QUA				
5,7603(0,0417)	6,4206(0,0388)	6,0296(0,0570)	6,8259(0,0275)	4,4309(0,0943)
INF				
4,0922(0,0790)	5,4714(0,0631)	3,3607(0,1518)	6,3840(0,0602)	2,0710(0,0894)
PRO				
4,8833(0,0463)	6,1393(0,0386)	6,0577(0,0492)	6,8270(0,0312)	3,3986(0,0853)

Dal ‘profile plot’ (*Figura 3.2*) possiamo notare come sono divisi i cluster per quanto riguarda i giudizi sui vari aspetti analizzati. Il grafico mostra come gli aspetti di frequenza delle visite e la qualità della formazione dei rappresentanti siano molto importanti per tutti i medici. L’aspetto meno importante risulta essere quello legato all’informazione sulle attività dell’azienda.

Figura 3.2: Profile Plot



3.2.2 Criterio d'informazione AIC3

Ripetiamo ora l'analisi utilizzando l'indice AIC3 per determinare il miglior modello. Come per l'analisi precedente il modello che meglio descrive i dati viene determinato dal valore minore.

La *Tabella 3.2.4* presenta i risultati principali dei vari modelli analizzati.

Tabella 3.2.4: stime del modello per varie classi e cluster

CLUSTER	CLASSI	AIC3	Log-likelihood	Number of parameters
1	1	56857,5303	-28365,7651	42
1	2	56860,5303	-28365,7651	43
1	3	56863,5303	-28365,7651	44
1	4	56866,5303	-28365,7651	45
1	5	56869,5303	-28365,7651	46
1	6	56872,5303	-28365,7651	47
1	7	56875,5303	-28365,7651	48
2	1	52579,3085	-26214,6542	50
2	2	52433,5915	-26138,7958	52
2	3	52429,3163	-26133,6582	54
2	4	52435,2653	-26133,6327	56
2	5	52441,1932	-26133,5966	58
2	6	52447,1478	-26133,5739	60
2	7	52453,1157	-26133,5578	62
3	1	51261,3932	-25543,6966	58
3	2	51096,3239	-25456,6619	61
3	3	51012,2879	-25410,1440	64
3	4	51000,2731	-25399,6366	67
3	5	50997,4537	-25393,7269	70
3	6	50997,6497	-25389,3249	73
3	7	51006,6529	-25389,3264	76
4	1	50806,2763	-25304,1381	66
4	2	50648,6410	-25219,3205	70
4	3	50543,8444	-25160,9222	74
4	4	50508,6563	-25137,3282	78
4	5	50504,8621	-25129,4311	82
4	6	50510,3723	-25126,1861	86
5	1	50701,4535	-25239,7268	74
5	2	50438,8102	-25100,9051	79
5	3	50358,2615	-25053,1307	84
5	4	50264,7511	-24998,8755	89
5	5	50241,5910	-24979,7955	94
5	6	50242,3517	-24972,6758	99
6	1	50599,5090	-25176,7545	82
6	2	50338,6503	-25037,3252	88
6	3	50130,4786	-24924,2393	94
6	4	50142,5609	-24921,2805	100
7	1	50527,8701	-25128,9351	90
7	2	50239,9269	-24974,4634	97
7	3	50052,5824	-24870,2912	104
7	4	49966,8935	-24816,9468	111
7	5	49864,5816	-24755,2908	118
7	6	49847,3272	-24736,1636	125
7	***7***	***49841,8927***	***-24722,9464***	***132***

Il più basso valore AIC3 è assunto dalla combinazione 7-7 (AIC3=49841,8927), essa è però composta da troppe classi/cluster, nella realtà considerare i bisogni e le caratteristiche di sette gruppi di medici diviene troppo dispendioso per l'azienda. Analizzando la

struttura del modello, notiamo che è formato da cluster di dimensioni inferiori all'8% (Cluster 5, Cluster 6, Cluster 7). Con classi troppo piccole vengono meno le caratteristiche di consistenza e dimensione dei segmenti.

Profile 7 Clusters- 7 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7
0,2626	0,2272	0,1721	0,1026	0,0807	0,0806	0,0741

Le stesse considerazioni emergono nel modello multilivello a 6 cluster di primo livello e 3 di secondo, e nella combinazione 5-5. Entrambi i modelli non rispettano le proprietà di dimensione dei segmenti di mercato.

Profile 6 Clusters- 3 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
0,3704	0,2481	0,1439	0,1000	0,0889	0,0487

Profile 5 Clusters- 5 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
0,3928	0,3075	0,1509	0,0986	0,0502

Il modello che presenta le caratteristiche adatte è quello con 4 cluster di giudizi e 5 classi di medici. La dimensione è, per tutti i cluster, superiore al 10%.

Profile 4 Clusters- 5 Class

CLUSTER:

Cluster1(s.e.)	Cluster2(s.e.)	Cluster3(s.e.)	Cluster4(s.e.)
0,3958(0,0173)	0,3350(0,0191)	0,1562(0,0151)	0,1131(0,0112)

Oltre ad osservare la dimensione dei cluster, è necessario verificare se anche le classi sono di dimensione adeguata. Nel modello appena scelto notiamo che è presente un gruppo formato dal 7% di medici (Class 5). Questo gruppo è troppo piccolo, soprattutto se confrontato con la classe che contiene la maggioranza dei medici (41%). Il modello non permette di soddisfare la proprietà di dimensione dei segmenti, la classe con 7%, essendo molto piccola è probabile che non sia stabile nel tempo, provocando un dispendio di risorse finanziarie, che con un altro modello potrebbe essere contenuto.

Profile 4 Clusters- 5 Class

CLASSI:

Class1(s.e.)	Class2(s.e.)	Class3(s.e.)	Class4(s.e.)	Class5(s.e.)
0,4136(0,0607)	0,2319(0,0588)	0,1799(0,0469)	0,1035(0,0378)	0,0711(0,0293)

Passiamo, quindi, ad osservare la nuova coppia di cluster/classi che ha il miglior indice AIC3, ossia il più basso, e si trova nella combinazione 4-4:

Profile 4 Clusters- 4 Class

CLUSTER:

Cluster1(s.e.)	Cluster2(s.e.)	Cluster3(s.e.)	Cluster4(s.e.)
0,3974(0,0174)	0,3392(0,0180)	0,1526(0,0142)	0,1109(0,0111)

CLASS:

Class1(s.e.)	**Class2(s.e.)**	**Class3(s.e.)**	**Class4(s.e.)**
0,4567(0,0562)	0,2519(0,0442)	0,1826(0,0450)	0,1089(0,0376)

Come si vede dal *Profile 4 Clusters-4 Classi*, le proprietà di consistenza e dimensione possono ora essere verificate, la diminuzione da 5 a 4 classi di secondo livello ha riequilibrato i gruppi.

-Analisi del Modello

Per aiutarci ad interpretare la soluzione, osserviamo la *Tabella 3.2.5* in cui possiamo vedere le relazioni tra le classi di primo e secondo livello.

Tabella 3.2.5: modello multilivello LC

	Class1(s.e.)	**Class2(s.e.)**	**Class3(s.e.)**	**Class4(s.e.)**
SIZE	0,4567(0,0562)	0,2519(0,0442)	0,1826(0,0450)	0,1089(0,0376)
Clusters:				
Cluster1	0,3107(0,0294)	0,3032(0,0325)	0,8232(0,0592)	0,2646(0,1077)
Cluster2	0,5731(0,0346)	0,2146(0,0373)	0,0751(0,0414)	0,0890(0,0456)
Cluster3	0,1044(0,0179)	0,0868(0,0180)	0,0853(0,0434)	0,6200(0,1171)
Cluster4	0,0118(0,0132)	0,3954(0,0402)	0,0164(0,0137)	0,0265(0,0246)

La Class 1 è legata al Cluster 2 (*0,5731*), la Class 3 è legata al Cluster 1 (*0,8231*), Class 4 con Cluster 3 (*0,6200*), è interessante osservare la Class 2 che risulta essere in relazione sia con Cluster 1 che con Cluster 4 (i valori sono simili, *0,3032* e *0,3954*).

La *Tabella 3.2.6*, invece, identifica le caratteristiche dei vari gruppi di medici che sono stati raggruppati. La Class 4 (11% dei medici) raggruppa i medici molto interessati agli aspetti in considerazione, infatti hanno dato a tutti gli aspetti dei giudizi attorno al valore massimo (valori tutti maggiori a 6). La Class 3 (18%) raggruppa medici che sono abbastanza interessati a tutti gli aspetti, mentre la Class 1 raggruppa quasi la metà dei medici che ha partecipato al sondaggio, essi risultano essere mediamente interessati agli aspetti considerati, tranne per l'elemento che riguarda l'informazione che ha ottenuto il punteggio minore (*3,9943*). La Class 2 (25%) si suddivide in due fazioni, l'una che non sembra

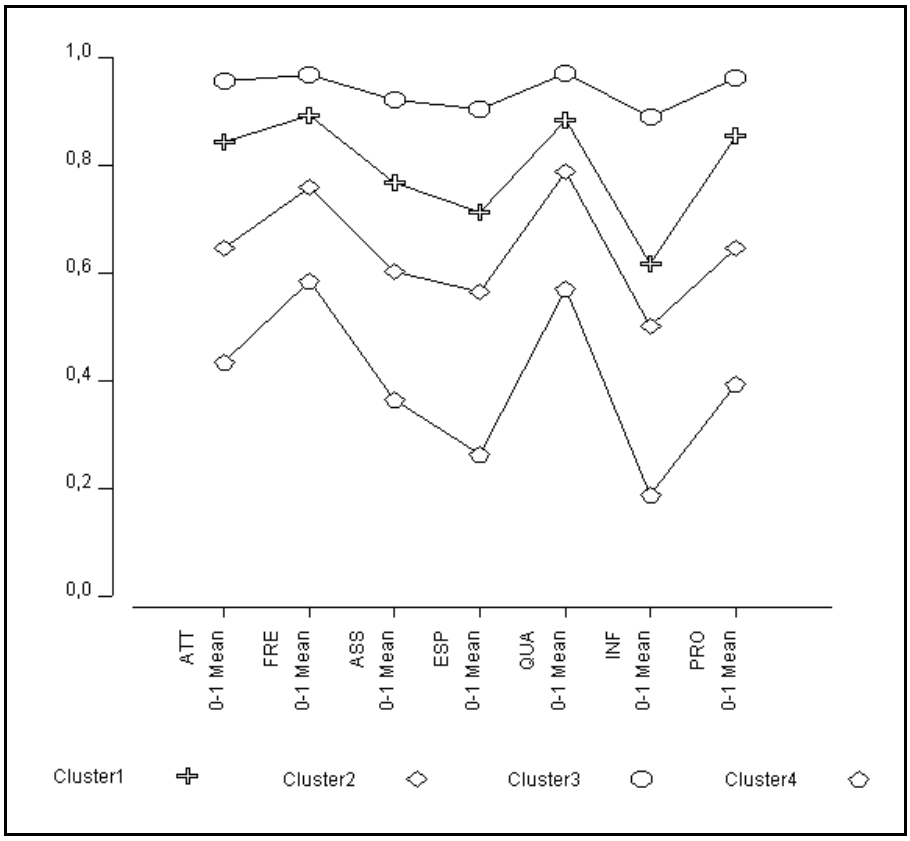
essere interessata agli aspetti considerati e un'altra che invece dà importanza a tutti.

Tabella 3.2.6: stime dei risultati e standard error

	Cluster1(s.e.)	Cluster2(s.e.)	Cluster3(s.e.)	Cluster4(s.e.)
SIZE	0,3974(0,0174)	0,3392(0,0180)	0,1526(0,0142)	0,1109(0,0111)
Indicators:				
ATT	6,0508(0,0370)	4,8734(0,0536)	6,7374(0,0352)	3,5998(0,0945)
FRE	6,3415(0,0318)	5,5552(0,0479)	6,7926(0,0306)	4,5093(0,1034)
ASS	5,5921(0,0421)	4,6082(0,0531)	6,5131(0,0458)	3,1826(0,0965)
ESP	5,2615(0,0514)	4,3842(0,0610)	6,4113(0,0524)	2,5688(0,1028)
QUA	6,2912(0,0322)	5,7185(0,0429)	6,8118(0,0273)	4,4104(0,0969)
INF	4,6986(0,0623)	3,9943(0,0717)	6,3291(0,0638)	2,1246(0,0980)
PRO	6,1197(0,0334)	4,8740(0,0496)	6,7726(0,0294)	3,3530(0,0861)

Il grafico (*Figura 3.3*) mostra come gli aspetti più importanti siano di nuovo la frequenza e regolarità della visita e la formazione dei rappresentanti farmaceutici mentre non si dà importanza all'aspetto riguardante l'informazione sulle aziende, che ha ottenuto in tutti i cluster i giudizi più bassi. Dal grafico è molto visibile la netta separazione in 4 gruppi dei giudizi, come è emerso dall'analisi della *Tabella 3.2.6*. Dal grafico vediamo che il Cluster 3 ha i valori più alti in riferimento all'attività di promozione, il Cluster 4 contiene i giudizi più bassi, mentre gli altri due cluster contengono i giudizi medi.

Figura 3.3:Profile Plot



3.2.3 Considerazioni

Dall'analisi con il modello CL multilivello emerge che l'utilizzo di un determinato criterio informativo incide nella scelta del modello finale. Dalle analisi con l'utilizzo dell'indice BIC per la valutazione dell'adattamento, il modello scelto ha 5 classi di primo livello e 4 di secondo, mentre nell'analisi con il criterio AIC3 il modello risultante ha 4 classi di livello inferiore e superiore.

In entrambi i modelli i gruppi di medici hanno delle caratteristiche specifiche che li distingue ma tutti danno molta importanza alla frequenza e alla regolarità delle visite dei rappresentanti e poca importanza all'aspetto riguardante l'informazione sulle attività delle imprese.

Il motivo per cui non vi è corrispondenza tra i modelli CL in cui si è usato il criterio AIC3 e BIC può essere legato al fatto che la scelta del modello, nel caso dell'indice AIC3, anche se basata sulla soddisfazione delle proprietà dei segmenti, è soggettiva. Infatti il modello con 5 classi di primo livello e 4 di secondo è stato rigettato, in quanto si è cercato di ottenere delle classi di dimensione maggiore al 10% e non si ritenuto significativamente grande la classe formata dal 7% di medici.

3.3 Modello a Classi Latenti Multilivello: Metodo a 3 Passi

In questa parte verrà utilizzata la procedura di valutazione dell'adattamento introdotta da Lukočienė, Varriale e Vermut (Lukočienė, Varriale e Vermut, 2010). Questo metodo elimina la dipendenza tra le decisioni di scelta tra il numero di livelli dovuta alla struttura gerarchica, infatti nella procedura classica la scelta delle classi di primo livello influenza il numero di classi di secondo. La scelta del numero di livelli del modello non è mutuamente indipendente, la decisione del livello superiore dipende dalle decisioni di livello inferiore, i modelli CL classici ignorano questa dipendenza.

Questo nuovo metodo, prevede di scegliere il modello seguendo tre passi nell'analisi. Il primo passo prevede di ignorare la struttura multilivello del modello, fissando ad 1 il numero di classi di secondo livello, e determinare il numero ottimo di classi di primo livello in base al criterio d'informazione. Il secondo passo consiste nel fissare il numero di classi di primo livello a quello trovato nel punto precedente e determinare il numero di classi di livello superiore. Nell'ultimo passo si fisserà il numero di classi di livello superiore al valore del passo due e si determinerà nuovamente il numero di quelle di livello inferiore, verificando se c'è una variazione nel numero scelto in precedenza. In questo modo si tiene conto della dipendenza tra le unità di livello inferiore.

Per la determinazione del modello migliore, si utilizzerà in un caso l'indice BIC e in un altro l'indice AIC3.

3.3.1 Criterio d'Informazione BIC

Iniziamo l'analisi determinando il miglior modello con l'utilizzo del criterio d'informazione BIC.

Il metodo a 3 passi prevede di scegliere il modello in tre fasi. Inizialmente si partirà scegliendo l'appropriato numero di classi di primo livello senza tener conto della struttura a più livelli. In questo primo passo, per non considerare la struttura multilivello, verrà fissato a 1 il numero di classi di secondo livello e si procede aumentando quelle di livello inferiore fino ad arrivare ad un massimo di 7 cluster.

- **Primo Passo**

Tabella 3.3.1: stime del modello per varie classi e cluster

CLUSTER	CLASSI	BIC	Log-likelihood	Number of parameters
1	1	57060,7573	-28365,7651	42
2	1	52821,2453	-26214,6542	50
3	1	51542,0400	-25543,6966	58
4	1	51125,6330	-25304,1381	66
5	1	51059,5201	-25239,7268	74
6	1	51012,1423	-25184,6829	82
7	***1***	***50963,3565***	***-25128,9351***	***90***

I risultati della *Tabella 3.3.1* mostrano che il numero di classi di primo livello da scegliere è 7, essendo la soluzione che dà l'indice BIC con valore minore (50963,3565).

Osservando i risultati che emergono con questa soluzione notiamo però che il suo profilo risulta essere problematico. Le classi risultano essere di dimensione inferiore al 4% per due classi (Cluster 6 e Cluster 7), la soluzione va quindi scartata perché non verificherebbe una delle condizioni più importanti nella strategia di

marketing, la dimensione della classi, esse devono rappresentare il mercato reale e garantire la profittabilità dei programmi di marketing.

Profile 7 Clusters- 1 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7
0,3583	0,2883	0,1161	0,1159	0,0730	0,0358	0,0127

Scartando questa soluzione andiamo a ritroso per trovare una nuova soluzione, che dia un valore BIC minore tra le rimanenti.

La scelta ricade nel modello con 6 classi di livello inferiore. Anch'esso presenta le stesse problematiche della soluzione precedente (la dimensione dei Cluster 5 e Cluster 6 è esigua). Andando ancora a ritroso ci fermiamo nella scelta di 5 classi di primo livello che ci porta nuovamente a scartare la soluzione.

Qui di seguito sono riportati i profili per 6 e 5 classi di primo livello.

Profile 6 Clusters- 1 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
0,3747	0,2884	0,1753	0,0976	0,0479	0,0160

Profile 5 Clusters- 1 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
0,3869	0,3522	0,1252	0,1238	0,0118

La soluzione finale sta nella scelta di 4 classi di livello inferiore, essa infatti presenta i profili migliori delle classi, tutte di dimensione adeguata al mercato che devono rappresentare.

Profile 4 Clusters- 1 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4
0,3941	0,3564	0,1271	0,1224

Passiamo al secondo passo dell'analisi fissando a 4 le classi di primo livello.

- **Secondo Passo**

Il secondo passo prevede di fissare il numero di classi di livello inferiore a quello identificato al passo precedente e aumentare il numero di quelle di secondo livello, sempre fino ad un massimo di 7.

In questo caso fissiamo a 4 le classi di primo livello.

Tabella 3.3.2: stime del modello per varie classi e cluster

CLUSTER	CLASSI	BIC	Log-likelihood	Number of parameters
4	1	51125,6330	-25304,1381	66
4	2	50987,3526	-25219,3205	70
4	3	50901,9109	-25160,9222	74
4	***4***	***50886,0778***	***-25137,3282***	***78***
4	5	50901,2550	-25129,2393	82

In questo passo (*Tabella 3.3.2*) fermiamo l'analisi a 4 classi di primo livello e 5 classi di secondo, in quanto nel passaggio da 4 a 5 classi di livello inferiore c'è un aumento del valore dell'indice BIC (l'indice passa da 50886,0778 a 50901,2550). La soluzione è quindi data da 4 cluster e 4 classi.

Come per il primo passo, verifichiamo che la soluzione rispetti la proprietà riguardante la dimensione delle classi.

Profile 4 Clusters- 4class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4
0,3974	0,3392	0,1526	0,1109

Tutte le classi sono di dimensione superiore al 10%, passiamo al terzo e ultimo passo fissando a 4 le classi di secondo livello.

- Terzo Passo

L'ultimo passo di questo metodo consiste nel fissare il valore identificato al passo due, per le classi di livello superiore, e ripetere l'analisi variando il numero di quelle di livello inferiore, sempre fino ad avere un massimo di 7 classi.
Fissiamo quindi a 4 le classi di livello superiore e determiniamo il giusto numero di cluster.

Tabella 3.3.3: stime del modello per varie classi e cluster

CLUSTER	CLASSI	BIC	Log-likelihood	Number of parameters
1	4	57084,2735	-28365,7651	45
2	4	52706,2346	-26133,6327	56
3	4	51324,4686	-25399,6366	67
4	4	50886,0778	-25137,3282	78
5	4	50695,3987	-24998,8755	89
6	***4***	***50543,9357***	***-24880,0310***	***100***
7	4	50593,6491	-24861,7746	111

Da quanto emerge dai risultati (*Tabella 3.3.3*), la scelta ricade nella combinazione di 6 classi di primo livello e 4 classi di secondo livello.

La soluzione però presenta una classe di dimensione inferiore al 5% (Cluster 6), per questo la soluzione viene rigettata e si passa a individuare un nuovo modello che presenti un indice BIC minore degli altri.

Profile 6 Clusters- 4 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
0,3653	0,2637	0,1445	0,0976	0,0818	0,0471

Andando a ritroso la scelta ricade nella combinazione tra cluster e classi 5-4. Analizzando il profilo di questo modello vediamo che il problema della soluzione precedente viene a meno, la dimensione delle classi è ora maggiore del 10%, la proprietà dei segmenti è verificata.

Profile 5 Clusters- 4 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
0,3261	0,2851	0,1477	0,1254	0,1157

CLASSI:

Class1(s.e.)	Class2(s.e.)	Class3(s.e.)	Class4(s.e.)
0,3368(0,0421)	0,2621(0,0396)	0,2194(0,0296)	0,1817(0,0290)

- **Analisi del Modello**

Come per l'analisi standard, anche in questa procedura osserviamo come è formato il modello scelto.

Tabella 3.3.4: modello multilivello LC

	Class1(s.e.)	**Class2(s.e.)**	**Class3(s.e.)**	**Class4(s.e.)**
SIZE	0,3368(0,0421)	0,2621(0,0396)	0,2194(0,0296)	0,1817(0,0290)
Clusters:				
Cluster1	0,6176(0,0361)	0,1168(0,0308)	0,2229(0,0356)	0,2123(0,0335)
Cluster2	0,2550(0,0392)	0,6078(0,0319)	0,0006(0,0026)	0,2191(0,0336)
Cluster3	0,0427(0,0209)	0,0003(0,0014)	0,5881(0,0348)	0,0229(0,0201)
Cluster4	0,0833(0,0157)	0,2575(0,0326)	0,0606(0,0141)	0,0910(0,0207)
Cluster5	0,0013(0,0056)	0,0175(0,0101)	0,1278(0,0218)	0,4547(0,0376)

Dalla *Tabella 3.3.4* possiamo vedere le relazione tra le classi di primo livello e quelle di secondo. Le associazioni sono ben definite, infatti Class 1 è associato col Cluster 1 (*0,6176*), Class 2 con Cluster 2 *(0,6078)*, Class 3 con Cluster 3 (*0,5881*) e Class 4 con Cluster 5 (*0,4547*).

Class 4 contiene il 18% dei medici che non danno importanza agli caratteristiche presentate, come si può vedere dalla *Tabella 3.3.5*, infatti il cluster a cui è associato contiene per ogni aspetto i giudizi più bassi. I medici più leali e interessati al lavoro dei rappresentanti, i più difficili da soddisfare, risultano essere quelli appartenenti alla Class 2 (26%). La Class 1 raggruppa la maggioranza dei medici (34%) che hanno dato giudizi medio bassi (tutti attorno a 4). Class 3 (22%) contiene i medici molto esigenti, i giudizi sono molto alti per tutti gli aspetti, tranne per quanto riguarda l'aspetto relativo all'informazione sull'attività delle industrie in esame (*3,3607*).

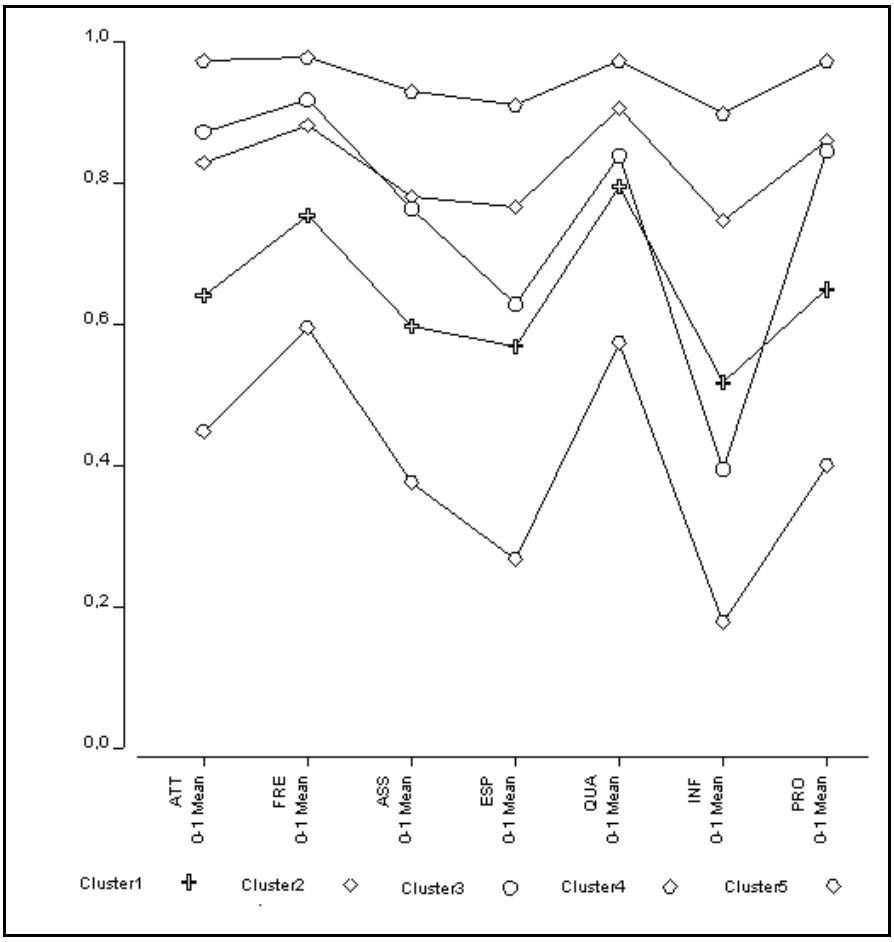
Molto strano risulta essere il Cluster 4 che, sebbene contenga i giudizi più elevati, non è associato ad un specifico gruppo di medici.

Tabella 3.3.5: stime dei risultati e standard error

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
SIZE				
0,3261(0,0171)	0,2851(0,0185)	0,1477(0,0173)	0,1254(0,0112)	0,1157(0,0111)
Indicators:				
ATT				
4,8336(0,0498)	5,9724(0,0491)	6,2229(0,0575)	6,8293(0,0318)	3,6691(0,0914)
FRE				
5,5131(0,0476)	6,2877(0,0428)	6,4976(0,0468)	6,8536(0,0276)	4,5562(0,1014)
ASS				
4,5675(0,0520)	5,6749(0,0506)	5,5678(0,0694)	6,5701(0,0486)	3,2374(0,0924)
ESP				
4,4130(0,0611)	5,5819(0,0578)	4,7533(0,0940)	6,4532(0,0571)	2,5907(0,1008)
QUA				
5,7603(0,0417)	6,4206(0,0388)	6,0296(0,0570)	6,8259(0,0275)	4,4309(0,0943)
INF				
4,0922(0,0790)	5,4714(0,0631)	3,3607(0,1518)	6,3840(0,0602)	2,0710(0,0894)
PRO				
4,8833(0,0463)	6,1393(0,0386)	6,0577(0,0492)	6,8270(0,0312)	3,3986(0,0853)

Dal profile plot (*Figura 3.4*) possiamo notare che gli aspetti che hanno ottenuto giudizi alti, quindi positivi, da parte dei medici sono la frequenza delle visite e la formazione dei rappresentanti farmaceutici, si dà meno importanza alle informazioni riguardanti l'azienda e all'esperienza dei medici. Confrontando i valori ottenuti per i vari cluster, tutti hanno dato a questo aspetto il giudizio più basso.

Figura 3.4:Profile Plot



3.3.2 Criterio d'Informazione AIC3

Ripetiamo l'analisi osservando, in questo caso, l'indice AIC3 che indicherà quale modello scegliere e a che livello fissare le classi ad ogni passo dell'analisi.

Come nel caso precedente il primo passo consiste nel determinare il numero di livelli inferiori non considerando la struttura multilivello.

- **Primo Passo**

Fissiamo a 1 le classi di secondo livello e incrementiamo il numero di classi di primo, fino ad un massimo di 7.

Tabella 3.3.6: stime del modello per varie classi e cluster

CLUSTER	CLASSI	AIC3	Log-likelihood	Number of parameters
1	1	56857,5303	-28365,7651	42
2	1	52579,3085	-26214,6542	50
3	1	51261,3932	-25543,6966	58
4	1	50806,2763	-25304,1381	66
5	1	50701,4535	-25239,7268	74
6	1	50615,3659	-25184,6829	82
7	***1***	***50527,8701***	***-25128,9351***	***90***

L'indice BIC porterebbe a fissare il numero di classi di livello inferiore a 7, ma come si nota dalla *Tabella 3.3.6*, questa soluzione non può essere accettata in quanto presenta delle classi troppo piccole, anche del 1% (Cluster 7).

Profile 5 Clusters- 4 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7
0,3583	0,2883	0,1161	0,1159	0,0730	0,0358	0,0127

Le considerazione fatte per la soluzione di 7 classi di primo livello si trovano anche nelle soluzioni che prevedono 6 e 5 classi. Queste scelte non possono essere accettate, in quanto presentano ciascuna una classe troppo piccola (rispettivamente del 2% e 1%).

Profile 6 Clusters- 1 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
0,3747	0,2884	0,1753	0,0976	0,0479	0,0160

Profile 5 Clusters- 1 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
0,3869	0,3522	0,1252	0,1238	0,0118

La migliore soluzione è data dal modello con 4 classi di livello inferiore in cui tutte le varie classi sono di dimensione maggiore al 10%.

PROFILE 4 CLUSTERS- 1 CLASS

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4
0,3941	0,3564	0,1271	0,1224

Fissiamo quindi a 4 le classi di primo livello e passiamo a determinare il numero di quelle di secondo.

- **Secondo Passo**

In questo punto dell'analisi determiniamo il numero di classi di secondo livello fissando il primo al valore scelto al passo precedente, ossia a 4.

Tabella 3.3.7: stime del modello per varie classi e cluster

CLUSTER	CLASSI	AIC3	Log-likelihood	Number of parameters
4	1	50806,2763	-25304,1381	66
4	2	50648,6410	-25219,3205	70
4	3	50543,8444	-25160,9222	74
4	4	50508,6563	-25137,3282	78
4	5	50504,4785	-25129,2393	82
4	6	50500,6413	-25121,3206	86
4	***7***	***50499,4925***	***-25114,7462***	***90***

Il miglior indice BIC è relativo al modello con 7 classi di secondo livello, come si può notare dalla *Tabella 3.3.7*. Come per l'analisi precedente, possiamo fissare a 7 le classi di livello superiore solo se, analizzandone la struttura, i cluster risultano di dimensione adeguata. Come si può vedere dal profile per il modello con 7 classi latenti, la dimensione è adeguata, mentre per quanto riguarda i gruppi di medici notiamo che in 2 classi sono raggruppati il 4% dei medici (Class 6 e Class 7). Questa soluzione va quindi scartata.

Profile 4 Clusters- 7 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4
0,3943	0,3327	0,1580	0,1150

CLASSI:

Class1	Class2	Class3	Class4	Class5	Class6	Class7
0,4039	0,2200	0,1419	0,0885	0,0649	0,0427	0,0382

Lo stesso problema emerge anche nei modelli con 6 e 5 classi di secondo livello. Anche in questi modelli i cluster sono di dimensione adeguata ma non le classi, che sono formate dal 4% dei medici. Queste soluzioni vengono scartate.

Profile 4 Clusters- 6 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4
0,3950	0,3330	0,1579	0,1140

CLASSI:

Class1	Class2	Class3	Class4	Class5	Class6
0,4315	0,2135	0,1512	0,0942	0,0652	0,0444

Profile 4 Clusters- 5 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4
0,3950	0,3365	0,1550	0,1135

CLASSI:

Class1	Class2	Class3	Class4	Class5
0,4374	0,2526	0,1448	0,1222	0,0430

Il modello migliore consiste in 4 classi di primo e secondo livello, in cui tutte le classi sono di dimensione superiore al 10%.

Profile 4 Clusters- 4 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4
0,3974	0,3392	0,1526	0,1109

CLASSI:

Class1	Class2	Class3	Class4
0,4567	0,2519	0,1826	0,1089

Fissiamo ora il numero delle classi di secondo livello del nostro modello a 4 e passiamo al terzo passo dell'analisi.

- **Terzo Passo**

In quest'ultimo passo utilizziamo la soluzione precedente per determinare il modello finale, fissiamo quindi a 4 il numero di classi di livello superiore.

Tabella 3.3.8: stime del modello per varie classi e cluster

CLUSTER	CLASSI	BIC	Log-likelihood	Number of parameters
1	4	56866,5303	-28365,7651	45
2	4	52435,2653	-26133,6327	56
3	4	51000,2731	-25399,6366	67
4	4	50508,6563	-25137,3282	78
5	4	50264,7511	-24998,8755	89
6	4	50060,0619	-24880,0310	100
7	***4***	***50056,5492***	***-24861,7746***	***111***

Dalla *Tabella 3.3.8*, la soluzione è data dal modello con 7 classi di primo livello ma, come per i casi precedenti, non prendiamo in considerazione questo caso dato che presenta dei gruppi di dimensione esigua (Cluster 7 1%).

Profile 7 Clusters- 4 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7
0,3130	0,2356	0,1984	0,0948	0,0901	0,0507	0,0173

Le stesse considerazioni si ritrovano anche nel modello con 6 classi di primo livello, infatti è presente una classe piccola (Cluster 6 5%).

Profile 6 Clusters- 4 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
0,3653	0,2637	0,1445	0,0976	0,0818	0,0471

La soluzione ottima è data dal modello con 5 classi di livello inferiore e 4 classi di livello inferiore.

Profile 5 Clusters- 4 Class

CLUSTER:

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
0,3261	0,2851	0,1477	0,1254	0,1157

CLASSI:

Class1	Class2	Class3	Class4
0,3368	0,2621	0,2194	0,1817

La soluzione ottima risulta essere quella definita da 5 classi di livello inferiore e 4 per quello superiore. Il modello scelto è il medesimo di quello determinato tramite il criterio d'informazione BIC.

3.3.3 Considerazioni

Dall'analisi utilizzando la procedura a tre passi, il modello che meglio descrive il mercato farmaceutico italiano è il modello multilivello con 5 classi di livello inferiore e 4 di livello superiore, sia per il criterio BIC che per il criterio AIC3. Questa soluzione coincide con la soluzione emersa nella prima analisi multilivello in cui si è utilizzata la procedura classica e si è considerato il criterio d'informazione BIC. Le procedure portano al medesimo risultato.

Il modello multilivello con 5 classi di livello inferiore e 4 classi latenti sembra spieghi molto bene il mercato in analisi, raggruppando i medici di medicina generale in 4 distinti gruppi.

CONCLUSIONI

Un mercato sempre più competitivo ha portato le aziende a incentrare le loro strategie di marketing per ottenere il maggior vantaggio competitivo, investendo minori risorse finanziarie. Per ottenere tale scopo viene data sempre più importanza alla conoscenza del mercato di riferimento e della propria domanda, per differenziare al meglio la produzione e trovare il posizionamento più adatto all'interno del mercato. Lo strumento migliore, che permette di raggiungere questo scopo, è la segmentazione. Questa tecnica permette di suddividere il mercato in gruppi omogenei, in base a caratteristiche simili. I segmenti che si formano sono accomunati da simili richieste per un bene, l'azienda riesce così ad individuare il o i segmenti più vantaggiosi a cui proporre i prodotti, potendo così soddisfare i clienti. Oltre ad aiutare a capire i consumatori, la segmentazione permette all'azienda di conoscere la concorrenza, fattore molto importante da considerare per ottenere il vantaggio competitivo. La strategia di segmentazione è sempre più usata poiché, oltre a definire il mercato in modo statico, lo rappresenta in modo dinamico, permettendo di carpire eventuali mutamenti.

Sono molti i tipi di segmentazioni che si possono applicare, ognuno di loro è determinato da specifiche caratteristiche che lo rendono più adatto a certi tipi di analisi e meno ad altre. Nel caso in analisi, con informazioni fornite dai medici sulle case farmaceutiche da loro conosciute, risulta più appropriato l'utilizzo della tecnica CL multilivello.

Questa tecnica permette di dividere il mercato in più livelli, in quanto è possibile riassumere delle caratteristiche osservate nelle variabili in caratteristiche latenti. I modelli CL riducono in questo modo una popolazione eterogenea in sottogruppi omogenei al loro

interno e eterogenei tra loro, senza imporre restrizioni ai dati (ad esempio non impone la normalità delle distribuzioni). Vengono, quindi, diminuite le distorsioni relative alla non conformità tra ipotesi iniziali e dati.

L'ipotesi fondamentale dei modelli CL è l'indipendenza locale, che consiste nel considerare le osservazioni indipendenti, data l'appartenenza alla specifica classe latente. Nel caso specifico, questo assunto viene in parte violato, in quanto i soggetti in analisi, medici di medicina generale, assegnano voti multipli a più aziende. Il modello che ammette la dipendenza delle osservazioni è il modello CL Mulitilivello. In questo modello le unità statistiche sono raggruppate nel primo livello e nel secondo le classi latenti a cui le unità sono associate.

In questo studio si è partiti rianalizando il lavoro di Bassi (2009). L'interesse era incentrato nella comprensione delle determinanti della domanda dei farmaci da parte dei medici di medicina generale in Italia, dando importanza alla soddisfazione degli stessi. I dati che si sono utilizzati provengono da un sondaggio di medicina generale in cui venivano espressi vari giudizi sulle strategie di promozione delle industrie farmaceutiche. Il mercato è stato segmentato individuando i gruppi di medici con atteggiamenti simili, utilizzando, appunto, il modello gerarchico di mistura finita per data set a tre vie (Vermut, 2007).

Il lavoro di Bassi è stato rianalizzato utilizzando sempre la tecnica a classi latenti multigruppo ma estendendo l'analisi fino a 7 classi sia per il primo che per il secondo livello. La medesima analisi è stata ripetuta sugli stessi dati ma utilizzando il *metodo a tre passi* introdotto da Lukočienė, Varriale e Vermut (2010). Questo metodo è stato introdotto per eliminare il problema dei modelli CL multilivello, dovuto alla dipendenza sulle decisione riguardante il numero di classi di livello superiore e inferiore. La dipendenza delle unità deriva dalla

struttura gerarchica, la procedura a tre passi tiene in considerazione la dipendenza tra le unità di livello inferiore. Essa consiste nel stabilire il numero di classi di primo livello, al primo passo, ignorando la struttura multilivello; al passo 2 si fissa il numero delle classi di primo livello a quello scelto al passo precedente e si determina il numero di quelle di secondo tramite l'utilizzo dei criteri d'informazione; l'ultimo punto di questa procedura consiste nel scegliere nuovamente il numero delle classi di livello inferiore, fissando quelle di livello superiore al valore del secondo passo. Nell'analisi classica i passi fatti dall'analista sono invece due, implicando una dipendenza tra il numero delle classi scelte a livello inferiore con quello di livello superiore. Vengono anche valutate le performance dei criteri d'informazione, confrontando gli indici BIC e AIC3, in quanto nell'articolo di Lukočienė, Varriale e Vermut (2010), AIC3 viene preferito a BIC.

Confrontando i procedimenti, notiamo che il metodo proposto a tre passi risulta molto più veloce rispetto a quello classico, e di facile applicabilità. Dall'analisi, la scelta del modello ricade su quello che prevede 5 classi di livello inferiore e 4 di livello superiore. Solo nell'analisi classica, in cui si è usato il criterio d'informazione AIC3, il modello scelto è formato da 4 classi di livello superiore e inferiore. Il motivo per cui non vi è corrispondenza tra i modelli è legato al fatto che è una scelta soggettiva, anche basata sulla verifica delle proprietà dei segmenti, soprattutto sulla dimensione. Il modello con 5 classi di primo livello e 4 di secondo è stato rigettato, nella procedura classica con l'indice AIC3, in quanto si è cercato di ottenere delle classi di dimensione maggiore al 10%, e non si ritenuto significativamente grande nella classe formata dal 7% di medici. Il criterio BIC risulta essere un buon indice anche in analisi in cui sono presenti risposte categoriali, a differenza di quanto è stato affermato da Lukočienė, Varriale e Vermut (2010). In tutti e due i modelli vi

sono segmenti ben separati, c'è un gruppo interessato a tutti gli aspetti, che ha dato giudizi alti, e uno formato da medici poco interessati, i giudizi sono tutti bassi. I giudizi maggiori sono in merito agli aspetti che riguardano la frequenza delle visite dei rappresentanti farmaceutici e la qualità della loro formazione; mentre sono stati dati i giudizi minori in merito all'aspetto che riguarda l'informazione dei rappresentanti.

Il nuovo procedimento a tre passi è un buon metodo di analisi che rende la procedura più veloce e di facile applicabilità, ma in sostanza i risultati ottenuti, sia con questa procedura che con quella tradizionale e sia facendo riferimento a BIC che a AIC3, sono gli stessi.

BIBLIOGRAFIA

AKAIKE H., (1974), "*A New Look at the Statistical Model Identication*", IEEE Transactions on Automatic Control, 19, pp. 716-723.

ANDREWS R. L., CURRIM I. S., (2003), "A Comparison of Segment retention criteria for Mixture Logit Models", Journal of Marketing, 40, pp. 235-243.

BASSI F., (2007), "*Latent Class Models for MarkeTing Strategies. An Application to the Italian Pharmaceutical Market*", Journal of Statistical Methods and Applications, 2, pp. 279-287.

BASSI F., (2009), "*Latent Class Factor Models for Market Segmentation: an Application to Pharmaceuticals*", Methodology, 5, pp. 40-45.

BOZDOGAN H., (1987), "*Model selection and Akaike's Information Criterion(AIC): The General Theory and Its Analytical Extensions*". Psychometrika 52, pp. 345-370.

BOZDOGAN H., (1993), "*Choosing the Number of Component Clusters in the Mixture-model Using a New Informational Complexity Criterion of the Inverse-fisher Information Matrix*", Studies in Classication, Data Analysis,and Knowledge Organization, pp. 40-54.

BRASINI S., TASSINARI F., TASSINARI G., (1993), "*Marketing e pubblicità Approccio statistico all'analisi dei mercati di consumo*", Il Mulino, Bologna.

CLOGG C. C., GOODMAN L. A., (1984), "*Latent Structure Analysis of a Set of Multidimensional Contingency Tables*", Journal of the American Statistical Association, 79, pp.762-771.

FABRIS G., (1992), "*La pubblicità, Teorie e Prassi*", Franco Angeli, Milano.

DIAS J. M.G., (2004), " *Finite Mixture Models. Review, Applications, and Computerintensive Methods*", Groningen, The Netherlands: Research School Systems, Organisation and Management, University of Groningen.

GOODMAN L. A., (1974a), "*The Analysis of Systems of Qualitative Variables when some of the Variables are Unobservable: Part I. A Modified Latent Structure Approach*", American Journal of Sociology, 79, pp.1179-1259.

GOODMAN L. A., (1974b), "*Exploratory Latent Structure Analisys Using Both Identifiable and Unidentifiable Models*", Biometrika, 61, pp. 215-231.

GRANDINETTI R., (2002), "*Concetti e Strumenti di Marketing. Il Ruolo del Marketing tra Produzione e Consumo*", Etas, Milano.

HALEY R.I., (1968), "*Benefit Segmentation: a Decision Oriented Research Tool*", Journal of Marketing, july.

HEINEN T., (1996), "*Latent Class and Discrete Latent Trait Models. Similarities and Differencies*", Sage Publications, Thousand Oaks.

LAZARSFELD P. F., HENRY N. W., (1968), "*Latent Structure Analysis*", Boston, Hughton Muffin.

LUKOČIENĖ O., VARRIALE R., e VERMUNT J. K., (2010), "*The Simultaneous Decision About the Number of Lower- And Higher-Level Classes in Multilevel Latent Class Analysis*", Sociological Methodology, Volume 40,pp. 247-283.

MADIGSON J., VERMUNT J. K. (2001), "*Latent Class Factor and Cluster Models, Bi-plots, and Related Graphical Displays*", Sociological Methodology, vol.31, pp. 223-264.

MAGIDSON J., VERMUT J. K., (2004), "*Latent Class Models*", The Sage Handbook of Quantitative Methodology for the Social Sciences, Section III/Models for Categorical Data, Capitolo 10, pp. 175-198.

MCLACHLAN G, PEEL D., (2000), "*Finite Mixture Models*", New York: J.Wiley and Sons. Inc.

PRANDELLI E., VERONA G., (2006), "*Marketing in Rete. Oltre Internet Verso il Nuovo Marketing*", McGraw-Hill, Milano.

SCHWARZ G. E., (1978), "*Esti*m*ating the Dimension of a Model*", Annals of Statistics 6, pp. 461-464.

VERMUNT J. K., (2003a), "*Application of Latent Class Analysis in Social Science Research*", Lecture Notes on Artificial Intelligence, 2711, pp. 22-36.

VERMUNT J. K., (2003b), "*Multilevel Latent Class Models*", Sociological Methodology, 33, pp 213-239.

VERMUNT J. K., (2007), "*A Hierarchical Mixture Model for Clustering Three-way Data Sets*", Computational Statistics & Data Analysis, 51, pp. 5368-5376.

VERMUNT J. K., (2007), "*Latent Class and Finite Mixture Models for Multilevel Data Sets*", Statistical Methods in Medical Research, pp.1-19.

VERMUNT J. K., MADIGSON J. (2002), "*Latent Class Cluster Analysis*", in HAGENAARS J. A., McCUTCHEON A. L., Applied latent class analysis, pp.89-106, Cambridge, UK: Cambridge University Press.

VERMUNT J. K., MADIGSON J. (2005), "*Latent GOLD 4.0 User's Guide*", Statistical Innovations Ic., Belmont (MA).

RINGRAZIAMENTI

I ringraziamenti sono pochi perché...il tempo stringe!!

Ringrazio la mia famiglia per la pazienza, MOLTA, e il sostegno, MOLTO, che mi ha dato in questi anni di studio.

Ringrazio i miei amici, vecchi e nuovi, per la simpatia e l'allegria, per tutti i momenti belli passati assieme.

Ringrazio la professoressa Bassi per l'aiuto e la pazienza con cui mi ha seguito in questo periodo di lavoro.

Ringrazio, infine, Google che nel momento del bisogno era sempre pronto a risolvere i miei innumerevole problemi con un semplice click.