UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

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

SEGMENTAZIONE DEL MERCATO EUROPEO DEI PRODOTTI FINANZIARI: UN APPROCCIO A CLASSI LATENTI.

RELATORE: CH.MA PROF.SSA FRANCESCA BASSI

LAUREANDA: ELISA VISENTIN

ANNO ACCADEMICO 2009/2010

Alla nonna Pia.

Ai miei genitori.

Indice

Introduzione.

Con la liberalizzazione dei mercati e la loro conseguente internazionalizzazione, è sempre più pressante l'esigenza di strategie che permettano un vantaggio competitivo in ambito sovranazionale, anche senza dover diversificare il portafoglio prodotti. La competizione è più intensa, perciò conoscere in modo approfondito il proprio mercato diventa, sul piano competitivo, una risorsa indispensabile per l'azienda che espande il proprio *business*.

Il mercato dei prodotti finanziari, negli ultimi decenni, ha conosciuto una forte internazionalizzazione. Le aziende vendono i propri prodotti al di là dei confini nazionali, anche grazie a fusioni, acquisizioni e alleanze. L'offerta a disposizione dei consumatori è, pertanto, sempre più varia e la competizione del settore più intensa. Per le aziende europee (e per quelle del resto del mondo), pertanto, si profila la necessità conoscere approfonditamente la struttura della domanda non solo del paese in cui si opera, ma anche del mercato europeo e perfino globale.

A questo proposito, si è dimostrato che l'analisi della proprietà familiare di prodotti finanziari porta informazioni rilevanti per supportare le decisioni di *marketing* (Bijmont, Paas e Vermunt, 2004). In particolare, conoscere similarità, differenze ed evoluzione di questo mercato in ambito internazionale si rivela cruciale nel formulare strategie di *marketing* sovranazionali. Non stupisce, quindi, che la letteratura economica e di *marketing* abbia rivolto la propria attenzione a questi argomenti.

Lo strumento che permette di approfondire la conoscenza della domanda che si serve è la segmentazione. Essa classifica gli individui di un mercato sulla base delle loro domande individuali in gruppi omogenei al loro interno ed eterogenei fra di loro. Lo scopo è adattare specifiche combinazioni degli strumenti del *marketing mix* alle esigenze espresse dai singoli segmenti e alle loro peculiarità. La segmentazione è alla base del *target marketing*, che si propone di individuare il segmento, o i segmenti, che l'azienda può servire in maniera più

efficiente ed efficace, date le sue risorse e competenze specifiche, e posizionarsi su di essi. La strategia competitiva di base implementabile, nota la segmentazione del mercato, è la differenziazione, che consiste nell'offrire ai segmenti un prodotto mirato, così da poter spuntare un *premium price*.

La statistica offre numerose tecniche di segmentazione, che si propongono di costruire gruppi omogenei internamente ed eterogenei tra di loro. Alcune sono a priori e altre a posteriori, ovvero alcune prevedono di determinare il numero e la tipologia dei segmenti prima dell'analisi, altre con i risultati della stessa.

L'obiettivo di questa tesi è segmentare il mercato europeo di alcuni prodotti finanziari (conto corrente e libretto di risparmio, titoli di stato o obbligazioni, azioni o partecipazioni, fondi comuni di investimento o gestioni patrimoniali, pensioni integrative private e assicurazioni sulla vita), con riferimento alla popolazione degli ultracinquantenni. I dati utilizzati sono desunti dall'indagine SHARE del 2006 e riguardano capofamiglia (e relative famiglie) di quattordici paesi europei, dall'Europa del nord, fino all'Europa mediterranea. Affinché sia possibile, una volta individuati, descrivere i segmenti in modo utile per l' utilizzo dei risultati in ambito di *marketing*, si impiegano anche informazioni demografiche e di carattere generale sulle unità statistiche, quali l'età del capofamiglia, il sesso, il livello di istruzione, le capacità cognitive, lo stato civile e occupazionale, l'avere o meno un compagno, la dimensione del nucleo familiare e l'eventuale proprietà dell'abitazione.

L'analisi di segmentazione sarà condotta utilizzando i modelli a classi latenti (*latent class model*), e in particolare le loro varianti classica e multilivello. Già dagli anni ottanta dello scorso secolo, si sono adoperati modelli a mistura finita per la segmentazione e, fra tutti, i modelli a classi latenti sono forse i più famosi. I modelli a classi latenti si rendono molto utili quando si presenta la necessità di individuare delle similarità in una popolazione omogenea, infatti recentemente sono stati riconosciuti come metodi di segmentazione (Vermunt e Madigson,

2002). Tuttavia la loro applicazione si dimostra utile anche nel campo delle scienze sociali e naturali. La procedura di segmentazione è probabilistica, perché *model-based*, e quindi molto più flessibile di tecniche, come ad esempio la *cluster analysis*, che raggruppano le unità statistiche sulla base delle loro distanze o similarità rispetto a una variabile. L'idea di fondo è riassumere l'eterogeneità osservata sulle unità statistiche, in base a loro caratteristiche latenti. In altre parole si utillizza una variabile latente discreta, le cui modalità definiscono delle classi omogenee. Sulla base delle probabilità condizionate di avere una determinata caratteristica, data l'appartenenza a una classe latente, le unità sono assegnate ai *cluster* latenti individuati, formando così dei segmenti. Le osservazioni sono assunte indipendenti, data la loro appartenenza a una classe latente (indipendenza locale).

Nell'analisi multilivello si ha la possibilità di tenere conto della struttura gerarchica o annidata della popolazione, su due o più livelli, nonché di rilassare l'ipotesi di indipendenza locale. Oltre ad individuare classi a livello uno (unità statistiche), si individuano delle classi latenti a livello due, che raggruppano le unità di secondo livello. Nel nostro lavoro, si utilizzerà il modello multilivello per classificare i capofamiglia in segmenti e i paesi in gruppi. Ad ogni gruppo verranno assegnati i segmenti di individui che presentano maggiori probabilità di appartenervi.

Nel primo capitolo, si presenterà la segmentazione, sia dal punto di vista delle implicazioni di *marketing*, sia dal punto di vista statistico. In particolare si esporranno alcune tecniche statistiche di classificazione.

Nel secondo capitolo, si presenteranno i modelli a classi latenti, nelle loro tre specificazioni tradizionale, fattoriale e multilivello.

Nel terzo capitolo si faranno delle analisi preliminari sui dati SHARE, per poter individuare già delle omogeneità nel campione. Si utilizzeranno strumenti della statistica descrittiva consoni alla natura delle variabili a disposizione e allo scopo di questo lavoro, quali distribuzioni di frequenza, mediane, medie e test chi quadrato di

indipendenza, e degli strumenti grafici, quali *boxplot*, *bubbleplot*, diagrammi a bastoncini.

Nel quarto capitolo, si segmenterà il mercato, implementando il modello classico. Con l'analisi dei risultati si costruiranno dei profili dei segmenti, di cui si verificherà la validità dal punto di vista del *marketing*.

Lo schema adottato nel quarto capitolo, si riproporrà, in parte, nel quinto, dove, però, si utilizzeranno i modelli a classi latenti multilivello. Infatti, dapprima si segmenterà il mercato, poi verranno costruiti i profili dei segmenti individuati. Saranno, quindi, analizzati i gruppi di paesi determinati dall'analisi e assegnati loro i *cluster* di consumatori. Infine si valuteranno efficienza ed efficacia dei segmenti.

Nelle conclusioni si commenteranno le due diverse analisi di segmentazione, individuando quale si adatta meglio alla struttura del nostro mercato. Si proporranno, inoltre, degli spunti per ricerche future.

Capitolo 1

La segmentazione.

Difficilmente un'impresa riesce a raggiungere con la propria offerta l'intero mercato in cui opera, o intende operare, per numerosi motivi, tra i quali la distanza geografica dai propri clienti, l'eterogeneità delle loro preferenze, la non disponibilità delle risorse necessarie a raggiungere tutti. Per questo motivo molte aziende scelgono di servire solo alcune porzioni di mercato, più attraenti e vantaggiose, date le loro risorse. Uno strumento utile per una maggiore conoscenza del mercato e un impiego più efficace del *marketing mix* è la segmentazione. Essa consiste nel suddividere la popolazione eterogenea dei consumatori di una tipologia di prodotti o servizi in gruppi di consumatori, detti segmenti, che richiedono prodotti o *marketing mix* differenziati.

La segmentazione è alla base delle strategie di posizionamento (Figura 1.1) e differenziazione e quindi si inserisce nell'insieme più ampio delle strategie di *marketing* aziendali. Kotler (2004) individua nella segmentazione la prima fase del *target marketing*, che si articola anche con la selezione di uno o più segmenti a cui rivolgere la propria offerta (*targeting*) e la definizione e articolazione dei benefici distintivi del prodotto sull'obiettivo scelto (posizionamento). La differenziazione, invece, prevede l'offerta di un prodotto percepito come unico dal consumatore in base ad alcune variabili che egli ritiene importanti e per il quale è disposto a pagare un *premium price*.

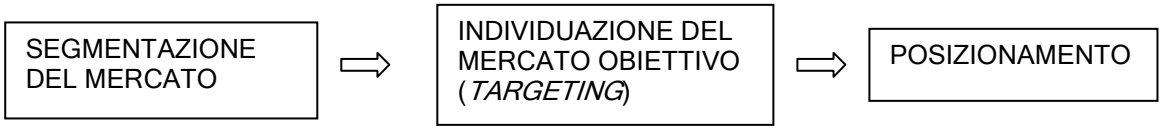

Figura 1.1 Fasi del target marketing.

Ogni consumatore di un mercato possiede una funzione di domanda individuale. La segmentazione si basa sull'identificazione delle domande individuali, la ripartizione dell'intero mercato in gruppi di consumatori accomunati da una specifica domanda e la sollecitazione di domande diverse da parte dei consumatori, in modo da adattare razionalmente i prodotti e le attività di marketing agli specifici bisogni che queste domande esprimono. L'analisi statistica di segmentazione di fatto suddivide un mercato in segmenti attraverso metodi con approcci molto diversi, come si vedrà in seguito. Resta necessaria, pur disponendo di strumenti di segmentazione, la capacità di interpretare il mercato in cui si opera e la domanda che si serve da parte degli analisti e del *management*: le funzioni di domanda individuali non sono osservabili, pertanto le aziende possono percepire in modo diverso l'eterogeneità della domanda, interpretandola con criteri differenti, più o meno efficaci.

1.1 Vantaggi della segmentazione.

Una corretta segmentazione porta numerosi vantaggi all'azienda.
Innanzitutto permette una migliore e approfondita conoscenza del mercato, definendolo in termini di bisogni del cliente e permettendone la percezione dei mutamenti, se non avviene *una tantum*. L'azienda che conosce bene il proprio mercato, effettivo o potenziale, possiede uno strumento molto potente per valutare i punti di forza e debolezza della propria offerta e di quella dei concorrenti; è in grado pertanto di definire con maggiore precisione ed efficacia gli obiettivi da perseguire e di razionalizzare il proprio portafoglio prodotti in modo profittevole. Si aggiunga che un posizionamento di successo, che non può esimere da una corretta analisi di segmentazione, di fatto eleva una barriera all'entrata di nuovi concorrenti non facilmente minabile. Infine con l'analisi di segmentazione, un'azienda guadagna la capacità di misurare gli effetti sulle vendite delle strategie di marketing implementate.

1.2 Fasi operative della segmentazione.

Dal punto di vista operativo, la segmentazione si può riassumere in quattro fasi consecutive: (i) definizione del problema e scelta del modello, (ii) indagine sul campo, (iii) scelta della metodologia di analisi dei dati raccolti, (iv) elaborazione ed interpretazione dei risultati ottenuti.
Affrontando la prima fase si pongono due problemi: la scelta tra condurre un'unica indagine o una serie di studi ripetuti e la scelta del modello di segmentazione. Spesso l'analisi di segmentazione è intesa come il risultato di una serie complessa di attività, alcune delle quali molto dispendiose, da attuare in un'unica soluzione. Le aziende che scelgono questa via, di solito, si propongono di conoscere in modo approfondito il mercato in cui operano o nel quale vorrebbero entrare. Tuttavia è chiaro che ripetere l'analisi in periodi successivi, sebbene possa essere molto dispendioso, permette di conoscere l'evolversi del mercato nel tempo, con la possibilità di adeguare le proprie strategie ai cambiamenti, benché questo richieda una flessibilità da parte dell'azienda non sempre raggiungibile.
Specificare un modello di segmentazione significa specificare le variabili che generano il processo di classificazione dei consumatori (basi della segmentazione) e quelle che descrivono i segmenti (i descrittori, che entrano in gioco nella fase di interpretazione dei profili individuati con l'analisi). Esistono diversi criteri per la segmentazione, che si distinguono proprio per le basi scelte. Possiamo distinguere la segmentazione geografica, demografica, psicografica, comportamentale e la *benefit segmentation*, che saranno approfondite nel paragrafo successivo. Dal punto di vista prettamente statistico, poi, è necessario scegliere se utilizzare modelli di segmentazione a priori o a posteriori.
Scelto il modello e tenendo ben presenti gli obiettivi, a questo punto bisogna definire il piano dell'indagine. L'analisi di segmentazione spesso richiede la raccolta di dati primari *ad hoc*, attraverso strumenti di indagine, metodi qualitativi (*focus groups*, interviste in profondità e

tecniche proiettive sono particolarmente utili per indagare soprattutto atteggiamenti, motivazioni e opinioni) e osservazione. In questo caso è necessario pianificare una vera e propria indagine, affrontando numerosi problemi. La scelta dell'unità di analisi più appropriata non è una questione banale (per esempio, all'interno di una famiglia, più individui, con le loro peculiarità, intervengono nelle scelte e bisogna tenerne conto) e può avvenire in anticipo o in corso d'opera, con i cosiddetti campioni a valanga o palla di neve. Anche nella scelta delle variabili bisogna prestare molta attenzione, perché la loro definizione operativa impatta sulle procedure analitiche e su ampiezza e composizione dei segmenti. Inoltre se si vuole andare oltre una semplice descrizione degli intervistati, bisogna estrarre un campione probabilistico, garantendosi così la possibilità di estendere i risultati all'intero universo di riferimento. Infine, è necessario predisporre una strategia per ovviare alla rilevazione incompleta. Tuttavia le aziende hanno a disposizione sia internamente che esternamente una grande quantità di dati secondari, che potrebbero essere utili ai fini di un'analisi di segmentazione e che sono reperibili velocemente e a bassi costi. È chiaro che la scelta di quali informazioni usare deve risultare da un'attenta valutazione di costi, benefici ed esigenze dell'azienda.

La scelta della tecnica statistica da utilizzare per l'analisi dei dati tiene conto di numerosi aspetti, come il tipo e la qualità dei dati a disposizione, gli obiettivi prefissati, le risorse finanziarie disponibili e, soprattutto, il modello di segmentazione scelto. Alcune tecniche, infatti, si adattano meglio a certi tipi di basi e al criterio di segmentazione.

In sede di interpretazione dei risultati, come anche di definizione del problema, entrano in gioco differenti competenze aziendali che devono interagire. Sicuramente, affinché un'analisi di segmentazione abbia successo, analista e *management* devono definire dettagliatamente il problema da risolvere, individuando le basi idonee a fornirne una soluzione e i descrittori che permettono il giusto grado di generalizzazione nella rappresentazione dei segmenti, tenendo sempre ben presenti gli obiettivi strategici aziendali. I risultati dell'analisi, allora,

diventano utili linee guida per scegliere le successive strategie da adottare, come la scelta del numero e dei segmenti su cui posizionarsi. Questa fase è molto delicata: perché i segmenti scelti portino un vantaggio competitivo all'azienda devono possedere una serie di requisiti, illustrati in seguito, da valutare attentamente. A questo scopo il *management* ha a disposizione numerosi strumenti, oltre ai risultati dell'analisi di segmentazione, come ad esempio la matrice di segmentazione e l'analisi di attrattività di Porter (lo schema a cinque forze applicato a livello segmento anzichè settore), che non sono oggetto di questo studio.

1.3 Criteri di segmentazione fondati sulla scelta delle basi.

In precedenza si è già accennato all'esistenza di cinque diverse tipologie di segmentazione, che si distinguono per le basi scelte, ovvero segmentazione geografica, demografica, psicografica, comportamentale e per benefici attesi.
Il criterio di segmentazione geografico suddivide il mercato in aree territoriali, presupponendo che le preferenze dei consumatori varino con le caratteristiche del luogo di residenza; alcune variabili base potrebbero essere città, Paese di residenza, densità di popolazione, condizioni infrastrutturali della zona di residenza, caratteristiche climatiche. La segmentazione demografica utilizza come basi variabili come sesso, età, reddito, professione, numero dei componenti del nucleo familiare, che chiaramente descrivono la condizione demografica dei consumatori. Con questi due criteri si ottiene un'ottima conoscenza del profilo dei consumatori e delle modalità per raggiungerli, ma non si ottiene alcuna informazione sui loro desideri, sulle loro aspettative, sulle loro motivazioni. La segmentazione di tipo geografico e demografico, pertanto, è molto efficiente, ma poco efficace, non permettendo la conoscenza del processo decisionale del consumatore. A questo problema si può ovviare con gli altri criteri di segmentazione.

La segmentazione psicografica, infatti, cerca di individuare segmenti con stili di vita[1]simili, integrando diverse discipline, come la psicologia, la sociologia, l'antropologia culturale e il behaviorismo[2]. Le variabili scelte come basi riguardano attività, interessi, opinioni e convinzioni dei consumatori .

La segmentazione comportamentale si focalizza sul comportamento del consumatore, concentrandosi sui suoi obiettivi e sulle caratteristiche ricercate nei prodotti e fornendo informazioni molto utili al posizionamento. Essa ha come basi variabili che descrivono aspetti quali l'occasioni d'uso, i vantaggi ricercati, lo *status* del consumatore (ovvero non consumatore, consumatore potenziale ma non effettivo, consumatori abituali, ex consumatori), l'intensità d'uso, la propensione all'acquisto, l'atteggiamento e la fedeltà di marca.

Infine la *benefit segmentation* raggruppa i consumatori in segmenti omogenei per i benefici e i vantaggi ricercati in un prodotto o servizio. Di solito questo tipo di segmentazione si sviluppa definendo l'insieme di benefici ricercabili in un prodotto o servizio e identificando successivamente gruppi di consumatori interessati al medesimo sottoinsieme di attributi. I dati necessari per segmentare il mercato secondo gli ultimi tre criteri sono generalmente più difficili e costosi da ottenere, giacché riguardano la sfera intima e personale dei consumatori; tuttavia esse forniscono informazioni strategicamente molto rilevanti.

[1] Giampaolo Fabris definisce gli stili di vita "insiemi di personeche 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 contesto. (Prandelli, Verona, 2006).

1.3.1 Microsegmenti e macrosegmenti.

Il problema da affrontare con la segmentazione, talvolta, è così complesso da dover essere articolato in due livelli: la definizione di segmenti sulla base del criterio ritenuto più opportuno e la loro qualificazione mediante informazioni utili per il marketing. Ad esempio le segmentazioni psicografica, comportamentale e per benefici attesi necessitano di descrittori demografici e\o geografici per profilare i segmenti, altrimenti difficilmente raggiugibili e identificabili. Da qui la necessità di adottare criteri aggiuntivi di segmentazione. In questo caso è utile distinguere i macrosegmenti, ovvero gli aggregati più ampi, dai microsegmenti, più circoscritti. Ad esempio, quando si segmenta un mercato internazionale, spesso i diversi paesi, o gruppi di paesi, rappresentano mercati con caratterizzazione nettamente specifica. Potrebbe allora risultare utile, ai fini di una ricerca approfondita, risegmentare ciascun paese o gruppo analizzandone l'eterogeneità interna della domanda, ottenendo così macrosegmenti rappresentati dai paesi e microsegmenti, interni a ciascun macrosegmento. Pure i descrittori, utilizzati per profilare i segmenti, possono offrire criteri di risegmentazione. Ad esempio Haley (1968)[3], analizzando il mercato dei dentifrici con la *benefit segmentation*, individuò quattro categorie di consumatori: sensoriali, socievoli, apprensivi ed autonomi. Esaminando le caratteristiche demografiche dei consumatori socievoli, scoprì che essi erano prevalentemente adolescenti e giovani. In casi di questo tipo, può essere utile risegmentare con ulteriori criteri ognuno dei segmenti individuati in una prima analisi.

I modelli a classi latenti, oggetto di questa tesi, sono particolarmente utili anche nei casi in cui si profila la necessità di segmentare in più livelli i mercati, grazie all'opportunità di inserire covariate nell'analisi,

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

che fungono proprio da variabili descrittori, e di analizzare strutture gerarchiche con la variante multilivello.

1.3.2 Caratteristiche dei segmenti.

Si potrebbe, erroneamente, pensare che segmentando per fasi successive un mercato, fino ad esaurire i criteri adottabili, si possa ottenere una fotografia esaustiva della varietà della domanda, una sorta di segmentazione perfetta. In realtà, come si è già accennato, affinché la domanda sia rappresentata in modo strategicamente funzionale per il *marketing*, i segmenti individuati devono possedere una serie di caratteristiche, che la segmentazione su più livelli non garantisce affatto.

Innanzitutto i segmenti devono essere identificabili, omogenei al loro interno ed eterogenei fra di loro, in altre parole devono raggruppare consumatori con domande individuali simili e risposte agli strumenti del *marketing mix* uniformi, ma ben distinte rispetto a quelle di consumatori appartenenti ad altri segmenti. Inoltre le basi di segmentazione devono essere pertinenti rispetto agli obiettivi dell'analisi.

I segmenti, poi, devono essere consistenti (*substantiality*) e profittabili, ovvero devono avere un'ampiezza e/o una capacità di assorbimento tali da garantire un profitto all'azienda che li serve.

Ancora, i segmenti devono essere sufficientemente stabili nel tempo; a questo proposito è utile ricordare che la specificità delle basi di segmentazione è direttamente proporzionale alla volatilità.

L'accessibilità, ovvero la possibilità di essere raggiunti dagli strumenti del *marketing mix*, è un'ulteriore caratteristica fondamentale, assieme alla misurabilità della loro dimensione e del loro potere d'acquisto (per questo devono essere misurabili le variabili che definiscono i segmenti).

Infine i segmenti devono avere una certa capacità di risposta e propositività.

Le prime due caratteristiche garantiscono che le strategie implementate siano efficaci (*meaningful segmentation*), ovvero che i risultati

conseguiti corrispondano agli obiettivi preposti; le altre invece garantiscono l'efficienza delle strategie (*actionable segmentation*), ovvero che i risultati conseguiti giustifichino (e superino) le risorse impiegate. Di fatto, a livello strategico, non è sempre possibile perseguire sia la segmentazione *actionable* che *meaningful*, perché ogni segmento può possedere un grado diverso di ogni caratteristica.

1.4 Modelli e tecniche statistiche di segmentazione.

Nella prima fase operativa, si devono scegliere sia il criterio che lo schema di segmentazione. I modelli utili a questo scopo si suddividono in a priori e a posteriori. Nei primi, la popolazione è suddivisa secondo le modalità di una o più basi, scelte a priori. Il numero e la tipologia dei segmenti sono anch'essi prefissati. Ad esempio, nell'analisi del mercato italiano si può scegliere di usare come basi le regioni italiane, piuttosto che la distinzione tra nord, centro e sud del paese. Nei modelli a posteriori, al contrario, basi, numero e tipologia dei segmenti non sono prefissati, ma determinati con l'analisi, in base a criteri di dissomiglianza rispetto alle variabili scelte. Questo tipo di modelli è utilizzato soprattutto nella segmentazione psicografica, comportamentale e per benefici attesi. Un'ulteriore distinzione nell'ambito dei modelli statistici si fa tra tecniche di segmentazione per omogeneità, per obiettivi e flessibili.
Nella terza fase, infatti, si sceglie proprio quale di queste tecniche adottare per l'analisi, chiaramente tenendo conto di quanto deciso ed attuato in precedenza. Le tecniche per omogeneità, come la *cluster analysis*, suddividono le unità statistiche in gruppi, che per costruzione hanno un'elevata omogeneità interna e un'ampia variabilità esterna, sulla base della similarità rispetto ad un insieme di variabili.
Le tecniche flessibili (come la *conjoint analysis*) suddividono i consumatori sulla base della similarità dei loro profili, in termini di preferenze accordate a prodotti esistenti o in fase di progettazione.
Le tecniche statistiche per obiettivi, invece, suddividono le unità sulla base di una o più variabili dipendenti note a priori, sulle quali hanno

influenza delle variabili esplicative che descrivono le caratteristiche dei segmenti. Esse sono *Automatic Interaction Detection* (AID), *Chi Squared Automatic Interaction Detection* (CHAID) e la regressione logistica.

1.5 Tecniche di analisi dei dati a priori.

Gli scopi delle tecniche a priori sono classificare le unità in segmenti e determinarne i profili.
La suddivisione delle unità avviene secondo le modalità di una variabile scelta a priori oppure, se le basi sono due o più, con la classificazione incrociata degli individui, sempre fatta secondo le modalità delle variabili. Se le basi sono di tipo continuo, è opportuno trasformarle in variabili categoriali con un numero ridotto di modalità, per limitare l'impatto dell'uso di metriche diverse. Un'obiezione a quest'ultima procedura, oltre alla possibile soggettività nella scelta dei valori, riguarda il fatto che in tal modo si può ostacolare l'individuazione delle cause di interazione tra le variabili, qualora ve ne siano di significative. Tuttavia, l'arbitrarietà è la filosofia alla base della segmentazione a priori, con le distorsioni che può portare: gli stessi confini dei segmenti non sono oggettivi.
Quelle a priori sono tecniche di segmentazione per obiettivi, pertanto si propongono di legare una serie di variabili esplicative potenzialmente rilevanti alla variabile dipendente, *focus* dell'analisi, in modo da spiegare le cause del comportamento del consumatore o, più semplicemente, descrivere i segmenti individuati. Le variabili esplicative di solito sono di tipo geografico e demografico.
La procedura di segmentazione per obiettivi si articola diverse fasi. Innanzitutto si selezionano le basi, quindi si scelgono le variabili esplicative, o concomitanti, che covariano con la base e descrivono le caratteristiche dei segmenti. Successivamente si suddividono iterativamente le unità statistiche in due gruppi, esaustivi e mutuamente esclusivi, secondo le modalità di una delle variabili esplicative. Si

valuta, ad ogni passo, la segmentazione migliore, sulla base di una regola di ottimalità, che tiene conto dell'omogeneità entro e dell'eterogeneità tra i sottoinsiemi per la variabile criterio. Il processo di suddivisione dei gruppi si ferma al raggiungere delle condizioni richieste dalla regola di arresto. Infine è utile rappresentare graficamente la segmentazione con un dendrogramma, i cui i nodi rappresentano i passi, i rami le condizioni di suddivisione e le foglie i nodi terminali.

1.5.1 *Automatic Interection Detection* (AID).

AID è una procedura gerarchica di segmentazione binaria: ad ogni passo, si prendono in considerazione tutte le possibili suddivisioni dicotomiche in gruppi disgiunti, secondo le modalità di una variabile esplicativa. Di volta in volta si sceglie la variabile che soddisfa il criterio di ottimalità, ovvero che bipartisce il collettivo in due sottogruppi con la massima devianza tra di loro e la minima devianza al loro interno, rispetto alla variabile base. La procedura iterativa termina quando un gruppo è troppo piccolo, o inferiore a una soglia prefissata, per garantire l'attendibilità delle stime statistiche o per essere di interesse concreto; oppure quando il gruppo originario è così omogeneo da non rendere opportuna un'ulteriore suddivisione (in altre parole non raggiunge la soglia minima di devianza tra i gruppi); quando non si individua una suddivisione che provochi un incremento della devianza tra i gruppi superiore a una soglia minima; quando si raggiunge il numero massimo di passi del processo, che corrisponde al numero di segmenti fissato a priori.

1.5.2 *Chi Squared Automatic Interaction Detection* (CHAID).

CHAID, invece, è una procedura gerarchica di segmentazione multipla, ovvero ad ogni passo si suddividono i gruppi in *s* sottoinsiemi disgiunti, sempre sulla base delle modalità delle variabili esplicative. Il criterio di ottimalità valuta l'omogeneità interna e l'eterogeneità esterna ai

segmenti con un test chiquadro. In pratica, si testa l'ipotesi nulla di indipendenza dei caratteri, confrontando le frequenze osservate e quelle teoriche, calcolate sotto H_0; se il valore del test risulta alto, si rifiuta l'ipotesi nulla, evidenziando così una forte dipendenza tra le variabili e di conseguenza una forte omogeneità interna. In altre parole CHAID individua quali variabili sono più connesse alla base e quindi le loro modalità più adatte a descrivere il profilo del segmento.
Tra le tecniche di segmentazione a priori, Brasini, Tassinari, Tassinari (1993) inseriscono anche l'analisi discriminante multipla. Anch'essa esamina la relazione tra la variabile base, che deve essere categorica, e le variabili predittive, che descrivono gli individui. L'obiettivo, però, è l'individuazione di una regola che predica quale modalità della variabile criterio presenta un individuo, sulla base di una funzione lineare che massimizza il rapporto di devianza tra ed entro i segmenti per la variabile criterio. In pratica l'analisi discriminante multipla permette di classificare le unità di cui si conosce il solo profilo, di verificare l'esistenza di differenze significative tra i valori medi delle esplicative all'interno delle classi e di individuare quali variabili caratterizzano le differenze tra i profili medi in modo migliore.

1.6 Tecniche di analisi dei dati a posteriori.

Questi metodi non fanno assunzioni a priori sulle tipologie esistenti nelle unità esaminate, ma usano un procedimento empirico di classificazione. Tra le tecniche di analisi a posteriori distinguiamo una procedura di segmentazione per omogeneità, che delega all'analisi statistica la definizione della partizione che garantisce la massima omogeneità interna e la minima omogeneità esterna, e una flessibile.

1.6.1 Una tecnica per omogeneità: la *cluster analysis*.

La più diffusa tecnica per omogeneità è la *cluster analysis* (CA). Si tratta di una tecnica multivariata ed esplorativa, che scompone una realtà

complessa di osservazioni plurime in tipologie specifiche. In altre parole suddivide un insieme eterogeneo in sottoinsiemi mutuamente esclusivi e omogenei all'interno. La CA permette di migliorare la comprensione dei comportamenti d'acquisto, valutare l'opportunità di sviluppare nuovi prodotti, valutare se i prodotti incontrano la concorrenza di altri ed infine selezionare mercati di prova per effettuare *test* di mercato, con risultati estendibili a tutti i mercati di quel gruppo. Il punto di partenza è la disponibilità di un campione di *n* unità rappresentato da *p* variabili; i dati sono raccolti in una matrice $n \times p$, in cui ogni riga rappresenta il profilo di un'unità statistica, che verrà sottoposta a elaborazioni successive.

Per comprendere la filosofia di base della CA, si può pensare a una rappresentazione in uno spazio metrico *p*-dimensionale. Ad ogni individuo è associato un vettore con *p* osservazioni, che rappresenta un punto in questo spazio. Gli eventuali addensamenti di punti, e quindi di unità, rappresentano i segmenti in cui viene suddivisa la popolazione.

Dal punto di vista procedurale, i primi passi sono la selezione degli elementi da analizzare e delle variabili di segmentazione. Dopo ciò, i raggruppamenti avvengono in modo diverso a seconda che si scelga un algoritmo aggregativo gerarchico o non gerarchico (detto anche di partizionamento iterativo)[4].

Nel primo caso a partire da *n* gruppi di 1 unità, si aggregano di volta in volta i due gruppi meno dissimili, ottenendo *n* partizioni concatenate. La classificazione in *g* gruppi è vincolata da quella in *g+1*, perché due unità, una volta fuse nello stesso gruppo, non possono essere separate. Questo è lo svantaggio del metodo gerarchico: l'ottimalità della partizione vale solo in riferimento alla partizione precedente. Esistono cinque algoritmi gerarchici che si distinguono per il criterio di valutazione delle distanze tra i gruppi, ovvero metodo del legame singolo, del legame completo, del legame medio del centroide e di Ward.

[4] In realtà esistono molte procedure di *clustering*, gli algoritmi gerarchici e non gerarchici sono i più importanti.

Negli algoritmi non gerarchici, invece, da una partizione iniziale delle unità in *G* gruppi, si spostano le unità, fino ad ottenere la partizione con la massima omogeneità interna e la minima omogeneità tra i gruppi. Di fatto gli algoritmi, come quello di McQueen e Foggy, allocano le unità al gruppo con il centroide più vicino, minimizzando implicitamente la devianza entro relativamente alle *p* variabili. Gli ottimi di questa procedura sono locali, a causa dell'arbitrarietà della partizione iniziale, perciò è buona norma ripetere l'applicazione dell'algoritmo con più partizioni e scegliere la migliore secondo la regola di ottimizzazione locale.

L'obiettivo della CA è individuare gruppi omogenei, ma non esiste alcuna regola generale che guidi l'analista nella scelta del numero di gruppi ottimale. Tuttavia alcuni metodi empirici, come la rappresentazione grafica dei gruppi e delle loro misure di dissomiglianza, le differenze tra le dissomiglianze in passi successivi e il *pseudo F*, possono aiutare nella scelta.

Come di norma, conclusa l'analisi, è necessario verificare la congruenza dei risultati ed interpretarli. Per esaminare se i *cluster* hanno un significato concreto, ovvero se i valori medi dei gruppi differiscono significativamente tra loro, rispetto ai valori medi usati per l'analisi dei raggruppamenti, ci si può avvalere del *test* di Arnold e, ma solo per algoritmi gerarchici, del coefficiente di correlazione cofenetico. Inoltre è utile confrontare i valori medi delle variabili di raggruppamento e i valori medi delle variabili utilizzate per descrivere i profili assunti nei *g* gruppi.

Tuttavia, anche applicando correttamente la procedura statistica, non è detto che il *clustering* sia poi concretizzabile; solo una conoscenza profonda della realtà che si sta studiando lo può garantire.

1.6.2 La tecnica flessibile: la *conjoint analysis* (COA).

La segmentazione flessibile si basa su valutazioni psicometriche di un set di prodotti alternativi, descritti come specifiche combinazioni di

modalità o livelli di attributi del prodotto in esame, sia esso presente sul mercato o in fase di progettazione. Essa ricorre all'integrazione di *conjoint analysis* e della simulazione del comportamento di scelta del consumatore. In pratica si chiede agli intervistati di ordinare profili di uno stesso prodotto, che differiscono per almeno una modalità di un attributo. In questo modo si riesce a stimare l'utilità associata a ciascuna caratteristica (utilità *parthworth*) e l'importanza relativa. Si ipotizza, a tal proposito, che i consumatori agiscano in modo razionale quando acquistano un prodotto, ovvero che scelgano l'alternativa che massimizza la loro utilità, rispettando i propri vincoli di bilancio. Inoltre si ipotizza che l'utilità di ogni bene derivi dagli attributi che lo compongono; da un punto di vista matematico, quindi, l'utilità complessiva di un bene è pari alla somma delle utilità *parthworth*. La segmentazione flessibile risulta più utile quando applicata a mercati con prodotti a forte coinvolgimento psicologico nella fase di acquisto, per i quali i consumatori valutano attentamente vantaggi e svantaggi di ogni alternativa presente nel mercato. La COA è una tecnica statistica multivariata che permette di comprendere e misurare i compromessi che i consumatori accettano scegliendo un'alternativa di prodotto. Infatti l'obiettivo è individuare la combinazione ottima di caratteristiche, potendo così prevedere le preferenze dei consumatori e individuare segmenti di mercato potenziali.

Da un punto di vista operativo, dopo aver selezionato un campione di consumatori, si individuano gli attributi rilevanti[5], che possibilmente devono essere incorrelati ed avere lo stesso numero di livelli. Quindi si definiscono i profili di prodotto, detti anche stimoli, da sottoporre al giudizio del campione, attraverso la determinazione di un piano degli esperimenti. Si può scegliere, infatti, di far giudicare tutti i possibili

[5] Spesso il prezzo è considerato dai consumatori come indice di qualità, perciò è bene porre molta attenzione quando lo si sceglie come attributo da sottoporre a giudizio. Ad esempio, se prezzo e qualità sono inseriti entrambi nello schema di rilevazione si rischia che il prezzo sia sottovalutato, oppure si rischia che in beni dal consto unitario elevato sia sopravvalutato.

stimoli (disegno fattoriale) o solo una selezione (disegno fattoriale frazionato), tenendo conto che, per garantire l'affidabilità delle stime, al minimo si possono somministrare un numero di profili pari al numero totale dei livelli – il numero degli attributi + 1. Ai consumatori nel campione si chiede di ordinare per la preferenza accordata i profili, rilevando i punteggi assegnati ad ogni attributo. Successivamente si stimano le utilità parziali attraverso un modello di utilità scelto in precedenza tra i modelli vettore, punto ideale e *parthworth*, in modo da garantire la massima corrispondenza tra i punteggi di preferenza rilevati sugli intervistati e quelli previsti[6]. Il modello può utilizzare dati individuali (si ha così a disposizione una base per applicare la CA) o aggregati. Si può, quindi, calcolare l'importanza relativa associata ad ogni attributo ed, infine, valutare l'utilità totale di ogni alternativa di prodotto.

Con la simulazione del comportamento di scelta del consumatore e in particolare con i criteri *first choice* e Bratford, Terry, Luce, è possibile ricostruire le preferenze, anche per le alternative non valutate direttamente.

La *conjoint analysis* permette di valutare la disponibilità dei consumatori a combinare tra loro diverse modalità o livelli degli attributi di un prodotto con procedure di rilevazione e stima abbastanza semplici. Essa si rivela particolarmente efficace nel lancio di nuovi prodotti, nel rilancio di prodotti già esistenti e nell'individuazione di nicchie. Tuttavia la COA presenta numerosi svantaggi. Essa può rappresentare il processo d'acquisto solo approssimativamente, infatti prodotti e servizi a forte contenuto di immagine non sono valutati analiticamente dai consumatori. Le ipotesi su cui si fonda, cioè lo schema addittivo che collega preferenze e utilità totale e l'assenza di interazioni tra gli attributi, condizionano fortemente i risultati. L'estrapolazione dei risultati, come l'estensione a modalità e caratteri non inclusi nell'indagine, non sempre è giustificata. Per esempio, la valutazione

[6] Per valutare l'adattamento si possono utilizzare la correlazione tra scelte previste e scelte osservate R di Pearson e τ di Kendall.

dell'attributo prezzo non è facile: intuitivamente a un livello basso corrisponde un'utilità maggiore, tuttavia esiste una soglia oltre la quale il prodotto può essere considerato di qualità scadente. Questo modifica le preferenze. Un valore di importanza relativa basso non sempre significa scarsa rilevanza per i consumatori, per esempio se le modalità di un attributo sono percepite simili hanno poco valore discriminante nella scelta. Infine la COA individua la combinazione ideale solo in relazione agli attributi utilizzati; inserire ulteriori attributi, o diversi, può variare i risultati, ma soprattutto, se non sono stati inseriti nel disegno sperimentale alcuni attributi chiave del prodotto, è bene essere molto cauti in fase previsiva.

Capitolo 2

I modelli a classi latenti.

1.1 I modelli a classi latenti.

I modelli a classi latenti[7] (d'ora in poi mcl) appartengono alla più ampia famiglia dei modelli a variabili latenti. Essi sono molto simili ai modelli fattoriali, ma si applicano a variabili di tipo categoriale.

Furono introdotti inizialmente da Lazarsfeld e Henry (1968) per misurare variabili latenti attitudinali a partire da *item* dicotomici. La novità che, al contrario dell' analisi fattoriale che utilizza solo variabili continue[8], fossero applicabili a dati dicotomici portò ad un raggio d' azione più ampio, ma solo più tardi l'utilizzo di questi modelli si diffuse, con i lavori di Goodman (1974°, 1974b) che formalizzò la metodologia dei mcl, estendendone l'applicazione anche a variabili nominali ed elaborando l'algoritmo di stima di massima verosimiglianza, usato anche nei moderni *software*. Negli anni, poi, furono introdotte estensioni per variabili ordinali (Heinen, 1996), indicatori continui e variabili su scale differenti, ovvero nominali, ordinali e continue (Vermunt e Madigson 2001) e covariate. Recentemente l'applicazione dei mcl ha conosciuto un'ulteriore diffusione grazie ai moderni *software* che ne permettono una stima agevole.

L'ipotesi di base dei mcl è che le caratteristiche osservate nelle variabili a disposizione possano essere riassunte da ulteriori caratteristiche latenti di queste. I mcl, quindi, mettono in relazione una serie di variabili

---

[7] Alcuni autori, fra cui Madigson e Vermunt (2003), identificano i modelli a classi latenti con i modelli mistura (*finite mixture* m*o*del), altri invece considerano i mcl come una particolare specificazione dei modelli di mistura.

[8] In pratica, come afferma Lazersfeld (1951), l'analisi delle classi latenti fa con le variabili categoriali ciò che l'analisi fattoriale fa con quelle cardinali, ovvero ne applica gli stessi principi senza violare la natura delle variabili cardinali (in altre parole non è necessario che le relazioni tra variabili manifeste abbiano distribuzione multinormale).

osservate discrete categoriali multivariate con un insieme di variabili latenti discrete categoriali, le cui modalità sono definite classi. Ogni classe è caratterizzata da un insieme di probabilità condizionate che indicano la probabilità che le variabili assumano un determinato valore. I casi, ovvero le unità statistiche, sono assegnate alle classi su base probabilistica, ovvero sulla base della loro probabilità di appartenere a una determinata classe, creando così dei gruppi che sono mutuamente indipendenti. Pertanto i mcl permettono di ridurre una popolazione eterogenea in sottogruppi di unità omogenei al loro interno ed eterogenei fra di loro, senza fare assunzioni restrittive sui dati (linearità della relazione, normalità, omogeneità), divenendo così meno soggetti a distorsioni dovute a non conformità dei dati con le ipotesi iniziali. Inoltre, la possibilità di inserire nell'analisi delle covariate permette un'efficace descrizione dei gruppi, senza dover utilizzare in un secondo momento l'analisi discriminante, per stabilire un criterio secondo il quale assegnare correttamente ulteriori unità ai segmenti, precedentemente individuati.

I vantaggi dell'utilizzo dei modelli a classi latenti rispetto a modelli più tradizionali sembrano evidenti: costruiscono gruppi internamente omogenei e mutuamente indipendenti, non fanno assunzioni restrittive sui dati e permettono l'uso di covariate. Tutto ciò, unito alla creazione di *software ad hoc* per la stima, è forse il motivo della recente diffusione del loro impiego nelle indagini di *marketing*, e in particolare in quelle di segmentazione, ma anche sociali e biomediche. Non è difficile pensare che le classi di una variabile latente possano essere i segmenti di un mercato.

La tecnica di segmentazione dei mcl è predittiva a posteriori, pertanto i segmenti individuati non sono il risultato di un disegno degli analisti precedente all'analisi, ma sono il risultato dell'analisi stessa. Inoltre tali modelli permettono di stimare simultaneamente la scelta dell'individuo e

l'appartenenza ad un segmento latente, con dei risultati facilmente interpretabili e, evidentemente, informativi.
In questo capitolo saranno descritte tre particolari specificazioni dei mcl, ovvero i modelli tradizionali, fattoriali e multigruppo.

2.2 I modelli a classi latenti tradizionali.

L'analisi a classi latenti tradizionale (Goodman, 1974) ha l'obiettivo di individuare il minor numero di classi latenti *T*, in grado di spiegare le associazioni osservate tra le variabili manifeste, a partire dai dati presenti in una tabella di contingenza ad entrata multipla (una per ogni variabile osservata).
Il mcl assume che ogni osservazione appartenga a una e solo una delle *T* classi latenti e che tra le variabili osservate esista l'indipendenza locale, ovvero che condizionatamente all'appartenenza alla classe latente le variabili osservate siano mutuamente indipendenti (in altre parole all' interno di ogni classe latente, le variabili manifeste sono indipendenti e l'associazione fra di esse è spiegata dalle classi della variabile latente).

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* individui soggetti di un'indagine (indicizzati con *i=1,...,N*) e una variabile latente *X*, con *T* (*t=1,...,T*) classi; allora il mcl può essere formulato come segue:

(1)

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

dove $P(Y_i)$ indica la probabilità che l'individuo *i* abbia il vettore di risposte Y_i, $P(Xi = t)$ è la probabilità che l' individuo *i* appartenga alla classe *t* e $P(y_{ik}|X_i = t)$ è la probabilità che l' individuo *i* abbia dato la risposta y_{ik} alla *k-sima* variabile manifesta, dato che appartiene alla classe latente *t*. I parametri da stimare ϑ_k definiscono la distribuzione della variabile osservata *k* all' interno della classe latente *t*.
Il mcl tradizionale può anche essere rappresentato graficamente con un *path diagram* (Figura 2.1), nel quale le variabili manifeste sono connesse tra loro solo attraverso la variabile latente *X*, che spiega tutte le associazioni tra le variabili osservate.

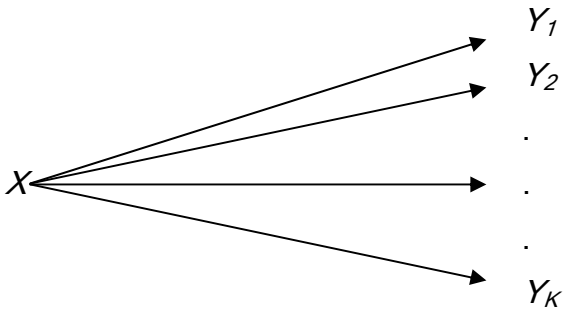

Fig.2.1: *Path diagram* di un mcl tradizionale con *K* variabili manifeste e una variabile latente.

Questo corrisponde all'ipotesi di indipendenza locale, secondo la quale all'interno di ogni classe latente le variabili osservate sono mutuamente indipendenti (e quindi non associate); ovvero nella classe latente *t*, la probabilità di ottenere la risposta *s* nella *k-sima* variabile è indipendente dalla probabilità di ottenere la risposta *r* nella variabile *j*, $k \neq j$. Nel modello (1) l'indipendenza locale è evidente, poiché la probabilità congiunta di risposta è il risultato del prodotto delle *K* singole probabilità condizionate di risposta.

2.2.1 Procedura di stima di un modello a classi latenti tradizionale.

L'analisi ha inizio con la stima di un modello, detto H_0, in cui le variabili sono mutuamente indipendenti, il che corrisponde a fissare *T=1*, ovvero

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

Assumendo che il modello nullo non fornisce un'adeguata stima dei dati, si stima un secondo modello con *T=2* classi. Si procede, quindi, a stimare modelli, aumentandone di volta in volta la dimensione, incrementando di 1 il numero delle classi latenti, finché non si perviene al modello più adeguato ai dati.

2.2.2 Specificazione delle distribuzioni.

La forma distributiva delle y_{ik} dipende dalla scala delle variabili osservate incluse nel modello, che possono essere categoriali (nominali o ordinali), continue o di conteggio.
Per le variabili di conteggio si utilizza, solitamente, una distribuzione di Poisson o Binomiale, per quelle continue una distribuzione Normale e, infine, per quelle categoriali si utilizza la distribuzione multinomiale, in particolare per le nominali si utilizza la logistica multinomiale e per le ordinali la logistico ordinale per categorie adiacenti.
Per l'assunzione di indipendenza locale le variabili osservate sono mutuamente indipendenti, condizionatamente alla variabile latente *X*, perciò di seguito si presenterà la forma distributiva per una generica variabile y_{ik}, sapendo che la distribuzione congiunta della variabile y_k sarà data dal prodotto di tutte le y_{ik} per *i=1,...,N*.
Per variabili manifeste di tipo nominale o ordinale, la distribuzione sarà di tipo multinomiale, ovvero

(2) $$P(y_{ik} = s | X_i = t) = \frac{\exp\left\{\eta^k_{s|t}\right\}}{\sum_{s'=1} \exp\left\{\eta^k_{s'|t}\right\}}$$

dove *s* (*s=1,...,*S^k) indica una particolare categoria di y_{ik} e S^k numero di categorie di ogni y_k. $P(y_{ik} = s|X_i = t)$ è la probabilità di rispondere esattamente *s*, data la variabile latente *X*, mentre $\eta^k_{s|t}$ è il termine lineare, la cui diversa specificazione distingue i modelli logit mulltinomiale e logit ordinale per categorie adiacenti.

Infatti, per il modello logit multinomiale, il predittore lineare è dato da:

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

dove il primo termine a destra è l'intercetta e il secondo è specifico per ogni classe latente *t*, mentre per il modello logit ordinale per categorie adiacenti il predittore è dato da (Agresti,2002):

$$\eta^k_{s|t} = \beta^k_{s0} + \beta^k_{.t0} y^{k*}_s$$

dove y^{k*}_s è un punteggio assegnato alla categoria *s* della *k-sima* variabile osservata.

L'analista è chiamato ad interpretare gli S^k*-1* logit per categorie adiacenti:

$$\log\left(\frac{P(y^k = s+1|X = t)}{P(y^k = s|X = t)}\right) = \eta^k_{s+1|t} - \eta^k_{s|t} = \beta^{.k}_{s0} + \beta^k_{.t0}(y^{k*}_{s+1} + y^k_s)$$

dove $\beta^{.k}_{s0} = \beta^k_{s+1,0} - \beta^{s0}_k$.

Come visto nell'equazione (1), la funzione di probabilità corrispondente alle risposte del soggetto *i-simo* è composta da due probabilità, una per la variabile latente e una per la dipendente. Quindi, dopo aver visto la forma distributiva per la probabilità condizionata, bisogna definire la forma distributiva della variabile latente *X*, anch'essa legata alla sua natura, nominale o ordinale.

La variabile latente ha ancora forma multinomiale ed è parametrizzata come segue:

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

Se si dispone di una sola variabile latente, si ha un modello logit multinomiale standard, dove 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 $\sum_{t=1}^{T} \gamma_{t0}$.

2.2.3 L'uso di covariate.

Un'importante estensione del mcl tradizionale è la possibilità di inserire delle covariate, sia per la classe latente sia per le variabili risposta, molto utili per la descrizione delle classi.

Le forme distributive per modelli contenenti covariate per classi latenti e variabili osservate sono regressione logistica multinomiale per la variabile latente *X* e una regressione appartenente alla famiglia dei modelli lineari generalizzati (glm) per le variabili risposta.

Essendo Z_i il vettore contenente *R* covariate per l'individuo *i*, l'espressione più generale per un modello con covariate per entrambe le variabili è

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

Essendo z_{ir} la r-sima covariata per l'individuo *i* (*r=1,...,R*), se gli indicatori sono nominali, i predittori lineari per la distribuzione condizionata sono

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

e per la probabilità della variabile latente sono

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

con i rispettivi vincoli per i parametri (Vermunt e Madigson, 2005).

Ciò che distingue gli indicatori dalle covariate è il fatto che la variabile latente spiega le associazioni tra gli indicatori, ma non quelle tra le covariate.

2.2.4 Misure di valutazione dell'adattamento della stima del modello.

Esistono diversi approcci complementari per giudicare l'adeguatezza del modello ai dati.

L'approccio più usato è la statistica rapporto di verosimiglianza L^2, che misura quanto le stime di massima verosimiglianza per le frequenze attese, $\hat{F}_i$, differiscono dalle frequenze osservate corrispondenti, f_i.

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

Secondo il criterio L^2, un modello si adatta ai dati se il valore di L^2 è sufficientemente basso da essere attribuibile al caso (generalmente 0.5).

Le stime di massima verosimiglianza per le frequenze attese $\hat{F}_i$ si ottengono con il seguente processo a due stadi. Innanzitutto, sostituendo sulla parte destra dell'equazione (1) le stime ML per i parametri del modello, si ottengono le stime ML delle probabilità di appartenere a una classe latente. Le probabilità stimate sono, poi, sommate per ogni classe latente, ottenendo così le stime della probabilità per ogni cella, e moltiplicate per la numerosità campionaria *N*. Si hanno, quindi, le stime ML per le frequenze attese.

Nel caso in cui le frequenze attese corrispondano perfettamente a quelle osservate, il modello stima perfettamente i dati e L^2 sarà uguale a 0; nel caso in cui $L^2>0$, esso misura la mancanza di adeguatezza del modello, quantificando l'associazione (non indipendenza) non spiegata dal modello.

Sotto opportune condizioni di regolarità, L^2 asintoticamente ha distribuzione χ^2 con gradi di libertà pari al numero di celle nella tabella multientrata meno il numero di parametri distinti del modello *M* meno 1.[9]

[9] Più specificamente, il numero di celle nella tabella multientrata è dato dal prodotto di tutte le modalità di ognuna delle *k* variabili manifeste, ovvero se ogni variabile

Nel caso di dati sparsi[10], la distribuzione χ^2 non va usata per calcolare il *p-value*, perché L^2 non potrebbe non essere ben approssimato, ma si può usare il *bootstrap approach* (Vermunt e Madigson,2002).
Un approccio alternativo al rapporto di verosimiglianza è l'uso di criteri informativi che tengano conto sia della bontà della stima sia della parsimonia del modello, ovvero il criterio informativo di Akaike (AIC) e il criterio informativo Bayesiano (BIC).
Essendo *gl* i gradi di libertà della statistica L^2, BIC si definisce:

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

Una seconda formulazione più generale di BIC si basa sulla log-verosimiglianza LL e sul numero di parametri *M* ed è

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

In generale, un modello con un valore di BIC basso è preferibile ad uno con BIC più alto.
Analogamente a BIC, un modello che presenta AIC più basso è preferibile.
Se il modello H_0 fornisce un'adeguata stima dei dati, non è necessaria nessuna analisi a classi latenti. Nella maggior parte dei casi, tuttavia, il modello nullo non si adatta bene ai dati, ma il valore di L^2 ad esso associato è usato, comparato a quello associato con i successivi mcl (con *T>1*), come indice dell'adattamento della stima ai dati. La

---

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*, invece è dato da $M = T - 1 + T\sum_{k=1}^{K}(n_k - 1)$. Infine $gl = N_k - M - 1.$

La regola generale dice che se *gl<0*, il modello non è identificato.

[10] Si hanno dati sparsi quando il numero di variabili osservate o il numero delle loro categorie è molto alto, oppure quando il modello è esteso a variabili continue.

percentuale di riduzione di L^2 tra modello nullo e uno con *T>1* rappresenta l'associazione totale spiegata da quest'ultimo modello. Il modello H_0, infatti, è il modello in cui l'associazione fra le variabili non è affatto spiegata e L^2 ad esso associato quantifica l'associazione totale nei dati.
Quest'ultima misura di adattamento è un approccio meno formale rispetto agli altri presentati, ma può integrare misure più statisticamente precise, come L^2 e BIC.

2.2.5 La significatività degli effetti.

Come in ogni altra analisi statistica, trovato il modello con il numero di classi che meglio si adatta ai dati, si valutano gli effetti delle variabili nel modello. Tutte quelle variabili che non sono significative, o meglio che tra le classi non variano in modo significativo, vengono eliminate.
Per verificare la significatività di una variabile *k*, si testa l'ipotesi nulla che la distribuzione delle sue *s* categorie sia identica in ognuna delle *t* classi:

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

Per implementare questo test, si usa la relazione fra le probabilità di risposta condizionata e i parametri log-lineari.
Una statistica utilizzabile a tal scopo è la differenza di L^2, ΔL^2 che è computato come la differenza di L^2 del modello ristretto e quello non ristretto, ovvero senza e con la variabile sotto esame. Sotto l'ipotesi nulla che il modello non ristretto sia vero, la statistica si distribuisce secondo un χ^2 con gradi di libertà pari al prodotto della differenza delle categorie n_k della variabile sotto esame meno 1 e il numero di classi meno 1, ovvero $df = (n_k-1)(T-1)$.
Un altro modo per testare la significatività dei parametri degli indicatori è il test di Wald, che verifica se i coefficienti di regressione sono uguali tra le classi. Tuttavia, il test di Wald è meno potente di ΔL^2. Sotto

l'ipotesi nulla che il modello non ristretto sia vero, il test di Wald ha la medesima distribuzione del test ΔL^2.

2.2.6 Classificazione.

Il passo finale dell'analisi a classi latenti è utilizzare i risultati ottenuti per classificare gli individui (casi) nella classe latente appropriata. A questo scopo si ricorre al Teorema di Bayes, calcolando la probabilità a posteriori che un individuo appartenga alla classe *t*, dato il suo *pattern* di risposte, ovvero

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

dove numeratore e denominatore sono ottenuti sostituendo ai parametri del modello (1) le loro stime.
I casi sono assegnati alle classi per le quali la probabilità a posteriori è maggiore.
Madigson e Vermunt (2001) e Vermunt e Madigson (2002) si riferiscono al mcl tradizionale come *latent class cluster model*, per sottolineare che l'obiettivo dell'analisi, classificare le unità in *T* gruppi omogenei, è comune alla *cluster analysis*. Mentre in quest'ultima per definire l'omogeneità fa uso di misure di distanze *ad hoc*, l'analisi a classi latenti la definisce in termini di probabilità: i casi appartenenti alla medesima classe latente sono simili, perché generati dalla stessa distribuzione di probabilità.

2.3 I modelli a classi latenti non tradizionali.

Talvolta può accadere che un mcl tradizionale con *T* classi sia rifiutato per mancanza di adeguatezza, dovuta al non sussistere dell'ipotesi di indipendenza locale. In tal caso la strategia usuale per ovviare al problema è aggiungere una classe latente, stimando un modello con *T+1* classi. Tuttavia esistono delle strategie alternative che portano a un modello più parsimonioso e più congruente alle ipotesi iniziali. Esse

consistono nell'aggiungere uno o più effetti diretti, eliminare *item* (y_{ik}) oppure aumentare il numero delle variabili latenti.

La prima alternativa prevede di includere nel modello dei parametri di effetti diretti, che spieghino le associazioni residue fra le variabili osservate, responsabili della dipendenza locale. È particolarmente indicato quando alcuni fattori esterni, incorrelati con la variabile latente, creano una rilevante associazione tra due variabili.

La seconda alternativa è particolarmente indicata nelle situazioni in cui due variabili sono responsabili della dipendenza locale, che si può eliminare semplicemente depennando una delle due variabili. Se le variabilisono ridondanti, la strategia è ancora più efficace.

La terza alternativa, infine, si adatta a quei casi in cui un gruppo di variabili è responsabile della dipendenza. Secondo Madigson e Vermunt (2001), il modello a classi latenti fattoriale, che si ottiene aumentando il numero di variabili latenti, anziché di classi, spesso si adatta meglio ai dati, senza bisogno di parametri aggiuntivi. Inoltre il modello fattoriale è identificato in casi in cui il modello tradizionale non lo è.

Una statistica diagnostica molto utile per determinare quale strategia scegliere è *BVR* (*Bivariate Residual*), che aiuta a definire le relazioni bivariate non adeguatamente spiegate dal modello: BVR misura quanta associazione osservata tra due variabili è spiegata dal modello. Partendo da una tabella a due entrate per le due variabili oggetto di studio, la BVR si calcola dividendo la statistica chi quadrato di Pearson, usata per il test di indipendenza, per i gradi di libertà, quindi confrontando le frequenze osservate con le corrispondenti frequenze attese del mcl implementato. Se BVR è minore di 1, il modello non spiega l'associazione tra le due variabili considerate.

2.4 **Modelli a classi latenti fattoriali.**

I modelli a classi latenti fattoriali (mclf) furono inizialmente proposti da Goodman (1974a) nel contesto dell'analisi a classi latenti confermatoria

e poi riproposti da Madigson e Vermunt (2001) come alternativa al modello a classi latenti tradizionale esplorativo.

Essi consistono nell'includere più di una variabile latente nel modello, aumentando, quindi, i fattori, che possono essere interpretati come una variabile congiunta (Goodman 1974b). I mclf sono usati come nell'analisi fattoriale tradizionale, nella quale variabili latenti multiple sono utilizzate per modellare le relazioni multidimensionali esistenti fra le variabili manifeste.

Per illustrare questo tipo di mcl si userà un esempio; supponiamo, allora, di avere una variabile latente *X* con *T=4* classi. *X* può essere espressa attraverso due variabili latenti dicotomiche $V = \{1, 2\}$ e $W = \{1, 2\}$, usando la corrispondenza:

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

Si ha, quindi, che *X=1* corrisponde a *V=1* e *W=1*, *X=2* a *V=1* e *W=2*, *X=3* a *V=2* e *W=1* e *X=4* a *V=2* e *W=2*.

Madigson e Vermunt (2001) considerano varie tipologie di modelli a classi latenti fattoriali, che si distinguono in base a diverse restrizioni.

Il mclf base è un mcl che contiene due o più fattori dicotomici e mutuamente indipendenti e che esclude interazioni di ordine superiore nelle probabilità condizionate di risposta. Il mclf base con *R* fattori ha esattamente lo stesso numero di parametri distinti di un mcl tradizionale con *R+1* classi, quindi si possono specificare mclf con 2^R classi ed esattamente lo stesso numero di parametri di un modello tradizionale con *R+1* classi latenti. Il mclf, quindi, risulta più parsimonioso e più facile da interpretare nonché più adatto all'analisi esplorativa e identificato in molte più situazioni rispetto al mcl tradizionale.

Inoltre, anche il mclf può essere esteso all'uso di covariate e si può usare con variabili di tipo categoriale, di conteggio e combinazioni di queste.

Formalmente, un mclf per quattro variabili manifeste *A*, *B*, *C*, *D*, e due fattori *V* e *W* si può formulare così:

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

dove π_{rs}^{VW} rappresenta la probabilità che *V* e *W* assumano rispettivamente i valori *r* (*r=1,...,R*) e *s* (*s=1,...,S*) e $\pi_{irs}^{A|VW}$ la probabilità condizionata che nell'*item A* si risponda i, dato che *V=r* e *W=s*, e $\pi_{jrs}^{B|VW}$, $\pi_{krs}^{C|VW}$, $\pi_{lrs}^{D|VW}$ sono le probabilità condizionate per gli items *B*, *C*, *D* rispettivamente.

2.5 Modelli a classi latenti multilivello.

I modelli a classi latenti multilivello sono un'estensione dei mcl utile quando le osservazioni non sono indipendenti e presentano una struttura gerarchica, ovvero quando i casi sono raggruppabili in gruppi (per esempio pazienti raggruppati in ospedali, dati economici per paesi europei). Nall'analisi a classi latenti può risultare utile ed efficace tener conto di questa struttura gerarchica, combinando i modelli a classi latenti con l'analisi multilivello standard.

Supponiamo di avere dei dati che presentino una struttura gerarchica a due livelli, sebbene possano presentare anche strutture più articolate; distinguiamo gli individui, o unità di livello uno, dai gruppi, o unità di livello due.

In un mcl tradizionale, si assume che i parametri del modello non varino per gli individui (unità di livello uno), mentre nei mclm si suppone che alcuni parametri possano variare tra i gruppi e lo facciano secondo due diverse modalità: con un approccio *fixed-effects*, che prevede l'introduzione di *dummies* di gruppo nel modello (Clogg e Goodman,

1984), e con un approccio *random-effects*, che prevede che i coefficienti specifici di gruppo seguano una particolare distribuzione, i cui parametri devono essere stimati.

2.5.1 Modelli a classi latenti multilivello a effetti fissi.

I modelli *fixed-effects* sono paragonabili ai modelli a classi latenti multigruppo (Clogg e Goodman, 1974).
Un modello a classi latenti semplice per dati multilivello con *J* gruppi, indicizzati da *j=1,...,J*, è:

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

dove Y_{ij} è il vettore di risposte dell'individuo *i* appartenente al gruppo *j*, *s* un possibile *pattern* di risposte, X_{ij} è una variabile latente con *T* classi, y_{ijk} è la risposta dell'individuo *i* appartenente al gruppo *j* all'*item k* e s_k ($s_k=1,...,S_k$) un particolare livello dell'*item k*. La probabilità di osservare un particolare pattern di risposte è la media ponderata delle probabilità specifiche di classe per la probabilità che l'unità *i* del gruppo *j* appartenga alla *t-sima* classe latente. Le osservazioni y_{ijk} sono supposte indipendenti, condizionatamente all'appartenenza a una determinata classe latente (indipendenza locale).
Ciò che rende questo modello un modello multilivello sono le diverse assunzioni sulla forma distributiva dei parametri del modello:

(4) $$P(X_{ij}=t)=\frac{\exp\{\gamma_{tj}\}}{\sum_{r=1}^{T}\exp\{\gamma_{rj}\}} \text{ e } P(y_{ijk}=s_k\mid X_{ij}=t)=\frac{\exp\{\gamma_{s_k}^{k}\}}{\sum_{r=1}^{S_k}\exp\{\gamma_{rtj}^{k}\}}.$$

Come si può notare nella specificazione *logit* delle probabilità, rispetto alle equazioni (2) e (3) compare l'indice di gruppo *j*, perché in questo caso i parametri del modello variano tra i gruppi.

Per vedere meglio il legame dei modelli con la variabile di secondo livello, si può considerare una seconda specificazione del modello. Sia *G* la variabile osservata di secondo livello, che identifica il gruppo di appartenenza delle unità statistiche, si ha allora:

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

Senza ulteriori restrizioni, questo modello è equivalente a un mcl multigruppo non ristretto (Clogg e Goodman, 1984). Un modello più ristretto si ottiene fissando i parametri della probabilità condizionata nei gruppi.

In questo modello il numero di parametri da stimare è pari al numero dei gruppi; appare, quindi, chiaro che se il numero dei gruppi è elevato (e il numero di individui per gruppo esiguo) il modello diviene piuttosto complesso: oltre al numero di parametri da stimare, che cresce velocemente con il numero di unità di livello due, le stime risultano instabili, a causa delle ampiezze dei gruppi (caratteristica tipica delle analisi multilivello) soprattutto quando i casi nei gruppi sono in numero esiguo (in generale se $J \geq 50$ e $n_j \leq 30$). Inoltre, poiché tutte le differenze tra i gruppi sono spiegate dalle *dummies* di gruppo, non è possibile determinare gli effetti delle covariate dei gruppi sulla probabilità di appartenere a una determinata classe latente.

2.5.2 Modelli a classi latenti multilivello ad effetti casuali.

Per ovviare ai problemi dell'approccio ad effetti fissi, si può adottare l'approccio a effetti casuali, che prevede che gli effetti specifici dei gruppi seguano una certa distribuzione. Nel modello a classi latenti si introduce una variabile latente, che può essere continua, ovvero uno o più effetti casuali a livello gruppo, o discreta, dove i parametri variano tra le diverse classi latenti dei gruppi (Vermunt, 2007). Nel primo caso si ha un mcl multilivello di tipo parametrico, perché la variabile si assume Normale, nel secondo non parametrico, perché la variabile latente si assume Multinomiale.

L'ipotesi sottostante a questo tipo di modelli è che le osservazioni nei gruppi siano correlate, perché i membri dello stesso gruppo tendono ad appartenere alla stessa classe latente.

Per quanto riguarda l'approccio non parametrico, quindi con variabile latente a livello due continua, si ha che il predittore ricavato dalla probabilità condizionata (4) ha forma $\gamma_{tj} = \gamma_t + \tau_t u_j$ con $u_j \sqcup N(0,1)$ e due restrizioni zero, una per τ_t e una per γ_t. L'assunzione di base è che i componenti casuali nei γ_{tj} sono perfettamente correlati, in particolare lo stesso effetto casuale u_j è scalato in maniera differente per ogni *t* dal parametro non noto τ_t. Questa formulazione suppone che ogni categoria nominale sia correlata a una tendenza di risposta latente sottostante.

L'approccio parametrico fa delle assunzioni distributive molto forti sul modello degli effetti casuali. Un'alternativa meno impegnativa è utilizzare una distribuzione discreta non specificata, che comporta definire una variabile latente anche per le unità di secondo livello, oltre che per quelle di primo. Tale modello fa assunzioni distributive meno forti, è più semplice dal punto di vista computazionale e spesso si adatta in modo migliore al problema di ricerca (Vermunt e Van Dijk, 2001). In molti casi sembra più naturale classificare i gruppi in pochi *cluster*, piuttosto che definirli su scala continua.

L'idea di fondo all'approccio non parametrico è che ogni gruppo appartenga a una fra le *M* classi latenti della variabile latente di secondo livello *D*. Sia D_j la classe di appartenenza del gruppo *j* e *m* (*m=1,...,M*) una particolare classe latente con $1 \leq D_j = m \leq M$ (Vermunt, 2003b).

Un modello multilivello a classi latenti può essere visto come la combinazione di due misture di componenti: una al primo livello e una al secondo.

A livello individui, in generale la probabilità di risposta Y_{ij}, condizionatamente all'appartenenza del gruppo *j* alla classe latente *m*, si può modellare come segue:

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

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

Si nota che i diversi gruppi differiscono sia per la probabilità che i loro componenti appartengano a una classe latente *t* sia per i parametri che definiscono le probabilità di risposta condizionate.

A livello gruppo, i legami fra gli individui appartenenti allo stesso gruppo è dato da:

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

ipotizzando che le n_j risposte degli individui siano mutuamente indipendenti, condizionatamente all'appartenenza del gruppo *j* alla classe latente *m*.

Combinando le due espressioni precedenti, si ottiene un modello a classi latenti multilivello a effetti casuali discreti:

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

.

Il lato a destra dell'equazione è formata da tre componenti, ovvero 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 dia una determinata risposta y_{ijk}, data la sua appartenenza alla classe latente *t*. Quindi la probabilità di osservare una risposta è la media pesata, dove i pesi sono le probabilità di appartenenza a una classe latente dei gruppi e degli individui.

Per quanto riguarda i predittori, si ha:

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

con la possibilità di scrivere $\gamma_{tm} = \gamma_t + u_{tm}$, dove u_{tm} segue una distribuzione non specificata.

Di fatto la struttura gerarchica dei dati ha tre livelli, che sono rispettivamente risposte multiple, individui e casi, così come un mcl classico ha due livelli (risposte multiple e individui).

È possibile estendere ulteriormente il modello parametrico, includendo covariate per predire l'appartenenza al livello uno e due. Supponendo di avere covariate di secondo livello, che chiameremo Z_{1j}, e di primo livello, che chiameremo Z_{2ij}, si ha

$$P(X_{ij} = t \mid Z_{1j}, Z_{2j}) = \frac{\exp\{\gamma_{0tj} + \gamma_{1t} Z_{1j} + \gamma_{2t} Z_{2tj}\}}{\sum_{r=1}^{T} \exp\{\gamma_{0rj} + \gamma_{1r} Z_{1j} + \gamma_{2r} Z_{2ij}\}}.$$

Questo modello è un'estensione del modello a classi latenti con variabili concomitanti, che contiene effetti fissi (quelli delle covariate di primo livello) e casuali (intercetta e i coefficienti delle covariate di secondo livello). Sostituendo γ_{2t} con γ_{2tj} e facendo delle assunzioni distribuzionali su γ_{2tj}, si ottiene un modello con anche la pendenza casuale.

Anche il modello non parametrico può essere esteso a covariate di livello uno e due. Un esempio è

$$P(X_{ij} = t \mid Z_{1j}, Z_{2j}) = \frac{\exp\{\gamma_{0tm} + \gamma_{1t} Z_{1j} + \gamma_{2tm} Z_{2ij}\}}{\sum_{r=1}^{T} \exp\{\gamma_{0rm} + \gamma_{1r} Z_{1j} + \gamma_{2rm} Z_{2ij}\}}.$$

In questo modello, sia l'intercetta che la pendenza delle covariate di livello uno sono assunte dipendenti da D_j.

Capitolo3

Presentazione del *dataset* e analisi preliminari sui dati.

3.1 Introduzione.

L'obiettivo di questa tesi è segmentare con l'utilizzo dei modelli a classi latenti il mercato europeo dei prodotti finanziari.
Si utilizzano a tal fine dati provenienti dalla seconda indagine SHARE, relativa al 2006.

3.2 L'indagine SHARE.

SHARE (acronimo dall'inglese Survey of Health, Ageing, and Retirement) è una bancadati multidisciplinare e multipaese su salute, invecchiamento e pensioni in Europa. In particolare raccoglie dati su salute, *status* socioeconomico, relazioni sociali e familiari degli ultracinquantenni. SHARE è coordinata a livello centrale presso il *Mannheim Research Institute for the Economics of Aging*. Il progetto nasce nel 2004 con un'indagine riguardante undici paesi europei, tra i quali alcuni scandinavi (Danimarca, Svezia), alcuni centrali (Austria, Francia, Germania, Svizzera, Belgio, Olanda) e alcuni mediterranei (Spagna, Grecia e Italia). Successivamente si sono aggiunti Israele nel 2005/2006, Repubblica Ceca, Polonia (nuovi entrati nell'UE) e Irlanda nel 2006. Questi quattordici paesi hanno partecipato al secondo studio nel 2007/2008, anno in cui si è aggiunta anche la Slovenia, che partecipa al terzo studio del 2008/2009. Il disegno dell'indagine prende a modello la statunitense HRS (*Health and Retirement Study*) e l'inglese ENSA (*English Longitudinal Study of Ageing*). L'innovazione di SHARE, rispetto a HRS ed ENSA, è la sua struttura multipaese, che permette di coprire politiche di *welfare*, culture e storie diverse.

I dati comprendono variabili di salute (ad esempio stato di salute percepito, utilizzo di strutture mediche, funzionalità fisica e cognitiva, presenza di comportamenti a rischio), psicologiche (benessere, livello di soddisfazione), economiche (occupazione pre e post pensionamento, fonti e composizione del reddito, istruzione, ricchezza e consumo) e di interazione sociale (assistenza all'interno della famiglia, relazioni sociali, attività di volontariato).

3.2.1 Popolazione di interesse.

La popolazione di interesse è definita in termini di individui e famiglie. In particolare la popolazione degli individui è definita come "l'insieme degli individui nati prima del 1954[11] che parlano la lingua ufficiale del paese e che non vivono durante il periodo di indagine all'estero o in un'istituzione come una prigione, e le loro spose/partner indipendentemente dall'età.", mentre quella delle famiglie "l'insieme delle famiglie con almeno una persona nata prima del 1954, che parla la lingua ufficiale del paese e che non vive, durante il periodo di indagine, all'estero o in un'istituzione come una prigione." (Börsch-Supan, Jürges, 2005). Alcuni paesi sono riusciti a includere nel campione individui che abitano in case per anziani. Nella seconda indagine, da cui sono tratti i dati per questa tesi, la popolazione obiettivo è la medesima.

3.2.2 Il campionamento.

Ogni paese partecipante è responsabile del disegno di campionamento riguardante la propria nazione. Sono utilizzati tre diversi tipi di campionamento: estrazione di un campione casuale semplice (stratificato) usando registri nazionali, campionamento a più stadi usando registri di popolazione locali o regionali (ad esempio in Italia si utilizzano le liste elettorali), campionamento a uno o più stadi basato sugli elenchi telefonici e seguito da uno *screening* sul campo. Per alcuni

[11] In Germania 1953.

paesi le unità finali di selezione sono le famiglie, per altri gli individui. Nonostante il disegno non sia univoco, ogni campione è estratto in modo da essere rappresentativo del paese a cui si riferisce.

3.2.3 Il questionario.

Il questionario SHARE si compone di tre parti. La prima, detta *coverscreen* o reperimento, è condotta sia a livello familiare che individuale e prevede la comunicazione da parte di un intervistatore dell'oggetto di indagine e la raccolta delle prime informazioni anagrafiche, allo scopo di definire l'idoneità dei contattati all'intervista. La seconda parte, la principale, è eseguita faccia a faccia da un intervistatore, che utilizza anche cartellini contenenti le opzioni di risposta. L'intervista principale è strutturata in ventitre moduli, che riguardano l'individuo o la famiglia, e per i quali si necessitano di quattro diversi rispondenti, individuati durante il reperimento (rispondente principale, rispondente per la parte finanziaria, rispondente per la parte riguardante la famiglia e rispondente per la parte relativa all'abitazione). Le informazioni raccolte comprendono variabili di salute, variabili psicologiche, economiche e di interazione sociale, e sui decessi avvenuti dopo la prima rilevazione. Le prime due fasi fanno uso di strumenti CAPI (*Computer Assisted Personal Interview*). L'ultima parte è a sua volta composta da tre questionari autocompilati su salute fisica e psicologica, assistenza sanitaria e rete locale.

3.3 Le variabili utilizzate per lo studio.

Il campione SHARE conta 23238 unità statistiche, sulle quali sono state rilevate numerose variabili. Al fine di indagare il possesso di prodotti finanziari, ne sono state selezionate diciassette, di seguito classificate secondo la tipologia di informazioni che forniscono.
Due variabili sono identificative: "hh_id" identifica la famiglia intervistata con id unico e "id_country" identifica il paese di appartenenza della

famiglia. “id_country”, infatti, è categoriale e si compone di quattordici modalità, quanti sono i paesi partecipanti all’indagine. Sei variabili, tutte dicotomiche, sono di tipo economico e indicano il possesso di beni finanziari: “account”, che riguarda il possesso o meno di almeno uno tra conto corrente bancario o postale o libretto di deposito; “bonds”, che indica il possesso di titoli di stato o obbligazioni; “stocks”, che riguarda il possesso di azioni o partecipazioni; “mutual_funds”, che riguarda il possesso di fondi comuni di investimento o gestioni patrimoniali; “ira”, che riguarda il possesso di pensioni integrative private (*individual retirement accounts*); “life_insurance”, che, infine, riguarda il possesso di assicurazioni sulla vita. Alcune variabili danno indicazioni demografiche e di carattere più generale sul capofamiglia: “female”, dicotomica, indica se si tratta di maschio o femmina; “age”, continua, ne riporta l’età; ”isced97”, categoriale, che ne riporta il livello di istruzione, secondo la classificazione internazionale ISCED97[12]; “partner”, dicotomica, che indica se l’intervistato ha, o meno, un compagno; “marital_status”, variabile con quattro categorie, che riporta lo stato civile; “occupation”, variabile categoriale, indica se l’intervistato è occupato e che tipo di occupazione ha. Ci sono due variabili che riguardano la famiglia, ovvero “hhsize”, che riporta il numero di membri del nucleo familiare, e “house_own”, che indica se la famiglia è proprietaria dell’abitazione. Un’ultima variabile, “fluency_test”, riporta i risultati di un *test* cognitivo sul capofamiglia, che consiste nel ricordare il maggior numero possibile di nomi di animali, o specie e razze, in un minuto, ed è continua.

[12] La classificazione ISCED (*International Statistic Classification of Education*) dell’UNESCO nasce negli anni settanta per rendere confrontabili i dati riguardanti il livello di istruzione su scala internazionale. La codificazione usata in questa tesi si riferisce alla modifica del 1997 dell’ISCED, denominata ISCED97 appunto. In particolare, poi, si considera solo la classificazione del livello di istruzione, non quella dell’ambito.

Per condurre l'analisi di segmentazione si considerano le variabili "account", "bonds", "stocks", "mutual_funds", "ira" e "life_insurance" come variabili indicatori e tutte le altre come covariate.

3.4 Analisi esplorative delle variabili di possesso dei prodotti finanziari.

Innanzitutto si prendono in considerazione gli indicatori: nella tabella 3.1 sono riportate le loro modalità, le relative decodifiche, i dati mancanti e le frequenze assolute, relative e percentuali[13].
Il conto corrente e il libretto postale sono il prodotto finanziario più usato: circa il 77% delle famiglie ne possiede almeno uno, mentre per il possesso di tutti gli altri prodotti le percentuali scendono notevolmente. Un leggero picco si riscontra con le assicurazioni per la vita, possedute da circa il 23% delle famiglie, e con le pensioni integrative private. I prodotti meno frequenti sono titoli di stato e obbligazioni, presenti solo in circa l'8% delle famiglie.
Per vedere come la proprietà dei prodotti finanziari si distribuisce tra i vari paesi coinvolti nell'indagine, si calcolano le frequenze congiunte marginali della variabile "id_country" con le variabili "account", "bonds", "stocks", "mutual_funds", "ira", "life_insurance". I diagrammi a bolle, riportati in seguito, le rappresentano graficamente: in corrispondenza di ogni paese, per ogni modalità della variabile dipendente, è raffigurato un cerchio, la cui dimensione è proporzionale al numero delle unità che presentano le due modalità congiuntamente.

[13] La frequenza relativa conta quante volte nel campione si osserva una specifica modalità di una variabile, per esempio 4846 famiglie intervistate non possiedono conti correnti e libretti di risparmio (modalità 0 della variabile "account"); le frequenze relative sono date dal rapporto tra la frequenza assoluta e la numerosità campionaria e indicano la proporzione di unità assegnabili a una modalità (per esempio 0,21 è la proporzione di famiglie che non possiedono conti correnti e libretti di risparmio); le frequenze percentuali sono date dal prodotto delle frequenze relative per cento e indicano la percentuale campionaria di unità con una data caratteristica (per esempio il 21% delle famiglie intervistate non possiede conti correnti e libretti di risparmio).

variabile	modalità	decodifica	frequenze assolute	frequenze relative	frequenze percentuali
account	0	non possesso	4846	0,21	20,85
	1	possesso	17848	0,77	76,81
	NA	dato mancante	544	0,02	2,34
		totale	23238	1,00	100,00
bonds	0	non possesso	20593	0,89	88,62
	1	possesso	1931	0,08	8,31
	NA	dato mancante	714	0,03	3,07
		totale	23238	1,00	100,00
stocks	0	non possesso	19023	0,82	81,86
	1	possesso	3543	0,15	15,25
	NA	dato mancante	672	0,03	2,89
		totale	23238	1,00	100,00
mutual_funds	0	non possesso	2678	0,85	85,38
	1	possesso	19840	0,12	11,52
	NA	dato mancante	720	0,03	3,10
		totale	23238	1,00	100,00
ira	0	non possesso	18233	0,78	78,46
	1	possesso	4456	0,19	19,18
	NA	dato mancante	549	0,02	2,36
		totale	23238	1,00	100,00
life_insurance	0	non possesso	17557	0,76	75,55
	1	possesso	5280	0,23	22,72
	NA	dato mancante	401	0,02	1,73
		totale	23238	1,00	100,00

Tabella 3.1 Variabili risposta: modalità e distribuzioni di frequenza.

Dal *bubbleplot* di “id_country” e “account” (Figura 3.1), si nota che il possesso di conto corrente e libretto di risparmio non è diffuso uniformemente. Nell’Europa centro-settentrionale è molto diffuso, mentre nei paesi mediterranei, in particolare in Grecia, e nei paesi più orientali meno. La tabella 3.2, che riporta le frequenze congiunte percentuali, conferma queste osservazioni. In tutti i paesi centro-settentrionali, ad esclusione dell’Irlanda con il 79% e l’Austria con l’88%,

la percentuale di famiglie che possiede almeno un conto corrente o libretto di risparmio è superiore o uguale al 90%. Nell'Europa mediterranea, la percentuale scende all'80% circa per Spagna e Italia e al 45% in Grecia, mentre nell'Europa dell'est la percentuale scende ancora con il 55% circa della Polonia e il 25% della Repubblica Ceca, percentuale più bassa in assoluto. Si potrebbe ipotizzare che queste differenze siano dovute a un minore sviluppo economico del Sud e dell'Est Europa o a politiche finanziarie poco favorevoli.

A questo punto sembra naturale pensare che "account" e "id_country" siano dipendenti. Il test Chiquadrato di Pearson (vedi Appendice A) conferma questa ipotesi, riportando un valore pari a 6938,273 con 13 gradi di libertà e un *p-value* prossimo allo zero, che porta a rifiutare l'ipotesi nulla di indipendenza delle variabili.

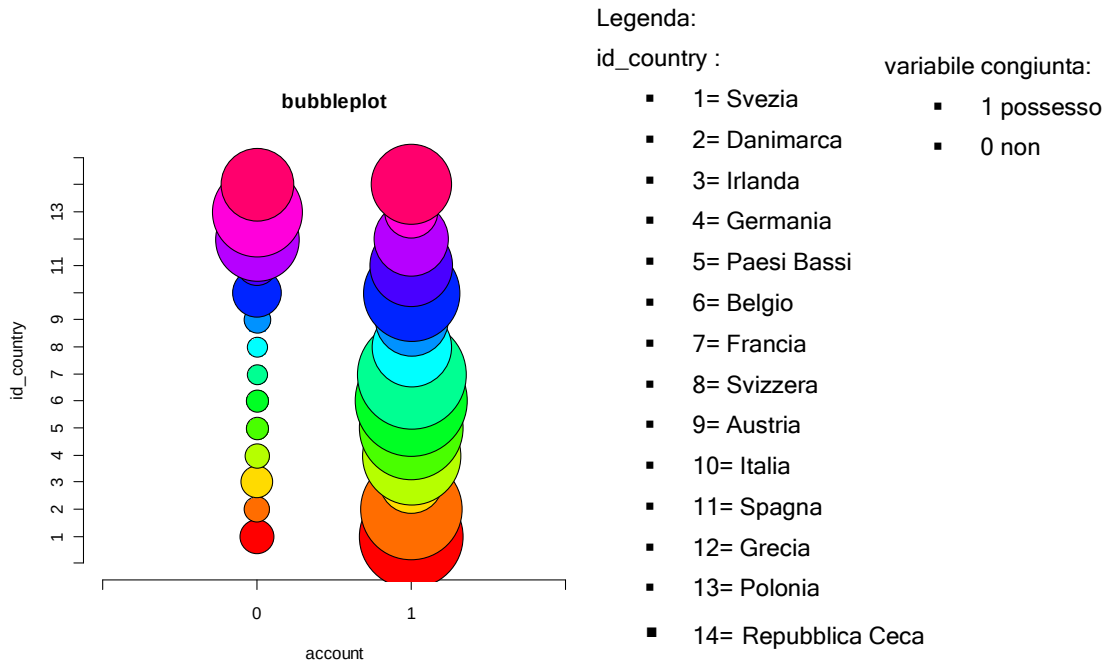

Figura 3.1 Grafico a bolle di "account" e"id_country".

Il grafico a bolle di "bonds" e "id_country", figura 3.2, evidenzia innanzitutto un ribaltamento tra le percentuali di possesso e non possesso, nel senso che titoli di stato e obbligazioni sono decisamente meno diffusi rispetto a conto corrente e libretto di risparmio.

id_country	account	
	0	1
Svezia	9,77	90,23
Danimarca	5,80	94,20
Irlanda	20,60	79,40
Germania	5,98	94,02
Paesi Bassi	4,21	95,79
Belgio	3,80	96,20
Francia	3,26	96,74
Svizzera	6,02	93,98
Austria	11,65	88,35
Italia	20,37	79,63
Spagna	19,42	80,58
Grecia	56,04	43,96
Polonia	74,84	25,16
Repubblica Ceca	45,36	54,64

Tabella 3.2 Frequenze congiunte percentuali di “id_country” e “account”.

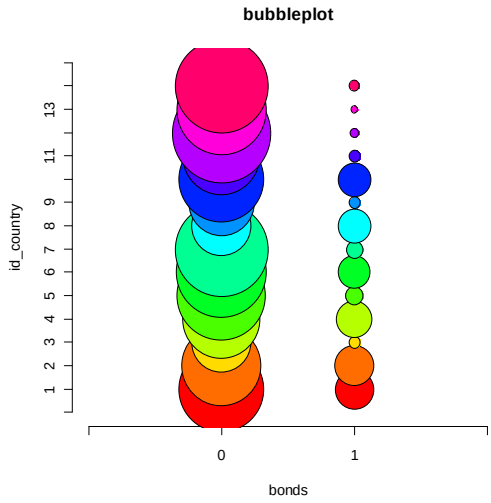

Figura 3.2 Grafico a bolle di “bonds” e “id_country”.

In effetti dalla tabella 3.3 si evince che i paesi in cui la percentuale di possesso è superiore o uguale al 20% sono solo Danimarca e Svizzera.

Svezia, Germania, Belgio e Italia si attestano su una percentuale di proprietà che varia tra il 17% e l'11%, comunque superiore al 10%. Infine tutti gli altri paesi hanno una percentuale inferiore al 5% e addirittura all'1% (Grecia e Polonia); di fatto questi ultimi paesi sono quelli mediterranei, ad esclusione dell'Italia, e dell'Est insieme a Francia, Irlanda e Austria. L'andamento delle frequenze, quindi, sembra meno legato alla zona europea di rilevazione. Il test chi quadrato evidenzia ancora dipendenza delle variabili, ma minore rispetto ad "account" e "id_country" (X^2=1680,292 con 13 gradi di libertà e *p-value* prossimo a zero).

Il *bubbleplot* che rappresenta le frequenze congiunte di "stocks" e "id_country" (figura 3.3) evidenzia che le percentuali di possesso di azioni e partecipazioni sono inferiori a quelle di non possesso, tranne per Svezia e Danimarca, per le quali le percentuali sembrano simili. Inoltre "stocks" presenta percentuali di possesso maggiori nei paesi centrosettentrionali, ad eccezione di Irlanda e Austria, rispetto a quelli sudorientali.

	bonds	
id_country	0	1
1	82,76	17,24
2	80,31	19,69
3	96,07	3,93
4	85,11	14,89
5	96,06	3,94
6	88,63	11,37
7	96,84	3,16
8	76,13	23,87
9	96,99	3,01
10	86,61	13,39
11	97,89	2,11
12	99,14	0,86
13	99,31	0,69
14	98,73	1,27

Tabella 3.2 Frequenze congiunte percentuali di "bonds e "id_country".

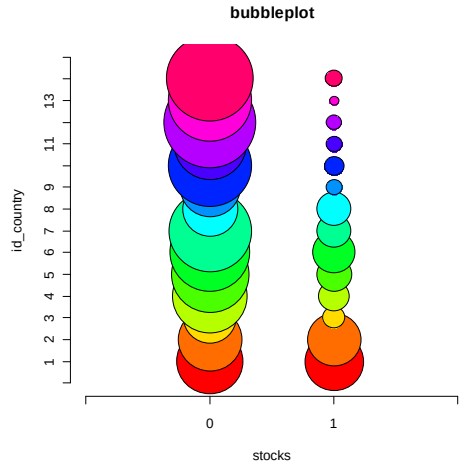

Figura 3.3 Diagramma a bolle di “id_country” e “stocks”.

Queste osservazioni sono confermate dalle frequenze congiunte percentuali, riportate nella tabella 3.4.

id_country	stocks	
	0	1
1	56,42	43,58
2	58,24	41,76
3	85,23	14,77
4	85,93	14,07
5	83,28	16,72
6	78,21	21,79
7	85,61	14,39
8	72,92	27,08
9	93,06	6,94
10	94,68	5,32
11	95,34	4,66
12	97,34	2,66
13	98,73	1,27
14	96,30	3,70

Tabella 3.4 Frequenze congiunte percentuali di “id_country” e “stocks”.

Si nota, infatti, che nei primi otto paesi, ovvero Svezia, Danimarca, Irlanda, Germania, Paesi Bassi, Belgio, Francia e Svizzera rispettivamente, le percentuali di possesso non scendono mai sotto il 14%, mentre nei restanti paesi, Austria, Italia, Spagna, Grecia, Polonia, Repubblica Ceca, le percentuali sono tutte al di sotto del 7%. In particolare poi nel primo gruppo, quello dalle percentuali maggiori, si nota che Svezia e Danimarca hanno percentuali simili a quelle di non possesso, rispetto agli altri paesi, Belgio e Svizzera si attestano al di sopra del 20%, mentre Irlanda, Germania Paesi Bassi e Francia si attestano intorno al 15%. Nel secondo gruppo, invece, si nota che la Polonia ha una percentuale davvero bassa (1,27%). Nel caso di "stocks" sembra netta la divisione fra centronord e sudest dell'Europa; sebbene in realtà si possano individuare tre gruppi in base al possesso: un primo gruppo formato da Svezia e Danimarca (che con l'Irlanda rappresentano il nord), un secondo gruppo formato da Irlanda, Germania, Paesi Bassi, Belgio, Francia e Svizzera (il centro) e un terzo gruppo con Austria, Italia, Spagna, Grecia, Polonia e Repubblica Ceca (sud-est). Il test chi quadrato di Pearson per "id_country" e "stocks" è pari a 3229,964 e il *p-value* associato è prossimo a zero (i gradi di libertà sono tredici). Da ciò si evince che esiste dipendenza tra la variabile geografica e il possesso di azioni e partecipazioni, sebbene in forma minore rispetto al possesso di conto corrente e libretto di risparmio e in forma maggiore rispetto al possesso di titoli di stato ed obbligazioni.

Per la distribuzione di fondi di investimento e gestioni patrimoniali, il diagramma a bolle (figura 3.4) sembra evidenziare una situazione di possesso simile a quella appena descritta: le bolle di area maggiore, per "mutual_funds" pari a uno, sono in corrispondenza dei primi otto paesi, ad eccezione del terzo, l'Irlanda. Inoltre il primo paese, ovvero la Svezia, sembra presentare una percentuale nettamente superiore al resto dell'Europa. Ancora una volta, le percentuali di possesso sono nettamente inferiori a quelle di non possesso, tranne per la Svezia.

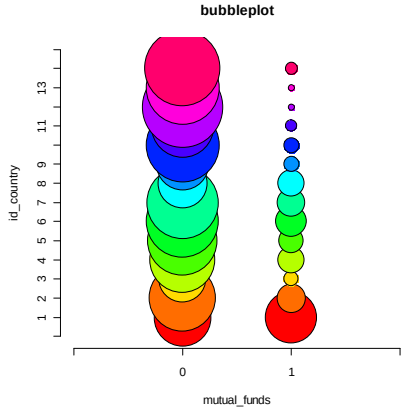

Figura 3.4 Diagramma a bolle di “id_country” e “mutual_funds”.

Le frequenze congiunte percentuali (Tabella 3.5) confermano le osservazioni, evidenziando, però, che in generale le percentuali di possesso di fondi di investimento e gestioni patrimoniali sono inferiori quelle di possesso di azioni e partecipazioni. Si nota che la Svezia ha una percentuale di possesso molto alta rispetto agli altri paesi (45% circa). Per i primi otto paesi, ad eccezione dell’Irlanda, le percentuali restano al di sopra del 12%, scendono all’ 8% e 9% per Irlanda e Austria e al di sotto del 5% per Italia, Spagna, Grecia, Polonia e Repubblica Ceca. La suddivisione dell’Europa in due gruppi sembra confermata.

Il test chi quadrato conferma la dipendenza tra “id_country “ e “mutual_funds” (X^2=2966,809 con tredici gradi di libertà e *p-value* prossimo a zero).

Il grafico a bolle delle frequenze congiunte di “id_country” e “ira” (Figura 3.6) evidenzia ancora una volta che il possesso del prodotto finanziario ha percentuali inferiori del non possesso e che nei primi otto paesi, ad eccezione di Irlanda e Paesi Bassi, le percentuali di possesso sono maggiori rispetto agli ultimi sei, tra i quali, però, la Repubblica Ceca ha una frequenza nettamente più alta. La Svezia, poi, sembra ancora avere percentuali di possesso e non possesso simili.

id_country	mutual_funds	
	0	1
1	54,87	45,13
2	84,30	15,70
3	91,39	8,61
4	86,00	14,00
5	88,35	11,65
6	84,56	15,44
7	87,00	13,00
8	78,02	21,98
9	92,12	7,88
10	95,58	4,42
11	96,61	3,39
12	99,38	0,62
13	99,25	0,75
14	97,15	2,85

Tabella 3.5 Frequenze congiunte di “id_country” e “mutual_funds”.

Le frequenze percentuali (Tabella 3.6) confermano le prime osservazioni. Si nota però che tra i paesi del gruppo centrosettentrionale anche la Germania, oltre all’Irlanda e ai Paesi Bassi, ha una percentuale inferiore. Infatti in questi ultimi il possesso varia intorno al 10%, mentre negli altri del gruppo le percentuali sono tutte superiori al 27%. Nel secondo gruppo le percentuali si abbassano notevolmente, tranne per Spagna e Austria che si attestano intorno al 10% e la Repubblica Ceca con il 36%.

La differenza netta tra i paesi del centro nord e del sud est nella diffusione delle pensioni integrative non è così regolare come per i prodotti precedentemente indagati. Tuttavia il possesso di pensioni integrative è molto più legato alle politiche (pensionistiche) statali rispetto agli altri prodotti. Si può, quindi, presumere che percentuali sorprendentemente basse (Paesi Bassi e Germania) si trovino in paesi in cui la politica pensionistica statale non rende necessarie integrazioni

private, al contrario percentuali sorprendentemente alte (Repubblica Ceca) si trovino in paesi in cui è necessario integrare.

Il test chi quadrato evidenzia dipendenza anche per “id_country” e “ira” (X^2=3503.930 con tredici gradi di libertà e *p-value* prossimo a zero), più forte che per tutti gli altri prodotti finanziari analizzati, tranne che per conto correnti e libretti di risparmio.

Il *bubbleplot* che rappresenta le frequenze congiunte di “id_country” e “life_insurance” (Figura 3.6) evidenzia che le percentuali di possesso di assicurazioni sulla vita sono più alte degli altri prodotti finanziari (esclusi conto corrente e libretto di risparmio). In particolare i primi nove paesi e gli ultimi due presentano una diffusione maggiore. Le frequenze percentuali (Tabella 3.7), inoltre, evidenziano una percentuale di possesso molto alta per la Svezia (43%) e per Danimarca, Irlanda, Germania e Paesi Bassi, in cui percentuali variano tra il 30% e il 37%. Frequenze abbastanza alte si riscontrano anche in Belgio, Francia, Svizzera e Austria (tra il 19% della Francia e il 24% del Belgio). Negli ultimi cinque paesi le percentuali sono notevolmente più basse, ad eccezione di Polonia (35%) e Repubblica Ceca (15.5%).

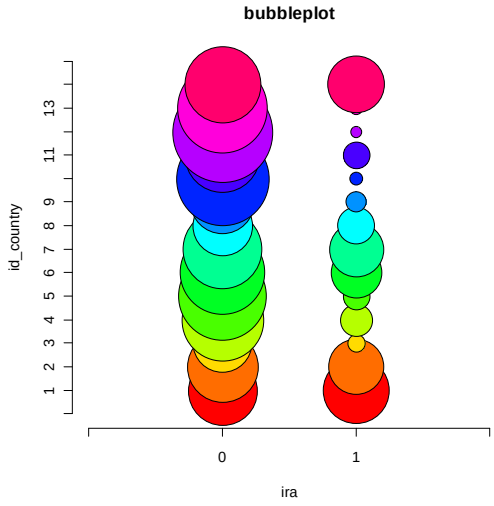

Figura 3.5 Diagramma a bolle di “id_country” e “ira”.

id_country	ira	
	0	1
1	52,13	47,87
2	62,07	37,93
3	91,66	8,34
4	86,42	13,58
5	91,57	8,43
6	73,13	26,87
7	67,67	32,33
8	72,67	27,33
9	91,30	8,70
10	97,98	2,02
11	89,18	10,82
12	98,87	1,13
13	98,12	1,88
14	63,88	36,12

Tabella 3.6 frequenze congiunte percentuali di “id_account” e “ira”.

Il test chi quadrato di Pearson ha valore 1709.115 con tredici gradi di libertà e *p-value* prossimo a zero, quindi conferma la dipendenza tra le variabili, sebbene sia più debole che per quasi tutti gli altri prodotti finanziari (è più forte solo rispetto a titoli di stato e obbligazioni (“bonds”)).

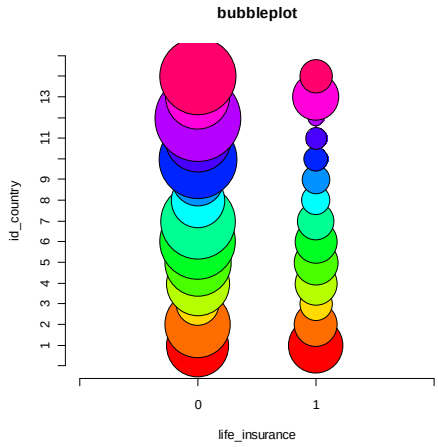

Figura 3.6 Diagramma a bolle di “id_country” e “life_insurance”.

id_country	life_insurance	
	0	1
1	56,74	43,26
2	70,25	29,75
3	63,26	36,74
4	68,98	31,02
5	69,59	30,41
6	75,81	24,19
7	80,75	19,25
8	77,40	22,60
9	77,23	22,77
10	91,32	8,68
11	90,55	9,45
12	96,47	3,53
13	64,93	35,07
14	84,44	15,56

Tabella 3.7 Frequenze congiunte percentuali di “id_country” e “life_insurance”.

Riassumendo, dall’analisi delle frequenze congiunte della variabile geografica “id_country”, indicante il paese europeo di residenza delle famiglie, e delle variabili che riportano il possesso o meno dei prodotti finanziari, si evidenzia una distribuzione dei prodotti dipendente dalla

zona europea di residenza. In particolare si delineano come gruppi l'Europa del nord, centrale, mediterranea e dell'est. Attraverso un modello a classi latenti multilivello è possibile verificare l'esistenza di questi gruppi, che fungerebbero da macrosegmenti.

3.4 Analisi esplorative delle variabili descrittive.

Nella tabella 3.9 sono riportate le frequenze assolute,relative e percentuali di tutte la variabili che saranno usate come covariate.
La variabile "female" assume i valori 0, se il capofamiglia è maschio, e 1, se è femmina. Il campione è formato per la maggior parte da femmine (55%).
La variabile "age" indica l'età del capofamiglia ed assume valori nell'intervallo [24, 104], con un'età media pari a 65.59 anni. Suddividendo "age" in classi, si evince che il 34% degli intervistati è al di sotto dei sessant'anni[14], il 32% ha un'età compresa tra i sessanta e sessantanove anni, il 22% tra settanta e settantanove anni e il restante 12% è ultraottantenne. Dal *boxplot* (figura 3.7) si nota che l'età mediana è sessantaquattro anni, mentre il primo e terzo quartile sono rispettivamente cinquantasette e settantatre anni, pertanto un quarto del campione ha meno di cinquantasette anni, metà ha meno di sessantaquattro anni e tre quarti ha meno di settantatre anni. La distribuzione di "age" sembra leggermente asimmetrica verso il basso e sono segnalati numerosi *outlier*, sia verso il basso che verso l'alto. In effetti vi sono dei valori, come ad esempio 24, 28, 104, 101 anni, che si discostano notevolmente dall'età mediana. Infatti, in seguito sarà

[14] In particolare nel campione 16 individui hanno un'età compresa tra i ventiquattro e i trentanove anni, mentre 219 tra i quaranta e i quarantanove anni. Questi dati sono anomali, come segnala anche il *boxplot* in figura 3.7, e potrebbe sorgere il dubbio di doverli considerare mancanti. Tuttavia non è infrequente avere un partner molto più giovane (che risulterebbe essere il capofamiglia), perciò questi dati saranno considerati reali. Nello stesso modo non saranno considerati dati mancanti età molto alte, visto che l'età media di vita è in continua crescita.

tracciato un profilo del capofamiglia tipo, utilizzando l'età mediana, meno sensibile ai valori *outlier*. I capofamiglia sono relativamente giovani. L'età relativamente bassa e la maggioranza di femmine nel campione si spiegano con il fatto che il capofamiglia è considerato il più giovane di una coppia e, a parità di età, la femmina. Infatti, come si deduce dalle definizioni della popolazione di interesse, solo alle femmine minori di cinquant'anni è permesso partecipare all'indagine.

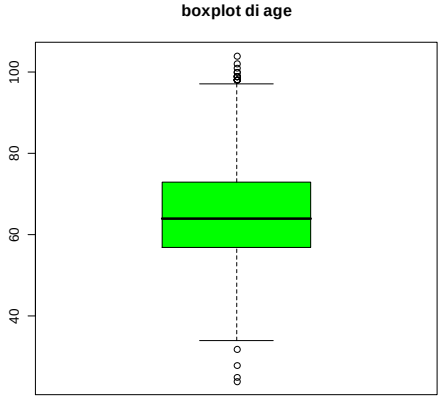

Figura 3.7 Boxplot di "age".

La variabile "partner" assume valori 0, se l'intervistato non ha un compagno, e 1, se ha un compagno. Il 64% dei capofamiglia dichiara di avere un compagno.

La variabile "marital_status" indica lo stato civile dell'intervistato. Il 61% del campione dichiara di essere sposato o di convivere legalmente con il compagno, il 21% è vedovo. Solo l'11% è separato o divorziato e il 7% celibe o nubile. La maggior parte del campione, quindi, è o è stata sposata, segno, forse, di una mentalità diversa da quella corrente, dovuta all'età.

Si nota che il 64% del campione dichiara di avere un partner, ma solo il 61% è sposato o convive legalmente. La tabella 3.8 riporta le frequenze percentuali congiunte di "partner" e "marital_status". Si nota che del 64% degli intervistati che hanno un compagno, l'1,5 % è separato, lo 0,5% è celibe o nubile e lo 0.65% è vedovo. Quindi, la variabile

“partner” riporta solo se il capofamiglia è fidanzato, non se ha un compagno convivente, mentre “marital_status” riporta solo lo stato civile del capofamiglia. In prima analisi, perciò, le due variabili non sembrano ridondanti.

	marital_status				
partner	1	2	3	4	totale marginale di "partner"
0	0,00	9,04	6,18	20,14	35,36
1	61,23	1,48	0,58	0,65	63,93

Tabella 3.8 Frequenze congiunte percentuali di “partner” e “marital_status”.

La variabile “hhsize” indica il numero di componenti del nucleo familiare, che nel campione variano tra uno e quattordici. Il numero medio di componenti familiari è 2,117, il numero mediano e il terzo quartile sono entrambi due, mentre il primo quartile è uno. Nel boxplot di “hhsize” (figura 3.8), infatti, si nota l’asimmetria e l’appiattimento verso il basso della scatola, segno, appunto, che i valori più bassi sono i più frequenti, e l’uguaglianza di mediana e terzo quantile. La maggioranza delle famiglie considerate, quindi, è composta solo da una o due persone, tant’è vero che i valori più alti di “hhsize” sono segnalati come *outlier*. Le frequenze di “hhsize”, suddivisa in classi, confermano quanto detto; infatti il 29% delle famiglie ha un solo componente, il 48%, quasi la metà, ne ha due, il restante 23% ne ha almeno tre. Trattandosi di un’indagine su ultracinquantenni non stupiscono questi dati: è ragionevole pensare che, quando i genitori siano in età da pensione, i figli si stiano formando una propria famiglia, al di fuori del nucleo familiare d’origine.

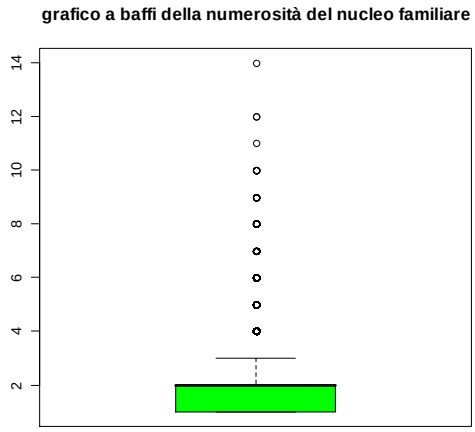

Figura 3.8 Diagramma a baffi di “hhsize”.

La variabile “occupation” riporta lo stato occupazionale del capofamiglia. Le sue frequenze sono riportate nel grafico a torta della figura 3.9, in cui l’area dei settori circolari è proporzionale alla frequenza della rispettiva modalità. La maggioranza del campione, 52%, è in pensione dal lavoro, mentre il 28% lavora ancora. La percentuale di disoccupati è, ovviamente, bassa, come quella dei malati cronici e disabili.

La variabile ”house_own” vale 0, se la famiglia non è proprietaria della casa in cui risiede, e 1, se lo è. La modalità più frequente è 1, pertanto la maggioranza delle famiglie è possiede la propria abitazione.

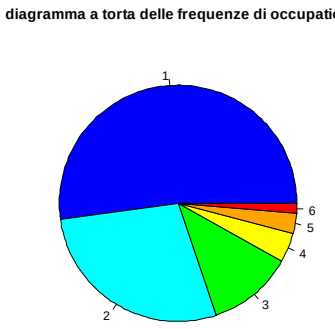

Figura 3.9 Grafico a torta di “occupation”.

La variabile “isced97” riporta il grado di istruzione del capofamiglia, secondo l’omonimo standard UNESCO. Le modalità sono sei e corrispondono ai rispettivi livelli della classificazione internazionale[15]. Il livello 0 corrisponde all’istruzione pre-elementare (la scuola dell’infanzia nel sistema scolastico italiano), il livello 1 all’istruzione elementare o primo stadio dell’istruzione base (in Italia la scuola primaria, ovvero le scuole elementari), il livello 2 all’istruzione secondaria inferiore o secondo stadio dell’istruzione base (scuola secondaria di primo grado o scuole medie inferiori), livello 3 all’istruzione secondaria superiore (scuola secondaria di secondo grado o medie superiori), livello 4 all’istruzione post-secondaria non terziaria, livello 5 al primo stadio dell’istruzione terziaria (laurea e laurea magistrale) e livello 6 al secondo stadio dell’istruzione terziaria.

Il primo quartile di “isced97” è 1, il che significa che almeno un quarto degli intervistati ha un’istruzione elementare, la mediana e il terzo quartile, invece, sono entrambi 3, quindi almeno tre quarti del campione non ha frequentato scuole che garantissero un livello di istruzione oltre a quello superiore. Nella figura 3.10, infatti, si nota l’asimmetria verso il basso delle distribuzione: la scatola del boxplot comprende i valori bassi di “isced97”. Inoltre si nota che mediana e terzo quantile coincidono. Le frequenze confermano queste osservazioni. La moda è livello 3, ovvero il 28% circa degli intervistati ha un’istruzione secondaria superiore, tuttavia il livello due è raggiunto solo dal 27% dei capofamiglia. Gli intervistati sembrerebbero non avere studiato moltissimo, visto le percentuali dei livelli 0, 1 e 2, ma questo non stupisce se si considera che parte degli intervistati è stato bambino e giovane prima e durante la Seconda Guerra Mondiale, periodi in cui l’obbligo scolastico, se c’era, era basso e probabilmente non rispettato.

[15] In realtà i dati presentano due ulteriori modalità, che sono “95” con frequenza 12 e “97” con frequenza 146. Poiché queste due modalità non hanno senso nella classificazione ISCED, si è scelto di considerarle dati mancanti.

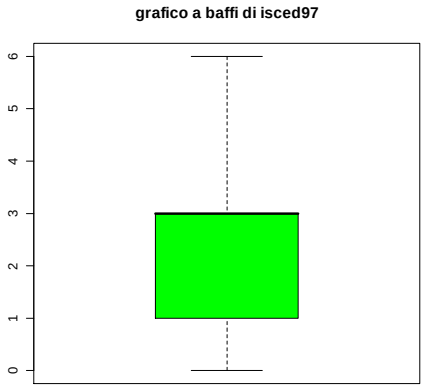

Figura 3.10 Diagramma a baffi di isced97 con i soli valori dell'indice.

La variabile "fluency_test" ha primo quartile pari a quattordici, mediana diciotto e terzo quartile ventitre. Inoltre in media gli intervistati hanno saputo dire in un minuto 18,85 nomi di animali. Il grafico in figura 3.11 rappresenta con un diagramma a bastoncini le frequenze del test. Si evince che la moda è 15.

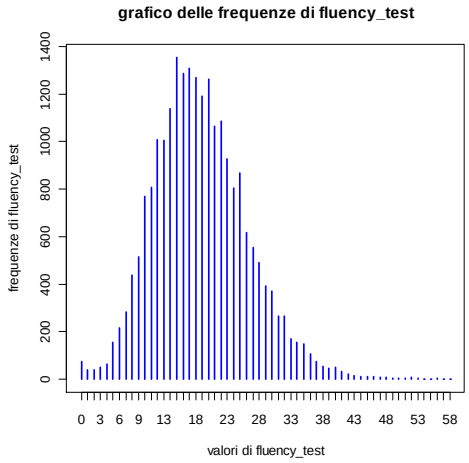

Figura 3.11 Grafico a bastoncini delle frequenze di "fluency_test".

Tracciando un profilo del capofamiglia tipo del campione, esso è una donna di sessantaquattro anni (età mediana), pensionata, con un'istruzione di livello secondario superiore (livello mediano). Ha un compagno con cui vive, perché è sposata o convive legalmente. Il suo

nucleo familiare è composto da due persone e vive in un'abitazione di proprietà. Ricorda in un minuto diciotto nomi di animali.

Come per la variabile "id_country", anche per le altre si è calcolato il test chi quadrato di Pearson, al fine di rilevare la presenza di associazione con "account", "bonds", "stocks", "mutaul_funds", "ira" e "life_insurance". I valori sono riportati nella tabella 3.10. Non si è rilevato alcun caso di non associazione, perché tutti i *p-value* sono prossimi a zero (e per questo non sono riportati in tabella). Giudicando la grandezza dei valori del test, si nota che la variabile "account" ha maggiore dipendenza da "isced97", così come "bonds", "stocks", e "mutual_funds". "ira", invece, mostra maggiore associazione con "occupation" e "life_insurance" con la variabile "age", suddivisa in classi, come spiegato in precedenza. Inoltre "ira" ha una misura di associazione nettamente superiore alle altre anche con "age", sempre suddivisa in classi e "isced97", mentre "life_insurance" con "occupation". Sembra, quindi, che mentre il possesso di conto corrente o libretto di risparmio, titoli di stato o obbligazioni, azioni o partecipazioni e fondi di investimento o gestioni patrimoniali sia legato al livello di istruzione, il possesso di pensione integrativa e assicurazione sulla vita sia più legato all'età e all'occupazione.

variabile	modali	decodifica	frequen	frequen	frequenz
female	0	maschio	10505	0,45	45,21
	1	femmina	12733	0,55	54,79
	NA	dato mancante	0	0,00	0,00
totale			23238	1	100
age	60-		7943	0,34	34,18
	[60,69]		7343	0,32	31,60
	[70,79]		5194	0,22	22,35
	80+		2758	0,12	11,87
	NA		0	0,00	0,00
totale			23238	1	100

partner	0	non ha un compagno	8250	0,36	35,50
	1	ha un compagno	14985	0,64	64,48
	NA		3	0,00	0,01
totale			23238	1	100
marital_sta	1	sposato o in convivenza legalmente	14289	0,61	61,49
	2	separato o divorziato	2444	0,11	10,52
	3	celibe o nubile	1570	0,07	6,76
	4	vedovo o vedova	4830	0,21	20,78
	NA		105	0,00	0,45
			23238	1	100
hhsize	1		6676	0,29	28,73
	2		11201	0,48	48,20
	[3,4]		4462	0,19	19,20
	[5,14]		899	0,04	3,87
	NA		0	0,00	0,00
totale			23238	1	100
occupation	1	pensionato	12023	0,52	51,74
	2	occupato	6444	0,28	27,73
	3	casalinga o casalingo	2667	0,11	11,48
	4	malato cronico o disabile	927	0,04	3,99
	5	disoccupato	626	0,03	2,69
	6	altro	325	0,01	1,40
	NA		226	0,01	0,97
totale			23238	1	100
house_own	0	non proprietario dell'abitazione	5973	0,26	25,70
	1	proprietario dell'abitazione	16888	0,73	72,67
	NA		377	0,02	1,62
totale			23238	1	100
isced97	0	istruzione pre-elementare	916	0,04	3,94
	1	istruzione elementare	6244	0,27	26,87
	2	istruzionesecondaria infariore	4083	0,18	17,57
	3	istruzione secondaria superiore	6502	0,28	27,98
	4	istruzione post secondaria non terziaria	815	0,04	3,51
	5	primo stadio di istruzione terziaria	4080	0,18	17,56
	6	secondo stadio di istruzione terziaria	86	0,00	0,37
	NA		512	0,02	2,20
totale			23238	1	100

Tabella 3.9 Frequenze assolute, relative e percentuali delle variabili covariate.

variabile	account	bonds	stocks	mutual_funds	ira	life_insurance	gradi di libertà
female	132,12	41,48	189,76	95,79	69,16	75,04	1
age in classi	81,86	17,40	181,34	117,46	1626,06	1591,94	3
partner	262,07	80,10	352,73	196,15	452,16	617,16	1
marital_status	318,16	85,84	359,38	211,02	605,79	613,80	3
hhsize in classi	447,65	93,52	304,14	205,28	405,26	600,27	3
occupation	370,16	91,61	434,86	313,92	2180,29	1544,99	5
house_own	45,51	55,02	267,33	99,96	208,50	79,67	1
isced97	1235,34	389,27	1286,68	724,57	1227,98	987,79	6

Tabella 3.10 Valori del test chi quadrato di Pearson e gradi di libertà.

3.4 I dati mancanti.

Nelle tabelle 3.1 e 3.9 sono riportate anche le frequenze dei dati mancanti. Nessuna variabile supera il 5% di dati mancanti, nemmeno “hh_id” e “id_country”, che non ne hanno, e “fluency_test”, che ne ha trecentocinque, pari all’1.31%. Con queste percentuali esigue, si ritiene opportuno non considerare i dati mancanti, perché non si corre il rischio di distorcere le stime. Di *default*, Latent Gold, il *software* utilizzato per la stima di modelli a classi latenti, cancella i *record* con dati mancanti.

Capitolo 4

Segmentazione del mercato con il modello a classi latenti classico.

4.1 Introduzione.

Obiettivo di questa tesi è individuare nell'eterogeneità della popolazione europea di ultracinquantenni dei segmenti di individui che differiscono significativamente tra loro e al contempo sono internamente omogenei rispetto al possesso di beni finanziari, quali concorrenti o libretti di risparmio, titoli di stato o obbligazioni, azioni o partecipazioni, fondi di investimento o gestioni patrimoniali, pensioni integrative private e assicurazioni sulla vita. A tal scopo si stima un modello classico a classi latenti utilizzando il *software* Latent GOLD 4.0, creato *ad hoc* da Madigson e Vermunt (2000).

Come indicatori sono utilizzate le variabili "account", "bonds", "stocks", "mutual_funds", "ira" e "life_insurance". Come covariate, utili per profilare i segmenti individuati dall'analisi, si utilizzano "id_country", "female", "age", "partner", "marital_status", "hhsize", "house_own", "occupation", "fluency_test" e "isced97". In particolare la variabile "age" è stata suddivisa come segue:

age= { 1 se l'età è minore di sessant'anni
2 se l'età è compresa tra i sessanta e sessantanove anni
3 se l'età è compresa tra i settanta e i settantanove anni
4 se l'età è superiore o uguale ad ottant'anni

La variabile "hhsize", invece, è stata suddivisa così:

hhsize= { 1 se nella famiglia c'è un solo componente
2 se nella famiglia ci sono due componenti
3 se nella famiglia ci sono tre o quattro componenti
4 se nella famiglia ci sono da cinque a quattordici componenti

4.1.1 Stima del modello.

Sono stati stimati i modelli a classi latenti con un numero di classi da uno a sette, utilizzando per la stima l'algoritmo EM (*Expetation Maximization*)[16].

Latent GOLD fornisce in *output* il risultato del *test* di Wald per la significatività delle variabili nel modello. In tutti i modelli, tutte le covariate sono significative ad esclusione di "hhsize", che non raggiunge mai la significatività all'1%. Ciò significa che la variabile non discrimina tra le classi, nel senso che assume valori molto simili, pertanto va eliminata. In effetti la stima dei modelli senza "hhsize" è migliore dal punto di vista dell'adattamento.

Eliminata "hhsize", ogni modello è stato stimato più volte, con insiemi di valori iniziali per le procedure iterative di stima differenti. Quindi è stato scelto il modello con il migliore adattamento.

In tabella 4.1 sono riportati i valori di alcune statistiche utili a giudicare la bontà dell'adattamento, ovvero BIC e L^2. Vi sono, inoltre il numero dei parametri nel modello, i gradi di libertà relative alla distribuzione X^2 della statistica rapporto di verosimiglianza (L^2) e il *p-value* associato.

Quando le variabili nel modello sono numerose, come nel nostro caso, si rischia che la tabella di contingenza osservata sia sbilanciata. I dati, pertanto, potrebbero essere sparsi. In tal caso non sono soddisfatte le condizioni di regolarità sotto le quali L^2 ha distribuzione asintotica X^2 e il *p-value* in tabella 4.1 non va giudicato.

[16] L'algoritmo EM è un algoritmo iterativo per stime di massima verosimiglianza. Si articola in due fasi: la fase E (*Expectation*), in cui si calcola il valore atteso della log-verosimiglianza, la fase M (*Maximization*), in cui si calcolano i parametri massimizzando la log-verosimiglianza attesa trovata al punto E. Le stime trovate con il passo M sono utilizzate nel successivo passo E. Nella stima di modelli a classi latenti classici, nel passo E si stimano le probabilità che un caso appartenga ad un dato *cluster* e nel passo M si utilizzano queste stime come pesi nella massimizzazione del valore atteso della log-verosimiglianza.

La scelta del modello che meglio si adatta ai dati, quindi, si basa su BIC e L^2. Il primo tiene conto anche della parsimonia del modello, il secondo indica quanta relazione tra le variabili non è spiegata. Sono preferibili modelli con valori bassi di entrambe le statistiche, rispetto ad altri con valori maggiori.

	BIC(LL)	Npar	L^2	df	p-value
1 classe	111908,76	6	104018,42	20999	6,0e-10735
2 classi	95221,68	41	86982,89	20964	5,3e-7862
3 classi	90560,73	76	81992,11	20929	1,4e-7057
4 classi	88435,89	111	79500,54	20894	6,2e-6667
5 classi	87210,21	170	74155,05	28835	3,5e-5838
6 classi	86246,62	211	72783,41	20794	6,1e-5636
7 classi	86262,07	252	72410,81	20753	2,6e-5589

Tabella 4.1 Indici di bontà di adattamento dei modelli a classi latenti stimati.

A giudicare dal BIC, il modello migliore è quello con sette classi latenti. L'ampiezza di alcune classi, però, è esigua (in ordine decrescente è 33%, 16,5%, 12%, 11%, 10%, 10% e 7%) e si corre il rischio che i segmenti meno estesi non siano consistenti e, quindi, profittabili[17]. Da un punto di vista di *marketing*, quindi, meglio scegliere un modello che garantisca che i segmenti abbiano le caratteristiche che li rendono strategicamente funzionali, presentate nel primo capitolo.

BIC e L^2 associati al modello a sette classi non sono di molto inferiori a quelli associati al modello a sei classi. Sembra, quindi, che il miglioramento apportato dal modello a sette classi non sia significativo. È possibile testare statisticamente questa supposizione con il test chi quadrato per modelli annidati. La statistica test ΔL^2 è data dalla differenza tra gli L^2 associati ai due modelli e si distribuisce secondo un

[17] I segmenti di un mercato sono consistenti quando la loro ampiezza e/o capacità di assorbimento sono tali da garantire un profitto all'azienda. Sono profittabili quando un'azienda guadagna ricavi dal servirli.

X^2 con gradi di libertà pari alla differenza dei gradi di libertà dei modelli a confronto. [18]

Abbiamo, quindi, $\Delta L^2 = L^2_{6classi} - L^2_{7classi} = 86262,02 - 86246,62 = 15,35$ e $df = df_{7classi} - df_{6classi} = 20794 - 20753 = 41$. Il *p-value* associato al test è 1, quindi il modello a sette classi non apporta miglioramenti significativi rispetto a quello a sei classi. Pertanto, per i motivi statistici e di *marketing* presentati, il modello a sette classi va rifiutato.

Il modello a sei classi ha BIC e L^2 notevolmente minori rispetto a quello a cinque classi, pertanto è migliore. In particolare spiega una quantità maggiore di associazione tra le variabili. Questa supposizione è confermata dal fatto che il test ΔL^2 ha *p-value* prossimo a zero. Inoltre, l'ampiezza delle classi, presentata in seguito, non pone dubbi sulla consistenza dei segmenti. Per questi due motivi, è scelto il modello a sei classi latenti.

4.2 Costruzione del profilo dei segmenti.

I segmenti individuati sono sei. Il primo segmento comprende il 30% degli ultracinquantenni europei, il secondo il 19%, il terzo il 14%, il quarto il 14,5%, il quinto il 12,5% e il sesto il 10% (vedi tabella 4.2). In tabella 4.2 e 4.3 sono riportate le probabilità di rispondere y_{ik} nell'*item k*, data l'appartenenza alla *t-sima* classe latente. Ad esempio se prendiamo la probabilità nella quarta riga della seconda colonna della tabella 4.2, essa esprime la probabilità che i capofamiglia rispondano 1 nell'*item* che riporta il possesso di conto corrente o libretto di risparmio, dato che appartengono alla prima classe latente. Le probabilità condizionate sono utili per descrivere i segmenti individuati.

Nei prossimi paragrafi si traccerà un profilo dei segmenti, basandosi sulle probabilità condizionate e utilizzando prima gli indicatori e poi le covariate. Così facendo, dapprima, si individuano le tendenze sul

[18]Le condizioni di regolarità, sotto le quali ΔL^2 ha distribuzione asintotica chi quadrato, sono meno stringenti rispetto a quelle per il semplice test L^2.

possesso dei prodotti finanziari di ogni segmento, al quale verrà assegnato un nome, e in un secondo momento si descrivono i segmenti in modo da renderli individuabili.

4.2.1 Profilo dei segmenti sulla base degli indicatori.

Per semplicità, si prendono in considerazione solo le modalità 1 degli indicatori, presentati in tabella 4.2, che indicano il possesso del relativo prodotto finanziario.
Nella prima classe latente, si nota una probabilità molto alta per la variabile “account” e molto bassa per tutte le altre. Pertanto, tutte le unità in questa classe possiedono quasi solo conto corrente o libretto di risparmio.
Nella seconda classe, invece, le probabilità condizionate sono molto basse per tutti gli indicatori: solo per “life_insurance” (0.12) la probabilità condizionata sale al di sopra del 10%. In sostanza, le unità classificate nel secondo segmento non possiedono alcun prodotto finanziario.
Le probabilità condizionate sono più articolate nella terza classe: é molto alta quella associata ad “account” (0.98), alta quella associata ad “ira” (0.89) e superano il 50% quelle associate a “stocks” e “life_insurance”. La probabilità per “mutual_funds” non supera lo 0,5, ma vi è molto vicina, quella associata a “bonds” supera il 20%. Gli ultracinquantenni del terzo segmento, quindi, possiedono numerosi prodotti finanziari: conti correnti o libretti di risparmio e pensioni integrative con alta probabilità, azioni o partecipazioni e assicurazioni con buona probabilità, ma anche fondi comuni di investimento o gestioni patrimoniali.
Nella quarta classe latente si nota una probabilità molto alta per “account” (0.95) e alta per “life_insurance” (0.65). I casi che appartengono alla quarta classe latente, quindi, possiedono conto corrente o libretto di risparmio e assicurazioni per la vita.
Nella quinta classe solo ad “account” è associata una probabilità molto alta. Tuttavia essa differisce dalla prima, perché presenta probabilità più

alte per tutte le altre variabili, esclusa “ira”. Quindi, in questo segmento gli ultracinquantenni possiedono con alta probabilità conto corrente o libretto di risparmio, con probabilità superiori al 30% azioni o partecipazioni, titoli di stato o obbligazioni e fondi comuni di investimento o gestioni patrimoniali (presentati secondo l’ordine decrescente delle probabilità condizionate) e con bassa probabilità assicurazioni sulla vita.

Infine nella sesta classe le probabilità condizionate maggiori sono associate ad “account” (0.85) e “ira” (0.68). Si distingue anche “life_insurance”, che ha probabilità pari a 0.24. In questa classe, quindi, è alta la probabilità che gli ultracinquantenni possiedano conto corrente o libretto di risparmio, buona che possiedano una pensione integrativa privata, discreta che possiedano un’assicurazione sulla vita.

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6
ampiezza delle classi	0,30	0,19	0,14	0,14	0,13	0,10
account						
0	0,03	0,93	0,02	0,05	0,02	0,15
1	0,97	0,07	0,98	0,95	0,98	0,85
bonds						
0	0,98	1,00	0,78	0,97	0,65	0,99
1	0,02	0,00	0,22	0,03	0,35	0,01
stocks						
0	0,99	1,00	0,41	0,91	0,57	0,95
1	0,01	0,00	0,59	0,09	0,43	0,05
mutual_funds						
0	0,99	1,00	0,54	0,92	0,70	0,94
1	0,01	0,00	0,46	0,08	0,30	0,06
ira						
0	1,00	1,00	0,20	0,90	0,97	0,32
1	0,00	0,00	0,80	0,10	0,03	0,68
life_insurance						
0	1,00	0,88	0,43	0,35	0,86	0,76
1	0,00	0,12	0,57	0,65	0,14	0,24

Tabella 4.2 Probabilità condizionate relative agli indicatori.

Per identificare in modo più immediato i sei segmenti nelle analisi presentate in seguito, si assegna loro un nome, che rispecchi le loro descrizioni e peculiarità. Chiameremo, allora, “conto correntisti” il primo segmento, perché ha probabilità alta solo per “account”; “poveri” il secondo, di cui si evidenzia più che altro il non possesso di prodotti finanziari; “ricchi” il terzo, perché è il segmento con probabilità buone di possesso per la maggior parte dei prodotti finanziari; “assicurati” il quarto, in quanto si distingue dal primo per il possesso di assicurazioni sulla vita; “benestanti” il quinto, in cui le probabilità di possesso sono discrete per molti prodotti finanziari; “previdenti” il sesto, che si distingue dal primo per la buona probabilità associata al possesso di una pensione integrativa.

4.2.2 Profilo dei segmenti sulla base delle covariate.

Grazie all’inserimento delle covariate nel modello, è possibile individuare una descrizione dei segmenti, che li renda più identificabili. Infatti, basandosi solo sul profilo appena tracciato, non si può capire quali siano gli ultracinquantenni in ogni *cluster*. In tabella 4.3 sono riportati i valori delle probabilità di risposta condizionate per le covariate e le relative modalità.

Il segmento dei concorrentisti ha probabilità maggiori associate alle modalità 10 (Italia), 11 (Spagna), 12 (Grecia), 6 (Belgio) e 7 (Francia) di “id_country” (in ordine decrescente per valore di probabilità). La probabilità di essere femmina è maggiore rispetto a quella di essere maschio, così come quella di avere un partner è maggiore di quella di non averlo e quella di essere proprietari della propria abitazione è maggiore di quella di non esserlo. I valori, però, non sono molto elevati. Per “age” la probabilità maggiore è associata alle modalità 3 (settantenni) e 2 (sessantenni); per “marital_status” è associata alle modalità 1 (sposato) e 4 (vedovo); per “occupation” alle modalità 1 (pensionato) e 3 (casalingo); per “fluency_test” alle classi [14,17] e [1,13]; per “isced97” ai livelli 1 (elementare) e 3 (secondaria superiore).

Per il segmento dei poveri, seconda classe latente, le probabilità condizionate maggiori sono associate alle modalità 13 (Polonia), 12 (Grecia) e 14 (Repubblica Ceca) di "id_country"; all'essere femmina (modalità 1 di "female"); alle classi 1 (minori di sessant'anni) e 2 (sessantenni) di "age"; all'avere un compagno (modalità 1 di "partner"); alle modalità 1 (sposato) e 4 (vedovo) di "marital_status"; 1 (pensionato), 3 (casalingo) e 2 (occupato) di "occupation"; al possedere la propria abitazione; ai valori minori di 17 di "fluency_test"; ai livelli 1 (istruzione elementare) e 3 (istruzione secondaria superiore) di "isced97".

Per il terzo segmento, ricchi, le probabilità maggiori sono associate a Svezia (1), Danimarca (2) e Belgio (6) per "id_country"; all'essere maschio (modalità 0 di "female"); alle classi dei minori di sessant'anni (1) e dei sessantenni (2) per "age"; all'avere un compagno[19]; all' essere sposati (1) per "marital_status"; all'essere ancora occupati (2), ma anche all'essere pensionati (1) per "occupation"; al possedere la propria abitazione (1) per "house_own"; ai valori maggiori di 22 in "fluency_test"; ai livelli 5 (istruzione terziaria) e 3 (istruzione secondaria superiore) per "isced_97".

Nel segmento degli assicurati, quarto, abbiamo probabilità condizionate maggiori per Paesi Bassi (5), Germania (4), Polonia (13) e Irlanda (3); per essere femmina; per avere meno di sessant'anni o essere sessantenni; per avere un *partner*; per essere sposati; per essere lavoratori o pensionati; per essere proprietari della propria abitazione; per dare tra 18 e 58 risposte nel *fluency test*; per avere un livello di istruzione secondario di secondo livello, terziario o secondario di primo livello.

Il segmento dei benestanti, quinta classe latente, ha maggior probabilità di abitare in Danimarca, Svezia, Belgio, Svizzera, Italia e Germania; di essere maschio; di avere sessanta o settanta anni; di avere un compagno; di essere sposati o vedovi; di essere pensionati; di possedere la propria abitazione; di rispondere con un numero di animali

[19] È la classe con probabilità maggiore di avere un compagno.

tra il 18 e il 58 nel *fluency test* e di avere un livello di istruzione secondario superiore o terziario.

Infine, nella classe dei previdenti (sesta), con maggiore probabilità si proviene da Repubblica Ceca, Francia e Belgio; si è femmina; si ha meno di sessant'anni o si è sessantenni; si ha un compagno; si è sposati, ma anche separati o divorziati; si lavora o si è pensionati; si è proprietari della propria abitazione; si riporta un valore tra 18 e 58 nel test cognitivo e si ha un livello di istruzione secondario o terziario.

Queste osservazioni, che sembrano sterili, acquisteranno maggiore significato nel prossimo paragrafo, dove si uniranno le descrizioni fatte con gli indicatori e con le covariate, per profilare in modo definitivo e più consono ad esigenze di *marketing* i segmenti.

	conto correntisti	poveri	ricchi	assicurati	benestanti	previdenti
id_country						
1	0,01	0,02	0,34	0,04	0,16	0,03
2	0,04	0,01	0,21	0,03	0,18	0,06
3	0,03	0,04	0,02	0,10	0,02	0,00
4	0,08	0,01	0,06	0,14	0,12	0,02
5	0,10	0,01	0,03	0,21	0,09	0,02
6	0,11	0,01	0,13	0,07	0,13	0,11
7	0,10	0,00	0,10	0,08	0,02	0,26
8	0,03	0,01	0,07	0,03	0,12	0,05
9	0,07	0,02	0,02	0,09	0,02	0,00
10	0,15	0,09	0,00	0,04	0,12	0,00
11	0,12	0,05	0,02	0,02	0,01	0,06
12	0,11	0,26	0,00	0,02	0,01	0,00
13	0,00	0,33	0,00	0,12	0,00	0,00
14	0,04	0,16	0,02	0,00	0,00	0,38
female						
0	0,42	0,38	0,59	0,48	0,52	0,41
1	0,58	0,62	0,41	0,52	0,48	0,59
age						
1	0,21	0,30	0,54	0,58	0,12	0,56
2	0,31	0,29	0,39	0,30	0,36	0,30
3	0,32	0,26	0,06	0,09	0,35	0,10
4	0,17	0,15	0,01	0,03	0,16	0,04
partner						
0	0,43	0,46	0,13	0,22	0,34	0,35
1	0,57	0,54	0,87	0,78	0,66	0,65
marital_status						
1	0,55	0,52	0,83	0,74	0,64	0,61
2	0,10	0,11	0,08	0,10	0,07	0,19
3	0,08	0,07	0,05	0,05	0,08	0,07
4	0,27	0,30	0,04	0,11	0,21	0,12
occupation						
1	0,62	0,58	0,28	0,33	0,78	0,39
2	0,14	0,14	0,61	0,46	0,11	0,50
3	0,17	0,15	0,03	0,10	0,07	0,05
4	0,03	0,07	0,03	0,06	0,02	0,03
5	0,02	0,04	0,03	0,03	0,00	0,03
6	0,01	0,02	0,02	0,01	0,01	0,00
house_own						
0	0,32	0,30	0,12	0,28	0,22	0,24
1	0,68	0,70	0,88	0,73	0,78	0,76
fluency_test						
1-13	0,25	0,40	0,02	0,11	0,08	0,09
14 - 17	0,27	0,28	0,07	0,17	0,17	0,15
18 - 21	0,24	0,19	0,19	0,25	0,24	0,23
22 - 25	0,13	0,08	0,23	0,21	0,23	0,22
26 - 58	0,11	0,05	0,48	0,26	0,28	0,30
media	16,69	14,36	24,59	20,68	21,10	21,54
isced97						
0	0,07	0,07	0,00	0,02	0,00	0,02
1	0,35	0,49	0,08	0,14	0,20	0,18
2	0,22	0,13	0,13	0,20	0,17	0,23
3	0,24	0,23	0,32	0,37	0,31	0,36
4	0,02	0,03	0,06	0,04	0,05	0,04
5	0,10	0,05	0,41	0,23	0,26	0,16
6	0,00	0,00	0,01	0,00	0,00	0,01

Tabella 4.3 Probabilità condizionate per le covariate.

4.3 Profilo definitivo dei segmenti e analisi della loro efficacia ed efficienza.

Di seguito daremo una descrizione più usufruibile dei segmenti individuati.

Osservando i *profile plot* (Figura 6.1 e Figura 6.2) dei sei segmenti è possibile farsi una prima idea dei profili e di alcune loro caratteristiche. In questi diagrammi sono rappresentati in ordinata le probabilità condizionate e in ascissa indicatori e covariate. Per le variabili binarie è rappresentata la modalità 1, che indica il possesso di beni finanziari o della caratteristica registrata nella relativa covariata.

Nel definire i profili, con i termini "sessantenni" e "settantenni" si fa riferimento alle modalità 2 e 3 di "age".

Il segmento dei conto correntisti, così chiamato perché manifesta un altissimo tasso di penetrazione solo per conto corrente o libretto di risparmio, è composto da abitanti dell'Europa Mediterranea, della Francia e del Belgio, ma soprattutto da italiani e spagnoli. La conto correntista tipo è una signora minore di settant'anni, principalmente con un compagno, il marito, ma potrebbe essere anche vedova. È pensionata, o casalinga, e con buona probabilità abita in una casa di proprietà. Non ha capacità cognitive brillanti, perché al di sotto della media europea, né ha studiato molto: è probabile si sia limitata a un'istruzione elementare. Tuttavia potrebbe aver concluso l'istruzione secondaria di primo o secondo livello.

Il povero tipo, che non possiede prodotti finanziari, ad eccezione di qualche assicurazione sulla vita, è anch'esso una signora, giovane: con buona probabilità ha meno di sessant'anni, ma potrebbe essere sulla sessantina. Abita per lo più nei paesi dell'Est o in Grecia. Ha un compagno ed è sposata o vedova. È una pensionata, che possiede la propria abitazione. Neanche le sue capacità cognitive sono eccezionali: sono al di sotto della media europea e della conto correntista. Ha un livello di istruzione elementare o al massimo secondario superiore.

Il segmento dei ricchi, formato da proprietari di tutti i prodotti finanziari, sebbene con probabilità diverse, è presente nei paesi scandinavi e in Belgio (con minore probabilità). I ricchi sono giovani signori, con un'età inferiore ai sessant'anni o sulla sessantina. Hanno una compagna, che è molto probabilmente la moglie. Lavorano ancora, ma alcuni, forse i più vecchi del segmento, sono pensionati. Possiedono, con probabilità massima fra tutti i segmenti, la propria abitazione, a conferma del fatto che siano i più ricchi. Hanno ottime capacità cognitive, superiori alla media europea, e un alto livello di istruzione (terziario per lo più o secondario di secondo livello).

Gli assicurati, che possiedono conto corrente o libretto postale e assicurazioni sulla vita (e per questo si distinguono dai conto correntisti), sono belgi e tedeschi e, con probabilità minore, polacchi e irlandesi. Sono delle giovani signore, con meno di sessant'anni o sessantenni. Hanno un compagno, presumibilmente il marito. Sono lavoratrici o pensionate, che verosimilmente possiedono la propria abitazione. Le capacità cognitive sono buone, superiori alla media europea, ma non di molto. Il loro grado di istruzione è superiore al livello elementare.

Nel segmento dei benestanti, che possiedono conto corrente e con probabilità inferiori gli altri prodotti finanziari, eccetto le pensioni integrative, sono presenti soprattutto i paesi scandinavi, ma anche Germania, Belgio e Svizzera. I benestanti sono verosimilmente signori, ma rispetto ai ricchi sono più anziani: sessantenni o settantenni, con probabilità pressoché uguale. Infatti sono pensionati. Hanno una compagna, ma non con altissima probabilità, tanto che sono sposati, ma potrebbero essere anche vedovi. Possiedono la propria abitazione. Hanno buone capacità cognitive, al di sopra della media, e un livello di istruzione secondario superiore o terziario.

Infine i previdenti, che possiedono conto corrente o libretto postale, pensione integrativa e con probabilità più bassa assicurazione sulla vita, sono francesi, cechi e in parte belgi. Sono femmine di età inferiore ai sessant'anni, ma anche sessantenni, con un compagno.

Probabilmente sono sposate, ma qualcuna è separata o divorziata. Lavorano ancora, ma alcune sono pensionate. Possiedono la propria abitazione. Hanno buone capacità cognitive, superiori alla media, sebbene non di molto e un livello di istruzione secondario o terziario.

Individuare dei segmenti non è sufficiente perché la segmentazione sia efficiente ed efficace. Perciò in seguito si verificherà se i segmenti possiedono le caratteristiche presentate nel primo capitolo.

I sei segmenti, per come sono costruiti, sono omogenei al loro interno ed eterogenei tra di loro, caratteristica desumibile anche dai *profile plot*. Eventuali somiglianze in realtà sono smentite dalle probabilità condizionate. Ad esempio, in tutti i segmenti la probabilità di avere un compagno è maggiore a quella di non averlo, ma, confrontando le probabilità condizionate di ogni classe, si nota quanto siano diverse. L'eliminazione dal modello di "hhsize", che assume valori molto simili nelle classi e per questo non è significativa, aumenta l'eterogeneità tra i segmenti. Pertanto questa scelta ha una duplice rilevanza: statistica, perché migliora il modello, e di *marketing*, perché migliora l'efficacia della segmentazione. Le basi utilizzate (gli indicatori) sono pertinenti all'obiettivo di segmentare il mercato europeo dei prodotti finanziari, con riferimento alla popolazione degli ultra sessantenni. Pertanto l'efficacia dei segmenti individuati è garantita.

Descrivere i segmenti utilizzando delle covariate ne garantisce l'identificabilità e l'accessibilità: il profilo dettagliato dei *cluster* facilita la raggiungibilità da parte delle aziende. L'ampiezza non troppo esigua assicura la consistenza, pertanto è presumibile che i segmenti siano profittevoli. Le basi, su cui sono costruiti, sono misurabili e non particolarmente specifiche. Di conseguenza i *cluster* sono altrettanto misurabili (la loro ampiezza è già stata misurata) e poco volatili. Inoltre, la proprietà dei prodotti finanziari è ben definita nelle classi e, per alcuni prodotti in particolare, è legata alle politiche statali di *welfare* o finanziarie (come per esempio la tassazione sul possesso di determinati beni), che, di solito, non variano repentinamente (almeno quelle di *welfare*). Tuttavia ogni azienda dovrebbe monitorare l'andamento della

proprietà di prodotti finanziari con dati in suo possesso o rifacendosi alle pubblicazioni degli istituti statistici nazionali ed europei, che spesso riguardano questo argomento. La capacità di risposta e la propositività difficilmente sono analizzabili a priori. Ciò nonostante, anche l'efficienza dei segmenti è garantita.

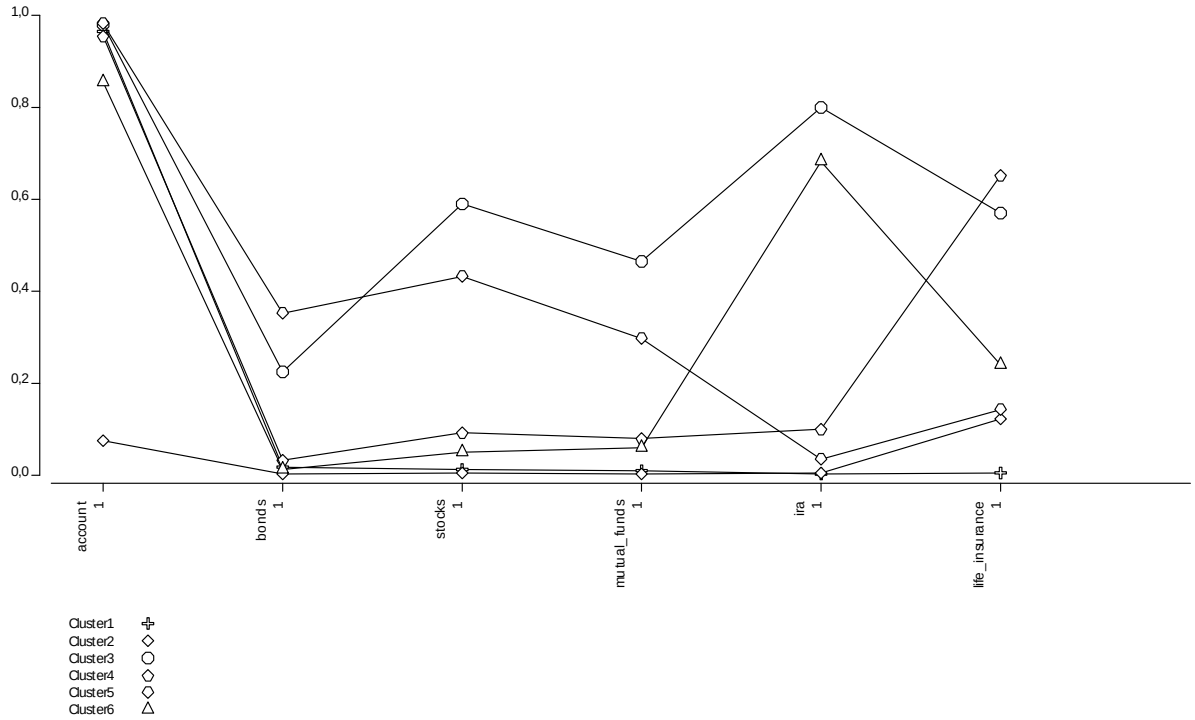

Figura 4.1 *Profile plot* dei sei segmenti con i soli indicatori.

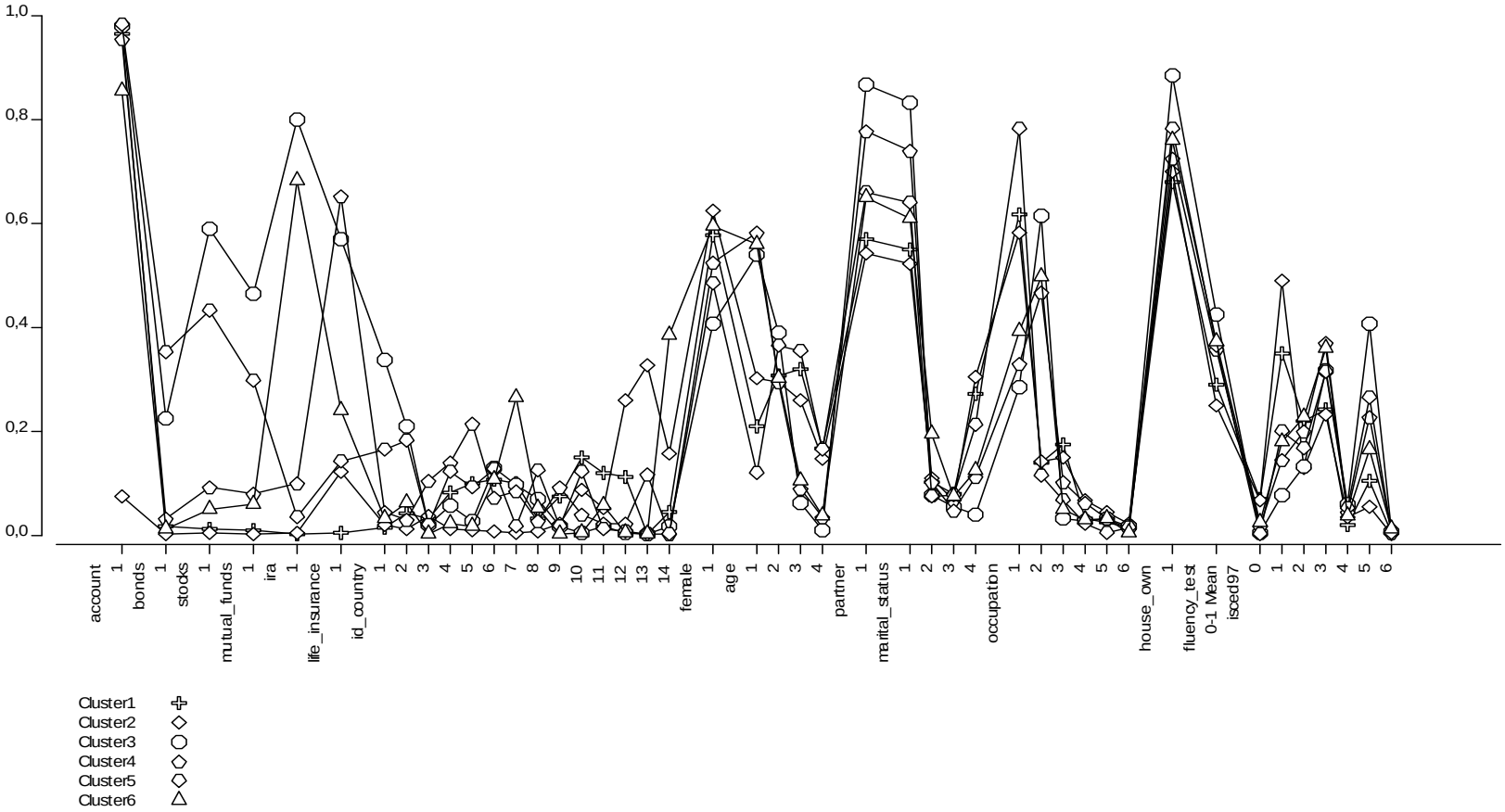

Figura 4.2 *Profile plot* dei segmenti con indicatori e covariate

4.4 Analisi dei residui bivariati.

Dai punti di vista del *marketing* e statistico, il modello a sei classi latenti sembra buono. L'analisi delle *Bivariate Residuals*, però, fa sorgere qualche dubbio: nella matrice dei residui bivariati (Tabella 4.4) molti valori sono maggiori di uno. Pertanto alcune coppie di variabili osservate sono dipendenti, invalidando l'ipotesi di indipendenza locale.
Una delle alternative possibili per rilassare l'indipendenza locale è stimare un nuovo modello, in cui è permesso alle variabili associate di essere dipendenti. Tra di esse, quindi, si introduce un effetto diretto. I modelli stimati con effetti diretti non presentano un adattamento migliore a quello classico: i valori di BIC non diminuiscono di molto e molte BVR hanno ancora valore maggiore a uno. Ad esempio nel modello a sei classi latenti con un effetto diretto tra "bonds" e "stocks", a cui è associato il valore di BVR maggiore, BIC è pari a 86241,1439 e L^2 a 72767,9788.
Stimando il modello con i soli indicatori, nessuna BVR è maggiore a uno (Tabella 4.5). Sembra, quindi, che l'uso di numerose covariate appesantisca il modello e ne peggiori l'adattamento. D'altra parte senza covariate verrebbe meno l'efficienza dei segmenti. Si è di fronte, quindi, a un *trade-off* tra esigenze statistiche e di *marketing*, nel senso che il modello *cluster* con tutte le covariate sarebbe da rifiutare, ma così facendo i risultati sarebbero più difficili da utilizzare. Poiché il profilo dei segmenti corrisponde alle aspettative su di essi, basate sulle analisi nel terzo capitolo, si può ritenere valida la segmentazione implementata dal modello classico.
L'adattamento non del tutto soddisfacente di questi modelli potrebbe essere dovuto alla presenza di una struttura gerarchica a due livelli (individui nei paesi), di cui il modello classico a classi latenti non tiene conto. Si presume, pertanto che la stima di modelli a classi latenti multilivello, che considerano anche la gerarchia insita nei dati, porti alla stima di un modello migliore, sotto il profilo dell'adattamento.

	account	bonds	stocks	mutual_funds	ira	life_insurance
account	.					
bonds	0,32	.				
stocks	7,15	9,38	.			
mutual_funds	4,78	0,19	2,82	.		
ira	0,29	1,87	0,10	0,60	.	
life_insurance	0,00	0,57	0,82	1,17	0,01	.

Tabella 4.4 Matrice delle BVR per il modello a sei classi latenti con le covariate.

	account	bonds	stocks	mutual_funds	ira	life_insurance
account	.					
bonds	0,0008	.				
stocks	0	0	.			
mutual_funds	0,0187	0,0269	0,0582	.		
ira	0	0,023	0,1316	0,0167	.	

Tabella 4.5 Matrice delle BVR per il modello a sei classi latenti senza covariate.

Capitolo 5

Segmentazione del mercato con il modello a classi latenti multilivello.

5.1 Introduzione.

I dati dell'indagine SHARE sembrano presentare una struttura gerarchica a due livelli, in cui gli individui sono raggruppabili per i paesi di appartenenza. Il modello a classi latenti più adatto a spiegare strutture annidate dei dati è quello multilivello. Esso prevede l'introduzione di una seconda variabile latente, a livello due (gruppi), e la possibilità che i parametri varino tra i gruppi. Inoltre presuppone che le osservazioni nei gruppi siano correlate, per la loro tendenza ad appartenere alla stessa classe latente di livello uno.

Per evitare instabilità delle stime, l'approccio adottato è a effetti casuali: i parametri tra i gruppi variano secondo una distribuzione non specificata a priori. In particolare si tratta di modelli *random effects* di tipo non parametrico, perché la variabile latente introdotta a livello gruppo ha distribuzione discreta. Il modello multilivello a effetti casuali discreti di tipo non parametrico fa assunzioni distributive meno stringenti, è più semplice computazionalmente e si adatta meglio ai problemi di ricerca.

Distinguiamo le unità in due gruppi: le unità di livello uno sono gli individui, le unità di livello due sono i paesi di appartenenza.

L'obiettivo che ci si pone è di classificare le unità di primo livello in classi latenti, che tengano conto della loro tendenza a possedere prodotti finanziari, e di classificare i paesi in gruppi, secondo la loro tendenza ad appartenere ai *cluster* individuati al primo livello. A livello uno, poi, sono introdotte delle covariate, che aiutano a disegnare un profilo dei segmenti individuati dall'analisi.

D'ora in poi si farà riferimento a *cluster* o segmenti, per le classi latenti individuate a livello uno (consumatori), e a gruppi, per le classi latenti individuate a livello gruppo (paesi).

5.1.1 Stima del modello.

Come per l'analisi a classi latenti classica, si sono stimati numerosi modelli con l'algoritmo EM[20], con insiemi di valori iniziali per le procedure iterative di stima differenti. In particolare si sono stimati modelli con diverse combinazioni di valori del numero dei *cluster* (livello uno, *T* seguendo la nomenclatura del capitolo primo) e dei gruppi (livello due, *M*). Per ogni tipologia, si è scelto il migliore per adattamento. Un riepilogo dei risultati ottenuti si trova nella tabella 5.1, che contiene il valore della statistica BIC del modello che, per ogni combinazione di *T* e *M*, ne presenta il valore minore.

gruppi	1	2	3	4	5	6	7	8	9
cluster									
1	111908,8	11918,71	111928,7	11938,62	119948,6	119958,5	111968,5	111978,4	111988,4
2	98478,49	96326,03	95647,19	95404,72	95365,06	95320,52	95340,43	95360,15	95376,23
3	96660,87	93050,42	91802,44	91456,62	91314,02	90970,38	91157,68	91030,53	90992,55
4	95556,2	91538,21	90335,31	89847,97	89341,22	89567,82	89037,13	89435,55	89371,56
5	95178,72	90877,56	89448,59	88876,4	88628,67	88153,24	87951,66	88202,53	87934,2
6	95152,71	91256,09	89174,21	88438,67	88416,37	87916,33	87559,38	87031,78	88108,05

Tabella 6.1 Valori di BIC dei modelli multilivello stimati per diverse combinazioni di *T*, *cluster*, e *M*, gruppi. Le celle verdi individuano i modelli con il miglior BIC di riga, quelle gialle con il miglior BIC di colonna. In celeste è evidenziato il modello con BIC minore in assoluto.

[20] L'algoritmo EM, usato per la stima di massima verosimiglianza dei modelli multilivello, nel passo E utilizza l'algoritmo *upward-downward* , che si basa sull'ipotesi di indipendenza condizionale (osservazioni a livello più basso sono indipendenti dalle altre, data l'appartenenza alle classi di livello più alto). Per maggiori informazioni si vedano Vermunt (2003b, 2007) e Vermunt e Madigson (2005).

Come suggeriscono Vermunt (2003b, 2007) e Vermunt e Madigson (2005), tra tutti, si sceglie il modello che presenta il minimo valore di BIC, ovvero quello con 6 *cluster* e 8 gruppi. Analizzandone la numerosità delle classi, non si sono rilevate ampiezze troppo esigue. Inoltre il valore di BIC è notevolmente migliore rispetto agli altri modelli. Pertanto la scelta è soddisfacente, dal punto di vista sia di *marketing* che statistico.

5.2 I *cluster*:

Le classi latenti a livello gruppo hanno un'ampiezza tale da non metterne in dubbio la consistenza: la prima comprende il 35% degli europei ultracinquantenni, la seconda ne comprende il 16,5%, la terza il 14%, la quarta il 13%, la quinta l'11% e la sesta il 10%.
Adottando lo schema proposto nel capitolo precedente, in un primo momento si profilerà la tendenza dei segmenti a possedere prodotti finanziari. In seguito la descrizione sarà definita con le informazioni tratte dalle covariate.

5.2.1 Profilo dei *cluster* sulla base degli indicatori.

In figura 5.1 sono rappresentati i profili dei segmenti con riferimento ai soli indicatori. Emerge una sostanziale differenza tra i profili, che sembrano assimilabili a quelli ottenuti con l'analisi a classi latenti standard. Per confermare queste supposizioni, si analizzano i profili dei *cluster*, utilizzando le probabilità di risposta, condizionate all'appartenenza ad una data classe latente (Tabella 5.2). Si prendono in considerazione le modalità 1 degli indicatori, che indicano il possesso di prodotti finanziari.
La prima classe latente, la più ampia, presenta una probabilità molto elevata solo in "account". Ha, quindi, un alto tasso di penetrazione per conto corrente e libretto di risparmio.

La seconda classe, invece, presenta probabilità molto basse per tutte le variabili, ad eccezione di un esiguo 12% per “life_insurance”. Per gli ultracinquantenni di questa classe, pertanto, vi sono tassi di penetrazione praticamente nulli per tutti i prodotti finanziari.
La terza classe ha probabilità molto elevata per “account” ed elevata per “life_insurance”. Le probabilità per “ira”, “stocks” e “mutual_funds” sono tra il 15% e il 10%. I tassi di penetrazione, pertanto, sono alti per conto corrente o libretto di risparmio e assicurazioni sulla vita, bassi per assicurazioni sulla vita, azioni o partecipazioni e fondi comuni di investimento o gestioni patrimoniali.
Nella quarta classe le probabilità condizionate di possesso sono più articolate: oltre a quella molto alta per “account”, le probabilità per “stocks”, “bonds” e “mutual_funds” sono maggiori del 30%, quindi discrete. “life_insurance” ha probabiltà 0.13. Questa classe, perciò, ha tassi di penetrazione alti per conto corrente o libretto di risparmio, discreti per azioni o partecipazioni e fondi comuni di investimento o gestioni patrimoniali, molto bassi per le assicurazioni sulla vita.
Nella quinta classe le probabilità sono alte per “account” e “ira”, buone per “stocks”, “life_insurance” e “mutual_funds” e discrete per “bonds”. I tassi di penetrazione, perciò, sono alti per conto corrente e pensioni integrative, scendono, ma restano buoni, per azioni o partecipazioni, assicurazioni sulla vita e fondi comuni di investimento, calano ancora per titoli di stato e obbligazioni, ma sono discreti.
La sesta classe latente ha probabilità alta per “account”, buone per “ira” e discrete per “life_insurance”. Qundi i tassi di penetrazione di conto corrente o libretto di risparmio sono alti, buoni per le pensioni integrative e più che discreti per le assicurazioni sulla vita.
Rispetto ai profili individuati con il modello a classi latenti tradizionale, le tendenze nel possesso dei prodotti finanziari sono confermate, pertanto i profili delle classi sono robusti rispetto al modello utilizzato e la denominazione dei segmenti può essere utilizzata anche per i *cluster*. Cambia, invece, l’ampiezza di alcune classi latenti, ma senza stravolgimenti. Le classi dei conto correntisti (*cluster* 1) e dei poveri

(*cluster* 2) si confermano le più numerose, ma la prima aumenta del 5% la propria ampiezza, la seconda la diminuisce del 3%. Gli assicurati (*cluster* 3) guadagnano ampiezza, ma solo nel senso cha passano da quarta a terza classe: in realtà la percentuale di ultracinquantenni assicurati non cambia. La quarta classe, i benestanti, conferma anch'essa l'ampiezza del modello tradizionale, ma passa da quinta a quarta classe. Questi ultimi due cambiamenti sono dovuti alla diminuzione del 3% dell'ampiezza della classe dei ricchi (*cluster* 5). Infine la classe dei previdenti si conferma la meno ampia, con la medesima percentuale.

	cluster 1	*cluster* 2	*cluster* 3	*cluster* 4	*cluster* 5	*cluster* 6
ampiezza dei *cluster*	0,35	0,16	0,14	0,13	0,11	0,10
account						
0	0,02	0,97	0,04	0,02	0,02	0,14
1	0,98	0,03	0,96	0,98	0,98	0,86
bonds						
0	0,98	1,00	0,94	0,68	0,77	0,99
1	0,02	0,00	0,06	0,32	0,23	0,01
stocks						
0	1,00	1,00	0,86	0,57	0,39	0,95
1	0,00	0,00	0,14	0,43	0,61	0,05
mutual_funds						
0	0,99	1,00	0,89	0,71	0,52	0,94
1	0,01	0,00	0,11	0,29	0,48	0,06
ira						
0	1,00	0,99	0,85	0,97	0,18	0,32
1	0,00	0,01	0,15	0,03	0,83	0,68
life_insurance						
0	0,95	0,88	0,30	0,87	0,44	0,77
1	0,05	0,12	0,70	0,13	0,56	0,23

Tabella 5.2 Probabilità di risposta condizionate per gli indicatori.

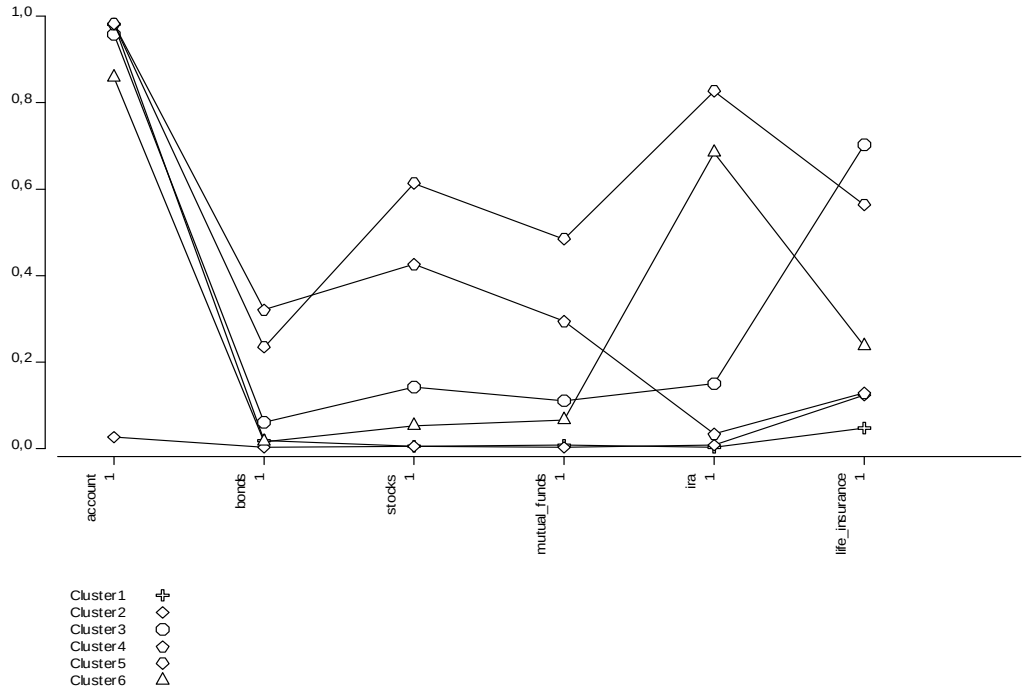

Figura 5.1 *Profile plot* dei sei segmenti con riferimento solo agli indicatori.

5.2.2 Profilo dei *cluster* sulla base di indicatori e covariate.

Nella tabella 5.3 sono riportate le probabilità condizionate per le covariate, usate per costruire un profilo dei segmenti individuati, che li renda riconoscibili. Nella descrizione, quando si elencano le caratteristiche, si prendono sempre in considerazione le modalità delle variabili in ordine decrescente di probabilità condizionata, se non specificato diversamente. Inoltre quando si parlerà di sessantenni e settantenni si farà riferimento rispettivamente alle modalità 2 e 3 di "age".

Il primo *cluster*, i contocorrentisti, è formato con buona probabilità da femmine di età compresa tra i sessanta e i sessantanove anni o tra i settanta e i settantanove anni. Hanno un compagno, ma non con altissima probabilità (0.57). Sono sposate (0.55) o vedove (0.27), molto probabilmente pensionate. Tuttavia potrebbero essere casalinghe o occupate. Verosimilmente possiedono la propria abitazione. Le capacità cognitive non sono brillanti: il numero di risposte più probabili nel

fluency test varia tra le quattordici e le diciassette, una e tredici, diciotto e ventuno. Il livello di istruzione più probabile è quello elementare, seguito dal secondario di secondo livello e di primo livello. Il loro nucleo familiare è composto da due persone (con probabilità 0.45) o una (0.36), raramente da tre e quattro (0.17).

Il segmento dei poveri è composto da femmine (con probabilità 0.62), con un'età inferiore di sessant'anni, ma anche sessantenni e settantenni. I poveri hanno con buona probabilità un compagno, con cui sono sposati o convivono legalmente (0.52 la probabilità associata). Potrebbero essere verosimilmente anche vedovi (0,30). Sono pensionati: è bassa la probabilità che siano casalinghi o occupati, molto basse le altre. Sono proprietari delle loro abitazione. Le capacità cognitive non brillano: con probabilità maggiore il numero di animali ricordati variano tra uno e tredici e con probabilità minori tra quattordici e diciassette e tra diciotto e ventuno. Il livello di istruzione è verosimilmente elementare (0.48), ma potrebbe essere secondario superiore o inferiore. Il numero di componenti della famiglia dei poveri è due, uno o poco probabilmente tre e quattro.

Il segmento degli assicurati è composto con uguale probabilità da maschi e femmine. L'età è minore di sessant'anni con alta probabilità, tra i settanta e i sessanta con discreta probabilità. Gli assicurati hanno un partner (0.82), che è lo sposo o il convivente legalmente riconosciuto (0.78). Lavorano, del resto sono piuttosto giovani, tuttavia potrebbero essere pensionati. Sono proprietari della loro abitazione (0.75 è la probabilità associata). Hanno buone capacità cognitive: ricordano tra i ventisei e i cinquantotto nomi di animali, ma anche tra i diciotto e ventuno e i ventidue e venticinque. Con buona probabilità hanno un livello di istruzione secondario di secondo livello, con minore probabilità terziario e con minore ancora secondario di primo livello. I componenti del nucleo familiare sono verosimilmente due, ma potrebbero essere anche tre o quattro.

I benestanti sono con maggiore probabilità maschi (0.51), tra gli ottanta e i sessanta anni. Potrebbero essere anche ultra ottantenni (0,17).

Ovviamente sono per lo più pensionati: molto bassa è la probabilità di essere occupati (0.10). Verosimilmente hanno un compagno, che è lo sposo o convivente legalmente riconosciuto, tuttavia potrebbero essere vedovi. Possiedono la propria abitazione. Nonostante l'età avanzata rispetto agli altri segmenti, le capacità cognitive sono buone, perché i valori del *fluency test* più probabili sono tra i ventisei e i cinquantotto, tra i diciotto e i ventuno e tra i ventidue e venticinque. Il livello di istruzione più probabile è il secondario superiore, seguito da terziario e secondario inferiore. Il numero di componenti della famiglia di un benestante è due o uno.

I ricchi sono più probabilmente maschi, al di sotto dei sessant'anni o sessantenni, che lavorano ancora. Potrebbero essere anche pensionati (0.29 la probabilità condizionata di esserlo). Verosimilmente hanno una compagna, che è la moglie o la convivente legalmente riconosciuta. Possiedono la propria abitazione con altissima probabilità (0.89). Le capacità cognitive sono davvero brillanti: la probabilità condizionata di ricordare tra i ventisei e i cinquantotto nomi di animali è 0.49, molto più bassa la probabilità di ricordarne tra i ventidue e venticinque e tra i diciotto e ventuno. Il livello di istruzione è elevato: il loro grado è terziario (0.41), secondario superiore (0.31). Il nucleo familiare è composto da due, o al massimo tre e quattro, persone.

La classe dei previdenti, infine, è composta più probabilmente da femmine, di età inferiore ai sessanta anni o sessantenni. Molto bassa la probabilità di essere sessantenni. Verosimilmente hanno un compagno, sono sposate o convivono legalmente; tuttavia potrebbero essere separate o vedove. Molto probabilmente lavorano o sono pensionate e possiedono la propria abitazione. Hanno buone capacità cognitive: ricordano tra ventisei e cinquantotto nomi di animali con molto buona probabilità (0.31) e tra i diciotto e ventuno e tra i ventidue e venticinque con buona probabilità (0.23). Il livello di istruzione è verosimilmente secondario superiore, ma potrebbe essere secondario inferiore e con probabilità pressoché uguale, non molto alta, primario o terziario. Il

numero di componenti del nucleo familiare è molto probabilmente due, tuttavia potrebbe essere uno oppure tre o quattro.

	cluster 1	*cluster* 2	*cluster* 3	*cluster* 4	*cluster* 5	*cluster* 6
female						
0	0,42	0,38	0,50	0,51	0,60	0,41
1	0,58	0,62	0,50	0,49	0,40	0,59
age						
1	0,22	0,30	0,67	0,10	0,53	0,56
2	0,31	0,29	0,27	0,37	0,40	0,30
3	0,31	0,26	0,04	0,37	0,06	0,11
4	0,16	0,15	0,02	0,17	0,01	0,04
partner						
0	0,43	0,46	0,18	0,35	0,13	0,35
1	0,57	0,54	0,82	0,65	0,87	0,65
marital_status						
1	0,56	0,52	0,78	0,63	0,83	0,61
2	0,10	0,11	0,10	0,07	0,07	0,19
3	0,08	0,07	0,04	0,08	0,05	0,08
4	0,27	0,30	0,08	0,22	0,04	0,12
occupation						
1	0,61	0,58	0,26	0,79	0,29	0,39
2	0,15	0,14	0,54	0,10	0,61	0,50
3	0,17	0,15	0,09	0,07	0,03	0,05
4	0,03	0,07	0,07	0,02	0,02	0,03
5	0,02	0,04	0,03	0,01	0,03	0,03
6	0,01	0,02	0,01	0,01	0,02	0,00
house_own						
0	0,32	0,30	0,25	0,22	0,11	0,24
1	0,68	0,70	0,75	0,78	0,89	0,76
fluency_test						
1-13	0,25	0,40	0,08	0,08	0,02	0,09
14 - 17	0,27	0,28	0,15	0,17	0,07	0,15
18 - 21	0,24	0,19	0,24	0,24	0,18	0,23
22 - 25	0,14	0,08	0,22	0,23	0,24	0,23
26 - 58	0,11	0,05	0,30	0,28	0,49	0,31
media	16,70	14,40	21,47	21,06	24,66	21,62
isced97						
0	0,07	0,07	0,01	0,00	0,00	0,02
1	0,35	0,49	0,11	0,21	0,08	0,18
2	0,22	0,13	0,19	0,17	0,13	0,22
3	0,25	0,23	0,38	0,31	0,31	0,36
4	0,02	0,03	0,05	0,05	0,06	0,04
5	0,10	0,05	0,26	0,26	0,41	0,17
6	0,00	0,00	0,01	0,01	0,01	0,01
hhsize						
1	0,36	0,33	0,13	0,34	0,11	0,26
2	0,45	0,36	0,52	0,58	0,64	0,50
3	0,17	0,22	0,30	0,08	0,22	0,22
4	0,03	0,09	0,06	0,00	0,02	0,03

Tabella 5.3 Probabilità condizionate per le covariate.

Anche il profilo completo dei segmenti (Figura 5.2), a conferma della sua validità e robustezza, non varia granché rispetto a quello individuato col modello classico, tranne che per la presenza della variabile “hhsize” e quindi di informazioni sul numero di componenti del nucleo familiare.

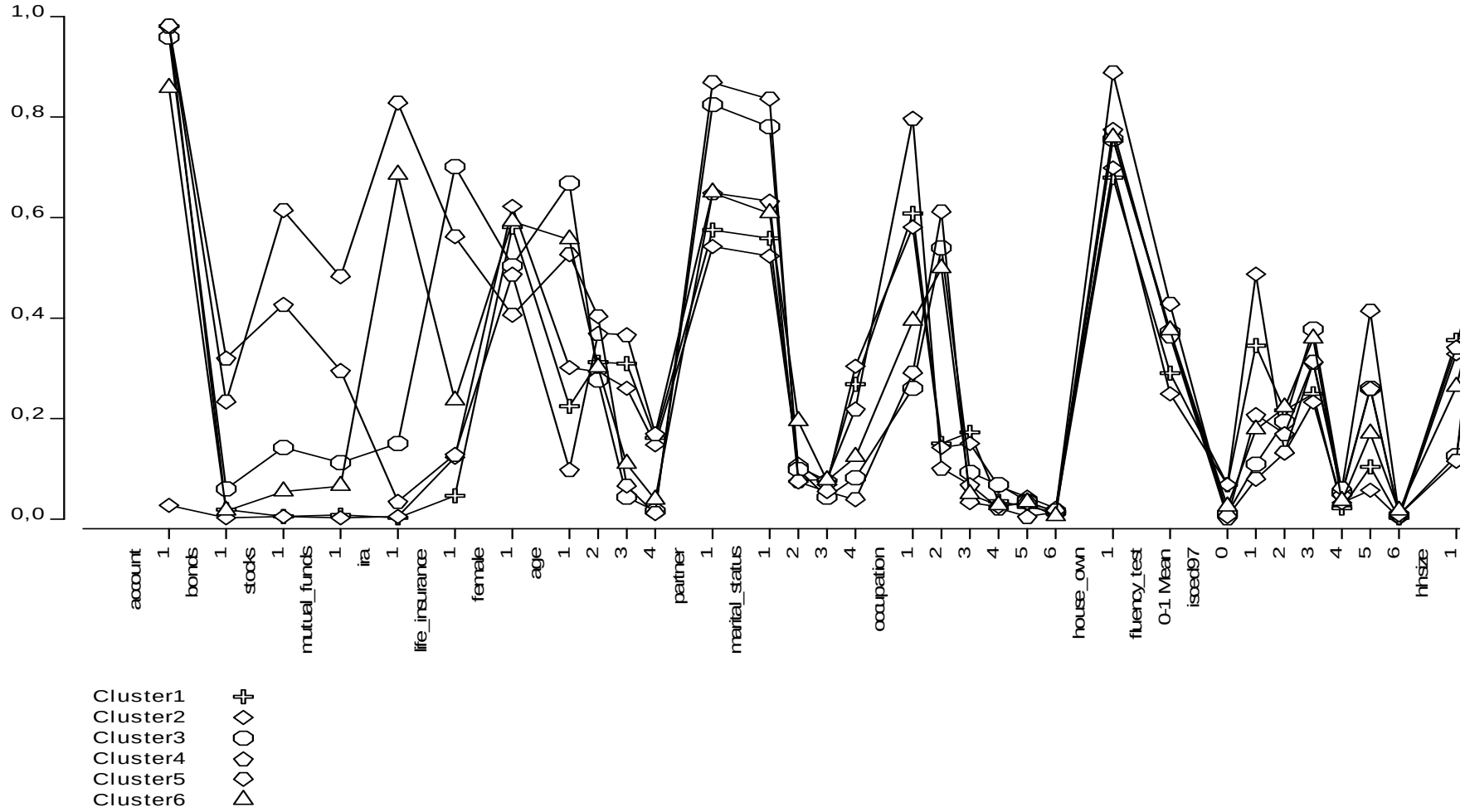

Figura 5.2 *Profile plot* dei *cluster* relativo agli indicatori e alle covariate.

5.3 I gruppi.

Agli otto gruppi di unità di livello due sono assegnati, attraverso le probabilità di appartenenza a posteriori, i paesi che ne fanno parte. Nel nostro modello, tutti i paesi sono assegnati al proprio gruppo con probabilità pari a uno.

Il primo gruppo è formato da Danimarca, Belgio e Svizzera; il secondo da Germania e Paesi Bassi; il terzo da Irlanda e Austria; il quarto da Grecia e Polonia; il quinto da Italia e Spagna. Seguono tre gruppi formati da un solo paese, ovvero Francia (gruppo 6), Svezia (gruppo 7) e Repubblica Ceca (gruppo 8).

La suddivisione potrebbe sembrare sfilacciata, tuttavia nelle precedenti analisi Francia, Repubblica Ceca e Svezia presentano dei comportamenti caratteristici. In particolare la Svezia si è evidenziata come il paese con maggiori tassi di penetrazione per tutti i prodotti finanziari, anche rispetto alla Danimarca, a cui spesso è stata accomunata. La Francia si distingue dagli altri paesi perché presenta un alto tasso di penetrazione per le pensioni integrative. La Repubblica Ceca per alcuni aspetti sembra assimilabile a Polonia e Grecia, ma per altri, ad esempio il possesso di conti correnti o libretti di risparmio e di pensioni integrative, manifesta delle peculiarità. Questi tre gruppi mono-paese, quindi, hanno ragione d'esserlo.

Le coppie Grecia e Polonia, Irlanda e Austria, Germania e Paesi Bassi hanno manifestato, finora, tendenze molto simili al loro interno. In particolare, la prima manifesta bassi tassi di penetrazione, mentre la seconda e la terza riportano spesso frequenze di possesso analoghe nell'analisi preliminare. Infatti Irlanda e Austria hanno percentuali di possesso simili per "bonds", "mutual_funds" e "ira" e differenze non sostanziali negli altri indicatori; Germania e Paesi Bassi hanno percentuali simili in "account", "stocks", "mutual_funds" e "life_insurance", non sostanzialmente dissimili in "bonds" e "ira".

Il raggruppamento di Danimarca, Belgio e Svizzera potrebbe sembrare poco verosimile. Nel modello tradizionale, tuttavia, sono assegnati allo

stesso segmento, i benestanti. Inoltre hanno manifestato buoni tassi di penetrazione per molti prodotti finanziari e l'assenza di tendenze peculiari nettamente definite al loro interno: nessuno di questi paesi, infatti, si distingue per un alto tasso di penetrazione per particolari prodotti, ma piuttosto per la diffusione di tutti i prodotti.
Alla luce di queste osservazioni, che tengono conto delle analisi precedentemente svolte, nessun raggruppamento sembra poco significativo.

5.4 Gruppi e *cluster*.

In tabella 5.4 sono presentate le probabilità per ogni gruppo di appartenere ai *cluster*.

	gruppo1	gruppo 2	gruppo 3	gruppo 4	gruppo 5	gruppo 6	gruppo 7	gruppo 8
conto correntisti	0,31	0,45	0,44	0,31	0,53	0,37	0,07	0,17
poveri	0,04	0,04	0,14	0,56	0,14	0,01	0,07	0,34
assicurati	0,09	0,29	0,31	0,11	0,06	0,09	0,09	0,00
benestanti	0,25	0,16	0,05	0,01	0,18	0,03	0,23	0,00
ricchi	0,20	0,05	0,04	0,00	0,04	0,14	0,49	0,03
previdenti	0,11	0,02	0,02	0,00	0,05	0,35	0,04	0,46

Tabella 5.4 Probabilità di ogni gruppo di appartenere a un determinato *cluster*.
In celeste sono evidenziate le probabilità maggiori a 0.10.

Prendendo in considerazione le probabilità maggiori al 10% in ordine decrescente di valore, al primo gruppo appartengono i segmenti dei conto correntisti, dei benestanti, dei ricchi e dei previdenti; al secondo i segmenti dei conto correntisti, degli assicurati e dei benestanti; al terzo dei conto correntisti, degli assicurati e dei poveri; al quarto dei poveri, dei conto correntisti e degli assicurati; al quinto dei conto correntisti, dei benestanti e dei poveri; al sesto dei conto correntisti, dei previdenti e dei ricchi; al settimo dei ricchi e dei benestanti; all'ottavo dei previdenti, dei poveri e dei conto correntisti.

Nessun gruppo appartiene in modo univoco a un segmento, ma le osservazioni fatte in precedenza sulla suddivisione dei gruppi sono rispettate.

I paesi del primo gruppo presentano tassi di penetrazione alti o buoni per molti prodotti. Nell'analisi classica, sono assegnati agli stessi segmenti[21]: benestanti, ricchi e conto correntisti. L'appartenenza a ben tre diversi *cluster*, tra i più ricchi (la probabilità di appartenere al segmento dei previdenti non è molto alta, sebbene superiore al 10%), conferma la tendenziale ricchezza di prodotti finanziari di Danimarca, Belgio e Svizzera.

Germania e Paesi Bassi, oltre a tendenze simili nelle analisi preliminari, nel modello classico sono stati assegnati entrambi al segmento degli assicurati e mostrano un andamento delle probabilità condizionate somigliante. Con l'assegnazione ai *cluster* dei conto correntisti, degli assicurati e dei benestanti si confermano, quindi, paesi con tassi di penetrazione molto alti per i conto corrente e i libretti di risparmio, alti per le assicurazioni della vita e con una ricchezza generale (in termini di prodotti finanziari) leggermente inferiore agli altri paesi dell'Europa Centrale, Austria esclusa, ma superiore ai paesi dell'Europa mediterranea.

Irlanda e Austria presentano tendenze diverse rispetto alle rispettive zone geografiche: l'Irlanda non dimostra andamenti simili ai paesi del Nord, l'Austria ai paesi dell'Europa Centrale. In particolare, nel confronto con le aree geografiche cui appartengono, sembrano avere tassi di penetrazione inferiori per molti prodotti o per prodotti diversi a quelli della tendenza generale. Nelle analisi preliminari dimostrano percentuali di possesso pressoché uguali per "bonds", "mutual_funds" e

[21] Sulla base delle probabilità di risposta condizionate per il modello classico, in tabella 4.3, la Svizzera è assegnata solo al segmento dei benestanti. Tuttavia, considerando le *probmeans*, è assegnata ai segmenti dei benestanti (0.34), dei conto correntisti (0.22) e dei ricchi (0.21). Le *probmeans*, rispetto alle probabilità condizionate, sommano a uno per riga e non per colonna. I dati forniti tra parentesi, quindi, vanno considerati come la proporzione di svizzeri appartenenti alle classi 5, 1, 3.

“ira”. Nelle probabilità condizionate del modello classico, in riferimento in particolare a “id_country”, Irlanda e Austria mostrano tendenze simili, nel senso che per il segmento degli assicurati hanno probabilità 0.10 la prima e 0.9 la seconda e basse per tutti gli altri segmenti. Solo l’Austria ha probabilità leggermente superiore (0.7 *versus* 0.3) per il segmento dei conto correntisti. Questo raggruppamento e la relativa assegnazione ai *cluster*, quindi, sembrano essere significativi.

Grecia e Polonia (gruppo 5) si sono dimostrati fin dall’inizio i paesi più poveri in termini di possesso di prodotti finanziari. In particolare hanno percentuali bassissime per “stock”, “bonds”, “mutual_funds” e “ira”. La Grecia, però, sembra possedere più conto correnti e libretti di risparmio, la Polonia più assicurazioni sulla vita. Tuttavia, sia nel modello classico, sia nel multilivello, entrambi appartengono al segmento dei poveri. Pertanto, anche questo raggruppamento e la relativa assegnazione nei *cluster* sono validi.

Italia e Spagna hanno dimostrato una maggiore ricchezza rispetto alla Grecia, l’altro paese del Sud Europa, ma inferiore rispetto ai paesi più a nord. Nelle analisi preliminari hanno evidenziato percentuali molto simili per ben quattro prodotti finanziari. Unica differenza: l’Italia ha una percentuale maggiore di titoli di stato e obbligazioni, la Spagna di pensioni integrative (la dissomiglianza, comunque, non è nettissima, ma di una decina di punti percentuali). Nel modello classico sono stati assegnati entrambi al segmento dei conto correntisti, a conferma del fatto che entrambe presentano percentuali altissime per conto corrente o libretto di risparmio. Lo stesso è avvenuto nel modello multilivello. Le differenze, quindi, non sono così forti da portare a separare i due paesi.

Più volte si è notato, nel corso delle precedenti analisi, che la Francia mostra alti tassi di penetrazione per le pensioni integrative, tanto da ipotizzarne come causa le politiche pensionistiche. Rispetto agli altri paesi anche la Repubblica Ceca ha mostrato questa tendenza. Infatti entrambe, sia nel modello classico che multilivello, appartengono al segmento dei previdenti. Tuttavia la Francia ha tassi di penetrazione maggiori per gli altri prodotti finanziari, in particolare per conto corrente

o libretto di risparmio. Pertanto la classificazione in gruppi singoli e distinti dei due paesi è assolutamente valida, date le peculiarità rispetto ad altri paesi e le differenze tra di loro.
La Svezia è, in assoluto, il paese con tassi di penetrazione maggiori rispetto agli altri paesi. Non stupisce che sia classificata in un gruppo a sé stante, anche rispetto alla Danimarca, che ha percentuali minori di possesso per “mutual_funds”, “ira” e “life_insurance”.
Tutti i segmenti sembrano essere transnazionali. In particolare, i conto correntisti sono presenti in quasi tutti i gruppi, sebbene con probabilità diverse. Il segmento con bassi di penetrazione per tutti i prodotti finanziari, esclusi conto corrente o libretto di risparmio, sembra essere paneuropeo, mentre quelli con delle peculiarità, o tassi di penetrazione più alti, coinvolgono meno paesi. Se si considerano le probabilità maggiori del 20% questa tendenza è molto evidente. Infatti, al primo segmento appartengono i primi sei gruppi, per un totale di dodici paesi, mentre ai restanti cinque segmenti appartengono due gruppi, coinvolgendo al massimo quattro paesi.

5.5 Analisi dell’efficienza e dell’efficacia dei segmenti.

Il fatto di avere suddiviso i paesi in gruppi e di aver assegnato loro dei segmenti di consumatori, aumenta l’eterogeneità fra i segmenti rispetto al modello classico. I *cluster* dimostrano tendenze nel possesso dei prodotti finanziari davvero peculiari, perciò anche l’omogeneità nei segmenti è rispettata. I profili tracciati con l’uso delle covariate sono omogenei all’interno (per costruzione) ed eterogenei fra di loro. Le basi della segmentazione, invariate rispetto al modello classico, sono pertinenti all’obiettivo di segmentare il mercato europeo dei prodotti finanziari, con riferimento agli ultracinquantenni. I segmenti, quindi, sono efficaci.
L’ampiezza dei segmenti, anche di quelli meno estesi, e la transnazionalità ne garantiscono la consistenza e, quindi verosimilmente anche la profittabilità. Segmenti nazionali appartenenti a

un paese poco numeroso avrebbero messo in dubbio la consistenza. L'uso di covariate per tracciare i profili e la possibilità di assegnarli a gruppi di paesi facilitano l'identificabilità dei segmenti, ma anche ne migliorano l'accessibilità da parte delle aziende. Per esempio, se in uno dei segmenti se ne dimostrasse alto il tasso d'uso (verosimile per i segmenti più giovani), internet potrebbe essere usato per aggredire i segmenti. Le basi della segmentazione sono misurabili e non particolarmente volatili, caratteristiche che i segmenti mantengono. Tuttavia, come regola generale, è sempre bene monitorare la stabilità dei segmenti con dati primari o secondari. La capacità di risposta e la propositività non sono facilmente misurabili a priori, tuttavia un'azienda prima di posizionarsi e aggredire un segmento fa delle analisi ulteriori, che potrebbero chiarire eventuali dubbi. Anche l'efficacia dei segmenti individuati, quindi, sembra garantita.

Conclusioni.

Lo scopo di questa tesi è segmentare il mercato europeo di prodotti finanziari, con riferimento alla popolazione degli ultra sessantenni europei.

Nel primo capitolo sono presentate la segmentazione e alcuni fra i più famosi modelli e tecniche statistiche per implementarla.

La segmentazione consiste nel suddividere un mercato in gruppi di consumatori, detti segmenti, omogenei al loro interno ed eterogenei fra loro, ed è alla base del *target marketing*. L'analisi di segmentazione può avvenire secondo diversi criteri, in linea con gli obiettivi preposti. Può essere condotta su base geografica, demografica, psicografica, comportamentale o sulla base dei benefici attesi da un dato prodotto. Nel nostro caso la segmentazione è condotta su base comportamentale (possesso o non possesso di prodotti finanziari), geografica e demografica. Inoltre il mercato è segmentato su più livelli: segmenti diversi per tendenze di consumo di prodotti finanziari sono descritti anche su base geografica e demografica.

Le tecniche statistiche di segmentazione si suddividono in a priori e a posteriori, a seconda se basi di segmentazione e numero e tipologia dei gruppi siano o meno determinati con l'analisi. Sono a priori AID e CHAID, a posteriori *cluster analysis* e *conjoint analysis*.

I dati utilizzati nel nostro studio provengono dall'indagine SHARE relativa al 2006, che indaga salute, invecchiamento e pensioni degli ultracinquantenni europei, estraendo campioni rappresentativi di ogni paese partecipante (quattordici nel nostro caso). Da questa bancadati multidisciplinare sono state estratte delle variabili che riguardano il possesso di alcuni prodotti finanziari, ovvero conto corrente o libretto di risparmio, titoli di stato o obbligazioni, azioni o partecipazioni, fondi comuni di investimento o gestioni patrimoniali, pensioni integrative e assicurazioni sulla vita. Sono state desunte, inoltre, delle variabili riportanti informazioni sul capofamiglia intervistato o sulla famiglia, utili per descrivere i segmenti individuati. Esse riguardano il sesso, l'età,

l'avere o meno un compagno, lo stato civile, lo stato occupazionale, la proprietà o meno dell'abitazione, il livello di istruzione, le capacità cognitive, il numero di componenti del nucleo familiare e il paese di appartenenza.

Dalle prime analisi esplorative è emerso che i prodotti posseduti con maggiore frequenza sono conto corrente o libretto di risparmio, seguiti con percentuali nettamente inferiori da assicurazione sulla vita, pensione integrativa, azioni o partecipazioni, obbligazioni o partecipazioni aziendali e titoli di stato o obbligazioni.

Dalle frequenze marginali per paese delle variabili finanziarie, si è evinto che in tutti i paesi, esclusi Grecia, Polonia e Repubblica Ceca, più del 79% degli intervistati possiede conto corrente o libretto di risparmio. Per titoli di stato e obbligazioni le percentuali maggiori di possesso si riscontrano in Svezia, Danimarca e Svizzera, ma anche Germania, Italia e Belgio. Per azioni e partecipazioni le percentuali maggiori si concentrano nei paesi centro-settentrionali, ad eccezione di Irlanda e Austria. Si distinguono Svezia e Danimarca, Svizzera e Belgio. La frequenza di possesso di fondi comuni di investimento e partecipazioni è alta solo per la Svezia, ma anche la Svizzera riporta un tasso di penetrazione che si distingue da quello degli altri paesi. In generale, nella zona centro-settentrionale le percentuali di possesso per "mutual_funds" sono più alte che per la zona mediterranea e dell'Est, ma non di molto. A macchia di leopardo qualche paese si differenzia dagli altri. Per le pensioni integrative le percentuali maggiori si riscontrano in Svezia, Danimarca, Francia, Repubblica Ceca, ma anche in Svizzera e Belgio. Infine per le assicurazioni sulla vita la percentuale maggiore è in Svezia, ma spiccano anche quelle di Irlanda e Polonia, Germania, Paesi Bassi e Svezia.

In sintesi, la Svezia è il paese più ricco di prodotti finanziari, seguito da Danimarca e Svizzera. L'Europa Centrale mostra percentuali di possesso dei prodotti finanziari superiori all'Europa mediterranea, dell'Est e dell'Austria. Tuttavia, oltre all'alto tasso di possesso dei conto correnti o libretti di risparmio, non si nota una tendenza univoca nel

possedere tutti i prodotti finanziari o nel possederne alcuni: Germania e Paesi Bassi hanno frequenze maggiori per le assicurazioni sulla vita, la Francia per le pensioni integrative, Belgio e Svizzera per molti prodotti. Pure l'Europa mediterranea mostra tendenze diverse. La Grecia ha frequenze basse per tutti i prodotti finanziari, mentre la Spagna e l'Italia hanno frequenze alte per conto correnti o libretti di risparmio. Inoltre la Spagna, rispetto all'Italia, ha percentuali maggiori per le pensioni integrative e minori per titoli di stato e obbligazioni. Per il resto, però, sono simili. I paesi dell'Est, Polonia e Repubblica Ceca, mostrano propensioni diverse: la Repubblica Ceca ha frequenze maggiori per i conto correnti o libretti di risparmio e pensioni integrative, minori per le assicurazioni sulla vita. In generale, però, non mostrano frequenze di possesso molto alte.

Attraverso l'analisi delle frequenze delle variabili non finanziarie, si è tracciato un profilo del capofamiglia tipo. Si tratta di una donna di sessantaquattro anni, pensionata. Ha un compagno, che è il marito o il convivente legalmente riconosciuto. Vive sola con il *partner* in una casa di proprietà. Ha un livello di istruzione secondario superiore e ricorda diciotto nomi di animali.

Per suddividere in segmenti il campione, si sono utilizzati i *latent class model*, classico e multilivello. L'ipotesi di base dei modelli è che le caratteristiche delle variabili utilizzate per l'analisi possano essere riassunte da loro ulteriori caratteristiche latenti.

Il modello tradizionale utilizza una variabile latente discreta, le cui modalità definiscono delle classi latenti, a cui sono assegnate in modo probabilistico e univoco le unità statistiche. Presuppone indipendenza locale tra le variabili. È possibile tracciare un profilo dei segmenti individuati, utilizzando le probabilità di risposta condizionate relative sia agli indicatori (variabili usate come basi della segmentazione), sia alle covariate.

Il modello multilivello è utile nel caso in cui le osservazioni siano dipendenti e presentino una struttura gerarchica, per esempio a due livelli. Esso raggruppa le unità sia di primo che di secondo livello, con

approcci diversi a seconda che si tratti di un modello *multilevel* a effetti fissi o casuali, e ancora parametrico o non parametrico. Nel nostro lavoro si è utilizzato un modello a effetti casuali non parametrico, che utilizza due variabili latenti discrete. La prima individua delle classi latenti a livello uno, la seconda, con procedimento simile, clusterizza le unità di secondo livello. In modo probabilistico sono assegnate le unità e i gruppi alle classi di primo livello. In particolare, in questo lavoro, si sono suddivise le unità sulla base del possesso di prodotti finanziari in segmenti (livello uno) e i paesi di appartenenza in gruppi (livello due). Si è poi verificato quali segmenti fossero presenti nei vari gruppi.

Il tipo di segmentazione attuabile con i *latent class model* è a posteriori, nel senso che l'analisi determina il numero e la tipologia dei segmenti individuabili in un mercato, e flessibile, nel senso che l'analisi definisce la ripartizione che garantisce massima omogeneità interna e minima omogeneità esterna. Per costruzione, quindi, i segmenti individuati sono efficaci dal punto di vista del *marketing*.

Nel quarto capitolo si sono presentati i risultati relativi all'analisi latente classica. Si sono stimati modelli con diversi numeri di classi latenti e, fra tutti, si è scelto il modello a sei classi, perché con migliore adattamento. Si sono individuati, quindi, sei segmenti nel mercato dei prodotti finanziari europei degli ultrasessantenni e sulla base delle probabilità condizionate di risposta si sono tracciati i loro profili. Essi sono i segmenti dei conto correntisti, dei poveri, dei ricchi, degli assicurati, dei benestanti e dei previdenti, in ordine decrescente di ampiezza. Il nome assegnato ai segmenti rispecchia il loro profilo di possesso dei prodotti finanziari. Alcuni segmenti dimostrano una ricchezza di prodotti generalmente diffusa, altri una spiccata tendenza a possedere particolari prodotti finanziari, uno una povertà generale. I ricchi hanno tassi di penetrazione molto alti, alti o buoni per tutti i prodotti finanziari, mentre i benestanti tassi alti solo per conto corrente e libretto di risparmio e buoni o discreti per tutti gli altri prodotti (pensioni integrative escluse). I conto correntisti hanno una tasso di penetrazione alto solo per conto correnti o libretti di risparmio. Gli assicurati e i previdenti

hanno entrambi un alto tasso per conto corrente o libretto di risparmio, ma anche per assicurazione sulla vita i primi e pensione integrativa i secondi; bassi per tutti gli altri prodotti. Infine i poveri hanno tassi di penetrazione molto bassi per tutti i prodotti finanziari.

Nell'analisi sono state introdotte anche le covariate, utili per rifinire il profilo dei segmenti con informazioni che li rendano identificabili. Al di là della descrizione di ogni singolo segmento, riportata nel capitolo quarto, è interessante notare che ai segmenti più ricchi appartengono per lo più i paesi scandinavi con Svizzera e Belgio e al più povero i paesi dell'Est e Grecia. Ai segmenti con tassi di penetrazioni per particolari prodotti, invece, si assegnano tendenzialmente i paesi centrali dell'Europa con Italia, Spagna e Francia. Il centro-sud europeo, però, non presenta tendenze omogenee: Italia e Spagna possiedono soprattutto conto corrente o libretto postale, Germania e Paesi Bassi conto corrente e assicurazioni sulla vita, la Francia conto corrente e pensione integrativa. La Repubblica Ceca, infine, mostra rispetto agli altri paesi "poveri", un tasso di penetrazione per le assicurazioni sulla vita elevato. Queste osservazioni trovano conferma nel modello multilivello stimato.

Si riscontrano altre tendenze: i segmenti con tassi di penetrazione da discreti ad alti (benestanti e ricchi) hanno una maggior probabilità di comprendere maschi, sposati e proprietari dell'abitazione. Sono formati da individui con capacità cognitive maggiori della media europea e con i più elevati livelli di istruzione. Pertanto la probabilità di possedere la propria abitazione, un alto livello di istruzione e capacità cognitive brillanti tende a crescere coll'aumentare della ricchezza di prodotti finanziari.

I segmenti individuati si sono dimostrati da un punto di vista di *marketing* efficienti ed efficaci, sebbene andrebbero condotte ulteriori analisi sulla stabilità (ma questa è una regola generale della segmentazione), sulla propositività e sulla capacità di risposta.

I dati a disposizione presentano una struttura gerarchica su due livelli: gli individui sono raggruppabili nei paesi. Inoltre, alcune tendenze riscontrate nel possesso di prodotti finanziari tendono ad essere legate

al paese d'appartenenza delle unità assegnate loro. Sembra, quindi, ragionevole stimare un modello multilivello, che tenga conto di queste strutture. Anche in questo caso si sono stimati numerosi modelli, con diverse combinazioni di numero di *cluster* (classi latenti a livello individuo) e di gruppi (classi latenti a livello paese). Si è scelto il modello con sei *cluster* e otto gruppi, perché presenta il miglior adattamento in termini di BIC. Il profilo dei segmenti per il modello multilivello, costruito sulla base delle probabilità condizionate di risposta per gli indicatori, è identico a quello per il modello classico, cosicché i nomi assegnati sono stati mantenuti. Allo stesso modo la descrizione dei segmenti non varia tra i modelli. Certo le probabilità condizionate cambiano, ma in nessun caso si stravolgono le osservazioni fatte per il modello tradizionale. Pertanto il profilo dei segmenti è robusto rispetto al modello utilizzato.

Al secondo livello della gerarchia, sono stati raggruppati i paesi, sulla base delle similarità delle loro strutture interne, rispetto ai segmenti di individui. I gruppi individuati sono Danimarca, Belgio e Svizzera (gruppo 1); Germania e Paesi Bassi (2); Irlanda e Austria (3); Grecia e Polonia (4); Italia e Spagna (5); Francia (6); Svezia (7); Repubblica Ceca (8). Sulla base delle probabilità di appartenenza a posteriori dei gruppi ai *cluster*, si è individuato quali segmenti sono assegnabili ad ogni gruppo. Il segmento dei poveri è assegnato ai gruppi 4, 8, 3 e 5. Il segmento dei conto correnti a tutti i gruppi tranne alla Svezia (7). Il segmento degli assicurati a Irlanda e Austria, Germania e Paesi Bassi, Grecia e Polonia. Il segmento dei previdenti a Francia, Repubblica Ceca, Danimarca, Belgio e Svizzera. Il segmento dei benestanti a Danimarca, Belgio e Svizzera, Svezia, Italia e Spagna, Germania e Paesi Bassi. Il segmento dei ricchi a Svezia, Danimarca, Belgio e Svizzera, Francia. Le osservazioni fatte in precedenza, commentando quali paesi il modello tradizionale assegnava ai vari segmenti, valgono anche per il modello multilivello. Si conferma, quindi, che la Svezia è il paese più ricco di prodotti finanziari.

Francia e Repubblica Ceca hanno entrambe alti tassi di penetrazione per le pensioni integrative, ma tassi di penetrazione diversi per tutti gli altri prodotti, tanto da essere assegnate a due gruppi distinti.
Grecia e Polonia hanno bassi tassi di penetrazione per tutti i prodotti finanziari e per questo si distinguono da tutti gli altri paesi europei.
Belgio, Svizzera e Danimarca si confermano tendenzialmente più ricchi rispetto agli altri paesi europei, ma non tanto da essere assimilabili alla Svezia. Infatti formano un gruppo a sé stante.
L'Europa centro-meridionale conferma delle tendenze dissimili fra paesi e il possesso di tassi di penetrazione elevati solo per particolari prodotti, come le assicurazioni sulla vita per Germania e Paesi Bassi, Irlanda e Austria. Questi ultimi però, hanno tassi di penetrazione per gli altri prodotti più bassi rispetto ai paesi del gruppo 2, tanto da formare un gruppo a sé stante.
Spagna e Italia confermano la tendenza ad avere alti tassi di penetrazione solo per conto correnti e libretto di risparmio.
Il modello classico è un ottimo strumento per la segmentazione, ma si comporta meglio con dati a struttura semplice. Il suo impiego dovrebbe riguardare mercati nazionali o limitati, come per esempio un singolo paese o una regione, oppure gli utenti di un servizio pubblico locale.
Il modello multilivello si dimostra uno strumento molto utile per la segmentazione internazionale: rispetto al modello classico i risultati sono interpretabili in modo più immediato, quando i dati hanno una struttura gerarchica. Potrebbe essere utilizzato, quindi, per la segmentazione di mercati sovraregionali, ma anche di negozi di un *franchising* o che vendono prodotti diversi, della clientela di un'azienda, suddivisa in regioni. Poiché i dati SHARE sono longitudinali, sarebbe molto interessante valutare gli effetti della crisi economica sulla segmentazione del mercato dei prodotti finanziari. A tal scopo si può utilizzare il modello a classi latenti multilivello, aggiungendo un livello, il tempo, alla gerarchia già presente nei dati.
Per garantire l'identificabilità dei segmenti è necessario l'uso di covariate, che sembrano appesantire i modelli. In futuro, allora, si

potrebbe chiarire il ruolo delle covariate da un punto di vista statistico, in modo da garantire che la stima dei modelli con indicatori e covariate non presenti problemi di adattamento.

Appendice A.

Metodi, indicatori, misure usate nella segmentazione a priori e a posteriori.

Siano n la numerosità del campione e n_j la numerosità del *j-simo* sottocampione, Y la variabile criterio e X_1, ..., X_p le variabili esplicative (o predittori), Y_{ij} il valore assunto dall'individuo i appartenente al sottocampione j e $\overline{Y}_j$ il valore medio della variabile criterio nel gruppo j.

AID e CHAID.

In AID il criterio di ottimalità, usato per scegliere ad ogni passo la bipartizione migliore, prevede di confrontare la devianza tra gruppi $n_1(\overline{Y}_1 - \overline{Y}_2)^2 + n_2(\overline{Y}_2 - \overline{Y})^2$

con la devianza entro i gruppi $\sum_{i=1}^{n_1}(Y_{i1} - \overline{Y}_1)^2 + \sum_{i=1}^{n_2}(Y_{i2} - \overline{Y}_2)^2$,

a parità di devianza totale $\sum_{i=1}^{n_1}(Y_{i1} - \overline{Y})^2 + \sum_{i=1}^{n_2}(Y_{i2} - \overline{Y})^2$.

In CHAID il criterio di ottimalità prevede di scegliere tra gli s sotttoinsiemi disgiunti quello con il valore maggiore del test X quadrato. Di fatto ad ogni passo si costruisce una tabella di contingenza con le modalità della variabile criterio e di una delle esplicative. Quindi si testa l'ipotesi nulla di indipendenza dei caratteri, confrontando le frequenze di ogni cella, f_{ij}, della tabella con quelle teoriche, F_{ij}, ottenute sotto l'ipotesi nulla.

Se il valore del test $X^2 = \sum_{ij} \frac{(f_{ij} - F_{ij})^2}{F_{ij}}$ è alto, si rifiuta H_0, ovvero si accetta l'ipotesi di dipendenza tra le variabili. Il valore del test è proporzionale alla forza della dipendenza tra le variabili, quindi se alto garantisce una maggiore omogeneità entro il segmento, perché a specifiche modalità di Y corrispondono specifiche modalità di X.

Asintoticamente $X^2 \sim X^2_{(I-1)(J-1)}$, dove *I* e *J* sono rispettivamente le categorie della variabile criterio e della esplicativa.

Cluster analysis.

Nella CA, le informazioni di partenza sono raccolte in una matrice $n \times p$

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & x_{ik} & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

dove ogni riga *i=1,...,n* rappresenta il profilo di un'unità statistica e ogni colonna *k=1,...,p* le determinazioni delle variabili osservate.

	individuo *j*		
		1	0
individuo *i*	1	a	b
	0	c	d

Purché le variabili siano misurate sulla stessa scala o siano standardizzate (questo garantisce la confrontabilità), le dissomiglianze tra i gruppi si possono misurare con i coefficienti di associazione (variabile espressa su scala nominale binaria) o le distanze (variabile su scala intervallare o rapporto).
I coefficienti di associazioni si calcolano a partire da una tabella che sintetizza i dati sugli individui, riportando il numero dei caratteri simultaneamente presenti (a) o non presenti (d).

Dalle frequenze a, b, c, d si possono calcolare:

- Coefficiente di Jaccard $J_{ij} = \frac{a}{a+b+c}$
- Coefficiente di Dice $D_{ij} = \frac{2a}{2a+b+c}$
- Coefficiente semplice di somiglianza $s_{ij} = \frac{a+d}{a+b+c+d}$, che rapporta le frequenze di accordo e quelle totali.
- Indice di dissomiglianza $d_{ij} = 1 - s_{ij}$
- Coefficiente di Gower $G_{ij} = \frac{\sum_{k=1}^{p} w_k s_{kij}}{\sum_{k=1}^{p} w_k}$

  dove s_{kij} vale 1 se la variabile è nominale o ordinale e vi è conocomitanza di presenza o assenza per *i* e *j*; 0 se la variabile è nominale o ordinale e non c'è concomitanza di assenza e presenza per *i* e *j*; $1 - \frac{|x_{ik} - x_{jk}|}{R_k}$ se la variabile è quantitativa e R_k è il suo campo di variazione, mentre w_k è un sistema di pesi.

Tranne l'indice di dissomiglianza, sono misure di somiglianza. Esse variano tra zero e uno.

Le distanze, invece, sono:

- Distanza di Minkowsky $^{r}d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^{r} \right)^{\frac{1}{r}}$ per r=1,2, ...
- Distanza di Mahalanobis $d_{ij} = \sqrt{\sum_{k=1}^{p} \sum_{h=1}^{p} s^{hk} (x_{ik} - x_{jk})(x_{ih} - x_{jh})}$ con s^{hk} generico elemento della matrice inversa delle varianze e covarianze tra le *p* variabili. Questo indicatore è utile nel caso in cui le variabili siano correlate, perché depura dalla correlazione.

Calcolate le dissomiglianze o le distanze, si costruisce una matrice *D* che le contenga, utile nella fase di aggregazione.

$$D = \begin{bmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \dots & \dots & \dots & \dots \\ d_{n1} & d_{n2} & \dots & 0 \end{bmatrix}$$

Nell'algoritmo gerarchico si utilizzano i seguenti criteri di valutazione delle distanze tra i gruppi:

- Metodo del legame singolo: la distanza tra i gruppi è pari alla più piccola delle distanze istituibili a due a due tra tutti gli elementi dei gruppi
- Metodo del legame completo: la distanza tra i gruppi è pari alla maggiore delle distanze istituibili a due a due tra tutti gli elementi dei gruppi
- Metodo del legame medio: la distanza tra i gruppi è pari alla media aritmetica delle distanze istituibili a due a due tra tutti gli elementi dei gruppi
- Metodo del centroide: la distanza tra i gruppi è pari alla distanza tra i rispettivi centroidi, ovvero i vettori con i valori medi delle p variabili, calcolati gruppo per gruppo
- Metodo di Ward: ad ogni passo si riuniscono i due gruppi la cui fusione provoca il minimo incremento della devianza entro

A questo punto è utile costruire le matrici *T* della devianza totale, *W* della devianza entro i gruppi e *B* della devianza tra i gruppi. Siano *G* il numero di gruppi (*g=1, ..., G* indicizza i gruppi) e n_g la numerosità del gruppo *g*, x_{igk} il valore dell'esplorativa *k* per l'unità *i* nel gruppo *g*, $\overline{\overline{x_k}} = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n} x_{igk}$ il valor medio della variabile *k* nell'intero collettivo e $\overline{x}_{gk} = \frac{1}{n_g}\sum_{i=1}^{n_g} x_{igk}$ il valor medio della variabile *k* nel gruppo *g*, allora si definiscono

$$T = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\sum_{k=1}^{p}(x_{igk} - \overline{x_k})^2 , \qquad W = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\sum_{k=1}^{p}(x_{igk} - \overline{x_{kg}})^2 \quad \text{e}$$

$B = \sum_{g=1}^{G}\sum_{k=1}^{p} n_g(\bar{x}_{kg} - \bar{x}_k)^2$, con $T=W+B$.

Un metodo empirico per scegliere il numero ottimale dei gruppi è la rappresentazione grafica con in ordinata il numero dei gruppi e in ascissa i valori di dissomiglianza corrispondenti. Il numero ottimale di *cluster* è quello in corrispondenza del quale la spezzata si appiattisce notevolmente. In tal modo non si raggruppano gruppi troppo dissimili. In alternativa si possono utilizzare le differenze delle misure di dissomiglianza tra due gruppi che si fondono: $\delta_g = {}_{g-1}d - {}_g d$. Il numero di gruppi ottimale è quello per cui δ_g è massimo. Un ultimo metodo è lo Pseudo F $F = \dfrac{\dfrac{tr(b)}{g-1}}{\dfrac{tr(W)}{n-g}}$, che va calcolato per tutti i gruppi. Si sceglie *g* per cui *F* è massimo.

Per valutare la validità dei risultati si utilizza il test di Arnold, che testa la significatività di una partizione. L'ipotesi nulla è che i dati provengano da popolazioni unimodali o uniformi. La statistica test è data da $C = \log\left(\dfrac{\det(T)}{\det(W)}\right)$ e tutti i possibili valori sono stati tabulati. Il coefficiente di correlazione cofenetico, invece, indica la qualità della rappresentazione con il dendrogramma della procedura gerarchica. Esso è pari a $R_c = \dfrac{\sum_{i=1}^{n}\sum_{j>i}(d_{ij} - \overline{d})(d_{ij}^* - \overline{d}^*)}{\sum_{i=1}^{n}\sum_{j>i}(d_{ij} - \overline{d})^2 \sum_{i=1}^{n}\sum_{j>i}(d_{ij}^* - \overline{d}^*)^2}$ con d_{ij}^* distanze cofenetiche, indicate nell'albero, e d_{ij} dissomiglianze originarie. Il valore di R_c è inversamente proporzionale alla distorsione e di solito ha come campo di variazione $[0.60, 0.95]$.

Conjoint analysis.

Siano f_{ik} il livello di intensità dell'attributo k ($k=1,...,$ p) nello stimolo j ($j=1, ..., J$), y_j la preferenza o utilità espressa per lo stimolo j, w_k un peso di importanza assegnato al profilo k, i_k il livello dell'attributo k nel profilo ideale del bene, s una funzione discontinua, u_{ij} l'utilità assegnata allo stimolo j dal consumatore i e V_k l'utilità parziale dell'attributo k.

La *conjoint analysis* mette in corrispondenza biunivoca preferenza accordata ad un profilo e utilità, nel senso che tanto più un profilo è gradito, tanto più la sua fruizione fornisce utilità. I modelli per descrivere questa utilità sono:

- Modello vettore: $y_i = \sum_{k=1}^{p} w_k f_{jk}$
- Modello punto-ideale: $d_j^2 = \sum_{k=1}^{p} w_k (f_{jk} - i_k)^2$
- Modello *parth-worth*: $y_i = \sum_{k=1}^{p} s_k f_{jk}$

Sono modelli additivi, in cui l'utilità complessiva associata ad un profilo è data dalla somma delle utilità parziali dei singoli livelli degli attributi. In particolare, per il modello vettore, la preferenza è data dalla somma ponderata degli attributi per dei pesi che possono essere specifici per consumatore, in modo da tener conto della struttura delle preferenze. Il modello punto-ideale, invece, considera la distanza del profilo reale considerato dal profilo ideale, che si costruisce sommando i livelli i_k di ogni attributo. Intuitivamente, l'utilità è indirettamente proporzionale alla distanza dall'ideale. Il modello *parth-worth*, infine, è il più generale e ricorre all'uso di una funzione discontinua s_k, definita per un insieme opportunamente selezionato di livelli di attributi.

L'importanza relativa di ogni attributo si calcola come rapporto tra la differenza del valore massimo e del valore minimo dell'utilità parziale dell'attributo k per la somma delle differenze per tutti gli attributi:

$$IRA_k = \frac{\max V_k - \min V_k}{\sum_{k=1}^{p} (\max V_k - \min V_k)}.$$

Nella simulazione del comportamento di scelta di un consumatore, che costituisce la seconda fase della segmentazione flessibile, si utilizzano dei criteri per prevedere quale scelta farà il consumatore, ovvero le sue preferenze. Questo si può fare anche considerando alternative non valutate direttamente.

Il criterio *first choice*, di tipo deterministico, si basa sul presupposto che il consumatore scelga l'alternativa cui è associato il valore maggiore di utilità totale. La quota di preferenza del prodotto *j*, QP_j,è data dal rapporto tra numero di quanti scelgono il profilo j sul totale. Questo criterio è poco realistico quando due alternative hanno valore elevato o simile; inoltre sulla scelta intervengono fattori, come ad esempio promozioni e informazioni possedute dal consumatore, di cui non si tiene conto. Per ovviare a questi svantaggi, si possono utilizzare i prossimi indicatori. Nel criterio Bratford, Terry, Luce la quota di preferenza del prodotto *j* è data dalla media aritmetica, estesa a tutti gli intervistati nel campione, della probabilità di scelta p_{ij} associata a ciascuna alternativa:

$$QP_j = \frac{1}{n}\sum_{i=1}^{n} p_{ij} \text{ con } p_{ij} = \frac{u_{ij}}{\sum_{j=1}^{J} u_{ij}}.$$

Nel criterio logit, rispetto al Bratford, Terry, Luce, varia solo la forma della probabilità di scelta delle alternative: $p_{ij} = \frac{\exp(u_{ij})}{\sum_{j=1}^{J}\exp(u_{ij})}$.

Infine, per valutare l'adattamento dei risultati della *conjoint analysis*, si possono usare la correlazione tra le scelte previste $\overline{y_j}$ e osservate y_j *R* di Pearson:

$$R = \frac{\frac{1}{nJ}\sum_{j=1}^{nJ} \overline{\overline{y}}_j y_j - \overline{y}_j^{\,2}}{\sqrt{\left[\frac{1}{nJ}\sum_{j=1}^{nJ}(\overline{\overline{y}}_j - \overline{y})^2\right]\left[\frac{1}{nJ}\sum_{j=1}^{nJ}(y_j - \overline{y}_j)^2\right]}}$$

e il τ di Kendall:

$\tau = \frac{\sum_{j=1}^{nJ-1}\sum_{h=1}^{nJ} segno\left\{\left(r(\overline{\overline{y}}_j) - r(\overline{\overline{y}}_h)\right)\left(r(y_j) - r(y_h)\right)\right\}}{nJ(nJ-1)}$. Entrambi gli indici variano nell'intervallo $[0,1]$ e indicano una forte relazione tra scelte previste e osservate se assumono valori prossimi a uno e una relazione debole per valori vicini a zero.

Bibliografia.

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

BIJMOLT T. H. A., PAAS L. J., VERMUNT J. K. (2004), “Country and consumer segmentation: multi-level latent class analysis of financial product ownership”, International Journal of Research Marketing, 21, pp. 323-340

BÖRSCH-SUPAN et al. (2008), *Health, Ageing and Retirement in Europe (2004-2007). Starting the longitudinal dimension*, Mannheim, MEA

BÖRSCH-SUPAN A., JÜRGES H. (eds.) (2005), *The Survey of Health, Ageing and Retirement in Europe. Methodology*, Mannheim, MEA

BRASINI S., TASSINARI F., TASSINARI G., (1996), *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

DEL GIOVINE C., (2008), Modello multilevel a classi latenti: estensione al modello multidimensionale, Tesi di dottorato, Università degli Studi di Bologna

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

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 e unidentifiable models", Biometrika, 61, 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.

IACUS S. M., MASAROTTO G., (2003), *Laboratorio di statistica con R*, McGraw-Hill, Milano.

KOTLER P., (2004), *Marketing Management*, Pearson Education Italia, Milano.

LAZARSFELD P. F., HENRY N. W., (1968), "Latent structure analysis", Boston, Hughton Muffin

MADIGSON J., VERMUNT J. K. (2005), “Hierarchical Mixture Model for nested data structures”, in WEIHS G., GAUL W. (eds), *Classification: The Ubiquitous Challenge*, pp. 176-183, Springer, Heidelberg

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

MADIGSON J., VERMUNT J. K. (2004), “Latent Class Models”, in KAPLAND D., *The sage Handbook of Quantitative Methodology for Social Science*, Chapter 10, pp. 175-198, Thousand Oaks.

PRANDELLI E., VERONA G., (2006), *Marketing in rete. Oltre internet verso il nuovo Marketing*, McGraw-Hill, Milano

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

VERMUNT J. K., (2007),”Latent class and finite mixture models for multilevel data sets”, Statistical Methods in Medical Research, pp.1-

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

VERMUNT J. K., VAN DIJK L., (2001), “A non parametric random coefficients approach: the latent class regression model.”, Multilevel Modelling Newsletter, 13, pp. 6-13

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. (2003), ”Latent class models for classification”, Computational Statistics & Data Analysis, 41, pp. 531-537

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

ZANI S. (2000), *Analisi dei dati statistici II. Osservazioni multidimensionali*, Giuffrè Editore, Milano.

Ringraziamenti.

Alla conclusione del mio percorso di studi, vorrei ringraziare alcune persone che mi sono state vicine e mi hanno sostenuto.

In primis, un infinito grazie ai miei genitori: non avete mai smesso di credere in me e sostenermi, soprattutto nelle grosse difficoltà. Devo a voi il raggiungimento di questo traguardo e se sono la persona che sono.

Grazie alla Cri, perché sei la mia consigliera, e a Giovi, per la tua genuinità e il tuo sostegno.

Grazie alla nonna e a Carlo, per avermi spinta a fare sempre di più

Grazie a Paolo, la mia medicina e molto molto di più. Grazie perchè sei il mio risolvitore di problemi! Grazie perché ci sei!

Grazie alla Lemma e a Paolo, perché ogni volta, con voi, è come se non fosse mai passato un secondo dall'ultimo incontro e perché mi accettate sempre, nonostante le mie mancanze!

Grazie alla Chiara e alla Cristina, mie compagne di viaggio nell'avventura della specialistica, per tutte le chiacchiere in giardino e il sostegno nello studio!

Grazie alla prof. Bassi, che mi ha dato l'opportunità di approfondire un argomento di grande interesse e di conoscere cose nuove. Grazie anche per la disponibilità.

Grazie al prof. Paccagnella per la grande disponibilità e la gentilezza.

Infine, grazie a Dio, a Maria e a Don Bosco!