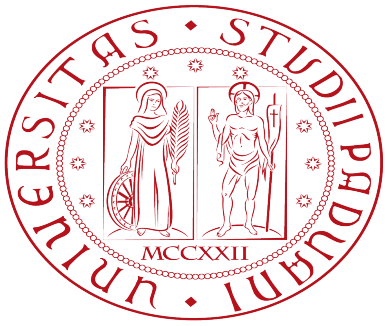
Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche



Sviluppi nelle ricerche sulla fecondabilità: un modello a Classi Latenti Multilivello.

Relatore: Prof.ssa Francesca Bassi
Dipartimento di Scienze Statistiche

Laureanda: Chiara Giangrande
Matricola N. 1013678

Anno Accademico 2013/2014

INTRODUZIONE

Questa tesi nasce per onorare la memoria del professor Bernardo Colombo, istitutore della Facoltà di Scienze Statistiche a Padova. Egli ha svolto molti progetti di ricerca, ma il più grande, a cui ha dedicato i suoi ultimi 30 anni di vita, è quello sulla *biometria del ciclo mestruale e fecondità*. I risultati di questo progetto di ricerca, ottenuti e ancora in corso, rappresentano elementi di conoscenza di grande interesse, scientifico e sociale.

Come coordinatore, in questo lavoro ha coinvolto ricercatori del Dipartimento di Scienze Statistiche.

Dopo averci lasciato, nel 2012, il Dipartimento ha voluto preparare articoli di ricerca sull'analisi dei database di altissima qualità che lui ha raccolto: si tratta di una raccolta di database diversi che per qualità e completezza sono unici e rappresentano una risorsa fondamentale per lo studio della biometria del ciclo femminile.

In questa tesi viene rappresentata prima un'analisi a Classi Latenti Classico e poi un'analisi a Classi Latenti Multilivello svolta sul database *Billings*.

Nel primo capitolo viene descritto in cosa consiste il processo di segmentazione e viene posta particolare attenzione alla segmentazione per omogeneità, vengono descritti specificatamente i metodi statistici di segmentazione della *Cluster Analysis* e della *Conjoint Analysis*.

Nel secondo capitolo vengono descritti i *Modelli a Classi Latenti* e più specificatamente i *Modelli a Classi Latenti Multilivello*.

Nel terzo capitolo vengono descritte innanzitutto le ricerche sulla *fecondabilità* della donna che Colombo ha condotto. Viene poi descritto il dataset *Billings* e vengono effettuate le prime analisi descrittive, sui file "*Donna*" e "*Ciclo*", presi prima separatamente e poi uniti, ed infine viene fatta un'analisi descrittiva prendendo solo le informazioni sul *muco*. Obiettivi di questa tesi sono andare a vedere

come si evolve il muco nei giorni precedenti e successivi al giorno di picco e creare dei gruppi omogenei di cicli e di donne.

Nel quarto capitolo vengono rappresentate le analisi condotte che hanno portato prima alla segmentazione di cicli e poi alla segmentazione di donne, utilizzando il *Modello a Classi Latenti Classico.* Dai risultati ottenuti è emerso che l'adattamento del Modello che ha portato alla segmentazione di cicli mestruali non è del tutto soddisfacente, in quanto vi è la presenza di una struttura gerarchica a due livelli, di cui il *Modello Classico a Classi Latenti* non tiene conto.

Nel quinto capitolo viene quindi presentata una stima di *Modelli a Classi Latenti Multilivello* per la segmentazione di cicli mestruali, che considerano anche la gerarchia insita nei dati e che ha portato alla stima di un modello migliore sotto il profilo dell'adattamento.

Infine, nel sesto capitolo viene presentata una stima di *Modelli a Classi Latenti Multilivello* per la segmentazione del muco cervicale.

CAPITOLO 1.

"La segmentazione"

Nel mercato dei beni di largo consumo un'azienda deve individuare e soddisfare domande diverse. A partire dal riconoscimento dell'eterogeneità dei consumatori, le procedure di segmentazione hanno l'obiettivo di definire tipologie di consumatori che esprimono esigenze diverse, nei confronti dei quali predisporre prodotti e politiche di vendita specifiche. D'altra parte, da un punto di vista strategico, per un'azienda è altrettanto importante verificare periodicamente qual è la percezione della propria offerta da parte dei consumatori attuali e potenziali.

1.1 COS'E' LA SEGMENTAZIONE DEL MERCATO

La segmentazione è la suddivisione del mercato in gruppi omogenei e distinti dei consumatori che si presume richiedano specifici prodotti. La segmentazione di qualunque tipo di mercato, *market segmentation,* consiste nell'adeguare tanto i prodotti quanto le strategie di marketing alle differenze individuabili entro l'insieme delle esigenze manifestate dai consumatori e/o utilizzatori.

I gruppi che si formano dalla segmentazione sono detti segmenti.

L'efficace attuazione delle politiche di marketing esige che ogni segmento presenti particolari connotazioni riguardanti:

- la tendenziale uniformità delle risposte degli acquirenti potenziali rispetto agli strumenti tipici del *marketing mix;*
- la *profittabilità,* nel senso che la dimensione del segmento o il suo livello di assorbimento devono essere tali da garantire un profitto;

- l'*accessibilità,* ossia la possibilità per l'azienda di raggiungere il segmento di interesse tramite gli strumenti di marketing attualmente disponibili, con costi addizionali o perdite minime.

Alle azioni di segmentazione del mercato sono complementari quelle di posizionamento di un prodotto o di una marca, il cui obiettivo consiste nel verificare qual è la percezione dell'offerta da parte dei consumatori. La segmentazione guida la selezione del mercato obiettivo nel quale operare, ossia l'identificazione e la scelta di uno o più segmenti di mercato da servire, e rappresenta la base di partenza delle iniziative di posizionamento.

Figura 1.1 "Fasi del target marketing"

1.2 FASI DI SEGMENTAZIONE

La procedura di segmentazione è formata dalle seguenti fasi:

- definizione del problema e selezione della procedura di segmentazione;
- messa a punto del programma dell'indagine sul campo al fine di raccogliere le informazioni necessarie alla realizzazione delle operazioni di segmentazione;
- elaborazioni, interpretazione e impiego dei risultati.

Per la prima fase vi sono due soluzioni alternative a disposizione, fare ricorso a un modello di segmentazione *a priori* oppure a un modello di segmentazione *a posteriori*.

Per applicare queste procedure di analisi sono necessarie informazioni riguardanti gli aspetti demografici, economici, sociali e psicografici

dei consumatori ed inoltre informazioni riguardanti le caratteristiche legate alle situazioni specifiche di consumo.

Qualunque sia il modello di segmentazione utilizzato, prima di tutto bisogna definire le variabili, che possono avere il ruolo di *basi*, ovvero fungere come caratteri rispetto ai quali viene eseguita la segmentazione, oppure avere il ruolo di *descrittori*, se entrano in gioco solo nella fase di interpretazione del profilo dei segmenti.

Nei modelli di *segmentazione a priori* si procede alla suddivisione del collettivo in esame a seconda delle modalità presentate da una o più basi, specificate a priori. Prima le basi erano soprattutto associate alle caratteristiche socio-demografiche dei consumatori, adesso invece si privilegia il ruolo di variabili direttamente collegate alle situazioni di acquisto o del consumo del prodotto.

Per individuare i descrittori dei profili dei segmenti si ricorre a tecniche statistiche di segmentazione binaria o multipla, ovvero l'*Automatic Interaction Detection* (AID) e il *Chi-squared Automatic Interaction Detection* (CHAID).

Tra le tecniche di *segmentazione a priori* Brasini, Tassinari F. e Tassinari G. (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.

I modelli di *segmentazione a posteriori* si basano sull'applicazione di

algoritmi di raggruppamento (*clustering*). I segmenti sono determinati appunto a posteriori attraverso la classificazione delle unità statistiche a seguito dei risultati di una *Cluster Analysis*, cioè a partire dal grado di dissomiglianza rispetto ad un insieme prescelto di variabili, che esprimono generalmente i comportamenti dei consumatori, i loro bisogni, le loro attitudini o altre caratteristiche di tipo psicologico. In questo caso manca una scelta a priori, e non sono prefissati né il numero né le tipologie dei gruppi da formare.

1.3 CRITERI DI SEGMENTAZIONE

Esistono cinque tipologie di segmentazione che si differenziano per le *basi* scelte:

- la *segmentazione geografica,* che permette di dividere il mercato in aree territoriali, poiché si presume che le preferenze dei consumatori dipendono dal territorio in cui si trovano;
- la *segmentazione demografica*, che ha come basi ad esempio l'età e il sesso;
- la *segmentazione psicografica*, che permette di individuare dei segmenti con stili di vita[1] simili, integrando diverse discipline, come la psicologia, la sociologia, l'antropologia culturale e il behaviorismo[2]. Le *basi* scelte riguardano gli interessi, le opinioni, le attività e le convinzioni dei clienti;

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

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

• la *segmentazione comportamentale*, che è centrata sul comportamento del consumatore, focalizzandosi sugli obiettivi e sulle caratteristiche richieste dal consumatore al prodotto.

• la *benefit segmentation*, che raggruppa i consumatori in segmenti omogenei sulla base di benefici e vantaggi che sono richiesti al prodotto o al servizio.

La *segmentazione geografica* e la *segmentazione demografica* permettono all'azienda di conoscere in modo approfondito il profilo dei loro consumatori e mostra come poterli raggiungere. Non sono conosciuti però i desideri del cliente e le sue aspettative, e non è possibile ottenere informazioni sul loro processo decisionale.

Con la *segmentazione psicografica*, con la *segmentazione comportamentale* e con la *benefit segmentation* si ha bisogno di dati più difficili da ottenere, e anche più costosi, essendo relativi ad aspetti personali dei consumatori, ma allo stesso tempo permettono di ottenere informazioni strategicamente molto importanti.

1.4 MODELLI E TECNICHE STATISTICHE DI SEGMENTAZIONE

Tra i modelli di segmentazione abbiamo il modello di *segmentazione a priori* e il modello di *segmentazione a posteriori*, che sono già stati presentati nel *Paragrafo 1.2*.

Le tecniche di segmentazione sono invece per *omogeneità* o per *obiettivi* e *flessibili*.

Nella tecnica per *omogeneità* i consumatori vengono divisi in gruppi con elevata omogeneità interna ed eterogeneità esterna, in base alla similarità di determinate variabili (per questa tecnica viene utilizzata principalmente la *Cluster Analysis*).

Nelle tecniche *flessibili*, le unità vengono divise in base alla similarità

dei profitti in termine di preferenze per i prodotti (per questa tecnica viene utilizzata principalmente la *Conjoint Analysis*).

Le tecniche per *obiettivi* raggruppano le unità statistiche in base a una o più variabili dipendenti, definite a priori, da cui sono influenzate le variabili esplicative che descrivono le caratteristiche dei segmenti (tra le tecniche per obiettivi ritroviamo l'*AID*, il *CHAID* e la *regressione logistica*).

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

1.4.1 TECNICHE PER OMOGENEITA': LA *CLUSTER ANALYSIS*

La *Cluster Analysis* si configura come uno strumento di classificazione capace di scomporre una realtà complessa di varie osservazioni in tipologie esplicite. Ovvero, da un insieme eterogeneo di clienti si possono ottenere sottoinsiemi omogenei al loro interno.

La *Cluster Analysis* rientra tra le tecniche di tipo esplorativo e pertanto non è necessaria alcuna assunzione a priori. Si configura quindi come un procedimento puramente empirico di classificazione.

Punto di partenza di ogni applicazione di *Cluster Analysis* è la disponibilità di un collettivo statistico di n elementi, ciascuno rappresentato da p variabili. Dati di questo tipo si possono organizzare in una matrice come la seguente:

$$\begin{vmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{vmatrix}$$

Nella quale ogni riga i=1,2,...,n riproduce il profilo individuale di

un'unità statistica attraverso le p variabili considerate e ogni colonna k=1,2,...,p contiene le determinazioni assunte da una variabile osservata nelle n unità in esame.

Per applicare le tecniche di *Cluster Analysis* a problemi di segmentazione del mercato si devono prima di tutto selezionare gli elementi da sottoporre ad analisi, poi scegliere i caratteri di segmentazione e loro eventuale trasformazione, come ad esempio la standardizzazione, per tenere conto di aspetti riconducibili alla scala e unità di misura, poi bisogna selezionare un criterio per valutare la dissomiglianza esistente tra gli elementi osservati e scegliere un algoritmo di raggruppamento delle unità, dopo è necessario determinare il numero dei gruppi che si formano tra gli elementi, ovvero individuare la partizione ottimale, ed infine verificare la congruenza dei risultati ed interpretarli.

La scelta delle variabili di segmentazione è un'operazione cruciale: bisogna identificare le caratteristiche più importanti in relazione alle finalità della segmentazione e determinare i valori o le modalità per ogni unità statistica considerata.

Una volta scelte le variabili, bisogna scegliere un criterio di misura della dissomiglianza tra i consumatori rispetto all'insieme delle variabili stesse. Per scegliere la misura della dissomiglianza sono disponibili numerosi indicatori, ma quelli utilizzati più frequentemente sono i coefficienti di associazione e le misure di distanza.

I coefficienti di associazione misurano la somiglianza tra unità quando i caratteri sono espressi su scala nominale binaria. I dati rilevati per ciascuna coppia di unità statistiche possono essere disposti su una tabella di contingenza 2x2 del tipo:

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

A partire dalle frequenze *a,b,c* e *d* si procede al calcolo di indici di somiglianza.

- *Coefficiente di Jaccard* (assume valori compresi tra 0 e 1)

$_{J}s_{ij} = a / a+b+c$

- *Coefficiente di Dice* (assume valori compresi tra 0 e 1)

$_{D}s_{ij} = 2a / 2a+b+c$

Questi due coefficienti non considerano *d*, dato l'assunto che la mancanza simultanea di un carattere non dovrebbe concorrere alla determinazione della misura di somiglianza tra le unità.

- *Coefficiente semplice di somiglianza*

$_{S}s_{ij} = a+d / a+b+c+d$

- *Coefficiente di Gower*

$$_{G}s_{ij} = \sum_{k=1}^{p} w_k s_{kij} / \sum_{k=1}^{p} w_k$$

s_{kij} è un indicatore elementare di somiglianza tra le unità *i* e *j* rispetto alla variabile *k* e vale 1 se la variabile è di tipo nominale o ordinale e vi è concomitanza di presenza per *i* e *j*; vale 0 se la variabile è di tipo nominale o ordinale e non vi è concomitanza di presenza per *i* e *j;* e vale $1 - |x_{ik} - x_{jk}| / R_k$ se la variabile è quantitativa (R_k è il campo di variazione della variabile *k*); s_{kij} è sempre compreso tra 0 e 1; w_k è un sistema di pesi da applicare eventualmente alle variabili.

Le misure di distanza invece sono:

- la *distanza di Minkowski*: $^{r}d_{ij} = [\sum_{k=1}^{p} |x_{ik}-x_{jk}|^{r}]^{1/r}$

(r assume qualunque valore intero non inferiore ad 1)

- la *distanza Euclidea:* è la distanza di Minkowski quando r=2, ed è la radice quadrata della somma dei quadrati degli scarti delle determinazioni omologhe assunte rispetto alle *p* variabili dalle unità *i* e *j* in questione.

- la *distanza di Mahalanobis* che prende in considerazione le correlazioni tra le variabili operando una ponderazione inversa rispetto

alle varianze e covarianze di queste.

Dopo aver utilizzato le misure di dissomiglianza appena descritte, si può determinare una matrice simmetrica della forma:

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

dove i termini della diagonale principale sono nulli. E' sufficiente quindi considerare la matrice triangolare superiore, o inferiore.

Poiché la *Cluster Analysis* ha l'obiettivo di individuare gruppi di unità statistiche con un grado elevato di omogeneità al loro interno, è necessario scegliere un idoneo algoritmo di raggruppamento. Gli algoritmi *gerarchici aggregativi* assumono come situazione di partenza una configurazione in cui ciascuna unità costituisce un gruppo a sé stante. Se le unità osservate sono in numero pari a *n*, si parte da *n* gruppi formati da una sola unità per procedere alla aggregazione dei due gruppi meno dissimili. Si ottiene quindi una ripartizione delle *n* unità in *n-1* gruppi, di cui *n-2* composti da una sola unità ed uno composto da due. Il processo viene iterato per *n-1* volte fondendo assieme di volta in volta i due gruppi meno dissimili, finché non si riuniscono tutte le *n* unità statistiche in un solo gruppo.

Una rappresentazione delle aggregazioni delle unità statistiche è il *dendrogramma*, che riproduce in scala anche i valori della misura di dissomiglianza o distanza per i due gruppi che si fondono assieme ad ogni passo.

Con i *metodi gerarchici,* ogni partizione individuata comporta l'allocazione ottimale delle unità solamente rispetto alla partizione di ordine immediatamente precedente.

Gli *algoritmi gerarchici* sono:

- il metodo del *legame singolo*, in cui la distanza è posta pari alla più piccola delle distanze istituibili a due a due tra tutti gli elementi

dei due gruppi;

- il metodo del *legame completo*, in cui la distanza è posta pari alla maggiore distanza istituibile a due a due tra tutti gli elementi dei due gruppi;
- il metodo del *legame medio*, in cui la distanza è posta pari al valore medio di tutte le distanze istituibili a due a due tra tutti gli elementi dei due gruppi;
- il metodo del *centroide*, in cui vanno determinati i valori (centroidi) contenenti i valori medi delle *p* variabili, gruppo per gruppo, e la distanza tra i gruppi viene assunta pari alla distanza tra i relativi centroidi;
- il metodo di *Ward*, che riunisce ad ogni tappa del processo i due gruppi dalla cui fusione deriva il minimo incremento possibile della devianza entro.

Se la *matrice di dissomiglianze* è costituita utilizzando la d*istanza euclidea*, si hanno generalmente buoni risultati sia con il metodo del *legame completo*, sia con il metodo di *Ward*. Un metodo idoneo con qualsiasi tipo di dissomiglianze è quello del *legame completo*, preferibile quando i caratteri sono qualitativi. Il metodo del *legame singolo* invece genera concatenamenti e può essere utilizzato quando le unità sono in sequenza.

Gli *algoritmi non gerarchici* mirano a classificare direttamente le *n* unità osservate in un numero *G* di gruppi generando una sola partizione. Un algoritmo non gerarchico è quello di *McQueen* (o delle *k medie*), che, data una prima partizione ottenuta a priori, procede a riallocare le unità al gruppo con centroide più vicino, fino a che per nessuna unità si verifica che sia minima la distanza rispetto al centroide di un gruppo diverso da quello a cui essa appartiene. Questa procedura minimizza la devianza entro i gruppi relativamente alle *p* variabili.

La *Cluster Analysis* mira all'individuazione di gruppi quanto più

omogenei, per cui è necessario trovare una procedura per scegliere il numero ottimale dei gruppi utili ai fini dell'interpretazione dei risultati. Gli indicatori che permettono di identificare il numero di gruppi ottimale sono:

- lo *pseudo F*: $F = [tr(B) / (g-1)] / [tr(W) / (n-g)]$
- il *Cubic Clustering Criterion*: $R^2 = 1 - [tr(W) / tr(T)]$

T, W e B sono le matrici delle devianze e codevianze rispettivamente totali, entro e tra i gruppi.

1.4.2 TECNICHE PER OMOGENEITA': LA *CONJOINT ANALYSIS*

La *Conjoint Analysis* fa parte della segmentazione flessibile e la sua applicazione richiede la selezione di un campione appartenente alla popolazione dei consumatori potenziali, al quale si sottopone un elenco di versioni alternative del prodotto o servizio, dette anche *stimoli*, descritte sulla base delle modalità o intensità presentate da alcuni attributi importanti.

I possibili modelli della *Conjoint Analysis* sono il *modello vettore*, il *modello punto ideale* e il *modello part-worth*.

Il *modello part-worth* è il modello più utilizzato, il quale esprime la preferenza y_j per lo stimolo j-esimo attraverso una funzione discontinua s definita per un insieme selezionato di livelli degli attributi quantitativi o per le modalità degli attributi qualitativi:

$$y_j = \sum_{k=1}^{K} s_k(f_{jk})$$

Questo modello presuppone uno schema additivo che tiene conto solo degli *effetti principali* di ciascuna modalità e non anche delle possibili interazioni tra queste.

Per realizzare uno studio di *Conjoint Analysis* si devono innanzitutto

individuare gli attributi rilevanti del prodotto o servizio e definirne i profili, selezionare poi un campione casuale di consumatori ai quali chiedere valutazioni di preferenza di ciascun profilo, stimare i valori delle utilità parziali associate ad ogni modalità o livello degli attributi, determinare poi l'importanza relativa di ciascun attributo ed infine valutare l'utilità totale corrispondente ai profili di prodotto o servizio non compresi nel piano della rilevazione.

L'obiettivo della *Conjoint Analysis* è dunque quello di identificare gruppi di rispondenti, e quindi *segmenti di mercato*, omogenei al loro interno, che si differenziano per il diverso valore di importanza degli attributi o di utilità dei vari livelli o modalità di questi ultimi.

CAPITOLO 2: "*Modelli a Classi Latenti Multilivello*"

2.1 MODELLI A CLASSI LATENTI

I modelli a classi latenti[3] appartengono alla più ampia famiglia dei modelli a variabili latenti e 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[4], fossero applicabili a dati dicotomici portò ad un raggio di azione più ampio, ma, solo più tardi, l'utilizzo di questi modelli si diffuse, con i lavori di Goodman (1974) che formalizzò la metodologia dei modelli a classi latenti, 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 modelli a classi latenti ha conosciuto un'ulteriore diffusione grazie ai moderni *software* che ne permettono

3 Alcuni autori, tra cui Madigson e Vermunt (2003), identificano i modelli a classi latenti con i modelli mistura (finite mixture model), altri invece considerano i modelli a classi latenti come una particolare specificazione dei modelli di mistura.

4 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.

una stima agevole.

L'idea di base dei modelli a classi latenti è che ciascuna unità studiata appartiene ad una delle *T* classi, dove il numero *T* di classi e la loro ampiezza non sono note a priori (Vermunt e Magidson, 2002).

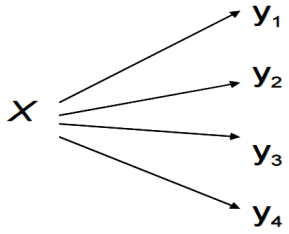
Il modello a classi latenti può essere formulato nel modo seguente:

$$P(Y_i) = \sum_{t=1}^{T} P(X_i=t)P(Y_i|X_i=t) = \sum_{t=1}^{T} P(X_i=t)P(Y_i;\vartheta_t)$$

La funzione di probabilità $P(Y_i)$, corrispondente alle risposte del soggetto *i*-esimo, è ottenuta dalle probabilità a priori $P(X_i=t)$ che la persona *i*-esima appartenga ad una certa classe *t* e dalle probabilità specifiche per ogni classe $P(Y_i;\vartheta_t)$, dove ϑ_t indicano i rispettivi parametri da stimare.

La natura della relazione tra gli indicatori e la variabile latente può essere rappresentata da un *path diagram,* dove le y variabili sono gli indicatori.

Figura 2.1 "Path diagram di un modello a classi latenti con una variabile latente e quattro indicatori"



Come si può osservare dal grafico, non vi è alcuna relazione diretta tra gli indicatori stessi. Essi sono tra di loro associati solo perché ciascuno di loro è direttamente associato alla *X*.

L'assunzione di indipendenza locale che caratterizza i modelli a classi latenti dice che le relazioni tra le variabili manifeste sono spurie (Lazarsfeld, 1955), ciò significa che nel momento in cui si introduce la variabile latente *X* nel modello, la relazione tra le variabili osservate si annulla. L'indipendenza locale implica che entro ciascuna classe

latente *t*, la probabilità di ottenere la risposta *k* in una determinata variabile sarà indipendente dalla probabilità di ottenere la risposta *k* ad una qualunque delle altre variabili osservate, condizionatamente all'appartenenza alla classe latente *t*. In altre parole, all'interno di ciascuna classe gli indicatori sono indipendenti l'uno dall'altro. Grazie a questa assunto, è possibile specificare una distribuzione univariata per ciascuna delle variabili manifeste data la classe di appartenenza, invece di una distribuzione multivariata:

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

dove ϑ_{kt} sono i parametri che definiscono la distribuzione della variabile risposta *k* nella classe latente *t*.

Quindi otteniamo:

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

$$= \sum_{t=1}^{T} P(X_i=t) \prod_{k=1}^{K} P(y_{ik};\vartheta_{kt})$$

La forma distributiva per y_{ik} dipende dalla scala delle variabili osservate incluse nel modello. Tali variabili possono essere categoriche (nominali o ordinali), continue o di conteggio.

Per le variabili continue si utilizza solitamente una distribuzione normale, per quelle di conteggio una distribuzione di Poisson o Binomiale. Quando invece le variabili sono categoriche, si utilizza una distribuzione multinomiale, in particolare, un modello logistico multinomiale per variabili di risposta nominali, un modello logistico ordinale per categorie adiacenti per variabili ordinali.

Un'importante estensione del modello a classi latenti descritta è quella di includere delle *covariate* nel modello sia per la classe latente che per le variabili risposta.

Il modello che comprende anche le *covariate* sarà un modello di regressione logistica multinomiale per X_i oppure un modello di

regressione appartenente alla famiglia dei modelli lineari generalizzati (GLM) per Y_{it}; più precisamente, per variabili di risposta ordinali si utilizzerà un modello di regressione ordinale.

L'espressione più generale per la struttura di probabilità di un modello a classi latenti comprendente anche le *covariate*, sia per la variabile latente che per quella dipendente è:

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

dove Z_i è il vettore delle *R covariate*.

2.2 MODELLI A CLASSI LATENTI MULTILIVELLO

Negli ultimi anni, l'interesse da parte dei ricercatori sociali è sempre più rivolto ad analizzare le relazioni tra gli individui e la società, poichè si è osservato che alcuni aspetti degli individui sono influenzati dal contesto socio-culturale a cui appartengono. Si può pensare che la società e gli individui facciano parte di un sistema gerarchico, in cui i primi si collocano ad un livello più alto, i secondi ad un livello successivo più basso (Goldstein, 2003).

In molti casi ci si trova ad analizzare dati che presentano una struttura gerarchica o multilivello, come ad esempio pazienti raggruppati negli ospedali, individui entro regioni, impiegati nelle aziende, bambini entro famiglie.

Anche nei modelli a classi latenti può risultare quindi interessante prendere in considerazione tale struttura nell'analisi dei dati: è possibile combinare i modelli a classi latenti con l'analisi standard multilivello. Definiamo quindi i modelli a classi latenti multilivello.

Supponiamo che il nostro sistema gerarchico sia costituito da due livelli e che il soggetto *i*-esimo appartenga ad uno dei *J* gruppi, in cui *j*

identifica un particolare gruppo con 1<=*j*<=*J* e n_j indica il numero di soggetti in ogni gruppo *j*. L'individuo *i*-esimo rappresenta l'unità di livello-1, il gruppo *j*-esimo l'unità di livello-2. Indichiamo con y_{ijk} le risposte dell'individuo *i*, entro il gruppo *j*, per l'indicatore *k*. La variabile a classi latenti sarà indicata con X_{ij}, una particolare classe latente con *t* e il numero complessivo di classi latenti con *T*. Il vettore di tutte le risposte del soggetto *i* nel gruppo *j* sarà indicata da Y_{ij}. Con *G* indicheremo la variabile osservata di secondo livello identificativa del gruppo di appartenenza.

Nelle analisi multilivello i parametri del modello differiscono tra i gruppi e la stessa cosa può essere applicata per la variante multilivello nel modello a classi latenti. In un approccio a effetti fissi la probabilità di risposta del soggetto *i*-esimo può essere definita come:

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

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

dove $P(X_{ij}=t|G=j)$ è la probabilità che il soggetto *i* del gruppo *j* appartenga alla classe latente *t*, così per ogni gruppo si avrà una probabilità di appartenenza; $P(y_{ijk}|X_{ij}=t,G=j)$ è la probabilità condizionata di y_{ijk} con ϑ_{jkt} parametri.

Quando il numero di gruppi è elevato (e il numero di soggetti all'interno di ogni gruppo è esiguo), un modello ad effetti fissi risulta piuttosto complesso, non solo per il numero di parametri da stimare che cresce rapidamente all'aumentare delle unità di secondo livello, ma anche perchè le stime risulterebbero piuttosto instabili date le ampiezze del gruppo, caratteristica tipica delle analisi multilivello.

Il problema associato al modello ad effetti fissi può essere risolto adottando un approccio ad effetti casuali. Invece di stimare tanti parametri quanti sono i gruppi osservati, si assume che gli effetti specifici del gruppo provengano da una precisa distribuzione, i cui

parametri devono tuttavia essere stimati. Questa è la stessa procedura impiegata nell'analisi multilivello semplice dove, infatti, le differenze di gruppo sono trattate attraverso effetti casuali invece che effetti fissi. Un modello a classi latenti multilivello è quindi ottenuto introducendo una variabile latente continua nel modello, cioè uno o più effetti casuali a livello del gruppo, oppure una variabile latente discreta, dove i parametri differiscono tra le classi latenti dei gruppi (Vermunt, 2006). Nel primo caso avremo un approccio ad effetti casuali di tipo parametrico, poiché la variabile latente si assume abbia una distribuzione normale, mentre nel secondo caso si tratterà di un approccio non parametrico poiché la distribuzione si assume multinomiale.

2.2.1 APPROCCIO NON PARAMETRICO

Utilizzare un approccio parametrico nei modelli a classi latenti multilivello comporta fare una forte assunzione sul modello distributivo degli effetti casuali. Meno impegnativo sarebbe invece adottare una distribuzione discreta non specificata.

In quest'ultimo caso si definiscono variabili a classi latenti anche le unità di livello-2, oltre che per quelle di livello-1: può essere più naturale classificare i gruppi, come per esempio le scuole, in un numero più piccolo di cluster, che definirli su una scala continua.

L'idea di base dell'approccio discreto è che i gruppi appartengono ad una delle T classi della variabile latente di secondo livello che indicheremo con D. Indichiamo con D_j la classe di appartenenza del gruppo j e con m una particolare classe, con $1<=D_j=m<=M$ (Vermunt, 2003).

La forma generale della probabilità di risposta Y_{ji} alle variabili osservate per un individuo i del gruppo j, condizionatamente all'appartenenza del gruppo j alla classe m, può essere così definita:

$$P(Y_{ji}|D_j=m) = P(Y_{ji};\vartheta_m)$$

$$= \sum_{t=1}^{T} P(X_{ji}=t|D_j=m) \prod_{k=1}^{K} P(y_{jik}|X_{ij}=t,D_j=m)$$

$$= \sum_{t=1}^{T} P(X_{ji}=t|D_j=m) \prod_{k=1}^{K} P(y_{jik},\vartheta_{ktm})$$

Da un lato le classi dei gruppi differiscono per le probabilità che i loro membri appartengano alla classe latente t e, dall'altro lato, per i parametri che definiscono le probabilità delle variabili risposta.

A livello del gruppo viene specificato il legame tra i casi che appartengono allo stesso gruppo:

$$P(Y_j) = \sum_{m=1}^{M} P(D_j=m) \prod_{i=1}^{nj} P(y_{ji}|D_j=m)$$

Tale funzione di probabilità espressa per l'intero vettore delle risposte date da tutti i soggetti del gruppo *j*, $P(Y_j)$, è ottenuta assumendo che le risposte degli n_j individui siano indipendenti tra di loro, condizionatamente all'appartenenza del gruppo *j* alla classe *m*, e dalla successiva marginalizzazione delle classi latenti per i gruppi.

Combinando le ultime due espressioni si ottiene:

$$P(Y_j) = \sum_{m=1}^{M} P(D_j=m) \prod_{i=1}^{nj} \sum_{t=1}^{T} P(X_{ij}=t|D_j=m) \prod_{k=1}^{K} P(y_{ijk}|X_{ij}=t,D_j=m)$$

che rappresenta il modello a classi latenti multilivello con effetti casuali discreti.

Il modello può essere ulteriormente generalizzato, includendo anche le *covariate*, con effetti su D_j , X_{ij} , o y_{ijk}.

2.3 STIMA DEL MODELLO

Le variabili latenti D_j e X_{ji} del modello a classi latenti multilivello sono trattate come dati mancanti in fase di stima del modello.

Molti metodi di trattamento dei dati incompleti si basano sull'assunzione di un modello esplicito e sulla stima di massima

verosimiglianza rispetto ai parametri del modello.

2.3.1 STIMA DI MASSIMA VEROSIMIGLIANZA

La verosimiglianza a dati osservati è una funzione complicata dei parametri e di rado le equazioni di massima verosimiglianza possono essere risolte in modo analitico. L'algoritmo *Expectation-Maximization* (EM) è un metodo che consente, per mezzo di un procedimento iterativo, di effettuare le stime di massima verosimiglianza dei parametri in presenza di dati incompleti, riconducendo il problema ad un problema standard di stima per dati completi.

Partendo da una stima iniziale dei parametri, l'algoritmo consiste, ad ogni iterazione del procedimento, nella applicazione di due passi:

- lo step E in cui si calcola il valore atteso della verosimiglianza rispetto alla distribuzione dei dati mancanti, condizionatamente ai dati osservati e alle stime correnti dei parametri;
- lo step M in cui viene massimizzato il valore atteso rispetto ai parametri.

A causa dell'elevata dimensionalità legata al problema dei dati mancanti, nell'implementazione dello step E per la stima del modello a classi latenti multilivello si utilizza un processo simile all'algoritmo forward-backward.

Verrà quindi impiegato un algoritmo EM con uno step E che è specificatamente adattato al problema in questione e che tiene conto di una condizione tipica dei modelli multilivello. Più precisamente, nello step E le osservazioni entro una unità di livello-2 sono assunte essere mutuamente indipendenti, data la classe di appartenenza del gruppo. Inoltre, lo step E utilizza l'algoritmo upward-downward per calcolare

la probabilità marginale a posteriori per entrambi i livelli.

2.4 MISURE DI ADATTAMENTO DEL MODELLO

Per valutare l'adattamento dei modelli a classi latenti vi sono diverse tecniche statistiche. Quella utilizzata maggiormente è la statistica data dal rapporto di verosimiglianza chi-quadro L^2 che confronta le stime di massima verosimiglianza ottenute con le frequenze attese.

Tuttavia, nel caso di dati "sparsi", la distribuzione chi-quadro non dovrebbe essere impiegata per calcolare i p-value poiché L^2 potrebbe non essere approssimato bene.

Un approccio alternativo è dato dal criterio di informazione di Akaike (*Akaike Information Criteria*, CAIC) e dal criterio di informazioni Bayesiano (*Bayesian Information Criteria*, BIC). La più usata nelle analisi a classi latenti è la statistica BIC:

$BIC_{L2} = L^2 - \log(N)df$

mentre il CAIC può essere espresso come:

$CAIC_{L2} = L^2 - [\log(N) + 1]df.$

Il modello con il valore più basso di BIC si preferisce ad altri con valori elevati. Una definizione più generale del BIC si basa sulla log-verosimiglianza e il numero di parametri (M) invece dell'L^2 e i gradi di libertà df ; cioè

$BIC_{LL} = -2LL + \ln(N)M.$

Un ulteriore approccio per la valutazione dell'adattamento del modello è dato dal confronto delle differenze dei valori della log-verosimiglianza tra due modelli tra loro annidati.

La riduzione percentuale misurata tra i due modelli rappresenta l'ammontare totale di associazione spiegata dal modello.

2.5 VALUTAZIONE DELLA SIGNIFICATIVITA' DEGLI EFFETTI

Nella fase di valutazione degli effetti delle variabili nel modello, quando la presenza o assenza di una variabile non risulta apportare differenze significative tra le classi del modello, tale variabile viene eliminata dal modello stesso. Per valutare se eliminare una variabile *k* dal modello, si testa l'ipotesi nulla che la distribuzione delle *s* categorie di *k* sia identica entro ciascuna classe *t*.

A tal fine si fa uso della relazione tra le probabilità di risposta condizionate e i parametri log lineari.

Una tecnica statistica utilizzata per tale scopo è il test dato dalla differenza degli L^2, ottenuti in corrispondenza del modello con e senza la variabile interessata. Con ΔL^2 viene perciò calcolata la differenza tra le due statistiche L^2.

Un altro modo per testare la significatività dei parametri degli indicatori è il test di *Wald*, il quale verifica se i coefficienti di regressione sono uguali tra le classi. Tuttavia, quest'ultimo test è meno potente della differenza tra gli L^2.

Sotto l'assunzione che il modello senza vincoli sia vero, entrambe le statistiche si distribuiscono asintoticamente come un Chi-quadro con numero di gradi di libertà uguale al numero di vincoli.

2.5 CLASSIFICAZIONE

Il passaggio finale nell'analisi dei modelli a classi latenti è quello di utilizzare i risultati del modello per classificare i casi nelle rispettive classi latenti individuate. Per ogni dato pattern di risposte, le stime della probabilità di appartenenza a posteriori possono essere ottenute usando il teorema di Bayes:

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

Vermunt e Magidson (2002) ritengono questo modello come un tipo di modello a cluster in quanto l'idea è quella di classificazione nelle *T* classi omogenee come avviene nell'analisi dei cluster. La differenza principale tra i due metodi è nella definizione di omogeneità in termini probabilistici nel modello a classi latenti e non in base a misure di distanze. I casi nella stessa classe latente sono simili tra di loro poiché le rispettive risposte provengono dalla stessa probabilità distributiva.
Ciascun caso viene classificato nella classe latente per cui presenta più elevata probabilità di appartenenza, e anche i gruppi di secondo livello, nel caso di un modello multilivello a classi latenti, vengono assegnati alla classe con maggiore probabilità a posteriori di appartenenza.

CAPITOLO 3. "*Le ricerche sulla fecondabilità della donna*"

3.1 INTRODUZIONE

La temperatura base del corpo (*Basal Body Temperature*) e i sintomi legati al muco cervicale (*Cervical Mucus Symptoms*) sono gli indicatori più utilizzati per identificare l'ovulazione e la fase fertile in un ciclo mestruale.

In particolare, l'inizio della fase fertile può essere individuato dal *Cervical Mucus Symptoms*, mentre la fine di questa fase può essere individuata da entrambi gli indicatori.

Per determinare la temperatura basale ci sono diversi metodi, ma le applicazioni di questi metodi per uno stesso insieme di dati possono differire tra loro, sia nell'individuare un ciclo come bifasico, oppure no, sia nella esatta collocazione dell'eventuale rialzo termico. Nei riguardi dell'accertamento di presenza o meno di muco cervicale, o di semplice sensazione, la valutazione è resa più complessa dalla fenomenologia tipologica, e dalla sua traduzione in simboli oggettivamente comparabili in osservazioni ripetute.

Per questo vengono definiti 14 differenti tipi di muco.

Dalle osservazioni del muco cervicale si cercano di capire vari aspetti della fecondabilità di una donna. E' molto importante per capire quindi la probabilità di concepimento nei giorni di picco del muco e nei giorni precedenti e successivi al giorno di picco.

Se sulla variabilità nell'interpretazione da parte di diversi osservatori della temperatura basale e del momento del suo rialzo esistono diversi lavori (ad esempio, Bauman, 1981), si è trovato solo uno studio sul muco cervicale e il suo picco (Kambic e Gray, 1989).

3.2 OBIETTIVI

Gli obiettivi principali di questo lavoro sono innanzitutto cercare di individuare gruppi omogenei di cicli e di donne, che abbiano quindi le stesse caratteristiche legate ad indicatori che verranno esposti nel successivo Capitolo.

Si andranno ad effettuare delle analisi prima sul file "*Donna*", dove vi sono variabili sulle caratteristiche delle donne, successivamente sul file "*Ciclo*" dove ci sono le caratteristiche del ciclo per ogni donna e poi si metteranno assieme questi due file.

Un altro obiettivo è quello di stabilire quali siano alcune importanti caratteristiche sulla fecondabilità di una donna e la probabilità di concepimento nei vari giorni del ciclo mestruale, andando a studiare le caratteristiche del muco nel giorno di picco e vedere come questo si evolve nei giorni precedenti e nei giorni successivi del picco.

3.3 IL METODO BILLINGS

Il metodo dell'ovulazione *Billings* è un metodo naturale di regolazione della fertilità e può essere utilizzato per la conoscenza dei periodi fertili o per la pianificazione familiare.

Il metodo si basa sull'osservazione delle modificazioni del muco cervicale, che appare più fluido e filante in prossimità dell'ovulazione, e sulle interpretazioni delle diverse sensazioni a livello vulvare che la donna deve imparare a riconoscere. Per una corretta applicazione del metodo è pertanto necessario che sia insegnato e appreso correttamente, poiché è affidato ad osservazioni soggettive.

Le osservazioni del muco vengono annotate su un'apposita cartella, consentendo l'individuazione della fase fertile di un ciclo. Il comportamento conseguente, nell'uso del metodo, dipenderà ovviamente dalle intenzioni della coppia: in generale, per evitare una

gravidanza, è richiesta l'astensione dai rapporti sessuali secondo una serie di regole prestabilite. Il metodo permette di suddividere ciascun ciclo mestruale in quattro fasi: i giorni del flusso mestruale, considerati potenzialmente fertili per la impossibilità di osservare l'eventuale presenza del sintomo del muco; una fase non fertile preovulatoria, detta quadro non fertile di base (QNFB); la fase fertile; la fase non fertile postovulatoria.

L'inizio della fase fertile è individuato dalla comparsa del sintomo del muco. In generale, le sue caratteristiche non sono le medesime per tutte le donne. Il QNFB può essere di due tipi: il più comune è caratterizzato da sensazione di asciutto e assenza di muco ed è identificabile fin dal primo ciclo di osservazione, il secondo è caratterizzato da sensazione di umido e/o perdita continua. Dopo circa tre cicli mestruali ciascuna donna avrà imparato ad individuare la tipologia del proprio QNFB di perdita continua in base al fatto che le caratteristiche di sensazione, aspetto e consistenza della perdita si mantengano stabilmente invariate giorno dopo giorno. La comparsa successiva di sintomo del muco cervicale con caratteristiche diverse dalla perdita tipica del QNFB segna l'inizio, per quel ciclo, del periodo di potenziale fertilità.

Un indicatore importante secondo il metodo *Billings* è il sintomo del picco del muco. In generale, esso si fa coincidere con l'ultimo giorno in cui la donna avverte la sensazione di bagnato o di lubrificazione e/o osserva la presenza di muco fluido filante o acquoso. Si considera che l'ovulazione avvenga entro il secondo giorno dopo il picco, perciò esso è utilizzato come riferimento per individuare la fine della fase fertile, tanto che, nel caso si voglia evitare una gravidanza, il metodo *Billings* richiede l'astensione dai rapporti sessuali fino a tre giorni successivi al picco. La mancata individuazione del picco del muco in un ciclo impedisce di considerare avvenuta l'ovulazione e dunque di identificare la fase non fertile postovulatoria. Tale evenienza è indice

di un prolungamento della fase preovulatoria e richiede pertanto l'applicazione delle regole d'uso proprie di tale fase.

Il metodo *Billings* può essere utilizzato efficacemente anche per facilitare l'evento di una gravidanza, infatti esso consente di individuare i giorni potenzialmente fertili di ciascun ciclo mestruale, che sono i giorni fino al picco.

Lo studio effettuato dal Dipartimento di Scienze Statistiche, con il coordinamento del professor Colombo, riguarda lo studio fatto da quattro Centri Italiani nei quali si fa riferimento al metodo dell'Ovulazione Billings: il Centro Lombardo Metodo Billings (CLOMB), a Milano, il Centro Piemontese Metodo Billings (CEPIMB), a Saluzzo, l'Associazione Metodo Billings Emilia Romagna (AMBER), a Parma, e il Centro Studi e Ricerche per la Regolazione della Fertilità dell'Università Cattolica del S. Cuore, a Roma.

Ognuno dei quattro centri ha inviato proprie schede relative a ciascun ciclo, le quali consistono nella codifica operata a posteriori da parte di una insegnante del metodo *Billings* delle informazioni registrate dalla donna in base alla propria osservazione. Esse contengono informazioni sul ciclo (inizio, fine, giorni interessati dal flusso mestruale); tipologia di muco osservato giornalmente secondo le categorie riportate nella *Tabella 3.1* e concordate tra i quattro centri; picco del muco (se individuato); rapporti sessuali; eventuale gravidanza; eventuali disturbi che possono avere alterato l'osservazione del muco (visite ginecologiche, terapie locali, stress, ecc.). Pur rimanendo le schede assolutamente anonime, si raccolgono anche alcune informazioni sulle caratteristiche demografiche della donna e del partner (età, data del matrimonio) e sulla storia ginecologica della donna (gravidanze precedenti, precedente assunzione di contraccettivi ormonali).

Per effettuare in modo corretto lo studio veniva posta particolare

attenzione su tutte le informazioni necessarie ad individuare le caratteristiche del muco cervicale, e quei problemi che potevano dare osservazioni distorte sull'evoluzione del muco, come ad esempio l'aver fatto una visita ginecologica, oppure la presenza di stress nella donna, ecc.

Tabella 3.1: "Classificazione del muco cervicale"

CODE	**SENSATION**	**APPEARANCE**
0	*No information*	*No information*
1	*No sensation or dry sensation*	*No mucus, no discharge*
2	*No more dry sensation*	*No mucus, or insubstantial discharge*
3	*Damp sensation*	*Thick, creamy, whitish, yellowish, sticky, stringy mucus*
4	*Wet, liquid (not slippery) sensation*	-
5	*Wet-lubrificated, slippery sensation*	*Clear, stretchy, liquid, watery mucus, blood trails*

All'interno dello studio si è chiesto a ciascuno dei quattro centri di inviare agli altri le fotocopie di un campione di 10 schede, corrispondenti ad altrettanti cicli mestruali, allo scopo di valutare la variabilità nell'interpretazione delle registrazioni effettuate dalle donne. A ciascun centro è stato chiesto di codificare, non solo i propri 10 cicli, ma anche i 30 ricevuti dagli altri, secondo la classificazione e le definizioni concordate e descritte sopra. La codifica è stata affidata, in ciascun centro, al "*Principal Investigator*", e dunque ad una insegnante esperta del metodo *Billings*.

Ogni centro ha quindi mandato i 40 cicli in proprio possesso al Dipartimento di Scienze Statistiche di Padova, in forma strettamente anonima, che ha poi creato i database sotto il coordinamento del

professor Colombo.

3.4 DESCRIZIONE DEL DATASET

In questa tesi viene svolto uno studio su due file presi dal dataset *Billings*. I due file sono "*Donna"* e *"Ciclo"*.

Le variabili contenute nel file *"Donna"* sono 17 e sono informazioni anagrafiche della donna e del suo partner, informazioni su precedenti gravidanze, data di nascita e sesso del figlio, informazioni sull'uso in passato di contraccezioni ormonali, data di entrata e di uscita dallo studio e le motivazioni dell'uscita, ecc.

Tutte le variabili sono riassunte nella *Tabella 3.2.*

Tabella 3.2: Descrizione delle variabili nel file "Donna", dataset Billings

Field progressive number	**Field name**	**Field type**	**Field length**	**Content**
1	*CODICE*	*Numerical*	*8*	*Woman's code: 2 figures for centre 3 figures for teacher 3 figures for woman*
2	*INGRESSO*	*Numerical*	*1*	*Progressive number of woman's entry in the study*
3	*TOT_SPEZ*	*Numerical*	*1*	*Total number of groups of consecutive cycles in each entry*
4	*NASCITA*	*Numerical MM-YY*	*4*	*Woman's birth date*
5	*NAS_PAR*	*Numerical*	*4*	*Partner's date of birth*

		MM-YY		
6	*GRAV_PRE*	*Numerical*	*2*	*Number of pregnancies before entering the study*
7	*D_UL_EV*	*Date YYYY-MM-DD*	*8*	*Date of last event before entering the study*
8	*TIPO_U_E*	*Numerical*	*1*	*Last event type 1=miscarriage 2=end of breastfeeding 3=child birth*
9	*CONTR*	*Numerical*	*1*	*Hormonal contraception 0=Missing data 1=Yes 2=No*
10	*D_CONTR*	*Date YYYY-MM-DD*	*8*	*Date when last pill was taken*
11	*D_USC_PR*	*Date YYYY-MM-DD*	*8*	*Date of study exit, i.e, day of last information recorded*
12	*MOT_USC*	*Numerical*	*1*	*Reason for leaving the study 1=pregnancy 2=miscarriage no later than 60 days since the beginning of the last period 3=drop out 4=study end*
13	*GRAV_CON*	*Numerical*	*1*	*Confirmation of pregnancy (with reference to the first day of last period) 0=missing data 1=pregnancy going on 2=pregnancy not going on*

14	*D_TEST*	*Date YYYY-MM-DD*	*8*	*Date of positive pregnancy test*
15	*D_PARTO*	*Date YYYY-MM-DD*	*8*	*Date of childbirth*
16	*SESSO*	*Numerical*	*1*	*Pregnancy result 0=missing data 1=boy 2=girl 3=miscarriage after 60 days since beginning of last period 4=twins boy & boy 5=twins girl & girl 6=boy & girl*
17	*D_MATR*	*Date YYYY-MM-DD*	*8*	*Marriage or beginning of relationship date*

Le variabili contenute nel file "Ciclo" sono al massimo 309 e sono informazioni sulle caratteristiche del ciclo, numero di cicli registrati, inizio e fine del ciclo, durata del ciclo, giorno del picco del muco, caratteristiche del muco, problemi nella donna, ecc.
Nella *Tabella 3.3* vengono riassunte tutte le variabili.

Tabella 3.3: Descrizione delle variabili nel file "Ciclo", dataset Billings

Field progressive number	**Field name**	**Field type**	**Field length**	**Content**
1	*CODICE*	*Numerical*	*8*	*Woman's code: 2 figures for centre 3 figures for teacher 3 figures for woman*
2	*INGRESSO*	*Numerical*	*1*	*Progressive number of woman's*

				entry in the study
3	*N_CARTELLA*	*Numerical*	*3*	*Cycle progressive number*
4	*P_SPEZ*	*Numerical*	*1*	*Total number of groups of consecutive cycles in each entry*
5	*DATA*	*Date YYYY-MM-DD*	*8*	*Date of cycle beginning*
6	*LUN_TOT*	*Numerical*	*2*	*Total cycle length in days*
7	*PICCOD*	*Numerical*	*2*	*Cycle day in which the woman identifies mucus peak*
8	*PICCO*	*Numerical*	*2*	*Cycle day in which there is mucus peak 0=peak non identifiable*
9	*BIOL*	*Numerical*	*2*	*Cycle day in which there is biological mucus peak 0=peak non identifiable*
10	*PERIOD*	*Numerical*	*2*	*Period length in days 0=missing data*
11	*QNFB*	*Numerical*	*1*	*Basic infertile pattern 0=not identified 1=dry (or first type) mucus 2=unchanging mucus 3=unchanging mucus*
12	*QUALIFI*	*Numerical*	*1*	*Cycle qualification 3=complete information on*

				mucus *6=missing data, disturbances, stress do not allow peak identification* *9=only total length available (no info on mucus or unprotected intercourse)*
13	*M1*	*Numerical*	*1*	*Mucus type* *0= no information* *1= no or dry sensation* *2= not anymore dry, no mucus, nor loss or insubstantial loss* *3=damp sensation, thick, creamy, whitish, yellowish sticky, stringy mucus* *4=wet, liquid sensation* *5=wet-slippery sensation, transparent, ropy, liquid, watery mucus, blood trails* *6=previous cycle without peak*
14	*D1*	*Numerical*	*1*	*Disturbances* *0=no* *1=yes*
15	*R1*	*Numerical*	*1*	*Unprotected intercourse* *0=no* *1=yes*
16	*M2*	*Numerical*	*1*	*Mucus type*
17	*D2*	*Numerical*	*1*	*Disturbances*
18	*R2*	*Numerical*	*1*	*Unprotected*

				intercourse
...	...	...	...	...
	M99	*Numerical*	*1*	*Mucus type*
	D99	*Numerical*	*1*	*Disturbances*
	R99	*Numerical*	*1*	*Unprotected intercourse*

3.5 ANALISI DESCRITTIVE

Nei paragrafi successivi sono rappresentate le analisi descrittive del file *"Donna"* e del file *"Ciclo"* presi separatamente, poi analisi descrittive dei due file uniti e infine vengono effettuate analisi descrittive usando solo il muco.

3.5.1 ANALISI DESCRITTIVE DEL FILE "*Donna*"

La variabile CODICE è una variabile numerica, dove le prime due cifre identificano i quattro Centri. Per ogni centro si è andato a vedere il numero di donne, il numero di spezzoni, il numero di gravidanze identificate, l'età media delle donne e relativa deviazione standard e l'età media dei partner e relativa deviazione standard.

I casi entrati nello studio sono stati scelti casualmente e sono diversi per i vari centri.

In particolare, sono entrate nello studio 50 donne seguite da diverse istruttrici del primo centro, codificato con il numero 80, per un totale di 71 spezzoni, quindi in media 1,39 per donna. Nei 71 spezzoni sono

state identificate 41 gravidanze.

Del secondo centro, codificato col numero 90, sono entrate nello studio 98 donne, per un totale di 114 spezzoni, in media 1,16 per donna. Nei 114 spezzoni sono state identificate 87 gravidanze.

Dal terzo centro, codificato col numero 10, sono entrate 17 donne, per un totale di 23 spezzoni, in media 1,35 per donna. Nei 23 spezzoni sono state identificate 13 gravidanze.

Infine, dal quarto centro, 28 donne per un totale di 36 spezzoni, in media 1,23 per donna. Nei 36 spezzoni sono state identificate 21 gravidanze.

Il numero minimo di spezzoni per donna è un solo spezzone, mentre il numero massimo è di 5 spezzoni.

Tabella 3.4: Numero di donne, di spezzoni e di gravidanze identificate nei quattro centri

CENTRI	**NUMERO DI DONNE**	**NUMERO DI SPEZZONI**	**NUMERO DI GRAVIDANZE IDENTIFICATE**
CODICE 80	*50*	*71*	*41*
CODICE 90	*98*	*114*	*87*
CODICE 10	*17*	*23*	*13*
CODICE 11	*28*	*36*	*21*
TOTALE	*193*	*244*	*162*

Per ogni spezzone, l'età media delle donne del primo centro è 29,59 anni, con una deviazione standard pari a 3,97, l'età media degli uomini è 32,52 anni, con deviazione standard pari a 4,74.

L'età media delle donne del secondo centro è 28,6 anni, con deviazione standard pari a 3,51, mentre l'età media degli uomini è 31,26 anni, con deviazione standard pari a 4,18.

L'età media delle donne del terzo centro è 29 anni, con deviazione standard 4,36, l'età media degli uomini è 31,26 anni con deviazione

standard pari a 4,39.

Infine, l'età media delle donne del terzo centro è pari a 33,61 anni con deviazione standard pari a 3,4, mentre l'età media degli uomini è pari a 35,83 anni, con deviazione standard pari a 5,05.

Tabella 3.5: Età media delle donne ed età media degli uomini con relative deviazioni standard, per ogni spezzone

CENTRI	ETA' DELLA DONNA		ETA' DELL'UOMO	
	MEDIA	DEVIAZIONE STANDARD	MEDIA	DEVIAZIONE STANDARD
CODICE 80	*29,59*	*3,64*	*32,75*	*4,75*
CODICE 90	*28,6*	*3,51*	*31,26*	*4,18*
CODICE 10	*29*	*4,36*	*33*	*4,39*
CODICE 11	*33,61*	*3,4*	*35,83*	*5,05*
TOTALE	*29,68*	*3,97*	*32,52*	*4,74*

3.5.2 ANALISI DESCRITTIVE DEL FILE "*Ciclo*"

Il numero totale di cicli studiati è 2901, di cui 992 provenienti dal primo Centro di Ovulazione, 1127 provenienti dal secondo Centro, 269 provenienti dal terzo Centro e 513 provenienti dall'ultimo centro.

Il numero medio di cicli per donna del primo centro è di 19,84 cicli; del secondo centro è di 11,5 cicli; del terzo centro è di 15,8 cicli e del quarto centro è di 18,32 cicli.

Il numero minimo di cicli che una donna ha segnato è 1, mentre il numero massimo è 73 cicli.

La lunghezza media del ciclo in giorni è pari a 28,9 giorni con una deviazione standard di 4,81.

Per ogni ciclo, la donna doveva segnare le caratteristiche del muco, e

quindi il giorno di picco del muco, ma dai dati è emerso che nel 21,34% dei cicli, ella non è riuscita ad identificare il giorno di picco.
Per i cicli in cui invece il giorno di picco è stato identificato, è risultato che il giorno del ciclo in cui la donna identifica il picco del muco è in media il diciassettesimo giorno.
Sul totale dei cicli, nel 91,9% dei casi la donna è riuscita a dare informazioni sul muco; nel 2,6% dei casi invece non si ha alcuna informazione sul muco, e questo è dovuto a dati mancanti, oppure a disturbi nella donna, quali ad esempio stress, che quindi non le hanno permesso di identificare alcuna informazione sul picco. Nel restante 5,5% dei casi, la donna non è riuscita a dare informazione sulle caratteristiche del muco, ma solamente sul numero di giorni di durata della perdita di muco.

3.5.3 ANALISI DESCRITTIVE DEI DUE FILES UNITI

Prendendo assieme i due file "*Donna*" e "*Ciclo*", possiamo andare a vedere insieme quali sono il numero di donne, il numero di spezzoni, il numero di cicli e il numero di gravidanze identificate per ogni Centro di Ovulazione Billings.

Tabella 3.6: Numero di donne, numero di spezzoni e numero di cicli per ogni centro

CENTRI	NUMERO DI DONNE	NUMERO DI SPEZZONI	NUMERO DI CICLI	NUMERO DI GRAVIDANZE IDENTIFICATE
CODICE 80	*50*	*71*	*992*	*41*
CODICE 90	*98*	*114*	*1127*	*87*
CODICE 10	*17*	*23*	*269*	*13*
CODICE 11	*28*	*36*	*513*	*21*
TOTALE	*193*	*244*	*2901*	*162*

3.5.4 ANALISI DESCRITTIVE USANDO SOLO IL MUCO

Un obiettivo di questo lavoro, come detto in precedenza, è andare a vedere le caratteristiche del muco nel giorno del picco, e come si evolve nei giorni precedenti e nei giorni successivi al picco.

Il muco è classificato in 7 tipi: 0=nessuna informazione; 1=nessuna sensazione o sensazione di asciutto; 2=non più asciutto, niente muco, niente perdite o niente perdite inconsistenti; 3=sensazione di umido, muco spesso, cremoso, biancastro, giallastro appiccicoso, filante; 4=sensazione di liquido bagnato; 5=sensazione di liquido scivoloso, muco trasparente, viscoso, liquido, acquoso, tracce di sangue; 6=ciclo precedente senza picco.

Nel giorno del picco, nello 0,9% dei casi non si ha alcuna informazione sul muco; nel 2,4% dei casi non vi è alcuna sensazione oppure si ha una sensazione di asciutto; nel 2,1% dei casi non si hanno perdite; nel 6% dei casi si ha una sensazione di umido, muco spesso, cremoso, biancastro, giallastro, filante; nell' 8% dei casi si ha una sensazione di liquido bagnato; nell' 80,6% dei casi si ha una

sensazione di liquido scivoloso, muco trasparente, viscoso, liquido, acquoso, tracce di sangue; e in nessun caso non si è avuto il ciclo precedente senza picco.

Tabella 3.7: Caratteristiche del muco nel giorno di picco

CODICE MUCO	PERCENTUALE CICLI
0	*0,90%*
1	*2,40%*
2	*2,10%*
3	*6,00%*
4	*8,00%*
5	*80,60%*
6	*0,00%*

Inoltre nel giorno di picco, nel 99,2% dei cicli la donna non ha nessun disturbo e nel 92,6% dei casi non ha rapporti non protetti.
Nei giorni precedenti al giorno di picco il muco si presenta in modo differente. Fino a 3 giorni prima del picco, ha sempre una sensazione di liquido scivoloso, e il muco si presenta trasparente, viscoso, liquido, acquoso o con tracce di sangue. Andando indietro nei giorni la donna non ha più questa sensazione, ma più una sensazione di umido.

Tabella 3.8: Caratteristiche del muco nei giorni precedenti al giorno di picco

CODICE MUCO	PERCENTUALE CICLI						
	GIORNO						
	-14	**-13**	**-12**	**-11**	**-10**	**-9**	**-8**
0	*11,10%*	*13,00%*	*13,50%*	*15,40%*	*15,10%*	*11,50%*	*7,70%*
1	*14,80%*	*15,90%*	*19,00%*	*21,90%*	*23,40%*	*27,60%*	*29,20%*
2	*4,30%*	*5,60%*	*5,80%*	*7,60%*	*10,20%*	*12,30%*	*12,60%*
3	*11,90%*	*13,10%*	*16,70%*	*17,70%*	*22,50%*	*26,80%*	*32,30%*
4	*1,60%*	*2,00%*	*1,30%*	*1,70%*	*1,70%*	*1,80%*	*2,40%*

5	*3,20%*	*3,30%*	*3,50%*	*3,50%*	*3,50%*	*4,30%*	*7,30%*
6	*53,10%*	*47,10%*	*40,20%*	*32,20%*	*24,60%*	*15,70%*	*8,50%*

CODICE MUCO	**PERCENTUALE CICLI**						
	GIORNO						
	-7	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**
0	*4,60%*	*2,70%*	*1,90%*	*1,50%*	*1,60%*	*1,10%*	*1,10%*
1	*25,80%*	*18,80%*	*12,20%*	*6,80%*	*3,70%*	*2,80%*	*2,00%*
2	*14,00%*	*14,50%*	*11,10%*	*8,10%*	*5,20%*	*2,90%*	*2,70%*
3	*37,80%*	*40,00%*	*40,70%*	*35,30%*	*25,40%*	*17,10%*	*8,90%*
4	*3,30%*	*5,00%*	*8,00%*	*11,70%*	*13,30%*	*10,40%*	*6,70%*
5	*10,30%*	*17,20%*	*25,20%*	*36,30%*	*50,70%*	*65,60%*	*78,60%*
6	*4,20%*	*1,80%*	*0,90%*	*0,30%*	*0,10%*	*0,10%*	*0,00%*

Dalla tabella si può notare che nella prima settimana del ciclo la maggior parte delle donne non ha ancora identificato il picco, e non ha alcuna sensazione di bagnato, o al più ha una sensazione di umido.

7, 6 e 5 giorni prima del giorno del picco aumenta la percentuale di donne che provano una sensazione di umido e inizia a crescere lentamente la percentuale di sensazione di liquido bagnato e di presenza di muco trasparente, viscoso, acquoso, che diventa poi elevata nei giorni subito precedenti al giorno di picco, quindi da 3 giorni prima in poi.

La percentuale di coloro che non hanno dei disturbi nel periodo precedente al giorno di picco rimane sempre costante attorno al 90% circa. La percentuale di coloro che hanno rapporti non protetti diminuisce nei giorni, parte da circa il 15% il primo giorno e poi diminuisce fino al 5,4% il giorno subito precedente al giorno di picco.

Dal giorno successivo al giorno di picco invece, si ha subito principalmente una sensazione di umido, e in percentuale meno significativa non si ha alcuna sensazione, oppure si ha una sensazione di asciutto.

Tabella 3.9: Caratteristiche del muco nei giorni successivi al giorno di picco

CODICE MUCO	PERCENTUALE CICLI						
	GIORNO						
	1	2	3	4	5	6	7
0	*1,20%*	*1,30%*	*1,50%*	*1,90%*	*3,70%*	*4,20%*	*5,10%*
1	*17,50%*	*23,90%*	*30,80%*	*35,60%*	*37,30%*	*39,00%*	*38,60%*
2	*17,30%*	*16,80%*	*16,50%*	*18,60%*	*19,40%*	*19,40%*	*20,30%*
3	*56,30%*	*51,80%*	*46,60%*	*39,80%*	*36,00%*	*33,60%*	*32,60%*
4	*2,60%*	*2,20%*	*1,60%*	*1,70%*	*1,80%*	*1,90%*	*1,40%*
5	*5,10%*	*4,00%*	*3,00%*	*2,40%*	*1,80%*	*1,90%*	*2,00%*
6	*0,00%*	*0,00%*	*0,00%*	*0,00%*	*0,00%*	*0,00%*	*0,00%*

CODICE MUCO	PERCENTUALE CICLI						
	GIORNO						
	8	9	10	11	12	13	14
0	*5,70%*	*5,60%*	*5,30%*	*5,00%*	*5,70%*	*5,90%*	*6,90%*
1	*38,10%*	*37,70%*	*38,90%*	*38,30%*	*36,40%*	*34,90%*	*34,30%*
2	*20,30%*	*20,50%*	*19,60%*	*19,80%*	*21,00%*	*21,30%*	*18,00%*
3	*32,90%*	*33,10%*	*32,60%*	*32,90%*	*32,20%*	*32,90%*	*34,30%*
4	*1,30%*	*1,40%*	*1,60%*	*1,90%*	*2,20%*	*2,60%*	*3,20%*
5	*1,70%*	*1,70%*	*2,00%*	*2,20%*	*2,50%*	*2,40%*	*3,30%*
6	*0,00%*	*0,00%*	*0,00%*	*0,00%*	*0,00%*	*0,00%*	*0,00%*

La percentuale di coloro che non hanno disturbi anche dopo il giorno di picco è sempre costante attorno al 99%, e la percentuale di coloro che non hanno rapporti non protetti diminuisce nei giorni successivi.

CAPITOLO 4: "*Segmentazione dei cicli e delle donne con il modello a classi latenti classico*"

4.1 INTRODUZIONE

L'obiettivo di questa tesi è quello di andare ad individuare dei segmenti di donne che siano internamente omogenei rispetto alle caratteristiche del ciclo mestruale. Quindi, prima di tutto, è necessario individuare proprio dei segmenti di cicli mestruali.

A tale scopo si stima un modello classico a classi latenti utilizzando il *software Latent GOLD 4.5*, creato *ad hoc* da Madigson e Vermunt (2005).

Come indicatori sono state utilizzate le variabili *"LUN_TOT", "PICCO"*, *"DURATA"*, *"QNFB"* e *"QUALIFI"*.

In particolare la variabile *"PICCO"* è stata suddivisa nel seguente modo:

$$PICCO = \begin{cases} 0 \text{ se il picco non è identificabile} \\ 1 \text{ se il giorno del picco è identificato prima del 15° giorno compreso} \\ 2 \text{ se il giorno del picco è identificato dal 16° al 19° giorno} \\ 3 \text{ se il giorno del picco è identificato dal 20° giorno in poi} \end{cases}$$

La variabile *"PERIOD"* è invece stata suddivisa così:

$$PERIOD = \begin{cases} 0 \text{ dato mancante} \\ 1 \text{ se la durata della mestruazione è minore o uguale a 4 giorni} \\ 2 \text{ se la durata della mestruazione è compresa tra 5 e 6 giorni} \\ 3 \text{ se la durata della mestruazione è maggiore o uguale a 7 giorni} \end{cases}$$

La variabile *"QUALIFI"* è stata suddivisa come segue:

$$QUALIFI = \begin{cases} 1 \text{ se assume valore 3} \\ 2 \text{ se assume valore 6} \\ 3 \text{ se assume valore 9} \end{cases}$$

Infine la variabile *"LUN_TOT"* è stata suddivisa nel seguente modo:

$$LUN\_TOT = \begin{cases} 1 \text{ se la durata del ciclo è minore o uguale a 26 giorni} \\ 2 \text{ se la durata del ciclo è compresa tra 27 e 32 giorni} \\ 3 \text{ se la durata del ciclo è maggiore o uguale a 33 giorni} \\ 4 \text{ se siamo in presenza di una gravidanza (valore 99)} \end{cases}$$

Gli indicatori *"LUN_TOT"*, *"PICCO"* e *"PERIOD"* sono quindi trattati come qualitative ordinali, mentre *"QNFB"* e *"QUALIFI"* come qualitative nominali.

4.2 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*), illustrato nel *Capitolo 2*.

Latent GOLD fornisce in *output* il risultato del *test di Wald* per la significatività delle variabili nel modello.
In tutti i modelli gli indicatori sono considerati significativi.
Ogni modello è stato stimato più volte, con insiemi di valori iniziali per le procedure iterative di stima differenti. Per evitare di incorrere in massimi locali, quindi è stato scelto il modello con il migliore adattamento.
Nella *Tabella 4.1* sono riportati i valori di alcune statistiche, tra cui *BIC* ed L^2, utili a giudicare la bontà dell'adattamento: sono preferibili modelli con valore sia del *BIC,* sia dell' L^2 basso.
Il *BIC* tiene conto della parsimonia del modello, mentre l' L^2 indica quanta relazione tra le variabili non è spiegata.
Nella *Tabella 4.1* inoltre troviamo il numero dei parametri del modello e i gradi di libertà relative alla distribuzione X^2 della statistica rapporto di verosimiglianza (L^2).

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

	BIC(LL)	**Npar**	**L^2**	**df**
1 cluster	*28120,1488*	*14*	*3996,7982*	*753*
2 cluster	*26378,4559*	*20*	*2207,2396*	*747*
3 cluster	*25949,2691*	*26*	*1730,1870*	*741*
4 cluster	*25486,3162*	*32*	*1219,3683*	*735*
5 cluster	*25217,3271*	*38*	*902,5135*	*729*
6 cluster	***24990,9108***	*44*	***628,2315***	*723*
7 cluster	*25152,6072*	*50*	*742,0621*	*717*

Osservando i valori del *BIC* e i valori dell' $L^{2,}$ il modello migliore sembrerebbe essere senza dubbio quello a sei cluster; infatti i valori più bassi di tali indicatori sono quelli assegnati al modello a sei cluster, rispetto a quelli assegnati agli altri modelli.
Si è scelto quindi di stimare un modello a sei classi latenti e si è

proseguito a costruire i profili dei segmenti di cicli mestruali.

4.3 INDIVIDUAZIONE DEL PROFILO DEI SEGMENTI DI CICLI

I segmenti individuati sono sei. Il primo segmento comprende il 43,85% dei cicli delle donne, il secondo il 27,01%, il terzo l' 11,63%, il quarto l' 8,02%, il quinto il 5,46% e il sesto il 4,04%, come è possibile osservare dalla *Tabella 4.2*.

Tabella 4.2 "Probabilità condizionate totali"

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
Cluster Size	*0,4385*	*0,2701*	*0,1163*	*0,0802*	*0,0546*	*0,0404*

Nei prossimi paragrafi verranno descritti i segmenti sulla base delle probabilità condizionate.

4.3.1 SEGMENTI SULLA BASE DEGLI INDICATORI

Nella *Tabella 4.3* sono presentate le probabilità condizionate relative agli indicatori, che permetteranno di individuare le caratteristiche del ciclo per ogni segmento.

Tabella 4.3 "Probabilità condizionate relative agli indicatori"

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
Indicators						
LUN_TOT	0,4493	0,0702	0,0011	0,0020	0,2555	0,0000
1	0,5454	0,8468	0,2817	0,3675	0,7278	0,0000
2	0,0053	0,0822	0,5960	0,5511	0,0167	0,0078
3	0,0000	0,0008	0,1212	0,0795	0,0000	0,9921
4						
PICCO						
0	*0,1869*	*0,0000*	*0,0000*	*0,9970*	*0,9986*	*0,0013*
1	*0,8086*	*0,0294*	*0,0000*	*0,0030*	*0,0014*	*0,6165*
2	*0,0045*	*0,9691*	*0,0154*	*0,0000*	*0,0000*	*0,3822*
3	*0,0000*	*0,0015*	*0,9846*	*0,0000*	*0,0000*	*0,0000*
PERIOD						
0	*0,0001*	*0,0001*	*0,0000*	*0,0000*	*0,9980*	*0,0001*
1	*0,2185*	*0,1521*	*0,1196*	*0,1137*	*0,0020*	*0,1828*
2	*0,6162*	*0,6130*	*0,5971*	*0,5927*	*0,0000*	*0,6184*
3	*0,1651*	*0,2349*	*0,2833*	*0,2936*	*0,0000*	*0,1987*
QNFB						
0	*0,2103*	*0,1238*	*0,1057*	*0,1357*	*0,3007*	*0,1563*
1	*0,6858*	*0,6606*	*0,6389*	*0,6708*	*0,6458*	*0,6821*
2	*0,0523*	*0,0825*	*0,0904*	*0,0776*	*0,0325*	*0,0697*
3	*0,0516*	*0,1331*	*0,1650*	*0,1159*	*0,0211*	*0,0919*
QUALIFI						
1	*0,9865*	*1,0000*	*0,9999*	*0,7535*	*0,0000*	*0,9998*
2	*0,0134*	*0,0000*	*0,0001*	*0,2415*	*0,0020*	*0,0002*
3	*0,000*	*0,0000*	*0,0000*	*0,0051*	*0,9980*	*0,0000*

Nel primo cluster si notano buone percentuali delle modalità di risposta 1 e 2 per l'indicatore *"LUN_TOT"*, un'alta percentuale della modalità di risposta 1 per l'indicatore *"PICCO",* una percentuale buona di risposta 2 per l'indicatore *"PERIOD"*, una percentuale buona di risposta 1 per l'indicatore *"QNFB"* e una percentuale alta della

modalità di risposta 1 per l'indicatore *"QUALIFI"*. Il primo cluster è quindi caratterizzato da cicli lunghi meno di 32 giorni e in cui il giorno del picco è identificato entro il 15° giorno del ciclo, la durata della mestruazione va da 5 a 6 giorni, il quadro non fertile di base è caratterizzato da un muco asciutto e si ha una informazione completa sul muco.

Nel secondo cluster si notano un'alta percentuale della modalità di risposta 1 per l'indicatore *"LUN_TOT"*, una percentuale alta della modalità di risposta 2 per l'indicatore *"PICCO"*, una buona percentuale di risposta 2 per l'indicatore *"PERIOD"*, una buona percentuale di risposta 1 per l'indicatore *"QNFB"* e il 100% della modalità di risposta 1 per l'indicatore *"QUALIFI"*. Quindi è possibile affermare che il secondo cluster è caratterizzato da cicli lunghi da 27 a 32 giorni e in cui il giorno del picco è esclusivamente identificato tra il 16° e il 19° giorno del ciclo, la durata della mestruazione va da 5 a 6 giorni, il quadro non fertile di base è caratterizzato maggiormente da un muco asciutto e si hanno informazioni complete sul muco.

Nel terzo cluster si notano una buona percentuale della modalità di risposta 3 per l'indicatore *"LUN_TOT"*, un'alta percentuale della modalità di risposta 3 per l'indicatore *"PICCO"*, buone percentuali della risposta 2 per l'indicatore *"PERIOD"* e di risposta 1 per l'indicatore *"QNFB"* e un'alta percentuale della modalità di risposta 1 per l'indicatore *"QUALIFI"*. Dunque, il terzo cluster è caratterizzato da cicli lunghi più di 32 giorni e in cui il giorno del picco è identificato dopo il 19° giorno del ciclo, la durata della mestruazione va da 5 a 6 giorni, il quadro non fertile di base è caratterizzato maggiormente da un muco asciutto e si hanno informazioni complete sul muco.

Nel quarto cluster invece si notano una buona percentuale della modalità di risposta 3 per l'indicatore *"LUN_TOT",* un'alta percentuale della modalità di risposta 0 per l'indicatore *"PICCO"*,

buone percentuali della risposta 2 per l'indicatore *"PERIOD"* e della risposta 1 per l'indicatore *"QNFB"* e un'alta percentuale della modalità di risposta 1 per l'indicatore *"QUALIFI"*. Allora, il quarto cluster è caratterizzato da cicli lunghi più di 32 giorni e in cui il giorno del picco non è stato identificato, con una durata della mestruazione che va da 5 a 6 giorni, con un quadro non fertile di base caratterizzato da muco asciutto e si hanno informazioni complete sul muco.

Nel quinto cluster si notano un'alta percentuale della modalità di risposta 2 per l'indicatore *"LUN_TOT"*, un'alta percentuale della modalità di risposta 0 per l'indicatore *"PICCO"*, un' alta percentuale per la modalità di risposta 0 per l'indicatore *"PERIOD"*, una buona percentuale per la modalità di risposta 1 per l'indicatore *"QNFB"* e un'alta percentuale della modalità di risposta 3 per l'indicatore *"QUALIFI"*. E' possibile affermare quindi che il quinto cluster è caratterizzato da cicli maggiormente lunghi dai 27 ai 32 giorni, in cui il giorno del picco non è stato identificato e neanche la durata della mestruazione, con il quadro non fertile di base caratterizzato da muco asciutto e non si hanno informazioni sulle caratteristiche del muco, ma solo sulla lunghezza totale del ciclo; risultati abbastanza prevedibili dato quanto scritto sopra.

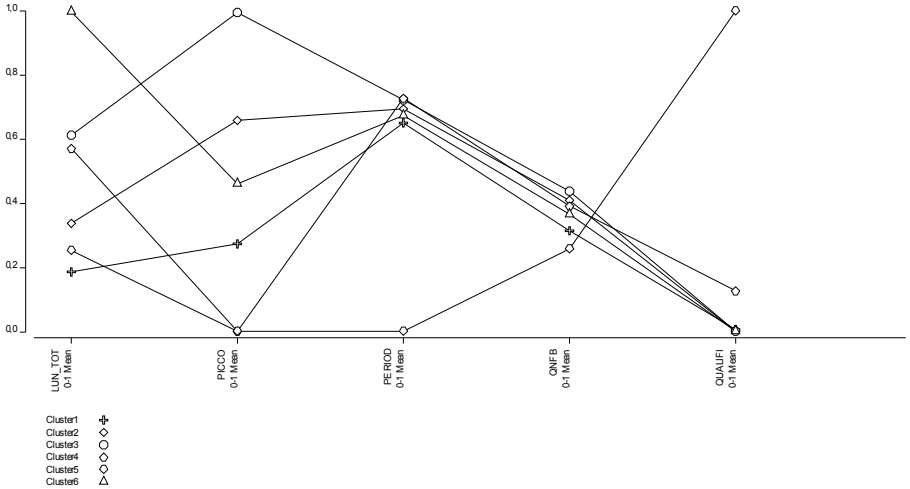
Infine, nel sesto cluster si notano un'alta percentuale di modalità di risposta 4 per l'indicatore *"LUN_TOT"*, una buona percentuale della modalità di risposta 1 per l'indicatore *"PICCO"* e buone percentuali della risposta 2 per l'indicatore *"PERIOD"* e della risposta 1 per l'indicatore *"QNFB"* e un'alta percentuale della modalità di risposta 1 per l'indicatore *"QUALIFI"*. Si può quindi affermare che il sesto cluster è caratterizzato principalmente da cicli con gravidanze, in cui il giorno del picco è stato individuato prima del 15° giorno del ciclo, in cui la durata della mestruazione è compresa tra i 5 e i 6 giorni, il quadro non fertile di base è caratterizzato maggiormente da muco asciutto e si hanno informazioni complete sul muco.

Per identificare in modo più immediato i sei segmenti si può fare una descrizione sintetica: il primo segmento è caratterizzato da cicli di lunghezza breve, con picco identificato e *QNFB* asciutto; il secondo segmento è caratterizzato da cicli di lunghezza normale, con picco identificato e *QNFB* asciutto; il terzo segmento è caratterizzato da cicli lunghi, con picco identificato e *QNFB* asciutto; il quarto segmento è caratterizzato da cicli lunghi, con picco non identificato e *QNFB* asciutto; il quinto segmento è caratterizzato da cicli di lunghezza normale, con picco non identificato e *QNFB* asciutto; il sesto segmento è caratterizzato da cicli con gravidanze, con picco identificato e *QNFB* asciutto.

4.3.2 DESCRIZIONE DEI SEGMENTI INDIVIDUATI

Osservando il *profile plot* (*Grafico 4.1*) dei sei segmenti è possibile farsi una prima idea dei profili e di alcune loro caratteristiche. In questo diagramma sono rappresentati in ordinata le probabilità condizionate e in ascissa gli indicatori.

Grafico 4.1 "Profile plot dei segmenti"



Osservando il *profile plot* è possibile affermare che il ciclo del primo cluster è mediamente lungo 24 giorni, il giorno del picco del muco è individuato mediamente all' 8° giorno circa, la durata della mestruazione è di 5 giorni e mezzo in media, il quadro non fertile di base è caratterizzato da muco asciutto e si hanno informazioni complete sulle caratteristiche del muco.

Il ciclo del secondo cluster è invece lungo mediamente circa 29 giorni, il giorno del picco del muco è individuato intorno al 17° giorno del ciclo, la durata della mestruazione è di quasi 6 giorni in media, il quadro non fertile di base è caratterizzato da muco asciutto e si hanno informazioni complete sul muco.

Il ciclo del terzo cluster è lungo mediamente 36 giorni, il giorno del picco del muco è individuato attorno al 24° giorno del ciclo in media, la durata della mestruazione è di 6 giorni, il quadro non fertile di base è caratterizzato da muco asciutto e si hanno informazioni complete sul muco.

Il ciclo del quarto cluster è mediamente lungo 34 giorni, il giorno del picco in questo cluster non è identificato, la lunghezza media della mestruazione è di 6 giorni, il quadro non fertile di base è muco asciutto e si hanno informazioni complete sul muco.

Il ciclo del quinto cluster è lungo mediamente 28 giorni, il giorno del picco del muco non è identificato e neanche la durata della mestruazione, il quadro non fertile di base è caratterizzato da muco asciutto e si hanno, chiaramente, informazioni solo sulla lunghezza del ciclo e nessuna informazione sulle caratteristiche del muco.

Infine, fanno parte del sesto cluster le gravidanze, il giorno del picco del muco è stato identificato mediamente attorno al 14° giorno, la durata della mestruazione è mediamente di 6 giorni, il quadro non fertile di base è caratterizzato da muco asciutto e si hanno informazioni complete sulle caratteristiche del muco.

4.4 ANALISI DEI RESIDUI BIVARIATI

Da un punto di vista statistico, il modello a sei classi latenti sembra buono. L'analisi delle *Bivariate Residuals* fa sorgere qualche dubbio, infatti nella matrice del *residui bivariati* (*Tabella 4.4*) c'è un valore maggiore di 1.

Tabella 4.4 "Matrice delle Bivariate Residuals per il modello a 6 classi latenti"

Indicators	**LUN_TOT**	**PICCO**	**PERIOD**	**QNFB**	**QUALIFI**
LUN_TOT	.				
PICCO	*0,0279*	.			
PERIOD	*0,0187*	*0,2700*	.		
QNFB	*0,9400*	*0,0127*	***3,8289***	.	
QUALIFI	*0,8478*	*0,1789*	*0,1400*	*0,6875*	.

Si può notare che la coppia *"QNFB"-"PERIOD"* è dipendente, 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. Il modello a sei classi latenti stimato con effetto diretto tra *"QNFB"* e *"QUALIFI"* non presenta un adattamento migliore rispetto a quello classico. Infatti il valore del *BIC* del modello a effetti diretti è pari a *25112,7300*, maggiore del valore del *BIC* del modello classico (*24990,9108*), e il valore dell' L^2 del modello a effetti diretti è pari a *742,0730*, maggiore del valore dell' L^2 del modello classico (*628,2315*).

La segmentazione implementata dal modello classico si può ritenere quindi valida, anche perché il profilo dei segmenti corrisponde alle aspettative su di essi, in base alle analisi descrittive presentate nel

Capitolo 3.

L'adattamento non del tutto soddisfacente di questo modello potrebbe essere dovuto alla presenza di una struttura gerarchica a due livelli, 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.

Nel *Capitolo 5* verrà dunque presentata una stima di modelli a classi latenti multilivello per l'individuazione di segmenti di cicli mestruali.

4.5 INDIVIDUAZIONE DEI SEGMENTI DI DONNE

Nell'andare ad individuare segmenti di donne sono state utilizzate come indicatori le variabili *"ETA_DONNA"*, *"ETA_UOMO"*, *"GRAV_PRE"* e *"GRAV_CON"*.

Le variabili *"ETA_DONNA"* ed *"ETA_UOMO"* sono state ottenute confrontando la data di inizio di ogni ciclo mestruale con la data di nascita delle donne e del partner rispettivamente[5]. Queste variabili sono state poi suddivise nel seguente modo:

5 F. Comodo, E. Moretti e G. Collodel nell'articolo "Età e riproduzione. Il tramonto della fertilità" affermano che in entrambi i sessi la funzione riproduttiva raggiunge la massima efficienza tra i 20 e i 25 anni, per poi iniziare a diminuire gradualmente. Il tempo che passa, tuttavia, incide in misura assai diversa nei due sessi. La capacità di concepimento nel caso dell'uomo ha un andamento assai più stabile e tende a conservarsi con il passare degli anni, pur mostrando una tendenza al peggioramento delle caratteristiche degli spermatozoi dopo i 50 anni di età. Nella donna mantiene una certa stabilità solo fino ai 30 anni per poi precipitare, con un primo netto calo sopra i 35 anni ed una diminuzione ancora più drastica dopo i 40 anni.

$$\text{ETA\_DONNA}^{6} = \begin{cases} 1 \text{ se l'età è compresa tra i 21 e i 27 anni} \\ 2 \text{ se l'età è compresa tra i 28 e i 30 anni} \\ 3 \text{ se l'età è compresa tra i 31 e i 33 anni} \\ 4 \text{ se l'età è compresa tra i 34 e i 40 anni} \end{cases}$$

$$\text{ETA\_UOMO}^{7} = \begin{cases} 1 \text{ se l'età è compresa tra i 22 e i 29 anni} \\ 2 \text{ se l'età è compresa tra i 30 e i 32 anni} \\ 3 \text{ se l'età è compresa tra i 33 e i 37 anni} \\ 4 \text{ se l'età è compresa tra i 38 e i 49 anni} \end{cases}$$

Inoltre la variabile *"GRAV_PRE"* è stata suddivisa così:

$$\text{GRAV\_PRE} = \begin{cases} 0 : \text{nessuna gravidanza} \\ 1 : 1 \text{ gravidanza} \\ 2 : 2 \text{ gravidanze} \\ 3 : 3 \text{ o più gravidanze} \end{cases}$$

Gli indicatori *"ETA_DONNA", "ETA_UOMO"* e *"GRAV_PRE"* sono stati trattati come qualitative nominali, mentre *"GRAV_CON"* come qualitativa nominale.

6 I tassi di fertilità della popolazione femminile in relazione all'età sono: 100% per età compresa tra i 20 e 24 anni, 80-100% per età compresa tra i 25 e i 29 anni, 50-55% per età compresa tra i 30 e i 34 anni, 18-25% per età compresa tra i 35 e i 39 anni, 5-7% per età compresa tra i 40 e i 44 anni, e nulla dopo i 45 anni. (F. Comodo, E. Moretti e G. Collodel)

7 L'età ha un impatto meno negativo sulla fertilità maschile. (F. Comodo, E. Moretti e G. Collodel)

4.5.1 STIMA DEL MODELLO

Sono stati stimati i modelli a classi latenti con un numero di classi da uno a cinque. In tutti i modelli tutti gli indicatori sono significativi, fatta eccezione per *"GRAV_CON",* che non raggiunge mai la significatività all'1%, ovvero la variabile non discrimina tra le classi, nel senso che assume valori molto simili, pertanto va eliminata.

La stima dei modelli senza *"GRAV_CON"* è migliore dal punto di vista dell'adattamento.

Eliminata *"GRAV_CON",* ogni modello è stato stimato più volte con insiemi di valori iniziali per le procedure iterative di stima differenti.

Nella *Tabella 4.5* sono riportati i valori del *BIC*, dell'L^2, il numero dei parametri nel modello e i gradi di libertà relative alla distribuzione X^2 della statistica rapporto di verosimiglianza.

Tabella 4.5 "Indici di bontà di adattamento dei modelli a classi latenti stimati"

	BIC(LL)	**Npar**	**L^2**	**df**
1 cluster	*1771,1945*	*9*	*465,1260*	*54*
2 cluster	*1481,3712*	*13*	*153,8764*	*50*
3 cluster	***1455,6926***	*17*	***106,7714***	*46*
4 cluster	***1455,9721***	*21*	***85,6246***	*42*
5 cluster	*1463,0618*	*25*	*71,2880*	*38*

A giudicare dal *BIC* il modello migliore è quello a tre cluster. Il valore associato al modello a tre cluster però non è di molto inferiore a quello associato al modello a quattro cluster. Per testare se il miglioramento apportato dal modello a quattro cluster sia significativo, possiamo utilizzare 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 X^2 con gradi di libertà pari alla differenza dei gradi di libertà dei modelli a confronto.

Abbiamo, quindi, $\Delta L^2 = L^2_{3classi} - L^2_{4classi} = 106,7714 - 85,6246 = 21,1468$ e $df = df_{3classi} - df_{4classi} = 46 - 42 = 4$. Il *p-value* associato al test è prossimo allo zero, quindi il modello a quattro classi apporta dei miglioramenti significativi rispetto a quello a tre classi.
Si è scelto quindi di stimare un modello a quattro classi latenti e si è proseguito a descrivere i segmenti individuati.

4.5.2 DESCRIZIONE DEI SEGMENTI INDIVIDUATI SULLA BASE DI INDICATORI

I segmenti individuati sono quattro. Il primo comprende il 43,13% delle donne, il secondo il 34,28%, il terzo il 16,97% e il quarto il 5,63%, come è possibile vedere dalla *Tabella 4.6*.

Tabella 4.6 "Probabilità condizionate totali"

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Cluster size	*0,4313*	*0,3428*	*0,1697*	*0,0563*

Nella *Tabella 4.7* sono presentate le probabilità condizionate relative agli indicatori, che permetteranno di individuare le caratteristiche del ciclo per ogni segmento.

Tabella 4.7 "Probabilità condizionate relative agli indicatori"

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Indicators				
ETA_DONNA				
1	*0,3902*	*0,0000*	*0,0033*	*0,0176*
2	*0,5353*	*0,0001*	*0,1454*	*0,3200*
3	*0,0738*	*0,0415*	*0,6463*	*0,5859*
4	*0,0007*	*0,9584*	*0,2050*	*0,0766*
ETA_UOMO				
1	*0,5608*	*0,0000*	*0,0074*	*0,0384*
2	*0,4024*	*0,0026*	*0,2012*	*0,4013*
3	*0,0366*	*0,2238*	*0,6961*	*0,5323*
4	*0,0001*	*0,7737*	*0,0954*	*0,0280*
GRAV_PRE				
0	*0,7096*	*0,0000*	*0,2813*	*0,0000*
1	*0,2419*	*0,0100*	*0,3932*	*0,0005*
2	*0,0484*	*0,5884*	*0,3224*	*0,1861*
3	*0,0001*	*0,4015*	*0,0031*	*0,8134*

Nel primo cluster troviamo con una buona percentuale donne di età compresa tra i 22 e i 30 anni e uomini di età compresa tra i 22 e i 29 anni, donne che con un'alta percentuale non hanno avuto gravidanze precedenti.

Nel secondo cluster troviamo con alte percentuali donne di età compresa tra i 34 e i 40 anni e uomini di età compresa tra i 38 e i 49 anni, donne che con una buona percentuale hanno già avuto almeno due gravidanze precedenti.

Nel terzo cluster invece troviamo con buone percentuali donne che hanno un'età compresa tra i 31 e i 33 anni e uomini che hanno un'età compresa tra i 33 e i 37 anni, donne che hanno avuto in precedenza già una o due gravidanze.

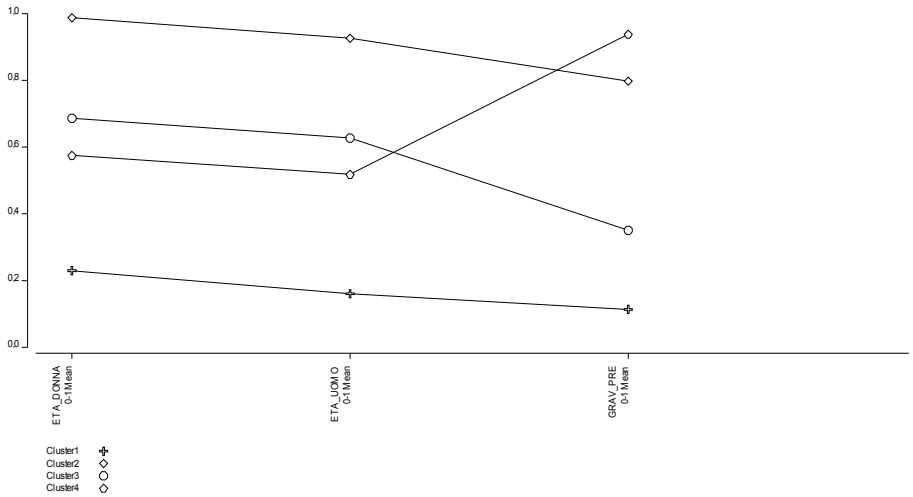
Infine, nel quarto cluster troviamo ancora con buone percentuali donne e uomini della stessa età delle donne e degli uomini del terzo

cluster, ma queste sono donne che con alta percentuale in precedenza hanno già avuto tre o più gravidanze.

Dunque, possiamo affermare che il primo cluster è caratterizzato da coppie di donne e uomini molto giovani che non hanno avuto nessun figlio prima di entrare nello studio; il secondo cluster è caratterizzato da coppie di donne e uomini maturi, in cui la donna ha avuto già almeno due gravidanze prima di entrare nello studio; il terzo e il quarto cluster invece sono caratterizzati entrambi da coppie di donne e uomini di età media, ma le donne del terzo cluster prima di entrare nello studio avevano in media già avuto una o due gravidanze, mentre le donne del quarto cluster tre o più.

Si possono avere dei profili dei segmenti più dettagliati grazie al *profile plot* (*Grafico 4.2*).

Grafico 4.2 "Profile plot"



Osservando il *profile plot* si può affermare che il primo cluster è caratterizzato da donne e uomini che in media hanno, entrambi, 27 anni, donne che non hanno mai avuto una gravidanza.

Il secondo cluster è caratterizzato da donne che in media hanno 40 anni e uomini che in media hanno 45 anni, donne che hanno avuto in

media almeno già due gravidanze.
Il terzo cluster è caratterizzato da donne che in media hanno 33 anni e uomini che in media hanno 35 anni, donne che in precedenza hanno avuto mediamente una gravidanza.
Il quarto cluster è caratterizzato da donne che in media hanno 31 anni e uomini 33 anni, donne che mediamente hanno già avuto tre o più gravidanze.
I segmenti individuati sembrerebbero essere buoni, infatti rispettano le aspettative in base alle analisi descrittive presentate nel *Capitolo 3;* questo è confermato dall'analisi dei *residui bivariati* (*Tabella 4.8),* infatti tutti i valori sono inferiori a 1.

Tabella 4.8 "Matrice delle Bivariate Residuals per il modello a 4 classi latenti"

Indicators	**ETA_ DONNA**	**ETA_ UOMO**	**GRAV_PRE**
ETA_DONNA	.		
ETA_UOMO	*0,0074*	.	
GRAV_PRE	*0,0230*	*0,0838*	.

Quindi la segmentazione implementata dal modello classico si può ritenere valida, sia perché il profilo dei segmenti corrisponde alle aspettative su di essi, in base alle analisi descrittive presentate le *Capitolo 3*, sia perché l'adattamento di questo modello sembrerebbe essere soddisfacente.

CAPITOLO 5: *"Segmentazione dei cicli con il modello a classi latenti multilivello"*

5.1 INTRODUZIONE

I dati del dataset *Billings* presentano una struttura gerarchica a due livelli, in cui i cicli mestruali sono raggruppabili per donna. 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 nelle stime, viene adottato l'approccio a effetti casuali: i parametri tra i gruppi variano secondo una distribuzione non specificata a priori.

Si distinguono le unità in due gruppi: le unità di livello uno sono i cicli mestruali, le unità di livello due sono le donne.

L'obiettivo che ci si pone è di classificare le unità di primo livello in classi latenti e di classificare le donne in gruppi, secondo la loro tendenza ad appartenere ai *cluster* individuati al primo livello.

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

5.2 STIMA DEL MODELLO

Come per l'analisi a classi latenti classica, si sono stimati numerosi modelli con l'algoritmo *EM*, 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) e dei gruppi (livello due). Per ogni tipologia, si è scelto il migliore per adattamento.

In tutti i modelli l'indicatore *"QUALIFI"* risulta essere non significativo, non raggiunge mai la significatività all'1%, ovvero questa variabile non discrimina tra le classi, nel senso che assume valori molto simili. Eliminato questo indicatore, la stima dei modelli risulta effettivamente migliore dal punto di vista dell'adattamento.

Nella *Tabella 5.1* viene rappresentato un riepilogo dei risultati, che contiene il valore della statistica *BIC* del modello.

Tabella 5.1 "Valori di BIC dei modelli multilivello stimati per diverse combinazioni di cluster e gruppi"

	Gruppi 1	Gruppi 2	Gruppi 3	Gruppi 4	Gruppi 5
Cluster 1	*2442,0624*	2447,7254	2453,3884	2459,0513	2464,7143
Cluster 2	2440,5582	*2305,6438*	2312,0521	2323,4810	2334,6936
Cluster 3	2416,3245	2317,5629	*2284,5439*	2301,5913	2318,5740
Cluster 4	2417,0548	2325,9263	2289,5819	***2278,8779***	2325,6668
Cluster 5	2439,5832	2321,0915	*2295,3698*	2313,5686	2333,3900
Cluster 6	2464,8970	2335,1615	*2312,5110*	2319,9990	2351,2100

Le celle con i valori in corsivo indicano i modelli con il miglior *BIC* di riga, le celle con i valori sottolineati indicano i modelli con il miglior *BIC* di colonna e in grassetto è indicato il modello con *BIC* minore in assoluto.

Come suggeriscono Vermunt (2003, 2007) e Vermunt e Madigson (2005), tra tutti, si sceglie il modello che presenta il minimo valore di

BIC.

Il modello che presenta il minimo valore di *BIC* risulta essere quello con quattro *cluster* e quattro *gruppi*. Osservando i profili dei cluster e dei gruppi, si nota che gli indicatori *"LUN_TOT"*, *"PICCO"* e *"PERIOD"* non discriminano bene tra le quattro classi e tra i quattro gruppi. Analizzandone le numerosità delle classi, inoltre, le ampiezze della terza e della quarta classe sono abbastanza esigue.

Il modello che presenta il minimo valore di *BIC* subito dopo quello a quattro *cluster* e quattro *gruppi* risulta essere quello a tre *cluster* e a tre *gruppi*. Per questo modello si possono fare le stesse osservazioni fatte per il modello precedente: gli indicatori non discriminano bene tra le classi e le ampiezze delle classi sono esigue.

Il modello che presenta il minimo valore di *BIC* subito dopo quello a tre *cluster* e tre *gruppi* è quello a quattro *cluster* e tre *gruppi*. Gli indicatori sono tutti significativi e sembrano discriminare abbastanza bene tra le classi e tra i gruppi. Il valore del *BIC* comunque non è di molto maggiore rispetto al valore del *BIC* più basso, e risulta notevolmente inferiore rispetto ai valori associati a tutti gli altri modelli. Pertanto la scelta sembrerebbe essere soddisfacente.

5.3 INDIVIDUAZIONE DEI SEGMENTI

Le classi latenti a livello gruppo hanno un'ampiezza tale da non metterne in dubbio la consistenza. Come è possibile notare dalla *Tabella 5.2,* la prima comprende il 49,53%, la seconda il 28,76%, la terza il 13,62% e la quarta l' 8,09%.

Tabella 5.2 "Probabilità totali"

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster Size	*0,4953*	*0,2876*	*0,1362*	*0,0809*

Si prosegue con la profilazione dei segmenti di cicli mestruali.

5.3.1 SEGMENTI SULLA BASE DEGLI INDICATORI

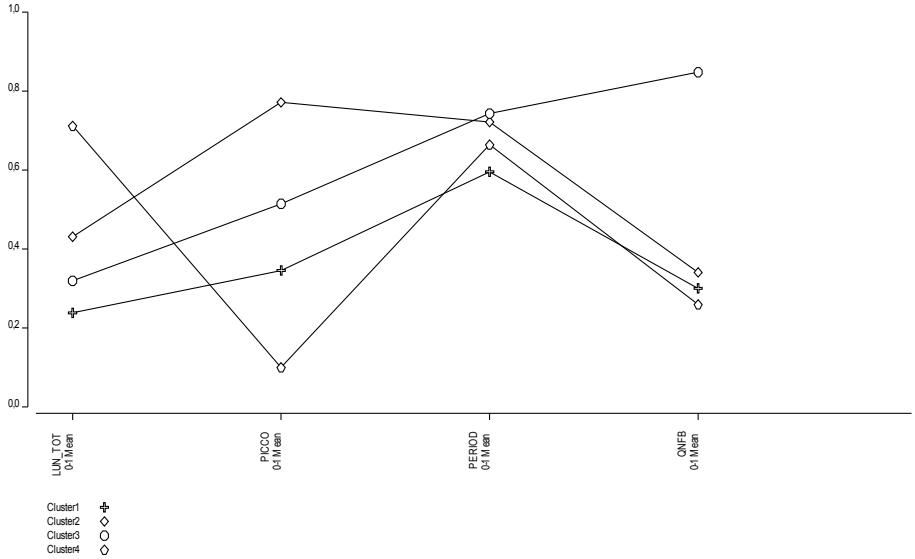
Nella *Tabella 5.3* sono riportate le percentuali per ogni modalità di tutti gli indicatori.

Tabella 5.3 "Probabilità condizionate sulla base degli indicatori"

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Indicators				
LUN_TOT				
1	*0,3353*	*0,0698*	*0,1739*	*0,0088*
2	*0,6254*	*0,6548*	*0,7168*	*0,2815*
3	*0,0360*	*0,1899*	*0,0913*	*0,2795*
4	*0,0032*	*0,0854*	*0,0180*	*0,4303*
PICCO				
0	*0,1967*	*0,0007*	*0,0438*	*0,7149*
1	*0,5780*	*0,0575*	*0,4130*	*0,2717*
2	*0,2199*	*0,5743*	*0,5039*	*0,0134*
3	*0,0054*	*0,3675*	*0,0394*	*0,0000*
PERIOD				
0	*0,0110*	*0,0020*	*0,0014*	*0,0045*
1	*0,3187*	*0,1368*	*0,1123*	*0,2092*
2	*0,5476*	*0,5611*	*0,5433*	*0,5780*
3	*0,1227*	*0,3001*	*0,3430*	*0,2083*
QNFB				
0	*0,1055*	*0,0122*	*0,0000*	*0,2265*
1	*0,8919*	*0,9610*	*0,0245*	*0,7725*
2	*0,0027*	*0,0267*	*0,4136*	*0,0009*
3	*0,0000*	*0,0001*	*0,5619*	*0,0000*

Il primo cluster è caratterizzato con una buona percentuale da cicli lunghi dai 27 ai 32 giorni, con una buona percentuale da cicli in cui il giorno del picco è identificato prima del 15° giorno del ciclo, in cui la durata della mestruazione è con buona percentuale di 5 o 6 giorni e il quadro non fertile di base è, con un'alta percentuale, muco asciutto.
Il secondo cluster è caratterizzato con una buona percentuale da cicli lunghi dai 27 ai 32 giorni, con una buona percentuale da cicli in cui il giorno del picco è identificato tra il 16° e il 19° giorno del ciclo, in cui la durata della mestruazione è con una buona percentuale di 5 o 6 giorni e il quadro non fertile di base è, con un'alta percentuale, muco asciutto.
Il terzo cluster è caratterizzato, con un'alta percentuale, da cicli lunghi tra i 27 e i 32 giorni, con una buona percentuale da cicli in cui il giorno del picco è identificato prima del 19° giorno del ciclo, in cui la durata della mestruazione è con una buona percentuale di 5 o 6 giorni e il quadro non fertile di base è caratterizzato, con un'alta percentuale, da muco immutabile.
Il quarto cluster è caratterizzato, con buone percentuali, da cicli di gravidanze, con una buona percentuale da cicli in cui il giorno del picco non è stato identificato, in cui la durata della mestruazione è con una buona percentuale di 5 o 6 giorni e il quadro non fertile di base è caratterizzato, con un'alta percentuale, da muco asciutto.
Osservando il *profile plot* (*Grafico 5.1*) si possono ricavare informazioni più dettagliate sul profilo dei cluster.

Grafico 5.1 "Profile plot"



Osservando il *Grafico 5.1* si evince che nel primo cluster la lunghezza del ciclo è in media di 25 giorni, il giorno del picco è identificato in media il 10° giorno del ciclo, la durata della mestruazione è in media di 5 giorni e il quadro non fertile di base è muco asciutto; nel secondo cluster la lunghezza del ciclo è in media di 30 giorni, il giorno del picco è identificato in media il 20° giorno del ciclo mestruale, la durata della mestruazione è in media di 5 giorni e il quadro non fertile di base è muco asciutto; nel terzo cluster il ciclo è lungo mediamente 29 giorni, il giorno del picco è identificato in media il 16° giorno del ciclo mestruale, la durata della mestruazione è in media di 6 giorni e il quadro non fertile di base è muco immutabile; infine, fanno parte del quarto cluster le gravidanze, il giorno del picco non è identificato, la durata della mestruazione è in media di 6 giorni e il quadro non fertile di base è muco asciutto.

Per identificare in modo più immediato i quattro segmenti possiamo descrivere loro con poche parole: fanno parte del primo cluster cicli di breve durata; fanno parte del secondo cluster cicli di media durata e con un quadro non fertile di base caratterizzato da muco asciutto; fanno parte del terzo cluster cicli di media durata e con un quadro non

fertile di base caratterizzato da muco immutabile; e fanno parte del quarto cluster le gravidanze.

Rispetto ai segmenti individuati con il modello a classi latenti tradizionali, si può affermare che il primo segmento individuato con il modello a classi latenti multilivello ha caratteristiche molto simili al primo segmento individuato con il modello a classi latenti classico, il secondo segmento individuato con il modello a classi latenti multilivello ha caratteristiche molto simili al secondo segmento individuato con il modello a classi latenti tradizionale, il quarto cluster individuato con il modello a classi latenti multilivello ha caratteristiche che sembrano simili agli ultimi due segmenti individuati con il modello a classi latenti tradizionale. Il terzo segmento individuato con il modello a classi latenti multilivello è differente da tutti i segmenti individuati con il modello tradizionale, infatti il quadro non fertile di base del ciclo di questo segmento è muco immutabile, e, nessuno dei sei segmenti individuati nel *Capitolo 4* era caratterizzato da cicli con lo stesso quadro non fertile di base, al contrario erano tutti caratterizzati da quadro non fertile di base muco asciutto. I segmenti terzo e quarto individuati con il modello tradizionale erano caratterizzati da cicli lunghi, mentre con il modello su due livelli non sono stati individuati segmenti con cicli simili.

5.4 I GRUPPI

Il primo gruppo ha una ampiezza pari al 53,32%, il secondo pari al 32,79% e il terzo pari al 13,89%. (*Tabella 5.4*).

Tabella 5.4 "Ampiezze totali dei tre gruppi"

	Gclass 1	Gclass 2	Gclass 3
Gclass Size	*0,5332*	*0,3279*	*0,1389*

Il primo gruppo è formato da quelle donne che hanno un ciclo breve; il secondo gruppo da quelle donne che hanno un ciclo con un quadro non fertile di base caratterizzato da muco asciutto, e con percentuale molto più bassa da donne in gravidanza; il terzo gruppo invece da quelle donne che hanno un ciclo di media durata e con un quadro non fertile di base caratterizzato da muco immutabile. (*Tabella 5.5*)
Nella *Tabella 5.5* sono riportate le probabilità per ogni gruppo di appartenere ai cluster.

Tabella 5.5 "Probabilità di ogni gruppo di appartenere ad un determinato cluster"

	Gclass 1	Gclass 2	Gclass 3
Clusters			
Cluster 1	*0,9118*	*0,0270*	*0,0024*
Cluster 2	*0,0213*	*0,8392*	*0,0075*
Cluster 3	*0,0006*	*0,0052*	*0,9657*
Cluster 4	*0,0663*	*0,1285*	*0,0244*

Le celle in cui sono presenti i valori sottolineati rappresentano le probabilità maggiori di 0,10. Quindi, come già affermato, prendendo in considerazione tali probabilità, al primo gruppo appartiene il primo cluster, al secondo gruppo i cluster secondo e quarto e al terzo gruppo il terzo cluster.
Il primo e il terzo gruppo appartengono entrambi in modo univoco ad un segmento, cosa che invece non avviene per il secondo gruppo.

Tabella 5.6 "Probabilità in base agli indicatori"

	Gclass 1	Gclass 2	Gclass 3
Indicators			
LUN_TOT			
1	*0,3079*	*0,0696*	*0,1695*
2	*0,6033*	*0,6064*	*0,7055*
3	*0,0555*	*0,1968*	*0,0965*
4	*0,0333*	*0,1272*	*0,0285*
PICCO			
0	*0,2268*	*0,0981*	*0,0602*
1	*0,5465*	*0,1009*	*0,4072*
2	*0,2140*	*0,4923*	*0,4918*
3	*0,0128*	*0,3088*	*0,0408*
PERIOD			
0	*0,0104*	*0,0025*	*0,0015*
1	*0,3074*	*0,1509*	*0,1153*
2	*0,5499*	*0,5628*	*0,5443*
3	*0,1323*	*0,2837*	*0,2206*
QNFB			
0	*0,1114*	*0,0422*	*0,0059*
1	*0,8849*	*0,9300*	*0,0518*
2	*0,0033*	*0,0248*	*0,3997*
3	*0,0003*	*0,0030*	*0,5427*

Osservando la *Tabella 5.6,* è possibile affermare che fanno parte del primo gruppo le donne che hanno un ciclo lungo, con buone percentuali, da 27 a 32 giorni, in cui, con buone percentuali, il giorno del picco del muco è stato identificato prima del 16° giorno, la durata della mestruazione è con buone percentuali di 5 o 6 giorni, il quadro non fertile di base è con un'alta percentuale muco asciutto.

Fanno parte del secondo gruppo le donne che hanno un ciclo in cui, con buone percentuali, la durata è compresa tra i 27 e i 32 giorni, in cui il giorno del picco è stato identificato, con buone percentuali, dopo

il 16° giorno, la durata della mestruazione è di 5 o 6 giorni, il quadro non fertile di base è muco asciutto con un'alta percentuale.
Fanno parte del terzo gruppo le donne che hanno un ciclo lungo, con un'alta percentuale, tra i 27 e i 32 giorni, in cui, con buone probabilità, il giorno del picco è identificato prima del 19° giorno, la durata della mestruazione è di 5 o 6 giorni, il quadro non fertile di base è muco immutabile.

5.5 ANALISI DELL'EFFICIENZA E DELL'EFFICACIA DEI SEGMENTI

Il fatto di aver suddiviso le donne in gruppi e aver assegnato loro dei segmenti di cicli, aumenta l'eterogeneità tra i segmenti rispetto al modello classico.
Le basi della segmentazione, di poco invariate rispetto al modello classico, sono pertinenti all'obiettivo di segmentazione di cicli mestruali; quindi, i segmenti sono efficaci.
L'ampiezza dei segmenti, anche di quelli meno estesi, ne garantiscono la profittabilità.
Inoltre, le basi della segmentazione sono misurabili e mantengono sempre queste caratteristiche.
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, quindi prima di posizionarsi e aggredire un segmento si fanno delle analisi ulteriori che potrebbero chiarire eventuali dubbi.
Anche l'efficacia dei segmenti individuati sembra, quindi, garantita.

5.6 SEGMENTI SULLA BASE DEGLI INDICATORI CONFRONTATI CON IL MODELLO A CLASSI LATENTI CLASSICO

Con il modello a classi latenti tradizionale (*Capitolo 4*) il numero di segmenti individuati per i cicli erano sei, mentre il numero di segmenti individuati per le donne erano quattro. Si vuole adesso andare a verificare se, stimando il modello a due livelli con sei *cluster* e quattro *gruppi*, i profili individuati con il modello a classi latenti multilivello confermino i profili individuati con il modello a classi latenti tradizionali.

Tabella 5.7 "Probabilità condizionate sulla base degli indicatori per il modello a 6 cluster e 4 gruppi"

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Cluster Size	*0,4495*	*0,2850*	*0,0945*	*0,0788*	*0,0587*	*0,0335*
Indicators						
LUN_TOT						
1	*0,3587*	*0,0094*	*0,0010*	*0,0097*	*0,5844*	*0,2977*
2	*0,6371*	*0,7224*	*0,2982*	*0,7265*	*0,4145*	*0,6963*
3	*0,0042*	*0,2031*	*0,3170*	*0,2007*	*0,0011*	*0,0060*
4	*0,0000*	*0,0650*	*0,3837*	*0,0631*	*0,0000*	*0,0001*
PICCO						
0	*0,1341*	*0,0002*	*0,6771*	*0,0037*	*0,1526*	*0,9513*
1	*0,6459*	*0,0426*	*0,3128*	*0,1930*	*0,6512*	*0,0486*
2	*0,2172*	*0,6159*	*0,0101*	*0,7047*	*0,1940*	*0,0002*
3	*0,0028*	*0,3413*	*0,0000*	*0,0986*	*0,0022*	*0,0000*
PERIOD						
0	*0,0006*	*0,0003*	*0,0005*	*0,0003*	*0,0000*	*0,2010*

1	*0,2668*	*0,1773*	*0,2292*	*0,1918*	*0,0396*	*0,7832*
2	*0,5785*	*0,5852*	*0,5862*	*0,5871*	*0,4181*	*0,0158*
3	*0,1541*	*0,2372*	*0,1841*	*0,2207*	*0,5423*	*0,0000*
QNFB						
0	*0,1063*	*0,0131*	*0,2467*	*0,0000*	*0,0000*	*0,0121*
1	*0,8923*	*0,9731*	*0,7528*	*0,0005*	*0,0225*	*0,9729*
2	*0,0014*	*0,0138*	*0,0004*	*0,2050*	*0,7405*	*0,0150*
3	*0,0000*	*0,0000*	*0,0000*	*0,7945*	*0,2370*	*0,0000*
QUALIFI						
1	*0,9999*	*0,9999*	*0,7700*	*0,9997*	*0,9522*	*0,4544*
2	*0,0001*	*0,0001*	*0,1905*	*0,0003*	*0,0459*	*0,3367*
3	*0,0000*	*0,0000*	*0,0395*	*0,0000*	*0,0019*	*0,2089*

Il primo segmento comprende il 44,95%, il secondo il 28,50%, il terzo il 9,45%, il quarto il 7,88%, il quinto il 5,87% e il sesto il 3,35%. Le ampiezze dei segmenti sono quindi molto simili alle ampiezze dei segmenti del modello a classi latenti tradizionale.

Nello specifico, il primo cluster è caratterizzato con una buona percentuale da cicli lunghi tra i 27 e i 32 giorni, in cui il picco è identificato con una buona percentuale prima del 16° giorno del ciclo, la durata della mestruazione è, con una buona percentuale, di 5 o 6 giorni, il quadro non fertile di base è, con un'alta percentuale, muco asciutto e si hanno informazioni complete sulle caratteristiche del muco con una percentuale altissima.

Il secondo cluster è caratterizzato con un'alta percentuale da cicli lunghi tra i 27 e i 32 giorni, in cui il picco è identificato con una buona percentuale tra il 16° e il 19° giorno del ciclo, la durata della mestruazione è, con una buona percentuale, di 5 o 6 giorni, il quadro non fertile di base è, con un'alta percentuale, muco asciutto e si hanno informazioni complete sulle caratteristiche del muco con una percentuale altissima.

Il terzo cluster è caratterizzato con buone percentuali da cicli lunghi

più di 32 giorni, o da gravidanze, in cui il picco, con una buona percentuale, non è identificato, la durata della mestruazione è, con una buona percentuale, di 5 o 6 giorni, il quadro non fertile di base è, con un'alta percentuale, muco asciutto e si hanno informazioni complete sulle caratteristiche del muco con una percentuale alta.

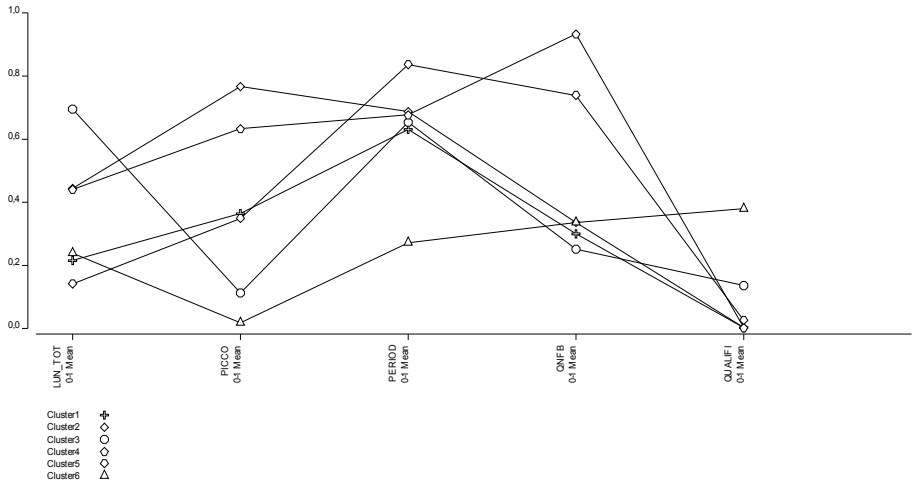
Il quarto cluster è caratterizzato con un'alta percentuale da cicli lunghi tra i 27 e i 32 giorni, in cui il picco è identificato con un'alta percentuale tra il 16° e il 19° giorno del ciclo, la durata della mestruazione è, con una buona percentuale, di 5 o 6 giorni, il quadro non fertile di base è, con un'alta percentuale, muco immutabile e si hanno informazioni complete sulle caratteristiche del muco con una percentuale altissima.

Il quinto cluster è caratterizzato con una buona percentuale da cicli lunghi meno di 26 giorni, in cui il picco è identificato con una buona percentuale prima del 16° giorno del ciclo, la durata della mestruazione è, con una buona percentuale, maggiore o uguale a 7 giorni, il quadro non fertile di base è, con un'alta percentuale, muco immutabile e si hanno informazioni complete sulle caratteristiche del muco con una percentuale altissima.

Infine, il sesto cluster è caratterizzato con una buona percentuale da cicli lunghi tra i 27 e i 32 giorni, con un'alta percentuale il giorno del picco del muco non è identificato, la durata della mestruazione è, con un'alta percentuale, inferiore o uguale a 4 giorni, il quadro non fertile di base è, con un'alta percentuale, muco asciutto e con buone percentuali si hanno informazioni complete sul muco, oppure si hanno dati mancanti dovuti a problemi nella donna, come per esempio stress, che non hanno permesso di identificare il picco del muco.

Osserviamo adesso il *profile plot* per avere informazioni di profilazione più dettagliate.

Grafico 5.2 "Profile plot del modello a 6 cluster e 4 gruppi"



Il primo cluster è caratterizzato da cicli con una lunghezza media di 25 giorni, in cui il giorno del picco è identificato in media l'11° giorno del ciclo, la durata della mestruazione è mediamente di 5 giorni, il quadro non fertile di base è muco asciutto e si hanno informazioni complete sulle caratteristiche del muco; il secondo cluster è caratterizzato da cicli con una lunghezza media di 30 giorni, in cui il giorno del picco è identificato in media il 20° giorno del ciclo, la durata della mestruazione è mediamente di 6 giorni, il quadro non fertile di base è muco asciutto e si hanno informazioni complete sulle caratteristiche del muco; il terzo cluster è caratterizzato da cicli con una lunghezza media di 36 giorni, in cui il giorno del picco non è identificato, la durata della mestruazione è mediamente di 5 giorni, il quadro non fertile di base è muco asciutto e si hanno informazioni complete sulle caratteristiche del muco; il quarto cluster è caratterizzato da cicli con una lunghezza media di 30 giorni, in cui il giorno del picco è identificato in media il 17° giorno del ciclo, la durata della mestruazione è mediamente di 6 giorni, il quadro non fertile di base è muco immutabile e si hanno informazioni complete sulle caratteristiche del muco; il quinto cluster è caratterizzato da cicli con una lunghezza media di 21 giorni, in cui il giorno del picco è

identificato in media il 10° giorno del ciclo, la durata della mestruazione è mediamente di 8 giorni, il quadro non fertile di base è muco immutabile e si hanno informazioni complete sulle caratteristiche del muco; infine, il sesto cluster è caratterizzato da cicli con una lunghezza media di 26 giorni, in cui il giorno del picco non è identificato, la durata della mestruazione è mediamente di 2 giorni, il quadro non fertile di base è muco asciutto e si hanno problemi che non permettono di identificare il picco.

Rispetto ai profili individuati con il modello a classi latenti tradizionali, le caratteristiche dei cicli sono confermate solo per tre classi. Infatti le prime due classi del modello a classi latenti con due livelli sono molto simili, rispettivamente, alle prime due classi del modello a classi latenti tradizionale; la terza classe del modello a classi latenti con due livelli è molto simile alla quarta classe del modello a classi latenti classico; le ultime tre classi invece non sono simili a nessuna delle classi individuate con il modello tradizionale. Nel modello a classi latenti con due livelli si ritrovano due classi caratterizzate da cicli con un quadro non fertile di base caratterizzato da muco immutabile, cosa che non è stata riscontrata nel modello a classi latenti tradizionale, in cui il quadro non fertile di base era con buone percentuali caratterizzato da muco asciutto per tutti e sei i segmenti, e questo aveva portato a dubitare sulla significatività di quell'indicatore. Inoltre, nel modello a classi latenti tradizionale si era individuato un segmento che comprendesse i cicli delle gravidanze, mentre non è stato individuato nel modello a classi latenti su due livelli.

5.7 SEGMENTI SULLA BASE DELLE COVARIATE

Per poter individuare gruppi omogenei di donne in base alle caratteristiche del ciclo mestruale, si è provato ad introdurre le covariate relative alle donne per poter descrivere il modello.

Si sono inserite, quindi, le variabili *"ETA_DONNA"*, *"ETA_UOMO"*, *"GRAV_PRE"* e *"GRAV_CON"* tra le covariate.

Le covariate sono tutte non significative, quindi si distribuiscono in modo uniforme nei gruppi.

CAPITOLO 6: *"Un modello a classi latenti multilivello per la segmentazione del muco cervicale"*

6.1 INTRODUZIONE

Come si era detto nel *Capitolo 3,* un obiettivo di questo lavoro è andare a vedere le caratteristiche del muco nel giorno del picco e come si evolve nei giorni precedenti e nei giorni successivi al picco.

Le caratteristiche del muco non sono state introdotte nelle analisi precedenti, in quanto, utilizzate insieme alle caratteristiche del ciclo mestruale, risultavano essere non significative.

Poiché i dati rappresentano una struttura gerarchica a due livelli, si raggruppa il muco per donna. Per evitare instabilità nelle stime, viene adottato l'approccio a effetti casuali.

Si distinguono le unità in due gruppi: le unità di livello uno sono le caratteristiche del muco, le unità di livello due sono le donne.

L'obiettivo che ci si pone è di classificare le unità di primo livello in classi latenti e di classificare le donne in gruppi, secondo la loro tendenza ad appartenere ai cluster individuati al primo livello.

Come indicatori vengono utilizzate le variabili *"M-3"*, *"M-2"*, *"M-1"*, *"M0"*, *"M+1"*, *"M+2"* e *"M+3"*, che rappresentano, rispettivamente, la tipologia di muco tre, due e un giorno prima del giorno del picco, il giorno del picco, uno, due e tre giorni dopo del giorno del picco.

6.2 STIMA DEL MODELLO

Come per le analisi fatte nei *Capitoli precedenti,* si sono stimati numerosi modelli con l'algoritmo *EM*, 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) e dei gruppi (livello due). Per ogni tipologia si è scelto il migliore per adattamento.

Nella *Tabella 6.1* viene rappresentato un riepilogo dei risultati, che contiene il valore della statistica *BIC* del modello.

Tabella 6.1 "Valori di BIC dei modelli multilivello stimati per diverse combinazioni di cluster e gruppi"

	Gruppi 1	Gruppi 2	Gruppi 3	Gruppi 4	Gruppi 5
Cluster 1	*3928,1164*	3933,4777	3938,8390	3944,2003	3949,5616
Cluster 2	3725,9819	*3679,7742*	3690,5128	3701,2194	3711,9494
Cluster 3	3665,0051	*3587,2408*	3603,2386	3619,3902	3635,4396
Cluster 4	3643,4350	*3553,6528*	3572,8339	3593,3675	3614,8076
Cluster 5	3630,2005	***3541,5887***	3563,8698	3589,7260	3616,0904
Cluster 6	3634,8826	*3554,4018*	3579,8251	3611,6598	3641,7546

Le celle con i valori in corsivo indicano i modelli con il miglior *BIC* di riga, le celle con i valori sottolineati indicano i modelli con il miglior *BIC* di colonna e in grassetto è indicato il modello con *BIC* minore in assoluto.

Il modello che presenta il minimo valore di *BIC* è quello con cinque cluster e due gruppi. Il valore del *BIC* è notevolmente inferiore rispetto ai valori associati a tutti gli altri modelli e, osservando il profilo dei segmenti e dei gruppi, tutti gli indicatori sono significativi e sembrano discriminare bene tra le classi e tra i gruppi. Pertanto la scelta sembrerebbe essere soddisfacente.

6.3 I SEGMENTI

Come è possibile notare dalla *Tabella 6.2*, la prima classe ha un'ampiezza del 48,74%, la seconda del 21,84%, la terza del 15,05%, la quarta dell'11,07% e la quinta del 3,30%.

Tabella 6.2 "Probabilità totali"

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster size	*0,4874*	*0,2184*	*0,1505*	*0,1107*	*0,0330*

Nella *Tabella 6.3* sono riportate le percentuali per ogni modalità di tutti gli indicatori.

Tabella 6.3 "Probabilità condizionate sulla base degli indicatori[8]"

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Indicators					
M-3					
0	*0,0000*	*0,0085*	*0,0042*	*0,1037*	*0,0466*
1	*0,0008*	*0,0925*	*0,0568*	*0,3891*	*0,2628*
2	*0,0067*	*0,1474*	*0,1137*	*0,2149*	*0,2180*
3	*0,1222*	*0,5317*	*0,5158*	*0,2688*	*0,4096*
4	*0,1021*	*0,0882*	*0,1076*	*0,0155*	*0,0354*
5	*0,7681*	*0,1317*	*0,2019*	*0,0080*	*0,0275*
M-2					
0	*0,0000*	*0,0063*	*0,0060*	*0,0968*	*0,4020*
1	*0,0000*	*0,0649*	*0,0629*	*0,3455*	*0,4531*
2	*0,0001*	*0,1503*	*0,1475*	*0,2755*	*0,1141*

8 Si ricorda la codifica del muco, descritta nel *Capitolo 3*: 0=nessuna informazione sulla sensazione e nessuna informazione sull'aspetto del muco; 1=nessuna sensazione oppure sensazione di asciutto, assenza di muco e perdita oppure perdita inconsistente; 2=sensazione di non più asciutto, assenza di muco e perdita oppure perdita inconsistente; 3=sensazione di umido, muco denso, cremoso, biancastro, giallastro, appiccicoso, filoso; 4=sensazione di bagnato, liquido; 5=sensazione di bagnato-lubrificato, scivolosità, muco trasparente, filante, liquido, acquoso, tracce di sangue.

3	*0,0044*	*0,3476*	*0,3459*	*0,2194*	*0,0287*
4	*0,0502*	*0,2141*	*0,2159*	*0,0465*	*0,0019*
5	*0,9453*	*0,2169*	*0,2218*	*0,0162*	*0,0002*
M-1					
0	*0,0000*	*0,0001*	*0,0003*	*0,1030*	*0,4556*
1	*0,0000*	*0,0030*	*0,0110*	*0,4581*	*0,4734*
2	*0,0005*	*0,0186*	*0,0466*	*0,2546*	*0,0615*
3	*0,0136*	*0,1343*	*0,2288*	*0,1640*	*0,0093*
4	*0,0442*	*0,1214*	*0,1411*	*0,0132*	*0,0002*
5	*0,9416*	*0,7226*	*0,5723*	*0,0070*	*0,0000*
M0					
0	*0,0000*	*0,0000*	*0,0029*	*0,0400*	*0,3345*
1	*0,0008*	*0,0000*	*0,0263*	*0,1713*	*0,4157*
2	*0,0068*	*0,0007*	*0,0874*	*0,2668*	*0,1878*
3	*0,0394*	*0,0086*	*0,1881*	*0,0694*	*0,0550*
4	*0,0528*	*0,0255*	*0,0943*	*0,0633*	*0,0038*
5	*0,9002*	*0,9652*	*0,6010*	*0,1892*	*0,0033*
M+1					
0	*0,0180*	*0,0382*	*0,0003*	*0,0040*	*0,4966*
1	*0,1616*	*0,2500*	*0,0086*	*0,0608*	*0,4352*
2	*0,2062*	*0,2330*	*0,0392*	*0,1313*	*0,0543*
3	*0,5341*	*0,4409*	*0,3621*	*0,5763*	*0,0138*
4	*0,0163*	*0,0099*	*0,0395*	*0,0299*	*0,0000*
5	*0,0639*	*0,0281*	*0,5503*	*0,1977*	*0,0000*
M+2					
0	*0,0064*	*0,0112*	*0,0000*	*0,0005*	*0,4427*
1	*0,2594*	*0,3540*	*0,0022*	*0,0507*	*0,5436*
2	*0,2026*	*0,2136*	*0,0131*	*0,0994*	*0,0128*
3	*0,4833*	*0,3939*	*0,2383*	*0,5948*	*0,0009*
4	*0,0258*	*0,0162*	*0,0970*	*0,0795*	*0,0000*
5	*0,0226*	*0,0110*	*0,6494*	*0,1751*	*0,0000*
M+3					
0	*0,0056*	*0,0235*	*0,0001*	*0,0035*	*0,1886*
1	*0,2577*	*0,5286*	*0,0122*	*0,1976*	*0,7514*

2	*0,1949*	*0,1946*	*0,0428*	*0,1805*	*0,0490*
3	*0,5085*	*0,2473*	*0,5183*	*0,5689*	*0,0110*
4	*0,0177*	*0,0042*	*0,0838*	*0,0239*	*0,0000*
5	*0,0156*	*0,0018*	*0,3429*	*0,0255*	*0,0000*

Il primo cluster è caratterizzato con alte percentuali da una sensazione di liquido scivoloso, muco trasparente, viscoso, liquido, acquoso e tracce di sangue nei tre giorni precedenti al giorno del picco e nel giorno del picco, mentre è caratterizzato, con buone percentuali, da una sensazione di umido, muco spesso, cremoso, biancastro, giallastro, appiccicoso e filante nei tre giorni successivi al giorno del picco.

Il secondo cluster è caratterizzato, con buone percentuali, da una sensazione di umido, muco spesso, cremoso, biancastro, giallastro, appiccicoso e filante tre e due giorni prima del giorno del picco, con alte percentuali da una sensazione di liquido scivoloso, muco trasparente, viscoso, liquido, acquoso e tracce di sangue il giorno precedente al giorno del picco e il giorno del picco, con buone percentuali da una sensazione di umido, muco spesso, cremoso, biancastro, giallastro, appiccicoso e filante nei due giorni successivi al giorno del picco, e con buona percentuale il terzo giorno successivo al giorno del picco non si ha alcuna sensazione o al più una sensazione di asciutto.

Il terzo cluster è caratterizzato con buone percentuali da una sensazione di umido, muco spesso, cremoso, biancastro, giallastro, appiccicoso e filante tre e due giorni prima del giorno del picco, con buone percentuali da una sensazione di liquido scivoloso, muco trasparente, viscoso, liquido, acquoso e tracce di sangue il giorno precedente al giorno del picco, il giorno del picco e i due giorni successivi al giorno del picco, mentre il terzo giorno dopo il giorno del picco da una sensazione di umido, muco spesso, cremoso, biancastro,

giallastro, appiccicoso e filante.

Il quarto cluster è caratterizzato, nei giorni precedenti al giorno del picco, con discrete percentuali, da nessuna sensazione o al più sensazione di asciutto, oppure da niente muco, niente perdite e niente perdite inconsistenti, nel giorno del picco con discrete percentuali da niente muco, niente perdite e niente perdite inconsistenti, oppure da una sensazione di umido, muco spesso, cremoso, biancastro, giallastro, appiccicoso e filante, nei giorni successivi al giorno del picco è caratterizzato invece, con buone percentuali, da una sensazione di umido, muco spesso, cremoso, biancastro, giallastro, appiccicoso e filante.

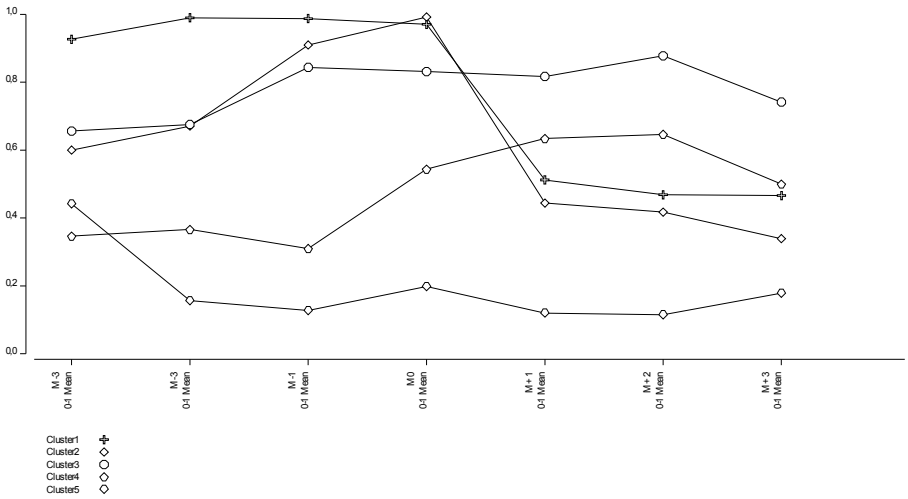
Infine, il quinto cluster è caratterizzato, tre giorni prima del giorno del picco, con una discreta percentuale da una sensazione di umido, muco spesso, cremoso, biancastro, giallastro, appiccicoso e filante, mentre nei due giorni precedenti al giorno del picco, nel giorno del picco e nei tre giorni successivi al giorno del picco è caratterizzato, con buone percentuali, da nessuna sensazione o al più sensazione di asciutto oppure non si hanno proprio informazioni.

Per semplicità possiamo descrivere il primo cluster come muco trasparente e acquoso fino al giorno del picco e spesso e cremoso nei giorni successivi; il secondo cluster come muco trasparente e acquoso il giorno del picco e il giorno precedente, come muco spesso e cremoso nei due giorni precedenti e nei due giorni successivi e con nessuna sensazione nel terzo giorno successivo; il terzo cluster come muco trasparente e acquoso nel giorno del picco, nel giorno precedente e nei due giorni successivi e come muco spesso e cremoso negli altri giorni; il quarto cluster come niente muco o nessuna sensazione nei giorni precedenti al giorno del picco, nessun muco, o al più muco cremoso nel giorno del picco e come muco cremoso nei giorni successivi al picco; infine il quinto cluster come muco cremoso tre giorni prima del giorno del picco e come nessuna sensazione o

nessuna informazione negli altri giorni.

Queste caratteristiche si possono osservare anche dal *profile plot* (*Grafico 6.1*).

Grafico 6.1 "Profile plot"



6.4 I GRUPPI

Il primo gruppo ha un'ampiezza pari al 70,17% e il secondo gruppo pari al 29,83%. (*Tabella 6.4*)

Tabella 6.4 "Ampiezze totali dei due gruppi"

	Gclass 1	Gclass 2
Gclass size	*0,7017*	*0,2983*

Il primo gruppo è formato da quelle donne che hanno un muco trasparente, viscoso, liquido, acquoso, tracce di sangue e una sensazione di liquido scivoloso nel giorno del picco e nei giorni immediatamente vicini al giorno del picco, e negli altri giorni subito prima o subito dopo hanno un muco spesso, cremoso, biancastro, giallastro, appiccicoso, filante e una sensazione di umido.

Il secondo gruppo, invece, è caratterizzato da donne che hanno un muco sia trasparente e acquoso, sia spesso e cremoso nei giorni del picco e nei giorni immediatamente precedenti e immediatamente successivi al giorno del picco, oppure non hanno nessuna sensazione, o al più una sensazione di asciutto, oppure nessun muco, oppure non hanno proprio alcuna informazione.
Nella *Tabella 6.5* sono riportate le probabilità per ogni gruppo di appartenere ai cluster.

Tabella 6.5 "Probabilità di ogni gruppo di appartenere ad un determinato cluster"

	Gclass 1	Gclass 2
Cluster 1	*0,6802*	*0,0341*
Cluster 2	*0,3102*	*0,0023*
Cluster 3	*0,0010*	*0,5018*
Cluster 4	*0,0079*	*0,3526*
Cluster 5	*0,0006*	*0,1092*

Le celle in cui sono presenti i valori sottolineati rappresentano le probabilità maggiori di 0,10. Quindi, prendendo in considerazione tali probabilità, al primo gruppo appartengono i primi due cluster e al secondo gruppo appartengono gli ultimi tre cluster.
Entrambi i gruppi non appartengono in modo univoco ad un segmento.

Tabella 6.6 "Probabilità in base agli indicatori"

	Gclass 1	Gclass 2
Indicators		
M-3		
0	*0,0035*	*0,0438*
1	*0,0326*	*0,1946*
2	*0,0522*	*0,1572*

3	*0,2510*	*0,4037*
4	*0,0971*	*0,0670*
5	*0,5636*	*0,1336*
M-2		
0	*0,0030*	*0,0811*
1	*0,0232*	*0,2030*
2	*0,0491*	*0,1840*
3	*0,1129*	*0,2550*
4	*0,1011*	*0,1272*
5	*0,7107*	*0,1498*
M-1		
0	*0,0011*	*0,0862*
1	*0,0049*	*0,2188*
2	*0,0082*	*0,1199*
3	*0,0525*	*0,1744*
4	*0,0680*	*0,0773*
5	*0,8653*	*0,3234*
M0		
0	*0,0006*	*0,0521*
1	*0,0022*	*0,1190*
2	*0,0072*	*0,1586*
3	*0,0318*	*0,1968*
4	*0,0444*	*0,0719*
5	*0,9139*	*0,4016*
M+1		
0	*0,0245*	*0,0565*
1	*0,1882*	*0,0793*
2	*0,2136*	*0,0795*
3	*0,5050*	*0,4056*
4	*0,0144*	*0,0309*
5	*0,0543*	*0,3481*
M+2		
0	*0,0081*	*0,0488*
1	*0,2870*	*0,0880*

2	*0,2049*	*0,0504*
3	*0,4559*	*0,3468*
4	*0,0233*	*0,0776*
5	*0,0208*	*0,3884*
M+3		
0	*0,0112*	*0,0221*
1	*0,3413*	*0,1678*
2	*0,1945*	*0,0976*
3	*0,4276*	*0,4798*
4	*0,0136*	*0,0511*
5	*0,0117*	*0,1816*

Osservando la *Tabella 6.5*, è possibile affermare che fanno parte del primo gruppo le donne che nel giorno del picco hanno, con un'alta percentuale, un muco trasparente, viscoso, liquido, acquoso, tracce di sangue e una sensazione di liquido scivoloso, con alte percentuali anche nei due giorni precedenti al giorno del picco e, con una buona percentuale, tre giorni prima del giorno del picco, mentre nei tre giorni successivi al giorno del picco hanno, con buone percentuali, un muco spesso, cremoso, biancastro, giallastro, appiccicoso, filante e una sensazione di umido.

Fanno parte del secondo gruppo, invece, quelle donne che il giorno del picco hanno, con una discreta percentuale, un muco trasparente, viscoso, liquido, acquoso, tracce di sangue e una sensazione di liquido scivoloso, nei tre giorni successivi al picco hanno, con discrete percentuali, un muco spesso, cremoso, biancastro, giallastro, appiccicoso, filante e una sensazione di umido, o al più, solo nei due giorni successivi, un muco trasparente, viscoso, liquido, acquoso, tracce di sangue e una sensazione di liquido scivoloso, mentre, nei tre giorni precedenti al giorno del picco hanno, con discrete percentuali, un muco spesso, cremoso, biancastro, giallastro appiccicoso, filante e una sensazione di umido, oppure nessun muco, nessuna perdita o

nessuna perdita inconsistente, oppure nessuna sensazione o una sensazione di asciutto.
Poiché le donne sono state suddivise in gruppi e sono stati loro assegnati segmenti di muco, i segmenti sono omogenei. I segmenti sono, inoltre, efficaci ed efficienti.

6.5 SEGMENTI SULLA BASE DELLE COVARIATE

Per poter individuare gruppi omogenei di donne in base alle caratteristiche del muco cervicale, si è provato ad introdurre le covariate relative alle donne per poter descrivere il modello.
Si sono inserite, quindi, le variabili *"ETA_DONNA"*, *"ETA_UOMO"*, *"GRAV_PRE"* e *"GRAV_CON"* tra le covariate.
Queste covariate però sono tutte non significative e questo significa che si distribuiscono in modo uniforme nei gruppi.

CONCLUSIONI

Il metodo di ovulazione *Billings* è un metodo naturale di regolazione della fertilità, che si basa sull'osservazione delle modificazioni del muco cervicale. Le osservazioni del muco vengono annotate su un'apposita cartella, consentendo l'individuazione della fase fertile di un ciclo. Nello studio sono entrati campioni di cicli di donne provenienti da quattro Centri Italiani nei quali si fa riferimento al metodo dell'Ovulazione *Billings*. Ogni centro ha inviato proprie schede relative a ciascun ciclo, le quali consistono nella codifica operata a posteriori da parte di una insegnante del metodo *Billings* delle informazioni registrate dalla donna in base alla propria osservazione. Esse contengono informazioni sul ciclo (inizio, fine, giorni interessati dal flusso mestruale), tipologia di muco osservato giornalmente, picco del muco (se individuato), rapporti sessuali, eventuale gravidanza ed eventuali disturbi che possono aver alterato l'osservazione del muco.

Pur rimanendo le schede strettamente anonime, si raccolgono anche alcune informazioni sulle caratteristiche demografiche della donna e del partner (età, data del matrimonio) e sulla storia ginecologica della donna (gravidanze precedenti, precedente assunzione di contraccettivi ormonali).

Dalle prime analisi descrittive è emerso che lo studio è stato effettuato su donne con età media di 30 anni, con partner con età media di 33 anni. Il numero totale di donne è 193, per un numero totale di spezzoni pari a 244 e un numero totale di 2901 cicli; il numero di gravidanze identificate è invece 162.

Il muco nel giorno di picco è, nell'80,6% dei casi, trasparente e liquido e si avverte una sensazione di liquido scivoloso. Nei giorni precedenti al giorno di picco, il muco si presenta in modo differente: fino a tre giorni prima del picco, ha sempre una sensazione di liquido scivoloso e il muco è trasparente e liquido, mentre andando indietro nei giorni,

la donna non ha più questa sensazione, ma più una sensazione di umido. Dal giorno successivo al giorno di picco, invece, si ha subito principalmente una sensazione di umido, e in percentuale meno significativa non si ha alcuna sensazione, oppure una sensazione di asciutto.

Gli obiettivi principali di questo lavoro sono individuare gruppi omogenei di cicli e di donne e andare a studiare le caratteristiche del giorno di picco e vedere la sua evoluzione nei giorni precedenti e successivi.

A tali scopi, si sono stimati modelli a classi latenti, prima col metodo tradizionale e poi con quello multilivello: i modelli a classi latenti si rendono molto utili quando si presenta la necessità di individuare delle similarità in una popolazione omogenea, infatti sono riconosciuti come metodi di segmentazione (Vermunt e Madgiston, 2002).

L'idea di fondo è riassumere l'eterogeneità osservata sulle unità statistiche, in base a loro caratteristiche latenti. In altre parole, si utilizza una variabile latente, 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, si individuano delle classi latenti a livello due, che raggruppano le unità di secondo livello.

In questo lavoro, si sono individuati, prima di tutto, segmenti di cicli mestruali con il modello a classi latenti tradizionale. Come indicatori (variabili usate come basi per la segmentazione), sono state utilizzate caratteristiche del ciclo, quali la lunghezza, il picco, la durata della

mestruazione, il quadro non fertile di base (QNFB), ovvero la fase non fertile preovulatoria, e le informazioni sul muco. I modelli sono stati stimati più volte, con insiemi di valori iniziali per le procedure di stima differenti (per la stima è stato utilizzato l'algoritmo EM, *Expetation Maximization,* che è un metodo che consente, per mezzo di un procedimento iterativo, di effettuare le stime di massima verosimiglianza dei parametri in presenza di dati incompleti, riconducendo il problema ad un problema standard di stima per dati completi). Il miglior modello, sulla base dell'adattamento (il modello con il minimo valore di *BIC*), è risultato essere quello a sei segmenti: il primo segmento è caratterizzato da cicli di lunghezza breve, con picco identificato e QNFB asciutto; il secondo segmento è caratterizzato da cicli di lunghezza normale, con picco identificato e QNFB asciutto; il terzo segmento è caratterizzato da cicli lunghi, con picco identificato e QNFB asciutto; il quarto segmento è caratterizzato da cicli lunghi, con picco non identificato e QNFB asciutto; il quinto segmento è caratterizzato da cicli di lunghezza normale, con picco non identificato e QNFB asciutto; il sesto segmento è caratterizzato da cicli con gravidanze, con picco identificato e QNFB asciutto, in ordine decrescente per l'ampiezza delle classi.

Per mezzo dell'analisi dei residui bivariati, è emerso che gli indicatori QNFB e durata della mestruazione sono tra loro dipendenti, invalidando l'ipotesi di indipendenza locale del modello a classi latenti classico, e questo è dovuto alla presenza di una struttura gerarchica a due livelli, di cui il modello classico a classi latenti non tiene conto, per cui si è scelto di stimare un modello a classi latenti multilivello, che, considerando anche la gerarchia insita nei dati, porta alla stima di un modello migliore, sotto il profilo dell'adattamento.

Prima di ottenere una segmentazione di cicli, raggruppabili per donna, con il modello a classi latenti multilivello, si sono individuati dei segmenti di donne con il modello a classi latenti tradizionale. Come

indicatori sono stati utilizzati le età della donna e del partner e il numero di gravidanze precedenti. Il miglior modello, sotto il profilo dell'adattamento, è risultato quello a quattro segmenti: il primo segmento è caratterizzato da coppie giovani, che non hanno avuto nessun figlio; il secondo segmento è caratterizzato da coppie mature, in cui le donne hanno già avuto almeno due gravidanze; gli ultimi due segmenti sono caratterizzati entrambi da coppie di età media, ma le donne del terzo cluster prima di entrare nello studio avevano già avuto una o due gravidanze, mentre le donne del quarto cluster tre o più.

Successivamente si è utilizzato il modello a classi latenti multilivello per stimare gruppi di cicli mestruali raggruppati per donna. Nel modello a classi latenti su due livelli, la prima variabile individua delle classi latenti a livello uno (cicli mestruali), la seconda, con procedimento simile, clusterizza le unità di livello due (donne). In modo probabilistico sono assegnate le unità e i gruppi alle classi di primo livello.

Il miglior modello è risultato essere quello a quattro cluster e tre gruppi. Fanno parte del primo cluster cicli di breve durata; fanno parte del secondo cluster cicli di media durata e con QNFB asciutto; fanno parte del terzo cluster cicli di media durata e con QNFB immutabile e fanno parte del quarto cluster le gravidanze. Rispetto ai segmenti individuati con il modello a classi latenti tradizionale, il primo segmento individuato con il modello a classi latenti multilivello ha caratteristiche molto simili al primo segmento individuato con il modello a classi latenti classico, il secondo segmento individuato con il modello a classi latenti multilivello ha caratteristiche molto simili al secondo segmento individuato con il modello a classi latenti tradizionale, il quarto cluster individuato con il modello a classi latenti multilivello ha caratteristiche che sembrano simili agli ultimi due segmenti individuati con il modello a classi latenti tradizionale. Il terzo segmento individuato con il modello a classi latenti multilivello

è differente da tutti i segmenti individuati con il modello tradizionale, infatti ha QNFB immutabile, e, nessuno dei sei segmenti individuati con il modello tradizionale è caratterizzato da cicli con lo stesso QNFB, al contrario erano tutti caratterizzati da QNFB asciutto. I segmenti terzo e quarto individuati con il modello tradizionale erano caratterizzati da cicli lunghi, mentre con il modello su due livelli non sono stati individuati segmenti con cicli simili.

Per quanto riguarda i gruppi, invece, il primo è formato da quelle donne che hanno un ciclo breve; il secondo da quelle donne che hanno un ciclo con QNFB asciutto, e con percentuale molto più bassa da donne in gravidanza; il terzo gruppo invece da quelle donne che hanno un ciclo di media durata e con QNFB immutabile.

Il tipo di segmentazione attuabile con i modelli a classi latenti è a posteriori, nel senso che l'analisi determina il numero e la tipologia dei segmenti individuabili, 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.

Inoltre, si è provato ad introdurre le covariate relative alle donne per poter descrivere il modello. Le covariate sono risultate tutte non significative, quindi si distribuiscono in modo uniforme nei gruppi.

Successivamente si è stimato un modello a classi latenti su due livelli, in cui si è raggruppato il muco per donna. Le unità di livello uno sono i cicli e le unità di livello due sono le donne. Come indicatori sono state utilizzate le variabili che descrivono le caratteristiche del muco da tre giorni prima del picco a tre giorni dopo del picco. Il miglior modello, sulla base dell'adattamento, è risultato essere quello a cinque cluster e due gruppi: il primo cluster è caratterizzato da muco trasparente e acquoso fino al giorno del picco e spesso e cremoso nei giorni successivi; il secondo cluster da muco trasparente e acquoso il giorno del picco e il giorno precedente, da muco spesso e cremoso nei

due giorni precedenti e nei due giorni successivi e da nessuna sensazione nel terzo giorno successivo; il terzo cluster da muco trasparente e acquoso nel giorno del picco, nel giorno precedente e nei due giorni successivi e da muco spesso e cremoso negli altri giorni; il quarto cluster da niente muco o nessuna sensazione nei giorni precedenti al giorno del picco, nessun muco, o al più muco cremoso nel giorno del picco e da muco cremoso nei giorni successivi al picco; infine il quinto cluster da muco cremoso tre giorni prima del giorno del picco e da nessuna sensazione o nessuna informazione negli altri giorni. Il primo gruppo, invece, è formato da quelle donne che hanno un muco trasparente, viscoso, liquido, acquoso, tracce di sangue e una sensazione di liquido scivoloso nel giorno del picco e nei giorni immediatamente vicini al giorno del picco, e negli altri giorni subito prima o subito dopo hanno un muco spesso, cremoso, biancastro, giallastro, appiccicoso, filante e una sensazione di umido; il secondo gruppo è caratterizzato da donne che hanno un muco sia trasparente e acquoso, sia spesso e cremoso nei giorni del picco e nei giorni immediatamente precedenti e immediatamente successivi al giorno del picco, oppure non hanno nessuna sensazione, o al più una sensazione di asciutto, oppure nessun muco, oppure non hanno proprio alcuna informazione.

La segmentazione consiste nel creare dei segmenti omogenei al loro interno ed eterogenei tra di loro. 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.

Nel nostro lavoro la segmentazione si basa sulle caratteristiche biometriche dei cicli. Si sono individuati gruppi di donne sulla base del ciclo mestruale e sulla base delle caratteristiche del muco cervicale.

Per un lavoro futuro, sarebbe interessante continuare un'analisi di

segmentazione delle donne sulla base comportamentale. Inserire nelle analisi alcune informazioni sui comportamenti della donna, ad esempio se la donna ha assunto contraccettivi ormonali nel passato, se la donna ha avuto rapporti non protetti durante il ciclo, se la donna ha avuto problemi di stress, oppure visite ginecologiche, se la donna ha avuto aborti spontanei nel passato e andare a vedere come queste variabili incidono nel ciclo mestruale e nel confermare o meno una gravidanza.

BIBLIOGRAFIA

BASSI F, (2012), "*Analysing markets within the latent class approach: an application to the pharma sector*", Research Article, Applied Stochastic Models in Business and Industry

BASSI F., MION A., COLOMBO B. (2004), "*Interobserver variation in interpreting cervical mucus as an indicator of the fertile phase in a menstrual cycle*", Genus, LIX (No. 3-4), pp 91-102

BRASINI S., TASSINARI F., TASSINARI G., (1996), *"Marketing e pubblicità. Approccio statistico all'analisi dei mercati del consumo"*, Il Mulino, Bologna

BRASINI S., FREO M., TASSINARI F., TASSINARI G., (2010), *"Marketing e pubblicità. Strumenti e modelli di analisi statistica"*, Il Mulino, Bologna

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

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

GOODMAN L. A., (1974), "*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., (1974), "*Exploratory latent structure analysis using both identifiable and unidentifiable models*", Biometrika, 61, pp. 215-231

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

LUKOCIENE O., VARRIALE R., VERMUNT J.K., (2009), *"The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis",* Sociological Methodology, American Sociological Association

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

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

SCARPA B., (2004), "*Sviluppi nelle ricerche sulla fecondabilità*", Gruppo di lavoro su Fertilità, Coordinatore prof. Bernardo Colombo

TRIVELLATO U., (2013), "*La lezione scientifica e umana di Bernardo Colombo*"

VERMUNT J.K., MAGIDSON J., (2003), "*Latent GOLD. User's Guide*", Belmont, MA: Statistical Innovations Inc

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., MAGIDSON J., (2005), "*Latent GOLD 4.0. User's Guide*", Belmont, MA: Statistical Innovations Inc

VERMUNT J.K., MAGIDSON J., (2013), "*Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*", Belmont, MA: Statistical Innovations Inc

VERMUNT J.K., MAGIDSON J., (2013), "Latent GOLD 5.0 Upgrade Manual", Belmont, MA: Statistical Innovations Inc

VERMUNT J.K., (2003), "*Multilevel latent class models"*, Sociological Methodology, 33, pp 213-239

VERMUNT J.K., (2008), *"Multilevel Latent Variable Modeling: An Application in Education Testing"*, Austrian Journal of Statistics, Number 3&4, pp 285-299

RINGRAZIAMENTI

Se sono riuscita a raggiungere la fine del mio percorso di studi devo dire grazie a diverse persone.

Prima di tutto ringrazio lei, Giorgia, la mia meravigliosa bambina, perché se non fosse così brava io probabilmente non sarei mai arrivata a questo punto. Se ho sempre avuto la forza di andare avanti l'ho fatto principalmente per lei e per il suo futuro.

Dopo ringrazio Matteo, il mio compagno di vita, che è molto fiero di me e mi ha sempre incoraggiata ad andare avanti perché anche con l'impegno di mamma ce la potevo fare...e ce l'ho fatta!

Ringrazio i miei splendidi genitori che mi hanno sempre appoggiata in tutte le mie scelte e mi hanno supportata, e spero che possano essere sempre orgogliosi di me. Li ringrazio inoltre per l'aiuto che mi hanno dato facendo i nonni a tempo pieno, finché siamo stati vicini.

Ringrazio anche il mio nonnino per essere un nonno e un bisnonno meraviglioso, perché anche lui, nonostante i suoi 90 anni, mi ha dato una mano per farmi studiare.

Ringrazio poi Alessia, mia sorella, per essere stata sempre presente e disponibile, ed è proprio grazie a lei che ho potuto fare i miei ultimi esami!

Ringrazio anche Lina e Antonio, che sono stati anche loro dei nonni sempre presenti e mi hanno sempre aiutato quando avevo bisogno.

Ringrazio infine la professoressa Francesca Bassi per avermi trasmesso entusiasmo nel fare questa tesi, e per essere stata sempre molto disponibile.