INDICE

PREFAZIONE

Scopo del presente lavoro è l'individuazione di un efficace metodo per la segmentazione della clientela di una compagnia telefonica.

La variabile di classificazione scelta è il *customer value*, ossia una misura che valuta la redditività mensile di ciascun utente. La variabile è ricavata dal data-set aziendale, che racchiude un'insieme di informazioni per un campione di 32.524 utenti. Per ognuno di essi sono disponibili alcune variabili di natura anagrafica e 5 serie storiche relative ai consumi telefonici per il periodo che va da novembre 2004 a marzo 2006. Da quest'ultimo mese è stata ottenuta la variabile di classificazione, moltiplicando i consumi mensili totali di ogni cliente per il costo in euro degli stessi. In un primo momento si suppone che questo dato sia ignoto, in modo da stimarlo con metodi di classificazione, per poi verificare se la suddivisione ottenuta si avvicina a quella nota.

Vengono stimati più modelli previsivi, costruiti combinando gli strumenti dell'analisi di classificazione con lo studio delle serie storiche.

Visto l'elevato numero di serie a disposizione, non sono stati analizzati i grafici di autocorrelazione per identificare tanti modelli diversi, ma si è adattato un modello comune a tutte le serie, con una procedura automatica.

Oltre ai metodi classici di analisi multidimensionale (analisi discriminante, albero di classificazione e modello multinomiale) è stato implementato e ampliato il criterio della distanza *auto-regressiva*, introdotto da Domenico Piccolo nel 1984. In quella ricerca venivano stimati dei modelli stocastici di tipo $ARIMA(p,d,q)$ su serie storiche relative ad alcuni indici di produzione industriale. Queste serie venivano raggruppate rispetto alla similarità della struttura dinamica stimata su ognuna di esse.

In un primo momento è stata ripetuta la stessa metodologia sulle 32.524 serie a disposizione, poi è stata introdotta un'integrazione per tener conto del trend deterministico che caratterizza i consumi di molti utenti.

1. SEGMENTAZIONE COMPORTAMENTALE

La segmentazione comportamentale consiste nella suddivisione del mercato in "segmenti" di clienti, caratterizzati dall'omogeneità per ciò che riguarda le modalità e la cadenza di uso dei prodotti o servizi offerti dall'azienda (CAMILLO, TASSINARI, 2002).

Il raggruppamento della clientela, realizzato impiegando metodologie per l'individuazione di legami non espliciti, rappresenta un importante strumento per il raggiungimento di un vantaggio competitivo sulle imprese concorrenti. Scoprire legami inattesi tra variabili esplicative di diversa natura (economiche e socio-demografiche) e variabile risposta, consente di definire opportune strategie di marketing, che indirizzino l'azione pubblicitaria a determinati soggetti con particolari caratteristiche. L'impiego di questi strumenti permette di razionalizzare le risorse aziendali a disposizione, creando un "profilo" per ciascun cliente, in modo da anticipare e soddisfare le sue esigenze.

Lo studio del comportamento del cliente, definito *customer behaviour*, è il fondamento per la definizione delle strategie di *Customer relationship management* (CRM). Obiettivo fondamentale di un sistema di CRM è la conoscenza dei singoli clienti. Questa conoscenza è espressa da un profilo associato ad ogni cliente, che ne descrive il valore per la società, i gusti, le preferenze e le attitudini.

Il profilo fornisce le informazioni per guidare le attività relazionali fra l'azienda e il cliente in questione (es. promozioni, campagne di informazione, interventi di assistenza).

Il marketing deve quindi pianificare apposite strategie, soprattutto nel caso di un mercato saturo come quello della telefonia mobile.

Al giorno d'oggi, infatti, tutti possiedono un telefono cellulare e l'attenzione delle compagnie telefoniche non è più concentrata sulle metodologie per attirare nuovi utenti, ma sulla ricerca di miglioramenti del servizio per quei clienti che già sono legati all'azienda, tramite un contratto di abbonamento o una scheda ricaricabile. Conoscendo i comportamenti dinamici e le aspettative del soggetto utente si è in grado di soddisfare le sue esigenze e di fidelizzarlo, limitando le possibilità che decida di cambiare operatore.

Per costruire un algoritmo di classificazione, è necessario disporre di un data base contenente le caratteristiche socio-demografiche dei clienti (id, sesso, età, ecc.) e dei dati sul loro traffico telefonico (numero di sms inviati, chiamate, ecc.). Queste variabili sono strettamente connesse al fatturato di una società, che rappresenta una delle componenti principali del suo bilancio.

L'ammontare di traffico effettuato e la quantità di sms e mms inviati, permette all'azienda di valutare quali sono gli utenti più "profittevoli" e quali sono gli utilizzatori che andrebbero maggiormente stimolati con azioni promozionali.

La novità dell'approccio della segmentazione comportamentale sta nel suddividere la clientela su informazioni provenienti da serie storiche e non a partire da dati medi di periodo. Ciò significa provvedere ogni anno all'operazione di assegnazione della clientela a un segmento. Si è in grado allora di misurare gli spostamenti di clientela da un segmento all'altro e di controllare a posteriori l'efficacia della politica commerciale applicata dall'azienda, che avrà ben operato se i segmenti poco remunerativi si saranno svuotati a favore di quelli redditizi.

2. DESCRIZIONE DATI

Il campione oggetto dello studio è stato estratto dal database della clientela di un'azienda di telefonia mobile. Contiene 32524 clienti, scelti casualmente tra coloro che hanno sottoscritto il contratto con l'azienda nel 2003 e, nell'aprile 2006, erano ancora attivi. Per ciascun cliente è noto il traffico telefonico mensile per un periodo che va da novembre 2004 a marzo 2006, oltre ad alcune caratteristiche socio-demografiche. Il data-set è già stato analizzato da Maela Bonetto (2007) allo scopo di individuare quali variabili influiscono la disattivazione del contratto telefonico (*churn*).
Di seguito viene esposto il procedimento per la costruzione della variabile risposta, che rappresenta una misura idonea a quantificare il "valore del cliente" per l'azienda.
Nel paragrafo successivo verranno elencate le variabili a disposizione, per poi effettuare le prime analisi esplorative, soprattutto grafiche, allo scopo di individuare il tipo di relazione tra i dati presenti nel *data-set* e la variabile risposta.

2.1 VARIABILE RISPOSTA E VARIABILI STATICHE

2.1.1 "Valore del cliente"

La variabile risposta per l'analisi di classificazione è il valore mensile del cliente, ovvero il consumo totale in euro (suddiviso in classi) di ciascun soggetto nel mese di marzo 2006, l'ultimo mese di cui si hanno dati disponibili.
La suddivisione in classi facilita le azioni di marketing per l'azienda, in quanto vengono definiti pochi gruppi sui quali attuare specifiche azioni promozionali sulla base delle caratteristiche che identificano ogni gruppo.
Una compagnia telefonica mette a disposizione un numero definito di tariffe, cercando di diversificare le proprie offerte per conseguire un utile e allo stesso tempo soddisfare il cliente. Esistono quindi diverse tipologie di utenti, a seconda delle loro preferenze di consumo. Lo scopo di questo lavoro non è tanto quello di studiare il comportamento di coloro che usufruiscono del servizio, ma creare una misura di sintesi per la variabile risposta, che verrà in seguito prevista con degli algoritmi di classificazione. Nel data-set

compaiono 8 diverse tariffe, ma non si conoscono tutti i costi ad esse associati perché la variabile è schermata. Tuttavia sono noti i costi medi per ciascun servizio offerto nel periodo di osservazione:

- 0,165€/min per le chiamate verso tutti;
- 0,15€ per gli sms;
- 0,55€ per gli mms;

Il valore del cliente per il mese di marzo 2006 è rappresentato dal prodotto fra il numero di utenze richieste da ogni soggetto e questi 3 costi medi.
Il dominio della nuova variabile è $[0,+\infty)$, che deve essere suddiviso in classi per separare le unità rispetto a questa nuova grandezza. Le soglie scelte sono:

consumi da 0 a 10€ $\longrightarrow$ clienti che generano un valore basso (classe "a");

consumi da 11 a 25€ $\longrightarrow$ clienti che generano un valore medio (classe "b");

consumi oltre i 26€ $\longrightarrow$ clienti che generano un valore alto (classe "c").

Oltre i 2/3 dei soggetti stanno al di sotto della prima soglia, mentre la classe centrale ha una fetta di consumatori pari al 10,73%, l'ultima al 5,59%.
Queste due soglie sono state scelte allo scopo di ottenere una prima fascia di clienti molto numerosa ma con consumi ridotti, e due fasce con pochi clienti profittevoli. Si pensa che un cliente con consumi ridotti abbia comportamenti e caratteristiche diverse rispetto a uno che usa frequentemente il cellulare.

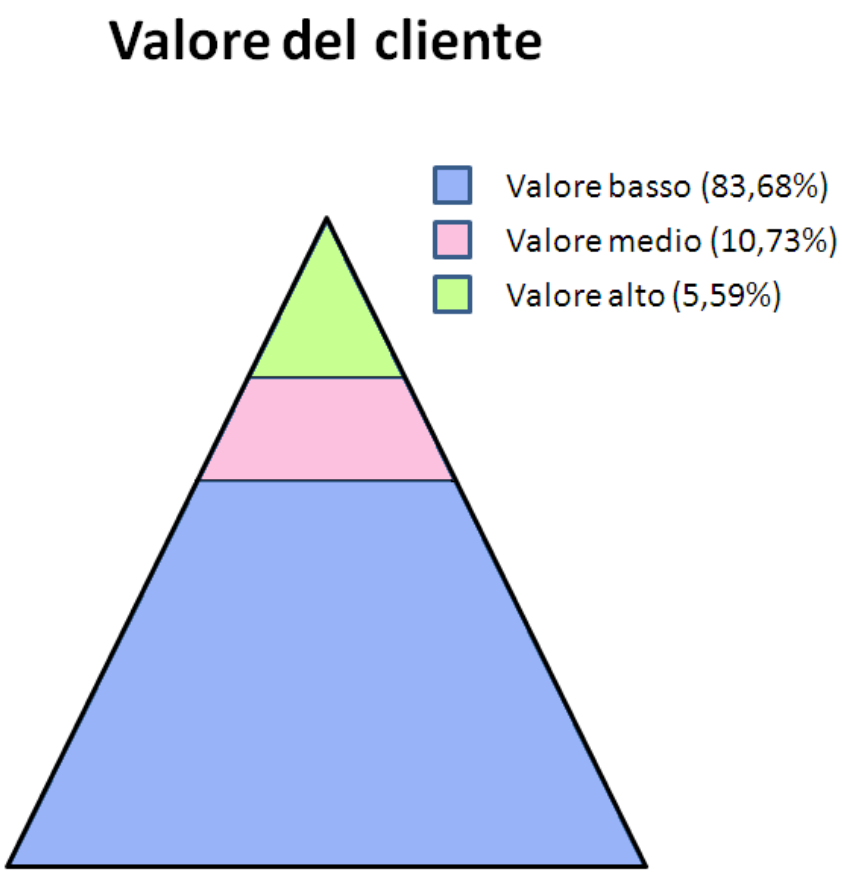

Figura 2.1: "Valore del cliente"

Le analisi svolte in questo elaborato hanno lo scopo di avvicinarsi alla corretta classificazione della clientela, costruendo dei modelli di previsione con le informazioni note fino a febbraio, per prevedere il valore per il mese successivo. Gli stessi modelli verranno poi aggiornati per prevedere il mese di aprile, maggio e così via, utilizzando come variabili in input i nuovi dati disponibili di mese in mese. Si presume quindi una stabilità temporale per l'algoritmo che stima la variabile risposta.

Le 3 classi non hanno numerosità omogenea, e ciò comporta difficoltà nell'effettuare una buona previsione. Nella classificazione questa caratteristica è frequente e l'analisi deve essere più accurata possibile per non escludere le classi meno numerose, che in questo caso sono quelle che apportano un maggior ricavo per l'azienda. I consumi per la prima classe rappresentano circa il 22% sul valore del cliente totale di marzo, mentre il restante 78% è apportato dai clienti con valore medio e alto. Per ricavare questi dati è stata calcolata la media per ciascuna fascia di consumo, ed ognuna è stata poi moltiplicata per la numerosità della fascia.

2.1.2 Variabili statiche

In tabella 2.1 vengono elencate le variabili e i valori che esse assumono:

Tabella 2.1: Elenco variabili statiche

	Nome	Tipo variabile	Ordinata	Possibili valori
1-	Valore marzo 2006	fattore a 3 livelli	SI	a, b, c
2-	Piano tariffario	fattore a 8 livelli	NO	A, B, C, D, E, F, G, H
3-	Marca	fattore a 5 livelli	NO	LG, Motorola, NEC, Nokia, Sony Ericsson
4-	Età	variabile quantitativa	SI	17…90
5-	Sesso	fattore a 2 livelli	NO	Maschio, Femmina
6-	Provincia	fattore a 104 livelli	NO	BG, BL, PD, RO…
7-	Zona	fattore a 5 livelli	NO	Nord-Est, Nord-Ovest, Centro, Sud, Isole

Ognuna di esse verrà messa a confronto con la variabile d'interesse, con l'ausilio di istogrammi e boxplot:

1. Valore del cliente marzo 2006: variabile risposta;

2. Piano Tariffario: indica la tipologia del contratto firmato dall'utente, alla quale corrispondono la tariffa applicata ed il tipo di pagamento. L'azienda telefonica ha diversi tipi di piani tariffari, ma non ci sono informazioni su di essi, perché la variabile è schermata, codificata cioè con lettere dell'alfabeto, rendendo impossibile risalire al nome del piano. Se l'analisi fosse eseguita in azienda sarebbe molto semplice collegare ogni piano tariffario alle sue caratteristiche.

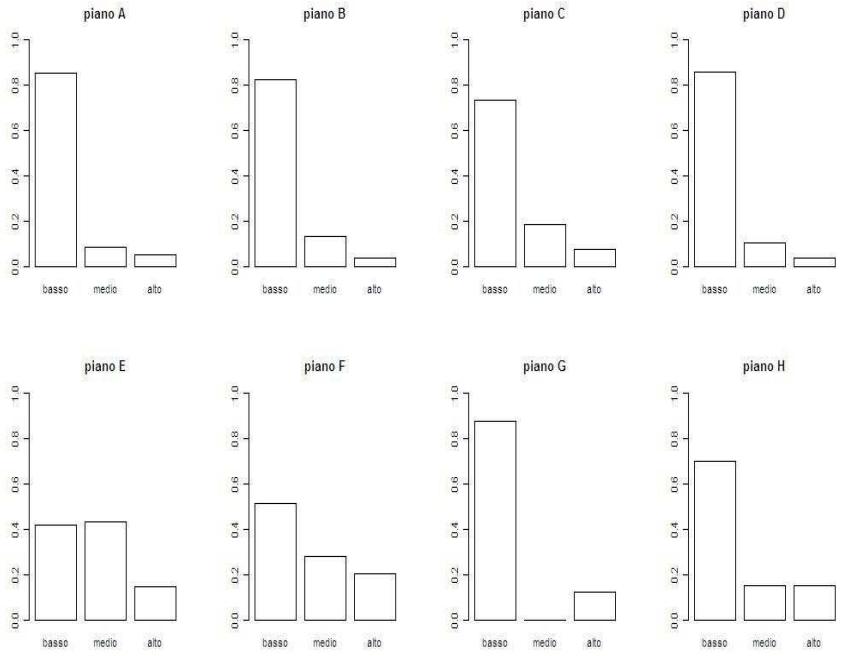

Figura 2.2: Istogrammi piano tariffario

Gli istogrammi di fig. 2.2 rappresentano le frequenze relative delle tariffe, allo scopo di individuare la possibile dipendenza con la variabile risposta. Nel caso di perfetta indipendenza le 3 barre che riportano le frequenze congiunte tra valore del cliente e tipo di piano dovrebbero avere la stessa struttura per tutti gli otto casi. I piani A,B,C,D,H hanno grafici abbastanza simili, mentre i restanti hanno strutture diverse. Ad esempio sapere che un consumatore utilizza il piano G dà la "certezza" (sulla base dei clienti osservati) che non appartiene alla classe "valore medio"; questa informazione potrebbe essere rilevante per l'analisi di classificazione.

È immediato notare che i clienti con valori bassi sono la categoria più frequente per tutte le otto tariffe, tranne per gli utenti con il piano E.

Calcolando le frequenze cumulate, è risultato che circa l'80% degli utenti utilizza il piano A o il piano B. Questo significa che gli istogrammi delle altre tariffe sono ottenuti dal rimanente 20% delle osservazioni, quindi sono meno informativi per la classificazione.

Un'altra cosa importante è che questa esplicativa non ha valori mancanti, in quanto non viene rilevata attraverso interviste e/o indagini, ma è un'informazione interna all'azienda.

Per verificare l'eventuale indipendenza fra il piano tariffario e la variabile risposta si propone di seguito il test chi quadro. Questa statistica mette a confronto la distribuzione condizionata di una variabile qualitativa rispetto alla variabile risposta, con la distribuzione stimata sotto l'ipotesi di indipendenza tra le due.

Variabile	Percentile	G.d.l.	P_value	Indipendenza
Piano tariffario	477,95	21	<0,001	NO

Il test, come ci si poteva aspettare, rifiuta l'ipotesi di indipendenza, quindi il piano tariffario è utile per la classificazione.

3. Marca e modello del telefonino: sono dati presenti solo per gli utenti che hanno attivato il contratto all'acquisto di un nuovo telefono, e non per quelli che hanno cambiato gestore mantenendo il proprio cellulare, perciò non è disponibile per la maggior parte delle persone. Tuttavia si riporta in figura 2.3 la distribuzione delle marche per i pochi valori a disposizione.

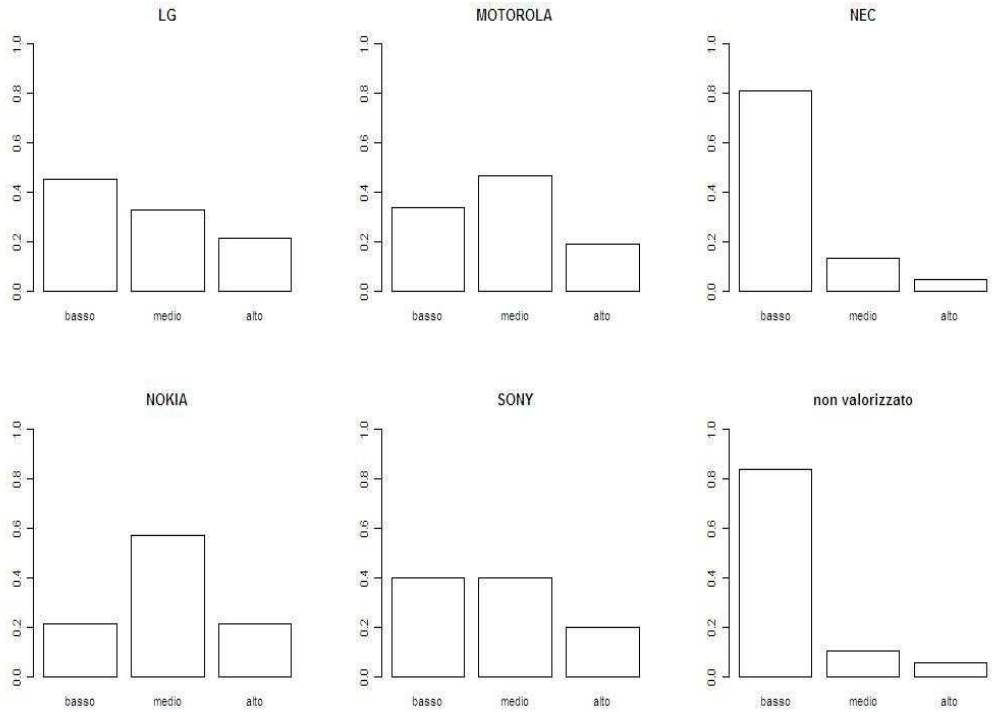

Figura 2.3: Istogrammi marca

Gli istogrammi per marca sono molto differenti tra loro, tuttavia soltanto 774 utenti sui 32524 totali hanno attivato il contratto acquistando un nuovo telefono. Il grafico in basso a destra rappresenta quindi la quasi totalità del campione, e rispetta le proporzioni già calcolate nell'analisi della variabile risposta. La marca e il modello di cellulare potrebbero essere informazioni rilevanti per il livello di consumo, ai fini di questo tipo di analisi potrebbe essere utile un'indagine per ricavare questo dato.

Il test chi-quadro rifiuta l'ipotesi nulla di indipendenza:

Variabile	Percentile	G.d.l.	P_value	Indipendenza
Marca	121,26	12	<0,001	NO

4. Età: Nel dataset è presente la data di nascita per ciascun utente, dalla quale si può ricavare l'età al 31 marzo 2006. L'età è approssimata all'anno, quindi una persona che ha 30 anni e 7 mesi viene registrata con età 31, mentre chi ha 30 anni e 6 mesi avrà età 30. Fra le osservazioni ci sono alcuni individui con età superiore ai 105 anni, che rappresentano evidentemente degli errori di rilevazione. Per non escludere questi utenti dalle analisi successive, si è deciso di abbassare la loro età alla media generale, pari a 41 anni.

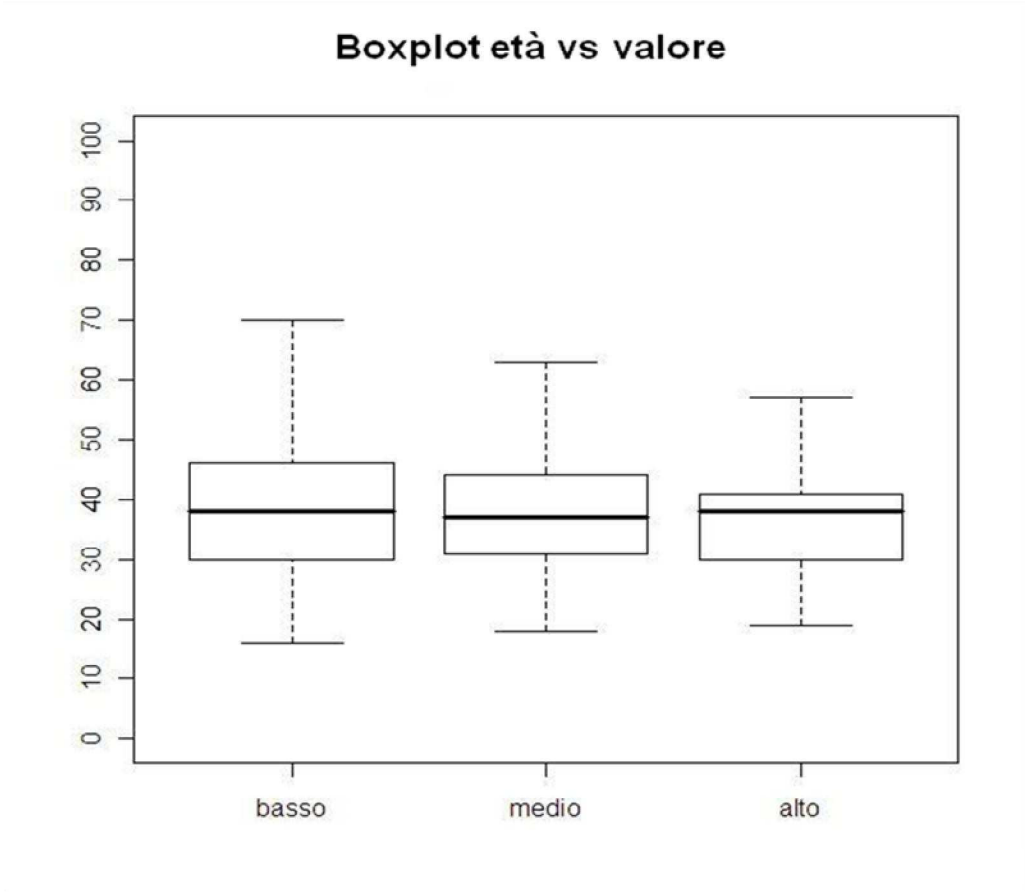

Figura 2.4: Boxplot tra età e valore del cliente

Le mediane e il primo quartile dei 3 boxplot sono abbastanza simili. I soggetti che generano un basso valore hanno la scatola meno schiacciata rispetto alle altre due classi. Si ricorda che l'età è valorizzata per tutti i soggetti nel campione, quindi può portare un contributo rilevante per la classificazione. Le scatole e i punti di massimo e di minimo (escludendo gli outlier) sono più concentrati attorno al valore mediano per gli utenti con consumi medi e alti.

Gli utenti con consumi medi e alti si concentrano nella fascia di età che va dai 30 ai 45 anni, e questo fenomeno può dipendere dal frequente utilizzo del cellulare durante la giornata lavorativa. Il campo di variazione di questa variabile va da un minimo di 16 anni fino ad un massimo di 91 anni.

Si verifica con un test *F* la dipendenza tra l'età media e il valore del cliente:

Variabile	Percentile	G.d.l.	P_value	Indipendenza
Età	21,83	2	<0,001	NO

Il test rifiuta l'ipotesi nulla di uguaglianza delle medie per l'età nei 3 gruppi individuati dalla variabile risposta.

5. Al momento della sottoscrizione del contratto ciascun cliente ha indicato il proprio sesso e il 75% del campione è costituito da maschi. In figura 2.5 vengono indicate le frequenze relative:

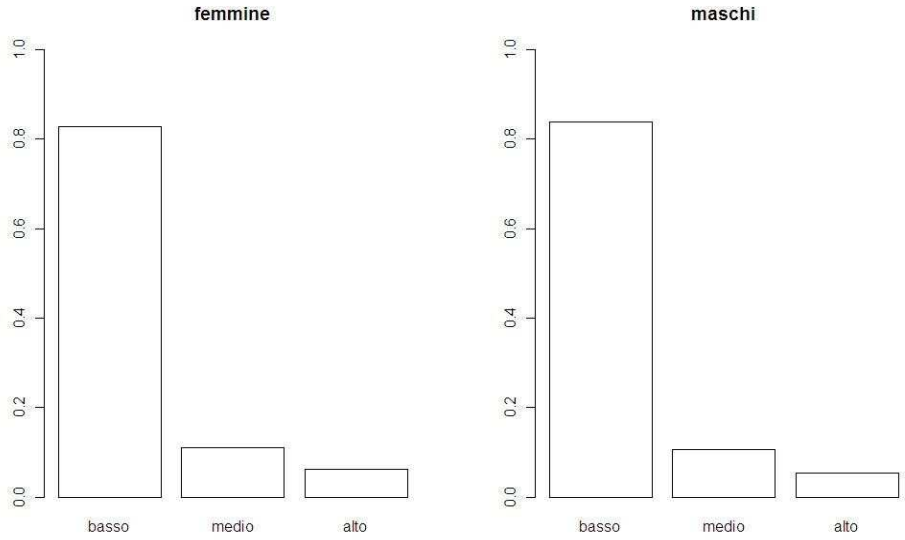

Figura 2.5: Istogrammi sesso

Per la variabile sesso le differenze sono minime rispetto al valore del cliente. La prima classe resta la più numerosa, la somma delle altre in entrambi i grafici è sotto il 20% del totale utenti.

Variabile	Percentile	G.d.l.	P_value	Indipendenza
Sesso	4,05	3	0,2563	SI

Il test chi quadro conferma l'informazione apportata dai grafici di figura 5, il sesso non sembra essere legato marginalmente al valore del cliente.

6. La provincia di provenienza è una variabile fattore a 104 livelli. Per facilitare lo studio della dipendenza si è deciso di raggruppare le province in 5 zone: nord-

ovest, nord-est, centro, sud e isole. Le zone dove si registra una maggiore diffusione del servizio sono il centro e il nord-est.

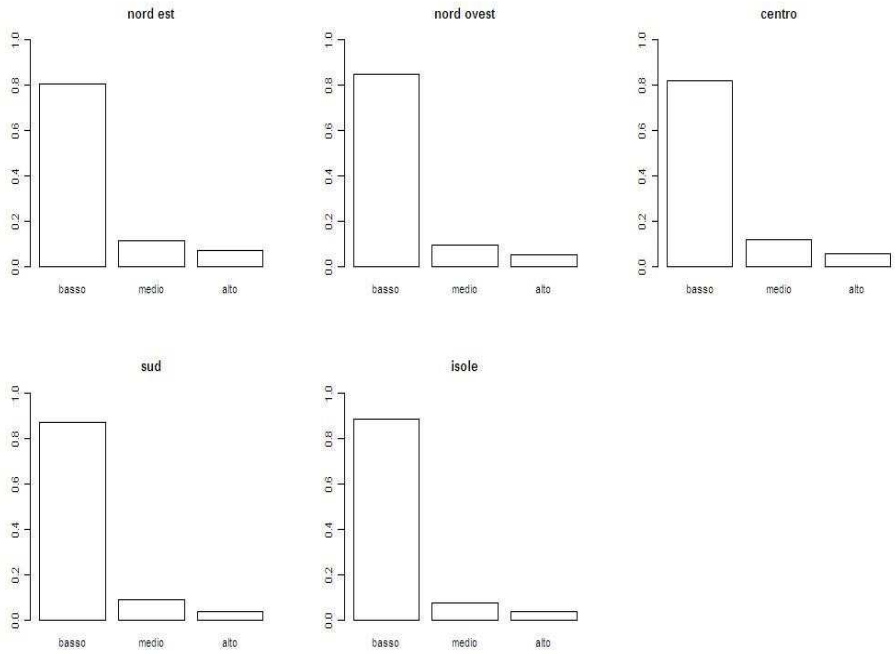

Figura 2.6: Istogrammi zona di provenienza

Non sembra ci siano differenze fra le 5 zone, in tutti i grafici la barra del valore basso si attesta attorno all'80%. Il nord-est ed il centro sono le zone con più fatturati medi e alti.

Variabile	Percentile	G.d.l.	P_value	Indipendenza
Zona	106,34	12	<0,001	NO

Rispetto ai grafici il chi-quadro rileva la dipendenza con la variabile risposta, quindi la variabile è rilevante per la classificazione.

2.2 VARIABILI LONGITUDINALI

Oltre alle caratteristiche socio-demografiche, il data-set contiene 5 serie storiche mensili per ciascun utente, che descrivono i consumi per il periodo che va da novembre 2004 a marzo 2006:

- Chiamate verso lo stesso operatore: per ogni mese sono rilevati i minuti di chiamate verso soggetti che hanno un contratto con la stessa compagnia telefonica;
- Chiamate verso fisso: indica i minuti mensili totali di chiamate verso i telefoni di rete fissa;
- Chiamate verso altri: riporta i minuti mensili totali di chiamate verso i telefonini con contratto telefonico di un'altra compagnia telefonica;
- Sms: numero di sms spediti verso tutti;
- Mms: numero di mms spediti verso tutti;

2.2.1 Serie storica dei consumi totali

Questa serie storica descrive l'andamento dei consumi a cadenza mensile per ciascun utente rilevato. Il criterio seguito per la costruzione di questa nuova matrice è lo stesso che viene applicato per la variabile risposta: avendo a disposizione i 5 *data-set* contenenti le serie storiche che descrivono l'evoluzione dei consumi per i 32.524 soggetti studiati, si somma il numero di servizi richiesti. Per rendere più chiaro il concetto vengono riportate di seguito le 5 serie per il primo utente e la serie calcolata dei consumi mensili:

Tabella 2.2: Serie storiche dei consumi per il primo utente

	nov-04	dic-04	gen-05	feb-05	mar-05	apr-05	mag-05	giu-05	lug-05	ago-05	set-05	ott-05	nov-05	dic-05	gen-06	feb-06
ch vs stesso	1	2	0	2	0	0	0	0	0	0	1	0	0	0	5	11
ch vs fisso	6	17	2	11	12	9	1	1	4	5	2	11	15	10	7	9
ch vs altri	0	1	0	2	5	0	1	0	2	3	1	3	5	7	11	8
sms	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mms	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2
consumi tot.	7	20	2	15	17	9	2	1	6	8	4	14	20	17	32	34

È stato scelto di non moltiplicare il numero di servizi per il costo degli stessi, come invece avviene per la variabile risposta. Le chiamate e gli sms hanno un costo molto simile (0,165 €/min e 0,15€), mentre gli mms hanno un costo triplo rispetto alle altre

utenze. Quindi la matrice che riporta il numero di mms inviati viene moltiplicata per 3, per poi essere sommata alle altre.

Come si vede dalla tabella 2.2 , la nuova serie calcolata sintetizza le informazioni a disposizione, passando da 5 righe a 1 riga soltanto.

Durante la fase di stima dei modelli dei prossimi capitoli, verranno tolte le serie dei consumi mensili che hanno molti zeri consecutivi nel periodo preso in esame, mentre le analisi di questo capitolo sono riferite all'intero campione osservato.

Come prima analisi, viene riportato il diagramma a scatola con baffi dei consumi totali divisi per mese, per il periodo che va da novembre del 2004 fino a marzo del 2006. La *scatola* riporta per ciascun mese il valore del cliente mediano, il primo e il terzo quartile, mentre i *baffi* restituiscono il valore massimo e minimo.

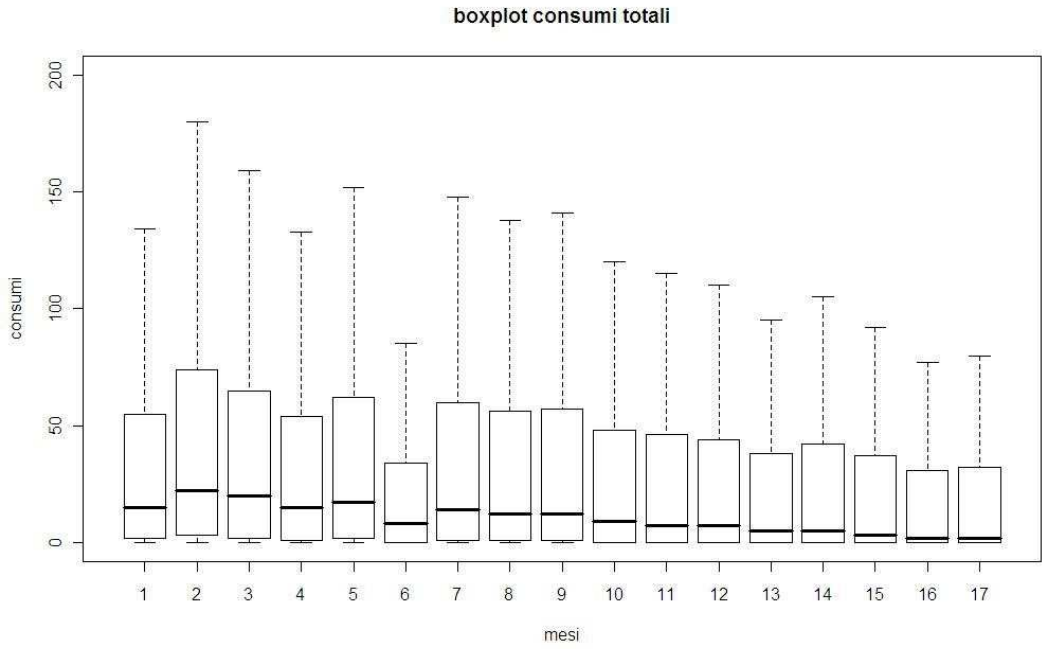

Figura 2.7: Boxplot consumi totali per il periodo osservato

Per la rappresentazione sono state escluse le osservazioni anomale che stanno al di sopra del baffo superiore, pari circa al 5% dei soggetti ogni mese, allo scopo di far risultare i boxplot meno schiacciati. Tali osservazioni non verranno scartate nelle analisi successive. In seguito si riportano le medie dei consumi per ciascun periodo:

Tabella 2.3: Medie dei consumi per l'intero campione

	nov-04	dic-04	gen-05	feb-05	mar-05	apr-05	mag-05	giu-05	lug-05
medie dei consumi	20,76	23,35	22,72	20,73	21,43	17,15	19,96	19,44	19,23
	ago-05	set-05	ott-05	nov-05	dic-05	gen-06	feb-06	mar-06	
	18,03	16,85	16,33	15,31	15,16	14,16	13,39	13,12	

Confrontando le diverse colonne della tab. 2.3 con la linea mediana dei boxplot non si notano grosse differenze; i consumi nell'arco del periodo osservato sono piuttosto esigui, soprattutto per la parte finale relativa all'inizio del 2006. Si passa dal picco di dicembre 2004, per poi abbassarsi mediamente di 10 tra messaggi e chiamate nell'ultimo mese osservato. Anche i picchi di massimo si registrano esclusivamente entro marzo del 2005. Se inizialmente i minimi erano staccati dalla scatola, negli ultimi 8 mesi minimo e primo quartile coincidono.

Per semplicità di esposizione ci si concentra su un piccolo campione di soggetti scelti in modo casuale, che comunque rappresentano gli andamenti tipici dell'intero campione.

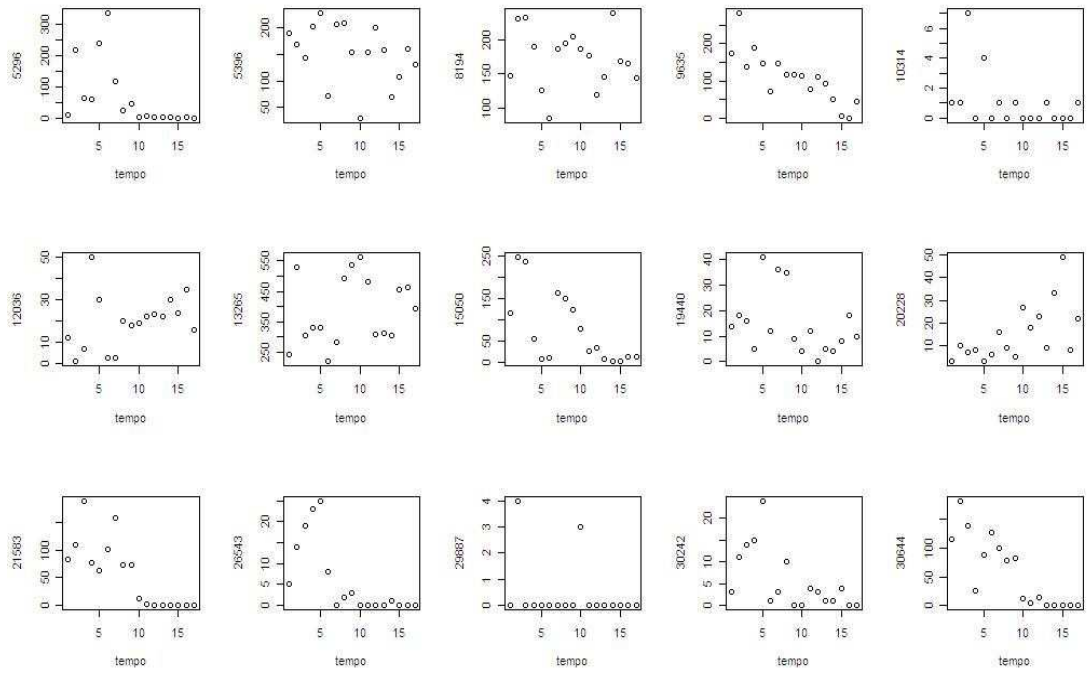

Figura 2.8: Consumi totali per 15 utenti scelti casualmente

I numeri alla sinistra dei grafici rappresentano l'utente, in ascissa sono riportati i mesi di osservazione e in ordinata i consumi in euro.

Dai grafici sembra che l'andamento non possa essere colto da una semplice funzione matematica (es. retta o parabola) per la maggior parte delle figure, escludendo quelle

serie che hanno i consumi uguali a zero per buona parte del periodo di osservazione. Per queste ultime la dipendenza seriale sarà prossima a uno per gli "zeri consecutivi". Inoltre c'è da osservare che la scala dei grafici è diversa per ciascun utente. Nel caso della settima serie, per esempio, i consumi si aggirano attorno a una media di 400 fra chiamate e messaggi, mentre tutte le altre hanno consumi molto più esigui.

Le 32524 serie del campione saranno utilizzate come strumento, al fine di raggruppare i soggetti in categorie definite dalla variabile "valore del cliente". L'idea alla base di questo studio sta nel riassumere il comportamento di queste serie con l'ausilio di pochi parametri che colgono sia la componente deterministica che la componente stocastica delle serie. Perciò in un primo momento verranno adattati dei semplici modelli con l'intento di individuare il trend deterministico (lineare o curvilineo). I residui risultanti da questa stima saranno prima ipotizzati serialmente incorrelati, poi nei capitoli successivi verranno studiati con opportuni modelli stocastici.

2.2.2 Modello lineare

Generalmente l'analisi di serie storiche prevede lo studio di un'unica serie rispetto al tempo, in questo caso bisogna cercare una forma funzionale comune a molte serie, alcune con andamento lineare, altre con andamento curvilineo. Si parte quindi con la stima del modello più semplice, quello della regressione lineare. Questo metodo permette di stimare il valore atteso condizionato della variabile dipendente, dati i valori di altre variabili indipendenti o esogene. In questo caso la variabile risposta è il numero di chiamate effettuate e messaggi inviati in un mese, mentre l'unica variabile indipendente è il tempo. Si tratta quindi del modello di regressione lineare semplice. La forma funzionale è:

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

L'intercetta rappresenta i consumi stimati di ottobre 2004, mentre il coefficiente angolare denota l'evoluzione degli stessi, che può essere crescente, decrescente o costante. Per la stima dei parametri si utilizza il metodo dei minimi quadrati.

Fra le ipotesi principali per stimare il valore atteso condizionato della variabile risposta, si ricorda la media condizionata degli errori uguale a zero ($E(\varepsilon_t \mid \underline{x}_t) = 0$) e l'indipendenza degli stessi rispetto al tempo ($E(\varepsilon_t \varepsilon_s \mid \underline{x}_t) = 0$ per $t \neq s$).

Si procede quindi con la stima dei due parametri per le 15 serie prese in considerazione e se ne riportano i grafici:

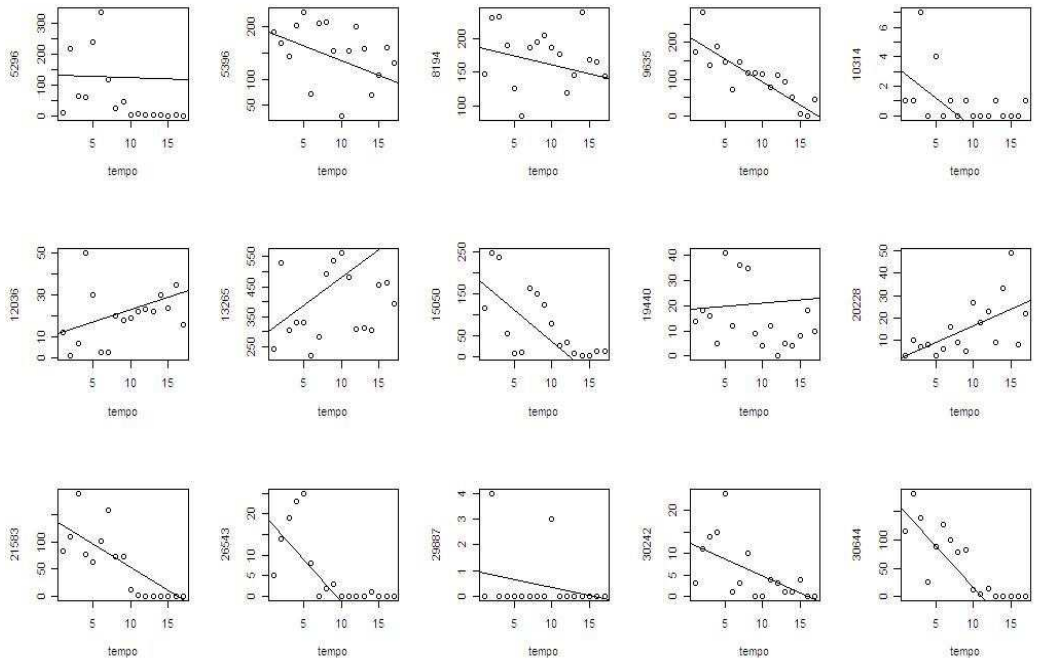

Figura 2.9: Modello lineare per le 15 serie dei consumi totali

Da questa prima analisi si può vedere come la retta stimata non interpoli adeguatamente i punti del grafico, tuttavia è utile per avere una prima idea sull'evoluzione (crescente, decrescente, costante) delle serie prese in esame.

La fig. 2.9 evidenzia come la maggior parte dei consumi abbiano un trend decrescente, dovuto in particolare al calo delle chiamate negli ultimi mesi osservati.

Per rendere più evidente questo fenomeno in fig. 2.10 si riportano gli istogrammi riferiti all'intero campione per entrambi i parametri stimati:

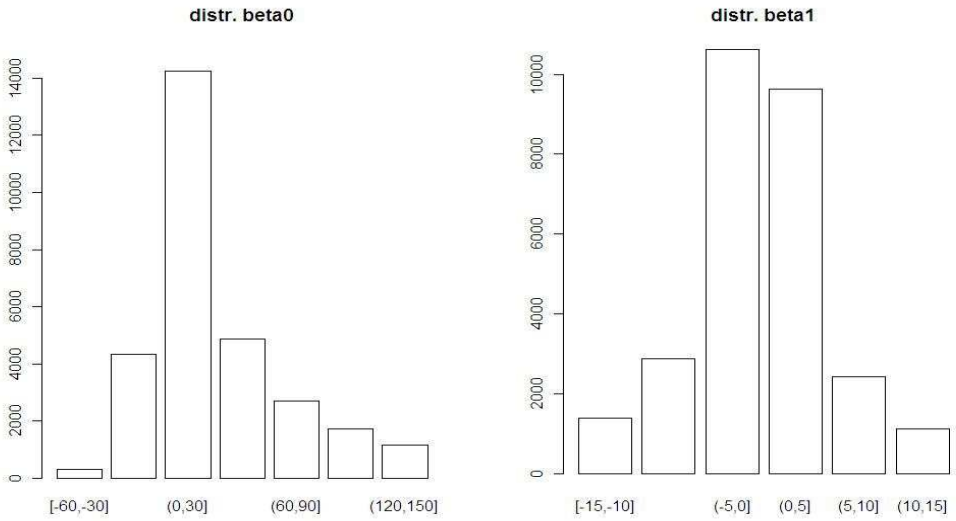

Figura 2.10: Distribuzione dei coefficienti di regressione

L'istogramma a sinistra rappresenta i consumi stimati al tempo zero, mentre quello di destra riporta le frequenze delle variazioni mensili degli stessi.

Buona parte delle pendenze sono nulle, ciò sta a significare che molti utenti non utilizzano il cellulare nel periodo osservato e quindi una retta orizzontale stima perfettamente questo andamento. Dato che la devianza residua è nulla, il coefficiente di determinazione R^2 dato dal rapporto tra la devianza spiegata dal modello e quella residua non può essere calcolato. Dal momento però che la retta orizzontale si adatta perfettamente alle osservazioni si pone il coefficiente pari a 1, cioè il massimo valore che può assumere.

Tabella 2.4: Frequenze dell'R^2 per il modello lineare

R^2	Serie	Freq. relative
R^2>=0,5	6.143	18,89%
R^2<0.,5	26.381	81,11%
Tot. osservazioni	32.524	100,00%

Un altro problema è legato alla significatività dei coefficienti, da verificare con il test *t* di student. Tale procedura permette di appurare se un coefficiente (β_0 *o* β_1) può essere considerato pari a zero, oppure se i dati portano a rifiutare questa congettura. Viene formulato quindi per ciascun cliente un sistema di ipotesi:

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases}$$

Con l'ausilio della statistica *t*:

$$t = \frac{\hat{\beta}_0 - \beta_{0,H_0}}{\sqrt{S^2}}$$

se il t_{oss} è compreso tra i percentili della distribuzione *t* di student sotto la nulla, significa che i dati la supportano, perciò $\beta_0 = 0$. Di seguito vengono riportate le frequenza assolute e relative dei *p-value* del test per entrambi i coefficienti di tutta la base clienti.

Tabella 2.5: Frequenze assolute e relative dei p_value associati ai coefficienti di regressione

P_VALUE	BETA_0 freq. ass.	BETA_0 freq. rel.	BETA_1 freq. ass.	BETA_1 freq. rel.
P<=0,5	5073	15,60%	17735	54,53%
P>0,5	27451	84,40%	14789	45,47%
Tot.oss	32524	100,00%	32524	100,00%

Per l'84,4% dei coefficienti β_0 non si rifiuta l'ipotesi nulla, mentre più della metà dei coefficienti β_1 sono significativamente diversi da zero.

Si procede con l'analisi dei residui per i 15 utenti precedentemente selezionati con il campionamento casuale:

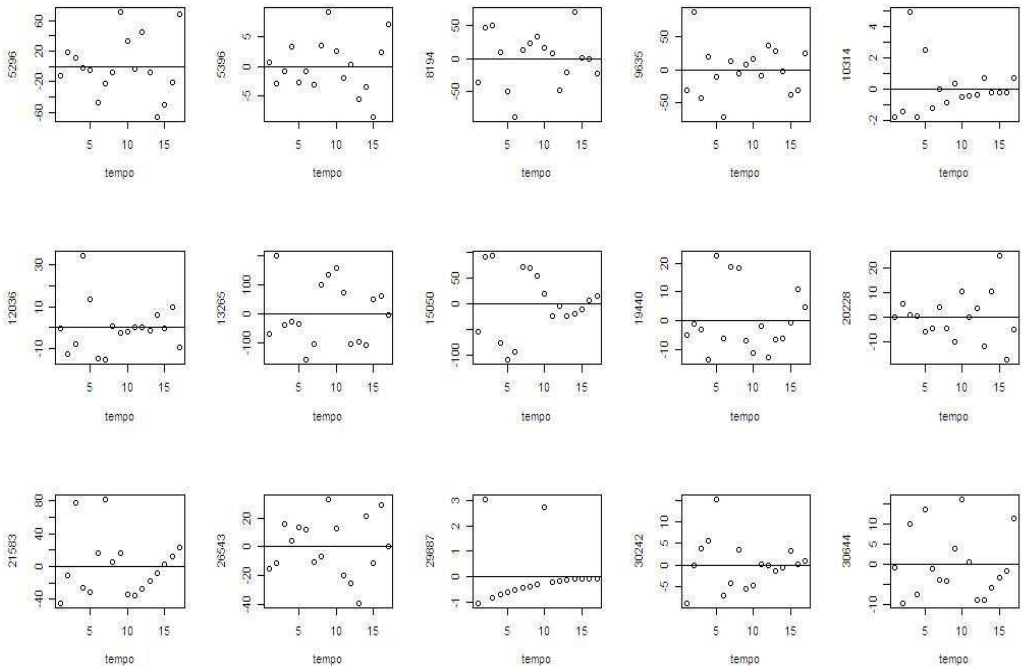

Figura 2.11: Residui per i 15 utenti

Come nel caso delle serie osservate, anche i residui seguono degli andamenti diversi, alcuni sembrano dispersi in modo casuale, altri hanno andamenti crescenti/decrescenti, altri ancora curvilinei. I residui, secondo le ipotesi del modello lineare, dovrebbero avere media zero, oltre ad essere serialmente incorrelati. Il grafico dovrebbe quindi riprodurre circa la metà dei residui al di sotto della retta tracciata, e i restanti al di sopra. Oltre a questo non dovrebbero esserci degli andamenti facilmente descritti da una funzione matematica (tipo una retta o una curva), il che starebbe a significare che il modello non è ben specificato e necessita dell'aggiunta di altri parametri.

Queste ipotesi non vengono rispettate, sarà quindi opportuno aggiungere delle componenti per migliorare l'adattamento.

I grafici sotto riportati verificano se viene rispettata l'ipotesi di normalità dei residui:

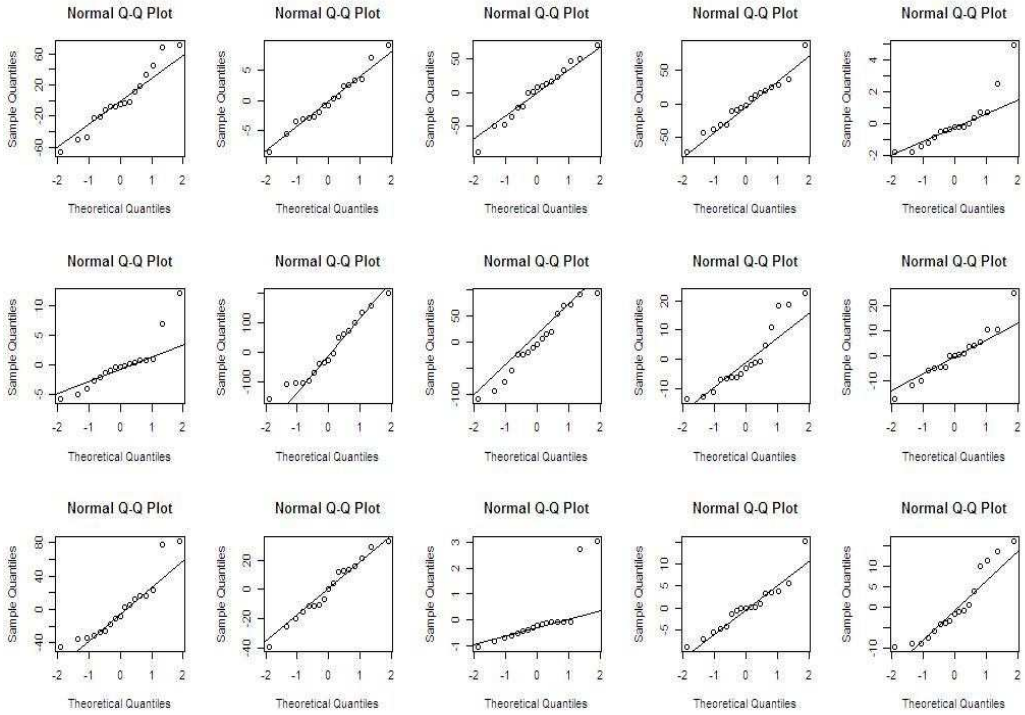

Figura 2.12: Qqnorm e qqline per i 15 utenti analizzati

Se i residui seguissero un processo gaussiano i punti dovrebbero allinearsi lungo la bisettrice del grafico, che in ascissa riporta i residui teorici e in ordinata i residui osservati. Per buona parte delle serie i punti sono in prossimità della "retta teorica", quindi per questo sottocampione si può accettare l'ipotesi di normalità dei residui.

2.2.3 Aggiunta di un parametro legato al tempo

Sembra che il modello scelto sia troppo parsimonioso per scopi previsivi. Si aggiunge quindi un parametro legato al tempo, dal momento che è l'unica esplicativa per ora a disposizione. Il nuovo modello è:

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$$

Questo modello è più generale perché include la retta adattata precedentemente, i due modelli sono quindi *annidati*.
Vengono ripetuti i ragionamenti fatti al paragrafo 2.2.2.

Restando sempre sui 15 clienti analizzati, viene verificato l'eventuale miglioramento nella bontà del modello. In tab. 2.6 si riportano i coefficienti stimati e i *p_value* associati:

Tabella 2.6: Coefficienti stimati e p_value associati per i 15 utenti. In grigio sono indicati i coefficienti significativi (p_value<0,05)

	BETA_0	PVALUE_B0	BETA_1	PVALUE_B1	BETA_2	PVALUE_B2
1	141,6640	0,0144	-8,0123	0,3703	0,4387	0,3638
2	197,8344	0,3793	-0,4298	0,6746	0,0506	0,3661
3	187,9412	0,0001	-2,6324	0,7814	0,0833	0,8707
4	219,0735	0,0000	-12,7076	0,1502	0,0571	0,9009
5	3,1618	0,0453	-0,3982	0,2975	0,0135	0,5066
6	12,5735	0,0046	-2,6451	0,0152	0,1294	0,0252
7	294,7353	0,0071	18,4623	0,4534	-0,7239	0,5843
8	187,5147	0,0042	-15,2189	0,2973	0,2419	0,7547
9	18,7059	0,0785	0,2343	0,9273	-0,0599	0,6668
10	1,5882	0,8573	1,4907	0,5125	0,0005	0,9966
11	137,6029	0,0008	-8,4421	0,3224	-0,0597	0,8950
12	18,8824	0,2337	-2,9307	0,5251	0,3487	0,1730
13	1,1912	0,2508	-0,1333	0,6085	0,0040	0,7751
14	12,6765	0,0220	-0,7896	0,5406	-0,0004	0,9955
15	149,9412	0,0561	1,1045	0,5567	-0,1210	0,2420

La maggior parte dei coefficienti sono poco significativi (*p_value*>=0.05), soprattutto le colonne di β_1 e β_2. Le intercette sono metà significative e metà non significative. I valori presenti in tabella sono difficilmente interpretabili per capire se il cliente ha registrato consumi crescenti o decrescenti. Infatti bisogna distinguere se nel range che va da 1 a 17 la parabola stimata si trova alla sinistra o alla destra del vertice.

Per questo motivo vengono riportati i grafici con le curve stimate:

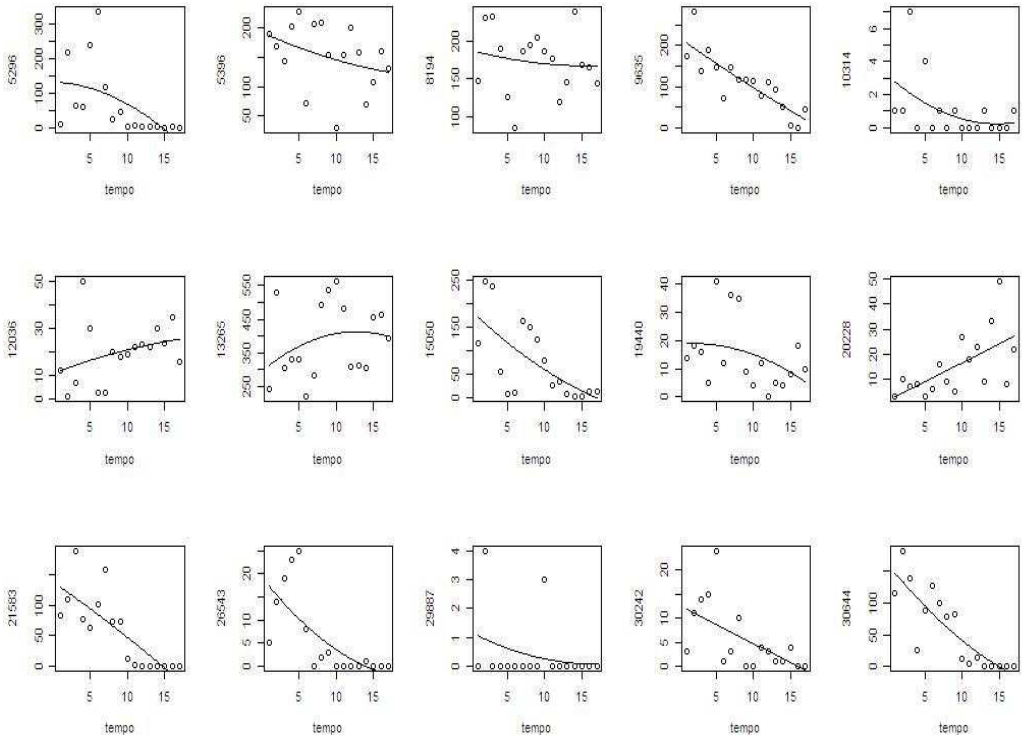

Figura 2.13: Modello con t^2 per le 15 serie dei consumi totali

La curva stimata si adatta leggermente meglio all'evoluzione delle chiamate, in quanto dispone di un parametro aggiuntivo rispetto alla retta. Per alcune serie (per esempio la quinta e la dodicesima) gli effetti dell'aggiunta del parametro β_2 sono notevoli e i modelli stimati seguono molto bene l'andamento dei consumi. Per le restanti la nuova forma funzionale non apporta miglioramenti.

Tabella 2.7: Frequenze dell'R^2 per il modello con t^2

R^2	Serie	Freq. relative
R^2>=0,5	7.003	21,53%
R^2<0,5	25.521	78,47%
Tot. osservazioni	32.524	100,00%

Gli R^2 maggiori della soglia posta a 0.5 sono leggermente aumentati, come spesso avviene a seguito dell'aggiunta di un nuovo parametro al modello. Come prima viene verificato l'eventuale non significatività dei coefficienti, e vengono riportati gli istogrammi solo per β_1 e β_2, visto che β_0 stima i consumi al tempo zero e non è importante per il trend delle chiamate.

Tabella 2.8: Frequenze assolute e relative dei p_value associati ai coefficienti

P_VALUE	BETA_0 freq. ass.	BETA_0 freq. rel.	BETA_1 freq. ass.	BETA_1 freq. rel.	BETA_2 freq. ass.	BETA_2 freq. rel.
P<=0,5	15.979	49,13%	7.535	23,17%	6.557	20,16%
P>0,5	16.545	50,87%	24.989	76,83%	25.967	79,84%
Tot.oss	32.524	100,00%	32.524	100,00%	32.524	100,00%

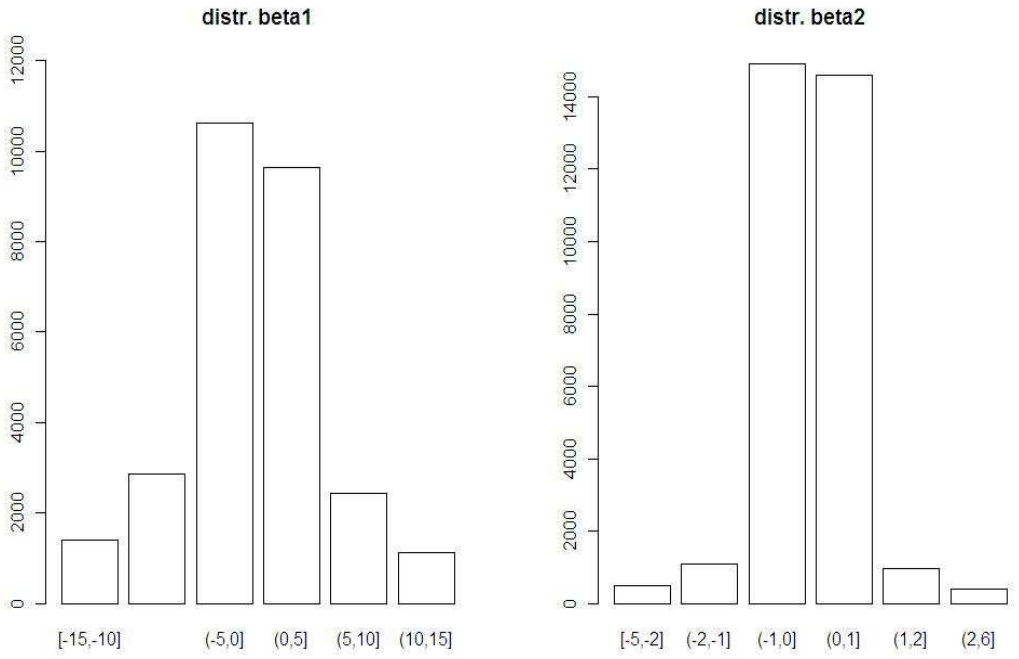

Figura 2.14: Distribuzione coefficienti per il modello con t^2

Il coefficiente $\hat{\beta}_1$ ha sia valori positivi che negativi in egual misura, mentre per $\hat{\beta}_2$ prevalgono valori negativi. Riguardo alla significatività la situazione non è migliorata con l'aggiunta del termine di secondo grado, in quanto sono diminuiti i coefficienti significativi per il coefficiente angolare. Tuttavia aumentano i $\hat{\beta}_0$ significativi, passando da 5073 (15,60%) con p_value<=0.5 a 15979 (49,13%).

Ancora una volta bisogna verificare l'eventuale autocorrelazione nei residui, in questo caso però non si utilizza il grafico dei residui osservati ma l'ACF. Questo strumento permette di verificare la struttura di dipendenza temporale di una serie storica, in questo caso dei residui della stessa. Sulla linea orizzontale è riportato il tempo delle osservazioni e le due linee tratteggiate riportano gli intervalli di confidenza per appurare l'ipotesi nulla di assenza di correlazione. In questo caso la distribuzione asintotica dell'autocorrelazione stimata $\hat{\rho}(k)$ è $N(0,1/n)$ e quindi le bande che

delimitano la regione di accettazione sono $t_{\alpha/2}/\sqrt{n}$ e $t_{1-\alpha/2}/\sqrt{n}$, dove t indica il percentile della distribuzione t di Student.

Tutte le serie a disposizione contengono 17 valori, troppo pochi per un'analisi accurata della struttura del fenomeno nel tempo. Le bande di confidenza sono molto ampie, quindi si accetterà quasi sempre l'ipotesi nulla, ma ciò è dovuto al motivo esposto sopra.

In figura 15 si riportano i grafici dell'ACF dei residui:

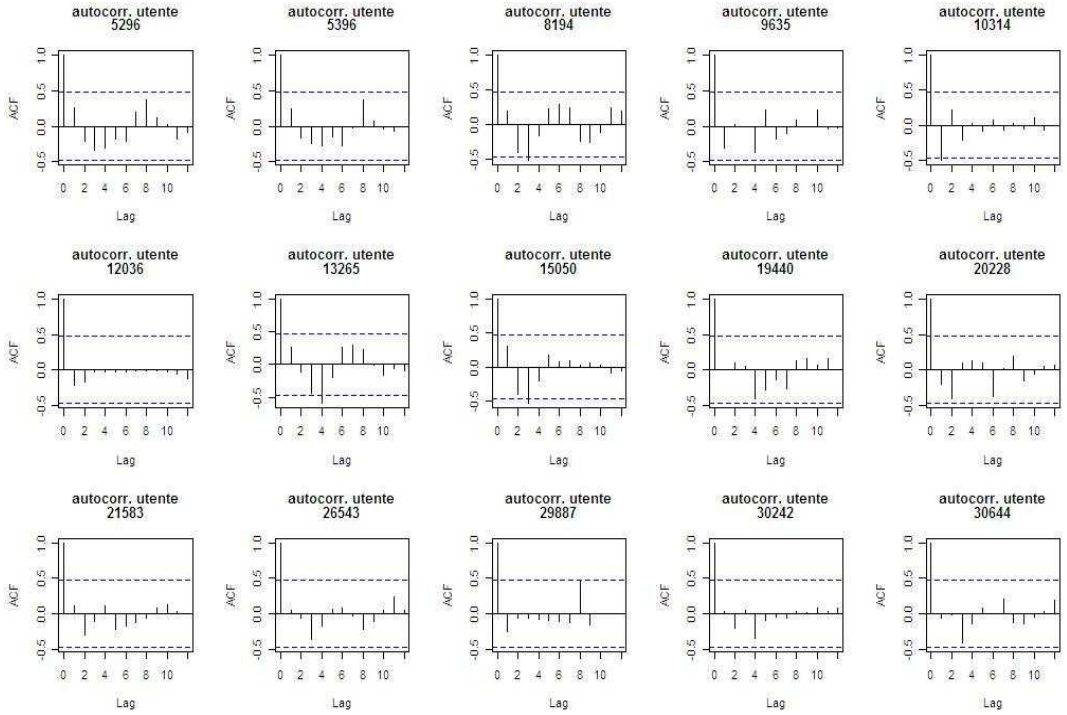

Figura 2.15: ACF per i 15 utenti analizzati

Il grafico riporta la correlazione dei residui per i medesimi 15 utenti. La correlazione seriale può variare fra zero e uno, due valori che stanno ad indicare rispettivamente minima e massima dipendenza. La prima banda di ciascuna figura descrive la dipendenza dal tempo presente, per questo è pari a uno.

Andando indietro con il *lag* temporale le correlazioni sono, a parte poche eccezioni, interne alle bande di confidenza.

2.3 ALCUNI ACCORGIMENTI

2.3.1 Variabili longitudinali

Uno degli scopi di questo lavoro è studiare il comportamento dei clienti (in termini di numero di chiamate e invio di messaggi) per capire se esistono delle relazioni tra la variabile risposta e le variabili legate alle serie storiche dei consumi. I clienti con consumi pari a zero per tutto il periodo di osservazione si possono considerare "non clienti". Su di essi è inutile stimare un modello rispetto al tempo, basterebbe effettuare un controllo preliminare sulla somma dei consumi per il periodo di osservazione. Se risulta uguale a zero, quei clienti andranno direttamente nella classe "a".

Inoltre esiste una grossa differenza tra chi non effettua chiamate e chi ha un traffico telefonico ridotto. Un soggetto che non consuma potrebbe aver smarrito la tessera SIM oppure aver cambiato operatore. Tali utenti non apportano informazioni utili alla compagnia telefonica, che ricerca le determinanti che portano ad un aumento del valore del cliente. Chi invece effettua poche chiamate potrebbe essere stimolato dall'azienda con promozioni, al fine di aumentare il suo valore.

Oltre ad un problema di sostanza, esiste un problema di natura statistica. Nei prossimi capitoli verranno stimati dei modelli stocastici di tipo $ARMA(p,q)$ che, fra le ipotesi di base, richiedono che la correlazione tra le osservazioni sia strettamente minore di uno ($|\rho_i| < 1 \quad \forall \; i$). Se la serie storica di un cliente presenta soltanto valori nulli, la correlazione risulta massima per qualsiasi lag temporale scelto.

Per questi motivi si è deciso di escludere dal campione quei soggetto con consumi "troppo bassi". Nella tabella successiva viene elencato il numero di utenti rapportato al numero totale di zeri osservati:

Tabella 2.9: Suddivisione utenti in base al numero totale di zeri presenti nella serie storica dei consumi totali

Numero zeri	Utenti	Freq. cumulate	% cumulate
0	10973	10973	33,74%
1	2433	13406	41,22%
2	1756	15162	46,62%
3	1725	16887	51,92%
4	1645	18532	56,98%
5	1590	20122	61,87%
6	1662	21784	66,98%
7	1650	23434	72,05%
8	1546	24980	76,80%
9	1492	26472	81,39%
10	1478	27950	85,94%
11	1377	29327	90,17%
12	1117	30444	93,60%
13	794	31238	96,05%
14	597	31835	97,88%
15	407	32242	99,13%
16	204	32446	99,76%
17	78	32524	100,00%
Tot. utenti	32524		

Sembra ragionevole tollerare un numero massimo di 7 zeri nella serie storica di ogni utente. Ne vengono quindi esclusi 9090 (27,95%) dalle successive analisi.

Osservando le singole colonne del data-set, il mese con il maggior numero di consumi uguali a zero è proprio marzo 2006, seguito da febbraio 2006 e gennaio 2006, quindi c'è un trend crescente di "non utilizzatori", che sarebbe interessante da studiare, ma che esula dagli scopi del presente lavoro.

Inoltre bisogna trattare in modo diverso i soggetti che hanno degli zeri sparsi su tutto il periodo studiato, rispetto a coloro che hanno degli zeri consecutivi. I comportamenti ipotizzabili sono molto diversi: gli zeri sparsi indicano un soggetto che per qualche mese durante l'anno non ha usufruito del cellulare ma comunque continua ad utilizzarlo; gli zeri consecutivi indicano che il consumatore non usa il telefonino e forse intende cambiare operatore.

Per queste considerazioni anche le persone che hanno consumi nulli per almeno 3 mesi consecutivi vengono scartate dall'analisi di classificazione.

A seguito di questo doppio criterio di eliminazione le percentuali di valore del cliente per il mese di marzo sono così modificate:

- valore basso: 77,59% (-6,09%)
- valore medio: 14,73% (+4%);
- valore alto: 7,68% (+2,09%).

La base clienti si è ridotta da 32524 clienti a 18464.

2.3.2 Periodo di osservazione

Il mese di marzo 2006 è stato inserito nelle analisi precedenti per avere una visione generale dell'andamento dei consumi. Tuttavia la variabile risposta è calcolata sull'ultimo mese a disposizione, quindi per gli studi successivi non può essere sfruttato come variabile esogena.
Il periodo di osservazione si riduce a 16 mesi, da novembre 2004 a febbraio 2006.

3. METODI UTILIZZATI

Le metodologie di analisi multivariata si pongono come obiettivo la sintesi delle osservazioni, ovvero la semplificazione della loro struttura (riduzione del numero delle variabili), l'ordinamento e il raggruppamento di osservazioni (classificazione), nello studio di interdipendenze tra le variabili.

In questa tesi vengono discusse e confrontate 5 diverse metodologie statistiche per risolvere problemi di classificazione:

- regressione multinomiale con risposta non ordinata;
- regressione multinomiale con risposta ordinata;
- alberi di classificazione;
- analisi discriminante lineare;
- analisi discriminante utilizzando la distanza AR.

Dopo aver definito le 3 classi e i vettori di osservazioni $\underline{x}_1, \underline{x}_2, ... \underline{x}_n$, relativi agli *n* soggetti del campione, si vuole costruire una funzione, detta regola di assegnazione, che a ciascun vettore $\underline{x}_i$ associ una delle 3 classi, sulla base di un criterio ragionevole.
In tutti i 5 metodi di classificazione si parte da una base dati, costituita da 18464 individui, dei quali si conosce a priori la corretta classificazione. Si può decidere di dividere questo data-set in 2 sottoinsiemi, in modo da usare il primo per creare una regola di classificazione e il secondo per verificare la validità del criterio, confrontando le previsioni svolte con la classificazione delle unità conosciuta a priori (AZZALINI, SCARPA, 2004). I 2 sottoinsiemi prendono rispettivamente il nome di *training set* e *validation set*. Entrambi sono frutto di una selezione casuale: il *training set* comprende 2/3 delle osservazioni mentre il *validation set* 1/3.
L'analisi discriminante lineare e il modello multinomiale sono metodi parametrici, in cui bisogna identificare dei parametri con i quali costruire la soglia di discriminazione, per poter poi effettuare la classificazione. L'albero di classificazione fa parte delle procedure non parametriche, cioè non presuppone la conoscenza della funzione

parametrica da utilizzare per stimare la variabile risposta, bensì utilizza una funzione approssimante a gradini.

3.1 MODELLO MULTINOMIALE CON VARIABILE RISPOSTA NON ORDINATA

Il modello può essere descritto come segue. Si suppone che un individuo i possa essere assegnato a una tra j possibili alternative o classi, ognuna delle quali ha una probabilità associata che dipende da alcune variabili esogene osservate sull'individuo. La variabile casuale che rappresenta la scelta dell'individuo *i*-esimo ha distribuzione multinomiale (FRANSES, PAAP, 2001):

$$Y_i \sim MN(1;\pi_1,...\pi_J)$$

Tale distribuzione implica che la probabilità che venga scelta l'alternativa *j* è $P[Y_i = j] = \pi_j$, con *j=1…J* e $\pi_1 + \pi_2 + ... + \pi_J = 1$.

Ciascuna probabilità è legata a un vettore *1x(q+1)* di variabili esogene $X_i = (1, x_{1,i} ... x_{q,i})$ osservate per ogni soggetto *i*, da una relazione del tipo:

$$\pi_j = F(X_i \beta_j) = F(\mu_{ij})$$

dove β_j è un vettore di coefficienti stimati per il gruppo *j,* che quantificano l'effetto di ciascuna esplicativa per quel gruppo sulla probabilità π_j. Rispetto al modello di regressione classico il predittore lineare μ_{ij} e il valore atteso condizionato $E(Y \mid X_i)$ non coincidono.

Dal momento che la probabilità associata a ciascuna classe deve essere compresa tra 0 e 1 e la somma di tutte le probabilità deve sommare a 1, una scelta idonea per F è la funzione *multilogit*.

In questo caso la probabilità che l'individuo i scelga l'alternativa *j* diventa:

$$w_i = P[Y_i = j \mid X_i] = \frac{exp(X_i\beta_j)}{\sum_{l=1}^{J} exp(X_i\beta_l)} \text{ per } j = 1,...J$$

La funzione legame tra predittore e valore atteso condizionato è il *logit*:

$$mlogit(w_i) = X_i\beta_j \qquad \text{dove } mlogit(\underline{x}) = \left[log\left(\frac{x_j}{x_0}\right)\right] \text{ per } j = 1...J \quad \text{(AZZALINI, SCARPA, 2004)}$$

Per evitare il problema dell'identificazione dei coefficienti, è indispensabile che una categoria venga scelta come base. Questo può essere fatto assumendo una "classe zero", ossia:

$$P[Y_i = 0 \mid X_i] = \frac{1}{1 + \sum_{l=1}^{J-1} exp\left(X_i\beta_l\right)}$$

Mentre per le altre classi:

$$P[Y_i = j \mid X_i] = \frac{exp\left(X_i\beta_j\right)}{1 + \sum_{l=1}^{J-1} exp\left(X_i\beta_j\right)} \quad \text{per } j = 1,...J-1$$

Non è possibile dare un'interpretazione diretta dei parametri β, perché l'effetto marginale delle *x* sulla risposta è chiaramente una funzione non lineare nei parametri. Come accade per il modello binomiale, si può dare un'interpretazione in termini di *odds ratio*.

L'odds ratio della categoria *j* rispetto alla categoria scelta come base, è definito come:

$$\frac{P[Y_i = j \mid X_i]}{P[Y_i = 0 \mid X_i]} = exp(X_i\beta_j)$$

ciò significa che l'effetto dei parametri β_j è di tipo esponenziale rispetto alla classe base.

I coefficienti $\beta_1 ... \beta_J$ vengono stimati massimizzando la funzione di verosimiglianza. Definendo una variabile dicotomica d_{ij}, che assume valore 1 quando il soggetto sceglie l'alternativa *j* e valore 0 altrimenti, la funzione di log verosimiglianza risulta:

$$\sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij} log P(Y_i = j \mid X_i)$$

e va massimizzata derivando rispetto a $\beta_1 ... \beta_J$.

3.2 MODELLO MULTINOMIALE CON VARIABILE RISPOSTA ORDINATA

Il modo più semplice per introdurre il modello di regressione "multinomiale" ordinato è pensare alla variabile risposta Y come determinazione di una variabile latente Y^*.
Si suppone che questa variabile sia legata a un vettore *1x(q+1)* di variabili esplicative X_i, dove l'indice *i* indica l'individuo *i-esimo:*

$$Y_i^* = X_i \beta + \varepsilon_i$$

Per il momento non viene specificata la distribuzione di ε_i.
Il dominio di Y non è continuo, ma categorizzato in J classi, con $J > 2$. A differenza del modello multinomiale a risposta non ordinata, si assume che la variabile risposta abbia un ordinamento naturale nelle alternative (es. basso, medio, alto) e al crescere del predittore lineare $X_i \beta$ corrisponda un livello più elevato per la risposta Y. Più formalmente il modello può essere scritto in questo modo:

$$Y_i = 1 \text{ se } \alpha_0 < Y_i^* < \alpha_1$$
$$Y_i = j \text{ se } \alpha_{j-1} < Y_i^* < \alpha_j \quad \text{per } j = 2,...J-1$$
$$Y_i = J \text{ se } \alpha_{J-1} < Y_i^* < \alpha_J$$

Dove $\alpha_0, ... \alpha_J$ rappresentano delle soglie ignote che servono per determinare a quale gruppo j appartiene Y e vanno stimate assieme ai parametri β. Non conoscendo il

campo di variazione della variabile latente la soglia minima α_0 viene posta a $-\infty$ e la massima α_J a $+\infty$.

Per garantire l'ordinamento, le soglie devono soddisfare la disuguaglianza: $\alpha_0 < \alpha_1 < \ldots < \alpha_J$.

Le equazioni indicate possono essere riassunte dicendo che l'individuo *i* viene assegnato alla categoria *j* se $\alpha_{j-1} < Y_i^* < \alpha_j$, quindi va determinata la probabilità che la variabile risposta cada in quell'intervallo (FRANSES, PAAP, 2001):

$$P[Y_i = j \mid X_i] = P[\alpha_{j-1} < Y_i^* < \alpha_j] = P[\alpha_{j-1} - X_i\beta < \varepsilon_i < \alpha_j - X_i\beta]$$
$$= F(\alpha_j - X_i\beta) - F(\alpha_{j-1} - X_i\beta)$$

per $j = 2 \ldots J-1$, e

$$P[Y_i = 1 \mid X_i] = F(\alpha_1 - X_i\beta) \quad \text{e} \quad P[Y_i = J \mid X_i] = 1 - F(\alpha_{J-1} - X_i\beta).$$

Sono le probabilità per la prima e l'ultima classe.

Questa specificazione implica che i parametri α e β non sono identificati. Per ovviare a questo problema la soglia che divide il primo gruppo dal secondo viene posta uguale a zero: $\alpha_1 = 0$.

Esistono diverse scelte possibili per la funzione di ripartizione, ma quelle maggiormente utilizzate sono la funzione di ripartizione della normale standard (*ordered probit model)* e la funzione di ripartizione logistica standard (*ordered logit model)*.

In entrambi i casi la varianza del termine d'errore σ_ε^2 è impostata al valore 1.

C'è da notare che i coefficienti β, che quantificano l'effetto di ogni esogena sulla risposta, non sono diversi per ogni classe *j*, come accadeva invece per il modello multinomiale non ordinato. Il numero di parametri da stimare è dato quindi dal numero di variabili esplicative, unito alle soglie α_j. Nel caso di 3 gruppi c'è un'unica soglia da stimare:

$$Y_i = \begin{cases} 1 \text{ se } Y^*_i \leq 0 \\ 2 \text{ se } 0 < Y^*_i \leq \alpha_1 \\ 3 \text{ se } Y^*_i > \alpha_1 \end{cases}$$

Anche per questo modello, i parametri vengono stimati con il metodo della massima verosimiglianza:

$$\sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij} log P(Y_i = j \mid X_i) \quad d_{ij} = 1 \text{ se l'individuo i sceglie l'alternativa j}$$

3.3 ALBERO DI CLASSIFICAZIONE

Un albero di classificazione è composto da un insieme di elementi detti nodi, che riportano un'affermazione di tipo logico sulle variabili esplicative *x*, relative a ciascun individuo analizzato (AZZALINI, SCARPA, 2004). Da ogni nodo parte un arco, ovvero una suddivisione in 2 vie: quella di sinistra viene percorsa se l'affermazione del nodo è vera, in caso contrario viene percorsa quella di destra. Il nodo da cui parte l'albero è il nodo radice, mentre gli elementi terminali sono chiamati foglie e riportano la classe dove l'individuo analizzato verrà classificato.

Ogni suddivisione corrisponde alla separazione dello spazio n-dimensionale delle variabili indipendenti X_i in rettangoli n-dimensionali non sovrapposti. Lo scopo è di creare dei rettangoli il più possibile *puri*: un rettangolo è puro se contiene soltanto osservazioni di una classe. Una misura di impurità, che indica la bontà della suddivisione ottenuta, è l'indice di Gini. Se *C* è il numero di classi e p_k la frazione di osservazioni nel rettangolo *R* che appartengono alla classe di indice *k*, allora l'indice di Gini è:

$$I(R) = \sum_{k=1}^{C} p_i(1 - p_i) = 1 - \sum_{k=1}^{C} {p_k}^2$$

Un indice equivalente, maggiormente utilizzato, è l'entropia:

$$H(R) = \sum_{k=1}^{C} p_k log_2 p_k$$

Se tutte le osservazioni in *R* appartengono alla stessa classe, allora $I(R)$ e $H(R)$ sono uguali a zero. Se le osservazioni in R sono equidistribuite fra le varie classi, allora entrambi gli indici raggiungono il loro valore massimo:

- l'indice di Gini: *(C-1)/C*;
- l'entropia: $log_2 C$.

Definita una misura per l'impurità di un rettangolo, la riduzione di impurità apportata da una suddivisione si può misurare come la differenza fra l'impurità del rettangolo originale e la media pesata dei rettangoli risultanti della sua suddivisione, dove i pesi sono le frequenze relative dei rettangoli generati rispetto a quello originale. Questa misura è detta *impurità attesa*. La sua formula è :

$$D = \sum_i \frac{|R_i|}{|R|} I(R_i)$$

|.| è il numero di osservazioni in un rettangolo.
Perciò la riduzione di impurità dovuta alla suddivisione è:

$$I(R) - D$$

La stessa formula può essere applicata con $H(R)$.
L'albero creato con questo criterio arriva fino alla separazione in classi completamente pure. Questo risultato sembra ideale, tuttavia non è adatto a prevedere nuove osservazioni.
Per questo soltanto una parte del campione viene utilizzata per la crescita, e la restante per la potatura. Producendo alberi impuri si tenta di lasciare spazio alla

generalizzazione su nuove osservazioni, al prezzo di un maggior rischio di errata classificazione.

L'algoritmo per la potatura valuta simultaneamente la misura di impurità e un costo complessità dell'albero:

$$C(\alpha) = D + \alpha J$$

dove D è l'impurità attesa scritta sopra e αJ è il costo complessità.

3.4 ANALISI DISCRIMINANTE

3.4.1 Analisi discriminante lineare

Si suppone di avere *K* gruppi $D_1, D_2, ..., D_K$: obiettivo dell'analisi discriminante è di collocare un individuo in uno di questi gruppi sulla base di un insieme di osservazioni $x = x_1, x_2, ..., x_p$ ad esso relative. Convenzionalmente tali variabili rilevate per ciascun individuo vengono chiamate *fattori*.

Ad ogni gruppo viene associata una funzione di densità definita sui *fattori (*AZZALINI, SCARPA, 2004*):* $p_1(x), ... p_K(x)$, oltre ad una serie di pesi $\pi_1, ... \pi_K$ rispetto al totale della popolazione ($\sum_k \pi_k = 1$).

Quindi la densità per la popolazione complessiva è:

$$p(x) = \sum_{k=1}^{K} \pi_k p_k(x)$$

A priori la probabilità che un soggetto non ancora classificato appartenga alla *k*-esima classe è data da π_k. Se per quel soggetto è noto il valore assunto da X, per il teorema di Bayes la probabilità a posteriori che quel soggetto appartenga al gruppo *k* è:

$$P(y = k \mid X = x_0) = \frac{\pi_k p_k(x_0)}{p(x_0)}$$

O equivalentemente il confronto tra classe *k* e classe *m* avviene sulla base di:

$$log\frac{P\{y=k \mid X=x_0\}}{P\{y=m \mid X=x_0\}} = log\frac{\pi_k}{\pi_m} + log\frac{p_k(x_0)}{p_m(x_0)}$$

quindi le varie classi vengono confrontate con la seguente *funzione discriminante*:

$$d_k(x_0) = log\pi_k + logp_k(x_0)$$

Quel valore di *k* che massimizza la funzione discriminante individua il gruppo nel quale l'individuo verrà classificato. Serve ora individuare la funzione di densità per calcolare d.

L'ipotesi parametrica più semplice è quella in cui ciascuna densità p_k è normale multipla con parametri dipendenti da *k*, tale per cui risulta:

$$p_k(x) = \frac{1}{(2\pi)^{p/2} det(\Sigma_k)^{1/2}} exp\left\{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right\}$$

nel caso semplificato, in cui tutte le matrici di varianza e covarianza siano uguali, la funzione discriminante prende la forma

$$d_k(x) = log\pi_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + x^T\Sigma^{-1}\mu_k$$

3.4.2 Analisi discriminante bayesiana

Nella statistica *bayesiana* si assume generalmente di poter attribuire delle probabilità a priori a ciò che costituisce l'oggetto di inferenza. Si parla di analisi discriminante *bayesiana* se è possibile assegnare in maniera soggettiva delle probabilità a priori $\pi_1, \pi_2, ..., \pi_K$ alle sottopopolazioni in modo tale che $\pi_k = P(x \in X_k)$ per *k=1…K*.

In tal caso, dette p_1,..., p_K le distribuzioni del carattere X nelle sottopopolazioni, è possibile utilizzare il teorema di Bayes per calcolare la probabilità a posteriori di ciascuna sottopopolazione. Si assume che le *k* distribuzioni *p*-dimensionali di X siano completamente specificate nella forma e nei parametri (per es. delle normali multiple come nel caso della analisi discriminante lineare). L'obiettivo è attribuire l'individuo con caratteristiche x alla sottopopolazione che ha maggiore probabilità di averlo generato, ovvero nel determinare il valore di *k** tale che:

$$p(X_{k^*} \mid x) = \max_k \left(\frac{\pi_k p_k(x)}{\sum_{k=1}^{K} \pi_k p_k(x)} \right)$$

3.5 DISTANZA AR

In molte situazioni le serie storiche possono essere raggruppate rispetto alla similarità della struttura dinamica che le ha generate.

Si consideri un processo $ARMA(p,q)$ Z_t di media zero e invertibile (CORDUAS M., PICCOLO D. , 2007):

$$\varphi(B) Z_t = \vartheta(B) a_t$$

dove a_t è white noise di media zero e varianza σ^2, B è l'operatore ritardo: $B^k Z_t = Z_{t-k}$, i polinomi $\varphi(B) = \phi(B)\Phi(B^s) = (1 - \phi_1 B ... - \phi_p B^p)(1 - \Phi_1 B ... - \Phi_P B^{SP})$ e $\vartheta(B) = \vartheta(B)\Theta(B^s) = (1 - \vartheta_1 B ... - \vartheta_q B^q)(1 - \Theta_1 B ... - \Theta_Q B^{SQ})$ non hanno fattori comuni e tutte le radici dell'equazione $\varphi(B)\vartheta(B) = 0$ sono in modulo maggiori di 1.

Inoltre si assume che tutte le componenti deterministiche vengono preventivamente rimosse dalla serie. Un processo di questo tipo può essere scritto nella forma

$$\pi(B) Z_t = a_t$$

Con $\pi(B) = \varphi(B)\vartheta^{-1}(B)$ e $\sum_{j=1}^{\infty}|\pi_j| < \infty$. Con questi assunti Piccolo(1984,1990) ha introdotto la distanza euclidea tra i coefficienti π della formulazione $AR(\infty)$:

$$d = \sqrt{\sum_{j=1}^{\infty}(\pi_{xj} - \pi_{yj})^2}$$

come una misura della dissimilarità strutturale tra due processi ARIMA X_t e Y_t. Questa distanza ha un'interessante interpretazione in termini di funzione di previsione. Per un processo $ARIMA(p,d,q)$ la stima della serie l passi in avanti $\hat{Z}_t(l)$, è data dalla somma dei valori passati di Z_t pesati coi coefficienti della rappresentazione $AR(\infty)$. Quindi se la distanza tra 2 processi è zero, significa che quei modelli produrranno le stesse previsioni.

3.5.1 Regola discriminante

La distanza autoregressiva (CORDUAS M., PICCOLO D., 2007) è uno strumento per classificare una serie di valori osservati rispetto al tempo all'interno di un gruppo caratterizzato da una specifica struttura stocastica di tipo $ARIMA(p,d,q)$.

Per spiegare l'algoritmo di calcolo viene considerato il caso più semplice in cui la popolazione è suddivisa in 2 classi, rappresentate dalle ipotesi H_1 e H_2. Ciascuna di esse è descritta dai coefficienti $\pi_1 = \{\pi_{1,k}, k = 1,...,m\}$ e $\pi_2 = \{\pi_{2,k}, k = 1,...,m\}$, ovvero le serie di coefficienti delle 2 strutture $AR(\infty)$, rispettivamente sotto l'ipotesi H_1 e H_2.

Quanto più l'insieme dei coefficienti stimati per una serie si avvicina ad una struttura H_j, tanto più si può pensare che quella serie sia stata generata da una struttura stocastica simile ad H_j. Questo è il criterio adottato da Piccolo (1984, 1990) per l'analisi discriminante su dati longitudinali.

Nel caso di 2 sole classi la suddivisione può avvenire sulla base di $D = d^2(\hat{\pi}_x, \pi_2) - d^2(\hat{\pi}_x, \pi_1)$. Se $D > 0$ significa che la distanza tra coefficienti stimati e coefficienti teorici del gruppo *2* è più grande rispetto alla distanza dai coefficienti del

gruppo *1*, quindi l'unità verrà assegnata alla classe più "vicina", ovvero la classe *1*; in caso contrario l'unità andrà al gruppo *2*. La distribuzione della trasformata:

$$\frac{\sqrt{n}}{2}\left\{D+(-1)^j d^2(\pi_1,\pi_2)\right\}\overset{a}{\sim} N(0,v_j^2)$$

dove $v_j^2=(\pi_1-\pi_2)'\Sigma_j(\pi_1-\pi_2), j=1,2$ con Σ_j matrice di varianza e covarianza dei coefficienti π_1 e π_2 è utile per calcolare le probabilità di errata classificazione:

$$P(2\,|\,1)=P(D\le 0\,|\,H_1)=\Phi\left(-\frac{\sqrt{n}}{2}\frac{d^2(\pi_1,\pi_2)}{v_1}\right) \qquad \text{e}$$

$$P(1\,|\,2)=P(D>0\,|\,H_2)=1-\Phi\left(\frac{\sqrt{n}}{2}\frac{d^2(\pi_1,\pi_2)}{v_2}\right)$$

Le 2 probabilità tendono a zero per $n\to\infty$ mentre quando $d^2(\pi_1,\pi_2)=0$ entrambe sono pari a 0.5.

Nel caso studiato le possibili classi di valore del cliente sono 3: valore basso (classe "a"), medio (classe "b") e alto (classe "c"). Con un adattamento è immediato ottenere la regola discriminante: la serie x_t viene assegnata alla classe "a" se contemporaneamente $D_{ab}<0$ e $D_{ac}<0$, dove: $D_{ab}=d^2(\hat{\pi}_x,\pi_a)-d^2(\hat{\pi}_x,\pi_b)$ e $D_{ac}=d^2(\hat{\pi}_x,\pi_a)-d^2(\hat{\pi}_x,\pi_c)$.

In maniera analoga ad altre situazioni (es. SHUMWAY, 1982) è possibile calcolare un limite inferiore per la probabilità di corretta classificazione. Utilizzando la disuguaglianza di Bonferroni e assumendo vera per esempio la classe "a":

$$P(a\,|\,a)=P(D_{ab}<0\cap D_{ac}<0\,|\,a)>1-P(D_{ab}>0\,|\,a)-P(D_{ac}>0\,|\,a)$$

Considerando la distribuzione approssimata precedentemente indicata, la probabilità di corretta classificazione è maggiore di:

$$1 - P(D_{ab} > 0 \mid a) - P(D_{ac} > 0 \mid a) = \Phi\left(\frac{\sqrt{n}}{2}\frac{d^2(\pi_a,\pi_b)}{v_a}\right) - 1 + \Phi\left(\frac{\sqrt{n}}{2}\frac{d^2(\pi_a,\pi_c)}{v_a}\right)$$

3.6 CALCOLO COEFFICIENTI AR

Il criterio della distanza AR è basato sulle differenze fra i coefficienti della rappresentazione $AR(\infty)$. Le 18.464 serie a disposizione non possono essere analizzate una ad una per l'individuazione dell'ordine di ciascun modello.
È stato scelto il modello stocastico $AR(2)$ sui residui delle rette adattate al capitolo 2, per i seguenti motivi:

1- la forte dipendenza tra i consumi di marzo 2006 e i due mesi precedenti;
2- il modello $AR(2)$ è già espresso nella formulazione $AR(\infty)$.

Per i modelli AR è cruciale l'ipotesi di stazionarietà della serie, che viene garantita se la correlazione fra osservazioni a tempi diversi è in modulo minore di uno.
A questo scopo sono già state eliminate alcune serie (paragrafo 2.3). Oltre a ciò, sulle serie rimaste è stato adattato un trend deterministico rappresentato da una funzione lineare del tempo ($\beta_0 + \beta_1 t$). In questo modo le serie dei residui sono stazionarie in media, e proprio su di esse verrà adattato il modello stocastico $AR(2)$.

4. ANALISI

In questo capitolo vengono adattati i metodi esposti nel capitolo 3 al campione oggetto di studio. I modelli vengono costruiti sul campione di stima e per ognuno di essi viene indicato il criterio seguito per individuare il miglior modello per la classificazione.

4.1 MODELLO MULTINOMIALE

4.1.1 Modello multinomiale con variabile risposta non ordinata

Il modello multinomiale con variabile risposta non ordinata può essere visto come un modello di regressione multinomiale. La differenza principale sta nel fatto che la variabile risposta è suddivisa in classi e la probabilità associata ad ognuna di esse è legata alle esplicative X dalla funzione legame *multilogit,* già indicata al capitolo 3.
Il modello viene costruito sul campione di stima.
Come prima analisi, nel predittore μ_{ij} vengono inserite soltanto le variabili non dipendenti dal tempo, vale a dire il piano tariffario, la zona di provenienza e l'età. Si esclude la variabile sesso vista l'indipendenza marginale con la variabile risposta e perché si pensa non possa apportare un'informazione utile per determinare il "valore del cliente" (capitolo 1).
Il metodo utilizzato per selezionare il miglior predittore lineare per il modello multinomiale è il criterio di informazione AIC , che suggerisce di minimizzare la seguente funzione obiettivo (AZZALINI,SCARPA,2004):

$$IC = -2logL + 2p$$

Tale funzione valuta l'adeguatezza del modello secondo 2 criteri contemporaneamente:

1. la massimizzazione della funzione di verosimiglianza;
2. la minimizzazione del numero di variabili esplicative (*p*).

Il *trade off* proposto dalla funzione permette di ottenere l'adattamento di un buon modello parsimonioso.

Nella tabella successiva si riportano le variabili che compongono il predittore lineare μ_{ij}, il numero *p* di parametri stimati, il criterio di informazione AIC e la devianza associata ad ogni modello:

Tabella 4.1: Elenco modelli multinomiali con variabili statiche e criteri per la selezione del miglior modello

	ESPLICATIVE INCLUSE	p	AIC	DEV_RES
1	*zona*piano*età*	160	18344.43	18056.43
2	*zona*piano+età*	82	18498.33	18346.33
3	*zona+piano*età*	40	18430.60	18350.60
4	*zona*età+piano*	34	18261.10	18193.10
5	*zona+piano+età*	26	18471.15	18419.15
6	*zona+piano*	24	18601.92	18553.92
7	*zona+età*	12	18640.41	18616.41
8	*piano+età*	18	18514.27	18478.27
9	*zona*	10	18765.02	18745.02
10	*piano*	16	18659.04	18627.04
11	*età*	4	18688.89	18680.89

Durante la procedura di stima, le variabili qualitative zona e piano vengono automaticamente codificate con delle variabili binarie (*dummy)*. Si utilizzano *m-1* dummy per ciascuna esogena qualitativa, dove *m* indica il numero di modalità. Quindi il *piano tariffario* ha sette coefficienti stimati sulle dummy associate e la *zona* ne ha quattro.

Lo stesso criterio viene applicato per la variabile risposta, perciò il numero totale di coefficienti stimati è il risultato di tutte le possibili combinazioni fra modalità della variabile risposta e modalità delle esplicative.

Il primo modello comprende 160 parametri, i successivi sono ottenuti togliendo una esplicativa (o un'interazione) alla volta, fino ad arrivare agli ultimi 3 che includono solo gli effetti marginali delle varibili statiche.

Il miglior modello sembra essere il quarto, perché fra quelli considerati ha AIC più basso.

A questo punto tale modello, costruito sul campione di stima, viene utilizzato per effettuare delle previsioni sul campione di verifica. Per avere una misura della capacità previsiva si utilizza la tabella di errata classificazione. In diagonale si trovano le osservazioni correttamente previste, mentre tutti gli elementi fuori dalla diagonale costituiscono l'errore totale di classificazione:

Multinomiale non ordinato solo var. statiche

		Osservati		
		a	b	c
Previst	a	4441	960	530
	b	7	11	12
	c	43	73	78

errore totale: 0,2640
falsi positivi & falsi negativi: 0.0269 0.9441

Dei 6155 clienti del campione di verifica, 5931 sono previsti nella prima classe. Più dei due terzi degli utenti analizzati hanno un consumo mensile basso, per questo era giusto attendersi una classificazione di questo tipo. Con le informazioni ottenute al capitolo 2 si può auspicare a un migliore risultato.

Aggiunta parametri legati al trend

A questo punto viene inserita nella regressione multinomiale anche la retta che descrive il trend dei consumi per il periodo che va da novembre 2004 a febbraio 2006. I tre parametri che identificano tale funzione sono l'intercetta β_0, il coefficiente angolare β_1 e l'indice di bontà di adattamento R^2 (stimati considerando errori *i.i.d.*, capitolo 2), che vengono impiegati in aggiunta alle variabili socio-demografiche descritte in precedenza allo scopo di migliorare la classificazione nel campione studiato. Quindi per ciascun utente ci sono 3 informazioni in più, per verificare se esiste un legame di dipendenza tra la variabile risposta e i parametri del trend deterministico.

Il miglior modello scelto dalla tabella 4.1 viene preso come base per aggiungere le componenti citate che, a differenza delle esplicative statiche, vengono inserite senza interazioni. Questa scelta è causata dalla eventuale difficoltà di interpretare

un'interazione tra le nuove variabili esaminate. In tab. 4.2 si osservano i miglioramenti apportati dai 3 nuovi regressori:

Tabella 4.2: Elenco modelli multinomiali e criteri per la selezione del miglior modello

	ESPLICATIVE INCLUSE	p	AIC	DEV_RES
1	*zona*età+piano+beta_0+beta_1+R^2*	40	10793.87	10713.87
2	*zona*età+piano+beta_0+beta_1*	38	10808.22	10732.22
3	*zona*età+piano+beta_0+R^2*	38	16586.07	16510.07
4	*zona*età+piano+beta_1+R^2*	38	17181.96	17105.96
5	*zona*età+piano+beta_0*	36	16637.17	16565.17
6	*zona*età+piano+beta_1*	36	17178.04	17106.04
7	*zona*età+piano+R^2*	36	18235.94	18163.94

Aggiungendo 3 variabili quantitative il numero di parametri non aumenta considerevolmente come era accaduto per le variabili qualitative. È evidente la notevole riduzione del criterio AIC che passa da 18.261 a 10.793. Questo sta ad indicare l'importanza dei nuovi regressori in funzione della variabile risposta.

Analisi parametri del modello

I parametri del modello vengono stimati massimizzando la funzione di verosimiglianza. I singoli β_{ij} (dove *i* identifica la variabile alla quale il coefficiente è riferito e *j* la classe della variabile valore del cliente), rappresentano la variazione nel log-rapporto di quote tra una generica classe *j* della variabile risposta e la classe presa come riferimento (nel caso studiato la classe *a*)

Coefficients:

	(Intercept)	*zonaiso*	*zonane*	*zonano*	*zonasud*	*eta*
b	-3,3448	-0,4839	-0,6021	-0,3077	-0,4059	-0,0048
c	-7,2524	0,0285	-0,6308	0,1326	0,0788	0,0044

	pianoB	*pianoC*	*pianoD*	*pianoE*	*pianoF*	*pianoG*
b	0,3961	0,4601	0,3447	0,8377	1,0917	-8,8406
c	0,3286	0,5795	0,0715	0,0122	1,3330	2,3271

	pianoH	*b0s*	*b1s*	*r2s*	*zonaiso:eta*	*zonane:eta*
b	-0,6366	0,0354	0,5224	0,4181	0,0051	0,0144
c	0,4048	0,0534	0,7695	1,1299	-0,0060	0,0150

	zonano:eta	*zonasud:eta*
b	0,0078	0,0106
c	-0,0038	-0,0038

Residual Deviance: 10713,8700
AIC: 10793,8700

Figura 4.1: Output di R per il modello multinomiale scelto

Si individuano le variabili che hanno maggior peso in valore assoluto nella determinazione della risposta. Per esempio, per un utente con alto valore rispetto ad uno con valore basso, ad una variazione unitaria di R^2 corrisponde un aumento moltiplicativo nel rapporto di quote (*odds ratio*) di:

$$\beta_{c,R\hat{}2} \cdot exp(\hat{\mu}_{i,c}) = 1{,}1299 \cdot exp(\hat{\mu}_{i,c})$$

Oltre all' R^2, altre variabili come i piani tariffari F e G rispetto agli altri piani, e il coefficiente angolare β_1, hanno forte impatto negli *odds ratio*.

Con la tabella di errata classificazione è possibile avere un'idea più chiara del miglioramento apportato dai nuovi regressori.

Multinomiale non ordinato +b0,b1,r2

		Osservati		
		a	b	c
Previsti	a	4297	528	48
	b	163	463	223
	c	31	53	349

errore totale: 0,1699
falsi positivi & falsi negativi: 0.0493 0.4960

L'errore totale è diminuito del 35% e i falsi negativi, presenti sulla parte superiore della diagonale, sono quasi dimezzati, passando dal 94,4% al 49,5%. L'informazione portata dalle tre nuove variabili è sicuramente indispensabile per ottenere una classificazione soddisfacente. Rispetto alla precedente classificazione, ancora una grossa fetta di consumatori del campione di verifica viene prevista nella prima classe, ma c'è un guadagno di circa 1000 utenti fra i 5931 che venivano prima inseriti nella classe "a".

4.1.2 Modello multinomiale con risposte ordinate

Nel modello "multinomiale" a risposte ordinate (FRANSES, PAAP 2005) si assume l'esistenza di una sola variabile latente per il livello di valore mensile. Anche in questo caso la media condizionale della variabile risposta viene categorizzata, suddividendo il

dominio $\Re$ in 3 sottoinsiemi. Per risolvere il problema di identificazione, la soglia che delimita la scelta tra la prima e la seconda classe è il valore zero, mentre per discriminare tra la seconda e la terza viene stimato il parametro α assieme ai coefficienti delle esplicative. Per semplicità si considera il modello *probit ordinato* (la distribuzione degli errori è $N(0,\sigma_\varepsilon^{\,2})$), imponendo $\sigma_\varepsilon^{\,2}$=1.

Nel caso di risposte non ordinate il miglior modello comprendeva l'interazione tra la zona di provenienza e l'età del soggetto, oltre agli effetti marginali del piano tariffario, intercetta, coefficiente angolare e R^2 della regressione lineare.

Coefficients:

	Value	Std. Error	t value
zonaiso	-0,1555	0,1634	-0,9516
zonane	-0,3482	0,0930	-3,7448
zonano	-0,0941	0,1510	-0,6234
zonasud	-0,1047	0,1255	-0,8348
eta	-0,0008	0,0018	-0,4424
pianoB	0,1642	0,0346	4,7428
pianoC	0,2534	0,0484	5,2378
pianoD	0,0946	0,1301	0,7272
pianoE	0,2694	0,1359	1,9819
pianoF	0,4866	0,0937	5,1954
pianoG	1,0465	0,0002	4297,7937
pianoH	0,1895	0,2865	0,6613
b0s	0,0165	0,0002	68,5179
b1s	0,2403	0,0038	63,1900
r2s	0,2364	0,0732	3,2315
zonaiso:eta	0,0007	0,0039	0,1806
zonane:eta	0,0080	0,0021	3,8308
zonano:eta	0,0010	0,0035	0,2824
zonasud:eta	0,0019	0,0030	0,6194

Residual Deviance: 11125,24
AIC: 11167,24

Figura 4.2: Output di R per il modello multinomiale a risposta non ordinata

Il numero di parametri è dimezzato al costo di un leggero aumento della varianza residua e del criterio AIC.

Il *t-value* verifica la significatività dei coefficienti. Come per il modello multinomiale a risposte non ordinate non è possibile valutare i parametri singolarmente per le variabili qualitative, ma è necessario considerarli congiuntamente.

L'effetto marginale dell'età non è significativo per il modello, ma assume importanza in alcune iterazioni con la zona di provenienza.

Infine le componenti che descrivono il trend lineare hanno un impatto importante nella determinazione della risposta, in particolare i parametri legati all'intercetta e al coefficiente angolare.

Adattamento

Tabella 4.3: tabelle di errata classificazione per il confronto fra multinomiale a risposta non ordinata e multinomiale a risposta ordinata

Multinomiale non ordinato +b0,b1,r2

Previsti \ Osservati	a	b	c
a	4297	528	48
b	163	463	223
c	31	53	349

errore totale: 0,1699
falsi positivi & falsi negativi: 0.0493 0.4960

Multinomiale ordinato statiche +b0,b1,r2

Previsti \ Osservati	a	b	c
a	4362	633	69
b	92	344	189
c	37	67	362

errore totale: 0,1766
falsi positivi & falsi negativi: 0.0399 0.5579

In termini previsivi il miglior modello resta quello con risposte non ordinate. A livello computazionale la stima di 20 parametri aggiuntivi non richiede un eccessivo tempo di elaborazione.

4.2 ALBERO DI CLASSIFICAZIONE

L'albero viene creato sul campione di stima, suddividendolo in due parti casuali: una per la crescita e una per la potatura.

L'algoritmo utilizzato per la crescita si basa sulla minimizzazione dell'entropia, realizzata scegliendo passo per passo tutte le variabili esplicative e tutti i possibili punti di suddivisione delle stesse, per poi selezionare quella variabile e quel punto di suddivisione che portano a un maggior guadagno in termini di impurità. La procedura termina nel momento in cui il numero delle foglie coincide col numero delle osservazioni.

Con la seconda parte del campione di stima si procede alla potatura, togliendo i nodi che non apportano una riduzione significativa dell'impurità per ciascun gruppo.

Considerando in un primo momento le sole esplicative statiche, si riporta a sinistra la crescita dell'albero e a destra la spezzata che permette di individuare il numero ottimale di nodi:

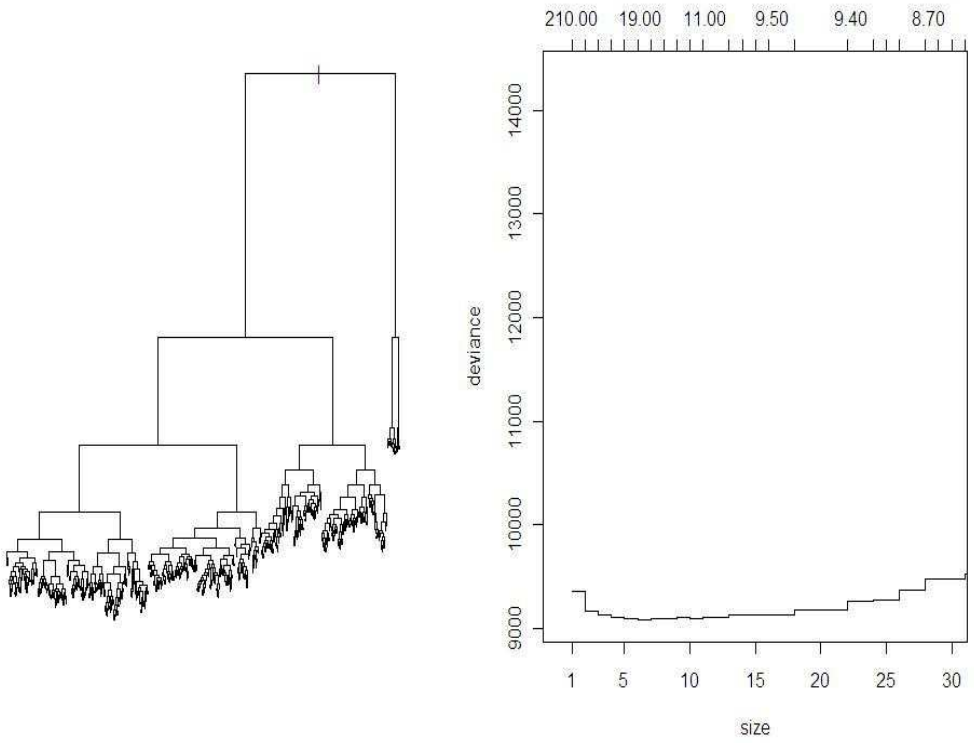

Figura 4.3: Crescita dell'albero e funzione C(.) per la potatura

Nella fase di stima sono state scelte 2 osservazioni come misura minima per generare un nodo. La lunghezza di ogni ramo è proporzionale alla diminuzione di impurità tra i nodi, maggiore è la lunghezza, maggiore è il guadagno in termini di impurità attesa; i rami più vicini alle foglie sembrano quindi inutili. Per decidere a che livello "tagliare" l'albero si utilizza la funzione illustrata in fig. 4.3 che riporta una misura di impurità per una serie di alberi stimati che si differenziano per il numero di nodi. La funzione è calcolata sulla seconda parte del campione di stima, ovvero quella riservata alla potatura.

In ordinata la *deviance*, nel caso di variabile risposta qualitativa, è intesa come la media delle entropie fra i nodi (paragrafo 3.2), una misura equivalente all'indice di Gini, più un costo complessità per l'albero:

$$C(\alpha) = \sum D_j + \alpha J$$

D_j è l'entropia all'interno di ogni foglia *j* e α è un parametro di penalizzazione. Per un fissato α si seleziona l'albero che minimizza $C(\alpha)$.

La funzione consiglia di utilizzare un albero con 6 foglie:

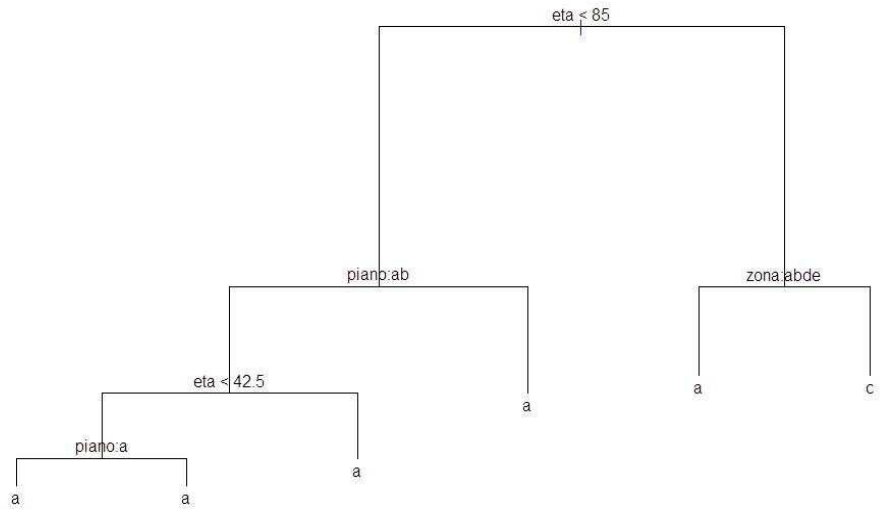

Figura 4.4: Albero con esplicative statiche

Quest'albero risulta migliore del precedente, ma tutte le suddivisioni della parte sinistra sono prive di effetto, dal momento che tutte terminano con la classe "a". La rappresentazione corretta con le variabili esplicative statiche comprende soltanto due nodi: il primo sfrutta la variabile età mentre il secondo la zona; viene esclusa la variabile piano, anche se durante le analisi descrittive al capitolo 2 presentava una forte dipendenza marginale con la risposta.

A differenza dei modelli parametrici non è possibile stabilire quali siano le variabili maggiormente legate alla risposta. La selezione avviene sulla base di una procedura automatica che non permette di capire l'ordine di importanza delle variabili mantenute nell'albero potato. Inoltre non è possibile fare inferenza sulle variabili scelte per la classificazione.

L'albero considera diverse interazioni tra le esplicative. Ogni foglia infatti è il risultato di più condizioni che devono essere soddisfatte simultaneamente. Inoltre, come si può

notare dal nodo di destra, il dominio di ogni variabile nominale viene suddiviso in opportuni sottoinsiemi con più elementi (es. zona:A,B,D,E) con lo scopo di ridurre l'indice di impurità fra le classi.

Guardando il grafico si nota la mancanza della classe intermedia, che comporterà un grosso errore nella tabella di errata classificazione.

Adattamento

Albero con le variabili statiche

		Osservati		
		a	b	c
Previst	a	4439	970	541
	b	0	0	0
	c	52	74	79

errore totale: 0,2660
falsi positivi & falsi negativi: 0,0276 0,9503

La matrice ottenuta è molto lontana dalla classificazione sperata. Nessuna persona viene catalogata nella classe centrale, causando delle percentuali molto elevate per gli errori di primo e secondo tipo.

Come accadeva per il modello multinomiale, le previsioni per la prima classe sono in numero troppo elevato rispetto alle classi "b" e "c", per cui i falsi negativi sono quasi il 100%.

Viene ripetuto lo stesso procedimento considerando ora anche le variabili che descrivono il trend dei consumi per ogni utente, già descritte nel secondo capitolo.

Con l'introduzione di intercetta e coefficiente angolare come variabili esogene per ciascun cliente, ci si aspetta una riduzione nella parte superiore della tabella di errata classificazione, in ragione del fatto che il valore del cliente dipende fortemente da queste nuove esplicative. Si ricorda infatti che la variabile risposta non è altro che il valore in euro, suddiviso in classi, dei consumi per l'ultimo mese a disposizione.

In seguito si riporta la spezzata che permette di ottenere il numero di nodi ottimale per il nuovo albero, con il principio utilizzato precedentemente.

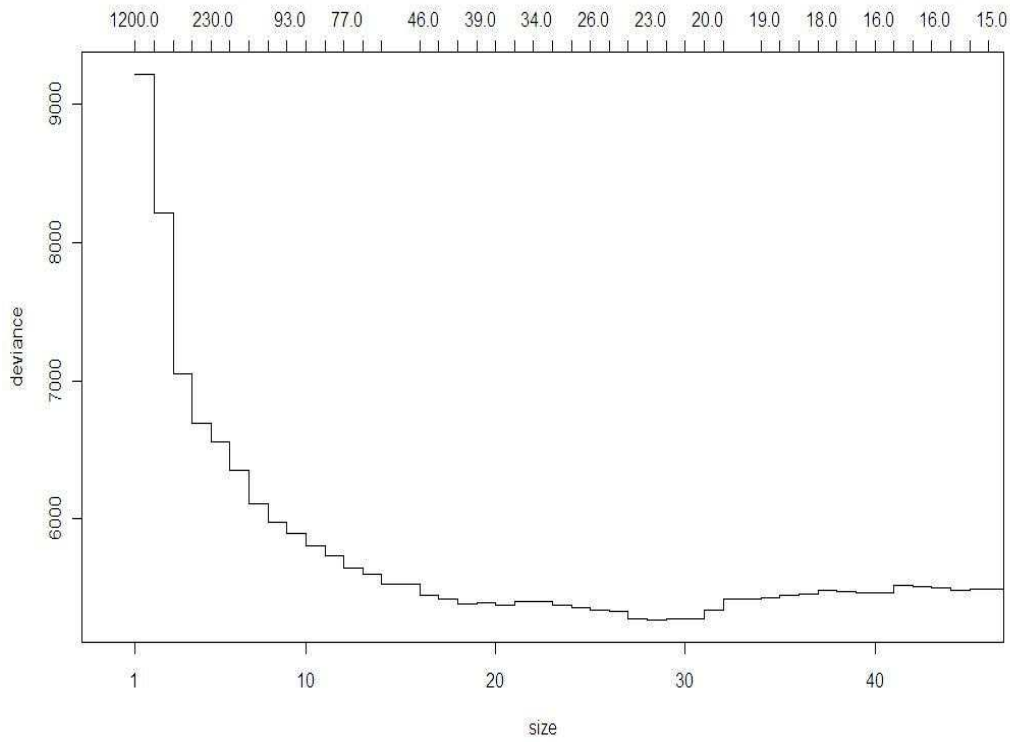

Figura 4.5: Algoritmo per la potatura dell'albero

In questo caso il criterio porta alla scelta di un albero con 28 foglie. Una parte dei nodi comportano però delle suddivisioni inutili, terminando con la stessa foglia. L'albero corretto ha 17 nodi e viene riportato in seguito:

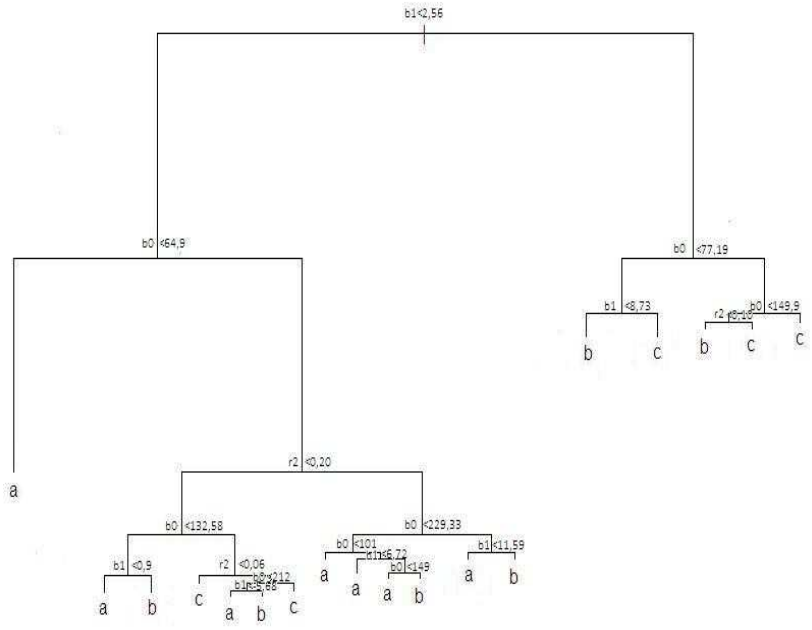

Figura 4.6: Albero con esplicative statiche e variabili legate al tempo

A differenza di prima tutte le 3 classi sono presenti, e probabilmente si otterrà una migliore previsione. Cadere nella parte destra dell'albero aumenta la probabilità di effettuare alti consumi per il mese preso in esame.

Dalla fig. 4.6 si vede che i nodi che portano una riduzione significativa di impurità sono i primi due, che riportano come affermazione logica: "beta_1 è minore di 2,56?" e "beta_0 è minore di 64,9?". Ciò sta a confermare l'importanza delle nuove esplicative introdotte, oltre al fatto che nessuna delle 17 scelte si basa su una variabile statica.

Adattamento

Tabella 4.4: Confronto tra alberi di regressione

Albero con le variabili statiche

Previsti	Osservati a	b	c
a	4439	970	541
b	0	0	0
c	52	74	79

errore totale: 0,2660
falsi positivi & falsi negativi: 0,0276 0,9503

Albero con var. statiche e coeff. del trend

Previst	Osservati a	b	c
a	4207	489	63
b	225	442	166
c	59	113	391

errore totale: 0,1812
falsi positivi & falsi negativi: 0,0787 0,4629

Nella precedente classificazione circa il 97% delle previsioni cadeva nella classe a, e nessuna osservazione apparteneva alla fascia centrale. Ora la situazione è decisamente migliorata: l'82% del campione di verifica è correttamente classificato e i falsi negativi si sono dimezzati, dal 95% al 46,29%.

Aggiunta dei coefficienti autoregressivi

Oltre ai parametri $\hat{\beta}_0, \hat{\beta}_1$ e R^2, sono disponibili i coefficienti $\hat{\phi}_1$ e $\hat{\phi}_2$ stimati sulla matrice dei residui, calcolata a seguito dell'adattamento del modello lineare. Con queste esplicative si cerca di modellare la componente stazionaria legata alle serie di ogni utente. Come per i parametri del modello lineare anche i valori stimati $\hat{\phi}_1$ e $\hat{\phi}_2$ sono disponibili per ogni soggetto studiato. L'aggiunta di queste due nuove informazioni serve per capire se utenti della stessa classe, individuata dalla variabile risposta, hanno anche dei coefficienti *AR* simili.
In seguito si riporta soltanto la matrice di confusione relativa al campione di verifica, ottenuta dopo aver costruito l'albero con le nuove variabili esogene:

Albero con l'aggiunta di ϕ_1 **e** ϕ_2

		Osservati		
		a	b	c
Previsti	a	4179	416	52
	b	277	535	226
	c	35	93	342

errore totale: 0,1786
falsi positivi & falsi negativi: 0,0791 0,4418

I risultati sono ulteriormente migliorati, il rapporto fra gli elementi della diagonale e i 6155 utenti del campione di verifica è diminuito dal 18,12% al 17,86%. I falsi positivi sono leggermente aumentati, ma i falsi negativi hanno registrato una riduzione del 5%.
Dall'elenco delle condizioni presenti su tutti i nodi, soltanto tre variabili vengono sfruttate per la costruzione dell'albero: $\hat{\beta}_0, \hat{\beta}_1, \hat{\phi}_1$.
Questo risultato consiglia di non tener conto delle variabili statiche per l'algoritmo di classificazione.
In effetti, seguendo quest'ipotesi, la matrice di errata classificazione diventa:

Albero con solo le variabili legate al tempo

		Osservati		
		a	b	c
Previsti	a	4132	373	47
	b	330	602	244
	c	29	69	329

errore totale: 0,1774
falsi positivi & falsi negativi: 0,0829 0,4163

In questo caso anche il coefficiente $\hat{\phi}_2$ è presente su alcune condizioni dell'albero. L'errore totale diminuisce nuovamente e i falsi negativi, che nel primo albero di fig. 4.4 erano il 95%, ora sono ridotti al 41%.

Di seguito si riporta l'albero costruito con le sole variabili legate all'evoluzione dei consumi nel periodo studiato:

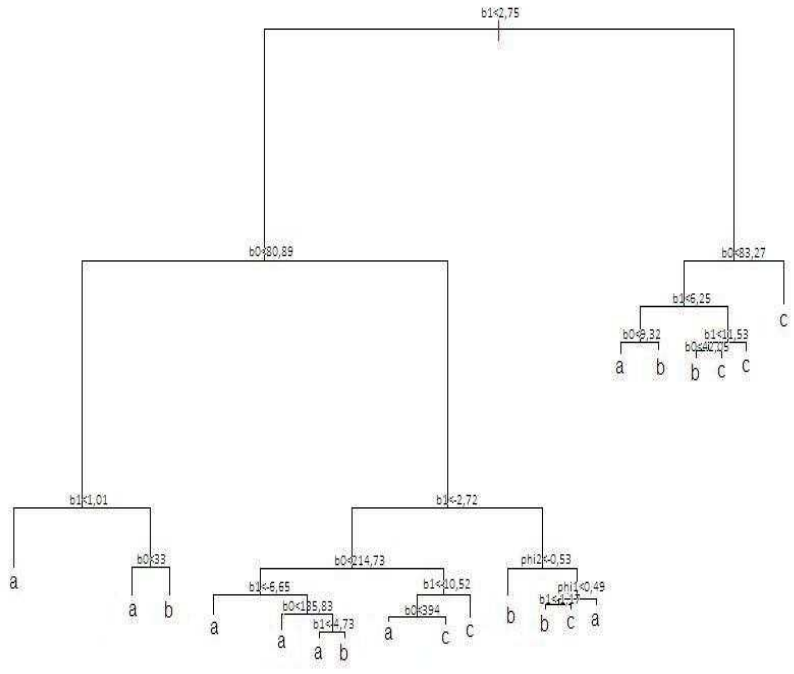

Figura 4.7: Albero con esplicative i coefficienti legati al tempo

Il numero ottimale di nodi è aumentato da 17 a 19 e le variabili legate ai residui del modello lineare sono presenti in 2 condizioni.

I clienti con valore medio e alto si collocano principalmente nella parte destra dell'albero di figura 2.7. Le condizioni che "fanno cadere" una nuova unità nelle foglie di destra richiedono che i coefficienti $\hat{\beta}_0, \hat{\beta}_1$ stimati sugli utenti restino al di sopra di determinate soglie. I due rami che portano a una maggior riduzione della *deviance* pongono come soglia il valore 2,75 per $\hat{\beta}_1$ e il valore 80,89 per $\hat{\beta}_0$. Questa è sicuramente un'informazione rilevante per capire se un nuovo cliente apporterà un valore medio/alto.

4.3 "SEGMENTAZIONE CLASSICA"

Generalmente in azienda non vengono impiegati dei metodi per sintetizzare l'informazione apportata dalle serie storiche studiate, ma si utilizzano direttamente i valori che le compongono, ed eventualmente dei tassi di variazione calcolati sui valori di due mesi successivi. Questo approccio, rispetto a quelli dei precedenti paragrafi, annulla di fatto il tempo necessario alla stima dei modelli, tuttavia se si hanno a disposizione serie storiche molto lunghe, l'algoritmo di classificazione potrebbe richiedere molto tempo per essere elaborato.
Nel campione esaminato le serie hanno soltanto 16 valori, quindi può essere ragionevole applicare questa "segmentazione classica" utilizzando l'analisi discriminante lineare e in seguito l'analisi discriminante logistica.

4.3.1 Analisi discriminante lineare

Nell'intero campione il 72,57% sono clienti con valore basso, 17,97% con valore medio e 9,45% con valore alto (dopo aver escluso le serie con molti zeri, par. 2.3). Il campione di stima, utilizzato per stimare l'algoritmo dell'analisi discriminante, è risultato di un'estrazione casuale dal data-set totale. In questo sottocampione, i pesi $\pi_1, \ldots \pi_K$ (par. 3.3.1) sono leggermente alterati rispetto all'intero campione. In un primo momento si utilizzano comunque i pesi del *training set* per la stima, poi gli stessi verranno modificati per verificare l'eventuale miglior adattamento (par. 3.3.2).

Stima

Prior probabilities of groups:

a	*b*	*c*
0,7238	*0,1848*	*0,0914*

Group means:

	nov04	*dic04*	*gen05*	*feb05*	*mar05*	*apr05*	*mag05*
a	*39,7965*	*52,81749*	*47,79571*	*39,64811*	*45,11258*	*24,12156*	*43,91234*
b	*85,96791*	*110,14901*	*100,07429*	*88,61407*	*102,59429*	*56,76835*	*106,4611*
c	*169,41333*	*213,16533*	*192,46044*	*175,39644*	*206,49956*	*112,42489*	*214,83644*
	giu05	*lug05*	*ago05*	*set05*	*ott05*	*nov05*	*dic05*
a	*41,61443*	*42,11449*	*37,55034*	*35,12426*	*32,50297*	*27,88169*	*29,75261*
b	*104,91429*	*107,07121*	*97,90945*	*102,33714*	*101,33363*	*94,18681*	*108,42462*
c	*214,75911*	*224,22133*	*198,82933*	*215,16*	*229,01511*	*219,528*	*247,87556*
	gen06	*feb06*					
a	*24,65103*	*18,76866*					
b	*100,10418*	*89,69231*					
c	*244,05689*	*228,19022*					

Coefficients of linear discriminants:

	LD1	*LD2*
nov04	*0,000914104*	*0,00069693*
dic04	*-0,000101265*	*0,001092411*
gen05	*-0,00017471*	*-0,001478823*
feb05	*0,000634027*	*-0,000885979*
mar05	*0,000692838*	*0,003971325*
apr05	*-0,001526322*	*-0,00844705*
mag05	*0,001772849*	*-0,001831617*
giu05	*-0,001310595*	*-0,003708806*
lug05	*0,001042163*	*0,008045463*
ago05	*0,000224697*	*-0,007454387*
set05	*0,000203809*	*-0,009405307*
ott05	*0,000764283*	*0,007059312*
nov05	*-0,001197161*	*0,008881606*
dic05	*0,002309006*	*-0,012450109*
gen06	*0,002579984*	*0,003683844*
feb06	*0,013047553*	*0,006843896*

Figura 4.8: Output di R per l'analisi discriminante con esplicative i valori della serie dei consumi

Per ogni mese è indicata la media del valore del cliente per le tre classi considerate. Per buona parte del periodo osservato le medie della classe “b” e “c” (valore medio e alto) sono pressoché invariate. Questo suggerisce che il coefficiente angolare stimato per ogni utente potrebbe non essere un’indicazione rilevante.

Viene valutata la capacità previsiva sul campione di verifica:

An. discr. con i consumi mensili

		Osservati		
		a	b	c
Previst	a	4423	678	51
	b	50	317	206
	c	18	49	363

errore totale: 0,1709
falsi positivi & falsi negativi: 0,0241 0,5789

L'errore totale è il più basso trovato fino a questo punto, tuttavia sono aumentati i falsi negativi.

4.3.2 Analisi discriminante bayesiana

Ora viene costruito lo stesso algoritmo cambiando soltanto la distribuzione a priori delle tre classi. Per dare maggior peso alle fasce di clienti che generano un valore medio/alto è stata scelta una distribuzione uniforme discreta: $Ud(a,b,c)$.

Adattamento

An. discr. con i consumi mensili e a priori uniforme

		Osservati		
		a	b	c
Previsti	a	4135	250	7
	b	334	716	204
	c	22	78	409

errore totale: 0,1454
falsi positivi & falsi negativi: 0,0821 0,2907

L'indice più rilevante è ulteriormente diminuito con un guadagno in termini assoluti di 156 utenti sui 6155 totali del campione di verifica. I falsi positivi sono leggermente aumentati mentre i falsi negativi sono praticamente dimezzati.
Un risultato importante riguarda le celle di coordinate (a,c) e (c,a), che esprimono la capacità dell'algoritmo di distinguere tra un individuo molto profittevole da uno poco profittevole. Soltanto 7 utenti previsti nella classe "a" consumano molto nel mese di marzo 2006 e 22 previsti "c" hanno consumi bassi. Chiaramente l'errore commesso nel primo caso è più grave rispetto al secondo, cioè è più importante individuare i

"soggetti redditizi". Rispetto ai primi metodi implementati in questo capitolo gli individui sopra alla diagonale si sono ridotti fino al 29%.

4.3.2 Albero di classificazione con la "segmentazione classica"

Ripetendo la stessa metodologia applicata per l'analisi discriminante, si riporta l'albero di classificazione che ha come esplicative i consumi mensili per ogni cliente e di seguito la matrice di confusione.

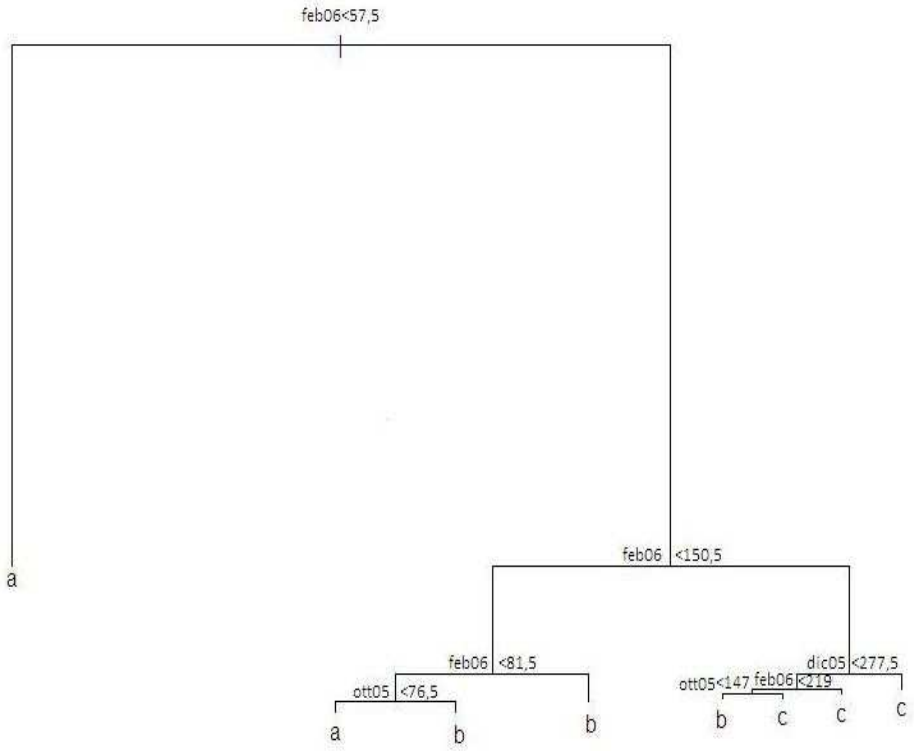

Figura 4.9: Albero con esplicative i valori delle serie dei consumi

Albero con i consumi mensili

		Osservati		
		a	b	c
Previst	a	4292	387	13
	b	181	593	240
	c	18	64	367

errore totale: 0,1467
falsi positivi & falsi negativi: 0,0511 0,4000

Non c'è stato un miglioramento rispetto all'analisi discriminante bayesiana. L'errore totale e i falsi negativi sono di poco aumentati. Nonostante ciò, l'albero di fig 4.9 suggerisce che i consumi del mese di febbraio 2006 ed ottobre 2005 contribuiscono ad una riduzione notevole della devianza, per questo motivo è interessante costruire un albero che sfrutta contemporaneamente questa informazione e le variabili stimate al capitolo 2.

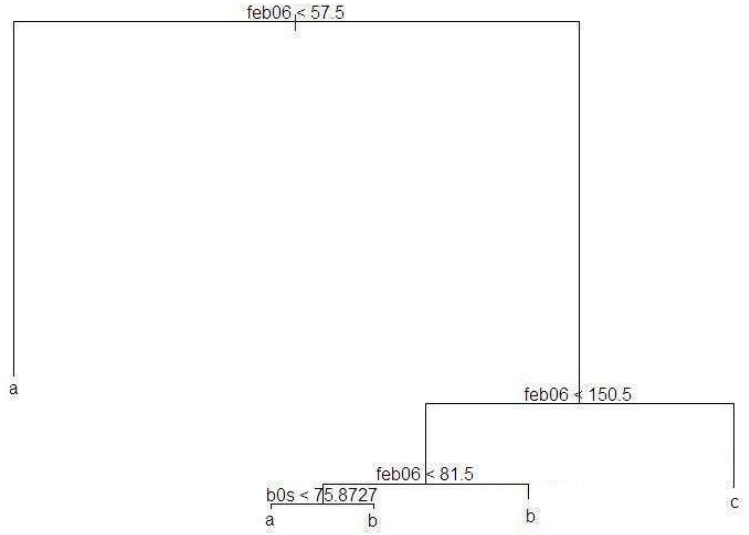

Figura 4.10: Albero "misto"

Albero con feb06 e ott06 più $\beta_0, \beta_1, \phi_1, \phi_2,$

		Osservati		
		a	b	c
Previst	a	4292	387	13
	b	173	557	180
	c	26	100	427

errore totale: 0,1428
falsi positivi & falsi negativi: 0,0581 0,3708

Quest'albero è leggermente migliore del precedente, ma anche se per la sua crescita si utilizzano $\beta_0, \beta_1, \phi_1, \phi_2$, soltanto il coefficiente β_0 viene mantenuto dopo la potatura. Ciò vuol dire che per questo campione di consumatori è importante conoscere soltanto il livello dell'intercetta e non:

- l'evoluzione dei consumi: crescente, decrescente o costante (β_1);
- l'andamento della componente stazionaria rappresentata dai residui del modello lineare (ϕ_1, ϕ_2).

5. DISTANZA AR

Nel corso del capitolo 2 è stato adattato un modello lineare alla serie dei consumi totali in modo da eliminare il trend deterministico e, ai residui risultanti, essendo stazionari in media, è stato applicato un modello stocastico $AR(2)$ per cogliere anche la componente stazionaria delle serie storiche.

L'idea è impiegare la distanza AR per ottenere una buona regola di classificazione sulla base della similarità della struttura dinamica delle serie. Ci si aspetta che i clienti che generano valori differenti abbiano anche diversi coefficienti AR.

Per utilizzare questo strumento è necessario calcolare nel campione di stima i coefficienti medi appartenenti alle 3 classi individuate dalla variabile valore del cliente. Per ogni utente vengono stimati i coefficienti $\hat{\phi}_1$ e $\hat{\phi}_2$ sulla serie dei residui, dopodiché ciascun utente verrà assegnato al gruppo che possiede i coefficienti medi $\overline{\hat{\phi}_1}$ e $\overline{\hat{\phi}_2}$ più prossimi (in termini di distanza euclidea) ai suoi coefficienti stimati. Nella tabella 5.1 vengono riportate le medie dei 3 gruppi:

Tabella 5.1: Stima delle medie per i coefficienti AR divisi per gruppo

	Coefficienti AR	Medie coeff.
Gruppo fatturati bassi	ϕ_1	0,1101
	ϕ_2	-0,1974
Gruppo fatturati medi	ϕ_1	0,0955
	ϕ_2	-0,1777
Gruppo fatturati alti	ϕ_1	0,1245
	ϕ_2	-0,1794

Non c'è grossa differenza fra i coefficienti medi delle 3 classi e ciò sta a significare che l'andamento stocastico della serie dei residui è simile per tutti i clienti studiati, indipendentemente dal "valore" che hanno per l'azienda. A questo punto vanno calcolate le distanze D_{ab}, D_{ac}, D_{bc} per tutti i 18464 utenti.

Come accennato al capitolo 3 l'unità verrà assegnata:

- alla classe "a" se $D_{ab} < 0$ e $D_{ac} < 0$;
- alla classe "b" se $D_{ab} > 0$ e $D_{bc} < 0$;
- alla classe "c" se $D_{ac} > 0$ e $D_{bc} > 0$.

Di seguito si verifica se questa regola costituisce un buon criterio per suddividere le unità del campione di verifica.

Adattamento

Distanza AR con ϕ_1 e ϕ_2

		Osservati		
		a	b	c
Previsti	a	899	211	112
	b	1920	464	259
	c	1672	369	249

errore totale: 0,7381
falsi positivi & falsi negativi: 0,7440 0,4494

L'errore totale è decisamente alto. L'algoritmo utilizzato classifica erroneamente quasi i 3/4 dei clienti del campione di verifica.

La tabella 5.1 indica che le medie delle variabili discriminanti utilizzate non sono molto diverse nei 3 gruppi, quindi diventa difficile stabilire delle soglie di separazione tra le classi basate su queste esplicative. Questo metodo non può essere confrontato con gli algoritmi del capitolo precedente, dal momento che sfrutta soltanto dei coefficienti stimati sulla serie dei residui per spiegare la variabile risposta. È già stato appurato che la componente di trend deterministico è un strumento indispensabile al fine di individuare una buona classificazione.

Per questo motivo la distanza D viene integrata con le esogene che descrivono il trend dei consumi. Come per i coefficienti $\hat{\phi}_1$ e $\hat{\phi}_2$, si calcolano nel campione di stima le medie dei coefficienti legati al trend: $\overline{\hat{\beta}}_0, \overline{\hat{\beta}}_1$.

Anche in questo caso l'unità viene assegnata alla classe che ha i 4 coefficienti medi più vicini ai coefficienti stimati su di essa.

Per rendere più chiara l'esposizione, viene riportato il calcolo della nuova distanza D per un individuo:

$$D_{ab} = d^2(\underline{\hat{\phi}}, \underline{\hat{\beta}}; \underline{\phi_a}, \underline{\beta_a}) - d^2(\underline{\hat{\phi}}, \underline{\hat{\beta}}; \underline{\phi_b}, \underline{\beta_b})$$

Dove ogni parametro indicato è un vettore di due elementi:

$\underline{\hat{\phi}} = (\hat{\phi}_1, \hat{\phi}_2)^T$; $\underline{\hat{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$; $\underline{\phi_a} = (\phi_{1,a}, \phi_{2,a})^T$; $\underline{\beta_a} = (\beta_{0,a}, \beta_{1,a})^T$

L'operazione è analoga per ottenere D_{ac} e D_{bc}.

Il principio seguito per la costruzione di questa distanza D è simile all'analisi discriminante di Fisher. Si cerca infatti di separare il più possibile i gruppi sulla base di una misura definita su quattro variabili esogene. Se questi gruppi differiscono rispetto ai valori assunti dalle esplicative è possibile definire una regola di decisione da applicare a una nuova unità statistica, di cui non si conosce la classe di appartenenza.

Il vettore $\begin{pmatrix} \underline{\phi} \\ \underline{\beta} \end{pmatrix}$ individua il centroide per ogni classe, dato dalle 4 medie dei coefficienti.

La nuova osservazione viene assegnata al centroide più vicino (con d^2 più basso) in modo da massimizzare la distanza tra le medie dei gruppi e minimizzare la dispersione delle unità all'interno di ogni gruppo. L'analisi discriminante di Fisher non richiede alcuna assunzione sulla forma distributiva delle variabili esplicative $\begin{pmatrix} \underline{\hat{\phi}} \\ \underline{\hat{\beta}} \end{pmatrix}$ che descrivono l'unità.

Adattamento

Distanza AR con ϕ_1, ϕ_2, β_0, β_1

		Osservati		
		a	b	c
Previsti	a	3612	482	135
	b	522	330	164
	c	357	232	321

errore totale: 0,3074
falsi positivi & falsi negativi: 0,2199 0,5454

La previsione sul campione di verifica è migliorata, tuttavia i metodi utilizzati nel capitolo 4 davano risultati decisamente più affidabili per classificare una nuova osservazione.

In seguito viene calcolata la probabilità di corretta classificazione con i risultati ottenuti per tutte le tre classi:

$$P(a \mid a) = \frac{3612}{4492} = 0{,}8041;$$
$$P(b \mid b) = \frac{330}{1044} = 0{,}3161;$$
$$P(c \mid c) = \frac{321}{620} = 0{,}5177;$$

Come per i metodi studiati al capitolo precedente i clienti più difficili da prevedere restano quelli della classe “b” e “c”.

6. CONFRONTO MODELLI

Nelle analisi dei capitoli 4 e 5 sono state presentate le tabelle di errata classificazione utilizzando di volta in volta variabili esplicative di diversa natura, allo scopo di non ripetere la stessa sequenza per tutti i modelli di classificazione studiati e risultare quindi ripetitivi. In questo capitolo vengono riassunti tutti gli esiti per ogni modello stimato, per poi dare una valutazione globale e scegliere il miglior metodo previsivo per il campione esaminato.
I modelli di classificazione vengono raggruppati in 5 categorie:

1. modelli con variabili statiche: le esplicative sono il piano tariffario, la zona di provenienza, l'età e le interazioni tra le stesse;
2. modelli con variabili statiche più i coefficienti legati al trend lineare e i $\hat{\phi}_1, \hat{\phi}_2$ stimati sulla serie dei residui del trend;
3. modelli con solo i coefficienti legati alle variabili longitudinali: $\hat{\beta}_0, \hat{\beta}_1, R^2, \hat{\phi}_1, \hat{\phi}_2$;
4. modelli con esplicative i consumi del periodo osservato;
5. modelli misti.

La tabella 6.1 riassume tutti i risultati ottenuti.

Tabella 6.1: indici di bontà di adattamento per i modelli stimati, ottenuti sul campione di verifica di 6155 utenti

	unità correttamente previste	falsi positivi	falsi negativi
Esplicative: zona, piano,età			
Modello multinomiale a risposta non ordinata	73,60%	2,69%	94,41%
Modello multinomiale a risposta ordinata	73,37%	2,80%	95,03%
Albero	73,40%	2,76%	95,03%
Analisi discriminante	72,98%	0,47%	98,68%
Esplicative: zona, piano, età, β_0, β_1, R^2, ϕ_1, ϕ_2			
Modello multinomiale a risposta non ordinata	83,22%	4,66%	49,57%
Modello multinomiale a risposta ordinata	82,70%	3,88%	54,63%
Albero	81,88%	7,87%	46,29%
Analisi discriminante	80,65%	3,31%	64,85%
Esplicative: β_0, β_1, R^2, ϕ_1, ϕ_2			
Modello multinomiale a risposta non ordinata	82,99%	4,70%	50,34%
Modello multinomiale a risposta ordinata	82,83%	3,82%	54,21%
Albero	82,26%	8,29%	41,63%
Analisi discriminante	80,42%	2,97%	66,25%
Esplicative: valori della serie dei consumi			
Modello multinomiale a risposta non ordinata	85,80%	4,94%	38,77%
Modello multinomiale a risposta ordinata	85,81%	4,42%	42,12%
Albero	85,33%	5,11%	40,00%
Analisi discriminante	82,91%	2,40%	57,89%
Analisi discriminante con a priori uniforme	85,46%	8,21%	29,07%
Esplicative: febo6, otto6, β_0, β_1, R^2, ϕ_1, ϕ_2			
Albero	85,72%	5,81%	37,08%
Distanza AR			
con ϕ_1 e ϕ_2	26,19%	74,40%	44,94%
con ϕ_1 ,ϕ_2, β_0, β_1	69,26%	21,99%	54,54%

La variabile risposta valore del cliente, essendo una trasformata dei consumi totali mensili per ogni soggetto, ha un forte legame di dipendenza con i coefficienti che descrivono le serie storiche dei consumi. I primi modelli, ottenuti con le sole variabili statiche, non suddividono opportunamente il campione di verifica, infatti i falsi negativi sono attorno al 95% per tutte le quattro metodologie applicate.

I falsi negativi sono quella "fetta" di consumatori che sono stati classificati con valore medio o basso, quando invece fanno parte della fascia medio/alta della variabile risposta:

$$falsi\ neg. = \left(\frac{P(a \mid c) + P(a \mid b) + P(b \mid c)}{P(a \mid c) + P(a \mid b) + P(b \mid c) + P(b \mid b) + P(c \mid c)} \right)$$

Questa percentuale rappresenta l'errore più grave per il problema analizzato, dal momento che l'azienda preferisce riservare gli investimenti pubblicitari per la porzione di clienti profittevoli che, nonostante siano in numero inferiore, da soli apportano oltre il 70% dei ricavi (dato ottenuto al capitolo 2). Non prevedere correttamente questa fascia di consumatori, significa perdere l'opportunità di conseguire un maggior guadagno.

Con l'aggiunta dei parametri legati al trend, la percentuale di falsi negativi si riduce notevolmente, passando da 95,78% al 53,83% (dati medi per i quattro metodi). I consumatori prima mal classificati si spostano sulla diagonale della matrice di confusione, aumentando così le unità correttamente previste, dal 73,34% all'82,36%.

Togliendo le variabili esplicative statiche dal predittore lineare, gli indici non si discostano molto dai valori precedentemente indicati. In ambito statistico, a parità di risultato, è sempre preferibile utilizzare l'algoritmo meno complesso, in cui vengono considerati meno parametri.

Per verificare se effettivamente i modelli con le sole variabili legate al tempo fossero i migliori in termini previsivi, nel paragrafo 4.3 è stata realizzata la segmentazione generalmente utilizzata in azienda. Gli indici si sono ulteriormente ridotti, con un guadagno medio del 3,5% sulla percentuale di unità correttamente previste.

Il dato che più sorprende sono i falsi negativi, che nel caso dell'analisi discriminante bayesiana si abbassano al di sotto del 30%. La percentuale di falsi positivi è maggiore rispetto ai modelli precedenti, tuttavia questo tipo di errore viene tollerato dall'azienda, il cui principale obiettivo è abbassare la soglia dei clienti profittevoli erroneamente classificati. Sempre nel paragrafo 4.3, l'albero di classificazione stimato è stato un utile strumento per individuare quali variabili esplicative vengono maggiormente utilizzate nelle condizioni di ciascun nodo. Integrando i parametri stimati sulle serie con i consumi effettivi del mese di febbraio 2006 e ottobre 2005 l'algoritmo ottenuto effettua una buona classificazione, anche se leggermente peggiore dell'analisi discriminante bayesiana.

L'impiego di un modello senza alcun parametro stimato sulle serie storiche è meno interpretabile rispetto ad uno che include le componenti che descrivono l'evoluzione dei consumi.

Inoltre è difficile pensare ad una stabilità temporale per un modello di questo tipo. Nel caso studiato i consumi di marzo 2006 hanno una forte dipendenza con il mese precedente, ma la logica potrebbe non essere la stessa per i mesi successivi. In quest'ottica potrebbe essere interessante ristimare l'albero con le stesse variabili calcolate su un nuovo insieme informativo, composto dai mesi che vanno da dicembre 2004 a marzo 2006, per prevedere il valore del cliente del mese di aprile. Nel caso in cui le variabili legate alle serie dei consumi vengano incluse nel predittore, tale modello potrebbe essere più affidabile per le successive previsioni.

Utilizzare sette variabili esogene, ridotte a tre a seguito della potatura dell'albero, rappresenta indubbiamente un vantaggio rispetto a considerarne diciassette (consumi mensili del periodo di osservazione).

La classificazione basata sulla distanza AR evidentemente non può essere adattata su un campione così numeroso. Sarebbe necessario studiare l'autocorrelazione sulla serie dei consumi di ogni utente, un lavoro impensabile per oltre 32000 consumatori. In ogni caso le serie dovrebbero comprendere un periodo di osservazione più ampio, in modo che le bande di confidenza per ACF e PACF (*auto-correlation function e partial auto-correlation function*) siano più contenute e permettano di individuare la corretta struttura di dipendenza per ciascuna serie.

Per gli utenti del data-set analizzato, non è rilevante sapere se i consumi sono crescenti/decrescenti o con una specifica dipendenza stocastica, ma è sufficiente conoscere il livello di consumi all'inizio del periodo, individuato da $\hat{\beta}_0$, e per altri due mesi, ottobre 2005 e febbraio 2006.

7. CONCLUSIONI

In questa tesi sono stati costruiti alcuni modelli statistici per prevedere il *customer value*, una misura che quantifica la redditività mensile di un utente. L'obiettivo principale era ottenere una buona previsione sfruttando gli strumenti classici per l'analisi delle serie storiche, ovvero i modelli $ARMA(p,q)$ e la regressione lineare per cogliere la componente di trend.

Sulla base degli indici ottenuti dalle tabelle di errata classificazione, il miglior modello non comprende buona parte dei parametri stimati nel capitolo 4. Tuttavia, i metodi che considerano soltanto i parametri legati al tempo hanno delle percentuali di errata classificazione abbastanza simili a quelle del miglior modello.

Per una variabile risposta di tipo fattore a tre livelli, ottenere un errore globale al di sotto del 15% può essere considerato un buon risultato. Rispetto ad una variabile risposta binaria la probabilità di incorrere in un errore di primo tipo (falsi negativi) o di secondo tipo (falsi positivi) è sicuramente più alta.

Si potrebbe allargare l'analisi considerando altri tipi di modelli per l'analisi delle serie, come ad esempio il filtro di Kalman, un algoritmo che prevede un aggiornamento automatico dei parametri sulla base dei valori assunti dalla serie.

Un'analisi di questo tipo potrebbe portare a risultati migliori, tuttavia aumenterebbe la complessità dei modelli. L'albero ottenuto è facilmente interpretabile, soprattutto perché utilizza poche variabili esplicative per la classificazione ed ha soltanto quattro nodi decisionali. Con queste quattro informazioni quasi nove clienti su dieci vengono assegnati alla classe alla quale realmente appartengono.

Altrimenti sarebbe interessante analizzare le singole serie di ogni cliente del campione. Potrebbe risultare che soltanto alcune serie influenzino la previsione, oppure ci siano dei comportamenti diversi per ogni tipologia di consumo (sms, mms, ch. vs. stesso, ch. vs. altri, ch. vs. fisso), che considerati assieme portino a una segmentazione migliore.

L'analisi potrebbe essere più dettagliata, ma andrebbe studiato un metodo per eliminare le serie che hanno esclusivamente valori nulli, in modo da poter stimare dei modelli stocastici di tipo $ARMA(p,q)$ sulle serie restanti.

APPENDICE: COMANDI IN R

Questa appendice è dedicata alla formulazione per R di alcuni modelli stimati nel corso della tesi. Vengono proposti alcuni comandi utilizzati per costruire i modelli dei capitoli 4 e 5.
Il testo di riferimento consultato è Venables, Ripley (2002) e le risorse on line fornite dal programma. Per la spiegazione dei modelli si rimanda invece al testo.

Modello multinomiale a risposta non ordinata

Il modello è stato stimato con il comando `multinom()` appartenente alla libreria `nnet`. Di seguito viene riportato un esempio sulla stima del modello multinomiale con le sole variabili statiche, costruito al paragrafo 4.1.1:

```
fit<-multinom(valore~zona*eta+piano,data=st.ca)
```

st.ca è il data-set che contiene i valori delle variabili statiche (zona, piano, eta) e la variabile risposta (valore) per il campione di stima, estratto casualmente dal campione totale di 32524 utenti. Il simbolo * sta ad indicare che nel modello vengono stimati dei parametri per l'interazione tra le esplicative zona ed eta.
Per verificare la capacità previsiva del modello è stata utilizzata la funzione `tabella.sommario()` che genera una matrice di errata classificazione tra le previsioni del modello stimato e i dati osservati (SCARPA A.A. 2008 2009, Lezioni di Tecniche statistiche di classificazione).

```
tabella.sommario <- function(previsti, osservati){
  n <-  table(previsti,osservati)
  err.tot <- 1-sum(diag(n))/sum(n)
  fn <- (n[1,2]+n[1,3]+n[2,3])/(n[1,2]+n[2,2]+n[1,3]+n[2,3]+n[3,3])
  fp <- (n[2,1]+n[3,1]+n[3,2])/(n[1,1]+n[2,1]+n[3,1]+n[2,2]+n[3,2])
  print(n)
  cat("errore totale: ", format(err.tot),"\n")
  cat("falsi positivi & falsi negativi: ",format(c(fp, fn)),"\n")
  invisible(n)}
```

Fn e *fp* sono i falsi positivi e i falsi negativi per la matrice 3x3 di errata classificazione, mentre *err.tot* è l'errore totale calcolato come 1- quoziente tra le unità sulla diagonale e il totale degli utenti del campione di verifica. La funzione tabella.sommario è stata impiegata per valutare la bontà di previsione di tutti i modelli considerati: modello multinomiale, albero di classificazione, analisi discriminante e distanza AR.

Albero di classificazione

In R esistono 2 librerie aggiuntive per la costruzione di alberi di classificazione, `tree` e `rpart`. Per questa tesi sono stati utilizzati i comandi contenuti nella prima, con la seguente sintassi:

```
library(tree)
fit<tree(valore~eta+piano+zona,data=st.ca[parte1,],control=tree.contro
l(nobs=length(parte1),minsize=2,mindev=0.0001)))
```

Il comando indicato si riferisce alla crescita dell'albero ottenuto al paragrafo 4.2. Per l'algoritmo di crescita viene utilizzata una parte del campione di stima (parte1) e i controlli successivi indicano il minimo numero di osservazioni che deve contenere un nodo (minsizer) e la minima riduzione di devianza per creare un nuovo nodo decisionale (mindev).
A seguito della crescita, l'albero viene potato con la seconda parte del campione di stima (funzione per la potatura $C(\alpha)$ a pag. 56):

```
fit.p<-prune.tree(fit,newdata=st.ca[parte2,])
```

Con la funzione prune.tree si ottiene il numero di nodi ottimale per minimizzare la devianza (par. 4.2):

```
j<-fit.p$size[fit.p$dev==min(fit.p$dev)]
```

con il numero di nodi ottimale l'albero definitivo si ottiene ancora con il comando `prune.tree`:

```
mdp<-prune.tree(fit,best=j)
```

Analisi discriminante bayesiana

Il comando per la stima di un modello di analisi discriminante bayesiana è `lda`, appartenente alla libreria MASS:

```
fit <- lda(valore~zona+piano+eta, prior=c(0.34,0.33,0.33),data=st.ca)
```

l'opzione `prior()` permette di indicare una distribuzione a priori soggettiva per le tre classi che costituiscono il dominio della variabile risposta. Se tale opzione non viene utilizzata, l'analisi discriminante considera come pesi per l'a-priori le frequenze relative della variabile risposta nel campione di stima.

Distanza AR

Per il calcolo della distanza AR è stato adattato al caso di tre gruppi il criterio indicato nell'articolo di Domenico Piccolo e Marcella Corduas (2007): Time series clustering and classification by the autoregressive metric.
Il campione di stima viene suddiviso secondo la variabile valore del cliente e per ciascun gruppo viene calcolata la media dei coefficienti autoregressivi:

```
ia<-which(mat.st$valore=="a")
ib<-which(mat.st$valore=="b")
ic<-which(mat.st$valore =="c")
p1a<-mean(phi1s[ia])
p2a<-mean(phi2s[ia])
p1b<-mean(phi1s[ib])
p2b<-mean(phi2s[ib])
p1c<-mean(phi1s[ic])
p2c<-mean(phi2s[ic])
```

mat.st è la matrice che contiene tutti i valori delle variabili esplicative indicizzate per numero utente e nella colonna valore comprende la classe della variabile risposta (“a”, “b”, “c”).

Le sei variabili p1a,p2a,… indicano le medie dei coefficienti ϕ_1, ϕ_2 per ciascuna “classe di utenti” del campione di stima. Queste medie, prese a due a due, costituiscono i centroidi per i tre gruppi. L’unità viene assegnata al centroide più vicino, per questo si calcola la distanza D per ogni unità del campione di verifica (cap. 5):

```
D_ab<-((phi1v-p1a)^2+(phi2v-p2a)^2)-((phi1v-p1b)^2+(phi2v-p2b)^2)
D_ac<-((phi1v-p1a)^2+(phi2v-p2a)^2)-((phi1v-p1c)^2+(phi2v-p2c)^2)
D_bc<-((phi1v-p1b)^2+(phi2v-p2b)^2)-((phi1v-p1c)^2+(phi2v-p2c)^2)
```

Avendo a disposizione le tre distanze per ciascuna unità è possibile definire la seguente regola di assegnazione:

```
cl<-((D_ab<0)&(D_ac<0))*1+((D_ab>0)&(D_bc<0))*2+((D_bc>0)&(D_ac>0))*3
```

Il vettore cl riporta per ogni unità la classificazione ottenuta da questo algoritmo:

- se 1: il cliente viene assegnato alla classe “a”;
- se 2: il cliente viene assegnato alla classe “b”;
- se 3: il cliente viene assegnato alla classe “c”;

BIBLIOGRAFIA

- AZZALINI A. (2001), Inferenza Statistica: una presentazione basata sul concetto di verosimiglianza. 2a edizione, Springer-Verlag, Milano. 366 pp.;
- AZZALINI A., SCARPA B. (2004), Analisi dei dati e data mining, Springer-Verlag, Milano. 232 pp.;
- BONETTO M. (2007), Prevedere il churn: un approccio longitudinale, tesi di laurea specialistica, Facoltà di Scienze Statistiche, Università degli studi di Padova, 98 pp.;
- BORDIGNON S (A.A. 2006-2007), Serie storiche economiche c.p., dispensa delle lezioni. Corso di Laurea Specialistica, Facoltà di Scienze Statistiche, Università degli studi di Padova.
- CAMILLO F., TASSINARI G. (2002), Data Mining, web mining e CRM, Tipomonza, Milano. 280 pp.;
- CAPPUCCIO N., ORSI R. (2005), Econometria, il Mulino, Bologna. 770 pp.;
- CORDUAS M., PICCOLO D. (2007), Time series clustering and classification by the autoregressive metric, Computational Statistics & Data Analysis 52, 1860-1872;
- FRANSES P., PAAP R. (2005), Quantitative Models in Marketing Research, Cambridge University press, Cambridge. 206 pp.;
- SCARPA B. (A.A. 2008-2009), Tecniche Statistiche di Classificazione, dispensa delle lezioni. Corso di Laurea Triennale, Facoltà di Scienze Statistiche, Università degli Studi di Padova;
- SHUMWAY (1982), Discriminant analysis for time series in Krishnaiah e Kanal eds. Handbook of Statistics, vol.2, p8;
- VENABLES RIPLEY (2002), Modern Applied Statistics with S, Springer, New York. 495 pp.;
- VENTURA L. (A.A. 2004-2005), Modelli Statistici II, dispensa delle lezioni. Corso di Laurea Triennale, Facoltà di Scienze Statistiche, Università degli Studi di Padova.