UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE
CORSO DI LAUREA IN STATISTICA E GESTIONE DELLE IMPRESE

Tesi di Laurea triennale

LA SEGMENTAZIONE DELLA CLIENTELA ATTRAVERSO LA CLUSTER ANALYSIS: IL CASO ELETTROINGROSS.

Relatore: Prof.ssa Francesca Bassi

Laureando: Tomas Stievano
Matr. 438315/GEI

ANNO ACCADEMICO 2005/2006

Indice.

1. Capitolo introduttivo 1

1.1 Presentazione di Elettroingross..... 1

1.2 Il settore della distribuzione di materiale elettrico. 3

1.3 La redditività nel commercio di materiale elettrico..... 4

1.4 Organizzazione e competitività. 6

1.5 Descrizione della clientela. 7

2. Segmentazione empirica della clientela..... 9

2.1 Indagine preliminare ed esplorativa..... 9

2.2 Il contatto con il cliente. 11

2.3 Strategie operative di segmentazione. 21

2.3.1 Segmentazione sulle variazioni di fatturato. 22

2.3.2 Segmentazione in base alla frequenza d'acquisto. 31

2.4 Conclusioni. 38

3. La Cluster Analysis. 39

3.1 Introduzione alla Cluster Analysis..... 39

3.2 Misure di prossimità e distanza fra unità statistiche. 42

3.3 Misure di similarità per dati binari. 43

3.4 Misure di similarità per dati categorici non binari..... 46

3.5 Dissimilarità e misure di distanza per dati continui..... 46

3.6 Misure di similarità per dati misti. 50

3.7 Clustering gerarchico. 51

3.7.1 Metodo del *legame singolo (single linkage):* 55

3.7.2 Metodo del *legame completo (complete linkage):* 56

3.7.3 Metodo del *legame medio (average linkage):* 57

3.7.4 Metodo del *Centroide:* 57

3.7.5 Metodo di *Ward:* 58

3.8 Algoritmo generale per le tecniche gerarchico-agglomerative..... 59

4. Applicazione pratica della Cluster Analysis. 61

4.1 La segmentazione della clientela con la Cluster Analysis. 61

4.2 Conclusioni. 87

5. La Cluster Analysis con SPSS. 89

5.1 Introduzione 89

5.2 L'utilizzo di SPSS per la Cluster Analysis. 90

Bibliografia. 95

1.

Capitolo introduttivo

1.1 Presentazione di Elettroingross.

Elettroingross, azienda leader nella distribuzione di materiale elettrico del Triveneto, nasce a Padova nel 1978 dalla fusione dei quattro maggiori grossisti locali: Canas, Fanton, Corazza, Vanotti.

Inizialmente la sede centrale e le filiali minori coprono il territorio tra Padova e Mestre assicurandosi il 13% del mercato locale. I risultati sono notevoli già a quel tempo con 27 miliardi di lire di fatturato, 140 dipendenti ed 12.000 m² totali di esposizione.

Nell'aprile del 1988 si giunge al momento di svolta, quando, con un fatturato ormai ben oltre i 100 miliardi di lire, l'azienda cede la quota di maggioranza al grande Gruppo francese SONEPAR, indiscusso leader mondiale nel settore distributivo di materiale elettrico.

Unisce le due aziende un'identità non solo sullo scopo, ma anche sulla filosofia del lavorare:

1) la qualità come fattore strategico assoluto;
2) la spinta continua verso la totale soddisfazione del cliente;
3) la professionalità degli operatori;
4) la cultura del "noi" aziendale.

Questi i principi che hanno portato l'azienda e il Gruppo alla realizzazione nel gennaio 2000 della attuale sede di Padova in via Riviera Maestri del Lavoro, 24. Oltre 350 dipendenti, più di 40 i venditori dislocati sul territorio del Triveneto e di buona parte dell' Emilia Romagna. La realizzazione copre 24.000 m², con ampia facoltà di parcheggio e un magazzino che raggiunge i 12 metri di altezza, per ospitare oltre 30.000 referenze.

Sono oltre 1.000 le spedizioni fatte giornalmente dalla sede che, con il contributo dei 15 punti vendita portano l'azienda oltre i 100 milioni di Euro di fatturato annui attuali.

Al piano terra fa spicco un moderno Self Service realizzato su un'area di circa 2.000 m² con ben 25.000 articoli direttamente disponibili al cliente che può facilmente rifornirsi di materiale elettrico ed elettronico, automazione, antennistica, cavi, illuminazione.

Elettroingross, consapevole del continuo e veloce sviluppo tecnologico e volendo contribuire alla crescita professionale e all'efficienza dei suoi clienti, offre consulenza nella progettazione e nella realizzazione di impianti di automazione e di illuminotecnica mettendo a disposizione la propria competenza nella scelta dei prodotti più adatti.

Sede e Filiali Elettroingross

1 - Padova – Sede
2 - Ravenna
3 - Marghera (VE)
4 - Fiume Veneto (PN)
5 - Verona
6 - Badia Polesine (Ro)
7 - Treviso
8 - Trieste
9 - Rimini
10 - Villanova di Castenaso (BO)
11 - Belluno
12 - Udine
13 - Schio (VI)
14 - Padova
15 - Rubano (PD)
16 - S.Maria di Sala (VE)

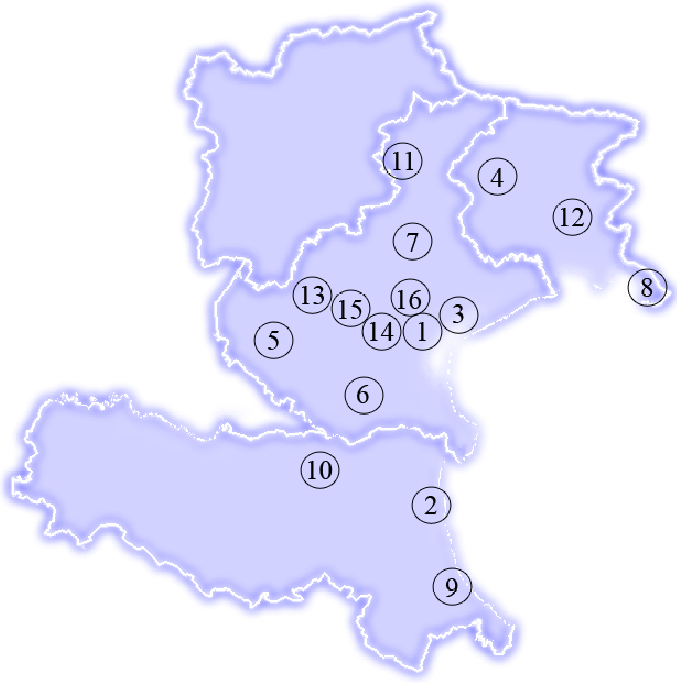


Figura 1.1 – dislocazione dei punti vendita Elettroingross.

1.2 Il settore della distribuzione di materiale elettrico.

La distribuzione commerciale è l'attività attraverso la quale i prodotti dell'industria (o di tutti i processi produttivi in genere) vengono immessi nella rete commerciale al fine di effettuarne la vendita. La distribuzione può essere diretta dal produttore al cliente, quando non prevede figure intermedie, oppure indiretta, quando si avvale di una rete organizzata e più o meno ramificata nel territorio, che contempla intermediari (come ad esempio i grossisti) e rivenditori (commercianti al dettaglio, grandi magazzini, centri commerciali).
La distribuzione di materiale elettrico è un'attività prevalentemente commerciale priva di qualsiasi processo produttivo, quindi le aziende come Elettroingross di fatto sono società che erogano servizi nei confronti dei:

- produttori, garantendo la diffusione nel territorio dei propri prodotti;
- clienti, per soddisfare i loro bisogni.

I produttori svolgono di rado anche la distribuzione (distribuzione diretta) perché è un'attività che richiede un grande impegno economico e umano, lunghi trasferimenti, ampi magazzini e un'efficace organizzazione logistica e burocratica.
I produttori, che concentrano le proprie risorse nell'attività produttiva, generalmente non intendono implementare una rete di distribuzione e per tali motivi richiedono l'iniziativa di terzi: questa sostanzialmente è la ragione dell'esistenza dei grossisti.
In un sistema distributivo moderno le aziende commerciali svolgono oltre al tradizionale ruolo logistico anche un ruolo di marketing.
Le finalità del marketing distributivo sono: innovare, differenziare e comunicare i propri prodotti e servizi in modo da indurre il cliente a preferirlo rispetto alle alternative offerte dai concorrenti (Ziliani, 1999).

Il fondamento del marketing distributivo risiede nell' eterogeneità della domanda di servizi commerciali e nella gamma di prodotti da selezionare da rendere accessibili oltre che convenienti per i consumatori.
Gli strumenti attraverso i quali l'azienda commerciale fa marketing sono la gestione degli assortimenti e dello spazio espositivo, la marca dell'insegna, la marca commerciale, le attività promozionali di prezzo e non di prezzo.
Le diverse leve del marketing distributivo possono essere manovrate a livello macro oppure a livello micro a seconda della possibilità tecnologica di segmentare la domanda.
Elettroingross, ad esempio, era solita avviare promozioni di prezzo "a pioggia" nel senso che non potendo individuare profili di clienti diversi, la promozione catturava effettivamente l'interesse di chi non avrebbe acquistato, ma allo stesso tempo, anche il cliente abitudinario che normalmente era disposto a pagare il prezzo pieno.
Pertanto si avverte la concreta necessità di segmentare la domanda, i clienti, al fine di gestire in maniera mirata ed ottimale le vendite.

1.3 La redditività nel commercio di materiale elettrico.

La marginazione, intesa come differenza tra ricavi di vendita e costo del venduto, è piuttosto ridotta in questo settore in cui la concorrenza è numerosa e rappresenta un mercato molto aggressivo sui prezzi e dove in particolare la disponibilità di prodotti succedanei è molto elevata.
I produttori verso i grossisti decidono ed applicano i prezzi indicati su un "listino grossisti", quest'ultimi applicano una percentuale di ricarica che rappresenta il loro guadagno e in tale maniera si determina il "listino pubblico". Bisogna precisare che dal "listino pubblico" generalmente si applicano gli sconti ai clienti in base alla loro importanza commerciale e talvolta sono così consistenti che pur di mantenere un cliente importante si pratica un prezzo di poco superiore al costo d'acquisto.

Per raggiungere il profitto i grossisti possono percorrere due strade:

- la marginazione di cui sopra;
- ottenere i premi di vendita dai produttori.

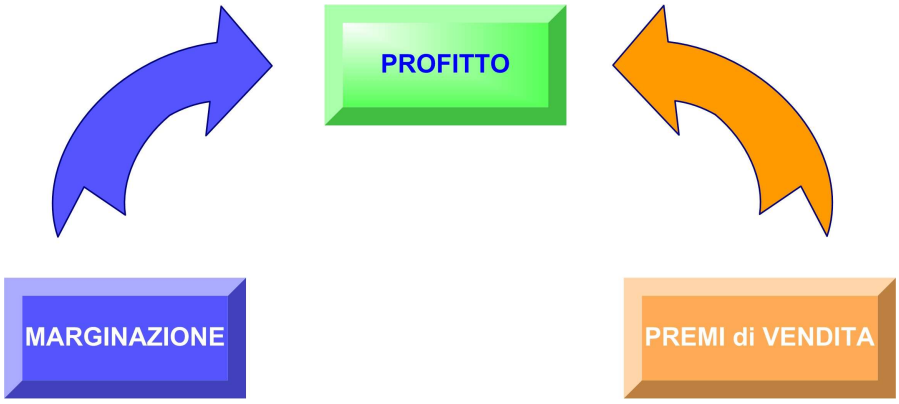


Figura 1.2 – Il conseguimento del profitto

I premi di vendita sono degli incentivi economici che i produttori e/o fornitori di materiale elettrico elargiscono ai grossisti se i volumi di vendita superano un livello di vendite concordato. Le soglie per l'ottenimento dei premi si stabiliscono a priori e vengono pattuite con cadenza annuale. Le contrattazioni si svolgono individualmente tra produttore e grossista e dato che in queste circostanze i produttori hanno maggior potere decisionale succede che i livelli di soglia tendenzialmente aumentano negli anni: significa impegnarsi a vendere sempre di più e superare di volta in volta i risultati ottenuti.

Nonostante la marginazione sia limitata come è stato precisato precedentemente, la redditività raggiunge livelli elevati con l'ottenimento dei premi; ma visto che è sempre più difficile raggiungerli dato che le soglie si spostano in modo sfavorevole per i grossisti, a quest' ultimi, per domare gli eventi, non resta che cercare e formare nuovi equilibri di mercato: per aumentare la capacità di vendita i maggiori grossisti effettuano acquisizioni di altre aziende analoghe riuscendo a ottenere più agevolmente i premi e imponendo nuovi "ritmi" alla concorrenza. Le

fusioni tra società comportano un altro vantaggio molto importante; quello di diventare più forti dal punto di vista contrattuale con i produttori-fornitori. Per queste ragioni i grossisti minori si trovano in un mercato sempre più aggressivo dove generalmente le situazioni sfavorevoli li portano a essere acquisiti o spazzati fuori dal mercato.

1.4 Organizzazione e competitività.

La competitività di un'azienda che opera nel settore del commercio di materiale elettrico non è dovuta solamente al prezzo e ai servizi praticati, ma è legata fortemente all'impostazione e all'organizzazione dell'azienda.
Elettroingross, sin dalla nascita, ha saputo impostare la propria struttura aziendale in modo da essere presente sul territorio in modo agile e veloce. L'idea vincente è stata quella di centralizzare il magazzino in un'unica struttura dalla quale tutti i punti vendita fanno riferimento.
I costi di gestione in questo modo sono minimizzati poiché i punti vendita non hanno scorte da gestire ma solamente la merce destinata alla vendita. Con il procedere delle vendite il software gestionale provvede automaticamente al riordino della merce qualora venga raggiunta una soglia minima prestabilita.
Tutta la concorrenza, al contrario, per ogni punto vendita ha annesso anche il magazzino: se, da un lato, c'è il vantaggio di avere una struttura capillarmente autonoma, dall'altro, si è penalizzati dai costi di gestione maggiori e dalla rigidità strutturale.
Per rigidità strutturale si intende la scarsa capacità di implementare nuovi punti vendita o di spostarli al verificarsi di opportunità commerciali.

1.5 Descrizione della clientela.

L'installatore:
è l'azienda o il professionista il cui business comprende l'implementazione, revisione e messa in funzione di impianti elettrici.
Nell'anagrafe clienti aziendale vi è un'ulteriore suddivisione in: installatore civile, commerciale, industriale.

In realtà è una suddivisione poco realista perché ciascun installatore non svolge in modo esclusivo una delle opzioni ma se gli è possibile una combinazione delle tre. Infatti, negli ultimi anni caratterizzati da una grande incertezza economica, in cui i risparmiatori hanno azzerato gli investimenti e rivolto gran parte dei risparmi sul "mattone", si è creata una profonda crisi nell'installazione industriale data la mancanza di fondi e una grande crescita nell'istallazione civile.
Con queste premesse era ovvio che gli installatori orientati verso l'installazione industriale spostassero il proprio baricentro verso l'installazione civile principalmente per due motivi:

A) per non soccombere e ottenere nuove commesse di lavoro;
B) per rifarsi degli investimenti (talvolta perdite) sostenute nell settore industriale.

È la categoria più interessante da analizzare perché genera i più grandi volumi di fatturato e implicitamente rappresenta un indice di concentrazione del business di materiale elettrico. Assumendo che gli installatori utilizzano sia il materiale acquistato direttamente sia il materiale acquistato da terzi che richiedono il loro intervento, è possibile analizzare dal punto di vista geografico come sono concentrate le residenze degli installatori per stimare l’intero valore economico di una zona e di conseguenza la quota di mercato.

Industria:

categoria rappresentata da aziende di produzione che svolgono investimenti in impianti e macchinari industriali. Acquistano totalmente il proprio fabbisogno di materiale elettrico e svolgono al loro interno le manutenzioni ordinarie, mentre, richiedono l'intervento esterno di installatori specializzati per le manutenzioni straordinarie o nuove installazioni provvedendo loro il materiale necessario.

Quadrista:

professione simile all'installatore ma circoscritta alla realizzazione dei pannelli di controllo, detti appunto quadri. Come per gli installatori si possono suddividere in civile, commerciale, industriale e, in merito a questa la suddivisione, valgono le stesse considerazioni fatte per la categoria installatore.

Rivenditori:

svolgono come i grossisti attività di distribuzione di materiale elettrico con la differenza di vendere capillarmente i prodotti non essendo così generalmente d'ostacolo o concorrenti pericolosi ma al contrario una categoria di clienti importante: i loro acquisti sono contraddistinti da una certa regolarità sia in valore sia in termini di frequenza d'acquisto. In certi casi possono diventare una vera e propria filiale: in tal caso vengono definiti “Terzisti”.

Enti, servizi:

rappresentati da enti pubblici come l'università, ferrovie dello stato, ospedali eccetera.

Occasionali e speciali:

clienti non riconducibili a nessuna delle categorie precedenti e che producono acquisti in modo occasionale e non ripetitivo. In questa categoria rientrano privati, negozi in allestimento, ecc.

2.

Segmentazione empirica della clientela.

2.1 Indagine preliminare ed esplorativa.

L'obiettivo del presente lavoro è di ottenere informazioni per rendere più profittevoli le relazioni con i clienti.
In particolare segmentando i clienti in base al comportamento d'acquisto si vuole conoscere quali stanno dimostrando una probabile tendenza all'abbandono.

Alla luce di quanto detto sull'importanza di generare ampi volumi di fatturato per ottenere i premi di vendita, e ovvio che interessa contrastare tutti quei fenomeni che possono ridurre tale grandezza: nel caso in esame è più vantaggioso mantenere i clienti attivi che conquistarne di nuovi.

Gli studi dimostrano che acquisire un nuovo cliente ha un costo dalle 5 alle 10 volte superiore a quello necessario per mantenerne uno esistente e che il cliente fedele comprerà di più durante la sua vita e sarà disposto a pagare un premio di prezzo a chi ha conquistato la sua fiducia

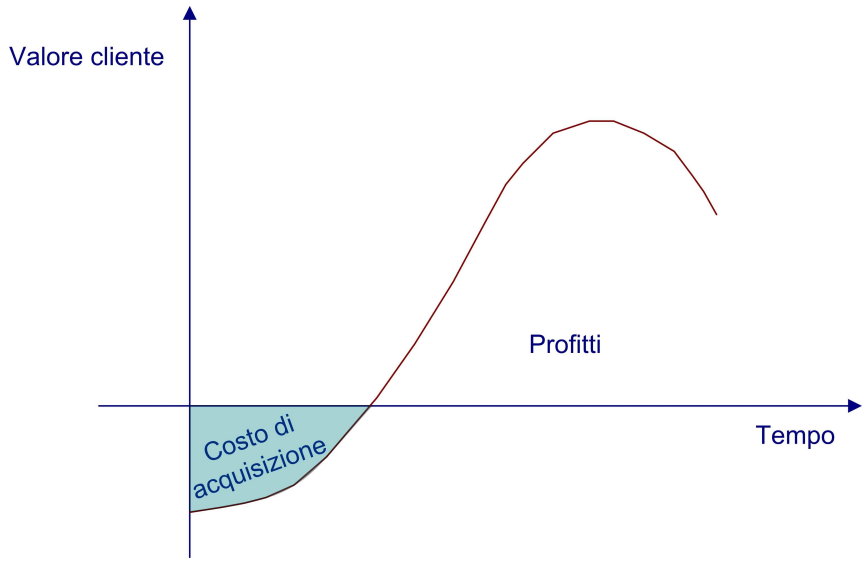


Figura 2.1 – Il valore del cliente nel tempo.

Il grafico 2.1 evidenzia la curva che descrive il ciclo di vita del cliente in relazione al tempo e al valore generato dall'impresa.

L'area negativa rappresenta il costo di acquisizione del cliente per l'impresa, l'area positiva rappresenta il profitto derivante dal l'interazione nel tempo con il cliente (Cuomo, 2000).

In questa fase del lavoro si stanno raccogliendo tutte le informazioni che possono risultare utili per decidere secondo quali criteri impostare la segmentazione; in particolare per fare una segmentazione efficace bisogna comprendere come si manifesta l’abbandono da parte dei clienti (punto di maggiore interesse): solo facendo le necessarie considerazioni e comportandosi di conseguenza si potranno raggiungere i risultati sperati.

2.2 Il contatto con il cliente.

Elettroingross, come tutti i grossisti di materiale elettrico in genere, propone alla clientela vari canali d'acquisto:

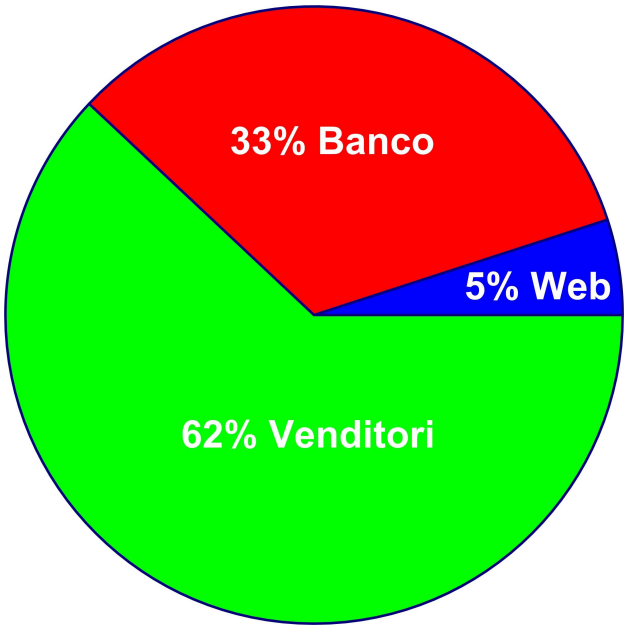


Figura 2.2 – Percentuale di utilizzo dei diversi canali di vendita.

Il grafico in figura 2.2 evidenzia un fattore molto importante: tutta la popolazione clienti di Elettroingross predilige acquistare con i canali che prevedono il contatto, la relazione umana.

Le preferenze del canale d'acquisto da parte della clientela varia da soggetto a soggetto; ad esempio i clienti con una configurazione ben strutturata (ufficio acquisti, magazzino) sono ben disponibili ad orientarsi verso sistemi più veloci e automatizzati come il Web, anche se risulta essere ancora uno strumento molto giovane, mentre altri preferiscono il contatto diretto con il venditore oppure rivolgersi al banco.

Il commercio di materiale elettrico è un campo complesso, caratterizzato da una costante evoluzione tecnologica, e dalla presenza di numerosi marchi. Gli installatori in particolare confidano nel parere e nei consigli del commerciante

per i loro acquisti, e da questo si determina l'importanza del contatto e della relazione umana. Risulta essere molto profittevole un forte legame da entrambe le parti. Uno dei fattori più importanti per la clientela, molto più importante del prezzo, è appunto la fiducia di ottenere precise indicazioni per svolgere adeguatamente il proprio lavoro scegliendo bene i prodotti, venendo indirizzati verso marchi o articoli alternativi.

Il meccanismo del contatto con il cliente può essere rappresentato con la seguente matrice:

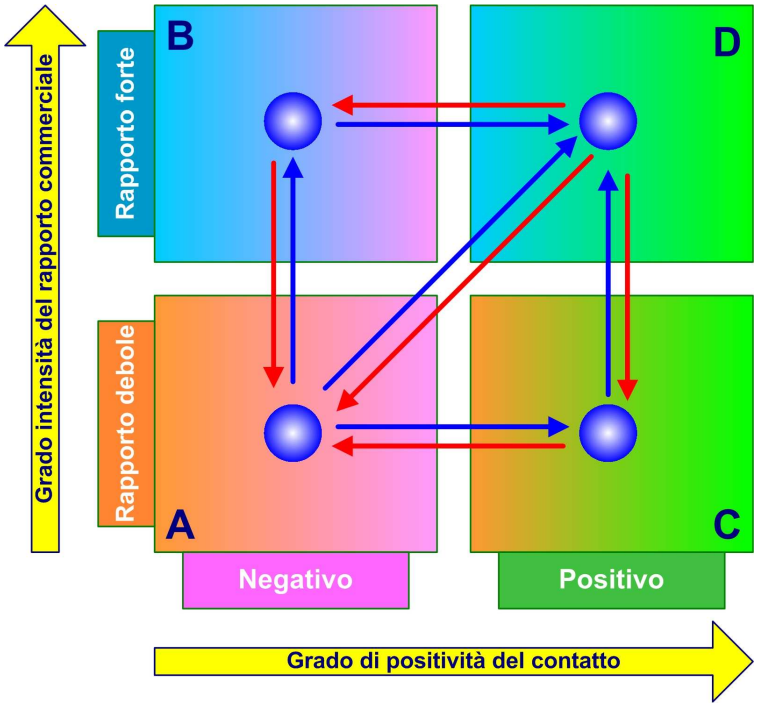


Figura 2.3 – Matrice dei rapporti Cliente / Elettroingross.

Questa matrice evidenzia che una relazione può avere uno spostamento ascendente o discendente in termini di intensità del rapporto commerciale e muoversi sul versante di positività o negatività del contatto.
Le due variabili, forza della relazione e grado di positività, combinandosi individuano un quadrante dei rapporti.

Se consideriamo il quadrante A (contatto negativo / rapporto debole) troviamo i clienti che risultano essere molto difficili e problematici da gestire:

- dal punto di vista contrattuale;
- dal punto di vista dei pagamenti che non avvengono regolarmente.

Poiché l'intensità commerciale è debole potrebbe trattarsi di clienti piccoli oppure di clienti di importanza maggiore ma che realizzano acquisti modesti.
In passato Elettroingross evitava i clienti con queste caratteristiche per ovvie ragioni cautelative, non sempre però e stata la strategia migliore. Ci sono stati casi in cui clienti di questo genere, apparsi da poco sul mercato, sono stati "abbandonati" da Elettroingross e subito accolti dalla concorrenza ottenendo da quest' ultima la fiducia e la comprensione necessaria per avviarsi. Successivamente si sono affermati col passare del tempo, diventando clienti importanti che si potrebbero posizionare nel quadrante C o D.
Attualmente con questi clienti si procede valutando le caratteristiche di ogni singolo, impegnandosi solamente con quelli che risultano più promettenti.
Il quadrante B è composto sempre da clienti difficili con la differenza di intrattenere relazioni commerciali più intense in termini di valore: sono commercialmente importanti ma possono rivelarsi molto pericolosi in quanto effettuano ordini impegnativi e i pagamenti potrebbero non essere corrisposti regolarmente.
Il quadrante C è caratterizzato da clienti con i quali si ha un contatto positivo, sono quindi facili da gestire, l'unico parametro migliorabile è l'intensità commerciale che risulta essere debole: potrebbero essere clienti di piccole dimensioni oppure clienti con una bassa penetrazione commerciale (legati alla concorrenza) e andrebbero incentivati ad acquistare con opportune strategie.
Il quadrante D è il migliore dei quattro ed è composto da clienti con ottime relazioni di contatto e ovviamente con forte intensità commerciale. Sono clienti tenuti in grande considerazione poiché si sono sempre rivelati affidabili e effettuano ordinativi importanti.

In generale si può affermare che i clienti posizionati sul versante di contatto negativo sono clienti che rendono l'attività di vendita molto onerosa e rischiosa, caso contrario ovviamente i clienti sul versante del contatto positivo.
Ogni cliente si colloca in un quadrante in base alla qualità della relazione e della convenienza del contatto con Elettroingross rispetto alla concorrenza, e di conseguenza, adeguerà l'intensità del rapporto commerciale.

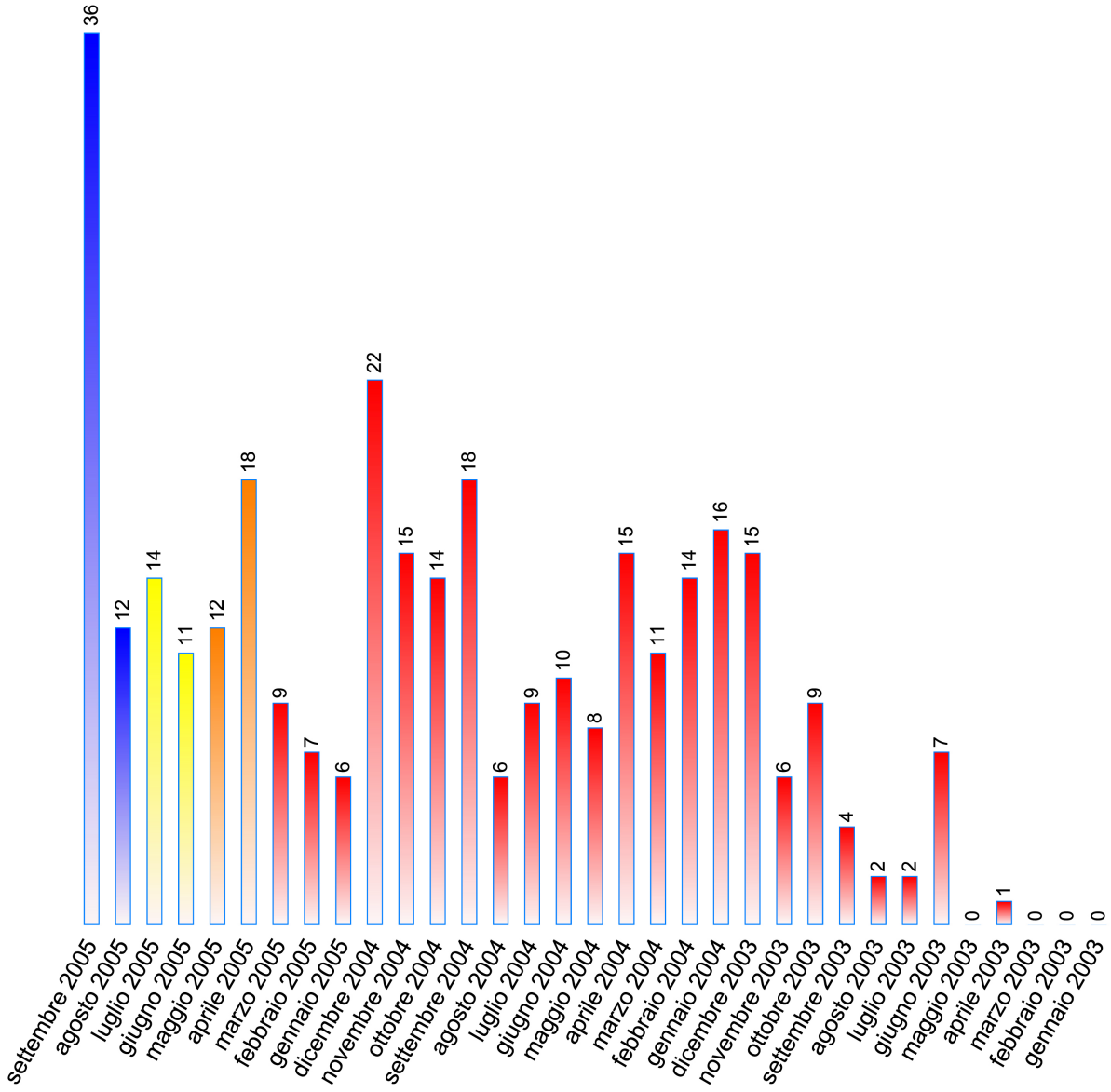


Figura 2.4 – Conteggio dei clienti che hanno smesso di acquistare.

Il grafico 2.4 è molto importante ed è un'utile rappresentazione per capire come si manifesta nel tempo il fenomeno della scomparsa dei clienti. Sono stati presi in esame solamente i clienti attivi di Elettroingross del 2003 che rappresentavano l'80% del fatturato: i clienti sono stati ordinati in senso decrescente in relazione al fatturato generato e sono stati considerati solamente i primi che concorrono alla formazione dell' 80% del fatturato totale. In questo modo sono stati considerati solo i clienti che acquistavano somme importanti: si tratta di un totale di 1.235 clienti.

L'altezza delle barre rappresenta il numero di clienti che ha effettuato l'ultimo acquisto nel mese indicato.

Questa analisi si è svolta nel mese di ottobre 2005, si può notare che l'asse del tempo procede a ritroso e parte dal mese di settembre 2005. Non è stato rappresentato anche il mese di ottobre 2005 perchè con il valore elevato di 908 avrebbe compresso la scala e pregiudicato la rappresentazione.

Dalle indicazioni ottenute da persone con forti competenze di vendita, possiamo considerare come clienti scomparsi quelli che non acquistano da molto tempo, nello specifico tutta la coda di destra, mentre i clienti che non acquistano nel passato recente (ultimi 5 – 6 mesi) sono quelli a rischio di abbandono. Solo gli ultimi 2 mesi (3 se conteggiamo anche la barra di ottobre non rappresentata) possiamo ancora ritenerli con prudenza ancora attivi. Dalle indicazioni pervenute si evince inoltre che i clienti si riforniscono in media da 3 grossisti, quindi, anche se il fabbisogno di materiale è molto frequente nel tempo non è raro incontrare clienti che acquistano a intervalli regolari di qualche mese, dato che "ruotano" gli approvvigionamenti su più fornitori.

Questa rappresentazione grafica se ripetuta nel tempo subirà degli aggiornamenti nella parte di sinistra mentre resterà probabilmente immutata nella coda di destra. Ad esempio, se i 12 clienti di agosto 2005 tornano tutti ad acquistare in novembre, il grafico aggiornato avrà 0 in agosto perchè quei clienti hanno spostato l'ultimo acquisto in novembre.

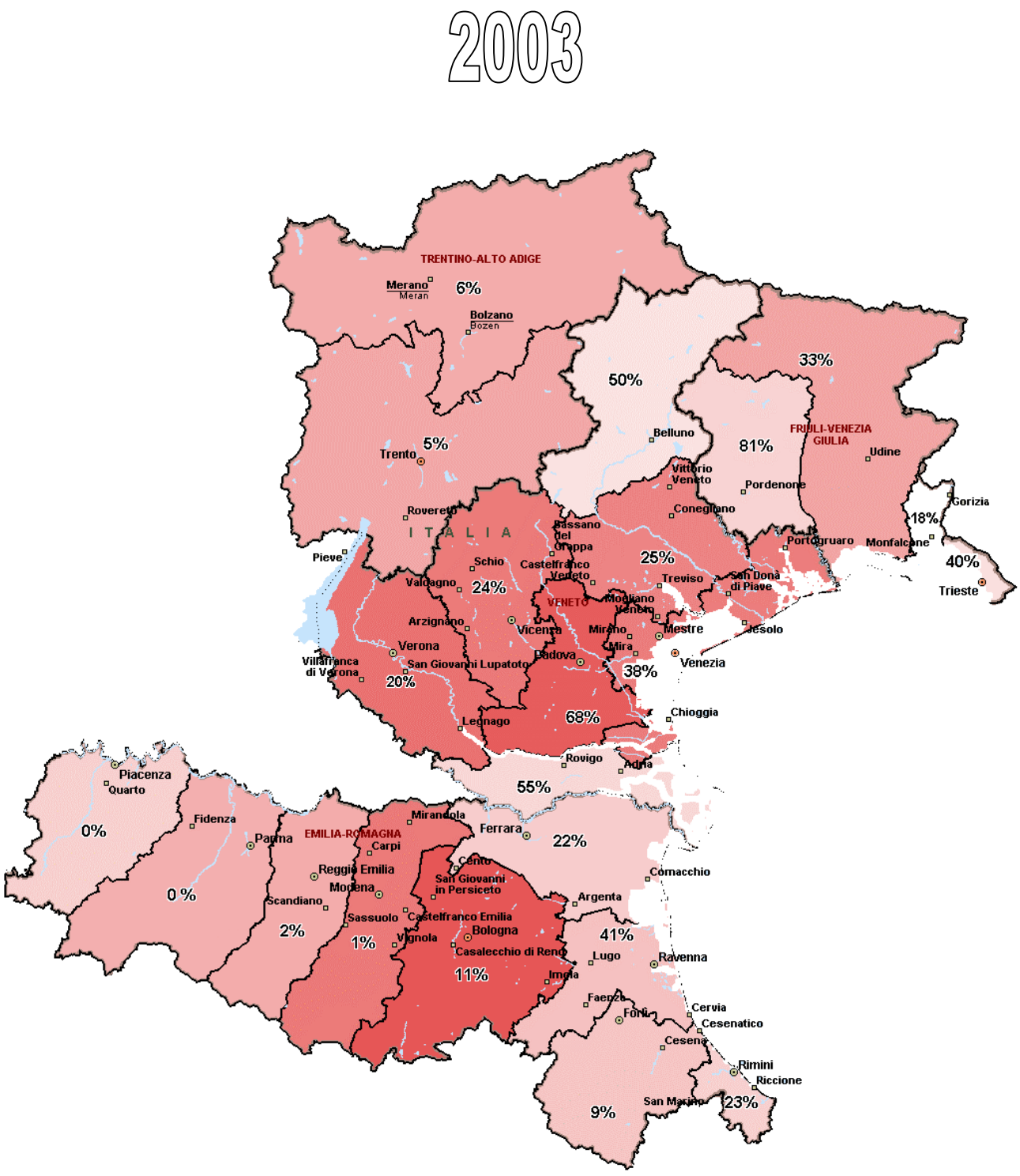
Le indicazioni del grafico non sono abbastanza tempestive per poter reagire con strategie, infatti, viene segnalato un cambiamento del comportamento del cliente

quando è già avvenuto, ma, in sede esplorativa, fanno capire che per studiare questi fenomeni bisogna prendere in considerazione il passato e con un orizzonte temporale ampio (almeno 3 anni): se dei clienti dimostrano una buona propensione all'acquisto prima di sparire probabilmente lanceranno dei segnali. Se questa supposizione è corretta e se si registrano questi segnali allora sarà possibile praticare delle procedure di recupero. Per esempio si potrebbe accusare una riduzione in valore o in frequenza degli acquisti (controllo eseguibile sui dati aziendali), oppure considerare se la qualità del contatto con il cliente sta peggiorando (informazione difficilmente reperibile e misurabile).
Il secondo passo è stato valutare e misurare sul territorio il contatto con il cliente, provincia per provincia dal 2003 al 2005 e per darne una valida rappresentazione è stato adottato il seguente indice di *copertura* del mercato:

$$\mathrm{K} = \frac{{}_{P}C_{a}}{{}_{P}C_{t}}$$

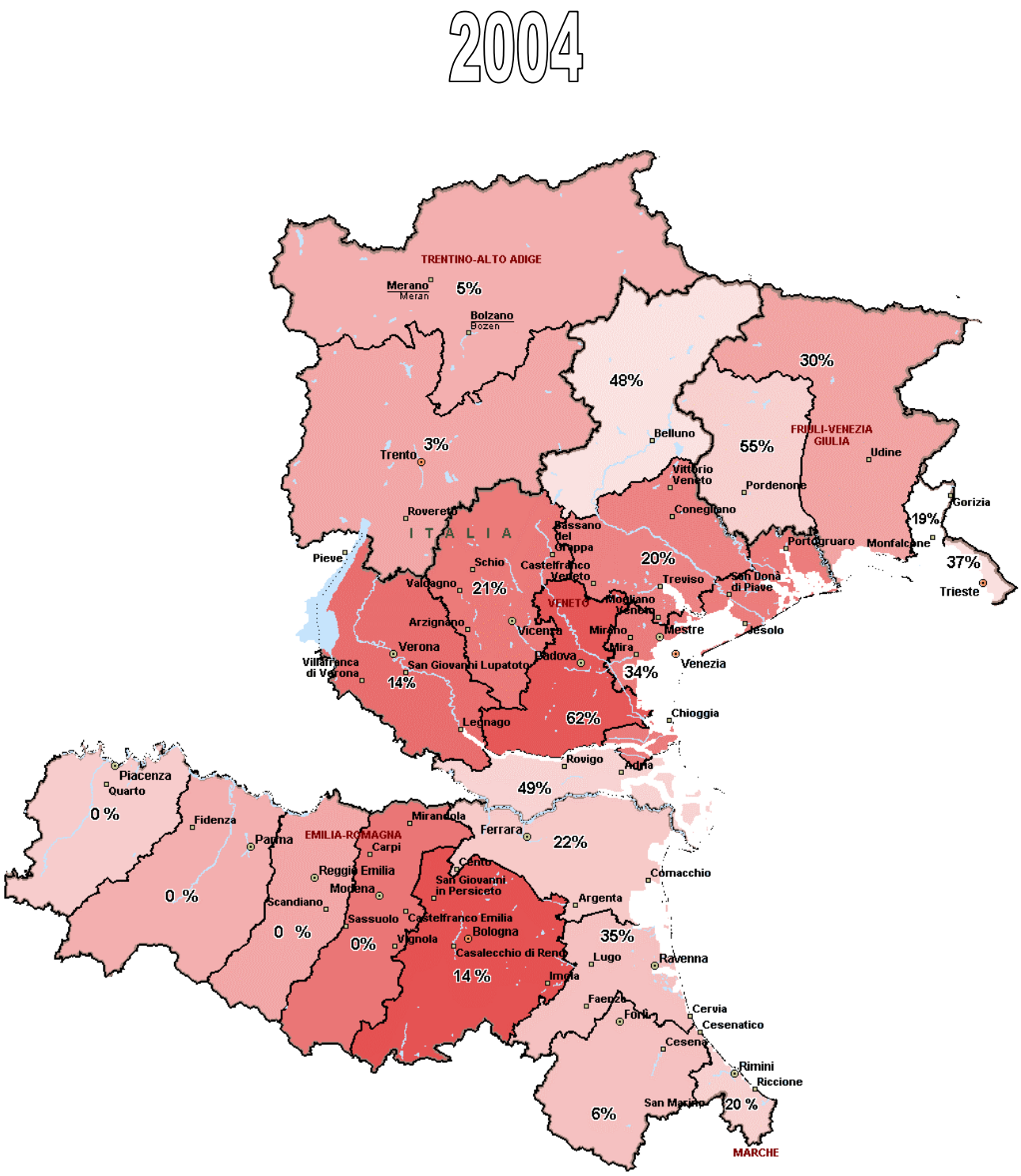
L'indice K esprime la percentuale di contatto dove ${}_{P}C_{a}$ è inteso come il numero di clienti attivi di Elettroingross dell'anno considerato della provincia *P* e ${}_{P}C_{t}$ i clienti potenziali totali della provincia *P*.
Per il calcolo di questo indice è stata presa in considerazione la sola categoria degli installatori di materiale elettrico, l'unica di cui si disponeva dei dati necessari: dati aziendali interni e un database precedentemente acquistato per soli scopi promozionali dalla camera di commercio, su tutti gli installatori del Triveneto ed Emilia.
Gli indici ottenuti di ogni provincia e per i tre anni considerati sono stati rappresentati graficamente in figura 2.5 a,b,c in modo tale da avere un immediato impatto visivo e di facilitarne il confronto, in particolare:

- ogni provincia assume una colorazione diversa in base alla concentrazione di clienti potenziali in numero assoluto;
- per ogni provincia viene espresso il valore dell'indice K.



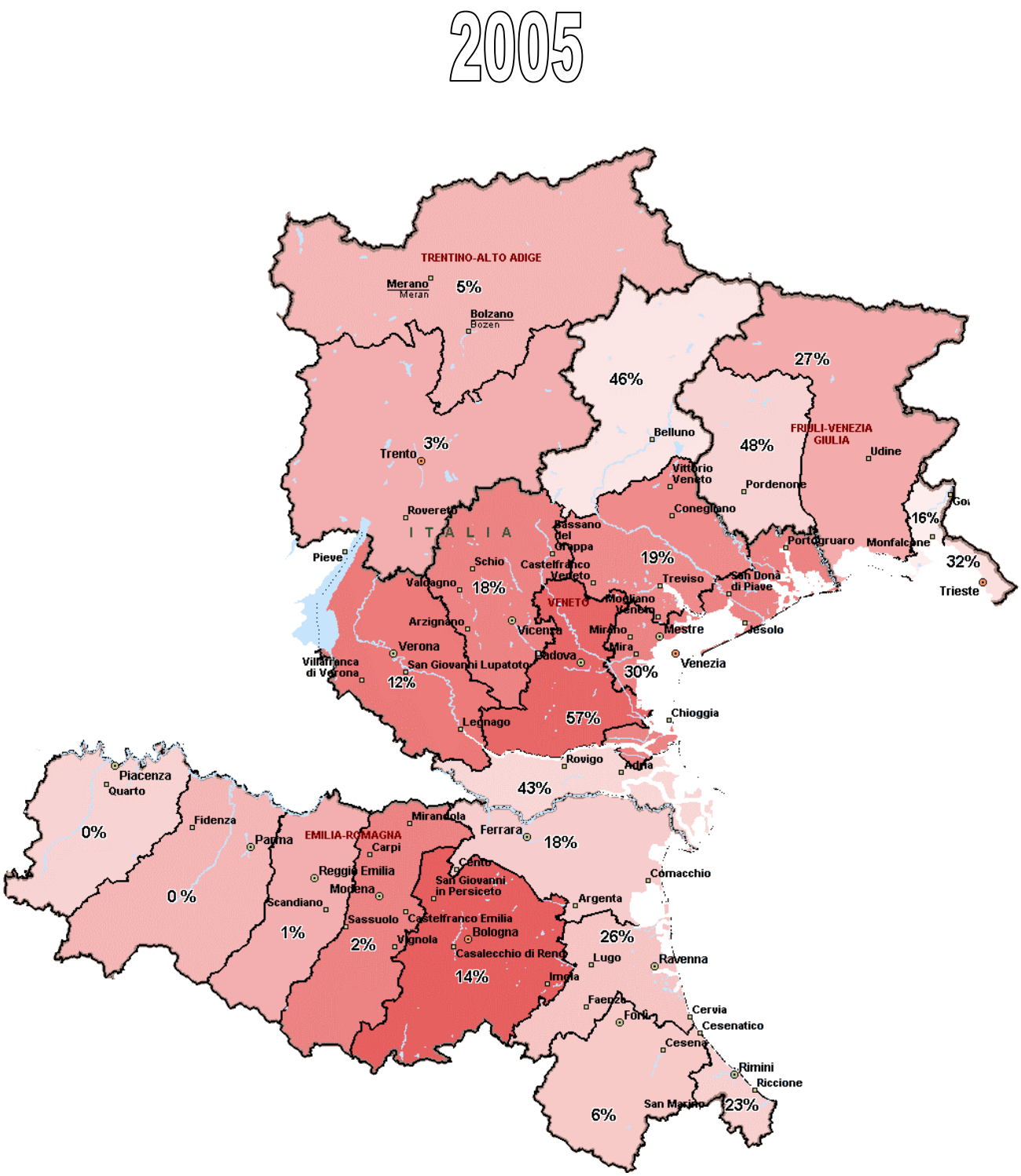
Numerosità clienti potenziali							
Da 1.313 a 1.500	Da 1.125 a 1.312	Da 938 a 1.124	Da 750 a 937	Da 563 a 749	Da 375 a 562	Da 188 a 374	Da 0 a 187

Figura 2.5a – Percentuale di contatto per provincia nel 2003.



Numerosità clienti potenziali							
Da 1.313 a 1.500	Da 1.125 a 1.312	Da 938 a 1.124	Da 750 a 937	Da 563 a 749	Da 375 a 562	Da 188 a 374	Da 0 a 187

Figura 2.5b – Percentuale di contatto per provincia nel 2004.



Numerosità clienti potenziali							
Da 1.313 a 1.500	Da 1.125 a 1.312	Da 938 a 1.124	Da 750 a 937	Da 563 a 749	Da 375 a 562	Da 188 a 374	Da 0 a 187

Figura 2.5c – Percentuale di contatto per provincia nel 2005.

Come si può notare per i soli installatori è avvenuto un generale calo di contatto, vale a dire che alcuni di questi clienti nel tempo non hanno più preso in considerazione Elettroingross per i loro acquisti. Pordenone, ad esempio, passa dal 81% del 2003 al 55% del 2004 fino al 48% del 2005.

Non è possibile capire effettivamente quali possano essere le cause scatenanti di questo fenomeno, probabilmente può dipendere dalla concorrenza che diventa più competitiva, forse da disguidi e inefficienze da parte di Elettroingross nei confronti dei clienti, oppure certi clienti potrebbero trovarsi in difficoltà e di conseguenza non acquistano, o magari più probabilmente è un mix di concause.

Di certo la % di contatto è un valore che è influenzato da molti fattori, come la qualità del servizio, i prezzi applicati, la distanza dal punto vendita, la presenza della concorrenza, ecc.

Il grafico in figura 2.6 evidenzia la percentuale di copertura del territorio che nei primi 5 – 6 km di raggio dal Punto vendita di Padova è pressoché totale, mentre su distanze superiori mostra uno scostamento tra numero di clienti attivi sui clienti potenziali dovuto anche dalla presenza dei concorrenti che mano a mano aumentano di numero con l'aumentare della distanza.

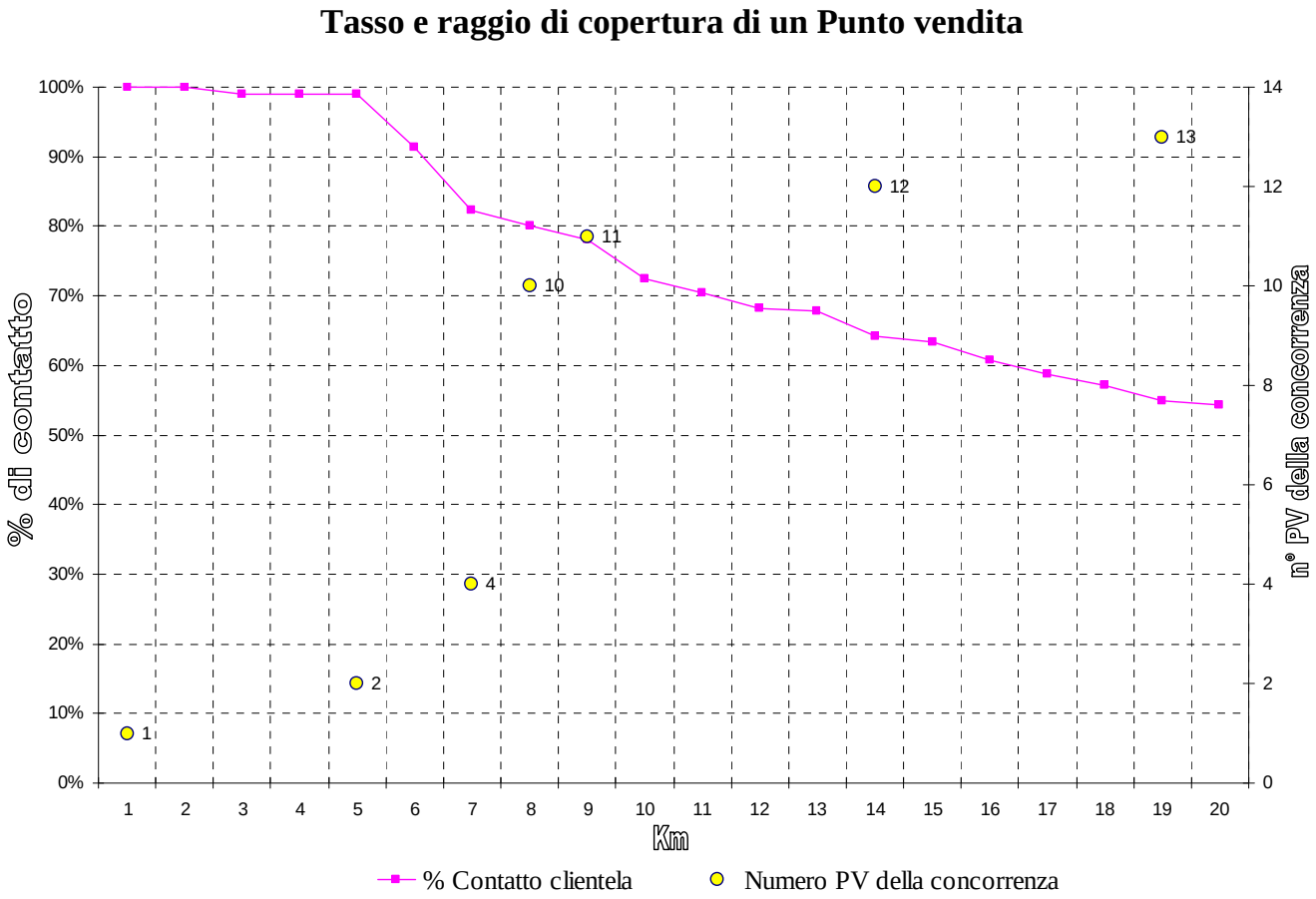


Figura 2.6 – Percentuale di contatto in base alla distanza in Km dal punto vendita.

Secondo queste considerazioni una delle possibili spiegazioni sulla generalizzata diminuzione di percentuale di contatto riscontrata in figura 2.5 a,b,c può essere la nuova apertura di punti vendita della concorrenza che hanno alterato l'equilibrio preesistente.

2.3 Strategie operative di segmentazione.

Per segmentare in modo opportuno la clientela si vuole considerare come discriminanti due fattori: le variazioni sul valore di fatturato e sulla frequenza degli acquisti per ogni cliente. Si decide di eseguire la segmentazione solamente sulla categoria installatori, poiché sono disponibili maggiori informazioni e limitando la numerosità della popolazione si procede più agevolmente.

2.3.1 Segmentazione sulle variazioni di fatturato.

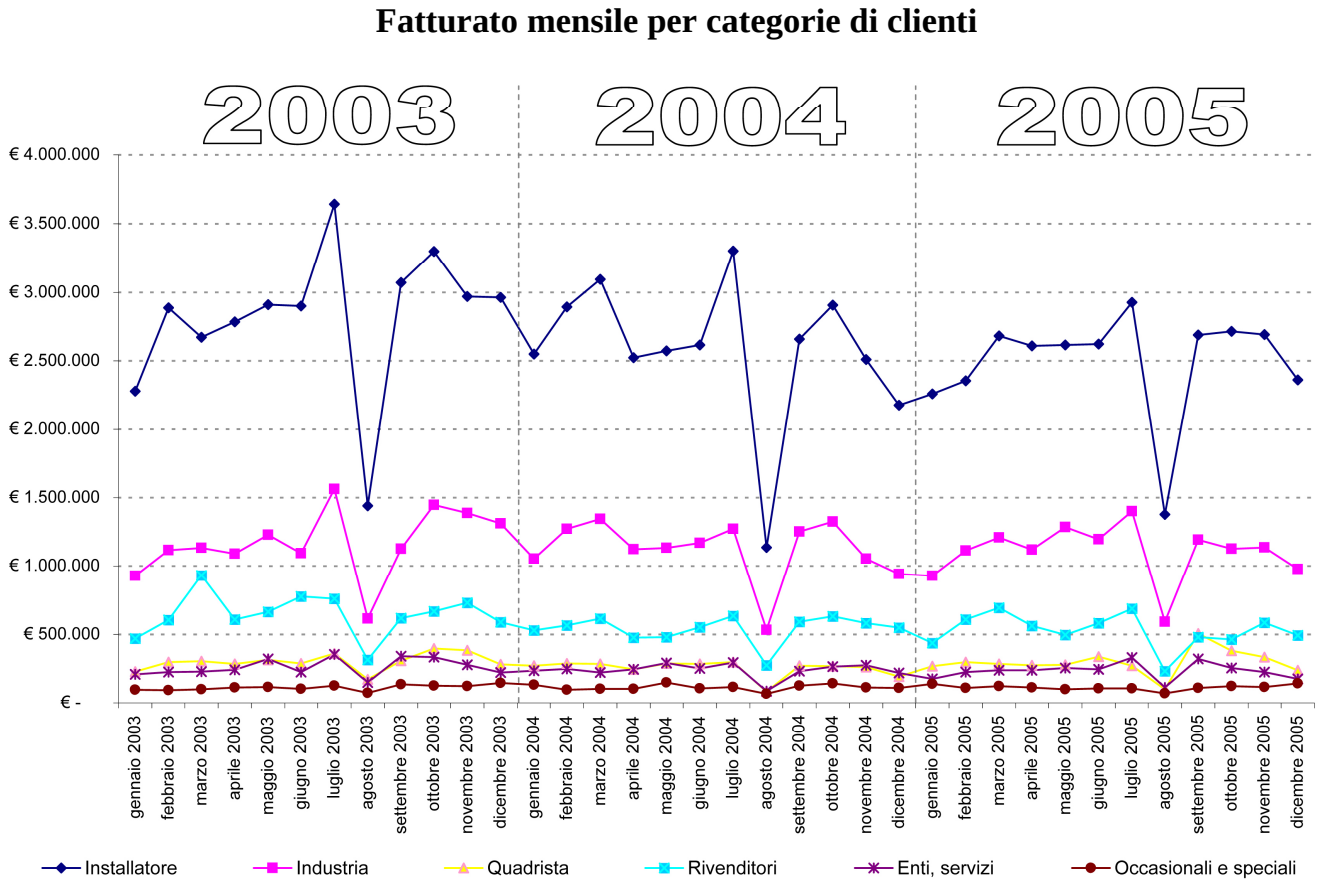


Figura 2.7 – Fatturato mensile per ogni categoria di clienti.

Questo grafico rappresenta il fatturato mensile per categorie di clienti, si vuole sfruttare una particolarità emersa in sede di indagine preliminare: la categoria installatori come la categoria industria dimostrano una certa regolarità di comportamento nel mese di luglio. Generalmente il mese di luglio rappresenta il picco massimo di fatturato procapite perchè si concentrano molte vendite nei confronti di queste categorie. Ciò è dovuto dal fatto che in agosto, mese feriale e di chiusura degli stabilimenti industriali, si approfitta per compiere installazioni, manutenzioni sia in ambito civile, commerciale e industriale. Bisogna insomma aspettarsi in tutto il secondo quadrimestre un aumento medio delle vendite.

Generalmente avviene che il primo e terzo quadrimestre come fatturato sono in media equivalenti mentre il secondo quadrimestre ha un fatturato maggiore.

Si vuole organizzare un metodo che spieghi se un cliente sta variando e in che modo il valore degli acquisti, ma il criterio non deve essere arbitrario e uguale per tutti bensì deve tenere conto della propensione all'acquisto dimostrata individualmente dal cliente.

Dal punto di vista operativo le serie storiche di vendita di ogni cliente non verranno destagionalizzate né trattate in alcun modo, in quanto non si deve fare nessuna previsione o manipolazione di sorta ma semplicemente esprimere un giudizio su come ogni singola serie storica si sia formata ed evoluta.

Sfruttando la particolarità del mercato di materiale elettrico in cui in uno dei quadrimestri (principalmente il secondo) si concentrano gli acquisti individuali sono stati considerati quattro momenti di osservazione per costruire degli indici:

- ${}^{1}Q_{t-1}$ media mensile degli acquisti in Euro del primo quadrimestre dell'anno *t-1*
- l_{t-1} gli acquisti in Euro del mese di luglio dell'anno *t-1*
- ${}^{3}Q_{t-1}$ media mensile degli acquisti in Euro dell'ultimo quadrimestre dell'anno *t-1*
- ${}^{1}Q_{t}$ media mensile degli acquisti in Euro del primo quadrimestre dell'anno *t*

Gli indici sono dei pesi costruiti in questo modo:

$$indice1 = \frac{{}^{1}Q_{t-1}}{{}^{1}Q_{t-1} + l_{t-1} + {}^{3}Q_{t-1} + {}^{1}Q_{t}} \qquad indice3 = \frac{{}^{3}Q_{t-1}}{{}^{1}Q_{t-1} + l_{t-1} + {}^{3}Q_{t-1} + {}^{1}Q_{t}}$$

$$indice2 = \frac{l_{t-1}}{{}^{1}Q_{t-1} + l_{t-1} + {}^{3}Q_{t-1} + {}^{1}Q_{t}} \qquad indice4 = \frac{{}^{1}Q_{t}}{{}^{1}Q_{t-1} + l_{t-1} + {}^{3}Q_{t-1} + {}^{1}Q_{t}}$$

La somma degli indici è uguale a 1 e ogni singolo indice rappresenta il peso degli acquisti dei rispettivi momenti considerati.
Ora l'algoritmo discriminante progettato in figura 2.8 valuta cliente per cliente se l'ultimo indice relativo al peso del primo quadrimestre medio dell'anno *t* (*indice4*) sia:

- inferiore a tutti gli altri, ovvero il cliente all'inizio dell'anno *t* sta riducendo la spesa in termini di valore (fatturato), allora lo consideriamo cliente in DIMINUZIONE;
- maggiore a tutti gli altri, il cliente sta aumentando la spesa in termini di valore all'inizio dell'anno *t*, allora lo consideriamo cliente in CRESCITA;
- inferiore al max{*indice1,indice2,indice3*} e superiore al min{*indice1,indice2,indice3*} , allora lo consideriamo cliente COSTANTE.

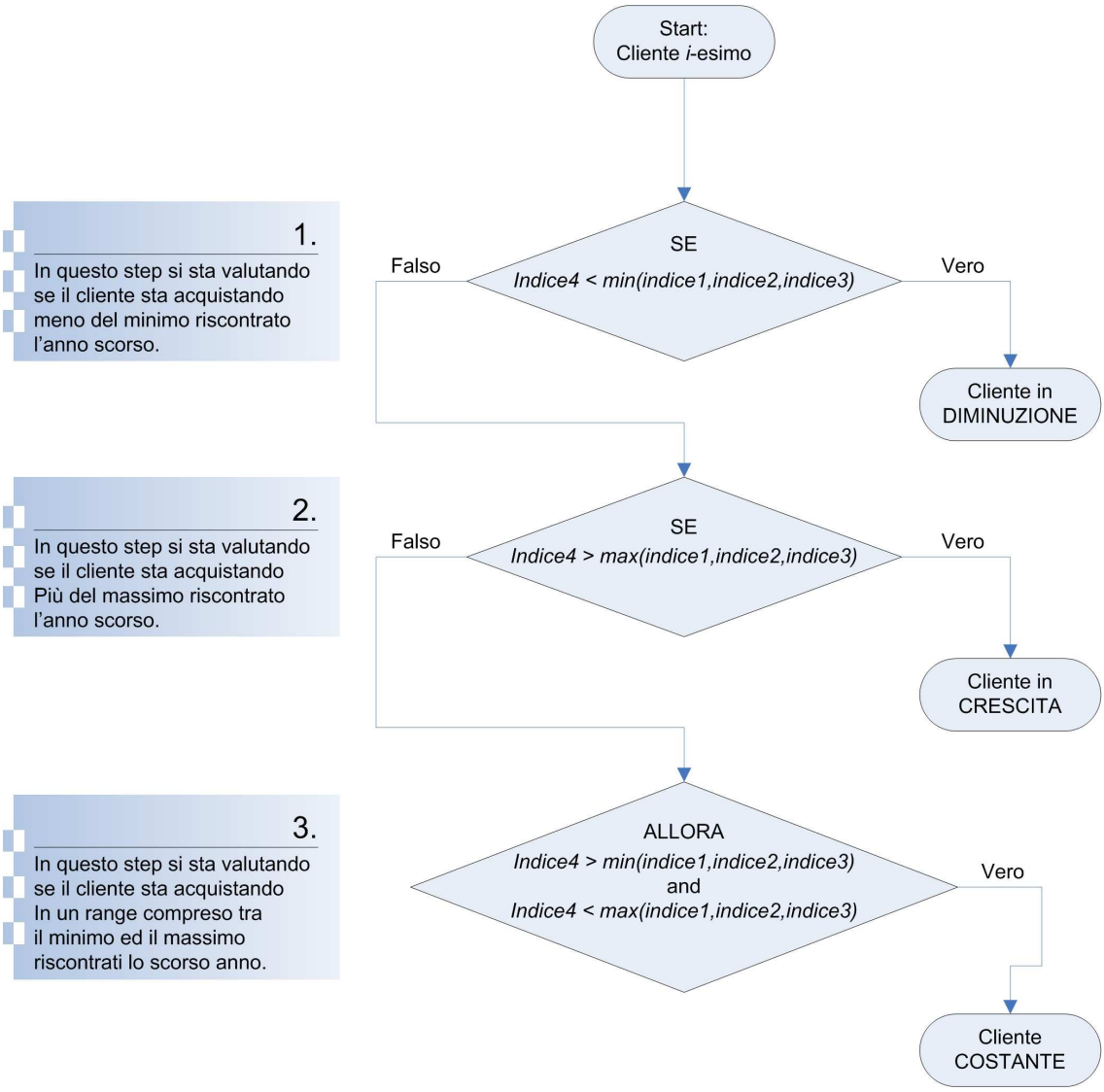


Figura 2.8 – Flowchart dell'algoritmo implementato.

L'analisi è stata applicata alla categoria installatori comprensiva di 3787 elementi misurando gli indici sul 2003 e sul primo quadrimestre 2004.

Risultato.

Sono state ottenute le 3 fasce desiderate con:

- 682 clienti per la fascia CRESCITA;
- 2875 clienti per la fascia COSTANTE;
- 230 clienti per la fascia DIMINUZIONE.

Fatturato mensile dal 2003 al 2005 per fasce clienti individuate.

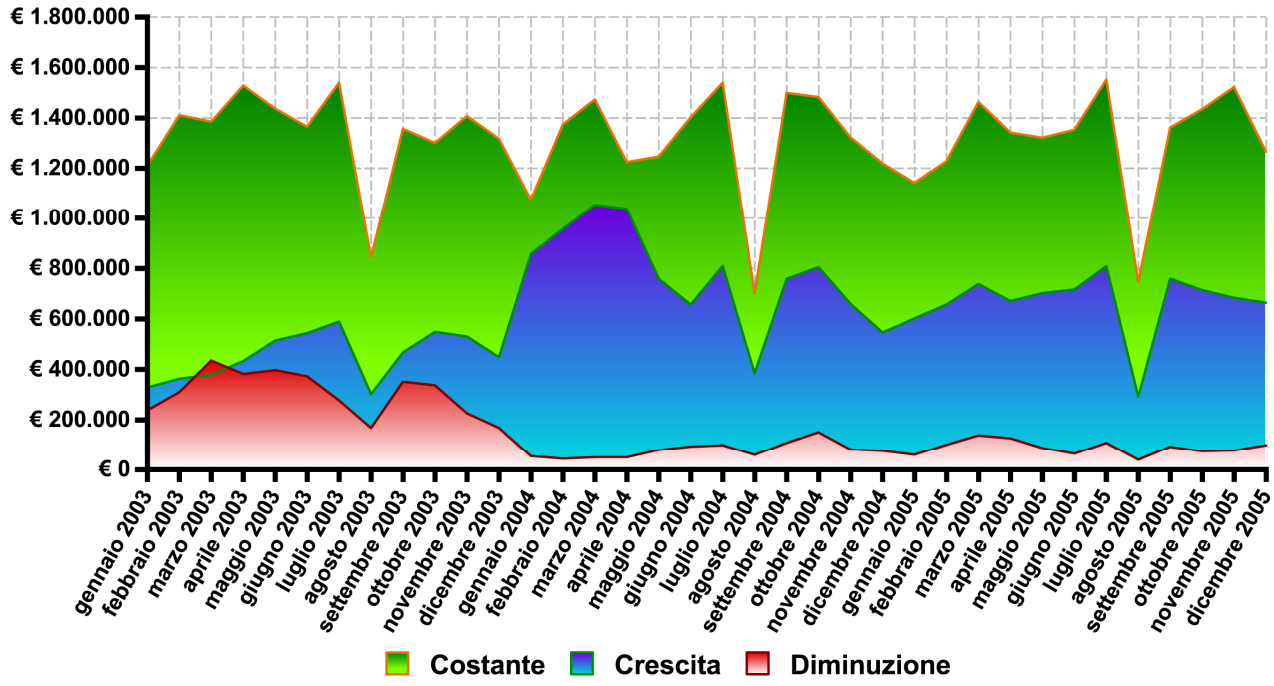


Figura 2.9 – Fatturato dei gruppi ottenuti.

Il grafico mostra la segmentazione applicata dal momento del calcolo degli indici (tutto il 2003 e il primo quadrimestre 2004) fino a tutto il 2005 per constatare che tale suddivisione ha individuato comportamenti d'acquisto che sono rimasti stabili nell'arco del tempo considerato.

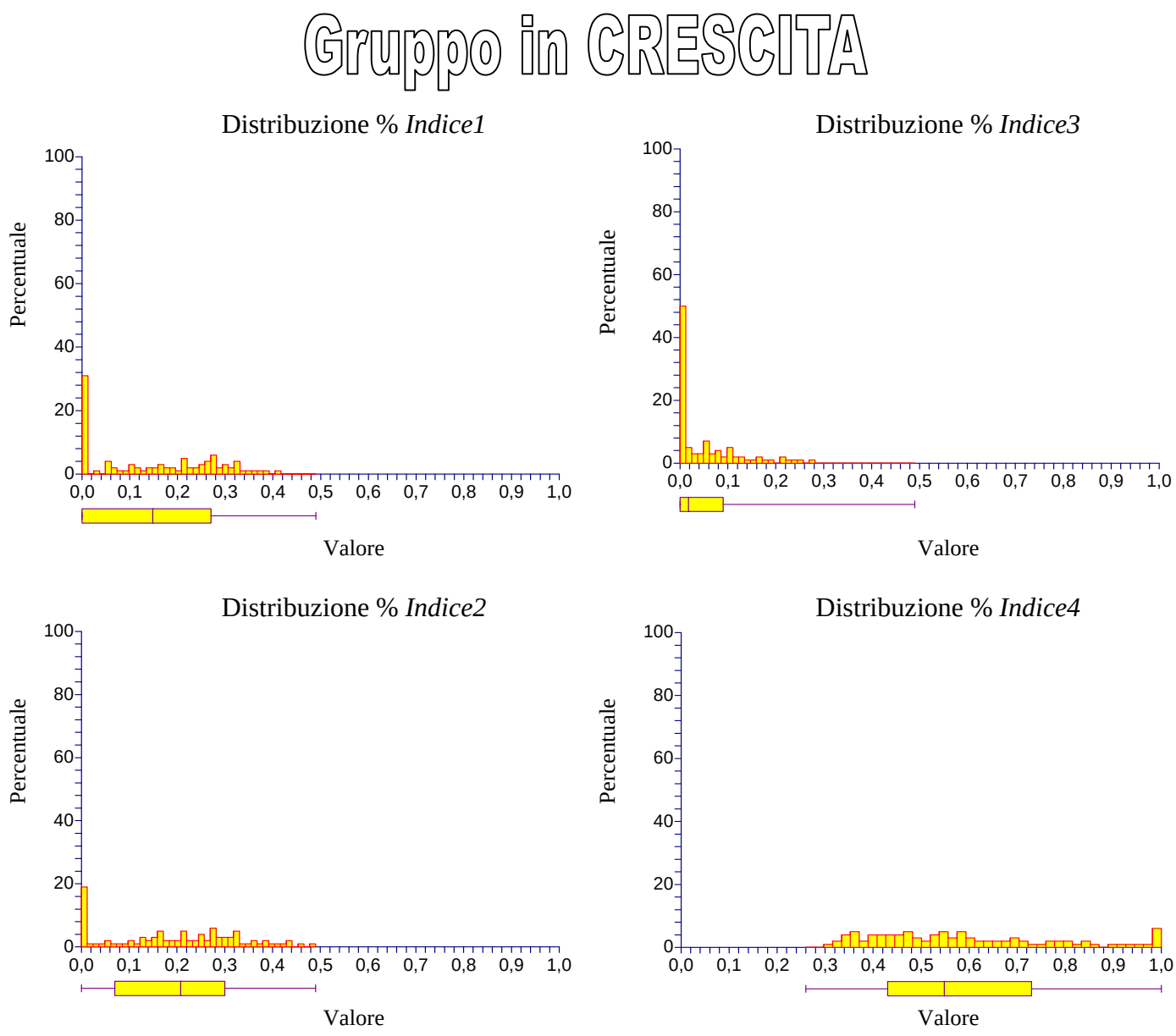
Il risultato sembra buono perché sono state rispettate le nostre intenzioni: si può notare la fascia clienti in DIMINUZIONE che hanno dimostrato di acquistare per un certo livello nel 2003 e successivamente dal 2004 hanno appiattito verso il basso il volume di fatturato, evidentemente è stato scelto un altro fornitore. Una cosa interessante è che non c'è più stato un recupero per questa fascia clienti, ma nemmeno la totale scomparsa o l'azzeramento degli acquisti e la motivazione è che sostanzialmente i clienti non si riforniscono come prima scelta da Elettroingross ma solo ed esclusivamente se non trovano merce particolare dal nuovo fornitore abituale. Elettroingross, infatti, ha il pregio di fornire anche materiale difficilmente reperibile da altri grossisti.

La fascia di clienti in CRESCITA ha dimostrato di acquistare con prudenza, quasi come se volesse testare Elettroingross in tutto il 2003, per poi esplodere nel primo quadrimestre 2004 e attestarsi successivamente con costanza ad un livello medio molto superiore al 2003.

La fascia di clienti COSTANTI ha invece dimostrato sin da subito una certa regolarità nella spesa ed è il segmento più numeroso e di valore maggiore.

Le figure 2.10 a,b,c mostrano le distribuzioni degli indici calcolati suddivise per ogni fascia individuata.

Per ogni indice viene rappresentato congiuntamente un istogramma ed un boxplot.

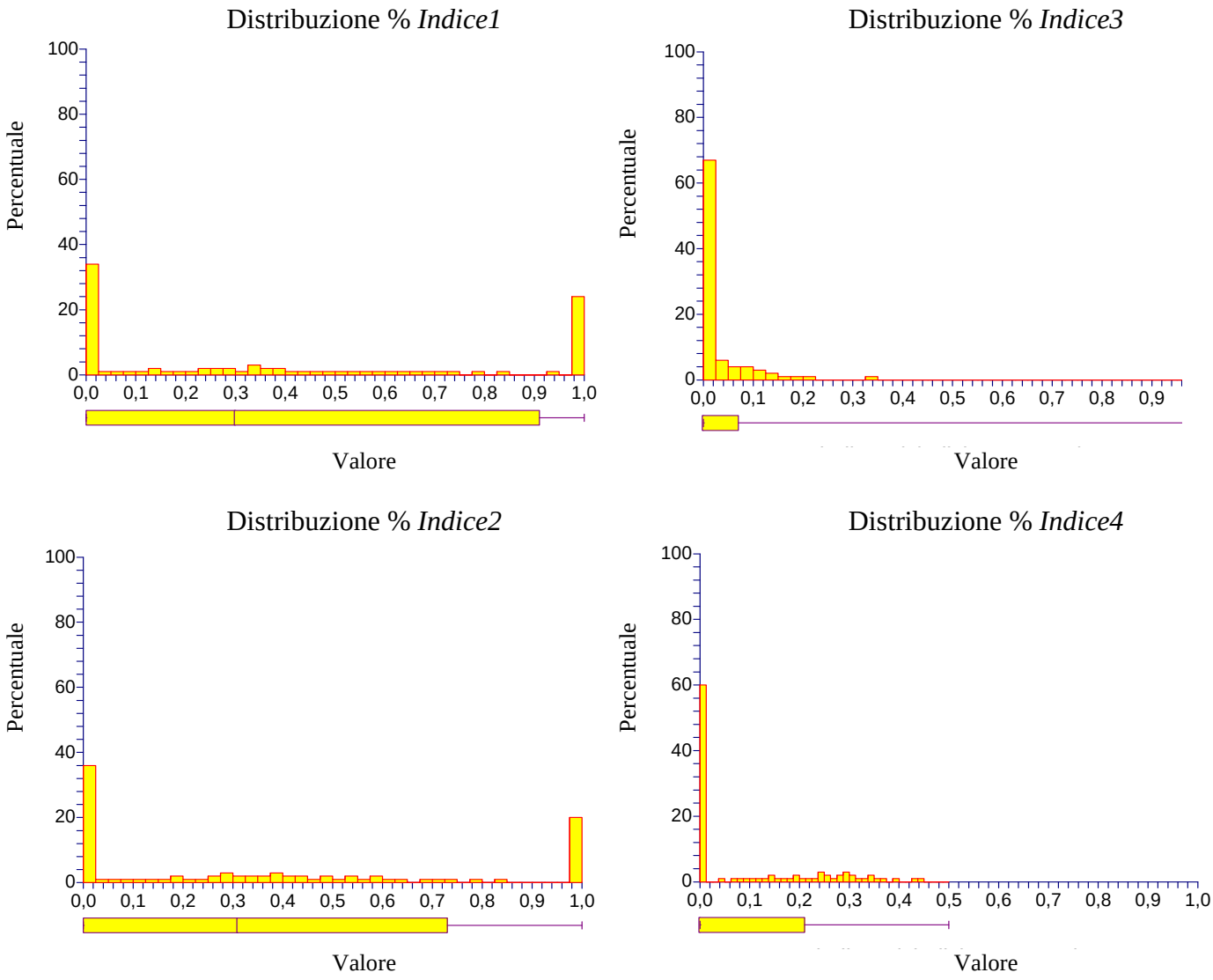


Crescita n°= 682						
	Mean	Standard Deviation	Median	Minimum	Maximum	Range
indice1	0,15	0,13	0,15	0	0,49	0,49
indice2	0,20	0,14	0,21	0	0,49	0,49
indice3	0,06	0,08	0,02	0	0,49	0,49
indice4	0,59	0,20	0,55	0,26	1	0,74

Figura 2.10a – Distribuzione indici del gruppo in Crescita e statistiche descrittive.

I clienti in Crescita dimostrano di avere superato il massimo dello scorso anno in maniera decisa. Il massimo del 2003 era come previsto l'*indice2* relativo al mese di luglio.

Gruppo COSTANTE

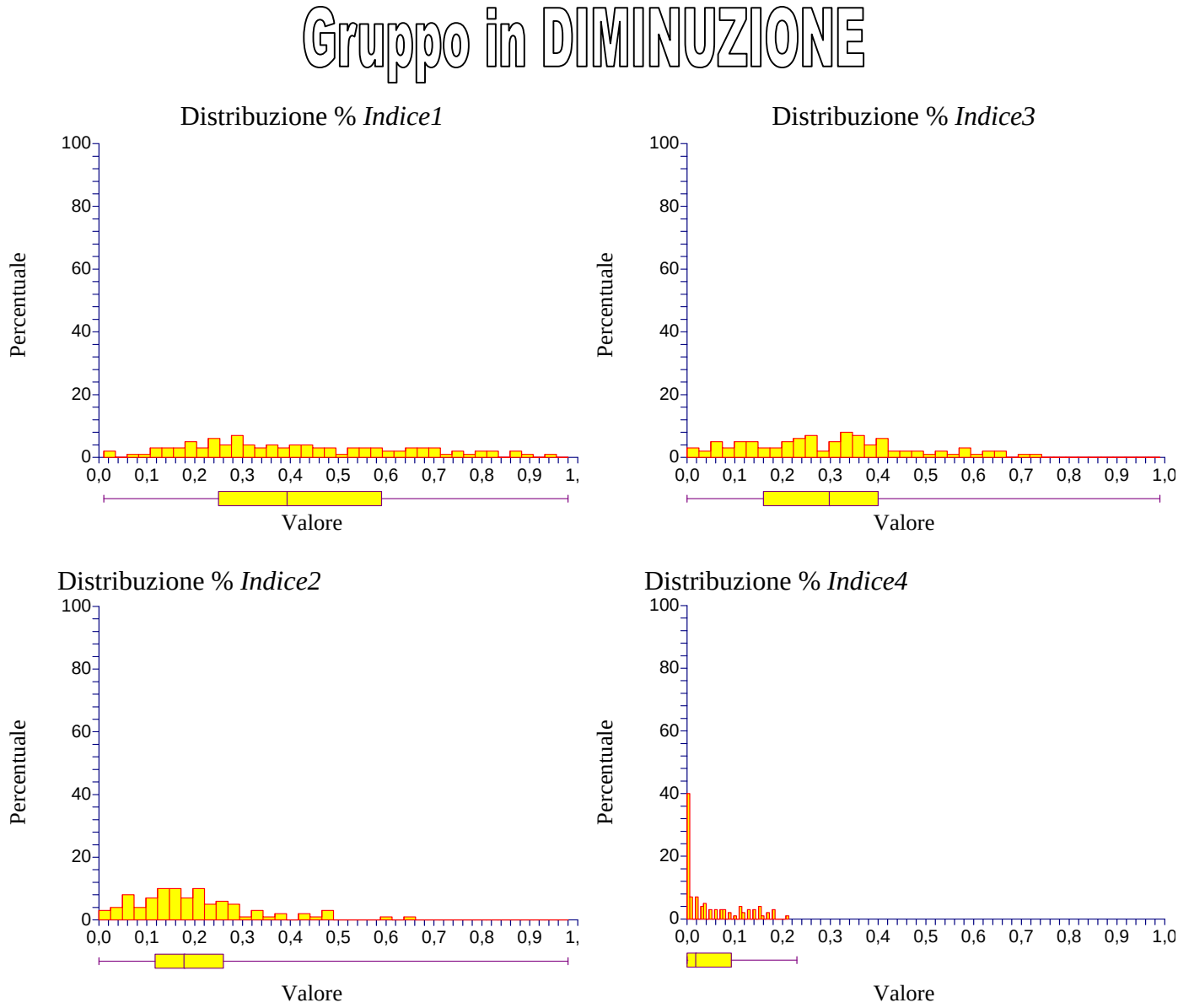


Costante n° = 2875						
	Mean	Standard Deviation	Median	Minimum	Maximum	Range
indice1	0,41	0,40	0,30	0	1	1
indice2	0,39	0,39	0,31	0	1	1
indice3	0,09	0,25	0,00	0	1	1
indice4	0,10	0,14	0,00	0	0,5	0,5

Figura 2.10b – Distribuzione indici del gruppo Costante e statistiche descrittive.

Come si può notare il gruppo Costante ha in media un *indice4* leggermente superiore all'*indice3* che risulta essere il minimo tra *indice1*,*indice2*,*indice3* definendo cosi la categoria Costante. Un'altra considerazione da fare è che l'*indice2* che ci saremmo aspettati fosse superiore agli altri indici dello stesso

anno è invece mediamente poco al di sotto dell'*indice1*.



	Diminuzione n°= 230					
	Mean	Standard Deviation	Median	Minimum	Maximum	Range
indice1	0,43	0,22	0,40	0,01	0,98	0,97
indice2	0,22	0,16	0,18	0	0,98	0,98
indice3	0,31	0,19	0,30	0	0,99	0,99
indice4	0,05	0,06	0,02	0	0,23	0,23

Figura 2.10c – Distribuzione indici del gruppo in Diminuzione e statistiche descrittive.

La categoria clienti in diminuzione ha ridotto notevolmente gli acquisti, infatti, l'*indice4* è circa ¼ dell'*indice2* che rappresenta il minimo dell'anno precedente.

Un dettaglio importante è proprio l'*indice2* che è il minore tra i tre del 2003 dove sappiamo che nel mese di luglio ci si aspetterebbe un incremento degli acquisti.
I clienti di questa categoria già nel mese di luglio 2003 fanno capire di non comportarsi come i clienti delle altre due categorie ribassando gli acquisti: questo è un chiaro segnale.

2.3.2 Segmentazione in base alla frequenza d'acquisto.

Viene costruito un indice che dà una valutazione del cliente in base alla sua frequenza d'acquisto e strutturato in modo tale da attribuire maggiore importanza al passato recente.
L'impostazione del business dei clienti è tale che il fabbisogno di approvvigionarsi è molto frequente nel tempo.
L'installatore, ad esempio, per ogni commessa di lavoro ottenuta provvede all'individuazione e all'acquisto del materiale necessario, contando solo in minima parte sul suo magazzino personale.
Infatti per questa categoria in particolare, ma non è l'unica, è normale aspettarsi acquisti con frequenza addirittura giornaliera.
L'approccio è di tipo just in time; si provvede all'acquisto nel momento in cui si manifesta il fabbisogno, garantendo il contenimento dei costi in quanto si tiene in magazzino il minimo indispensabile.
Per questa analisi si considera il mese come elemento unitario d'osservazione; è un buon compromesso perché da un lato rappresenta un buon dettaglio e dall'altro la mole di dati da utilizzare è gestibile. Se usassimo come elemento unitario d'osservazione il giorno si guadagnerebbe sicuramente in dettaglio ma a discapito della effettiva governabilità dei dati.
L'obiettivo non è semplicemente creare una grandezza che conteggi quante mensilità sia comparso un cliente ma valutare in che modo è stato presente.

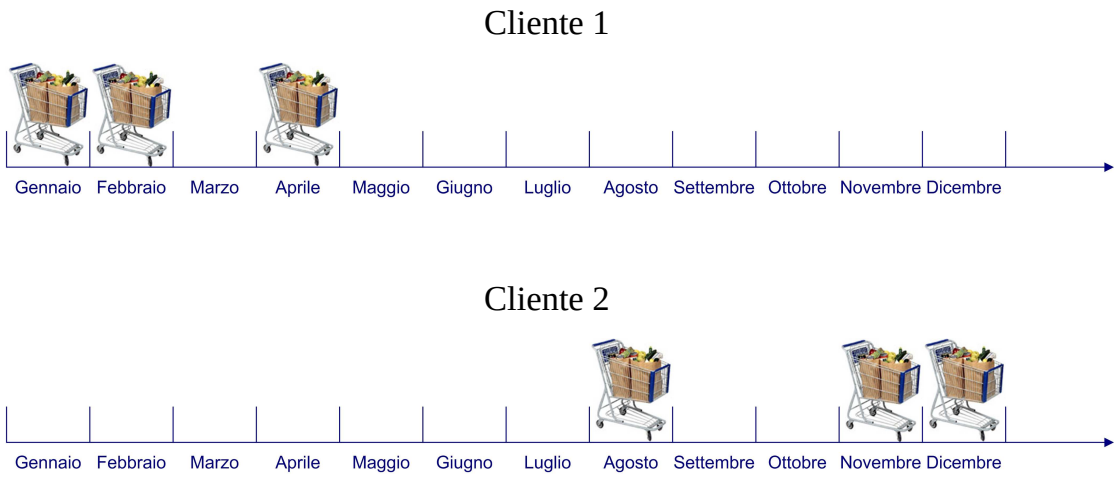


Figura 2.11 – Momenti d'acquisto di due clienti.

Ad esempio un semplice conteggio dei mesi in cui i due clienti in figura 2.11 hanno acquistato li metterebbe alla pari, mentre è chiaro che il secondo cliente è presumibilmente più attivo del primo che ormai non acquista da tempo.

Definiamo l'**Indice frequenza Acquisti** in questo modo:

$$I_f = \sum_{i=1}^{12} \frac{a \times i}{12} \qquad \text{con} \begin{cases} a = +1 & \text{se è stato acquistato nel mese } i\text{-esimo} \\ a = -1 & \text{se non è stato acquistato nel mese } i\text{-esimo} \end{cases}$$

Quindi il range di I_f è:

$$-6{,}5 \le I_f \le 6{,}5$$

Per renderlo maggiormente comprensibile si vuole che il range di validità sia $0 \le I_f \le 1$ e si procede modificando I_f in questo modo:

$$I_f = \frac{\sum_{i=1}^{12} \frac{a \times i}{12} + 6{,}5}{13}$$

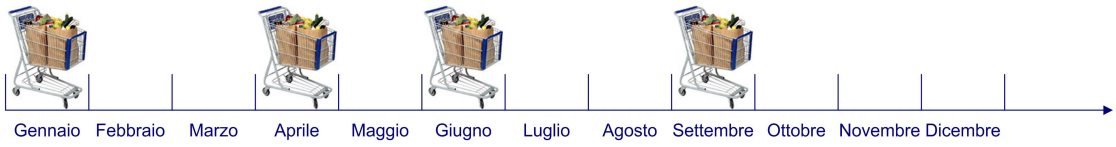
Nel caso dei 2 clienti del precedente esempio avremo che:

$$I_f(Cliente1) = \frac{\frac{1+2-3+4-5-6-7-8-9-10-11-12}{12}+6,5}{13} = 0,09$$

$$I_f(Cliente2) = \frac{\frac{-1-2-3-4-5-6-7+8-9-10+11+12}{12}+6,5}{13} = 0,40$$

Viene dimostrato che I_f è in grado di esprimere in modo quantitativo la qualità di presenza di un cliente.

Ecco degli opportuni esempi:



Indice frequenza acquisti pari a 0,26

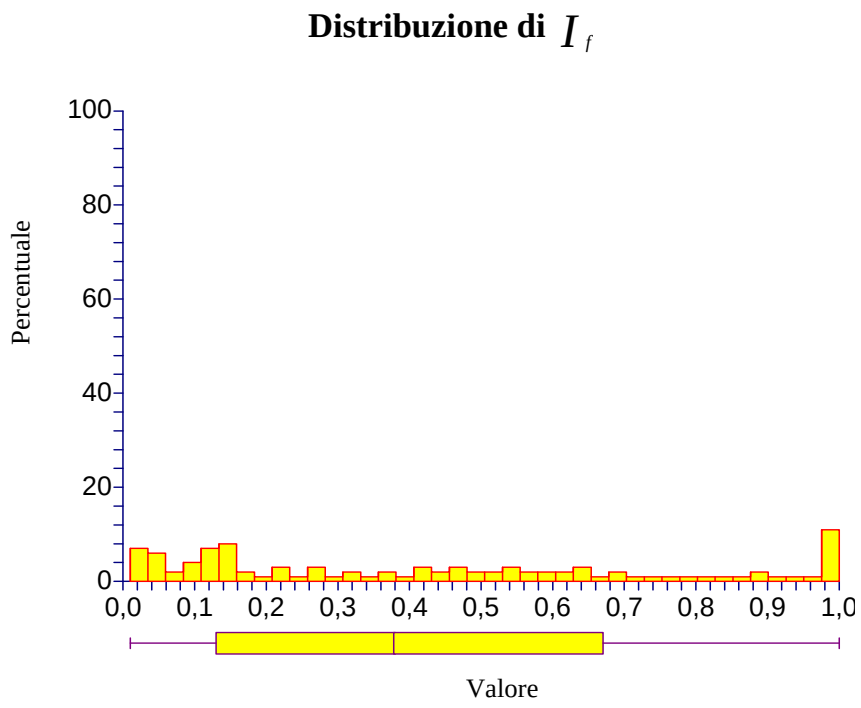
Indice frequenza acquisti pari a 0,50

Indice frequenza acquisti pari a 0,76

Indice frequenza acquisti pari a 1,00

Figura 2.12 – Esempi di valore dell' Indice frequenza acquisti in casi di combinazioni di momenti d'acquisto.

Un indice che si avvicina allo zero evidenzia una tendenza a non acquistare e a scomparire da Elettroingross, un valore prossimo a 0,5 indica un comportamento sostanzialmente neutro (il ritmo d'acquisto è largo e cadenzato) mentre un valore vicino a 1 specifica una completa presenza del cliente.



Count	Mean	Standard Deviation	Median	Minimum	Maximum	Range
3787	0,42	0,33	0,38	0,01	1	0,99

Figura 2.13 – Distribuzione dell' Indice frequenza acquisti per l'intera popolazione e statistiche descrittive.

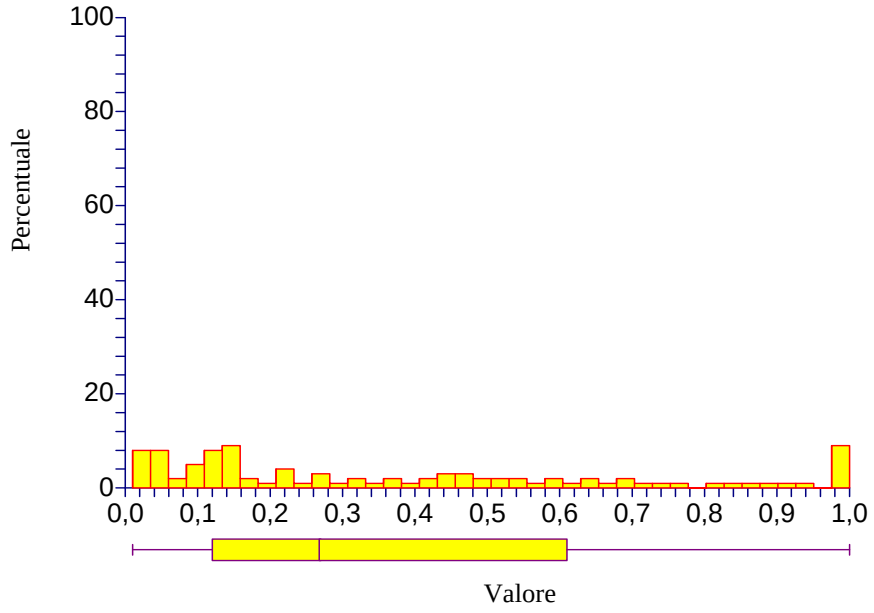
Osservando gli indici di tendenza centrale si scopre che la clientela è leggermente posta sotto il valore di neutralità 0,5.

La distribuzione di I_f in figura 2.13 dimostra anche che il 25% dei clienti attivi è compresa tra 0,68 e 1,00 ovvero trattasi di clienti con alte frequenze d'acquisto per cui possiamo stabilire che sono clienti fidelizzati, un altro 25% risiede tra il 0,40 e 0,68 intorno alla soglia di neutralità, il 50% rimanente si trova in un range molto basso.

Queste indicazioni, se sfruttate da un punto di vista manageriale, fanno pensare che mediamente ci siano ampi margini di crescita sui clienti già conosciuti e attivi.

Le tre fasce individuate precedentemente discriminando per valore sono caratterizzate da queste distribuzioni dell'indice di frequenza:

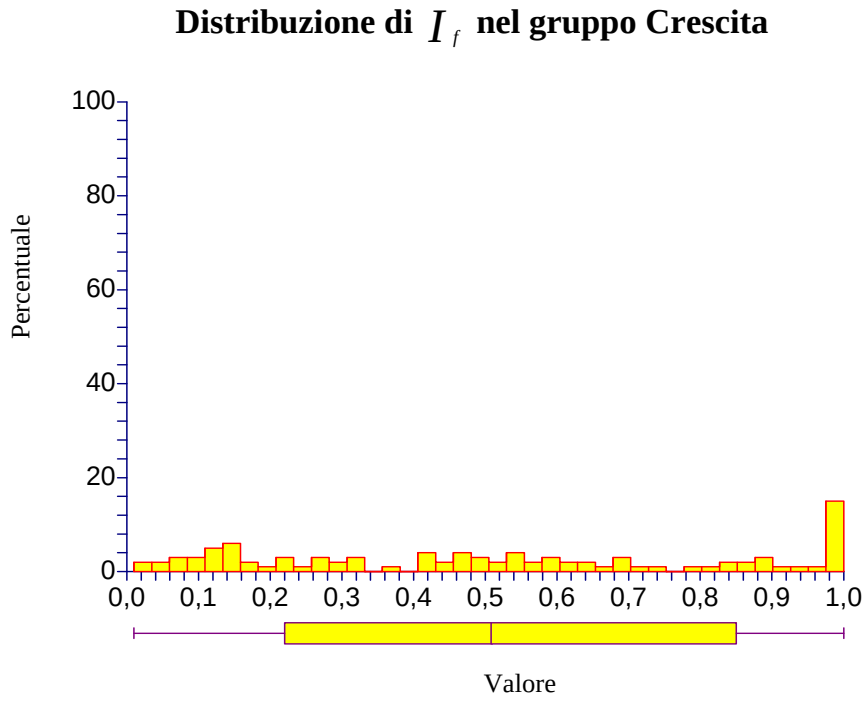
Distribuzione di I_f nel gruppo Costante



Count	Mean	Standard Deviation	Median	Minimum	Maximum	Range
2875	0,38	0,32	0,27	0,01	1	0,99

Figura 2.14a – Distribuzione dell' Indice frequenza acquisti per il gruppo costante e statistiche descrittive.

Osservando la distribuzione per il gruppo costante si incominciano a trovare i limiti di tutta la segmentazione empirica applicata fino a questo punto: se da un lato abbiamo fatto considerazioni favorevoli sui risultati ottenuti dall'altro constatiamo che per questo gruppo l'indice di frequenza è distribuito in modo molto simile alla popolazione originaria, quindi non sono state individuate peculiarità proprie di questo profilo di clienti individuato.



Count	Mean	Standard Deviation	Median	Minimum	Maximum	Range
682	0,52	0,32	0,51	0,01	1	0,99

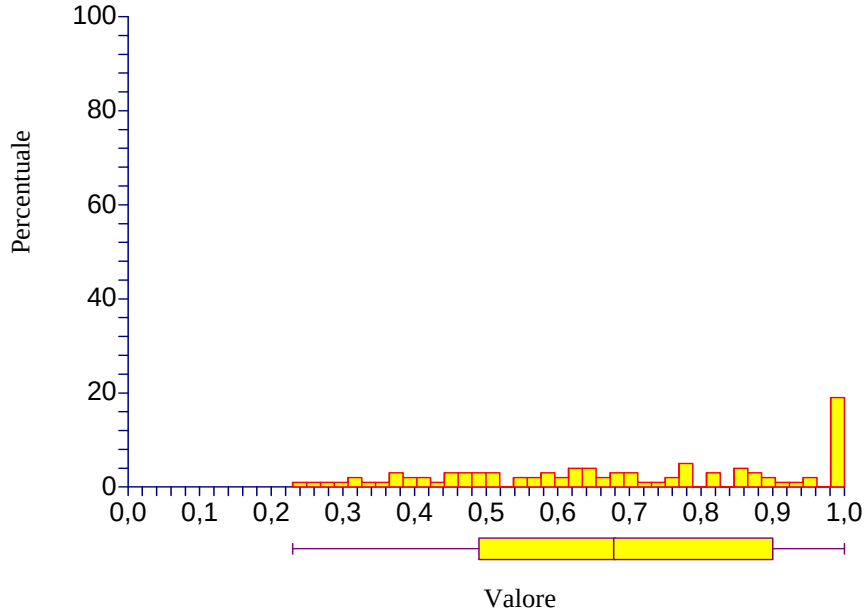
Figura 2.14b – Distribuzione dell' Indice frequenza acquisti per il gruppo in crescita e statistiche descrittive.

La situazione del gruppo in crescita è migliore, la media e la mediana ed il boxplot indicano che la distribuzione è concentrata più in alto.

Osservando il grafico viene da pensare che i clienti in crescita non siano individui che stanno aumentando di valore i loro singoli acquisti ma probabilmente la loro frequenza e conseguentemente il fatturato globale aumenta. Si nota comunque che è presente un picco massimo in corrispondenza del valore 1,00.

In tal caso sono clienti che stanno aumentando il valore dei singoli acquisti dato che per quanto riguarda la frequenza d'acquisto sono sempre presenti: si stà passando da un rapporto commerciale debole ad un rapporto forte, argomento trattato nel paragrafo 2.2.

Distribuzione di I_f nel gruppo Diminuzione



Count	Mean	Standard Deviation	Median	Minimum	Maximum	Range
230	0,69	0,23	0,68	0,23	1	0,77

Figura 2.14c – Distribuzione dell' Indice frequenza acquisti per il gruppo in diminuzione e statistiche descrittive.

Il gruppo in diminuzione ha una distribuzione concentrata ancora più in alto dei clienti in crescita e I_f si muove in un range più ristretto rispetto le altre distribuzioni.

Poiché stanno diminuendo l'ammontare dei loro acquisti si sta verificando molto probabilmente il fenomeno contrario di quello che avviene per i clienti in crescita: si stanno riducendo il numero degli acquisti.

Analogamente al gruppo in crescita è presente un picco massimo in corrispondenza del valore 1,00 ma probabilmente con intenzioni opposte: se da un lato la presenza di questi clienti è totale, dall'altro si riduce la spesa.

Significa che su questi clienti si sta perdendo penetrazione commerciale, sta avvenendo un passaggio sul versante grado di intensità commerciale da rapporto forte a rapporto debole visibile nella figura 2.3 a pagina 12.

2.4 Conclusioni.

La segmentazione empirica implementata per il caso Elettroingross ha dato dei risultati buoni per quanto concerne la segmentazione in valore.
Sono stati individuati i clienti con una tendenza a ridurre notevolmente il valore degli acquisti, rendendo possibile impostare le prime misure commerciali per "invogliare" ad acquistare meglio e di più.
Successive analisi interne all'azienda permetteranno di capire invece per quali motivi i clienti in crescita stanno aumentando i consumi e si cercherà di rendere sistematiche le peculiarità necessarie per fidelizzare questa particolare clientela.
Quando è stata introdotta la variabile frequenza, i gruppi precedentemente creati non hanno beneficiato di ulteriori conferme di validità dei raggruppamenti effettuati.
L'indice frequenza acquisti, si è visto, funziona molto bene e come è stato detto precedentemente attraverso un' espressione quantitativa viene stabilita la qualità della presenza di un cliente.
La difficoltà riscontrata in questa fase del lavoro è stata l'impossibilità di far convergere contemporaneamente gli indici sul valore di fatturato con l'indice di frequenza acquisti nell'algoritmo discriminante.
Per rimediare a questa limitazione si potrebbe molto semplicemente "spaccare" i tre gruppi in due parti ciascuno:

1. gli elementi del gruppo con buona frequenza d'acquisto ($I_f \geq 0,50$);
2. gli elementi del gruppo con cattiva frequenza d'acquisto ($I_f < 0,50$).

In tal caso l'apporto informativo dell'analisi migliorerebbe di poco, ecco perché si è cercato una strada diversa per trovare un algoritmo che nell'azione di stratificare permettesse di valutare contemporaneamente molte variabili e la scelta è ricaduta sulla *Cluster Analysis*.

3.

La Cluster Analysis.

3.1 Introduzione alla Cluster Analysis.

Per *cluster analysis* si intende un insieme di procedure e metodologie utili a ricavare da una popolazione[1] di dati una struttura a gruppi.

La *cluster analysis* è stata proposta e sperimentata soprattutto dagli anni ’60 in poi, anche se la prima comparsa avviene nel 1939 quando Tryon R. C. pubblica una monografia proprio con il titolo di *Cluster Analysis*.

La prima esposizione sistematica di tali tecniche risale al 1963 ad opera di Sokal e Sneath che in quell’anno pubblicano *Principles of numerical taxonomy*.

Le tecniche di *clustering* sono state applicate ad un'ampia varietà di problemi di ricerca.

Hartigan (1975). fornisce un eccellente sommario dei molti studi pubblicati che segnalano i risultati delle analisi compiute con la *cluster analysis*. Per esempio, nel campo della medicina, le malattie, le cure per le malattie o i sintomi delle malattie possono condurre a tassonomie[2] molto utili.

Nel campo della psichiatria, la *cluster analysis* ha contribuito alla diagnosi corretta dei sintomi della paranoia, della schizofrenia, ecc. Tutto ciò si rivela essenziale per la buona riuscita della terapia da intraprendere.

[1] Popolazione: insieme finito o infinito di unità statistiche oggetto d’indagine.

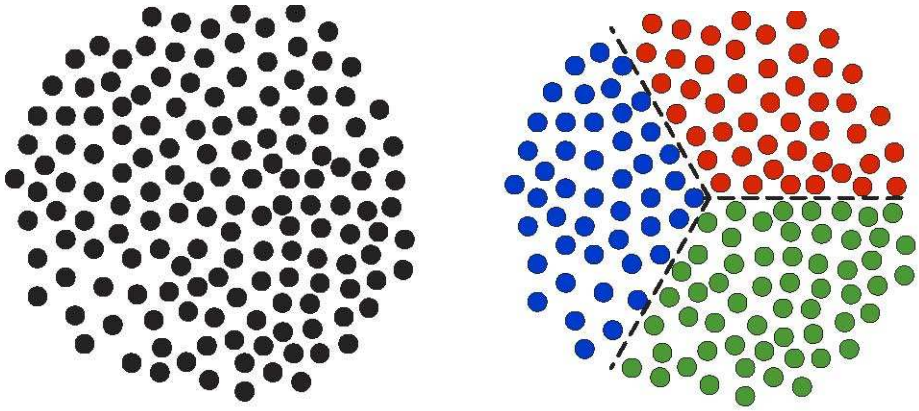
[2] Con il termine Tassonomia (dal greco ταξινομία (taxinomia) dalle parole taxis = ordine e nomos = regole) ci si riferisce sia alla classificazione gerarchica di concetti, sia al principio stesso della classificazione. Tutti i concetti, gli oggetti animati e non, i luoghi e gli eventi possono essere classificati seguendo uno schema tassonomico. La tassonomia è una struttura ad albero di istanze (o categorie); è composta inizialmente da un'istanza singola, il nodo radice, le cui proprietà si applicano a tutte le altre istanze sottostanti della gerarchia (sotto-categorie). I nodi sottostanti alla radice costituiscono categorie più specifiche le cui proprietà caratterizzano il sotto-gruppo.

In archeologia, i ricercatori hanno tentato di stabilire le tassonomie degli attrezzi di pietra, degli oggetti funerei, e di altri reperti ancora applicando le tecniche di *clustering*.

In generale, ogni qualvolta si deve classificare una grande mole di informazioni in gruppi espressivi e trattabili, l'analisi dei gruppi è un ottimo strumento.

Anche in ambito manageriale la *cluster analysis* è di grande aiuto; per definire gli obiettivi e le modalità di una strategia commerciale, bisogna essere in grado di valutare caratteristiche, bisogni e comportamenti degli acquirenti in modo da limitare il mercato in cui si intende operare e adeguare l'offerta a ciascun gruppo di clienti individuato presumendo vengano richiesti specifici prodotti e servizi, verso i quali dovranno essere indirizzate politiche di vendita altrettanto specifiche.

Con il termine *cluster analysis*, o analisi dei gruppi o delle classi, si intendono le procedure che permettono di individuare, all'interno di un insieme di oggetti di qualsiasi natura, alcuni sottoinsiemi, i clusters appunto, mutuamente esclusivi e tendenzialmente omogenei al loro interno.



Insieme di dati Dati riorganizzati in gruppi

Figura 3.1 – Rappresentazione dei dati e dei gruppi ottenuti con la cluster analysis.

Le tecniche di *cluster analysis* creano i gruppi in modo tale che ogni osservazione sia molto simile a tutte le altre che appartengono allo stesso gruppo, in funzione di alcuni criteri prestabiliti dal ricercatore.

Alla fine del procedimento, i *cluster* finali dovrebbero esibire un'alta omogeneità interna (intra-cluster) ed un'alta eterogeneità esterna (inter-cluster). Quindi, se la classificazione ha successo, gli oggetti all'interno dei *cluster* saranno vicini tra loro, mentre gli oggetti che appartengono a differenti *cluster* saranno più lontani tra loro (Barbarito, 1999).

Punto di partenza di ogni applicazione di *cluster analysis* è la disponibilità di un collettivo statistico (anche campionario) di *n* elementi ciascuno rappresentato da *p* variabili.

I dati si possono organizzare in una matrice come la seguente:

$$X = \begin{pmatrix} x_{11} & x_{22} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

La *cluster analysis* rientra fra le tecniche di tipo esplorativo e pertanto non è necessaria alcuna assunzione a priori, impone però una serie di decisioni da parte del ricercatore, prima, durante e dopo l'analisi.

PRIMA	• Scelta delle variabili • Criteri di similarità-distanza
DURANTE	• Tecniche di aggregazione • Numero dei gruppi da ottenere
DOPO	• Valutazione della qualità della soluzione • Scelta fra le diverse possibili soluzioni alternative

Tabella 3.2 – Le decisioni nella Cluster Analysis

Ovviamente scelte diverse portano a risultati diversi, pertanto, questa componente di arbitrarietà è stata fonte di notevoli critiche, ma evidentemente

nelle scienze il fattore di soggettività accomuna tutti i procedimenti di analisi multivariata dei dati. È infatti tipico dei procedimenti di conoscenza scientifica un processo di riduzione e di semplificazione controllata delle informazioni disponibili, per favorire la comprensione dei fenomeni (Massart et al., 1988).

Le fasi del processo di analisi dei gruppi sono le seguenti:

− Scelta delle unità di osservazione;

− scelta delle variabili; omogeneizzazione delle scale di misura;

− scelta della misura di similarità o di diversità tra le unità statistiche (misure di somiglianza e di diversità);

− scelta del numero dei gruppi;

− scelta dell'algoritmo di classificazione: cluster analysis gerarchica; cluster analysis non gerarchica;

− interpretazione dei risultati ottenuti.

3.2 Misure di prossimità e distanza fra unità statistiche.

Per classificare e raggruppare le unità statistiche in gruppi omogenei è necessario introdurre la nozione di prossimità o similarità. Gli indici di prossimità tra coppie di unità statistiche forniscono le informazioni preliminari indispensabili per poter individuare gruppi di unità omogenee.

Un indice di prossimità tra due generiche unità statistiche x_i e x_j è definito come una funzione dei rispettivi vettori riga nella matrice dei dati:

$$IP_{ij} = f(x_i', x_j') \quad i, j = 1,2,\ldots,n$$

Due individui sono "vicini" quando la loro dissimilarità o distanza è piccola o, equivalentemente, quando la loro similarità è grande.

Generalmente nel caso in cui le variabili considerate siano qualitative gli indici di prossimità utilizzati sono gli indici di similarità, se invece i caratteri sono di tipo quantitativo verranno utilizzati gli indici di dissimilarità e le distanze.
Esistono infine indici di prossimità che vengono utilizzati nel caso in cui le variabili siano miste, ovvero alcune qualitative e altre di tipo quantitativo.
Di seguito illustreremo molte delle possibili misure di prossimità:

- misure di prossimità per le variabili a categorie (discrete)
- misure di prossimità per le variabili continue
- misure di prossimità su insiemi di dati contenenti sia variabili discrete che continue (insiemi misti).

3.3 Misure di similarità per dati binari.

La più comune tipologia di dati a categorie è quella in cui tutte le variabili sono binarie. Per questo motivo un gran numero di misure di similarità sono state proposte per dati binari. Tutte le misure sono definite in termini della computazione delle concordanze e discordanze delle *p* variabili per due individui. La tabella 3.3 evidenzia che:

- gli individui *i* e *j* assumono lo stesso valore 1 su *a* variabili;
- gli individui *i* e *j* assumono lo stesso valore 0 su *d* variabili;
- su *b* variabili l'individuo *i* assume il valore 0, mentre l'individuo *j* assume il valore 1;
- su *c* variabili l'individuo *i* assume il valore 1, mentre l'individuo *j* assume il valore 0.

	Individuo *i*			
Individuo j	Esito	1	0	Totale
	1	*a*	*b*	*a* + *b*
	0	*c*	*d*	*c* + *d*
	Totale	*a* + *c*	*b* + *d*	*p* = *a* + *b* + *c* + *d*

Tabella 3.3 – Tabella di contingenza, computazione dell'esito binario per due individui.

Gli indici di similarità sono definiti con riferimento agli elementi di un insieme (unità statistiche), anziché ai corrispondenti vettori riga, e assumono valori nell'intervallo chiuso [0, 1], anziché un qualunque valore non negativo (come accade invece a una distanza).

Due individui, *i* e *j*, hanno un coefficiente di similarità s_{ij} pari a 1 se possiedono valori identici per tutte le variabili. Un valore di similarità pari a 0 indica viceversa che i due individui differiscono il più possibile nel valore di tutte le variabili.

La tabella 3.4 elenca alcune delle misure di similarità che sono state proposte per dati binari. La ragione per una tale moltitudine di possibili misure è da attribuirsi alla apparente incertezza di come debbano essere trattate le corrispondenze (0,0) o assenza-assenza (*d* nella tabella 3.3). In alcuni casi le corrispondenze (0,0) sono del tutto equivalenti alle corrispondenze (1,1) e debbono essere incluse nel calcolo della misura di similarità. Un esempio è il genere (maschile, femminile) in cui non esiste alcuna preferenza nella scelta di quale delle due categorie debba essere codificata con zero o uno.

In altri casi, tuttavia, l'inclusione o meno delle *d* corrispondenze (0,0) è molto problematica, come quando il valore zero corrisponde all'effettiva assenza di qualche proprietà (esempio: assenza di ali in uno studio sugli insetti). Infatti la domanda che ci dobbiamo porre è se la co-assenza contenga informazioni utili sulla similarità di due oggetti: attribuire un alto grado di similarità ad una coppia di individui semplicemente perché a entrambi manca un certo numero di attributi

può non essere significativo in molte situazioni. In questo caso conviene utilizzare le misure, tra quelle riportate in tabella 2.3, che ignorano il conteggio della co-assenza *d*, per esempio S2 o S4.

Quando invece alla co-assenza di un fattore può essere associato un contenuto informativo si utilizza usualmente il coefficiente di corrispondenza S1. Le misure S3 ed S5 sono esempi di coefficienti simmetrici che prendono in considerazione anche le corrispondenze negative, pur assegnando *pesi* diversi nei due casi.

Non esistono regole veloci e rigide per stabilire se le corrispondenze negative debbano essere incluse o meno: la decisione spetta all'investigatore dei dati, che deve effettuare una scelta affidandosi al proprio livello di esperienza e familiarità con il materiale trattato (Sokal and Sneath, 1963).

La scelta della misura di similarità è molto importante, dal momento che l'utilizzo di diversi coefficienti di similarità può condurre a risultati diversi.

	Misura	Formula
S1	Coefficiente di Corrispondenza di Sokal e Michener	$_{SM}S_{ij} = \frac{a+d}{a+b+c+d}$
S2	Coefficiente di Jaccard	$_{J}S_{ij} = \frac{a}{a+b+c}$
S3	Coefficiente di Rogers e Tanimoto	$_{RT}S_{ij} = \frac{a+d}{a+2(b+c)+d}$
S4	Coefficiente di Sokal e Sneath	$_{SS}S_{ij} = \frac{a}{a+2(b+c)}$
S5	Coefficiente di Gower e Legendre	$_{GL}S_{ij} = \frac{a+d}{a+\frac{1}{2}(b+c)+d}$
S6	Coefficiente di Dice	$_{D}S_{ij} = \frac{2a}{2a+b+c}$
S7	Coefficiente di Russel e Rao	$_{RR}S_{ij} = \frac{a}{a+b+c+d}$

Tabella 3.4 – Misure di similarità per dati binari.

Si può dimostrare che alcuni coefficienti possono condurre ad uno stesso ordinamento: Gower e Legendre, (1986) evidenziano che S2, S4, S6 sono tra loro in relazione di monotonicità, così come S1, S3, S5, S7, ma S1 ed S2 possono condurre a diverse valutazioni delle similarità relative di un insieme di oggetti.

3.4 Misure di similarità per dati categorici non binari.

Dati a categorie in cui le variabili hanno più di due possibili livelli (esempio: il colore degli occhi) potrebbero essere trattate in modo analogo ai dati binari, considerando ogni livello di ciascuna variabile come una singola variabile binaria. Naturalmente questo approccio non è conveniente, a causa del grande numero di corrispondenze negative che verrebbero inevitabilmente generate. Un metodo migliore è il seguente: considerati due individui *i* e *j* a *p* dimensioni e la variabile k-esima ($k \in [1, p]$), si pone $s_{ijk} = 1$ se i due

individui assumono lo stesso valore per la variabile *k*, in caso contrario si pone $s_{ijk} = 0$. Il valore di similarità tra l'individuo *i* e *j* è poi semplicemente calcolato effettuando una media su tutte le *p* variabili:

$$s_{ij} = \frac{1}{p} \sum_{k=1}^{p} s_{ijk}$$

3.5 Dissimilarità e misure di distanza per dati continui.

Quando tutte le variabili sono continue, le prossimità tra gli individui sono calcolate utilizzando misure di dissimilarità che sono anche misure di distanza.
Un valore di similarità s_{ij} può essere facilmente convertito in un valore di dissimilarità d_{ij} dato da $d_{ij} = 1 - s_{ij}$.

Osserviamo che il complemento a uno di un indice di similarità è detto indice di dissimilarità e rappresenta una classe di indici di prossimità più ampia delle distanze, che devono soddisfare anche la disuguaglianza triangolare.

La **distanza tra due punti** corrispondenti ai vettori riga i,j $\in \Re^p$ è una funzione d_{ij} che gode delle seguenti proprietà:

non negatività:	$d_{ij} \geq 0$		$\forall x, y \in \Re^p$
identità:	$d_{ii} = 0$	$\Leftrightarrow$	$x = y$
simmetria:	$d_{ij} = d_{ji}$		$\forall x, y \in \Re^p$
disuguaglianza triangolare:	$d_{ij} \leq d_{ik} + d_{kj}$		$\forall x, y, z \in \Re^p$

Per il raggruppamento delle unità statistiche, generalmente si considera la distanza tra tutte le unità statistiche presenti nella matrice dei dati. L'insieme di tali distanze viene rappresentato in una matrice delle distanze.
Per tutte le coppie di individui (i, j), (i,m) e (m, j). Una matrice $n \times n$ di dissimilarità

$$D = \left\{ d_{ij} \right\} \quad d_{ij} = 0 \quad \forall i$$

è detta *metrica* se la disuguaglianza triangolare vale per tutte le triplette (i, j,m).
Dalla disuguaglianza triangolare segue anche che la matrice D è simmetrica:

$$d_{ij} = d_{ji} \quad \forall i, j$$

Una generica **matrice delle distanze** è strutturata nel modo seguente:

$$D = \begin{pmatrix} 0 & \dots & \dots & d_{1j} & \dots & \dots & d_{1n} \\ \dots & 0 & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & 0 & \dots & \dots & \dots & \dots \\ d_{i1} & \dots & \dots & 0 & \dots & \dots & d_{in} \\ \dots & \dots & \dots & \dots & 0 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & 0 & \dots \\ d_{n1} & \dots & \dots & d_{nj} & \dots & \dots & 0 \end{pmatrix}$$

dove il generico elemento d_{ij} è la misura della distanza tra le entità *i* e *j*.

	Misura	Formula
D1	Distanza Euclidea	$d_{ij} = \left[\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \right]^{\frac{1}{2}}$
D2	Distanza City Block (o Rettilinea o di Manhattan)	$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|$
D3	Distanza di Minkowski	$d_{ij} = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^r \right)^{1/r} ;\ r \geq 1$
D4	Distanza di Camberra	$d_{ij} = \begin{cases} 0 & ;x_{ij=} x_{jk} = 0 \\ \sum_{k=1}^{p} \frac{\left| x_{ik} - x_{jk} \right|}{\left| x_{ik} \right| + \left| x_{jk} \right|} & ;altrimenti \end{cases}$

Tabella 3.5 – Distanze per dati continui

Da un insieme di osservazioni multivariate si può ricavare una matrice di dissimilarità utilizzando una delle misure riportate in Tabella 3.5. Tutti i tipi di distanza possono essere pesate in modo non uniforme. Ad esempio la distanza Euclidea pesata da $w_1, w_2, ..., w_p$ risulta essere:

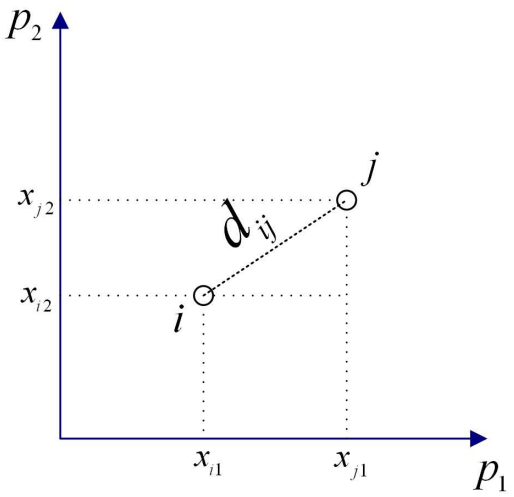
$$d_{ij} = \sqrt{w_1 \left| x_{i1} - x_{j1} \right|^2 + w_2 \left| x_{i2} - x_{j2} \right|^2 + ... + w_p \left| x_{ip} - x_{jp} \right|^2}$$

In altri termini:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} w_k \left(x_{ik} - x_{jk} \right)^2}$$

La distanza D1 (distanza Euclidea) ha la piacevole proprietà che d_{ij} può essere interpretata come la distanza tra due punti $x_i = \left(x_{i1}, ..., x_{ip} \right)$ e $x_j = \left(x_{j1}, ..., x_{jp} \right)$ in uno spazio a *p* dimensioni.

La figura 3.6 presenta l'esempio nel caso in cui *p* sia uguale a 2 nel quale è facile intuire che la distanza euclidea è a tutti gli effetti una distanza fisica.



Distanza euclidea, $p = 2$

$$d_{ij} = \left[\left(x_{i1} - x_{j1} \right)^2 + \left(x_{i2} - x_{j2} \right)^2 \right]^{\frac{1}{2}}$$

Figura 3.6 – Distanza euclidea per *p*=2.

Le distanze D1 e D2 riportate nella tabella 2.4 sono un caso particolare della distanza D3, con $r = 2$ ed $r = 1$ rispettivamente.

La distanza di Camberra è molto sensibile a piccole variazioni intorno a $x_{ik} = x_{jk} = 0$.

Una proprietà desiderabile delle matrici di dissimilarità è che siano Euclidee:

una matrice di dissimilarità quadrata $D_{(n \times n)} = \{ d_{ij} \}$ è detta Euclidea se presi due individui qualunque *i* e *j* tra gli *n* individui, questi sono posti nello spazio *p*-dimensionale a distanza Euclidea pari a d_{ij}. La proprietà Euclidea è utile perché, allo stesso modo della misura di distanza Euclidea, consente di interpretare le dissimilarità come distanze fisiche. Se una matrice è Euclidea allora è anche metrica, il viceversa non è vero.

Tra le misure di distanza elencate solo D1 produce matrici di dissimilarità Euclidee.

3.6 Misure di similarità per dati misti.

Sono state proposte varie misure di similarità per dati di tipo misto, che contengono cioè variabili sia di tipo continuo che di tipo categorico. Noi ci concentriamo soltanto su quella introdotta da Gower[3].
Gower ha proposto nel 1971 la seguente misura di similarità per dati misti:

$$s_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} s_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$$

dove s_{ijk} è la similarità tra gli individui *i* e *j*, misurata sulla k-esima variabile, e w_{ijk} è il relativo peso.

Abbiamo che:

$$w_{ijk} = \begin{cases} 0\ ; & \text{se il valore della k-esima variabile è mancante per uno od entrambi gli individui} \\ 0\ ; & \text{se, in caso di variabili binarie, si vogliano escludere corrispondenze negative} \\ 1\ ; & \text{altrimenti} \end{cases}$$

Il valore di s_{ijk} è valutato in modo diverso a seconda della natura delle variabili. Per variabili binarie o categoriche con più di due possibili valori, $s_{ijk} = 1$ se i due individui hanno lo stesso valore della variabile *k*, in caso contrario $s_{ijk} = 0$.

[3] per avere informazioni sulle altre tipologie di misura: Everitt S., Landau S., *Cluster Analysis*. Oxford University Press, fourth edition, 2001

Per variabili continue abbiamo:

$$s_{ijk} = 1 - \frac{\left| x_{ik} - x_{jk} \right|}{R_k}$$

dove R_k è il *range* della k-esima variabile (in pratica si utilizza la distanza *Manhattan* dopo aver scalato la k-esima variabile all'unità).

3.7 Clustering gerarchico.

Quando si effettua una classificazione di tipo gerarchico i dati non risultano partizionati in un certo numero di *cluster* in un passo solo, bensì sono previste una serie di fasi successive. Il punto di partenza della classificazione può essere un singolo *cluster* contenente tutti gli individui oppure *n cluster* contenenti ciascuno un solo individuo. Nel primo caso l'algoritmo di *clustering* gerarchico procede per partizioni successive, nel secondo per fusioni.
Nella figura 3.7 possiamo vedere come si rappresenta graficamente le varie fasi di *clustering* attraverso un diagramma bidimensionale detto dendrogramma che illustra le fusioni/divisioni verificatesi ad ogni stadio dell'analisi.
Per semplicità si supponga che esistano solamente 5 unità statistiche a disposizione.

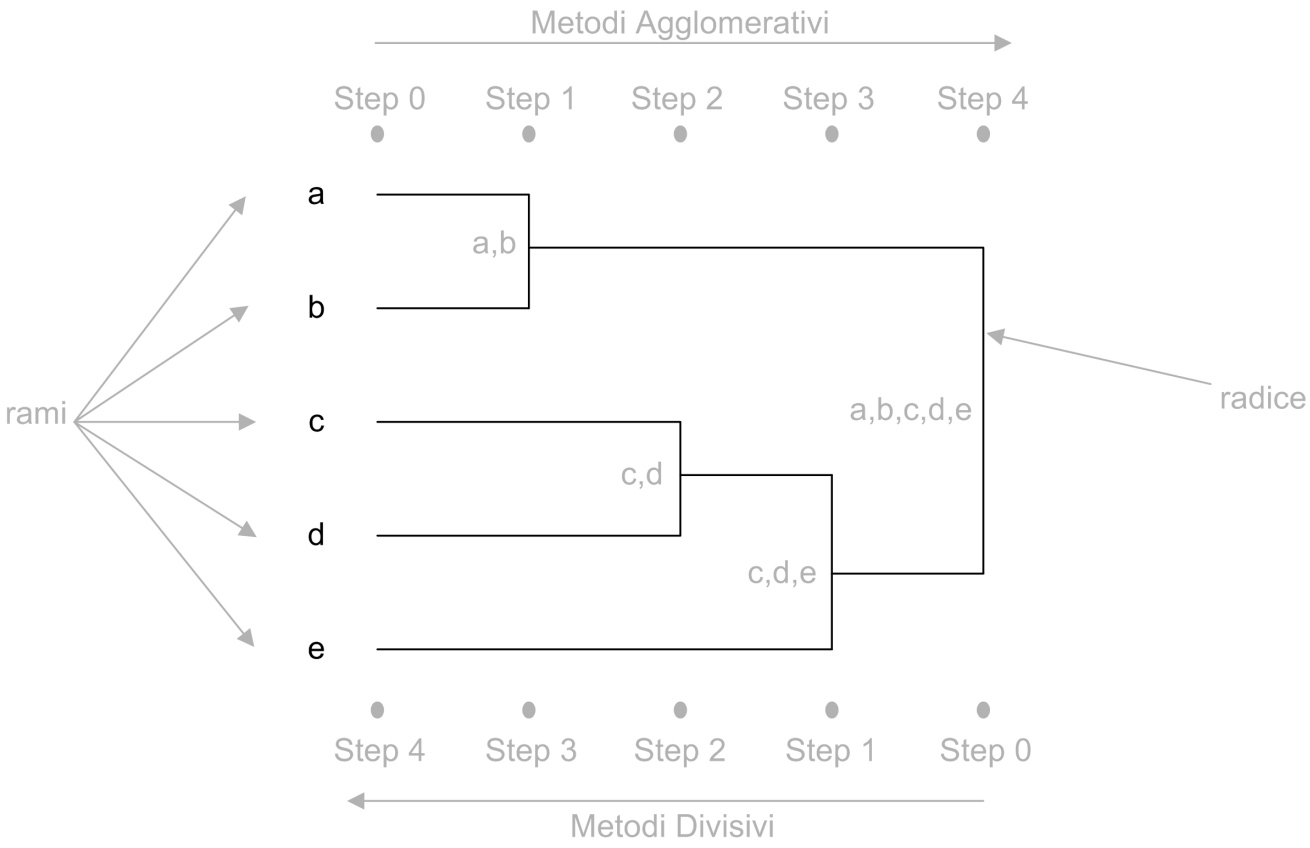


Figura 3.7 – Struttura del dendrogramma.

I rami dell'albero descrivono classificazioni successive delle unità statistiche. Alla radice dell'albero, tutte le unità statistiche sono contenute in una sola classe. Le successive divisioni in rami individuano divisioni successive delle unità in *cluster*. Infine, i rami terminali indicano la partizione finale delle unità statistiche. Se la formazione dei gruppi avviene dai rami alla radice (nella Figura 3.7, da sinistra verso destra), vale a dire, se si parte dalla situazione in cui ogni unità statistica appartiene a un gruppo a sé stante e si procede a un raggruppamento di tali unità, i metodi di classificazione gerarchica vengono detti **agglomerativi**. Invece, se la costruzione dei *cluster* avviene dalla radice ai rami dell'albero i corrispondenti metodi gerarchici vengono detti **scissori o divisivi**. Questo modo di procedere ha il vantaggio di ridurre il numero di partizioni da confrontare, rendendo la procedura computazionalmente più efficiente, ma anche lo svantaggio di non poter "correggere" errori di classificazione se commessi nei passi precedenti.

E’ chiaro quindi che una prima suddivisione dei metodi di *clustering* gerarchico risulta essere la seguente:

- **metodi agglomerativi** : prevedono un *input* costituito da un insieme di *n cluster*, ciascuno contenente inizialmente un solo individuo, ed effettuano una serie di successive fusioni, che aumentano la dimensione dei *cluster* e ne riducono il numero fino al valore desiderato (strategia *bottom-up*)
- **metodi divisivi** : prevedono un *input* costituito da un unico *cluster*, contenente inizialmente tutti gli individui, ed effettuano una serie di successive partizioni, che diminuiscono la dimensione dei *cluster* e ne aumentano il numero fino al valore desiderato (strategia *top-down*).

Lo scopo di entrambe le varianti è il medesimo: arrivare, tramite suddivisioni o sintesi successive, a una partizione ottimale, operando su una matrice di prossimità di qualche tipo. Le successive partizioni individuate da un dendrogramma sono "nidificate". Ciò significa che, nei metodi gerarchici, gli elementi che vengono uniti (o divisi) a un certo passo resteranno uniti (divisi) fino alla fine del processo di classificazione. Infatti, la caratteristica principale degli algoritmi gerarchici è che le divisioni o fusioni sono irrevocabili: i metodi agglomerativi consentono solo fusioni, quelli divisivi solo partizioni e non sono ammessi algoritmi gerarchici di tipo misto; quindi se due individui sono stati assegnati a due *cluster* diversi non potranno in seguito essere nuovamente membri di uno stesso *cluster*. L’obiettivo dichiarato è di ottenere gruppi esclusivi ed omogenei che espresso in altri termini implica che le intersezioni tra gruppi diversi danno come risultato l’insieme vuoto.

Poiché le tecniche gerarchiche agglomerative proseguono finché i dati non siano racchiusi in un unico *cluster* contenente tutti gli individui, mentre le tecniche divisive suddividono l’insieme dei dati fino ad avere *n* gruppi, ognuno dei quali contenente un singolo individuo, l’investigatore deve possedere un criterio che induca la terminazione dell’algoritmo nel momento in cui si sia raggiunto il numero ottimale di *cluster*.

Le tecniche di classificazione gerarchiche sono molto usate in marketing, biologia, sociologia, medicina e in tutti quei settori in cui è implicita una struttura gerarchica nei dati. Sebbene il *clustering* gerarchico possa rivelarsi molto utile anche dove non sia presente una sottostante struttura gerarchica, allo scopo di affrontare problematiche prestazionali, evidenziamo che i metodi gerarchici non devono essere usati se non sono chiaramente necessari.

A questo punto possiamo descrivere, con riferimento ai metodi agglomerativi, la procedura statistica per ottenere un dendrogramma. Risulta opportuno schematizzare la procedura nelle fasi riassunte di seguito.

1. **Inizializzazione**: date *n* unità statistiche da classificare, ogni elemento rappresenta un gruppo (si hanno, in altri termini, *n cluster*). I *cluster* verranno indicati con un numero che va da 1 a *n*.
2. **Selezione**: vengono selezionati i due *cluster* più "vicini" rispetto alla misura di prossimità fissata inizialmente. Per esempio, rispetto alla distanza euclidea.
3. **Aggiornamento**: si aggiorna il numero dei *cluster* (che sarà pari a *n* - 1) attraverso l'unione, in un unico *cluster*, dei due gruppi selezionati nel punto precedente. Conseguentemente, si aggiorna la matrice delle distanze, sostituendo, alle due righe (colonne) di distanze relative ai due *cluster*, nei confronti di tutti gli altri, una sola riga di distanze, "rappresentativa" del nuovo gruppo. I metodi agglomerativi differiscono per il modo in cui viene definita tale rappresentatività.
4. **Ripetizione**: si eseguono i passi (2) e (3) per (*n* – 1) volte.
5. **Arresto**: la procedura si arresta quando tutti gli elementi vengono incorporati in un unico *cluster*.

In base ai diversi modi in cui vengono calcolate le distanze fra il gruppo neo-formato e le altre unità statistiche, si distinguono diversi metodi gerarchici di classificazione.

Introduciamo i diversi metodi considerando la distanza fra gruppi.

Per prima cosa bisogna distinguere fra i metodi che richiedono esclusivamente, come *input*, la matrice di distanza, e i metodi che richiedono anche la matrice dei dati.
Cominciamo con il primo tipo.

3.7.1 Metodo del *legame singolo (single linkage):*

Il primo metodo è il più semplice tra i metodi di clustering gerarchico ed è detto metodo del *legame singolo (single linkage)* o del *confinante più vicino (nearest-neighbour technique).*
L'assunto base di questa tecnica è identificare la distanza (o similarità) fra due gruppi con quella fra i loro membri più vicini (o più simili).
Il grado di vicinanza fra due gruppi è stabilito prendendo in considerazione solo le informazioni relative a due oggetti più vicini, ignorando quelle che si riferiscono a tutti gli altri oggetti appartenenti ai gruppi.
La distanza tra due gruppi A e B è definita come la distanza minore rilevata tra la coppia di individui *(i,j)* con $i \in A, j \in B$, in altri termini si considera il minimo delle $n_A \times n_B$ distanze tra ciascuna delle unità del gruppo A e ciascuna delle unità del gruppo B .

$$d_{AB} = \min_{i \in A,\, j \in B} d_{ij}$$

La tecnica del legame singolo gode di diverse importanti proprietà matematiche. Si può dimostrare infatti che la sequenza di partizioni che si ottengono è invariante rispetto a trasformazioni monotoniche delle variabili (Jardine e Sibson).
Questa tecnica è quindi una delle poche che risultano insensibili a trasformazioni delle variabili che conservano l'ordine dei valori nella matrice di similarità.

Anche se gode di importanti proprietà matematiche la tecnica del legame singolo si rivela di regola di scarsa utilità. Essa ha infatti la tendenza a concatenare quasi tutti i casi in un unico grande gruppo; si mantengono separati solo piccoli gruppi o casi isolati.

3.7.2 Metodo del *legame completo (complete linkage):*

L'assunto base del metodo del *legame completo*, chiamato anche del *confinante più lontano* (*farthest neighbor technique)*, è opposto a quella della tecnica del legame singolo: la distanza/similarità fra due gruppi è identificata con quella fra i membri più lontani (o meno simili).
La distanza tra due gruppi A e B è definita come la distanza maggiore rilevata tra la coppia di individui *(i,j)* con $i \in A, j \in B$, in altri termini si considera il massimo delle $n_A \times n_B$ distanze tra ciascuna delle unità di un gruppo e ciascuna delle unità dell'altro gruppo.

$$d_{AB} = \max_{i \in A,\, j \in B} d_{ij}$$

La tecnica manifesta la tendenza a identificare gruppi relativamente compatti, che risultano – nello spazio multidimensionale delle variabili – di forma ipersferica, composti da oggetti fortemente omogenei rispetto alle variabili impiegate.

3.7.3 Metodo del *legame medio (average linkage):*

Con questa tecnica per determinare la distanza fra due gruppi A e B si prende in considerazione tutte le distanze fra gli n_A oggetti membri del primo rispetto a tutti gli n_B oggetti membri del secondo.

Con la tecnica del *legame medio* la distanza fra due gruppi si computa in base alla media aritmetica di tali distanze (Sokal e Michener,1958; McQuitty,1964).

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d_{ij}$$

Il secondo tipo di metodi richiede anche la matrice dei dati. Illustriamo in seguito i principali.

3.7.4 Metodo del *Centroide:*

La tecnica del *centroide* fa riferimento ad una rappresentazione spaziale degli oggetti da classificare, infatti, per ogni gruppo si definisce *centroide* il punto nello spazio multidimensionale che ha come coordinate la media aritmetica di tutti gli oggetti appartenenti al gruppo. La distanza fra i gruppi è in questo caso identificata dalla distanza fra i rispettivi centroidi.

La distanza tra due gruppi A e B di numerosità n_A e n_B è definita come la distanza tra i rispettivi centroidi (medie aritmetiche), $\overline{x}_A$ e $\overline{x}_B$.

$$d_{AB} = d\left(\overline{x}_A, \overline{x}_B\right)$$

Dopo la fusione dei gruppi A e B il centroide del nuovo gruppo formato AB è dato da:

$$\overline{x}_{AB} = \frac{n_A \overline{x}_A + n_B \overline{x}_B}{n_A + n_B}$$

II metodo del *centroide* e il metodo del *legame medio* presentano delle interessanti analogie da considerare: il metodo del *legame medio* considera la media delle distanze tra le unità di ciascuno dei due gruppi, mentre il metodo del *centroide* calcola le medie di ciascun gruppo, e in seguito misura le distanze tra di esse.

3.7.5 Metodo di *Ward:*

La tecnica di Ward si propone di realizzare una classificazione gerarchica minimizzando la varianza delle variabili entro ciascun gruppo. Ad ogni stadio, vengono pertanto fusi i gruppi che producono il minimo aumento della varianza totale entro i gruppi (Ward, 1963). Questa tecnica permette di generare gruppi di dimensioni relativamente equivalenti e di forma tendenzialmente sferica.
Il metodo di *Ward* minimizza, nella scelta dei gruppi da aggregare, una funzione obiettivo che parte dal presupposto che una classificazione ha l'obiettivo di creare gruppi che abbiano la massima coesione interna e la massima separazione esterna.

La *devianza totale* delle *p* variabili viene scomposta in *devianza nei gruppi* e *devianza fra i gruppi*:

$$Dev_{\text{Totale}} = Dev_{\text{Entro}} + Dev_{\text{Tra}}$$

Formalmente, data una partizione di *G* gruppi di numerosità vaiabile n_g con *(g=1,2,…,G)*:

- la devianza totale delle *p* variabili corrisponde alla somma delle devianze delle singole variabili rispetto alla corrispondente media generale $\overline{x}_k$:

$$Dev_{\text{Totale}} = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\sum_{k=1}^{p}\left(x_{ikg} - \overline{x}_k\right)^2 = \sum_{s=1}^{n}\sum_{k=1}^{p}\left(x_{sk} - \overline{x}_k\right)^2$$

- la devianza nei gruppi è data dalla somma delle devianze di ciascun gruppo:

$$Dev_{\text{Entro}} = \sum_{g=1}^{G}\sum_{i=1}^{n_g}\sum_{k=1}^{p}\left(x_{ikg} - \overline{x}_{kg}\right)^2 = \sum_{s=1}^{n}\sum_{k=1}^{p}\left(x_{sk} - \overline{x}_{kg}\right)^2$$

- la devianza fra i gruppi è data dalla somma delle devianze (ponderate) delle medie di gruppo rispetto alla corrispondente media generale:

$$Dev_{\text{Tra}} = \sum_{g=1}^{G}\sum_{k=1}^{p} n_g\left(\overline{x}_{kg} - \overline{x}_k\right)^2$$

Nel metodo di Ward, ad ogni passo della procedura gerarchica si aggregano tra loro i gruppi che comportano il minor incremento della devianza nei gruppi Dev_{Entro}, e maggior incremento di Dev_{Tra}, per ottenere la maggiore coesione interna possibile, quindi di conseguenza, la maggior separazione esterna.

3.8 Algoritmo generale per le tecniche gerarchico-agglomerative.

Lance e Williams (1967) hanno dimostrato che tutte le tecniche di classificazione gerarchico-agglomerative che abbiamo presentato possono essere considerate come varianti di un’unica procedura aggregativa, che si può esprimere in forma compatta e ricorsiva.

Tale procedura può essere definita formalmente nei termini seguenti:

a. si parte da una partizione in *n* gruppi formati da un sol oggetto;
b. si uniscono i due gruppi *i* e *j* che minimizzano la misura di dissimilarità d_{ij};
c. si ripete il passo b) finché tutti gli oggetti non formano un solo gruppo.

La misura di dissimilarità fra gruppi può essere calcolata ricorsivamente.
Allo stadio a) della procedura di aggregazione, le dissimilarità fra gli *n* gruppi coincidono ovviamente con le dissimilarità fra gli *n* oggetti.
Nei successivi passi della procedura, la misura di dissimilarità fra il grupo *k* e il gruppo *(IJ)* derivante dalla fusione dei gruppi *i* e *j* si calcola sulla base della seguente espressione:

$$d_{k,ij} = \alpha(i)\, d_{ki} + \alpha(j)\, d_{kj} + \beta d_{ij} + \Gamma \left| d_{ki} - d_{kj} \right|$$

i parametri $\alpha(i)$, $\alpha(j)$, β, Γ si possono determinare, a seconda della tecnica preferita, in base ai valori riportati nella tabella 3.8.

Tecnica	$\alpha(i)$	$\alpha(j)$	β	Γ
Legame singolo	1/2	1/2	0	–1/2
Legame completo	1/2	1/2	0	1/2
Legame medio	$n_i / (n_i + n_j)$	$n_j / (n_i + n_j)$	0	0
Ward	$\frac{(n_i + n_k)}{(n_k + n_i + n_j)}$	$\frac{(n_k + n_j)}{(n_k + n_i + n_j)}$	$\frac{-n_k}{(n_k + n_i + n_j)}$	0

Tabella 3.8 – Parametri per il calcolo delle misure di dissimilarità.

4.

Applicazione pratica della Cluster Analysis.

4.1 La segmentazione della clientela con la Cluster Analysis.

In questa fase del lavoro viene presentata l'applicazione della *Cluster Analysis* nel caso specifico di Elettroingross.

Come è stato precisato più volte l'obiettivo è di segmentare i clienti in base al comportamento d'acquisto e individuare profili di clienti da trattare commercialmente in modo diverso.
Le variabili discriminanti utilizzate rimangono: *indice1, indice2, indice3, indice4* e l'indice frequenza acquisti I_f.
Sono state testate diverse misure di distanza e diversi algoritmi gerarchici ma parleremo solamente della combinazione che ha generato i risultati migliori: la distanza euclidea e l'algoritmo di Ward.
Come sappiamo la *Cluster Analysis* non ha bisogno di nessuna assunzione a priori e non si potrebbe fare altrimenti perché la scelta di utilizzare la *Cluster analysis* avviene quando non ci sono sufficienti indicazioni preliminari.

Ad esempio non si conosce esattamente in quanti gruppi conviene suddividere la popolazione oggetto di studio, e per ottenere con precisione questa importante informazione si procede in questo modo:

- si avvia un processo di *clustering* con l'obiettivo di ottenere un singolo gruppo;
- si analizza la scheda di aggregazione che mostra la distanza che intercorre tra i gruppi formati in successione nei vari stadi del processo di *clustering*.

Il dendrogramma del processo di *clustering* per un solo gruppo è il seguente:

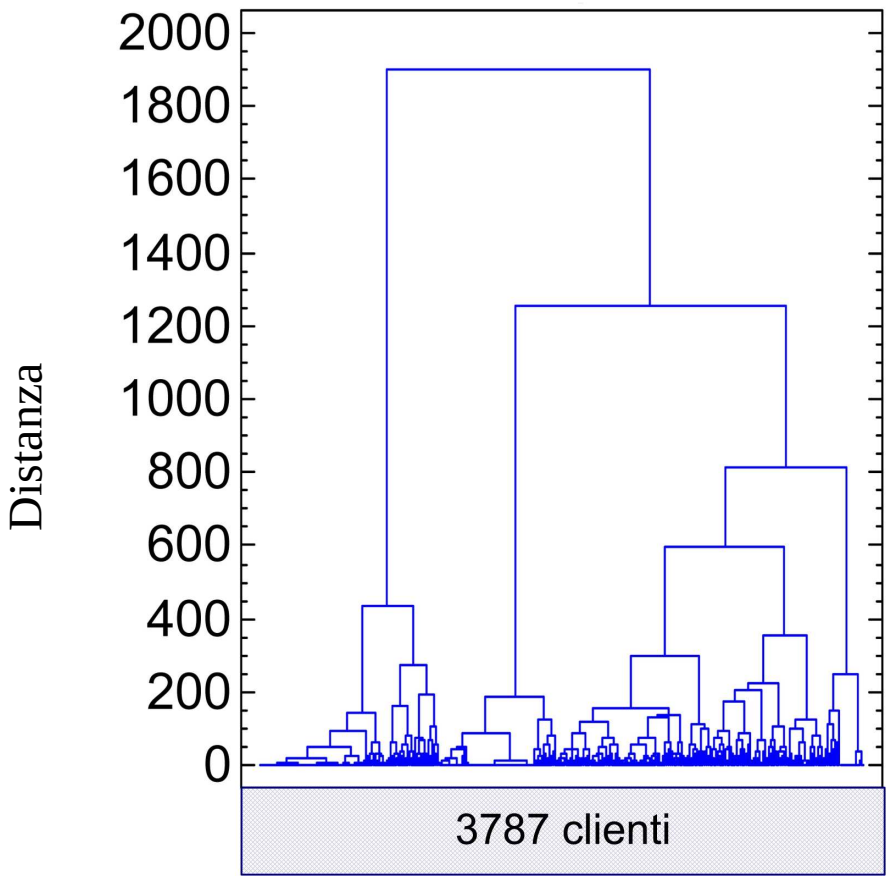


Figura 4.1 – Dendrogramma ottenuto con l'algoritmo di Ward.

Scheda di aggregazione						
Passo	**Gruppi che si uniscono**		**Distanza tra i gruppi**	**Passo nel quale il gruppo appare per la prima volta**		**Passo successivo nel quale appare**
	Cluster 1	**Cluster 2**		**Cluster 1**	**Cluster 2**	
1	3.780	3.786	0,00	0	0	6
2	3.781	3.785	0,00	0	0	5
3	3.773	3.784	0,00	0	0	9
4	3.585	3.782	0,00	0	0	54
5	64	3.781	0,00	0	2	32
6	61	3.780	0,00	0	1	19
7	3.767	3.777	0,00	0	0	11
8	3.701	3.775	0,00	0	0	19
9	139	3.773	0,00	0	3	25
10	3.732	3.771	0,00	0	0	14
11	9	3.767	0,00	0	7	40
12	3.697	3.759	0,00	0	0	20
13	3.607	3.755	0,00	0	0	39
14	40	3.732	0,00	0	10	17
15	3.717	3.729	0,00	0	0	17
16	3.683	3.725	0,00	0	0	22
17	40	3.717	0,00	14	15	31
18	3.145	3.705	0,00	0	0	330
19	61	3.701	0,00	6	8	273
20	1	3.697	0,00	0	12	189
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
3768	48	101	128,60	3.734	3.735	3.773
3769	4	177	133,74	3.763	3.721	3.774
3770	3	117	139,28	3.756	3.725	3.776
3771	8	23	144,83	3.744	3.764	3.777
3772	5	15	152,48	3.760	3.765	3.779
3773	78	83	161,12	3.751	3.755	3.779
3774	2	48	171,03	3.748	3.767	3.785
3775	4	24	182,01	3.768	3.761	3.778
3776	31	139	193,53	3.766	3.701	3.783
3777	3	86	205,35	3.769	3.752	3.781
3778	8	59	218,63	3.770	3.737	3.783
3779	4	19	235,55	3.774	3.758	3.780
3780	5	78	255,48	3.771	3.772	3.782
3781	4	27	281,40	3.778	3.762	3.782
3782	1	3	308,69	3.733	3.776	3.786
3783	4	5	396,83	3.780	3.779	3.784
3784	8	31	562,42	3.777	3.775	3.784
3785	4	8	794,66	3.782	3.783	3.785
3786	2	4	1.210,45	3.773	3.784	3.786
3787	1	2	1.899,91	3.781	3.785	0

Tabella 4.2 – Scheda di aggregazione ottenuta con 3787 elementi e l'algoritmo di Ward.

La scheda di aggregazione contiene molte informazioni che spiegheremo qui di seguito:

La prima colonna rappresenta il **passo** del processo di *clustering*; si parte con il numero 1 in cui ci sono *n* gruppi distinti composti da un solo elemento (nel nostro caso *n* è uguale 3787) e si arriva al passo numero *n* con 1 gruppo singolo contenente tutti gli elementi.
I gruppi che si uniscono sono rappresentati su due colonne e spiegano per ogni passo quali sono i gruppi che si fondono.
La distanza tra i gruppi è l'informazione cruciale che serve per poter quantificare il miglior numero di *cluster* da ottenere e nella scheda di aggregazione viene riportata la distanza tra i gruppi che si uniscono nelle varie fasi.
Per i due gruppi che si uniscono ad ogni passo si indica **in quale passo i gruppi sono comparsi** individualmente **per la prima volta** e l'ultima colonna indica in quale **passo successivo** comparirà il nuovo gruppo formato con la fusione dei due.

Utilizzando le ultime 20 righe della scheda di aggregazione è stata creata la tabella 4.3 nella quale andremo a considerare quale sia il numero più conveniente di *cluster* da formare.
Sono stati calcolati gli incrementi percentuali di distanza da 20 gruppi a 1 e bisogna individuare in quale passo si concentra un'improvviso incremento di distanza.

Numero cluster	Distanza tra gruppi	Variazione % della distanza rispetto al passo precedente
20	5,12	4,15
19	5,14	4
18	5,54	4,15
17	5,55	3,98
16	7,65	5,28
15	8,64	5,66
14	9,91	6,15
13	10,98	6,42
12	11,52	6,33
11	11,82	6,11
10	13,28	6,46
9	16,92	7,74
8	19,93	8,46
7	25,92	10,14
6 ◄	27,29	9,7
5	88,14	28,55
4	165,58	41,73
3	232,24	41,29
2	415,79	52,32
1	689,47	56,96

Tabella 4.3 – Distanze e variazioni percentuali delle distanze tra i gruppi che si uniscono per le soluzioni da 20 gruppi a 1 gruppo.

Per stabilire il miglior numero di gruppi da considerare si scorre la tabella 4.3 fino a quando non si registra un grande aumento della distanza: in quel caso il numero di gruppi del passo precedente è ottimale perché si utilizzano il numero di gruppi la cui formazione ha comportato la minor distanza di fusione.

Come si può notare nel passare da 6 a 5 gruppi si verifica un incremento di distanza del 28,55% e di 88,14 in valore assoluto.

Il grande aumento di distanza indica che si andrebbero ad unire gruppi molto distanti fra loro: in altri termini il compromesso da accettare per formare 5 gruppi va a discapito dell'espressività che i gruppi devono rappresentare.

Questa valutazione ci permette di stabilire che il numero ottimale di gruppi è 6 indicato in tabella con il simbolo ◄.

Graficamente la distanza che intercorre tra il numero di gruppi è rappresentata qui di seguito in figura 4.4.

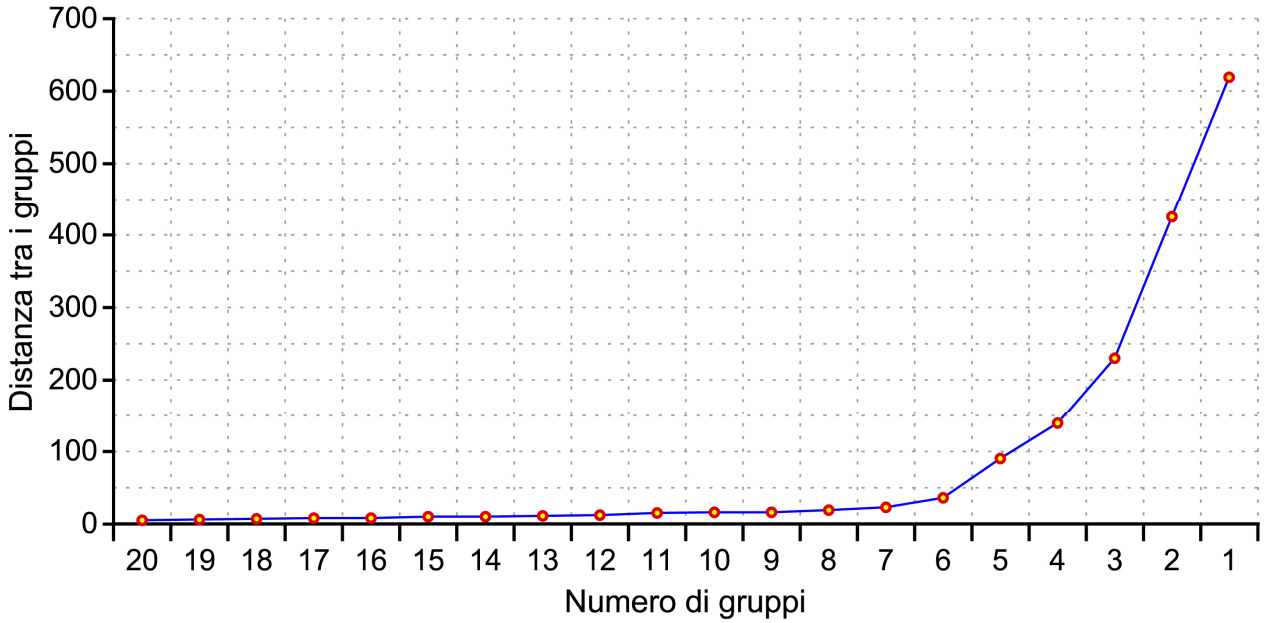


Figura 4.4 – Rappresentazione della distanza tra gruppi.

Graficamente è facile notare che fino a 6 gruppi si uniscono gruppi molto "vicini", mentre già al quinto si presenta uno sbalzo di distanza, al ricercatore giunge quindi la conferma che la popolazione è rappresentata molto efficientemente suddividendola in 6 gruppi che conferiscono la significatività oggetto d'indagine.

In via sperimentale sono state testate tutte le soluzioni da 6 a 2 gruppi ed effettivamente la migliore è stata la scelta di formare 6 gruppi: le altre soluzioni con un numero di gruppi inferiori generavano gruppi con un andamento di fatturato generico e molto similari tra un gruppo e l'altro; inoltre le distribuzioni degli indici erano "stirate" denunciando che non si stavano considerando gruppi con significatività e concentrazioni particolari.

A questo punto del lavoro è stata praticata la *cluster analysis* impostandola per l'ottenimento di 6 gruppi con i dati a disposizione utilizzando sempre la distanza euclidea e l'algoritmo di Ward.

Il dendrogramma risultante dalla composizione di 6 gruppi è il seguente:

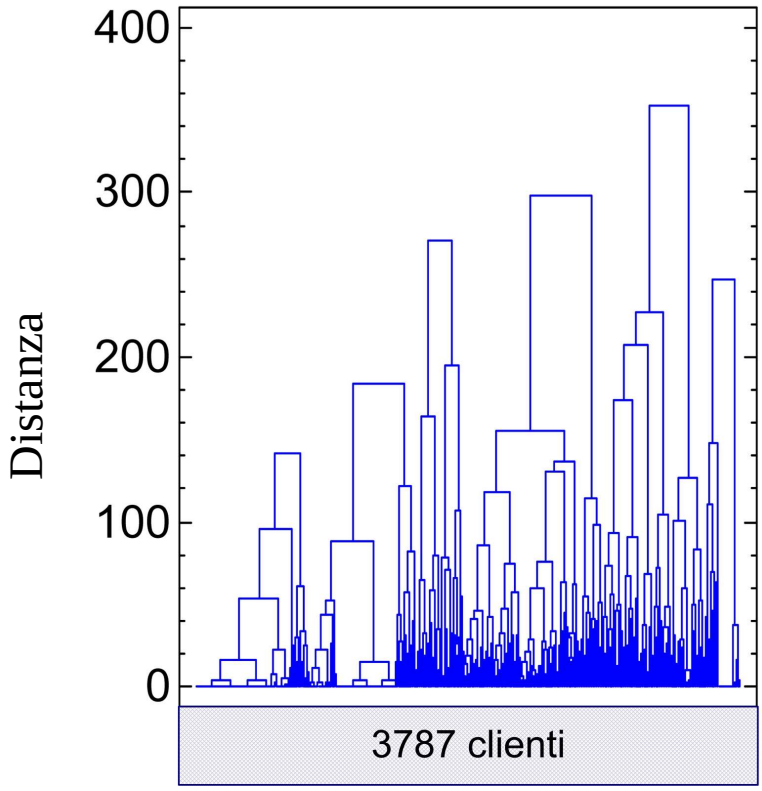


Figura 4.5 – Dendrogramma rappresentante 6 gruppi ottenuti con l'algoritmo di Ward.

Sono stati ottenuti i 6 cluster con:

- 792 clienti per il Cluster 1;
- 644 clienti per il Cluster 2;
- 1033 clienti per il Cluster 3;
- 536 clienti per il Cluster 4;
- 485 clienti per il Cluster 5;
- 297 clienti per il Cluster 6.

Nel quarto capitolo si spiegava che la *cluster analysis* impone molte decisioni al ricercatore e che ovviamente scelte diverse comportano risultati diversi; ora per delineare in che modo è avvenuto tutto il processo di clustering vengono riproposte e riassunte le scelte intraprese:

PRIMA	• Scelta delle variabili	*indice1, indice2, indice3, indice4*; I_f .
	• Criteri di similarità-distanza	Distanza euclidea
DURANTE	• Tecniche di aggregazione	Algoritmo di Ward
	• Numero dei gruppi da ottenere	6 gruppi
DOPO	• Valutazione della qualità della soluzione	Valutazione del fatturato e delle distribuzioni degli indici
	• Interpretazione dei risultati	Ricavare un giudizio commerciale dei gruppi ottenuti.

Tabella 4.6 – Scelte intraprese per svolgere la cluster analysis.

Di seguito sono presentati i gruppi ottenuti con una breve descrizione, fornendo per ognuno le valutazioni del fatturato e delle distribuzioni degli indici e ancora l' interpretazione commerciale che stabilisce se il gruppo comprende clienti in DIMINUZIONE, CRESCITA o COSTANTI.

Cluster 1

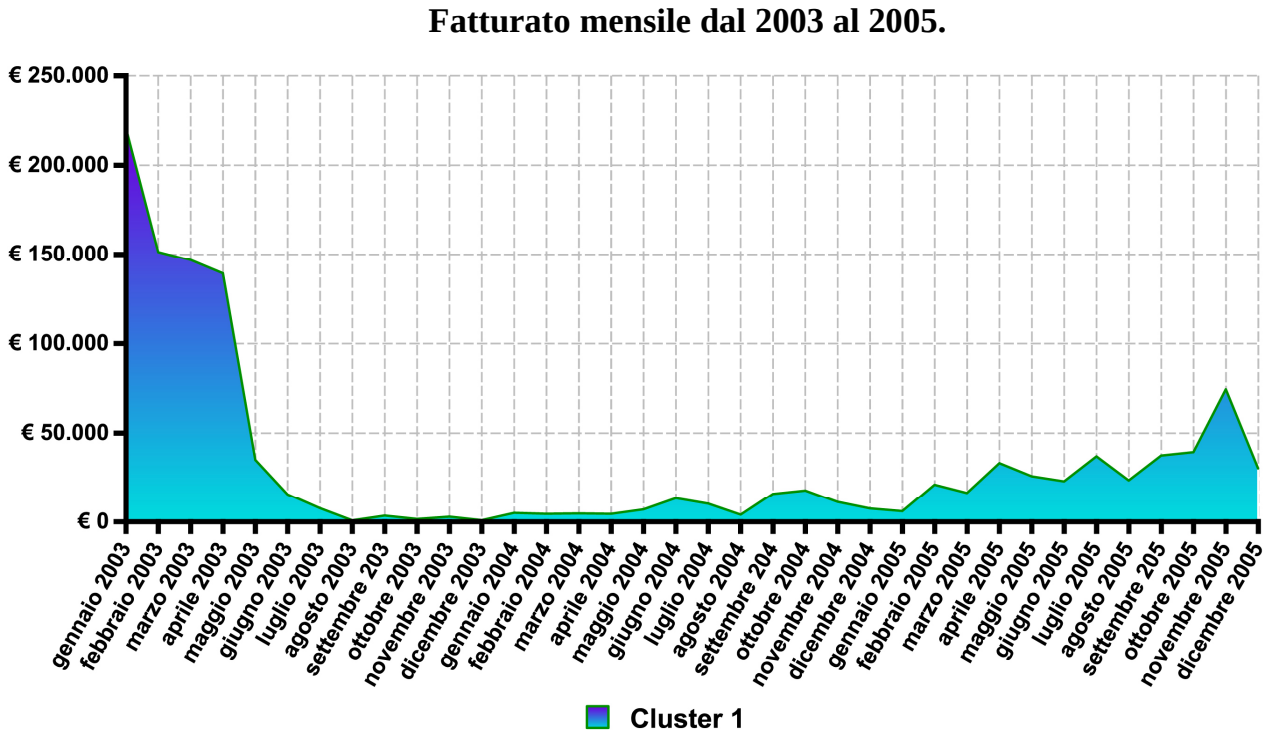


Figura 4.7a – Grafico fatturato dei clienti del Cluster 1.

Il primo *cluster* individuato evidenzia un notevole valore d'acquisto generato nel primo quadrimestre 2003 e successivamente un netto calo fino ad azzerarsi nell'ultimo quadrimestre 2003 e nel primo del 2004.

Segue un lento aumento del valore degli acquisti per riprendere solo alla fine del 2005 un livello posto approssimativamente al 25% del livello del primo quadrimestre 2003

I clienti appartenenti a questo *cluster* hanno dimostrato di poter acquistare in maniera rilevante ma non c'è stata probabilmente da parte dell'azienda la padronanza di gestire questi clienti in modo appropriato dato che le vendite nei loro confronti hanno subito un pesante tracollo.

Anche l'indice di frequenza è molto basso, quindi negativo, il quale ci spiega che la tendenza di questi clienti è di non acquistare presso Elettroingross.

Un ulteriore dettaglio è che questi clienti sono rimasti attivi nel mercato fino al momento attuale e si delineano timidi segnali di ripresa, ciò nonostante consideriamo il Cluster 1 un gruppo di clienti in DIMINUZIONE.

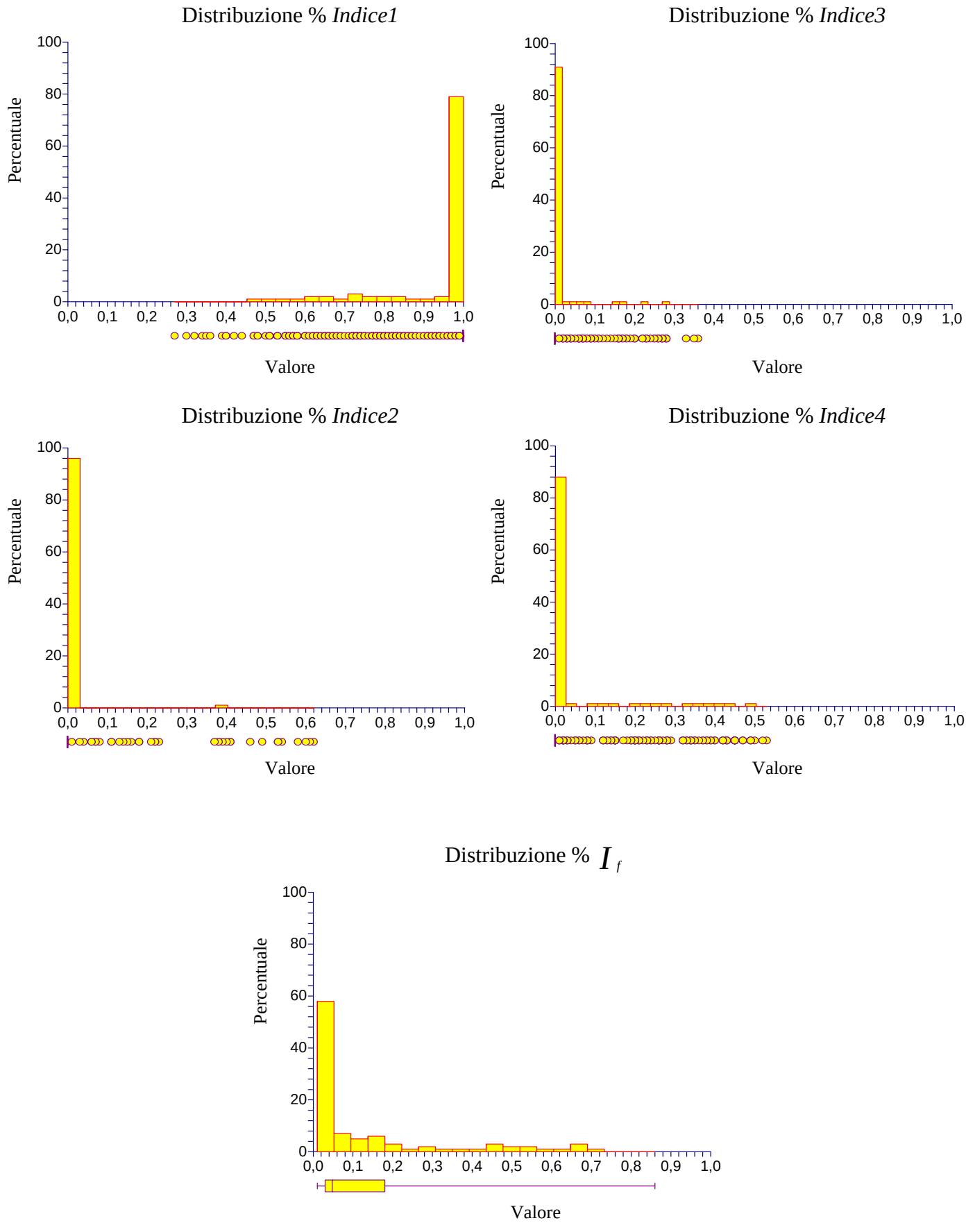


Figura 4.7b – Distribuzioni degli indici del Cluster 1.

	Cluster 1, n°792					
	Mean	**Deviation**	**Median**	**Minimum**	**Maximum**	**Range**
indice1	0,94	0,14	1	0,27	1	0,73
indice2	0,01	0,07	0	0	0,62	0,62
indice3	0,01	0,05	0	0	0,36	0,36
indice4	0,03	0,10	0	0	0,53	0,53
I_f	0,15	0,19	0,05	0,01	0,86	0,85

Tabella 4.7c – Statistiche descrittive degli indici del Cluster 1.

Cluster 2

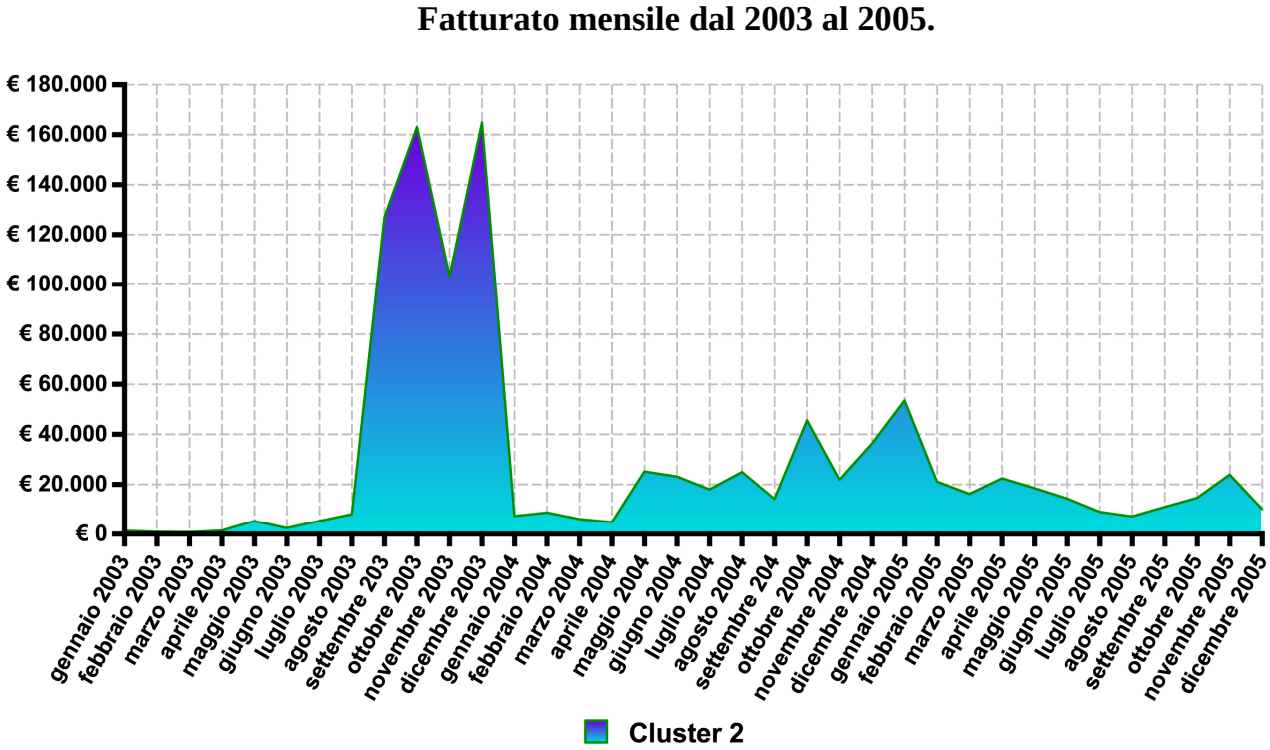


Figura 4.8a – Grafico fatturato dei clienti del Cluster 2.

Anche questo *cluster* rivela un picco di acquisti in un lasso temporale ristretto ma a differenza del cluster 1 in questo le vendite elevate sono posizionate sul terzo quadrimestre 2003.

A parte questa differenza valgono le medesime considerazioni fatte per il primo *cluster*: questi clienti hanno dimostrato di poter spendere in un certo modo ma Elettroingross non ha saputo approfittarne per fidelizzare concretamente questi clienti.

Effettivamente l'indice di frequenza è molto basso pertanto la valutazione di questo gruppo è ancora clienti in DIMINUZIONE, ovviamente la procedura di *clustering* ha formato gruppi distinti perché le distribuzioni degli indici discriminanti sono effettivamente diverse (vedi figura 4.7b e 4.8b) ma dal punto di vista manageriale questi due gruppi hanno similarità di comportamento.

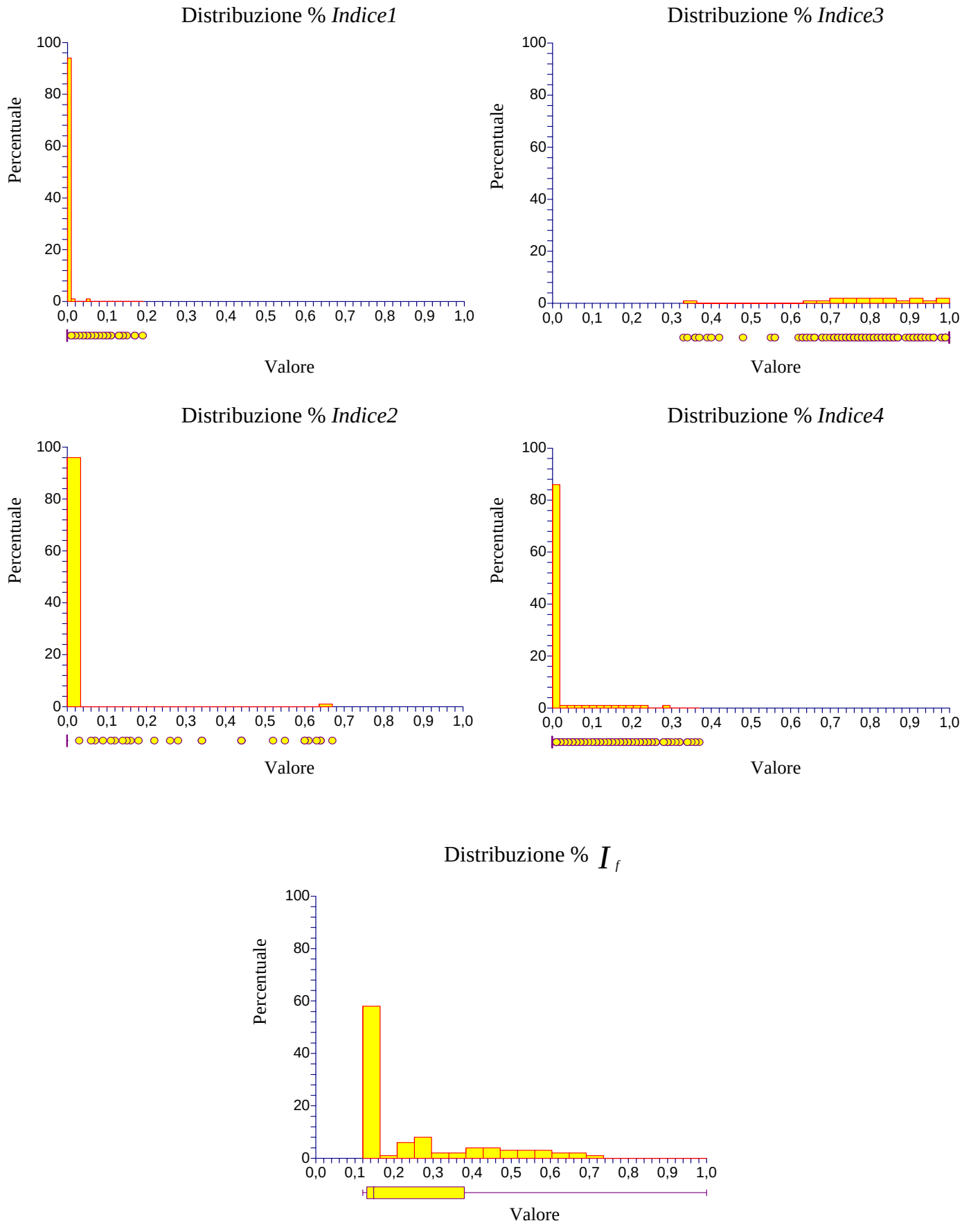


Figura 4.8b – Distribuzioni degli indici del Cluster 2.

	Cluster 2, n°644					
	Mean	**Standard Deviation**	**Median**	**Minimum**	**Maximum**	**Range**
indice1	0,00	0,02	0	0	0,19	0,19
indice2	0,02	0,09	0	0	0,67	0,67
indice3	0,96	0,11	1	0,33	1	0,67
indice4	0,02	0,07	0	0	0,37	0,37
I_f	0,26	0,18	0,15	0,12	1	0,88

Tabella 4.8c – Statistiche descrittive degli indici del Cluster 2.

Cluster 3

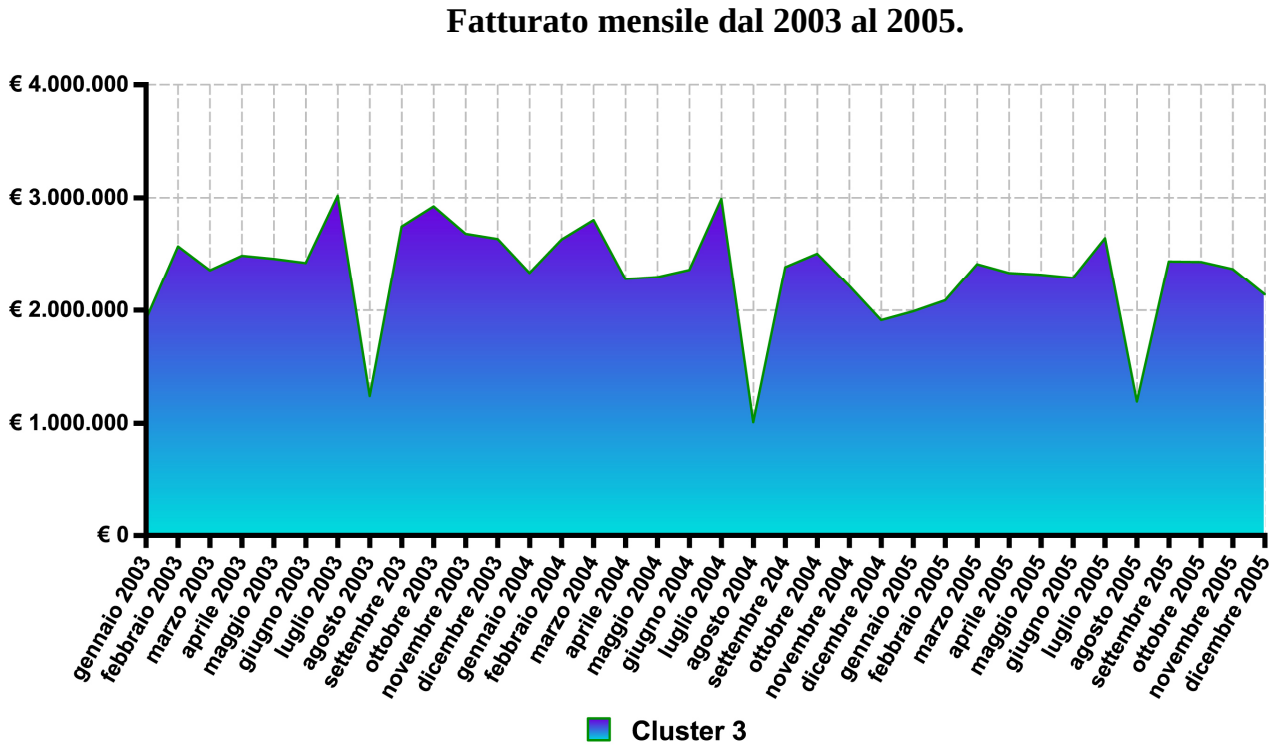


Figura 4.9a – Grafico fatturato dei clienti del Cluster 3.

Questo *cluster* è da definirsi il gruppo dei clienti costanti per eccellenza: si possono notare, infatti:

- che le distribuzioni degli indici di valore sono molto simili tra di loro, indicando quindi un comportamento stabile nel tempo;
- mediamente si delinea un andamento orizzontale.

Questo comportamento, che è stato rilevato dagli indice nel 2003 e nel primo quadrimestre 2004, si è protratto fino ad oggi; già con la segmentazione empirica si era riscontrato che le rilevazioni esercitate in un periodo relativamente ristretto permettono di esprimere giudizi che conservano la loro validità anche in tempi successivi. Dal punto di vista della frequenza d'acquisto si notano valori elevati, quindi, con tutte queste indicazioni si deduce che i clienti del Cluster 3 sono clienti COSTANTI in valore e clienti abituali dato che sono sempre presenti, in una parola: **fidelizzati.**

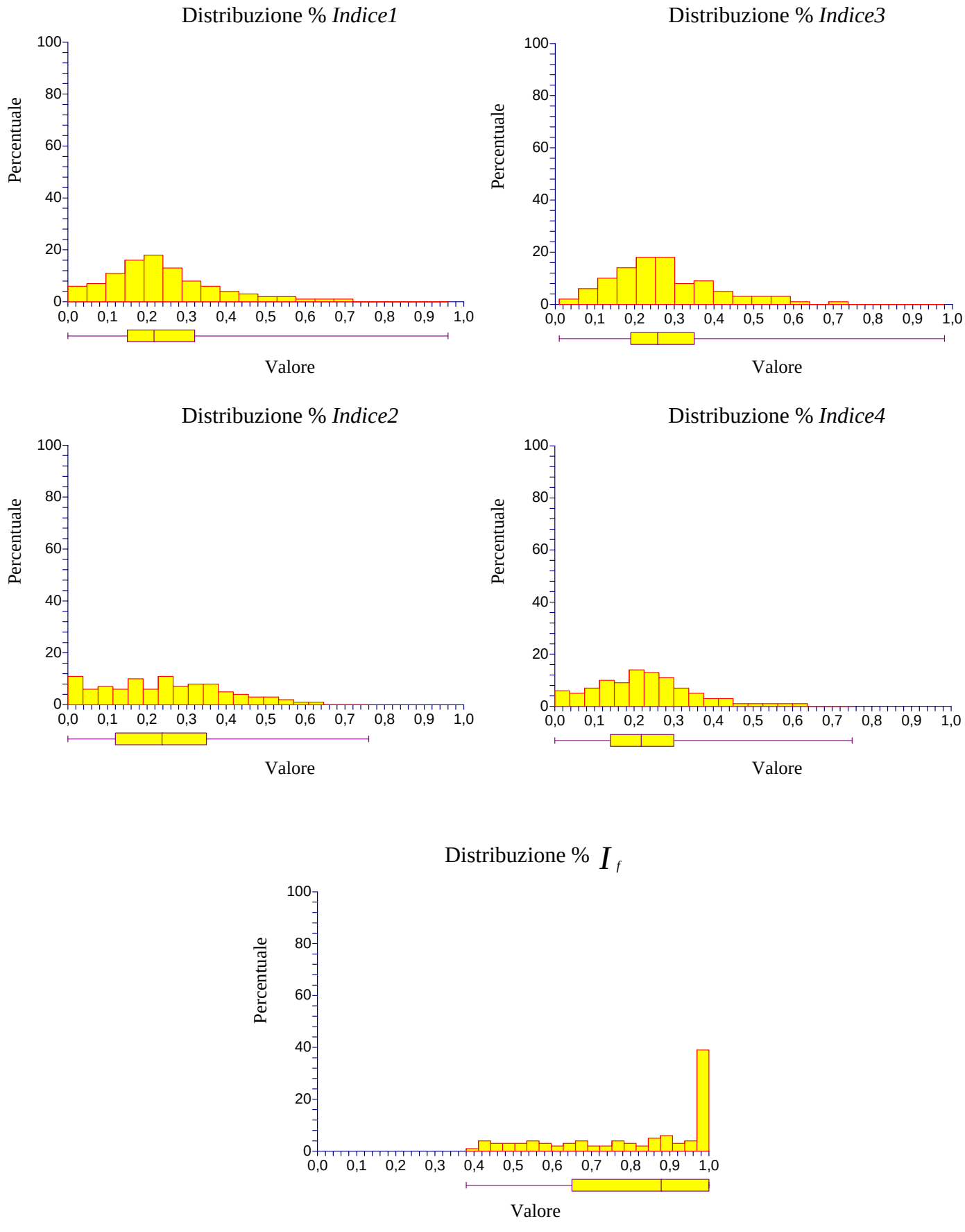


Figura 4.9b – Distribuzioni degli indici del Cluster 3.

	Cluster 3, n°1033					
	Mean	**Standard Deviation**	**Median**	**Minimum**	**Maximum**	**Range**
indice1	0,25	0,15	0,22	0	0,96	0,96
indice2	0,25	0,16	0,24	0	0,76	0,76
indice3	0,28	0,14	0,26	0,01	0,98	0,97
indice4	0,23	0,13	0,22	0	0,75	0,75
I_f	0,81	0,20	0,88	0,38	1	0,62

Tabella 4.9c – Statistiche descrittive degli indici del Cluster 3.

Cluster 4

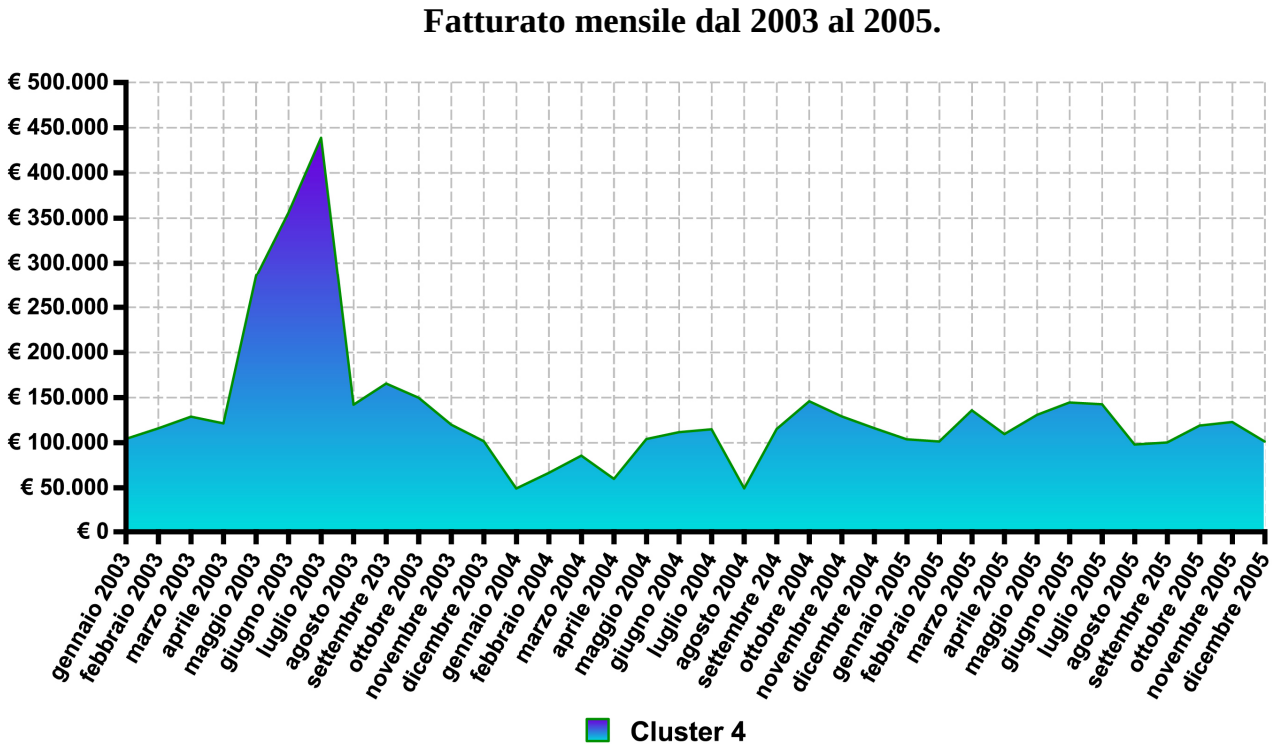


Figura 4.10a – Grafico fatturato dei clienti del Cluster 4.

Il quarto *cluster* si presenta con una tendenza di fondo stabile che si muove attorno ai 100.000 €.

Nel mese di luglio 2003 si riscontra il picco massimo di acquisti prossimo ai 450.000 € che i clienti di questa categoria hanno effettuato e una valutazione da un punto di vista manageriale porta a considerare questo livello come l'espressione della spesa massima che questi clienti potrebbero esercitare presso Elettroingross.

La distribuzione dell'indice di frequenza è concentrata in un range che varia tra 0,24 e 0,66 si muove quindi in un versante negativo-neutrale che fa presupporre che con questi clienti ci sia un ampio margine di miglioramento e che in ogni caso è stato raggiunto un equilibrio stabile.

I clienti di questa categoria sono da considerarsi COSTANTI.

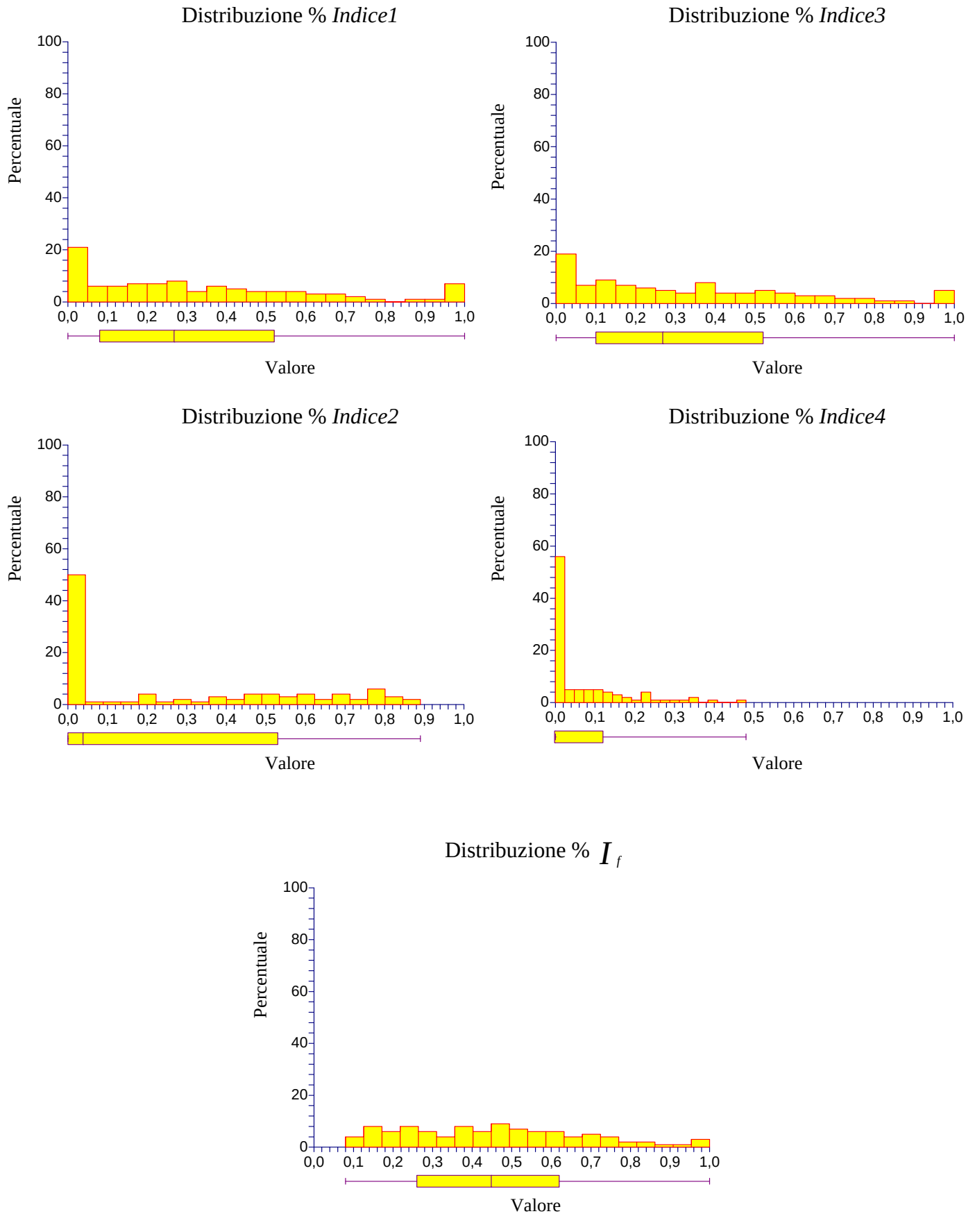


Figura 4.10b – Distribuzioni degli indici del Cluster 4.

	Cluster 4, n°536					
	Mean	**Standard Deviation**	**Median**	**Minimum**	**Maximum**	**Range**
indice1	0,26	0,31	0,04	0	0,89	0,89
indice2	0,34	0,29	0,27	0	1	1
indice3	0,33	0,28	0,27	0	1	1
indice4	0,07	0,11	0	0	0,48	0,48
I_f	0,46	0,23	0,45	0,08	1	0,92

Tabella 4.10c – Statistiche descrittive degli indici del Cluster 4.

Cluster 5

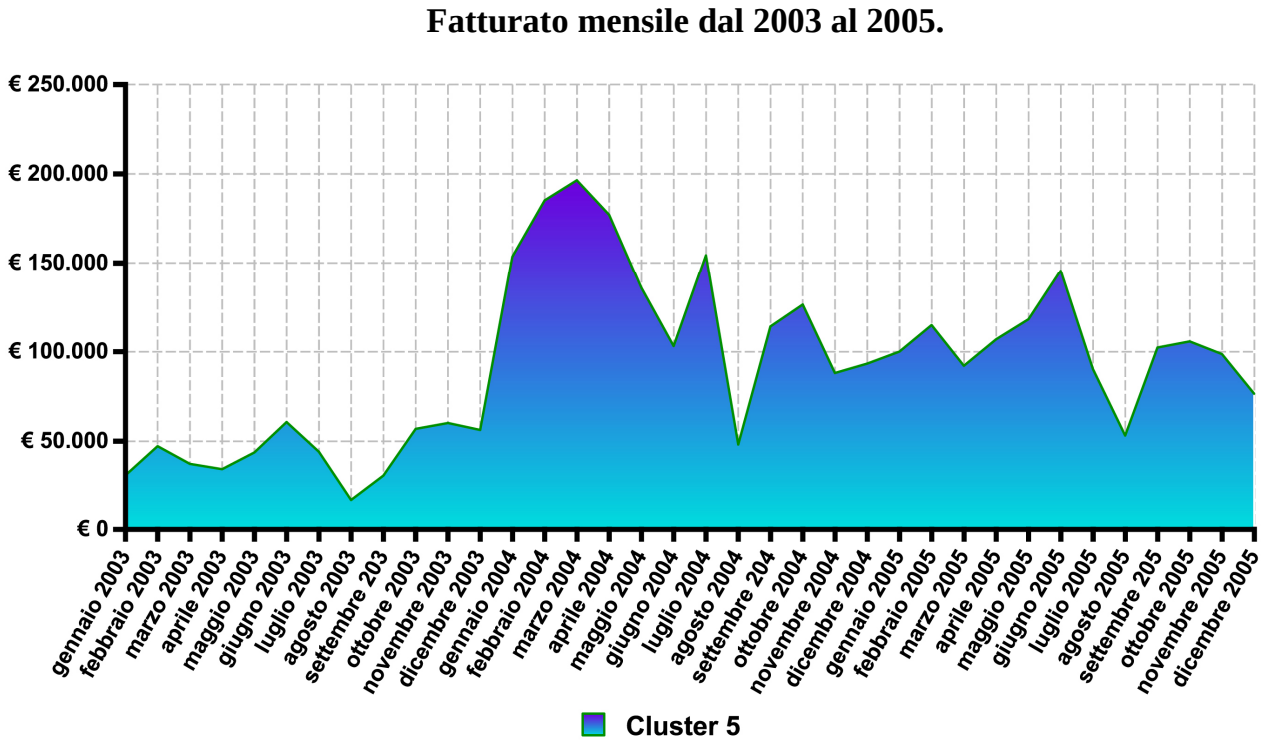


Figura 4.11a – Grafico fatturato dei clienti del Cluster 5.

La caratteristica di questo gruppo è che per tutto il 2003 il livello di spesa è stato prudenzialmente basso (intorno ai 45.000 €) per poi aumentare in modo deciso nel primo quadrimestre sfiorando i 200.000 € e assestarsi dal giugno 2004 fino alla fine del 2005 ad un livello medio di 100.000 € circa.

Osservando le distribuzioni degli indici si nota che l'*indice4* è la distribuzione che concentra i valori più elevati dimostrando che la spesa del primo quadrimestre 2004 ha una rilevanza maggiore rispetto a tutto il 2003.

L'indice di frequenza è piuttosto basso con una distribuzione asimmetrica verso destra confermando il fatto che i clienti che dimostrano una tendenza ad aumentare il valore di spesa in realtà stanno semplicemente aumentando il numero degli acquisti spostando verso l'alto il valore di I_f.

I clienti di questa categoria sono clienti in CRESCITA.

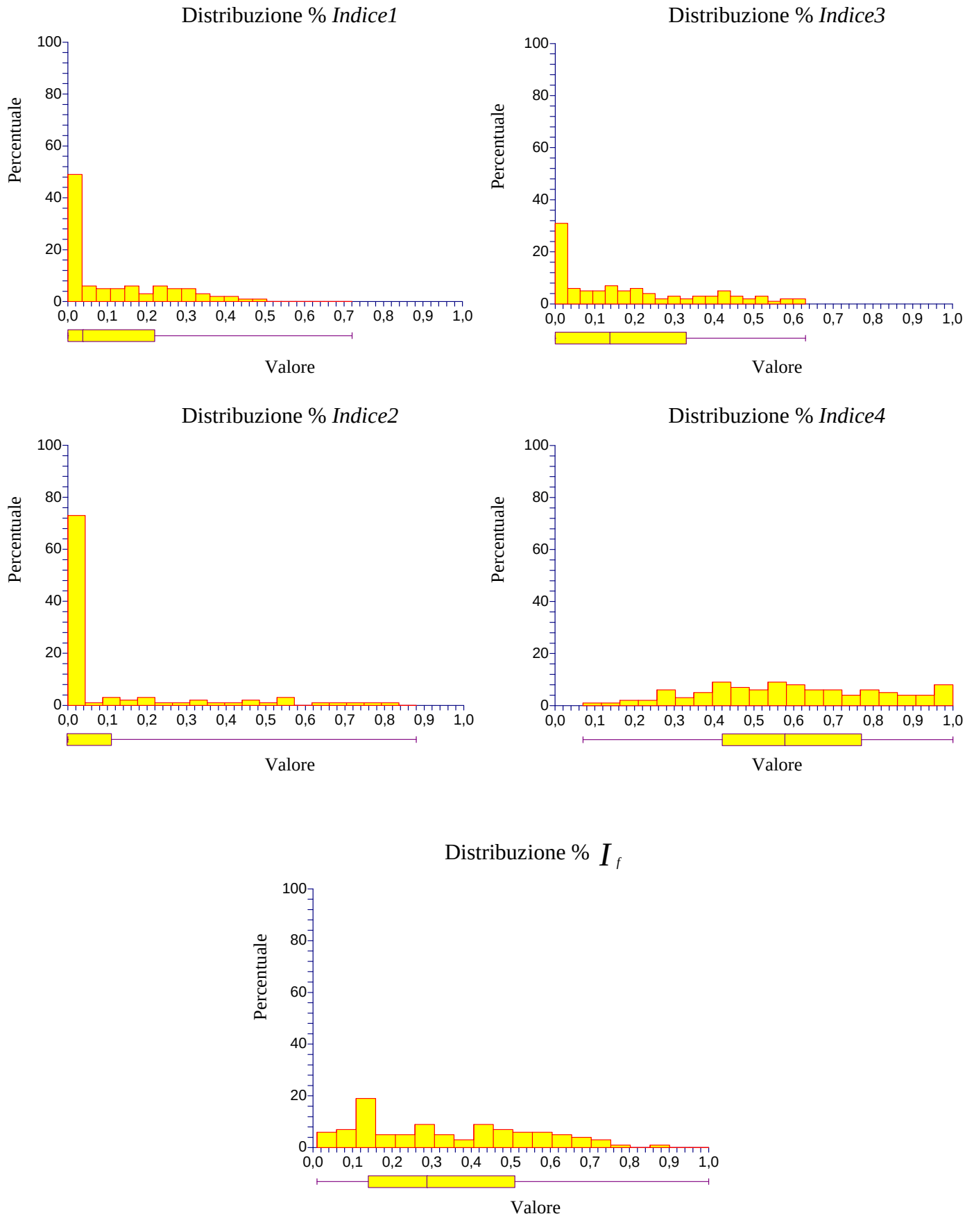


Figura 4.11b – Distribuzioni degli indici del Cluster 5.

	Cluster 5, n°485					
	Mean	**Standard Deviation**	**Median**	**Minimum**	**Maximum**	**Range**
indice1	0,11	0,14	0,04	0	0,72	0,72
indice2	0,11	0,22	0	0	0,88	0,88
indice3	0,19	0,18	0,14	0	0,63	0,63
indice4	0,59	0,23	0,58	0,07	1	0,93
I_f	0,34	0,21	0,29	0,01	1	0,99

Tabella 4.11c – Statistiche descrittive degli indici del Cluster 5.

Cluster 6

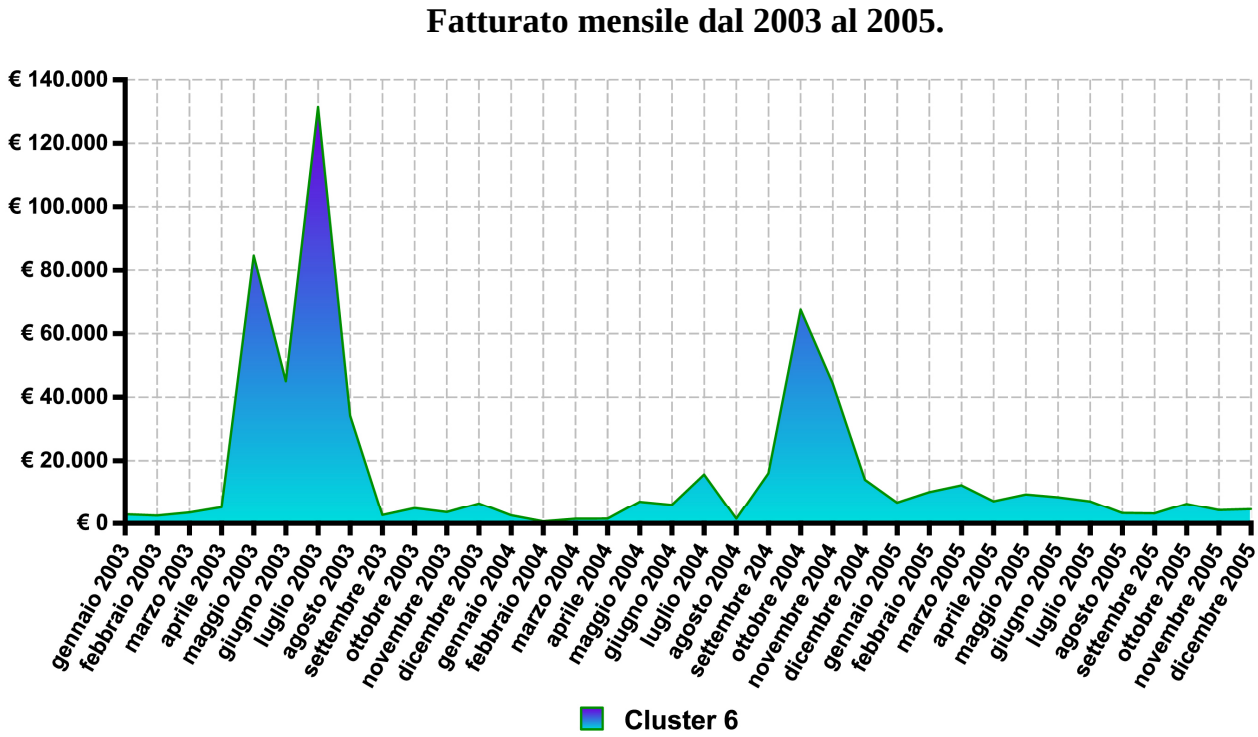


Figura 4.12a – Grafico fatturato dei clienti del Cluster 6.

Il sesto *cluster* rivela un altro gruppo di clienti che acquistano in modo scostante sia intermini di valore sia di frequenza.

Con questi clienti Elettroingross non sa esercitare un rapporto commerciale adeguato. La serie presenta dei picchi massimi decrescenti e comunque è predominante la tendenza ad azzerare gli acquisti, sembrano clienti fortemente decisi ad abbandonare Elettroingross salvo per tornare in qualche momento in modo del tutto imprevisto. Appare chiaro che questi clienti sono legati a qualche altro grossista e riservano solo acquisti sporadici in Elettroingross.

Il margine di miglioramento con questi clienti è solo potenzialmente elevato perché è ovvio che il contatto se esiste è molto debole.

Anche questi clienti li consideriamo in DIMINUZIONE.

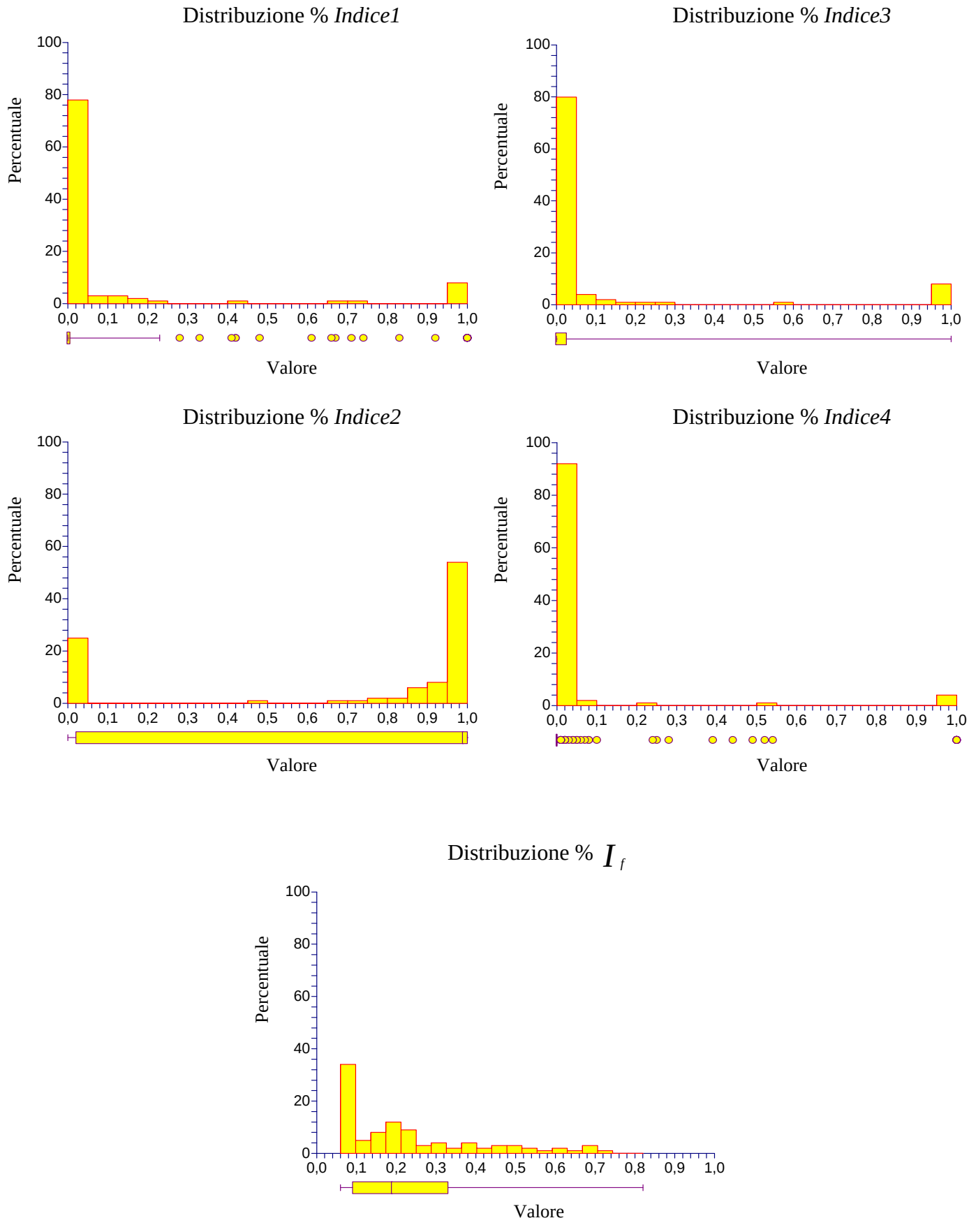


Figura 4.12b – Distribuzioni degli indici del Cluster 6.

	Cluster 6, n°297					
	Mean	**Standard Deviation**	**Median**	**Minimum**	**Maximum**	**Range**
indice1	0,12	0,29	0	0	1	1
indice2	0,71	0,42	0,99	0	1	1
indice3	0,11	0,29	0	0	1	1
indice4	0,05	0,21	0	0	1	1
I_f	0,24	0,18	0,19	0,06	0,82	0,76

Figura 4.12c – Statistiche descrittive degli indici del Cluster 6.

Riassumendo le valutazioni commerciali svolte fino a questo punto si possono studiare 3 diverse strategie commerciali per:

- 1.733 clienti considerati in DIMINUZIONE;
- 1.569 clienti considerati COSTANTI;
- 485 clienti considerati in CRESCITA.

Ci si potrebbe chiedere se non valga la pena studiare azioni mirate per ogni *cluster*, ed effettivamente lo si potrebbe fare ma, poiché i cluster vengono interpretati con una prospettiva manageriale non sono richieste suddivisioni ulteriori a quelle di cui sopra.

Un' altra domanda potrebbe essere per quale motivo non si sia predisposto direttamente la *cluster analysis* per l'ottenimento di soli 3 gruppi e la risposta è semplicemente perché non si sarebbero ottenuti i medesimi risultati: si sarebbero fusi gruppi "simili" secondo l'algoritmo di *clustering*, ma in realtà, da un punto di vista commerciale sarebbero stati associati e valutati in modo differente.

Nel caso di questo lavoro si è preferito ottenere i gruppi con la massima espressività e in sede di interpretazione dei risultati formulare le associazioni ritenute rilevanti dal punto di vista strategico – commerciale.

4.2 Conclusioni.

La *Cluster Analysis* ben si presta a individuare clienti accomunati da comportamenti d'acquisto simili.
In questo lavoro si può confrontare una segmentazione empirica che da dei risultati parziali mentre la *cluster analysis* è in grado di individuare in maniera molto efficace dei gruppi di clienti che acquistano in un modo particolare e diverso dagli altri gruppi.
Il difetto riscontrato nella segmentazione empirica in cui la categoria costante racchiudeva in realtà clienti anche molto diversi tra loro con la *Cluster Analysis* questo genere di problema non è stato minimamente riscontrato.
Sono stati ottenuti infatti 6 gruppi con numerosità non troppo elevate.
L'efficienza è attribuibile sostanzialmente ai seguenti fattori:

- l'opportuna e buona costruzione degli indici d'acquisto e di frequenza per essere utilizzati come variabili discriminanti dato che non esisteva niente di adatto e disponibile;
- la scelta della distanza euclidea che come sappiamo è una distanza "fisica" tra due elementi;
- l'algoritmo di Ward che unisce gli elementi contraddistinti dalla distanza minore;
- le distribuzioni degli indici discriminanti sono risultate più concentrate rispetto alla segmentazione empirica;
- l'andamento molto preciso del fatturato e molto particolare per ogni cluster individuato;
- la possibilità di dare a posteriori una valutazione commerciale e non a priori come avviene nella segmentazione empirica.

La segmentazione ottenuta è stata molto utile per impostare misure di incentivazione all'acquisto o di recupero dei clienti in base ai profili individuati.

Della *Cluster Analysis* si è apprezzato molto la capacità di isolare e raggruppare clienti con tendenze d'acquisto particolari che altrimenti non si sarebbe potuto prevedere o immaginare data la limitata conoscenza iniziale del fenomeno, quindi, possiamo affermare che tutte le analisi a priori sono "viziate" dalle conoscenze, opinioni e pregiudizi del ricercatore che vanno benissimo nello studio di un fenomeno già conosciuto almeno parzialmente, ma nel caso di un fenomeno nuovo e magari mai misurato possono limitare fortemente i risultati ottenibili.

Ovviamente anche i risultati della tecnica dei gruppi sono influenzati da molti fattori e il successo o meno di questa tecnica dipende dalle scelte del ricercatore, dai dati, ma senza ombra di dubbio dalla capacità di individuare o creare, come nel caso di questo lavoro, le variabili discriminanti.

5. La Cluster Analysis con SPSS.

5.1 Introduzione

Praticare la *Cluster Analysis* senza l'ausilio di strumenti software appropriati renderebbe molto difficile l'elaborazione dei dati e l'ottenimento dei risultati rapidamente.

La *Cluster Analysis*, come molte altre discipline statistiche, ha beneficiato solo in tempi relativamente recenti di hardware e software utili a trattare grandi mole di dati e di conseguenza l'uso in larga scala di queste tecniche avviene in tempi successivi a quelli di formulazione e divulgazione.

Il presente lavoro è stato il pretesto per provare vari software che rendono possibile la tecnica dei gruppi, e il più adatto, di quelli testati, che ha risposto meglio in termini di parametrizzazioni (scelte rese disponibili per l'utente) e rapidità di elaborazione è stato SPSS.

L'unica caratteristica che non è piaciuta di SPSS è il modo di rappresentare il dendrogramma che in presenza di una numerosità elevata di elementi (come nel nostro caso) tende a "spaccare" la rappresentazione su più fogli (o schermate) rendendo molto difficile la comprensione di come i gruppi si sono formati e più generalmente di come l'algoritmo di clustering si è comportato.

Nel presente lavoro per la rappresentazione dei dendrogrammi è stato utilizzato un altrettanto valido programma di statistica: MINITAB.

MINITAB rappresenta in un'unica schermata l'intero dendrogramma che nel caso di numerosità elevata non è possibile cogliere il dettaglio del singolo

elemento ma viene comunque rappresentata l'intera sequenza di aggregazioni riuscendo a trasmettere nel complesso la costruzione dei gruppi.

5.2 L'utilizzo di SPSS per la Cluster Analysis.

Dopo aver importato i dati in SPSS, che devono essere strutturati in *casi* per riga e *variabili* su colonne, si richiama la procedura di clustering in questo modo:

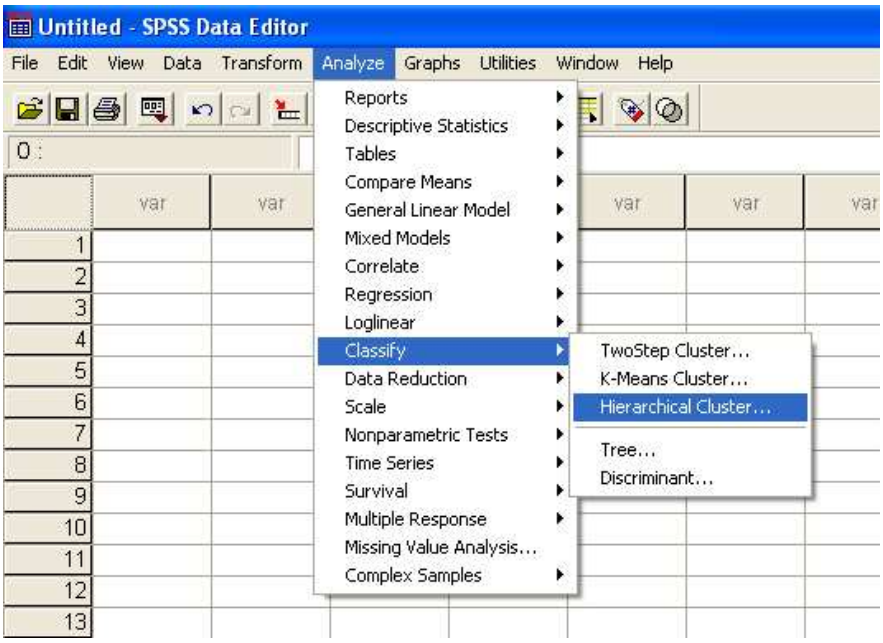


Figura 5.1 – Richiamare la procedura di clustering in SPSS.

Dal menù si seleziona: *Analyze >>> Classify >>> Hierarchical Cluster* e appare la seguente finestra.

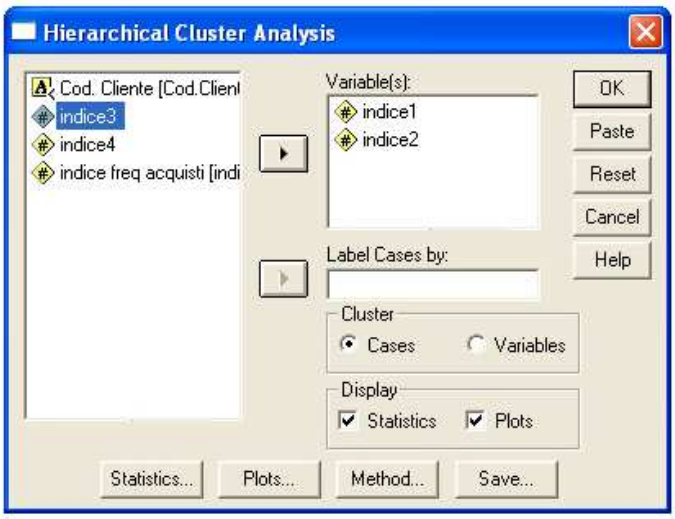


Figura 5.2 – Finestra generale per l'impostazione della Cluster Analysis.

In questa finestra di figura 5.2 si stabiliscono le variabili discriminanti da utilizzare spostandole dalla lista di sinistra sulla casella *Variable(s)*.

Il frame *Cluster* permette di stabilire se si intende praticare l' analisi dei gruppi per casi oppure per variabili selezionando l'opzione corrispondente.

Nel frame *Display* si stabilisce invece cosa deve essere visualizzato sul report che scaturisce da ogni attività di SPSS; selezionando *Statistics* e *Plots* si visualizzerà tutto ciò che è stato impostato nei rispettivi pulsanti situati nella parte inferiore della finestra.

Cliccando sul pulsante *Statistics* appare la seguente finestra:

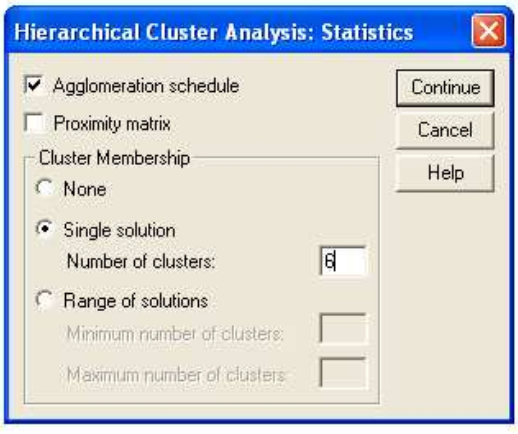


Figura 5.3 – Impostazioni per l'ottenimento dei gruppi.

Selezionando *Agglomeration schedule* verrà generata la scheda di aggregazione (vedi capitolo quarto, figura 4.2) mentre *Proximity matrix* è la matrice di dissimilarità che nel nostro caso non è stato possibile rappresentare perché non vi erano abbastanza risorse disponibili.
Per rappresentarla, infatti, doveva risultare una matrice di 3787×3787 elementi[4].
Da qui è possibile stabilire quanti gruppi devono essere ottenuti inserendone il numero su *single solution, number of clusters* oppure un range di soluzioni su *range of solutions* indicando il numero minimo di cluster ed il massimo che si è interessati ad ottenere.

Ciccando sul pulsante *plots* in figura 5.2 compare la finestra:

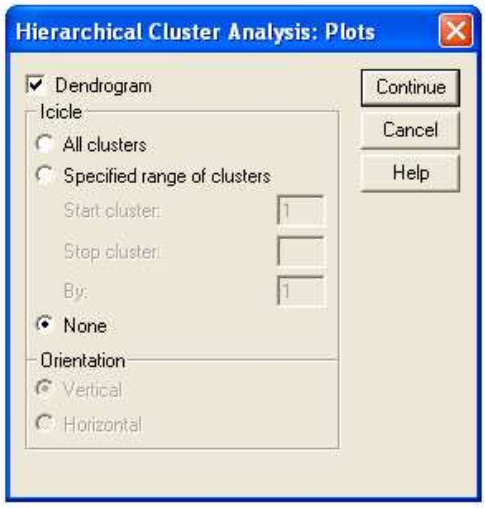


Figura 5.4 – Impostazioni grafiche per l'ottenimento dei gruppi.

In questa finestra è possibile richiedere la rappresentazione del dendrogramma e dell'icicle.

[4] In MS excel ad esempio la dimensione massima rappresentabile è una matrice di 65.536×256 celle.

In figura 5.2 selezionando *method* appare:

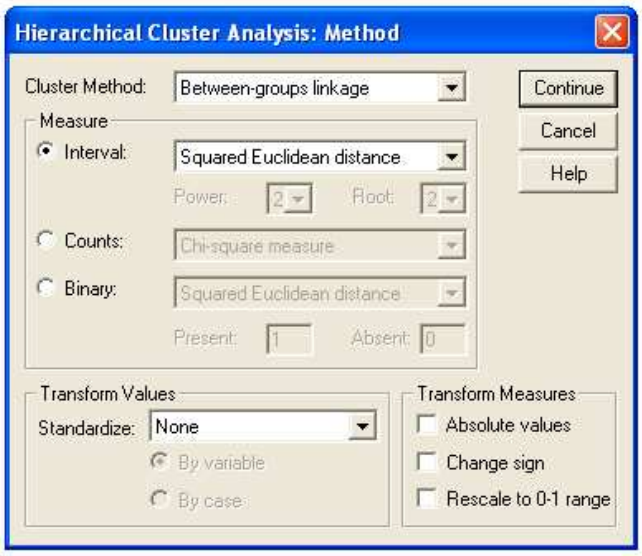


Figura 5.5a – Impostazioni della distanza e dell'algoritmo.

In questa finestra è possibile scegliere tra vari algoritmi di *clustering* come è evidenziato in figura 5.5b

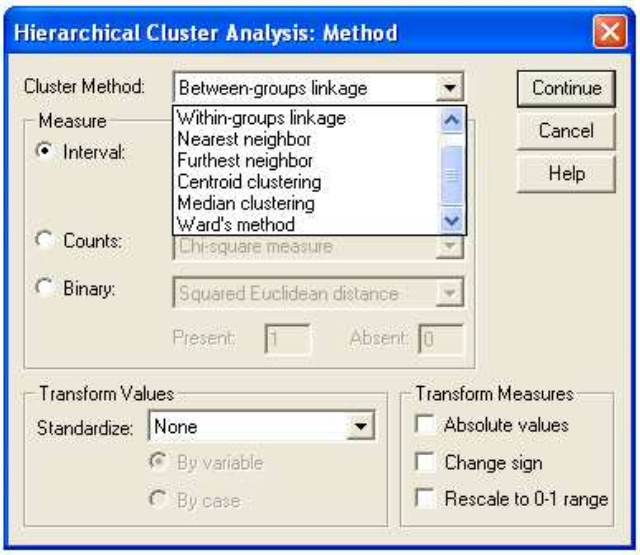


Figura 5.5b – Impostazioni dell'algoritmo.

Sono disponibili ovviamente anche diverse misure di distanza:

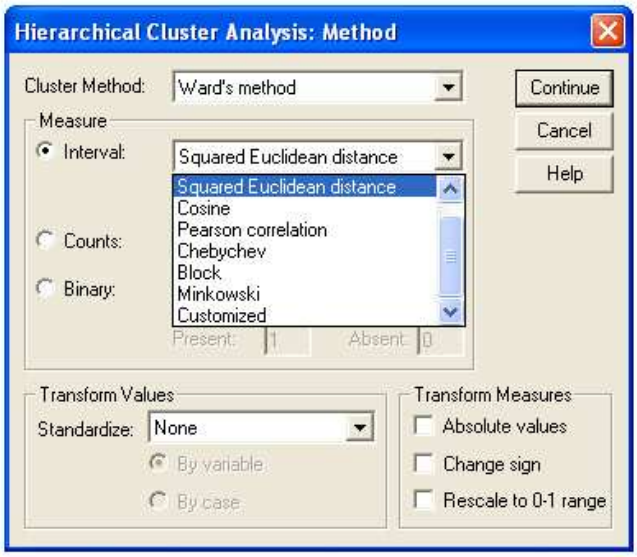


Figura 5.5c – Impostazioni della distanza.

Volendo è possibile anche standardizzare i dati qualora sia doveroso portare le diverse variabili ad una scala di misura comune.

Bibliografia.

Barbarito L.

1999 *L'analisi di settore: metodologia e applicazioni*, Milano, Franco Angeli,

Biorcio R.

1993 *L' analisi dei gruppi*, Milano, Franco Angeli.

Brasini S., Freo M., Tassinari F., Tassinari G.

2002 *Statistica Aziendale e analisi di mercato*, Bologna, Il Mulino.

Caroli M. G.

1999 *Il Marketing territoriale*, Milano, Franco Angeli.

Cuomo M. T.

2000 *La customer satisfaction. Vantaggio competitivo e creazione di valore*, Cedam.

East R.

2003 *Comportamento del consumatore*, Milano, Apogeo.

Everitt B. S.

2001 *Cluster Analysis*, Oxford University Press, fourth edition.

Gower J. C. and Legendre P.

1986 *Metric and euclidean properties of dissimilarity coefficients*, Journal of Classification.

Hartigan J. A.

1975 *Clustering algorithms*, New York, Wiley.

Jaccard P.

1963 *An Introduction of Numerical Classification*, New York, Academic Press.

Jardine E. e Sibson R.

1971 *Mathematical Taxonomy*, New York, John Wiley.

Lance G. N. e Williams W. T.

1967 *A General Theory of Classification Sorting Strategies, hierarchical systems*, in "Computer Journal", n. 10, pp. 271-277.

Mariani P.

2005 *Prendersi cura del proprio prodotto*, Milano, Franco Angeli.

Massart D.L., Vandeginste B.G.M., Deming S.N., Michotte Y., Kaufman L.

1988 *Chemometrics: a textbook*, Elsevier.

McQuitty L. L.

1964 *Capabilities and Improvements of Linkage Analysis as a Clustering Method*, in "Educational and Psychological Measurement", n. 24, pp. 441-456.

Sokal R. e Michener C. D.

1958 *A statistical method for evaluating systematic relationships*, in "University of Kansas Scientific Bulletin", n. 38, pp. 1409-1438.

Sokal R. e Sneath P.

1963 *Principles of Numerical Taxonomy*, San francisco, W. H. Freeman.

Tryon R. C.

1939 *Cluster Analysis*, New York, Mc Graw-Hill

Ward J.

1963 *Hierarchical grouping to optimize an ojective function*, in "The journal of the American Statistical Association" n. 58, pp. 236-244.

Ziliani C.

1999 *Micromarketing: Le carte fedeltà della distribuzione in europa*, Milano,Egea.