AIR MILES
un case study di customer segmentation

Da: G. Saarenvirta, “Mining customer data”, DB2 magazine on line, 1998

http://www.db2mag.com/db_area/archives/1998/q3/98fsaar.shtml

Customer clustering & segmentation

- due delle metodologie più importanti usate nel marketing
- si usano le transazioni d'acquisto dei clienti per
  - tracciare il comportamento di acquisto
  - creare iniziative strategiche di business
  - dividere i clienti in segmenti, sulla base di variabili:
    - redditività del cliente per l'impresa
    - misure di rischio
    - misure del Customer Lifetime Value
    - probabilità di fedeltà

Life Cycle e LifeTime Value

- Il comportamento dei clienti varia nel tempo
  - Tali modifiche sono indicazioni per il futuro
- Le modifiche nel tempo sono note come *Customer LifeCycle*
- LTV Calculation e Customer Acquisition Cost Calculations
  - Supponiamo che un cliente medio faccia acquisti per 2 anni, quindi si interrompa per almeno 1 anno
  - Il ciclo di vita del cliente è di 2 anni
  - Sui 2 anni il cliente medio effettua 16 acquisti
  - 16 x $1.20 profitto per acquisto = $19.20 LTV del cliente medio
- `http://www.jimnovo.com/LTV.htm`

Customer segments

- Esempio: clienti high-profit, high-value, low-risk
  - Tipicamente il 10-20% dei clienti che creano il 50-80% dei profitti di un'azienda
- Potrebbe essere attraente anche un segmento di clienti low-profit, high-value, low-risk
  - l'iniziativa strategica per il segmento è di aumentare i profitti
  - cross-selling (vendita di nuovi prodotti)
  - up-selling (vendere in quantità maggiore ciò che i clienti già acquistano)

Segmenti di comportamento/demografici

- All'interno di insiemi di comportamento, si possono creare sotto-insiemi demografici
- I dati demograifici dei clienti tipicamente *non* sono usati per la creazione di cluster
- La (sub)segmentazione demografica è utilizzata per selezionare tattiche opportune (pubblicità, canali/campagne di marketing) per cercare di soddisfare le iniziative strategiche dei segmenti comportamentali

Il Loyalty Group in Canada

- Presenta un AIR MILES Reward Program (AMRP) per un'alleanza di oltre 125 compagnie in diversi settori industriali – finanza, carte di credito, vendita, dettaglio, gas, telecomunicazioni
- Include il 60% delle ditte Canadesi
- AMRP è un programma per frequent-shopper:
  - il cliente colleziona punti che possono essere usati per ottenere premi (viaggi aerei, pernottamenti in hotel, noleggio autovetture, biglietti di teatro, eventi sportivi, …)

Raccolta dati

- I partner dell'alleanza raccolgono i dati relativi alle transazioni dei clienti e li trasmettono a The Loyalty Group, che
  - memorizza tali dati sulle transazioni in un DB per iniziative di marketing dirette ai partner della coalizione
- Il DW di The Loyalty Group contiene
  - più di 6.3 milioni di record di prodotti
  - 1 miliardo di record di transazioni

Prima del Data Mining

- The Loyalty Group ha utilizzato tecniche standard di analisi
  - Analisi Recency, Frequency, Monetary
  - Strumenti OLAP
  - Metodi statistici lineari

  per analizzare il successo delle varie iniziative di marketing intraprese dalla coalizione e dai partner

Perché l'analisi RFM funziona?

- I tre principi alla base dell'analisi RFM:
  - I clienti che hanno acquistato recentemente sono più ricettivi alle promozioni successive rispetto ai clienti il cui ultimo acquisto è lontano nel tempo
  - I clienti abituali sono più ricettivi rispetto a quelli saltuari
  - Coloro che spendono molto sono più ricettivi rispetto a chi spende poco
- Principi universali validi in diversi campi: assicurazioni, banche, dettaglio, viaggi, …
- L'analisi RFM semplicemente permette di "quantificare" i tre principi esposti
- Si codificano i clienti in "celle" e si esaminano le risposte dei clienti nelle varie celle alle stesse promozioni
- In genere, i clienti più ricettivi sono quelli nelle celle con punteggio RFM migliore

Obiettivi dell'analisi RFM

- Identificare pattern di migrazioni tra le celle RFM nel tempo
- Determinare se i pattern di migrazione dei clienti si distribuiscono in cluster
- Identificare strategie di investimento e di marketing appropriate per ciascun cluster di migrazione
- Stabilire l'efficacia dell'analisi RFM rispetto ad altre strategie disponibili

Progetto Data Mining alla AMRP

- Scopo:
  - creare una segmentazione dei clienti usando uno strumento DM
  - confrontare i risultati con quelli ottenuti in precedenza utilizzando l'analisi RFM
- Piattaforma Data Mining:
  - DB2 Universal Database Enterprise parallelizzato su un sistema RS/6000 SP a 5 nodi
  - Intelligent Miner for Data
    - Motivi: possiede algoritmi di clustering categorico e scoperta di regole associative

Modello dei dati

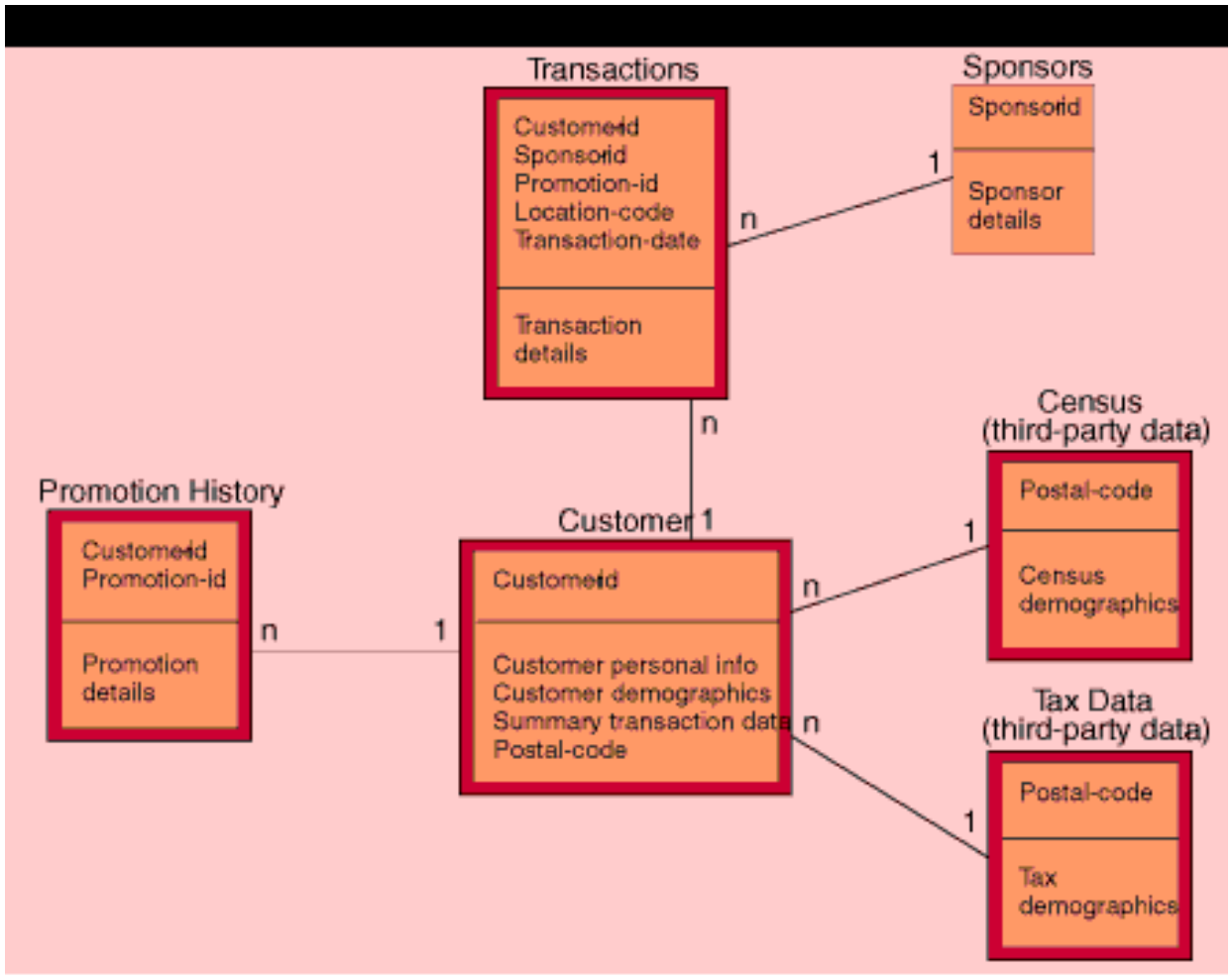


Figure 2. *AIR MILES case study data model.*

- ~ 50,000 clienti e relative transazioni per un periodo di 12 mesi

Preparazione dei dati

- variabili “shareholder value”
  - reddito
  - tempo di sopravvivenza del cliente
  - numero di compagnie “sponsor” utilizzate nel tempo di sopravvivenza del cliente
  - numero di compagnie “sponsor” utilizzate negli ultimi 12 mesi
  - distanza (in mesi) dell’ultima transazione
- calcolate aggregando i dati delle transazioni ed aggiungendoli ai record di ciascun cliente

Preparazione dei dati (2)

- Dataset ottenuto tramite un join dei dati delle transazioni con la tabella clienti per creare l'input dell'algoritmo di clustering
- 84 variabili =
  - 14 categorie per le compagnie sponsor ×
  - 3 variabili per categoria ×
  - 2 trimestri (primi due trimestri del 1997)

Data Cleaning – valori mancanti

- Dati demografici
  - Normalmente dati categorici
  - hanno un'alta % di valori mancanti
  - i valori mancanti possono essere resi come "sconosciuto" o "non risponde"
- Se una grande frazione dei dati di un attributo è mancante, può convenire eliminare l'attributo
- Nel case study, i valori numerici mancanti sono posti a 0

Trasformazione dei dati

- Variabili rapporto
  - Es.: profitability = profitto/tempo di vita
- Variabili tempo-differenziali
  - Es.: profitto 2o trimestre – profitto 2o trimestre
- Discretizzazione usando quantili
  - Es.: break point a 10, 25, 50, 75, and 90%.
- Discretizzazione usando intervalli predefiniti
  - Es.: quelli usati nei censimenti
- Trasformazioni logaritmiche
  - Es.: per distribuzioni non uniformi

- Cleaning e trasformazione effettuati a livello DB (SQL)

Perché trasformare i dati?

- Per rimuovere l'effetto degli outlier
  - gli algoritmi che minimizzano l'errore quadratico medio sono estremamente sensibili ai valori al di fuori degli intervalli di maggioranza (99%)
  - in taluni casi, gli outlier sono i pattern interessanti (es.: fraud detection)
- per rendere i dati meglio interpretabili
  - dati non uniformi
  - discretizzazione

Distribuzione dei dati discreti

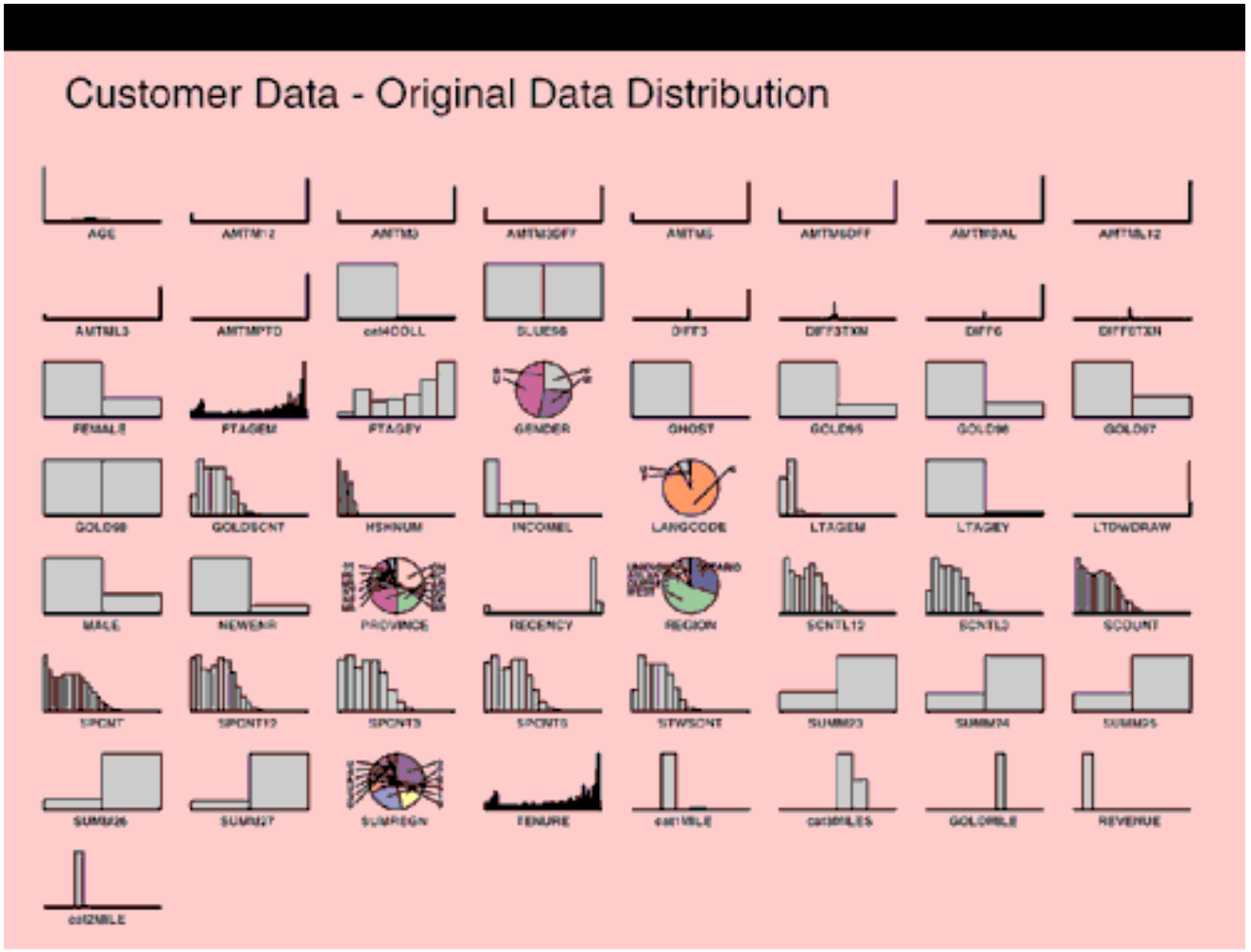


Figure3. *Original data.*

Distribuzione originale dei dati

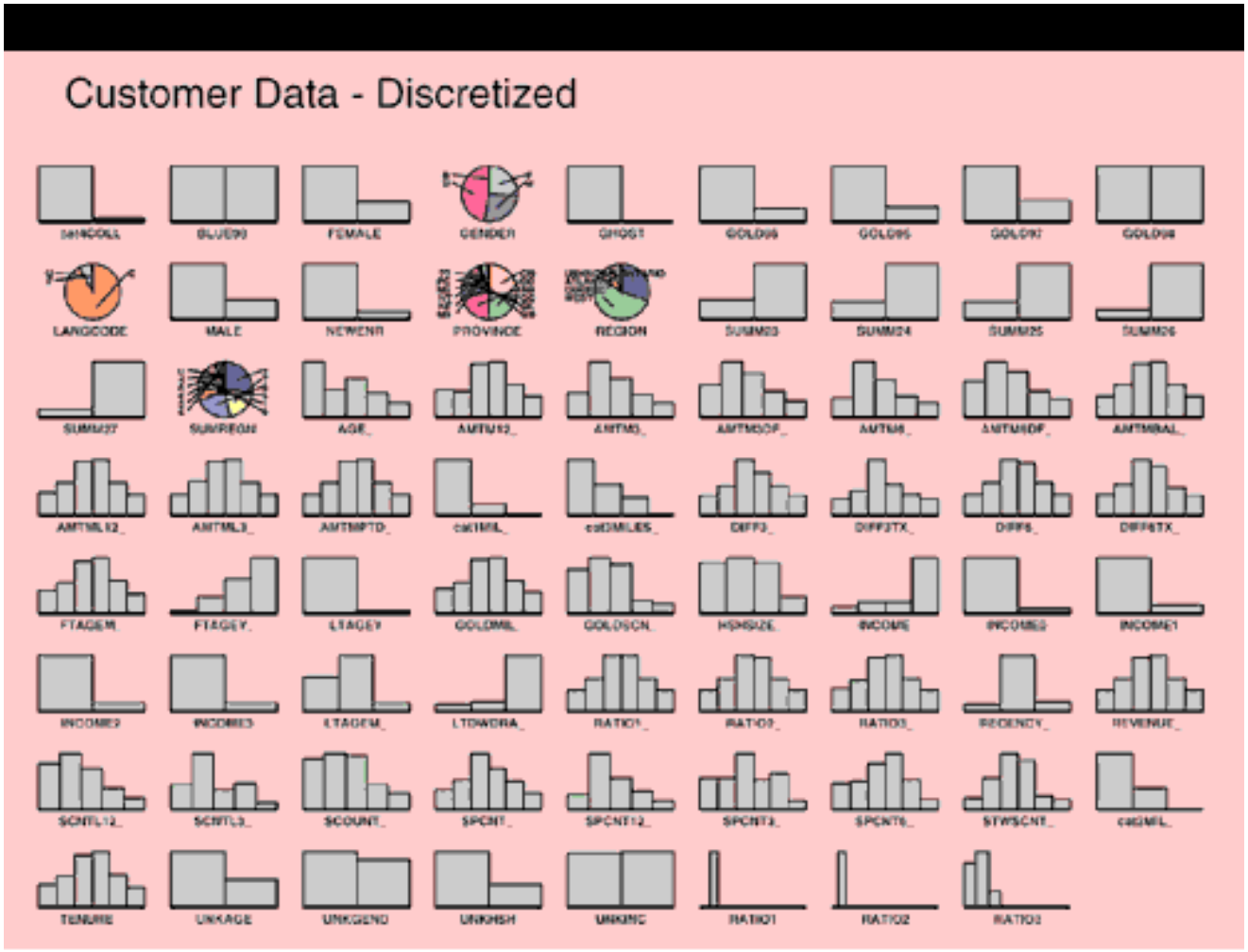


Figure 4. *Discretized data.*

Prima/dopo la discretizzazione

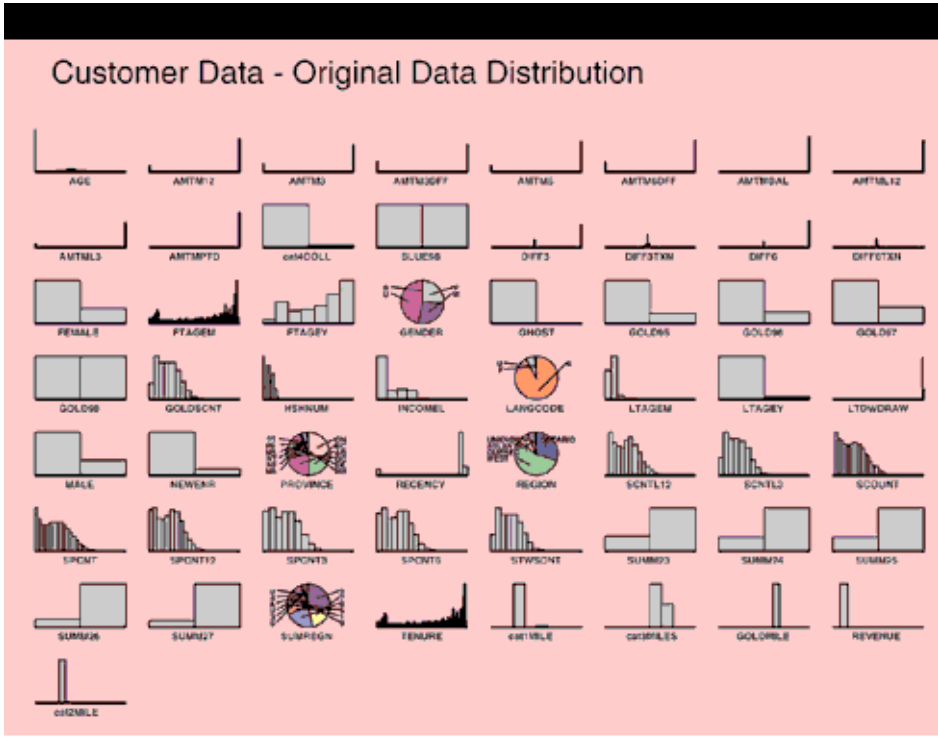


Figure3. *Original data.*

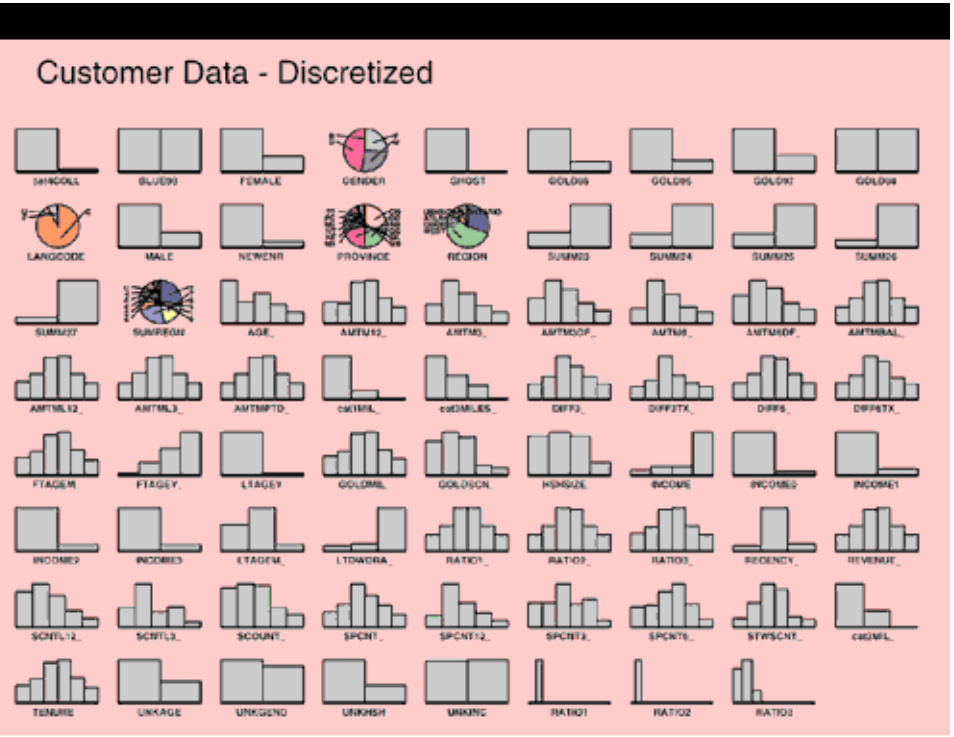


Figure 4. *Discretized data.*

Tecnica di clustering/segmentazione

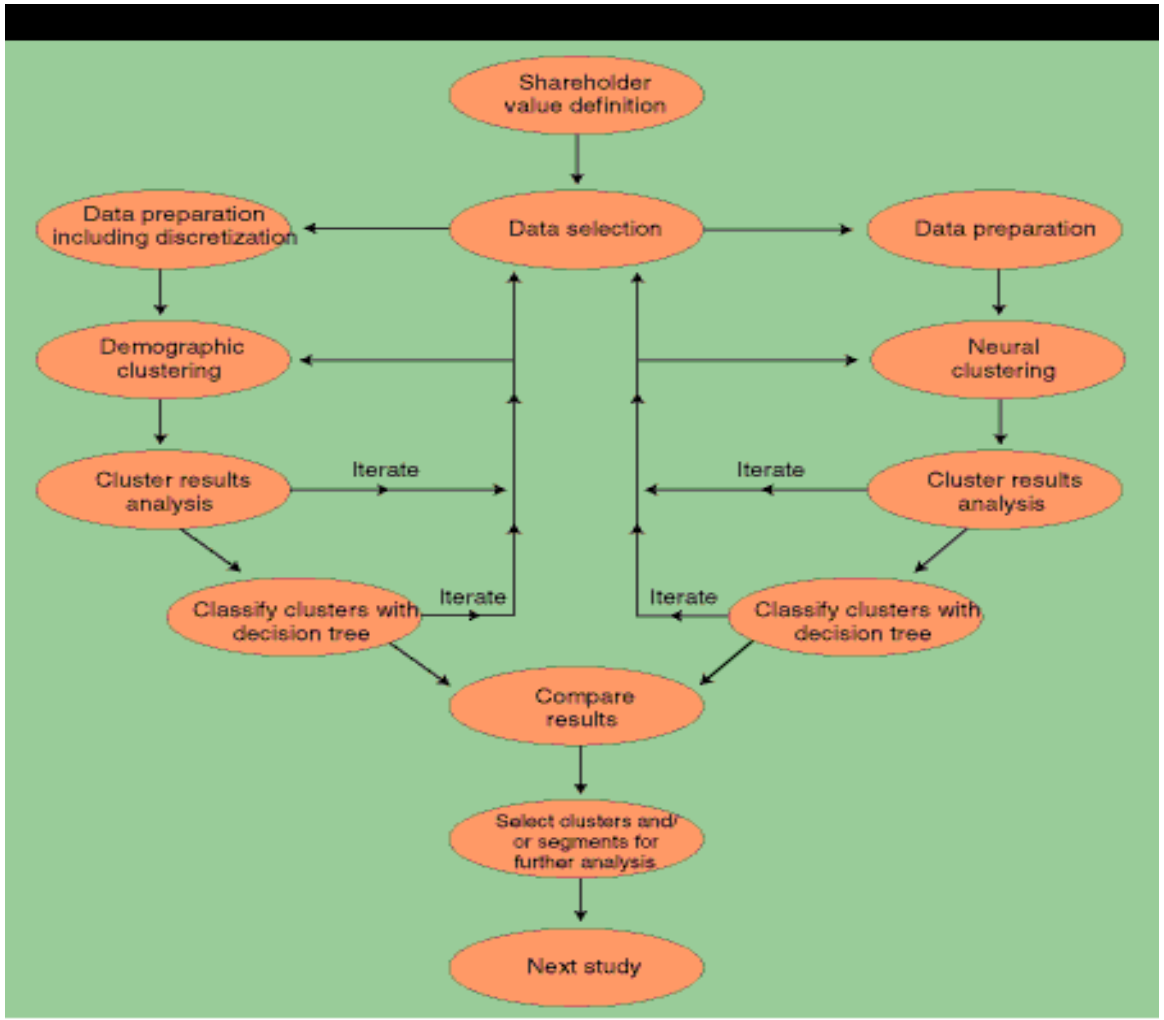
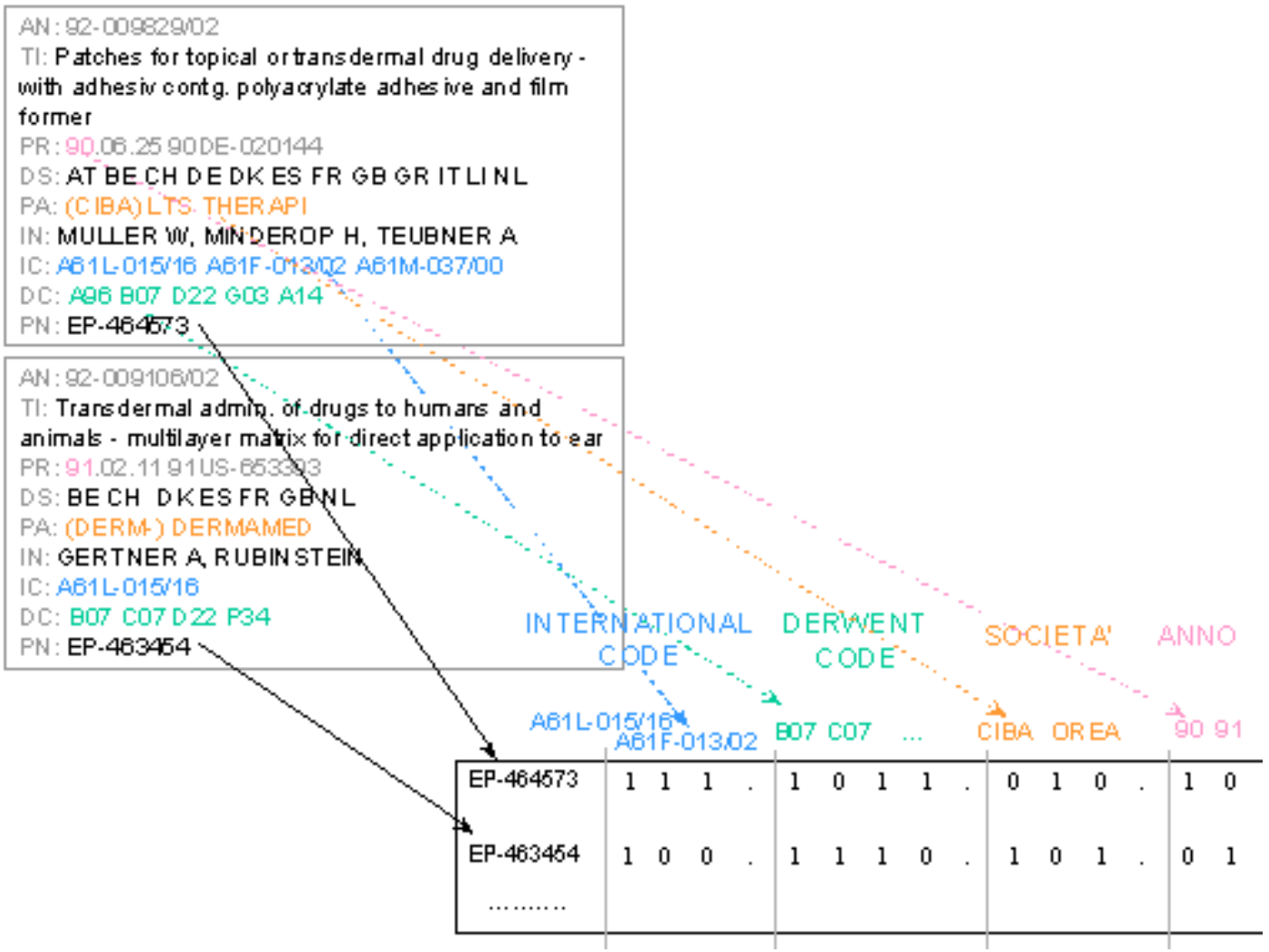


Figure 6. *Clustering workflow.*

Clustering demografico IBM-IM

- Progettato per variabili categoriche
- Indice di similarità
  - aumenta col numero di valori uguali su attributi comuni
  - diminuisce col numero di valori diversi su attributi comuni
- # di cluster non fissato a priori
  - solamente un limite superiore
- massimo numero di iterazioni (→precisione)
- precisione

Clustering demografico: strutture dati



Clustering demografico: parametri

	W_1	W_2	...								W_n
Doc i	1	1	1	1	0	1	1	0	1	0	1	0
Doc j	1	0	0	1	1	1	0	1	0	0	0	1

$$N_{11} = \sum_{k=1}^{m} x_{ik}\, x_{jk}$$

$$N_{10} = \sum_{k=1}^{m} x_{ik}\,(1 - x_{jk})$$

$$N_{01} = \sum_{k=1}^{m} (1 - x_{ik})\, x_{jk}$$

$$N_{00} = \sum_{k=1}^{m} (1 - x_{ik})\,(1 - x_{jk})$$

- Indice di Somiglianza

$$s(i,j) = \frac{a\, N_{11}}{b\, N_{11} + c\,(N_{10} + N_{01})}$$

  - **Condorcet** $a = b = 1$ $c = 1/2$
  - **Dice** $a = b = 1$ $c = 1/4$

- Soglia di Somiglianza

se $s(i,j) > \alpha$ ⇒ Doc_i e Doc_j sono simili α in $[0,1]$

  - default: $\alpha = 0.5$

- Sistema di ponderazione

$$N_{11} = \sum_{k=1}^{m} x_{ik}\, x_{jk}\, w_k \quad (N_{10} = .. N_{01} = ...)$$

  - $w_k = 1 / x_{.k}$
  - $w_k = \log(N / x_{.k})$

Clustering demografico: indice di similarità

- proporzionale a 1-1
- inversamente proporzionale a 0-1 ed 1-0
- indipendente da 0-0
- Indice di Condorcet =
  - N11 / (N11 + 1/2(N01 + N10))
- Indice di Dice =
  - N11 / (N11 + 1/4(N01 + N10))
- Dice è più “lasco” rispetto a Condorcet
  - appropriato quando ci sono oggetti molto diversi

Clustering demografico: indice di similarità

- Soglia di similarità α
  - i,j considerati simili se $s(i,j)>\alpha$
  - valori bassi (<0.5) appropriati se esistono oggetti molto diversi
- Pesatura per gli attributi
  - l'importanza degli attributi nell'indice di similarità può essere variata modificando i pesi
  - peso standard = 1

Clustering demografico IBM-IM

- parametri base:
  - Numero massimo di cluster
  - Numero massimo di iterazioni sui dati
  - Precisione: criterio di arresto prematuro dell'algoritmo
    - se il cambiamento del criterio di Condorcet tra due iterazioni successive è minore della precisione (in %), l'algoritmo viene arrestato
  - Criterio di Condorcet è un valore in [0,1]:
  - 1 indica una segmentazione perfetta
    - tutti i cluster sono omogenei, e
    - completamente diversi dagli altri cluster

… altri parametri

- Soglia di similarità
  - definisce la soglia di similarità tra due valori in termini di distanza
  - se la soglia di similarità è 0.5, allora due valori sono considerati uguali se la loro differenza in valore assoluto è ≤ 0.5
- Nel case study:
  - # massimo di cluster: 9
  - # massimo di iterazioni: 5
  - precisione: 0.1

Input dataset

- dataset: variabili continue discretizzate
- variabili in input:
  - # di prodotti acquistati nel tempo di vita del cliente
  - # di prodotti acquistati negli ultimi 12 mesi
  - Contributo del cliente ai profitti nel tempo
  - Tempo di vita del cliente (in mesi)
  - Rapporto profitto/tempo di vita
  - Rapporto # di prodotti/tempo di vita
  - Regione di provenienza
  - Distanza ultimo acquisto
  - Anzianità (# di mesi dall'entrata del cliente nel programma).

Input dataset

- Altre variabili discrete e/o categoriche ed alcune variabili continue sono fornite come variabili supplementari:
- Variabili usate per fornire il profilo dei cluster ma non per definirli
- Forniscono una più semplice interpretazione dei cluster rispetto alle variabili di input

Output del clustering demografico

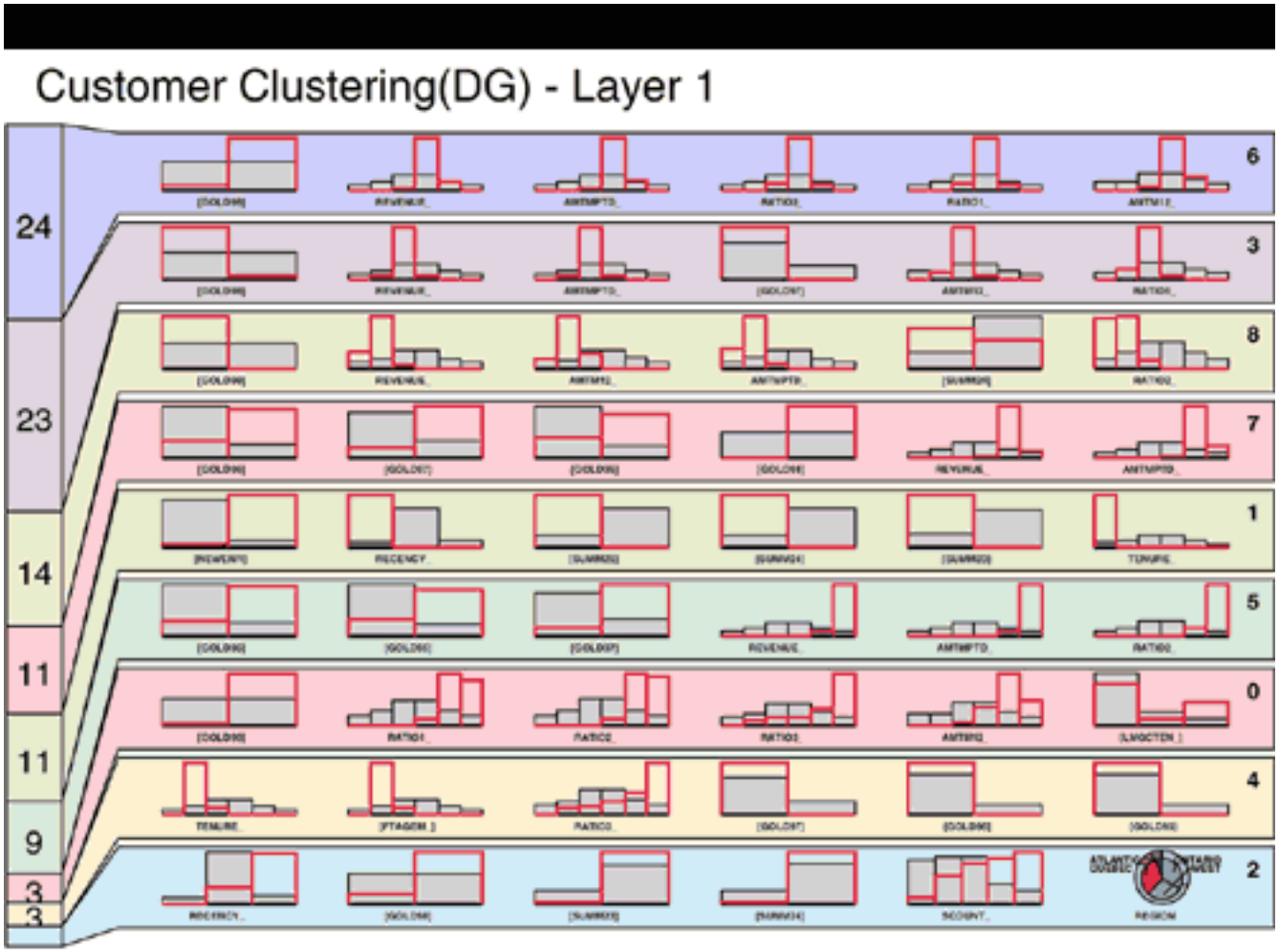


Figure 7. *Demographic clustering output.*

Visualizzazione dei cluster

- Ogni linea orizzontale = un cluster
- I cluster sono elencati dall'alto al basso in ordine di dimensione (# di clienti)
- Le variabili sono elencate da sinistra a destra, in ordine di importanza per il cluster, secondo un test chi-quadro tra la variabile e l'ID del cluster
- Altre metriche includono
  - l'entropia
  - il criterio di Condorcet, e
  - l'ordine all'interno del DB

Visualizzazione dei cluster

- Le variabili usate per definire i cluster sono indicate senza parentesi
- Le variabili supplementari sono indicate tra parentesi
- Le variabili numeriche, discrete, binarie, e continue vengono visualizzate come istogrammi rappresentanti la distribuzione (frequenza)
  - barre rosse = distribuzione della variabile all'interno del cluster
  - barre grigie = distribuzione della variabile all'interno del dataset

Visualizzazione dei cluster

- Le variabili categoriche sono visualizzate come grafici a torta
  - cerchio interno = distribuzione della variabile all'interno del cluster
  - anello esterno = distribuzione della variabile all'interno del dataset

- Maggiore è la differenza tra la distribuzione all'interno del cluster e quella globale, più "interessante" è il cluster

Output del clustering demografico

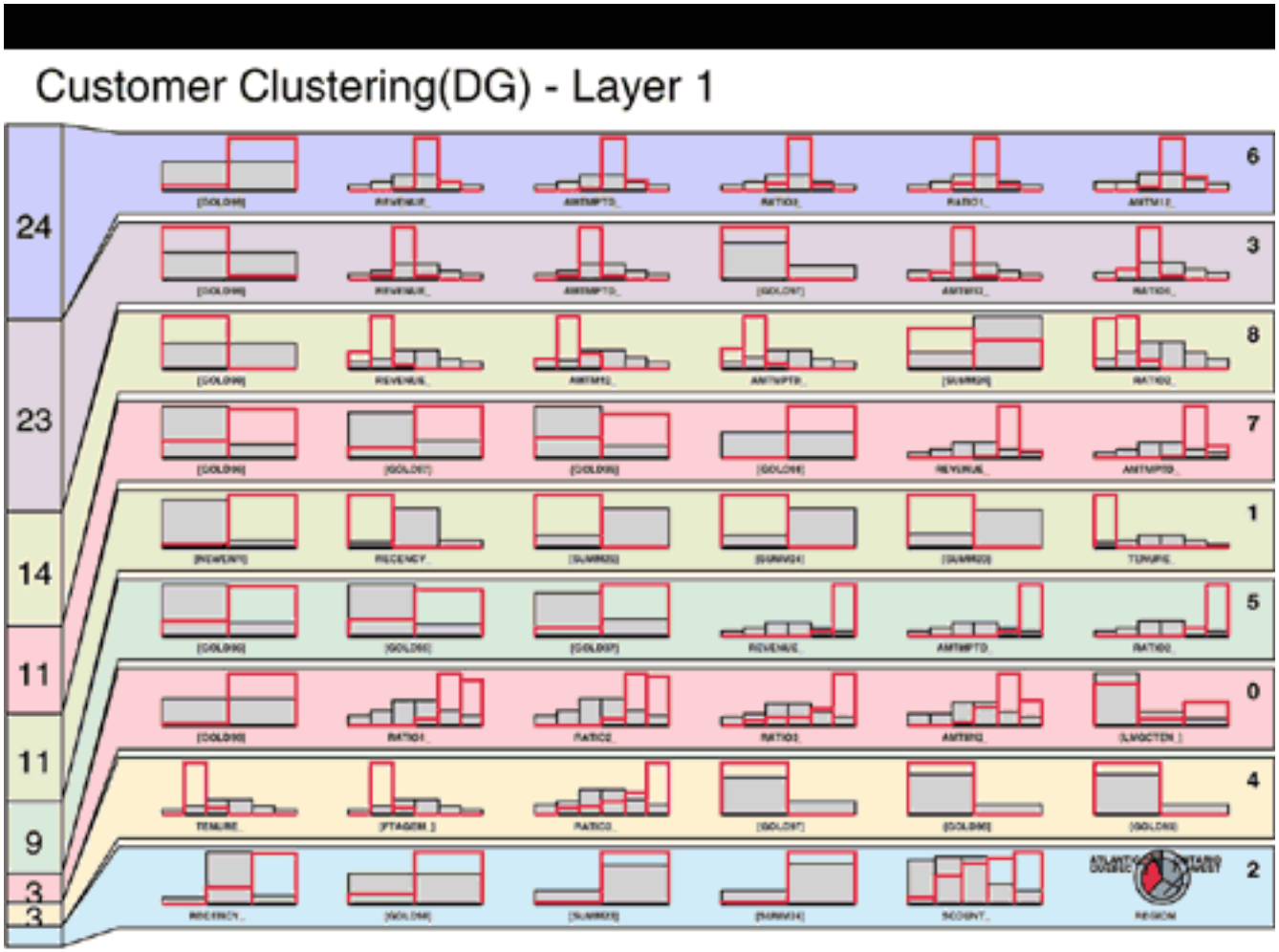


Figure 7. *Demographic clustering output.*

Analisi qualitativa dei cluster

- Gold98 è una variabile binaria che indica i clienti “migliori” nel DB
- Creata in precedenza dai business analyst usando l’analisi RFM (Recency, Frequency, Monetary)
- Il modello di clustering si accorda ottimamente con la definizione:
- La maggior parte dei cluster possiede o quasi tutti clienti Gold o quasi nessuno
- Il segmento Gold attuale è confermato!

Analisi qualitativa dei cluster

- Risultati del clustering
  - utili non solo per validare il concetto esistente di clienti Gold
  - estendono l'idea di cliente Gold creando cluster all'interno della categoria Gold98
  - esiste un gruppo di clienti platinum
- Cluster 6
  - quasi tutti clienti Gold98, per i quali il reddito, i punti totali ottenuti nel tempo,
    il profitto mensile, ed il tempo di vita
    sono tutti tra il 50° ed il 75° percentile

Analisi qualitativa dei cluster

- Cluster 3:
  - Nessun cliente Gold98
  - Il reddito, i punti ottenuti, ed il profitto mensile sono tutti tra il 25° ed il 50° percentile
- Cluster 5:
  - 9 % della popolazione
  - Reddito e punti ottenuti sono oltre 75°, distribuiti quasi tutti oltre il 90° percentile
  - Sembra un cluster estremamente “attraente”

Vista dettagliata del cluster 5

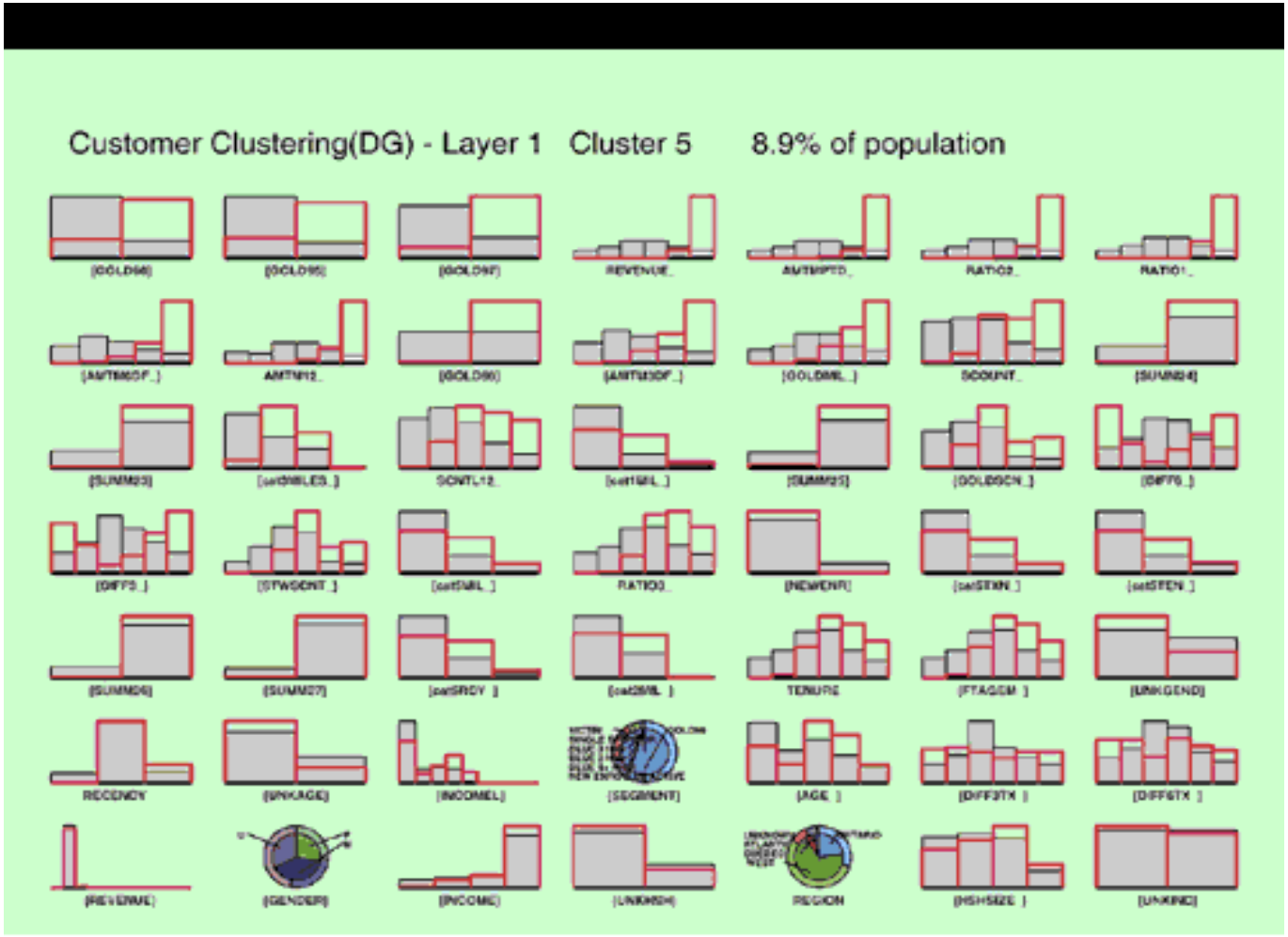


Figure 8. *Cluster 5 output.*

Profili dei cluster

- Scopo: scoprire il valore potenziale di ciascun cluster per il business, mostrando i valori aggregati delle variabili “shareholder value” per ogni cluster

CLUSTERID	REVENUE	CUSTOMERS	PRODUCT INDEX	LEVERAGE	TENURE
5	34.74%	8.82%	1.77	3.94	60.92
6	26.13%	23.47%	1.41	1.11	57.87
7	21.25%	10.71%	1.64	1.98	63.52
3	6.62%	23.32%	.73	.28	47.23
0	4.78%	3.43%	1.45	1.40	31.34
2	4.40%	2.51%	1.46	1.75	61.38
4	1.41%	2.96%	.99	.48	20.10
8	.45%	14.14%	.36	.03	30.01
1	.22%	10.64%	.00	.02	4.66

Table 1. *Profiling a cluster.*

Profili dei cluster

- leverage = rapporto tra profitto e # di clienti
- Il cluster 5 è quello potenzialmente più interessante
- Aumentando la profitability, aumenta anche il numero di prodotti acquistati
- product index = rapporto tra il numero medio di prodotti acquistati dai clienti all'interno del cluster ed il numero medio di prodotti acquistati in totale
- La profitability aumenta con il tempo di vita

Opportunità per il business

- I migliori clienti sono nei cluster 2, 5, e 7:
  - indicazione: mantenimento
- Cluster 2, 6, e 0:
  - indicazione: cross-selling in contrasto con i cluster 5 e 7
  - I cluster 2, 6, e 0 hanno un product index vicino a quello dei cluster 5 e 7, che però hanno il maggior numero di prodotti acquistati
  - Cercare di portare clienti dai cluster 2, 6, e 0 ai cluster 5 e 7
  - Confrontando quali prodotti sono acquistati possiamo trovare quali prodotti siano buoni candidati per il cross-selling

Opportunità per il business

- Cluster 3 e 4
  - indicazione: cross-selling verso i cluster 2, 6, e 0
- Cluster 1
  - indicazione: wait and see
  - sembra essere un gruppo di nuovi clienti
- Cluster 8
  - indicazione: non sprecare investimenti in marketing

Follow-up

- Reazioni da The Loyalty Group
  - visualizzazione dei risultati disponibile per un'analisi
  - validazione della tecnica originaria di segmentazione
    - però i raffinamenti alla segmentazione originaria possono risultare utili
  - decisione di intraprendere ulteriori progetti di Data Mining, tra cui:
    - modelli predittivi per direct mail targeting,
    - ulteriore analisi sulla segmentazione utilizzando dati comportamentali più dettagliati
    - Identificazione di opportunità usando algoritmi per la scoperta di regole associative all'interno dei segmenti individuati