Data Mining per il CRM
Customer segmentation

Dino **Pedreschi**
Pisa KDD Lab, ISTI-CNR & Univ. Pisa
http://www-kdd.isti.cnr.it/

MAINS, 23 marzo 2012

Data Mining Technologies for CRM

Clustering

Customer segmentation

Clustering

Problem definition

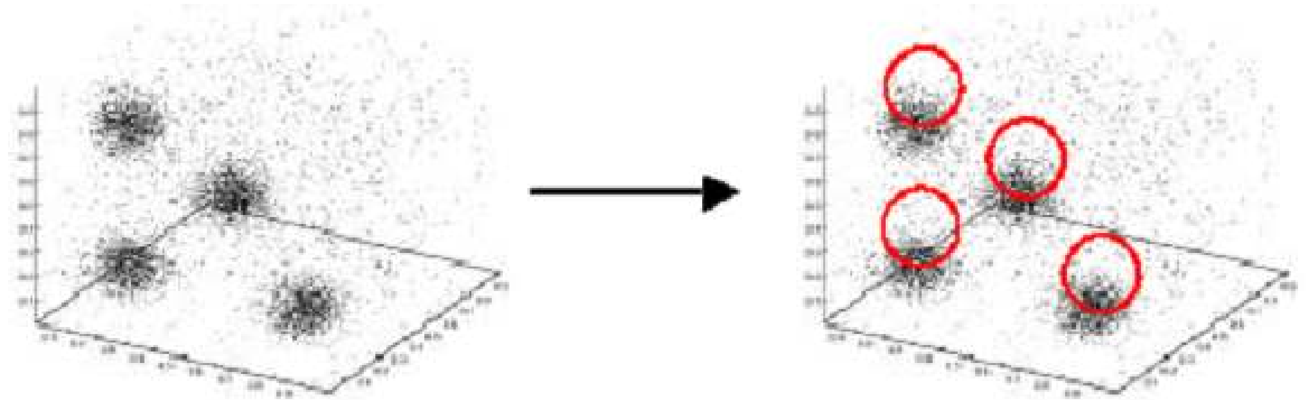
Problem definition

- Given:
  - A data set with *N* *d*-dimensional data items.

- Task:
  - Determine a natural partitioning of the data set into a number of clusters (*k*) and noise.
  - Clusters should be such that:
    - items in same cluster are similar
      ➔intra-cluster similarity is maximized
    - items from different clusters are different
      ➔inter-cluster similarity is minimized

Example – clustering in 3D

- Data: points in the 3D space
- Similarity: based on (Euclidean) distance



Application: customer segmentation

- Given:
  - Large data base of customer data containing their properties and past buying records
- Goal:
  - Find groups of customers with similar behavior

Case study: customer segmentation

AIR MILES
a case-study on customer segmentation

From: G. Saarenvirta, "Mining customer data", DB2 magazine on line, 1998

http://www.db2mag.com/98fsaar.html

Customer clustering & segmentation

- two of the most important data mining methodologies used in marketing
- use customer-purchase transaction data to
  - track buying behavior
  - create strategic business initiatives.
  - divide customers into segments based on "shareholder value" variables:
    - customer profitability,
    - measure of risk,
    - measure of the lifetime value of a customer,
    - retention probability.

Customer segments

- Example: high-profit, high-value, and low-risk customers
  - typically 10% to 20% of customers who create 50% to 80% of a company's profits
  - strategic initiative for the segment is retention
- A low-profit, high-value, and low-risk customer segment may be also attractive
  - strategic initiative for the segment is to increase profitability
  - cross-selling (selling new products)
  - up-selling (selling more of what customers currently buy)

Behavioral vs. demographic segments

- Within behavioral segments, a business may create demographic subsegments.
- Customer demographic data are not typically used together with behavioral data to create segments.
- Demographic (sub)segmenting is used to select appropriate tactics (advertising, marketing channels, and campaigns) to satisfy the strategic behavioral segment initiatives.

The Loyalty Group in Canada

- runs an AIR MILES Reward Program (AMRP) for a coalition of more than 125 companies in all industry sectors - finance, credit card, retail, grocery, gas, telecom.
- 60% of Canadian households enrolled
- AMRP is a frequent-shopper program:
  - the consumer collects bonuses that can then redeem for rewards (air travel, hotel accommodation, rental cars, theatre tickets, tickets for sporting events, …)

Data capture

- The coalition partners capture consumer transactions and transmit them to The Loyalty Group, which
- stores these transactions and uses the data for database marketing initiatives on behalf of the coalition partners.
- The Loyalty Group data warehouse currently contains
  - more than 6.3 million household records
  - 1 billion transaction records.

Before data mining

- The Loyalty Group has employed standard analytical techniques
  - Recency, Frequency, Monetary value (RFM) analysis
  - online analytic processing tools
  - linear statistical methods
- to analyze the success of the various marketing initiatives undertaken by the coalition and its partners.

Data mining project at AMRP

- Goal: create a customer segmentation using a data mining tool and compare the results to an existing segmentation developed using RFM analysis.
- data mining platform
  - DB2 Universal Database Enterprise parallelized over a five-node RS/6000 SP parallel system.
  - Intelligent Miner for Data (reason: has categorical clustering and product association algorithms which are not available in most other tools)

Data model

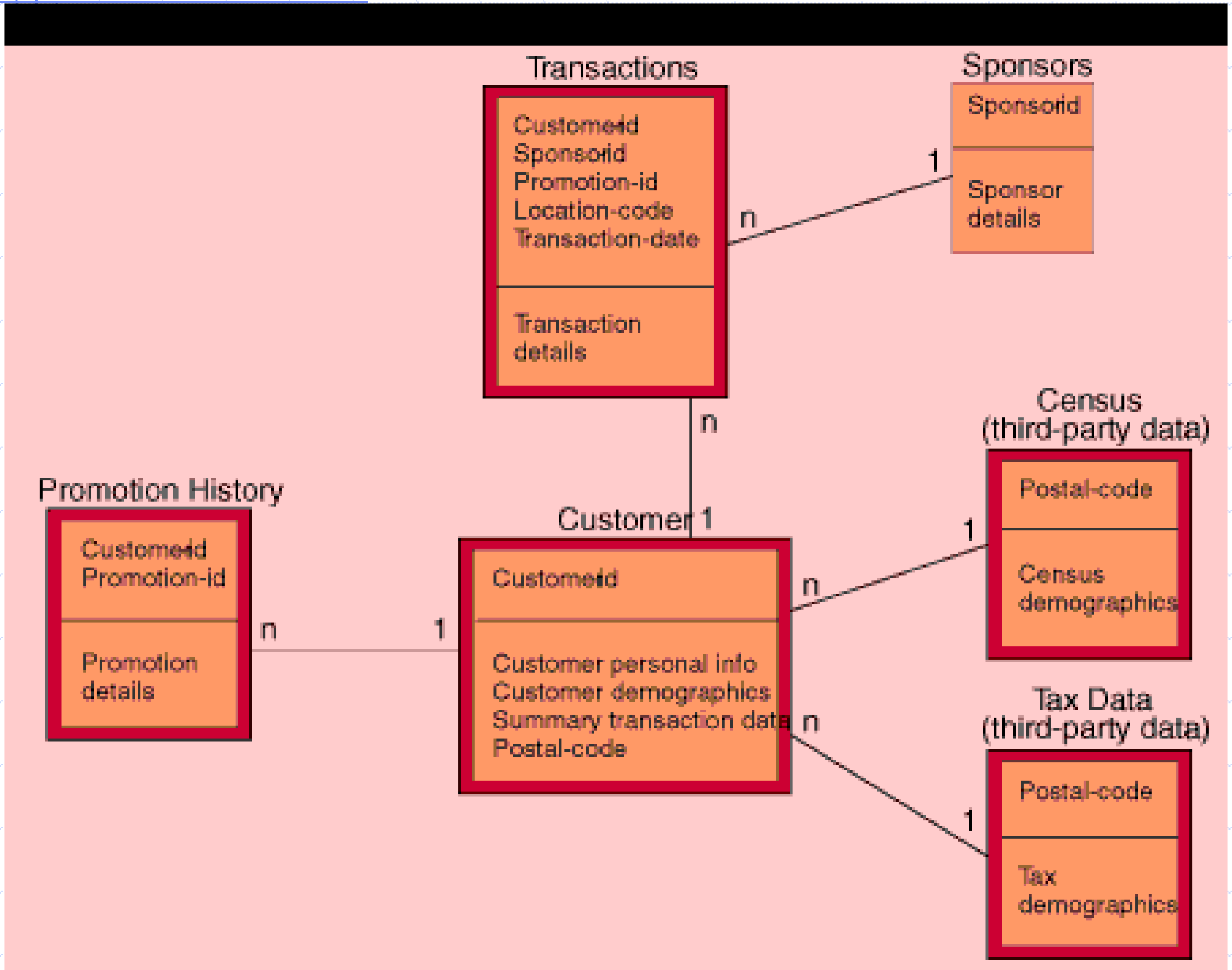


Figure 2. *AIR MILES case study data model.*

- ~ 50,000 customers and their associated transactions for a 12-month period.

Data preparation

- "shareholder value" variables
  - revenue
  - customer tenure
  - number of sponsor companies shopped at over the customer tenure
  - number of sponsor companies shopped at over the last 12 months,
  - recency (in months) of the last transaction
- calculated by aggregating the transaction data and then adding then to each customer record

Data preparation (2)

- Dataset obtained by joining the transaction data to the customer file to create the input for clustering algorithms
- 84 variables =
  - 14 categories of sponsor companies ×
  - 3 variables per category ×
  - 2 quarters (first two quarters of 1997)

Data cleansing - missing values

- demographic data
  - is usually categorical
  - has a high % of missing values
  - the missing values can be set to either unknown or unanswered (if result of unanswered questions)
- if a large portion of the field is missing, it may be discarded.
- In the case study, missing numeric values set to 0

Data transformation

- Ratio variables.
  - E.g.: profitability = profit / tenure
- Time-derivative variables.
  - E.g.: profit 2nd quarter - profit 1st quarter
- Discretization using quantiles.
  - E.g., break points at 10, 25, 50, 75, and 90.
- Discretization using predefined ranges.
  - E.g., those used in census
- Log transforms.
  - E.g., for very skewed distributions

Distribution of original data

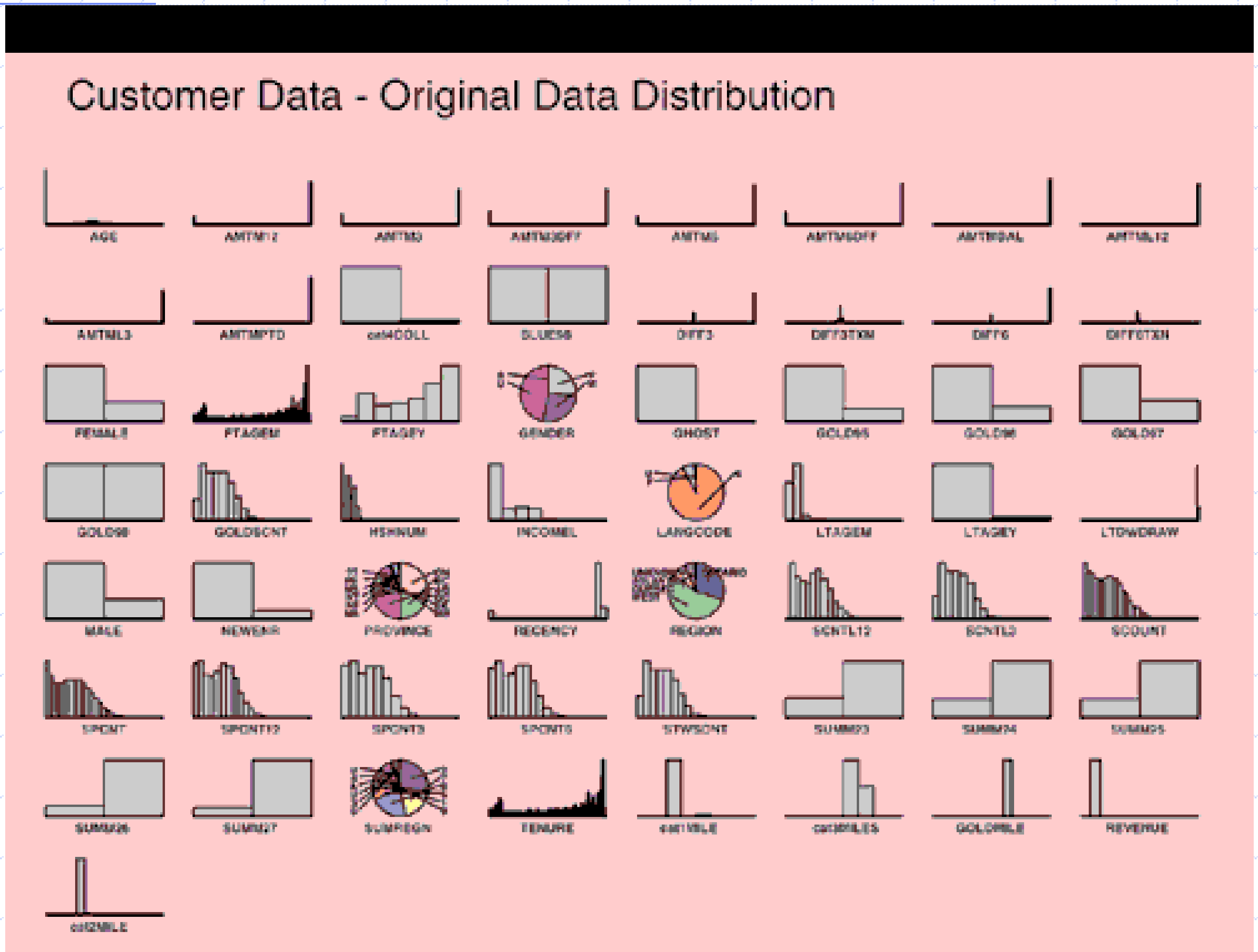


Figure3. *Original data.*

Distribution of discretized data

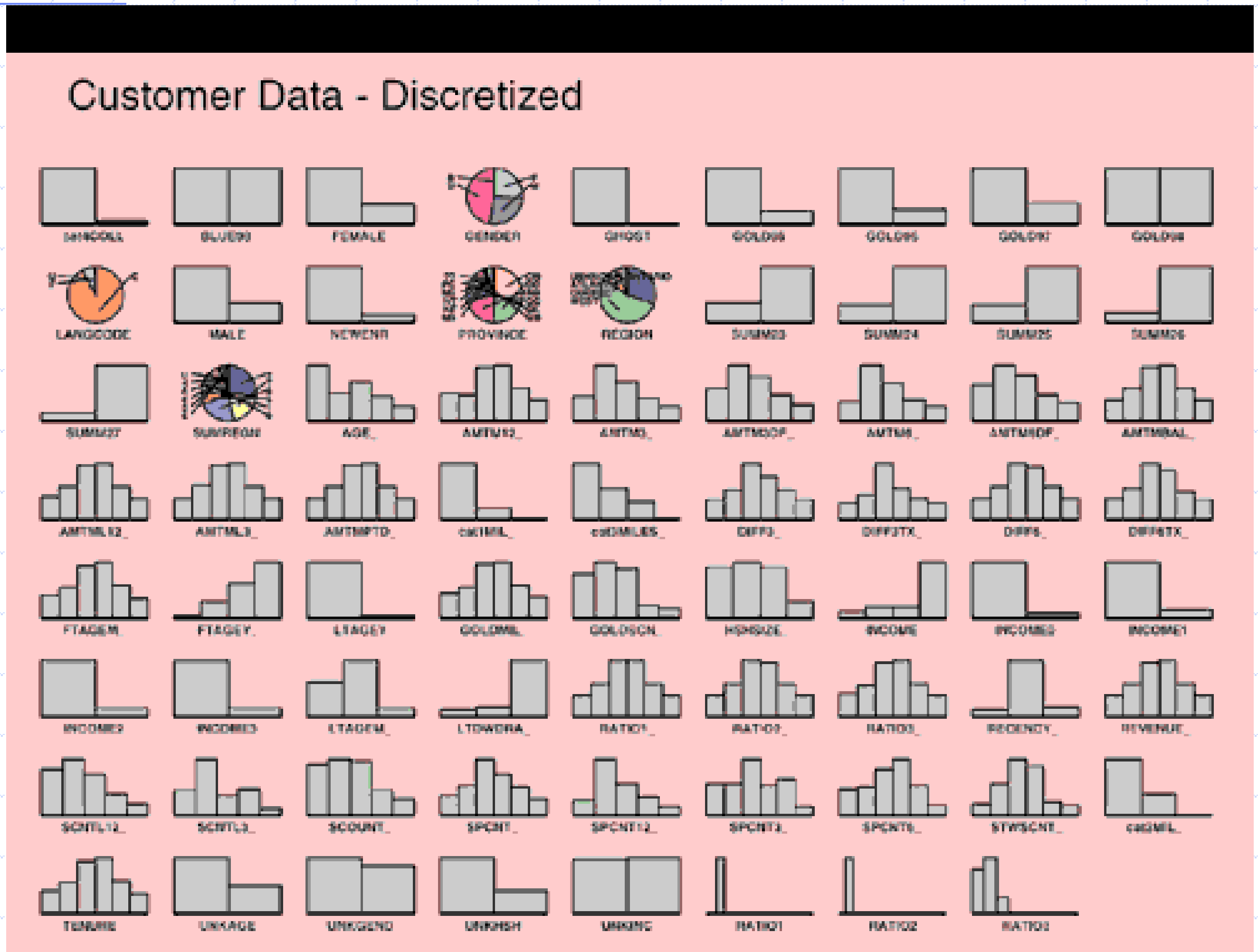


Figure 4. *Discretized data.*

Before/after discretization

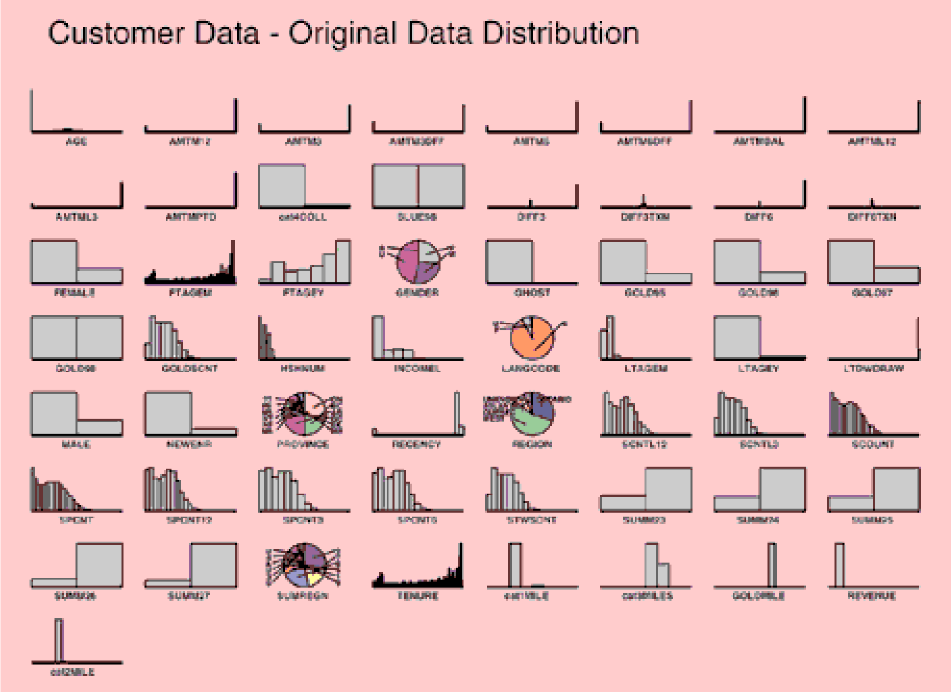


Figure3. *Original data.*

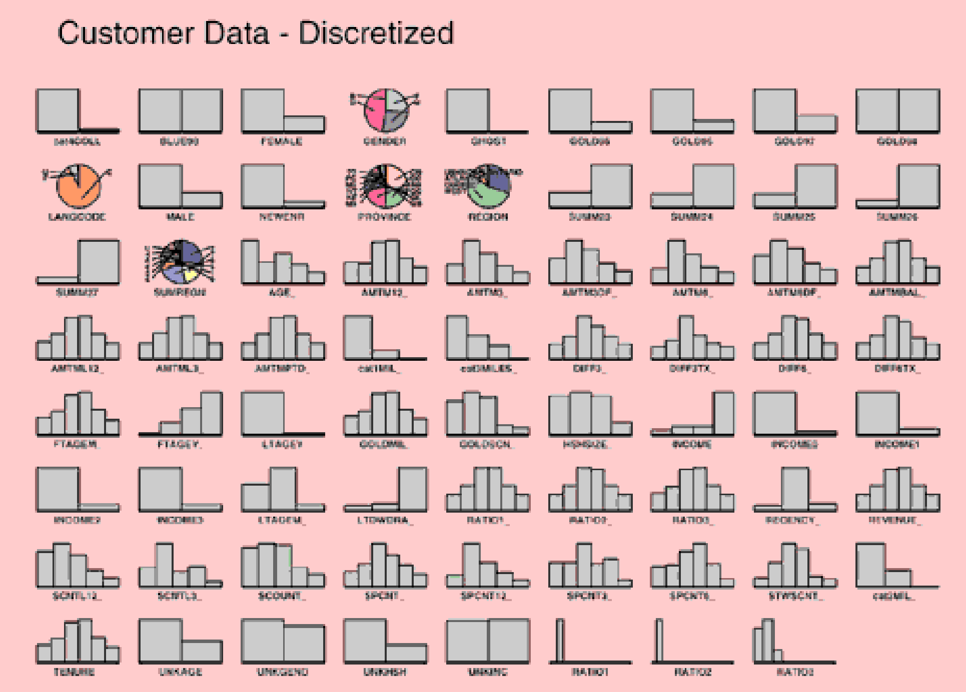


Figure 4. *Discretized data.*

Clustering/segmentation methodology

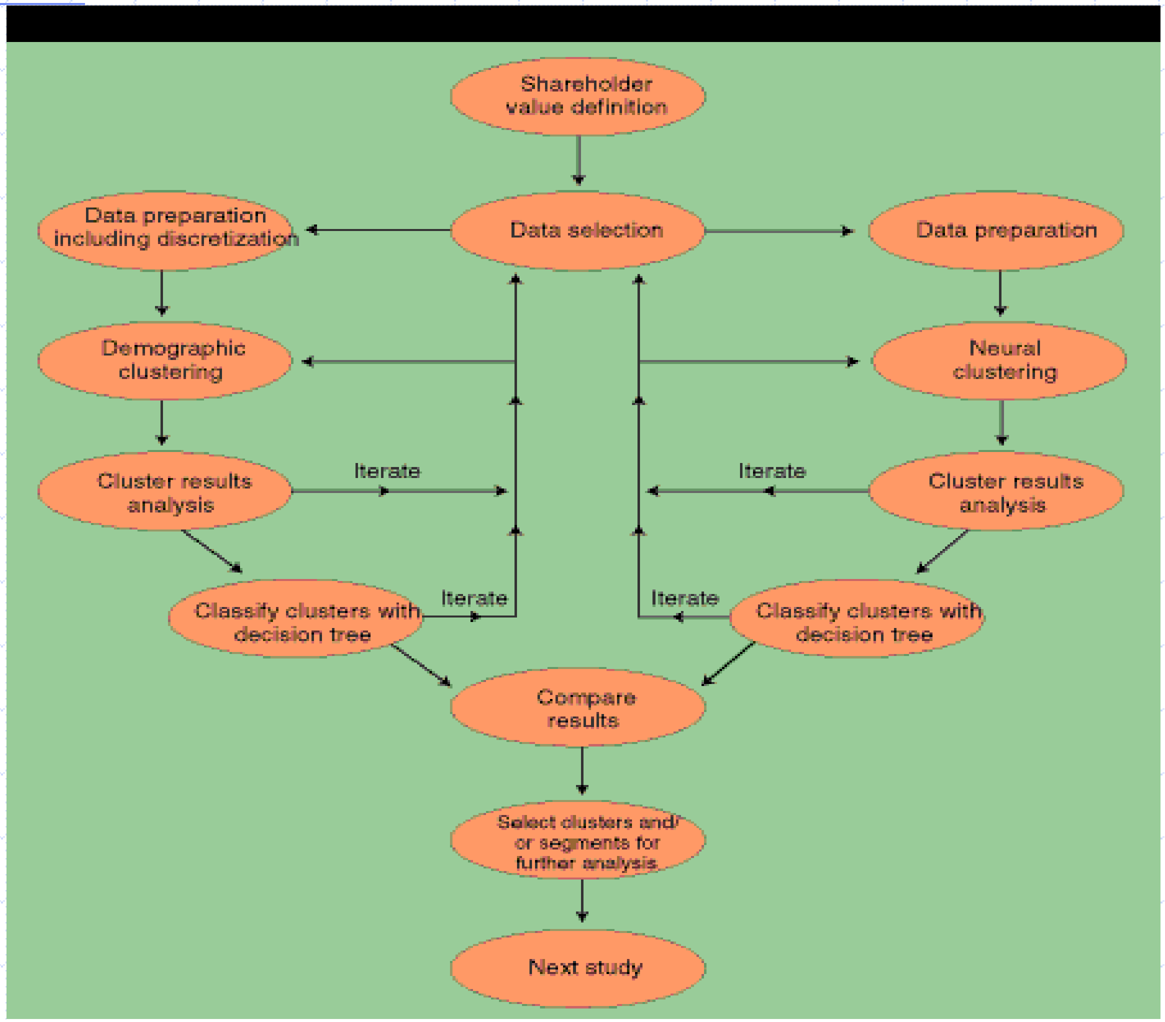
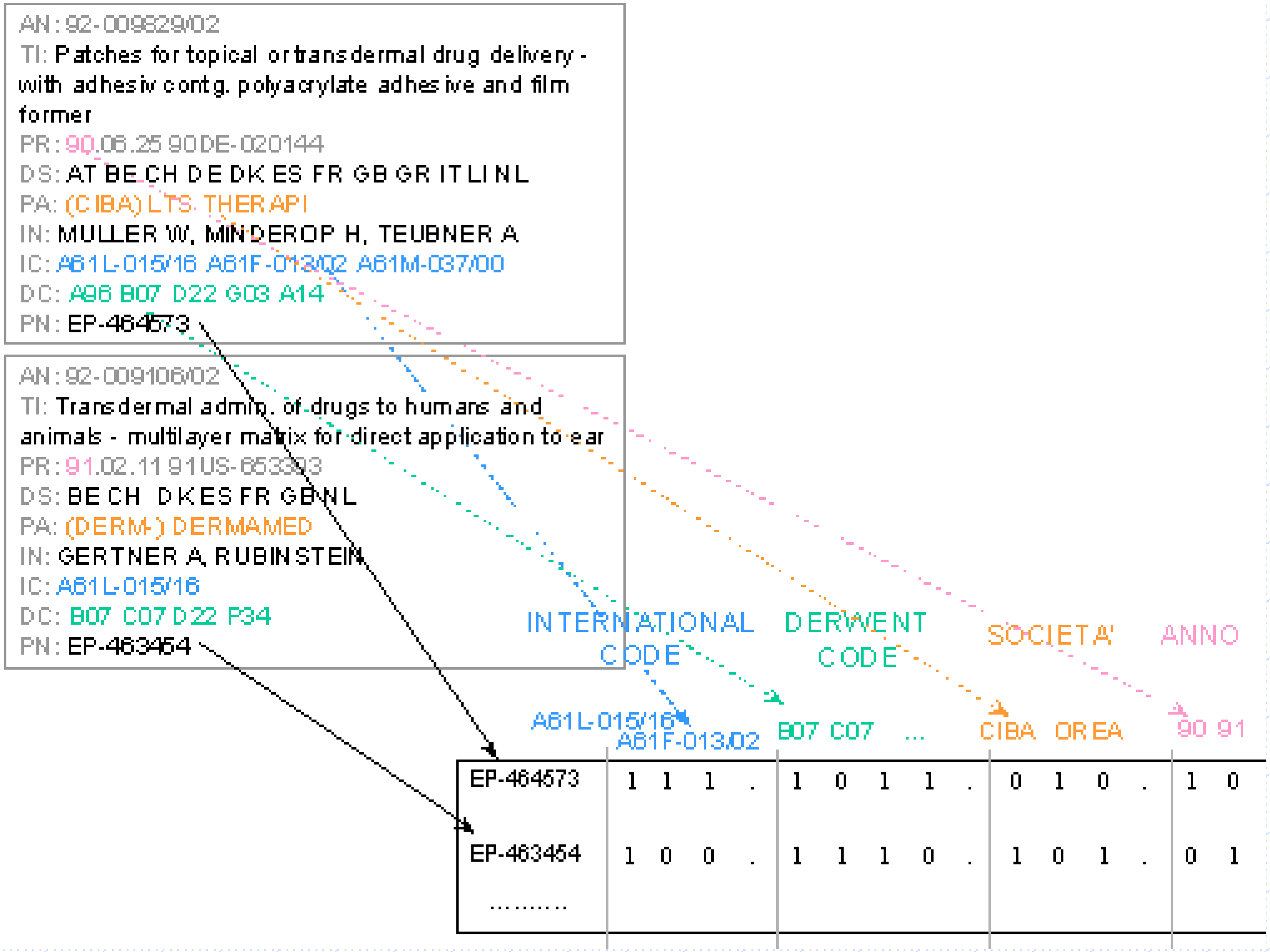


Figure 6. *Clustering workflow.*

IBM-IM demographic clustering

- Designed for categorical variables
- Similarity index:
  - increases with number of common values on same attribute
  - decreases with number of different values on same attribute
- # of clusters is not fixed a priori
  - only upper bound set

Demographic clustering: data structure



Demographic clustering: parameters

	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

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

  - **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

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

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

Demographic clustering: similarity index

- proportional to 1-1
- inversely proportional to 0-1 and 1-0
- unaffected by 0-0
- Condorcet index=
  - $N_{11} / (N_{11} + {}^{1}/_{2}(N_{01} + N_{10}))$
- Dice index=
  - $N_{11} / (N_{11} + {}^{1}/_{4}(N_{01} + N_{10}))$
- Dice looser then Condorcet
  - appropriate with highly different objects

Demographic clustering: similarity index

- Similarity threshold α
  - i,j assumed similar if $s(i,j)>\alpha$
  - low values (<0.5) appropriate with highly different objects

- Weights for attributes
  - importance of attributes in the similarity index may be varied with different weights
  - default weight = 1

IM Demographic clustering

- basic parameters:
  - Maximum number of clusters.
  - Maximum number of passes through the data.
  - Accuracy: a stopping criterion for the algorithm. If the change in the Condorcet criterion between data passes is smaller than the accuracy (as %), the algorithm will terminate.
  - The Condorcet criterion is a value in [0,1], where 1 indicates a perfect clustering -- all clusters are homogeneous and entirely different from all other clusters

… more parameters

- Similarity threshold.
  - defines the similarity threshold between two values in distance units.
  - If the similarity threshold is 0.5, then two values are considered equal if their absolute difference is less than or equal to 0.5.
- In the case study:
  - maximum # of clusters: 9
  - maximum # of passes: 5
  - accuracy: 0.1

Input dataset

- dataset: all continuous variables discretized.
- input variables :
  - # of products purchased over customer's lifetime
  - # of products purchased in the last 12 months
  - Customer's revenue contribution over lifetime
  - Customer tenure in months
  - Ratio of revenue to tenure
  - Ratio of number of products to tenure
  - Region
  - Recency
  - Tenure (# of months since customer first enrolled in the program).

Input dataset

- Other discrete and categorical variables and some interesting continuous variables were input as supplementary variables:
- variables used to profile the clusters but not to define them.
- easier interpretation of clusters using data other than the input variables.

Output of demographic clustering

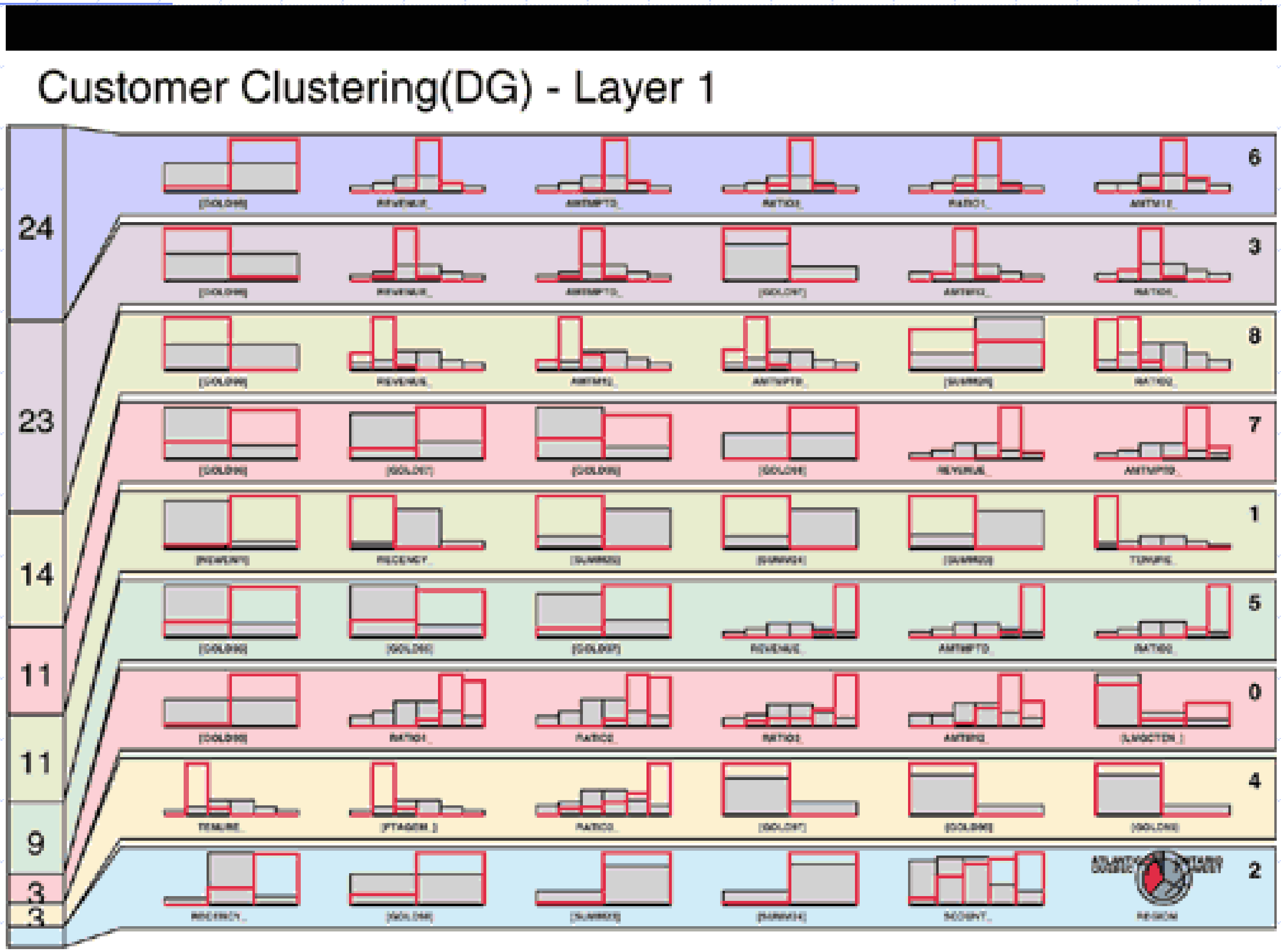


Figure 7. *Demographic clustering output.*

Visualization of clusters

- horizontal strip = a cluster
- clusters are ordered from top to bottom in order of size
- variables are ordered from left to right in order of importance to the cluster, based on a chi-square test between variable and cluster ID.
- other metrics include entropy, Condorcet criterion, and database order.

Visualization of clusters

- variables used to define clusters are without brackets, while the supplementary variables appear within brackets.
- numeric (integer), discrete numeric (small integer), binary, and continuous variables have their frequency distribution shown as a bar graph.

- red bars = distribution of the variable within the current cluster.
- gray solid bars = distribution of the variable in the whole universe.

Visualization of clusters

- Categorical variables are shown as pie charts.
- inner pie = distribution of the categories for the current cluster
- outer ring = distribution of the variable for the entire universe.

- The more different the cluster distribution is from the average, the more interesting or distinct the cluster.

Output of demographic clustering

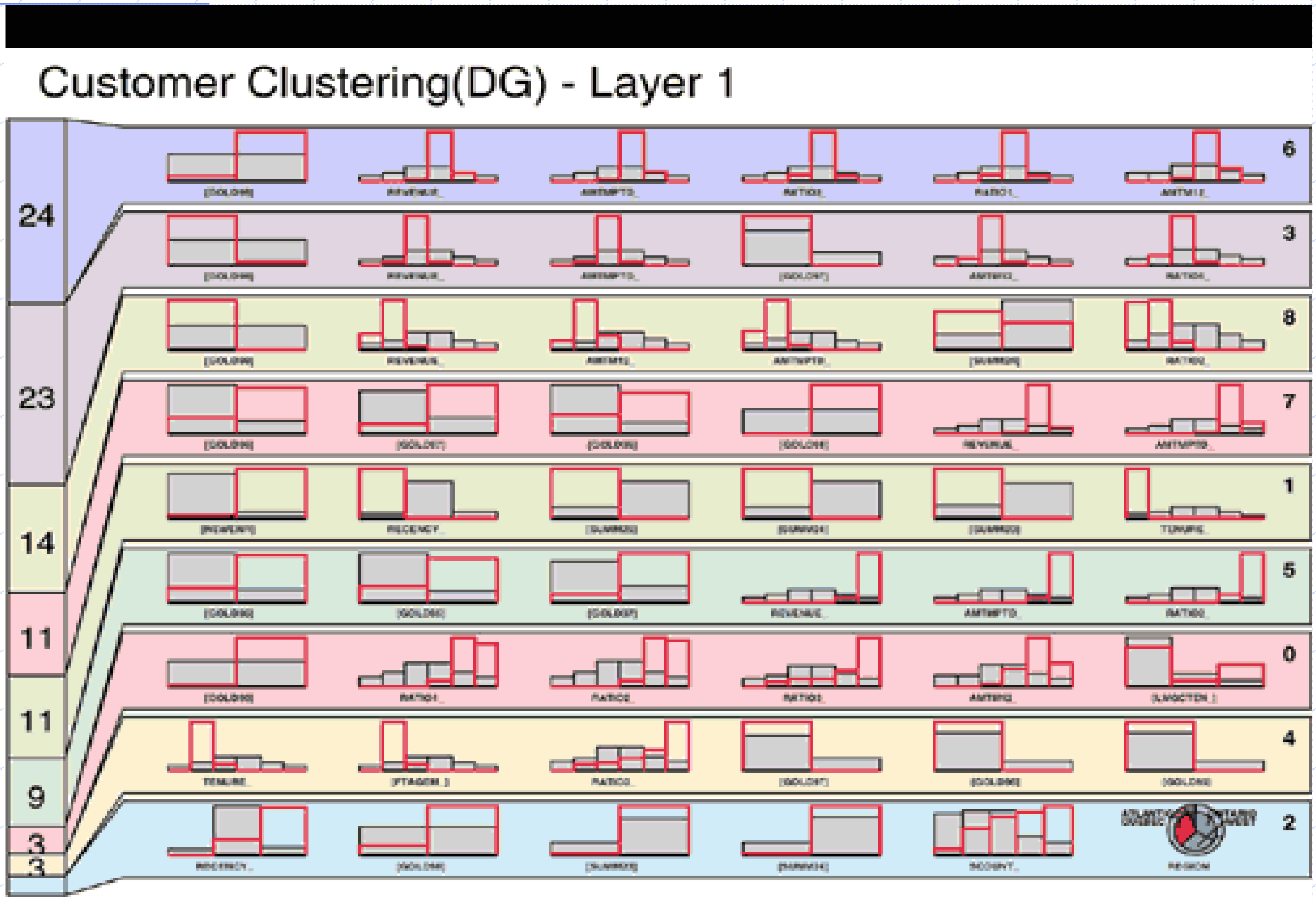


Figure 7. *Demographic clustering output.*

Qualitative characterization of clusters

- Gold98 is a binary variable that indicates the best customers in the database, created previously by the business using RFM analysis.
- The clustering model agrees very well with this existing definition: Most of the clusters seem to have almost all Gold or no Gold customers.
- Confirmed the current Gold segment!

Qualitative characterization of clusters

- Our clustering results
  - not only validate the existing concept of Gold customers,
  - they extend the idea of the Gold customers by creating clusters within the Gold98 customer category.
  - A platinum customer group
- Cluster 5
  - almost all Gold98 customers, whose revenue, bonus collected lifetime to date, revenue per month, and lifetime to date per month are all in the 50th to 75th percentile.

Qualitative characterization of clusters

- Cluster 3:
  - no Gold98 customers. Its customer revenue, bonus collected, revenue per month, are all in the 25th to 50th percentile.
- Cluster 5:
  - 9 %of the population.
  - revenue, bonus collected are all in the 75th percentile and above, skewed to almost all greater than the 90th percentile.
  - looks like a very profitable cluster

Detailed view of cluster 5

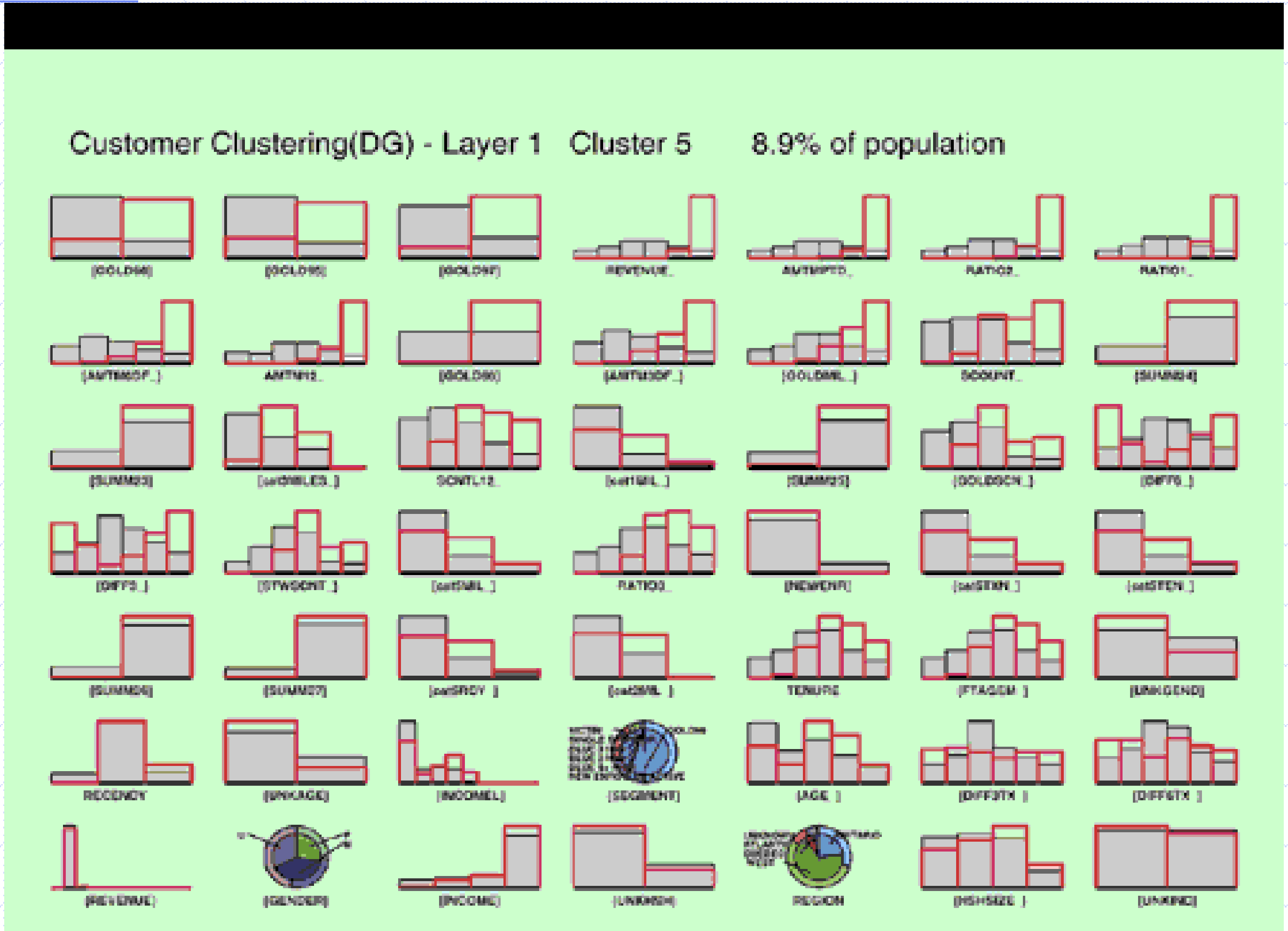


Figure 8. *Cluster 5 output.*

Profiling clusters

- Goal: assess the potential business value of each cluster quantitatively by profiling the aggregate values of the shareholder value variables by 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.*

Profiling clusters

- leverage = ratio of revenue to customer.
- cluster 5 is the most profitable cluster.
- as profitability increases, so does the average number of products purchased.
- product index = ratio of the average number of products purchased by the customers in the cluster divided by the average number of products purchased overall.
- customer profitability increases as tenure increases.

Business opportunities

- Best customers in clusters 2, 5, and 7. :
  - indication: retention
- clusters 2, 6, and 0
  - indication: cross-selling by contrasting with clusters 5 and 7.
  - Clusters 2, 6, and 0 have a product index close to those of clusters 5 and 7, which have the highest number of products purchased.
  - Try to convert customers from clusters 2, 6, and 0 to clusters 5 and 7. By comparing which products are bought we can find products that are candidates for cross-selling.

Business opportunities

- Clusters 3 and 4
  - indication: cross-selling to clusters 2, 6, and 0 •
- Cluster 1
  - indication: wait and see. It appears to be a group of new customers
- Cluster 8
  - indication: no waste of marketing dollars

Follow-up

- Reactions from The Loyalty Group
  - visualization of results allowed for meaningful and actionable analysis.
  - original segmentation methodology validated, but that refinements to the original segmentation could prove valuable.
  - decision to undertake further data mining projects, including
    - predictive models for direct mail targeting,
    - further work on segmentation using more detailed behavioral data,
    - opportunity identification using association algorithms within the segments discovered.