BEHIND MARKETING DATA IN FINANCIAL INDUSTRY

Behavioural Analysis of Retail Customers:
A Case Study from Banca Commerciale Italiana

Flavio Addolorato - Luca Bodio - Luigi Ferrari - Luigi Guastalla
Banca Commerciale Italiana

Abstract: *the objective of this work is to test the use of several analytical instruments ("data mining tools") in order to pinpoint some of the behavioural characteristics of the bank's retail clientele* **.**

DEFINITION OF SAMPLE

The sample was constructed in accordance with two stratification variables: the combinations of products held by retail customers ("Current Accounts", "Savings Passbooks / Certificates of Deposit", "Loans" and "Investments"), and their distribution over five geographic areas.

The data were sourced from the corresponding product files. We then reclassified the data on the basis of a "client-physical person" logic, with the use of the General Customer Number (GCN) assigned to each customer; i.e. the client is the logical unit of every analysis.

In the case of joint account relationships, we opted to duplicate the products utilized for each of the individual GCNs involved, treating each GCN as an individual customer.

The following files were considered:

Customer information	(103,267 GCN)
Current account statistics	(40,427)
Passbook savings statistics	(32,931)
Credit card statistics	(30,290)
Securities statistics	(51,551)
Loan statistics	(30,672)
Insurance statistics	(3,816)
Certificate of deposit statistics	(33,736)

The extraction of data was made with reference to 31 December 1995. The data in each file were relative to the July-December 1995 period, and were considered as both stock and flow information.

We attempted to cover the entire analytical process **-** from the extraction of the data to the identification of the selling tools **-** in order to highlight the potential of the statistical tools and any related implementation problems.

In particular, we limited our examination to holders of **ordinary current accounts**, and these customers thus represent the universe of reference for our analyses. We also believed it was important to use only fully significant data, and that, in our opinion, has the greatest explicatory power.

Having identified the GCNs through the Current Account File, we reclassified the products used by any individual customer. We thus obtained a sample of 31,110 individuals which, on the basis of the "weights" assigned for the initial extraction of data, are representative of just under 900,000 customers overall.

ANALYSES

Following are the analyses carried out using the data sample identified above.

FACTORIAL ANALYSIS

The factorial analysis is aimed at the research and identification of the structure (the factors) underlying a group of observed variables.

In order to describe the behaviour of our customers, we initially relied only on the available current accounts data, as well as on information regarding the securities deposits linked to the current account themselves.

The following variables were considered for the accounts held by each customer:

- average age of the relationship
- average balance
- number of cheques issued
- number of cheques paid
- number of outgoing bank transfers (debits)
- number of incoming bank transfers (credits)
- number of withdrawals
- number of deposits.

Other data considered for securities deposit accounts included:

- average age of the relationship
- average balance
- number of buy / sell transactions.

Within this phase, we were able to move from the original variables, as initially reported, to six standardized and non-correlated factors which are still highly representative of the variance of behaviours.

In the next table we report the normalized factor loadings (their value ranges from -1 to +1 depending on their importance) obtain from the analisys. To give a sketch of their relevance we only reported those which scored the highest.

Using the table it is possible to spot the economic content represented by each factor. The first factor seems to represent the typical transaction enacted by the layman as opposed to the second one which refers to transaction of the professional customer. The next two factors are instead linked to the fidelity of the customer (the age of the relationship) and to his wealth. Finally the last two factors are deemed to represent the frequency and the dimension of the transaction.

	FACTOR1	FACTOR2	FACTOR3
average age of the relationship			0.850
average account balance			
number of cheques issued	0.503	0.591	
number of cheques paid	0.820		
number of outgoing bank transfers (debits)		0.695	
number of incoming bank transfers (credits)		0.794	
number of withdrawals			
number of deposits	0.850		
average age of securities			0.719
average balance - securities			
number of buy / sell transactions			

	FACTOR4	FACTOR5	FACTOR6
average age of the relationship			
average account balance	0.741	0.355	
number of cheques issued			
number of cheques paid			
number of outgoing bank transfers (debits)			
number of incoming bank transfers (credits)			
number of withdrawals			0.985
number of deposits			
average age of securities			
average balance - securities	0.756	0.339	
number of buy / sell transactions		0.874	

Note: The values less than 0.3 are indicated by '.'.

CLUSTER ANALYSIS

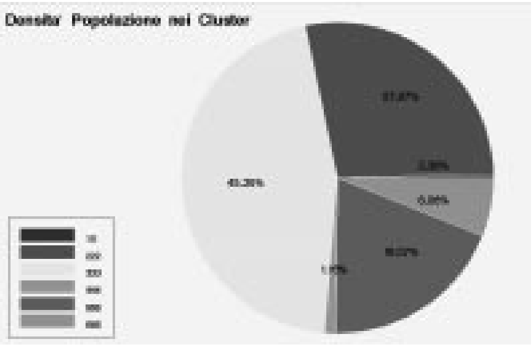
With the framework of marketing studies, cluster analyses are used for:

- segmentation analyses, as a tool for automatic classification;
- the development and research of markets for new products - in determining competitive niches within broader market structures by classifying brand names, products and customer needs;
- the choice of test-market areas. This application regards the identification of relatively standard market areas (and not only geographic areas), so that a generalization of the results obtained relative to a test area can be applied to the remaining data belonging to the same cluster.

On the basis of the factors set forth above, a first subdivision of the population in clusters evidenced 16 customers whose behaviour was extremely anomalous. After appropriately verifying such behaviour, these customers were no longer taken into consideration for the rest of the analysis .

The remainder of the population was subdivided again, with another cluster analysis. As a result, the sample was divided into six behavioural clusters (segments), as indicated below:

Cluster	Frequency	Standard Deviation	Nearest Cluster	Distance between Cluster Centroids
111	174	2.3662	3	8.8049
222	7106	0.6156	3	1.9538
333	13924	0.4511	5	1.7614
444	275	1.7973	2	4.8560
555	7177	0.6495	3	1.7614
666	2438	0.8796	3	3.0143



As indicated, the size of the clusters varies significantly. Taken together, the clusters 111 and 444 account for less than 2% of the population, while cluster 333 represent 45%. This disparity suggests that yet another segmentation could be feasible.

CORRESPONDENCE ANALYSIS

The correspondence analysis is a multivariate analytical technique which allows for representing relationships existing between various qualitative variables on a two-dimensional plane. (In this case, the variables are the clusters, and the information being used to describe them.)

In order to describe the clusters obtained, we went back to the original sample data (and not only information on current accounts and securities deposit accounts), and we re-classified the data, assigning codes as indicated in the data record below.

In detail, the data were classified in the following categories:

- **CUSTOMER INFORMATION**

age

0- 24	= ETA0024
25 - 35	= ETA2535
36 - 45	= ETA3645
46 - 60	= ETA4660
oltre 60	= ETAOV60

geographic area

northwest	= A
northeast	= B
central	= C
south and islands	= D
large cities	= E

- **CURRENT ACCOUNT STATISTICS**

average balance for the period

overdraft	= ROSSO
0 - 10 mn lire	= VERDE
10-34 mn lire	= ALTO
more than 35 mn lire	=ALTO++

total number of transaction for the period

none	= 0000
up to 24	= NNNN
more than 24	= !!!!

- **SECURITIES DEPOSIT STATISTICS**

value of portfolio

less than 20 mn lire	= TIT--
greater than 20 mn lire	= TIT++

tot. num. of transaction for the period

up to 2 transactions per 6 months	= TIT||,
more than 2 transaction per 6 months	= TIT**

- **CARD STATISTICS**

client with no cards	= NOBM
Bancomat:	
less than 6 transactions	= BM
6 or more transactions	= BM++
POS user	= POS
cartaSi user	= CSI

- **PASSBOOK SAVING STATISTICS**

balance < 10 mn lire and transaction < 3	= LRO
balance >= 10 mn lire and transaction < 3	= LROS
balance < 10 mn lire and transaction >= 3	= LRO-M
balance >= 10 mn lire and transaction >= 3	= LROSM

- **INSURANCE STATISTICS**

users of AssiBa products	= ASO

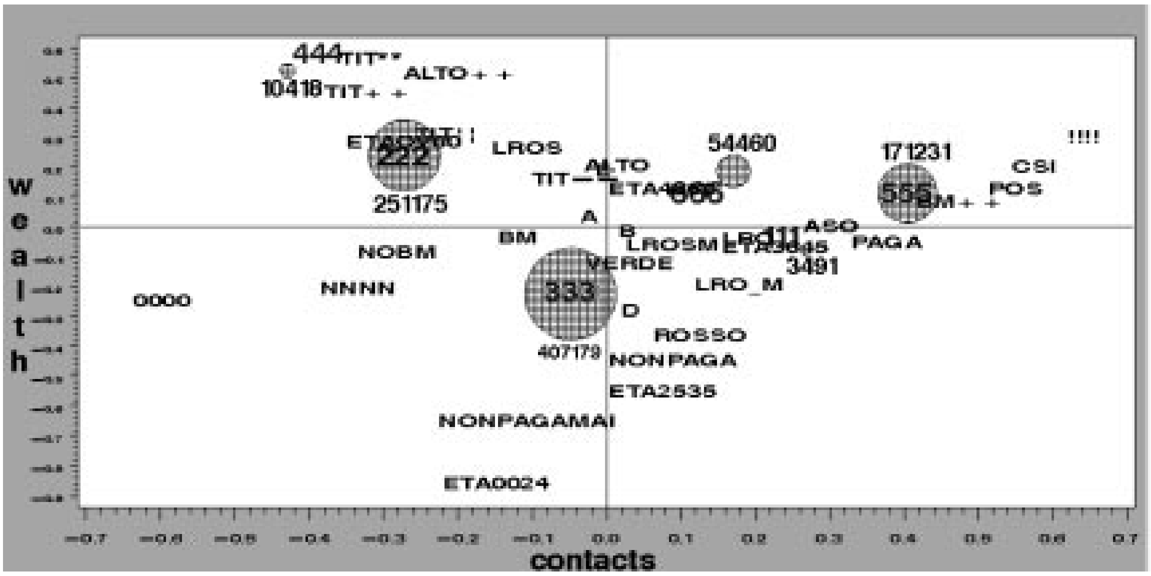
- **LOAN STATISTICS**

repayments on schedule	= PAGA,
up to two payments past due	= NONPAGA,
more than two payments past due	= NONPAGAMAI

The results of the analysis indicated the possibility to represent some 88% of the data with only two variables (see table below).

```
                           Inertia and Chi-Square Distribution

     Exact       Principal     Chi-Square       Perc.         10    20    30    40    50
    Values.       Inertia                                 ----+----+----+----+----+-
  Principal


    0.24202      0.05857        21855.4        52.11%     **************************

   0.20271      0.04109        15331.8        36.55%     ******************

   0.08560      0.00733        2734.22         6.52%     ***

   0.07032      0.00494        1845.09         4.40%     **

   0.02178      0.00047        177.042         0.42%

     TOTAL       0.11241        41943.6          Degrees of Freedom = 200
```

Using the data above, we were able to come up with the following diagram:



The diagram highlights the six clusters with the number of customers included therein. The behavioural classification symbols are also reported.

At first sight, it would seem that the horizontal axis is strongly correlated to the number of "contacts" (in a broad sense - in other words, the number of products used, the number of transactions carried out, etc.). Meanwhile, the vertical axis would seem to express a dimension of customer "wealth" (in other words, the level of the assets of the customers placed with the bank).

Furthermore, when considering the age variable, it would seem possible to identify a "customer life-cycle" trend on the diagram. In order to verify such an assumption, an analysis of the migratory flows between clusters over time would be necessary, but this would require a sample extracted from a subsequent period.

The positioning of the clusters, instead, may be read in the following way:

The cluster 444 is made up of wealthy, older customers who do not use POS terminals or Bancomat. On the other hand, these customers tend to have passbook savings accounts, and high volumes of securities transactions.

The cluster 222 is very similar to 444, but there is less volume in securities transactions, and the levels of wealth are on average lower.

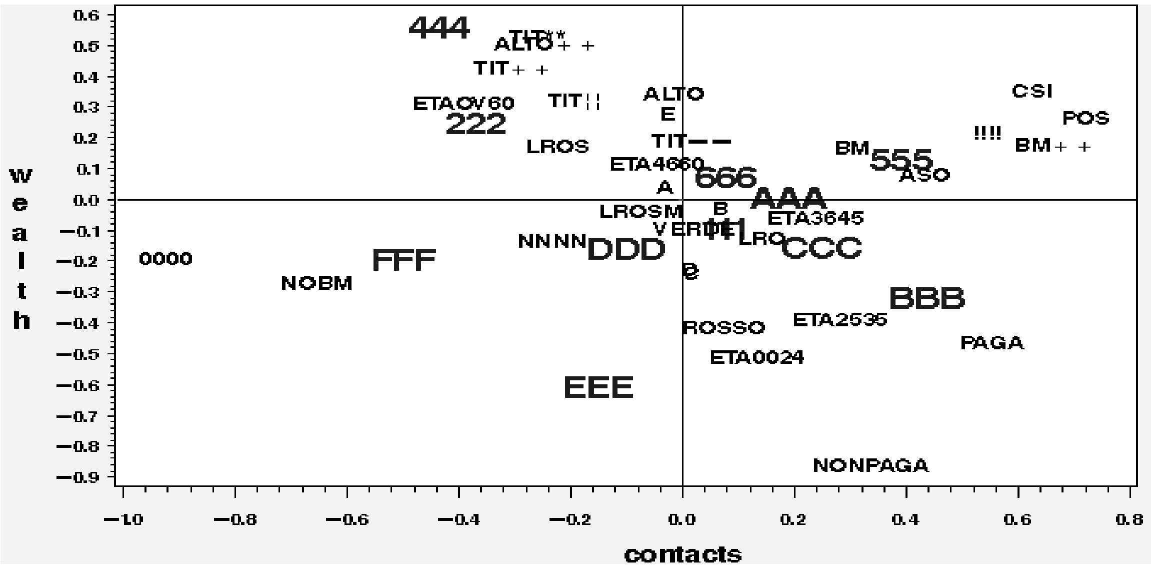
Cluster 555 consists of individuals who have very active relationships with the bank, and whose age ranges between 36 and 60. The level of wealth is considered average.

Cluster 111 is very similar to 555, though the level of wealth is somewhat lower. The customer relationships in this case also include a high percentage of passbook savings accounts.

Cluster 666 represents a group of customers who are very active with the bank, and who have a sound economic situation.

There are no particular traits evident at first glance with final cluster, **333**. If anything, the group appears to consist of younger customers and customers having greater difficulty in repaying loans. Considering the magnitude of the group, this cluster warrants more in-depth examination (Attachment 1).

The new segmentation of cluster 333 was appended to the overall sample. The new segmentation was used for another run of correspondence analysis which yielded the diagram below.



An examination of the graph above indicates the following:

- the axes have the same significance as in the previous analysis;
- in an outcome similar to what occurred with the first analysis, the new clusters are positioned in the lower quadrants where “wealth” is below the average for the entire population.

In what follows we report some of the findings resulting from a comparison of each individual cluster with the overall cluster 333:

The cluster **333 / AAA** is a younger cluster than the others. Confirming its position on the graph, we note: a smaller number of overdrafts and dormant accounts as well as the total absence of loans; furthermore, the customers in this cluster have bank cards which are used with a frequency that is almost double the frequency reported for cluster 333. This applies to all three bank cards (Bancomat, CartaSi' and POS).

Cluster **333 / BBB** is also young, with current-account balances and securities deposit accounts which are lower than the average of cluster 333. The clientele in this cluster uses bank loans and has no great difficulty in meeting the repayment schedule. The entire cluster has bank cards which are used with a frequency that is almost double. This applies to all three bank cards (Bancomat, CartaSi' and POS).

The customers belonging to cluster **333 / CCC** all have an insurance product, and have activity in securities which is greater than the activity represented by cluster 333. Passbook savings accounts are widespread; there is generally a good transaction volume in the current accounts, and dormant accounts are almost non-existent. The possession and use of Bancomat is in line with the average for cluster 333, but the credit cards and POS are used to a greater extent.

Cluster **333 / DDD** is the one which is closest to the original. The capital position of the customers is better than average, while the bank cards are used and the age is somewhat high.

Cluster **333 / EEE** is a credit cluster in which all customers have loans. There is little activity in the other products. Balances in current accounts and securities deposits are well below the average, and there are no bank cards at all. This age is young, and the group is mainly concentrated in the southern part of the country.

Cluster **333 / FFF** consists of customers having only current accounts, passbook savings accounts and securities deposit accounts (in line with cluster 333). The percentage of dormant accounts is high; there are no bank cards, and the age is high.

MARKETING POLICY IMPLICATIONS OF THE RESULTS

Since these analyses have segmented customers by their behaviour, the practical application of the results can be seen in the following areas.

- **Instruments to supplement the sales effort.** The possibility of classifying each customer profile within a cluster is a situation which could provide the banker with an immediate vision of the areas for possible development (Attachment 1).
- **Creation of customer portfolios.** Upon identifying the clusters, it would be possible to assign commercial customer portfolios to bank managers in accordance with the managers' professional credentials.
- **Targeting.** A well established behavioural segmentation of customers should allow the bank to improve the effectiveness of further analyses: socio-demographic and geographic analyses; customer life-cycle analyses; the analyses of migration between clusters (for example, to verify results of special marketing campaigns); definition of test areas (for launching new products); and identification of data categories used in other analyses (credit-scoring analyses).
- **Cross-selling of products, and efforts to enhance customer loyalty**. With an average behavioural profile (the potential) of each cluster, the bank can differentiate marketing efforts, thus offering the mix of products deemed most appropriate.

In our example the data exploratory analysis may be read in a suggestive way along the following lines:

111 Tradition *These customers probably appreciate a personalized relationship with a branch, while it does not seem they would be a well suited target for high-technology products.*

222 Wealth *The customers in this group would appear to be natural candidates for* *institutionally managed savings products; in this case, the bank could also propose some technologically sophysticated services (home banking).*

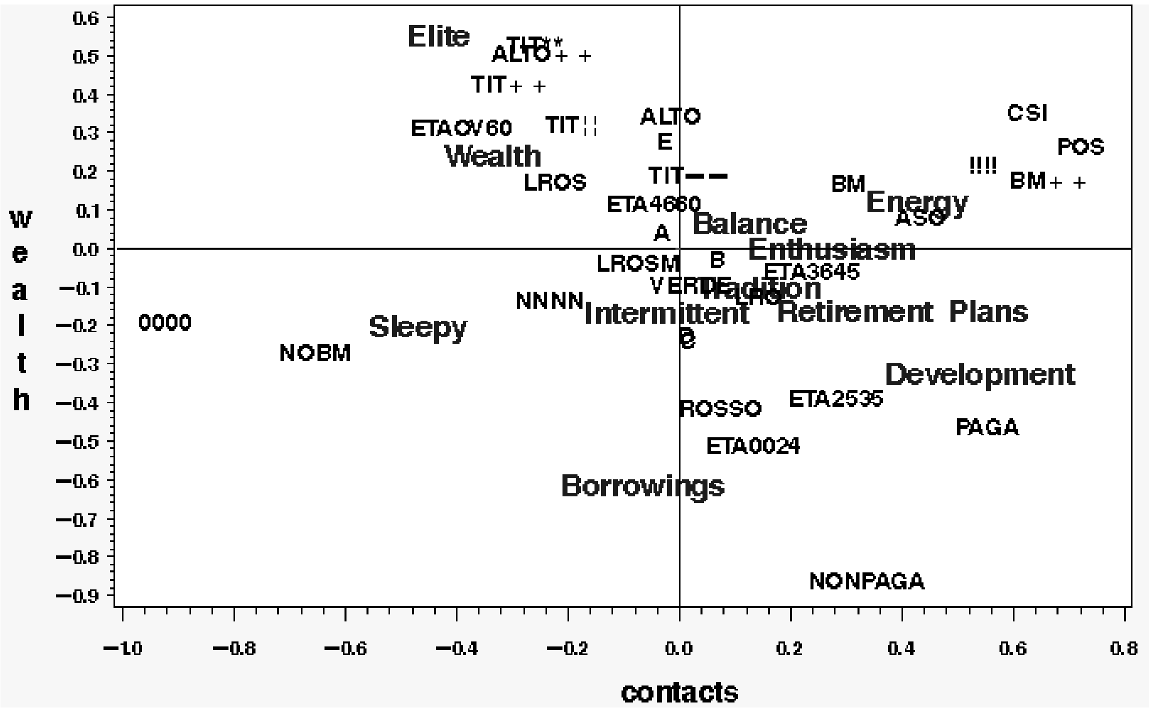
333 Interested *These are customers to be developed in order to convert them into members of the other clusters.*

- ***AAA Enthusiasm*** *Though young, this customer group is open to technological innovation and could have the future need for loans or retirement plans.*
- ***BBB Development*** *Also open to technological innovation, this group could be interested in instruments which allow for building up capital over time.*
- ***CCC Retirement Plans*** *These customers could be interested in the base products in the area of institutionally managed savings, such as the "Conto Piu'" (Plus Account), and in some cases the "Conto Dinamico" (Dynamic Account) and the "GPM Speciale" (Special Asset Management Account).*
- ***DDD Intermittent*** *Improvements in customer loyalty are needed in this segment, and could be achieved, for example, with efforts to facilitate the use of bank cards.*
- ***EEE Borrowings*** *Starting from the outstanding loans, the bank could work at developing base products with this group. Such products include the automatic payment of utility bills and bank cards.*
- ***FFF Sleepy or Poised to Flee*** *These are customers to be cultivated; they transact only marginal business with the bank. They could be the target of initiatives to revive relationships.*

444 Elite *These are customers with highly personalized traditional products (essentially, private banking).*

555 Energy *This is fertile ground for innovative products across the entire spectrum of bank services.*

666 Balance *These customers are rather loyal, and open to the possibilities of new products and new types of distribution channels.*



CONCLUDING REMARKS

Though purely a theoretical exercise, the survey turned out to be interesting not only for assessing the efficiency of the statistical techniques involved, but also because it ended up indicating possible practical applications from the results obtained.

Further availability of periodic samples will allow to perform other types of studies based on the data used in this analysis. Such analyses could include an investigation of migrations between clusters, and the monitoring of sales and marketing campaigns.

To facilitate the implementation of such studies, an integrated information-processing environment is needed. On one hand, this sort of environment would provide periodic access to updated data sample (useful for giving the most simple responses) while on the other hand, it would allow for reporting the results back to the branch system. However, the advantages of such a system fully repay the costs necessary for its implementation.

FURTHER READING

Addolorato F., Cuzzocrea G., Saccardi A. (1995). La progettazione di uno Scoring System per il Direct Marketing, *Finanza, Marketing e Produzione,* Anno XIII - Numero 2 - Giugno 1995, EGEA, Milano

Addolorato F., Bonati G., Cuzzocrea G., Saccardi A. (1993). SAS per la ricerca dei migliori clienti, Atti del convegno “SUGItalia’93”, Trieste 20-22 ottobre 1993.

Addolorato, F. (1987). Il settore delle vendite per corrispondenza. SDA Bocconi. Milano.

Agresti, A. (1990). Categorical Data Analysis, John Wiley & Sons, New York

Bouroche, J.M. - Saporta, G. (1980). L'analyse des données, C.L.U. Editrice, Napoli.

Bonati, G. (1991), La comunicazione diretta, *Giornale di marketing,* Giugno, 1991.

Coppi R. - Bolasco S. (Editors) (1989). Multiway data analysis, Elsevier Science Publishers B.V. (North-Holland).

Corbetta, P. (1992). Metodi di analisi multivariata per le scienze sociali, Bologna, Il Mulino, 1992

Cramer, J.S. (1991). The logit model for economist, Edward Arnold a div. of Hodder & Stoughton, London, 1991

Fabbri, G. - Orsini R. (1993), Reti neurali per le scienze economiche, Franco Muzzio Editore.

Fabris, G. (1995). Consumatore & Mercato. Le nuove regole, Sperling & Kupfer Editori, Milano

Fienberg, S.E. (1983). The Analysis of Cross-Classified Categorical Data, The MIT press (Massachusetts).

Jobson, J.D. (1992). Applied Multivariate Data Analysis, Springer-Verlag New York.

Kohonen, T. (1995). Self-Organizing Maps, Springer-Verlag, Berlin Heidelberg New York.

Kohonen, T. (1984). Self-Organization and Associative Memory, Springer-Verlag, Berlin Heidelberg New York.

Lauro, N.C. - Siciliano, R. (1989). Exploratory methods and modelling for contingency tables analysis: an integrated approach. *Statistica applicata,* Vol.1 n.1, 5-32.

Molteni, L. (1993). L’analisi multivariata nelle ricerche di marketing. EGEA. Milano

SAS Institute Inc., *SAS/STAT User's Guide, Version 6, Fourth Edition,* Cary, NC: SAS Institute Inc., 1989.

Siciliano, R. (1991). Reduced-Rank Models for dependence analysis of contingency tables, *International Workshop on Multidimensional Data Analysis,* Meeting of Dutch & Italian Schools, October 2-5, 1991, Anacapri, Italy.

Tabachnick, B.G. - Fidell, L.S. (1989). Using Multivariate Statistics, Harper & Row, Publishers, New York.

Wasserman, P.D. (1993), Advanced Methods in Neural Computing, New York: Van Nostrand Reimhold.

Wasserman, P.D. (1989), Neural Computing Theory and Practice, New York: Van Nostrand Reimhold.

Yoon, Swales, Margavio (1993), A comparison of Discriminant Analysis versus Artificial Neural Networks, Jour.Opl.Res.Soc. 44, n.1, pag.51-60

ATTACHMENT 1

From a statistical perspective, cluster 333 is the one which has the largest number of customers and the least level of diversification with respect to the population average. Given this situation, we settled for performing another analysis on the individuals belonging to this group in an attempt to identify significant groups for the purpose of behavioural segmentation. We decided further deeper the analysis by first using the entire data base, even in the "clustering" phase, and by relying on a neuronal network.

NEURAL NETWORKS

Neural networks are aggregates of parallel calculation units, known as neurons, that are interconnected through links known as synapses, with each having a parameter known as the weight. Each neuron has the capacity to perform simple operations: to weigh and to sum inputs, to filter them through several activation functions, and to pass the value assumed to other neurons or to an outside environment in the form of a result. The interaction of a sufficient number of neurons, organized in the correct number of layers, will allow to use neural networks as trainable universal approximators of functions, regardless of complexity. The training is obtained through iterative algorithms whose objective is the selection of the optimal set of weights of the network. For the purpose of the analysis herein, we might distinguish between networks which operate with and without supervision. In the case of supervision, the network receives as inputs the series of a certain number of variables with the relative sequences of output variables. The topology and the weights of the network are then determined in order to minimize the distance between the actual and the calculated outputs, according to a certain metric, and they may then be used in forecasting (regressors). Without supervision, no indication is given regarding the result which is intended to be obtained, since it is not known; organizing input profiles into standard classes is assigned to the network, thereby leaving it up to the analyst to make the interpretation (classifiers). As a result, these networks are labelled as the self-organizing type, and have the particular ability to create an internal representation of the data. The Kohonen maps are a particular type of self-organizing networks; they are not suitable for tackling interpolation problems, but they are very efficient for identifying relationships, even non-linear relationships, between variables. On the basis of these relationships, the populations of the individual clusters can be regrouped.

We thus took into account the 11 variables relative to current accounts and securities deposit accounts, and added information available on insurance, passbook savings accounts, loans and credit cards, to come up with a total of 35 variables. These variables represent the input neurons relative to a Kohonen map of 16 classes (4 x 4). Accordingly, with each input neuron connected to all classes of the map, the network should assign weights to 560 synapses, and then with iterations, make a fit against 13,924 available profiles (the customers classified in the target 333).

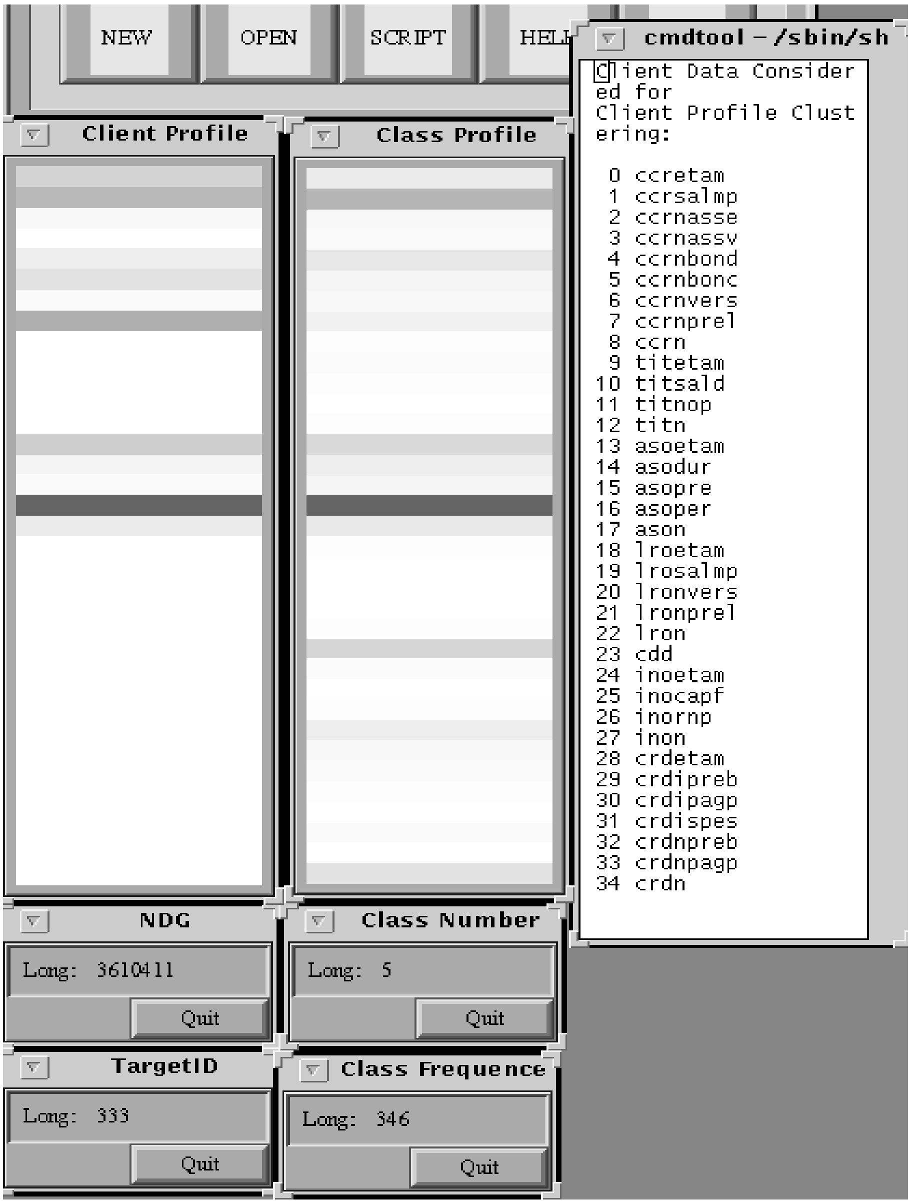
The 16 classes imposed on the Kohonen map, given similarity criteria, were regrouped into the following six classes:

Cluster	Number in Cluster	Cluster	Number in Cluster
AAA	176299	BBB	14356
CCC	5501	DDD	20188
EEE	16429	FFF	174406

As shown, the segmentation is more refined due to the abundant amount of information used, and to the neural network's capacity to grasp non-linear subdivisions of the multi-dimensional input space.

The analysis of the output weights provides an interesting and immediate use of the Kohonen map. It is possible to compare the value of the features input for the individual customers (variables used for training the neuronal network), with the value of the weights relative to the cluster to which the individual belongs. The latter data represent, de facto, the commercial potential of the cluster.

For example, the GCN 3610411 does not have any certificates of deposit (CDD), does not have any loans (INO) and has no credit cards (CRD) or Bancomat; nonetheless, this customer is part of a group in which such products appear.



```
Client Data Consider
ed for
Client Profile Clust
ering:

 0 ccretam
 1 ccrsalmp
 2 ccrnasse
 3 ccrnassv
 4 ccrnbond
 5 ccrnbonc
 6 ccrnvers
 7 ccrnprel
 8 ccrn
 9 titetam
10 titsald
11 titnop
12 titn
13 asoetam
14 asodur
15 asopre
16 asoper
17 ason
18 lroetam
19 lrosalmp
20 lronvers
21 lronprel
22 lron
23 cdd
24 inoetam
25 inocapf
26 inornp
27 inon
28 crdetam
29 crdipreb
30 crdipagp
31 crdispes
32 crdnpreb
33 crdnpagp
34 crdn
```

NDG — Long: 3610411

Class Number — Long: 5

TargetID — Long: 333

Class Frequence — Long: 346