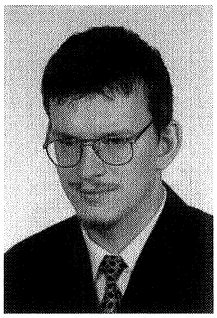
IN THIS ISSUE

ON THE CHOICE OF SAMPLE SIZES IN STRATIFIED DOUBLE SAMPLING FOR NONRESPONSE

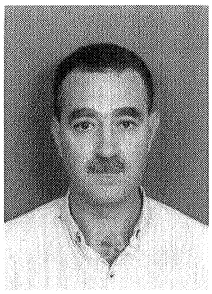
The two-phase (or double) sampling is one of well-known methods for reduction of the unit nonresponse bias. According to this procedure, a subsample is drawn for the set of nonrespondents, and units belonging to the subsample are recontacted in order to collect data. The population parameter estimates evaluated on the basis of the sample and the subsample are then combined. This allows reduction in nonresponse rate, and under some conditions it also leads to unbiased estimates of population parameters. In this paper, a two-phase stratified sampling strategy for estimation of the population mean is considered from randomization theory point of view. A method of establishing initial sample and subsample sizes utilizing J.B Rosen's gradient projection nonlinear programming algorithm is studied.



Wojciech Gamrot was born on 1975, in Gliwice, Poland. In 1999 he received M.Sc degree in computer science at the Faculty of Automatic Control, Electronics and Computer Science of the Silesian Technical University. In 2000 he received licentiate diploma in finance and banking at the University of Economics in Katowice. In 2000 he started his PhD at the University of Economics in Katowice. Since 2002 he is employed as assistant in the Department of Statistics. Current activities include teaching statistics and mathematical statistic, as well as research in the area of survey sampling. Actually, his research interests concentrate on nonresponse problems. He is preparing PhD thesis devoted to methods of statistical inference under nonresponse.

APPLICATION OF THE CART METHOD TO A NEW PRICING CLASSIFICATION IN AUTOMOBILE INSURANCE

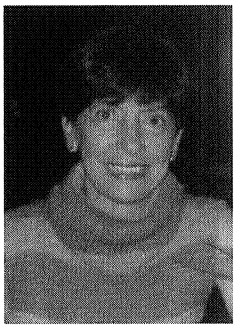
The aim of this work is to propose a new pricing classification for the "third party guarantee premium" in automobile insurance. For this, by applying the CART's method (Classification And Regression Trees), we have built two regression trees; one is based on the prediction the variable "the ratio *s/p*" (disaster victims number / production), the other is based on the prediction variable "disaster victims number". The prediction variable used for the new pricing classification is chosen so as to minimize the standard error. After pruning the two trees, the results show that the tree obtained with "disaster victims number" as the prediction variable gives a minimal error. The modeling of the pricing problem by a mathematical program (non linear objective function with linear constraints) gives the tariffs for the new classification.



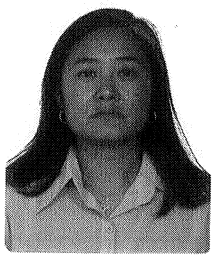
Adjabi Smail was born on 1956 in Kherrata (Bejaia), Algeria and received a D.E.S in Probability-Statistics from Constantine University (1979), DEA (1981) and Doctorat third Cycle (1984) from University of Paris XI – Orsay, France in Statistics, in the laboratory "Probability-Statistics and Modelisation" of Professor D. Dacunha-Castelle. He is associate Professor in the Department of Operations Research, Faculty of Science and Engineering Science, University of Bejaia, Algeria. He is the head of "Statistical Quality Control" team, of the laboratory LAMOS (Laboratory of Modelisation and Optimisation of Systems) which was founded since 1985 at Bejaia University.

EFFECT OF PROCESS VARIABILITY ON THE EFFICIENCY OF $\overline{X}$-CONTROL CHART DESIGNS

Traditionally the efficiency of $\overline{X}$ control charts is compared by ARLs (Average Run Length) values considering shifts only the process mean. In this paper, values of ARL's were computed and plotted for a selected set of parameters in two scenarios: shifts only in process mean and shifts in the process mean and process variability. Values of ARLs for the second scenario are not always lower than for the first one, but they are surely more realistic. The outcomes reveal that erroneous conclusions about the efficiency of the control chart can be taken when shifts in the variability are not considered.



Osiris Turnes holds a MsC in Mathematics and a Doctor in Industrial Engineering, both from the Federal University of Santa Catarina, Brazil. Her research interests and previous papers have been in Economic Design of Control Charts and Modelling Farm Plans for Cropping System. She works as collaborator in University of Brasilia Hospital on Statistical Quality Control Methods in Infection Control.



Linda Lee Ho received her PhD in 1995 from the Department of Industrial Engineering at Universidade de São Paulo, Brazil. Since 1990 she has been serving on the faculty of the same department. Her major research interests are statistical quality control, process capability index and reliability analysis.

PARTIAL REWEIGHTED LEAST SQUARES

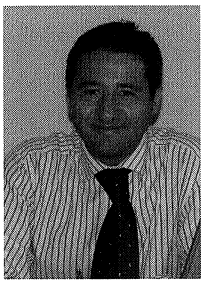
Partial Least Squares Regression is a derived components regression method which has recently attired great attention especially in application fields like chemometrics. The author proposes here the Partial Reweighted Least Squares (PRLS) algorithm illustrated on regression problems. The use of PRLS regression as a less sensitive to outliers alternative is demonstrated. Finally, the Akaike Information Criterion (AIC) is used for model selection.

Athanassios Kondylis was born on the 1st of novemver 1976 in Athens, Greece. In 2000 he obtained his graduation from the statistics department at the University of Piraeus. During the academic years 2002-2003 he followed a master degree in statistics at the University of Neuchâtel where he actually prepares his PhD thesis.

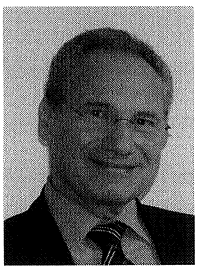
DATA & STATISTICS

STATISTICAL ANALYSIS OF THE FINANCIAL STRUCTURE IN ITALY

The main objective of the authors is supplying the needed information for analysing regional financial unbalances in Italy by using techniques of multivariate analysis and classification. The particular analysis of the clusters (cluvar method) allowed to draw of a picture of the differences of the financial system, highlighting the relations between the regions and the most important variables that have contributed to the determination of such relations.



Benedetto Torrisi graduated in Statistics at the University of Palermo and was PhD in Applied Statistics of the Territory at the University of Naples; since 1999 he has been teaching Quantitative Methods at the Faculty of Economics and at the Faculty of Political Science of the University of Catania. At the present he is a researcher of the Economical Statistics Department. His methodological interests are mainly based upon the experimentation of quantitative methods for dynamic and static classification of the statistical units by selecting explicit variables.



Giorgio Skonieczny is a Professor of Economical Statistics at the University of Catania. At the present he teaches in different departments such as the Faculties of Economics, Medicine, Psychology and Political Science of the Universities of Catania and Enna. He is involved in numerous university institutional activities. He has been working and researching principally on statistical classification and spatial and temporal distributions. He is author of international publications for quantitative methods applied in tourism, infrastructures, welfare, finance and agriculture.

Student (2004), Vol. 5 No. 2, 77-83

On the Choice of Sample Sizes in Stratified Double Sampling for Nonresponse

Wojciech Gamrot

Department of Statistics,
University of Economics
Bogucicka 14
40-226 Katowice, Poland
Email: gamrot@ae.katowice.pl

Abstract : The two-phase (or double) sampling is one of well-known methods for reduction of the unit nonresponse bias. According to this procedure, a subsample is drawn from the set of nonrespondents, and units belonging to the subsample are re-contacted in order to collect data. The population parameter estimates evaluated on the basis of the sample and the subsample are then combined. This allows reduction in nonresponse rate, and under some conditions it also leads to unbiased estimates of population parameters. In this paper, a two-phase stratified sampling strategy for estimation of the population mean is considered from randomization theory point of view. A method of establishing initial sample and subsample sizes utilizing J.B Rosen's gradient projection nonlinear programming algorithm is studied.

1 Introduction - Double Sampling

Let us assume that a population U of the size N is divided into M non-overlapping strata $U_1, \ldots, U_M$, of the sizes respectively equal to $N_1 ... N_M$. It is assumed that nonresponse is deterministic, and response rates vary between strata. Hence, each stratum U_i can be divided into two non-overlapping subsets U_{1i} and U_{2i} of the sizes equal to N_{1i} and N_{2i} respectively. Sampling units belonging to each of the subsets U_{1i}, would respond, if included in the sample. Units belonging to each of the subsets U_{2i} would not respond. Let us also denote $W_{1i} = N_{1i}/N_i$, $W_{2i} = N_{2i}/N_i$. The aim of the survey is to estimate the mean value of some population characteristic Y.

In the first phase of the survey a simple random sample s_i of the size n_i is drawn without replacement from U_i, according to the sampling design:

$$P_1(s_i) = \binom{N_i}{n_i}^{-1}. \tag{1}$$

Each sample s_i can be divided into two subsets $s_{1i} \subset U_{1i}$ and $s_{2i} \subset U_{2i}$ of the sizes respectively equal to n_{1i} and n_{2i}, where $s_{1i} \cup s_{2i} = s_i$, $s_{1i} \cap s_{2i} = \emptyset$, $n_{1i} + n_{2i} = n_i$. The sizes of these subsets are random variables with hypergeometric distribution function given by:

$$P_n(n_{1i}) = P_n(n_{2i}) = \frac{\binom{N_{1i}}{n_{1i}}\binom{N_{2i}}{n_{2i}}}{\binom{N_i}{n_i}} \tag{2}$$

where:

$$\max(0, n_i - N_{2i}) \leq n_{1i} \leq \min(n_i, N_{1i})$$

$$\max(0, n_i - N_{1i}) \leq n_{2i} \leq \min(n_i, N_{2i})$$

For large N_{1i}, N_{2i} and n_i, we may assume, that none of the subsets is empty with the probability close to 1. Units belonging to subsets s_{1i} respond to questions. Units from subsets s_{2i} do not respond.

In the second phase of the survey, simple subsamples u_i of the size $n_{ui} > 0$ are drawn without replacement from among n_{2i} units of each subset s_{2i}, with the conditional probability of selection given by:

$$P_2(u_i \mid n_{2i}) = \binom{N_{2i}}{n_{ui}}^{-1} \tag{3}$$

Units from subsamples u_i are re-contacted in order to collect data. As it has been proven by Hansen and Hurwitz (1946), if all the second-phase contact attempts are succesful, the following statistic is an unbiased estimator of the mean value of Y in the stratum U_i.

$$\overline{y}_{wi} = \frac{1}{n_i}(n_{1i}\overline{y}_{1i} + n_{2i}\overline{y}_{ui}) \tag{4}$$

where $\overline{y}_{1i}$ is the mean of the characteristic under study in s_{1i} and $\overline{y}_{ui}$ is the mean of the characteristic under study in the subsample u_i. As it has been shown by Cochran (1977), the variance of this estimator is given by:

$$V(\overline{y}_{wi}) = \frac{N_i - n_i}{N_i n_i} S_i^2 + \left(\frac{n_{2i}}{n_i}\right)^2 \frac{N_{2i} - n_{ui}}{N_{2i} n_{ui}} S_{2i}^2 \tag{5}$$

where S_i^2 is the variance of the characteristic under study in the stratum U_i and S_{2i}^2 is its variance in the subsample U_{2i}.

Hence each statistic $\overline{y}_{wi}$ is an unbiased estimator of the mean value in i-th stratum, it is easy to show that the following statistic is an unbiased estimator of the population mean:

$$\overline{y}_w = \sum_{i=1}^{M} \frac{N_i}{N} \overline{y}_{wi} \tag{6}$$

And its variance is equal to:

$$V(\overline{y}_w) = \sum_{i=1}^{M} \frac{N_i^2}{N^2} \left[\frac{N_i - n_i}{N_i n_i} S_i^2 + \frac{n_{2i}^2}{n_i^2} \frac{n_{2i} - n_{ui}}{n_{2i} n_{ui}} S_{2i}^2 \right] \tag{7}$$

because all the statistics $\overline{y}_{wi}$ are independent random variables. There are several possible rules governing the subsample size (see Hansen and Hurwitz 1946 or Rao 1973). Let us consider the procedure similar to the one proposed by Srinath (1971). According to this procedure, the size of subsample n_{ui} in every stratum is always set to minimum value, for which the conditional variance (given fixed n_i and n_{2i}) of the estimator (4) does not exceed some predefined limit V_{0i}. Consequently, minimum subsample size n_{ui} in the i-th stratum is given by[1]:

$$n_{ui} = \frac{n_{2i}^2}{z_i(n_i) + n_{2i}} \tag{8}$$

where

$$z_i(n_i) = n_i \left(\frac{n_i}{n_i'} - 1 \right) \frac{S_i^2}{S_{2i}^2} \tag{9}$$

$$n_i' = \frac{N_i S_i^2}{N_i V_{0i} + S_i^2} \tag{10}$$

and for the solution $n_{ui} \geq 0$ to exist it is required that $n_i \geq n_i'$. The quantity n_i' represents the minimum initial sample size in the i-th stratum required to achieve the variance limit V_{0i} without nonresponse.

2 The cost and its expectation

Let us assume that the cost of surveying the i-th stratum may be expressed as:

$$K_i = C_{0i} n_i + C_{1i} n_{1i} + C_{2i} n_{ui} \tag{11}$$

where C_{0i} is the per-unit cost of making the first contact attempt in the i-th stratum, C_{1i} is the per-unit cost of processing the data obtained during the first contact attempt in the i-th stratum and C_{2i} is the per-unit cost of the second attempt to collect and process data. The overall survey cost in the population is then equal to:

$$K = \sum_{i=1}^{M} K_i = \sum_{i=1}^{M} (C_{0i} n_i + C_{1i} n_{1i} + C_{2i} n_{ui}) \tag{12}$$

Hence for fixed V_{0i}, n_i and n_{2i} the cost K_i grows with the subsample size n_{ui}, then it is minimized for the subsample size given by the expression (8). The expected cost in the i-th stratum is then equal to:

$$E(K_i) = C_{0i} n_i + C_{1i} E(n_{1i}) + C_{2i} E\left(\frac{n_{2i}^2}{z_i(n_i) + n_{2i}} \right) \tag{13}$$

Let us assume, as in the paper of Srinath, that $n_{2i} = E(n_{2i}) + \Delta_i$ where $E(\Delta_i) = 0$ and derive the following:

[1] In practice, it would be rounded up to the nearest integer

$$E\left(\frac{n_{2i}^2}{z_i(n_i)+n_{2i}}\right) = E\left(n_{2i} - z_i(n_i) + \frac{z_i^2(n_i)}{z_i(n_i)+n_{2i}}\right) = \tag{14}$$

$$= E(n_{2i}) - z_i(n_i) + E\left(\frac{z_i^2(n_i)}{z_i(n_i)+n_{2i}}\right) = \tag{15}$$

$$= E(n_{2i}) - z_i(n_i) + E\left(\frac{z_i^2(n_i)}{z_i(n_i)+E(n_{2i})+\Delta_i}\right) = \tag{16}$$

$$= E(n_{2i}) - z_i(n_i) + E\left(\frac{z_i^2(n_i)}{z_i(n_i)+E(n_{2i})} \cdot \frac{z_i(n_i)+E(n_{2i})}{z_i(n_i)+E(n_{2i})+\Delta_i}\right) = \tag{17}$$

Further, we may write:

$$\frac{z_i(n_i)+E(n_{2i})}{z_i(n_i)+E(n_{2i})+\Delta_i} = \left[1 - \frac{\Delta_i}{z_i(n_i)+E(n_{2i})} + \frac{\Delta_i^2}{(z_i(n_i)+E(n_{2i}))^2} - ...\right] \tag{18}$$

For any $n_i \geq n_i'$, $z_i(n_i)$ is nonnegative. Assume that Δ_i is small in comparison to $z_i(n_i) + E(n_{2i})$ which means that the fraction n_{2i}/n_i is close to the stratum non-respondent fraction W_{2i}. For sufficiently large n_i this assumption is true with the probability close to one, hence n_{2i} has hypergeometric distribution given by formula (2). Consequently, all the elements of the sum in squared brackets except the first one may be ignored, which leads to the derivation:

$$E\left(\frac{n_{2i}^2}{z_i(n_i)+n_{2i}}\right) \doteq E(n_{2i}) - z_i(n_i) + \frac{z_i^2(n_i)}{z_i(n_i)+E(n_{2i})} = \frac{E^2(n_{2i})}{z_i(n_i)+E(n_{2i})} \tag{19}$$

and the approximate expected cost of the survey in the i-th stratum is:

$$E(K_i) = C_{0i}n_i + C_{1i}E(n_{1i}) + C_{2i}\frac{E^2(n_{2i})}{z_i(n_i)+E(n_{2i})} \tag{20}$$

Replacing $z_i(n_i)$ by expressions (9) and (10) and assuming $E(n_{1i}) = W_{1i}n_i$, $E(n_{2i}) = W_{2i}n_i$ we obtain:

$$E(K_i) = C_{0i}n_i + C_{1i}w_{1i}n_i + \frac{C_{2i}S_{2i}^2W_{2i}^2N_in_i}{n_i(N_iV_{0i}+S_i^2) - N_i(S_i^2 - W_{2i}S_{2i}^2)} \tag{21}$$

In particular, for $n_i = n_i'$ we have:

$$E(K_i) = n_i'\left(C_{0i} + C_{1i}w_{1i} + C_{2i}w_{2i}\right) \tag{22}$$

If V_{0i} is predefined, the expected cost in the i-th stratum takes minimum value for:

$$n_i = \begin{cases} N_i & \text{for} \quad 0 < V_{0i} < V_{0i}' \\ n_i'\left(W_{2i}\frac{S_{2i}^2}{S_i^2}(\beta_i - 1) + 1\right) & \text{for} \quad V_{0i}' \leq V_{0i} \leq V_{imax} \end{cases} \tag{23}$$

where

$$V_{0i}' = \frac{1}{N_i}(\beta_i - 1)W_{2i}S_{2i}^2 \tag{24}$$

$$\beta_i = \max\left(1, \sqrt{\frac{C_{2i}(S_i^2 - W_{2i}S_{2i}^2)}{(C_{0i} + C_{1i}W_{1i})S_{2i}^2}}\right) \tag{25}$$

$$V_{imax} = \frac{N_i - 2}{2N_i}S_i^2. \tag{26}$$

The quantity V_{imax} represents the theoretical maximum variance of $\overline{y}_{wi}$ corresponding to the minimum initial sample size $n_i = 2$ in i^{th} stratum[2]. Applying the formulae (21)

[2] For any stratum, sample size can not be less than two, because according to previous assumptions

and (23) the minimum expected cost in the i-th stratum may be expressed as the following function of V_{0i}:

$$E'(K_i) = \begin{cases} N_i & \text{for} \quad 0 < V_{0i} < V'_{0i} \\ n'_i \left(W_{2i} \frac{S^2_{2i}}{S^2_i} (\beta_i - 1) + 1 \right) & \text{for} \quad V'_{0i} \leq V_{0i} \leq V_{imax} \end{cases} \tag{27}$$

3 The optimization problem

For any set of predefined stratum variance limits $V_{01}...V_{0M}$, formulae (8) and (23) allow to minimize the expected total cost:

$$E'(K) = \sum_{i=1}^{M} E'(K_i) \tag{28}$$

However stratum variance limits may be set in many ways. Hence the variance $V(\overline{y}_w)$ may be expressed as a weighted sum of variances $V(\overline{y}_{w1})...V(\overline{y}_{wM})$ and each $E'(K_i)$ decreases with the increase of V_{0i}, then, by assuming $V(\overline{y}_{wi}) = V_{0i}$ for $i = 1...M$ we can formulate the problem of establishing limits $V_{01}...V_{0M}$ minimizing the expected cost for required precision V_0 of the estimator $\overline{y}_w$:

$$\begin{cases} E'(K) \longrightarrow \min \\ V(\overline{y}_w) \leq V_0 \\ 0 \leq V_{0i} \leq V_{imax} \end{cases} \tag{29}$$

By denoting:

$$x_i = \frac{N_i^2}{N^2} V(\overline{y}_{wi}) = \frac{N_i^2}{N^2} V_{0i} \tag{30}$$

the program above may be transformed to the form:

$$\begin{cases} f(x_1, ..., x_M) \longrightarrow \min \\ x_1 + ... + x_M \leq V_0 \\ 0 \leq x_i \leq x_{imax} \end{cases} \tag{31}$$

where:

$$f(x_1, ..., x_M) = \sum_{i=1}^{M} e(x_i) \tag{32}$$

$$e(x_i) = \begin{cases} a(x_i) & \text{for} \quad x_i \leq x'_i \\ b(x_i) & \text{for} \quad x_i > x'_i \end{cases} \tag{33}$$

$$a(x_i) = N_i \left(C_{0i} + C_{1i} W_{1i} + \frac{C_{2i} N_i W^2_{2i} S^2_{2i}}{N^2 x_i + N_i W_{2i} S^2_{2i}} \right) \tag{34}$$

$$b(x_i) = \frac{N_i^2 \left(W_{2i} S^2_{2i} (\beta_i - 1) + S^2_i \right)}{N^2 x_i + N_i S^2_i} \left(C_{0i} + C_{1i} W_{1i} + \frac{1}{\beta_i} C_{2i} W_{2i} \right) \tag{35}$$

$$x'_i = \frac{N_i^2}{N^2} V'_{0i} = \frac{N_i^2}{N^2} (\beta_i - 1) W_{2i} S^2_{2i} \tag{36}$$

$$x_{imax} = \frac{N_i^2}{N^2} V_{imax} \tag{37}$$

$n_{1i} > 0$ and $n_{2i} > 0$

When the solution $x_1, ..., x_M$ is found, initial sample sizes may be calculated according to the formula:

$$n_i = \begin{cases} N_i \quad \text{for} \quad x_i \leq x_i' \\ \frac{N_i^2(W_{2i}S_{2i}^2(\beta_i - 1) + S_i^2)}{N^2 x_i + N_i S_i^2} \quad \text{for} \quad x_i' \geq x_i \geq x_{imax} \end{cases} \tag{38}$$

It's worth noting, that the limits in the program (31) are linear, but the objective function is not. However it is easy to notice, that $e(x_i)$ is a convex and differentiable function of x_i, because $a(x_i)$ and $b(x_i)$ are decreasing, differentiable functions of x_i, their derivatives towards x_i increase with x_i and it can be shown, that $a(x_i') = b(x_i')$ and derivatives of $a(x_i)$ and $b(x_i)$ towards x_i are also equal in x_i'.

Hence the objective function in the program (31) is convex and differentiable and limits are linear, an iterative nonlinear programming algorithm developed by J.B.Rosen (1969) may be applied to solve it. The description of the algorithm is also given by Findeisen (1980).

4 An example

The following table shows an example solution obtained by applying this method. It was assumed that the population of the size $N = 100000$ is divided into 10 strata of the sizes N_i and response rates W_{1i} and W_{2i} shown in the first three columns. Stratum variances S_i^2, nonrespondent variances S_{2i}^2 and per-unit costs C_{0i}, C_{1i} and C_{2i} vary between strata as it is shown in the columns 4-8. Required precision V_0 was arbitrarily set to 0,05. Optimum values of x_i computed by using Rosen's algorithm, expected within-stratum costs $e(x_i)$ calculated according to the expression (33) and corresponding initial sample sizes n_i in each stratum, obtained according to formula (38) are shown in columns 9-11. Hence the computed sample sizes n_i are not integers, they were rounded up to obtain integer values.

1	2	3	4	5	6	7	8	9	10	11
N_i	W_{1i}	W_{2i}	S_i^2	S_{2i}^2	C_{0i}	C_{1i}	C_{2i}	x_i	$e(x_i)$	n_i
8150	0,6	0,4	1000	1000	0,5	2,0	6,0	0,0039	5540,8	1655
11500	0,8	0,2	1950	1455	0,7	1,5	4,0	0,0053	8871,2	3672
12500	0,55	0,45	950	2000	0,1	0,3	1,5	0,0022	4037,4	4296
5000	0,1	0,9	1500	500	0,1	0,2	9,0	0,0026	3862,3	5000
13000	0,45	0,55	1450	1550	0,3	1,0	3,4	0,0055	8580,5	3965
11700	0,45	0,55	1500	1230	0,6	2,0	8,0	0,008	11339	2962
7100	0,5	0,5	2000	1340	0,6	2,5	6,5	0,0049	7384,3	2057
13250	0,65	0,35	1950	1790	0,7	1,0	7,5	0,0076	11856	4483
8750	0,85	0,15	1250	2300	0,8	2,5	5,0	0,0044	6407,7	1744
9050	0,85	0,15	2540	1540	0,5	2,5	7,5	0,0056	9188,0	2900

Table 1: An example

It is worth noting, that according to the results, all units in the fourth stratum should be included in the first-phase sample, because the relation between stratum

variance and cost is relatively attractive. The total expected cost for the entire population was equal to 77068,99.

References

[1] Cochran W.G. (1977) *Sampling Techniques* New York, Wiley.
[2] Findeisen W. Szymanowski J. Wierzbicki A. (1980) *Teoria i metody obliczeniowe optymalizacji*, PWN, Warsaw.
[3] Rao J.N.K.(1973) On Double Sampling for Stratification and Analytical Surveys, *Biometrika* 60, 125-133
[4] Rosen J.B. (1969) The gradient projection method for nonlinear programming, *Journal of Society for Industrial and Applied Mathematics* vol.8, No.1, 181-217.
[5] Srinath K.P. (1971) Multiphase Sampling in Nonresponse Problems, *Journal of the American Statistical Association* 33, 583-586.

Student (2004), Vol. 5 No. 2, 85-96

Application of the CART Method to a New Pricing Classification in Automobile Insurance

Smaïl Adjabi and Soraya M'Sili

Laboratory of Modelisation and Optimization of Systems
LAMOS
University of Bejaia
Algeria

Abstract: The aim of this work is to propose a new pricing classification for the 'third party guarantee premium' in automobile insurance. For this, by applying the CART method (Classification And Regression Trees), we have built two regression trees; one is based on the prediction variable 'the ratio s/p' (disaster victims number/production), the other is based on the prediction variable 'disaster victims number'. The prediction variable used for the new pricing classification is chosen so as to minimize the standard error. After pruning the two trees, the results show that the tree obtained with 'disaster victims number' as the prediction variable gives a minimal error. The modelling of the pricing problem by a mathematical program (non linear objective function with linear constraints) gives the tariffs for the new classification.

Key words: CART method, regression trees, mathematical programming, pricing, classification, automobile insurance.

1 Introduction

Insurance is a technique of covering casualties risks affecting the goods of a person or an entity (fire, theft, water, natural disasters ...) or resulting from the occurrence of a random event striking the person itself (diseases, disability, industrial accident, death) and injuring him by affecting his ability to get the incomes allowing him to live and take his family into his charge. The current segmentation is quite old and obsolete. Recent studies has shown that certain classes are without interest in terms of production.

The aim of this work, is consequently, elaborating a new pricing system for the third party guarantee premium in automobile insurance based on a new segmentation with respect to new tariff criteria, that takes into account statistical data about the disaster victims.

Regression methods with trees lead to yield a binary tree with the data while partitioning them into homogenous subset until it's not possible to continue (increasing of tree) after that, pruning the branches of that tree in order to simplify without loosing good adequation of the data. This property allows us to define the new segmentation while applying the CART method [1] .

2 Presentation of the CART method

2.1 Describing the method

CART, abbreviation of 'Classification And Regression Trees', is a decision tree algorithm. It was introduced in 1984 by L. Breiman, J. Friedman, R. Olshen and C.J. Stone [1] .

In CART strategy for regression of Breiman and al., a greedy search algorithm is used to construct a binary tree in the independent variables. The goal is to produce nodes as homogeneous as possible with respect to the dependent variable. Consider the multiple regression problem $y_i = f(x_{i1}, \ldots, x_{ip}) + \varepsilon_i$, $i = 1, \ldots, n$, where f is unknown and not easily parameterized, the x_{ij} are known independent variables, and the ε_i are random error terms with zero means. A node t is a subset of the indices $\{1, \ldots, n\}$. The deviance of node t is defined as

$$D(t) = \sum_{i \in t} (y_i - \bar{y}(t))^2 \ ,$$

where $\bar{y}(t)$ is the mean over the observations in node t. The algorithm is recursive. The root node consists of all observations. At each stage, an attempt is made to divide a parent node t into two child nodes, a left node t_{left} and right node t_{right}, so as minimize $D(t_{\text{left}}) + D(t_{\text{right}})$.

For each independent variable, all possible splits are considered according to the appropriate rule, and the deviance for the node following the split, $D(t_{\text{left}}) + D(t_{\text{right}})$, is calculated. The split with the smallest such deviance is saved as a candidate. The candidate splits are calculated for each independent variable, and the variables whose best split produces the smallest deviance is the one selected to partition node t. The algorithm proceeds recursively until no further splitting is possible according to predetermined criteria. Typically, an a priori minimum node size is specified, or splitting stops when the deviance of a node drops below a certain level, e.g. 1% of the deviance of the root node.

The steps for building the decision tree with the CART method could be summarized in what follows:

- Choice of a variable and a partition at every node;
- Stopping criterion;
- Pruning of a tree;
- Importance of the variables;
- Validating the model.

The details of these steps are given in [1] , [3] .

2.2 Binarizing the variables

The description variables are seldom initially binary. If $S(v)$ is the set of the values taken by the variable v, a fundamental aspect of research, consists of the way of building the bipartition (partition into 2 classes) of $S(v)$, in order to obtain binary attributes associated with v. Such a structure must depend on the structure of $S(v)$.

2.2.1 Qualitative variables case

The binarization of a nominal qualitative variable with L levels generates $(2^{L-1}-1)$ binary attributes, corresponding to all the partitions into two non empty classes of the L modalities.

2.2.2 Quantitative variables case

Continuous variable. For each continuous variable X, the set of the binary questions (tests) is of the form (is $X \leq s$), for all $s \in]-\infty, +\infty[$. There is then an infinity of possible tests.

Discrete variable. The CART method, allows the treatment of quantitative variables with integer values as continuous variables. Let $d_1, d_2, \ldots, d_N$ be the set of the values taken by the discrete variable X ordered increasingly. The set of the questions (tests) is (is $X \leq s_i$?), with $S_i = \frac{d_i + d_{i+1}}{2}$; for $i = 1, 2, \ldots, N-1$.

2.3 Construction of the sequence of trees

A sequence $T_1 \succ T_2 \succ \ldots \succ T_K = \{t_l\}$, of decreasing trees is built [5] , [4] , where t_l indicates the root of the tree. Each of the trees T_i is obtained from T_{i-1} while pruning one of its branches. The pruned branch is chosen as follows: a measurement cost-complexity of a tree T is considered, it is defined by:

$$R_\alpha(T) = R(T) + \alpha|\widetilde{T}| = \sum_{t \in \widetilde{T}} [R(t) + \alpha] = \sum_{t \in \widetilde{T}} R_\alpha(t)\,,$$

where $\widetilde{T}$ represents the number of final nodes of the tree T, $R_\alpha(T)$ is the total cost of T for which a penalty equals to α when a terminal node is added. This measurement

can be extended to a branch T^i for which the node root is t, the cost will be then $R_\alpha(T^t) = R(T^t) + \alpha|\widetilde{T}^t|$. For each value of α, we note $T(\alpha)$ the smallest tree of T_{max} that minimizes cost-complexity measurement. The smaller α is the weaker penalty due to high number of final nodes is, and consequently $T(\alpha)$ has a high number of final nodes. As α increases, the tree $T(\alpha)$ becomes increasingly small. For a sufficiently large value of α, $T(\alpha)$ is reduced to the root of the tree T_{max}.

2.4 Regression tree

We deal about a regression tree in the CART method when it to give a certain variable Y in terms of some other explicative variables $X_1, X_2, \ldots, X_p$. In this case, the variable Y is continuous.

The regression tree and the classification tree have the same structure. The space of description X is partitioned by a sequence of binary questions on the terminal nodes. The predicted value in each terminal node is:

$$\bar{y}(t) = \frac{1}{N(t)} \sum_{y_n \in t} y_n \ ,$$

with $R(t) = \frac{1}{N(t)} \sum_{y_n \in t} (y_n - \bar{y}(t))^2$ which represents the squared mean error (deviation) at the node t. $N(t)$ is the cardinal of the set t (the number of observations at a node t).

3 The new classification

3.1 Introduction

In this work, based on the CART method, we propose a new complete classification for the pricing of the third party guarantee premium of an automobile insurance, which takes in consideration all the power codes, all the codes use, as well as the most frequent kind codes.

We will thus try to define the following points:

- Which criterion influences more the the sinistrality?
- Which are the profitable classes in terms of production and the nonprofitable classes?
- Which are the combinations that yield the same degrees of sinistrality?
- Which prediction variable must be considered in order to propose a new classification, 'the ratio s/p' (disaster victims number/production) or 'disaster victims number'?

We will first build two regression trees, one is based on the prediction variable 'the ratio s/p', and the other is based on the prediction variable 'disaster victims number', then, we will define a new classification, that will be based on the prediction variable giving a minimal standard error. The used data are recent and related to the production and to the sinistrality. Each class is identified by a tariff code which is described as follows: the driver's age, driving licence, age of vehicle, use of vehicle, power of vehicle.

3.2 Results and interpretation

The number of trees obtained in the phase of pruning for the prediction variable 'the ratio s/p', as it is seen on table Tab. 1 , is 16 trees, the last tree in the sequence is reduced to 4 nodes and the optimal tree contain 25 final nodes, with standard error equal to 0.845. The number of trees obtained in the phase of pruning for the prediction variable 'disaster victims number', as it is shown in table Tab. 2 , is 19 trees. The last tree in the sequence is reduced to 5 nodes and the optimal tree contains 28 final nodes, with standard error equal to 0.714. Both two trees were validated by a sample test, but as the tree constructed with the variable 'disaster victims' has given the minimal error, the new segmentation is the one given by this tree.

The optimal tree of regression (Fig. 1) contains 28 final nodes, where each final node describes a class. The use and power attributes take integer values defined as follows: use code $= \{0, 1, 2, 3, 4, 5, 7\}$, power code $= \{0, 1, 2, 3, 4, 5, 7\}$. The details of each class and the conditions of membership are given in table Tab. 3.

3.3 Importance of variables

Another advantage of CART, is that it quantifies the importance of the variables [6] . This latter is measured by the sum of all decrements of deviance yielded at every node of the tree when replacing for each node the optimal division by the substitution one based on the variable. This measure depends upon the tree from which it is calculated. In general, the importance is scaled to the interval [0 , 100] by division by their maximum and multiplying by 100. Indeed, as we note it on table Tab. 4, the variable power is more significant in this regression. In second position, we find the use code, then the age of the driver, the age of the driving licence and in last position the age of the vehicle.

4 Modelling the pricing problem

The pricing problem could be modelled by a mathematical program (nonlinear function objective with linear constraints).

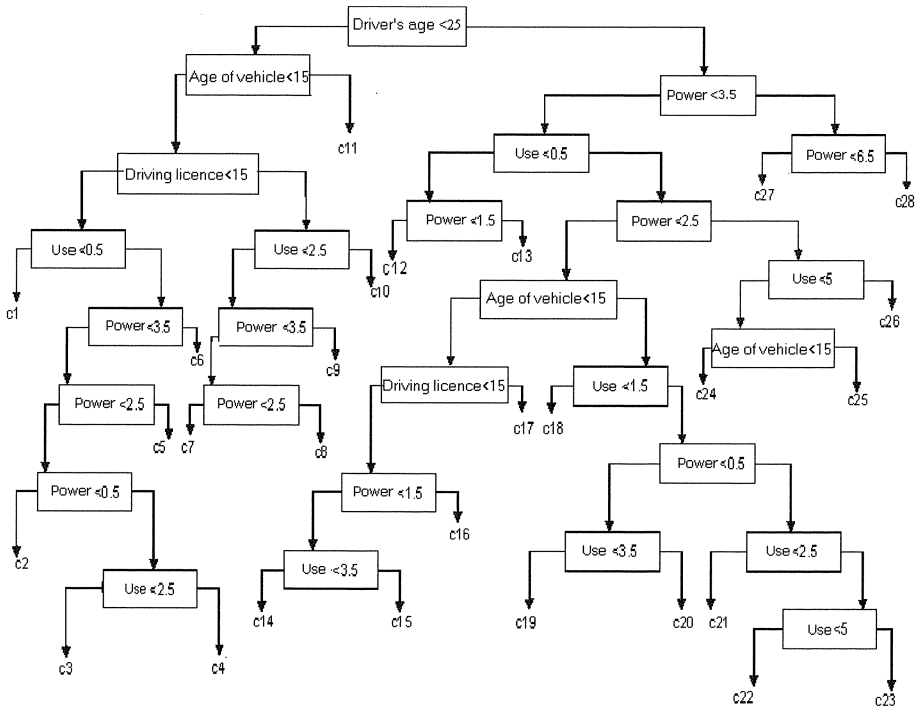


Figure 1. Classification tree.

Tree standard N^o	Nbr of nodes terminals	Error of the tree
1	25	0.845
2	24	0.876
3	22	0.874
4	22	0.873
5	19	0.87
6	18	0.899
7	17	0.894
8	16	0.873
9	13	0.910
10	12	0.893
11	9	0.887
12	8	0.900
13	7	0.901
14	6	0.953
15	5	0.902
16	4	0.971

Table 1: Sequence of trees obtained in the phase of pruning for the variable 'the ratio s/p'.

4.1 The constraints of the model

Let us denote

d_j duration of the contract, $d_j = 30,\ 60,\ 180,\ 365$ days, $j = 1, 2, 3, 4$.

N_{ij} the number of people of class i having taken the contract of duration d_j, $i = 1, \ldots, 28$, $j = 1, 2, 3, 4$.

P_{ij} the price of the automobile insurance of class i for the contract duration d_j, $i = 1, \ldots, 28$, $j = 1, 2, 3, 4$.

S_i The total cost of the disaster victims of the class i, $i = 1, \ldots, 28$.

- The sum of the receipts (incomes) of each class is higher than the sum of the disaster victims costs of the same class:

$$\sum_{j=1}^{4} N_{ij} P_{ij} \geq S_i \quad i = 1, \ldots, 28.$$

- The insurance company supports much more contracts of longer period than those of short period. Thus, the respective ratio prices of each period does not imply

Tree standard N^o	Nbr of nodes terminals	Error of the tree
1	40	0.723
2	38	0.725
3	36	0.724
4	35	0.754
5	34	0.724
6	33	0.724
7	32	0.716
8	28	0.714
9	25	0.717
10	24	0.723
11	22	0.725
12	20	0.726
13	18	0.783
14	15	0.783
15	12	0.791
16	10	0.98
17	9	0.974
18	8	0.725
19	5	0.723

Table 2: Sequence of trees obtained in the phase of pruning for the variable 'number disaster victims'.

the equality with the ratio of the duration of successive periods:

$$\frac{P_{ij}}{P_{i(j-1)}} \leq \frac{d_j}{d_{(j-1)}} \quad i = 1, \ldots, 28 \quad j = 2, 3, 4.$$

- the ratio of the prices of contract of two successive durations have to be not lower than the ratio of these durations scaled by a parameter θ:

$$\frac{P_{ij}}{P_{i(j-1)}} \geq \theta \, \frac{d_j}{d_{(j-1)}} \quad i = 1, \ldots, 28 \quad j = 2, 3, 4.$$

θ is a parameter chosen by the decision maker. The current tariffs register fixes the value of θ within the interval [0.7 , 0.9].

4.2 The objective function

The main goal of a given insurance company is to carry out a balance between its receipts and its expenditure. If $(1 - \beta)$ represents the rate of charge applied by the company, composed by an intermediate commission, overhead and benefit; then ob-

Class N 1	Class N 2	Class N 3
Driver's age < 25 years Driving licence< 1 year Age of vehicle< 15years Use< 0.5	Driver's age < 25 years Driving licence< 1 year Age of vehicle< 15years Use>0.5 Power< 0.5	Driver's age <25 years Driving licence< 1 year Age of vehicle< 15years 0.5<Use< 2.5 0.5 <Power< 2.5
Class N 4	Class N 5	Class N 6
Driver's age < 25 years Driving licence< 1 year Age of vehicle< 15years Use > 2.5 0.5 <Power< 2.5	Driver's age < 25 years Driving licence< 1 year Age of vehicle< 15years Use> 0.5 2.5 <Power< 3.5	Driver's age < 25 years Driving licence< 1 year Age of vehicle> 15years Power> 3.5
Class N 7	Class N 8	Class N 9
Driver's age < 25 years Driving licence< 1 year Age of vehicle> 15years Use< 2.5 Power< 2.5	Driver's age < 25 years Driving licence< 15 years Age of vehicle> 1 year Use< 2.5 2.5 <Power< 3.5	Driver's age < 25 years Driving licence< 15 year Age of vehicle> 15years Use< 2.5 Power > 3.5
Class N 10	Class N 11	Class N 12
Driver's age < 25 years Driving licence< 1year Age of vehicle> 15 years Use< 0.5	Driver's age < 25 years Driving licence> 1 year	Driver's age > 25 years Use< 0.5 Power< 1.5
Class N 13	Class N 14	Class N 15
Driver's age > 25 years Use< 0.5 1.5 <Power< 3.5	Driver's age > 25 years Driving licence> 1 year Age of vehicle< 15years 0.5 <Use< 3.5	Driver's age > 25 years Driving licence< 1 year Age of vehicle< 15years Use< 0.5 Power< 21.5
Class N 16	Class N 17	Class N 18
Driver's age > 25 years Driving licence< 1year Age of vehicle< 15years Use>0.5 1.5 <Power< 2.5	Driver's age > 25 years Driving licence> 1 year Age of vehicle> 15years	Driver's age > 25 years Driving licence> 1year 0.5 < Use< 1.5 Power< 2.5
Class N 19	Class N 20	Class N 21
Conducting age > 25 years Driving licence> 1year Power< 0.5 1.5 <Use< 3.5	Driver's age > 25 years Driving licence>1year Power< 0.5 Use> 3.5	Driver's age > 25 years Driving licence> 1year 0.5 <Power< 2.5 1.5 <Use< 2.5
Class N 22	Class N 23	Class N 24
Driver's age > 25 years Driving licence> 1year 0.5 <Power< 2.5 1.5 <Use< 5	Driver's age > 25 years Driving licence> 1year 0.5 <Power< 2.5 Use> 5	Driver's age 0 > 25 years Driving licence< 1 year 2.5 <Power< 3.5 0.5<Use< 5
Class N 25	Class N 26	Class N 27
Driver's age > 25 years Driving licence> 1year 2.5 <Power< 3.5 0.5<Use< 5	Driver's age > 25 years 2.5 <Power< 3.5 Use>5	Driver's age > 25 years 3.5 <Power< 6.5
Class N 28		
Driver's age > 25 years 2.5 <Power< 6		

Table 3: Details of the obtained class.

Variables	Power	Use	Driving licence	Driver's Age	Age of the vehicle
Importance	100.00	84.89	26.95	23.02	0.77

Table 4: Importance of variables.

jective function to be minimized is:

$$\sum_{i=1}^{28}\left[\frac{S_i+(1-\beta)\sum_{j=1}^{4}N_{ij}P_{ij}}{\sum_{j=1}^{4}N_{ij}P_{ij}}\right]^2=\sum_{i=1}^{28}\left[\frac{S_i}{\sum_{j=1}^{4}N_{ij}P_{ij}}-\beta+1\right]^2.$$

It is then sufficient to minimize

$$\sum_{i=1}^{28}\left[\frac{S_i}{\sum_{j=1}^{4}N_{ij}P_{ij}}-\beta\right]^2.$$

4.3 The final Model

The model of pricing the automobile insurance premium is:

$$\begin{cases} \min\sum_{i=1}^{28}\left[\frac{S_i}{\sum_{j=1}^{4}N_{ij}P_{ij}}-\beta\right]^2 & \\ \frac{P_{ij}}{P_{i(j-1)}}\leq\frac{d_j}{d_{(j-1)}} & i=1,\ldots,28,\ j=2,3,4 \\ \frac{P_{ij}}{P_{i(j-1)}}\geq\theta\frac{d_j}{d_{(j-1)}} & i=1,\ldots,28,\ j=2,3,4 \\ \sum_{j=1}^{4}N_{ij}P_{ij}\geq S_i & i=1,\ldots,28,\ j=1,2,3,4 \\ P_{ij}>0 & i=1,\ldots,28,\ j=1,2,3,4 \end{cases}$$

The resolution of this model was obtained by Solver of Excel it uses the nonlinear code of optimization 'Generalised Reduced Gradient' developed by Leon Lasdon and Alain Waren [9] , on a sample of 58106 data production and 2913 sinister data. Percentage of loading applied by the company is 0.75 and parameter $\theta = 0.70$ (to increase the interval of realizable search for solution). The obtained tariffs are presented in the Table 5.

5 Conclusion

CART is a non-parametric method, it builds a partition in basic rectangular paving stones of the space of the explanatory variables, and assigns a value of the explained variable to each set of the partition. The partition is represented by a binary tree,

Class number	Pricing 365 days	Pricing 180 days	Pricing 90 days	Pricing 30 days
1	3815.53	2119.71	1065.18	467.08
2	4003.36	2002.68	1011.45	481.64
3	3920.40	1990.13	995.02	406.49
4	1800.20	900.90	455.92	178.09
5	2150.32	1081.64	547.94	197.10
6	1908.16	960.99	493.88	204.04
7	2671.63	1343.20	839.50	201.10
8	1587.35	814.02	407.01	157.75
9	4022.34	2012.17	1010.73	370.23
10	3001.13	1500.54	761.70	299.88
11	2987.52	1780.40	1013.89	357.00
12	788.65	450.65	265.09	113.87
13	978.06	589.19	392.79	175.35
14	2203.56	1115.35	564.52	241.36
15	3874.36	1939.11	1104.28	431.35
16	2013.63	1213.76	606.88	236.14
17	2238.38	1565.30	801.89	283.35
18	488.64	344.35	242.50	58.54
19	893.33	607.70	394.61	170.82
20	794.48	472.9	271.78	155.30
21	694.48	415.85	231.03	104.53
22	2017.56	1024.14	547.66	203.59
23	978.56	666.14	374.23	136.58
24	3704.77	2107.37	1053.68	367.68
25	3120.42	1651.01	922.87	376.13
26	1954.20	1045.00	527.79	226.51
27	1008.34	695.40	451.56	213.00
28	1203.32	636.67	340.46	162.22

Table 5: Pricing of new classification.

which is optimal for the squared sum errors criterion of the intra-classes. This property seems to be well adapted to the problem of searching a new pricing classification for the third party guarantee premium in automobile insurance. The application of CART method, has allowed to obtain a tree with 28 classes of customers. The modelling of the pricing problem by a mathematical program gave the new pricing satisfying the desired criteria.

References

[1] Breiman, L., Freidman, J., Olshen, R. and Stone, C.J. (1984). *Classification and regression trees*, Wadswoth.

[2] Bowman, A.W. and Azzalini, A. (1997). *Applied smothing techniques for data analysis*, Oxford Science Publication.

[3] Breiman, L. (1996). Heuristic of instability and stabilization in model selection. *Annals of Statistics*, 24, 2350–2383.

[4] Gueguen, A. and Nakache, J.P. (1988). Méthode de discrimination basée sur la

construction d'un arbre binaire. *Revue de Statistique Appliquée*, 36, 19-37.

[5] Ghattas, B. (2000). Agrégation d'arbres de classification. *Revue de Statistique Appliquée*, 48, 85-98.

[6] Ghattas, B. (1999). Importance des variables dans les méthodes CART. *Modulad*, 24, 29-39.

[7] Ghattas, B. (1999). Prévision des pics d'ozone par arbre de régression simples et agrégés par bootstrap', *Revue de Statistique Appliquée*, 47, 61-88.

[8] Tomassone, R., Danzart, M., Daudin, J.J. and Masson, J.P. (1988). *Discrimination et classement*. Masson.

[9] Lasdon, L.S. and Waren, A.D. (1978). Generalized Reduced Gradient Software for linearly and nonlinearly constrained problems. In *Design and implementation of optimisation software*, Greenberg, Sitjhoff and Noordhoof.

Student (2004), Vol. 5 No. 2, 97-111

Effect of Process Variability on the Efficiency of $\overline{X}$-Control Chart Designs

Osiris Turnes

Departamento de Estatística
Universidade de Brasília
Brasília - DF, Brazil
E-mail: osiris@unb.br

Linda Lee Ho

Departamento de Engenharia de Produção
EPUSP
São Paulo - SP, Brazil
E-mail: lindalee@usp.br

Abstract: In this paper the efficiency of $\overline{X}$-control charts is compared by ARL (Average Run Length) values in two scenarios: shifts only in process mean and shifts in the process mean and process variability. The ARLs were computed and plotted for a selected set of parameters. ARLs for the second scenario are not always lower than for the first one, but they are surely more realistic. The outcomes reveal that erroneous conclusions about the efficiency of the control chart can be taken when shifts in the variability are not considered.

Key words: Statistical Process Control, $\overline{X}$-control chart, power of the test, ARL.

1 Introduction

Design of control charts is still a subject of great interest. Many authors have made important contributions making general reviews about this matter. Gibra (1975) has made an extensively review about control charts from 50's to 70's. According to Gibra (1975), until 1975 most of economic design was focused on $\overline{X}$-charts. And tasks to formulate economic models for p, c and R was still to be accomplished. Montgomery (1980) has made a review and literature survey on economic design procedures for

control charts. He has stated that most of standard control charts, such as $\overline{X}$ charts, p-chart and cumulative sum chart have been extensively investigated and several different process models have yet been studied. Vance (1983) has made a general review about statistical quality control charts techniques (including economic design) for the period of 1970-1980. Similarly Svoboda (1991) and Ho & Case (1994) have contributed with a literature review about economic design of control charts. In the same direction Keats *et al.* (1997) have addressed recent contributions on the economic modelling of control charts. They have described about the development and advantages of such methods, discussed the main barriers to implementation, and the future of such approaches. Finally Woodall (1997) has presented a comprehensive bibliography of references on control chart for attribute data. A brief overview, perspectives and suggestion for future research is given in this area. The above cited papers do not exhaust the subject but they are good references.

When control charts are designed using economic criteria, a sample size (n), a sampling interval (h) and control limits coefficient (k) are determined such that the on-line monitoring procedure costs are minimized. This type of design has received considerable attention. Since Duncan's (1956) pioneering work, many authors have made contributions on this subject. For instance, in Gibra (1971), $\overline{X}$-charts are used to evaluate the quality of a current process and he has determined the optimal design for such chart assuming an exponential distribution for the time in which the process remains in control. Baker (1971) developed semi-economic models for the discrete time considering the Poisson distribution as an alternative to the geometric one. Ladany and his colleagues (1973, 1975, 1976) developed economic designs for Shewhart's charts considering p-chart case, to control short production runs, as in a job-shop mode. Del Castillo and Montgomery (1993) extended the latter results to design $\overline{X}$-charts. A general method to determine an economic model to control charts was developed by Lorenzen and Vance (1986). Such method can be applied to design all types of control charts, regardless of the statistic used to draw the charts. Collani (1987) developed an economic model to design $\overline{X}$-chart that reduces number of parameters necessary to obtain the optimal sample size (n) and the control limit coefficient (k), since he had already determined the most efficient duration of the sampling interval in Collani (1985). A general approach to optimal control which incorporates monitoring inspection and replacement/renewal activities was also proposed by Collani(1988). About the joint design of $\overline{X}$-R charts, we have contributions from Saniga (1989); Saniga *et al.* (1995) and McWillians (1994). In Saniga (1989), a cost per time unit model is minimized subject to constraints on error probabilities (type I and type II).

One measurement of efficiency for a control chart design is ARL (Average Run Length) which represents the average number of samples to be taken until a shift in the process is detected. When control charts for variables are designed not taking into account economic criteria, the most common is to use jointly $\overline{X} - R$ charts or $\overline{X} - S$ charts. Traditionally, the expression of ARL for $\overline{X}$-chart does not take into account a shift in the variability of the process, although no shift in the variability of a process

is rarely observed. Thus, a comparative study of the efficiency values of $\overline{X}$-control-chart designs is developed in this paper considering two scenarios: no changes in the variability and with changes in the variability. Such comparison intends to show that the effect of the process variability should not be neglected, even when simple $\overline{X}$-charts are used. In Section 2, a brief theoretical description of $\overline{X}$-charts and the efficiency of the plans are presented. The results are illustrated using a set of parameters. In Section 3 the conclusions and final remarks are drawn.

2 Efficiency of $\overline{X}$-control chart designs

Measurements of a quality characteristic of interest from successive items are taken in a specific production run. When the process is under control, they are independent normal random variables with mean μ_0 and standard deviation σ_0 . An $\overline{X}$-control chart is obtained by plotting the values of sample means $\overline{X}_i$ from successive samples of size n . A process is considered out of control when the first $\overline{X}_i$ does not fall within the lower (LCL) and upper (UCL) control limits which are respectively:

$$LCL = \mu_0 - k\sigma_0/\sqrt{n}$$

and

$$UCL = \mu_0 + k\sigma_0/\sqrt{n}$$

where k is the number of standard deviations which determines the critical region of the chart. There is a direct connection between control charts and tests of hypothesis. Basically, a control chart verifies the hypothesis: the process is under control versus the process is out of control. As in test of hypotheses, let us consider Type I error, represented by α and Type II error, represented by β. The former occurs when the process is under control, but this hypothesis is rejected; and the latter occurs when the process is out of control, but the hypothesis that it is under control is considered. In the present work, α is seen as "the manufacturer risk" and β is "the consumer risk". Determining the control limit coefficient (k), the sample size (n) and the sampling interval (h) is a critical decision in designing a control chart. If the limits determined by k are too narrow, many false alarms will occur. On the other hand, if the limits are too wide, then the sample size n and the sample cost will necessarily increase, otherwise Type II error will occur frequently. The problem consists of finding the balance between these two risks which move in opposite directions. Here, Type II error will be considered as the control chart performance measurement. Consider then, a process with standard deviation σ_0 . If the process mean shifts from the target value μ_0 to $\mu_1 = \mu_0 + \sigma_0$, then the probability of no shift detection is

$$\beta = P(LCL \leq \overline{x} \leq ULC \mid \mu_1 = \mu_0 \pm \sigma_0).$$

Because $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and LCL and UCL are known, the probability of taking a correct decision (power) is:

$$p_1 = 1 - \beta = 1 - [\Phi(k - \Delta\sqrt{n}) - \Phi(-k - \Delta\sqrt{n}], \tag{1}$$

with $\Delta = \frac{|\overline{X} - \mu|}{\sigma_0}$ and $\Phi(.)$ denotes the cumulative standard normal distribution function. The expected number of samples to be taken until the shift is detected, is known as Average Run Length (ARL) and it is given by:

$$ARL = \sum_{k=1}^{\infty} k\beta^{k-1}(1-\beta) = (1-\beta)^{-1} \tag{2}$$

Observe that the expression (2) does not consider a shift in the variability of the process. Unfortunately, the great majority of production processes experience shifts in variability. These shifts can lead to false values of β and can also lead to mistaken conclusions on the power of the control chart design adopted. Now, let consider the quality characteristic normally distributed with parameters μ_0 and σ_0 - when the process is under control - and also normally distributed with parameters μ_1 and σ_1 - when a special cause with expected value is present, then the probability of making the correct decision written in function of the four parameters is

$$p_2 = 1 - \beta = 1 - \left[\Phi\left(\frac{\mu_0 - \mu_1}{\sigma_1}\sqrt{n} + \frac{k\sigma_0}{\sigma_1}\right) - \Phi\left(\frac{\mu_0 - \mu_1}{\sigma_1}\sqrt{n} - \frac{k\sigma_0}{\sigma_1}\right)\right] \tag{3}$$

The following set of parameter values:

- $n = 2, 3, 4, 5, 7, 10$;
- $\delta = \frac{\mu_0 - \mu_1}{\sigma_0} = 0.5; 1.0; 1.5; 2.0$ and 2.5 ;
- $k = 2.5; 3.0; 3.5$ and 4.0;
- $q = \frac{\sigma_0}{\sigma_1} = 0.001; 0.005; 0.01; 0.05; 0.25; 0.5; 0.75$ and 1.0;

was used to analyse the effect of the variability of the process in control $\overline{X}$-chart designs. Values of (2) using (3) are the power of the sampling plan with shifts in the process mean and in the process variability. These values are denoted as ARL_2 and they are displayed in Tables 1-6. When $q = \frac{\sigma_0}{\sigma_1} = 1$, values of (2) are the power of the sampling plan with shifts only in the process mean. These values are bold-faced in Tables 1-6 and columns denoted by *ratio* is the quotient between ARL_2 and these bold-faced values (in the text, these values will be referred as ARL_0). In Figures 1-6 values of ARL_2 were plotted against q varying n; k and δ.

We observe that as k increases, values of ARL_2 decrease as expected. When the standard deviation of the out-of-control process is much higher than the under control

process ($\sigma_1 \gg \sigma_0$), values of ARL_2 are more realistic than ARL_0, detecting quickly a shift in the process, for any value of k. In this case, ARL_2 underestimates ARL_0 and $\overline{X}$-chart considering shifts in a variability is recommendable.

As δ increases, the effectiveness of $\overline{X}$-chart design considering shift in a process variability increases and the pattern of the figures becomes similar. When $\delta \cong k$ (the number of standard deviation of the process out-of-control is similar to the coefficient of the chart control limits) values of ARL_2 become similar to ARL_0.

3 Conclusions and final remarks

Economic design of control charts has given emphasis to $\overline{X}$-charts in isolation. The efficiency of these plans, especially the ARL selected plans, are calculated by means of a procedure that does not take into account the effects of the shifts in the process variability. Although the set of parameters used in this paper is limited, some conclusions can be drawn. The present study reveals that when the existence of shifts in the process variability is not considered, erroneous results about the efficiency of the control chart can occur. It can also be observed that, ARL's obtained for large shifts in the process variability are smaller than those obtained in absence of shift. This result seems reasonable, that is, $\overline{X}$-charts will quickly detect current shifts, regardless of the control limits and shifts in process mean.

Acknowledgments: The authors would like to thank CNPq/Brasil for the financial support of this research. Also, our sincere gratitude is extended to the referee for the suggestions.

References

[1] Baker, K.R. (1971). Two Processes in the Economic Design of an $\overline{X}$-chart. *AIIE Transactions*, 3, 257-263.

[2] Von Collani, E., (1985) *Optimal Inspection Intervals*, Mathematische Institut der Universität Würzburg, Preprint 31.

[3] Von Collani, E., (1986) A Simple Procedure to Determine the Economic Design of a Control Chart. *Journal of Quality Technology*, 18, 145-151.

[4] Von Collani, E., (1987) The Economic Design of -Control Charts. *Proceedings of IASTED International Symposium on Reliability and Quality Control.* Paris, 186-189.

[5] Von Collani, E. (1988) Unified Approach to Optimal Process Control. *Metrika*, 35, 145-159.

[6] Del Castillo, E. and Montgomery, D.C. (1993) Optimal Design of Control Charts for Monitoring short production runs. *Journal and Newsletter for Quality and Reliability*, 8, 225-240.

[7] Duncan, A.J. (1971) The Economic Design of X-Charts when there is a Multiplicity of Assignable Causes. *Journal of American Statistical Association*, 66, 107-121.

[8] Gibra, I.N. (1969) Economically Optimal Determination of the Parameters of Control Chart. *Management Science*, 17, 635-646.

[9] Gibra, I.N. (1975). Recent developments in Control Chart Techniques. *Journal of Quality Technology*, 7, 183-192.

[10] Ho, C.; Case, K.E. (1994). Economic Design of Control Charts: A Literature Review for 1981-1991. *Journal of Quality Technology*, 26, 39-53.

[11] Keats, J.B.; Del Castillo, E; von Collani, E. and Saniga, E.M. (1997), Economic Modeling for Statistical Process Control, in: A Discussion on Statistically-Based Process Monitoring and Control. *Journal of Quality Technology*, 29, 144-162.

[12] Ladany, S.P. (1973) Optimal Use of Control Charts for Controlling Current Production. *Management Science*, 19, 763-772.

[13] Ladany, S.P. and Alperovitch, Y. (1975) An Optimal Set-up Policy for Control Charts. *Omega*, 3, 113-118.

[14] Ladany, S.P. and Bedy, D.N. (1976) Selection of the Optimal Set-up Policy. *Naval Research Logistics Quarterly*, 23, 219-233.

[15] Lorenzen, T.J. and Vance, L.C. (1986) The Economic Design of Control Charts: a Unified Approach. *Technometrics* 28, 3-10.

[16] Montgomery, C.C. (1980), The Economic Design of Control charts: A Review and Literature Survey. *Journal of Quality Technology*, 12, 75-87.

[17] Saniga, E.M. (1989) Economic Statistical Control Chart Designs with an Application to $\overline{X}$ and R Charts. *Technometrics*, 31, 313-320.

[18] Saniga, E.M.; Davis, D.J. and McWilliams, T.P. (1995). Economic, Statistical and Economic-Statistical design of attribute charts. *Journal of Quality Technology*, 27, 56-73.

[19] Svoboda, L., (1991) Economic Design of Control Charts: A Review and Literature Survey (1979-1989). In: *Statistical Process Control in Manufacturing*, edited by J.B. Keats and D.C. Montgomery, Marcel Dekker, New York, NY.

[20] Vance, L.C. (1983) A Bibliography of Statistical Quality Control Chart Techniques, 1970-1980. *Journal of Quality Technology*, 15, 59-62.

[21] Woodall, W.H. (1997), Control Charts Based in Attribute Data: Bibliography and Review, *Journal of Quality Technology*, 29, 172-183.

Table 1: Values of ARL_2 and *ratio* for $\delta = 0.5$

n	q	$k = 2.5$		$k = 3.0$		$k = 3.5$		$k = 4.0$	
		ARL2	Ratio	ARL2	Ratio	ARL2	Ratio	ARL2	Ratio
2	0.001	1	26.85	1	90.43	1	379.91	1	2005.38
2	0.05	1.11	24.23	1.14	79.84	1.16	328.08	1.19	1693.09
2	0.1	1.25	21.61	1.31	69.32	1.38	276.96	1.45	1387.93
2	0.15	1.41	19.08	1.53	59.33	1.66	229.2	1.82	1108
2	0.25	1.86	14.48	2.17	41.72	2.57	148.19	3.08	653.5
2	0.5	4.18	6.44	6.34	14.3	10.1	37.72	16.91	118.96
2	0.75	10.26	2.62	22	4.12	52.91	7.2	143.48	14.02
2	1	**26.91**	1	**90.65**	1	**380.97**	1	**2011.8**	1
3	0.001	1	19.37	1	60.54	1	235.97	1	1155.25
3	0.05	1.11	17.48	1.14	53.46	1.16	203.79	1.19	975.41
3	0.1	1.24	15.59	1.31	46.43	1.37	172.11	1.45	799.98
3	0.15	1.41	13.78	1.53	39.78	1.66	142.61	1.81	639.59
3	0.25	1.85	10.51	2.16	28.14	2.55	92.91	3.04	380.79
3	0.5	3.95	4.91	5.9	10.29	9.22	25.66	15.14	76.54
3	0.75	8.62	2.25	17.66	3.44	40.59	5.83	105.2	11.02
3	1	**19.41**	1	**60.69**	1	**236.64**	1	**1158.95**	1
4	0.001	1	14.89	1	43.79	1	160.5	1	738.28
4	0.05	1.11	13.43	1.14	38.67	1.16	138.62	1.19	623.38
4	0.1	1.24	11.99	1.31	33.59	1.37	117.12	1.45	511.5
4	0.15	1.41	10.6	1.52	28.81	1.66	97.16	1.81	409.56
4	0.25	1.84	8.12	2.14	20.51	2.52	63.78	3.01	246.1
4	0.5	3.75	3.98	5.51	7.96	8.48	18.97	13.7	54.08
4	0.75	7.43	2.01	14.67	2.99	32.5	4.95	81.22	9.12
4	1	**14.92**	1	**43.89**	1	**160.95**	1	**740.64**	1
5	0.001	1	11.93	1	33.32	1	115.79	1	504.42
5	0.05	1.11	10.77	1.14	29.42	1.16	100.01	1.19	425.94
5	0.1	1.24	9.61	1.31	25.57	1.37	84.53	1.45	349.66
5	0.15	1.41	8.51	1.52	21.95	1.65	70.21	1.8	280.39
5	0.25	1.83	6.55	2.12	15.72	2.5	46.43	2.98	170.01
5	0.5	3.57	3.35	5.18	6.45	7.86	14.78	12.49	40.5
5	0.75	6.52	1.83	12.49	2.67	26.83	4.33	64.97	7.79
5	1	**11.96**	1	**33.4**	1	**116.11**	1	**506.03**	1
7	0.001	1	8.34	1	21.33	1	67.67	1	268.47
7	0.05	1.11	7.53	1.14	18.84	1.16	58.45	1.19	226.73
7	0.1	1.24	6.72	1.31	16.38	1.37	49.44	1.45	186.3
7	0.15	1.4	5.96	1.52	14.09	1.65	41.17	1.8	149.83
7	0.25	1.81	4.63	2.09	10.21	2.46	27.62	2.91	92.46
7	0.5	3.27	2.56	4.63	4.62	6.84	9.92	10.61	25.38
7	0.75	5.24	1.59	9.54	2.24	19.46	3.49	44.72	6.02
7	1	**8.36**	1	**21.38**	1	**67.86**	1	**269.33**	1
10	0.001	1	5.57	1	12.79	1	36.26	1	128.05
10	0.05	1.11	5.03	1.13	11.3	1.16	31.33	1.19	108.16
10	0.1	1.24	4.49	1.3	9.84	1.37	26.53	1.44	89
10	0.15	1.4	3.99	1.51	8.49	1.64	22.17	1.79	71.89
10	0.25	1.78	3.14	2.05	6.25	2.39	15.19	2.82	45.49
10	0.5	2.91	1.92	4	3.21	5.74	6.33	8.63	14.89
10	0.75	4.06	1.38	6.95	1.85	13.31	2.73	28.7	4.48
10	1	**5.58**	1	**12.83**	1	**36.36**	1	**128.46**	1

Table 2: Values of ARL_2 and *ratio* for $\delta = 1.0$

n	q	$k = 2.5$		$k = 3.0$		$k = 3.5$		$k = 4.0$	
		ARL2	Ratio	ARL2	Ratio	ARL2	Ratio	ARL2	Ratio
2	0.001	1	7.19	1	17.69	1	53.9	1	205.19
2	0.05	1.11	6.49	1.13	15.62	1.16	46.56	1.19	173.3
2	0.1	1.24	5.79	1.3	13.59	1.37	39.4	1.44	142.47
2	0.15	1.4	5.14	1.52	11.7	1.65	32.85	1.79	114.74
2	0.25	1.8	4.01	2.08	8.52	2.44	22.2	2.88	71.42
2	0.5	3.14	2.3	4.39	4.04	6.43	8.41	9.86	20.88
2	0.75	4.78	1.51	8.5	2.09	16.95	3.19	38.09	5.4
2	1	**7.2**	1	**17.73**	1	**54.06**	1	**205.9**	1
3	0.001	1	4.51	1	9.74	1	25.88	1	85.44
3	0.05	1.11	4.07	1.13	8.6	1.16	22.36	1.19	72.18
3	0.1	1.24	3.64	1.3	7.5	1.37	18.95	1.44	59.45
3	0.15	1.39	3.24	1.51	6.48	1.63	15.87	1.78	48.16
3	0.25	1.76	2.57	2.02	4.82	2.35	11.02	2.77	30.97
3	0.5	2.72	1.66	3.68	2.66	5.19	5	7.66	11.18
3	0.75	3.53	1.28	5.85	1.67	10.81	2.4	22.48	3.81
3	1	**4.52**	1	**9.76**	1	**25.95**	1	**85.72**	1
4	0.001	1	3.23	1	6.29	1	14.93	1	43.82
4	0.05	1.11	2.92	1.13	5.56	1.16	12.9	1.19	37.02
4	0.1	1.24	2.61	1.3	4.85	1.37	10.95	1.44	30.55
4	0.15	1.39	2.33	1.5	4.2	1.62	9.22	1.77	24.89
4	0.25	1.72	1.88	1.97	3.2	2.28	6.56	2.66	16.5
4	0.5	2.42	1.34	3.18	1.98	4.36	3.44	6.25	7.03
4	0.75	2.82	1.15	4.41	1.43	7.67	1.95	14.97	2.94
4	1	**3.24**	1	**6.3**	1	**14.97**	1	**43.96**	1
5	0.001	1	2.52	1	4.48	1	9.67	1	25.64
5	0.05	1.11	2.28	1.13	3.96	1.16	8.36	1.19	21.67
5	0.1	1.24	2.04	1.3	3.46	1.36	7.11	1.44	17.92
5	0.15	1.38	1.83	1.49	3.01	1.61	6.01	1.75	14.68
5	0.25	1.69	1.5	1.92	2.34	2.21	4.38	2.57	10.01
5	0.5	2.19	1.15	2.81	1.6	3.76	2.58	5.27	4.88
5	0.75	2.37	1.07	3.53	1.27	5.83	1.66	10.76	2.39
5	1	**2.53**	1	**4.5**	1	**9.7**	1	**25.73**	1
7	0.001	1	1.79	1	2.76	1	5.08	1	11.35
7	0.05	1.11	1.62	1.13	2.44	1.16	4.39	1.19	9.6
7	0.1	1.24	1.45	1.3	2.14	1.36	3.74	1.43	7.96
7	0.15	1.37	1.31	1.48	1.87	1.59	3.19	1.73	6.59
7	0.25	1.63	1.1	1.84	1.5	2.09	2.43	2.41	4.73
7	0.5	1.87	0.96	2.31	1.2	2.98	1.71	4.01	2.84
7	0.75	1.84	0.97	2.53	1.09	3.83	1.33	6.46	1.76
7	1	**1.79**	1	**2.77**	1	**5.09**	1	**11.39**	1
10	0.001	1	1.34	1	1.77	1	2.71	1	4.96
10	0.05	1.11	1.21	1.13	1.56	1.16	2.35	1.19	4.19
10	0.1	1.23	1.09	1.29	1.37	1.35	2.01	1.42	3.5
10	0.15	1.36	0.99	1.45	1.22	1.57	1.74	1.69	2.94
10	0.25	1.55	0.86	1.73	1.02	1.94	1.4	2.2	2.26
10	0.5	1.58	0.85	1.87	0.94	2.31	1.18	2.96	1.68
10	0.75	1.45	0.93	1.82	0.97	2.5	1.09	3.77	1.32
10	1	**1.34**	1	**1.77**	1	**2.72**	1	**4.97**	1

Table 3: Values of ARL_2 and *ratio* for $\delta = 1.5$

n	q	$k = 2.5$		$k = 3.0$		$k = 3.5$		$k = 4.0$	
		ARL2	Ratio	ARL2	Ratio	ARL2	Ratio	ARL2	Ratio
2	0.001	1	2.83	1	5.26	1	12	1	33.07
2	0.05	1.1	2.56	1.1	4.64	1.16	10	1.2	27.94
2	0.1	1.2	2.29	1.3	4.05	1.37	8.7	1.4	23.08
2	0.15	1.4	2.05	1.5	3.52	1.62	7.4	1.8	18.85
2	0.25	1.7	1.66	2	2.7	2.25	5.3	2.6	12.68
2	0.5	2.3	1.24	3	1.77	4.04	3	5.7	5.8
2	0.75	2.6	1.1	3.9	1.34	6.64	1.8	13	2.63
2	1	**2.8**	1	**5.3**	1	**11.91**	1	**33**	1
3	0.001	1	1.85	1	2.9	1	5.4	1	12.39
3	0.05	1.1	1.67	1.1	2.56	1.16	4.7	1.2	10.47
3	0.1	1.2	1.5	1.3	2.24	1.36	4	1.4	8.69
3	0.15	1.4	1.35	1.5	1.97	1.6	3.4	1.7	7.18
3	0.25	1.6	1.13	1.9	1.57	2.11	2.6	2.4	5.13
3	0.5	1.9	0.97	2.4	1.23	3.06	1.8	4.1	3.01
3	0.75	1.9	0.98	2.6	1.11	4.01	1.4	6.8	1.82
3	1	**1.9**	1	**2.9**	1	**5.45**	1	**12**	1
4	0.001	1	1.44	1	2	1	3.2	1	6.28
4	0.05	1.1	1.3	1.1	1.76	1.16	2.8	1.2	5.31
4	0.1	1.2	1.17	1.3	1.55	1.36	2.4	1.4	4.43
4	0.15	1.4	1.06	1.5	1.37	1.58	2.1	1.7	3.7
4	0.25	1.6	0.92	1.8	1.13	1.99	1.6	2.3	2.78
4	0.5	1.7	0.87	2	1	2.49	1.3	3.2	1.95
4	0.75	1.6	0.93	2	1	2.83	1.2	4.4	1.43
4	1	**1.5**	1	**2**	1	**3.24**	1	**6.3**	1
5	0.001	1	1.24	1	1.56	1	2.3	1	3.85
5	0.05	1.1	1.12	1.1	1.38	1.16	2	1.2	3.26
5	0.1	1.2	1.01	1.3	1.22	1.35	1.7	1.4	2.72
5	0.15	1.4	0.92	1.5	1.08	1.55	1.5	1.7	2.3
5	0.25	1.5	0.82	1.7	0.93	1.89	1.2	2.1	1.81
5	0.5	1.5	0.83	1.8	0.89	2.12	1.1	2.7	1.44
5	0.75	1.4	0.92	1.7	0.95	2.19	1	3.2	1.21
5	1	**1.2**	1	**1.6**	1	**2.26**	1	**3.9**	1
7	0.001	1	1.07	1	1.2	1	1.5	1	2.04
7	0.05	1.1	0.97	1.1	1.06	1.16	1.3	1.2	1.73
7	0.1	1.2	0.88	1.3	0.94	1.34	1.1	1.4	1.46
7	0.15	1.3	0.81	1.4	0.85	1.52	1	1.6	1.26
7	0.25	1.4	0.75	1.6	0.76	1.73	0.9	1.9	1.07
7	0.5	1.3	0.83	1.5	0.82	1.69	0.9	2	1.01
7	0.75	1.2	0.93	1.3	0.92	1.57	0.9	2	1.01
7	1	**1.1**	1	**1.2**	1	**1.47**	1	**2.1**	1
10	0.001	1	1.01	1	1.04	1	1.1	1	1.29
10	0.05	1.1	0.91	1.1	0.92	1.16	1	1.2	1.1
10	0.1	1.2	0.83	1.3	0.82	1.33	0.8	1.4	0.93
10	0.15	1.3	0.78	1.4	0.76	1.47	0.8	1.6	0.83
10	0.25	1.3	0.76	1.4	0.72	1.56	0.7	1.7	0.76
10	0.5	1.2	0.88	1.2	0.84	1.36	0.8	1.6	0.84
10	0.75	1.1	0.97	1.1	0.94	1.21	0.9	1.4	0.92
10	1	**1**	1	**1**	1	**1.12**	1	**1.3**	1

Table 4: Values of ARL_2 and *ratio* for $\delta = 2.0$

n	q	$k = 2.5$		$k = 3.0$		$k = 3.5$		$k = 4.0$	
		ARL2	Ratio	ARL2	Ratio	ARL2	Ratio	ARL2	Ratio
2	0.001	1	1.59	1	2.31	1	3.97	1	8.26
2	0.05	1.1	1.43	1.13	2.04	1.2	3.44	1.19	6.99
2	0.1	1.2	1.29	1.29	1.79	1.4	2.94	1.43	5.81
2	0.15	1.4	1.16	1.47	1.58	1.6	2.52	1.71	4.83
2	0.25	1.6	0.99	1.8	1.29	2	1.95	2.33	3.55
2	0.5	1.8	0.91	2.14	1.08	2.7	1.47	3.58	2.31
2	0.75	1.7	0.95	2.23	1.04	3.3	1.22	5.27	1.57
2	1	**1.6**	1	**2.32**	1	**4**	1	**8.29**	1
3	0.001	1	1.2	1	1.47	1	2.05	1	3.37
3	0.05	1.1	1.08	1.13	1.3	1.2	1.78	1.19	2.85
3	0.1	1.2	0.98	1.29	1.15	1.4	1.53	1.42	2.39
3	0.15	1.4	0.89	1.44	1.02	1.6	1.33	1.67	2.02
3	0.25	1.5	0.8	1.67	0.88	1.9	1.11	2.09	1.61
3	0.5	1.5	0.82	1.69	0.87	2	1.02	2.54	1.33
3	0.75	1.3	0.92	1.57	0.94	2	1.01	2.91	1.16
3	1	**1.2**	1	**1.47**	1	**2.1**	1	**3.38**	1
4	0.001	1	1.07	1	1.19	1	1.44	1	1.99
4	0.05	1.1	0.97	1.13	1.05	1.2	1.25	1.18	1.69
4	0.1	1.2	0.88	1.28	0.93	1.3	1.08	1.4	1.42
4	0.15	1.3	0.81	1.42	0.84	1.5	0.95	1.63	1.23
4	0.25	1.4	0.75	1.57	0.76	1.7	0.84	1.91	1.05
4	0.5	1.3	0.83	1.45	0.82	1.7	0.87	2	1
4	0.75	1.2	0.93	1.29	0.92	1.6	0.93	2	1
4	1	**1.1**	1	**1.19**	1	**1.5**	1	**2**	1
5	0.001	1	1.02	1	1.07	1	1.19	1	1.46
5	0.05	1.1	0.93	1.13	0.95	1.2	1.04	1.18	1.24
5	0.1	1.2	0.84	1.27	0.85	1.3	0.9	1.39	1.05
5	0.15	1.3	0.78	1.39	0.77	1.5	0.81	1.59	0.92
5	0.25	1.4	0.75	1.48	0.73	1.6	0.74	1.77	0.83
5	0.5	1.2	0.86	1.3	0.83	1.5	0.82	1.69	0.87
5	0.75	1.1	0.95	1.16	0.93	1.3	0.92	1.57	0.94
5	1	**1**	1	**1.08**	1	**1.2**	1	**1.47**	1
7	0.001	1	1	1	1.01	1	1.04	1	1.11
7	0.05	1.1	0.91	1.13	0.89	1.2	0.9	1.18	0.94
7	0.1	1.2	0.83	1.26	0.8	1.3	0.79	1.37	0.81
7	0.15	1.3	0.79	1.35	0.75	1.4	0.73	1.52	0.73
7	0.25	1.3	0.79	1.36	0.74	1.5	0.71	1.57	0.71
7	0.5	1.1	0.92	1.14	0.88	1.2	0.85	1.35	0.82
7	0.75	1	0.98	1.04	0.97	1.1	0.95	1.2	0.92
7	1	**1**	1	**1.01**	1	**1**	1	**1.11**	1
10	0.001	1	1	1	1	1	1	1	1.01
10	0.05	1.1	0.91	1.13	0.89	1.2	0.87	1.18	0.86
10	0.1	1.2	0.84	1.24	0.81	1.3	0.78	1.35	0.75
10	0.15	1.2	0.81	1.3	0.77	1.4	0.74	1.43	0.7
10	0.25	1.2	0.84	1.24	0.81	1.3	0.77	1.38	0.73
10	0.5	1	0.97	1.05	0.95	1.1	0.92	1.14	0.89
10	0.75	1	1	1.01	0.99	1	0.99	1.04	0.97
10	1	**1**	1	**1**	1	**1**	1	**1.01**	1

Table 5: Values of ARL_2 and *ratio* for $\delta = 2.5$

n	q	$k = 2.5$		$k = 3.0$		$k = 3.5$		$k = 4.0$	
		ARL2	Ratio	ARL2	Ratio	ARL2	Ratio	ARL2	Ratio
2	0.001	1	1.17	1	1.42	1	1.94	1	3.1
2	0.05	1.11	1.06	1.13	1.25	1.16	1.68	1.18	2.6
2	0.1	1.23	0.96	1.29	1.11	1.35	1.44	1.41	2.2
2	0.15	1.34	0.88	1.44	0.99	1.54	1.26	1.66	1.9
2	0.25	1.5	0.79	1.65	0.86	1.84	1.06	2.07	1.5
2	0.5	1.43	0.82	1.65	0.86	1.97	0.99	2.45	1.3
2	0.75	1.28	0.92	1.52	0.93	1.96	0.99	2.75	1.1
2	1	**1.18**	1	**1.42**	1	**1.94**	1	**3.11**	1
3	0.001	1	1.03	1	1.1	1	1.25	1	1.6
3	0.05	1.11	0.93	1.13	0.97	1.16	1.08	1.18	1.3
3	0.1	1.22	0.85	1.27	0.86	1.33	0.94	1.4	1.1
3	0.15	1.31	0.79	1.4	0.79	1.49	0.84	1.6	1
3	0.25	1.39	0.75	1.51	0.73	1.65	0.76	1.81	0.9
3	0.5	1.22	0.85	1.34	0.82	1.51	0.83	1.77	0.9
3	0.75	1.09	0.95	1.19	0.93	1.36	0.92	1.67	1
3	1	**1.03**	1	**1.1**	1	**1.26**	1	**1.59**	1
4	0.001	1	1	1	1.02	1	1.07	1	1.2
4	0.05	1.11	0.91	1.13	0.9	1.16	0.93	1.18	1
4	0.1	1.21	0.83	1.26	0.81	1.32	0.81	1.38	0.9
4	0.15	1.29	0.78	1.36	0.75	1.45	0.74	1.54	0.8
4	0.25	1.31	0.77	1.4	0.73	1.51	0.71	1.64	0.7
4	0.5	1.12	0.9	1.19	0.86	1.29	0.83	1.45	0.8
4	0.75	1.03	0.98	1.07	0.95	1.15	0.93	1.29	0.9
4	1	**1.01**	1	**1.02**	1	**1.07**	1	**1.19**	1
5	0.001	1	1	1	1	1	1.02	1	1.1
5	0.05	1.11	0.91	1.13	0.89	1.15	0.88	1.18	0.9
5	0.1	1.2	0.83	1.25	0.8	1.31	0.78	1.37	0.8
5	0.15	1.26	0.79	1.33	0.75	1.41	0.72	1.49	0.7
5	0.25	1.25	0.8	1.32	0.76	1.41	0.72	1.51	0.7
5	0.5	1.07	0.94	1.11	0.91	1.17	0.87	1.27	0.8
5	0.75	1.01	0.99	1.03	0.98	1.06	0.96	1.13	0.9
5	1	**1**	1	**1**	1	**1.02**	1	**1.06**	1
7	0.001	1	1	1	1	1	1	1	1
7	0.05	1.1	0.91	1.13	0.89	1.15	0.87	1.18	0.9
7	0.1	1.19	0.84	1.24	0.81	1.29	0.78	1.34	0.8
7	0.15	1.22	0.82	1.28	0.78	1.34	0.75	1.41	0.7
7	0.25	1.16	0.86	1.21	0.83	1.27	0.79	1.34	0.8
7	0.5	1.02	0.98	1.04	0.96	1.06	0.94	1.11	0.9
7	0.75	1	1	1	1	1.01	0.99	1.03	1
7	1	**1**	1	**1**	1	**1**	1	**1**	1
10	0.001	1	1	1	1	1	1	1	1
10	0.05	1.1	0.91	1.12	0.89	1.15	0.87	1.17	0.9
10	0.1	1.17	0.85	1.21	0.83	1.25	0.8	1.3	0.8
10	0.15	1.18	0.85	1.22	0.82	1.27	0.79	1.32	0.8
10	0.25	1.09	0.92	1.12	0.89	1.15	0.87	1.19	0.8
10	0.5	1	1	1.01	0.99	1.01	0.99	1.03	1
10	0.75	1	1	1	1	1	1	1	1
10	1	**1**	1	**1**	1	**1**	1	**1**	1

Table 6: Values of ARL_2 and *ratio* for $\delta = 3.0$

n	q	$k = 2.5$		$k = 3.0$		$k = 3.5$		$k = 4.0$	
		ARL2	Ratio	ARL2	Ratio	ARL2	Ratio	ARL2	Ratio
2	0.001	1	1.04	1	1.12	1	1.29	1	1.67
2	0.05	1.11	0.94	1.13	0.99	1.16	1.12	1.18	1.42
2	0.1	1.22	0.85	1.28	0.88	1.34	0.97	1.4	1.2
2	0.15	1.32	0.79	1.4	0.8	1.5	0.87	1.61	1.05
2	0.25	1.4	0.74	1.52	0.74	1.67	0.78	1.84	0.91
2	0.5	1.24	0.84	1.36	0.82	1.55	0.84	1.82	0.92
2	0.75	1.11	0.94	1.21	0.92	1.41	0.92	1.75	0.96
2	1	**1.04**	1	**1.12**	1	**1.3**	1	**1.68**	1
3	0.001	1	1	1	1.01	1	1.04	1	1.13
3	0.05	1.11	0.91	1.13	0.9	1.16	0.91	1.18	0.96
3	0.1	1.21	0.83	1.26	0.8	1.32	0.8	1.38	0.82
3	0.15	1.28	0.78	1.35	0.75	1.44	0.73	1.53	0.74
3	0.25	1.29	0.78	1.37	0.74	1.47	0.71	1.59	0.71
3	0.5	1.1	0.91	1.16	0.88	1.25	0.84	1.38	0.82
3	0.75	1.02	0.98	1.05	0.96	1.11	0.94	1.23	0.92
3	1	**1**	1	**1.01**	1	**1.05**	1	**1.13**	1
4	0.001	1	1	1	1	1	1	1	1.02
4	0.05	1.11	0.91	1.13	0.89	1.15	0.87	1.18	0.87
4	0.1	1.2	0.83	1.25	0.8	1.3	0.77	1.36	0.76
4	0.15	1.25	0.8	1.31	0.76	1.38	0.73	1.46	0.7
4	0.25	1.21	0.83	1.27	0.79	1.35	0.75	1.43	0.71
4	0.5	1.04	0.96	1.07	0.93	1.12	0.9	1.19	0.86
4	0.75	1	1	1.01	0.99	1.03	0.98	1.07	0.95
4	1	**1**	1	**1**	1	**1.01**	1	**1.02**	1
5	0.001	1	1	1	1	1	1	1	1
5	0.05	1.1	0.91	1.13	0.89	1.15	0.87	1.18	0.85
5	0.1	1.19	0.84	1.23	0.81	1.28	0.78	1.34	0.75
5	0.15	1.22	0.82	1.28	0.78	1.34	0.75	1.4	0.71
5	0.25	1.16	0.86	1.2	0.83	1.26	0.79	1.33	0.76
5	0.5	1.02	0.98	1.03	0.97	1.06	0.95	1.1	0.92
5	0.75	1	1	1	1	1.01	0.99	1.02	0.98
5	1	**1**	1	**1**	1	**1**	1	**1**	1
7	0.001	1	1	1	1	1	1	1	1
7	0.05	1.1	0.91	1.12	0.89	1.15	0.87	1.17	0.85
7	0.1	1.17	0.85	1.21	0.83	1.25	0.8	1.3	0.77
7	0.15	1.17	0.85	1.22	0.82	1.27	0.79	1.32	0.76
7	0.25	1.09	0.92	1.12	0.89	1.15	0.87	1.19	0.84
7	0.5	1	1	1.01	0.99	1.01	0.99	1.03	0.98
7	0.75	1	1	1	1	1	1	1	1
7	1	**1**	1	**1**	1	**1**	1	**1**	1
10	0.001	1	1	1	1	1	1	1	1
10	0.05	1.1	0.91	1.12	0.89	1.14	0.88	1.17	0.86
10	0.1	1.15	0.87	1.18	0.85	1.22	0.82	1.25	0.8
10	0.15	1.13	0.89	1.16	0.87	1.19	0.84	1.23	0.82
10	0.25	1.04	0.96	1.05	0.95	1.07	0.93	1.09	0.92
10	0.5	1	1	1	1	1	1	1	1
10	0.75	1	1	1	1	1	1	1	1
10	1	**1**	1	**1**	1	**1**	1	**1**	1

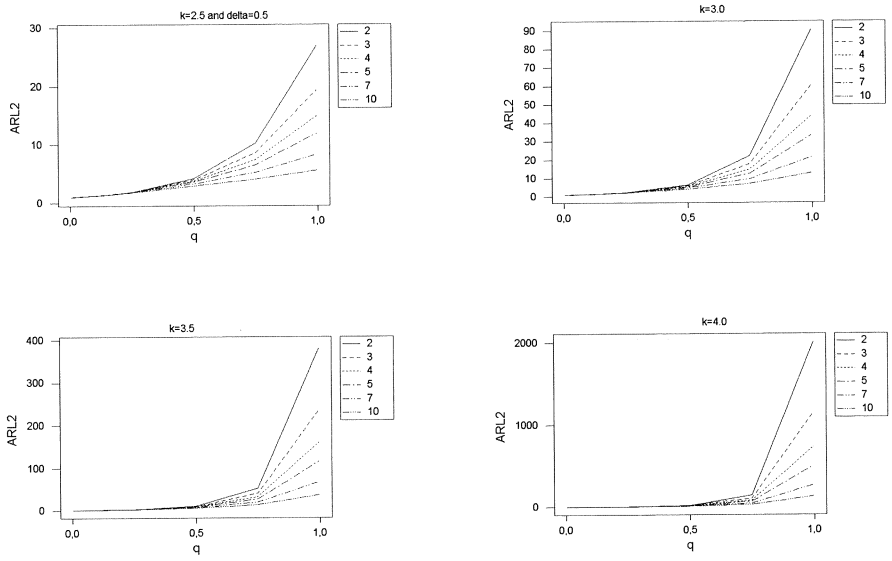


Figure 1: Values of ARL_2 versus q when $\delta = 0.5$

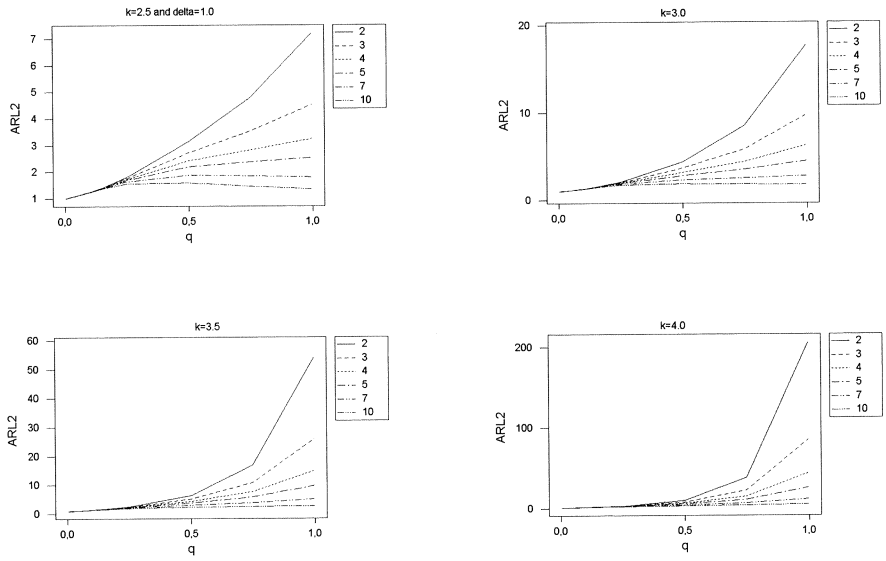


Figure 2: Values of ARL_2 versus q when $\delta = 1.0$

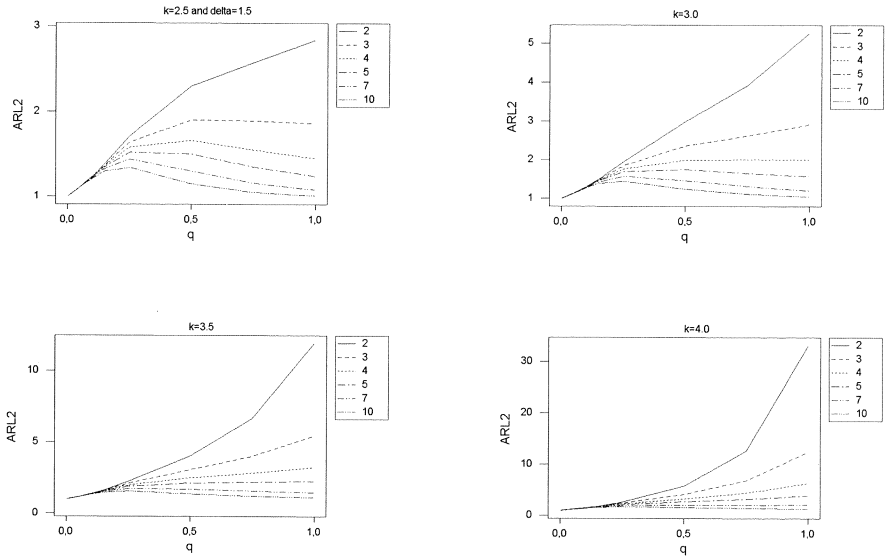


Figure 3: Values of ARL_2 versus q when $\delta = 1.5$

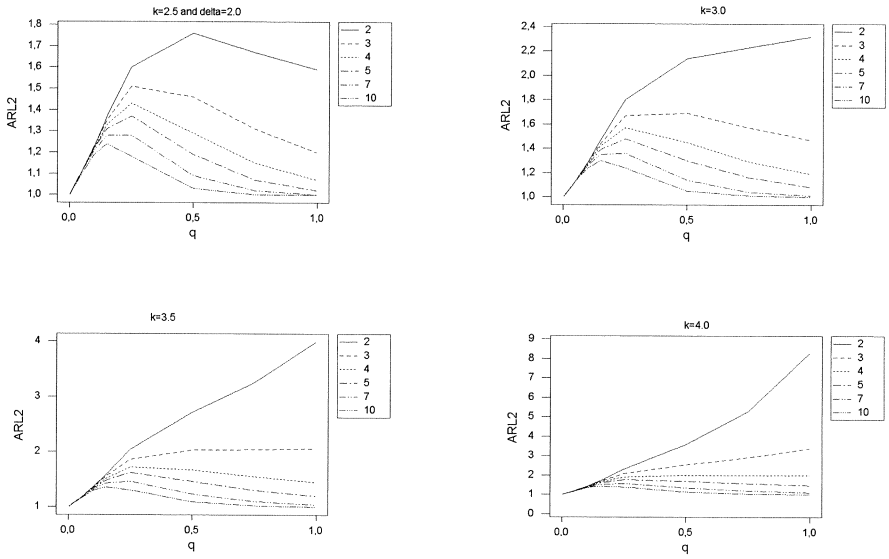


Figure 4: Values of ARL_2 versus q when $\delta = 2.0$

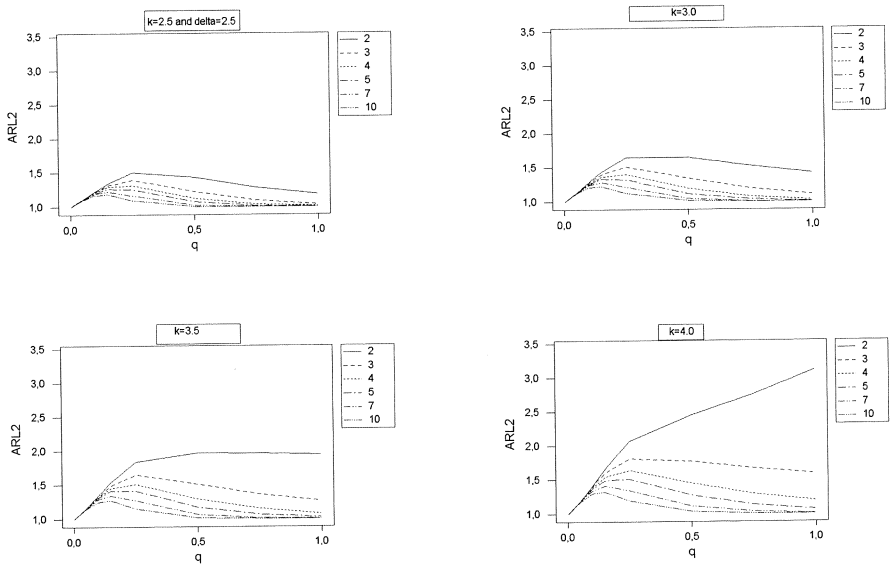


Figure 5: Values of ARL_2 versus q when $\delta = 2.5$

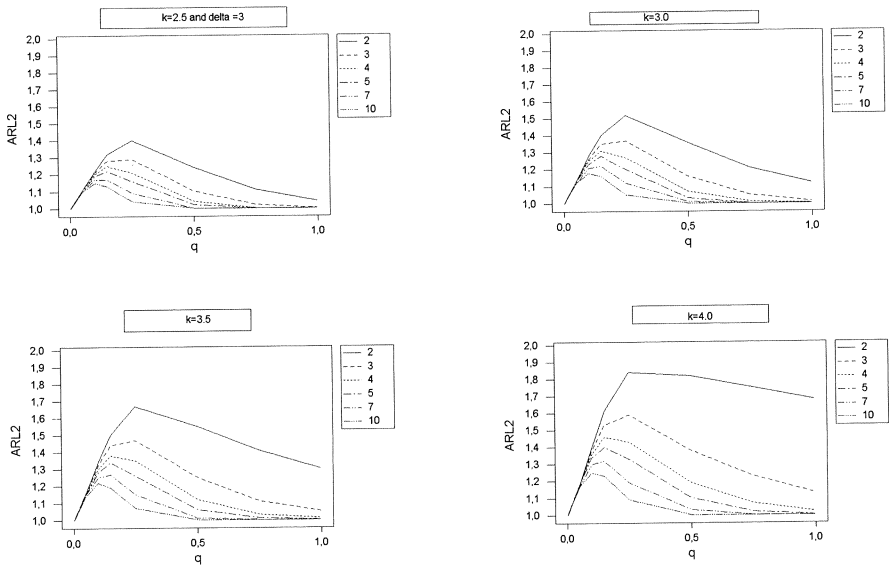


Figure 6: Values of ARL_2 versus q when $\delta = 3.0$

Student (2004), Vol. 5 No. 2, 113-123

Partial Reweighted Least Squares

Athanassios Kondylis

Groupe de Statistique
Université de Neuchâtel
Espace de l'Europe, 4
CH-2002 Neuchâtel, Suisse
Email: Atanassios.Kondylis@unine.ch

Abstract : Partial Least Squares Regression is a derived components regression method which has attracted a lot of attention especially in application fields like chemometrics. We propose here the Partial Reweighted Least Squares (PRLS) algorithm which we illustrate on regression problems. The use of PRLS regression as a less sensitive to outliers alternative is demonstrated. Finally, the Akaike Information Criterion (AIC) is used for model selection.

Key words : Partial Least Squares Regression, outliers, model selection.

1 Introduction

Partial Least Squares (PLS) were originated by Herman Wold [23] . Since then PLS attired mostly the interest of chemometricians who implemented PLS in complex and sophisticated settings (eg inverse regression problems and multivariate calibration). Numerous articles can be found in chemometric journals, amongst them we notice [22] as well as [11] and [7] . PLS however did not lack the attention of the statistical society. PLS properties and the new perspectives arising from its applications, either in classification or in regression setting, were the central interest of relatively recent work. Stone, M. and Brooks [18] introduced the *continuum regression* which is a generalized regression family where ordinary linear regression, principal components regression and partial least squares regression are special cases. Frank and Friedman [6] compared the three above methods together with Ridge Regression (RR). The latter was further motivated by the shrinkage properties of PLS regression illustrated in [8] . The relation of RR with continuum regression was further analyzed in [3] and

[20] . PLS, though is not only restricted to regression problems. PLS discriminant analysis and classification are the main interest of new researchers. For a global review of PLS, its properties, and its relation to other multivariate techniques one can see [21] .

In this article we are concerned with one response PLS regression modelling, denoted as PLS1. PLS1 is a special case of PLS. PLS methods use a linear multivariate model in order to relate two data matrices, that is, the predictors $\mathbf{X}$ and the responses $\mathbf{Y}$. PLS1 corresponds to the case where $\mathbf{Y}$ is a single column (noted from now on as $\mathbf{y}$), while $\mathbf{X}$ is the matrix of many interrelated variables. Such modelling strategies envisage to provide rigorous statistical models which mainly focus in reducing the space generated by the big number of correlated predictors in lower dimensional space generated by the derived components. The latter are extracted using the PLS algorithm.

PLS regression is sensitive to outliers. PLS regression's sensitivity to outliers is due to the use of least squares estimates in regression, and the use of Pearson's product-moment estimate for the covariance in PLS algorithm (analytically described in Section 2).

The need for less sensitive to outliers PLS methods has been initially motivated by the work done within chemometricians. For a quick review one can see [9] . Recently, the need for such procedures has given rise to new research and interest (see [12] and [4]). PRLS regression, as described throughout this paper, is further analyzed in [14] .

We propose in this paper the use of a modified algorithm which retains PLS structure while it protects the regression results from outliers. The latter is achieved by using robust regression, and by using weights in the estimation of the covariance. The Akaike Information Criterion is finally used to select the final model. The method described here is mainly applicable to fat data, that is, data sets where the number of the recorded variables exceed the number of the observations.

This article is organized as follows. Section 2 reviews PLS regression for one response $\mathbf{y}$, and PLS1 regression algorithm is described. Section 3 presents the Partial Reweighted Least Squares (PRLS) regression together with some theoretical considerations. PRLS algorithm is compared to PLS1. Section 4 discusses the model selection criteria to be used in Section 5 where real data sets test the efficiency of PRLS. Finally, in the last section conclusions are drawn.

Notation

The following notations will be used below to describe the PLS1 and PRLS methods:

-k denotes the current iteration of the algorithm,

-p is the number of the predictor labelled by $j = 1, 2, ..., p$,

-n is the number of observations labelled by $i = 1, 2, ..., n$,

-$E(\mathbf{x}_j)$ is the expected value for variable $\mathbf{x}_j$,

-$s_{\mathbf{x}_j}$ is the standard deviation for variable $\mathbf{x}_j$, and $s_{\mathbf{y}}$ the standard deviation for the response $\mathbf{y}$,

-$\mathbf{X}_k$ denotes the matrix of predictors $(\mathbf{x}_{1,k}, \mathbf{x}_{2,k}, ..., \mathbf{x}_{p,k})$ at iteration k,
-$\mathbf{y}_k$ denotes the response vector at iteration k,
-$\mathbf{q}$ denotes the regression parameter vector $(\mathbf{q}_1, ..., \mathbf{q}_k)$,
-$\mathbf{T}_{1:k}$ denotes the nxk matrix with each column corresponding to successive derived component vectors, we note $\mathbf{T}_{1:k} = (\mathbf{t}_1, ..., \mathbf{t}_k)$,
-$\mathbf{w}$ denotes the pxk weight matrix where, for each iteration k, its elements are calculated according to $\mathbf{w}_k = \{cov(\mathbf{x}_{1,k}, \mathbf{y}_k), ..., cov(\mathbf{x}_{p,k}, \mathbf{y}_k)\}$, with
$cov(\mathbf{x}_j, \mathbf{y}) = E(\mathbf{x}_j\mathbf{y}) - E(\mathbf{x}_j)E(\mathbf{y})$,
-$\mathbf{h}_\mathbf{k}^{\mathbf{x_j}}$ and $\mathbf{h}_\mathbf{k}^{\mathbf{y}}$ are weight vectors pooled in the matrix $\mathbf{H}_k$ for each iteration k, as it will be seen in PRLS regression algorithm,
-$E(\mathbf{y}|\mathbf{t}_k)$ is the least square fit resulting from the regression of $\mathbf{y}$ on the k derived component $\mathbf{t_k}$,
-$E(\mathbf{y}_k|\mathbf{t}_1, ..., \mathbf{t}_k) = E(\mathbf{y}_k|\mathbf{T}_{1:k})$ is the least square fit of $\mathbf{y}$ on the k mutually orthogonal derived components $(\mathbf{t}_1, ..., \mathbf{t}_k)$.

2 Review of PLS1

PLS1 proposes a new, flexible and efficient alternative method in regression modelling. In contrast to Multiple Linear Regression (MLR), PLS regression models are not linear on the explanatory variables $\mathbf{x}_j$, but on the orthogonal derived components which are called factors and are noted as $\mathbf{t}$. These components are extracted using the PLS algorithm. PLS1 thus produce orthogonal factors as linear combinations of the original predictor variables $\mathbf{x}_j$. So we get

$$t_{i,k} = w_{j,k} x_{ij,k-1},$$

where $x_{ij,k-1}$ represents the value of the j^{th} $\mathbf{X}$ variable for the i^{th} observation and $w_{j,k}$ the coefficients of the linear transformations in $\mathbf{X}$. The subscript k indicates the algorithm's iterative step and will be more clear at the description of the PLS1 algorithm that follows. The components $\mathbf{t}_k = (t_{i,k})$ are used as regressors in the regression model

$$\widehat{\mathbf{y}} = \mathbf{T}_{1:k}\widehat{\mathbf{q}}_k,$$

where $\widehat{\mathbf{y}}$ is the fitted values vector, $\widehat{\mathbf{q}}_k$ denotes the estimated regression parameter vector at iteration k, and $\mathbf{T}_{1:k}$ is the regression's design matrix where the $X's$ are replaced by the derived components. The index $(1 : k)$ indicates that we essentially build successively additive models of the form $E(\mathbf{y} \mid \mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_k)$ for each iteration k in the PLS1 algorithm.

PLS1 Algorithm

- scale or standardize all the variables and give the new set $(\mathbf{X}, \mathbf{y})$.
  $k \leftarrow 1$.

$\mathbf{X} \leftarrow \mathbf{X_0}$, $\mathbf{y} \leftarrow \mathbf{y_0}$.

Step 1: Calculate vector $\mathbf{w}_k = cov(\mathbf{x}_{j,k-1}, \mathbf{y}_{k-1})$ and normalize in order to have $\mathbf{w}_k'\mathbf{w}_k = 1$.

Step 2: Extract $\mathbf{t}_k = \mathbf{X}_{k-1}\mathbf{w}_k$.

Step 3: 1. Regress $\mathbf{y}_{k-1}$ to $\mathbf{t}_k$, calculate the residuals

$$\mathbf{y}_k = \mathbf{y}_{k-1} - E(\mathbf{y}_{k-1}|\mathbf{t}_k).$$

2. Regress each $\mathbf{x}_{j,k-1}$ to $\mathbf{t}_k$, calculate for each one the residuals

$$\mathbf{x}_{j,k} = \mathbf{x}_{j,k-1} - E(\mathbf{x}_{j,k-1}|\mathbf{t}_k),$$

where $j = 1, 2, .., p$.

Step 4: Use Ordinary Least Squares and regress the initial $\mathbf{y}$ on the extracted components (PLS1 regression step).
Give the regression model $\widehat{\mathbf{y}}_k = \mathbf{T}_{1:k}\widehat{\mathbf{q}}_k = E(\mathbf{y}_k|\mathbf{t}_1, ..., \mathbf{t}_k)$.

- If a certain stopping criterion (to be discussed in what follows) is satisfied, give the final PLS1 regression model $\widehat{\mathbf{y}} = \mathbf{T}_{1:k}\widehat{\mathbf{q}}_k$.

  If not, return to *Step 1* with $k \leftarrow k + 1$.

Brief description of the PLS1 algorithm

The algorithm above starts by properly transforming data. From the initial data set $(\mathbf{X}_0, \mathbf{y}_0)$ weights $\mathbf{w}_1$ are calculated as the covariance of each predictor with the response. The weights are then used to extract the first PLS component. Afterwards, successive regressions of the first component $\mathbf{t}_1$ on $\mathbf{y}$ and each $\mathbf{x}_j$ deflate the data. Yet, deflating for $\mathbf{y}$ is optional and the PLS1 algorithm is simplified by deletion of *Step 3.1*. The residuals of the regressions in *Step 4* are stored in a new data matrix which will serve as the starting point (*Step 1*) of the new iteration. The algorithm continuous for $k = 2$ and so forth. At the end of the PLS1 algorithm regressions of the extracted components $(\mathbf{t}_1, ..., \mathbf{t}_k)$ on $\mathbf{y}$ are run in order to build the final PLS regression model.

Example - An Orthogonal Design

PLS1 method is a generalization of MLR. The latter is justified using a regression problem where no correlation between the predictors occur, that is an orthogonal design.

Regression analysis for the data in Table 1 ($\mathbf{y}$ as the response and two explanatory variables $\mathbf{x}_1$ and $\mathbf{x}_2$) lead to the following least squares estimates:

$$\widehat{\mathbf{y}} = 11.40 - 1.8571\mathbf{x}_1 + 0.1408\mathbf{x}_2$$

$\mathbf{y}$	$\mathbf{x}_1$	$\mathbf{x}_2$
18	-2	4
12	1	3
10	0	-6
16	-1	-5
11	2	3
9	0	-3
11	-1	6
8	1	-1
7	1	0
12	-1	-1

Table 1: Orthogonal Design

Using PLS1 algorithm and regressing the response on the derived components we are led to exactly the same estimates with only one component used, that is, by $\widehat{\mathbf{y}} = \mathbf{t}_1\widehat{\mathbf{q}}_1$. We finally get

$$\widehat{\mathbf{y}} = \mathbf{T}_1\widehat{\mathbf{q}}_1 = \mathbf{X}\mathbf{w}_0\widehat{\mathbf{q}}_1 = \mathbf{X}\widehat{\beta}_{PLS}.$$

Thus for standardized data the implied regression coefficients are

$$\widehat{\beta}_{PLS} = (-0.68, 0.164),$$

which, transformed back, equal to $(11, 40, -1.8571, 0.1408)$. The latter are the MLR estimates.

3 Partial Reweighted Least Squares Regression

PLS1 is modified in order to create an algorithm less sensitive to outliers. This is achieved by two main operations. Firstly, by robust estimation of the covariances, and secondly by robust regression of $\mathbf{y}$ on the set of the derived components $(\mathbf{t}_1, ..., \mathbf{t}_k)$. The former is achieved by the use of a reweighted covariance estimate, and the latter by the use of a robust regression method.

PRLS regression uses a weight vector $\mathbf{h_k}$ in order to classify at each iteration the observations in *outliers* and *non outliers*. The latter is obtained while regressing the predictors and the response on the derived components using Least Trimmed Squares (LTS). The weight vector $\mathbf{h_k}$ is calculated according to

$$h_{i,k} = \begin{cases} 1 & \text{if } \epsilon_i \leq c, \\ 0 & \text{if } \epsilon_i \text{ otherwise,} \end{cases} \tag{1}$$

where ϵ_i are the scaled residuals, and c is a critical value, usually set to 2.5. For more information about LTS and the choice of the critical value one can see [13]. The weight vectors construct, at each iteration k, the weight matrix $\mathbf{H_k} = \{\mathbf{h_k^{x_1}}, ..., \mathbf{h_k^{x_p}}, \mathbf{h_k^{y}}\}$. Ma-

trix $\mathbf{H_k}$ is used in order to compute the weighted covariance estimate

$$cov(\mathbf{x}_j, \mathbf{y}) = \frac{\sum_{i=1}^{n} h_i^{\mathbf{x_j}}(x_{ij} - \bar{x}_j^c)\ h_i^{\mathbf{y}}(y_i - \bar{y}^c)}{(\sum_{i=1}^{n} h_i^{\mathbf{x_j}} h_i^{\mathbf{y}}) - 1}, \tag{2}$$

where

$$\bar{x}_j^c = \frac{\sum_{i=1}^{n} h_i^{\mathbf{x}_j}\, x_{ij}}{\sum_{i=1}^{n} h_i^{\mathbf{x}_j}} \quad \text{and} \quad \bar{y}^c = \frac{\sum_{i=1}^{n} h_i^{\mathbf{y}}\, y_i}{\sum_{i=1}^{n} h_i^{\mathbf{y}}}, \tag{3}$$

and to extract the PRLS components for the following iteration. The subscript k is omitted from (2) and (3) for notational convenience.

At the final step of the PRLS algorithm the response $\mathbf{y}$ is regressed on the whole set of the extracted components $\mathbf{T}_{1:k}$ using LTS instead of ordinary least squares which is the case for PLS1. The PRLS regression algorithm is sketched below.

PRLS Regression Algorithm

- scale or standardize all the variables and give the new set $(\mathbf{X}, \mathbf{y})$.
  $k \leftarrow 1$.
  $\mathbf{X} \leftarrow \mathbf{X_0}$, $\mathbf{y} \leftarrow \mathbf{y_0}$.

Step 1: Calculate vector $\mathbf{w}_k$ according to (2) and (3). Normalize in order to have $\mathbf{w}_k'\mathbf{w}_k = 1$.

Step 2: Extract $\mathbf{t}_k = \mathbf{X}_{k-1}\mathbf{w}_k$.

Step 3: 1a. Regress $\mathbf{y}_{k-1}$ on $\mathbf{t}_k$ using LTS and take $h_{i,k}^{\mathbf{y}}$ as in (1).
1b. Regress $\mathbf{y}_{k-1}$ on $\mathbf{t}_k$ using OLS and calculate the residuals

$$\mathbf{y}_k = \mathbf{y}_{k-1} - E(\mathbf{y}_{k-1}|\mathbf{t}_k).$$

2a. Regress each $\mathbf{x}_{j,k-1}$ on $\mathbf{t}_k$ using LTS and take $h_{i,k}^{\mathbf{x}_j}$ as in (1).
2b. Regress each $\mathbf{x}_{j,k-1}$ to $\mathbf{t}_k$ using OLS and calculate the residuals

$$\mathbf{x}_{j,k} = \mathbf{x}_{j,k-1} - E(\mathbf{x}_{j,k-1}|\mathbf{t}_k),$$

for each $j = 1, 2, .., p$.

Step 4: Use LTS to regress the initial $\mathbf{y}$ on the extracted components.
Give the regression model $\widehat{\mathbf{y}}_k = \mathbf{T}_{1:k}\widehat{\mathbf{q}}_k$.

- If a certain stopping criterion is satisfied, give the final PRLS regression model $\widehat{\mathbf{y}} = \mathbf{T}_{1:k}\widehat{\mathbf{q}}_k$.

  If not, return to *Step 1* with $k \leftarrow k + 1$.

Remarks on PRLS Regression Algorithm

The PRLS regression algorithm is a modified PLS1 algorithm which is more resistant to perturbations due to outliers. It retains orthogonal components since it deflates data using least squares, yet it uses the vector $\mathbf{h}_k$ in order to protect from outliers (on both X and y directions), and robust regression in order to obtain the regression model estimates.

At the first iteration the covariance ($cov(\mathbf{x_j}, \mathbf{y}) = E(\mathbf{x_j}\,\mathbf{y}) - E(\mathbf{x_j})E(\mathbf{y})$) is computed by leaving out of the $E(.)$ a proportion δ of the initial values. This trimming constant δ is a number less than 0.5 which gives the proportion trimmed in the calculations for $cov(\mathbf{x_j}, \mathbf{y})$. It should be a number larger than the suspected fraction of outliers. A value of $\delta = 0.15$ is generally accepted when no prior information is available. When prior information exists it should be taken into account for constructing δ.

The PRLS regression algorithm is mainly suitable in cases where a large number of variables $\mathbf{x}$ are available. PRLS is designed especially for fat data where the number of the variables exceed the number of the observations. The latter is commonly the case in PLS regression modelling.

Finally we notice that the algorithm above can be further extended by using a smooth function of weights $\mathbf{h}_k$ in order to downweight the outliers instead of a critical value c. Different robust regression methods could be employed in order to do so (for example M-regression).

4 Selecting the final model

By model selection we mean the process of selecting the number of components which will be retained in the final model. The latter should at least guarantee a generalization performance to new data sets. The assessment of this performance is very important in practice and various methods are implemented. Cross-validation, the bootstrap, and penalized maximum likelihood criteria are commonly used in order to select the final model.

In **cross-validation** the data set is split into the training set (denoted as S_{train}), and the test set (denoted as S_{test}). In the former models are constructed, while in the latter the models are tested. The model, built on the training set, which minimises a chosen loss function, computed in the test set, is finally selected. The Root Mean Squared Error (RMSE) is a commonly used loss function. For a model containing k components, the RMSE is computed according to

$$RMSE_k = \sqrt{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \widehat{y}_{i,k})^2},$$

where the value $\widehat{y}_{i,k}$ corresponds to the predicted value for observation $i \in S_{test}$, of a model constructed on the training set, and which contains k components. For the CV(1) the above formula is slightly modified since $n_{test} = 1$. A fundamental work in cross-validation and model assessment specially in linear regression areas can be found in [16] .

Many researchers on the other hand propose the use of **bootstrap** methods in validating the different models. In such cases the prediction error is estimated as an average of the error over all the bootstrap samples. They do claim that using bootstrap they do not split the data set (as in cross-validation) which is essential when the observations are important due to the small sample size. They moreover claim that data-splitting does not validate the final model, but rather a sub-model developed on a subset of the data. For a good review on bootstrap and statistical inference one can see [5] .

We propose here the use of a **penalized maximum likelihood** criterion for model selection. In particular we use the Akaike Information Criterion, denoted AIC [1] . AIC penalizes the log-likelihood of a fitted model to $2\times$ # of parameters. Thus, as the complexity of a model increases (more parameters in the model) AIC penalizes its likelihood by two times the number of parameters. Stone [17] discusses the asymptotic equivalence of model choice between AIC and cross-validation seen above.

It is commonly accepted that AIC performs well in relatively large data sets. However, caution must be taken when small data sets are analyzed, since AIC is an approximate criterion. For a small sample correction of AIC one can see [19] . Further discussion concerning the use of penalized maximum likelihood criteria in model selection and assessment is also provided by [10] .

5 Example

5.1 Fish data

Fish data is a relatively small data set arising from NIR instrumentation experiment on 45 fish samples. The data set contains 45 fish sample measurements on fat concentration at nine different wavelengths. The goal of the analysis is to model the relation between fat concentration, and these spectra. Fish data set was firstly used in [15] .

Naes reported in his work that observations 39 to 45 are outlying. We run both PLS1 and PRLS algorithms for the Fish data set. The algorithms correspond to two iterative routines implemented on **R 1.8**. For each iteration the resulting AIC for PLS1 and PRLS are stored and plotted in Figure 1. AIC for the PRLS was computed based on weighted least squares with weights obtained from the LTS. Moreover the trimming constant was set to 0.15 which reflects the report in [15] .

Figure 1 illustrates the decrease of the AIC values for PLS1 until the 4^{th} component. While it reaches its minimum at $k = 4$ one would argue that three components are enough since the decrease in AIC for the 4^{th} component is negligible (the values of the AIC are given at Table 2). The AIC for PRLS indicates with more certainty the third component as the final one. PRLS method retains three components in the final model and reduces efficiently the dimension of the regression problem.

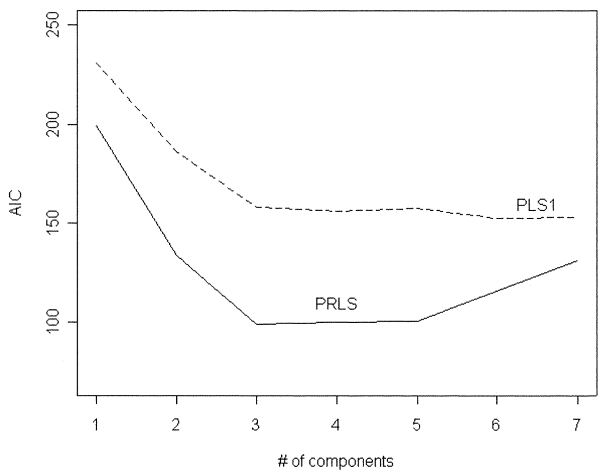


Figure 1. Fish data: AIC for both PLS1 and PRLS

From Figure 1 it is additionally remarked that AIC for PRLS stands below the AIC line for PLS1. The latter is worth noticing since AIC measures a distance from a true theoretical model. This distance is smaller for PRLS than for PLS1 regression model for each iteration k. This could have indicated a better model fit. Yet, the fact that we reweight the observations probably decreases the likelihood and consequently the AIC.

# components	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
AIC_{PLS1}	231.0619	186.1802	157.7920	156.1318	157.2012	152.2560
AIC_{PRLS}	199.1646	133.5048	98.65854	100.0978	100.4530	115.8881

Table 2: Fish Data. AIC for both PLS1 and PRLS Regression.

6 Conclusions

Outliers are data defects which often influence statistical procedures. Least squares estimates are sensitive to outliers. PLS regression models are misled and regression estimates are invalidated in the presence of outliers. In this work a schema of a less sensitive algorithm for PLS procedures was presented. Partial Reweighted Least Squares regression protects both covariance estimate and regressions from outliers. Penalized maximum likelihood (AIC) has been used in model selection. The latter revealed that PRLS reduce efficiently the dimension of the final regression. However there was no further reduction compared to PLS1.

Acknowledgments: I would like to thank the associated editor and the anonymous referee for their valuable comments.

Bibliography

[1] Akaike, H. (1973) Information Theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, Budapest: Akademiai Kiado, 267–281.

[2] de Jong, S. (1993) SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, 18, 251?-263.

[3] de Jong, S. and Farebrother, R.W. (1994) Extending the relationship between ridge regression and continuum regression. *Chemometrics and Intelligent Laboratory Systems*, 25, 179–181.

[4] Dodge, Y. and Whittaker, J. (2003) Partial Quantile Regression *to appear.*

[5] Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap.* Chapman and Hall, N.Y.

[6] Frank, I. and Friedman, J. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109–135.

[7] Geladi, P. and Kowalski, B. R. (1986) Partial Least-Squares Regression: A Tutorial. *Analytica Chimica Acta*, 185, 1–17.
[8] Goutis, C. (1996) Partial Least Squares Algorithm Yields Shrinkage Estimators. *The Annals of Statistics*, 24, 816–824.
[9] Griep,M. I. and Wakeling, I. N. and Vankeerberghen, P. and Massart,D. L. (1995) Comparison of semirobust and robust partial least squares procedures. *Chemometrics and Intelligent Laboratory Systems*, 29, 37–50.
[10] Hastie, T. and Tibshirani, R. and Friedman, J. (2001) *Elements of Statistical Learning.* Springer, N.Y.
[11] Helland, I. S. (2001) Some theoretical aspects of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 58, 97–107.
[12] Hubert, M. and Branden, K. (2003) Robust Methods for Partial Least Squares Regression. *Journal of Chemometrics*, 17, 537–549.
[13] Leroy, A. and Rousseeuw, P.J., Robust Regression & Outlier Detection, ,John Wiley & Sons, New York, 1987.
[14] Kondylis, A. (2003) *Partial Reweighted Least Squares* Master Thesis, University of Neuchâtel, Switzerland.
[15] Naes, T. (1985) Multivariate calibration when the error covariance matrix is structured. *Technometrics*, 27, 301–311.
[16] Stone, M. (1974) Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Ser. B*, 36, 111–147.
[17] Stone, M. (1977) An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Ser. B*, 39, 44–47.
[18] Stone, M. and Brooks, R. (1994) Joint continuum regression for multiple predictands. *Journal of the American Statistical Association*, 89, 1374–1377.
[19] Sugiura, N. (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics. Theory and Methods*, 7, 13–26.
[20] Sundberg, R. (1993) Continuum regression and ridge regression. *Journal of the Royal Statistical Society*, 55, 653–659.
[21] Tenenhaus, M. (1998) *La régression PLS. Théorie et pratique.* Technip, Paris.
[22] Wold, S. and Sjostrom, M. and Eriksson, L. (2001) PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58, 109–130.
[23] Wold, H. (1975) Soft modelling by latent variables: The nonlinear itrative partial least squares (NIPLS) approach. *Perspectives in Probability and Statistics, In Honor of M.S.Bartlett*, 117–144.

Student (2005), Vol. 5 No. 2, 127-149

Statistical Analysis of the Financial Structure in Italy

Giorgio Skonieczny and Benedetto Torrisi

Department of Economy and Territory
Economical statistics
Faculty of Economics
University of Catania
corso Italia 57
I-95127 Catania, Italy
E-mail: giorgios@unict.it; btorrisi@unict.it

Abstract: The process of reshaping and redefining the function of banking intermediation has caused deep changes in the Italian financial system. In this scenario, the banks have attempted a rapid structural reorganization with the aim of reaching a better settlement on the profitability level. The main objective of the research is that of supplying information about the regional financial imbalances in Italy, by using the techniques of multivariate analysis. The particular analysis of the clusters (cluvar) has allowed the drawing of a picture of the differences of the financial system of the regions of Italy, highlighting the relations between the regions and the most important variables that have contributed to the determination of such relations.

Key words : Banks, cluster analysis, PCA, regionalization.

1 Introduction

For years the financial and money market has been subjected to important changes ascribable to different factors. Among others, the so-called 'credit disintermediation', a process of reshaping and redefining the functions of banks' financial intermediation, with the consequent creation of new and specialized non-banking operators (Galli, 1990). In the meantime the gradual freeing of capital movements and the perspective of a sole European market have caused greater pression of competition among the financial systems (Parrillo, 1998). The banks have attempted a rapid reorganization with the aim of reaching a better settlement on the profitability plane (Alessandrini,

1992). The large number of fusion and incorporation operations, the growing diversification of the banking activity with the progressive opening of banking, financial and property services is the proof of this, as well as the resorting to diverse models of management. Competition favoured the uprising and the rapid spreading of innovative instruments for employing savings. All this in view of a greater efficiency of the system (Parrillo, 1995). The evolution of the financial markets constitutes a second field of research, particularly as far as relations between the real economy and financial intermediation are concerned, among credit, development and the articulation of the phenomenon in relation to the different territorial realities. This feature constitutes the theme of this work. The aim is to trace a map of territorial diversification of the Italian financial system on the base of specific indicators. In this way we will attempt to identify differences or similarities between the Italian regions, classifying them in groups with homogeneous characteristic and ascertain the presence and the nature of possible financial differences.

Cluvar is the acronym for cluster variables (groups with a relation of variables). The technique goal is to individuate groups and relations among unities (routes). These are characterized by a big degree of homogeneity among both unities and variables. They are characterized even by relation typologies among unities and among unities and Cluster, identifying among variables those that have contributed more to the links among unities couples. Particularly, it is given a data table where x_{ij} represents the determination of X_j variable for $j = 1 \ldots, m$ on the U_i unities for $i = 1, \ldots, n$. Let $S_{n,m}$ be the correspondent table of standardized values and d_j the most opportune distance determining m matrices of distances which lines are constituted by combinations of n unities and columns by combinations of m variables. Let's associate to every unity couple (line) and for every matrix one among the j variables combination (column) that has the smallest distance value. Starting from $k-1$ and for every m-matrix we need to consider the i-th unity couple and associate to this the couple that has the smallest distance because it has as first term of the combination the second term of the precedent couple. The process continues until one of the couples is selected for the second time. All unities forming the 'slip not' are part of the Cluster. The process starts again substituting the Cluster unities with its 'centre' and imposing the condition that all unities of the next 'slip notes' must have among themselves inferior distances compared to the distance between them and one of the 'centre' determined previously. At the end we will have m partitions. For every cluster will be underlined just the m variables that have made relations among every unity and all other unities.

2 Indicators

The territory subdivision into homogeneous areas in relation to the studied phenomenon assumes the identification of specific indicators capable of summarizing the statistical information that one wishes to obtain.

Proceeding in this way, an exhaustive list of variables in relation to this hypothesis has been identified, and macro indicators concerning the following aspects of the phenomenon have been created: real economy (7 simple indicators); banking density (5 simple indicators); banking productivity (5 simple indicators); financing of productive activity (6 simple indicators); credit brokerage (3 simple indicators); dimensionals of the credit companies (4 simple indicators); risk and interest tax (9 simple indicators); interregional flows of financing (4 simple indicators); families financial activity (4 simple indicators); payment system (8 simple indicators).

2.1 Indicators of real economy

The typical indices used for analysing the economic structure of a territory (Ardeni, 1995) concern, labour force, income per capita and product per employee[1]:

1. Unemployment rate
2. Work force /population
3. GDP /number of employed
4. GDP /population

In addition some indicators have been considered that are characteristic of the type of consumption of the different areas:

5. Consumption of electricity/ population

Finally to estimate the production potentiality of different regions in the sphere of international business the following ratios have been calculated:

6. Imports /GDP
7. Exports/GDP

2.2 Indicators of banking density

The territory in which the credit activity is carried out is countersigned by a group of geographical, demographical, economic and social factors, which it is necessary to bear in mind when making a correct analysis of the regional financial systems. In this respect the statistics of credit, on a first examination of the bankable territory[2], offers a series of indicators or structural ratios that allow the evaluation of the relative size of credit companies in comparison to the territorial area, the population served, the sum of the gross national product (Busetta, 1998):

1. GDP/ branches
2. Branches/ banking centres
3. Areas/branches
4. Population/branches

[1] Regional data to calculate in particular the indices for the real economy and banking density are taken from the Italian statistic annuals of ISTAT, from the national economic Accounts published by ISTAT.

[2] The word 'bankable' means 'territory made bankable' and corresponds at least to a market supplied with one bank. Bankable markets is a council in which at least one bank branch exists. For branch we mean a bank dependency thus called(branch, agency, etc.)

5. (Regional branches/national branches)(Regional population/national population[3])

2.3 Indicators of banking productivity

The statistics about the consistency[4] of jobs and deposits allow us to know, for each region, the productivity of the bank branches (from the dual aspect of collection and jobs)

1. Deposits /branches
2. Jobs/branches

Besides information about the physical production of the bank branches the following relationships can be drawn up:

3. Number of commitments/branches
4. Amount of currrent accounts/branches
5. Amount of savings deposits /branches

2.4 Indicators on financing of productive activity

An evaluation of the role of the credit institutes in favour of the real economy has been obtained by considering the sum of the gross national product of the area, the resident population and the size of the presence of credit companies:

1. Jobs/GDP
2. Jobs/population
3. Jobs/deposits

The last ratio mentioned in particular has supplied valid information about the action of supporting the regional economy or of the draining of financial resources. An analogous topic was useful for information about medium and long term credit granted by the Institutes of Special Credit:

4. Mortgages/GDP[5]

In order to make an over all evaluation of the financing of productive activity we have considered the relationship between :

5. (jobs/mortgages) GDP

Inside the same medium-long term credit it was interesting to examine the incidence of credit facilitation. Therefore the relation of composition [6] is of great help:

6. Credit facilitation/total credit [7]

---

[3] Index of balance of the branches

[4] Statistic report of the Bank of Italy, year 2000 and 2001

[5] The same ratio can be calculated for any single category of institute (building society, Section for the financing of public works, Institutes of property and building credit, institutes of land credit).

[6] The two indicators differ only for the multiplying factor (total national credit/ national facilitated credit) and they are closely linked ($r = 1$). It is, however, superfluous to consider them both.

[7] The same indices can be calculated for the individual categories of institute

2.5 Indicators of banking intermediation

From the analysis of the bank deposits it has been possible to observe and judge the phenomenon of banking intermediation, as well as explain the changes present in the financial structure and the different weight that credit companies have at a regional level[8].

1. Number of commitments/population
2. Bank deposits/population

The spreading of banking activity and the relative size of the clientele have been analysed by means of a third ratio, drawn from the disclosures of the Risks Centre on interest tax: amount of current accounts/ population

2.6 Dimensional indicators of credit companies

With the aim of evaluating, from a dual profile of deposits and jobs, the role of the smallest banks on the market of regional and national credit the following ratios have been calculated[9]:

1. Jobs small banks/tot. Jobs region
2. Jobs medium-big banks/ tot. Jobs region
3. Deposits small banks/tot. Deposits region
4. Deposits medium-big banks/tot. Deposits region

2.7 Indicators of risk in credit activity

Among the statistics published by the Bank of Italy, data about the Risk Centre[10] have a particular importance. The credit is divided by agreement, used and limits exceeded[11]. From the information about agreed credit, used credit and exceeding limits the following indicators have been obtained:

1. Credit used / credit agreed [12]
2. Excess limits/ credit agreed

---

[8] In order to reveal the extent of the function of transfer one can consider the ratio: SF/GDP where SF=(deposits+securities−jobs−mortgages). However the heterogenity of the data in question, related to different periods, does not permit an adequate use of the indicator.

[9] To simplify the analysis of small banks, the smaller banks and the smaller and minor savings banks have been classified according to the model adopted by the Bank of Italy in the Statistical Report in small banks. The medium-big banks include,instead, the major big and medium banks, and the major savings banks, medium and big indicated in the aforementioned classification.

[10] Published in the statistic report and in the supplement 'Statistical indicators drawn from the Risks Centre'.

[11] Credit agreed is any line of credit granted by the bank to an economic operator, who gives at the same time the capacity, even if only potential, to indebit himself for the company. For credit used it is intended, instead, the real exposition of the client, in other words the total sum of credit effectively supplied by the company of credit in favour of the same commitment. The excess of limit represents the difference between the effective use of credit and the sum put at disposition by the bank in favour of one client.

[12] The part of credit that is effectively used by the clientele in the light of commitments received.

The ratios in discussion have been calculated for the total of the credits, for the total of the credits for each cash desk and for each individual type of credit (direct loans, current accounts, transactions abroad, transactions with real guarantees). In particular, the Risks Centre of the Bank of Italy has disclosed, in recent years, other information about new instruments of financing: factoring and leasing and the guarantees granted to its own clients. From this information further indicators have been obtained:

3. Factoring used/factoring agreed
4. Leasing used/leasing agreed
5. Leasing in excess of limits/leasing agreed
6. Guarantees granted/population

Among the credits highlighted by banks, the statistics of the Risks Centre disclose the so-called 'overdue bills'[13]. Statistical data about overdue bills are useful for numerous calculations. They are widely used in the analysis on a territorial level and of sub-areas for examining the functioning of the banking system and the solidity of the productive sector (Galli, 1990). A particular feature concerns the incidence of overdue bills on jobs and the tax on financial lapsing for the cash desks:

7. Overdue bills/jobs
8. Tax on financial lapsing for cash desks[14]. The tax on overdue bills offers interesting information about the allocation ability of the banking system in the various areas of the nation, but at the same time it has some limitations, among which it is necessary to bear in mind a correct evaluation of the representability of data in the regional sphere. The general minimum limit of disclosure of credits is currently equal to 40'000 euros, the position of risk constituted by credit of overdue bills are subject to a particular limit of credit of 5'000 euros (and it is clear that in the most backward regions of the South the relative incidence of credits inferior to 40'000 euros is greater than in the more developed areas of the North). The most suitable indicator for understanding the regional differences in interest rates is given by the differential between active and passive rates (Notarstefano and Vassallo, 1999):
9. Active rates/passive rates.

This ratio is useful for a correct analysis of the phenomenon examined especially when geographic areas with different levels of development are compared.

2.8 Indices of interregional financial flows

The information on credit exchange is of significant interest for the study of the regional economies. They highlight, indeed, the main characteristics of territorial allocation of credit and of the flows of financial resources among regions, and they permit an

[13] In the overdue bills the totality of the ratios is included for cases implemented by persons in state of insolvency, that find themselves, in other words in the impossibility of satisfying normally the obligations assumed.

[14] The denominator of the tax for lapsing is constituted by the sum of credit used in the year T by all the persons registered at the Risks Centre and not considered in a situation of 'rectified overdue bills' the year T-1, the numerator is equal to the sum of credit used by those who, among these persons, have entered into rectified overdue bills during the year T.

evaluation of the relations between credit exchange and economic development in the observed area. The Bank of Italy carries out the disclosure of the agreed and used credits, of deposits and jobs, according to the localization of the bank branches and according to the residence of the clientele. Through a cross analysis of the two methods of disclosure it has been possible to obtain information about the exchange of credit between the regions, in order to identify the importing areas and those that export credit. The indices that have been used for the evaluation of the exchange of credit between the regions are as follows:

1. Inflow deposits/down flow deposits
2. In flow jobs/down flow jobs
3. (Inflow dep−down flow. Dep)/(inflow dep. + down flow dep)
4. (Inflow jobs−down flow jobs)/(inflow jobs. + down flow jobs)

These indices point to a dependence or a financial autonomy of the region. The values of the ratio of more than one indicate that the region is independent as far as the credit flows are concerned. Vice versa, if the ratio is less than one, the region is defined as being financially dependent.

2.9 Indices of the financial activity of families

The estimate of the sum of financial activities according to the methodology introduced by Brancati [3] allows for the calculation, for each region, of some interesting relations:

1. Bank deposits/ available deposits
2. Post office deposits/ available deposits

In particular with these ratios, it is interesting to reveal the degree of availability that the financial system has in the financing of family investments:

3. Financing of houses/jobs
4. Value of signed credits

The former index supplies the weight that the financing granted for the buying of property has on the total sum of jobs for the region. The latter considers the guarantees that the bank authorizes to the clientele (guarantees, endorsements, etc.).

2.10 Indicators of the payment system

In recent years, the Italian system of payments has recorded a gradual process of modernization and development, in line with the evolutive tendencies taking place in the major European countries. Firstly, the process of 'financialization' of the economy needs to be pointed out as well as the most intense competitivity between brokers and the financial institutions in general. Such a process supplies information about the supply of payment instruments, that go hand in hand with the growing innovations in the computer science fields and telecommunications (Silipo, 1998).

In Italy the evolution in the use of payment instruments seems different according to the specific features of the areas of the country, which go back to the structural conditions of the regional economy. The use of traditional means of payment (currency

and other paper instruments issued by the banking and postal system) is widespread in the southern regions. In the Central - Northern ones, the innovation of products finds a wider application and the operative procedures too (Messori and Silipo, 1995).

In order to simplify the analysis, because of the enormous quantity of available data, the payment system has been subdivided into two large blocks. On one hand, the group of transactions payment and cash supply have been put together, used mainly by families to meet their daily needs; on the other hand, the main procedures of collection and payment have been placed. With these companies settle their credits or their debits. The most traditional instruments of payment used by families are bank cheques, banker's cheques and savings deposits (Orsi and Pastorello, 1997). In relation to this information the following indicators have been obtained:

1. Sum of savings deposits/population
2. Indirect collection[15]/population
3. Interest bearing securities and certificates of deposit/population.

Moreover, in order to evaluate the transactions of cash supply the following ratios have been considered:

4. Sum of transactions cash supply/population
5. Sum of ATM transactions [16]/population
6. Sum of POS transactions/population
7. Home corporate banking /population
8. Phone banking/population

3 The methodology used

In order to sample the territorial division of Italy in relation to each of the aspects that countersign, according to the hypotheses made, the different types of behaviour of the phenomenon, groups of variables have been used individually according to what has been indicated beforehand. Thought the factorial weights obtained from the 55 variables have been obtained 10 macro compound indicators object of the successive elaboration with the Cluvar technique. This technique aim at the identification of homogeneous groups and the relations between the units (routes) (relations between

[15] The indirect collection considers securities or other property values, in net of passivity of issue,that are deposited at the bank in custody account, in administration or in connection with the running of the property estate.

[16] The growing competition among suppliers of payment instruments and the change in consumer habits, has determined a wide diffusion of magnetic debit cards. They are useable both at the POS points installed in the biggest retail centres, as well as for making payments and withdrawing cash from automatic teller machines (ATM). The 'POS' Point of Sale) are terminals linked to sales points, by which authorized persons can carry out automatic payment to pay for goods and services purchased.The equipment is activated by a magnetic card and the digiting of a personal code. They are, moreover, directly linked to the bank's elaboration centre. The ATM 'Automatic Teller machines' are, instead, automatic tellers, equipment automated for the carrying out of bank transactions, like withdrawing money,electronic payment of orders and information requests. The activation of ATM occurs by using a magnetic card (e.g., cash point card) owned by the user and the digiting of a secret personal code.

the individual units and the clusters identifying among the variables those that have contributed the most to the ties between the couples of units). Particularly, it is given a data table where x_{ij} represents the determination of X_j variable for $j = 1, \ldots, m$ on the unities U_i for $i = 1, \ldots, n$. Let $S_{n,m}$ be the correspondent table of standardized values and d_j the most opportune distance. The matrices m of distance are determined. The technique of grouping 'Cluvar' has been used for this aim (further details are given in (Skonieczny, 1995)).

In particular it was adopted the Cluvar technique with the intent not only to extrapolate the homogeneity and not-homogeneity among territorial units but to underline even the aspect of variables that interact in the system and the whole of the relations among unities groups.

Differently from classical techniques Cluvar gives the possibility to determine aggregations among statistic techniques underlining both the variables that significantly influence in cluster determination and the weight that those have on the aggregation definition.

The use of groups of variables has allowed a thematic analysis of the phenomenon and its articulations on the territory not allowing, however, a vision of the groups as a consideration of the single aspects and of the events caused by the integration between them (Torrisi, 2000). Often the accumulation of joint causes generate demonstrations and different behaviour to the usual interpretive models. With regard to this further details on the subject can be obtained by the application of alternative techniques but more sophisticated ones for example the neutral networks (RNA), the partial least squared (PLS) (Torrisi, 2001; Skonieczny and Torrisi, 2003).

4 Empiric analysis

The following results have been obtained on the basis of the adopted methodology and in relation to the underlined aspects:

4.1 Real economy

The Italian regions have been subdivided in two homogeneous groups, highlighting the existing dualism between north-central regions and those of south.

In the first Cluster ambit it is possible to distinguish six relational structures:

- A group made by Piedmont- Veneto- Tuscany. Each of the three regions is related to the others through all the seven considered variables.

- Chain relation made by Umbria - Marches. For this have been selected the following variables: unemployment rate, labour force/population, GDP, number of employed, GDP/ population, importations/ GDP and Marches-Tuscany with the use of all seven considered variables.

The chain among Umbria- Marches - Tuscany regions is to be related to the Cluster through Tuscany region. This region is part of he first group.

- Relation between Abruzzo- Marches that use all the seven considered variables.

- Relation between Friuli Venezia Giulia - Piedmont.

- Chain relation among Trentino Alto Adige - Emilia Romagna- Piedmont. (Piedmont acts as means with the previous region)

- Chain relation among Valle d'Aosta - Lombardy. - Veneto in which Valle d'Aosta and Lombardy are linked by: unemployment rate, labour force/population, GDP/employed number, GDP/population, exportation/GDP.

In complex the 11 regions forming the Cluster are in relation through a well equilibrated structure that has its pin around the nucleus Piedmont- Veneto- Tuscany. Even the second grouping is made by a cluster and two chain relations. These are in relation to the Cluster through Campania.

- the Cluster is formed by Campania- Sicily-Sardinia. In the relation between Sicily and Sardinia, Sardinia and Campania the algorithm has taken into consideration all the variables excepting the exportation/ GDP (number 7)

- the relation between Calabria and Campania with the exclusion of the variables: energetic consumptions electr./pop. and importations/ GDP.

The chain relation among Liguria - Latium - Molise - Basilicata - Apulia - Campania.

In the complex the Cluster resulted from the nine central- south regions appears little structured, of the three regions that made the relation pin just Campania results as relational spinner terminal while Sicily and Sardinia are isolated (see Table 1 in appendix).

4.2 Bank density

For this index three cluster have been obtained (see Table 2 in appendix) for identifying Italy as divided in three as far as the development of bank density is concerned. For what regards the bank density, Italian regions are divided in three Cluster of which two in the north. The regional north system results more structured respect to the south one.

4.3 Bank productivity

The third factor summarizes the information supplied by the indices of bank productivity. In comparison to the determining factor two clusters have been obtained (see Table 3 in appendix). These results highlight that in 2000 in almost all the regions in Italy a similarity in the capacity of bank productivity occured, an element that isn't highlighted for the regions of Umbria and Marches.

The regional structure related to the bank productivity index results complex. Both for what concerns unities and variables. Some units enter in relation with other with a reduced number of variables. It is the case of Friuli Venezia Giulia and Veneto that find similarities just for what concern: deposits/ windows, employees/ windows, savings deposits amount/ windows;

Molise with Sicily: number of commitments / branches, c.c. amount / windows, savings deposits amount/ windows.

Lombardy- Emilia Romagna deposits/ windows, number of commitments / windows, c/c amount / windows.

4.4 The financing of the economy

The structural component that groups the features of the indices of financing of the economy is positively correlated with all the indices examined, with the exception of the index of credit facilitation (inverse correlation).

Cluvar has produced 5 groups (see Table 4 in appendix). Such groups highlight a leopard spotted condition of the capacity of the credit system to finance productive activity of the companies. These results demonstrate on one hand the unhomogeneous distribution of the production sector among the Italian regions and on the other the correlated disequilibrium in financing activity. The partition is enough unregular following the index on the financing of productive activity. Five homogeneous cluster emerge for what regards the variables use.

4.5 Banking Brokerage

The fourth factor identifies the peculiar characteristics of the indices of banking brokerage. Even in this case cluvar has produced two clusters (see Table 5 in appendix). For banking brokerage Italy is presented as being divided in two, confirming for this feature dualistc factors too.

The intermediation banking index divides the Italian regions into two groups: North and South. From the used variables point of view the situation is homogeneous excluding the relation among Liguria, Veneto, Calabria and Molise that all exclude the variable number commitments / population.

4.6 The size of credit companies

The dimension index of credit companies was higher in those regions characterised by a stronger presence of small sized credit companies, and in order from: Trentino Alto Adige, Marches, Friuli Venezia Giulia, Umbria, Emilia Romagna, and Abruzzo. The phenomenon seems to have a greater diffusion in the areas of Central- North- Eastern Italy (NEC). Instead, the lowest values are observed in Piedmont, Latium, Calabria Sardinia and Val d'Aosta. The position of some southern regions (Sicily, Sardinia, Calabria and Campania) needs to be underlined, where the credit market seems characterised by the presence of a few medium-big sized banks, often with headquarters in north, capable of operating in conditions of relative advantage. This fact can contribute partially to explaining the conditions of relative backwardness of the southern financial system. The lack of strong competition between credit companies inhibits the

natural process of development of the credit structure and hinders the introduction of new financial instruments and more specialised brokers.

Two clusters have been obtained having submitted to clusterization the Italian regions in comparison to the credit companies index of size (see Table 6 in appendix).

There is not difference between North and South for what regards the dimension of credit enterprises.

In all the relations established variables are used with the exception of Sicily-Friuli Venezia Giulia that use (3.) small banks deposit/ total regional deposit and (4.) medium- big banks deposit/ total regional deposit.

4.7 Credit risks

The regions with the highest credit risks are those that have a higher level of active interest rate, as well as a greater differential between active rates and passive rate and a high ratio of regional active tax/average national active tax.

The considerations, about the regional differentials of risks in credit activity are also valid for an evaluation of the territorial distribution of interest rates.

A clear difference emerges between three areas of the country from a rapid examination of the formed clusters' structure (see Table 7 in appendix). Regions with the highest risk are all to be found in the South. In particular the position of Calabria, Basilicata and Apulia stands out. On the contrary inside the Central- Northern division Lombardy, Emilia Romagna, Piedmont, Liguria and Trentino Alto Adige, are the regions with the lowest degree of credit risk. Tuscany, which is classified among the regions with a higher degree of risk, the same gets recorded with a relatively low factor value.

4.8 Interregional finance flows

The factorial analysis, applied to the indices of interregional finance flows, permitted us to get a representative index of financial dependence of each region in comparison to all the others.

By using cluster only one cluster has been obtained made up from different aggregations (see Table 8 in appendix). For all the regions belonging to the above mentioned cluster, the financial dependency index was higher than the unit (tantamount to financial independence). The value of job down flow is also high in relation to the corresponding in flow, a further confirmation of the important role played by the credit companies of Central-North- Eastern Italy in the circuit of interregional financing.

The Italian finance system is structurally highly integrated. It has credit institutes operating throughout the national territory, and, in any case, not exclusively with the region of concern. The presence in the North of a large number of credit companies, with headquarters in other regions, which supply credit to receivers who live in the region but who substantially work in a multiregional context it can lead to an overesti-

mation of credit flows and an underestimation of the relative down flows (for example the case of Piedmont and Emilia Romagna).

There are not special subdivisions for what regards interregional financial fluxes. A unique complex regional structure gives the opportunity of examining the compartment of all fluxes.

4.9 Families' financial activity

The structural determination of the financial activity of families is drawn from information supplied about the ratios: financial activity/available income/available income gilts/financial activity, gilts/available income, bank deposits/available income, postal deposits/available income.

From an exam of the values assumed by the indicator it emerges that the investment in financial activity is particularly intense in the North. The highest scores observed are in: Trentino Alto Adige, Valle d'Aosta, and Lombardy, followed by Emilia Romagna, Veneto, Piedmont, Friuli Venezia Giulia and Liguria. Instead the southern regions are characterised by the lack of investment activity in financial activity, in the order is lowest in: Molise, Calabria, Sardinia, Basilicata and Campania. Three homogeneous groups have been obtained from cluvar. They are formed by different aggregations (see Table 9 in the appendix). Families' financial activities outline a classic typology of repartition of Italian regions.

4.10 The payment system

The application of factorial analysis to the payment system has allowed two summarial indicators to be obtained, which refer respectively to payment instruments used by families and those used by businesses.

The regions with the highest scores are to be found in the Central-Northern area. At the top there are: Valle d'Aosta, Piedmont, Lombardy, Emilia Romagna and Liguria. Typically southern the regions with the worst score are among Basilicata, Abruzzo, Calabria, Molise, Sicily and Apulia. Once again Trentino Alto Adige stands out, the relative lateness of which is perhaps to be found in the small size of its credit companies.

The second component of the payment system refers to the instruments typically used by businesses (RID, MAV, RIBA; cash orders, bills and other notes). From the examination of the data the position of relative backwardness of the South financial system emerges again.

Two clusters have been obtained (see Table 10 in the appendix). The financial innovation process in payment instruments of the business sector has been most advanced in the more developed areas of the country, where the competition between financial brokers is sharper, also as a result of the progressive financialization of the big industries. Lombardy, Veneto, Emilia Romagna, Tuscany, Fruili Venezia Giulia and Piedmont are the regions where the introduction of new and more refined means

of payment has occured with a greater speed. Generally the regions with the lowest scores are southern. Among these we can point out Sicily , Sardinia, Basilicata, Calabria Apulia and Campania.

The Italian payment system, in the case of families as well as businesses, points out the traditional division between the North and the South of the country, in particular as far as the newest instruments of collection and payment are concerned, introduced after a growing process of financial innovation linked to the development of the computer sector and of communication technology and telematic transmission of data.

The structural elements that we have identified, by means of cluster analysis applied to groups of financial indicators, summarize the plurality of the information supplied by the same indices.

The use of variables groups has permitted a thematic analysis of the phenomenon and of its articulation on the territory however not allowing, a vision of the groups as a consideration of single features and of the events caused by the integration among them. Often the accumulation of joint causes generates demonstrations and different behaviour from the normal explanatory models. The main problem of such an analysis lies in the large number of variables to be dealt with. Thus, the need to reduce their number in order to consent the easiest analysis in the shortest processing time is felt.

With this in mind composed indicators have been used. They are made by using the studied averages from factorial weights of the 55 variables considered (see Table 11 in the appendix).

Having, summarized the information generated by the individual variables in the 10 composed macro indicators, these have then been used to determine the territorial clusters (clusterization of the regions in Italy). To do this the method of classification for $k-$averages was used in order to generate groups of regions that are homogeneous among themselves.

The cluvar analysis was applied in order to give a picture of the territorial differences in the economic-financial system, that on the groups of variables (simple indicators) of each macro indicator, with the aim of constructing the regional differences related to the various aspects contemplated by each individual macro indicator (the distance estimated was the Euclidean one and the elimination of the missing cases of the casewise type): real sphere, the density and banking productivity, the capacity to finance productive activity, bank intermediation and so on.

5 Conclusions

The analysis of the clusters (applied to the main components obtained by means of factorial analysis carried out on 10 groups of regional indicators - see Table 11) has allowed us to draw a picture of the national financial system (see Fig.1).

The structural elements, that were quantified in previous phases, offer a summary of the various features of the financial system of the Italian regions. As a general rule, from an initial examination of the groups, also in the financial sphere, the traditional

division of the country in three clear areas is confirmed. North emerges as characterised by a competitive credit system in strong growth.

The classification technique applied on the total sum of the regional financial indicators has revealed an Italy with three speeds. The process of forming the groups changes slightly, but in general the results of the whole are the same. This fact contributes partly to confirm the validity of the determining factors, the concise indices of which are representative of the regional financial differences. It is necessary to add, besides this, that the technique adopted has presented an innovative and original representation in comparison to other techniques adopted by authors (see Skonieczny and Torrisi B, 2003), adding specifications of classification that have not been used before now and that are capable of highlighting inside the relations between the units closer spaces to the variables that explain such relations, considering therefore a reduced loss of information.

The first group contains the Central-North regions, countersigned as a whole by the presence of a large number of very productive banks. The regions of the NEC have a decisive weight, characterised by high values of the factor 'size of credit company'. High too is the activity of bank intermediation, which allows the clientele to choose from a vast range of financial instruments and methods of payment, also among the most innovative, and which guarantees a sole basis for the financial requirements of the numerous businesses present in the area. Diametrically opposite is the situation in the regions grouped in the second and third cluster. This cluster identifies the southern area. The financial system in this division is characterised by a relative gap between the real structure and the financial system. The credit companies are widely spread over the regional territory. Each branch serves on average a considerable number of people and a very large surface area. In area the large-sized banks' activity predominates, probably these banks have headquarters in the financial centres of the North. Moreover, bank productivity is low as is the intermediation of the credit companies. All this can contribute to explain the lack of activity of the economy financing, particularly risky, entrusted to the big financiers of the North (the dependence of the South on external credit is high), as well as the scars use of investment in financial activity (such as gilts, and still insufficient the use of the newest instruments of collection and payment).

The comparative analysis of the financial sector has confirmed the existence of a persistent division between the North and the South, highlighting a type of classification of the whole which is almost analogous to that obtained for economic development. This difference has been historically modified and does not show signs of changing.

The classification technique applied on the total of the regional financial indicators has led to the conclusion of an Italy with three different speeds in the financial system. The group formation process partially changes, but in general the results of the whole are the same. This fact in part contributes to confirm the validity of the determining factors as representative concise indices of regional financial differences.

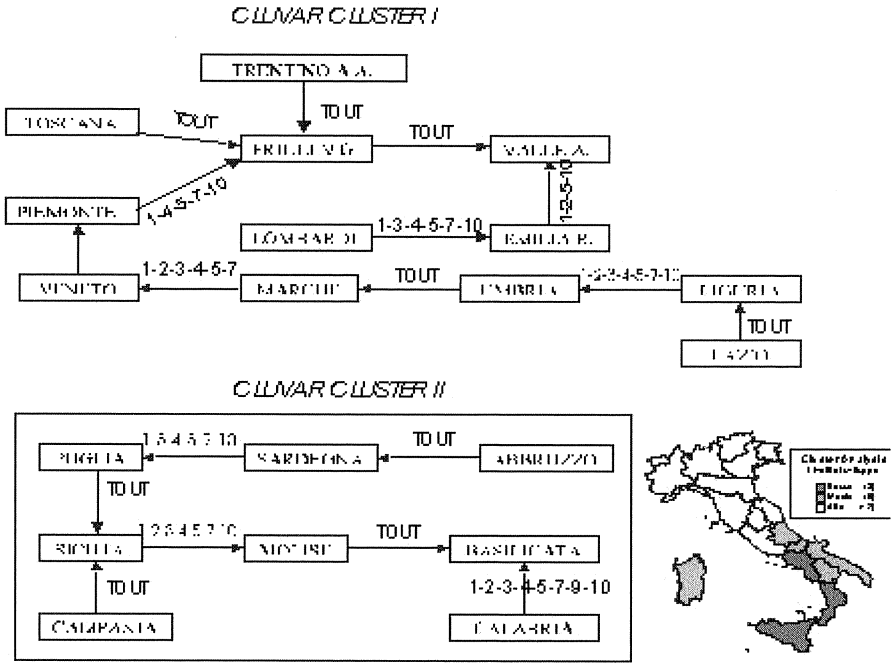


Figure 1. Diagram Cluver cluster

The number of groups was chosen by trying to differentiate the different areas of the country in the best way, by always maintaining a sufficient degree of homogenity inside each grouping.

The cartographic representation of the regional financial structure shows a clear difference between the two large areas in which Italy is traditionally divided in half.

The separation of the southern financial system, in general more backward than the rest of Italy, always remains clear from the different aspects considered (see the analysis of the individual indicators). Inside the southern division the behaviour of the regions seems more homogeneous notwithstanding it is considered in two groups: on one hand Abruzzo, Molise Apulia Basilicata and Sardinia and on the other Campania Calabria and Sicily.

Cluster n.1		**Weight**	**Variables**
Piedmont	Veneto	0.571	1-2-3-4-5-6-7
Veneto	Tuscany	0.286	1-2-3-4-5-6-7
Tuscany	Piedmont	0.857	1-2-3-4-5-6-7
	other relation:		
Umbria	Marches	0.429	1-2-3-4-6
Marches	Tuscany	0.571	1-2-3-4-5-6-7
	other relation:		
Abruzzo	Marches	0.429	1-2-3-4-5-6-7
	other relation:		
Friuli Venezia Giulia	Piedmont	0.429	1-2-3-4-5-6-7
	other relation:		
Trentino Alto Adige	Emilia Romagna	0.714	1-2-3-4-5-6-7
Emilia Romagna	Piedmont	0.429	1-2-3-4-5-6-7
	other relation:		
Valle d'Aosta	Lombardy	0.429	1-2-3-4-7
Lombardy	Veneto	0.429	1-2-3-4-5-6-7
Cluster n.2			
Campania	Sicily	0.857	1-2-3-4-5-6-7
Sicily	Sardinia	0.571	1-2-3-4-5-6
Sardinia	Campania	0.429	1-2-3-4-5-6
	other relation:		
Calabria	Campania	0.571	1-2-3-4-7
	other relation:		
Liguria	Latium	0.714	1-2-3-4-5-6-7
Latium	Molise	0.714	1-2-3-4-5-6-7
Molise	Basilicata	0.571	1-2-3-4-5-6-7
Basilicata	Apulia	0.714	1-2-3-4-5-6-7
Apulia	Campania	0.857	1-2-3-4-5-6-7

Table 1. Indicators for real economy

Cluster n.1		Weight	Variables
Piedmont	Liguria	0.400	1-3-4-5
Liguria	Tuscany	0.400	1-2-3-4-5
Tuscany	Piedmont	0.400	3-4-5
	other relation:		
Lombardy	Tuscany	0.400	3-4-5
Cluster n.2			
Friuli Venezia Giulia	Marches	0.800	1-2-3-4-5
Marches	Veneto	0.400	1-2-3-4-5
Veneto Friuli Venezia	Giulia	0.200	1-2-3-4-5
	other relation:		
Umbria	Marches	0.200	1-2-3-4-5
	other relation:		
Valle d'Aosta Emilia	Romagna	0.600	1-3-4-5
Emilia Romagna	Friuli Venezia Giulia	0.800	1-3-4-5
	other relation:		
Trentino Alto Adige	Valle d'Aosta	1.000	1-2-3-4-5
Cluster n.3			
Abruzzo	Basilicata	0.400	1-2-3-4-5
Basilicata	Sardinia	0.800	1-2-3-4-5
Sardinia	Abruzzo	0.600	1-2-3-4-5
	other relation:		
Apulia	Sicily	0.600	1-2-3-4-5
Sicily	Abruzzo	0.400	1-2-3-4-5
	other relation:		
Molise	Basilicata	0.800	2-3-4-5
	other relation:		
Campania	Calabria	1.000	1-2-3-4-5
Calabria	Molise	0.600	1-2-3-4-5
	other relation:		
Latium	Abruzzo	0.400	3-5

Table 2. Indices of bank density

Cluster n.1		Weight	Variables
Calabria	Apulia	0.800	1-2-3-4-5
Apulia	Abruzzo	0.800	1-2-3-4-5
Abruzzo	Calabria	0.600	1-2-3-4-5
	other relation:		
Campania	Calabria	0.200	1-2-3-4-5
	other relation:		
Piedmont	Tuscany	0.600	1-2-3-4-5
Tuscany	Emilia Romagna	0.400	1-2-3-4-5
Emilia Romagna	Veneto	0.400	1-2-3-4-5
Veneto	Friuli Venezia Giulia	0.600	1-2-3-5
Friuli Venezia Giulia	Sardinia	0.600	1-2-3-4-5
Sardinia	Molise	0.200	1-2-3-4-5
Molise	Sicily	0.600	3-4-5
Sicily	Calabria	0.400	1-2-3-4-5
	other relation:		
Basilicata	Sicily	0.600	1-2-3-4-5
	other relation:		
Latium	Piedmont	0.200	1-2-3-4-5
	other relation:		
Trentino Alto Adige	Veneto	0.200	1-2-3-4-5
	other relation:		
Lombardy	Emilia Romagna	0.600	1-3-4
	other relation:		
Valle d'Aosta	Liguria	0.800	1-2-3-4-5
Liguria	Friuli Venezia Giulia	0.600	1-2-3-4-5
Cluster n.2			
Umbria	Marches	0.600	1-2-3-5
Marches	Umbria	0.400	1-2-3-4-5

Table 3. Indices of bank productivity

Cluster n.1		Weight	Variables
Piedmont	Veneto	0.833	1-2-3-4-5-6
Veneto	Marches	0.500	1-2-3-4-5-6
Marches	Piedmont	0.667	1-2-3-4-5-6
	other relation:		
Tuscany	Piedmont	0.500	1-2-4-5-6
	other relation:		
Lombardy	Emilia Romagna	0.500	1-2-3-4-5-6
Emilia Romagna	Trentino Alto Adige	0.667	1-2-3-4-5-6
Trentino Alto Adige	Piedmont	1.000	1-2-3-4-5-6
	other relation:		
Latium	Lombardy	0.667	1-2-3-4-5-6
Cluster n.2			
Abruzzo	Basilicata	0.500	1-2-3-4-5
Basilicata	Abruzzo	0.333	1-2-3-4-5-6
	other relation:		
Sicily	Abruzzo	0.500	1-2-3-5-6
	other relation:		
Liguria	Sardinia	0.667	1-2-3-4-5
Sardinia	Basilicata	0.833	1-2-3-4-5-6
	other relation:		
Valle d'Aosta	Abruzzo	0.500	1-2-3-4-5-6
Cluster n.3			
Friuli Venezia Giulia	Umbria	0.500	1-2-4-5-6
Umbria	Friuli Venezia Giulia	0.500	1-2-3-4-5-6
Cluster n.4			
Molise	Calabria	0.500	1-2-3-4-5
Calabria	Molise	0.500	1-2-3-4-5-6
Cluster n.5			
Campania	Apulia	0.333	1-2-4-6
Apulia	Campania	0.500	1-2-3-4-5-6

Table 4. Index on the financing of productive activity

Cluster n.1		Weight	Variables
Friuli Venezia Giulia	Tuscany	1.000	1-2-3
Tuscany	Veneto	0.333	1-2-3
Veneto	Friuli Venezia Giulia	0.333	1-2-3
	other relation:		
Umbria	Marches	1.000	1-2-3
Marches	Veneto	0.667	1-2-3
	other relation:		
Liguria	Veneto	0.667	2-3
	other relation:		
Valle d'Aosta	Friuli Venezia Giulia	0.333	1-2-3
	other relation:		
Trentino Alto Adige	Emilia Romagna	1.000	1-2-3
Emilia Romagna	Valle d'Aosta	1.000	1-2-3
	other relation:		
Lombardy	Valle d'Aosta	0.333	1-2-3
	other relation:		
Latium	Lombardy	0.333	1-2-3
	other relation:		
Piedmont	Friuli Venezia Giulia	1.000	1-2-3
Cluster n.2			
Basilicata	Sicily	1.000	1-2-3
Sicily	Molise	0.667	1-2-3
Molise	Basilicata	0.667	1-2-3
	other relation:		
Calabria	Molise	0.667	2-3
	other relation:		
Abruzzo	Sardinia	1.000	1-2-3
Sardinia	Campania	0.333	1-2-3
Campania	Apulia	0.667	1-2-3
Apulia	Basilicata	0.667	1-2-3

Table 5. Indices of banking intermediation

Cluster n.1		Weight	Variables
Piedmont	Valle d'Aosta	0.500	1-2-3-4
Valle d'Aosta	Campania	0.500	1-2-3-4
Campania	Piedmont	0.500	1-2-3-4
	other relation:		
Liguria	Piedmont	0.500	1-2-3-4
	other relation:		
Lombardy	Campania	1.000	1-2-3-4
	other relation:		
Sardinia	Lombardy	0.500	1-2-3-4
	other relation:		
Latium	Lombardy	0.500	1-2-3-4
Cluster n.2			
Tuscany	Emilia Romagna	0.500	1-2-3-4
Emilia Romagna	Veneto	0.500	1-2-3-4
Veneto	Tuscany	0.500	1-2-3-4
	other relation:		
Calabria	Veneto	0.500	1-2-3-4
	other relation:		
Abruzzo	Basilicata	0.500	1-2-3-4
Basilicata	Emilia Romagna	1.000	1-2-3-4
	other relation:		
Trentino Alto Adige	Umbria	0.500	1-2-3-4
Umbria	Marches	0.500	1-2-3-4
Marches	Apulia	1.000	1-2-3-4
Apulia	Friuli Venezia Giulia	1.000	1-2-3-4
Friuli Venezia Giulia	Molise	1.000	1-2-3-4
Molise	Tuscany	0.500	1-2-3-4
	other relation:		
Sicily	Friuli Venezia Giulia	0.500	3-4

Table.6 Dimensional index of the credit companies

Cluster n.1		Weight	Variables
Emilia Romagna	Friuli Venezia Giulia	0.444	2-3-5-7-8-9
Friuli Venezia Giulia	Trentino Alto Adige	0.556	2-3-4-5-7-8-9
Trentino Alto Adige	Emilia Romagna	0.556	2-3-4-5-7-8-9
	other relation:		
Lombardy	Emilia Romagna	0.667	1-2-3-4-5-6-7-8-9
	other relation:		
Piedmont	Emilia Romagna	0.556	2-5-6-7-8-9
Cluster n.2			
Veneto	Marches	0.333	1-2-3-4-5-6-7-8-9
Marches	Veneto	0.444	1-2-3-4-5-6-7-9
	other relation:		
Tuscany	Veneto	0.333	1-2-3-4-5-6-7-8-9
	other relation:		
Liguria	Latium	0.556	1-2-3-7-8-9
Latium	Umbria	0.444	1-2-3-6-7-8-9
Umbria	Marches	0.333	3-4-5-7-9
	other relation:		
Apulia	Basilicata	0.333	1-2-3-4-5-6-7-8-9
Basilicata	Umbria	0.444	1-2-3-4-5-6
	other relation:		
Abruzzo	Umbria	0.222	2-3-4-5-7-8-9
	other relation:		
Molise	Abruzzo	0.667	1-2-3-4-5-6-7-8-9
	other relation:		
Valle d'Aosta	Veneto	0.333	1-2-3-4-5-6-7-8-9
Cluster n.3(Cl3)			
Campania	Sardinia	0.333	1-2-3-4-5-6-7-8-9
Sardinia	Campania	0.667	1-2-3-4-6-7-8-9
	other relation:		
Calabria	Sicily	0.778	1-2-3-4-5-6-7-8-9
Sicily	Sardinia	0.333	1-2-3-4-5-6-7-8-9

Table 7. Index of risk in credit activity

Cluster n.1		Weight	Variables
Emilia Romagna	Campania	0.250	1-2-3-4
Campania	Calabria	0.750	1-2-3-4
Calabria	Emilia Romagna	0.500	1-2-3-4
	other relation:		
Liguria	Trentino Alto Adige	0.250	1-2-3-4
Trentino Alto Adige	Campania	0.500	1-2-4
	other relation:		
Piedmont	Sicily	0.750	1-2-3-4
Sicily	Friuli Venezia Giulia	0.500	1-2-3-4
Friuli Venezia Giulia	Emilia Romagna	0.250	1-2-3-4
	other relation:		
Apulia	Sicily	0.750	1-2-3-4
	other relation:		
Veneto	Friuli Venezia Giulia	0.500	2-3-4
	other relation:		
Tuscany	Veneto	0.250	1-2-3-4
	other relation:		
Lombardy	Marches	0.750	1-2-3-4
Marches	Basilicata	0.500	2-3-4
Basilicata	Molise	0.250	1-2-3-4
Molise	Friuli Venezia Giulia	0.250	1-2-3-4
	other relation:		
Umbria	Molise	0.500	1-2-3-4
	other relation:		
Latium	Abruzzo	0.750	2-3-4
Abruzzo	Sardinia	0.500	1-2-3-4
Sardinia	Umbria	0.250	1-2-3-4
	other relation:		
Valle d'Aosta	Sicily	0.250	1-2-3-4

Table 8. Index of interregional finance flows

Cluster n.1		Weight	Variables
Liguria	Veneto	0.750	1-2-3
Veneto	Emilia Romagna	0.500	1-2
Emilia Romagna	Liguria	0.500	1-2
	other relation:		
Lombardy	Emilia Romagna	0.500	1-2-3-4
	other relation:		
Piedmont	Valle d'Aosta	0.500	1-2
Valle d'Aosta	Friuli Venezia Giulia	0.500	1-2
Friuli Venezia Giulia	Tuscany	0.500	1-2
Tuscany	Liguria	0.250	1-2-3-4
	other relation:		
Trentino Alto Adige	Piedmont	0.500	1-2-4
Cluster n.2			
Apulia	Sardinia	0.500	1-2-4
Sardinia	Apulia	0.500	1-2-3-4
	other relation:		
Latium	Sicily	0.500	1-2
Sicily	Sardinia	0.500	1-2-3-4
	other relation:		
Umbria	Marches	0.500	1-2-3-4
Marches	Apulia	0.500	1-2-3-4
Cluster n.3			
Basilicata	Calabria	0.500	1-2-3-4
Calabria	Molise	0.500	1-2-3-4
Molise	Basilicata	0.500	1-2-3-4
	other relation:		
Abruzzo	Campania	0.500	1-2-3-4
Campania	Basilicata	0.500	1-2-3-4

Table 9. Indices of the financial activity of families

Cluster n.1		Weight	Variables
Emilia Romagna	Friuli Venezia Giulia	0.875	1-2-3-5-6-7-8
Friuli Venezia Giulia	Veneto	0.750	1-2-3-5-6-7-8
Veneto	Emilia Romagna	0.375	2-5-6-7
	other relation:		
Tuscany	Veneto	0.375	1-2-3-6-7-8
	other relation:		
Lombardy	Veneto	0.375	1-2-3-4-5-6-7-8
	other relation:		
Latium	Liguria	0.375	1-2-5-6-7-8
	other relation:		
Umbria	Marches	0.750	1-2-3-4-5-6
Marches	Liguria	0.625	1-3-5-6-7-8
	other relation:		
Lombardy	Emilia Romagna	0.375	1-2-3-4-5-6-7-8
	other relation:		
Piedmont	Emilia Romagna	0.375	1-2-3-5-6-7-8
Cluster n.2			
Valle d'Aosta	Trentino Alto Adige	0.375	1-3-5-6-7-8
Trentino Alto Adige	Valle d'Aosta	0.375	1-3-4-5-6-7-8
Cluster n.3 (Cl3)			
Sicily	Campania	0.500	1-2-3-4-5-6-7-8
Campania	Sardinia	0.625	1-2-3-4-5-6-7-8
Sardinia	Sicily	0.750	1-2-3-4-5-6-7-8
	other relation:		
Calabria	Campania	0.500	1-2-3-4-5-6-7-8
	other relation:		
Molise	Basilicata	0.750	1-2-3-4-5-6-7-8
Basilicata	Apulia	0.375	1-2-3-4-5-6
Apulia	Sicily	0.375	1-2-3-4-5-6-7-8
	other relation:		
Abruzzo	Sicily	0.500	1-2-3-4-5-6-7-8

Table 10. Indices on the payment system

Indicator of the real economy									
Variables	Tasso-di	Tasso-At	GDP-Pop	GDP-Occ	X-Pil	M-Pil	Enel-Pop		
Factor weight	-0,94163	0,938274	0,97485	0,82356	0,77516	0,391043	0,78751		
Index of bank density									
Variables	GDP-spo	spo-pbnc	sup-spo	pop-spo	equi-spo				
Factor weight	0,8700	-0,1978	-0,9770	0,9596	-0,9806				
Index of bank productivity									
Variables	idrac-spo	dp-spo	imp-spo	Naff-spo	ncc-spo				
Factor weight	-0,5271	0,8856	0,9398	0,6094	0,9663				
Index on the financing of productive activity									
Variables	imp-pil	imp-pop	imp-db	mut-pil	fin.eco	r.comp			
Factor weight	0,9733	0,9493	0,8339	0,8680	0,6658	-0,6979			
Index of bank intermediation									
Variables	naff-pop	db-pop	ncc-pop						
Factor weight	0,8389	0,9413	0,9691						
Dimensional index of credit companies									
Variables	q.impbp	q.impbp	q.dbbp	q.dbbmg					
Factor weight	0,9529	-0,9528	0,9529	-0,9528					
Index of risk in credit activity									
Variables	uti-acc	uti-acc	sc-acc	grc-pop	futi-facc	uti-acc	scnf-acc	dif-tas	soff-imp
Factor weight	0,5243	0,6674	-0,2612	-0,7245	0,6506	0,8509	0,8808	0,8163	0,8882
Index of interregional finance flows									
Variables	afd-defd	afi-defi	afdfd-mv	afdfi-mv					
Factor weight	0,7245	-0,6501	-0,1769	0,7560					
Index of the financial activity of families									
Variables	db-dd	dp-dd	ful-imp	Crfir					
Factor weight	-0,9136	0,9136	-0,8541	-0,7225					
Index on the payment system									
Variables	rci-pop	idrac-pop	itotac-pop	bcd-pop	iatm-pop	ipos-pop	hcb/pop	phb/pop	
Factor weight	-0,7339	0,0351	-0,8761	-0,7106	0,8204	0,9031	0,9067	0,8123	

Table 11. PCA Factorial analysis

References

[1] Alessandrini, P. (1992), Squilibri regionali e dualismo finanziario in Italia, in *Mon-*

eta e Credito n.177.

[2] Ardeni, P.G. (1995), *Teorie della crescita endogena*, Giappichelli, Torino.

[3] Brancati, R. (1985), *Politiche regionali e politiche macroeconomiche*, F. Angeli, Milano, 1985.

[4] Busetta, P. (1998), *Statistiche finanziarie*, Cedam, Milano.

[5] Galli, G. Eds. (1990), *Il Sistema Finanziario nel Mezzogiorno*, Banca d'Italia, Roma.

[6] Messori, M. and Silipo, D. (1995), *Mercati del credito e sviluppo territoriale dell'economia italiana*, Banca d'Italia, Roma.

[7] Notarstefano, G. and Vassallo, E. (1999), *I tassi bancari nel Mezzogiorno: alcune verifiche empiriche*, Bancaria, 4.

[8] Orsi, R. and Pastorello, S. (1997), Composizione del portafoglio delle famiglie: un modello a fattori per variabili qualitative, in *Ricerche Qualitative Per La Politica Economica*, Minerva Bancaria.

[9] Skonieczny, G. (1995), Cluvar plus: un metodo di classificazione con la selezione delle variabili, in *Giornate di analisi dei dati multidimensinale.*

[10] Skonieczny, G. and Torrisi, B. (2003), PCA, PLS and ANN: diversification of the Italian financial system, in *PLS and related methods*, Decisia, Portugal.

[11] Silipo, D. (1998), Credito e sviluppo territoriale: un'analisi empirica del comportamento bancario, in *Mezzogiorno Senza Credito.*

[12] Torrisi, B. (2000), Economie di scala e di scopo. Un'applicazione della Hybrid traslog cost function al sistema bancario siciliano, in *Bancaria Editrice*, 7/8.

[13] Torrisi, B. and Beccari, A. (2001), New statistical metodology for variable selection: a chemiometric application, in *CISIA-CERESTA*, Montreuil France.