Discussiones Mathematicae
Probability and Statistics 31 (2011) 103–119
doi:10.7151/dmps.1132

STATISTICAL MODELLING: APPLICATION TO THE FINANCIAL SECTOR

Cláudia Roçadas

Center of Mathematics and Applications
Faculty of Sciences and Technology
Department of Mathematics, New University of Lisbon

e-mail: cvrlm@iol.pt

Teresa A. Oliveira

UAb-Universidade Aberta, Department of Sciences and Technology
CEAUL-Center of Statistics and Applications, University of Lisbon

e-mail: toliveir@uab.pt

AND

João T. Mexia

Center of Mathematics and Applications
Faculty of Sciences and Technology
Department of Mathematics, New University of Lisbon

e-mail: jtm@fct.unl.pt

Abstract

Our research is centred on the stochastic structure of matched open populations, subjected to periodical reclassifications. These populations are divided into sub-populations. In our application we considered two populations of customers of a bank: with and without account manager. Two or more of such population are matched when there is a 1-1 correspondence between their sub-populations and the elements of one of them can go to another, if and only if the same occurs with elements from the corresponding sub-populations of the other. So we have inputs and outputs of elements in the population and along with several sub-populations in which the elements can be placed. It is thus natural to use Markov chains to model these populations.

Besides this study connected with Markov chains we show how to carry out Analysis of Variance – like analysis of entries and departures to and from de populations of customers. Our purpose is to study the flows in and out of customers in classes for the two populations and to make research on the influence of the factors year, class and region. We used the Likelihood ratio tests for the hypotheses formulated on the basis of these factors. In our work we verified that major hypotheses were all rejected. This raises the question of what are the effects and interactions truly relevant. Looking for an answer to this problem, we present the first partition to a change in the log Likelihood. This partition is very similar to the analysis of variance for the crossing of the factors that allowed us to use algebraic established results, see Fonseca *et al.* (2003, 2006), for models with balanced cross.

Keywords: populations with periodic reclassification, likelihood ratio tests, Markov chains, isomorphism.

2010 Mathematics Subject Classification: 60J20, 62P05.

1. Inputs and outputs of customers in classes

1.1. Introduction

We will start with the separate analysis of flows to customers with and without account manager. So, we consider that:

n_{ijk} is the number of customers at the beginning of the year i, which belong to the class j and are located in the region k;

e_{ijk} is the number of customers who come in year i in class j and is based in the region k;

s_{ijk} is the number of customers who drop out in year i to class j being based in the region k;

The probability of entry and exit may be estimated by

$$\begin{cases} p_{ijk} = \frac{e_{ijk}}{n_{ijk}}, \\ p^*_{ijk} = \frac{s_{ijk}}{n_{ijk}}. \end{cases}$$

In what follows, we study the action of factors A – Year, C – Class and R – region.

For each of these factors combination we can test the hypothesis that there is no action of such factors. For example $H_0(\{A, C\})$ is the hypoth-

esis that the year and the class do not influence the flow. These hypotheses are distinct from those who consider the analysis of variance in effect $H_0(\{A, C\})$ would break down in cases of absence of the effects of factors A and C and the absence of interactions in which these factors participate. Further, consider a partition of the variation of the logarithm of the log-likelihood, similar to the analysis of variance. In this approach the parcels on which decomposes the variation correspond to effects and interactions. Note that portions of this partition correspond to the whole lattice of subsets of $\{A, C, R\}$:

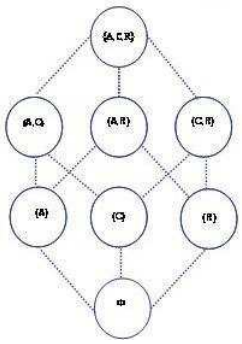

Figure 1. "Screening" of all the factors in study and their interaction.

Such a lattice is the underlying analysis of variance for balanced designs with three factors interactions, see Mexia (1988). In such models one takes the same number of observations for all combinations of factor levels.

We consider the pairing of the populations of customers with and without account manager for a comparative study, and we use Likelihood Ratio Tests.

2. Construction testing

The Likelihood Ratio Test is one of the most used techniques for testing composite hypotheses. Let $(x_1, \ldots, x_n)$ represent a sample from a population with density function or probability function $f(x; \theta)$, depending on k parameters $\theta_1, \theta_2, \ldots, \theta_k$. Denote by Ω the set of all values $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Let H_0 be the hypothesis which imposes certain restrictions on the values θ, determining a subset of Ω, say ω, being the null and alternative hypotheses

written as follows:

$$H_0 : \theta \in \omega \text{ versus } H_1 : \theta \in \omega_1, \text{ with } \omega_1 = \Omega - \omega, \ \ \theta \notin \omega.$$

The likelihood function for a given sample $(x_1, \ldots, x_n)$ is given by

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta), \ \ \theta \in \Omega.$$

Consider L_Ω as the supreme of $L(\theta)$ when $\theta \in \Omega$. Similarly L_ω shall be the supreme L when $\theta \in \omega$, and $\lambda = \frac{L_\omega}{L_\Omega}$ is the likelihood ratio. Note that $\lambda \leq 1$, since L_ω never exceeds L_Ω. Thus when λ is sufficiently close to 1 we are led to not reject H_0. To perform the tests, we use Likelihood Ratio (see Mood et al., 1963) Wilks's theorem:

Theorem 1 (Wilks's theorem)**.** *Suppose we wanted to test:*

$$H_0 : \theta_1 = \theta_1^0, \ldots, \theta_r = \theta_r^0, \theta_{r+1}^0, \ldots, \theta_k^0$$

against all alternatives using a sample $(x_1, \ldots, x_n)$ *with density or probability function* $f(x, \theta)$*.*

When H_0 *is true, (verifying regularity conditions) the distribution of* $-2 \log \lambda$ *converges, when* $n \to \infty$*, for the chi-square central with* $k - r$ *degrees of freedom. Although this statement the number* g *of degrees of freedom is the number* $(k - r)$ *of parameters not specified* H_0*, generally with* $g = dim(\Omega) - dim(\omega)$ *degrees of freedom. For the unilateral right test the critical value for the* α *level is the quantile* $\chi_{g,1-\alpha}$ *of distribution* χ_g^2 *for the probability* $1 - \alpha$*.*

3. The study case

Let us now see how to perform the Likelihood Ratio Test for the hypotheses $H_0(\{A\})$ a $H_0(\{A, C, R\})$ outlined above. The data that we consider relate to the years 2005, 2006 and 2007. Customers are assigned to classes set out above and are located in 33 regions. We consider separately:

- Input from customers with account manager;
- Departure of customers with account manager;

- Input from customers without account manager;
- Departure of customers without account manager.

To avoid duplication we have assumed the construction of tests to work with random variables with binomial distributions with parameters n_{ijk} and q_{ijk}, taking values x_{ijk}.

The likelihood is

$$L_\Omega = \prod_i \prod_j \prod_k \left(\begin{array}{c} n_{ijk} \\ x_{ijk} \end{array} \right) q_{ijk}^{x_{ijk}} (1 - q_{ijk})^{n_{ijk} - x_{ijk}}$$

with logarithm

$$l_\Omega = k + \sum_i \sum_j \sum_k (x_{ijk} \log q_{ijk} + (n_{ijk} - x_{ijk}) \log(1 - q_{ijk}))$$

being

$$k = \sum_i \sum_j \sum_k \log \left(\begin{array}{c} n_{ijk} \\ x_{ijk} \end{array} \right).$$

Since $\frac{\delta l_\Omega}{\delta q_{ijk}} = \frac{x_{ijk}}{q_{ijk}} - \frac{n_{ijk} - x_{ijk}}{1 - q_{ijk}}$, we obtain the maximum likelihood estimators $\hat{q}_{ijk} = \frac{x_{ijk}}{n_{ijk}}$, being

$$\hat{l}_\Omega = k + \sum_i \sum_j \sum_k \left(x_{ijk} \log \frac{x_{ijk}}{n_{ijk}} + (n_{ijk} - x_{ijk}) \log \frac{n_{ijk} - xijk}{n_{ijk}} \right).$$

On the other hand, we have:

$q_{ij.} = \frac{1}{33} \sum_{k=1}^{33} q_{ijk}$	$n_{ij.} = \sum_{k=1}^{33} n_{ijk}$	$x_{ij.} = \sum_{k=1}^{33} x_{ijk}$
$q_{i.k} = \frac{1}{4} \sum_{j=1}^{4} q_{ijk}$	$n_{i.k} = \sum_{j=1}^{4} n_{ijk}$	$x_{i.k} = \sum_{j=1}^{4} x_{ijk}$
$q_{.jk} = \frac{1}{3} \sum_{i=1}^{3} q_{ijk}$	$n_{.jk} = \sum_{i=1}^{3} n_{ijk}$	$x_{.jk} = \sum_{i=1}^{3} x_{ijk}$
$q_{i..} = \frac{1}{4 \times 33} \sum_{j=1}^{4} \sum_{k=1}^{33} q_{ijk}$	$n_{i..} = \sum_{j=1}^{4} \sum_{k=1}^{33} n_{ijk}$	$x_{i..} = \sum_{j=1}^{4} \sum_{k=1}^{33} x_{ijk}$
$q_{.j.} = \frac{1}{3 \times 33} \sum_{i=1}^{3} \sum_{k=1}^{33} q_{ijk}$	$n_{.j.} = \sum_{i=1}^{3} \sum_{k=1}^{33} n_{ijk}$	$x_{.j.} = \sum_{i=1}^{3} \sum_{k=1}^{33} x_{ijk}$
$q_{..k} = \frac{1}{3 \times 4} \sum_{i=1}^{3} \sum_{j=1}^{4} q_{ijk}$	$n_{..k} = \sum_{i=1}^{3} \sum_{j=1}^{4} n_{ijk}$	$x_{..k} = \sum_{i=1}^{3} \sum_{j=1}^{4} x_{ijk}$
$q_{...} = \frac{1}{3 \times 4 \times 33} \sum_{i=1}^{3} \sum_{j=1}^{4} \sum_{k=1}^{33} q_{ijk}$	$n_{...} = \sum_{i=1}^{3} \sum_{j=1}^{4} \sum_{k=1}^{33} n_{ijk}$	$x_{...} = \sum_{i=1}^{3} \sum_{j=1}^{4} \sum_{k=1}^{33} x_{ijk}$.

Our hypotheses that can be written as

$$
\begin{aligned}
&H_0(A,C,R) : q_{ijk} = q_{...}\\
&H_0(C,R) : q_{ijk} = q_{i..}\\
&H_0(A,R) : q_{ijk} = q_{.j.}\\
&H_0(A,C) : q_{ijk} = q_{..k}\\
&H_0(R) : q_{ijk} = q_{ij.}\\
&H_0(C) : q_{ijk} = q_{i.k}\\
&H_0(A) : q_{ijk} = q_{.jk}.
\end{aligned}
$$

Given the reproducibility of the binomial distribution, when the various hypotheses are verified, there are the log-likelihoods, can be written:

$$
\begin{aligned}
&l_{\{A,C,R\}} = k + x_{...}\log q_{...}(n_{...} - x_{...})\log(1 - q_{...})\\
&l_{\{C,R\}} = k + \textstyle\sum_i (x_{i..}\log q_{i..} + (n_{i..} - x_{i..})\log(1 - q_{i..}))\\
&l_{\{A,R\}} = k + \textstyle\sum_j (x_{.j.}\log q_{.j.} + (n_{.j.} - x_{.j.})\log(1 - q_{.j.}))\\
&l_{\{A,C\}} = k + \textstyle\sum_k (x_{..k}\log q_{..k} + (n_{..k} - x_{..k})\log(1 - q_{..k}))\\
&l_{\{R\}} = k + \textstyle\sum_i \sum_j (x_{ij.}\log q_{ij.} + (n_{ij.} - x_{ij.})\log(1 - q_{ij.}))\\
&l_{\{C\}} = k + \textstyle\sum_i \sum_k (x_{i.k}\log q_{i.k} + (n_{i.k} - x_{i.k})\log(1 - q_{i.k}))\\
&l_{\{A\}} = k + \textstyle\sum_j \sum_k (x_{.jk}\log q_{.jk} + (n_{.jk} - x_{.jk})\log(1 - q_{.jk}))
\end{aligned}
$$

with maximum

$$
\begin{aligned}
&\hat{l}_{\{A,C,R\}} = k + x_{...}\log\tfrac{x_{...}}{n_{...}} + (n_{...} - x_{...}) + \log\tfrac{n_{...} - x_{...}}{n_{...}}\\
&\hat{l}_{\{C,R\}} = k + \textstyle\sum_i x_{i..}\log\tfrac{x_{i..}}{n_{i..}} + (n_{i..} - x_{i..})\log\tfrac{n_{i..} - x_{i..}}{n_{i..}}\\
&\hat{l}_{\{A,R\}} = k + \textstyle\sum_j x_{.j.}\log\tfrac{x_{.j.}}{n_{.j.}} + (n_{.j.} - x_{.j.})\log\tfrac{n_{.j.} - x_{.j.}}{n_{.j.}}\\
&\hat{l}_{\{A,C\}} = k + \textstyle\sum_k x_{..k}\log\tfrac{x_{..k}}{n_{..k}} + (n_{..k} - x_{..k})\log\tfrac{n_{..k} - x_{..k}}{n_{..k}}\\
&\hat{l}_{\{R\}} = k + \textstyle\sum_i \sum_j x_{ij.}\log\tfrac{x_{ij.}}{n_{ij.}} + (n_{ij.} - x_{ij.})\log\tfrac{n_{ij.} - x_{ij.}}{n_{ij.}}\\
&\hat{l}_{\{C\}} = k + \textstyle\sum_i \sum_k x_{i.k}\log\tfrac{x_{i.k}}{n_{i.k}} + (n_{i.k} - x_{i.k})\log\tfrac{n_{i.k} - x_{i.k}}{n_{i.k}}\\
&\hat{l}_{\{A\}} = k + \textstyle\sum_j \sum_k (x_{.jk}\log\tfrac{x_{.jk}}{n_{.jk}} + (n_{.jk} - x_{.jk})\log\tfrac{n_{.jk} - x_{.jk}}{n_{.jk}}\,.
\end{aligned}
$$

To build the corresponding Likelihood ratio tests, the degrees of freedom (given by $\dim(\Omega)$) will be:

$g_{\{A,C,R\}} = 3 \times 4 \times 33 - 1 = 395$
$g_{\{C,R\}} = 3 \times 4 \times 33 - 3 = 393$
$g_{\{A,R\}} = 3 \times 4 \times 33 - 4 = 392$
$g_{\{A,C\}} = 3 \times 4 \times 33 - 33 = 363$
$g_{\{A\}} = 3 \times 4 \times 33 - 4 \times 33 = 264$
$g_{\{C\}} = 3 \times 4 \times 33 - 3 \times 33 = 297$
$g_{\{R\}} = 3 \times 4 \times 33 - 3 \times 4 = 384.$

3.1. Tests of hypotheses on input from customers with account manager

In this case we have

$\hat{l}_{\{\Omega\}} = -346248,001$
$\hat{l}_{\{A,C,R\}} = -358971,182$
$\hat{l}_{\{C,R\}} = -358863,179$
$\hat{l}_{\{A,R\}} = -349469,660$
$\hat{l}_{\{R\}} = -348931,459$
$\hat{l}_{\{A,C\}} = -357688,330$
$\hat{l}_{\{C\}} = -357325,653$
$\hat{l}_{\{A\}} = -347387,557$

and we obtained the test statistics presented in Table 1.

Table 1. Likelihood Ratio Tests on input from customers with account manager.

Factors	Statistical test	Degrees of Freedom	Tabulated Chi-Square value
$\{A, C, R\}$	25446,3625	395	442,3406
$\{C, R\}$	25230,3560	393	440,2233
$\{A, R\}$	6443,3180	392	439,1646
$\{R\}$	5366,9176	384	430,6919
$\{A, C\}$	22880,6597	363	408,4271
$\{C\}$	22155,3047	297	338,1930
$\{A\}$	2279,1122	264	302,8983

Therefore all hypotheses are rejected to $\alpha = 5\%$.

3.2. Testing hypotheses of the departure of classes from customers with account manager

From the data, we obtained:

$$\hat{l}_{\{\Omega\}} = -85661,169$$
$$\hat{l}_{\{A,C,R\}} = -94250,903$$
$$\hat{l}_{\{C,R\}} = -93453,429$$
$$\hat{l}_{\{A,R\}} = -87465,349$$
$$\hat{l}_{\{R\}} = -86281,823$$
$$\hat{l}_{\{A,C\}} = -93966,695$$
$$\hat{l}_{\{C\}} = -93026,854$$
$$\hat{l}_{\{A\}} = -87111,204$$

and we obtained the test statistics presented in Table 2.

Table 2. Likelihood Ratio Tests of the departure of classes from customers with account manager.

Factors	Statistical test	Degrees of Freedom	Tabulated Chi-Square value
$\{A, C, R\}$	17179,4680	395	442,3406
$\{C, R\}$	15584,5189	393	440,2233
$\{A, R\}$	3608,3605	392	439,1646
$\{R\}$	1241,3080	384	430,6919
$\{A, C\}$	16611,0521	363	408,4271
$\{C\}$	14731,3691	297	338,1930
$\{A\}$	2900,0698	264	302,8983

Therefore all hypotheses are rejected to $\alpha = 5\%$.

3.3. Tests of hypotheses on input from customers without account manager

From the data we obtained

$$\hat{l}_{\{\Omega\}} = -777828,669$$
$$\hat{l}_{\{A,C,R\}} = -792205,794$$

$\hat{l}_{\{C,R\}} = -791286,373$
$\hat{l}_{\{A,R\}} = -784228,704$
$\hat{l}_{\{R\}} = -782195,942$
$\hat{l}_{\{A,C\}} = -789259,040$
$\hat{l}_{\{C\}} = -787948,026$
$\hat{l}_{\{A\}} = -780549,942$

and we obtained the test statistics presented in Table 3.

Table 3. Likelihood Ratio Tests on input from customers without account manager.

Factors	Statistical test	Degrees of Freedom	Tabulated Chi-Square value
$\{A, C, R\}$	28754,2491	395	442,3406
$\{C, R\}$	26915,4082	393	440,2233
$\{A, R\}$	12800,0692	392	439,1646
$\{R\}$	8734,5452	384	430,6919
$\{A, C\}$	22860,7418	363	408,4271
$\{C\}$	20238,7134	297	338,1930
$\{A\}$	5442,5460	264	302,8983

Therefore all hypotheses are rejected to $\alpha = 5\%$.

3.4. Testing hypotheses of the departure of classes from customers without account manager

From the data we obtained

$\hat{l}_{\{\Omega\}} = -861900,408$
$\hat{l}_{\{A,C,R\}} = -897576,772$
$\hat{l}_{\{C,R\}} = -897381,957$
$\hat{l}_{\{A,R\}} = -866675,515$
$\hat{l}_{\{R\}} = -864720,165$
$\hat{l}_{\{A,C\}} = -896267,745$
$\hat{l}_{\{C\}} = -895738,495$
$\hat{l}_{\{A\}} = -864414,771$

and we obtained the test statistics presented in Table 4.

Table 4. Likelihood Ratio Tests of the departure of classes from customers without account manager.

Factors	Statistical test	Degrees of Freedom	Tabulated Chi-Square value
$\{A, C, R\}$	71352,7269	395	442,3406
$\{C, R\}$	70963,0982	393	440,2233
$\{A, R\}$	9550,2139	392	439,1646
$\{R\}$	5639,5127	384	430,6919
$\{A, C\}$	68734,6728	363	408,4271
$\{C\}$	67676,1729	297	338,1930
$\{A\}$	5028,7255	264	302,8983

Therefore all hypotheses are rejected to $\alpha = 5\%$.

4. Partition of the variation of log-likelihood

4.1. Algebraic treatment

As we observe from the Likelihood Ratio tests all hypotheses are rejected. This conduces us to the conclusion that perhaps there is too much information. We have the four cases:

- Input from customers with account manager;
- Departure of customers with account manager;
- Input from customers without account manager;
- Departure of customers without account manager.

And it is crucial to know what relevant effects and interactions there are. We use the word relevant because, for now, we will present the theory that arises at the level of descriptive statistics as it gives us the fraction of the variation in log-likelihood attributable to each of the sets of factors. There is a clear parallel between this technique and the partition of the sum of squares for ANOVA with factors interaction.

When you have L factors with $j_1, \ldots, j_L$ levels we have $n^0 = \prod_{l=1}^{L} j_l$ possible treatments.

We then have an orthogonal partition, see Fonseca *et al.* (2003, 2006).

$$R^{n0} = \boxplus_{h \in \Gamma} \nabla(h),$$

where $\boxplus$ indicates direct sum of orthogonal sub-spaces and $\Gamma(h; h_l = 0, 1; l = 1, \ldots, L)$.

In particular, $\nabla(0)$ will be the sub-space formed by vectors with equal components. So, given a vector v whose components have average $v.$, we will have

$$\nu - 1\nu_{\cdot} = \sum_{h \in \Gamma \setminus \{0\}} s(h)$$

with $s(h)$ denoting the square of the norm of the orthogonal projection of v on $\nabla(h)$. Thus, $s(h)$ will be proportional to the variation in the v components attributable to the factor or factors and the index is:

$$\varsigma(h) = \{l : h_l = 1\}.$$

Being $T(h)$ the vector components of the total v corresponding to combinations of factor levels with $\varsigma(h)$ indexes we have, see Fonseca *et al.* (2003, 2006)

$$s(h) = \sum_{0} (-1)^{()\varsigma(h) - \varsigma(k))} \frac{\|T(k)\|^2}{\prod_{l \notin \varsigma(h)} j_l}.$$

On the analysis of variance $s(h)$ divided by $d(h)$, the dimension of $\nabla(h)$, will give us the "mean square": $qm(h) = \frac{s(h)}{d(h)}$, being, see Fonseca *et al.* (2003, 2006), the dimensions given by $d(h) = \prod_{l \in \varsigma(h)} (j_l - 1)$. One can then calculate the mean squares:

$$qmr(h) = \frac{qm(h)}{\sum_{h \notin \Gamma}} qm(h),$$

which will measure the relevance of the various sets of factors.

Let us now try to clarify the relationship between the approaches of the Likelihood Ratio Test and the Partition of the variation of the logarithm of the likelihood. Represent by ζ^c the complement of ζ. For example, to $\zeta = \{A, R\}$, we will have ζ^c. Now we can show that

$$\omega(\zeta) = (\boxplus \nabla(\zeta'))^{\perp},\ \zeta' \subseteq \zeta^c,\ \text{where } \perp \text{ shows the orthogonal complement.}$$

The hypotheses for the Likelihood Ratio tests were of the form:

$$H_0(\zeta) : q \in \omega(\zeta).$$

Then we have the hypotheses:

$$H_0^0(\zeta) : q \in \nabla(\zeta)^{\perp}.$$

5. Applications

The results of this technique can be presented in a summary table similar to the analysis of variance. The effects of factors are indicated by their symbols in our application, namely: A, C and R and their interactions, AxC, AXR, CXR and $AxCxR$.

5.1. Input of customer with account manager

We have the summary table

Table 5. Summary of entries in the classes of customers with account manager.

Source of Variation	Sum of Squares "$s(h)$"	Dimension	Mean square "qm"	Relative mean square "qmr"
$\{A\}$	216,006	2	108,003	0,016
$\{C\}$	19003,044	3	6334,348	0,946
$\{A, C\}$	860,394	6	143,399	0,021
$\{R\}$	2565,703	32	80,178	0,012
$\{A, R\}$	509,349	64	7,959	0,001
$\{C, R\}$	1598,503	96	16,651	0,002
$\{A, C, R\}$	693,363	192	3,611	0,001

This results in the status table of the dominant class factor, and it was graphically:

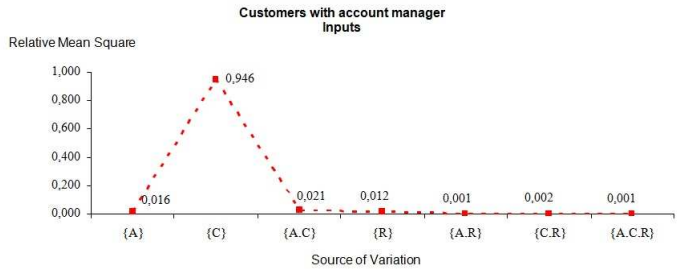

Figure 2. Adjustment of the mean square relative to the input of customers with account manager.

5.2. Outputs from customers with account manager

We have the summary table

Table 6. Summary table of departures of customers with account manager from classes.

Source of Variation	Sum of Squares "$s(h)$"	Dimension	Mean square "qm"	Relative mean square "qmr"
$\{A\}$	1594,949	2	797,475	0,146
$\{C\}$	13571,108	3	4523,703	0,826
$\{A, C\}$	772,103	6	128,684	0,024
$\{R\}$	568,416	32	17,763	0,003
$\{A, R\}$	284,734	64	4,449	0,001
$\{C, R\}$	139,875	96	1,457	0,000
$\{A, C, R\}$	248,283	192	1,293	0,000

This results in the status table of the dominant class factor, and graphically we have:

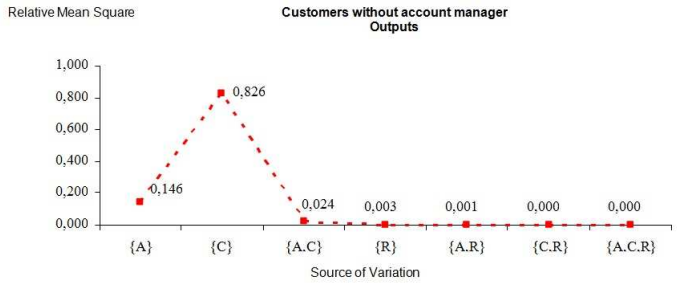

Figure 3. Adjustment of the mean square relative to the outputs from customers with account manager.

5.3. Inputs from customers without account manager

We have the summary table

Table 7. Summary table of incoming customers without account manager classes.

Source of Variation	Sum of Squares "$s(h)$"	Dimension	Mean square "qm"	Relative mean square "qmr"
$\{A\}$	1838,84	2	919,420	0,135
$\{C\}$	15954,18	3	5318,060	0,779
$\{A, C\}$	2226,68	6	371,114	0,054
$\{R\}$	5893,51	32	184,172	0,027
$\{A, R\}$	783,19	64	12,237	0,002
$\{C, R\}$	1464,02	96	15,250	0,002
$\{A, C, R\}$	593,83	192	3,093	0,000

This results in the status table of the dominant class factor. Graphically we have:

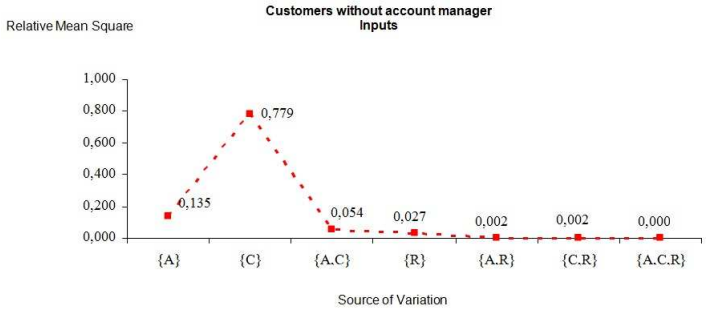

Figure 4. Adjustment of the mean square relative to the input of customers without account manager.

5.4. Outputs from customers without account manager

We have the summary Table 8.

Table 8. Summary table of outputs from customers without account manager.

Source of Variation	Sum of Squares "$s(h)$"	Dimension	Mean square "qm"	Relative mean square "qmr"
$\{A\}$	389,629	2	194,814	0,009
$\{C\}$	61802,513	3	20600,838	0,958
$\{A, C\}$	3521,073	6	586,845	0,027
$\{R\}$	2618,054	32	81,814	0,004
$\{A, R\}$	668,871	64	10,451	0,000
$\{C, R\}$	1903,434	96	19,827	0,001
$\{A, C, R\}$	449,153	192	2,339	0,000

This results in the status table of the dominant class factor. Graphically we have:

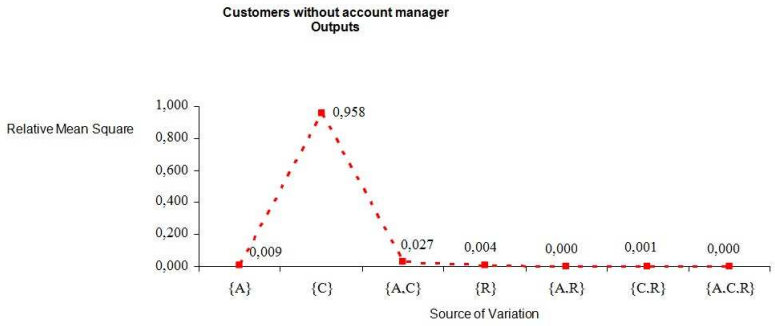

Figure 5. Adjustment of the mean square relative to the outputs of customers without account manager.

5.5. Paired comparison of populations

Given the matching, we compare for the different classes, the probability of abandonment. To lighten the writing we represent in each class by n_1 and n_2 the numbers of customers with and without account manager and by S_1 and S_2 the exit numbers and the probabilities of outputs. We want to test whether the probability of outputs and of customers with and without account manager are equal in each of the four classes. So, we then test the hypothesis:

$$H_0 : p_1 = p_2.$$

Reasoning as above we obtain:

$$\begin{cases} \hat{l}_\Omega = k + \sum_{i=1}^{2}(s_i log(\frac{s_1}{n_1}) + (n_1 - s_1)log(\frac{n_i - s_i}{n_i})), \\ \hat{l}_\omega = k + (s_1 + s_2)log(\frac{s_1+s_2}{n_1+n_2}) + (n_1 + n_2 - s_1 - s_2)log(\frac{n_1+n_2-s_1-s_2}{n_1+n_2}) \end{cases}$$

taking up the test statistic

$$V = -2(\hat{l}_\omega - \hat{l}_\Omega).$$

Given the Wilks theorem we can assume that when the hypothesis H_0 is true, V is distributed as a Chi-square central with $r = dim(\Omega) - dim(\omega) = 1$ degrees of freedom.

The results for the different classes are presented in Table 9.

Table 9. Likelihood Ratio Test for the outputs.

Class	Statistical test	Degrees of freedom	Tabulated Chi-Square value
1	3498,26660	1	3,8415
2	2717,4450	1	3,8415
3	3311,7620	1	3,8415
4	6505,8440	1	3,8415

Therefore all hypotheses are rejected to $\alpha = 5\%$.

To complete the analysis we performed a partition of Change in Logarithm of Likelihood. Let us now consider the factors class (C) and the account manager (E).

We obtained the following summary Table 10, showing the dominant character of the factor account manager. Be allocated to an account manager overrides the class in determining the degree of customer loyalty.

Table 10. Summary table of relevance of factors of class and with account manager in the outputs.

Source of Variation	Sum of Squares "$s(h)$"	Dimension	Mean square "qm"	Relative mean square "qmr"
$\{C\}$	74037,441	3	24679,147	0,25
$\{E\}$	60796,535	1	60796,535	0,61
$\{C, E\}$	44763,208	3	14921,069	0,15

6. Conclusion

We conclude that the dominant factor was either in class or in abandoned entries. Using the pairing of costumers we noticed the existence of account manager customer loyalty. This result is very interesting because it will allow us to (i) work together globally to customers without having to disassemble the same for regions, (ii) admit the homogeneity of Markov chains: the matrices of probabilities transition do not vary from year to year.

References

[1] J.T. Mexia, Introdução à Inferençia Estatística Linear (Centro de Estudos de Matemática Aplicada, Edições Universitárias Lusófonas, 1995).

[2] J.R. Norris, R. Gill, B.D. Ripley and S. Ross, Markov Chains (Cambridge Series in Statistical & Probabilistic Mathematics, 1998).

[3] N.E. Stewart and T.G. Kurtz, Markov Processes (New York, John Wiley, 1986).

[4] M. Fonseca, J.T. Mexia and R. Zmyślony, Estimators and tests for variance components in cross nested designs (StatLin, Będlewo, Poland, 2003).

[5] M. Fonseca, J.T. Mexia and R. Zmyślony, Estimação de components de variância em modelos lineares com OBS (XIV Congresso da SPE, Covilhã, Portugal, 2006).

Received 18 June 2011