# IDENTIFICAÇÃO DAS CARACTERÍSTICAS DOS CLIENTES ASSOCIADAS AO RISCO DE CRÉDITO

**LUÍS N. PEREIRA**
Mestre em Estatística e Gestão de Informação
Professor Adjunto na ESGHT – Universidade do Algarve
lmper@ualg.pt

**LUÍS R. CHORÃO**
Mestre em Estatística e Gestão de Informação
Director – Banco BPI
luis.ribeiro.chorao@bancobpi.pt

## RESUMO

O processo da tomada de decisão sobre a avaliação de uma solicitação de crédito comercial é por vezes difícil para o julgamento humano, devido à imensidão de variáveis que estão em jogo e das suas inter-relações. Neste artigo propomo-nos identificar as características dos clientes associadas a alto e a baixo risco, com recurso a um modelo aplicacional. A partir de uma base de dados de um cartão de crédito, formada por variáveis de natureza qualitativa e quantitativa, ajustámos um modelo *logit* binário, com o objectivo de tornar o processo de decisão mais objectivo e quantificável. Em seguida, identificámos oito classes de risco através da aplicação de um método de classificação não hierárquica (*K-means*) sobre o vector da pontuação do modelo *logit*. Aferimos temporalmente o comportamento de cada classe de risco ao longo de 70 meses, verificando-se que probabilidades baixas de *default* estão associadas a classes de risco baixo. As características dos clientes tipicamente associadas ao risco de crédito foram identificadas através de uma Análise Factorial das Correspondências.

## PALAVRAS CHAVE:

*Credit Scoring*, Regressão Logística, Classificação Hierárquica, Estimadores de Kaplan-Meier, Análise Factorial das Correspondências.

## ABSTRACT

The decision making process of evaluating the creditworthiness of a loan is sometimes difficult to the human mind because of the great number of variables and interrelations among them. What we propose here, is to identify the characteristics related to high and low risk, and this is made by using an applicant model. So, with a credit card database with categorical and continuous variables, in order to make the decision process more streamlined and quantifiable, we performed a binary logistic model. Applying the non-hierarchical clustering method (K-means) to the logit output vector we identified eight risk classes. Each class was evaluated temporarily by the product-limit estimators (Kaplan-Meier estimators) for 70 months, showing that low probability of default is indeed associated with low risk classes. The statistic technique applied to identify the client risk characteristics was the correspondence analysis.

## KEYWORDS:

Credit Scoring, Logistic Regression, Non-hierarchical Clustering Method, Kaplan-Meier Estimators, Correspondence Analysis.

## 1. INTRODUÇÃO

O processo da tomada de decisão quanto à concessão, ou não, de um crédito comercial a um determinado cliente é uma área em que o julgamento humano normalmente se precipita em subjectivismos contraproducentes para o negócio. Os dados do cliente, sejam eles de ordem qualitativa ou de ordem quantitativa, deverão ser considerados na avaliação de concessão de crédito sem mácula de ideias pré-concebidas. Ora, a capacidade humana de julgar correctamente um processo de solicitação de crédito é bastante limitada por força da imensidão de variáveis que estão em jogo e das interrelações que se estabelecem entre elas. Apesar da aplicação de técnicas lineares de discriminação contribuir para um melhor desempenho na avaliação do risco do cliente e da operação, o problema solicita outra

metodologia de resposta bem mais complexa, porque a metodologia da pergunta, apesar de ser um simples "Sim/Não", também o é. Surgem, assim, os modelos probabilísticos não-lineares que pretendem modelar a experiência humana, oferecendo de permeio a oportunidade de reduzir o tempo necessário à avaliação da concessão de um empréstimo, ao mesmo tempo que conferem à resposta qualidade e igualdade de critérios face a idênticas solicitações de diferentes clientes.

Numa área como o risco de crédito, a identificação de características associadas a perfis de risco é importantíssima para que a instituição financeira preste um serviço apercebido pelos seus clientes como sendo único. E com qualidade! Ora, é na exacta medida em que o banco consegue avaliar o risco do seu cliente, que lhe proporciona um serviço com qualidade: juros mais baixos, oportunidades de aquisição de novos produtos, etc. Não será, então, de admirar que o primeiro propósito deste trabalho seja o de tentar atribuir uma pontuação de risco ao cliente. Para isso, ajustou-se um modelo *logit* binário a dados provenientes de uma base de dados de um cartão de crédito de uma instituição financeira nacional. Após o cálculo da pontuação de risco associada ao reque-rente, procedeu-se à formação de grupos homogéneos (classes de risco) sobre o vector da pontuação do modelo *logit* tendo-se utilizado a classificação não-hierárquica (*K-means*). De seguida, aferiu-se se na constituição dos grupos é possível retirar critérios temporais de análise. Melhor explicitando: a entrada num estado de incumprimento dos compromissos assumidos (normalmente aferido por atraso superior a 90 dias na liquidação de uma prestação com pagamentos de base regular) e de ora em diante designado por *default* (vd. Secção 2.1. O que é o Risco), tem na sua génese duas dimensões: a dimensão do perfil do cliente - ao ser avaliado o cliente, ou é aceite, ou é rejeitado - e, a dimensão temporal - tendo sido aceite, por quanto tempo é suposto o cliente manter-se numa situação de regularidade? Ora, se a pontuação do modelo *logit* apenas se atém à primeira dimensão, importa testar se a formação de classes de risco consegue ir mais além e incorporar a dimensão temporal. Para isso, empregou-se a metodologia dos estimadores do produto-limite de Kaplan-Meier. Estes estimadores são baseados na estimação de probabilidades condicionadas, permitindo estimar a taxa de sobrevivência ou de *default* em cada momento do tempo de acordo com a diferente natureza dos dados (censurados e não censurados) e, portanto, certificar se a dimensão temporal está presente na classificação anteriormente encetada. Torna-se, assim, evidente a importância que tem para a instituição financeira o conhecimento das características associadas a cada classe de risco para a prossecução de uma rigorosa política de risco de crédito. E eis-nos, portanto, na parte nobre deste trabalho, isto é, na identificação das modalidades em confronto para a formação das diferentes classes. Para a selecção das modalidades chave fez-se uso da Análise Factorial das Correspondências. Por último, procedeu-se à formação de classes homogéneas de modalidades, através de uma classificação hierárquica com base nas componentes principais obtidas na Análise Factorial das Correspondências, tendo como objectivo a comparação dessas classes de modalidades com os conjuntos de modalidades associadas a classes de risco baixo e a classes de risco elevado.

## 2. *CREDIT SCORING*

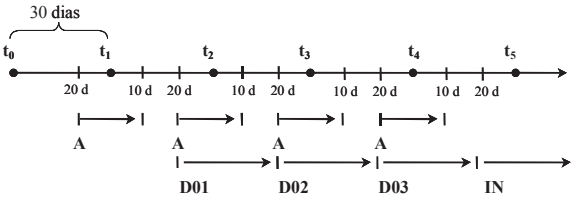
### 2.1. O QUE É O RISCO

Há indivíduos que na sua relação com a instituição financeira sempre primaram pelo cumprimento escrupuloso do pagamento das prestações a que ficaram sujeitos pela contratação de um determinado crédito e, outros há que, não tendo sido tão escrupulosos, nunca incorreram em situações mais graves do que alguns atrasos nos pagamentos das referidas prestações acrescidas das habituais penalizações (*late-charge*[1]) e nos juros de mora a que ficaram obrigados pela contratação do empréstimo. Todavia, existe um outro conjunto de indivíduos que ultrapassam todos os prazos para a liquidação das supra referidas prestações nunca vindo a inverter o seu plano de pagamento. Assim, desde que um indivíduo tenha um atraso de 90 dias no pagamento de uma prestação, diz-se que o indivíduo entrou em *default*. Ou seja, *default* é o estado de "insistência" de uma situação de delinquência/incumprimento. Na tabela 1 apresentam-se os *status* da conta-cartão e na figura 1 ilustra-se o percurso de entrada em *default*.

**Tabela 1: *Status* da conta-cartão**

| ***Status*** | **Descrição** |
|---|---|
| AA | Regular |
| D0Y | Delinquência; Y=1, 2, 3 |
| DX0 | *Overline*; X=1, 2, 3 |
| BXY | Delinquência e *overline* |
| IN | ***Default*** |

**Figura 1: Marcação em *status* de *default***




A - saída de extracto com pagamento até ao dia 10 do mês seguinte

Aquilo que melhor traduz a actividade bancária, no que diz respeito ao fenómeno da intermedição financeira, é a qualidade da carteira de empréstimos e não tanto a capacidade de recuperar créditos em *default*. E a carteira de empréstimos considera-se boa se a qualidade dos devedores o for igualmente. Assim, para aferir se um determinado indivíduo, aquando da sua avaliação, tem perfil de bom ou de mau pagador, empresta-se-lhe uma pontuação que resulta da aplicação de um determinado algoritmo. Tal algoritmo compreende a aplicação de um modelo que responde à seguinte questão: qual é a probabilidade de *default* associada ao requerente? A resposta a esta questão encontra-se na adopção do comummente designado *credit scoring*.

### 2.2. CREDIT SCORING

O *credit scoring* pode ser definido como um método de avaliação de risco de crédito dos clientes através da implementação de uma fórmula ou conjunto de regras. O paradigma em que se baseia o *credit scoring* é na aprendizagem dos clientes actuais, na identificação das características dos bons clientes, e na concessão de crédito a novos clientes com características semelhantes às daqueles. O método produz um "*score*" que o banco utiliza para posicionar o crédito em termos de risco e decidir quanto à concessão, ou não, do crédito (Loretta, 1997).

Vários métodos estatísticos são utilizados para desenvolver sistemas de *credit scoring*, dos quais se destacam os modelos de probabilidade linear, os modelos *logit*, os modelos *probit*, métodos baseados em árvores de decisão e modelos de análise discriminante. Mais recentemente "apareceram" modelos cujo suporte assenta em redes neuronais. Apresentam-se as vantagens e limitações associadas à utilização de um modelo de *scoring*. Algumas das vantagens que se podem enumerar pela utilização do *credit scoring* são as seguintes:

- Identificação das variáveis preditivas (ou mais discriminatórias);
- Capacidade de análise de um maior número de propostas;
- Redução do risco de crédito;
- Automatização (rapidez de decisão, precisão, menor carga de trabalho);
- Erro humano eliminado, pela objectividade do método;
- Consistência na análise;
- Redução do tempo de resposta, com impacto na qualidade de serviço.

As limitações do *credit scoring* prendem-se, sobretudo, com as questões estatísticas em que se baseiam os modelos. Podem-se apontar:

- A precisão. Apesar da redução dos custos no processo de avaliação, um modelo não suficientemente preciso poderá conduzir a situações danosas na concessão do crédito;
- A qualidade dos dados.

## 3. MODELO *LOGIT* PARA DADOS BINÁRIOS

Foi aplicado um modelo *logit* para a modelação dos dados por duas ordens de razão. Em primeiro lugar, decorrente de um ponto de vista histórico: a utilização do modelo *logit* e concomitante apuramento dos *odds* (rácio entre a probabilidade de entrar em *default* e o seu complementar) subjacentes à análise de risco ocorre desde a massificação dos cartões de crédito. Em segundo lugar porque o modelo *logit* tem muitas das propriedades desejáveis do modelo de regressão linear. Este modelo é linear nos seus parâmetros, tem domínio em $\Re$ e tem como contradomínio o intervalo ]0,1[, intervalo este que apresenta, em termos de risco de crédito, a probabilidade do cliente possuir perfil de *default*.

Seja $Y_i$ uma variável binária que representa a situação do i-ésimo indivíduo. Define-se que $y_i=1$ se o indivíduo i entra em *default* e $y_i=0$ se o indivíduo i não entra em *default*. Tem-se portanto que $y_i$ é a realização da variável aleatória dependente, $Y_i$, em que $P(Y_i=1)=\pi_i$ e $P(Y_i=0)=1-\pi_i$. O modelo *logit* caracteriza-se por estimar directamente a probabilidade de um indivíduo entrar em *default*, assumindo que a mesma tem a forma logística:

$$P(Y_i=1)=\pi_i=\frac{e^{X_{i\bullet}\beta}}{1+e^{X_{i\bullet}\beta}} \qquad (1)$$

onde $X_{i\bullet}=(X_{i1},\ldots,X_{ik})$ é um vector de dimensão k corresponde à i-ésima linha da matriz formada pelas k variáveis explicativas e $\beta=(\beta_1,\ldots,\beta_k)'$ é o vector dos coeficientes da regressão da mesma dimensão. Calculada a probabilidade de um indivíduo entrar em *default*, comparada com a probabilidade desse mesmo indivíduo não entrar em *default*, ou seja,

o *odd* de sucesso, e aplicando logaritmos a ambos os membros da equação daí resultante, obtém-se o seguinte modelo linear:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta \qquad (2)$$

Pode, portanto, concluir-se que usar uma função logística para modelar a probabilidade de um indivíduo entrar em *default* (sucesso), é equivalente a ajustar um modelo de regressão linear onde a variável dependente é substituída pelo logaritmo do *odd* de sucesso. O modelo (2) foi estimado pelo método da máxima verosimilhança.

Assim, sobre uma base de dados de cartões de crédito de uma instituição financeira nacional, foi estimado um modelo *logic* binário para particulares residentes. O conceito modelado, *default*, é definido como o *status* em que o cliente fica classificado após 90 dias de atraso no pagamento de uma prestação. Foram analisados cerca de 130000 contas-cartão, contudo, o desenvolvimento do modelo assentou em 23706 contas-cartão dada a "estabilização" requerida na janela de amostragem. Foram, ainda, retirados da amostra as delinquências de 3.º nível para melhor poder diferenciar os dois conjuntos em modelação.

O teste para o modelo completo, com todos os parâmetros contra o modelo apenas com a constante é estatisticamente aceitável, $\chi 2$ (g.l. = 24, N = 23706) = 912,542, *p-value* < 0,001, indicando que os parâmetros, tomados em conjunto, diferenciam entre clientes em *default* e clientes regulares[2].

**Tabela 2: Teste *Omnibus* aos coeficientes do modelo (passo 1)**

| $\chi^2$ | g.l. | p |
|---|---|---|
| 912,452 | 24 | 0,000 |

Este resultado é corroborado pela curva ROC (*Receiver Operating Characteristic*) que apresenta uma área igual a 72,4%. Além disso, o teste de Hosmer-Lemeshow configura uma boa modelação já que apresenta um valor $\chi 2$ (g.l. = 8, N = 23706) = 6,144, sendo os resultados apresentados nas tabelas 3 e 4.

**Tabela 3: Tabela de Contingência para o teste de Hosmer-Lemeshow (passo 1)**

| Decil | Regular | | *Default* | | Total |
|---|---|---|---|---|---|
| | Freq. Observada | Freq. Esperada | Freq. Observada | Freq. Esperada | |
| 1 | 2341 | 2346,0 | 30 | 25,0 | 2371 |
| 2 | 2331 | 2328,2 | 40 | 42,8 | 2371 |
| 3 | 2302 | 2312,4 | 69 | 58,6 | 2371 |
| 4 | 2295 | 2295,8 | 76 | 75,2 | 2371 |
| 5 | 2278 | 2277,4 | 93 | 93,5 | 2371 |
| 6 | 2256 | 2255,9 | 115 | 115,1 | 2371 |
| 7 | 2241 | 2228,6 | 130 | 142,4 | 2371 |
| 8 | 2202 | 2189,4 | 168 | 180,6 | 2371 |
| 9 | 2130 | 2125,3 | 241 | 245,7 | 2371 |
| 10 | 1927 | 1944,0 | 441 | 424,0 | 2368 |

**Tabela 4: Teste de Hosmer-Lemeshow (passo 1)**

| $\chi^2$ | g.l. | p |
|---|---|---|
| 6,144 | 8 | 0,631 |

Pigeon e Heyse (1999) notaram que a aproximação à distribuição do qui-quadrado depende de se terem frequências esperadas suficientemente grandes. Alguns problemas poderão advir quando as probabilidades estimadas se aproximam de zero ou de um. Todavia, a estratégia de ordenar os decis de risco (conforme são denominados os decis formados pelo teste de Hosmer-Lemeshow) agrupa deliberadamente todas as baixas probabilidades e as elevadas probabilidades para os acontecimentos em conjunto, pelo que é possível ter para os primeiros grupos, baixas frequências esperadas para os "sucessos" e para os últimos grupos, frequências esperadas baixas para os "insucessos". Estes autores afirmaram ainda, que em alguns casos estas frequências esperadas são menores do que um, o que invalida, claramente, a aproximação do teste estatístico à distribuição do qui-quadrado.

Assim, Pigeon e Heyse (1999) propuseram uma ordenação baseada nas covariáveis, mostrando vantagens relativamente ao teste de Hosmer-Lemeshow, uma vez que permite uma distribuição

mais equitativa das pequenas e elevadas probabilidades estimadas além de que, a aproximação à estatística do qui-quadrado se faz com (g - 1) graus de liberdade. Desta forma, ordenando as pontuações pela variável agregado familiar, obteve-se o seguinte valor $\chi 2$ (g.l. = 9, N = 23706) = 6,203, o que traduz bem a boa aderência da realidade observada.

Procedeu-se, ainda, a uma análise dos resíduos. Elegeu-se o teste de *Pearson residual*. Este teste identifica as observações com pouca aderência ao modelo estimado. Se na regressão linear múltipla se assume que o erro é independente da média condicional de Y, na regressão logística a variância do erro é função da média condicional. Eis por que, os resíduos deverão ser estandardizados e ajustados pelos seus desvios-padrões. Assim, o teste de *Pearson* mais não é do que a diferença entre os valores observados e as probabilidades estimadas divididos pelo desvio-padrão da probabilidade estimada (Menard, 2001). Para grandes amostras, o resíduo estandardizado, deverá seguir uma distribuição normal com média nula e desvio-padrão unitário, sendo que 95% das suas observações deverão estar compreendidas entre -1,96 e +1,96.

Por último, salienta-se ainda que os coeficientes de regressão obedeceram ao critério de inclusão subjacente à estatística de Wald com *p-value* < 0,05.

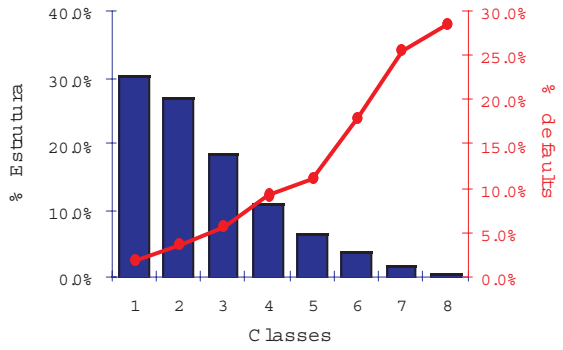
**Tabela 5: *Pearson residual***

| Estado | Frequência | Percentagem |
|---|---|---|
| Regular | 22563 | 95,2% |
| *Default* | 1143 | 4,8% |
| Total | 23706 | 100,0% |

## 4. Formação de Classes de Risco

Conforme referido na Secção 1, pretende-se agora agrupar os indivíduos com características comuns de risco. Elegeu-se a classificação não-hierárquica (*K-means*) para fazer a partição dos 23706 indivíduos em oito classes de risco por duas ordens de razão. Primeiro, por ser um meio expedito de criar classes homogéneas; e, segundo, porque a dimensão da base de dados, formada por 23706 indivíduos, torna inacessível a utilização de uma classificação hierárquica. A distância utilizada pelo método *K-means* é a distância Euclideana e o critério de convergência adoptado foi o de 0,02, isto é, a iteração para a formação de grupos pára quando uma iteração completa não move nenhum dos centros das classes por uma distância maior do que 2% da menor distância entre qualquer dos centros iniciais das classes. Na figura seguinte apresenta-se a taxa de *default* e a estrutura de efectivos em cada classe.

**Gráfico 1: Taxa de *default* por classe de risco**




Decorre da análise do gráfico 1 que a utilização da classificação não-hierárquica proporcionou a formação de classes de risco, em que se evidencia que a classes com menor pontuação resultante do modelo *logit* correspondem menores taxas de *default* e vice-versa, como era aliás mister encontrar.

## 5. Estimadores de Kaplan-Meier

Conforme atrás aludido, a pontuação do modelo *logit* apenas se atém à dimensão da classificação de risco, importando também verificar se a formação de classes de risco consegue ir mais além e incorporar a dimensão temporal. Uma vez que a análise comporta indivíduos censurados e não-censurados, empregou-se a metodologia dos estimadores do produto-limite de Kaplan-Meier (KM), a qual permite estimar a função de sobrevivência com dados do tempo de vida.

Seja $0 < t(1) < \ldots < t(r) < \infty$ um conjunto de momentos do tempo nos quais se observaram dados do tempo de vida, censurados ou não censurados. O estimador de KM, que dá a probabilidade de um indivíduo não entrar em *default* até ao momento $t_{(j)}$, é dado por (Kaplan e Meier, 1958):

$$\hat{S}(t)=\prod_{j:t_{(j)}}\left(\frac{n_j-d_j}{n_j}\right) \qquad (3)$$

onde $n_j$ é o número de indivíduos sob risco e $d_j$ é o número de incumprimentos no momento $t_{(j)}, j=1, \ldots, r$. A probabilidade acumulada de um indivíduo entrar em incumprimento até ao momento $t_{(j)}$ é então dada por $1-\hat{S}(t)$.

A observação dos indivíduos foi realizada durante um período de 70 meses. Por esta razão, os indivíduos encontram-se classificados em dois grandes grupos: os indivíduos sobre os quais foi possível observar o seu comportamento desde o momento da concessão de cartão até à entrada em *default*,

designados por não censurados; e, aqueles aos quais foi concedido cartão de crédito e que à data da extracção do ficheiro se encontravam numa situação regular, designados por censurados.

**Tabela 6: Função Cumulativa de *default* segundo o método de KM**

| Meses | Classe 1 | Classe 2 | Classe 3 | Classe 4 | Classe 5 | Classe 6 | Classe 7 | Classe 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% |
| 2 | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% |
| 3 | 0,00% | 0,00% | 0,00% | 0,04% | 0,00% | 0,00% | 0,00% | 0,00% |
| 4 | 0,00% | 0,00% | 0,00% | 0,04% | 0,00% | 0,00% | 0,00% | 0,00% |
| 5 | 0,01% | 0,03% | 0,05% | 0,07% | 0,06% | 0,00% | 0,21% | 0,00% |
| 6 | 0,07% | 0,11% | 0,25% | 0,37% | 0,35% | 0,19% | 0,83% | 0,79% |
| 7 | 0,09% | 0,24% | 0,39% | 0,63% | 0,65% | 0,66% | 1,46% | 3,17% |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 12 | 0,51% | 1,03% | 1,79% | 2,54% | 3,64% | 4,91% | 9,98% | 9,52% |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 24 | 1,42% | 2,94% | 4,19% | 6,92% | 8,35% | 13,14% | 19,75% | 23,81% |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 36 | 1,91% | 3,72% | 5,76% | 9,27% | 11,21% | 18,33% | 26,27% | 29,20% |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 48 | 1,91% | 3,79% | 6,12% | 9,44% | 12,01% | 18,80% | 30,31% | 31,56% |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 60 | 1,91% | 3,97% | 6,62% | 10,03% | 12,01% | 22,12% | 30,31% | 31,56% |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 66 | 1,91% | 4,20% | 6,62% | 11,23% | 12,01% | 23,97% | 30,31% | 31,56% |
| 67 | 1,91% | 4,20% | 6,62% | 11,23% | 12,01% | 23,97% | 30,31% | 31,56% |
| 68 | 1,91% | 4,20% | 6,62% | 11,23% | 12,01% | 23,97% | 30,31% | 31,56% |
| 69 | 1,91% | 4,20% | 6,62% | 11,23% | 12,01% | 23,97% | 30,31% | 31,56% |
| 70 | 1,91% | 4,20% | 6,62% | 11,23% | 12,01% | 23,97% | 30,31% | 31,56% |

Está-se, assim, em condições de calcular os estimadores do produto-limite. Para cada classe de risco foi modelada uma função de morte (incumprimento), $1-\hat{S}(t)$ .

Tem-se vantagem em calcular a função de incumprimento pelo método de KM e não pelas tradicionais tábuas de mortalidade, uma vez que o primeiro calcula tempos de incumprimento individuais enquanto o segundo os calcula de uma forma agrupada em intervalos. O método de KM pode, também, ser considerado um caso especial dos estimadores de uma tábua de mortalidade em que cada intervalo contém uma única observação. Apresenta-se, na tabela 6, a função cumulativa calculada pelo método de KM onde está patente a diferente probabilidade de *default* para cada classe. Da análise das funções de morte para as diferentes classes ressalta o seguinte:

- A classes de menor risco correspondem funções de "morte" mais baixas,
- Assim, a classificação não hierárquica anteriormente levada a cabo, além de associar pontuações semelhantes do modelo *logit*, legitima comportamentos distintos entre as classes de risco, e
- Confere uma dimensão temporal, como era propósito descortinar.

## 6. ANÁLISE FACTORIAL DAS CORRESPONDÊNCIAS

A Análise Factorial das Correspondências (AFC) de uma tabela de contingência pode ser vista como um processo de elaboração de uma "mensagem" que permite captar o sentido da informação da realidade observada. Segundo Crivisqui (1993) uma tabela de contingência comunica o resultado de uma observação simultânea de duas características numa dada população. Para captar o sentido da informação contido na tabela de contingência é necessário produzir uma "mensagem" que comunique essa informação. Tal mensagem comporta duas componentes:

- Componente "índice": valores de co-ocorrência das categorias observadas. A qualidade desta componente da informação veiculada pela

tabela de contingência depende da pertinência, da homogeneidade e da exaustividade com as quais se realizou a observação.

- Componente "ordem": comparação exaustiva dos elementos comparáveis da tabela. A qualidade deste aspecto da informação inerente à tabela de contingência depende do carácter exaustivo das comparações que são feitas e da pertinência com a qual se realizam essas comparações.

No presente estudo interessa, sobretudo, realçar este segundo aspecto de informação contida numa tabela de contingência. Também segundo Diday *et al* (1982), a técnica matemática utilizada na AFC é semelhante à da Análise em Componentes Principais. Difere desta última pela escolha de um tipo de distância particular: a distância do qui-quadrado. De facto, efectuar-se-ão dois tipos de análise paralelas, uma sobre a nuvem $N_{(I)}$ dos "perfis de linha", e outra sobre a nuvem $N_{(J)}$ dos "perfis de coluna", explorando a simetria dos conjuntos I e J. Apesar da representação gráfica da informação de uma tabela de contingência permitir analisar quais as modalidades de cada variável que contribuem para estabelecer uma certa associação entre as mesmas, o que se traduz graficamente por uma certa deformação (relativamente à esfericidade) das nuvens de pontos-perfil correspondentes (Lebart *et al*, 2000), não deverá o analista descurar uma análise detalhada dos quadros das contribuições absolutas e relativas para a interpretação rigorosa dos dados (Gomes, 1993).

## 6.1. Identificação das Características Chave no Risco de Crédito

Pela actuação do modelo *logit* em conjunção com a classificação não hierárquica, foi possível eleger oito classes de risco. Enceta-se agora, uma análise sobre a tabela de contingência que cruza as oito classes de risco (da classe de menor risco, a classe 1, até à classe de maior risco, a classe 8) com as cinquenta e quatro características dos 23706 indivíduos que fazem parte da base de dados. Tratando-se de uma tabela de contingência, cada par ordenado (classe; modalidade) representa o número de indivíduos detentores daquelas duas características. A AFC é o método adequado para estudar as associações existentes entre as classes de risco e as características dos indivíduos representadas na tabela de contingência.

### 6.1.1. Valores Próprios

A realização da AFC sobre a tabela de contingência anteriormente apresentada forneceu os seguintes resultados sobre os valores próprios, percentagem de inércia explicada e valores do qui-quadrado de contingência associados a cada um dos eixos. Da tabela 7 ressalta imediatamente o número máximo de sete dimensões, uma vez que a partir da diagonalização da matriz são obtidos tantos valores próprios quantas as dimensões menos um, assim como os respectivos vectores próprios associados. Observa-se que os dois primeiros eixos são suficientes para explicar a informação contida no quadro de dados uma vez que reproduzem 97,6% da inércia total, ou seja, que contribuem em 97,6% para o valor do qui-quadrado de contingência. O produto do traço t = 0,079004 pelo efectivo total tem como resultado o valor do qui-quadrado: 33711,69 = 426708 x 0,079004. Na hipótese de independência das linhas e colunas da tabela de contingência, este valor é uma realização de um qui-quadrado com 371 graus de liberdade. Neste caso, a hipótese da independência é claramente rejeitada.

**Tabela 7: Valores próprios, percentagem de inércia e valores do qui-quadrado**

| Eixo | Valor Próprio | Qui-Quadrado[a] | Percentagem de Inércia | |
|---|---|---|---|---|
| | | | Simples | Acumulada |
| 1 | 0,06974 | 29760,34 | 88,3% | **88,3%** |
| 2 | 0,00738 | 3150,32 | 9,3% | **97,6%** |
| 3 | 0,00117 | 500,00 | 1,5% | 99,1% |
| 4 | 0,00039 | 166,71 | 0,5% | 99,6% |
| 5 | 0,00018 | 76,33 | 0,2% | 99,8% |
| 6 | 0,00008 | 34,98 | 0,1% | 99,9% |
| 7 | 0,00005 | 23,01 | 0,1% | 100,0% |
| Total | **0,07900** | 33711,69 | 100,0% | - |

a. 371 graus de liberdade

Em geral, são suficientes dois ou três eixos principais para estudar a associação existente entre as modalidades (com estatuto de indivíduos) da nuvem $N_{(I)}$ e as modalidades da nuvem $N_{(J)}$. Relembre-se que a AFC não mede a intensidade da associação existente entre as variáveis, mas apenas identifica associações entre modalidades. Na interpretação de cada eixo principal de inércia foram seguidas as seguintes regras:

- Em primeiro lugar, identificaram-se as modalidades que mais contribuem para a formação de cada eixo principal, ou seja, as modalidades responsáveis pela construção de cada eixo. Consideraram-se como modalidades determinantes da nuvem $N_{(I)}$ para a formação do eixo aquelas que têm uma contribuição absoluta (CTA) superior a $1/n$ , isto é 1/8, e da nuvem $N_{(J)}$ as que têm um CTA superior a $1/p$ , isto é 1/54;
- De seguida identificaram-se as oposições (de sinal) entre as modalidades;
- Por último, estudou-se a qualidade da representação de cada modalidade no eixo principal através das contribuições relativas (CTR). As contribuições relativas mostram quais são as modalidades que caracterizam um dado eixo. Considerou-se que uma modalidade está bem representada num eixo principal se o seu CTR estiver compreendido entre 0,5 e 0,8 e muito bem representada se o seu CTR for superior.

### 6.1.2. COMPONENTES PRINCIPAIS NORMADAS AO VALOR PRÓPRIO DA NUVEM $N_{(I)}$, CTA E CTR

Na tabela 8 podem observar-se as coordenadas, CTA e CTR da nuvem $N_{(I)}$ associadas aos dois primeiros eixos principais de inércia. Nessa tabela pode ainda observar-se a inércia associada a cada modalidade da nuvem $N_{(I)}$, na coluna da "frequência relativa" o centro de gravidade da nuvem $N_{(J)}$ e na coluna seguinte encontra-se a distância de cada modalidade da nuvem $N_{(I)}$ ao centro de gravidade desta nuvem.

A análise da tabela 8 permite observar o seguinte:

- As classes de risco 1, 5 e 6 têm uma CTA superior à média para a formação do primeiro eixo principal de inércia, sendo também muito boa a qualidade da sua representação.
- No primeiro eixo, existe uma oposição entre a classe de risco 1 e as classes de risco 5 e 6.
- Para a formação do segundo eixo principal de inércia, são as classes 1, 2 e 3 que têm uma CTA superior à média. Todavia, a qualidade da sua representação sobre este eixo não é boa.
- O segundo eixo principal de inércia opõe a classe 1 às classes de risco 2 e 3.

**Tabela 8: Frequências relativas, distância ao centro de gravidade, coordenadas, inércia, CTA e CTR sobre os dois primeiros eixos factoriais**

| Classe de Risco | Frequência Relativa | Distância | Coordenadas | | Inércia | CTA | | CTR | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | | 1 | 2 | 1 | 2 | Total |
| 1 | 29,2% | 0,104 | **0,311** | **-0,083** | 0,030 | **0,405** | **0,274** | **0,932** | 0,067 | 0,999 |
| 2 | 26,6% | 0,014 | 0,078 | **0,083** | 0,004 | 0,023 | **0,251** | 0,427 | 0,486 | 0,913 |
| 3 | 18,6% | 0,018 | -0,103 | **0,077** | 0,003 | 0,029 | **0,151** | 0,602 | 0,336 | 0,938 |
| 4 | 11,5% | 0,057 | -0,235 | 0,023 | 0,006 | 0,091 | 0,008 | 0,975 | 0,010 | 0,985 |
| 5 | 7,2% | 0,142 | **-0,369** | -0,046 | 0,010 | **0,140** | 0,020 | **0,960** | 0,015 | 0,974 |
| 6 | 4,5% | 0,281 | **-0,511** | -0,130 | 0,013 | **0,167** | 0,102 | **0,927** | 0,060 | 0,988 |
| 7 | 2,0% | 0,419 | -0,605 | -0,204 | 0,008 | 0,106 | 0,114 | 0,872 | 0,099 | 0,971 |
| 8 | 0,5% | 0,729 | -0,718 | -0,331 | 0,004 | 0,039 | 0,079 | 0,707 | 0,151 | 0,858 |
| Total | 100,0% | - | - | - | **0,079** | 1,000 | 1,000 | - | - | - |

**Tabela 9: Contribuição absoluta (percentual) de cada classe de risco para a inércia total**

| Classe de Risco | Contribuição |
|---|---|
| 1 | 38,3% |
| 2 | 4,8% |
| 3 | 4,2% |
| 4 | 8,2% |
| 5 | 12,9% |

| Classe de Risco | Contribuição |
|---|---|
| 6 | 15,9% |
| 7 | 10,8% |
| 8 | 4,9% |
| Total | 100,0% |

A análise anterior permite concluir que o primeiro eixo opõe classes de baixo risco a classes de alto risco, enquanto o segundo eixo principal isola a 1.ª classe de risco e opõe-a às classes 2 e 3, podendo indicar uma análise mais específica. Portanto, quer no primeiro quer no segundo eixos, manifesta-se o isolamento da classe de risco mais baixa, a classe 1. É ainda de salientar que as classes de risco 1, 5 e 6, para além de serem as que têm uma CTA superior à media para a formação do primeiro eixo principal, são também as que têm uma contribuição absoluta mais elevada para a inércia total da nuvem: 38,3%, 12,9% e 15,9%, respectivamente.

### 6.1.3. Componentes Principais Normadas ao Valor Próprio da Nuvem $N_{(J)}$, CTA e CTR

Na tabela 10 podem observar-se as coordenadas, CTA e CTR da nuvem $N_{(J)}$ associadas aos dois primeiros eixos principais de inércia. Nessa tabela pode ainda observar--se a inércia associada a cada modalidade da nuvem $N_{(J)}$, na coluna da "frequência relativa" o centro de gravidade da nuvem $N_{(I)}$ e na coluna seguinte encontra-se a distância de cada modalidade da nuvem $N_{(J)}$ ao centro de gravidade desta nuvem.

**Tabela 10: Frequências relativas, distância ao centro de gravidade, coordenadas, inércia, CTA e CTR sobre os dois primeiros eixos factoriais**

| Modalidade | Frequência Relativa | Distância | Coordenadas | | Inércia | CTA | | CTR | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | | 1 | 2 | 1 | 2 | Total |
| X10_1 | 1,1% | 0,111 | -0,314 | -0,052 | 0,001 | **0,016** | 0,004 | 0,885 | 0,024 | 0,909 |
| X10_2 | 4,0% | 0,019 | 0,132 | 0,042 | 0,001 | 0,010 | 0,009 | 0,905 | 0,090 | 0,994 |
| X10_3 | **0,3%** | **0,727** | -0,722 | -0,331 | 0,002 | **0,022** | **0,045** | 0,717 | 0,151 | 0,868 |
| X10_4 | 0,1% | 0,202 | 0,428 | -0,093 | 0,000 | 0,003 | 0,001 | 0,907 | 0,042 | 0,949 |
| X20_1 | **0,6%** | **0,717** | -0,816 | -0,195 | 0,004 | **0,057** | **0,031** | 0,928 | 0,053 | 0,981 |
| X20_2 | 2,7% | 0,030 | -0,142 | 0,092 | 0,001 | 0,008 | **0,031** | 0,671 | 0,279 | 0,950 |
| X20_3 | 1,1% | 0,063 | 0,232 | 0,050 | 0,001 | 0,008 | 0,004 | 0,862 | 0,040 | 0,901 |
| X20_4 | 0,6% | 0,165 | 0,396 | -0,028 | 0,001 | 0,013 | 0,001 | 0,951 | 0,005 | 0,956 |
| X20_5 | **0,6%** | **0,573** | 0,697 | -0,295 | 0,003 | **0,040** | **0,068** | 0,848 | 0,152 | 0,999 |
| X300_1 | 1,0% | 0,160 | -0,396 | -0,052 | 0,002 | **0,023** | 0,004 | 0,977 | 0,017 | 0,994 |
| X300_2 | 1,5% | 0,006 | -0,047 | -0,060 | 0,000 | 0,001 | 0,007 | 0,343 | 0,558 | 0,901 |
| X300_3 | 2,5% | 0,019 | 0,131 | 0,042 | 0,001 | 0,006 | 0,006 | 0,894 | 0,091 | 0,985 |
| X300_4 | 0,5% | 0,067 | 0,247 | 0,071 | 0,000 | 0,005 | 0,004 | 0,908 | 0,075 | 0,983 |
| X300_5 | 0,1% | 0,058 | 0,223 | -0,031 | 0,000 | 0,000 | 0,000 | 0,862 | 0,017 | 0,879 |
| X40_1 | 1,7% | 0,303 | 0,497 | -0,234 | 0,005 | **0,059** | **0,124** | 0,815 | 0,181 | 0,996 |
| X40_2 | 2,4% | 0,092 | -0,289 | 0,087 | 0,002 | **0,029** | **0,025** | 0,912 | 0,083 | 0,994 |
| X40_3 | 1,5% | 0,026 | -0,086 | 0,121 | 0,000 | 0,002 | **0,029** | 0,284 | 0,559 | 0,843 |
| X50_1 | 0,8% | 0,261 | -0,487 | -0,143 | 0,002 | **0,027** | **0,022** | 0,908 | 0,078 | 0,986 |
| X50_2 | 4,8% | 0,007 | 0,082 | 0,024 | 0,000 | 0,005 | 0,004 | 0,908 | 0,078 | 0,986 |
| X60_1 | 3,1% | 0,030 | -0,165 | 0,049 | 0,001 | 0,012 | 0,010 | 0,915 | 0,080 | 0,995 |
| X60_2 | 2,5% | 0,048 | 0,210 | -0,062 | 0,001 | 0,015 | 0,013 | 0,915 | 0,080 | 0,995 |
| X70_1 | 1,8% | 0,009 | -0,089 | -0,024 | 0,000 | 0,002 | 0,002 | 0,863 | 0,065 | 0,928 |
| X70_2 | 3,7% | 0,002 | 0,044 | 0,012 | 0,000 | 0,001 | 0,001 | 0,863 | 0,065 | 0,928 |
| X80_1 | 1,8% | 0,003 | -0,017 | 0,052 | 0,000 | 0,000 | 0,007 | 0,094 | 0,862 | 0,956 |
| X80_2 | 1,0% | 0,365 | -0,581 | -0,157 | 0,004 | **0,046** | **0,031** | 0,923 | 0,067 | 0,990 |
| X80_3 | 0,6% | 0,112 | -0,309 | 0,046 | 0,001 | 0,008 | 0,002 | 0,851 | 0,019 | 0,870 |
| X80_4 | 1,1% | 0,048 | 0,176 | 0,122 | 0,001 | 0,005 | **0,022** | 0,646 | 0,309 | 0,955 |
| X80_5 | 1,1% | 0,261 | 0,500 | -0,094 | 0,003 | **0,041** | 0,014 | 0,958 | 0,034 | 0,992 |
| X90_1 | 1,4% | 0,119 | -0,331 | -0,091 | 0,002 | **0,022** | 0,016 | 0,920 | 0,069 | 0,989 |
| X90_2 | 2,3% | 0,016 | 0,124 | 0,010 | 0,000 | 0,005 | 0,000 | 0,983 | 0,006 | 0,989 |
| X90_3 | 1,0% | 0,005 | 0,029 | 0,060 | 0,000 | 0,000 | 0,005 | 0,172 | 0,719 | 0,891 |
| X90_4 | 0,9% | 0,033 | 0,171 | 0,055 | 0,000 | 0,004 | 0,004 | 0,883 | 0,093 | 0,976 |
| X100_1 | 3,8% | 0,069 | 0,261 | 0,021 | 0,003 | **0,037** | 0,002 | 0,991 | 0,007 | 0,998 |
| X100_2 | **0,5%** | **0,882** | -0,864 | -0,284 | 0,005 | **0,056** | **0,057** | 0,847 | 0,091 | 0,938 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| X100_3 | 1,3% | 0,191 | -0,421 | 0,055 | 0,002 | **0,032** | 0,005 | 0,930 | 0,016 | 0,945 |
| X110_1 | 3,9% | 0,021 | 0,136 | 0,050 | 0,001 | 0,010 | 0,013 | 0,875 | 0,120 | 0,995 |
| X110_2 | 0,1% | 0,143 | 0,292 | -0,218 | 0,000 | 0,001 | 0,006 | 0,596 | 0,334 | 0,930 |
| X110_3 | 1,6% | 0,133 | -0,347 | -0,109 | 0,002 | **0,028** | **0,026** | 0,907 | 0,089 | 0,996 |
| X120_1 | 1,8% | 0,489 | -0,683 | -0,129 | 0,009 | **0,123** | **0,041** | 0,953 | 0,034 | 0,987 |
| X120_2 | 3,7% | 0,120 | 0,338 | 0,064 | 0,005 | **0,061** | **0,020** | 0,953 | 0,034 | 0,987 |
| X130_1 | 1,0% | 0,004 | -0,045 | -0,014 | 0,000 | 0,000 | 0,000 | 0,477 | 0,048 | 0,524 |
| X130_2 | 4,6% | 0,000 | 0,010 | 0,003 | 0,000 | 0,000 | 0,000 | 0,477 | 0,048 | 0,524 |
| X140_1 | 2,1% | 0,002 | -0,040 | -0,012 | 0,000 | 0,001 | 0,000 | 0,829 | 0,071 | 0,900 |
| X140_2 | 3,5% | 0,001 | 0,023 | 0,007 | 0,000 | 0,000 | 0,000 | 0,829 | 0,071 | 0,900 |
| X150_1 | 2,6% | 0,014 | -0,114 | -0,030 | 0,000 | 0,005 | 0,003 | 0,916 | 0,062 | 0,978 |
| X150_2 | 1,8% | 0,001 | 0,005 | 0,020 | 0,000 | 0,000 | 0,001 | 0,026 | 0,418 | 0,444 |
| X150_3 | 1,2% | 0,063 | 0,245 | 0,034 | 0,001 | 0,010 | 0,002 | 0,959 | 0,019 | 0,978 |
| X160_1 | 3,1% | 0,000 | -0,004 | 0,000 | 0,000 | 0,000 | 0,000 | 0,046 | 0,000 | 0,046 |
| X160_2 | 2,4% | 0,000 | 0,005 | 0,000 | 0,000 | 0,000 | 0,000 | 0,046 | 0,000 | 0,046 |
| X170_1 | 3,7% | 0,000 | -0,003 | 0,006 | 0,000 | 0,000 | 0,000 | 0,048 | 0,141 | 0,189 |
| X170_2 | 1,8% | 0,001 | 0,006 | -0,011 | 0,000 | 0,000 | 0,000 | 0,048 | 0,141 | 0,189 |
| X400_1 | 0,6% | 0,299 | -0,536 | 0,038 | 0,002 | **0,026** | 0,001 | 0,961 | 0,005 | 0,966 |
| X400_2 | 3,6% | 0,044 | -0,172 | 0,118 | 0,002 | 0,015 | **0,069** | 0,673 | 0,319 | 0,992 |
| X400_3 | 1,3% | 0,647 | 0,726 | -0,343 | 0,009 | **0,100** | **0,210** | 0,815 | 0,182 | 0,997 |
| Total | 100,0% | - | - | - | **0,0790** | 1,000 | 1,000 | - | - | - |

A partir da análise da tabela 10 podem retirar-se as seguintes conclusões:

- As modalidades que têm uma CTA superior à média para a formação dos primeiro e segundo eixos principais de inércia, assim como os respectivos sinais das coordenadas nestes eixos. Estas modalidades estão sistematizadas nas tabelas 11 e 12, respectivamente para o primeiro e segundo eixos.
- A CTR do segundo eixo para cada uma das modalidades é muito baixa, tal como seria de esperar, dada a excelente qualidade de representação da esmagadora maioria das modalidades no primeiro eixo principal.

Observe-se ainda que as modalidades divorciado e separado judicialmente (X10_3), idade menor ou igual a 25 anos (X20_1), idade superior 56 anos (X20_5) e habitação arrendada (X100_2) apresentam frequências relativas muito baixas e distâncias ao centro de gravidade da nuvem $N_{(J)}$ muito elevadas: estes perfis dão a indicação que as características acima podem apresentar um comportamento atípico.

**Tabela 11: Modalidades que têm uma CTA superior à média para a formação do primeiro eixo principal de inércia**

| Modalidade | Descrição | Sinal |
|---|---|---|
| X20_5 | Idade – > 56 anos | + |
| X40_1 | Habilitações literárias – Curso médio e superior | + |
| X80_5 | Antiguidade no emprego – > 15 anos | + |
| X100_1 | Tipo de Habitação – Própria | + |
| X120_2 | *Pay-off*[3] – 100% | + |
| X400_3 | Situação profissional – Quadro superior (efectivo e a prazo), Quadro médio/técnico (efectivo e a prazo) e Reformado. | + |
| X10_1 | Estado civil – Solteiro | – |
| X10_3 | Estado civil – Divorciado e Separado judicialmente | – |
| X20_1 | Idade – ≤ 25 anos | – |
| X300_1 | Região – Ribatejo, Grande Porto e Madeira | – |
| X40_2 | Habilitações literárias – Preparatório/ Geral e Complementar | – |
| X50_1 | Sem automóvel | – |
| X80_2 | Antiguidade no emprego – 1 a 3 anos | – |
| X90_1 | Agregado familiar – 1 pessoa | – |
| X100_2 | Tipo de Habitação – Arrendada | – |
| X100_3 | Tipo de Habitação – Família e/ou Outro tipo | – |
| X110_3 | Sem regime de casamento (não se aplica) | – |
| X120_1 | *Pay-off* – 10%, 25% e 50% | – |
| X400_1 | Situação profissional – Outros profissionais liberais, Trabalhador não especializado (efectivo) e Comissionista (efectivo e a prazo ) | – |

A análise das tabelas 11 e 12 permite observar que 19 modalidades têm uma CTA superior à média para a formação do primeiro eixo e 16 têm uma CTA superior à média para a formação do segundo eixo. É de realçar que 12 daquelas modalidades têm uma CTA superior à média em ambos os eixos. Contudo, em geral, a CTR do segundo eixo para cada uma das modalidades é muito baixa. Para além disso, as modalidades da nuvem $N_{(J)}$ que têm um maior contributo para a inércia total da nuvem são algumas das referidas acima. Por ordem decrescente de contribuição, podem salientar-se o *pay-off* - 10%, 25% e 50%; a situação profissional – quadro superior (efectivo e a prazo), quadro médio/técnico (efectivo e a prazo) e reformado; as habilitações literárias – curso médio e superior; o tipo de habitação – arrendada; o *pay-off* – 100%; a idade máxima de 25 anos, etc. Em termos do plano principal 1-2, verifica-se que a contribuição relativa deste plano para cada modalidade é bastante elevada, ou dito de outra forma, a comunali-dade associada a cada modalidade está bastante pró-xima de 1. Pode, portanto, afirmar-se que a qualidade da representação da projecção destas modalidades, no referido plano, é muito boa.

**Tabela 12: Modalidades que têm uma CTA superior à média para a formação do segundo eixo principal de inércia**

| Modalidade | Descrição | Sinal |
|---|---|---|
| X20_2 | Idade – > 25 anos e ≤ 42 anos | + |
| X40_2 | Habilitações literárias – Preparatório/ Geral e Complementar | + |
| X40_3 | Habilitações literárias – Não tem, Básico e Desconhecido | + |
| X80_4 | Antiguidade no emprego – 7 a 15 anos | + |
| X120_2 | *Pay-off* – 100% | + |
| X400_2 | Situação profissional – Profissional liberal licenciado, Agricultor/pescador, Comerciante, Outros empresários, Trabalhador qualificado/administrativo (efectivo e a prazo) e Trabalhador não especializado (a prazo) | + |
| X10_3 | Estado civil – Divorciado e Separado judicialmente | – |
| X20_1 | Idade – ≤ 25 anos | – |
| X20_5 | Idade – > 56 anos | – |
| X40_1 | Habilitações literárias – Curso médio e superior | – |
| X50_1 | Sem automóvel | – |
| X80_2 | Antiguidade no emprego – 1 a 3 anos | – |
| X100_2 | Tipo de Habitação – Arrendada | – |
| X110_3 | Sem regime de casamento (não se aplica) | – |
| X120_1 | *Pay-off* – 10%, 25% e 50% | – |
| X400_3 | Situação profissional – Quadro superior (efectivo e a prazo), Quadro médio/técnico (efectivo e a prazo) e Reformado. | – |

### 6.1.4. Representação Simultânea

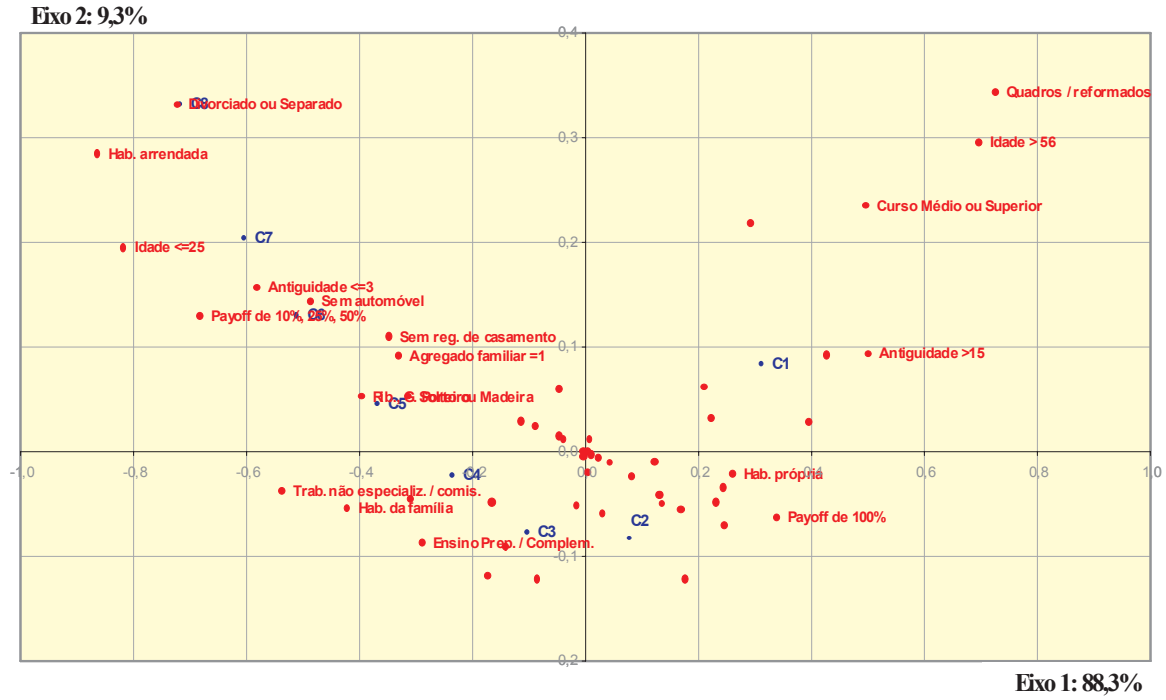
A representação das nuvens dos pontos linha (classes de risco) e dos pontos coluna (características dos clientes) no plano de projecção formado pelos dois primeiros eixos, permite observar um efeito Guttman, ou seja, as nuvens de pontos têm um efeito parabólico, que surge geralmente quando variáveis contínuas são transformadas em variá-veis nominais. Na situação da existência de um efeito Guttman (onde o 2.º eixo é uma função do 2.º grau do 1.º eixo), a mensagem transmitida pelo segundo eixo, traduz o mesmo fenómeno que o primeiro eixo. Contudo, a interpretação do primeiro eixo pode ser melhorada ou afinada com a interpretação do 2.º eixo, que geralmente representa os pontos linha e os pontos coluna que estão próximos da origem sobre o 1.º eixo.

Com base na representação simultânea dos pontos linha e dos pontos coluna, e nas ajudas à interpretação apresentadas nas secções anteriores, pode-se finalmente identificar as características dos clientes associadas ao risco de crédito. Os indivíduos da classe de risco mais baixa, a classe 1, têm associadas as modalidades de idade superior a 56 anos, curso médio e/ou superior, têm uma antiguidade no emprego superior a 15 anos, habitação própria,

escolheram um *pay-off* a 100% e detêm uma das seguintes situações profissionais: quadro superior (efectivo e a prazo), quadro médio/técnico (efectivo e a prazo) ou reformado. Em oposição estão os indivíduos das classes de risco 5 e 6 que têm associadas as modalidades de solteiro, divorciado e separado judicialmente e nas quais o regime de casamento não se aplica, idade até 25 anos, habitam na região do Ribatejo, Grande Porto e Madeira, têm como habilitações literárias o curso preparatório / geral ou complementar, não têm automóvel, a sua antiguidade no emprego varia entre 1 e 3 anos, o seu agregado familiar é composto por uma única pessoa, habitam em casa arrendada, de família ou outro tipo, escolheram um *pay-off* a 10%, 25% ou 50%, e a sua situação profissional é uma das seguintes: outros profissionais liberais, trabalhador não especializado (efectivo) e comissionista (efectivos e a prazo).

As oposições entre a classe de risco 1 e as classes de risco 5 e 6 também podem ser observadas no quadro de perfil de linha, $X_I$: por um lado, as modalidades associadas à classe 1 apresentam uma proporção maior de indivíduos nesta classe e menor nas classes 5 e 6, enquanto por outro lado, as modalidades associadas às classes 5 e 6 apresentam maiores proporções de indivíduos nestas classes e menores proporções na classe 1.

**Gráfico 2: Representação das modalidades da nuvem $N_{(I)}$ – a azul, e da nuvem $N_{(J)}$ – a vermelho, no plano formado pelos dois primeiros eixos.**




## 6.2. ANÁLISE COM ELEMENTOS SUPLEMENTARES

Adicionalmente à análise efectuada anteriormente, fez-se uma análise considerando elementos (indivíduos e/ou variáveis) suplementares, ou seja, elementos que não participam na AFC. Uma vez que os elementos suplementares não intervêm na formação dos eixos factoriais, a análise com este tipo de elementos permite que se faça o posicionamento dos elementos considerados suplementares sobre a nuvem dos indivíduos, $N_{(I)}$, ou das modalidades, $N_{(J)}$, através do cálculo *a posteriori* das suas coordenadas sobre os eixos factoriais. A análise que se realizou com os elementos suplementares, teve como objectivo a verificação se alguns dos elementos considerados atípicos (características dos clientes) incluídos anteriormente na análise de correspondências, influenciam a interpretação dos eixos principais de inércia, e se é diferente o seu posicionamento nas nuvens projectadas nos eixos factoriais quando esses elementos são considerados suplementares. Tal como já foi observado anteriormente, as modalidades divorciado e separado judicialmente, idade menor ou igual a 25 anos, idade superior 56 anos e habitação arrendada parecem indicar um perfil atípico. Será que a passagem destes elementos de activos para suplementares, vem alterar a interpretação dos dois primeiros

eixos principais de inércia? A resposta é negativa. Todas as oposições anteriormente observadas, sem excepção, entre classes de risco e entre modalidades nos dois primeiros eixos continuam a existir. A partir da realização da AFC sobre a tabela de contingência reduzida de quatro colunas correspondentes às quatro modalidades referidas acima, obtêm-se os seguintes resultados sobre os valores próprios, percentagem de inércia explicada e valores do qui-quadrado de contingência associados a cada um dos eixos.

**Tabela 13: Valores próprios, percentagem de inércia e valores do qui-quadrado**

| Eixo | Valor Próprio | Qui-Quadrado[a] | Percentagem de Inércia | |
|---|---|---|---|---|
| | | | Simples | Acumulada |
| 1 | 0,05919 | 24749,21 | **89,2%** | **89,2%** |
| 2 | 0,00602 | 2517,78 | **9,1%** | **98,2%** |
| 3 | 0,00059 | 248,00 | 0,9% | 99,1% |
| 4 | 0,00033 | 138,93 | 0,5% | 99,6% |
| 5 | 0,00013 | 54,42 | 0,2% | 99,8% |
| 6 | 0,00006 | 26,76 | 0,1% | 99,9% |
| 7 | 0,00005 | 22,71 | 0,1% | 100,0% |
| Total | **0,06638** | 27757,80 | 100,0% | - |

a. 343 graus de liberdade

Nesta situação, observa-se que os dois primeiros eixos explicam 98,2% da inércia total (mais 0,6% do que no caso em que todos as modalidades são consideradas activas), sendo que o primeiro eixo explica 89,2% e o segundo 9,1% da inércia. Considera-se mais uma vez que os dois primeiros eixos são suficientes para explicar a informação contida na tabela de contingência. A hipótese da independência das linhas e colunas da tabela de contingência continua a ser rejeitada. Na tabela 14 encontram-se as coordenadas, CTA e CTR da nuvem $N_{(I)}$ associadas aos dois primeiros eixos principais de inércia.

**Tabela 14: Frequências relativas, distância ao centro de gravidade, coordenadas, inércia, CTA e CTR sobre os dois primeiros eixos factoriais**

| Classe de Risco | Frequência Relativa | Distância | Coordenadas | | Inércia | CTA | | CTR | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | | 1 | 2 | 1 | 2 | Total |
| 1 | 29,3% | 0,104 | **0,291** | **0,074** | 0,026 | **0,419** | **0,267** | **0,939** | 0,061 | 1,000 |
| 2 | 26,8% | 0,014 | 0,066 | **-0,080** | 0,003 | 0,020 | **0,285** | 0,376 | 0,551 | 0,927 |
| 3 | 18,7% | 0,018 | -0,102 | **-0,067** | 0,003 | 0,033 | **0,140** | 0,649 | 0,282 | 0,931 |
| 4 | 11,4% | 0,057 | -0,221 | -0,016 | 0,006 | 0,094 | 0,005 | 0,980 | 0,005 | 0,985 |
| 5 | 7,1% | 0,142 | **-0,345** | 0,052 | 0,009 | **0,142** | 0,032 | **0,955** | 0,022 | 0,976 |
| 6 | 4,4% | 0,281 | **-0,473** | **0,135** | 0,011 | **0,165** | **0,131** | **0,919** | 0,074 | 0,993 |
| 7 | 2,0% | 0,419 | -0,537 | 0,168 | 0,006 | 0,095 | 0,092 | 0,885 | 0,087 | 0,972 |
| 8 | 0,5% | 0,729 | -0,612 | 0,242 | 0,002 | 0,032 | 0,049 | 0,799 | 0,125 | 0,924 |
| Total | 100,0% | - | - | - | **0,066** | - | - | - | - | - |

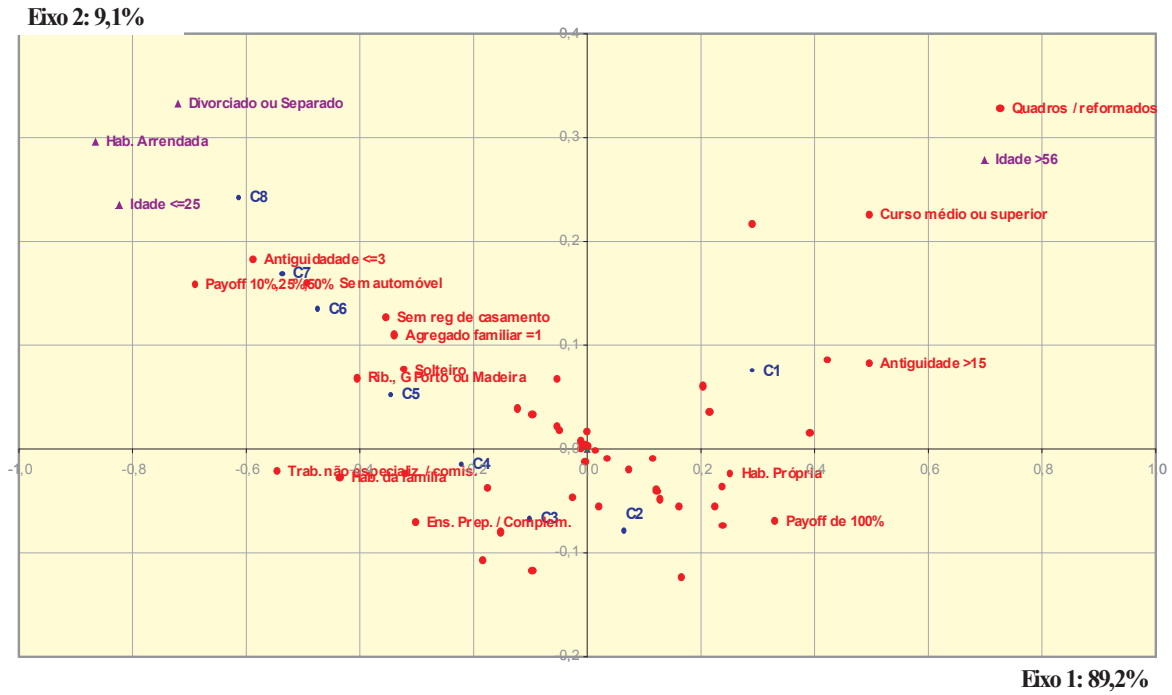
Uma análise minuciosa da tabela anterior permite concluir que os dois primeiros eixos continuam a transmitir-nos a mesma "mensagem" que nos transmitiam quando todas as classes de risco contribuíam para a formação dos eixos, ou seja:

- Existe uma oposição entre a classe de risco 1 e as classes de risco 5 e 6, tendo todas elas uma CTA superior à média para a formação do primeiro eixo principal de inércia;
- As classes 1, 2, 3 e 6 têm uma CTA superior à média para a formação do segundo eixo principal de inércia, existindo uma oposição entre a classe 1 e 6 e as classes de risco 2 e 3;

Omite-se uma análise semelhante à anterior para as modalidades da nuvem $N_{(J)}$ pelo facto dos resul-

tados obtidos nesta análise serem semelhantes aos obtidos na Secção 7.1. Contudo, o posicionamento das classes de risco da nuvem $N_{(I)}$, assim como o posicionamento das modalidades activas e suplementares da nuvem $N_{(J)}$ no plano 1-2, pode ser apreciado no gráfico 3. Facilmente se observa que o posicionamento de todos estes elementos no plano formado pelos dois primeiros eixos, permite retirar as mesmas conclusões que foram retiradas a partir do gráfico 2, quando todos os elementos foram considerados activos na AFC. Perante estes resultados, pode concluir-se que as modalidades divorciado e separado judicialmente, idade menor ou igual a 25 anos, idade superior 56 anos e habitação arrendada não apresentam um perfil atípico, uma vez que mesmo considerando estas modalidades como elementos suplementares na AFC, se mantém a interpretação dos dois primeiros eixos e o seu posicionamento nas nuvens projectadas nos dois primeiros eixos factoriais.

**Gráfico 3: Representação das modalidades da nuvem $N_{(I)}$ – a azul, e da nuvem $N_{(J)}$ – a vermelho, no plano formado pelos dois primeiros eixos**




## 6.3. FORMAÇÃO DE CLASSES HOMOGÉNEAS DE MODALIDADES

Para além da identificação das características dos clientes associadas a classes de risco baixo e a classes de risco elevado, pode ainda proceder-se à formação de classes homogéneas de modalidades (características dos clientes), com base nas componentes principais da nuvem $N_{(J)}$ obtidas na AFC. A análise conjunta dos resultados obtidos através destas duas abordagens permite a comparação das classes de modalidades, obtidas através da classificação, com os conjuntos de modalidades associadas a classes de risco baixo e a classes de risco elevado, deduzidas a partir da AFC. Os resultados da classificação hierárquica ascendente, segundo o critério de Ward, permitem retirar as seguintes conclusões:

- No processo hierárquico de agregação das modalidades, as associadas a classes de risco elevado agrupam-se todas na mesma classe;
- A identificação das características chave de risco pode ser obtida a partir da análise dos dois ou dos três primeiros eixos principais de inércia. Esta última conclusão é perfeitamente aceitável e expectável, não só porque os dois primeiros eixos restituem 97,6% da inércia total, mas também porque na presença de um efeito Guttman, a mensagem transmitida pelo segundo eixo e seguintes traduz o mesmo fenómeno que a mensagem transmitida pelo primeiro eixo.

Desta forma, pode-se afirmar que estas duas abordagens não só se complementam, mas também se validam.

## 7. CONCLUSÃO

A partir da interpretação dos eixos principais retidos após a aplicação da Análise Factorial de Correspondências à tabela de contingência que

cruza as oito classes de risco _ formadas pela aplicação do *K-means* sobre o vector de pontuação do modelo *logit* e certificadas pelos estimadores Kaplan-Meier _, com as cinquenta e quatro características dos clientes, foi possível identificar um retrato robô das características associadas às classes de risco baixo e risco elevado. Concluiu-se, assim, que as classes de risco baixo se caracterizam por conterem clientes com idade superior ou igual a 56 anos, habitação própria, antiguidade no emprego superior a 15 anos, habilitações literárias de nível médio ou superior, por serem quadros médios/técnicos, quadros superiores ou reformados e por terem um *pay-off* de 100%. As classes de risco elevado caracterizam-se por conterem clientes com idade inferior a 25 anos, antiguidade no emprego inferior ou igual a 3 anos, curso preparatório, geral ou complementar, serem solteiros, divorciados ou separados judicialmente com agregado familiar de uma pessoa, habitarem em habitação arrendada ou de família, residirem no Ribatejo, Grande Porto ou Madeira, não possuírem automóvel, serem trabalha-dores não especializados, comissionistas ou outros profissionais liberais e por terem um *pay-off* de 10%, 25% ou 50%.

## NOTAS FINAIS

1- Taxa cobrada por atraso no pagamento da prestação.
2- As variáveis contínuas que compõem o modelo econométrico foram sujeitas ao teste Box-Tidwell certificando-se que são lineares em *logit* não se subestimando, portanto, o erro tipo II.
3- Nos cartões de crédito diz respeito à percentagem de pagamento do montante em dívida acumulado escolhido pelo cliente para liquidação mensal. Normalmente são as seguintes opções: 10%, 25%, 50% e 100%. Só o *pay-off* de 100% se encontra isento de juros.

## BIBLIOGRAFIA


ALDRICH, J. H. e Nelson, F. D. (1984), *Linear Probability, Logit, and Probit Models*, Sage QASS Series, Iowa City.

AMEMIYA, T. (1981), "Qualitative response models: A survey", *Journal of Economic Literature*, 19, pp. 1483-1536.

BAILEY, M. (2001), *Credit Scoring – The Principles and Practicalities*, White Box, Southampton.

BERTIER, P. e Bouroche, J.-M. (1981), *Analyse des Données Multi-dimensionnelles*, PUF, Paris.

BOUROCHE, J. e Saporta, G. (2002), *L'Analyse des Données*, PUF, Que Sais-Je?, Paris.

CELEUX, G., Diday, E., Govaert, G., Lechevallier, Y., e Ralambondrainy, H. (1989), *Classification automatique des données - aspects statistiques and informatiques*. Dunod, Paris.

CRIVISQUI, E. M. (1993), *Analisis Factorial de Correspondencias – un instrumento de investigación en ciencias sociales*. Edición del Laboratorio de Informática Social – Universidad Catolica de Asunción, Asunción.

DEMARIS, A. (1992), *Logit Modeling-Practical Applications*, Sage QASS Series, Iowa City.

DIDAY, E., Lemaire, J., Pouget, J. e Testu, F. (1982), *Éléments d'analyse de données*, Dunod, Paris.

ESCOFIER, B. e Pagés, J. (1990), *Analyse Factorielles Simples et Multiples: Objectifs, Méthodes et Interprétation*, Dunod, Paris.

ESCOFIER. B. (2003), *Analyse des Correspondances: Recherches au coeur de l'analyse des données*, Presses Universitaires de Rennes, Rennes.

GOMES, P. J. (1993), *Análise de Dados*, Instituto Superior de Estatística e Gestão de Informação - Universidade Nova de Lisboa, Lisboa.

GREENE, W. H. (2000), *Econometric Analysis* (3ª ed.), Prentice-Hall International, Inc.

HOSMER, D. W. e Lemeshow, S. (1989), *Applied Logistic Regression*, John Wiley, New York.

JOHNSTON, J. e Dinardo, J. (1997), *Econometric Methods* (4ª ed.), McGraw-Hill International Editions, Inc.

KAPLAN, E. L. e Meier, P. (1958), "Non parametric estimation from incomplete observation". *Journal of the American Statistics Association*, 53, pp. 457-481.

LEBART, L., Piron, M. e Morineau, A. (2000), *Statistique exploratoire multidimensionnelle* (3ª ed.), Dunod, Paris.

LEE, E. T. (1992), *Statistical Methods for Survival Data Analysis* (2ª ed.), Wiley-Interscience.

LIAO, T. F. (1994), *Interpreting Probability Models – Logit*, Probit*, and Other Generalized Models*, Sage QASS Series, Iowa City.

LONG, J.S. (1997), *Regression Models for Categorical and Limited Dependent Variables*, Sage QASS Series, Thousand Oaks, California.

LORETTA, J. M. (1997), "What's the point of Credit Scoring", *Business Review*, Federal Reserve Bank of Philadelphia.

MAYS, E. (2001), *Handbook of Credit Scoring*, GPCo, Chicago.

MCNAB, H., e Wynn, A. (2001), *Principles and Practice of Consumer Credit Risk Management*, Financial World Publishing, Kent.

MENARD, S. (2001), *Applied Logistic Regression Analysis* (2ª ed.), Sage QASS Series, Thousand Oaks.

MICKEY, J. e Greenland, S. (1989), "A study of the impact of confounder-selection criteria on effect estimation", *American Journal of Epidemiology*, 129, pp. 125-137.

PIGEON, J. G. e Heyse, J. F. (1999), "A cautionary note about assessing the fit of logistic regression models", *Journal of Applied Statistics*, 26, (7), pp. 847-853.

POWERS, D. A. e Xie, Y. (2000), *Statistical Methods for Categorical Data Analysis*, Academic Press.

SAPORTA, G. (1990), *Probabilités. Analyse des Données et Statistique*, Éditions Technip, Paris.

SPSS (1991), *Statistical Algorithms* (2ª ed.), SPSS Inc.

TABACHNICK, B. G. e Fidell, L. S. (2001), *Using Multivariate Statistics* (4ª ed.), Allyn and Bacon.

THOMAS, L. C., Edelman, D. B. e Crook, J. N. (2002), *Credit Scoring and its Applications*, Siam, Philadelphia.