ESTUDO DE UMA CARTEIRA DE CRÉDITO AO CONSUMO

DE

UM BANCO DE CABO VERDE

por

José Moniz Lopes Fernandes

(Mestre)

Dissertação apresentada como requisito parcial para obtenção do grau de Doutor em Estatística e Gestão de Informação

pelo

Instituto Superior de Estatística e Gestão de Informação

da

Universidade Nova de Lisboa.

Lisboa

2012

José Moniz Lopes Fernandes

ESTUDO DE UMA CARTEIRA DE CRÉDITO AO CONSUMO DE UM BANCO DE CABO VERDE

Dissertação orientada por

Doutor Manuel Leote Tavares Inglês Esquível

Doutora Gracinda Rita Diogo Guerreiro

Doutora Maria de Rosário Fraga de Oliveira Martins

Doutora Patrícia Xufre Gonçalves da Silva Casqueiro.

Dissertação apresentada para obtenção do grau de Doutor em Estatística e Gestão de Informação na especialidade de Actuariado e Gestão de Risco, pela Universidade Nova de Lisboa, Instituto Superior de Estatística e Gestão de Informação.

Lisboa
2012

Aos meus filhos

Isaías e Larissa

Agradecimentos

Gostaria de manifestar os meus sinceros agradecimentos à Professora Doutora Maria de Rosário Martins, ao Professor Doutor Manuel L. Esquível, à Professora Doutora Gracinda R. Guerreiro e à Professora Doutora Patrícia Xufre, orientadores da dissertação, pelos conhecimentos científicos que me transmitiram, pelo apoio, pela paciência e total disponibilidade demonstrados ao longo deste trabalho.

Um agradecimento especial ao Professor Manuel L. Esquível e à Professora Gracinda R. Guerreiro, pelo acolhimento, pela possibilidade da utilização dos recursos internos do CMA e pela inscrição como membro colaborador do CMA.

Ao Instituto Português de Apoio ao Desenvolvimento pela concessão da bolsa de estudos e à Universidade de Cabo Verde pela dispensa de serviço docente, sem o apoio dos quais, este projecto, não teria sido possível.

Gostaria de agradecer à Caixa Económica de Cabo Verde, em especial, ao seu Presidente do Conselho de Administração, Economista Emanuel Miranda, por permitir o acesso à informação necessária à realização deste trabalho, e ao Sr. João Monteiro pela sua prontidão e disponibilidade no esclarecimento das dúvidas surgidas ao longo deste trabalho.

Agradeço ainda ao Danilson Semedo pela forma fascinante que sempre demostrou em torno deste tema e pelas secções de discussões da aplicação do software SAS.

Agradeço ainda aos colegas do CMA que me aconpanharam ao longo deste período.

À minha querida esposa Ângela e aos meus filhos, pelo companheirismo, pela paciência, pelo apoio, pelo amor e carinho que sempre demostraram ao longo deste percurso.

Aos meus Pais, pelo apoio, coragem e confiança que sempre depositaram em mim. Obrigado por tudo.

Aos meus irmãos, pelo incentivo e pela amizade que sempre demostraram.

Por último, os meus sinceros agradecimentos a todos os que, de uma forma ou de outra, colaboraram para que este trabalho se tornasse realidade.

Resumo

O objectivo central desta dissertação consiste na análise de uma carteira de crédito ao consumo de um banco de Cabo Verde, onde muito trabalho existe para desenvolver nestas áreas.

Com base num modelo de Regressão Logística, e recorrendo a variáveis socio-económicas e financeiras de cada cliente, estimou-se a probabilidade de incumprimento a *priori* para cada cliente. Esta estimação auxiliará na decisão de concessão de crédito e, caso o crédito seja concedido, constituirá uma ferramenta importante na estimação do *spread* a aplicar ao cliente.

Na estimação do *spread* é ainda necessário ter em conta a taxa de recuperação do crédito, para clientes em incumprimento. Nessa perspectiva, apresentamos uma solução para estimação da taxa de recuperação da carteira.

Neste trabalho propomos uma fórmula simples para estimação do *spread* a aplicar a um novo cliente, uma vez observadas as suas características.

Considerando que a probabilidade de incumprimento não é constante ao longo do tempo, analisámos o incumprimento da carteira utilizando, para tal, um modelo de Markov para populações abertas: o modelo Vórtices Estocásticos.

No que se refere ao modelo Vórtices Estocásticos, em termos teóricos generalizamos a forma funcional que modela os fluxos de entrada na população proposta nos estudos de Guerreiro et al. Desenvolvemos os resultados referentes à inferência estatística para fluxos de entrada com a nova forma funcional aqui proposta, o que permitiu obter desenvolvimentos relativos à estimação das intensidades de entrada de novos elementos para a população, bem como à análise da estrutura da mesma num qualquer período de tempo, inclusivamente numa perspectiva de longo prazo. Os resultados obtidos permitem-nos estimar a proporção de clientes nas várias classes de risco, através de estimativas pontuais e intervalos de confiança. Estes resultados serão úteis numa perspectiva de gestão de risco da carteira.

PALAVRAS CHAVE: Classes de Risco, Cadeias de Markov, Vórtices Estocásticos, Regressão Logística, Probabilidade de Incumprimento, *Spread*, Taxa de Recuperação.

Abstract

The purpose of this dissertation is the analysis of a portfolio of consumer credit from a bank of Cape Verde, where there is much work to develop in these areas.

Based on a logistic regression model, using socio-economic and financial variables responsibilities of each client, we estimated the default probability for each client. This estimation will assist in the decision to grant credit and if credit is granted, will constitute an important tool for estimating the spread to apply the customer.

In the estimation of the spread is still necessary to take into account the recovery rate of credit to customers in default. From this perspective, we present a solution to estimate the recovery rate of the portfolio.

In this work we propose a simple formula to estimate the spread, applicable to a new customer, since their observed characteristics.

Considering that the default probability is not constant over time, we analyzed the default probability of the portfolio using, for this, a Markov model to open populations: the Stochastic Vortices Model.

With regard to the Stochastic Vortices model, in theoretical terms, in this work it was generalized the functional form that models the flow into the population studies, proposed in Guerreiro et al. We develop the results of statistical inference for inflows to the new functional form proposed in this work, which have enabled developments relating to the estimation of the intensities of the entry of new elements to the population as well as the structural analysis of the same in any period time, even in the long term. The obtained results allow us to estimate the proportion of clients in the various risk classes through point estimates and confidence intervals.

KEYWORDS: Risk Classes, Markov Model, Stochastic Vortex, Logistic Regression, Probability of Default, Spread, Recovery Rate.

Índice

Lista de Tabelas

Lista de Figuras

Introdução

Motivação

Devido à regulação de Basileia II, e à recente crise financeira, a análise do risco de crédito tem merecido uma atenção especial por parte das instituições financeiras. Mais importante que isso, são factores como a competitividade dos mercados, a evolução dos sistemas e gestão de informação, a diminuição do pedido de crédito e o aumento da probabilidade de incumprimento condicionam as intuições financeiras à mitigação de melhores técnicas para a análise e gestão do risco de crédito.

Destacamos alguns dos estudos desenvolvidos para os países "considerados" em desenvolvimento, dos quais Cabo Verde faz parte: são eles os estudos de [Abdou et al., 2008] para o Egipto, [Dinh e Kleimeier, 2007], para o Vietname, [Schreiner, 2004] para a Bolívia e [Viganó, 1993] para o Burkina Faso. Particularmente, para o caso de Cabo Verde, o estudo de [Semedo, 2010] consistiu na comparação de Redes Neuronais e Regressão Logística aplicado a uma base de dados de crédito ao consumo de uma instituição financeira Caboverdiana. Isto mostra o quão é importante desenvolver estudos para estes países e para Cabo Verde em particular.

O risco que cada cliente representa para a instituição bancária deve ser avaliado e estimado *a priori*, aquando da concessão do crédito, usando técnicas adequadas. Pelas mais diversas razões, as condições socio-económicas do cliente, analisadas aquando da concessão do crédito, podem alterar-se durante o período do empréstimo, condicionando o cumprimento do pagamento das prestações, representando um risco acrescido ao estimado inicialmente. Assim, o risco que o cliente representa para a instituição bancária não é constante ao longo do tempo.

Além das razões acima referidas, esta dissertação também é motivada por problemas relacionados com a modelização de risco de crédito num contexto de crise que contemple a eventualidade de taxas de recuperação, r, quase nulas e altas probabilidades de incumprimento, p. Assim sendo, o resultado que permite estimar o *spread*, s, como aproximadamente igual a $(1-r)p$, válido em contexto normal (baixas probabilidades de incumprimento e altas taxas de recuperação), não se deve aceitar, devendo por isso procurar-se alternativas mais

exactas.

Esta dissertação enquadra-se no caso em que as instituições bancárias classificam os seus clientes segundo as classes de risco em função de incumprimento ou cumprimento no pagamento das suas prestações. Os clientes da carteira são periodicamente reclassificados e considera-se as entradas e saídas dos clientes observadas mensalmente.

Objectivos

Nesta dissertação estabelecemos os seguintes objectivos:

- estimar a probabilidade de incumprimento da carteira e do cliente através da Regressão Logística, em função das características socio-demográficas e comportamentais de cada cliente;
- apresentar uma generalização da forma funcional que modela os fluxos de entrada nas populações abertas sujeitas a reclassificações períodicas, proposta nos estudos de Guerreiro et al;
- estimar a evolução temporal da probabilidade de incumprimento da carteira de crédito ao consumo, numa perspectiva de longo prazo, através de um modelo para populações abertas sujeitas a reclassificações períodicas designado por Vórtices Estocásticos;
- propor um modelo de *spread* em função da probabilidade de incumprimento e da taxa de recuperação;
- estimar o *spread* da carteira e o *spread* a aplicar a um determinado cliente da carteira.

Metodologia

Agrupamos as metodologias utilizadas nesta dissertação de acordo com os capítulos desenvolvidos. Assim, no primeiro capítulo realiza-se uma pesquisa bibliográfica nas bases de dados, da web, das principais monografias dedicadas ao tema credit scoring (período de análise: 1992 a Março de 2010) dos artigos científicos relacionados com o tema (período de análise: até Março de 2010).

Nos restantes capítulos, baseados nos modelos que iremos propor e aplicar, nesta dissertação, estruturamos as metodologias da seguinte forma:

A Utilização da Regressão Logística

A Regressão Logística é muito utilizada na estimação da probabilidade de incumprimento e dos determinantes das mesmas. Nesta dissertação utilizamo-la para esse fim, admitindo que:

- a carteira de crédito é constituída pelas variáveis comportamentais, de modo a permitir a definição da variável resposta(cliente cumpridor ou incumpridor). Consideramos

ainda as variáveis socio-demográficas e financeiras como variáveis explicativas. Estas foram categorizadas através da abordagem "Weight of Evidence" (WOE). A capacidade preditiva das variáveis categorizadas foi calculada pelo Infomation Value (IV);

- o modelo foi validado através da curva ROC, estatística de Kolmogorov Smirnov, índice de Gini, matriz de classificação e *out-of-sample*. Utilizamos ainda um modelo de score-card para validar com o modelo desenvolvido através da Regressão Logística.

Os Vórtices Estocáticos

A metodologia dos Vórtices Estocásticos, aplicada e desenvolvida nesta dissertação, baseia-se nos estudos de [Guerreiro, 2008], [Guerreiro e Mexia, 2008] e [Guerreiro et al., 2010, 2012b]. Consideramos populações abertas, divididas num número finito de sub-populações. Os elementos da população são inicialmente colocados numa sub-população e periodicamente reclassificados, podendo, em cada reclassificação, ser colocados em qualquer das sub-populações existentes.

Como prolongamento dos estudos de Guerreiro et al, introduzimos novos resultados ao nível da generalização da convergência do modelo dos vórtices estocásticos e introduzimos também um estudo estatístico da modelação dos fluxos de entrada, utilizando uma forma sigmoidal $\lambda_i = (a + be^{-\theta i})^{-1}$. Esta modelação revelou-se ser a mais adequada aos dados da carteira em estudo.

A opção por estes modelos deve-se ao facto destes terem sido utilizados, com sucesso, nos estudos de Guerreiro et al, bem como no estudo de uma carteira de seguro automóvel de uma seguradora Cabo-verdiana.

A Metodologia Actuarial do Spread

Nesta dissertação, no que se refere à estimação do *spread* seguimos [McNeil et al., 2005], que fundamenta que existem duas metodologias de apreçamento de risco de crédito; as actuariais e financeiras. Assim, através da metodologia actuarial propomos dois métodos para a estimação do *spread* da carteira de crédito, baseados no:

- modelo a um período para um modelo de obrigações cupão zero;
- modelo de uma carteira a tempo discreto.

A determinação do *spread* para ambas as metodologias será em função da probabilidade de incumprimento e da taxa de recuperação. Dadas as característica da carteira de crédito em estudo, focalizamos a nossa metodologia no estudo do modelo de uma carteira a tempo discreto.

Softwares utilizados

Nesta dissertação utilizamos os seguintes softwares:

- Enterprize Miner Client 6.1 do *SAS*® para estimação da probabilidade de incumprimento;
- Wolfram Mathematica 7 para estimação da evolução temporal da probabilidade de incumprimento;
- SAS Enterprise Guide 4.3 do *SAS*® para estimação da taxa da recuperação.

Estrutura

No primeiro capítulo, apresentamos uma referência bibliográfica dos modelos de credit scoring focalizando na discussão e comparação dos estudos empíricos das técnicas baseadas em estatísticas tradicionais, metaheurísticas e modelos e conjuntos híbridos.

No capítulo 2 estima-se um modelo de probabilidade de incumprimento para os clientes de crédito ao consumo, através de uma técnica estatística multivariada, designadamente a Regressão Logística. O modelo da probabilidade de incumprimento estimado agrega as variáveis como entidade patronal, taxa nominal, actividade profissional, valor do empréstimo, valor da prestação, número de prestações pagas e a agência a que o cliente pertence. O modelo ajustado identifica as seguintes variáveis como sendo significativas: Número de prestações pagas, actividade profissional, entidade patronal, agência, taxa nominal, valor da prestação, idade, sexo, prazo, habilitações literárias e valor de empréstimo.

No terceiro capítulo, como prolongamento dos estudos realizados em Guerreiro e Guerreiro et al, introduzimos uma nova modelação dos fluxos de entrada e generalizamos a forma funcional que modela os fluxos de entradas na população. Esta generalização contempla o caso particular utilizado nesta dissertação e permite o ajustamento de muitas outras modelações dos fluxos de entrada. Obtemos também os estimadores de máxima verosimilhança para os parâmetros dos fluxos de entrada e de novas regiões de confiança para as dimensões absolutas e relativas das sub-populações, para o fluxo de entradas ajustado à carteira em estudo. Por fim, com base na informação da carteira de clientes de crédito ao consumo, estimamos as probabilidades de incumprimento da carteira numa perspectiva temporal de longo prazo, através do modelo de Markov para populações abertas.

No quarto capítulo, em prolongamento do estudo de [Vale, 2010], propomos dois modelos que determinam o *spread* adequado para a carteira e a cada cliente, em função da probabilidade de incumprimento e da taxa de recuperação. Propomos também um modelo que permite estimar a taxa de recuperação da carteira. A probabilidade de incumprimento utilizada para a determinação do *spread* da carteira é estimada no Capítulo 3 e para o *spread* mínimo de cada cliente é estimada no Capítulo 2.

No capítulo final serão apresentadas as conclusões, as limitações e os estudos futuros.

Contribuição

Nesta dissertação apresentamos duas contribuições sob os pontos de vista de desenvolvimentos teóricos e das aplicações ao risco do crédito.

Na perspectiva dos desenvolvimentos teóricos, em prolongamento dos estudos de Guerreiro e Guerreiro et al, esta dissertação contribui com a generalização da forma funcional que modela os fluxos de entrada numa população aberta e sujeitas às reclassificações periódicas. Obtivemos um resultado geral que caracteriza as condições sob as quais se garante a existência de estabilidade a longo prazo nas dimensões relativas das sub-populações, e consequente existência de um vórtice estocástico nessas sub-populações. A introdução de uma nova modelação dos fluxos de entrada introduz a necessidade de obtenção de estimadores de máxima verosimilhança para os parâmetros dos fluxos de entrada e de novas regiões de confiança para as dimensões absolutas e relativas das sub-populações.

Introduzimos também uma fórmula que permite estimar o *spread* em função da probabilidade incumprimento e da taxa de recuperação. Esta última contribuição é uma extensão natural do modelo a um período para obrigações de cupão zero.

Na perspectiva do risco de crédito, contribuímos com uma análise detalhada de uma carteira de crédito ao consumo de um banco de Cabo Verde, avaliando questões que são importantes e poucos exploradas na realidade da instituição, nomeadamente:

- cálculo da probabilidade de incumprimento em função do número de dias em atraso para a carteira;
- estimação da probabilidade de incumprimento através da Regressão Logística;
- estimação da probabilidade de incumprimento para a carteira de crédito ao consumo ao longo do tempo;
- estimação da taxa de recuperação do valor pago em função do valor de empréstimo e respectivo *spread* para a carteira;
- estimação do *spread* mínimo para cada cliente através da taxa de recuperação da carteira e da probabilidade de incumprimento.

Capítulo 1

Revisão Bibliográfica sobre Credit Scoring

1.1 Introdução

Os modelos de Credit Scoring são baseados em técnicas estatísticas que atribuem uma pontuação a cada pedido de atribuição de crédito. Esses modelos visam a identificação de características que permitam distinguir os pontencialmente bons dos maus créditos [Lewis, 1992].

Os primeiros indícios da existência da prática do crédito ao consumo datam de há 4000 anos, mas há evidências arqueológicas que apontam o seu início na época da civilização Suméria (por exemplo, existe o registo numa tábua suméria de argila do empréstimo de dinheiro a dois agricultores para compra de grãos, com a promessa de o devolverem na época da colheita). No entanto, somente nos últimos 50 anos, com o aparecimento dos cartões de crédito (emitidos pela primeira vez nos EUA em 1958 e em 1966 no Reino Unido), e com o crescimento da aquisição de casa própria e de empréstimos hipotecários, é que o crédito ao consumo se tornou tão generalizado.

Neste capítulo pretende-se pesquisar e analisar as principais monografias dedicadas ao tema credit scoring (periódo de análise: 1992 a Março de 2010) e artigos científicos relacionados com o tema (período de análise: até Março de 2010).

Neste capítulo, faz-se uma referência à definição de credit scoring de [Thomas et al., 2002], [Lewis, 1992] e [Mester, 1997] e descreve-se uma definição baseada em [Thomas, 2010], na secção 1.2. Na secção 1.3 enfatiza-se os tipos de credit scoring. Na secção 1.4 apresentam-se as principais limitações de credit scoring. Na secção 1.5 faz-se a pesquisa e a análise da literatura sobre credit scoring. As técnicas de credit scoring são apresentadas na secção 1.6, focalizando na discussão das várias técnicas de credit scoring. Apesar de actualmente várias

serem as técnicas discutidas e analisadas na literatura, iremos apenas descrever algumas das baseadas em estatísticas tradicionais, metaheurísticas e modelos e em conjuntos híbridos. Por fim, nas secções 1.7 e 1.8 faz-se a comparação e a discussão dessas técnicas, baseada nos estudos empíricos que se analisaram na secção 1.6.

1.2 Definição de Credit Scoring

Através da análise dos diferentes estudos que têm sido dedicados ao tema é possível encontrar diferentes definições de credit scoring. Neste trabalho, faz-se referência à definição dos seguintes autores:

- "Credit scoring é o conjunto de modelos de decisão e técnicas subjacentes que ajudam os credores na concessão de crédito ao consumo" [Thomas et al., 2002].
- "Credit scoring é um método estatístico (ou quantitativo) usado para prever a probabilidade de um cliente entrar em situação de incumprimento" [Mester, 1997].
- "Credit scoring é um processo em que a informação sobre o solicitante é convertida em números que, de forma combinada formam, um score" [Lewis, 1992].

Neste capítulo, considera-se a definição formal baseada em [Thomas, 2010]. A razão desta escolha prende-se com os desafios no contexto actual de crise e o desenvolvimento que é imposto pelos modelos de credit scoring.

Assume-se que cada cliente, seja ele o candidato, no caso de um score de aplicação ou um cliente actual, no caso de um scoring comportamental, pode ser descrito por um vector de características $\mathbf{x} = (x_1, \ldots, x_m), \mathbf{x} \in X$, onde X é o conjunto de todas as combinações possíveis das características do cliente. Estas características incluem variáveis sócio-económicas como idade, valor de empréstimo, taxa de juro, etc. e, no caso dos scores comportamentais, dados de desempenho como, por exemplo, o número de pagamentos em atraso nos últimos 12 meses. Se o que se pretende avaliar é o risco, diz-se que todos os clientes com mais de 90 dias em atraso no pagamento das suas prestações, nos primeiros 12 meses, são considerados "incumpridores" e os restantes "cumpridores". O score, $s(\mathbf{x})$, é então uma função das características $\mathbf{x}$ de um potencial cliente, que pode ser traduzido na estimativa da probabilidade do cliente ser "cumpridor".

O pressuposto fundamental em credit scoring consiste na previsão da probabilidade de incumprimento. É semelhante a uma estatística suficiente. Normalmente, o score assume uma relação monótona crescente com a probabilidade de ser "bom" cliente; daí que, se um cliente tem um score maior do que um outro, consequentemente, terá uma maior probabilidade ser melhor do que o segundo.

O score adequado, ou suficiente, $s(\mathbf{x})$ capta o máximo de informação necessária para prever a probabilidade de um Bom (B) e Mau (M) desempenho. Assim, quer a decisão seja tomada com base no vector inicial de dados original quer no score $s(\mathbf{x})$, o resultado final é o mesmo:

$$\mathbb{P}\{B \mid \mathbf{x}\} = \mathbb{P}(B \mid s(\mathbf{x})) = \mathbb{P}(B \mid s(\mathbf{x}), \mathbf{x}) = \mathbb{P}(B \mid \mathbf{x}), \forall \mathbf{x} \in X \tag{1.1}$$

em que $\mathbb{P}(B \mid s(\mathbf{x}))$ é a probabilidade de um cliente ser cumpridor, baseada no score $s(\mathbf{x})$. Quando adequado iremos deixar de escrever a dependência $\mathbf{x}$ do score:

$$\mathbb{P}(s) = \mathbb{P}(B \mid s(\mathbf{x}))$$

e

$$1 - \mathbb{P}(s) = 1 - \mathbb{P}(B \mid s(\mathbf{x}), \mathbf{x}) = \mathbb{P}(M \mid s(\mathbf{x})), \forall \mathbf{x} \in X$$

Uma forma de scoring é a probabilidade log score, onde:

$$s(\mathbf{x}) = ln \frac{\mathbb{P}(B \mid \mathbf{x})}{\mathbb{P}(M \mid \mathbf{x})} \tag{1.2}$$

$$com \quad \mathbb{P}(B \mid \mathbf{x}) + \mathbb{P}(M \mid \mathbf{x}) = 1, \quad \mathbf{x}\boldsymbol{I} \in X \tag{1.3}$$

Assim, o log score de chances tende para menos infinito, quando $\mathbb{P}(B|\mathbf{x}) = 0$, e para mais infinito, quando $\mathbb{P}(B|\mathbf{x}) = 1$. Log odds são produzidos quando se utiliza a Regressão Logística para determinar a classificação do scorecard[1], mas também podem ser obtidos a partir de outras abordagens. Especificar o score de um evento é equivalente a especificar a sua probabilidade; assim pode-se escrever a probabilidade em função do score da seguinte forma:

$$\mathbb{P}(B \mid \mathbf{x}) = \frac{e^{s(\mathbf{x})}}{1 + e^{s(\mathbf{x})}} = \frac{1}{1 + e^{-s(\mathbf{x})}} \tag{1.4}$$

Uma característica importante de log odds score consiste na separação completa entre a informação sobre a população, e a informação sobre o cliente individual que está a ser avaliado. Aplicando o Teorema de Bayes, no caso da probabilidade de um cliente ser classificado como bom ou mau, dado que possui características $\mathbf{x}$, com a proporção de clientes cumpridores e incumpridores na população dadas por $\mathbb{P}_B$ e $\mathbb{P}_M$, respectivamente, vem:

$$\mathbb{P}(B \mid \mathbf{x}) = \frac{\mathbb{P}(\mathbf{x} \mid B)\mathbb{P}_B}{\mathbb{P}(\mathbf{x})} \tag{1.5}$$

onde $\mathbb{P}(\mathbf{x})$ é a probabilidade do cliente ter características $\mathbf{x}$. Substituindo a expressão (1.5) na equação (1.2) vem:

[1] modelo de pontuação que permite classificar indivíduos. Para mais detalhes, ver [Siddiqi, 2006].

$$s(\mathbf{x}) = ln\frac{\mathbb{P}_B\mathbb{P}(\mathbf{x} \mid B)}{\mathbb{P}_M\mathbb{P}(\mathbf{x} \mid M)} = ln\frac{\mathbb{P}_B}{\mathbb{P}_M} + ln\frac{\mathbb{P}(\mathbf{x} \mid B)}{\mathbb{P}(\mathbf{x} \mid M)} = lnO_{pop} + lnI(\mathbf{x}) = S_{pop} + S_{Inf}(\mathbf{x}) \quad (1.6)$$

Assim, log odds score é a soma de dois termos, em que o primeiro depende apenas do odds da população ($S_{pop} = lnO_{pop}$) e o segundo depende da informação de um certo cliente. O primeiro termo de (1.6) é o score *a priori* - score de um indivíduo seleccionado aleatoriamente a partir da população; este score é aumentado ou reduzido pelo score que se baseia nos dados individuais de cada cliente. Mais detalhes sobre os conceitos básicos de credit scoring e sobre as diferentes abordagens para a construção de um scorecard, podem-se encontrar em referências tais como: [Thomas, 2009], [Anderson, 2007], [Thomas et al., 2005], [Mays, 2004] e [McNab e Wynn, 2000] e nos trabalhos de revisão: [Thomas, 2010], [Crook et al., 2007], [Thomas et al., 2005], [Thomas, 2000] e [Hand e Henley, 1997].

1.3 Tipos de Credit Scoring

De acordo com [Paleologo et al., 2010], os modelos de credit scoring podem ser classificados em quatro categorias :

- aplicação de scoring (application scoring): refere-se à avaliação da capacidade de crédito para novos clientes. Quantifica o risco associado ao crédito, avaliando os dados sociais, demográficos, financeiros e outros recolhidos no momento do pedido de crédito;
- scoring comportamental (behavioral scoring): este modelo incorpora variáveis que relacionam a história do cliente com a instituição. Este modelo tem como objetivo auxiliar o analista de crédito nas suas decisões sobre renovações de empréstimos de clientes, renegociações de dívidas, entre outros, ou seja, todas as decisões relativas à gestão do crédito de clientes que já possuem uma relação ou um histórico com a instituição.
- colecção de scoring (collection scoring): colecção de scoring é utilizado para dividir os clientes com diferentes níveis de incumprimento, separando aqueles que necessitam de acções mais decisivas, dos que não necessitam de uma intervenção imediata. Estes modelos são diferenciados de acordo com o seu grau de incumprimento (no início, no meio e na recuperação do atraso) e a utilização deste tipo de modelos de credit scoring permite uma melhor gestão de clientes incumpridores, logo desde os primeiros atrasos no pagamento das prestações (30-60 dias) para as fases subsequentes e emissão de dívidas;
- detecção de fraudes (fraud detection): modelos de scoring de fraudes classificam os candidatos de acordo com a probabilidade relativa de que um pedido possa ser fraudulento.

1.4 Limitações de Credit Scoring

É óbvio que a aplicação de modelos de credit scoring permite enormes benefícios para a instituição mas, no entanto, a sua construção possui algumas limitações. Uma das principais limitações que podem surgir na construção de um modelo de credit scoring relaciona-se com a base de dados utilizada no desenvolvimento desse modelo [Hand, 2001]. A maioria dos modelos de credit scoring utiliza somente os candidatos aceites (clientes a quem foi atribuído um empréstimo) e os que foram rejeitados não são, regra geral, incluídos na amostra para a construção do modelo. Assim, a amostra será parcial (i.e. diferente da população em geral) uma vez que "os bons clientes" estão muito bem representados. Os modelos de credit scoring construídos, baseados na situação acima referida, podem não ter um bom desempenho em observações que estejam mal representadas na amostra, ver [Lee e Chen, 2005].

O segundo problema que poderá ocorrer na construção de modelos de credit scoring baseia-se na mudança de padrões, ou seja, passagem de um estado para outro (por exemplo, de cumpridor para incumpridor) ao longo do empréstimo. O pressuposto fundamental para qualquer modelação preditiva é que o passado pode prever o futuro [Berry, 2000]. No credit scoring isso significa que as características dos candidatos, que são posteriormente classificados como "cumpridores" ou "incumpridores", podem ser utilizadas para prever o *status* de crédito de novos candidatos. Às vezes, a tendência para a mudança das características dos clientes é tão rápida que exige uma actualização constante do modelo de credit scoring para que este permaneça relevante.

Outra limitação do modelo de credit scoring é a possibilidade da total dependência de algumas instituições na utilização destes modelos o que, em alguns casos especiais, pode condicionar a utilização de um julgamento mais prudente. Em outros casos, as equipas de desenvolvimento de credit scoring podem involuntariamente aplicar mais recursos do que é necessário para trabalhar toda a carteira, ver [Lucas, 2000]. A maioria dos estudos apenas analisa o desempenho da previsão média dos modelos, sem considerar os erros de tipo I e de tipo II.

1.5 Recolha e Análise da Literatura

A recolha da literatura baseia-se nos critérios da descrição detalhada dos modelos e experiências realizadas em conjuntos de dados reais. Numa primeira fase fez-se a pesquisa nas bases de dados mais importantes da área (ScienceDirect, JSTOR, ProQuest Database, Business Source Complete Research Databases, IngentaConnect, SpringerLink, Blackwell Synergy, B-on, Science Citation Index, Wiley InterScience) onde a pesquisa foi realizada com as palavras-chave "credit scoring" ou "Credit risk evaluation (assessment/analysis)" quer

no título quer no resumo, em artigos publicados em jornais académicos, em proceedings de conferências, congressos e seminários durante o período entre 1970 e Março de 2010.

Numa segunda fase, para a selecção das monografias relativas a modelos de risco de crédito foram consultadas as duas maiores websites de livros online: Amazon (www.amazon.com) e Google books search (http://books.google.com).

Relativamente à palavra-chave "credit scoring", foram encontrados pouco mais de vinte livros ou monografias relevantes para este tópico, o que evidencia a pertinência de se continuar a desenvolver investigação nesta área. A compilação apresentada na Tabela 1.1, baseia-se nas websites acima referidas, no livro de [Anderson, 2007] e na página de Ross Gayler (http://sites.google.com/site/rgayler/creditscoringresources)[2].

Nos livros apresentados na Tabela 1.1, os modelos de risco de crédito focalizam-se principalmente em métodos estatísticos. A revisão baseia-se principalmente nos artigos, uma vez que todos os métodos descritos e desenvolvidos nesses livros encontram-se bem detalhados nos artigos científicos analisados. Na secção seguinte faz-se uma breve descrição das técnicas mais utilizadas na construção de um modelo de credit scoring.

1.6 Técnicas de Credit Scoring

O crescimento significativo da concessão de crédito proporcionou o desenvolvimento de uma enorme diversidade de métodos estatísticos e não estatísticos para a classificação dos clientes. Nesta secção, realizar-se-á uma breve descrição de algumas das técnicas mais utilizadas na construção de modelos de credit scoring: Modelos estatísticos, - Análise Discriminante (Discriminant Analysis, DA) e Regressão Logística (Logistic Regression, LR) - são apresentados num primeiro grupo. Num segundo grupo, apresentamos algumas metaheurísticas que podem ser utilizadas na construção dos modelos de Redes Neurais Artificiais (Artificial Neural Networs, ANN), Algoritmos Genéticos (Genetic Algorithms, GA), Classificação K-Vizinhos mais Próximos (k-Nearest-Neighbour Classifiers, KNN), Árvore de Decisão (Decision Trees, DT), Classificação da Rede Bayesiana (Bayesian Network Classifiers, BNC), Técnicas de Aprendizado de Máquinas (Support Vector Machines, SVM), Conjunto Rugosos/aproximativos (Rough Sets, RS). Por fim, apresentamos alguns Conjuntos de Métodos Híbridos (Hybrid and Ensemble Methods, HEM).

[2]consultada em Março de 2010

Tabela 1.1: *Livros sobre credit scoring*

Ano	Autor(es)	Título
2007	Anderson, R.	The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation
2004	Bailey, M. (Ed.)	Consumer credit quality: Underwriting, scoring, fraud prevention and collections.
2004	Bailey, M. (Ed.)	Credit scoring: The principles and practicalities (2nd ed.)
2004	Bailey, M.	Lies, damn lies and statistics in consumer credit.
2006	Bailey, M. (Ed.)	Consumer collections and recoveries: Operations and strategies (2nd ed.).
2006	Bailey, M.	Practical credit scoring: Issues and techniques.
2010	Breeden, J.L.	Reinventing retail lending analytics: Forecasting, stress testing, capital and scoring fpr a world of crises.
2006	Engelmann, B. & Rauhmeier, R. (Eds.)	The Basel II Risk parameters: Estimation, validation, and stress testing.
2008	Finlay, S.	The management of consumer credit: Theory and practice.
2009	Finlay, S.	Consumer credit fundamentals (2nd ed.).
1994	Lewis, E. M.	An introduction to credit scoring (2nd ed.).
1998	Mays, E. (Ed.)	Credit risk modeling: Design and application.
2001	Mays, E. (Ed.)	Handbook of credit scoring.
2003	Mays, E. (Ed.)	Credit scoring for risk managers: The handbook for lenders.
2008	McNab, H. & Taylor, P.	Principles and practice of consumer credit risk management (3rd ed.).
2005	Siddiqi, N.	Credit risk scorecards: Developing and implementing intelligent credit scoring.
2009	Thomas, L. C.	Consumer credit models: Pricing, profit and portfolios.
1992	Thomas, L. C., Crook, J. N. & Edelman, D. B. (Eds.)	Credit scoring and credit control.
2002	Thomas, L. C., Edelman, D. B. & Crook, J. N.	Credit scoring and its applications.
2004	Thomas, L. C., Edelman, D. B. & Crook, J. N. (Eds.)	Readings in credit scoring: Recent developments, advances, and aims.
2008	Van Gestel, T. & Baesens, B.	Credit risk management: Basic concepts.
1995	Hoyland, C.	Data-Driven Decisions for Consumer Lending.
2003	Thomas, L., Edelman, D., & Crook, J.	Readings in Credit Scoring.

1.6.1 Modelos Estatísticos

Análise Discriminante

A Análise Discriminante (DA) foi inicialmente proposta por Fisher (1936), como técnica de classificação. A sua utilização tem sido aplicada em várias áreas tais como: medicina, educação, biologia, engenharia, negócios, quimica, gestão, ver [Trevino e Daniels, 1995] e [Altman, 1968]. Nas décadas de 80 e 90, a aplicação da análise discriminante na construção de modelos de credit scoring foi utilizada nos trabalhos de [Bardos, 1998], [Desai et al., 1996], [Overstreet et al., 1992], [Titterington, 1992], [Reichert et al., 1983] e [Martell e Fitts, 1981].

A Análise Discriminante é utilizada geralmente para classificar as observações em dois ou mais grupos mutuamente exclusivos, usando as informações fornecidas por um conjunto de atributos preditivos. Quando apenas duas classes estão envolvidas, esta técnica é conhecida como análise discriminante de dois grupos. Porém, quando há três ou mais grupos, é chamada de análise discriminante múltipla, ver [Hair et al., 2006]. Dada a natureza do problema em estudo nesta dissertação, focalizamos a revisão da literatura apenas nos modelos que utilizam a análise discrimante de duas classes:

$$y = c + \sum_{i=1}^{n} w_i \, x_i$$

em que y é designado por score discriminante, c é uma constante, w_i é o peso de cada atributo e x_i é a característica independente.

Embora a Análise Discriminante seja uma das técnicas de data mining mais utilizada nos problemas de classificação, esta possui algumas limitações:

- não permite tratar de forma conveniente situações em que as variáveis independentes apresentem natureza categórica;

- assume que as observações se distribuem de forma idêntica pelos vários grupos considerados;

- assume também que as variáveis preditivas, para além de se relacionarem de forma linear, seguem uma distribuição normal e que a hipótese de homocedasticidade é válida.

Regressão Logística

A Regressão Logística (LR) é comummente utilizada para a análise de dados com resposta binária ou dicotómica e consiste em relacionar, através de um modelo funcional, a variável resposta com factores que influenciam ou não a probabilidade de ocorrência de determinado

evento. As variáveis independentes podem ser contínuas, categóricas, ou de ambos os tipos, ver por exemplo, [Cox e Snell, 1989] e [Hosmer e Lemeshow, 1989].

No âmbito da aplicação ao risco de crédito, a técnica de regressão logística é utilizada para a avaliação do risco de incumprimento. Assume-se que a probabilidade de incumprimento é logisticamente distribuída, com resultado binomial 0 ou 1. A Regressão Logística é utilizada para prever a probabilidade de ocorrência de um determinado evento (neste caso, a concessão de crédito) e assume que a relação de máxima verosimilhança (odds) é linear. Matematicamente, a regressão logística escreve-se da seguinte forma:

$$y = log(\frac{p}{1-p}) = w_0 + \sum_{i=1}^{n} w_i \, log \, x_i$$

onde p é a probabilidade de sucesso, y é a probabilidade dos resultados da classificação, w_i é o peso associado à característica e x_i é a variável explicativa. A Regressão Logística permite obter assim, uma variável preditiva $y = log(p/(1-p))$, que resulta da combinação linear das variáveis explicativas. Os valores que esta variável preditiva assume são, posteriormente, transformados em probabilidades através da aplicação de uma função logística. Este método tem sido amplamente utilizado nas aplicações de credit scoring devido à simplicidade da sua interpretação. As Tabelas 1.2 e 1.5 referem alguns desses trabalhos.

Tabela 1.2: *Aplicações da Regressão Logística na construção de modelos de credit scoring*

Área de aplicação	Referência
Credit scoring	[Laitinen, 1999], [Joanes, 1993] e [Wiginton, 1980]
Crédito comercial	[Leonard, 1993], [Westgaard e Van der Wijst, 2001] e [Dinh e Kleimeier, 2007]
Microfinanças	[Schreiner, 2004]

Relativamente às hipóteses subjacentes, a Regressão Logística é bastante mais flexível que a Análise Discriminante. Ao contrário da Análise Discriminante, a Regressão Logística não exige que as variáveis independentes sejam normalmente distribuídas, nem linearmente relacionadas e nem mesmo que exista igualdade da variância dentro de cada grupo, ver por exemplo [Tabachnick e Fidell, 1996]. Apesar da relação dessas hipóteses em [Harrell e Lee, 1985] verifica-se que Regressão Logística pode ser tão eficiente e precisa quanto a Análise Discriminante.

1.6.2 Metaheurísticas

Nesta subsecção selecionamos um conjunto de estudos que utilizam técnicas metaheurísticas no desenvolvimento dos modelos de credit scoring. Na Tabela 1.3, encontram-se alguns desses trabalhos.

Tabela 1.3: *Metaheurísticas*

Metaheuristicas	Referências
ANN	[Abdou et al., 2008] [West, 2000] [Piramuthu, 1999] [Desai et al., 1996] [Goldberg, 1989] e [Jensen, 1992]
GA	[Desai et al., 1996] [Yobas et al., 2000] [Chen e Huang, 2003] e [Varetto, 1998]
KNN	[Chatterjee e Barcun, 1970] e [Henley e Hand, 1996]
DT	[Frydman et al., 1985] [Davis et al., 1992] e [Zhou et al., 2008]
BNC	[Islam et al., 2007] [Baesens et al., 2002] e (Chang et al., 2000)
SVM	[Huang et al., 2007] [Lee, 2007] [Li et al., 2006] [Gestel et al., 2006] [Huang et al., 2004] e [Baesens et al., 2003]
RS	[Beynon e Peel, 2001]

OBS: Artificial Neural Networs, (ANN), Genetic Algorithms (GA), Decision Trees (DT), k-Nearest-Neighbour Classifiers (KNN), Bayesian Network Classifiers (BNC), Support Vector Machines (SVM) e Rough Sets (RS)

1.6.3 Modelos e Conjuntos Híbridos

Várias experiências têm demonstrado que a hibridização e conjuntos de dois ou mais modelos podem gerar resultados mais precisos do que a utilização individual de cada modelo. Hibridação e conjuntos de múltiplos métodos simples podem superar algumas das limitações de um único método e, assim, gerar uma classificação mais poderosa e um melhor sistema de previsão.

Nos últimos anos, vários trabalhos utilizaram métodos híbridos para a construção de modelos de credit scoring. Na Tabela 1.4 apresentam-se alguns exemplos de aplicação, em particular o estudo de [Hsieh, 2005] apresenta uma arquitectura híbrida para gerar um modelo de scoring. Fundamenta-se em duas técnicas amplamente utilizadas em data mining: Clustering e Redes Neuronais. Para a análise de clustering utilizam-se Self-Organized Maps (SOM), que permitem agrupar as observações que constituem o dataset em análise. De seguida, eliminam-se os clusters que não possuam amostras significativas. Numa segunda fase, utilizam-se as Redes Neuronais para a construção do modelo de scoring.

Embora a utilização destes métodos híbridos possa melhorar o desempenho do modelo de credit scoring, a forma como os vários métodos podem ser combinados para esse fim ainda não é clara.

Tabela 1.4: *Modelos e Conjuntos Híbridos*

Referência	Técnicas Comparadas
[Malhotra e Malhotra, 2003] e [Piramuthu, 1999]	Sistema fuzzy e ANN
[Wang et al., 2005]	Sistema fuzzy e SVM
[Yu et al., 2008] e [Ahn et al., 2000]	rough set e SVM
[Lee e Chen, 2005]	ANN e MARS
[Hsieh, 2005]	Clustering e ANN

1.7 Comparação das Técnicas

Na literatura são vários os trabalhos em que se comparam diferentes técnicas para o desenvolvimento dos modelos de credit scoring. Em geral, essa comparação é realizada ao nível da eficiência e do desempenho na previsão da probabilidade de incumprimento (clientes cumpridores e incumpridores). Nesta secção faz-se uma comparação das várias técnicas. Essa comparação é baseada no desempenho apresentado, na precisão de predição e na dimensão da amostra utilizada. A Tabela 1.5 apresenta alguns estudos que fizeram comparações entre as várias técnicas identificadas nas secções anteriores.

Tabela 1.5: *Comparação das Técnicas*

Referências	Técnicas
[Wiginton, 1980]	LR e DA
[Altman et al., 1994]	DA, LR, ANN e CT
[Tam e Kiang, 1992]	DA, LR, KNNs e DT
[Malhotra e Malhotra, 2003], [Piramuthu, 1999], [Desai et al., 1997], [Desai et al., 1996], [Jensen, 1992] e [Salchenberger et al., 1992]	ANNs, LR e DA
[Desai et al., 1996]	LDA, LR e ANN
[Yobas et al., 2000]	LDA, ANNs, GAs, e DT
[Galindo e Tamayo, 2000]	CART, KNNs e PA
[West, 2000]	ANN e DT
[Baesens et al., 2003]	SVMs, LR e LDA
[Schebesch e Stecking, 2005] e [Gestel et al., 2006]	SVMs, LR e LDA
[Huang et al., 2004]	SVM e ANN
[Huang et al., 2007]	ANN, DT e GA
[Abdou et al., 2008]	LR e ANN

O estudo de [Wiginton, 1980], foi uns dos pioneiros a comparar a Regressão Logística com

Análise Discriminante numa aplicação de classificação de crédito. Os resultados mostraram que Regressão Logística apresentava uma maior taxa de precisão, no entanto, nenhuma técnica foi considerada suficientemente boa para a rentabilidade do problema. O estudo de [Altman et al., 1994] compararou a Análise Discriminante Linear, Regressão Logística e Redes Neuronais na classificação dos maus clientes.

No início dos anos noventa, [Tam e Kiang, 1992] estudaram a aplicação das Redes Neuronais Artificiais para a previsão de insolvência bancária. Neste estudo, as Redes Neuronais Artificiais foram comparadas com a Análise Discriminante Linear, Regressão Logística, K-Vizinho mais Próximo e Árvore de Decisão. Os resultados desse trabalho sugerem que os modelos de credit scoring com base em Redes Neuronais Artificiais são mais precisos, seguidos pelos construídos através da Análise Discriminante Linear, Regressão Logística, Árvore de Decisão e K-Vizinho mais Próximo.

Também os trabalhos de [Malhotra e Malhotra, 2003], [Piramuthu, 1999], [Desai et al., 1997], [Desai et al., 1996], [Jensen, 1992] e [Salchenberger et al., 1992]) reforçaram essa ideia, e a principal razão apontada para tal facto, diz respeito à forma como as Redes Neuronais são hábeis a lidar com as relações não lineares existentes entre as variáveis.

Os estudos de [Abdou et al., 2008], [Huang et al., 2007], [Baesens et al., 2003], [Lee, 2007], [Gestel et al., 2006], [Schebesch e Stecking, 2005] e [Huang et al., 2004] compararam a utilização de Suport Vector Machine para a construção de modelos de credit scoring com técnicas estatísticas tradicionais. A principal conclusão a retirar da análise desses trabalhos é a de que nenhuma das técnicas analisada se destaca em termos de desempenho das restantes.

1.8 Discussão

A literatura sobre a análise de crédito e as técnicas estatísticas paramétricas e não paramétricas utilizadas para a construção de modelos de credit scoring foi revista na seção anterior. A Análise Discriminante foi uma das primeiras técnicas a ser utilizada na construção de scorecards [Reichert et al., 1983]. No entanto, a sua adequação para a construção de modelos de credit scoring é questionada, quer pela natureza categórica dos dados, quer pelo facto da natureza da covariância das classes de crédito não serem susceptíveis de serem iguais, ver [Hsieh, 2005].

De acordo com os estudos analisados conclui-se que uma técnica alternativa à Análise Discriminante é a Regressão Logística. A Regressão Logística é a técnica mais utilizada pela maioria das instituições financeiras na construção dos seus modelos de risco de crédito.

Outras técnicas, em particular, as não paramétricas e a combinação destas, têm sido bastante utilizadas por investigadores e profissionais da área.Nos últimos anos, vários estudos

têm demonstrado que as técnicas de Metaheurísticas, tais como Redes Neuronais Artificiais [Desai et al., 1996] e [West, 2000], Árvore de Decisão [Hung e Chen, 2009] e Support Vector Machine (SVM) [Schebesch e Stecking, 2005], [Hung e Chen, 2009] e [Baesens et al., 2003] podem ser utilizados de forma eficiente como métodos alternativos de credit scoring. A utilização de Algoritmos Genéticos ou de Redes Neuronais Artificiais é particularmente interessante em situações em que a variável dependente e as variáveis independentes apresentem relações não-lineares e complexas. Em contraste com as técnicas estatísticas, as técnicas de inteligência artificial não exigem que se formulem hipóteses sobre as distribuições de dados. Estas técnicas extraem automaticamente conhecimento a partir das amostras de treino.

Capítulo 2

Estimação da Probabilidade de Incumprimento

2.1 Introdução

Nas últimas décadas, tem havido um enorme interesse em utilizar modelos de credit scoring na avaliação e gestão de risco de crédito no sector bancário e no sistema financeiro em geral. Num ambiente cada vez mais competitivo e ancorado pelo cenário de crise no sistema financeiro, particularmente no sistema bancário, as técnicas de credit scoring tornaram-se, actualmente, uma das ferramentas mais importantes utilizadas na avaliação do risco de crédito dos empréstimos bancários. Além disso, nas últimas décadas, o credit scoring é considerado como uma das principais aplicações dos problemas de classificação, ver [Abdou et al., 2008].

Em Cabo Verde, com a publicação das novas regras bancárias (rácio de capital, fornecimento e empréstimo, classificação e avaliação de risco de crédito, concentração de crédito e a necessidade de quantificar o risco de crédito), em Novembro de 2007, pelo Banco de Cabo Verde (BCV), as instituições bancárias viram-se obrigadas a reestruturar os seus gabinetes de análise de risco no sentido de, por um lado, responder a estas novas exigências e, por outro, implementar um sistema que melhor permitisse assessorar os decisores na concessão, gestão e avaliação de crédito.

Os modelos de risco de crédito, em particular os de credit scoring, são procedimentos usuais e obrigatórios no sistema financeiro dos países desenvolvidos. Contrariamente, os sistemas financeiros dos países em desenvolvimento não possuem mecanismos avançados, ou os que existem, funcionam ainda de forma deficitária. Com o acordo Basileia II, estes países foram sujeitos a desenvolver diferentes modelos de risco de crédito.

Apesar das considerações acima descritas, importa destacar alguns estudos, em particu-

lar no contexto do crédito ao consumo, desenvolvidos para os países em vias de desenvolvimento, dos quais Cabo Verde faz parte: são eles os estudos de [Abdou et al., 2008] para o Egipto, [Dinh e Kleimeier, 2007], para o Vietname, [Schreiner, 2004] para a Bolívia e [Viganó, 1993] para o Burkina Faso. Nestes dois últimos trabalhos, as amostras analisadas apresentavam uma dimensão bastante reduzida para que as conclusões fossem relevantes (31 e 100 empréstimos, respectivamente).

O artigo de [Dinh e Kleimeier, 2007] foi o primeiro trabalho aplicado a um país não industrializado, em que a amostra estudada apresentava uma dimensão considerável (56 mil empréstimos) e em que os modelos de credit scoring desenvolvidos se aplicavam a todos os segmentos do mercado a retalho. Para o caso de Cabo Verde, existe o trabalho de [Semedo, 2010] que consistiu em comparar as Redes Neuronais e Regressão Logística aplicado a uma base de dados de crédito ao consumo de uma instituição financeira Cabo-verdiana.

A decisão de aceitar ou rejeitar o crédito a um cliente (ou futuro cliente), pode ser suportada por técnicas de julgamento e/ou modelos de credit scoring. As técnicas de julgamento dependem do conhecimento adquirido no passado, do contexto actual e da experiência dos gestores de crédito, que avaliam se um determinado cliente satisfaz ou não os requisitos necessários para a concessão de crédito. Tais requisitos podem ser, por exemplo, prestígio pessoal do cliente, a sua capacidade de reembolso do crédito, as garantias especiais disponibilizadas, entre outros, ver [Sarlija et al., 2004].

A Regressão Logistica é uma técnica estatística multivariada e é utilizada para prever a probabilidade de um evento (por exemplo, a probabilidade de ocorrer incumprimento) e assume que a relação de máxima verosimilhança (odds) é linear (a sua descrição será desenvolvida na secção seguinte). Na construção de modelos de credit scoring, a Regressão Logística foi explorada por diversos autores, entre os quais referimos [Westgaard e Van der Wijst, 2001], [Laitinen, 1999], [Joanes, 1993] e [Wiginton, 1980]. O modelo de Regressão Logística tem sido amplamente discutido em outras áreas como investigação social, investigação médica, previsão de falência, segmentação do mercado e comportamentos do cliente. Algumas referências desses trabalhos são, por exemplo, [Kay et al., 2000], [Laitinen e Laitinen, 2000], [Suh et al., 1999] e [Flagg et al., 1991].

As medidas de curva ROC, estatística de Kolmogorov Smirnov e o índice de Gini permitem-nos concluir se o modelo estimado através da Regressão Logística tem capacidade de, a partir dos dados do crédito ao consumo, discriminar de forma bastante satisfatória os clientes *incumpridores* dos clientes *cumpridores* e, desta forma, contribuir para o fortalecimento do processo da avaliação e estimação do crédito ao consumo, no sector bancário Cabo-verdiano.

Neste capítulo pretende-se estimar um modelo de probabilidade de incumprimento para os clientes de crédito ao consumo, utilizando a Regressão Logística e, simultaneamente, identificar quais as variáveis determinantes na probabilidade de incumprimento. O modelo

ajustado identifica como variáveis significativas o número de prestações pagas, a actividade profissional, a entidade patronal, a agência, a taxa nominal, o valor da prestação, a idade, o género, o prazo, as habilitações literárias e o valor de empréstimo.

Este capítulo está organizado da seguinte forma: na secção 2.2 descreve-se a metodologia baseada na fundamentação teórica da Regressão Logística como caso particular dos Modelos Lineares Generalizados (MLG). A descrição da base de dados encontra-se na secção 2.3. Ainda nesta secção descreve-se os procedimentos para a construção da variável dependente e as técnicas de identificação das variáveis independentes. Na secção 2.4 as técnicas de validação do modelo são apresentadas. Na secção 2.5 apresentam-se e discutem-se os resultados empíricos para a carteira de crédito ao consumo. Ainda nesta secção estima-se a probabilidade de incumprimento para cada cliente. Por fim, apresentam-se as conclusões e as principais linhas de investigação futuras.

2.2 Modelos Lineares Generalizados

Para a construção deste texto, que apresenta a metodologia de Modelos Lineares Generalizados (MLG's) pode, por exemplo, ser citado como referência [Turkman e Silva, 2000]. Os MLG's, introduzidos por [Nelder e Wedderburn, 1972], correspondem a uma síntese de vários modelos estatísticos, incluindo a regressão linear, regressão logística, modelo probit para estudos de proporção e regressão de Poisson, que vêm assim unificar, tanto do ponto de vista teórico como conceptual, a teoria da modelação estatística até então desenvolvida. A particularidade destes modelos prende-se com a apresentação duma estrutura de regressão linear e têm em comum o facto da variável resposta seguir uma distribuição dentro de uma família de distribuições com propriedades muito específicas: a família exponencial.

Seja Y a variável aleatória, de interesse primário, também designada por variável dependente ou variável resposta, e um vector $\mathbf{X} = (x_1, \ldots, x_k)^T$ de k variáveis explicativas, também designadas por covariáveis ou variáveis independentes, que se crê explicarem parte da variabilidade inerente a Y. A variável resposta Y pode ser contínua, discreta ou dicotómica. As covariáveis, determinísticas ou estocásticas, podem ser também de qualquer natureza: contínuas, discretas, qualitativas de natureza ordinal ou dicotómicas. Assume-se que os dados têm a forma

$$(y_i, \boldsymbol{x_i}), i = 1, \ldots, n, \tag{2.1}$$

resultantes da realização de $(Y, \boldsymbol{X})$ em n indivíduos, sendo as componentes Y_i do vector aleatório $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ independentes. Pode-se representar (2.1) na forma matricial,

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p}. \end{bmatrix} \tag{2.2}$$

Os modelos lineares generalizados são uma extensão do modelo linear clássico,

$$\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon} \tag{2.3}$$

ou

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \tag{2.4}$$

onde $\mathbf{Z}$ é uma matriz de dimensão $n \times p$ de especificação do modelo (em geral corresponde à matriz de variáveis explicativas $\mathbf{X}$ um primeiro vector unitário), associada a um vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ de parâmetros e $\boldsymbol{\varepsilon}$ um vector de erros aleatórios. A escolha da função de ligação depende do tipo de resposta e do estudo particular que se pretende realizar. Por exemplo, para dados binários utiliza-se a função de ligação Logit que será tratada no ponto seguinte, na exposição acerca do modelo de Regressão Logística.

Definição 2.1. *Diz-se que uma variável aleatória Y tem distribuição pertencente à família exponencial se a sua função densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) se puder escrever na forma:*

$$f(y|\theta, \phi) = exp(y\theta - b(\theta))/a(\phi) + c(y, \phi), \tag{2.5}$$

onde θ e ϕ são parâmetros escalares, $a(.), b(.)$ e $c(.)$ são funções reais conhecidas. Neste trabalho considera-se $\phi = 1$. Para aplicar a metodologia dos MLG's a um conjunto de dados existe a necessidade de após a formulação do modelo que se pensa adequado, de se proceder à realização de inferências sobre esse modelo. A inferência em MLG's baseia-se essencialmente na verosimilhança.

2.2.1 Regressão Logística

A Regressão Logística pode ser utilizada quando se deseja perceber a natureza do relacionamento entre a resposta média (probabilidade de ocorrência de um evento) e uma ou mais variáveis independentes, ou então com o objectivo preditivo, quando se deseja prever se determinado evento ocorrerá num prazo pré-definido, dado um conjunto de variáveis explicativas. A Regressão Logística é um caso particular dos modelos lineares generalizados, ver [McCullagh e Nelder, 1989], onde cada variável resposta é binomialmente distribuída

$Y_i \sim B(1, p_i)$ com probabilidade de "sucesso" p_i e de "fracasso" $(1 - p_i)$. A função probabilidade é dada por

$$f(y_i \mid p_i) = p_i^{y_i}(1 - p_i)^{1-y_i}, \quad y_i = 0, 1 \tag{2.6}$$

e a função de ligação é a função logit. Aplicando a transformação logística à equação (2.6) obtém-se

$$logit(p_i) = log(\frac{p_i}{1 - p_i}) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i, \quad i = 1, \ldots, n. \tag{2.7}$$

Sendo $Y_i \sim B(1, p_i)$ pertencente à família exponencial, temos que $E[Y_i] = \frac{e^{\theta_i}}{1+e^{\theta_i}} = p_i$.

Fazendo $\theta_i = \boldsymbol{z}_i^T \boldsymbol{\beta}$ obtém-se, pela transformação inversa

$$E[Y_i] = p_i = \frac{exp(\boldsymbol{z}_i^T \boldsymbol{\beta})}{1 + exp(\boldsymbol{z}_i^T \boldsymbol{\beta})} \tag{2.8}$$

em que $\pi_i = P[Y_i = 1|X = x_i]$ é a probabilidade de sucesso. A cada indivíduo i está associado um vector de especificação $\boldsymbol{z}_i$, que resulta do vector das variáveis independentes $x_i, i = 1, \ldots, n$.

No âmbito da aplicação ao risco de crédito, a técnica de Regressão Logística é utilizada para a avaliação do incumprimento de determinado grupo de clientes em situações relativas à concessão de crédito, assumindo que a probabilidade de incumprimento é logisticamente distribuída, com resultado binomial 0 ou 1.

2.2.2 Método de Estimação

De modo a poder aplicar a metodologia dos modelos lineares generalizados a um conjunto de dados, há necessidade, após a formulação do modelo que se pensa adequado, de se proceder à realização de inferências sobre esse modelo. A inferência com MLG, como referido atrás é, essencialmente baseada na verosimilhança. Com efeito, não só o método da máxima verosimilhança é o método de eleição para estimar os parâmetros de regressão, como também os testes de hipóteses sobre os parâmetros do modelo e de qualidade do seu ajustamento são, em geral, métodos baseados na verosimilhança, ver [Turkman e Silva, 2000].

Os procedimentos de estimação e inferência a serem utilizados em Regressão Logística são um caso particular da metodologia de MLG's já descritos. A função de verosimilhança para o modelo logístico é dada por

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \tag{2.9}$$

em que:

$$p_i = \frac{exp(\boldsymbol{z}_i^T \boldsymbol{\beta})}{1 + exp(\boldsymbol{z}_i^T \boldsymbol{\beta})}, \quad x_i, i = 1, \ldots, n \quad i.i.d. \tag{2.10}$$

Os estimadores são obtidos através da matriz hessiana da função de log verosimilhança.

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j^2} = -\sum_{i=1}^{n} x_{i,j}^2 p_i(1-p_i) \quad e \quad \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \boldsymbol{\beta}_l} = -\sum_{i=1}^{n} x_{i,j} x_{i,l}\, p_i(1-p_i)$$

Igualando estas expressões a zero, o sistema de equações obtido não é linear. Em virtude disso, essas equações não têm, em geral, solução analítica. Portanto, a sua resolução implica o recurso a métodos numéricos, ver [Turkman e Silva, 2000]. Como a verosimilhança depende da probabilidade de sucesso desconhecida p_i, que, por sua vez, depende dos parâmetros $\boldsymbol{\beta}'s$, a função de verosimilhança pode ser vista como função de $\boldsymbol{\beta}$.

2.2.3 Teste de Significância

Quando estamos perante um problema de selecção de covariáveis e queremos testar se um submodelo é melhor que o modelo original, é comum utilizar a Estatística de Wald, a Estatística de Wilks ou a Estatística de Razão de Verosimilhança (para uma descrição mais detalhada, ver [Turkman e Silva, 2000]). Estas estatísticas são deduzidas a partir das distribuições assimptóticas dos estimadores de máxima verosimilhança e de funções adequadas desses estimadores.

Considera-se o teste de hipóteses da forma:

$$H_0 : C\boldsymbol{\beta} = \xi \;\; vs \;\; H_1 : C\boldsymbol{\beta} \neq \boldsymbol{\xi}, \tag{2.11}$$

onde $\boldsymbol{C}$ é uma matriz $q \times p$, com $q \leq p$, de característica completa q e ξ é um vector de dimensão q, previamente especificado. Seja o caso particular:

$$H_0 : \mathbf{C}\boldsymbol{\beta}_j = 0 \;\; vs \;\; H_1 : \mathbf{C}\boldsymbol{\beta}_j \neq \mathbf{0}, \tag{2.12}$$

para algum j, sendo $q = 1$ e $\boldsymbol{C} = (0, \ldots, 0, 1, 0, \ldots, 0)$ e sendo 1 a j-ésima posição da matriz $\boldsymbol{\xi} = 0$. No caso em que uma variável é policotómica e toma $r + 1$ valores distintos, é aconselhável construir r variáveis dicotómicas para as representar havendo, nesse caso, r parâmetros $\boldsymbol{\beta}'s$ que lhe estão associados. Para se averiguar se essa variável deve ou não ser incluída no modelo, interessa testar se os r parâmetros são significativamente diferentes de zero. Também a importância da associação entre a variável dependente e cada uma das variáveis explicativas é avaliada. Essa avaliação é sustentada pelos seguintes testes:

Teste de Wald

A estatística de Wald, ver [Turkman e Silva, 2000] baseia-se na normalidade assimptótica do estimador de máxima verosimilhança $\hat{\boldsymbol{\beta}}$. Considere-se que a hipótese nula estabelece que $\boldsymbol{C\beta} = \boldsymbol{\xi}$, onde $\boldsymbol{C}$ é a matriz $q \times p$, de característica completa q. Seja $\mathbf{C\beta} = \boldsymbol{\xi}$ o estimador de máxima verosimilhança de $\boldsymbol{\beta}$, o qual tem uma distribuição assimptótica $N_p(\boldsymbol{\beta}, \Im^{-1}(\boldsymbol{C\beta}))$ (aqui o vector $\boldsymbol{\beta}$ já foi substituído pela sua estimativa admitindo que para grandes amostras $I(\boldsymbol{\beta}) \approx I(\hat{\boldsymbol{\beta}})$, onde $\Im^{-1}(\boldsymbol{\beta})$ é matriz de covariâncias). Dado que o vector $\boldsymbol{C}\hat{\boldsymbol{\beta}}$ é uma transformação linear de $\hat{\boldsymbol{\beta}}$ então, pelas propriedades da distribuição normal multivariada,

$$\boldsymbol{C}\hat{\boldsymbol{\beta}} \sim \boldsymbol{N}_q(\boldsymbol{C\beta}, \boldsymbol{C}\Im^{-1}\hat{\boldsymbol{\beta}}\boldsymbol{C}^T) \tag{2.13}$$

e consequentemente, sob a hipótese nula, a estatística

$$W = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \xi)^T[\boldsymbol{C}\Im^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C}^T](C\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}), \tag{2.14}$$

tem uma distribuição assimptótica de um χ^2 com q graus de liberdade. A estatística W, em (2.14), designa-se por estatística de Wald.

Deste modo, rejeita-se a hipótese nula, a um nível de significância α, se o valor observado da estatística de Wald for superior ao quantil da probabilidade $1 - \alpha$ de um χ^2_q. Geralmente, a Estatística de Wald mais frequentemente utilizada para testar hipóteses sobre coeficientes individuais, embora também se use para testar hipóteses nulas do tipo $\boldsymbol{\beta}_r = \mathbf{0}$ quando o subvector $\boldsymbol{\beta}_r$ representa o vector correspondente a uma recodificação de uma variável policotómica. A estatística de Wald é muito conhecida e útil na selecção/exclusão das variáveis explicativas como iremos ver no ponto seguinte e na subsecção 2.2.4.

Teste de Razão de Verosimilhança

A Estatística de Razão de Verosimilhança, também conhecida por estatística de Wilks, ver por exemplo, [Turkman e Silva, 2000], é definida por:

$$\Lambda = -2 \ln \frac{max_{H_0} L(\boldsymbol{\beta})}{max_{H_0 \cup H_1} L(\boldsymbol{\beta}} = -2 \left(\ell(\tilde{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}) \right), \tag{2.15}$$

onde $\tilde{\boldsymbol{\beta}}$, o estimador de máxima verosimilhança restrito, é o valor de $\boldsymbol{\beta}$ que maximiza a verosimilhança, sujeito às restrições impostas pela hipótese $\boldsymbol{C\beta} = \boldsymbol{\xi}$. O Teorema de Wilks estabelece que, sob certas condições de regularidade, a estatística Λ tem, sob H_0, uma distribuição assimptótica de um χ^2 onde o número de graus de liberdade é igual à diferença entre o número de parâmetros a estimar sob $H_0 \cup H_1$ (neste caso p) e o número de parâmetros a estimar sob H_0 (neste caso $p - q$). Assim, sob H_0,

$$\Lambda = -2 \left(\ell(\tilde{\beta}) - \ell(\hat{\beta}) \right) \overset{a}{\sim} \chi^2_q. \tag{2.16}$$

Com base no Teste de Razão de Verosimilhanças, rejeita-se a hipótese nula $H_0 : \boldsymbol{C\beta} = \boldsymbol{\xi}$, a um nível de significância α, se o valor observado da estatística for Λ superior ao quantil de probabilidade $1 - \alpha$ de uma distribuição χ^2_q. A Estatística de Razão de Verosimilhança é mais utilizada quando se pretende comparar modelos que estão encaixados, isto é, modelos em que um é submodelo do outro, ver [Turkman e Silva, 2000].

2.2.4 Selecção das Variáveis Explicativas

A abordagem *stepwise* é um dos procedimentos mais populares na literatura para a escolha das variáveis, ver por exemplo [Fan e Cheng, 2007]. Esta abordagem *stepwise* selecciona as variáveis que deverão ser incluídas ou excluídas do modelo. Essa selecção baseia-se exclusivamente na medida de AIC (*Akaike Information Criterion*). A medida AIC leva em consideração tanto a log-verosimilhança dos dados, quanto o número de parâmetros do modelo ajustado, sendo que um modelo é melhor do que o outro se apresentar menor valor da medida de AIC.

Duas principais versões da abordagem *stepwise* são a selecção *forward* seguido do teste de eliminação *backward*, ou eliminação *backward* seguido por selecção *forward*. A selecção *forward* inicia-se somente com o modelo nulo, ou seja, sem nenhuma variável explicativa, e de seguida selecciona a variável a incluir, com base no menor valor de *p-value* da Estatística de Wald. A eliminação das variáveis inicia-se com o modelo saturado, ou seja, considere-se todas as variáveis e elimine-se a variável com maior valor de *p-value* da Estatística de Wald. A abordagem *stepwise* combina esses dois passos que incluem ou excluem variáveis em cada iteração.

De acordo com [Turkman e Silva, 2000], quanto menor (ou maior) for o valor do *p-value* dado pelo teste de Wald mais (menos) importante é a variável considerada. Após a escolha da variável, faz-se uma segunda análise ao seu grau de importância através do valor do *p-value* do teste de razão de verosimilhança entre os modelos que a incluem e os que a excluem. A decisão final sobre a exclusão (ou inclusão) da variável no modelo final é tomada após estas análises. O valor de *p-value* para o teste de significância estatística das variáveis para inclusão e exclusão do modelo é geralmente definido como 0.05, mas esse limite deve ser discutido pelo gestor de risco, ver [Dreiseitl e Ohno-Machado, 2002].

2.3 Base de Dados

Normalmente, os dados do crédito ao consumo são uma mistura de variáveis contínuas, semi-contínuas e categóricas. Muitas vezes, a base de dados possui milhões de registos e centenas de variáveis. Consequentemente, os modelos têm sido tradicionalmente construídos

utilizando amostras em detrimento da população total, tal como referido em [Finlay, 2008].

2.3.1 Fonte, Descrição e Processamento da Base de Dados

Os dados utilizados neste estudo foram fornecidos por um banco comercial Cabo-verdiano, representando a totalidade da carteira de crédito ao consumo. Para cada cliente, o conjunto de características que o definem pode ser subdividido em dois grupos: variáveis sócio-demográficas, que caracterizam o cliente no momento do pedido de empréstimo, e variáveis financeiras. Este último grupo permite ao banco calcular um conjunto de indicadores financeiros, que poderão ser utilizados no desenvolvimento do modelo de credit scoring.

2.3.2 Selecção do Período Temporal/Janela de Amostragem

Os modelos de credit scoring são desenvolvidos tendo como pressuposto que o desempenho futuro irá de alguma forma ser reflexo do desempenho passado (princípio da continuidade). Partindo deste pressuposto, o desempenho dos clientes da carteira é analisado a fim de prever o desempenho dos futuros clientes. Para realizar esta análise, é preciso reunir os dados dos clientes durante um determinado período de tempo, e monitorizar o seu comportamento para um outro período de tempo específico, de forma a prever o comportamento de um cliente ser "cumpridor" ou "incumpridor", ver [Siddiqi, 2006].

A janela de performance corresponde ao período temporal em que o desempenho dos clientes é monitorizado para atribuir para cada cliente o valor da variável dependente (ou seja classificá-lo como: incumpridor, indeterminado, cumpridor ou excluído). Alguns autores, [Anderson, 2007], [Siddiqi, 2006], [Thomas et al., 2001] e [Bailey, 2001], argumentam que é necessário que os clientes que farão parte da amostra tenham sido monotorizados um período mínimo de 12 a 18 meses, tempo necessário para se consolidar o seu comportamento. Em alguns casos, tais como fraudes e falências, a classe de desempenho já é conhecida ou pré-determinada. [Siddiqi, 2006] argumenta ainda que é útil, entretanto, realizar a análise descrita e seguidamente determinar a janela de performance ideal.

Uma forma simples de estabelecer o desempenho e a janela de amostragem é analisar o pagamento ou o comportamento dos clientes e traçar o desenvolvimento dos casos ao longo do tempo de incumprimento (ser um cliente *cumpridor* ou *incumpridor*). Uma boa fonte para essa análise é o corte mensal ou trimestral ou o relatório de análise vintage produzidos na maioria dos departamentos de risco de crédito, ver [Siddiqi, 2006].

2.3.3 Discussão da Variável Dependente/Target

Nesta secção faz-se a discussão e a construção da variável dependente (clientes *incumpridores*, *indeterminados* e *cumpridores*) baseada no número de dias em atraso no pagamento das suas prestações. Na realidade, cada instituição tem a sua própria política de crédito e estes conceitos, de cliente *cumpridor* ou *incumpridor* podem mudar dependendo da instituição, das conjecturas económicas ou do próprio país.

Existem várias definições para esses segmentos, dos quais destacaríamos: um cliente é considerado *incumpridor* ("mau" ou em "*default*"), se se atrasar no pagamento de alguma das prestações do contrato por um período superior a 90 dias nos primeiros doze meses da vigência do contrato, ver [Siddiqi, 2006]. Ainda segundo [Siddiqi, 2006], definições como "três vezes 30 dias em atraso ou duas vezes 60 dias em atraso, ou uma vez 90 dias em atraso", que podem ser reflexões mais precisas, são muito mais difíceis de identificar e podem não ser apropriadas para todas as empresas. [Siddiqi, 2006] recomenda que a escolha de uma definição mais simples facilita a gestão e a tomada de decisão.

Uma vez definidos os critérios que classificam um cliente como incumpridor, clientes *incumpridores*, considerar-se-á um cliente *cumpridor* quando este tem no máximo 30 dias em atraso no pagamento das suas prestações. Na realidade, num esquema usual de prestações mensais, significa que o cliente não tem nenhuma prestação em atraso. Definição dos clientes *cumpridores* é menos analítica e, geralmente, mais óbvia, tal como refere [Siddiqi, 2006]. Por fim, os restantes são os *indeterminados*. Estes são considerados assim, porque não existe, ainda, uma posição clara sobre eles. Sendo assim, estes clientes, não têm histórico de desempenho suficiente para a classificação, ou têm alguma delinquência, com *roll rate* não suficientemente baixa que permita ser classificado como *cumpridor* e nem suficientemente alta que permita ser classificado como *incumpridor*. Para um estudo mais detalhado dos *indeterminados*, consultar [Siddiqi, 2006]. Também podem existir clientes considerados como excluídos, por possuirem alguma característica peculiar não devem ser considerados (por exemplo, funcionário da instituição).

A maioria da literatura dos modelos de credit scoring concentra-se nos elementos da definição do desempenho *cumpridor* ou *incumpridor*. *Incumpridor* é geralmente definido com base nos indicadores de desempenho negativos, tais como a falência, a fraude, a inadimplência, *write-off/charge off* e um valor actual líquido negativo (VPN) [Anderson, 2007]. Uma informação mais detalhada sobre este tema pode ser encontrada em [Anderson, 2007] e em [Siddiqi, 2006].

Na especificação do modelo, as instituições consideram apenas os clientes *cumpridores* e *incumpridores*, devido à maior facilidade de trabalhar com os modelos de resposta binária. Os trabalhos de [Thomas, 2009], [Finlay, 2008], [Thomas et al., 2002] e [Hand e Henley, 1997] utilizaram essa metodologia na definição dos clientes *cumpridores* e *incumpridores*.

Nesta dissertação optou-se por classificar um cliente como *incumpridor* quando este esteve, pelo menos uma vez durante o contracto, com mais de noventa dias de incumprimento

2.3.4 Categorização e Escolha das Variáveis Explicativas

De acordo com [Anderson, 2007] e [Siddiqi, 2006], a maioria dos modelos de credit scoring desenvolvidos, independentemente do tipo, tem um grande número de variáveis explicativas que poderiam ser utilizadas. No entanto, regra geral, há apenas entre 6 a 15 características que melhor explicam o comportamento do cliente. A maioria das bases de dados utilizadas na construção de modelos de credit scoring apresenta um grande número de variáveis que, normalmente, são uma mistura de variáveis contínuas, semi-contínuas e categóricas. Estas podem ser categorizadas antes da sua utilização, fazendo selecção das que fazem parte do modelo final, ver [Anderson, 2007].

Relativamente à categorização das variáveis, há duas opções de implementação das variáveis categóricas no modelo de scoring [Thomas, 2000]. A primeira foi aplicada nos estudos de [Van Gool et al., 2009] e [Crook et al., 1992]. Nesta implementação, uma variável binária (*dummy*) é criada para cada categoria possível de uma variável exploratória e permite modelar o comportamento não linear. A outra abordagem, "Weight of Evidence" (WOE), categoriza uma variável em duas ou mais categorias, ver por exemplo, [Van Gool et al., 2009].

A selecção das variáveis com maior grau de capacidade na discriminação de clientes *incumpridores* e *cumpridores* da carteira faz-se calculando o Information Value (IV) de cada variável e o WOE para cada categoria da variável exploratória ver, por exemplo, [Van Gool et al., 2009], [Thomas, 2000], [Hand e Henley, 1997] e [Crook et al., 1992]. O Information Value (IV) para além da função acima referida, também corrige a diferença entre os dois grupos da carteira.

O WOE e o IV são utilizados em vários estudos para a análise e modelação porque, para alguns autores, representam uma boa alternativa para aproximar a não-linearidade nos dados, ver [Van Gool et al., 2009]. Valores negativos elevados correspondem a alto risco e valores positivos elevados correspondem a baixo risco. O IV, como é conhecido em credit scoring, é tecnicamente referenciado como medida de divergência de Kullback. Este mede a diferença entre duas distribuições. O cálculo desses indicadores procede-se da seguinte forma:

- para as *variáveis contínuas*, o IV é definido como:

$$IV = \int (f_c - f_i)\ \ln\left(\frac{f_c}{f_i}\right)\ dx, \tag{2.17}$$

  onde f_c e f_i são as densidades de probabilidade condicional da variável de previsão, quando a "resposta" é, respectivamente, *cumpridor* ou "*non-default*" e *incumpridor* ou "*default*".

- para *variáveis discretas*, calcula-se em cada intervalo a percentagem dos *cumpridores* (zeros) e *incumpridores* (uns). O WOE e IV são calculados da seguinte forma, ver [Siddiqi, 2006]:

$$WOE_k = \ln\left(\frac{fc_k}{fi_k}\right) \tag{2.18}$$

em que, fc_k e fi_k são distribuições de *cumpridores* e *incumpridores*, respectivamente. Finalmente, o IV é determinado através de:

$$IV = \{\sum_{k=1}^{n}(fc_k - fi_k) \times WOE_k\}. \tag{2.19}$$

A regra empírica para avaliar o IV, ver [Siddiqi, 2006], é a seguinte: se IV for inferior a 0.02, então a variável não é preditiva; entre 0.02 e 0.1, a variável tem fraco poder preditivo; entre 0.1 e 0.3, a variável tem poder preditivo médio; por fim, se for superior a 0.3: a variável tem um forte poder preditivo.

2.4 Desenvolvimento e Validação de Modelos

No desenvolvimento de um modelo de credit scoring, a amostra em estudo poderá ser dividida, dependendo da técnica utilizada, em amostra de treino, validação e teste e após a decisão da proporção de *cumpridores*, *incumpridores* e *rejeitados*[1] a incluir na amostra final, ver [Siddiqi, 2006].

Existem várias formas de dividir o conjunto de dados da amostra de treino (amostra em que o modelo de credit scoring é desenvolvido) e amostra de validação (amostra em que o modelo é validado). Normalmente, 70% a 80% da amostra é utilizada para treinar os modelos e os restantes 20% a 30% são reservados para a fase de validação dos modelos. Quando o número de observações é pequeno, o modelo pode ser desenvolvido e validado através de validação cruzada, ver [Siddiqi, 2006].

Para a validação de um modelo de credit scoring, três requisitos fundamentais são considerados: a estabilidade, a clareza e o poder discriminatório, ver [Gestel et al., 2006].

Estabilidade: Um modelo estável exige coeficientes bem determinados e com grande nível de confiança e resultados semelhantes em características de desempenho, se testados dentro e fora da amostra.

Legibilidade: Um modelo é legível quando os seus coeficientes têm uma interpretação fácil.

Poder discriminatório: Esta característica é definida pelo Comité de Supervisão Bancária de Basileia (2005) como a capacidade de classificar correctamente as observações sobre a base de probabilidade de inadimplência através da atribuição de pontuações.

[1] Os rejeitados não fazem parte deste estudo, como foi referido anteriormente.

Para testar o poder discriminatório e a comparabilidade dos modelos, várias medidas de desempenho são utilizadas: Percentagem dos Correctamente Classificados (PCC), Sensibilidade ($SENS$), Especificidade (Esp), Kolmogorov-Smirnov (KS), Índice de Gini, AUC e Rácio de Precisão (AR). Nos modelos de credit scoring, essas medidas permitem-nos averiguar a discriminação ideal entre clientes *cumpridores* e *incumpridores*. As subsecções que se seguem descrevem de forma detalhada as medidas que serão utilizadas no estudo empírico. É de destacar: a estatística de Kolmogorov-Smirnov, a curva ROC e o índice de Gini.

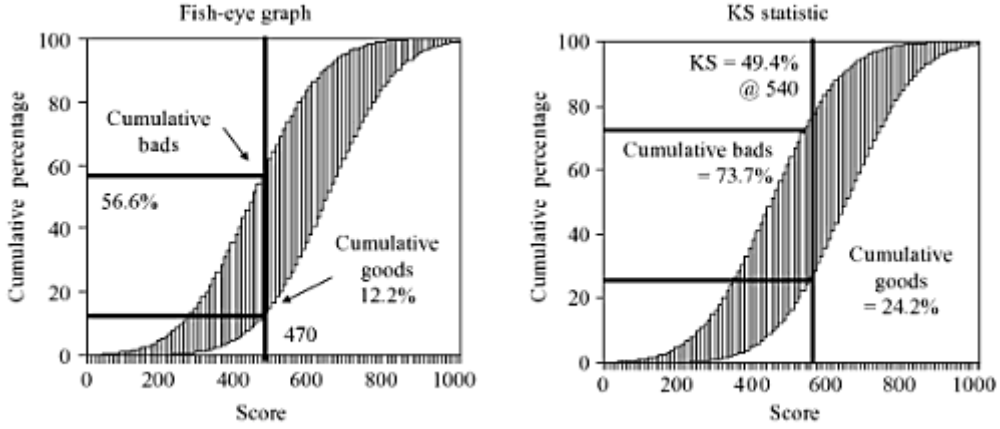
2.4.1 Estatística de Kolmogorov-Smirnov

A estatística de Kolmogorov-Smirnov (KS) é utilizada na teoria de estatística não paramétrica para testar se as funções de distribuição de uma variável são iguais em dois grupos, ver por exemplo, [Conover, 1999]. Este indicador é muito utilizado para avaliar o desempenho de modelos de credit scoring e baseia-se na ideia da distância entre as distribuições de probabilidades, s, dos clientes *incumpridores* e *cumpridores*. KS mede a máxima separação entre a frequência relativa acumulada dos clientes *incumpridores*, relativa acumulada dos clientes *cumpridores*, $F_b(s)$. A estatística de Kolmogorov-Smirnov é definida por:

$$KS = \max_{0 \leq s \leq \infty} |F_i(s) - F_c(s)|, \quad 0 \leq KS \leq 1. \tag{2.20}$$

Assim como o coeficiente de Gini, o KS varia entre 0 e 1 e valores mais altos indicam uma melhor perfomance. A Figura (2.1) ilustra um exemplo de cálculo do KS. Na Tabela 2.1

Figura 2.1: *Estatística de Kolmogorov-Smirnov (adaptado: Anderson,* 2007*)*



estão descritos intervalos de valores de KS, com os seus respectivos níveis de discriminação, usualmente adoptados como referência na modelação de credit scoring.

Tabela 2.1: *Valores de referência de KS(adaptado: Anderson,* 2007*).*

Valor de KS	**Níveis de Discriminação**
$[0, 0.25[$	Baixa
$[0.25, 0.35[$	Aceitável
$[0.35, 0.45[$	Bom
$[0.45, 1]$	Excelente

Apesar da estatística de KS ser a medida de avaliação mais utilizada, o uso isolado deste indicador não garante que, para valores altos desta medida, tenhamos modelos bem ajustados, pois existem situações em que se pode obter valores altos de KS quando clientes "cumpridores" e "incumpridores" estão apenas parâmetros numa faixa de score. É recomendável o uso desta medida em conjunto com pelo menos outros dois indicadores de desempenho, como por exemplo, a curva ROC e coeficiente de Gini.

2.4.2 Curva ROC e Coeficiente de Gini

A curva ROC (Receiver Operating Characteristic), também conhecida como curva de Lorenz [Hanley e McNeil, 1982], é baseada nos conceitos de sensibilidade e especificidade estatísticas (medidas da taxa de classificações correctas) que podem ser obtidas a partir da construção de matriz de classificação (2×2), ver [Johnson e Wichern, 2002], obtidas do resultado da classificação dos indivíduos gerada pelo modelo estimado.

Com o modelo ajustado, a partir de uma amostra de n clientes, atribui-se um score S a cada indivíduo. O i-ésimo indivíduo será classificado como "incumpridor" se $S_i \leq P_c$, (em que P_c é um ponto de corte para o score S_i, pré-determinado) e como "cumpridor", caso contrário. Para um determinado P_c, é possível determinar a matriz de classificação, também conhecida pela matriz de confusão ou tabela de contingência, como apresentada na Tabela 2.2.

Tabela 2.2: *Matriz de Classificação*

	Previsto		
Observado	Cumpridor	Incumpridor	Total
Cumpridor	n_{cc}	n_{ci}	$n_{c\bullet}$
Incumpridor	n_{ic}	n_{ii}	$n_{i\bullet}$
Total	$n_{\bullet c}$	$n_{\bullet i}$	$n_{\bullet\bullet}$

em que:

n_{cc} - Número de clientes "cumpridores" classificados como "cumpridores" - *Classificação*

correcta;

n_{ci} - Número de clientes "cumpridores" classificados como "incumpridores" - *Classificação incorrecta*;

n_{ic} - Número de clientes "incumpridores" classificados como "cumpridores" - *Classificação incorrecta*;

n_{ii} - Número de clientes "incumpridores" classificados como "incumpridores" - *Classificação correcta*;.

Através da matriz de classificação, é possível determinar as taxas de classificações correctas, que correspondem a medidas de especificidade (proporção dos clientes "incumpridores", classificados correctamente por terem score menor que um ponto de corte) e de sensibilidade (proporção de clientes "cumpridores", classificados correctamente por terem score igual ou superior a um ponto de corte), ou seja:

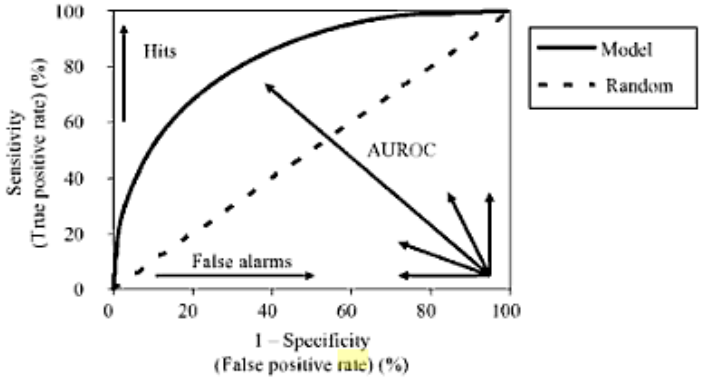
$$sensibilidade = \frac{n_{cc}}{n_{c\bullet}} \quad e \quad especificidade = \frac{n_{ii}}{n_{i\bullet}}. \tag{2.21}$$

Pode-se calcular também a precisão do modelo, que será dada pela proporção total da classificação correcta, ou seja:

$$precisão = \frac{n_{ii} + n_{cc}}{n}. \tag{2.22}$$

A curva ROC é construída a partir da união dos pontos formados pelos valores da *sensibilidade* e (1-*especificidade*), calculadas a partir de todas as matrizes de classificação, geradas pelas observações da amostra, considerando-se diferentes pontos de corte do modelo. A Figura 2.2 ilustra um exemplo de curva ROC.

Figura 2.2: *Curva ROC (adaptado: Anderson,* 2007*)*



Como a área sob a curva ROC varia entre 0.5 e 1, é mais adequado utilizar o coeficiente de Gini, ver [Thomas et al., 2002], que é dado por duas vezes a área entre a curva ROC e a recta

bissectriz ($y = x$). Tem-se um indicador de desempenho que varia entre 0 e 1. O cálculo de coeficiente de Gini resulta directamente da utilização da curva ROC. Observando-se o exemplo ilustrado na Figura 2.2 da curva ROC, descreve-se a interpretação do cálculo do coeficiente de Gini, que pode ser definido como sendo o quociente da área entre a recta e a curva (área A) sobre a área total acima da diagonal (soma da área A com a área B). Quanto mais a curva se afasta da recta, maior será o coeficiente de Gini e maior será a discriminação entre os "cumpridores" e "incumpridores" clientes, ou seja,

$$Gini = \frac{areaA}{areaA + areaB} \tag{2.23}$$

Como A é a diferença entre a área acima da diagonal e a área acima da curva e $A+B$ é toda a área acima da diagonal, sendo igual a metade da área do quadrado, (ou seja, $\frac{1}{2}$), pode-se obter o coeficiente de Gini da seguinte forma:

$$Gini = 2 \times (\textit{área acima da diagonal} - \textit{área acima da curva}), \tag{2.24}$$

ou ainda, directamente do valor obtido da curva ROC, como:

$$Gini = 2 \times (ROC - 0,5), \tag{2.25}$$

sendo ROC, neste caso, o valor obtido do cálculo da área sob a curva ROC.

A Tabela 2.3, apresenta-se os valores intervalares para avaliação do resultado da área sob a curva ROC, aplicadas em modelos de credit scoring, ver [Hosmer e Lemeshow, 1989].

Tabela 2.3: *Valores de referência da curva ROC*

Valor da Curva ROC	**Níveis de Discriminação**
$[0, 0.7[$	Baixa
$[0.7, 0.8[$	Aceitável
$[0.8, 0.9[$	Bom
$[0.9, 1]$	Excelente

2.5 Resultados e Discussões

Nesta secção, inicia-se uma análise estatística da carteira, dando especial relevância à categorização das variáveis, à validação dos diferentes modelos desenvolvidos, à identificação dos determinantes da probabilidade de incumprimento na subsecção **??** e, por fim, à estimação da probabilidade de incumprimento.

2.5.1 Análise Estatística da Carteira

A base de dados fornecida pelo banco é uma matriz de dados em painel, em que os indivíduos são registados no momento das decisões do empréstimo, associados a um conjunto de características socio-demográficas e financeiras. Fazem ainda parte da base de dados outras variáveis que a instituição calcula para fazer o seguimento dos clientes.

O principal pressuposto implícito na construção de um modelo de risco de crédito assenta em que o padrão de comportamento dos clientes se mantém ao longo do tempo. Por outro lado, os dados devem ser os mais actuais possíveis, tendo em conta as sucessivas mudanças a que a própria economia está sujeita. Assim, propomos um estudo com duas amostras com períodos temporais distintos, com o objectivo de, por um lado, tentar satisfazer estes pressupostos e, por outro, avaliar a robustez do modelo ajustado. A Amostra 1 contempla todos os clientes cujo crédito concedido entre Janeiro de 2003 e Março de 2011 e a Amostra 2 contém apenas os créditos concedidos num período mais recente, de Janeiro de 2006 a Março de 2011. Pretendemos, desta forma, avaliar se existem diferenças significativas nos resultados do modelo quando se observa o fenómeno num período temporal mais longo. A Tabela 2.4 caracteriza as duas amostras.

Para a preparação da base de dados, foram consideradas algumas restrições, de acordo com as instruções da instituição bancária, com o objectivo de eliminar possíveis erros ou mesmo valores atípicos. Rejeitaram-se alguns clientes de acordo com os seguintes critérios: valor emprestado inferior ou igual a 3.000 e superior ou igual a 2.500.000 escudos Cabo-verdianos (ECV); idade inferior ou igual a 17 anos e superior ou igual a 80 anos; valor utilizado na amortização da renda inferior ou igual a 2000 ECV; taxa nominal inferior ou igual a três (para eliminar os clientes com taxa nominal de $2,5\%$, uma vez que são funcionários da instituição e não devem ser utilizados para a estimação do modelo) e superior ou igual a 40%. O número de clientes excluídos de cada Amostra pode também ser consultado na Tabela 2.4.

Através da base de dados de performance dos clientes, definimos a variável resposta em duas fases: numa primeira fase, calculámos o número de dias em atraso na prestação e, de seguida, definimos os clientes "incumpridores" e "cumpridores". Definimos um cliente como sendo *incumpridor* se, pelo menos uma vez durante o período da Amostra ultrapassou os 90 dias em atraso. São considerados "cumpridores" todod os clientes que têm prestações em atraso com um máximo de 90.

Neste capítulo, consideramos como *indeterminados* os clientes com data de financiamento a partir de 01 de Abril de 2011. Essa restrição foi imposta com a finalidade de ter um período de tempo, neste caso seis meses, de modo que cada cliente tenha um razoável período de maturidade, de acordo com a prática usalmente utilizada na literatura. A Tabela 2.4 resume, para cada Amostra, os principais resultados da análise da base dos dados.

Tabela 2.4: *Definição da variável Target vs População/Amostra*

	Nº	Incumpridor	Indeterminado	Cumpridor	Excluído	Data Início	Data Fim	Total
População	37133	3676	2770	30687	1055	Set-2002	Out-2011	38188
				Amostra1				
Amostra	34215	3635	434	30146	963	Jan-2003	Março-2011	35178
				Amostra2				
Amostra	27872	2804	412	24656	1014	Jan-2006	Março-2011	28699

Na Tabela 2.5 identificam-se as variáveis explicativas que foram consideradas para o desenvolvimento deste estudo.

Tabela 2.5: *Definição das variáveis*

Variáveis	**Descrição**
	Variáveis Socio-Demográficas
Género	Sexo do Cliente
Estado Civil	Estado Civil do cliente
Idade	Idade do Cliente
Habilitações	Habilitações do cliente
Actividade Profissional	Profissão do cliente
Entidade Patronal	Entidade Patronal do cliente
Agência	Localização das Agência
	Variáveis de Relação Cliente Banco
Valor do Empréstimo	Valor emprestado pela instituição financeira
Tipo de Garantia	Garantia apresentada pelo cliente no acto da solicitação de crédito
Prazo	Número de Prestações Mensais
Taxa de Juro	Taxa de juro nominal
Prestações pagas	Nº de prestações pagas pelo cliente na data de extração da base de dados
Valor da Prestação	Valor que o cliente paga em cada prestação

Análise Descritiva das Variáveis Explicativas vs Variável Target

As Tabelas 2.6 e 2.7 ilustram o resultado da categorização das variáveis explicativas utilizadas no estudo. Categorizamos as variáveis por dois motivos: primeiro, para evitar categorias com poucas observações, pois tal pode conduzir a estimativas pouco robustas dos parâmetros associados. O segundo motivo tem a ver com a eliminação de parâmetros desnecessários para o desenvolvimento do modelo, ver [Van Gool et al., 2009], [Thomas, 2000], [Hand e Henley, 1997], [Crook et al., 1992]. As Tabelas 2.6 e 2.7 ilustram a relação entre a variável target e as variáveis explicativas categorizadas pelo método WOE. Os valores indicados nestas tabelas foram calculados a partir da Amostra 1, atavés do nó *Interactive Grouping* (ING) do SAS (versão Enterprise Miner Client6.1), e têm como objectivo fazer uma análise descritiva da carteira.

Tabela 2.6: *Variável Target versus WOE*

Variáveis	**c**ategoria	**G**rupo	**T**otal		**i**ncumpridor		**c**umpridor		**W**OE
			efectivos	%	**e**fectivos	%	**e**fectivos	%	
Nº Prestações Pagas	1	inferior a 18	6822	0,29	371	0,15	6451	0,30572	0,740
	2	entre 18 e 21	3226	0,14	293	0,12	2933	0,14	0,188
	3	entre 22 e 27	4715	0,20	557	0,22	4158	0,2	-0,105
	4	entre 28 e 37	5608	0,24	724	0,28	4884	0,23	-0,207
	5	superior a 37	3274	0,14	599	0,24	2675	0,13	-0,619
Agência	1	18, 25, 17 e 31	587	0,02	8	0,00	579	0,028	2,166
	2	2, 9, 14, 22, 23, 26, 29, 30 e 32	3775	0,16	163	0,06	3612	0,178	0,983
	3	3, 6, 8, 10, 11, 24 e 28	9362	0,40	886	0,35	8476	0,408	0,143
	4	1 e 19	5494	0,23	718	0,28	4776	0,23	-0,221
	5	4, 5, 7 e 12	4427	0,19	769	0,30	3658	0,178	-0,556
Prestações Totais	1	inferior a 24	4176	0,18	412	0,16	3764	0,18	0,097
	2	24 a 42	11459	0,48	1187	0,47	10272	0,47	0,042
	3	superior a 42	8010	0,34	945	0,37	7065	0,33	-0,104
Taxa Nominal	1	inferior a 11,5	2936	0,12	118	0,05	2818	0,13	1,058
	2	superior ou igual 11,5	20709	0,88	2426	0,95	18283	0,87	-0,096
Valor da Prestação	1	inferior a 2.952	1862	0,08	70	0,03	1792	0,085	1,127
	2	2.952 a 5.687	4315	0,18	331	0,13	3984	0,19	0,372
	3	5.687 a 8.908	4424	0,19	462	0,18	3962	0,19	0,033
	4	8.908 a 28.384	11682	0,49	1384	0,54	10298	0,4	-0,109
	5	superior ou igual a 28.384	1362	0,06	297	0,12	1065	0,05	-0,839
Valor do Empréstimo	1	inferior a 97.280	2960	0,13	165	0,06	2795	0,13	0,714
	2	97.280 a 200.000	6084	0,26	533	0,21	5551	0,26	0,228
	3	200.000 a 350.000	5769	0,24	634	0,25	5135	0,24	-0,024
	4	350.000 a 1.000.000	7507	0,32	946	0,37	6561	0,31	-0,179
	5	superior ou igual a 1.000.000	1325	0,06	266	0,10	1059	0,05	-0,734
Idade	1	inferior a 26	1858	0,08	298	0,12	1560	0,07	-0,460
	2	26 a 40	11326	0,48	1376	0,54	9950	0,47	-0,137
	3	40 a 46	4388	0,19	400	0,16	3988	0,19	0,184
	4	superior ou igual a 46	6073	0,26	470	0,18	5603	0,27	0,363
Actividade Profissional	1	Emp. escritório/comércio/serviço e Quadro médio	12280	0,52	923	0,36	11357	0,54	0,394
	2	Estudante, liberal/Quadro superior, outras, Operário especializado e não especializado	9173	0,39	1197	0,47	7976	0,38	-0,219
	3	Actividade desconhecida, Doméstica, Pequenas/mé dias empresas	2192	0,09	424	0,17	1768	0,08	-0,688
Género	1	Masculino	13982	0,59	1649	0,65	12333	0,59	-0,103
	2	Feminino	9663	0,41	895	0,35	8768	0,42	0,166

Tabela 2.7: *Variável Target versus WOE (cont.)*

Variáveis	**c**ategoria	**G**rupo	**T**otal		**I**ncumpridor		**C**umpridor		**W**OE
			efectivos	%	**e**fectivos	%	**e**fectivos	%	
Entidade Patronal	1	Ministérios e aposentados/pensionistas	9728	0,41	591	0,23	9137	0,43	0,623
	2	câmara Municipal, grandes empresas, Instituições financeiras e Institutos públicos	6190	0,26	530	0,21	5660	0,269	0,253
	3	Hotelaria/restuaração e Não declarada	4883	0,21	807	0,32	4076	0,19	-0,496
	4	outras, conta própria e pme's	2844	0,12	616	0,24	2228	0,11	-0,830
Estado Civil	1	solteiro, separado e união de facto	17429	0,74	1963	0,77	15466	0,73	-0,051
	2	casado, viuvo e divrociado	6216	0,26	581	0,23	5635	0,27	0,156
Habilitações	1	Habilitações desconhecidas	4310	0,18	624	0,25	3686	0,17	-0,339
	2	Escolaridade obrigatórias	12445	0,53	1362	0,54	11083	0,53	-0,019
	3	Curso superior e ensino secundário	5608	0,24	475	0,19	5133	0,24	0,265
	4	curso médio e formação superior	1282	0,05	83	0,03	1199	0,06	0,555
Tipo de Garantia	1	Hipoteca sem imóveis para habitação, depositos junto da instituição, outras hipotécas e penhor	1313	0,06	101	0,04	1212	0,06	0,369
	2	outras cauções	16687	0,71	1611	0,63	15076	0,71	0,121
	3	outras entidades	5645	0,24	832	0,33	4813	0,23	-0,360

Observando a Tabela 2.8 pode verificar-se que as três variáveis com maior capacidade preditiva, ou seja, com IV superiores são as mesmas para ambas as amostras. Contudo, verifica-se que a Amostra 2 apresenta melhores resultados na maioria das variáveis.

Tabela 2.8: *Information Value*

Amostra 1				**Amostra 2**		
Ordem	Variável	IV	Tipo	Variável	IV	Tipo
1	Entidade Patronal	0.315	N	Entidade Patronal	0.353	N
2	Agencia	0.25	N	Agencia	0.332	N
3	Nº Prestações Pagas	0.203	I	Nº Prestações Pagas	0.293	I
4	Valor da Prestação	0.148	I	Actividade Profissional	0.144	N
5	Actividade Profissional	0.146	N	Valor da Prestação	0.14	I
6	Valor do Emprestimo	0.111	I	Valor do Emprestimo	0.138	I
7	Taxa Nominal	0.101	I	Taxa Nominal	0.126	I
8	Idade	0.062	I	Idade	0.076	I
9	Tipo de garantia	0.057	N	Tipo de garantia	0.071	N
10	Habilitações	0.053	N	Habilitações	0.052	N
11	Género	0.017	N	Estado Civil	0.021	N
12	Estado Civil	0.012	N	Prazo	0.014	I
13	Prestações Totais	0.006	I	Género	0.011	N

I-Intervalar; N-Nominal

2.5.2 Descrição das Estratégias para a Construção do Modelo

Afim de testar a robustez do modelo da Regressão Logística, e seguidamente escolher um modelo para descrever os determinantes da probabilidade de incumprimento, e estimar a probabilidade de incumprimento, recorrendo às variáveis explicativas da Tabela 2.5, seguimos as seguintes abordagens:

1. **Abordagem 1**
   Construir modelo apenas com variáveis explicativas cujo IV é superior a 0.1, ou seja, com capacidade preditiva média-forte. Nesta abordagem utilizaremos as metodologias:
   - Regressão Logística (variáveis agrupadas pelo ING do Miner 6.1);
   - Scorecard (variáveis categorizadas pelo WOE do ING do Miner 6.1);
2. **Abordagem 2**
   Construir modelo com todas as variáveis explicativas. A utilização de variáveis com valores de IV extremamente baixos tem como objectivo o aumento das variáveis disponíveis para o modelo e, em particular, o aumento das variáveis sócio-demográficas. Note-se que este grupo de variáveis tendencialmente apresenta valores de IV mais baixos.

3. **Abordagem 3**
   Utilização de variáveis não categorizadas. Apesar de, na prática de credit scoring, as variáveis serem frequentemente categorizadas, nesta dissertação optámos por considerar uma abordagem em que tal prática não fosse utilizada. Por um lado, porque pretendemos averiguar o comportamento destas variáveis e, por outro, porque pretendemos também comparar os resultados obtidos com as abordagens anteriores. No caso das variáveis não categorizadas, estima-se apenas um coeficiente para cada uma das variáveis consideradas. Nesta abodagem optámos por não categorizar as variáveis idade, valor de empréstimo e prestações totais. Considera-se esta especificação neste estudo, uma vez que assim é possível testar a existência de uma relação não-linear em relação a essas variáveis quando categorizadas. Tal permite também averiguar a sensibilidade para diferentes níveis de regressores.

2.5.3 Validação do Modelo

Após a construção dos modelos referidos nas Abordagens 1, 2 e 3 é imperativo efectuar a validação desses modelos de forma a optar por aquele que revele melhor performance, de forma que este possa ser considerado como um modelo que identifique as variáveis explicativas e estime adequadamente a probabilidade de incumprimento de um cliente.

Para tal, e de acordo com o usualmente efectuado e recomendado por diversos autores, cada uma das Amostras 1 e 2 foi dividida em Amostra de treino e Amostra de validação, numa proporção de 70% e 30%, respectivamente.

A Tabela 2.9 ilustra a partição efectuada às Amostras.

Tabela 2.9: *Partição da Amostra para Treino e Validação*

	Amostra 1			*Amostra* 2		
	Incumpridor	**Cumpridor**	**Total**	**Incumpridor**	**Cumpridor**	**Total**
Treino	2544	21101	**23645**(70%)	1962	17258	**19220**(70%)
Validação	1091	9045	**10136**(30%)	842	739	**8240**(30%)
Total	**3635**(10,76%)	**30146**(89,24%)	**33781**(100%)	**2804**(10,21%)	**24656**(89,79%)	**27460**(100%)

Para cada uma das Amostras consideradas, foram construídos modelos com base em cada uma das Abordagens 1, 2 e 3. A escolha do modelo que se considerou ser o mais adequado como explicativo da probabilidade de incumprimento do cliente foi efectuda com base em medidas como a curva ROC, estatística de Kolmogorov-Smirnov e o índice de Gini.

Tanto a amostra de treino como na amostra de validação, o valor da curva ROC dos modelos estão próximos de 80% e os valores para o teste KS são maiores que 0.35, valor mínimo considerado para que um modelo apresente bom poder de discriminação, ver [Anderson, 2007].

Tabela 2.10: *Medidas de Desempenho dos Modelos*

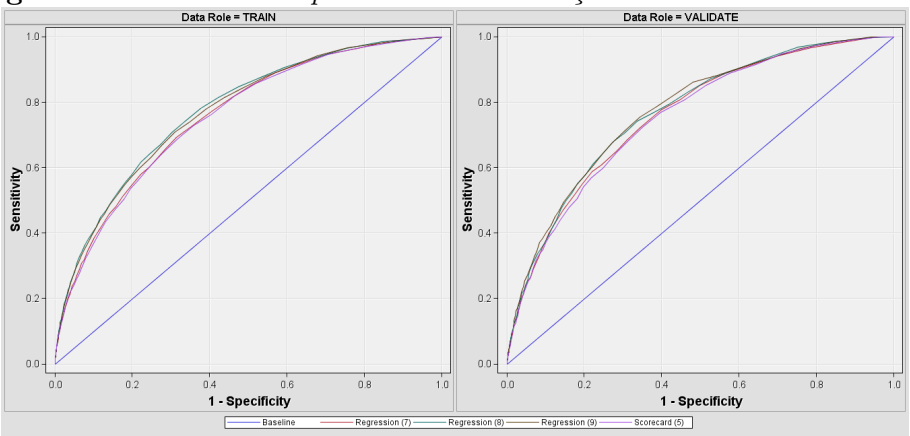
	Curva ROC		Valor de KS		Coeficiente de Gini	
Modelos	Treino	Validação	Treino	Validação	Treino	Validação
			Amostra 1			
Abordagem 3	0.775	0.768	0.406	0.405	0.55	0.536
Abordagem 2	0.769	0.773	0.398	0.409	0.539	0.546
Abordagem 1	0.758	0.759	0.378	0.377	0.516	0.518
Scorecard *	0.754	0.752	0.371	0.371	0.508	0.505
			Amostra 2			
Abordagem 3	0.786	0.761	0.438	0.393	0.573	0.523
Abordagem 2	0.785	0.766	0.427	0.394	0.57	0.531
Abordagem 1	0.775	0.755	0.407	0.366	0.551	0.509
Scorecard *	0.773	0.75	0.404	0.362	0.546	0.5

* Os modelos de Scorecards são aplicações do nó de scorecard do Miner 6.1

Recordamos que a Tabela 2.1 ilustra os valores referência para KS.

Ao analisar os valores das medidas, observa-se que os resultados obtidos para os modelos não variam muito quer na amostra do treino quer na amostra de validação, ver a Tabela 2.10. Essas medidas fornecem um resumo do desempenho geral do modelo em que, de acordo com [Thomas et al., 2002], na maioria das vezes apenas é importante obter uma boa performance em alguns scores fixados, .

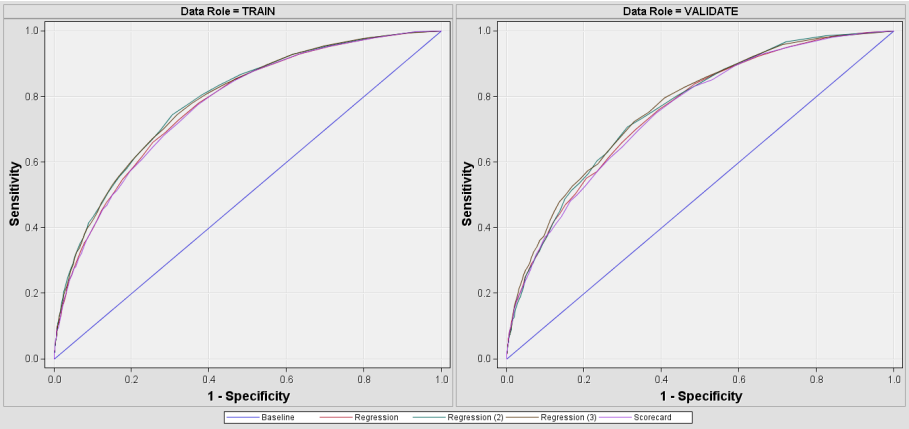
Figura 2.3: *Curva ROC para Treino e Validação dos Modelos. Amostra* 1



Relativamente à curva ROC, pode-se observar que nenhum dos modelos possui uma curva ROC com ordenada superior aos restantes, para todas as possíveis abcissas (ver as Figuras 2.3 e 2.4 da curva ROC, para as Amostras 1 e 2, respectivamente). Como nenhum modelo apresenta desempenho sempre superior aos demais, é interessante analisar outras medidas como o índice de Gini.

A capacidade de discriminação de cada um dos modelos considerados foi analisada também pela matriz de classificação. De acordo com a Tabela 2.11, os modelos foram capazes de clas-

Figura 2.4: *Curva ROC para Treino e Validação dos Modelos. Amostra* 2



sificar correctamente entre 89.5% e 90.24% dos clientes *cumpridores* e entre 55.03% e 63.22% dos *incumpridores* para a Amostra de treino e entre 89.42% e 90.24% dos clientes *cumpridores* e entre 50% e 60.53% dos *incumpridores*, para a Amostra de validação. Enquanto que o poder total de classificação correcta dos modelos está compreendido entre 88.06% e 89.63% para a Amostra de treino e entre 88.67% e 89.60% para a Amostra de validação. Observa-se, também, que todos modelos são mais eficientes em classificar os clientes cumpridores.

Tabela 2.11: *Matriz de Classificação*

	Treino		Validação		Precisão	
Modelos	Cumpridores	Incumpridores	Cumpridores	Incumpridores	Treino	Validação
			Amostra1			
Abordagem 3	89.73	57.09	89.71	51.28	88.79	88.67
Abordagem 2	89.63	59.68	89.55	51.28	88.92	88.86
Abordagem 1	89.57	60.13	89.53	51.35	88.97	88.88
Scorecard*	89.50	62.93	89.42	50	89.06	89
			Amostra2			
Abordagem 3	90.24	55.03	90.24	52.27	89.35	89.27
Abordagem 2	90.13	58.51	90.05	55.1	89.5	89.51
Abordagem 1	90.09	58.97	90.01	60.53	89.54	89.59
Scorecard*	90.03	63.22	90	59.46	89.63	89.60

* Os modelos de Scorecards são aplicações do nó de scorecard do Miner 6.1

Verifica-se que os modelos desenvolvidos pelo scorecard identificam melhor os clientes *incumpridores* e também apresentam melhores resultados na precisão da validação do modelo. Globalmente, os modelos desenvolvidos a partir da Amostra 2 apresentam melhores resultados em relação à Amostra 1. Este facto pode evidenciar a existência de maior consistência entre os dados mais recentes da base de dados, comparativamente com os dados obtidos para um período de tempo mais alargado. Através dos valores da Tabela 2.11 concluí-se que o modelo que melhor discrimina o cliente *cumpridor* do *incumpridor* é o modelo Abordagem 1 construído através da Amostra 2, com base no facto de identificar uma maior percentagem

de clientes *incumpridores*.

A Tabela 2.12 permite ainda verificar que todos os modelos construídos são globalmente significativas à luz do teste de ração de verosimilhança.

Tabela 2.12: *Teste de Razão de Verosimilhança*

Modelos	-2log Intercept Only	Likelihood Intercept and Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
		Amostra 1			
Abordagem 3	16147.17	13744.19	2402.97	50	< .0001
Abordagem 2	16147.17	13858.71	2288.45	31	< .0001
Abordagem 1	16147.17	14055.53	2091.64	22	< .0001
Scorecard	16147.17	14125.81	2021.35	7	< .0001
		Amostra 2			
Abordagem 3	12671.05	10699.34	1972.01	13	< .0001
Abordagem 2	12671.05	10671.3	1999.74	26	< .0001
Abordagem 1	12671.05	10819.56	1851.5	22	< .0001
Scorecard	12671.05	10857.95	20818.02	7	< .0001

2.5.4 Estimação da Probabilidade de Incumprimento

Nesta subsecção e, à luz das estratégias anteriormente propostas, optámos por analisar um modelo de Regressão Logística com maior número de variáveis explicativas, utilizando a Amostra 2 e a Abordagem 2, ou seja os dados mais recentes da base de dados e explicativas. A razão da escolha fundamenta-se na tese de que os dados mais recentes podem descrever melhor os dados futuros.

A Tabela 2.13 ilustra para cada variável, as estimativas dos coeficientes, os respectivos desvios padrão, as estatísticas de Wald, os graus de liberdade e os níveis descritivos dos testes de significância do modelo adoptado.

Para ambos os modelos constatamos, na Tabela 2.12, que a maioria dos coeficientes de todas as variáveis incluídas são estatisticamente diferentes de zero e que a maioria apresenta sinais e estimativas de acordo com o que seria expectável. Assim, de acordo com os níveis descritivos do teste, todas as variáveis consideradas são relevantes para a discriminação dos clientes "cumpridores" e "incumpridores".

Neste modelo é de destacar que as variáveis idade, habilitações literárias e género que, apesar de terem um valor preditivo fraco, de acordo com o IV, foram selecionadas pelo método de stepwise da Regressão Logística. É de referir que a idade é uma variável utilizada na maioria dos modelos de risco de crédito por ser considerada um indicador da etapa de ciclo de vida do cliente.

Seguidamente faremos uma análise detalhada das variáveis identificadas como explicativas.

Tabela 2.13: *Resultados da Estimação por Máxima Verosimilhança (Amostra 2 - Abordagem 2)*

					Teste de Wald	
Parâmetros	Grupo	DF	$\hat{\beta}$	Desvio−padrão σ	Qui−quadrado	Pr>ChiSq
Intercept		1	-2.7695	0.1042	707.03	<.0001
Nº Prestações Pagas	1	1	-1.0252	0.0564	330.15	<.0001
Nº Prestações Pagas	2	1	-0.4586	0.0646	50.41	<.0001
Nº Prestações Pagas	3	1	0.2476	0.0492	25.34	<.0001
Nº Prestações Pagas	4	1	0.3449	0.0465	55.12	<.0001
Agencia	1	1	-1.5168	0.2884	27.66	<.0001
Agencia	2	1	-0.3619	0.0988	13.40	0.0003
Agencia	3	1	0.3081	0.0807	14.57	0.0001
Agencia	4	1	0.7766	0.0823	88.95	<.0001
Prestações Totais	1	1	0.4677	0.0684	46.78	<.0001
Prestações Totais	2	1	-0.0964	0.0405	5.67	0.0172
Taxa Nominal	1	1	-0.4960	0.0756	43.05	<.0001
Valor da Prestação	1	1	-0.6576	0.1558	17.81	<.0001
Valor da Prestação	2	1	-0.2818	0.0741	14.48	0.0001
Valor da Prestação	3	1	-0.0604	0.0642	0.88	0.3471
Valor da Prestação	4	1	0.2213	0.0679	10.61	0.0011
Valor de Emprestimo	1	1	0.4374	0.1408	9.65	0.0019
Valor de Emprestimo	2	1	-0.1117	0.0674	2.74	0.0976
Valor de Emprestimo	3	1	-0.1329	0.0548	5.88	0.0153
Valor de Emprestimo	4	1	-0.1425	0.0681	4.38	0.0364
Idade	1	1	0.3611	0.0828	19.01	<.0001
Idade	2	1	0.1938	0.0391	24.57	<.0001
Idade	3	1	-0.1510	0.0516	8.56	0.0034
Actividade Profissional	1	1	-0.2761	0.0391	49.80	<.0001
Actividade Profissional	2	1	-0.0421	0.0329	1.64	0.2000
Género	1	1	0.1135	0.0237	22.96	<.0001
Entidade Patronal	1	1	-0.6762	0.0402	283.29	<.0001
Entidade Patronal	2	1	-0.3726	0.0413	81.31	<.0001
Entidade Patronal	3	1	0.3317	0.0415	63.83	<.0001
Habilitações	1	1	0.1503	0.0537	7.82	0.0052
Habilitações	2	1	0.0136	0.0425	0.10	0.7479
Habilitações	3	1	-0.0532	0.0511	1.08	0.2981

Género: é uma variável binária: esta variável tem um coeficiente estimado positivo, o que significa que, um cliente do sexo masculino tem uma maior probabilidade de ser incumpridor, quando comparado com um do sexo feminino, com todas as outras variáveis mantidas constantes. [Dinh e Kleimeier, 2007] incluiram, inicialmente, a variável Género decidindo depois com o auxílio de procedimentos de selecção de variáveis se essa variável deverá ou não permanecer no modelo de credit scoring final.

Número de Prestações Pagas: A interpretação do coeficiente associado a esta variável é interessante uma vez que há uma mudança do sinal nos coeficientes; clientes com menor número de prestações pagas, até 21 prestações, são menos propensos a entrar em *incumprimento* em comparação com os clientes com mais de 37 prestações pagas. Por outro lado, clientes com 22 a 37 prestações pagas são mais propensos quando comparados com clientes com mais de 37 prestações pagas, mantendo tudo o resto constante.

Valor da Prestação: Relativamente a esta variável verifica-se que há mudanças de sinal nos coeficientes estimados. Assim, os clientes que pagam até 5.687 ECV têm menor probabilidade de pagarem as suas prestações, quando comparados com os clientes com valor de prestação superior ou igual a 28.384 ECV. Por conseguinte, os que pagam mensalmente entre 8.908 ECV e 28.384 ECV são mais propensos ao incumprimento do que os que pagam mensalmente uma prestação igual ou superior a 28.384 ECV, mantendo tudo o resto constante. Sendo esta variável definida em função do montante e duração do empréstimo, poderá estar a revelar alguma informação acerca do nível socio-económico do cliente: se a prestação concedida é alta deve significar que o cliente tem boas condições económicas (caso contrário o empréstimo teria sido recusado), o que faz com que a probabilidade de incumprimento seja menor.

Prestações Totais: A mudança de sinal apresentada nos coeficientes das categorias desta variável evidencia que os clientes com menos de 24 meses são mais propensos a entrar em *incumprimento* do que aqueles com mais de 42 meses de prestações pagas. Os clientes com prestações pagas entre os 24 e 42 são os melhores clientes para a carteira, apresentando uma menor probabilidade de incumprimento, mantendo tudo o resto constante. Isto significa que clientes com prestações totais maiores têm maior porbabilidade de entrar em incumprimento durante o perído de empréstimo.

Entidade Patronal: Nesta variável, os clientes que possuem um vínculo laboral com a entidade patronal como: ministérios, câmara municipais, grandes empresas, aposentados/pensionistas, instituições financeiras e institutos públicos possuem menor probabilidade de serem "incumpridores" quando comparados com trabalhadores de outras entidades, trabalhadores de conta própria e pme's. No entanto, os que possuem como entidade patronal a hotelaria/restauração e entidades não declaradas possuem maior probabilidade de serem "incumpridores" em relação aos que trabalham nas entidades conta própria, outras entidades e pme's, mantendo tudo o resto constante.

Actividade profissional: Os clientes do banco em estudo, que exercem como profissão funcionários de escritório, serviços, comércio e quadros médios são menos propensos a *incumprimento* do que os Estudantes, Liberais/Quadros Superiores, Outras, Operários Especializados e Não Especializados Contudo, todas as actividades referidas são menos propensas ao incumprimento do que as Domésticas, Actividades Desconhecidas e funcionários de pequenas e Médias Empresas, mantendo tudo o resto constante. É importante referir que, no entanto, a categoria que representa os Estudantes, Liberais/Quadros Superiores, Outras, Operários Especializados e Não Especializados não se revelou estatísticamente significativa a 5%.

Idade: A mudança de sinal apresentada nos coeficientes das categorias desta variável evidencia que os clientes com menos de 40 anos são mais propensos a entrar em *incumprimento* do que aqueles com 46 ou mais anos de idade. Os clientes com idade entre os 40 e 46 são

os melhores clientes para a carteira, apresentando uma menor probabilidade de deixar de cumprir com o pagamento do seu empréstimo, mantendo tudo o resto constante. O resultado está em consonância com a observação de que os mais velhos, com idade superior a 46 anos, têm maior estabilidade financeira e logo, possuem maior potencial para reembolsar, dentro do prazo, os seus empréstimos, em comparação com os mais novos (inferiores a 40 anos).

Taxa Nominal: Esta variável, com duas categorias (taxa inferior a 11.5% e superior a 11.5%), captura o efeito da taxa nominal aplicada pelo banco no momento do empréstimo. Esta variável tem um coeficiente estimado negativo, o que indica que os clientes que possuem um empréstimo com maior taxa possuem maior probabilidade de entrarem em incumprimento, o que quer dizer que poderão não vir a ser bons clientes para o banco, mantendo todo o resto constante.

Valor de Empréstimo: O valor que o banco empresta aos clientes foi agrupado em cinco categorias. A categoria dos clientes que solicitam um crédito entre 200.000 e 350.000 ECV é a única que não se revelou ser significativa. É ainda de notar uma mudança no sinal dos coeficientes. Esta mudança significa que os clientes que contraem empréstimos menores são mais propensos a serem "incumpridores", comparativamente com os que requerem uma maior quantia. Verifica-se ainda que os que solicitam valores entre 200.000 e 1.000.000 ECV são menos proponsos a serem "incumpridores" do que aqueles que contraiem empréstimos de valor superior a 1.000.000 ECV. Os melhores clientes para esta carteira, segundo esta variável, são os que possuem um crédito com valor entre 200.000 e 1.000.000 ECV, mantendo tudo o resto constante.

Agência: Esta variável contribui para a inferência das agências que apresentam maior propensão dos clientes a serem *incumpridores*. A Agência por si só não deverá ser um factor determinante na probabilidade de incumprimento. No entanto, a Agência está relacionadas com diferentes zonas de residência, o que poderá trazer alguma informação estatística acerca do fenómeno de incumprimento. É de nota que todas as categorias da variável são significativas. Estima-se, assim, que os clientes das agências dos Grupos 1 e 2 têm menor probabilidade de incumprimento, comparando com os do Grupo 5. Contudo, os clientes dos Grupos 3 e 4 têm maior probabilidade de incumprimento, mantendo tudo o resto constante. Isto poderá indiciar que certas zonas de residência contêm pessoas com diferentes níveis socio-económicos.

Habilitações: Esta variável categórica apresenta somente uma categoria que se revela ser significativo. Em termos da probabilidade de incumprimento, significa que clientes com habilitações desconhecidas têm maior probabilidade de incumprimento comparando com clientes com um curso médio ou formação profissional, mantendo tudo o resto constante.

Tendo por objectivo a construção de um modelo com boa capacidade preditiva, optou-se

por reduzir o modelo anterior, retirando todas as variáveis explicativas que, embora sendo estatisticamente significativas, revelaram uma capacidade preditiva fraca, ou seja, para as quais se obteve um valor de IV inferior a 0.1.

Sob esta perspectiva, foram retirados do modelo anterior as variáveis Prestações Totais, Idade, Género e Habilitações, o que equivale a estimar o modelo com base na Abordagem 1. Estimaremos o modelo utilizando Amostra 2.

Os resultados obtidos encontram-se ilustrados na Tabela 2.14

Tabela 2.14: *Resultados da Estimação por Máxima Verosimilhança (Amostra 2 - Abordagem 1)*

					Teste de Wald	
Parâmetros	Grupo	DF	$\hat{\beta}$	Desvio−padrão	Qui−quadrado	Pr>ChiSq
Intercept		1	-3.114	0.0982	1006.22	<.0001
Nº Prestações pagas	1	1	-1.006	0.0561	321.48	<.0001
Nº Prestações pagas	2	1	-0.2581	0.0623	17.19	<.0001
Nº Prestações pagas	3	1	0.1861	0.0480	15.03	0.0001
Nº Prestações pagas	4	1	0.3275	0.0482	46.13	<.0001
Agência	1	1	-1.487	0.3076	23.37	<.0001
Agência	2	1	-0.5406	0.1135	22.67	<.0001
Agência	3	1	0.3570	0.0876	16.60	<.0001
Agência	4	1	0.7922	0.0890	79.25	<.0001
Taxa Nominal	1	1	-0.5103	0.0820	38.76	<.0001
Valor da Prestação	1	1	-0.5812	0.1326	19.21	<.0001
Valor da Prestação	2	1	-0.1590	0.0626	6.46	0.0110
Valor da Prestação	4	1	0.5005	0.0706	50.21	<.0001
Valor do Emprestimo	1	1	0.5294	0.1218	18.88	<.0001
Valor do Emprestimo	3	1	-0.2271	0.0584	15.11	0.0001
Valor do Emprestimo	4	1	-0.3112	0.0671	21.52	<.0001
Actividade Profissional	1	1	-0.1624	0.0293	30.82	<.0001
Entidade Patronal	1	1	-0.7585	0.0462	268.98	<.0001
Entidade Patronal	2	1	-0.3734	0.0475	61.68	<.0001
Entidade Patronal	3	1	0.3374	0.0452	55.71	<.0001

A estimativa da probabilidade de incumprimento de cada cliente, que designaremos por p_i, obtém-se, como visto anteriormente, utilizando a expressão (2.10).

A Tabela 2.15 ilustra a estimativa da probabilidade de incumprimento de sete clientes para cada uma das Abordagens utilizadas.

Tabela 2.15: *Probabilidade de default para os clientes (exemplos)*

Amostra 2 Abordagem 1

	Características							
Cliente	V. crédito	V. Prest	Act. Profissional	Ent. Patronal	Tx Nom	Agência	Prest. Pagas	$p_i(A1)$
27799	1.500.000	33.747	Serviços	PME's	12.5	1	18	**0.186**
32017	12.000	4.014	Serviços	Outras	12.5	1	36	**0.279**
33921	100.000	6.121	Serviços	Grandes Empresas	12.5	2	18	**0.03**
33955	94.695	5.731	Serviços	Ministérios	11	3	18	**0.014**
34008	404.000	13.515	Outros	Ministérios	12.5	8	21	**0.04**
34017	360.000	12.043	Serviços	PME's	12.5	1	21	**0.216**
34023	1.200.000	31.896	Outros	Não Declarada	12.5	11	30	**0.26**

Amostra 2 - Abordagem 2

	Características											
Cliente	V. crédito	V. Prest	Act. Profissional	Ent. Patronal	Tx Nom	Agência	Prest. Pagas	Prest. Totais	Género	Habilitações	Idade	$p_i(A2)$
27799	1.500.000	33.747	Serviços	PME's	12.5	1	18	60	F	H. Desconhecidas	50	**0.116**
32017	12.000	4.014	Serviços	Outras	12.5	1	36	36	M	H Obrigatória	41	**0.234**
33921	100.000	6.121	Serviços	Grandes Empresas	12.5	2	18	19	M	H Desconhecidas	40	**0.05**
33955	94.695	5.731	Serviços	Ministérios	11	3	18	18	F	H Obrigatória	50	**0.018**
34008	404.000	13.515	Outros	Ministérios	12.5	8	21	37	F	H Desconhecidas	37	**0.057**
34017	360.000	12.043	Serviços	PME's	12.5	1	21	37	M	H Obrigatória	34	**0.22**
34023	1.200.000	31.896	Outros	Não Declarada	12.5	11	30	48	M	H Obrigatória	48	**0.145**

$p_i(A1)$ - Probabilidade de incumprimento Amostra 2 Abordagem 1

$p_i(A2)$ - Probabilidade de incumprimento Amostra 2 Abordagem 2

As diferenças nas estimativas apresentadas para cada uma das Abordagens consideradas reflectem que alguma ou várias das variáveis retiradas do modelo revelam ter um peso significativo nas estimativas efectuadas.

O modelo com todas as variáveis significativas revela-se um modelo mais prudente do ponto de vista da instituição bancária, uma vez que sobrevaloriza as probabilidades de incumprimento. Por outro lado, pode revelar-se um modelo menos competitivo, uma vez que se traduzirá num *spread* superior e, consequentemente numa prestação mais elevada.

O modelo que contém apenas as variáveis com IV mais elevado estimam menores probabilidades de incumprimento sendo, como visto atrás (ver Tabela 2.11), o modelo que melhor identifica os clientes incumpridores.

As diferenças observadas nas estimativas das probabilidades de incumprimento indiciam que deve haver uma análise cuidada, sob diversos pontos de vista da instituição bancária, antes da escolha do modelo final.

2.6 Considerações Finais, Limitações e Estudos Futuros

Considerações Finais

Tem havido um enorme interesse nas últimas décadas na utilização de credit scoring para avaliar risco de crédito no sector bancário. Num ambiente competitivo para as instituições financeiras, particularmente para os bancos, as técnicas de credit scoring tornaram-se, actualmente uma das ferramentas mais importantes utilizadas na avaliação do risco de crédito dos empréstimos.

Com este estudo pretendeu-se desenvolver um modelo de credit scoring para a gestão de crédito ao consumo, num banco de Cabo Verde, através de uma técnica estatística multivariada, a Regressão Logística, identificando as variáveis determinantes da probabilidade de incumprimento e estimar a probabilidade de incumprimento para cada cliente da carteira.

A utilização da Amostra com dados mais recentes (Amostra entre Janeiro de 2006 e Março de 2011) permite, por um lado, comparar com os clientes mais antigos, que supostamente, na data da extração da base de dados, não fazem parte da carteira e, por outro, fazer um modelo com clientes mais recentes, em que a maioria ainda faz parte da carteira. É de referir, também, que dados mais recentes contêm informação que reflete a conjectura económica mais actual, o que permite aferir com mais precisão o comportamento futuro dos clientes. Os resultados ilustram que quando os dados são mais actuais obtemos melhores resultados, pelo que optámos por um modelo com essas características.

As Amostras 1 e 2 foram aleatoriamente divididas de modo que a amostra de treino contivesse

70% e a amostra de validação os restantes 30%, conforme descrito na Tabela 2.9. O resultado do estudo permite concluir que a Regressão Logística discrimina bem os clientes *cumpridores* e *incumpridores* em ambas as amostras utilizadas.

A capacidade de discriminação dos modelos foi analisada também pela matriz de classificação. Os modelos foram capazes de classificar entre 89.5% e 90.24% dos clientes "cumpridores" e entre 55.03% e 63.22% dos "incumpridores" para a amostra do treino e 89.42% e 90.24% dos clientes "cumpridores" e entre 50% e 60.53% dos "incumpridores" para a amostra de validação. Enquanto que, o poder total de classificação dos modelos estão compreendidos entre 88.06% e 89.63% para a amostra do treino e entre 88.67% e 89.60% para a amostra de validação. Observa-se também, que todos modelos são mais eficientes em classificar os clientes "cumpridores".

Observando as estimativas dos modelos ajustados verifica-se que a maioria dos coeficientes apresentam sinais em conformidade com o que seria espectável, sendo os coeficientes com sinais negativos os que correspondem às variáveis que menos contribuem para a propensão de um cliente ser "incumpridor".

A probabilidade de incumprimento estimada da carteira é de 10.8% e 10.2% na Amostra 1 e Amostra 2, respectivamente, de acordo com os dados da Tabela 2.9.

Os resultados obtidos neste capítulo, nomeadamente a probabilidade de incumprimento da carteira e a probabilidade de incumprimento de cada cliente, voltarão a ser utilizados e referidos nos Capítulos 3 e 4, respectivamente.

Limitações e Estudos Futuros

A base de dados utilizada neste estudo contém somente informações dos clientes a quem foram concedidos empréstimos e não contém informações sobre os candidatos rejeitados, ou seja, sobre clientes que solicitaram crédito, e cujo pedido foi indeferido. Assim, como um dos estudos futuros propomos a recolha dessa informação para elaboração dos modelos com esses clientes. No estudo de [Banasik et al., 2003] foi comparada a precisão das classificações realizadas por um modelo cuja estimação se baseou somente em candidatos aceites e as obtidas por um modelo estimado tendo por base uma amostra de todos os candidatos e o resultado do estudo mostrou que as diferenças foram mínimas. Por outro lado, [Hand e Henley, 1997] analisaram um processo de inferência dos rejeitados, ou seja, um processo de tentar inferir o verdadeiro *status* de crédito dos candidatos rejeitados. Os autores concluíram que uma inferência confiável na rejeição é impossível e as melhorias nos modelos de scoring alcançadas por inferência dos rejeitados são meros acasos. No entanto, à luz da realidade da população de Cabo Verde, consideramos que este estudo poderia e deveria ser realizado pois, poder-se-ão obter conclusões diferentes das dos autores referidos.

Também propomos a utilização de outras técnicas abordadas na revisão bibliográfica para

comparações, tais como para modelos híbridos. Por exemplo, no estudo de [Semedo, 2010] comparou-se a Regressão Logística com Redes Neuronais e conclui que não há evidencia estatística a 95% de confiança para afirmar que as redes neuronais são preferíveis ao modelo logit (ou vice versa). Neste estudo foram incluídos os rejeitados na amostra analisada.

Capítulo 3

Estimação da Evolução Temporal da Probabilidade de Incumprimento

3.1 Introdução

O risco que cada cliente representa para a instituição bancária deve ser avaliado e estimado *a priori*, aquando da concessão do crédito, usando técnicas adequadas como as desenvolvidas e aplicadas no Capítulo 2 pois, como já referido, este risco deve reflectir-se no *spread* que o cliente deverá pagar ao banco para que este assuma o risco de empréstimo.

No entanto, o risco que o cliente representa para a instituição bancária não é constante ao longo do tempo. Pelas mais diversas razões, muitas delas difíceis de observar e mensurar, os clientes podem, em qualquer instante do seu contrato, entrar em incumprimento, deixando de efectuar os pagamentos devidos ou começando a ter, sistematicamente, prestações em atraso. Estas situações devem ser medidas e acauteladas pela instituição bancária pois têm, naturalmente, custos implícitos. Por um lado, o pagamento não atempado das prestações implica uma cobrança de juros desadequeada de acordo com os termos do contrato, o que se pode traduzir numa perda monetária para a instituição, uma vez que se um cliente deixa de cumprir com as suas prestações, o banco terá que suportar custos judiciais de forma a recuperar o montante em dívida. Por outro lado, as condições socio-económicas do cliente, analisadas aquando da concessão do crédito, podem alterar-se ao longo do tempo, fazendo com que este possa deixar de cumprir com os pagamentos devidos, representando um risco superior ao inicialmente estimado.

Neste capítulo, com base na informação da carteira de clientes de crédito ao consumo já

referida e estudada no Capítulo 2, estimar-se-ão as probabilidades de incumprimento ao longo do tempo, com base num modelo de Markov para populações abertas, obtendo assim estimativas para as probabilidades de incumprimento da carteira num horizonte temporal de longo prazo o que, como veremos adiante, se pode obter em situações de estacionaridade da carteira.

3.2 Revisão da Literatura

A modelação de um fenómeno recorrendo aos modelos usuais de Cadeias de Markov pode ter algumas limitações quando o fluxo de entradas para a população em estudo é preponderante na evolução da população. Por exemplo, nos modelos usuais de cadeia de Markov, um dos pressupostos envolvidos considera que a carteira em estudo é fechada. Esta é uma limitação forte, mas necessária para que se garantam condições de estacionaridade a longo prazo. Essas limitações, para além de reflectir hipóteses pouco realistas do ponto de vista das aplicações, traduzir-se-ão necessariamente em diferentes estimativas no que diz respeito à lei de probabilidade da cadeia em cada instante e numa perspectiva de longo prazo, conforme referido por diversos autores, ver por exemplo [Vassiliou, 1998]. Em trabalhos como [Centeno e Silva, 2001] e [Guerreiro e Mexia, 2004], é possível observar aplicações onde se evidenciam diferenças significativas nas estimativas a longo prazo.

Para um melhor entendimento das limitações em causa, faremos uma pequena exposição acerca do estudo de populações recorrendo a modelos tradicionais de cadeias de Markov.

3.2.1 Abordagem segundo Cadeias de Markov

Considere-se um modelo populacional, modelado através de uma cadeia de Markov com espaço dos estados $E = \{1, 2, \ldots, k\}$ definida por uma distribuição inicial $\boldsymbol{c}^T = (c_1, c_2, \ldots, c_k)$ e uma matriz de transição $\boldsymbol{P} = [p_{ij}]_{(i,j)\in\{i,\ldots,k\}^2}$.

Após uma primeira transição, a proporção de elementos da população num dado estado i é a proporção dos que permanecem no estado i, adicionado à percentagem dos que transitam dos estados 2 a k para o estado i, isto é:

$$c_1\, p_{1i} + c_2\, p_{2i} + \ldots + c_i\, p_{ii} + \ldots + c_k\, p_{ki} = \sum_{j=1}^{k} c_j\, p_{ji}$$

e, assim, os valores das proporções em todos os estados, após uma transição, (a lei de probabilidade da cadeia após um passo), podem ser obtidos, como é sabido da teoria das cadeias de Markov, a partir de $\boldsymbol{c}^T\boldsymbol{P}$ e, após n transições, a lei de probabilidade da cadeia

será obtida a partir de

$$\boldsymbol{p}_n^T = \boldsymbol{c}^T \boldsymbol{P}^{(n)} \quad , \quad n \in \mathbb{N}.$$

com $\boldsymbol{P}^{(0)} = \boldsymbol{I}$, $\boldsymbol{P}^{(1)} = \boldsymbol{P}$ e, por indução matemática, $\boldsymbol{P}^{(n+1)} = \boldsymbol{P} \cdot \boldsymbol{P}^{(n)}$.

Na evolução da lei de probabilidade da cadeia assume-se que a cadeia é fechada, não se admitindo entradas de novos elementos ao longo do tempo, limitação forte do ponto de vista prático, mas necessária para o estudo da estacionaridade a longo prazo.

A análise assimptótica dos modelos usuais de cadeias de Markov de parâmetro discreto é sobejamente conhecida e estudada pelo que não nos alongaremos nessa exposição. Para mais detalhes consulte-se, por exemplo, [Parzen, 1965] ou [Ross, 1996].

Suponhamos agora que, recorrendo a cadeias de Markov, pretendemos analisar a evolução do número de indivíduos em cada estado da cadeia admitindo que, a cada data i, $i \in \{1, \ldots, n\}$, um número aleatório N_i de novos elementos entra na população.

Simultaneamente com a entrada do segundo grupo ocorre a primeira transição do primeiro grupo, sendo estes indivíduos "conduzidos" de acordo com as probabilidades de transição da cadeia de Markov. Em instantes posteriores, o processo repetir-se-á, admitindo-se novos elementos e reclassificando-se os existentes.

Na Tabela 3.1 ilustra-se este processo de contagem.

Tabela 3.1: *Contagem para n grupos*

Data	1	2	3	...	$n-1$	n
1	$N_1\mathbf{c}_1^T$	$N_1\mathbf{c}_1^T\mathbf{P}$	$N_1\mathbf{c}_1^T\mathbf{P}^{(2)}$	...	$N_1\mathbf{c}_1^T\mathbf{P}^{(n-2)}$	$N_1\mathbf{c}_1^T\mathbf{P}^{(n-1)}$
2	–	$N_2\mathbf{c}_2^T$	$N_2\mathbf{c}_2^T\mathbf{P}$	...	$N_2\mathbf{c}_2^T\mathbf{P}^{(n-3)}$	$N_2\mathbf{c}_2^T\mathbf{P}^{(n-2)}$
3	–	–	$N_3\mathbf{c}_3^T$	...	$N_3\mathbf{c}_3^T\mathbf{P}^{(n-4)}$	$N_3\mathbf{c}_3^T\mathbf{P}^{(n-3)}$
...	...	...	...	...	...	...
n	–	–	–	–	–	$N_n\mathbf{c}_n^T$

Após n períodos de tempo, o número total de elementos em cada estado da cadeia pode ser obtido a partir de

$$\begin{aligned}\mathbf{S}_n &= N_n\mathbf{c}_n^T + N_{n-1}\mathbf{c}_{n-1}^T\mathbf{P} + \cdots + N_2\mathbf{c}_2^T\mathbf{P}^{(n-2)} + N_1\mathbf{c}_1^T\mathbf{P}^{(n-1)} = \\ &= \sum_{k=1}^{n} N_k\mathbf{c}_k^T\,\mathbf{P}^{(n-k)} \,.\end{aligned}$$

Se, adicionalmente, se considerar que $N_1, \ldots, N_n$ são variáveis aleatórias com valores esperados $n_1, \ldots, n_n$, o valor esperado do número total de elementos em cada estado pode ser

obtido a partir de

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{S}_n\right] &= n_n\mathbf{c}_n^T + n_{n-1}\mathbf{c}_{n-1}^T\mathbf{P} + \cdots + n_2\mathbf{c}_2^T\mathbf{P}^{(n-2)} + n_1\mathbf{c}_1^T\mathbf{P}^{(n-1)} = \\ &= \sum_{k=1}^{n} n_k\mathbf{c}_k^T\ \mathbf{P}^{(n-k)}\ .\end{aligned} \tag{3.1}$$

mas, sem hipóteses adicionais, pouco mais pode ser dito.

Nesta formulação ficam em aberto questões como o comportamento assimptótico de $\mathbb{E}[\boldsymbol{S}_n]$, sobre o impacto das variáveis $N_1, \ldots, N_n$ na evolução de $\mathbb{E}[\boldsymbol{S}_n]$ e qual o grau de confiança das estimativas obtidas.

3.2.2 Abordagens segundo Modelos de Markov para populações abertas

De uma forma breve, nesta secção, faz-se uma revisão de alguns dos estudos acerca de populações abertas assentes em modelos de Markov.

Em [Gani, 1963], foi desenvolvido um modelo para estimar o número de matrículas de alunos e o número de graus concedidos em universidades australianas. A expressão aí obtida é semelhante à expressão (3.1), no caso especial de uma matriz com apenas duas diagonais não nulas. Este trabalho foi posteriormente estendido em [Stadje, 1999], considerando um processo estacionário de entradas não necessariamente independentes, obtendo como resultado a convergência da distribuição conjunta das dimensões dos estados e a determinação da distribuição assimptótica da transformada de Laplace correspondente.

[Stadje, 1999] observa ainda que, considerando as entradas para a população como uma sequência de variáveis aleatórias independentes e identicamente distribuídas, o vector aleatório do número de indivíduos em cada classe pode ser descrito por uma cadeia de Markov homogénea. [Staff e Vagholkar, 1971] consideram que o número de novas entradas segue uma distribuição geométrica e utilizam a distribuição Multinomial para a classificação inicial dos novos elementos. Nestas condições, e tendo por base um modelo de Markov, obtêm a distribuição estacionária do número de indivíduos em cada estado, recorrendo às funções geradoras de probabilidades.

[Pollard, 1967] considera que as entradas de novos elementos seguem uma distribuição de Poisson e, após mencionar a importância do papel da distribuição Multinomial na classificação inicial dos elementos, obtém a função característica conjunta do número de elementos em cada classe, reconhecendo estas variáveis como tendo distribuição de Poisson e sendo independentes entre si. J. H. Pollard desenvolveu ainda outros trabalhos relativamente ao estudo da evolução de populações abertas, ver [Pollard, 1966], [Pollard, 1969], [Pollard, 9679] e [Pollard e Sherris, 1980].

Pode ainda referir-se [Bartholomew, 1982] como uma referência bibliográfica que aborda

modelos para populações abertas, englobando alguns dos modelos de outros autores citados atrás.

Mais recentemente, [Yakasai, 2005], baseando-se no trabalho de Stadje, restringe a entrada de novos elementos a um único estado e procura identificar a distribuição estacionária das entradas na população de tal forma que a capacidade da população em estudo não seja excedida. Em [Centeno e Silva, 2001] é desenvolvido um modelo de cadeias de Markov não homogéneas para populações abertas, com o intuito de estimar a distribuição estacionária para um sistema de *Bonus Malus* numa carteira aberta.

Também motivados pela análise de um sistema de *Bonus Malus* em carteiras abertas, [Guerreiro e Mexia, 2004] apresentaram um dos estudos mais recentes acerca de modelos de Markov para populações abertas. Mais tarde, ver [Guerreiro e Mexia, 2008], os autores apresentaram a análise estocástica do modelo para populações com estruturas mais complexas, considerando cadeias com diversas classes de comunicação transientes e vários estados recorrentes. Desenvolvimentos posteriores deste modelo, ao nível do estudo estatístico das dimensões relativas e absolutas das classes da cadeia foram obtidos após a consideração de um número aleatório de novas entradas. Estes desenvolvimentos permitem a determinação de regiões de confiança para as estimativas obtidas a partir do modelo aberto para cadeias de Markov e podem ser vistos em [Guerreiro et al., 2010], [Guerreiro et al., 2012a] e [Guerreiro et al., 2012b].

3.3 Modelo Vórtices Estocásticos

Nesta dissertação, para análise da probabilidade de incumprimento da carteira de crédito ao consumo, numa perspectiva de longo prazo, utilizar-se-á um modelo para populações abertas sujeitas a reclassificações periódicas designado por Vórtices Estocásticos, a partir do qual se poderão obter estimativas pontuais e por intervalo de confiança para o número total e proporção de clientes nas diferentes classes de risco, em qualquer instante do tempo.

O modelo Vórtices Estocásticos foi inicialmente proposto por [Mexia, 2000] e assenta numa formulação que tem por base um modelo de cadeias de Markov, mas considerando populações abertas. Este modelo começou por ser desenvolvido nos estudos de [Guerreiro, 2001] e [Guerreiro e Mexia, 2004] como uma solução alternativa para a análise de Sistemas de *Bonus Malus* em sede de Seguro Automóvel, sob a perspectiva de uma carteira aberta, tendo sido aplicado a dados de uma Seguradora Portuguesa e, posteriormente, e no mesmo âmbito, aplicado a dados de uma Seguradora Cabo-verdiana em [Rodrigues, 2011].

Desenvolvimentos posteriores, ver [Guerreiro e Mexia, 2008] e [Guerreiro, 2008], visam con-

templar o estudo de populações com estrutura mais complexa, tendo sido aí obtidos resultados ao nível da estrutura estocástica das populações e estimação da evolução das dimensões das mesmas. O modelo tem continuado a ser desenvolvido, ver [Guerreiro et al., 2010], [Guerreiro et al., 2012a] e [Guerreiro et al., 2012b], tendo-se obtido resultados ao nível da inferência estatística para as dimensões da população e suas sub-populações.

De uma forma resumida e geral, o modelo dos Vórtices Estocásticos aplica-se a populações abertas, divididas em sub-populações, em que os elementos que a compõem são sujeitos a reclassificações periódicas, em instantes igualmente espaçados no tempo. A base do modelo que aqui apresentaremos é uma cadeia de Markov finita, homogénea, com s estados transientes e 1 estado absorvente, correspondente às saídas da população[1].

Em [Guerreiro e Mexia, 2008] e [Guerreiro, 2008] apresenta-se a estrutura estocástica para as populações e mostra-se como estimar o número e proporção de elementos em cada sub-população, numa perspectiva de curto, médio e longo prazo, considerando a possibilidade de vários estados transientes (que podem constituir mais do que uma classe de comunicação) e mais do que um estado recorrente.

No que se segue, definir-se-á dimensão absoluta e dimensão relativa de uma sub-população como o número total e a proporção, respectivamente, de elementos que, num dado instante, integram essa sub-população. Assim, a dimensão relativa é calculada como o quociente entre o número de elementos nessa sub-população e o número de elementos que compõem a totalidade da população (e que ocupam, portanto, uma qualquer das suas sub-populações). Quando a dimensão relativa de um conjunto de sub-populações estabiliza, ou seja, converge para um valor finito, dir-se-á que essas sub-populações estão integradas num vórtice estocástico. Esse vórtice estocástico estará sediado num conjunto maximal de estados para os quais as dimensões relativas são estáveis, ver [Guerreiro e Mexia, 2008] ou [Guerreiro, 2008].

A entrada de novos elementos na população, que se assume seguir uma distribuição de Poisson, está sujeita a uma classificação inicial, a qual determina a sub-população onde cada novo elemento irá ser inicialmente colocado. Esta sub-população será o ponto de partida e, após um período de tempo, o elemento será sujeito a uma reclassificação podendo manter a classificação anterior ficando, portanto, no mesmo estado da cadeia ou transitar para outra sub-população, de acordo a avaliação que dele é feita nesse instante.

Com base nos trabalhos desenvolvidos por [Guerreiro e Mexia, 2008], [Guerreiro et al., 2010] e [Guerreiro et al., 2012b], o modelo do Vórtices Estocásticos permite obter as estimações e previsões para as dimensões absolutas e relativas das sub-populações. Esta estimação pode ser efectuada pontualmente ou por intervalo de confiança, para um dado período de tempo,

[1]Nesta dissertação utilizar-se-á a versão mais simples do modelo, pois é a que melhor corresponde à estrutura das classes de risco da carteira de crédito que se irá analisar. Para populações com estrutura mais complexa, ver [Guerreiro e Mexia, 2008] ou [Guerreiro, 2008].

ou numa perspectiva de longo prazo.

Os fluxos de entrada na população têm, como veremos adiante, um papel preponderante na forma como estas irão evoluir ao longo do tempo e na possibilidade de existência de uma estabilidade assimptótica ao nível das dimensões absolutas e/ou relativas das sub-populações, ver, por exemplo, [Guerreiro, 2008] ou [Guerreiro et al., 2010]. Nos trabalhos até aqui desenvolvidos no âmbito dos Vórtices Estocásticos, os fluxos de entrada para a população têm sido modelados considerando que, em cada instante i, o número esperado de novos elementos que integram a população pode ser estimado a partir de $\lambda_i = a + b\theta^i$, $i \in \mathbb{N}$ e (a, b, θ) pertencentes a um espaço paramétrico que definiremos adiante. Esta forma funcional para o número esperado de novas entradas em cada instante contempla diversos tipos de fluxos de entrada. Note-se, tal como referido em [Guerreiro, 2008] ou [Guerreiro et al., 2010] que, por exemplo, diferentes valores para os parâmetros se traduzem em diferentes comportamentos para os fluxos de entrada:

- Se $b = 0$ ou $\theta = 1$, o fluxo de novas entradas é constante ao longo do tempo;
- Se $a = 0$, o número esperado de novas entradas evolui em progressão geométrica de razão θ;
- Se $a = -b$ e $\theta = e^{-\alpha}$, obter-se-á um número esperado de novas entradas com evolução assimptótica com fluxo de entradas crescente;
- Se $a = b$ e $\theta = e^{-\alpha}$, obter-se-á um número esperado de novas entradas com um evolução assimptótica com fluxo de entradas decrescente.

Uma das contribuições desta dissertação prende-se com a generalização da forma funcional que modela os fluxos de entrada na população. Após a obtenção de um resultado geral que caracteriza as condições sob as quais se garante a existência de estabilidade a longo prazo nas dimensões relativas das sub-populações, e consequente existência de um vórtice estocástico nessas sub-populações, centrar-nos-emos numa forma funcional não explorada anteriormente, e que se traduz num melhor ajustamento aos dados da carteira em estudo. A introdução de uma nova modelação dos fluxos de entrada introduz a necessidade de obtenção de estimadores de máxima verosimilhança para os parâmetros dos fluxos de entrada e de novas regiões de confiança para as dimensões absolutas e relativas das sub-populações.

Nesta secção, começaremos por estudar a estrutura da cadeia de Markov, com base nos trabalhos de [Guerreiro, 2001] e [Guerreiro e Mexia, 2004]. Dada a estrutura das classes de risco da carteira de crédito, debruçar-nos-emos sobre os vórtices estocásticos com suporte nos estados transientes. Obteremos resultados relativos à evolução, a longo prazo, da dimensão esperada das sub-populações. No que se refere aos fluxos de entrada na população, apresentaremos alguns resultados já obtidos em [Guerreiro, 2008], [Guerreiro et al., 2010] e

[Guerreiro et al., 2012b], bem como os resultados desenvolvidos no âmbito desta dissertação, identificando os casos em que a estabilidade a longo prazo se encontra definida, sob as hipóteses consideradas para as intensidades de entrada.

Apresentaremos, por fim, os resultados da aplicação do modelo a uma carteira de crédito ao consumo, estimando a evolução de cada classe de risco em termos assimptóticos. Para efeitos comparativos, utilizaremos duas possíveis modelações para os fluxos de entrada.

3.3.1 Estrutura da População

Consideremos que a carteira de crédito é constituída por s classes de risco, correspondendo a s estados transientes de uma cadeia de Markov. Considerando um estado adicional $s+1$, absorvente, correspondente às saídas da carteira, podemos facilmente verificar que a matriz de transição num passo pode ser definida por:

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{K} & \boldsymbol{q} \\ \mathbf{0} & 1 \end{bmatrix}$$

onde:

$\boldsymbol{K}$ - matriz de dimensão $s \times s$, cujos elementos representam as probabilidades de transição entre as classes transientes (as classes de risco);

$\boldsymbol{q}$ - vector coluna, de dimensão $s \times 1$, cujos elementos representam as saídas dos clientes das classes transientes.

É de notar que:

- a última linha da matriz representa o estado de saída dos clientes da carteira.
- a soma dos elementos de cada linha da matriz $\boldsymbol{K}$ com a correspondente componente do vector $\boldsymbol{q}$, será sempre igual a 1: $\sum_{i=1}^{s} k_{ij} + q_i = 1$.

Considere-se o seguinte lema:

Lema 3.1. *A matriz probabilidade de transição em n passos será da forma:*

$$\boldsymbol{P^n} = \begin{bmatrix} \boldsymbol{K}^n & \boldsymbol{q}_n \\ \mathbf{0} & 1 \end{bmatrix}$$

em que: $\boldsymbol{q}_n = \sum_{i=1}^{n-1} \boldsymbol{K}^i \boldsymbol{q}$, $n \in \mathbb{N}$

A demostração do Lema 3.1, e outros que caracterizam a matriz de probabilidades de transição pode ser consultada em [Guerreiro, 2001] ou [Guerreiro e Mexia, 2004].

3.3.2 Fluxos de Entrada nas Populações

Sejam E_i, $i \in \mathbb{N}$, o número total de novos clientes aos quais é concedido um crédito, no período i.

Admitiremos que o número total de entradas para a carteira, em cada período, ocorre no início desse período e, sem perda de generalidade, tomaremos os períodos como meses. Estas assumpções correspondem, na prática, a considerar que os novos contratos de crédito concedidos em cada mês são integrados na carteira no início desses mês.

Consideraremos, adicionalmente, que o número de novos clientes que integram a carteira no mês i segue uma distribuição de Poisson de valor esperado λ_i, ou seja,

$$E_i \sim \mathcal{P}(\lambda_i).$$

Nos estudos até aqui desenvolvidos, tal como referido anteriormente, tem sido considerado que a intensidade de entrada de novos clientes, λ_i, é modelada pela seguinte expressão:

$$\lambda_i = a + b\,\theta^i \ , \ (a, b, \theta) \in \Theta_1 \ , \ i \in \mathbb{N} \tag{3.2}$$

em que o espaço de parâmetros Θ_1, correspondente aos possíveis valores de a, b e θ, é definido por:

$$\Theta_1 = \big\{\{(a, b, \theta) : a, \theta \in \mathbb{R}^+, b \in \mathbb{R}, a + b\theta > 0\} \backslash \{(b, \theta) \in \mathbb{R}^- \times]1, +\infty[\}\big\}. \tag{3.3}$$

A adequabilidade destes fluxos de entrada aos dados da carteira pode ser testada recorrendo a vários testes de hipóteses que poderão ser consultados de forma detalhada em [Guerreiro, 2008] ou [Guerreiro et al., 2010].

É de salientar que os fluxos de entrada modelados por (3.2) e alguns dos seus casos particulares foram já aplicados nos estudos de [Guerreiro, 2001], [Guerreiro e Mexia, 2004], [Guerreiro e Mexia, 2008], [Guerreiro, 2008], [Guerreiro et al., 2010], [Guerreiro et al., 2012b] e [Rodrigues, 2011].

Ao longo desta dissertação, e tendo em conta a natureza dos dados da carteira em estudo, considerou-se uma nova formulação para o número médio de novas entradas em cada mês i, dada por:

$$\lambda_i = (a + b\, e^{-\theta i})^{-1} \ , \ (a, b, \theta) \in \Theta_2 \ , \ i \in \mathbb{N} \tag{3.4}$$

em que o espaço de parâmetros Θ_2, correspondente aos possíveis valores de a, b e θ é definido por:

$$\Theta_2 = \big\{\{(a, b, \theta) : a \in \mathbb{R}^+, b, \theta \in \mathbb{R}, a + b\, e^{-\theta i} > 0\} \backslash \{(a, b, \theta) : a + b\, e^{-\theta i} = 0 \ , \ i \in \mathbb{N}\}\big\}. \tag{3.5}$$

A modelação da evolução da carteira recorrendo a novos fluxos de entrada deixa em aberto a análise da estabilidade da carteira a longo prazo, ou seja, a análise da existência de Vórtices Estocásticos nos estados transientes da cadeia de Markov. Como veremos adiante, esta formulação pertencerá a uma classe de funções para as quais se garante a existência de Vórtices Estocásticos na cadeia, o que nos permitirá estimar a proporção de clientes em cada classe de risco, numa perspectiva de longo prazo.

A adequabilidade do fluxo de entradas considero pode ser testada recorrendo ao teste de hipóteses desenvolvido na secção 3.3.10.

3.3.3 Classificação Inicial

Tal como exposto anteriormente, consideraremos que o número de novos clientes que são integrados na carteira de crédito no mês i segue uma distribuição de Poisson de parâmetro λ_i.

Estes novos clientes serão distribuídos, à data de entrada, pelas classes de risco da carteira, de acordo com a classificação deles é inicialmente feita em termos de risco.

Seja $\boldsymbol{c}_i$, $i \in \mathbb{N}$ o vector de classificação inicial para o mês i, definido como:

$$\boldsymbol{c_i^T} = \big[\boldsymbol{t_i^T}|0\big] \tag{3.6}$$

em que:

- $\boldsymbol{t}_i$ representa um vector cujas componentes são as probabilidade de entrada de um novo cliente em cada uma das classes de risco;
- a última componente indica que a probabilidade de um novo cliente ser imediatamente colocado no estado de saída é nula.

Faz-se notar que o modelo permite que sejam considerados vectores de probabilidade distintos em cada mês i, tomando

$$\boldsymbol{c_i} = \boldsymbol{c} + \boldsymbol{\omega}\gamma^i = [c_{ij}] \ , \ j = 1, \ldots, s \ , \ 0 < \gamma < 1 \tag{3.7}$$

em que:

- $\boldsymbol{c}$ - vector fixo de probabilidades de classificação inicial cujas componentes verificam $\sum_{j=1}^{s} c_j = 1$;
- $\boldsymbol{\omega}$ - vector fixo cujas componentes verificam $\sum_{j=1}^{s} \omega_j = 0$.

Para efeitos assimptóticos, é importante notar que a formulação (3.7) se traduz no seguinte $\lim_{i \to +\infty} \boldsymbol{c}_i = \boldsymbol{c}$, o que equivale a considerar um vector de classificação inicial convergente.

A formulação (3.7) é, naturalmente, extensível ao vector $\boldsymbol{t}_i$, ver (3.6).

3.3.4 Amostragem Aleatória

Nesta secção, descrevemos de forma suscinta a ideia geral da amostragem aleatória que, para mais detalhes, pode ser consultada, por exemplo em [Feller, 1968, p. 216].

Suponhamos que um conjunto de N elementos, com N uma variável aleatória com distribuição de Poisson de parâmetros λ, são distribuídos por s classes, de acordo com a distribuição Multinomial com vector de probabilidades $\mathbf{c} = (c_1, \ldots, c_s)$.

Após a distribuição, os números N_j , $j = 1, \ldots, s$, de elementos na classe j, serão variáveis aleatórias independentes com distribuição de Poisson com parâmetros $\lambda c_1, \ldots, \lambda c_s$. Este resultado, que resume a ideia da amostragem aleatória, pode ser formalmente escrito como se segue.

Sejam $N \frown \mathcal{P}(\lambda)$ e $\mathbf{X} = (X_1, \ldots, X_s) \frown \mathcal{M}(N, \mathbf{c})$, isto é, se com $n = n_1 + \cdots + n_s$,

$$\mathbb{P}\left[X_1 = n_1, X_2 = n_2, \ldots, X_r = n_s \mid N = n\right] = \frac{n!}{n_1! \cdots n_s!} c_1^{n_1} \cdots c_s^{n_s},$$

então

$$\begin{aligned}\mathbb{P}\left[X_1 = n_1, \ldots, X_s = n_s \ , \ N = n\right] &= \mathbb{P}\left[X_1 = n_1, \ldots, X_s = n_s \mid N = n\right] \cdot \mathbb{P}\left[N = n\right] = \\ &= \frac{n!}{n_1! \cdots n_s!} c_1^{n_1} \cdots c_s^{n_s} e^{-\lambda} \frac{\lambda^n}{n!} = \frac{(c_1\lambda)^{n_1}}{n_1!} e^{-\lambda n_1} \times \cdots \times \frac{(c_s\lambda)^{n_s}}{n_s!} e^{-\lambda n_s},\end{aligned}$$

que corresponde ao produto das funções de probabilidade de variáveis aleatórias com distribuição de Poisson de parâmetros $\lambda c_1, \ldots, \lambda c_s$.

De alguma forma, este resultado é notável e especial. De facto, duas conclusões podem ser retiradas deste fenómeno:

1. Se o número de elementos numa população seguir uma distribuição de Poisson de parâmetro λ e for multinomialmente distribuído por um número fixo de classes, de acordo com um vector de probabilidades $\boldsymbol{c}$, então os números de elementos em cada classe serão v.a.'s aleatórias independentes com distribuição de Poisson de parâmetro λc_i , $i = 1, \ldots, s$.

2. Se uma população, com um número aleatório de elementos, é distribuída por um número fixo de classes de acordo com a distribuição multinomial, de tal forma que os números de indivíduos nas várias classes são v.a.'s independentes, então a população inicial terá necessariamente distribuição de Poisson.

Para completar as considerações acima, apresentaremos o seguinte teorema, baseado em [Dacunha-Castelle et al., 1970], e uma prova deste resultado inspirado num exercício apresentado em [Dacunha-Castelle et al., 1970].

Teorema 3.1. *Seja uma população de um número aleatório de N indivíduos a serem distribuídos por um conjunto de s classes, as sub-populações $\mathbf{X} = (X_1, \ldots, X_s)$, de acordo com a distribuição multinomial $\mathcal{M}(N, \mathbf{c})$ com um vector de probabilidade $\mathbf{c} = (c_1, \ldots, c_s)$. Isto é, $\mathbf{X} \frown \mathcal{M}(N, \mathbf{c})$. Isto significa que para $n, n_1, \ldots, n_s$ inteiros tais que $n_1 + \cdots + n_s = n$, se tem:*

$$\mathbb{P}\left[X_1 = n_1, X_2 = n_2, \ldots, X_s = n_s \mid N = n\right] = \frac{n!}{n_1! \cdots n_s!} c_1^{n_1} \cdots c_s^{n_s} \ .$$

Supondo que $\mathbb{E}[N] < +\infty$, pode afirmar-se que se as leis das sub-populações são independentes, então $N \frown \mathcal{P}(\lambda)$, para algum λ, isto é, a população inicial segue uma distribuição de Poisson.

Demonstração: Primeiro, suponhamos que $s = 2$ e assim $\mathbf{X} = (X_1, X_2) \frown \mathcal{M}(N, (c_1, c_2))$. Designando por N a v.a. da dimensão da população e por X_1 e X_2 as v.a.'s referentes à dimensão das sub-populações, tem-se que $X_1 + X_2 = N$. Assim, neste caso, a distribuição da população inicial pelas duas classes é binomial. Considere $B_1, B_2, \ldots, B_N, \ldots$ uma sequência de variáveis aleatórias identicamente distribuídas com B, uma variável aleatória de Bernoulli, com parâmetro c_1 e descrita por

$$X_1 = B_1 + B_2 + \cdots + B_N \quad \text{e} \quad X_2 = N - X_1 = (1 - B_1) + (1 - B_2) + \cdots + (1 - B_N) \ . \tag{3.8}$$

Consideremos φ_{X_1}, φ_{X_2}, φ_N e φ_B as funções geradoras de momentos das variáveis aleatórias X_1, X_2, N e B, respectivamente. Condicionando em N, e usando a representação na expressão (3.8), obteremos:

$$\forall t_1 \in \mathbb{R} \quad \varphi_{X_1}(t_1) = \varphi_N\left(\log(\varphi_B(t_1))\right) \ . \tag{3.9}$$

Da mesma forma, também podemos verificar que:

$$\forall t_2 \in \mathbb{R} \quad \varphi_{X_2}(t_2) = \varphi_N\left(\log(e^{t_2}\varphi_B(-t_2))\right) \ , \tag{3.10}$$

e novamente, também, para a lei conjunta do par (X_1, X_2), ter-se-á

$$\forall t_1, t_2 \in \mathbb{R} \quad \varphi_{(X_1,X_2)}(t_1, t_2) = \varphi_N \left(\log(e^{t_2}\varphi_B(t_1 - t_2)\right) . \tag{3.11}$$

Consideremos agora a função geradora de probabilidade da v.a. N:

$$g(z) = \mathbb{E}\left[z^N\right] = \sum_{k=0}^{+\infty} \mathbb{P}\left[N = k\right] \cdot z^k.$$

Tem-se então que

$$g(1) = \sum_{k=0}^{+\infty} \mathbb{P}\left[N = k\right] = 1 \, , \tag{3.12}$$

pelo que o lema clássico de Abel, para séries de potências, nos garante que g é uma função analítica no domínio $\{z \in \mathbb{C} : |z| < 1\}$. Além disso, pelo teorema de Abel para séries de potências, a expressão (3.12) implica que g é contínua em $[0, 1]$.

Provemos agora que, se suposermos que X_1 e X_2 são v.a.'s independentes, então g é uma função exponencial e, como consequência, N é uma v.a. com distribuição de Poisson.

Observemos, em primeiro lugar que, para t_1 e t_2 pertencentes a uma pequena vizinhança de zero, como:

$$\varphi_B(t_1) = e^{t_1}p_1 + (1 - p_1),$$

$$e^{t_2}\varphi_B(-t_2) = p_1 + (1 - p_1)e^{t_2}$$

e

$$e^{t_2}\varphi_B(t_1 - t_2) = e^{t_1}p_1 + (1 - p_1)e^{t_2}$$

se X_1 e X_2 são v.a. independentes, isto é, se $\varphi_{(X_1,X_2)}(t_1, t_2) = \varphi_{X_1}(t_1) \cdot \varphi_{X_2}(t_2)$ então, de acordo com as expressões (3.9), (3.10) e (3.11), temos que:

$$g(e^{t_1}p_1 + (1 - p_1)e^{t_2}) = g(e^{t_1}p_1 + (1 - p_1)) \cdot g(p_1 + (1 - p_1)e^{t_2}) \, .$$

Notemos que a expressão anterior pode ser reescrita como uma equação funcional do tipo $g(x + y - 1) = g(x) \cdot g(y)$, fazendo $x = e^{t_1}p_1 + (1 - p_1)$ e $y = p_1 + (1 - p_1)e^{t_2}$. Como g é uma função analitica em $\{z \in \mathbb{C} : |z| < 1\}$, derivando esta equação funcional em relação a y, obtemos

$$g'(x + y - 1) = g(x) \cdot g'(y).$$

Desta forma, para $\{z \in \mathbb{C} : |z| < 1\}$, tem-se que

$$g'(z) = \sum_{k=1}^{+\infty} k\mathbb{P}\left[N = k\right] z^{k-1}.$$

Como $\mathbb{E}[N] < +\infty$, recorrendo novamente ao teorema do Abel para séries de potências, podemos afirmar, para $y = 1$, que

$$g'(x) = g(x) \cdot g'(1). \tag{3.13}$$

A equação funcional, escrita na forma (3.13), leva-nos a concluir que

$$g(x) = c_0 \, e^{\lambda x}.$$

Como $g(1) = 1$ teremos ainda que

$$g(x) = e^{\lambda(x-1)}.$$

Para terminar, uma vez que

$$\varphi_N(t) = g(e^t)$$

ter-se-á então que

$$\varphi_N(t) = e^{\lambda(e^t-1)} \ ,$$

o que coincide com a função geradora de momentos de uma v.a. com distribuição de Poisson.

Como a função geradora de momentos determina univocamente uma distribuição, podemos afirmar que N tem distribuição de Poisson de parâmetro λ.

Consideremos agora o caso geral de uma população dividida num qualquer número s de classes. Tal como exposto atrás, a população distribuir-se-á pelas s classses de acordo com a distribuição Multinomial.

Denotemos por $X_1, \ldots, X_s$ o número de elementos em cada uma das s classes. Consideremos agora uma "super classe" constituída pelas $s-1$ primeiras classes. O número total de elementos nessa "super classe" será, naturalmente, $X_1 + \ldots + X_{s-1}$.

Se as v.a.'s X_i, $i = 1, \ldots, s$, forem independentes, então o número total de elementos na "super classe" e o número de elementos na classe s são ainda v.a.'s independentes.

Como a distribuição da população pelas s classes se efectua de acordo com uma lei de probabilidade Multinomial então, considerando apenas a distribuição dos elementos entre a "super classe" e a classe s, podemos afirmar que esta distribuição se faz de acordo com uma lei de probabilidade Binomial.

Aplicando o raciocínio efectuado atrás para o caso particular de $s = 2$, podemos concluir que a lei de probabilidade de N é Poisson. □

3.3.5 Evolução da Dimensão das Sub-Populações

Nesta secção, seguindo os trabalhos de Guerreiro et al. e ainda os de Pollard e Staff, anteriormente citados, iremos mostrar como aplicar o conceito de amostragem aleatória para estimar a evolução das populações abertas sujeitas a reclassificações periódicas, considerando que estas populações são modeladas por uma cadeia de Markov.

Começaremos por enunciar um Teorema, inspirado no Teorema 3.1 de [Guerreiro et al., 2010]. Este resultado é de extrema importância no modelo que iremos aplicar pois permite-nos estimar os parâmetros das leis de Poisson das classes da população, em qualquer data n , $n \in \mathbb{N}$.

Teorema 3.2. *Considere-se um sistema com s classes. Suponha-se que:*

1. *Em cada data $i \in \{0, 1, \ldots, T\}$, entram na população N_i elementos, com N_i seguindo uma distribuição de Poisson de parâmetro λ_i, e são distribuídos pelas classes, de acordo com uma lei Multinomial $\mathcal{M}(N_i, \mathbf{c}_i)$, com $\mathbf{c}_i^T = (c_{i1}, \ldots, c_{is})$;*
2. *Em cada data $i \in \{0, 1, \ldots, T\}$, e simultaneamente com a entrada de novos elementos no sistema, os elementos de cada classe evoluem de acordo com a lei de uma cadeia de Markov com matriz de transição $\mathbf{P} = [p_{ij}]_{(i,j)\in\{1,\ldots,s\}^2}$. Esta evolução traduz as probabilidades de reclassificação dos elementos pertencentes à população.*

À data $n \in \{0, 1, \ldots, T\}$, após o grupo de novos elementos, com N_n elementos, ter entrado no sistema e a população de cada classe ter sido reclassificada, de acordo com a lei da cadeia de Markov correspondente ao número de reclassificações efectuadas até à data n, a dimensão da população na classe $j \in \{1, \ldots, s\}$ seguirá uma distribuição de Poisson com parâmetro dado por

$$\sum_{i=1}^{n} \lambda_i \mathbf{c}_i^T \boldsymbol{\delta}_j^T (\mathbf{P}^{(n-i)})^T \,, \tag{3.14}$$

com $\boldsymbol{\delta}_j^T = (\delta_{j1}, \delta_{j2}, \ldots, \delta_{js})$ em que δ_{ji} é o delta de Kronecker.

Demonstração: Começaremos por descrever a ideia geral implícita na demonstração deste resultado.

Na data 1, o primeiro grupo é distribuído por cada uma das s classes, de acordo com a distribuição multinomial. Pelo resultado da secção 3.3.4 a população na classe $j \in \{1, \ldots, s\}$ terá uma distribuição de Poisson com parâmetro $\lambda_1 c_{1j}$.

À data 2, ocorrerá uma reclassificação dos indivíduos que já se encontram no sistema. Essa reclassificação será conduzida pela matriz de transição num passo da cadeia de Markov e, assim, a população da classe $j \in \{1, \ldots, s\}$, que tem distribuição de Poisson de parâmetro $\lambda_1 c_{1j}$, será redistribuída pelas classes do sistema, de acordo com uma lei multinomial com

vector de probabilidades $(p_{j1}, p_{j2}, \ldots, p_{js})$ que corresponde, naturalmente, à j-ésima linha da matriz de probabilidades de transição.

Consequentemente, após essa reclassificação, a população na classe $j \in \{1, \ldots, s\}$ terá ainda distribuição de Poisson, agora com parâmetro dado por

$$\lambda_1 c_{11}\, p_{1j} + \lambda_1 c_{12}\, p_{2j} + \cdots + \lambda_1 c_{1s}\, p_{sj}\ .$$

Observamos que, com $\boldsymbol{\delta}_j^T = (\delta_{j1}, \delta_{j2}, \ldots, \delta_{js})$, tem-se que

$$\lambda_1 \mathbf{c}_1^T \boldsymbol{\delta}_j^T \mathbf{P}^T = \lambda_1 c_{11}\, p_{1j} + \lambda_1 c_{12}\, p_{2j} + \cdots + \lambda_1 c_{1s}\, p_{sj}.$$

No entanto, na data 2, chegará à população um segundo grupo de N_2 novos elementos, independente do primeiro, e seguindo uma distribuição de Poisson de parâmetro λ_2. Este segundo grupo é distribuído pelas s classes do sistema, de acordo com a lei Multinomial $\mathcal{M}(N_2, \mathbf{c}_2)$. Após a incorporação dos elementos do segundo grupo, e dada a reprodutibilidade da distribuição de Poisson, podemos afirmar que o número total de elementos na classe $j \in \{1, \ldots, s\}$ terá uma distribuição de Poisson com parâmetro dado por:

$$\lambda_1 \mathbf{c}_1^T \boldsymbol{\delta}_j^T \mathbf{P}^T + \lambda_2 \mathbf{c}_2^T \boldsymbol{\delta}_j^T \mathbf{I}\,.$$

Novamente, à data 3, ocorrerá uma reclassificalção de cada elemento da população, de acordo com a matriz de transição da cadeia de Markov. Tal como na primeira reclassificação, para estimar a dimensão de cada classe $j \in \{1, \ldots, s\}$, será necessário ter em conta os elementos que estão nas restantes classes e que serão multinomialmente distribuídos com vector de probabilidades $\boldsymbol{\delta}_j^T \mathbf{P}^T$, que corresponde à j ésima linha para a matriz de transição transposta.

Facilmente se conclui que o número de elementos que pertencerão à classe j terão distribuição de Poisson com parâmetro

$$\left(\lambda_1 \mathbf{c}_1^T \boldsymbol{\delta}_j^T \mathbf{P}^T + \lambda_2 \mathbf{c}_2^T \boldsymbol{\delta}_j^T \mathbf{I}\right) \cdot \boldsymbol{\delta}_j^T \mathbf{P}^T = \lambda_1 \mathbf{c}_1^T \boldsymbol{\delta}_j^T (\mathbf{P}^{(2)})^T + \lambda_2 \mathbf{c}_2^T \boldsymbol{\delta}_j^T \mathbf{P}^T\ .$$

No entanto, novamente, à data 3, um grupo de N_3 novos elementos, seguindo uma distribuição de Poisson de parâmetro λ_3, independente de N_1 e N_2, dará entrada na população e será inicialmente classificado, ou seja, multinomialmente distribuído de acordo com $\mathcal{M}(N_3, \mathbf{c}_3)$.

Após as entradas destes elementos na classe $j \in \{1, \ldots, s\}$, novamente devido à reprodutibilidade da distribuição de Poisson, terá ainda distribuição de Poisson, agora com parâmetro dado por

$$\lambda_1 \mathbf{c}_1^T \boldsymbol{\delta}_j^T (\mathbf{P}^{(2)})^T + \lambda_2 \mathbf{c}_2^T \boldsymbol{\delta}_j^T \mathbf{P}^T + \lambda_3 \mathbf{c}_3^T \boldsymbol{\delta}_j^T \mathbf{I}\,.$$

Por indução matemática, é facilmente demonstrado o resultado da expressão (3.14). □

Observação 3.1. *Pode verificar-se que um vector com componentes dadas por* (3.14) *pode também ser escrito como*

$$\boldsymbol{\lambda}_n^{++T} = \sum_{i=1}^{n} \lambda_i \mathbf{c}_i^T \mathbf{P}^{(n-i)} \tag{3.15}$$

e, consequentemente, as expressões (3.1) *e* (3.14) *descrevem vectores idênticos.*

A principal contribuição deste teorema, estabelecido nos estudos de [Guerreiro, 2008] e [Guerreiro et al., 2010], prende-se com o facto de podermos concluir que as dimensões das populações presentes em cada classe do sistema, em cada data $i \in \{0, 1, \ldots, T\}$, seguem distribuições de Poisson independentes com parâmetros conhecidos.

Este resultado permitir-nos-á estimar a dimensão esperada de cada uma das sub-populações correspondentes às classes de risco da carteira, em qualquer instante de tempo. Poderemos, ainda, desenvolver técnicas estatísticas adequadas à estimação destes parâmetros, nomeadamente ao nível da estimação por intervalo de confiança e o desenvolvimento de testes de hipóteses.

Tendo em conta a expressão (3.1), estabelecemos ainda a seguinte proposição, cuja demonstração pode consultar-se em [Guerreiro, 2001] ou [Guerreiro e Mexia, 2004]:

Proposição 3.1. *Após n períodos de tempo, o vector médio do número de clientes nas sub-populações será dado por:*

$$\boldsymbol{\lambda}_n^{++T} = \left(\sum_{i=1}^{n} \lambda_i \boldsymbol{t}_i^T \boldsymbol{K}^{n-i} \mid \sum_{i=1}^{n} \lambda_i \boldsymbol{t}_i^T \boldsymbol{q}_{n-1} \right) \tag{3.16}$$

sendo a primeira componente do vector, que designaremos por $\boldsymbol{\lambda}_n^{+}$, a dimensão estimada das sub-populações correspondentes aos estados transientes (as classes de risco) e a segunda componente do vector correspondente ao número estimado de clientes que saíram da carteira de crédito.

3.3.6 A Estabilidade das Sub-Populações conduzidas por um modelo de Markov aberto

Nesta secção centrar-nos-emos na análise assimptótica das dimensões das sub-populações que constituirão as classes de risco da carteira, ou seja, os estados transientes da cadeia de Markov.

A estabilização das dimensões relativas das sub-populações, apesar dos fluxos de entrada, saída e reclassificações dos clientes, caracteriza-se pela existência de um vórtice estocástico estabelecido nos estados transientes.

Nos resultados que se seguem, e uma vez que a análise assimptótica da população se irá restringir ao estudo dos estados transientes, incidiremos a análise assimptótica da população na matriz $\boldsymbol{K}$ dos estados transientes, ver (3.1).

No que se segue, e como primeira hipótese restritiva, supõe-se que a matriz de transição entre estados trasientes, $\boldsymbol{K}$, é diagonalizável e, como tal, ver [Schott, 1997]:

$$\boldsymbol{K} = \sum_{j=1}^{s} \eta_j \boldsymbol{\alpha}_j \boldsymbol{\beta}_j^T \ ,$$

com $\{\eta_j \ , \ j = 1, \ldots, s\}$ os valores próprios da matriz $\boldsymbol{K}$, $\{\boldsymbol{\alpha}_j \ , \ j = 1, \ldots, s\}$ os vectores próprios à esquerda e $\{\boldsymbol{\beta}_j \ , \ j = 1, \ldots, s\}$ os vectores próprios à direita da matriz $\boldsymbol{K}$.

Notamos que $j \in \{1, \ldots, s\}$ corresponde a um estado transiente se e só se $\mid \eta_j \mid < 1$.

Podemos ainda verificar que, ver novamente [Schott, 1997],

$$\boldsymbol{K}^n = \sum_{j=1}^{s} \eta_j^n \boldsymbol{\alpha}_j \boldsymbol{\beta}_j^T \ , \tag{3.17}$$

e, como consequência da expressão (3.16), para o vector médio das dimensões das sub-populações correspondentes aos estados transientes, $\boldsymbol{\lambda}_n^{+T}$, tem-se:

$$\boldsymbol{\lambda}_n^{+T} := \sum_{i=1}^{n} \lambda_i \mathbf{t}_i^T \boldsymbol{K}^{(n-i)} = \sum_{j=1}^{s} \sum_{i=1}^{n} \lambda_i \eta_j^{n-i} \mathbf{t}_i^T \boldsymbol{\alpha}_j \boldsymbol{\beta}_j^T \ . \tag{3.18}$$

Como segunda hipótese geral, supõe-se que o vector de probabilidades de classificação inicial é constante, ou seja, para $i \geq 1$, $\boldsymbol{t}_i = \boldsymbol{t}_0 \neq \mathbf{0}$.

Desta forma, a expressão (3.18) pode reescrever-se como

$$\boldsymbol{\lambda}_n^{+T} = \sum_{j=1}^{s} \left(\sum_{k=1}^{n} \lambda_k \eta_j^{n-k} \right) \mathbf{t}_0^T \boldsymbol{\alpha}_j \boldsymbol{\beta}_j^T \ . \tag{3.19}$$

Esta hipótese é apenas colocada para efeitos de simplicidade dos cálculos que se seguem. Veremos que os resultados obtidos podem ser estendidos ao caso geral em que o vector de classificação inicial pode não ser constante, mas convergente para um vector não nulo, ou seja

$$\lim_{i \to +\infty} \mathbf{t}_i = \mathbf{t}_\infty \neq \mathbf{0}$$

como havíamos já suposto na subsecção 3.3.3.

Notamos ainda que na aplicação que iremos efectuar para análise da carteira de crédito, o vector de classificação inicial será constante em todos os períodos de tempo.

Os dois pontos seguintes introduzem novos resultados ao nível da convergência do modelo dos Vórtices Estocásticos. Os resultados aqui obtidos generalizam os anteriormente estudados nos trabalhos de Guerreiro et al.

O caso de uma sequência convergente de parâmetros

Numa primeira fase, analisamos o caso em que os parâmetros das v.a.'s de Poisson, referentes ao número de novas entradas na população, corresponde a uma sucessão convergente.

De acordo com a expresssão (3.19), o comportamento assimptótico de $\boldsymbol{\lambda}_n^+$ é determinado pelo comportamento assimptótico de

$$I_n = \sum_{i=1}^{n} \lambda_i \eta^{n-i} = \lambda_1 \eta^{n-1} + \cdots + \lambda_{n-1}\eta + \lambda_n \ , \tag{3.20}$$

com $\mid \eta \mid < 1$.

O facto de $\mid \eta \mid < 1$, e observando o segundo membro de (3.20), leva-nos a deduzir que, de uma forma geral, assimptoticamente, ter-se-á:

$$I_n \underset{n\to+\infty}{\asymp} \lambda_n \ . \tag{3.21}$$

De forma a detalhar este comportamento assimptótico, utilizaremos uma técnica standard de análise assimptótica, mais precisamente, a conhecida transformação de Abel, ver [Zorich, 2009], que afirma que:

$$\sum_{i=1}^{n} a_i b_i = \left(\sum_{k=1}^{n} a_k\right) b_n + \sum_{i=1}^{n-1} \left(\sum_{k=1}^{i} a_k\right) (b_i - b_{i+1}) \ . \tag{3.22}$$

Esta transformação é bastante útil sempre que o comportamento das somas com termos a_k são conhecidos e as oscilações dos termos b_i são controláveis.

Com $b_i = \lambda_i$ e $a_i = \eta^{n-i}$, para $i = 1, \ldots, n+1$, tem-se:

$$I_n = \sum_{i=1}^{n} a_i b_i = \left(\sum_{k=1}^{n} \eta^{n-k}\right) \lambda_n + \sum_{i=1}^{n-1} \left(\sum_{k=1}^{i} \eta^{n-k}\right) (\lambda_i - \lambda_{i+1}) \ .$$

Somando as progressões geométricas, obtém-se a seguinte expressão:

$$I_n = \frac{1-\eta^n}{1-\eta}\lambda_n + \sum_{i=1}^{n-1} \left(\eta^{n-i}\frac{1-\eta^i}{1-\eta}\right) (\lambda_i - \lambda_{i+1}) \ , \tag{3.23}$$

que se revela de grande importância para o que se segue.

Consideremos o seguinte teorema:

Teorema 3.3. *Se* $\lim_{n\to+\infty}\lambda_n = \lambda \in \mathbb{R}_+$ *então*

$$\lim_{n\to+\infty} I_n = \frac{\lambda}{1-\eta}\ .$$

Demonstração: Para demonstrar o resultado do teorema, é apenas necessário mostrar que o segundo membro da equação (3.23) tende para zero, quando n tende para $+\infty$.

Comecemos por notar que o segundo membro da equação (3.23) pode ser decomposto em duas parcelas, tendo-se, para a segunda, que:

$$-\sum_{i=1}^{n-1}\left(\frac{\eta^n}{1-\eta}\right)(\lambda_i-\lambda_{i+1}) = -\left(\frac{\eta^n}{1-\eta}\right)(\lambda_1-\lambda_n) \xrightarrow[n\to+\infty]{} 0\ . \tag{3.24}$$

Relativamente à primeira parcela, considere-se M tal que $\lambda_n \leq M$, algum $\varepsilon > 0$ arbitrário e algum índice n_0 tal que, para $n \geq n_0$, se tem $\mid \lambda_{n-1}-\lambda_n \mid < \varepsilon$.

Para $n \geq n_0$, teremos que:

$$\sum_{i=1}^{n-1}\left(\frac{\eta^{n-i}}{1-\eta}\right)(\lambda_i-\lambda_{i+1}) = \sum_{i=1}^{n_0}\left(\frac{\eta^{n-i}}{1-\eta}\right)(\lambda_i-\lambda_{i+1}) + \sum_{i=n_0+1}^{n-1}\left(\frac{\eta^{n-i}}{1-\eta}\right)(\lambda_i-\lambda_{i+1})\ . \tag{3.25}$$

Consideremos separadamente as duas somas de (3.25). Para a primeira parcela, ter-se-á que:

$$\left|\sum_{i=1}^{n_0}\left(\frac{\eta^{n-i}}{1-\eta}\right)(\lambda_i-\lambda_{i+1})\right| \leq 2M\ \mid\eta\mid^{n-n_0}\ \ \frac{1-\mid\eta\mid^{n_0}}{\mid 1-\eta\mid} \xrightarrow[n\to+\infty]{} 0\ . \tag{3.26}$$

Para a segunda parcela, notamos que

$$\left|\sum_{i=n_0+1}^{n-1}\left(\frac{\eta^{n-i}}{1-\eta}\right)(\lambda_i-\lambda_{i+1})\right| \leq \varepsilon\ \mid\eta\mid\ \frac{1-\mid\eta\mid^{n-1-n_0}}{\mid 1-\eta\mid} \leq \varepsilon\ \frac{2\mid\eta\mid}{\mid 1-\eta\mid}\ . \tag{3.27}$$

Como $\varepsilon > 0$ é arbitrário, a expressão (3.27), juntamente com (3.24) e (3.26), mostra que a segunda parcela da expressão (3.23) pode ser tão pequena quanto se queira, provando-se assim o resultado do teorema. □

Observação 3.2.

- *Fazemos notar que o Teorema 3.3 abrange o importante caso onde a sucessão dos parâmetros das distribuições de Poisson correspondentes às entradas na população converge para algum valor real. Esta é uma situação realista e comum numa variedade de situações em que as entradas estabilizam de uma forma clara ao fim de algum tempo.*

- *Outros casos ainda contemplados no Teorema 3.3 correspondem a fenómenos onde a sucessão dos parâmetros das entradas na população são limitados ou convergentes para alguma assimptota.*

- *O resultado do Teorema 3.3 inclui o caso em que os fluxos de entrada são modelados por* $\lambda_i = a + b\theta^i \ (a,b,\theta) \in \Theta_1$, *com* $\theta \in [0,1]$, *proposto e aplicado nos estudos de Guerreiro et al, bem como a forma sigmoidal que iremos adoptar na aplicação que apresentaremos adiante:* $\lambda_i = (a + b\, e^{-\theta i})^{-1}$, $(a,b,\theta) \in \Theta_2$.

- *Notamos ainda que o Torema 3.3 garante a estabilidade a longo prazo das populações cujos fluxos de entrada sejam modelados por sucessões mais complexas como* $\lambda_i = a + b\theta^{\frac{|i-c|^\alpha}{d^\beta}}$, *com* $\theta \in]0,1]$, $\alpha, \beta \in [1, +\infty[$ *e* a, b, c, d *parâmetros não negativos, ou outras expressões que sejam de interesse a aplicações práticas que respeitem as hipóteses do teorema.*

O caso de uma sequência controlada de parâmetros

Veremos, neste ponto, que nos casos em que a sequência dos parâmetros das entradas na população não se revela ser uma sucessão convergente, poderemos ainda, sob algumas condições, garantir a estabilidade das dimensões relativas a longo prazo, evidenciando-se, portanto, a existência de um Vórtice Estocástico nos estados transientes.

O próximo teorema complementa o Teorema 3.3, abrangendo os casos em que o número de novas entradas na população evolui de uma forma não limitada.

Referimos desde já que o resultado obtido corresponde ao que nos era expectável, de acordo com a expressão (3.21) e a extensão obtida no Teorema 3.3.

Teorema 3.4. *Suponhamos que:*

1. $\lim_{n\to+\infty} \lambda_n = +\infty$

2. *Existe uma constante* $C > 0$ *tal que*

$$\max_{1\le i\le n} \left| \frac{\lambda_i - \lambda_{i+1}}{\lambda_n} \right| \le C \quad .$$

Então temos que:

$$\lim_{n\to+\infty} \frac{|\, I_n \,|}{\lambda_n} = \frac{1}{1-\eta} \, .$$

Demonstração: A demostração é feita de forma análoga à do Teorema 3.3. Utilizando novamente a expressão (3.22) e a mesma decomposição para obter, à semelhança de (3.24),

$$-\sum_{i=1}^{n-1} \left(\frac{\eta^n}{1-\eta}\right) \left(\frac{\lambda_i - \lambda_{i+1}}{\lambda_n}\right) = -\left(\frac{\eta^n}{1-\eta}\right) \left(\frac{\lambda_1}{\lambda_n} - 1\right) \xrightarrow[n\to+\infty]{} 0 \, .$$

O segundo termo da equação pode ser escrito como

$$\sum_{i=1}^{n-1}\left(\frac{\eta^{n-i}}{1-\eta}\right)\left(\frac{\lambda_i-\lambda_{i+1}}{\lambda_n}\right)=\sum_{i=1}^{+\infty}\left(\frac{\eta^{n-i}}{1-\eta}\right)\left(\frac{\lambda_i-\lambda_{i+1}}{\lambda_n}\right)\mathbb{I}_{\{1,\ldots,n\}}(i)=\int_{\mathbb{N}} f_n(i)d\mu_c(i)\ , \tag{3.28}$$

com $\mathbb{I}_{\{1,\ldots,n\}}$ a função indicatriz sobre o conjunto $\{1,\ldots,n\}$, μ_c a medida de contagem sobre $\mathbb{N}$ e f_n uma função sobre os inteiros definida por:

$$f_n(i)=\left(\frac{\eta^{n-i}}{1-\eta}\right)\left(\frac{\lambda_i-\lambda_{i+1}}{\lambda_n}\right)\mathbb{I}_{\{1,\ldots,n-1\}}(i)\ .$$

Aplicaremos o teorema da convergência dominada de Lebesgue ao integral em (3.28), observando que se verifica a condição do teorema, obtendo-se:

$$|f_n(i))\,|\leq\frac{C}{|\,1-\eta\,|}\,|\,\eta\,|^{n-i}\,\mathbb{I}_{\{1,\ldots,n-1\}}(i)$$

e

$$\int_{\mathbb{N}}|\,\eta\,|^{n-i}\,\mathbf{1}_{\{1,\ldots,n-1\}}(i)d\mu_c(i)=\sum_{i=1}^{n-1}|\,\eta\,|^{n-i}\leq\frac{2\,|\,\eta\,|}{1-|\,\eta\,|}<+\infty\ .$$

Para todo $i\geq 1$, fixo, considerando que $|\,\eta\,|<1$ e a segunda hipótese do teorema, verifica-se ainda que

$$\lim_{n\to+\infty,\,n\geq i} f_n(i)=\lim_{n\to+\infty,\,n\geq i}\left(\frac{\eta^{n-i}}{1-\eta}\right)\left(\frac{\lambda_i-\lambda_{i+1}}{\lambda_n}\right)\mathbb{I}_{\{1,\ldots,n-1\}}(i)=0\ .$$

Concluí-se ainda que:

$$\sum_{i=1}^{n-1}\left(\frac{\eta^{n-i}}{1-\eta}\right)\left(\frac{\lambda_i-\lambda_{i+1}}{\lambda_n}\right)=\int_{\mathbb{N}} f_n(i)d\mu_c(i)\xrightarrow[n\to+\infty]{}0\ ,$$

demonstrando-se assim o resultado pretendido. □

Observação 3.3. *O Teorema 3.4 permite-nos uma generalização da Proposicão* 4.1 *em [Guerreiro et al., 2010], abrangendo, entre outros, o caso dos fluxos de entrada modelados por* $\lambda_i=a+b\theta^i$, $(a,b,\theta)\in\Theta_1$, *com* $\theta\in]1,+\infty[$.

O próximo teorema resume os resultados obtidos nesta secção afirmando que, sob hipóteses naturais, garante-se a existência de vórtices estocásticos nos estados transientes de uma cadeia de Markov.

Teorema 3.5. *Consideremos que um sistema, é modelado por uma cadeia de Markov em que a matriz de transição num passo entre estados transientes é diagonalizável.*

Suponhamos que as entradas no sistema são realizações de v.a's independentes com intensidades $\{\lambda_i\,,\,i\geq 1\}$ *e que o vector de classificação inicial nos estados transientes é convergente para um valor fixo, isto é,* $\lim_{i\to+\infty}\mathbf{t}_i=\mathbf{t}_\infty\neq\mathbf{0}$. *Então, com* $\boldsymbol{\lambda}_n^{+T}$ *o vector dos parâmetros de Poisson, correspondentes à dimensão das sub-populações, à data* $n\geq 1$, *teremos:*

1. *Se* $\lim_{n\to+\infty}\lambda_n = \lambda \in \mathbb{R}^+$ *então*

$$\lim_{n\to+\infty} \boldsymbol{\lambda}_n^{+T} = \sum_{j=1}^{s} \frac{\lambda}{1-\eta_j} \mathbf{t}_\infty^T \boldsymbol{\alpha}_j \boldsymbol{\beta}_j^T \ . \tag{3.29}$$

2. *Se* $\lim_{n\to+\infty}\lambda_n = +\infty$ *e existe um constante* $C > 0$ *de tal modo que*

$$\max_{1\le i\le n} \left| \frac{\lambda_i - \lambda_{i+1}}{\lambda_n} \right| \le C$$

então

$$\lim_{n\to+\infty} \frac{\boldsymbol{\lambda}_n^{+T}}{\lambda_n} = \sum_{j=1}^{s} \frac{1}{1-\eta_j} \mathbf{t}_\infty^T \boldsymbol{\alpha}_j \boldsymbol{\beta}_j^T \ . \tag{3.30}$$

Demonstração: O ponto 1 é uma consequência da prova do Teorema 3.3, considerando

$$\widetilde{I}_n^j := \sum_{i=1}^{n} \left(\lambda_i \mathbf{t}_i^T \boldsymbol{\alpha}_j\right) \eta^{n-i}$$

em vez de I_n e observando, sob as hipóteses do teorema, que se tem

$$\lim_{i\to+\infty} \lambda_i \mathbf{t}_i^T \boldsymbol{\alpha}_j = \lambda \mathbf{t}_\infty^T \boldsymbol{\alpha}_j.$$

O ponto 2 é uma consequência da prova do Teorema 3.4, considerando

$$\widetilde{\lambda}_n^j := \lambda_k \mathbf{t}_k^T \boldsymbol{\alpha}_j$$

em vez de λ_i e observando que, sob as hipótese do teorema, verifica-se ainda que

$$\lim_{n\to+\infty} \widetilde{\lambda}_n^j = +\infty$$

e que

$$\max_{1\le i\le n} \left| \frac{\widetilde{\lambda}_i^j - \widetilde{\lambda}_{i+1}^j}{\widetilde{\lambda}_n^j} \right| \le C \ .$$

□

Através da equação (3.29) do Teorema 3.5 garante-se, portanto, a existência de um vórtice estocástico nos estados transientes, uma vez que a equação (3.30) assegura a existência a estabilidade das dimensões relativas das sub-populações, numa perspectiva de longo prazo, apesar do número de indivíduos em cada sub-população não ser limitado.

De facto, observando-se que

$$\boldsymbol{\lambda}_n^{+T}/\lambda_n = (\lambda_{n,1}/\lambda_n, \ldots, \lambda_{n,s}/\lambda_n)$$

fica evidente que a proporção de indivíduos numa dada classe $j \in \{1, \ldots, s\}$, à data n, pode ser obtido a partir de

$$\pi_{n,j} = \frac{\lambda_{n,j}}{\sum_{k=1}^{s} \lambda_{n,k}} = \frac{\lambda_{n,j}/\lambda_n}{\sum_{k=1}^{s} (\lambda_{n,k}/\lambda_n)} \ , \tag{3.31}$$

que se prova ser convergente, de acordo com a expressão (3.30) do Teorema 3.5.

Uma vez garantida a existência de estabilidade a longo prazo nas dimensões relativas das sub-populações, e consequente existência de um vórtice estocástico sediado nos estados transientes (as classes de risco) iremos, nas secções que se seguem, obter estimadores de máxima verosimilhança e regiões de confiança para os parâmetros de interesse do modelo.

3.3.7 Estimação dos Parâmetros dos Fluxos de Entrada

Nesta secão estimar-se-ão, pelo método da máxima verosimilhança, os parâmetros dos fluxos de entrada de novos clientes na população.

Adoptaremos duas expressões funcionais que podem ser utilizadas na modelação dos fluxos de entrada em carteiras de crédito. Ambas se revelam úteis de um ponto de vista das aplicações. A primeira, dada por $\lambda_i = a + b\theta^i$, $(a, b, \theta) \in \Theta_1$ foi proposta e aplicada nos estudos de [Guerreiro, 2008], [Guerreiro et al., 2010] e [Guerreiro et al., 2012b] e será aqui apresentada como uma proposta viável de aplicação aos dados da carteira de crédito ao consumo que iremos estudar adiante.

Como contribuição desta dissertação, e uma vez que esta forma funcional respeita as condições do Teorema 3.3, apresentaremos ainda o estudo estatístico da modelação dos fluxos de entrada utilizando uma forma sigmoidal $\lambda_i = (a + be^{-\theta i})^{-1}$, $(a, b, \theta) \in \Theta_2$, que se revelou ser uma melhor opção para a modelação da carteira que iremos estudar.

Para ambas as formulações apresentaremos testes de hipóteses que nos permitam aferir sobre a validade da utilização de cada um dos fluxos para modelar as entradas de novos clientes na carteira.

Em ambos os casos, assumiremos que a matriz de transição num passo, $\boldsymbol{P}$, é conhecida e tendo em vista a aplicação que iremos efectuar na secção 3.5 consideramos, relativamente ao vector de classificação inicial, que $\boldsymbol{c}_i = \boldsymbol{c} = (c_1, \ldots, c_j)$ e obtemos o seu estimador de máxima verosimilhança.

3.3.8 Função de Verosimilhança na Ausência de Restrições

Consideremos uma amostra recolhida ao longo das datas $i \in \{1, \ldots, T\}$. Em cada data i é observado o resultado das v.a.'s independentes, correspondentes ao número de indivíduos que, nessa data, foram colocados em cada uma das classes do sistema.

Seja N_{ij} a v.a. correspondente ao número de indivíduos que, à data i, entram na carteira como novos clientes e são inicialmente colocados na classe de risco j. Como visto anteriormente,

$$N_{ij} \frown \mathcal{P}(\lambda_i c_{ij}) \equiv \mathcal{P}(\lambda_i c_j)\ ,\ i = 1, \ldots, T\ ,\ j = 1, \ldots, s.$$

Sejam $\boldsymbol{N}_i = [N_{ij}]\ ,\ i = 1, \ldots, T\ ,\ j = 1, \ldots, s$, e n_{ij} as realizações das v.a.'s N_{ij}, ou seja, os elementos que, no início do período i são colocados na sub-população j. Consideremos ainda $\lambda_{ij} = \lambda_i c_j\ ,\ i = 1, \ldots, T\ ,\ j = 1, \ldots, s$ as componentes de $\lambda_i \boldsymbol{c}$.

Diremos assim, que $N_{ij} \frown \mathcal{P}(\lambda_{ij})\ ,\ i = 1, \ldots, T\ ,\ j = 1, \ldots, s$ e $\boldsymbol{N}_i \frown \mathcal{P}(\lambda_i \boldsymbol{c})$.

Na ausência de restrições sobre os parâmetros dos fluxos de entrada, a função verosimilhança para $\boldsymbol{N} = [\boldsymbol{N}_i] = [N_{ij}]$, será dada por

$$L_\Omega(\boldsymbol{\lambda}) = \prod_{i=1}^{T} \prod_{j=1}^{s} e^{-\lambda_{ij}} \frac{\lambda_{ij}^{n_{ij}}}{n_{ij}!} \tag{3.32}$$

com $\boldsymbol{\lambda} = [\lambda_{ij}], i = 1, \ldots, T, j = 1, \ldots, s$.

Tomando $\boldsymbol{n} = [n_{ij}]\ , i = 1, \ldots, T, j = 1, \ldots, s$, a função de log-verosimilhança será dada por

$$\ell_\Omega(\boldsymbol{\lambda}) = a(\boldsymbol{n}) - \sum_{i=1}^{T} \sum_{j=1}^{s} \lambda_{ij} + \sum_{i=1}^{T} \sum_{j=1}^{s} n_{ij} log(\lambda_{ij}) \tag{3.33}$$

com

$$a(\boldsymbol{n}) = -\sum_{i=1}^{T} \sum_{j=1}^{s} log(n_{ij}!)$$

O máximo da função de log-verosimilhança, na ausência de restrições, obter-se-á a partir da equação

$$\frac{\partial l_\Omega(\boldsymbol{\lambda})}{d\lambda_{ij}} = -1 + \frac{n_{ij}}{\lambda_{ij}}, i = 1, \ldots, T, j = 1, \ldots, s$$

pelo que, os estimadores de máxima verosimilhança para λ_{ij}, na ausência de restrições, são dados por

$$\hat{\lambda}_{ij,\Omega} = N_{ij}\ , i = 1, \ldots, T, j = 1, \ldots, s \tag{3.34}$$

Assim, pelas equações (3.33) e (3.34), o máximo da função de log-verosimilhança, na ausência de restrições, será dado por

$$\hat{\ell}_{\Omega} = a(\boldsymbol{n}) - \sum_{i=1}^{T}\sum_{j=1}^{s} n_{ij} + \sum_{i=1}^{T}\sum_{j=1}^{s} n_{ij}\, log(n_{ij}) \tag{3.35}$$

Por uma questão de simplificação da escrita, utilizaremos a notação usual

$$n_{i\bullet} = \sum_{j=1}^{s} n_{ij} \quad , \quad n_{\bullet j} = \sum_{i=1}^{T} n_{ij} \quad , \quad n_{\bullet\bullet} = \sum_{i=1}^{T}\sum_{j=1}^{s} n_{ij}$$

bem como a correspondente notação para as v.a.'s $N_{i\bullet}, N_{\bullet j}$ e $N_{\bullet\bullet}$.

Assim, a equação (3.35), pode ser reescrita como

$$\hat{\ell}_{\Omega} = a(\boldsymbol{n}) - n_{\bullet\bullet} + \sum_{i=1}^{T}\sum_{j=1}^{s} n_{ij}\, log(n_{ij}) \tag{3.36}$$

3.3.9 Forma Exponencial $\lambda_i = a + b\theta^i$

Nesta subsecção apresentar-se-á o teste de hipótese que nos permitirá validar a hipótese do número de novos clientes que entram para a carteira, em cada data i, evoluir de acordo com $\lambda_i = a + b\theta^i, (a, b, \theta) \in \Theta_1$.

Este ajustamento foi estudado e aplicado em [Guerreiro et al., 2012b], [Guerreiro et al., 2012a], [Rodrigues, 2011], [Guerreiro et al., 2010] e [Guerreiro, 2008].

Consideremos as seguintes hipóteses:

$$\mathcal{H}_0 : \lambda_i = (a + b\theta^i)c_j \;\; vs \;\; \mathcal{H}_1 : \lambda_i \neq (a + b\theta^i)c_j, \;\; i \in \{1, \ldots, T\}, \;\; j \in \{1, \ldots, s\} \;\; (a, b, \theta) \in \Theta_1 \tag{3.37}$$

Sob a hipótese $\mathcal{H}_0$, a função de verosimilhança será dada por

$$L_{\omega_0}(a, b, \theta, \mathbf{c}) = \prod_{i=1}^{T}\prod_{j=1}^{s} e^{(a+b\theta^i)c_j} \frac{[(a + b\theta^i)c_j]^{n_{ij}}}{n_{ij}!} \,. \tag{3.38}$$

A função de log-verosimilhança, restrita a $\mathcal{H}_0$, ser-nos-á dada por

$$\ell_{\omega_0}(a, b, \theta, \mathbf{c}) = a(\boldsymbol{n}) - \sum_{i=1}^{T}\sum_{j=1}^{s}(a + b\theta^i)c_j + \sum_{i=1}^{T}\sum_{j=1}^{s} n_{ij}[log(a + b\theta^i) + log(c_j)] \tag{3.39}$$

em que $a(\boldsymbol{n}) = -\sum_{i=1}^{T}\sum_{j=1}^{s} log(n_{ij}!)$.

Sendo $\boldsymbol{c}$ o vector de classificação inicial, há que ter em conta a restrição:

$$\sum_{j=1}^{s} c_j = 1. \tag{3.40}$$

Introduzindo a restrição acima e recorrendo aos multiplicadores de Lagrange, trabalharemos com a seguinte função auxiliar

$$\begin{aligned} \ell^{\nu}_{\omega_0}(a,b,\theta,\mathbf{c}) &= a(\boldsymbol{n}) - \sum_{i=1}^{T}\sum_{j=1}^{s}(a+b\theta^i)c_j + \\ &+ \sum_{i=1}^{T}\sum_{j=1}^{s} n_{ij}[log(a+b\theta^i) + log(c_j)] + \nu\left(\sum_{j=1}^{s} c_j + 1\right) \end{aligned} \tag{3.41}$$

O máximo da função log-verosimilhança obtém-se a partir do seguinte sistema de equações:

$$\begin{cases} \frac{\partial \ell^{\nu}_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial a} = -\sum_{i=1}^{T}\sum_{j=1}^{s} c_j + \sum_{i=1}^{T}\sum_{j=1}^{s} \frac{n_{ij}}{a+b\theta^i} = 0 \\ \frac{\partial \ell^{\nu}_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial b} = -\sum_{i=1}^{T}\sum_{j=1}^{s} \theta^i c_j + \sum_{i=1}^{T}\sum_{j=1}^{s} \frac{\theta^i n_{ij}}{a+b\theta^i} = 0 \\ \frac{\partial \ell^{\nu}_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial \theta} = -\sum_{i=1}^{T}\sum_{j=1}^{s} bi\theta^{i-1} c_j + \sum_{i=1}^{T}\sum_{j=1}^{s} n_{ij}\frac{bi\theta^{i-1}}{a+b\theta^i} = 0 \\ \frac{\partial \ell^{\nu}_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial c_j} = -\sum_{i=1}^{T}\left(a+b\theta^i + \frac{n_{ij}}{c_j} + \nu\right) = 0 \\ \frac{\partial \ell^{\nu}_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial \nu} = \sum_{j=1}^{s} c_j - 1 = 0 \end{cases} \tag{3.42}$$

em que $i = 1, \ldots, T$, $j = 1, \ldots, s$.

Os estimadores de Máxima Verosimilhança $(\hat{a}, \hat{b}, \hat{\theta})$, para (a, b, θ), são as soluções das equações normais

$$\begin{cases} \sum_{i=1}^{T} \frac{N_{i\bullet}}{\hat{a}+\hat{b}\hat{\theta}^i} = T \\ \sum_{i=1}^{T} \hat{\theta}^i = \sum_{i=1}^{s} N_{i\bullet}\frac{\hat{\theta}^i}{\hat{a}+\hat{b}\hat{\theta}^i} \\ \sum_{i=1}^{T} i\hat{\theta}^{i-1} = \sum_{i=1}^{T} N_{i\bullet}\frac{i\hat{\theta}^{i-1}}{\hat{a}+\hat{b}\hat{\theta}^i} \end{cases} \tag{3.43}$$

Finalmente, $\hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_s)$, o estimador de máxima verosimilhança para $\mathbf{c}$, é obtido para $\nu = 0$, a partir da quarta equação do sistema (3.42), isto é, a partir de:

$$\nu + \frac{1}{T}\sum_{i=1}^{T}\left(\hat{a} + \hat{b}\,\hat{\theta}^i + \frac{n_{ij}}{\hat{c}_j}\right) = 0\,,$$

obtendo-se,

$$\forall i \in \{1,\ldots,s\} \quad , \quad \hat{c}_j = \frac{\frac{1}{T}\sum_{i=1}^{T}(\hat{a}+\hat{b}\hat{\theta}^i)}{\frac{1}{T}\sum_{i=1}^{T} n_{ij}} \ .$$

A partir dos resultados anteriores e com base no Teorema de Wilks, para mais detalhes ver [Mood et al., 1963], obter-se-á a estatística de teste para a qual, sob a hipótese $\mathcal{H}_0$, se tem

$$dim(\Omega) = T + s - 1$$

e

$$dim(\omega_0) = s + 2$$

pelo que $W = -2(\hat{\ell}_{\omega_0} - \hat{\ell}_{\Omega})$ terá distribuição assimptótica de um Qui-quadrado com graus de liberdade dados por $dim(\Omega) - dim(\omega_0)$, ou seja, $W \overset{a}{\frown} \chi^2_{T-3}$, o que nos permitirá testar a hipótese considerada, recordando que os testes de razão de verosimilhança são unilaterais direitos.

Com esta análise estatística, estamos em condições de ajustar a forma funcional exponencial aos dados de entrada na carteira, a partir dos quais se poderá prever a evolução da população.

3.3.10 Forma Sigmoidal $\lambda_i = \left(a + be^{-\theta i}\right)^{-1}$

Nesta secção obteremos os estimadores de máxima verosimilhança para os parâmnetros da forma sigmoidal $\lambda_i = (a + be^{-\theta i})^{-1}$ e desenvolveremos um teste de hipóteses para esta formulação.

$$\mathcal{H}_0 : \lambda_i = \left(a + be^{-\theta i}\right)^{-1} c_j \quad vs \mathcal{H}_1 : \lambda_i \neq \left(a + be^{-\theta i}\right)^{-1} c_j \quad , \ i \in \{1,\ldots,T\} \ , \ j \in \{1,\ldots,s\}$$

A função verosimilhança, sob a hipótese $\mathcal{H}_0$, é dada por:

$$L_{\omega_0}(a,b,\theta,\mathbf{c}) = \prod_{i=1}^{T}\prod_{j=1}^{s} \frac{\left(\frac{c_j}{a+be^{-\theta i}}\right)^{n_{ij}}}{n_{ij}!} e^{-\frac{c_j}{a+be^{-\theta i}}} \ . \tag{3.44}$$

A função de log-verosimilhança correspondente, sob a hipótese $\mathcal{H}_0$, é dada por

$$\ell_{\omega_0}(a,b,\theta,\mathbf{c}) = a(\mathbf{n}) - \sum_{i=1}^{T}\sum_{j=1}^{s} \frac{1}{a+be^{-\theta i}} c_j + \sum_{i=1}^{T}\sum_{j=1}^{s} n_{ij}\left[\log\left(\frac{1}{a+be^{-\theta i}}\right) + \log(c_j)\right] \tag{3.45}$$

com $a(\mathbf{n}) = -\sum_{i=1}^{T}\sum_{j=1}^{s}\log(n_{ij}!)$.

Devido à restrição $\sum_{j=1}^{s} c_j = 1$, necessitamos utilizar multiplicadores de Lagrange e assim teremos uma nova função objectivo para optimizar, dada por

$$\ell^{\nu}_{\omega_0}(a,b,\theta,\mathbf{c}) = \ell_{\omega_0}(a,b,\theta,\mathbf{c}) + \nu\left(\sum_{j=1}^{s} c_i - 1\right)$$

com o parâmetro auxiliar ν.

O conjunto das equações normais é dado por:

$$\begin{cases} \dfrac{\partial \ell^{\nu}_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial a} = \displaystyle\sum_{i=1}^{T}\sum_{j=1}^{s} \frac{1}{(a+be^{-\theta i})^2} c_j - \sum_{i=1}^{T}\sum_{j=1}^{s} \frac{n_{ij}}{a+be^{-\theta i}} = 0 \\ \dfrac{\partial \ell^{\nu}_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial b} = \displaystyle\sum_{i=1}^{T}\sum_{j=1}^{s} \frac{e^{-\theta i}}{(a+be^{-\theta i})^2} c_j - \sum_{i=1}^{T}\sum_{j=1}^{s} \frac{n_{ij}e^{-\theta i}}{a+be^{-\theta i}} = 0 \\ \dfrac{\partial \ell^{\nu}_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial \theta} = -\displaystyle\sum_{i=1}^{T}\sum_{j=1}^{s} \frac{bie^{-\theta i}}{(a+be^{-\theta i})^2} c_j + \sum_{i=1}^{T}\sum_{j=1}^{s} \frac{n_{ij}bie^{-\theta i}}{a+be^{-\theta i}} = 0 \\ \dfrac{\partial \ell^{\nu}_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial c_j} = \displaystyle\sum_{i=1}^{T} \left(-\frac{1}{a+be^{-\theta i}} + \frac{n_{ij}}{c_j} + \nu \right) = 0 \\ \dfrac{\partial \ell^{\nu}_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial \nu} = \displaystyle\sum_{j=1}^{s} c_j - 1 = 0 \, . \end{cases} \tag{3.46}$$

Em consequência, os estimadores de máxima verosimilhança $(\hat{a},\hat{b},\hat{\theta})$, para (a,b,θ), são as soluções das seguintes equações:

$$\begin{cases} \displaystyle\sum_{i=1}^{T} \frac{1}{(\hat{a}+\hat{b}e^{-\hat{\theta}i})^2} = \sum_{i=1}^{T} \frac{\left(\sum_{j=1}^{s} N_{ij}\right)}{\hat{a}+\hat{b}e^{-\hat{\theta}i}} \\ \displaystyle\sum_{i=1}^{T} \frac{e^{-\hat{\theta}i}}{(\hat{a}+\hat{b}e^{-\hat{\theta}i})^2} = \sum_{i=1}^{T} \frac{\left(\sum_{j=1}^{s} N_{ij}\right) e^{-\hat{\theta}i}}{\hat{a}+\hat{b}e^{-\hat{\theta}i}} \\ \displaystyle\sum_{i=1}^{T} \frac{\hat{b}ie^{-\hat{\theta}i}}{(\hat{a}+\hat{b}e^{-\hat{\theta}i})^2} = \sum_{i=1}^{T} \frac{\left(\sum_{j=1}^{s} N_{ij}\right) \hat{b}ie^{-\hat{\theta}i}}{\hat{a}+\hat{b}e^{-\hat{\theta}i}} \end{cases} \tag{3.47}$$

Finalmente, $\hat{\mathbf{c}} = (\hat{c}_1,\ldots,\hat{c}_s)$, o estimador de máxima verosimilhança para $\mathbf{c}$, é obtido para $\nu = 0$, a partir da quarta equação do sistema (3.46), isto é, apartir de:

$$\nu + \frac{1}{T}\sum_{i=1}^{T} \left(-\frac{1}{\hat{a}+\hat{b}e^{-\hat{\theta}i}} + \frac{N_{ij}}{\hat{c}_j} \right) = 0 \, ,$$

obtendo-se,

$$\forall j \in \{1,\ldots,s\} \quad , \quad \hat{c}_j = \frac{\frac{1}{T}\sum_{i=1}^{T} \frac{1}{\hat{a}+\hat{b}e^{-\hat{\theta}i}}}{\frac{1}{T}\sum_{i=1}^{T} N_{ij}} \, .$$

Com esta análise estatística, estamos em condições de ajustar a forma funcional sigmoidal aos dados de entrada na carteira, podendo a partir daqui prever a evolução da população.

3.4 Intervalos de Confiança para as dimensões das classes de risco

Nesta secção, centrar-nos-emos na obtenção de intervalos de confiança para a estimação das dimensões relativas de cada classe de risco, na situação em que o número esperado de novas entradas anuais é modelado de acordo com $\lambda_i = (a + be^{-\theta i})^{-1} : (a, b, \theta) \in \Theta_2$, apresentado na subsecção 3.3.10.

De acordo com os resultados obtidos na subsecção 3.3.10, e considerando novamente o vector de classificação inicial, $\boldsymbol{c}$, e a matriz de transição num passo, $\boldsymbol{P}$, como conhecidas, iremos obter a distribuição assimptótica dos estimadores dos parâmetros de entrada na carteira, o que nos permitirá obter intervalos de confiança assimptóticos para as proporções de clientes em cada classe de risco.

3.4.1 Distribuição assimptótica dos estimadores de máxima verosimilhança de (a, b, θ)

Quando o número esperado de novas entradas para a carteira, em cada data i, pode ser estimado de acordo com

$$\lambda_i = (a + be^{-\theta i})^{-1} \, , \, (a, b, \theta) \in \Theta_2 \, ,$$

a matriz de informação de Fisher, dada por

$$\mathbf{I}_{(\hat{a},\hat{b},\hat{\theta})} = \begin{pmatrix} \mathbb{E}\left[-\frac{\partial^2 l_{\omega_0}}{\partial a^2}\right] & \mathbb{E}\left[-\frac{\partial^2 l_{\omega_0}}{\partial a \partial b}\right] & \mathbb{E}\left[-\frac{\partial^2 l_{\omega_0}}{\partial a \partial \theta}\right] \\ \mathbb{E}\left[-\frac{\partial^2 l_{\omega_0}}{\partial b \partial a}\right] & \mathbb{E}\left[-\frac{\partial^2 l_{\omega_0}}{\partial b^2}\right] & \mathbb{E}\left[-\frac{\partial^2 l_{\omega_0}}{\partial b \partial \theta}\right] \\ \mathbb{E}\left[-\frac{\partial^2 l_{\omega_0}}{\partial \theta \partial a}\right] & \mathbb{E}\left[-\frac{\partial^2 l_{\omega_0}}{\partial \theta \partial b}\right] & \mathbb{E}\left[-\frac{\partial^2 l_{\omega_0}}{\partial \theta^2}\right] \end{pmatrix}$$

traduzir-se-á em

$$\mathbf{I}_{(\hat{a},\hat{b},\hat{\theta})} = \begin{pmatrix} \sum_{i=1}^{T} \lambda_i^3 & \sum_{i=1}^{T} e^{-i\theta} \lambda_i^3 & -b \sum_{i=1}^{T} i e^{-i\theta} \lambda_i^3 \\ \sum_{i=1}^{T} e^{-i\theta} \lambda_i^3 & \sum_{i=1}^{T} e^{-2i\theta} \lambda_i^3 & -b \sum_{i=1}^{T} i e^{-2i\theta} \lambda_i^3 \\ -b \sum_{i=1}^{T} i e^{-i\theta} \lambda_i^3 & -b \sum_{i=1}^{T} i e^{-2i\theta} \lambda_i^3 & b^2 \sum_{i=1}^{T} i^2 e^{-2i\theta} \lambda_i^3 \end{pmatrix} ,$$

uma vez que da derivada de (3.45), obtém-se:

$$\begin{cases} \dfrac{\partial l_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial a} = \displaystyle\sum_{i=1}^{T} \lambda_i^2 - \sum_{i=1}^{T} \frac{n_i}{\lambda_i} \\ \dfrac{\partial l_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial b} = \displaystyle\sum_{i=1}^{T} \lambda_i^2 e^{-\theta i} - \sum_{i=1}^{T} n_i \frac{e^{-\theta i}}{\lambda_i} \\ \dfrac{\partial l_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial \theta} = \displaystyle\sum_{i=1}^{T} \lambda_i^2 bie^{-\theta i} - \sum_{i=1}^{T} n_i \frac{bie^{-\theta i}}{\lambda_i} \end{cases}. \tag{3.48}$$

e aplicando as derivadas de ordem 2 na equação (3.39) teremos:

$$\begin{aligned}
-\tfrac{\partial^2 l_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial a^2} &= \sum_{i=1}^{T}\sum_{j=1}^{s} \lambda_i^2(2\lambda_i c_i - n_{ij}) \\
-\tfrac{\partial^2 l_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial a \partial b} &= \sum_{i=1}^{T}\sum_{j=1}^{s} e^{-\theta i}\lambda_i^2(2\lambda_i c_i - n_{ij}) \\
-\tfrac{\partial^2 l_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial a \partial \theta} &= \sum_{i=1}^{T}\sum_{j=1}^{s} bie^{-\theta i}\lambda_i^2(-2\lambda_i c_i + n_{ij}) \\
-\tfrac{\partial^2 l_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial b^2} &= \sum_{i=1}^{T}\sum_{j=1}^{s} e^{-2\theta i}\lambda_i^2(2\lambda_i c_i - n_{ij}) \\
-\tfrac{\partial^2 l_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial b \partial \theta} &= \sum_{i=1}^{T}\sum_{j=1}^{s} ie^{-\theta i}\lambda_i(-2e^{-\theta i}b\lambda_i^2 c_i + \lambda_i c_i + 2n_{i,j}e^{-\theta i}b\lambda_i - n_{ij}) \\
-\tfrac{\partial^2 l_{\omega_0}(a,b,\theta,\mathbf{c})}{\partial \theta^2} &= \sum_{i=1}^{T}\sum_{j=1}^{s} bi^2e^{-\theta i}\lambda_i(2be^{-\theta i}\lambda_i^2 c_i - \lambda_i c_i - bn_{i,j}e^{-\theta i}\lambda_i + n_{ij})
\end{aligned}$$

com $(\hat{a},\hat{b},\hat{\theta})$ os estimadores de máxima verosimilhança de (a,b,θ). As estimativas para (a,b,θ) serão obtidas a partir das observações nas datas $i \in \{1,\ldots,T\}$.

A matriz de variâncias-covariâncias de $(\hat{a},\hat{b},\hat{\theta})$ será dada por:

$$\boldsymbol{\Sigma}_{(\hat{a},\hat{b},\hat{\theta})} = \begin{bmatrix} \hat{\sigma}^2_{\hat{a}} & \hat{\sigma}_{\hat{a},\hat{b}} & \hat{\sigma}_{\hat{a},\hat{\theta}} \\ \hat{\sigma}_{\hat{b},\hat{a}} & \hat{\sigma}^2_{\hat{b}} & \hat{\sigma}_{\hat{b},\hat{\theta}} \\ \hat{\sigma}_{\hat{\theta},\hat{a}} & \hat{\sigma}_{\hat{\theta},\hat{b}} & \hat{\sigma}^2_{\hat{\theta}} \end{bmatrix} = \boldsymbol{I}^{-1}_{(\hat{a},\hat{b},\hat{\theta})}. \tag{3.49}$$

De acordo com as propriedades dos estimadores de máxima verosimilhança, ver por exemplo [Mood et al., 1963], podemos afirmar que $(\hat{a},\hat{b},\hat{\theta})$ têm distribuição assimptoticamente Normal com valor médio (a,b,θ) e matriz de variâncias-covariâncias dada por (3.49), ou seja

$$(\hat{a},\hat{b},\hat{\theta}) \overset{a}{\frown} \mathcal{N}\left((a,b,\theta), \boldsymbol{\Sigma}_{(\hat{a},\hat{b},\hat{\theta})}\right)$$

3.4.2 Distribuição Assimptótica de $\boldsymbol{\pi}_n$

Sendo $\pi_{n,j}$ a proporção de clientes que, à data n, se encontram na classe de risco $j \in \{1,\ldots,s\}$, tem-se que

$$\pi_{n,j} = \frac{\lambda_{n,j}^{+}}{\sum_{k=1}^{s} \lambda_{nk}^{+}}$$

sendo

$$\hat{\pi}_{n,j} = \frac{\hat{\lambda}_{n,j}^{+}}{\sum_{k=1}^{s} \hat{\lambda}_{nk}^{+}}$$

o seu estimador.

Observamos que $\pi_{n,j}$ é função de (a,b,θ), uma vez que é função de $\lambda_i = \lambda_i(a,b,\theta)$, com $a,b,\theta \in \Theta_2$.

De forma a obter a distribuição assimptótica de $\pi_{nj}(a,b,\theta)$, utilizaremos o método Delta, ver [Dacunha-Castelle e Duflo, 1983, p. 97] ou [Cramér, 1999, p. 353].

De forma a construir a matrix Jacobiana das transformadas ϕ_n, que nos permitem passar de (a,b,θ) para $\pi_{n1}(a,b,\theta),\ldots,\pi_{ns}(a,b,\theta)$, comecemos por definir os seguintes vectores:

$$\begin{cases} {}^{2}\boldsymbol{\lambda}_n^{+T} = \sum_{i=1}^{n} \lambda_i^2 \mathbf{q}_i^T \mathbf{P}^{(n-i)} \\ {}^{2\mathrm{e}}\boldsymbol{\lambda}_n^{+T} = \sum_{i=1}^{n} e^{-i\theta} \lambda_i^2 \mathbf{q}_i^T \mathbf{P}^{(n-i)} \\ {}^{2\mathrm{eb}}\boldsymbol{\lambda}_n^{+T} = b \sum_{i=1}^{n} i e^{-id} \lambda_i^2 \mathbf{q}_i^T \mathbf{P}^{(n-i)} \end{cases} \tag{3.50}$$

Relativamente aos vectores definidos em (3.50), podemos observar que

$$\frac{\partial \pi_{nj}}{\partial a} = \frac{-\,{}^{2}\lambda_{nj}^{+}\left(\sum_{j=1}^{s}\lambda_{nj}^{+}\right) + \lambda_{nj}^{+}\left(\sum_{j=1}^{s}{}^{2}\lambda_{nj}^{+}\right)}{\left(\sum_{j=1}^{s}\lambda_{nj}^{+}\right)^2},$$

$$\frac{\partial \pi_{nj}}{\partial b} = \frac{-\,{}^{2\mathrm{e}}\lambda_{nj}^{+}\left(\sum_{j=1}^{s}\lambda_{nj}^{+}\right) + \lambda_{nj}^{+}\left(\sum_{j=1}^{s}{}^{2\mathrm{e}}\lambda_{nj}^{+}\right)}{\left(\sum_{j=1}^{s}\lambda_{nj}^{+}\right)^2},$$

$$\frac{\partial \pi_{nj}}{\partial \theta} = \frac{{}^{2\mathrm{eb}}\lambda_{nj}^{+}\left(\sum_{j=1}^{s-1}\lambda_{nj}^{+}\right) - \lambda_{nj}^{+}\left(\sum_{j=1}^{s-1}{}^{2\mathrm{eb}}\lambda_{nj}^{+}\right)}{\left(\sum_{j=1}^{s-1}\lambda_{nj}^{+}\right)^2},$$

A matriz Jacobiana das transformadas ϕ_n,dada por

$$\boldsymbol{J}(\phi_n) = \begin{pmatrix} \frac{\partial \pi_{n1}}{\partial a} & \frac{\partial \pi_{n1}}{\partial b} & \frac{\partial \pi_{n1}}{\partial \theta} \\ \vdots & \ddots & \vdots \\ \frac{\partial \pi_{ns}}{\partial a} & \frac{\partial \pi_{ns}}{\partial b} & \frac{\partial \pi_{ns}}{\partial \theta} \end{pmatrix},$$

encontra-se perfeitamente definida e é facilmente determinada.

Como consequência do método-delta, podemos concluir que o vector $\hat{\boldsymbol{\pi}}_n^T = (\hat{\pi}_{n1}, \ldots, \hat{\pi}_{ns})$ tem distribuição assimptoticamente Normal com vector médio $\boldsymbol{\pi}_n^T = (\pi_{n1}, \ldots, \pi_{ns})$ e matriz de variâncias-covariâncias dada por

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}_{\boldsymbol{n}}} = \frac{1}{n}\left[\boldsymbol{J}(\phi_n)\Sigma_{(\hat{a},\hat{b},\hat{\theta})}\boldsymbol{J}^T(\phi_n)\right] \tag{3.51}$$

o que nos permitirá determinar intervalos de confiança assimptóticos para a proporção de clientes em cada classe de risco, aferindo assim acerca da precisão das estimativas efectudas.

3.5 Aplicação

Com o objectivo de estimar a evolução da temporal da probabilidade de incumprimento de uma carteira de crédito ao consumo de uma instituição financeira Cabo-verdiana apresentamos nesta secção, uma aplicação das abordagens propostas nas secções anteriores. A modelação baseia-se em duas formas funcionais de ajustamento, nomeadamente a $f_s(x) = (a + b\, e^{-\theta x})^{-1}$ e $f_e(x) = a + b\, \theta^x$. Iremos, ainda, analisar a evolução das dimensões absolutas e relativas para a carteira e para as classes de risco. Para terminar a secção, apresenta-se os intervalos de confiança para as proporções de acordo com abordagem sigmoidal.

3.5.1 Caracterização da Carteira

A carteira utilizada neste capítulo é constituída por dados históricos de todos os clientes cujo contractos foram financiados entre Janeiro de 2003 e Outubro de 2011, de uma carteira de empréstimos bancários para crédito ao consumo de um banco comercial Cabo-verdiano. O ponto seguinte descreve a forma como definimos as classes de risco. Tendo em conta a natureza da carteira de créditos ao consumo, tomemons como unidade temporal o mês, pelo que avaliaremos a evolução da carteira ao longo de todos os meses.

População e Classes de Risco

Nesta aplicação, consideramos apenas os clientes da base de dados para os quais dispunhamos de toda a informação relativa ao histórico do seu contrato. Desta forma, optámos por considerar apenas os clientes cujo o contrato havia terminado até 31 de outubro de 2011, o que se traduziu num conjunto de 23821 clientes, observados ao longo de 106 meses. Neste conjunto de clientes e neste período temporal verificam-se entradas e saídas, e consideramos que as probabilidades de transição entre as classes de risco, num mês, correspondem a uma cadeia

de Markov homogénea com cinco estados transientes e um estado recorrente (absorvente), correspondente ao estado de saída da carteira.

Na aplicação que iremos efectuar, definimos que os clientes da carteira seriam classificados em diferentes classes de risco, de acordo com o número de dias de incumprimento após a data da primeira prestação. Consideremos cinco clases de risco:

- $C_1 \equiv [0, 30]$ - contém os clientes sem prestações em dívida por um período entre 0 e 30 dias;
- $C_2 \equiv [31, 60]$ - contém os clientes com prestações em dívida por um período entre 31 e 60 dias;
- $C_3 \equiv [61, 90]$ - contém os clientes com prestações em dívida por um período entre 61 e 90 dias;
- $C_4 \equiv [91, 120]$ - contém os clientes com prestações em dívida por um período entre 91 e 120 dias;
- $C_5 \equiv [120, +]$ - contém os clientes com prestações em dívida por um período superior a 120 dias.

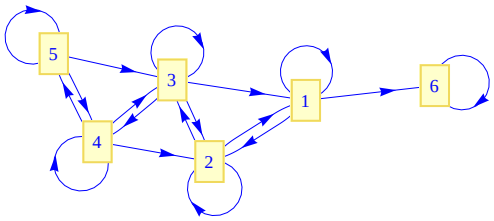
Tabela 3.2: *Sub-populações - Classes de Risco*

Sub-população	**Número de dias em incumprimento**
1	0-30
2	31-60
3	61-90
4	91-120
5	> 120
6 - Saídas	—

Fazemos notar que as classes de risco poderão ser construídas de com critérios bastante distintos, de acordo com a filosofia da instituição bancária no que respeita à classificação do risco dos clientes.

Por exemplo, as classes de risco podem agrupar os clientes de acordo com o *spread* que lhes foi atribuído. Esta aplicação será particularmente interessante se o *spread* do cliente for sendo ajustado ao longo do contrato, de acordo com diferenças de risco que a experiência do contrato vá transparecendo. Seria assim interessante estimar, a longo prazo, a proporção de clientes que pagará um dado *spread*. Referimos ainda que, neste contexto, o vector $\boldsymbol{c}$ de classificação inicial, poderá não ser constante ao longo do tempo.

Figura 3.1: *Grafo das Transições entre as Classes da Cadeia*



A Figura 3.1 ilustra as possibilidades de transição entre as classes de risco. Dada a forma como estas foram construídas, observamos que, todos os clientes entram directamente na primeira classe de risco. Os clientes são, em seguida, reclassificados, a cada mês, transitando para a classe de risco correspondente de acordo com o atraso, ou não, dos seus planos de reembolso. Sempre que um cliente termina o seu contrato na classe $[0, 30]$ dias, ele é classificado na sexta classe de risco, que corresponde a saída da carteira, indicando que o seu contrato terminou uma vez que o cliente efectuou todos os pagamentos necessários ao reembolso do empréstimo, de acordo com as condições do contrato. Sempre que um cliente tem mais de 30 dias de atraso até a data do término do seu contrato, ele é considerado cliente da carteira mantendo-se na classe correspondente ao número de dias de incumprimento, até que a última das prestações em dívida seja paga. De acordo com esta definição, apenas serão possiveis saídas da carteira, a partir da classe 1. Os clientes em incumprimento manter-se-ão na carteira até à liquidação da sua dívida.

3.5.2 Análise da Matriz de Transição

Definidas as classes de risco, construir-se-á a matriz de probabilidades de transição num passo para a carteira em causa. Note-se que, nesta aplicação, um passo corresponderá a um mês, uma vez que, regra geral, os planos de pagamentos de crédito ao consumo têm uma periodicidade mensal. A matriz de transição num passo, com as classes de risco definidas como na secção anterior, será dada por:

$$\boldsymbol{P} = \begin{bmatrix} p_{1,1} & p_{1,2} & 0 & \dots & 0 & p_{1,s} & q_1 \\ p_{2,1} & p_{2,2} & p_{2,3} & \dots & 0 & p_{2,s} & q_2 \\ p_{3,1} & p_{3,2} & p_{3,3} & \dots & 0 & p_{3,s} & q_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ p_{s,1} & p_{s,2} & p_{s,3} & \dots & p_{s,s-1} & p_{s,s} & q_s \\ 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{bmatrix}$$

em que:

- a soma dos elementos de cada linha será sempre igual a 1. Com efeito, um cliente, ao

fim de um período (mês), ou permanece na carteira ocupando uma dada classe de $\boldsymbol{K}$, ou, se o pagamento corresponde à última prestação, o cliente transita para a classe de saída.

- a última coluna da matriz representa o estado de saída dos clientes da carteira do crédito. De acordo com opcção efectuada à cerca das saídas dos clientes verifica-se, neste caso, que $q_i = 0$, $i = 2, \ldots, s$.

A matriz que se segue ilustra as probabilidades das migrações futuras, estimadas a partir dos dados históricos da carteira de crédito, no final de cada mês.

Tabela 3.3: *Matriz de Transição a um Passo*

Classes	$[0, 30]$	$[31, 60]$	$[61, 90]$	$[91, 120]$	$]120, +[$	**Saídas**
$[0, 30]$	0.934735	0.026566	0	0	0	0.038698
$[31, 60]$	0.518363	0.285733	0.195903	0	0	0
$[61, 90]$	0.009076	0.372018	0.248963	0.369943	0	0
$[91, 120]$	0	0.007835	0.335464	0.205361	0.450928	0
$]120, +[$	0	0	0.000820	0.052900	0.94628	0
Saídas	0	0	0	0	0	1

Analisando as linhas da matriz de transição podemos notar, por exemplo, que existe uma probabilidade de 93.47% dos clientes sem pagamentos em atraso continuarem na mesma classe de risco, 2.66% transitarem para a classe de risco C_2 e 3.87% saírem da carteira no mês seguinte.

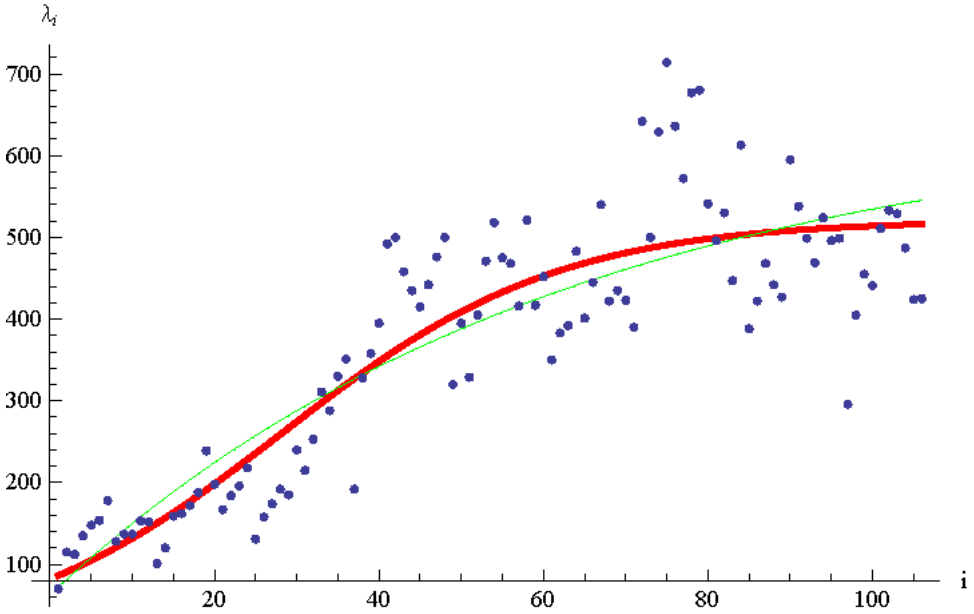
Podemos observar que os clientes da segunda, terceira e quarta classes de risco, têm 28.57%, 24.90% e 20.54% de probabilidade, respectivamente, de se manter na mesma classe. Isso indica que, à medida que o número de dias de incumprimento aumenta, vai diminuindo a probabilidade de se manter na mesma classe.

Observando a Tabela 3.3, pode verificar-se ques os clientes colocados na classe de risco mais gravosa, com 94.63% de probabilidade, no proximo mês manter-se-ão nessa classe. Isto significa que a probabilidade de um cliente com muitas prestações em atraso vir recuperar do incumprimento é pequena.

3.5.3 Ajustamento das Formas Funcionais

Na Figura 3.2 e na Tabela 3.4 pode analisar-se a evolução das duas formas de ajustamentos para o número de novos clientes que entrem mensalmente para a carteira. A forma funcional f_s corresponde a uma sigmoide e a forma funcional f_e corresponde a uma exponencial.

Figura 3.2: *Ajustamento das duas formas funcionais mensais para novos clientes*



Conforme se pode verificar, para ambas as formas funcionais, a vermelho a forma sigmoidal e a verde a forma exponencial, o número de novos clientes a entrar para a carteira tende a tornar-se estável. Essa estabilização reflete, naturalmente, a quota de mercado da instituição em relação à população de Cabo Verde, o que nos permite também considerar os modelos adoptados como representativos da realidade da carteira.

Os parâmetros das formas funcionais são obtidos pelos estimadores de máxima verosimilhança, apresentados na secção 3.3.10 e obtidos através da resolução das equações normais (3.46). Como se pode verificar, para esta definição de classe de riscotemos que $\boldsymbol{c}^T = (1, 0, \ldots, 0)$. Isto significa que os novos indivíduos são inicialmente classificados na primeira classe de risco, como não poderia deixar de ser, pois nenhum cliente terá inicialmente prestações em atraso. O detalhe dos cálculos do ajustamento são apresentados na Tabela 3.4.

Tabela 3.4: *Parâmetros Estimados e medidas de Ajustamentos*

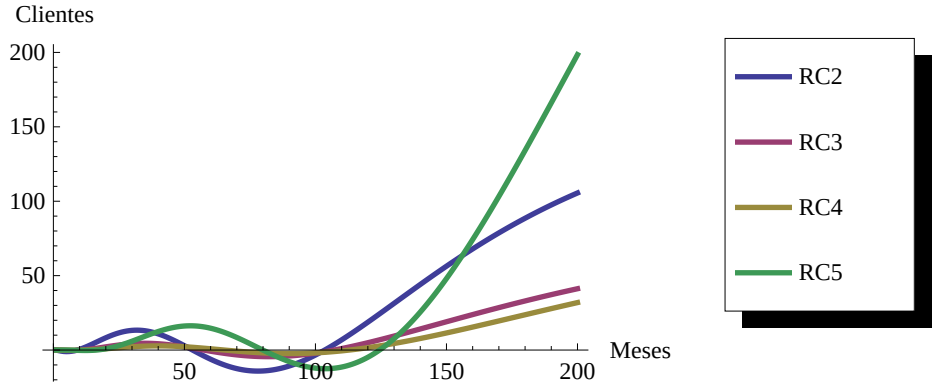
	$f_s(x) = (a + b\, e^{-\theta x})^{-1}$	$f_e(x) = a + b\, \theta^x$
$\hat{a}$	0.00191885	654.705
$\hat{b}$	0.0102673	−592.542
$\hat{\theta}$	0.0594656	0.984174
Σ^2	573.555	688.020
$\lim_{x\to+\infty}$	521.146	654.705

Observando a soma dos quadrados dos desvios entre os dados mensais e o valor correspondente de cada ajustamento, verifica-se que a forma funcional sigmoidal proporciona um

melhor ajustamento aos dados da carteira.

Na Figura 3.3 pode analisar-se a diferença entre os dois ajustamentos. É de notar que as diferenças tendem a ser estritamente crescentes a partir do mês 125, com maior relevância para a última classe, sendo esta uma classe importante na medida em que estima a proporção de clientes com mais prestações em atraso.

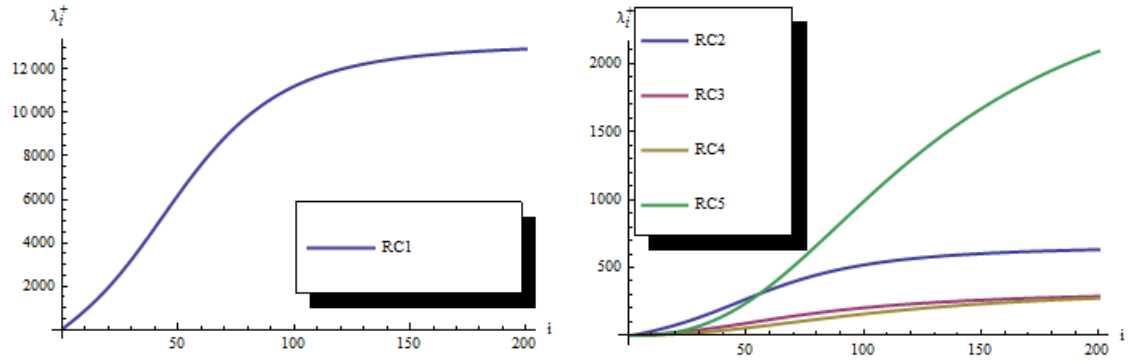
Figura 3.3: *Diferença entre o ajustamento das duas formas funcionais*



3.5.4 Evolução das Dimensões das Classes de Risco

Através do ajustamento da função sigmoide gerámos uma tabela de valores referente a 200 meses para novos clientes da carteira e a partir da expressão (3.15) foi estimado o número esperado de clientes em cada classe de risco da carteira ao longo de um período de 200 meses.

Figura 3.4: *Números de clientes nas classes de risco de 1 a 5 - forma sigmoidal*

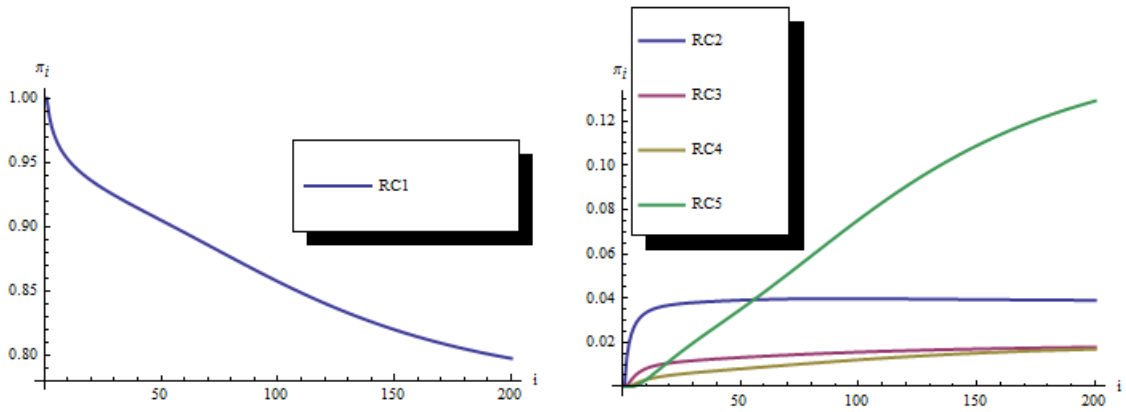


Como se pode verificar, na Figura 3.4, o número estimado de clientes nas classes de risco revela uma tendência crescente até que se atinja a estacionaridade, altura a partir da qual se verifica uma estabilização na evolução das classes de risco.

É de salientar que as classes de risco dos clientes com mais de cento e vinte dias de incumprimento cresce de forma mais acentuada em relação às classes de risco dos clientes que possuem entre sessenta e um a cento e vinte dias de incumprimento. Verifica-se, ainda que, a partir do mês setenta, ultrapassa a classe de risco dos clientes com trinta a sessenta dias de incumprimento. Esse comportamento pode ser explicado pelo facto de clientes com mais dias de incumprimento dificilmente pagarem todas as prestações em atraso, pelo que constituem um problema real para a instituição.

Analisando as estimativas para a dimensão relativa de cada classe de risco, na Figura 3.5, verifica-se uma estabilização ao longo do tempo. É de notar que, contrariamente às outras classes, o peso relativo da classe de risco de clientes com zero a trinta dias de incumprimento inicia-se com a totalidade dos clientes, e revela uma tendência decrescente até atingir a estacionaridade. O decréscimo inicial é reflexo da forma como os clientes são classificados inicialmente. Essa é também a razão pela qual as restantes classes de risco só após

Figura 3.5: *Evolução das proporções nas classes de risco de 1 a 5 - forma sigmoidal*



alguns meses passaram a ter clientes. Particularmente, a quinta classe de risco apresenta um crescimento mais acentuado, quando comparada com a terceira e a quarta de risco. Isto justifica-se, pelo facto desses clientes raramente cumprirem o plano de pagamento, o que faz com que quase todos se mantenham nessa classe de risco, à qual se irão adicionando os novos clientes bastante "incumpridores", a cada mês.

Havendo estabilização nas dimensões absolutas e relativas da populaçãao, pode verificar-se a existência de um vórtice estocástico nos estados transientes e conclui-se, portanto, que no caso das intensidades de entrada se efectuarem de acordo com um crescimento sigmoidal, a população atinge a estacionaridade.

Naturalmente, a última classe de risco atinje a estabilidade mais tarde comparativamente com as outras classes de risco. Verifica-se assim, em todas as classes, uma distribuição limite que

permite estimar todos os parâmetros numa situação de carteira estável. A estimação destes parâmetros, constituiem elementos importantes que poderão ser utilizados pelos gestores das carteiras de crédito na tomada de decisões. Por fim, uma conclusão importante, prende-se com a aplicação de uma política independente da dimensão da população da carteira.

Para efeitos de comparação com os resultados obtidos no Capítulo 2 agregamos as classes de risco 4 e 5 que correspondem a clientes com mais de 90 dias de incumprimento. O resultado encontra-se ilustrado na Figura 3.6.

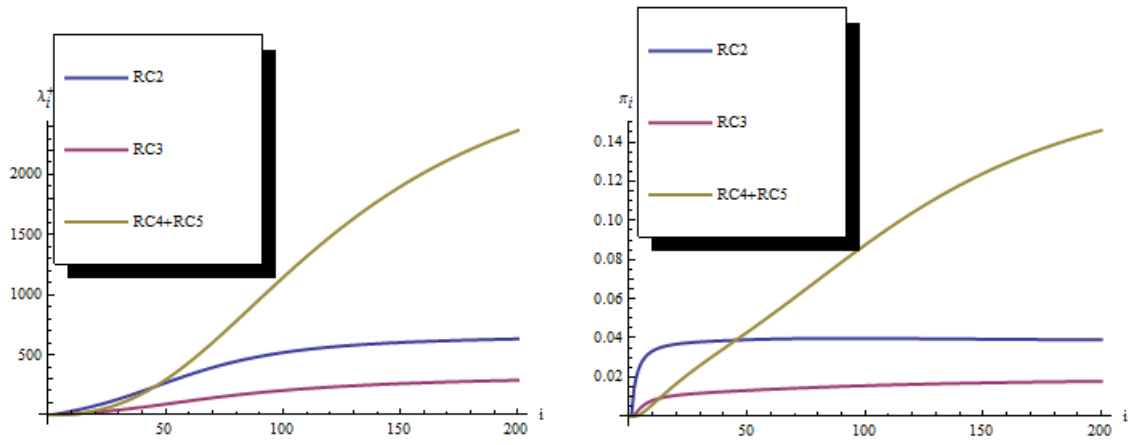


Figura 3.6: *Estimação da Evolução do Número e Proporção de Clientes - Forma Sigmoidal*

Utilizando os resultados sobre os intervalos de confiança para as proporções, obtidos na subsecção 3.3.10, obtemos os intervalos de confiança para estas proporções, dada pela expressão (3.15), à data correspondente para o mês 106 (Outubro de 2011). Os resultados são apresentados na Tabela 3.5.

Tabela 3.5: *Modelos de Proporções, no mês* 106*, para a forma funcional sigmoidal*

	RC1	**RC2**	**RC3**	**RC4 + RC5**
%	85.1926	3.96638	1.57805	9.2629
IC (95%)	85.1046,85.2806	3.96108,3.97168	1.57169,1.58441	6.6961,11.8299
Obs (%)	90.755	1.972	0.933	6.340

Podemos concluir que no último mês da carteira observado, mês 106 (Outubro de 2011), o modelo estima que a soma das proporções de clientes nessas duas classes é de 9.26298% (com um intervalo de confiança de 95% dos $[6.69607, 11.8299]$). Como se pode verificar, a estimativa da probabilidade de incumprimento da carteira obtido, através da Regressão Logística, pertence ao intervalo de confiança acima estimado pelo modelo aberto da cadeia de Markov.

A última linha da Tabela 3.5 ilustra as proporções observadas da carteira, à data 106. Como se pode verificar, a percentagem observada de clientes na classe de risco mais gravosa, encontra-se próxima do limite inferior do intervalo de confiança a 95%.

Isto indicia que ambos os modelos (Regressão Logística e Vórtices Estoicásticos) são robustos para a estimação da probabilidade de incumprimento e iremos utilizá-los no cálculo do *spread* do cliente e da carteira, respectivamente, no capítulo 4.

A Tabela 3.6 ilustra a aplicação do Teorema 3.5, cujos valores foram calculados pela expressão (3.29). Os parâmetros assimptóticos das leis de Poisson para cada classe de risco encontram-se na segunda linha da tabela . Na terceira linha desta tabela, temos as proporções relativas em relação às classes de risco de 1-5 para o mês 200 (Agosto de 2019). De acordo com as observações da carteira em estudo, o modelo estima que no mês 200 a carteira atingirá uma probabilidade de incumprimento de aproximadamente 17%.

Tabela 3.6: *Vórtices estocásticos nas classes de risco para a forma funcional sigmoidal - mês 200*

	RC1	**RC2**	**RC3**	**RC4**	**RC5**	**RC4+RC5**
#	13,230	661	320	326	2621	2947
%	77.108	3.853	1.868	1.898	15.273	17.171

3.6 Considerações Finais e Estudos Futuros

Considerações Finais

Neste capítulo, ao nível de desenvolvimentos teóricos, generalizámos a forma funcional que modela os fluxos de entrada na população proposta nos estudos de Guerreiro et al. Desenvolvemos os resultados referentes à inferência estatística para fluxos de entrada com distribuição de Poisson, com fluxo de entrada com forma sigmoidal, que permitiram obter desenvolvimentos relativos à estimação das intensidades de entrada de novos elementos para a população, bem como relativamente à análise da estrutura da mesma num dado período. Finalmente, também, em termos teóricos foi possível construir regiões de confiança e testes de hipóteses.

Numa perspectiva mais prática, e útil para a instituição que cedeu os dados para este estudo, foram ajustados dois exemplos de modelos de Markov para carteiras abertas, alimentadas por v.a's independentes de Poisson, apresentando estabilidade assintótica das proporções de indivíduos nas classes de risco. Mostrámos também que este estudo permite a estimação dos parâmetros relevantes e, numa fase posterior, fazemos a análise estatística dos resultados

do modelo. Aplicámos o modelo a uma carteira de crédito ao consumo de um banco Caboverdiano e conseguimos para ambos, indicar uma taxa de inadimplência actual, referente ao último mês de análise da base de dados, e uma probabilidade de incumprimento em concordância com o resultado obtido no Capítulo 2.

Por fim, estimámos a probabilidade de incumprimento do mês 200 (Agosto de 2019) para a carteira em estudo podendo, no entanto, esta estimativa ser efectuada para qualquer outro instante temporal.

Estudos Futuros

No prolongamento dos estudos feitos por Guerreiro et al, pretendemos futuramente:

- analisar a estrutura estocástica da carteira sob diferentes classes de risco. Analisar a evolução temporal do *spread* da carteira com base nas proporções de clientes em cada classe de risco;
- estudar com mais detalhes a estrutura dos Vórtices Estocásticos nos estados recorrentes, sob a hipótese dos fluxos de entrada modelados por uma função sigmoidal;
- estimar a probabilidade de incumprimento ao longo do tempo, através de um modelo de Markov para populações abertas, com base nas variáveis socio-económicas dos clientes, analisadas aquando da concessão do crédito.

Capítulo 4

Uma Abordagem Actuarial para a Estimação do *Spread*

4.1 Introdução

Ao contrair um empréstimo, cada cliente deverá ressarcir a instituição credora do montante emprestado, adicionado do montante dos juro decorrentes desse empréstimo.

Nos termos do contrato, o cliente fica sujeito ao pagamento de uma taxa de juro e de um *spread*. O *spread*, também denominado por "taxa de risco" corresponde à diferença entre a taxa de juro que as instituições financeiras pagam na aquisição do dinheiro e a que cobram aos clientes. O *spread* não é o mesmo para todos os clientes pois este é função do risco que o cliente representa para a instituição e a propensão para o incumprimento, como vimos anteriormente é parcialmente explicável pelas características do cliente.

O risco de incumprimento de cada cliente deve ser mensurado e o *spread* cobrado deve ser o adequado para fazer face a esse risco, e a proporcionar lucro à instituição. Para estimação do *spread*, estimação essa efectuada *a priori*, ou seja, aquando da concessão do crédito utilizam-se ferramentas como as descritas nos Capítulos 1 e 2.

Verifica-se, no entanto, como veremos adiante, que a estimação da probabilidade de incumprimento tem um papel preponderante, mas não suficiente na estimação do *spread* adequado a cada cliente.

Existem diversos autores e inúmeros trabalhos que propõem modelos para a estimação do *spread*. Não nos alongaremos em muitos detalhes. Propomos, para mais detalhes, a consulta de [Duffee, 1998], [Longstaff e Schwartz, 1995], [Bevan e Garzarelli, 2000], [Merton, 1974], [Davies, 2008] e [Morris et al., 1998].

Nesta dissertação, no prolongamento do estudo de [Vale, 2010], propomos um conjunto de modelos que definam o *spread* da carteira e de cada cliente, em função da probabilidade de incumprimento e da taxa de recuperação. A taxa de recuperação representa qual a proporção do crédito que a instituição bancária consegue recuperar, dado que o cliente entrou em *incumprimento*. A taxa de recuperação não é certamente a mesma para cada cliente e poderá talvez ser mensurada e estimada pelas características do cliente.

A carteira em estudo possui dados que permitiram estimar, esses elementos indispensáveis: as probabilidades de incumprimento e as taxas de recuperação. Na secção que se segue descrever-se-ão detalhadamente cada um desses elementos.

Este capítulo estrutura-se da seguinte forma: na secção 4.2 formulam-se os modelos para estimação do *spread* de acordo com a metodologia actuarial. Ainda nesta secção propõe-se a formulação teórica da taxa de recuperação da carteira. A discussão dos resultados empíricos do *spread* da carteira e do cliente faz-se na secção 4.3. Na secção 4.4 indentifica-se as variáveis preditivas para a recuperação do cliente incumpridor, através da Regressão Beta. Por fim, na secção 4.5 apresenta-se as conclusões e os trabalhos futuros.

4.2 Metodologia actuarial

Observamos que, de acordo com a prática habitual, ver [McNeil et al., 2005], existem duas metodologias de apreçamento de risco de crédito: as actuariais e financeiras. Estas metodologias são diferenciadas, respectivamente, pelo uso da medida de probabilidade natural e da medida martingala equivalente. Esta última metodologia que deve provar a existência de alguma medida de probabilidade, coincidindo com a probabilidade natural nos conjuntos de probabilidade zero e tal que, no processo da evolução dos *cash-flows* sob a medida de martingala equivalente ou medida neutra face ao risco é, agora, uma martingala. Neste estudo focalizamos na metodologia actuarial.

4.2.1 Generalidades

Iremos, nesta subsecção, começar por apresentar as definições, variáveis e parâmetros necessários para a modelação matemática do risco de crédito. A evolução do *cash-flow* pode ser escrito como um processo estocástico $(X_t)_{t\in\mathcal{I}}$ com o conjunto de tempo $\mathcal{I}$ que é qualquer subconjunto dos números inteiros, no caso do modelo em tempo discreto ou dos números reais não negativos, no caso do modelo em tempo contínuo. Em função da complexidade da evolução de *cash-flow*, podemos recorrer à modelação deste fenómeno por um processo de Markov ou mesmo uma martingala.

Para *cash-flows* dos clientes incumpridores (sujeitos ao risco de *incumprimento*) o tempo de

incumprimento τ corresponde ao momento da ocorrência de *incumprimento*, que deve ser um tempo de paragem. Associada a esta variável ter-se-á a probabilidade de incumprimento, que deverá ser um parâmetro do modelo;

A recuperação dos *cash-flows* dos incumpridores deve ser também representado por um processo de Markov. Note-se que a informação disponível sobre a recuperação nem sempre é fiável, o que sugere que este processo deve ser considerado como um parâmetro constante λ (os valores de λ registados por uma instituição financeira referem-se apenas ao valor recuperado pela mesma, no entanto, as empresas encarregues de recuperar o património podem recuperar $k > \lambda$ ou $k < \lambda$ mas devolvem apenas à instituição financeira a parte acordada com a mesma).

Considera-se que o *spread* deverá ser um processo de Markov. Adicionalmente, nos modelos mais completos deve considerar-se também a existência de alguma estrutura temporal (*term structure*). Teoricamente, ver [McNeil et al., 2005], o *spread* de crédito $s(t,T)$, no tempo t e com maturidade T, para uma obrigação de cupão zero com possibilidade de incumprimento (*defaultable zero coupon bond*) com o preço no tempo t é dado por $p_1(t,T)$ e maturidade no tempo T, é tal que

$$p_1(t,T)exp\left[(T-t)s(t,T)\right] = p_0(t,T), \tag{4.1}$$

com $p_0(t,T)$ o preço no tempo t de uma obrigação de cupão zero livre de incumprimento (*default-free zero coupon bond*) com vencimento em T. A expressão que descreve o *spread* pode ser obtida a partir da expressão 4.1:

$$s(t,T) = -\frac{1}{T-t} ln\left(\frac{p_1(t,T)}{p_0(t,T)}\right). \tag{4.2}$$

4.2.2 Modelo a um período para um modelo de *zero cupon bonds*

Considere-se uma obrigação, ou um empréstimo de $100u.m.$, com a taxa de juro constante r, a probabilidade de incumprimento p, a taxa de recuperação λ e o *spread* s. A Tabela 4.1, ilustra o valor esperado dos *cash-flows* a um período para uma obrigação ou um empréstimo *defaultable* (sujeitos a risco de *incumprimento*).

Suponhamos dois cenários:

- primeiro, em que uma obrigação está livre de *incumprimento* (*default-free*) ou um empréstimo, com $100u.m.$ que, na maturidade T, converter-se-ão em $100(1+r)u.m.$;
- segundo, em que um empréstimo ou uma obrigação *defaultable* (sujeitos ao risco de *incumprimento*) considera-se a probabilidade de incumprimento, a taxa de recuperação e o *spread*.

Assim, o *cash-flow* de uma obrigação, para $t = 1$, é 100 com probabilidade de incumprimento $(1-p)$ no caso de não *incumprimento* e $\lambda \times 100$ com probabilidade p no caso de ocorrência

de *incumprimento*. Neste caso, λ representa a recuperação após o *incumprimento* de uma obrigação.

A Tabela 4.1 ilustra o valor esperado dos *cash-fows* do empréstimo ou obrigação nos dois cenários.

Tabela 4.1: *Cash-Flows a um período*

Tempo	cumpridor	incumpridor
t=0	100	100
t=1	100(1+r)	[100(1-p)+λ100p](1+r)(1+s)

Observa-se que, para $\lambda < 1$, se tem:

$$100(1-p) + \lambda 100p < 100$$

daí o *spread* representar uma taxa de prémio de risco que relaciona o investimento de risco com o investimento sem risco. Assim, em termos de expectativas de mercado, teríamos que:

$$100(1+r) = [100(1-p) + \lambda 100p](1+r)(1+s) \tag{4.3}$$

caso contrário, ninguém investiria na obrigação menos rentável. Assim, consideremos o seguinte teorema:

Teorema 4.1. *No modelo discreto a um período, para o risco de crédito, o spread é função da taxa de recuperação e da probabilidade de incumprimento na medida natural de probabilidade,* $\mathbb{P}$*, e é dado por:*

$$s = \frac{(1-\lambda)p}{1-(1-\lambda)p} \tag{4.4}$$

Demonstração: Para demonstração do resultado é suficiente resolver a equação (4.3) em ordem a s.

Proposição 4.1. *Nas situações em que se verifica que* $(1-\lambda)p << 1$ *tem-se um resultado clássico:*

$$s \approx (1-\lambda)p$$

Demonstração: O resultado enunciado deduz-se facilmente a partir da equação 4.4.

4.2.3 Modelo de uma carteira a tempo discreto

Nesta subsecção, propomos um modelo para a determinação do *spread*, que será aplicado a uma carteira de crédito; este é uma extensão natural do modelo período clássico para obrigações de cupão zero descritos na subsecção 4.2.2. Antes descrevemos as seguintes definições e notações:

- seja $(X_n)_{n\in\{0,1,...,T\}}$ um processo estocástico que descreve os valores das obrigações da carteira. Considera-se $(\Omega, \mathcal{A}, \mathbb{P})$ um espaço de probabilidade em que as variáveis aleatórias X_n, para $n \in \{0, 1, ..., T\}$, são definidas para cada $\omega \in \Omega, X_n(\omega)$ é o valor da obrigação do cliente ω na data $n \in \{0, 1, ..., T\}$. Suponhamos que as obrigações na carteira, representada por $(X_n)_{n\in\{0,1,...,T\}}$ não estão sujeitas a incumprimento.
- introduzimos neste ponto a probabilidade de incumprimento para as obrigações nas carteiras. Assim, consideremos $(\tilde{X}_n)_{n\in\{0,1,...,T\}}$ variáveis aleatórias que denotam o processo estocástico descrito para os valores das obrigações, as quais estão agora sujeitas a incumprimento.
- incumprimento pode ocorrer como um tempo aleatório τ, que designaremos por **tempo de incumprimento da carteira**, de modo que para cada $n \in \{0, 1, ..., T\}$, se tem $\{\tau \geq n\} \in \mathcal{A}$. Adicionalmente, uma hipótese natural, é que τ é o tempo de paragem relativamente à filtração natural do processo de $(X_n)_{n\in\{0,1,...,T\}}$, ou seja $\{\tau \geq n\} \in \mathcal{A}_n$, com $\mathcal{A}_n$ a sigma-algebra gerada por X_k, para $0 \leq k \leq n$. O significado desta hipótese é que o tempo de início do incumprimento τ é perfeitamente definido por X_k, variáveis aleatórias da carteira até ao tempo presente para qualquer n. Devido à natureza do nosso modelo não iremos precisar desta hipótese.

Para o modelo em estudo, propomos os seguintes pressupostos:

1. Existe uma função $F : \mathbb{R} \times \Omega \to \mathbb{R}$ e algum parâmetro $\lambda \in [0, 1]$, chamado **taxa de recuperação da carteira** do modelo de tal forma que com o tempo de maturidade T, temos:
$$\mathbb{E}[F(X_T, \bullet)] = \lambda \mathbb{E}[X_T].$$
2. consideramos que, para esta função F no tempo de maturação T se tem:
$$\tilde{X}_T = X_t \mathbb{I}_{\{\tau>T\}} + F(X_T, \bullet)\mathbb{I}_{\{\tau>T\}}$$
3. Consideremos a taxa de juro de risco r e o *spread* s, na maturidade T denotado por s_T, são ambos constantes.
4. O processo da carteira de obrigações $(X_n)_{n\in\{0,1,...,T\}}$ e o tempo de incumprimento τ da carteira são independentes.

Com este conjunto de hipóteses mostramos que, pelo princípio do valor esperado da metodologia actuarial, o *spread* s_T é uma função da probabilidade de incumprimento e da taxa de recuperação da carteira.

Teorema 4.2. *No âmbito do princípio do valor esperado da metodologia actuarial, temos que,*

$$\mathbb{E}\left[\frac{X_t}{(1+r)^T}\right] = \mathbb{E}\left[\frac{\tilde{X}_T}{(1+r)^T}(1+s_T)\right] \tag{4.5}$$

e se $\mathbb{E}[X_T] \neq 0$*, o spread na data* T *é dado por:*

$$s_T = \frac{(1-\lambda)\mathbb{P}[\tau \leq T]}{1-(1-\lambda)\mathbb{P}[\tau \leq T]}. \tag{4.6}$$

Demonstração: Através das hipóteses em cima, temos que a expressão (4.5) implica que

$$\mathbb{E}[\tilde{X}_T] = \mathbb{E}[X_t]\,\mathbb{E}[\mathbb{I}_{\{\tau>T\}}] + \lambda\,\mathbb{E}[X_t]\,\mathbb{E}[\mathbb{I}_{\{\tau>T\}}(1+s_T)]$$

e efectuando algumas reorganizações, obteremos a expressão (4.6)

Proposição 4.2. *Nas situações em, que se verifica que* $(1-\lambda)\mathbb{P}[\tau \leq T] << 1$ *obtemos o seguinte resultado:*

$$s_T = (1-\lambda)\mathbb{P}[\tau \leq T],$$

que corresponde a uma fórmula semelhante à do modelo de uma obrigação de cupão zero.

Demonstração: O resultado enunciado deduz-se facilmente a partir da equação (4.6).

4.2.4 Modelo para a estimativa da taxa de recuperação da carteira

Para aplicação do modelo apresentado na subsecção 4.2.3, torna-se necessário estimar a taxa de recuperação da carteira, λ, e a probabilidade de incumprimento da carteira, $\mathbb{P}[\tau \leq T]$. Se, na ocorrência do incumprimento sabemos o montante pago na data de maturidade, podemos estimar λ, através de uma regressão linear simples dada por:

$$F(X_T, \bullet) = \lambda X_T + \varepsilon \tag{4.7}$$

em que

1. Modelo é linear;
2. $\mathbb{E}(\varepsilon_i|x_i) = 0$;
3. $\mathbb{V}ar(\varepsilon_i) = \sigma^2$ (Homocedasticidade);
4. $cov(\varepsilon_i,\varepsilon_j \neq 0)$ (Ausência de autocorrelação)

5. $\varepsilon_i \sim N(0, \sigma^2)$ (Normalidade);

em que ε representa o erro da regressão. Se considerarmos que $\mathbb{E}[\varepsilon] = 0$, é evidente que o pressuposto 1 do modelo da subsecção 4.2.3 é verificada.

De uma forma geral, em estatística, o diagnóstico do modelo é possivelmente o passo mais importante no processo da construção do modelo. A literatura especializada fornece-nos diferentes métodos de diagnóstico de modelos, nomeadamente: R^2, Cook's Distance, Q-Q Plot, resíduos de Jack-knife. Neste estudo, focalizamo-nos apenas nos métodos de R^2 e na análise dos resíduos.

Convém salientar que não constitui objectivo principal, neste trabalho, uma análise profunda desses métodos. Contudo, como a base de dados utilizada é real, a análise e o estudo desta torna-se imprescindível. Por outro lado, as suposições do modelo ajustado devem de ser diagnosticadas para que os resultados sejam fiáveis. Assim, de forma sucinta, descrevemos o coeficiente de regressão e os resíduos.

O coeficiente de determinação, R^2, é uma medida muito utilizada na regressão linear múltipla, quantificando a proporção da variação da variável dependente, Y, explicada pelas variáveis independentes. Seja $\mathbb{E}(Y_i|x_i) = m_i = \boldsymbol{\beta}\boldsymbol{x}_i^T$, em que X_i é o vector $1 \times p$ das variáveis explicativas para o indivíduo i e β é o vector $1 \times p$ dos coeficientes da regressão

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{m}_i - \hat{m})^2}{\sum_{i=1}^{n}(Y_i - \hat{m})^2}, \tag{4.8}$$

onde $\hat{m} = \hat{\boldsymbol{\beta}} x_i^T$, $\hat{\boldsymbol{\beta}}$ é a estimativa dos mínimos quadrados de $\boldsymbol{\beta}$ e $\hat{m} = \sum_i \frac{Y_i}{n}$.

Uma outra medida que iremos utilizar na validação do modelo é a análise dos resíduos. A análise dos Resíduos constitui um conjunto de técnicas utilizadas para investigar a adequabilidade de um modelo de regressão com base nos resíduos. O resíduo, ε, é dado pela diferença entre a variável resposta observada (X_i) e a variável resposta estimada (Y_i), isto é

$$\varepsilon_i = Y_i - \hat{Y}_i \ , \ i = 1, ..., n.$$

A ideia básica da análise dos resíduos é que, se o modelo for apropriado, os resíduos devem reflectir as propriedades impostas pelo termo de erro do modelo. A análise de resíduos é fundamental, não só para avaliar localmente a qualidade de ajustamento do modelo, no que diz respeito à escolha da distribuição, da função de ligação e de em termos do preditor linear, como também para ajudar na identificação de observações mal ajustadas (observações que não são bem explicadas pelo modelo). Um resíduo ε_i deve exprimir a discrepância entre o valor observado y_i e o valor $\hat{y}_i$ ajustado pelo modelo [Turkman e Silva, 2000].

4.3 Aplicação

Neste capítulo utilizamos o menu *Regression* da versão *Enterprise Miner* 4.3 do SAS Assim, nesta secção apresentamos os resultados empíricos da seguinte forma: a descrição da carteira em estudo faz-se na subsecção 4.3.1; na subsecção 4.3.2 estimamos a taxa de recuperação; na subsecção 4.3.3 apresentamos os resultados do cálculo do *spread* da carteira. Por fim, apresentamos os resultados do *spread* para os clientes na subsecção 4.3.4.

4.3.1 Carteira de crédito ao consumo

Nesta dissertação, como já tinhamos referido no Capítulo 2, a base de dados utilizada foi fornecida pelo um banco comercial de Cabo Verde, extraída em Outubro de 2011, com informações socio-económicas e financeiras dos clientes do crédito ao consumo no período comprendido entre Janeiro de 2003 e Outubro de 2011. Atendendo ao objectivo deste capítulo, que consiste em estimar o *spread* da carteira utilizando a metodologia proposta na subsecção 4.2.3, é necessário estimar a probabilidade de incumprimento e a taxa de recuperação da carteira.

Os dados utilizados neste capítulo, foram extraídos da base de dados acima referida.

Esta carteira é constituída por clientes que nunca entraram em incumprimento, e por clientes que pelo menos uma vez tiveram mais de 90 dias em atraso, aqui designados, por incumpridores, à semelhança da definição utilizada no Capítulo 2.

Enquanto que os clientes incumpridores viram os seus planos alterados pelo banco afim de cobrir as despesas causados pelo incumprimento, tal não sucede com os clientes cumpridores. A Tabela 4.2 faz uma partição dos clientes da carteira utilizado de acordo com a sua situação de incumprimento (em que consideramos no Caso I clientes em incumprimento na data de vencimento do crédito e caso II clientes em incumprimento até data de vencimento de crédito).

Tabela 4.2: *Amostra dos clientes*

	Cumpridores		Incumpridores	
	Nº	%	Nº	%
Caso I	22677	95.5	1066	4.5
Caso II	22275	95.8	1468	6.2

Considera-se, para este modelo todos os clientes com empréstimos entre 30.000 e 1.800.000 ECV e com prestações totais máximo de contrato de 48 meses. De modo a avaliar a consistência do modelo da taxa de recuperação, que iremos desenvolver na subsecção seguinte,

dividimos a amostra em quatro grupos: clientes com prestações totais de empréstimo até 12 meses, com prestações totais superior a 12 e inferior ou igual a 24 meses, prestações totais superior a 24 e inferior ou igual a 36 meses e, por fim, clientes com prestações totais de empréstimo superior a 36 meses.

A Tabela 4.3 ilustra os resultados dessa partição da amostra.

Tabela 4.3: *Distribuição dos incumpridores por prestações totais*

Prestações totais (em meses)	Caso I		Caso II	
	Nº	%	Nº	%
≤ 12	148	13,88	188	14,81
> 12 e ≤ 24	430	40,34	599	40,8
> 24 e ≤ 36	316	29,64	438	29,84
> 36	172	16,14	243	16,55
Total	1066	100	1468	100

4.3.2 Estimação da taxa de recuperação da carteira

Uma das questões que importa analisar ao estudar o risco de incumprimento prende-se com o montante de crédito não recuperado para os clientes que não pagam completamente os seus empréstimos. Assim, é fundamental que uma instituição financeira estime a taxa de recuperação da carteira de crédito, ou seja, que obtenha uma estimativa da proporção do montante a recuperar.

Considerou-se que, quando um cliente é incumpridor, o montante em dívida corresponde ao valor que o cliente deve à instituição bancária à data de encerramento do seu contrato.

Neste estudo, a estimação da taxa de recuperação é feita de acordo com o modelo proposto na secção 4.3.2 e segmentada por prestações totais como mostra a Tabela 4.3.

Os resultados das estimativas através dos Mínimos Quadrados Ordinais (MQO), para ambos os casos e para cada um dos modelos, encontram-se na Tabela 4.4. Nesta tabela ilustra-se ainda as estimativas para o Modelo Global, ou seja, não considerando a partição dos contratos de acordo com o prestações totais. Note-se que todos os modelos ajustados são estatísticamente significativos. O modelo que reflecte um melhor ajustamento é o modelo24, quer para o Caso I quer para o Caso II. O valor esperado da taxa de recuperação é mais elevado no modelo24 do caso II $(0,693)$. Isto significa que a taxa de recuperação da carteira, para os clientes com empréstimos entre 12 e 24 meses, é de $69,3\%$. Ainda, de acordo com a Tabela 4.4, o modelo global apresenta uma taxa de recuperação $63,2\%$ e $66,7\%$ para cada um dos casos, respectivamente.

Tabela 4.4: *Estimativas dos MQO*

Modelos	Estimativas		Standard Error		Pr>F	R^2	
	Caso I	Caso II	Caso I	Caso II		Caso I	Caso II
Modelo12	0,527	0,587	0,020	0,019	<.0001	0,832	0,842
Modelo24	0,637	0,693	0,012	0,010	<.0001	0,870	0,892
Modelo36	0,608	0,635	0,017	0,015	<.0001	0,811	0,811
Modelo48	0,655	0,681	0,021	0,019	<.0001	0,851	0,848
Modelo Global	**0,635**	**0,667**	**0,008**	**0,007**	**<.0001**	**0,843**	**0,847**

Relativamente ao Modelo Global, avaliámos também o modelo da estimação da taxa de recuperação através do erro quadrático Médio e da análise dos resíduos. Os resultados dos resíduos encontram-se na Figura 4.1.

O ajustamento obtido reflete um R^2 de $84,3\%$ e $84,7\%$ para ambos os casos, respectivamente. Pode concluir-se que o modelo apresenta uma boa qualidade de ajustamento. Assim, através do coeficiente de determinação com as características dos clientes agrupadas, verifica-se que as variáveis explicativas conseguem explicar aproximadamente 85% do percentual da recuperação, ou seja, outras variáveis explicativas podem estar influenciar a variável em estudo.

Após a análise das medidas de qualidade de ajustamento do modelo, tem que:

$$F(X_T, \bullet) = \lambda X_T + \varepsilon \tag{4.9}$$

com

- ε - erro de regressão;
- X_T é o valor de empréstimo;
- $F(X_T, \bullet)$ é o montante devido.

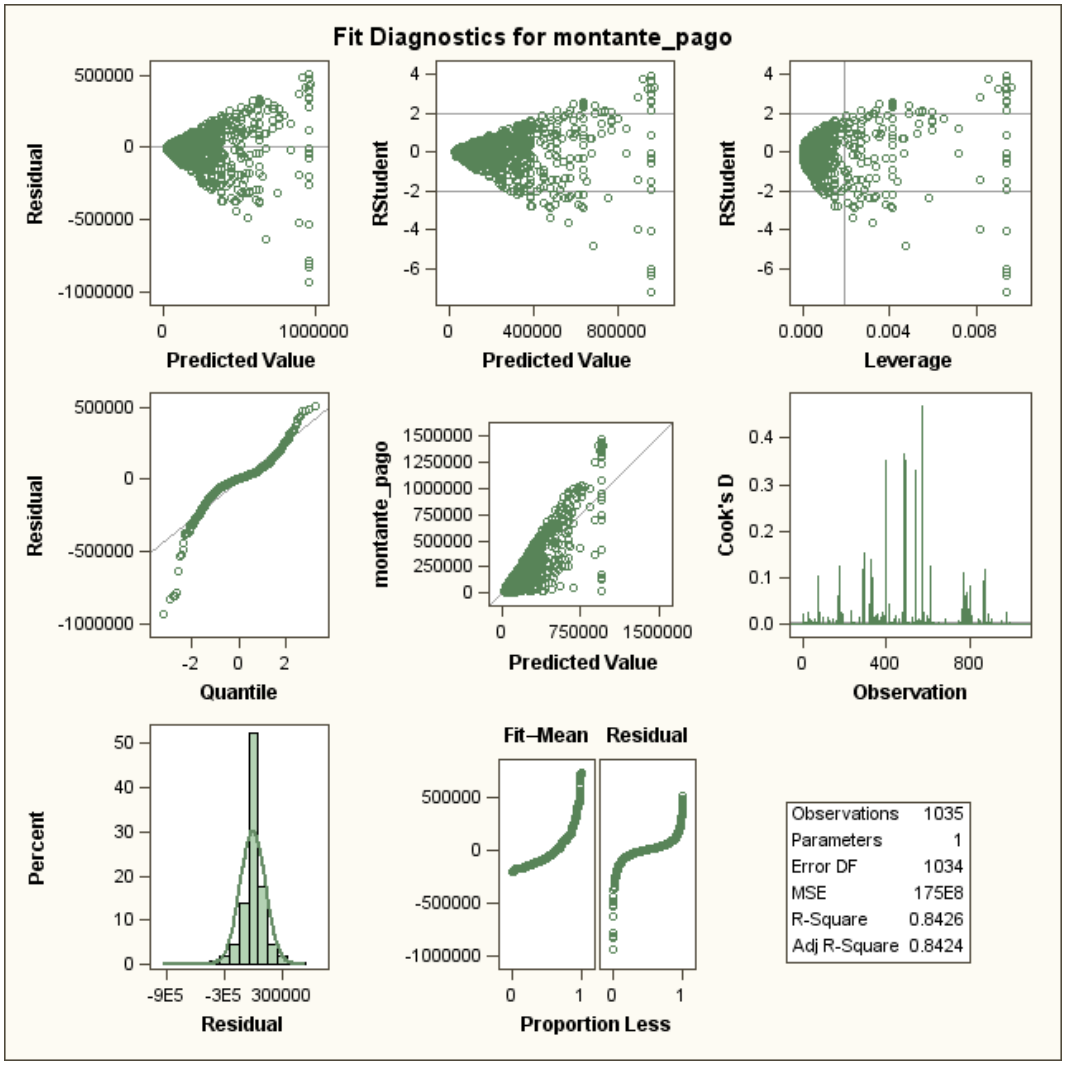
O modelo estimado toma a seguinte forma:

$\widehat{F(X_T, \bullet)} = 0.635$*valor de empréstimo e $\widehat{F(X_T, \bullet)} = 0.667$*valor de empréstimo para caso I e caso II, respectivamente. Assim, a estimativa da taxa de recuperação da carteira de $63,2\%$ e $66,7\%$ é aceitável, dadas as condições da carteira em estudo.

4.3.3 *Spread* da carteira

Através da fórmula da 4.6, da probabilidade de incumprimento estimada no capítulo 3 e da taxa da recuperação estimada neste capítulo o *spread* aplicada à carteira é aproximadamente

Figura 4.1: *Ajustamento do modelo da regressão e progonóstico da taxa de recuperação - caso I*



0.036 e pela fórmula usual da Proposição 4.2 é 0.034. Conclui-se que o erro relativo da fórmula do *spread* proposto, neste estudo, através da metodologia actuarial tem um erro relativo de aproximadamente $0,12\%$ em relação à fórmula usual da Proposição 4.2. Conclui-se que a diferença entre uma e outra fórmula é mínima. Assim, consideramos aceitável a aplicação da fórmula 4.6 proposto neste capítulo.

4.3.4 *Spread* do cliente

Nesta subsecção apresentamos, na Tabela 4.5 o valor do *spread* mínimo que se deve atribuir a cada cliente, estimado com base na taxa de recuperação constante igual a 0.635, estimada pelo modelo 4.9 e pela probabilidades de incumprimento, segundo a Equação 4.6. Ainda,

nesta Tabela 4.5 encontram-se as probabilidades de incumprimento estimadas no capítulo 2, conforme se tem vindo a descrever, para as caracteristicas socio-demográficas e financeiras dos clientes.

Tabela 4.5: *Probabilidade de Incumprimento e Spread Estimadas*

	Características								
Cliente	V. crédito	V. prestação	Actividade Profissional	Entidade Patronal	Taxa Nominal	Agência	Prestações Pagas	PD	*Spread*
1	387.000	12.930	Emp.Esc/serv/Comer	Câmara Municipal	12.5	8	36	0.105	0.039
2	210.000	12.857	Emp.Esc/serv/Comer	pme's	12.5	12	18	0.293	0.012
3	538.000	18.004	Emp.Esc/serv/Comer	Ministérios	12.5	10	36	0.078	0.029
4	300.000	18.370	Lib/ Q Sup	Não declarou	12.5	1	18	0.21	0.083
5	600.000	20.083	Outros	Grandes Empresas	12.5	4	36	0.198	0.078
6	300.000	14.197	Outros	PME's	12.5	8	24	0.351	0.147
7	400.000	10.633	Outros	PME's	12.5	5	48	0.471	0.207
8	368.500	22.551	Emp.Esc/serv/Comer	Grandes Empresas	12.5	1	18	0.094	0.036
9	505.000	13.416	Outros	Grandes Empresas	12.5	10	48	0.135	0.052
10	100.000	6.131	Estudante	Não declarou	12.5	12	18	0.209	0.083
11	1.500.000	39.873	Outros	conta própria	12.5	5	48	0.44	0.191
12	290.000	9.685	Emp.Esc/serv/Comer	Ministérios	12.5	1	36	0.113	0.043
13	581.000	19.420	Oper Eesp	Ministérios	12.5	8	36	0.096	0.036
14	1.500.000	39.651	Emp.Esc/serv/Comer	Não declarou	12.5	5	48	0.271	0.11
15	91.500	5.546	Emp.Esc/serv/Comer	pme's	11	5	18	0.108	0.041
16	145.200	8.896	Emp.Esc/serv/Comer	pme's	12.5	1	18	0.236	0.094
17	209.000	3.371	Outros	Não declarou	12.5	1	48	0.289	0.117
18	720.000	44.084	Peq / Med Empres	pme's	12.5	3	18	0.204	0.08

PD - Probabilidade de incumprimento

4.4 Modelo para a estimação da proporção de recuperação de crédito para o cliente

Em qualquer processo industrial, pode ser definido um conjunto de causas ou factores que produzem determinado efeito sobre uma ou mais características de qualidade de um produto. Esta secção apresenta um modelo de regressão com resposta contínua entre zero e um (modelo de Regressão Beta) a ser utilizado na estimação da recuperação de um cliente incumpridor de uma carteira de crédito ao consumo através das variáveis socio-demográficas e económicas. Ademais, na construção destes modelos as variáveis independentes ou fatores controláveis podem ser de natureza quantitativa ou qualitativa.

4.4.1 Regressão Beta

A análise de regressão é uma técnica estatística utilizada para investigar e modelar, com base em um banco de dados, a relação entre uma variável de interesse e um conjunto de variáveis explicativas. O modelo de regressão normal linear é bastante utilizado em análises empíricas. Contudo, tal modelo torna-se inadequado em situações em que a variável resposta é restrita ao intervalo (0,1), como ocorre com taxas e proporções.

A classe dos modelos de Regressão Beta, introduzida em [Ferrari e Cribari-Neto, 2004] é útil para modelar as variáveis contínuas y que assumem valores no intervalo (0,1). O objectivo é definir um modelo em que a variável de interesse assume uma distribuição beta. O modelo de Regressão Beta baseia-se numa parametrização alternativa da densidade beta em termos da média das variáveis e do parâmetro de precisão. A densidade da distribuição Beta é geralmente expressa como

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}\ ,\ 0<y<1, \tag{4.10}$$

onde $p, q > 0$ e $\Gamma(\cdot)$ é função Gama.

Consideremos $y_1, \ldots, y_n$ observações independentes, tais que cada y_i , $i = 1, \ldots, n$, possuem distribuição Beta. Supondo uma reparametrização em termos de média e precisão, podemos escrever

$$f(y_i|\mu_i;\phi) = \frac{\Gamma(\phi)}{\Gamma(\phi\mu_i)\Gamma(\phi(1-\mu_i))} y_i^{\phi\mu_i-1}(1-y_i)^{\phi(1-\mu_i)-1}\ ,\ 0<y_i<1, \tag{4.11}$$

onde $\mathbb{E}[y_i] = \mu_i$ e $Var(y_i) = [\mu_i(1-\mu_i)]/(\phi+1)$. Assumindo que a média $\mu_i (i = 1, \ldots, n)$ é uma função não-linear de variáveis explicativas conhecidas, daí escrevemos que

$$g(\mu_i) = \boldsymbol{x_i}\boldsymbol{\beta} = \eta_i.$$

onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ é um vector $k \times 1$ de parâmetros desconhecidos ($p < n$) e $\boldsymbol{x_i} = (x_{i0}, \ldots, x_{ip})$ representa o vector de covariáveis associadas à i-ésima observação. Assumimos

g como sendo a função de ligação logística. Também assumiremos a distribuição a priori normal para $\boldsymbol{\beta}$ e uma distribuição a priori log-normal para ϕ com média zero e variância σ_ϕ^2 conhecida, escolhida com o intuito de facilitar a análise feita através das aproximações determinísticas. Daí, o modelo pode ser escrito da seguinte forma:

$$y_i \sim Beta(\phi\mu_i, \phi(1-\mu_i)), \quad i = 1, \ldots, \quad n \quad \text{independentes}$$
$$log(\mu_i) = log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x_i\beta}$$
$$\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{cI})$$
$$\phi \sim Log - Normal(0, \sigma_\phi^2)$$

onde $\boldsymbol{X} = (x_1, \ldots, x_n)$.

Existem vários tipos de resíduos para avaliar o modelo de Regressão Beta. Neste estudo usamos o resíduo de Pearson, o qual [Ferrari e Cribari-Neto, 2004] denominou de *standardized ordinary residuals* e o definiu como

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{VAR}(y_i)}} \tag{4.12}$$

onde $\widehat{VAR}(y_i) = \frac{\hat{\mu}_i(1-\hat{\mu}_i)}{1+\hat{\phi}_i} \quad \hat{\mu}_i = g^{-1}(x_i\hat{\beta})$.

Muitos estudos, em diferentes áreas do conhecimento, como em [Brehm e Gates, 1993], [Hancox et al., 2010], [Kieschnick e McCullough, 2003] e [Smithson e Verkuilen, 2006], utilizam Regressão Beta ou outras abordagens para examinar como um conjunto de covariáveis se relaciona com alguma porcentagem ou proporção.

4.4.2 Resultados da estimação da recuperação

Nesta subsecção apresentemos alguns resultados preliminares da estimação da taxa de recuperação através da metodologia apresentada na subsecção 4.4.1.

Analogamente ao capítulo 2 todas as variáveis independentes foram categorizadas com o apoio do nó *Interactive Grouping* (ING) do software *SAS*® (*Enterprise Miner Client* 6.1, ver a Tabela 4.6). Foram utilizadas as mesmas variáveis utilizadas para a estimação da probabilidade de incumprimento no Capítulo 2, ver a Tabela 2.5.

Tabela 4.6: *Variáveis Categorizadas*

CodVariável	Variáveis	Grupo	Cumpridor	Incumpridor	Total
PRESTPAGASPP1	Prestações Pagas	1	3100	730	3830
PRESTPAGASPP2		2	2207	585	2792
PRESTPAGASPP3		3	3399	963	4362
PRESTPAGASPP4		4	3722	1115	4837
AGAG1	Agência	1	1005	554	1559
AGAG2		2	1289	570	1859
AGAG3		3	603	203	806
AGAG4		4	5912	1462	7374
AGAG5		5	3619	604	4223
TXNTXN1	Taxa Nominal	1	1663	558	2221
TXNTXN2		2	10765	2835	13600
PRESTPREST1	Valor Prestação	1	2912	850	3762
PRESTPREST2		2	3224	918	4142
PRESTPREST3		3	2507	663	3170
PRESTPREST4		4	3785	962	4747
EMPRESTIMOEMPT1	Valor Emprestimo	1	2352	812	3164
EMPRESTIMOEMPT2		2	2490	701	3191
EMPRESTIMOEMPT3		3	4984	1317	6301
EMPRESTIMOEMPT4		4	2602	563	3165
AGAG1	Idade	1	1364	527	1891
AGAG2		2	949	271	1220
AGAG3		3	1748	538	2286
AGAG4		4	4139	1067	5206
AGAG5		5	4228	990	5218
PROFPROF1	Actividade Profissional	1	1210	469	1679
PROFPROF2		2	3606	1227	4833
PROFPROF3		3	897	215	1112
PROFPROF4		4	6715	1482	8197
SEXF	Género	1	5212	1240	6452
SEXM		2	7216	2153	9369
ENTPATENTPAT1	Entidade Patronal	1	1111	181	1292
ENTPATENTPAT2		2	6541	1451	7992
ENTPATENTPAT3		3	1602	415	2017
ENTPATENTPAT4		4	1885	758	2643
ENTPATENTPAT5		5	1289	588	1877
ESTCIVILESTCIVIL1	Estado Civil	1	3405	734	4139
ESTCIVILESTCIVIL2		2	9023	2659	11682
HABILITHABIL1	Habilitações	1	2326	758	3084
HABILITHABIL2		2	6418	1757	8175
HABILITHABIL3		3	2285	553	2838
HABILITHABIL4		4	687	162	849
HABILITHABIL5		5	712	163	875
GARANTIAGARNATIA1	Tipo de Garantia	1	4340	1310	5650
GARANTIAGARNATIA2		2	8088	2083	10171
PRAZOPRZ1	Prestações Totais	1	2694	461	3155
PRAZOPRZ2		2	2180	508	2688
PRAZOPRZ3		3	3976	989	4965
PRAZOPRZ4		4	2496	886	3382
PRAZOPRZ5		5	1082	549	1631

Após a categorização das variáveis independentes, desenvolvemos um modelo de recuperação para os clientes que entraram em incumprimento, utilizando a Regressão Beta.

A Tabela 4.7 apresenta as estimativas dos parâmetros, o desvio padrão e as estatísticas Z de todas as variáveis utilizadas na estimação do modelo de recuperação. A estimação do modelo foi feito através dos estimadores de máxima verosimilhança.

Tabela 4.7: *Modelo com todas as variáveis*

Parâmetros	Estimativas	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	0.841669	0.110870	7.591	3.16e-14	***
PRESTPAGASPP2	1.032	0.064885	15.903	< 2e-16	***
PRESTPAGASPP3	1.558	0.059371	26.246	< 2e-16	***
PRESTPAGASPP4	2.744	0.067322	40.760	< 2e-16	***
TXNTXN2	-0.038434	0.045806	-0.839	0.401440	
PRESTPREST2	0.065727	0.050244	1.308	0.190818	
PRESTPREST3	0.098712	0.064241	1.537	0.124393	
PRESTPREST4	0.043284	0.078160	0.554	0.579724	
EMPRESTIMOEMPT2	-0.075639	0.046801	-1.616	0.106058	
EMPRESTIMOEMPT3	-0.278201	0.064846	-4.290	1.79e-05	***
EMPRESTIMOEMPT4	-0.393226	0.095285	-4.127	3.68e-05	***
IDAIDA2	-0.062096	0.058856	-1.055	0.291400	
IDAIDA3	-0.034018	0.049319	-0.690	0.490353	
IDAIDA4	-0.001008	0.044282	-0.023	0.981833	
IDAIDA5	0.076118	0.047510	1.602	0.109127	
PRAZOPRZ2	-0.605682	0.076309	-7.937	2.07e-15	***
PRAZOPRZ3	-1.045	0.073425	-14.238	< 2e-16	***
PRAZOPRZ4	-2.012	0.079918	-25.174	< 2e-16	***
PRAZOPRZ5	-1.988	0.086334	-23.030	< 2e-16	***
AGAG2	0.074957	0.049528	1.513	0.130171	
AGAG3	0.069993	0.070482	0.993	0.320683	
AGAG4	-0.002082	0.042377	-0.049	0.960815	
AGAG5	-0.041405	0.048112	-0.861	0.389467	
PROFPROF2	0.186860	0.045773	4.082	4.46e-05	***
PROFPROF3	0.189290	0.071458	2.649	0.008073	**
PROFPROF4	0.187121	0.050218	3.726	0.000194	***
SEXM	-0.010490	0.028866	-0.363	0.716303	
ENTPATENTPAT2	0.096773	0.063023	1.536	0.124654	
ENTPATENTPAT3	0.011359	0.071428	0.159	0.873652	
ENTPATENTPAT4	-0.091397	0.068687	-1.331	0.183311	
ENTPATENTPAT5	-0.151465	0.068615	-2.207	0.027281	*
ESTCIVILESTCIVIL2	-0.049570	0.036719	-1.350	0.177020	
HABILITHABIL2	0.112137	0.038297	2.928	0.003411	**
HABILITHABIL3	0.186088	0.049394	3.767	0.000165	***
HABILITHABIL4	0.166074	0.075594	2.197	0.028026	*
HABILITHABIL5	0.161171	0.076180	2.116	0.034373	*
GARANTIAGARNATIA2	0.037444	0.031553	1.187	0.235351	

As estimativas dos parâmetros do modelo final estão apresentadas na Tabela 4.8.

Tabela 4.8: *Modelo final*

Parâmetros	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	0.79385	0.05180	15.324	< 2e-16	***
PRESTPAGASPP2	1.040	0.06431	16.167	< 2e-16	***
PRESTPAGASPP3	1.567	0.05762	27.195	< 2e-16	***
PRESTPAGASPP4	2.736	0.06383	42.864	< 2e-16	***
PRAZOPRZ2	-0.61943	0.07457	-8.306	< 2e-16	***
PRAZOPRZ3	-1.076	0.06598	-16.314	< 2e-16	***
PRAZOPRZ4	-2.020	0.07013	-28.810	< 2e-16	***
PRAZOPRZ5	-1.994	0.07461	-26.729	< 2e-16	***
EMPRESTIMOEMPT3	-0.21845	0.03197	-6.833	8.30e-12	***
EMPRESTIMOEMPT4	-0.33668	0.04517	-7.454	9.06e-14	***
PROFPROF2	0.23432	0.04475	5.236	1.64e-07	***
PROFPROF3	0.24428	0.07021	3.479	0.000503	***
PROFPROF4	0.27921	0.04587	6.087	1.15e-09	***
ENTPATENTPAT5	-0.17956	0.03609	-4.976	6.50e-07	***
HABILITHABIL2	0.09689	0.03762	2.575	0.010016	*
HABILITHABIL3	0.18061	0.04773	3.784	0.000154	***
HABILITHABIL4	0.19398	0.07481	2.593	0.009516	**
HABILITHABIL5	0.15116	0.07540	2.005	0.044991	*

Assim propõe-se para a carteira de crédito ao consumo de um banco de Cabo Verde, um modelo de recuperação com as seguintes variáveis: número de prestações pagas, prestações totais, valor de empréstimo, actividade profissional, entidade patronal e habilitações literárias. Contudo, remetemos para estudos futuros uma análise aprofundada deste modelo.

4.5 Considerações Finais e Estudos Futuros

Considerações Finais

Encerramos este capítulo com uma síntese relativa à nossa contribuição para uma proposta simples que define o *spread* para uma carteira de crédito ao consumo de um banco de Cabo Verde. Também mostramos que pelo princípio do valor actuarial, a metodologia actuarial do *spread* S_T é uma função de probabilidade de incumprimento e da taxa da recuperação da carteira. Em anexo apresentamos as linhas gerais da possibilidades de continuação e aperfeiçoamento da estimação do *spread* para o cliente em que a taxa de recuperação estimada pelas características socio-demográficas, financeiras e comportamentais de cada cliente.

Em termos teóricos propomos uma extensão natural do modelo a um período para obrigações de cupão zero.

O modelo proposto para o *spread* da carteira de crédito ao consumo dos clientes de um banco de Cabo Verde apresenta os seguintes resultados:

- uma taxa de recuperação de 63.5% para clientes que entraram em incumprimento ao longo do empréstimo;
- o spread médio da carteira igual a 0.036 para o modelo proposto;
- através da Equação 4.6 calculamos o valor do *spread* mínimo que se deve atribuir a cada cliente, estimado com base na taxa de recuperação constante igual a 0.635 e pelas probabilidades de incumprimento estimadas no capítulo 2, conforme se tem vindo a descrever, para as caracteristicas socio-demográficas e financeiras dos clientes.

De acordo com o modelo de recuperação desenvolvido para os clientes incumpridores da carteira, as variáveis que melhor explicam, de entre todas analisadas são número de prestações pagas, prestações totais, valor de empréstimo, actividade profissional, entidade patronal e habilitações literárias.

Estudos Futuros

Remetemos para os estudos futuros a modelação do *spread* do cliente através da probabilidade de incumprimento e da recuperação de cada cliente em função das suas características sócio-demográficas e económicas.

Capítulo 5

Conclusão

Esta dissertação teve como objectivo geral, em termos práticos, estimar a probabilidade de incumprimento e do *spread* da carteira e do cliente de uma carteira de crédito ao consumo de um banco de Cabo Verde. Assim, estimou-se a probabilidade de incumprimento: da carteira através da definição do número de dias de incumprimento; dos clientes através da Regressão Logística; ao longo do tempo através de um modelo para populações abertas sujeitas a reclassificações períodicas designado por Vórtices Estocásticos. Estimou-se também um modelo de *spread* para a carteira e para o cliente de uma carteira de crédito ao consumo de um banco de Cabo Verde.

Os resultados dos modelos de probabilidade de incumprimento estimadas através da Regressão Logística e do scorecard, esse último para efeito de validação, apresentam-se resultados bastantes semelhantes para as medidas como a cruva ROC, estatísticas de KS, Índice de Gini e matriz de classificação. A carteira de crédito utilizada é constituída por um conjunto de treze variáveis socio-demográficas, financeiras e comportamentais. O modelo seleccionado agrega as seguintes variáveis: número de prestações pagas, agências, taxa nominal, valor da prestação, valor de empréstimo, actividade profissional e entidade patronal.

Em termos teóricos introduzimos novos resultados ao nível da convergência do modelo dos Vórtices Estocásticos. Os resultados aqui obtidos generalizam os anteriormente estudados nos trabalhos de Guerreiro e Guerreiro et al. Em termos práticos, propomos um novo ajustamento, a sigmoidal, aplicada a uma carteira de crédito ao consumo, em que verificamos pelos vários teste desenvolvidos, nesta dissertação, que os resultados da sigmoidal são melhores em relação ao ajustamento da exponencial utilizados nos estudos de Guerreiro e Guerreiro et al.

Ainda em termos teóricos, relativamente ao modelo do *spread* propomos também, um modelo a tempo discreto que permite estimar o *spread* da carteira em função da probabilidade de incumprimento, estimada através das variáveis socio-demográficas financeiras e comporta-

mentais e da taxa de recuperação de cada cliente em incumprimento. Além dos desenvolvimentos teóricos, evidencia-se a aplicação feita a uma carteira de crédito dos clientes com todas as prestações pagas.

Os estudos efectuados, e a análise das várias metodologias, permitiram-nos concluir que o risco de incumprimento da carteira e de cada cliente deve ser estimado e analisado ao longo do tempo e o *spread* cobrado deve ser o adequado para fazer face a esse risco, e a proporcionar lucro à instituição credora. Verificamos que a estimação da probabilidade de incumprimento tem um papel preponderante, mas não suficiente na estimação do *spread* adequado a cada cliente.

No decorrer da realização deste trabalho deparamo-nos com a necessidade de uma carteira que "fotografasse" a situação de cada cliente na data de prestação e de pagamento de cada prestação, de modo que permitisse, num trabalho futuro, aplicar Vórtices Estocásticos em relação às variáveis que evoluem ao longo do tempo e com as mudanças socio-económicas e financeiras dos clientes.

Bibliografia

[Abdou et al., 2008] Abdou, H., Pointon, J., e El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in egyptian banking. *Expert Systems with Applications*, 35 (3):1275–1292.

[Ahn et al., 2000] Ahn, B., Cho, S., e Kim, C. (2000). The integrated methodology of rough set theory and artificial neural network for business failure prediction. *ExpertSystems with Applications*, 18:65–74.

[Altman et al., 1994] Altman, E., Marco, G., e Varetto, F. (1994). Corporate distress diagnosis; comparisons using linear discriminant analysis and neural networks (the italian experience). *Journal of Banking and Finance*, 18:505–529.

[Altman, 1968] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, pages 589–609.

[Anderson, 2007] Anderson, R. (2007). *Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press, Oxford.

[Baesens et al., 2003] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., e Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54:1082–1088.

[Baesens et al., 2002] Baesens, B., Viaene, S., Poel, D. V. D., Vanthienen, J., e Dedene, G. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research*, 138:191–211.

[Bailey, 2001] Bailey, M. (2001). *Credit Scoring - The principles and Practicalities*. White Box Publishing.

[Banasik et al., 2003] Banasik, J., Crook, J., e Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, 54(8):822–832.

[Bardos, 1998] Bardos, M. (1998). Detecting the risk of company failure at the banque de france. *Journal of Banking and Finance*, 22:1405–1419.

[Bartholomew, 1982] Bartholomew, D. J. (1982). *Stochastic models for social processes*. John Wiley & Sons Ltd., third edition, Wiley Series in Probability and Mathematical Statistics.

[Berry, 2000] Berry, M. e Linoff, G. (2000). Mastering data mining: The art and science of customer relationship management. *John Wiley and Sons, Inc, New York*.

[Bevan e Garzarelli, 2000] Bevan, A. e Garzarelli, F. (2000). Corporate bond spreads and the business cycle. *Journal of Fixed Income*, 9(4):8–18.

[Beynon e Peel, 2001] Beynon, M. e Peel, M. (2001). Variable precision rough set theory and data discretisation an application to corporate failure prediction. *OMEGA*, 29:561–576.

[Brehm e Gates, 1993] Brehm, J. e Gates, S. (1993). Donut shops and speed traps: Evaluating models of supervision on police behavior. *American Journal of Political Science*, 37(2):555–581.

[Centeno e Silva, 2001] Centeno, M. e Silva, J. (2001). Bonus systems in an open portfolio. *Insurance, Mathematics and Economics*, 28:341–350.

[Chatterjee e Barcun, 1970] Chatterjee, S. e Barcun, S. (1970). A nonparametric approach to credit screening. *Journal of the American Statistical Association*, 65(329):150–154.

[Chen e Huang, 2003] Chen, M. e Huang, S. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications*, 24:433–441.

[Conover, 1999] Conover, W. J. (1999). *Practical Nomparametric Statistics* 3^{nd} *Edition*. New York: John Wiley and Sons.

[Cox e Snell, 1989] Cox, D. e Snell, E. (1989). *Analysis of Binary Data*. Chapman Hall, London.

[Cramér, 1999] Cramér, H. (1999). *Mathematical Methods of Statistics*. Princeton Landmarks in Mathematics, Princeton University Press, Reprint of the 1946 original, Princeton, NJ.

[Crook et al., 2007] Crook, J., Edelman, D., e Thomas, L. (2007). Recent developments in consumer credit risk assessment. *Eur J Opl Res*, 18:1447–1465.

[Crook et al., 1992] Crook, J., Hamilton, R., e Thomas, L. (1992). A comparison of discriminations under alternative definitions of credit default. *In: L.C. Thomas, J.N. Crook and D.B. Edelman, Editors, Credit scoring and credit control, Oxford University Press, Oxford*, pages 217–245.

[Dacunha-Castelle e Duflo, 1983] Dacunha-Castelle, D. e Duflo, M. (1983). *Probabilités et statistiques. Tome 2.* Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree] Problèmes à temps mobile. [Movable-time problems], Masson, Paris.

[Dacunha-Castelle et al., 1970] Dacunha-Castelle, D., Revuz, D., e Schreiber, M. (1970). *Recueil de problèmes de calcul des probabilités.* Deuxième édition, revue et augmentée. Préfaces de A. Tortrat Masson et Cie, Éditeurs, Paris.

[Davies, 2008] Davies, A. (2008). Credit spread determinants: An 85 year perspective. *Journal of Financial Markets*, 2:180–197.

[Davis et al., 1992] Davis, R., Edelman, D., e Gammerman, A. (1992). Machine-learning algorithms for credit-card applications. *IMA Journal of Management Mathematics*, 4(1):43–51.

[Desai et al., 1997] Desai, V., Conway, D., Crook, J., e Overstreet, G. (1997). Credit scoring models in the credit union environment using neural networks and genetic algorithms. *IMA Journal of Mathematics Applied in Business and Industry*, 8(4):323–3463.

[Desai et al., 1996] Desai, V., Crook, J., e Overstreet, G. (1996). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *European Journal of Operational Research*, 95:24–37.

[Dinh e Kleimeier, 2007] Dinh, T. e Kleimeier, S. (2007). A credit scoring model for vietnam's retail banking market. *International Review of Financial Analysis*, 16(5):471–495.

[Dreiseitl e Ohno-Machado, 2002] Dreiseitl, S. e Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 35:352–359.

[Duffee, 1998] Duffee, G. (1998). The relation between treasury yields and corporate bond yield spreads. *Journal of Finance*, 53:2225–2241.

[Fan e Cheng, 2007] Fan, T. H. e Cheng, K. F. (2007). Tests and variables selection on regression analysis for massive datasets. *Data and Knowledge Engineering*, 63 (4):811–819.

[Feller, 1968] Feller, W. (1968). *An introduction to probability theory and its applications. Vol. I.* Third edition, John Wiley & Sons Inc., New York.

[Ferrari e Cribari-Neto, 2004] Ferrari, S. e Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

[Finlay, 2008] Finlay, S. (2008). *The Management of Consumer Credit: Theory and Practice.* Palgrave Macmillan, Basingstoke, UK.

[Flagg et al., 1991] Flagg, J. C., Giroux, G. A., e Wiggins, C. E. (1991). Predicting corporate bankruptcy using failing firms. *Review of Financial Economics*, 1:67–78.

[Frydman et al., 1985] Frydman, H., Altman, E., e Kao, D. (1985). Introducing recursive partitioning for financial classification: The case of financial distress. *The Journal of Finance*, 40(1):269–291.

[Galindo e Tamayo, 2000] Galindo, J. e Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modelling applications. *Computational Economics*, 15:107–143.

[Gani, 1963] Gani, J. (1963). Formulae for projecting enrolments and degrees awarded in universities. *Journal of the Royal Statistical Society. Series A (General)*, 126(3):400–409.

[Gestel et al., 2006] Gestel, V., T., Baesens, B., Van Dijcke, P., Garcia, J., Suykens, J., e Vanthienen, J. (2006). A process model to develop an internal rating system: Sovereign credit ratings. *Decision Support Systems*, 42(2):1131–1151.

[Goldberg, 1989] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. MA: Addison-Wesley.

[Guerreiro, 2001] Guerreiro, G. (2001). *Uma Abordagem Alternativa para Bonus Malus*. Tese de mestrado, Instituto Superior de Economia e Gestão, Universidade Técnica de Lisboa.

[Guerreiro, 2008] Guerreiro, G. (2008). *Populações Sujeitas a Reclassificações Periódicas*. Tese de doutoramento, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa.

[Guerreiro e Mexia, 2004] Guerreiro, G. e Mexia, J. (2004). An alternative approach to bonus malus. *Discussiones Mathematicae, Probability and Statistics*, 24:197–213.

[Guerreiro e Mexia, 2008] Guerreiro, G. e Mexia, J. (2008). Stochastic vortices in periodically reclassified populations. *Discussiones Mathematicae, Probability and Statistics*, 28(2).

[Guerreiro et al., 2010] Guerreiro, G., Mexia, J., e Miguens, M. (2010). A model for open populations subject to periodical re-classifications. *Journal of Statistical Theory and Practice*, 4(2):303–321.

[Guerreiro et al., 2012a] Guerreiro, G., Mexia, J., e Miguens, M. (2012a). Preliminary results on coefidence intervals for open bonus systems. *Selected Papers of XVII Congress of Sociedade Portuguesa da Estística, Springer*.

[Guerreiro et al., 2012b] Guerreiro, G., Mexia, J., e Miguens, M. (2012b). Stable distributions for open populations subject to periodical re-classifications. *Journal of Statistical Theory and Practice, Aceite para publicação.*

[Hair et al., 2006] Hair, J., Black, B., Babin, B., Anderson, R., e Tatham, R. (2006). *Multivariate Data Analysis.* Upper Saddle River, NJ: Prentice-Hall.

[Hancox et al., 2010] Hancox, D., Hoskin, C., e Wilson, R. (2010). Evening up the score: Sexual selection favours both alternatives in the colour-polymorphic ornate rainbowfish. *Animal Behaviour*, 80(5):845–851.

[Hand, 2001] Hand, D. (2001). Modelling consumer credit risk. *IMA Journal of Management Mathematics*, 12(1):139–155.

[Hand e Henley, 1997] Hand, D. e Henley, W. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society*, Series A (Statistics in Society) 160 (3):523–541.

[Hanley e McNeil, 1982] Hanley, J. A. e McNeil, B. (1982). The meaning and use of the area under a reciver operating characteristic (roc) curve. *Radiology*, 143:29–36.

[Harrell e Lee, 1985] Harrell, F. e Lee, K. (1985). A comparison of the discrimination of discriminant analysis and logistic regression. *In: P.K. Se, Editor, Biostatistics: Statistics in biomedical, Public health, and environmental sciences, North-Holland, Amsterdam.*

[Henley e Hand, 1996] Henley, W. e Hand, D. (1996). A k-nearest neighbour classifier for assessing consumer risk. *Statistician*, 44:77–95.

[Hosmer e Lemeshow, 1989] Hosmer, D. e Lemeshow, S. (1989). *Applied Logistic Regression.* New York: Wiley.

[Hsieh, 2005] Hsieh, N.-C. (2005). Hybrid mining approach in the design of credit scoring models. *Expert Systems with Applications*, 28:655–665.

[Huang et al., 2007] Huang, C., Chen, M.-C., e Wang, C. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33:847–856.

[Huang et al., 2004] Huang, Z., Chen, H., Hsu, C., Chen, W., e Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems (Special issue: Data Mining for Financial Decision Making)*, 37(4):543–558.

[Hung e Chen, 2009] Hung, C. e Chen, J. (2009). A selective ensemble based on expected probabilities for bankruptcy prediction. *Expert Systems with Applications*, 36(3):5297–5303.

[Islam et al., 2007] Islam, M. J., Wu, Q. M. J., Ahmadi, M., e Sid-ahmed, M. A. (2007). Investigating the performance of naïve bayes classifiers and k-nearest neighbor classifiers. *In International conference on convergence information technology*, 21-23:1541–1546.

[Jensen, 1992] Jensen, H. (1992). Using neural networks for credit scoring. *Managerial Finance*, 18(1):15–26.

[Joanes, 1993] Joanes, D. N. (1993). Reject inference applied to logistic regression for credit score. *IMA Journal of Mathematics Applied in Business and Industry*, 5:35–43.

[Johnson e Wichern, 2002] Johnson, R. e Wichern, D. (2002). *Applied Multivariate Statistical Analisys, 5. ed.* New Jersey, Prentice Hall.

[Kay et al., 2000] Kay, O. W., Warde, A., e Martens, L. (2000). Social differentiation and the market for eating out in the uk. *International Journal of Hospitality Management*, 19 (2):173–190.

[Kieschnick e McCullough, 2003] Kieschnick, R. e McCullough, B. (2003). Regression analysis of variates observed on (0,1): Percentages, proportions and fractions. *Statistical Modelling*, 3(3):193–213.

[Laitinen, 1999] Laitinen, E. K. (1999). Predicting a corporate credit analyst's risk estimate by logistic and linear models. *International Review of Financial Analysis*, 8 (2):97–121.

[Laitinen e Laitinen, 2000] Laitinen, E. K. e Laitinen, T. (2000). Bankruptcy prediction: Application of the taylor's expansion in logistic regression. *International Review of Financial Analysi*, 9 (4):327–349.

[Lee e Chen, 2005] Lee, T. e Chen, I. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, 28(4):743–752.

[Lee, 2007] Lee, Y. (2007). Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications)*, 33(1):67–74.

[Leonard, 1993] Leonard, K. (1993). Empirical bayes analysis of the commercial loan evaluation process. *Statistics and Probability Letters*, 18:289–296.

[Lewis, 1992] Lewis, E. (1992). *An Introduction to Credit Scoring.* California.

[Li et al., 2006] Li, S., Shiue, W., e Huang, M. (2006). The evaluation of consumer loans using support vector machines. *Expert Systems with Applications)*, 30(4):772–782.

[Longstaff e Schwartz, 1995] Longstaff, F. e Schwartz, E. (1995). A simple approach to valuing risky fixed and floating rate debt. *Journal of Finance*, 50:789–820.

[Lucas, 2000] Lucas, P. (2000). Why recoveries are on the rise. *Credit Card Management*, pages 71–72.

[Malhotra e Malhotra, 2003] Malhotra, R. e Malhotra, D. (2003). Evaluating consumer loans using neural networks. *Omega*, 31(2):83–96.

[Martell e Fitts, 1981] Martell, T. e Fitts, R. (1981). A quadratic discriminant analysis of bank credit card user characteristics. *Journal of Economics and Business*, 33:153–159.

[Mays, 2004] Mays, E. (2004). *Credit Scoring for Risk Managers, The Handbook for Lenders*. Thomson South Western: Mason Ohio.

[McCullagh e Nelder, 1989] McCullagh, P. e Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall/CRC, USA.

[McNab e Wynn, 2000] McNab, H. e Wynn, A. (2000). *Principles and Practice of Consumer Credit Risk Management*. CIB Publishing: Canterbury.

[McNeil et al., 2005] McNeil, A., Frey, R., e Embrechts, P. (2005). *Quantitative Risk Management*. Princeton University Press, USA.

[Merton, 1974] Merton, R. (1974). On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance*, 29:449–470.

[Mester, 1997] Mester, L. (1997). What's the point of credit scoring. *Business Review, Federal Reserve Bank of Philadelphia*.

[Mexia, 2000] Mexia, J. (2000). *Vórtices Estocásticos de Parâmetro Discreto*. Comunicação, Comunicação nos III Colóquios Actuariais FCT-UNL.

[Mood et al., 1963] Mood, A., Graybill, F., e Boes, D. (1963). *Introduction to the Theory of Statistics, 3^{rd} Edition*. Mc-Graw Hill International Editions.

[Morris et al., 1998] Morris, C., Neale, R., e Rolph, D. (1998). Creditspreads and interest rates: a cointegration approach. Technical report, Federal Reserve Bank of Kansas City.

[Nelder e Wedderburn, 1972] Nelder, J. e Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, A(135).

[Overstreet et al., 1992] Overstreet, J. G., Bradley, J. E., e Kemp, R. (1992). The flat-maximum effect and generic linear scoring model: A test. *IMA Journal of Mathematics Applied in Business and Industry*, 95:97–109.

[Paleologo et al., 2010] Paleologo, G., Elisseeff, A., e Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research*, 201:490–499.

[Parzen, 1965] Parzen, E. (1965). *Stochastic Processes*. Holden-Day.

[Piramuthu, 1999] Piramuthu, S. (1999). Financial credit-risk evaluation with neural and neurofuzzy systems. *European Journal of Operational Research*, 112:310–321.

[Pollard, 1966] Pollard, J. H. (1966). On the use of the direct matrix product in analysing certain stochastic population models. *Biometrika*, 53:397–415.

[Pollard, 1967] Pollard, J. H. (1967). Hierarchical population models with Poisson recruitment. *Journal of Applied Probability*, 4:209–213.

[Pollard, 9679] Pollard, J. H. (19679). Matrix analysis of the cash flows and reserves of a life offices. *The Australian Journal of Statistics*, 21:315–324.

[Pollard, 1969] Pollard, J. H. (1969). Continuous-time and discrete-time models of population growth. *Journal of the Royal Statistical Society. Series A. General*, 132:80–88.

[Pollard e Sherris, 1980] Pollard, J. H. e Sherris, M. (1980). Application of matrix methods to pension funds. *Scandinavian Actuarial Journal*, (2):77–95.

[Reichert et al., 1983] Reichert, A., Cho, C., e Wagner, G. (1983). An examination of the conceptual issues involved in developing credit-scoring models. *Journal of Business and Economic Statistics*, 1:101–114.

[Rodrigues, 2011] Rodrigues, E. V. (2011). *Uma Abordagem Comparativa e Analítica de Dois Sistemas de Bónus Malus em Cabo Verde: O Sistema Actual e a Proposta da Garantia*. Dissertação de mestrado, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.

[Ross, 1996] Ross, S. (1996). *Stochastic Processes*. Wiley, New York.

[Salchenberger et al., 1992] Salchenberger, L., Cinar, E., e Las, N. (1992). Neural networks: A new tool for predicting thrift failures. *Managerial Finance*, 23(4):899–915.

[Sarlija et al., 2004] Sarlija, N., Bensic, M., e Bohacek, Z. (2004). Multinomial model in consumer credit scoring. *10th International Conference on Operational Research Croatia, Trogir*.

[Schebesch e Stecking, 2005] Schebesch, K. B. e Stecking, R. (2005). Support vector machines for classifying anddescribing credit applicants: Detecting typical and critical regions. *Journal of the Operational Research Society*, 56:1082–1088.

[Schott, 1997] Schott, J. (1997). *Matrix Analysis for Statistics*. Wiley Series in Probability and Statistics.

[Schreiner, 2004] Schreiner, M. (2004). Scoring arrears at a microlender in bolivia. *Journal of Microfinance*, 6:65–88.

[Semedo, 2010] Semedo, D. P. V. (2010). *Credit Scoring Aplicação da Regressão Logística vs Redes Neuronais Artificiais na Avaliação do Risco de Crédito no Mercado Cabo-Verdiano.* Dissertação de mestrado, Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa.

[Siddiqi, 2006] Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring.* John Wiley and Sons Inc., New Jersey.

[Smithson e Verkuilen, 2006] Smithson, M. e Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54–71.

[Stadje, 1999] Stadje, W. (1999). *Stationarity of a stochastic population flow model.* Journal of Applied Probability.

[Staff e Vagholkar, 1971] Staff, P. J. e Vagholkar, M. K. (1971). *Stationary distributions of open Markov processes in discrete time with application to hospital planning.* Journal of Applied Probability.

[Suh et al., 1999] Suh, E. H., Noh, K. C., e Suh, C. K. (1999). Customer list segmentation using the combined response model. *Expert Systems with Applications*, 17 (2):89–97.

[Tabachnick e Fidell, 1996] Tabachnick, B. e Fidell, L. (1996). *Using Multivariate Statistics.* New York: HarperCollins.

[Tam e Kiang, 1992] Tam, K. e Kiang, M. (1992). Managerial applications of neural networks: The case of bank failure predications. *Management Science*, 38(7):926–947.

[Thomas, 2009] Thomas, L. (2009). *Consumer Credit Models: Pricing, Profit, and Portfolios.* Oxford University Press Inc., New York.

[Thomas et al., 2002] Thomas, L., Edelman, D., e J.N., C. (2002). *Credit Scoring and its Applications.* SIAM, Philadelphia.

[Thomas et al., 2005] Thomas, L., Oliver, R., e Hand, D. (2005). A survey of the issues in consumer credit modelling research. *J Opl Res Soc*, 56:1006–1015.

[Thomas, 2000] Thomas, L. C. (2000). *A survey of Credit and Behavioral Scoring; Forecasting Financial Risk of Lending to Consumers.* University of Edinburgh, Edinburgh, U.K.

[Thomas, 2010] Thomas, L. C. (2010). Consumer finance: Challenges for operational research. *Journal of the Operational Research Society*, 61:41–52.

[Thomas et al., 2001] Thomas, L. C., J., H., e T, S. W. (2001). *Time will tell: Behavioral Scoring and the Dinamics of Consumer Credit Assessment.* University of Edinburgh, Edinburgh, U.K.

[Titterington, 1992] Titterington, D. (1992). Discriminant analysis and related topics. *In: J.N. Crook and D.B. Edelman, Editors, Credit scoring and credit control, Oxford University Press, Oxford*, pages 53–73.

[Trevino e Daniels, 1995] Trevino, L. e Daniels, J. (1995). Fdi theory and foreign direct investment in the united states: A comparison of investors and non-investors. *International Business Review*, 4:177–194.

[Turkman e Silva, 2000] Turkman, M. e Silva, J. (2000). Modelos lineares generalizados - da teoria à prática. Technical report.

[Vale, 2010] Vale, D. C. (2010). *Modelação e Estimação do Risco de Crédito Estudo de uma Carteira*. Tese de mestrado, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.

[Van Gool et al., 2009] Van Gool, J., Baesens, B., Sercu, P., e Verbeke, W. (2009). An analysis of the applicability of credit scoring for microfinance. *Academic and Business Research Institute Conference, Orlando (US)*, pages 24–26.

[Varetto, 1998] Varetto, F. (1998). Genetic algorithms applications in the analysis of insolvency risk. *Journal of Banking and Finance*, 22:1421–1439.

[Vassiliou, 1998] Vassiliou, P. (1998). The evolution of the teory of non-homogeneous markov systems. *Applied Stochastic Models and Data Analysis*, 13:159–176.

[Viganó, 1993] Viganó, L. C. (1993). A credit scoring model for development banks: An african case study. *Savings and Development*, 4:441–442.

[Wang et al., 2005] Wang, Y., Wang, S., e Lai, K. (2005). A new fuzzy support vector machine to evaluate creditrisk. *Transaction on Fuzzy Systems)*, 13:820–831.

[West, 2000] West, D. (2000). Neural network credit scoring models. *Computer and Operations Research*, 27:1131–1152.

[Westgaard e Van der Wijst, 2001] Westgaard, S. e Van der Wijst, N. (2001). Default probabilities in a corporate bank portfolio: A logistic model approach. *European Journal of Operational Research*, 135 (2):338–349.

[Wiginton, 1980] Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, 15:757–770.

[Yakasai, 2005] Yakasai, B. M. (2005). Stationary population flow of a semi-open Markov chain. *Journal of the Nigerian Association of Mathematical Physics*, 9:395–398.

[Yobas et al., 2000] Yobas, M., Crook, J., e Ross, P. (2000). Credit scoring using evolutionary techniques. *IMA Journal of Mathematics Applied in Business e Industry*, 11:111–125.

[Yu et al., 2008] Yu, L., Wang, S., Wen, F., Lai, K., e He, S. (2008). Designing a hybrid intelligent mining system for creditriskevaluation. *Systems Science and Complexity*, 21:527–539.

[Zhou et al., 2008] Zhou, X., Zhang, D., e Jiang, Y. (2008). A new creditscoring method based on rough sets and decisiontree. *Lecture Notes in Artificial Intelligence*, 5012:1081–1089.

[Zorich, 2009] Zorich, V. A. (2009). *Mathematical analysis I.* Springer-Verlag, Berlin.