Matilde Sobral Pinto Castro de Oliveira

Licenciada em Matemática Aplicada à Economia e à Gestão

Calibração e Simulação de um Modelo de Cadeias de *Markov* para um seguro *Long-Term Care*

Dissertação para obtenção do Grau de Mestre em

Matemática e Aplicações
Ramo Atuariado, Estatística e Investigação Operacional

Orientador: Professor Doutor Manuel Leote Esquível, Professor Associado, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

setembro, 2017

Calibração e Simulação de um Modelo de Cadeias de *Markov* para um seguro *Long-Term Care*

Este documento foi gerado utilizando o processador (pdf)LaTeX, com base no template "novathesis" [1] desenvolvido no Dep. Informática da FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

Agradecimentos

Este trabalho insere-se num projeto que está em desenvolvimentos há uns anos e foram muitas as colaborações que permitiram a sua realização.

Ao Dr. Hugo Lopes, agradeço a partilha de toda a informação, utilizada numa das partes deste trabalho; os seus esclarecimentos foram fundamentais.

À Professora Doutora Susana Nascimento, agradeço a introdução e contextualiazação de uma ferramenta muito importante (análise de *clusters*) para o desenvolvimento deste trabalho.

À Professora Doutora Gracinda Guerreiro, agradeço todos os esclarecimentos feitos, fulcrais, para a obtenção de bons resultados.

Um agradecimento muito especial ao Professor Doutor Manuel Leote Esquível, por toda a disponibilidade, ajuda, incentivo, sem ele nada seria possível, muito obrigada.

Agradeço à minha família e aos meus amigos por todo o apoio nesta etapa, com um agradecimento especial aos meus pais e ao meu irmão, por toda a compreensão e paciência.

Por fim, um agradecimento a todas as pessoas que, de uma forma ou de outra, tornaram possível a realização deste trabalho.

Resumo

O presente trabalho tem o objetivo de obter valores para prémios de um seguro de dependência ou *Long-Term Care*. Este consiste, essencialmente, no pagamento de um certo capital seguro consoante o grau de dependência do segurado, existindo já em diversos países (EUA, França, Alemanha).

Define-se pessoa dependente como alguém que não tem autonomia e necessita de ajuda de terceiros para realizar as atividades da vida diária.

Estudou-se a modulação matemática deste seguro por meio de uma cadeia de *Markov* a tempo contínuo, através de um modelo de estados múltiplos com cinco estados (saudável, dependência fraca, moderada, severa e morte). Estes estados foram determinados a partir da aplicação de técnicas de *clustering*, pelo método *CLARA* (*Clustering LARge Applications*) a uma base de dados de 2015, fornecida, pela Rede Nacional de Cuidados Continuados Integrados (RNCCI). Esta base de dados contém registos das avaliações dos utentes, nomeadamente das atividades da vida diária, da locomoção e do seu estado cognitivo. mérica de equações diferenciais de *Chapman-Kolmogorov*, escritas a

Posteriormente, recorrendo ao *software Mathematica*, foram resolvidas numericamente as equações diferenciais de *Chapman-Kolmogorov* do modelo, obtidas a partir da calibração de intensidades de transição pelos dados fornecidos.

Por fim, foram feitas simulações dos custos associados à dependência dos utentes e calculado o prémio do seguro, usando-se a título de exemplo, o princípio do valor esperado.

Palavras-chave: Seguro *Long-term Care*, Dependência, Modelos de estados múltiplos, *Clusters*, Simulação

-

Abstract

The present work has the objective of obtaining the Long-term Care Insurance's premium value which consists on the payment of a certain amount according to the insured's dependence degree; this insurance, already, exists in some countries like the USA, France and Germany.

Dependent person is someone that has no autonomy and needs help to carry out daily activities.

Thus, this insurance was adjusted to a continuous time Markov Chain, by a multiple state model with five states (healthy, weak dependence, moderate dependence, severe dependence and death). These states were obtained with the clustering application by the CLARA method (Clustering LARge Applications) to a database, provided by Rede Nacional de Cuidados Continuados Integrados. This database contains patient's evaluation variables, daily activities, locomotion and cognitive status.

Subsequently, with Mathematica software, Chapman-Kolmogorov's differential equations were numerically solved, obtained from the calibration of transition intensities by the data provided.

Finally, simulations of the costs associated with the patient's dependence were made and the insurance's premium was calculated.

Keywords: Long-term Care Insurance, Dependency, Multiple State Model, Clusters, Simulation

Índice

Lista de Figuras

Lista de Tabelas

Listagens

Capítulo 1

Introdução

Nos últimos anos em Portugal, tem-se verificado um progressivo envelhecimento da população, em consequência, existe um maior número de pessoas dependentes. A dependência, isto é, a falta de autonomia e/ou incapacidade de realizar tarefas, acarreta, por vezes, custos insuportáveis para os doentes. Assim, para fazer face a estes custos pode considerar-se fazer um Seguro de Dependência ou *Long-term Care*. Este seguro, ainda inexistente em Portugal, mas já comercializado em diversos países do mundo, tem o objetivo de cobrir qualquer tipo de custo relativo à dependência, após uma dada idade, pagando uma certa quantia, durante a vida ativa. Tem-se, então, que o objetivo desta dissertação é o cálculo, através da calibração e simulação de um modelo de cadeias de *Markov*, do prémio de um seguro desta natureza. Este é, então, ajustado a um modelo de estados múltiplos, uma aplicação das cadeias de *Markov*.

Partindo-se de uma base de dados da Rede Nacional de Cuidados Continuados Integrados (RNCCI) de 2015, que contém variáveis de avaliações de utentes, aplicou-se uma ferramenta estatística, chamada de *clustering*, para simplificar e dividi-la em conjuntos que representam estados de dependência. Uma vez que cada utente pode estar em vários estados, em diferentes períodos, também, é possível obter-se uma matriz de probabilidades de transições dos estados a tempo discreto. Com o intuito de obter as probabilidades a tempo contínuo é necessário resolver-se as equações diferenciais de *Chapman-Kolmogorov*. No entanto, para isso, é necessário terem-se intensidades de transição, por isso, através da minimização de uma função perda (diferença entre as probabilidades a tempo contínuo e a tempo discreto) estas intensidades foram calibradas. Posto isto, são simulados os custos, fazendo-se uma amostra e por fim, calculado um prémio com base na média desta.

Esta trabalho é composto por 12 capítulos, de seguida será descrito o que foi abordado em cada um.

No capítulo 2 irá ser feita uma contextualização de algumas definições importantes

para todo o trabalho, nomeadamente a de dependência, considerando [5], [22] e [4]. Serão apresentadas estatísticas que demonstram o progressivo envelhecimento da população portuguesa e apresentada a Rede Nacional de Cuidados Continuados, isto é, uma rede constituída por diversas instituições com o objetivo de prestar cuidados a pessoas dependentes. Também, se apresentará uma definição mais pormenorizada do seguro *Long-term Care* e diferentes produtos deste tipo de seguro e um pouco da história deste em alguns países, segundo [8].

O seguro de dependência foi ajustado a uma Cadeia de *Markov* com um número finito de estados, a tempo contínuo, isto é, um processo estocástico em que é verificada a propriedade de *Markov*. Assim, no capítulo 3, será feita uma abordagem sobre este tema, serão referidas diversas definições, nomeadamente, a propriedade referida anteriormente, matrizes de intensidades de transição, equações de *Chapman-Kolmogorov* e probabilidades de transição. Uma das aplicações deste tema é o modelo de estados múltiplos, ter-se-á, então, uma expedição deste modelos, mostrando-se alguns exemplos e esclarecidos os seus pressupostos.

No capítulo 4, será feita uma descrição de *clustering*, isto é, uma ferramenta estatística que tem o objetivo de simplificar uma base de dados, dividindo-a em conjuntos com elementos com caraterísticas semelhantes entre si e diferentes dos outros grupos, existindo diversos métodos para o fazer, nomeadamente, métodos por partição, hierárquicos, entre outros.

Para fazer uma análise da dependência da população portuguesa foi estudada uma base de dados da RNCCI que será descrita no capítulo 5. Esta contém variáveis dos utentes: os seus internamentos, caraterização demográfica, avaliações do estado cognitivo, das atividades da vida diária e da locomoção e dos óbitos da rede. Neste capítulo é apresentada uma outra base de dados, que inicialmente foi trabalhada. Para aplicação de ferramentas estatísticas aos dados é necessário tratá-los, mais especificamente, transformar as variáveis qualitativas em quantitativas. No sexto capítulo tem-se, então, o tratamento da base de dados e a análise descritiva das diversas variáveis.

No capítulo 7, serão apresentados os diferentes elementos representativos dos *clusters* (conjuntos com elementos com caraterísticas semelhantes entre si e diferentes dos outros grupos) chamados de *medoids* e admitiu-se que a base de dados pode ser dividida em 3, 4 e 5 conjuntos. Para o resto do trabalho, considerou-se que os dados eram divididos em 4 *clusters*, cada um destes com caraterísticas semelhantes entre si, correspondendo a diferentes estados: saudável, dependência fraca, moderada e severa.

De seguida, no capítulo 8, ter-se-á o cálculo das matrizes de transições de probabilidades. Tirando-se partido do *software R*, fizeram-se as contagens das transições dos diversos utentes, adicionando um novo estado: o da morte. Adequando a [24], foram feitas matrizes por conjuntos de idades, de maneira a ter-se, aproximadamente, o mesmo número de observações por matriz.

No capítulo 9, serão apresentadas as equações diferencias de *Chapman-Kolmogorov* do modelo de estados múltiplos do seguro de dependência, tal como feito em [3] e [18]. Para a

sua resolução numérica é necessário terem-se intensidades de transição, assim, a partir de uma função perda são calibradas e obtidas as probabilidades de transição a tempo contínuo. Pelas intensidades de transição também serão obtidas, numericamente, as distribuições dos tempos de permanência em cada estado.

No capítulo 10, apresentar-se-á a simulação dos custos de cada utente, pela simulação dos tempos de permanência e das transições entre estados, fazendo-se uma amostra de custos, e, de seguida, é calculado um prémio do seguro a partir da média dos custos.

No penúltimo capítulo, serão calculados diferentes prémios de *Long-Term Care* para o seguro, considerando-se diferentes cenários para a idade em que o indivíduo começa a ser segurado e em que começam a ser pagos os prémios.

Por fim, no último capítulo, será apresentada uma pequena conclusão de todo o trabalho.

Algumas dissertações de mestrado já abordaram o tema de *Long-Term Care*: [2] apresenta uma formulação financeira-atuarial do seguro de dependência; [21] cujo principal objetivo é a implementação de um simulador em *Excel* para o cálculo de prémios.

Capítulo 2

Dependência na 3ª Idade

O presente trabalho é um estudo sobre seguros de dependência, por isso, é necessário fazer-se um esclarecimento da definição de dependência, e por consequência, uma contextualização desta, em Portugal. Há poucos estudos sobre este tema na população portuguesa. O risco de dependência cresce, muitas vezes, com o aumento da idade da pessoa, assim, apresentar-se-á uma análise do envelhecimento em Portugal.

2.1 O Conceito de Dependência

Dependência deriva de depender (sufixação do verbo por *-ência*), este tem origem no latim *dependeo*, ou seja, "pender de ". Existem diversas definições para dependência, algumas destas serão apresentadas, seguidamente:

- Dicionário de Língua Portuguesa 2004, Porto Editora: *estado de dependente; sujeição; subordinação; falta de autonomia, maturidade e independência*;
- Conselho da Europa (1998): ser dependente é *a pessoa que, por razões ligadas à falta ou perda de capacidade física, psíquica ou intelectual, tem necessidade de assistência e/ou ajuda para a realização das actividades da vida diária*;
- Organização Mundial de Saúde: a pessoa dependente é *aquela que não é completamente capaz de cuidar de si mesma e manter uma elevada qualidade de vida de acordo com as suas preferências, com o maior grau de independência, autonomia, participação, satisfação e dignidade pessoal possível*;
- Segundo o Decreto Lei nº101 de 6 de junho de 2006 ([6]), a dependência é uma *situação em que se encontra a pessoa que, por falta ou perda de autonomia física, psíquica ou intelectual, resultante ou agravada por doença crónica, demência orgânica,*

sequelas pós traumáticas, deficiência, doença severa e ou incurável em fase avançada, ausência ou escassez de apoio familiar ou de outra natureza, não consegue, por si só, realizar as actividades da vida diária.

Assim sendo, dependência pode definir se como um estado em que a pessoa não tem autonomia e não consegue realizar, por si só, atividades, por falta de capacidades. Estas atividades denominam-se como atividades da vida diária, e são:

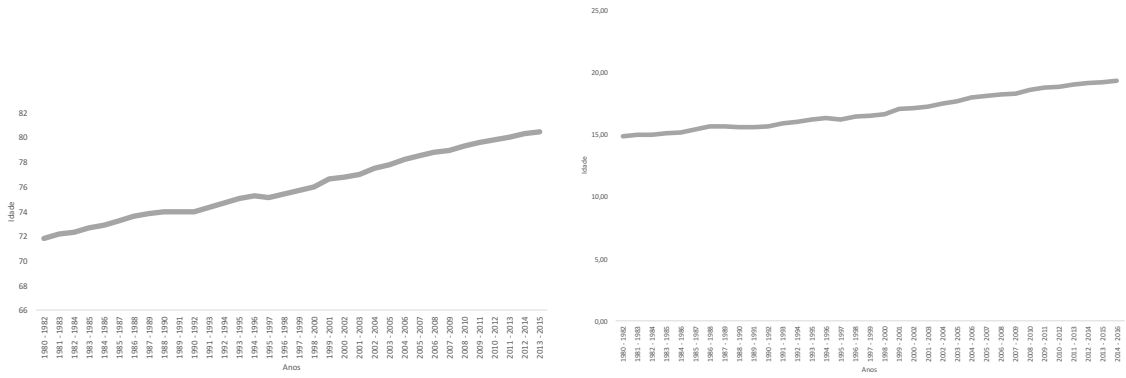
- Lavar - capacidade do indivíduo se lavar autonomamente;
- Vestir - capacidade do indivíduo se vestir e despir de forma autónoma;
- Ir à Sanita - capacidade das pessoas irem à casa-de-banho mantendo um nível higiénico satisfatório;
- Deitar - capacidade da pessoa se deitar e levantar da cama, sem ajuda;
- Sentar - capacidade do indivíduo se sentar e levantar de um sofá ou cadeira autonomamente;
- Continência urinária - capacidade de micção voluntária;
- Continência fecal - capacidade de controlar voluntariamente a emissão de fezes;
- Alimentar - capacidade de comer por si um alimento preparado;
- Mobilidade ou locomoção - capacidade de se movimentar, seja:
  - Dentro de casa;
  - Na rua;
  - Descer ou subir escadas.

2.2 Envelhecimento em Portugal

Portugal tem sofrido um envelhecimento da população significativo; existem vários índices e valores que o comprovam.

- Tem havido um permanente aumento na esperança média de vida; pode definir-se como esperança média de vida o número médio de anos que uma pessoa com uma certa idade possa esperar viver, mantendo-se as taxas de mortalidade observadas nesse momento. Normalmente são analisadas a duas idades, à nascença e aos 65 anos, podendo ser calculadas a qualquer idade.
  - À nascença, tem-se visto um aumento significativo, em Portugal. Em 20 anos, a esperança média de vida aumentou cerca de 8 anos, passando de 72 em 1980/1982, a 80 anos em 2013/2015; é possível observar esta evolução no gráfico 2.1a;

– A esperança média de vida aos 65 anos também tem aumentado, mas não tão significativamente. Em 14 anos, aumentou cerca de 4 anos e meio, sendo de 14.48 anos em 1980/1982 e de 19.31 em 2014/2016, veja-se o gráfico 2.1b.



a Esperança média de vida à nascença

b Esperança média de vida aos 65 anos

Figura 2.1: Esperança média de vida
Fonte: [1]

- A percentagem da população com mais de 60 anos tem aumentado, sendo as dos grupos etários 75 a 79, 80 a 84 e mais de 85 anos, aquelas em que se tem verificado uma maior evolução, conforme é possível observar na figura 2.2;
- O Índice de Envelhecimento corresponde ao número de pessoas com 65 ou mais anos por cada 100 com menos de 15 anos. Um valor maior que 100 significa que existem mais idosos do que jovens. Em Portugal, este índice está em constante crescimento, tendo aumentado de 43.8 para 143.8 em 35 anos, conforme se pode verificar na figura 2.4;
- Outro indicador importante para observar o envelhecimento da população é o Índice de Longevidade. Este consiste numa proporção entre o número de pessoas com 75 ou mais anos e as pessoas com 65 ou mais anos (normalmente multiplicado por 100). Assim, quanto mais envelhecida é a população mais elevado é o valor do índice. Em Portugal, este tem tido um aumento nos últimos 20 anos, como é possível observar na figura 2.4.
  - Futuramente, segundo as projeções, este índice irá aumentar ao longo do tempo, veja-se na figura 2.3:
- O Índice de Dependência de Idosos consiste no número de pessoas com 65 e mais anos por cada 100 em idade ativa, isto é, entre os 15 e 64 anos. Valores inferiores a 100 representam mais pessoas ativas do que idosos, valores superiores significam mais idosos do que pessoas entre os 15 e 64 anos. Em Portugal, apesar deste índice ser inferior a 100, tem vindo a aumentar ao longo do tempo. Verifica-se, por exemplo de 1961 a 2015 houve um crescimento para mais do dobro, ver figura 2.4.

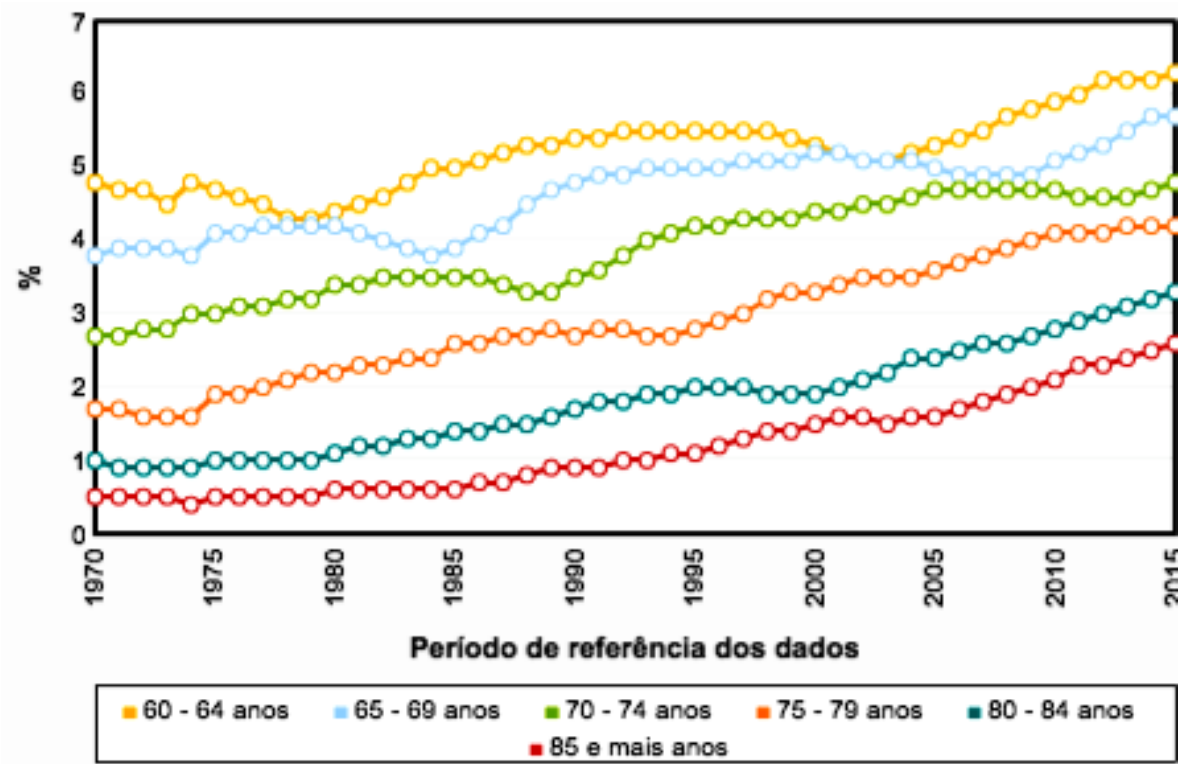


Figura 2.2: Distribuição da População residente (%) por grupo etário; Anual em Portugal
Fonte: [1]

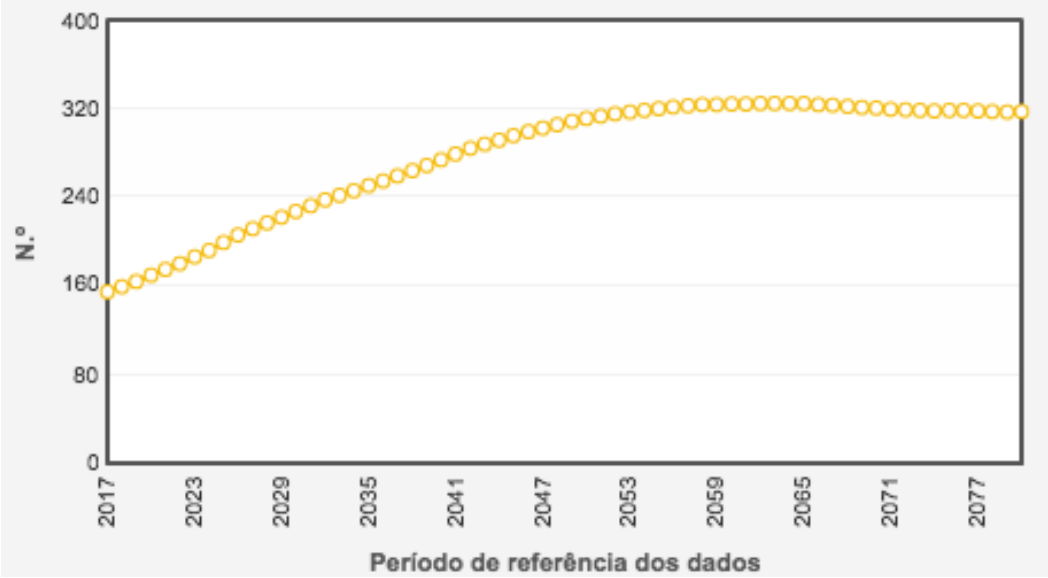


Figura 2.3: Índice de Envelhecimento, projeções 2015-2080
Fonte: [1]

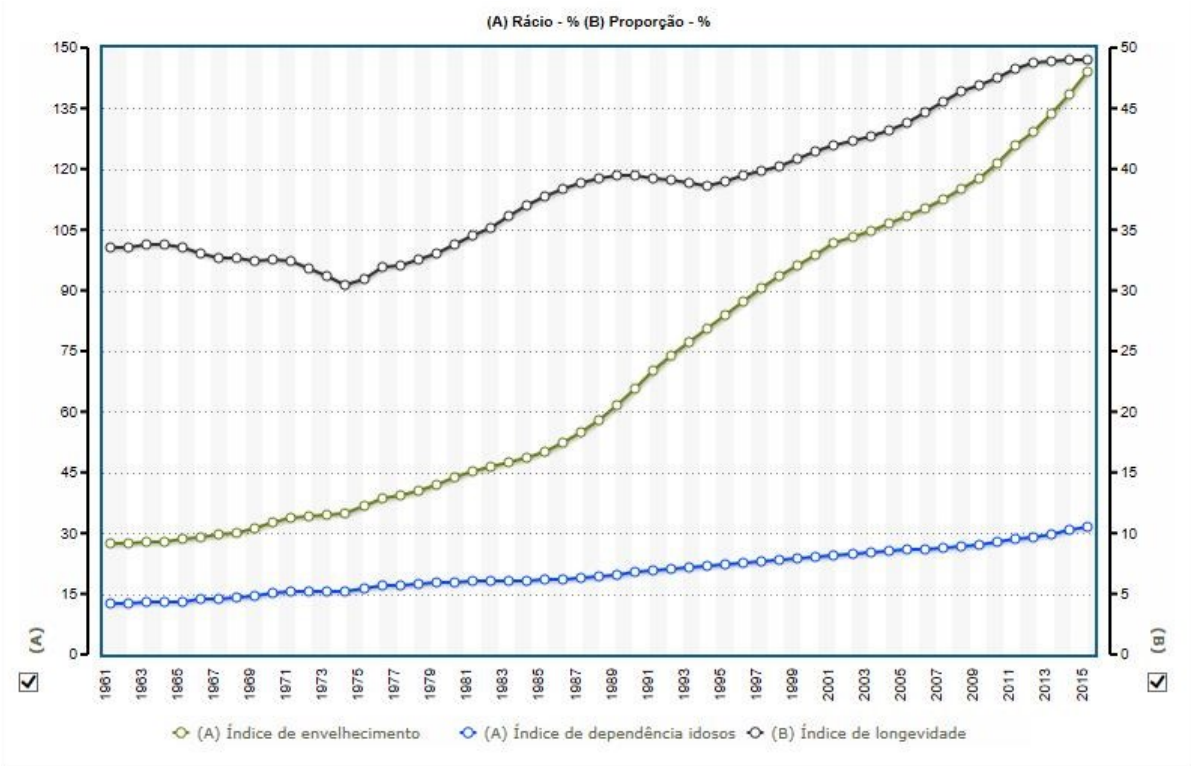
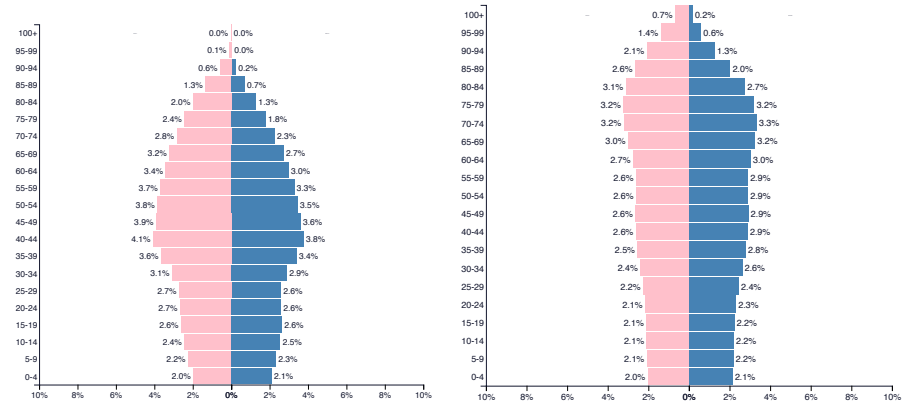


Figura 2.4: Indicadores de Envelhecimento
Fonte: [1]

- Por fim, observando-se as pirâmides etárias 2017 e 2080, observa-se um envelhecimento da população, pois a parte superior da pirâmide aumenta, associada a uma diminuição da parte inferior e central desta, gráfico 2.5.



a Pirâmide Etária em Portugal no Ano 2017

b Pirâmide Etária em Portugal no Ano 2080

Figura 2.5: Pirâmides Etárias em Portugal
Fonte: [15]

2.2.1 Rede Nacional de Cuidados Continuados Integrados

Para fazer face ao envelhecimento da população portuguesa, o Ministério do Trabalho e da Solidariedade Social e o Ministério da Saúde criaram, então, a Rede Nacional de Cuidados Continuados Integrados (RNCCI) através de [6], alterado pelo [7].

Segundo [6], a RNCCI tem como objetivo principal: *a prestação de cuidados continuados integrados a pessoas que*, (...), *se encontrem em situação de dependência.* Esta rede é um modelo organizacional constituída por *unidades e equipas de cuidados continuados de saúde, e ou apoio social, e de cuidados e acções paliativas, com origem nos serviços comunitários de proximidade, abrangendo os hospitais, os centros de saúde, os serviços distritais e locais da segurança social, a Rede Solidária e as autarquias locais.*

A RNCCI está dividida em várias unidades:

- Unidades de internamento:
  - Cuidados continuados de convalescença, esta unidade tem como objetivo a estabilização, avaliação e reabilitação da pessoa que sofreu uma perda de autonomia transitória, não necessitando de cuidados hospitalares de agudos. Esta serve para internamentos esperados até 30 dias consecutivos (pode existir em conjunto com a unidade de média duração e reabilitação);
  - Cuidados continuados de média duração e reabilitação é uma unidade de internamento vinculada com um hospital de agudos, com o objetivo de prestar cuidados, reabilitar, apoiar e avaliar pessoas que se encontram num processo de recuperação de uma situação aguda ou de uma descompensação de uma doença crónica, o período de internamento esperado desta unidade é entre 30 a 90 dias consecutivos;
  - Cuidados continuados de longa duração e manutenção têm a função de prestar apoio social e cuidados a pessoas com doenças crónicas, em diferentes graus de dependência, e que não podem ser cuidadas no domicílio, com o tempo de internamento, esperado, de mais de 90 dias consecutivos;
  - Cuidados paliativos consiste numa unidade de internamento localizada num hospital para o tratamento, acompanhamento e supervisão de indivíduos em situações complexas e de sofrimento devido a doenças severas, avançadas, incuráveis ou progressivas, não existe tempo esperado de internamento.
- Unidades de ambulatório - Unidades de dia e de promoção da autonomia servem para os cuidados integrados de suporte, para promoverem tanto o apoio social como a autonomia de indivíduos com diferentes graus de dependência, que não estão em condições de serem cuidados no domicílio. Este unidade funciona oito horas por dia e pelo menos nos dias úteis.
- Equipas hospitalares:

  - Equipas de gestão de altas consistem em unidades hospitalares multidisciplinares, constituídas, pelo menos, por um médico, um enfermeiro e um assistente. Estas equipas têm o objetivo de preparar e gerir as altas hospitalares, quer no domicílio quer nas unidades de convalescença quer nas unidades de média duração e reabilitação;
  - Equipas intra-hospitalares de suporte em cuidados paliativos são equipas multidisciplinares compostas, pelo menos, por um médico, um enfermeiro e um psicólogo, que têm o objetivo de prestar aconselhamento diferenciado nos cuidados paliativos a utentes internados em estado avançado ou terminal da doença.

- Equipas domiciliárias:

  - Equipas de cuidados continuados integrados são entidades que prestam cuidados domiciliários multidisciplinares. O público alvo são indivíduos com dependências funcionais, doenças terminais ou estão num processo de convalescença que os impede de se movimentarem autonomamente, mas não necessitam de ser internados;
  - Equipas comunitárias de suporte em cuidados paliativos são equipas com a finalidade de prestar apoio em cuidados paliativos e às outras equipas associadas a estes cuidados, são multidisciplinares e devem ser constituídas, no mínimo, por um médico e um enfermeiro.

A cada uma destas áreas está associado um custo, dado por [19]:

- A unidade de convalescença tem um custo de 105.46 euros diários por utente (3 163.80 €/mês);
- Um utente na unidade de cuidados paliativos por dia tem um custo de 105.46 euros (3 163.80 €/mês);
- A unidade de média duração e reabilitação acarreta um custo diário por utente de 87.56 euros (2 626.80 €/mês);
- Um utente estando hospitalizado na unidade de longa duração e manutenção tem um custo diário de 60.19 euros (1 805.70 €/mês);
- Por fim, a unidade de dia e de promoção da autonomia tem um custo diário de 9.58 euros por utente (287,40 €/mês).

Estes custos são de grandeza comparável aos descritos no Capítulo 10, mas não foram calculados com as mesmas justificações.

No entanto, só em certas situações, os indivíduos podem integrar a RNCCI, sendo este acesso apenas possível nas seguintes circunstâncias:

- Dependência funcional transitória;

- Dependência funcional prolongada;
- Idosos com critérios de fragilidade;
- Incapacidade grave;
- Doença severa.

A RNCCI têm tido um crescimento notório, veja-se a sua evolução, gráfico 2.6, de 2008 até novembro de 2015 houve um aumento de quase 5 000 camas na rede.

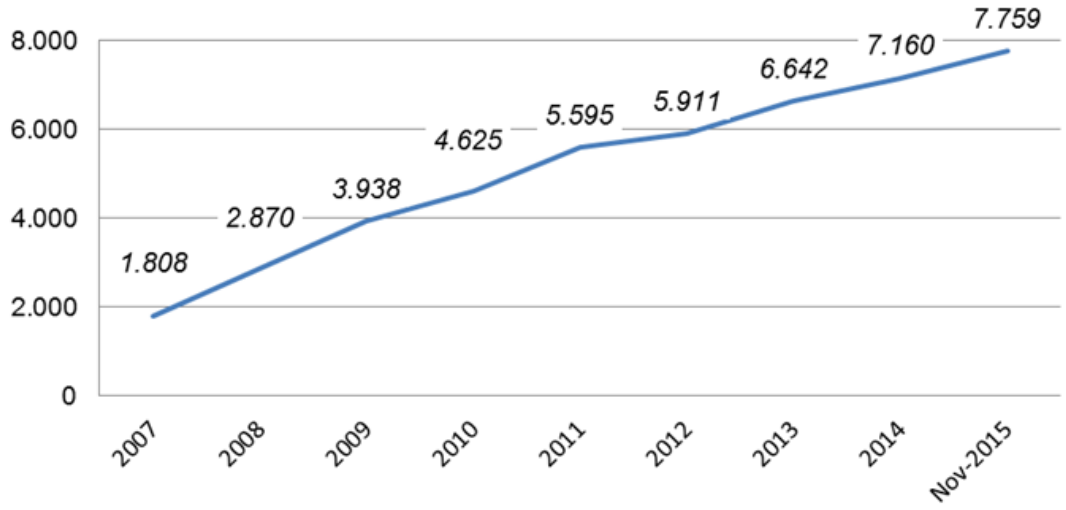


Figura 2.6: Número de camas da RNCCI
Fonte: [20]

2.3 Seguro *Long-Term Care*

O Seguro *Long-Term Care* (LTC) ou Seguro de Dependência consiste num produto que tem o objetivo de cobrir custos referentes a doenças crónicas ou incapacidades que tornam as pessoas dependentes. Existem diversas maneiras de o fazer, havendo vários tipos deste produto.

2.3.1 Tipos de Seguro de Dependência

- ***Stand-Alone Annuity***

Estes seguros são dirigidos a pessoas sem nenhuma dependência quando o adquirem tendo um período de diferimento, por vezes. Estes consistem no pagamento de uma renda vitalícia ou de uma renda quando a pessoa for dependente ou de uma por um período de tempo. O valor desta pode aumentar, conforme o grau de dependência do indivíduo.

- ***Rider Benefit Annuity***

Este produto é dirigido a pessoas em situação saudável, é uma combinação de uma renda temporária, se o indivíduo estiver dependente, com um capital, se o segurado vier a falecer. Este depende do valor que o indivíduo recebeu enquanto estava dependente.

- ***Enhanced Pension Annuity***

Um seguro deste tipo é ativado quando o indivíduo se reforma. Este pode escolher uma de duas situações. A primeira situação é o pagamento de uma renda vitalícia ao segurado (C). Na segunda, a seguradora garante o pagamento de uma renda, de termos C^a $(C > C^a)$, se o indivíduo estiver vivo e autónomo, e uma renda com termos C^d, $(C < C^d)$ se o indivíduo ficar dependente.

- ***Enhanced Annuity***

Por último, este tipo de seguro é comercializado apenas para indivíduos que já estejam dependentes. A seguradora garante o pagamento de uma renda imediata e o segurado o pagamento de um prémio único.

Este tipo de seguros já existe em diferentes países.

2.3.2 Seguro *Long-Term Care* no Mundo

- **Estados Unidos da América**

Os primeiros seguros de dependência surgiram nos Estados Unidos da América em 1974, por seguradoras de média dimensão. Estes tiveram um crescimento notório, em 1985 contavam com cerca de 100 000 apólices e, em 1999, este valor cresceu para 6.7 milhões. Em 1996, foi promulgada uma lei que beneficiou fiscalmente as pessoas com seguros de dependência. Nos EUA, este seguro tem uma oferta grande e diversificada.

- **França**

O primeiro seguro de dependência, no mercado francês surgiu em 1986. Neste país, estes seguros são um complemento a outros tipos de seguro. Também em França, saiu uma lei para beneficiar fiscalmente os indivíduos com apólices de seguros de dependência. Em 2001, havia cerca de 1.5 milhões de apólices deste tipo de seguros.

- **Alemanha**

A Alemanha foi o primeiro país europeu a iniciar a comercialização deste seguro, em 1985. Em 1994, foi aprovada a *Lei Federal do Seguro de Dependência* com o objetivo de tornar obrigatório este tipo de seguros, criando assim um sistema de financiamento de cobertura do risco de dependência.

O seguro de dependência já é muito comercializado no estrangeiro, no entanto, devido à falta de informação, nomeadamente, de dados, não existe em Portugal, apesar de tudo indicar que a dependência na população portuguesa está a aumentar.

Capítulo 3

Cadeias de *Markov* a Tempo Contínuo

Para a formalização matemática do seguro abordado na Secção 2.3 é necessário definir alguns conceitos e aplicações. Um destes são as cadeias de *Markov* a tempo contínuo. Estas cadeias, ver por exemplo [12], consistem num processo estocástico, em que a variável tempo é contínua, e o processo verifica a propriedade de *Markov*, ou seja, informalmente, a probabilidade de o processo estar num dado estado num momento futuro, depende apenas do presente e não do passado.

3.1 Definição

Considerar, daqui por diante, que x representa a idade de um segurado. O processo $\{S(t)\}_{t\geqslant 0}$ é uma Cadeia de Markov a Tempo Contínuo se:

- $\{S(t)\}_{t\geqslant 0}$ é um processo estocástico;
- $\mathbb{P}[S(x+t)=j|S(x)=i, S(u)=k, 0\leq u\leq x] = \mathbb{P}[S(x+t)=j|S(x)=i]$ para qualquer i e j pertencente a $S(t)$, ou seja, $S(t)$ verifica a propriedade de *Markov*.

A expressão $\mathbb{P}[S(x+t)=j|S(x)=i]$ representa a probabilidade de transição para o estado j, à idade $x+t$, sabendo que à idade x o indivíduo se encontra no estado i, representando-se também, por ${}_tp_x^{ij}$.

Estas probabilidades podem representar-se por uma matriz:

$$ {}_tP_x = \begin{bmatrix} {}_tp_x^{11} & \cdots & {}_tp_x^{1n} \\ \vdots & \ddots & \vdots \\ {}_tp_x^{n1} & \cdots & {}_tp_x^{nn} \end{bmatrix} $$

Esta matriz é estocástica, ou seja, as probabilidades das linhas somam um:

$$\sum_{j\in S} {}_tp_x^{ij} = 1 \quad \forall i,t$$

Estas probabilidades de transição verificam as equações de *Chapman-Kolmogorov*, ou seja:

$${}_tp_x^{ij} = \sum_{k\in S} {}_up_x^{ik}\, {}_{t-u}p_x^{kj} \quad \forall i,j \quad ;$$

Demonstração:
Segundo [12] e utilizando a propriedade de *Markov*, tem-se:

$$\begin{aligned}
{}_tp_x^{ij} &= \mathbb{P}\left[S(x+t)=j|S(x)=i\right] = \sum_{k\in S}\mathbb{P}\left[S(x+t)=j\wedge S(x+u)=k|S(x)=i\right] \\
&= \sum_{k\in S}\mathbb{P}\left[S(x+u)=k|S(x)=i\right]\mathbb{P}\left[S(x+t)=j|S(x+u)=k\wedge S(x)=i\right] \\
&= \sum_{k\in S}\mathbb{P}\left[S(x+u)=k|S(x)=i\right]\mathbb{P}\left[S(x+t)=j|S(x+u)=k\right] \\
&= \sum_{k\in S} {}_up_x^{ik}\, {}_{t-u}p_x^{kj}
\end{aligned}$$

■

3.2 Matriz de Intensidades de Transição

A função μ_x^{ij} corresponde à intensidade de transição do estado i para o j, à idade x, e considerando que estes dois estados não são o mesmo, define-se, sempre que este limite exista, por:

$$\mu_x^{ij} = \lim_{h\to 0^+} \frac{{}_hp_x^{ij}}{h} \quad \forall i,j;$$

A matriz, cujos elementos são as intensidades de transição entre cada dois estados à idade x, é denominada de Matriz de Intensidades de Transição:

$$M_x = \begin{bmatrix} \mu_x^{11} & \cdots & \mu_x^{1n} \\ \vdots & \ddots & \vdots \\ \mu_x^{n1} & \cdots & \mu_x^{nn} \end{bmatrix}$$

Esta matriz é quadrada e de ordem n, sendo n o número de estados que $S(t)$ pode assumir.

A intensidade total de transição do estado i, μ_x^i, obtém-se somando as intensidades de transições parciais do estado i para os outros estados, ou seja:

$$\mu_x^i = \sum_{\forall j\neq i} \mu_x^{ij} \quad \forall i$$

Pode ainda ter-se:

$$\mu_x^i = \lim_{h\to 0^+} \frac{1 - {}_hp_x^{ii}}{h} \quad \forall i$$

Demonstração:

Segundo [12], tem-se:

$$\mu_x^i = \sum_{\forall j\neq i} \mu_x^{ij} = \sum_{\forall j\neq i} \lim_{h\to 0^+} \frac{{}_hp_x^{ij}}{h} = \lim_{h\to 0^+} \frac{1 - {}_hp_x^{ii}}{h}$$

■

3.3 Equações Diferenciais de *Chapman-Kolmogorov*

As equações diferenciais de *Chapman-Kolmogorov* relacionam as probabilidades de transição entre estados e as respetivas intensidades de transição. A sua resolução permite obter as probabilidades de transição, dadas determinadas intensidades:

$$\frac{d\left({}_tp_x^{ij}\right)}{dt} = \sum_{k\neq j} {}_tp_x^{ik}\mu_{x+t}^{kj} - {}_tp_x^{ij}\mu_{x+t}^{j} \quad \forall i,j$$

Demonstração:

Admitindo que os limites indicados existem, pela definição de derivada e seguindo [12], tem-se:

$$\begin{aligned}
\frac{d\left({}_tp_x^{ij}\right)}{dt} &= \lim_{h\to 0^+} \frac{{}_{t+h}p_x^{ij} - {}_tp_x^{ij}}{h} \\
&= \lim_{h\to 0^+} \frac{\sum_{k\in S} {}_tp_x^{ik}\,{}_hp_{x+t}^{kj} - {}_tp_x^{ij}}{h} \\
&= \lim_{h\to 0^+} \frac{\sum_{k\neq j} {}_tp_x^{ik}\,{}_hp_{x+t}^{kj} + {}_tp_x^{ij}\,{}_hp_{x+t}^{jj} - {}_tp_x^{ij}}{h} \\
&= \sum_{k\neq j} \lim_{h\to 0^+} \frac{{}_tp_x^{ik}\,{}_hp_{x+t}^{kj}}{h} + \lim_{h\to 0^+} \frac{{}_tp_x^{ij}\left({}_hp_{x+t}^{jj} - 1\right)}{h} \\
&= \sum_{k\neq j} {}_tp_x^{ik} \lim_{h\to 0^+} \frac{{}_hp_{x+t}^{kj}}{h} - {}_tp_x^{ij} \lim_{h\to 0^+} \frac{\left(1 - {}_hp_{x+t}^{jj}\right)}{h} \\
&= \sum_{k\neq j} {}_tp_x^{ik}\mu_{x+t}^{kj} - {}_tp_x^{ij}\mu_{x+t}^{j}
\end{aligned}$$

■

3.4 Probabilidades de Permanência

Seja ${}_tp_x^{ii}$ a probabilidade de permanência no estado i entre as idades x e $x+t$ sem nunca sair do estado. Tem-se então:

$$\frac{d\left({}_tp_x^{ii}\right)}{dt} = -{}_tp_x^{ii}\mu_{x+t}^{i} \quad \forall i$$

Demonstração:
Admitindo que os limites indicados existem, como feito anteriormente e tendo em conta [12],

$$\begin{aligned}
\frac{d\left({}_tp_x^{ii}\right)}{dt} &= \lim_{h\to 0^+} \frac{{}_{t+h}p_x^{ii} - {}_tp_x^{ii}}{h} \\
&= \lim_{h\to 0^+} \frac{{}_tp_x^{ii}\,{}_hp_{x+t}^{ii} - {}_tp_x^{ii}}{h} \\
&= \lim_{h\to 0^+} \frac{{}_tp_x^{ii}({}_hp_{x+t}^{ii} - 1)}{h} \\
&= {}_tp_x^{ii} \lim_{h\to 0^+} \frac{{}_hp_{x+t}^{ii} - 1}{h} \\
&= -{}_tp_x^{ii} \lim_{h\to 0^+} \frac{1 - {}_hp_{x+t}^{ii}}{h} \\
&= -{}_tp_x^{ii}\mu_{x+t}^{i}
\end{aligned}$$

■

Uma vez que se tem uma expressão para a derivada da probabilidade de permanência interrupta pode obter-se uma para a probabilidade:

$${}_tp_x^{ii} = \exp\,\left(-\int_0^t \mu_{x+s}^{i} ds\right) \quad \forall i$$

Demonstração:
Segundo [12]:

$$\begin{aligned}
\frac{d\left({}_tp_x^{ii}\right)}{dt} = -{}_tp_x^{ii}\mu_{x+t}^{i} \quad &\Leftrightarrow \quad \frac{\frac{d\left({}_tp_x^{ii}\right)}{dt}}{{}_tp_x^{ii}} = -\mu_{x+t}^{i} \Leftrightarrow \\
&\Leftrightarrow \quad \frac{d\log\left({}_tp_x^{ii}\right)}{dt} = -\mu_{x+t}^{i} \Leftrightarrow \\
&\Leftrightarrow \quad \int_0^t \frac{d\log\left({}_tp_x^{ii}\right)}{dt} ds = -\int_0^t \mu_{x+s}^{i} ds \Leftrightarrow \\
&\Leftrightarrow \quad \log\left({}_tp_x^{ii}\right) - \log\left({}_0p_x^{ii}\right) = -\int_0^t \mu_{x+s}^{i} ds \Leftrightarrow \\
&\Leftrightarrow \quad \log\left({}_tp_x^{ii}\right) - \log\left(1\right) = -\int_0^t \mu_{x+s}^{i} ds \Leftrightarrow \\
&\Leftrightarrow \quad \log\left({}_tp_x^{ii}\right) = -\int_0^t \mu_{x+s}^{i} ds \Leftrightarrow \\
&\Leftrightarrow \quad {}_tp_x^{ii} = exp(-\int_0^t \mu_{x+s}^{i} ds)
\end{aligned}$$

■

3.5 Modelo de Estados Múltiplos

Uma das aplicações das cadeias de *Markov* a tempo contínuo é o modelo de estados múltiplos. Para [25], este modelo descreve as transições aleatórias de um indivíduo entre as condições que este possa ser ou ter ao longo do tempo. Considere-se a variável aleatória, $S(t)$, que representa o estado em que o indivíduo se encontra à idade t. O conjunto das variáveis aleatórias $\{S(t)\}_{t\geqslant 0}$ é então uma cadeia de *Markov* a tempo contínuo, se se verificarem as propriedades da definição apresentada em 3.1.

3.5.1 Exemplos

Vejam-se alguns exemplos deste tipo de modelos.

3.5.1.1 Seguro de Vida em Caso de Morte

Num seguro de vida em caso de morte, a seguradora garante o pagamento de um certo capital aquando da morte do segurado. É possível identificar dois estados: Vivo e Morto, assim a variável $S(t)$ pode assumir dois valores, neste modelo só é possível a transição do primeiro estado para o segundo, como representado na figura 3.1.

Figura 3.1: Seguro de Morte

Neste caso a matriz de transição é dada por:

$$ {}_tP_x = \begin{bmatrix} {}_tp_x^{11} & {}_tp_x^{12} \\ {}_tp_x^{21} & {}_tp_x^{22} \end{bmatrix} = \begin{bmatrix} {}_tp_x^{11} & 1 - {}_tp_x^{11} \\ 0 & 1 \end{bmatrix} $$

3.5.1.2 Seguro de Invalidez Temporária

Um contrato de um seguro de invalidez temporária consiste, por exemplo, no pagamento de um capital ou renda quando o indivíduo se encontra inválido. Nesta situação, uma vez que há três estados diferentes: saudável, inválido temporário e morte, a variável $S(t)$ pode assumir estes três valores. Neste modelo há a possibilidade de se entrar no mesmo estado mais de uma vez, como é possível observar na figura 3.2, ou seja, antes do indivíduo morrer, os momentos de invalidez podem alternar-se com momentos saudáveis.

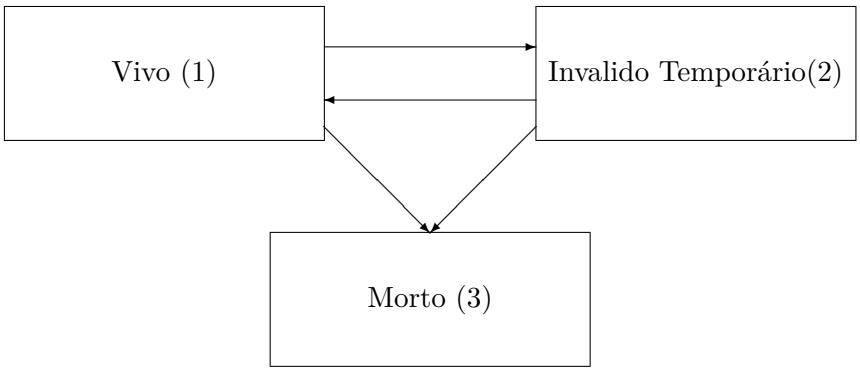


Figura 3.2: Seguro de Invalidez Temporário

Neste caso, a matriz de transição consiste em:

$$ {}_tP_x = \begin{bmatrix} {}_tp_x^{11} & {}_tp_x^{12} & {}_tp_x^{13} \\ {}_tp_x^{21} & {}_tp_x^{22} & {}_tp_x^{23} \\ {}_tp_x^{21} & {}_tp_x^{22} & {}_tp_x^{33} \end{bmatrix} = \begin{bmatrix} {}_tp_x^{11} & {}_tp_x^{12} & {}_tp_x^{13} \\ {}_tp_x^{21} & {}_tp_x^{22} & {}_tp_x^{23} \\ 0 & 0 & 1 \end{bmatrix} $$

3.5.2 Pressupostos

Segundo [25], os modelos de estados múltiplos têm três pressupostos:

1. Assume-se que o processo $\{S(t)\}_{t\geqslant 0}$ é uma cadeia de *Markov* a tempo contínuo, ou seja, para qualquer estado i e j, e momento x e $x+t$, a probabilidade $\mathbb{P}(S(x+t) = j|S(x) = i)$ não depende de qualquer informação anterior ao momento t.

Disto resulta que as probabilidades de futuros acontecimentos são completamente determinadas conhecendo-se o estado atual do indivíduo. Note-se que este pressuposto não é sempre necessário, repare-se no modelo de seguro de morte, quando o indivíduo se encontra no estado 1, vivo, sabe-se todo o seu passado.

2. Considera-se que a probabilidade de ocorrerem duas ou mais transições durante um intervalo de tempo h é infinitesimal de h, com $h \geqslant 0$, ou seja, $\mathbb{P}[\text{Ocorrerem duas ou mais transições em } h] = o(h)$, sendo $\lim\limits_{h\to 0^+} \dfrac{o(h)}{h} = 0$.

Este pressuposto nem sempre é necessário. Por exemplo, no seguro de morte há uma única transição do estado 1 para o estado 2, ou seja, nunca é possível haver mais de duas transições.

3. Assume-se que a função ${}_tp_x^{ij}$ é diferenciável em t, para qualquer $x \geq 0$ e estados i e j.

Este pressuposto é necessário para o cálculo das intensidades de transição, sendo mais técnico.

O seguro de dependência abordado no presente trabalho é um exemplo de um modelo de estados múltiplos, com 5 estados, que será detalhado mais à frente, ver capítulo 8.

Capítulo 4

Análise de *Clusters*

Tendo uma base de dados de grandes dimensões, muitas vezes é difícil uma análise eficaz desta, devido ao seu tamanho. Por isso, existem várias ferramentas para simplificação destes tipos de dados. Uma destas é a análise de *clusters*, também designada por *clustering*. Esta tem o objetivo de obtenção de informação inteligível e coerente dos dados.

4.1 Definição

Na literatura existem diversas definições de *clustering*, tais como:

- Para Han e Kamber, em [13], a análise de *cluster* ou *clustering* consiste num processo de partição de um conjunto de dados em subconjuntos, denominados *clusters*, de modo a que objetos de cada *cluster* sejam semelhantes entre si, mas diferentes dos objetos de outros *clusters*.

- Em [16], segundo Kaufman, a análise de *clustering* tem o objetivo de identificar grupos homogéneos de acordo com determinadas caraterísticas, em que exista homogeneidade dos elementos de cada grupo (homogeneidade intra-grupos) e diferenças entre grupos (heterogeneidade inter-grupos).

- Segundo Hastie et al., em [14], a segmentação de dados ou análise de *clusters* tem diversos objetivos. Todos se referem a agrupar ou segmentar uma coleção de dados em subconjuntos, de tal forma que objetos, que estão dentro de cada *cluster*, estão mais relacionados uns com os outros do que objetos atribuídos a diferentes *clusters*.

Assim sendo, a análise de *cluster* é um processo que tem o objetivo de dividir um conjunto de dados em subconjuntos, tais que:

- Existam semelhanças entre elementos do mesmo grupo;

- Grupos diferentes correspondem a dessemelhanças dos objetos.

Formalmente, considere-se um conjunto de n elementos, $\boldsymbol{X} = \{\boldsymbol{X_1}, \boldsymbol{X_2}, ...\boldsymbol{X_n}\}$, em que $\boldsymbol{X}_i \in \mathbb{R}^p$ é um vetor de p variáveis, estes devem ser particionados em k grupos distintos $\boldsymbol{C} = \{C_1, C_2, ...C_k\}$, sendo k o número de *clusters* tal que:

- $C_1 \cup C_2 \cup ... \cup C_k = \boldsymbol{X}$;
- $C_i \neq \varnothing$, $\forall i$, $1 \leq i \leq k$;
- $C_i \cap C_j = \varnothing$, $\forall i \neq j$, $1 \leq i \leq k$ e $1 \leq j \leq k$.

4.2 Medida de Similaridade

A semelhança entre objetos é analisada por medidas de similaridade; estas fornecem valores que exprimem distâncias entre certos aspetos de dois objetos, por isso, medidas de similaridade menores significam objetos semelhantes. Do mesmo modo, maiores distâncias resultam de objetos diferentes entre si. Defina-se como uma medida de similaridade ou distância a função $D(x, y)$, sendo $x, y, z \in \mathbb{R}^n$ tal que:

- Seja uma função positiva: $D(x, y) \geq 0$, $\forall x, y$;
- Seja simétrica, ou seja, $D(x, y) = D(y, x)$;
- Verifique a desigualdade triangular, isto é, $D(x, z) \leq D(x, y) + D(y, z)$, $\forall x, y, z$;
- Seja nula apenas com pontos coincidentes, $D(x, x) = 0$.

As medidas de similaridade utilizadas devem ser adequadas ao tipo de dados que se analisa.

Distância *Euclidiana*

Uma das medidas mais utilizada, para variáveis quantitativas, é a distância *Euclidiana*:

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4.1}$$

4.3 Requisitos Necessários a Algoritmos de Análise de *Clusters*

Em [13], um bom algoritmo de *clustering* deve verificar algumas caraterísticas, nomeadamente:

- O algoritmo deve ser escalável, isto é, se as observações se tornarem demasiado grandes (milhões de objetos) o método deverá conseguir suportá-las sem que o seu desempenho piore;

- Deve conseguir trabalhar com diversos tipos de dados, isto é, variáveis numéricas, binárias, nominais, ordinais ou uma mistura dos vários tipos;
- Um algoritmo deve conseguir obter *clusters* de forma arbitrária. Algoritmos baseados em distâncias como a *Euclidiana* tendem a encontrar *clusters* com densidade e tamanho semelhante, mas deveria ser possível obter grupos de qualquer tamanho, isto é, de forma arbitrária;
- Muitos dos algoritmos de *clustering* requerem que o utilizador determine alguns parâmetros de entrada (por exemplo, o número de *clusters*). Dados de grandes dimensões podem ser sensíveis a algumas alterações dos parâmetros iniciais, o que leva a que, para além de uma análise exaustiva dos resultados, haja uma difícil tarefa de determinação destes parâmetros; o que pode implicar um exigente controlo de qualidade do *clustering*;
- Os métodos devem ser robustos relativamente à presença de ruído, por exemplo, a maioria das bases de dados contém *outliers*, dados em falta, desconhecidos ou errados. A existência destes não deverá afetar a qualidade dos *clusters* obtidos;
- São necessários algoritmos insensíveis tanto à entrada de novos objetos, bem como à ordem dos elementos das bases de dados. Existem métodos que originam resultados muito diferentes, conforme a ordem dos objetos, o que não é desejável. Um mesmo conjunto de objetos, quando apresentado com diferentes ordens, deverá fornecer os mesmos resultados. Alguns algoritmos, aquando da entrada de novos objetos, têm de ser recalculados de novo, o que se torna num trabalho pesado, visto que, as bases de dados estão sempre em atualização;
- Deve haver uma boa capacidade de manuseamento de dados de alta dimensionalidade. Os métodos devem conseguir trabalhar, com eficiência, objetos de elevadas dimensões e fornecer resultados compreensíveis;
- A análise de *clusters* deve poder satisfazer algumas restrições, pois os problemas reais muitas vezes apresentam vários tipos de restrições e os métodos devem conseguir, para além de encontrarem os *clusters*, verificar as condições dadas pelos utilizadores;
- Dos algoritmos devem obter-se resultados interpretáveis, compreensíveis, utilizáveis e de representação simples.

4.4 Métodos de *Clustering*

Existem diversos algoritmos de análise de *clusters* que são organizados em conjuntos:

1. Métodos por partições (Secção 4.4.1);
2. Métodos hierárquicos (Secção 4.4.2);

3. Métodos baseados em densidade (Secção 4.4.3);

4. Outros métodos.

De seguida, será melhor explicado cada conjunto de métodos.

4.4.1 Métodos por Partições

Este tipo de métodos tem como objetivo, tendo-se um conjunto de dados, dividi-lo em subconjuntos, onde cada um destes é um *cluster*, tal que:

- Cada grupo tem pelo menos um elemento;
- Cada elemento pertence a um único grupo.

Estes algoritmos, iterativamente, encontram a melhor partição, de acordo com uma medida de similaridade, de modo a que objetos do mesmo *cluster* sejam o mais próximo possível, isto é, sejam semelhantes. Enquanto elementos de diferentes *clusters* devem ter distâncias maiores e ser diferentes. Existe sempre um elemento representativo do *cluster*, relativamente ao qual é calculado a medida de similaridade. Há duas possibilidades para este elemento que originam diferentes métodos:

- Se o centro for uma média dos elementos do *cluster*, tem-se então o algoritmo *K-means* (1967);
- Se o centro for o elemento mais representativo do *cluster*, está-se perante o algoritmo de cálculo dos *K-medoids*, *PAM* (4.4.1) e suas derivações: *CLARA* (4.4.1) (1990).

Na figura 4.1, tem-se um gráfico representativo deste tipo de métodos.

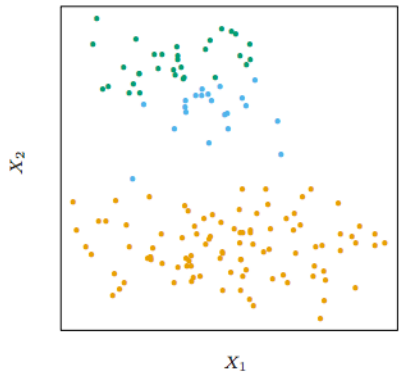


Figura 4.1: *Clustering* de dados usando algoritmo *k-means*, dados simulados

PAM

O *PAM* (*Partitioning around Medoids*, 1987) é um algoritmo usado para o cálculo dos *k-medoids*, ver [13]. Inicialmente, são escolhidos arbitrariamente objetos que serão representativos dos *clusters*, denominados *medoids*. Para todos os outros objetos são calculadas as suas distâncias a cada *medoid*, associando cada objeto ao *medoid* com menor distância. De seguida, de forma iterativa, fazem-se substituições de objetos representativos por não representativos. São feitas todas as substituições possíveis até não haver nenhuma melhoria na qualidade dos *clusters*.

No que se segue, representemos os *medoids* por $m_1, m_2, ..., m_k$ e os elementos da base de dados por $x_1, x_2, ..., x_n$. Para verificar se um objeto não representativo, $m_{aleatório}$, pode ser um bom substituto de um *medoid* $(m_1, m_2, ..., m_k)$, calcule-se a distância de cada objeto x_p $(p \leq n)$, a todos os objetos do conjunto $\{m_1, m_2, ..., m_{aleatório}, ..., m_k\}$. Pode verificar-se, um dos seguintes casos para cada elemento x_p, se m_j for substituído por $m_{aleatório}$ e:

1. x_p pertencia ao *cluster* com elemento representativo m_j:

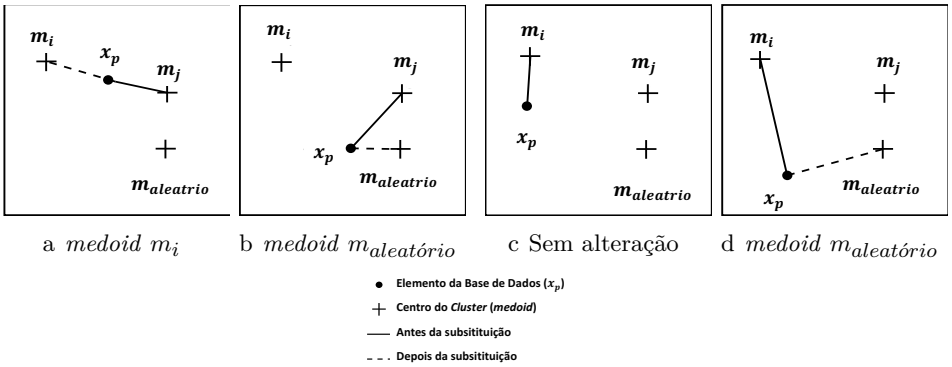
   a) Se $D(x_p, m_{aleatório}) > D(x_p, m_i)$, $i \neq j$, então x_p é associado ao *medoid* m_i; (Figura 4.2a)

   b) Se $D(x_p, m_{aleatório}) < D(x_p, m_i), i \neq j$, então x_p pertence ao *cluster* de $m_{aleatório}$; (Figura 4.2b)

2. x_p pertence ao grupo com o *medoid* m_i, $i \neq j$:

   a) Se $D(x_p, m_{aleatório}) > D(x_p, m_i)$, $i \neq j$, então a atribuição do *medoid* não se altera; (Figura 4.2c)

   b) Se $D(x_p, m_{aleatório}) < D(x_p, m_i)$, $i \neq j$, então x_p é associado ao *cluster* de $m_{aleatório}$. (Figura 4.2d)



a *medoid* m_i b *medoid* $m_{aleatório}$ c Sem alteração d *medoid* $m_{aleatório}$

Figura 4.2: Algoritmo de *k-medoids*

Seguidamente, é descrito o algoritmo de cálculo dos *k-medoids* passo a passo, sendo uma adaptação ao que é apresentado em [14].

Algoritmo de *PAM*

- k - número de *clusters*;
- $\{m_1, m_2, ..., m_k\}$ - conjunto de k *medoids*;
- $\{C_1, C_2, ..., C_k\}$ - conjunto dos k *clusters*;
- $\{x_1, x_2, ..., x_n\}$ - base de dados com n objetos;

1. Escolher, arbitrariamente, k objetos (*medoids* iniciais);
2. Associar cada objeto x_p ao *cluster* com o *medoid* mais próximo:

$$\forall \quad 1 \leq p \leq n \quad , \quad x_p \in C_j : j* = \underset{1 \leq j \leq k}{argmin}\{D(m_j, x_p)\}$$

3. Calcular a média das distâncias:

$$D_{média} = \frac{1}{n} \sum_{p=1}^{n} D(m_{j*}, x_p), \quad x_p \in C_{j*} \tag{4.2}$$

4. Arbitrariamente, escolher um novo *medoid* ($m_{aleatório}$) e substitui-lo por um existente, fazer o passo 1 e 2 e calcular a função custo:

$$C = D^2_{média} - D^1_{média} \tag{4.3}$$

   Sendo $D^1_{média}$ calculada com os primeiros *medoids* e $D^2_{média}$ calculada com o conjunto de *medoids* ao qual pertence $m_{aleatório}$ e não pertence o que foi substituído.

5. Fazer iterativamente os passos anteriores até a distância média (4.2) atingir o mínimo e a função custo (4.3) não se alterar.

Um exemplo gráfico do *PAM*, pode ser visto na figura 4.3, considerando um $k = 2$.

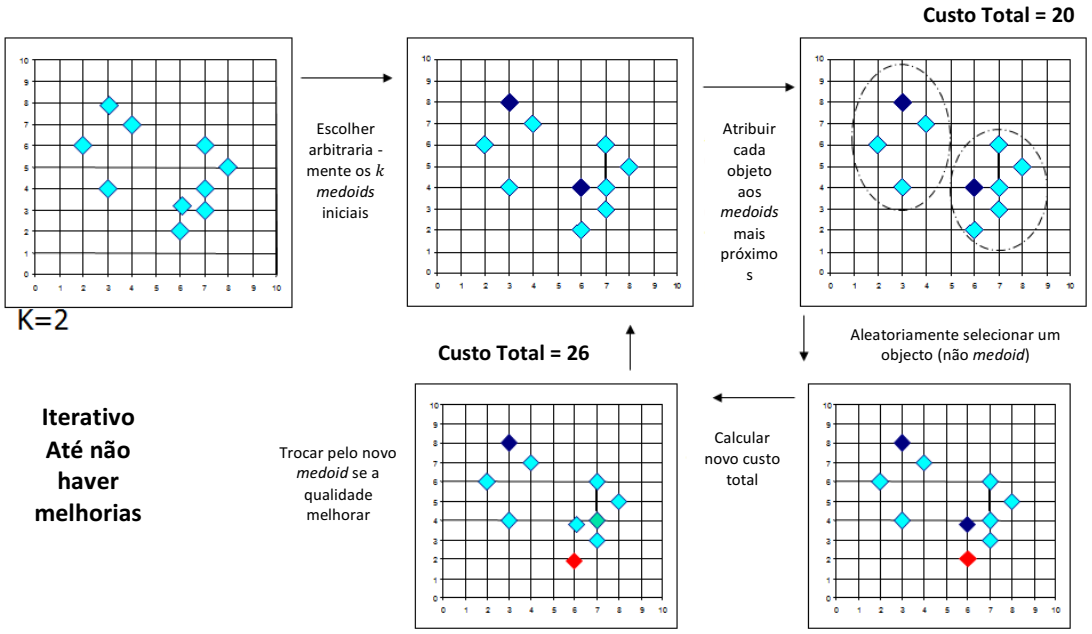


Figura 4.3: Exemplo do algoritmo *PAM*

CLARA

Em [13], Han e Kamber admitem que o método explicado anteriormente é eficaz para pequenas bases de dados, no entanto, o mesmo não se verifica para dados com elevadas dimensões. Por isso, existe um método para estudar conjuntos grandes de dados, baseado em amostragem chamado *CLARA* (*Clustering LARge Applicactions*). Este, em vez de aplicar o método *PAM* à base de dados toda, obtém diversas amostras aleatórias e calcula os *clusters* e respetivos *medoids* destas, pelo *PAM*, encontrando a melhor partição que corresponderá a de toda a base de dados. A eficiência de *CLARA* depende do tamanho da amostra e se esta é tendenciosa, ou seja, se forem diferentes as probabilidades de cada objeto ser selecionado para a amostra.

4.4.2 Métodos Hierárquicos

Os métodos hierárquicos criam uma decomposição hierárquica da base de dados que é representada por um dendograma, isto é, uma árvore que divide os dados em subconjuntos menores, veja-se uma exemplo na figura 4.4.

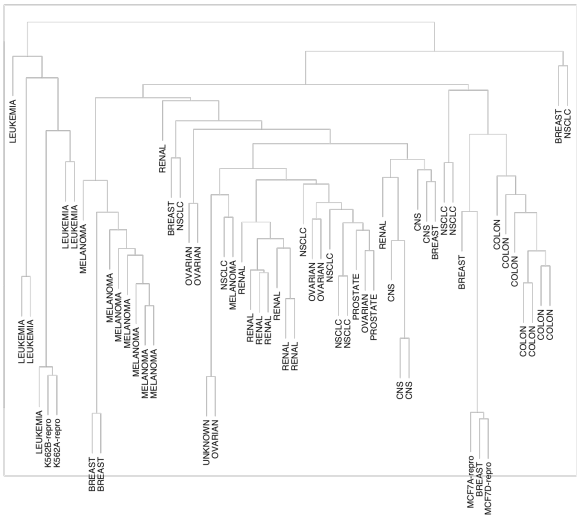


Figura 4.4: Dendograma

Existem dois tipos de métodos hierárquicos, os decisivos e os aglomerativos. Os primeiros algoritmos iniciam-se com um único *cluster* e, nas seguintes iterações, é escolhido um *cluster* que é dividido em dois mais pequenos; este processo termina quando o número de *clusters* for igual ao número de objetos ou até haver uma condição de término. DIANA (*DIvisive ANAlysis*) é um método deste tipo. Nos aglomerativos tem-se uma situação contrária à anterior: inicia-se com o número de *clusters* igual ao número de elementos da base de dados, nos passos seguintes, calculam-se as distâncias entre *clusters* e os que tiverem a distância mínima são agrupados até se ter um único ou existir uma condição de terminação válida, AGNES (*AGglomerative NESting*) é um método deste tipo. Na figura seguinte (4.5) é possível ver uma representação destes métodos:

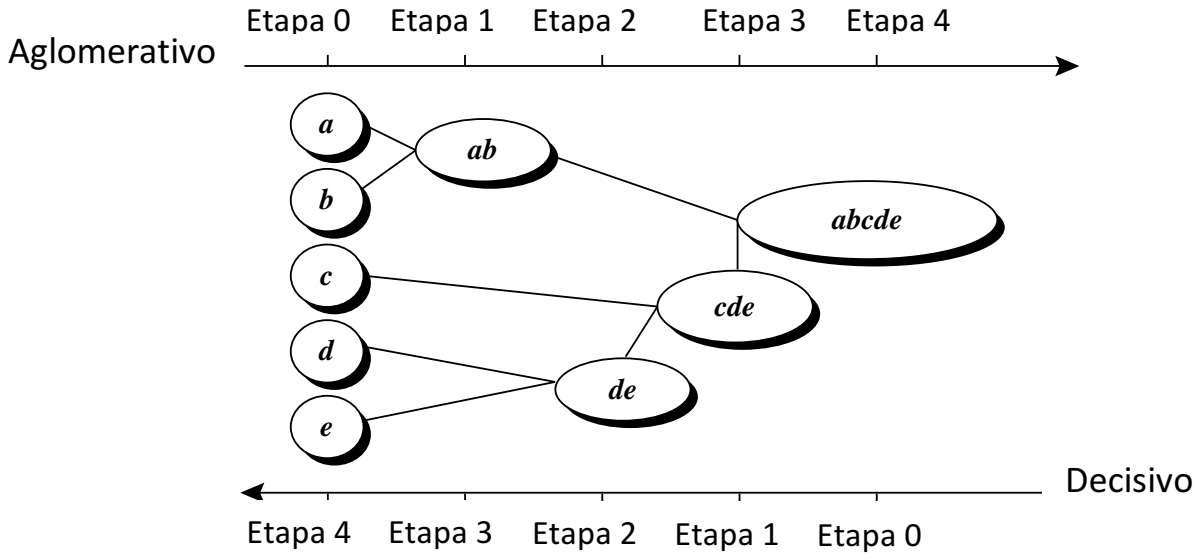


Figura 4.5: Métodos hierárquicos

4.4.3 Métodos Baseados em Densidades

Os métodos baseados em densidades particionam os dados com base na noção de densidade, isto é, pelo número de objetos de cada *cluster*, o que leva a construir *clusters* não esféricos, de forma arbitrária, conseguindo filtrar ruídos e/ou *outliers*. A ideia deste tipo de algoritmos é obter regiões mais densas com vizinhanças de menor densidade. Por exemplo, para cada objeto de um dado *cluster*, a vizinhança deve conter um número mínimo de objetos. DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), OPTICS (*Ordering points to identify the clustering structure*) e DENCLUE (*DENsity-based CLUstEring*) são exemplos de algoritmos baseados em densidade. Na figura 4.6, pode ver-se um exemplo gráfico destes métodos.

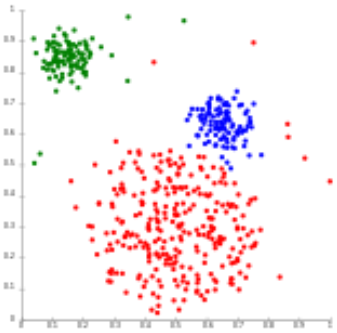


Figura 4.6: Método baseado em densidades

Existem muitos mais algoritmos de análise de *clusters*, pois esta é uma ferramenta fundamental para o estudo das bases de dados de grandes dimensões.

4.5 Índices de Validação

Muitas vezes é necessário fazer uma avaliação da qualidade do *clustering*, para, por exemplo, decidir o número de subconjuntos em que se deve dividir a base de dados. Assim, existem métodos intrínsecos que aferem a qualidade dos *clusters*, avaliando tanto a sua separação, bem como a sua compatação. Normalmente, estes métodos são medidas de similaridade. Segundo [13], uma destas medidas é o coeficiente de *Silhouette*.

4.5.1 Coeficiente de *Silhouette*

O coeficiente de *Silhouette* é uma medida de avaliação da qualidade dos *clusters*; a forma do seu cálculo será apresentada de seguida. Dado um conjunto de dados x, com n objetos, $\{x_1, x_2, ..., x_n\}$, considera-se que foi dividido em k *clusters*, $\{C_1, C_2, ..., C_k\}$. Para cada objeto, $x_p \in x$, calcule-se as medidas $a(x_p)$ e $b(x_p)$, que correspondem, respetivamente, à distância média entre x_p e todos os objetos do *cluster* a que x_p pertence e a distância

mínima média de x_p aos *clusters* que x_p não pertence. Supondo que $x_p \in C_i (1 \leq i \leq k)$, tem-se:

$$a(x_p) = \frac{\sum\limits_{p' \in C_i, p' \neq p} D(x_p, x'_p)}{|C_i| - 1} \tag{4.4}$$

$$b(x_p) = \min_{C_j, j \neq i, 1 \leq j \leq k} \left\{ \frac{\sum\limits_{p' \in C_j} D(x_p, x'_p)}{|C_j|} \right\} \tag{4.5}$$

Assim, o coeficiente de *Silhouette* do elemento x_p é calculado segundo a seguinte expressão:

$$s(x_p) = \frac{b(x_p) - a(x_p)}{\max\{a(x_p), b(x_p)\}} \tag{4.6}$$

Este coeficiente varia entre -1 e 1, em que o valor de $a(x_p)$ reflete o compatamento do *cluster* a que x_p pertence, pequenos valores correspondem a grandes compatamentos, ou seja, a distâncias pequenas entre elementos do *cluster*. Por outro lado, o valor de $b(x_p)$ representa o grau de separação entre o elemento x_p e os outros *clusters*, quanto maior $b(x_p)$, maior é a distância de x_p para os outros conjuntos. Assim, $s(x_p)$ próximo de 1 significa uma melhor situação, pois a distância entre elementos do *cluster*, a que x_p pertence, é pequena, e as distâncias entre os *clusters* são altas, no caso contrário, valores próximos de -1 representam um mau *clustering*, tem-se que $b(x_p) < a(x_p)$, ou seja, as distâncias entre *clusters* diferentes são menores do que as distâncias do mesmo *cluster*.

Para analisar toda a base de dados, e não apenas um elemento, utiliza-se a média dos coeficientes de *Silhouette*.

A qualidade do resultado final do trabalho poderá ser aferida à *posteriori* por meio de certos indicadores, tais como:

- A separação dos intervalos de confiança das médias das avaliações de cada *cluster* (7.2.4);
- A distância dos diferentes *medoids*, tabela 7.18;
- O valor médio para os tempos totais de permanência serem estáveis à medida que o tamanho da amostra aumenta (Capítulo 11).

A análise de *clusters* será utilizada numa base de trabalho para a modelação de um seguro *Long-Term Care*. Com uma base nos dados da Rede Nacional de Cuidados Continuados Integrados, utilizada esta dissertação, identificar-se-á os *clusters* que representam diferentes graus de dependência, o que permitirá um modelação do seguro com recurso aos Modelos de Estados Múltiplos.

Capítulo 5

Base de Dados

5.1 A Base de Dados de 2015

Foi fornecida uma base de dados da Rede Nacional de Cuidados Continuados Integrados (Secção 2.2.1). Neste capítulo será feita uma descrição desta. É de ter em conta que todas as observações e variáveis inviáveis e desnecessárias, para o presente trabalho, foram eliminadas. A base de dados, fornecida em *Excel*, é composta por seis ficheiros com dados, e mais um com informação sobre as variáveis:

1. *Resumo variáveis.xlsx* - apresenta as variáveis que os outros ficheiros têm;.
2. 00_*internamentos.csv* - este ficheiro tem 9 variáveis e 39 032 observações:
   a) *Contrato* - Nome encriptado da instituição que presta os cuidados;
   b) *ID Utente* - Número encriptado do processo clínico, ou seja, do utente;
   c) *ID Episódio* - Número encriptado do episódio do doente, aquando de uma entrada na rede do utente;
   d) *Data Início* - Data de admissão na instituição:
      - Data mais antiga: 30\09\2014;
      - Data mais recente: 13\01\2016;
      - Amplitude em dias: 470 (aproximadamente um ano e três meses e meio).
   e) *Data Alta* - Data da alta, ou seja, de saída da instituição:
      - Data mais antiga: 02\01\2015;
      - Data mais recente: 03\07\2016;
      - Amplitude em dias: 548 (aproximadamente um ano e meio).

f) *Dias de internamento* - Total de dias de internamento, ou seja, diferença das variáveis anteriores:

- Mínimo: 0 dias;
- Máximo: 446 dias;
- Média: 66 dias.

g) *Estado* - Estado em que se encontrava o utente:

- Internado: 30 931 observações;
- Alta: 8 101 observações.

h) *Região* - Local onde o utente vive:

- Norte: 13 105 observações;
- Centro: 7 924 observações;
- Lisboa Vale do Tejo: 10 736 observações;
- Alentejo: 3 476 observações;
- Algarve: 3 791 observações.

i) *Tipologia* - Tipo de cuidados que o doente recebeu:

- Equipa Intra-Hospitalar de Suporte em Cuidados Paliativos: 3 228 observações;
- Unidade de Média Duração e Reabilitação: 8 780 observações;
- Unidade de Longa Duração e Manutenção: 7 399 observações;
- Equipas de Cuidados Continuados Integrados: 11 018 observações;
- Unidade de Convalescença: 6 537 observações;
- Equipas Domiciliárias de Suporte em Cuidados Paliativos: 111 observações;
- Unidade de Cuidados Paliativos: 1 959 observações.

3. 02 _*caracterizaoscio-demografica .csv* - este ficheiro tem 5 variáveis e 70 618 observações:

a) *ID Utente* - Número do processo clínico;

b) *ID Episódio* - Número do episódio do doente;

c) *Concelho* - Local onde o utente vive, 18 concelhos diferentes;

d) *Género* - Género do utente:

- Feminino: 39 855 observações;
- Masculino: 30 763 observações.

e) *Idade* - Idade do utente aquando do episódio, na seguinte tabela (5.1) têm-se algumas medidas descritivas das idades dos indivíduos:

Tabela 5.1: Medidas descritivas das idades

Medidas	Valor
Mínimo	60
1ºQuartil	73
Mediana	80
Média	79.10
3ºQuartil	85
Máximo	107
Desvio Padrão	8.40
Amplitude	47

f) *Estado Civil* - Estado civil em que a pessoa se encontra aquando do episódio, esta variável apresenta alguns valores nulos; no entanto, esta variável não foi utilizada:

- Solteiro: 3 351 observações;
- Casado: 22 050 observações;
- Divorciado ou Separado: 19 39 observações;
- Viúvo: 15 161 observações;
- Desconhecido: 28 117 observações.

4. 09 *_evoluo_do_estado_cognitivo .csv* - este ficheiro tem 16 variáveis e 138 588 observações:

a) *ID Utente* - Número do processo clínico;

b) *ID Episódio* - Número do episódio do doente;

c) Existem 10 variáveis com a avaliação cognitiva dos utentes. Na tabela seguinte (5.2), é possível ver os seus valores;

Tabela 5.2: Número de observações das variáveis da avaliação cognitiva

Variáveis	Errado	Certo
Ano	62 931	75 657
Mês	61 585	77 003
Dia	75 189	63 399
Estação	59 290	79 298
D. Sem	69 545	69 043
País	45 303	93 285
Distrito	51 495	87 093
Terra	46 339	92 249
Casa	59 674	78 914
Andar	70 381	68 207
Nº. Utentes	40 132	

d) *Resultado* - Variável qualitativa que representa o resultado do estado cognitivo do utente:

- 0: 54 103 observações;
- 1: 15 594 observações;
- 2: 14 630 observações;
- 3: 54 261 observações.

e) *Denominação* - Variável quantitativa que representa a denominação da variável anterior, cada número corresponde a um nível do estado cognitivo:

- 0 → Mau;
- 1 → Insatisfatório;
- 2 → Satisfatório;
- 3 → Bom.

f) *Data IAI* [1] - Esta variável diz respeito à data de avaliação do indivíduo:

- Data mais antiga: 01\01\2015;
- Data mais recente: 31\12\2015;
- Amplitude em dias: 365 (um ano).

g) *Data do Episódio* - Data na qual o utente deu entrada na instituição:

- Data mais antiga: 01\01\2015;
- Data mais recente: 31\12\2015;
- Amplitude em dias: 365 (um ano).

5. 10_*evoluo_da_autonomia_fisica* .*csv* - este ficheiro tem 14 variáveis e 178 295 observações:

a) *ID Utente* - Número do processo clínico;

b) *ID Episódio* - Número do episódio do doente;

c) Existem 8 variáveis para a avaliação das atividades diárias dos utentes. A seguinte tabela (5.3) apresenta o número de observações que cada nível assume:

[1] Instrumento de Avaliação Integrado

Tabela 5.3: Número de observações das variáveis das atividades da vida diária

Variáveis	Incapaz	Dependente_3s	Meios	Independente
Lavar	45 974	112 079	9 629	10 613
Vestir	44 840	102 849	16 543	14 063
Sanita	62 687	76 600	14 394	24 614
Deitar	42 714	97 430	15 162	22 989
Sentar	41 439	95 013	17 428	24 415
Cont. Urina	71 485	21 728	41 885	43 197
Cont. Fezes	71 693	24 352	29 202	53 048
Alimentar	28 018	81 863	23 823	44 591
Nº. Utentes		40 022		

d) *Resultado* - Variável que representa o resultado numérico da avaliação das atividades da vida diária do utente:

- 0: 88542 observações;
- 1: 73444 observações;
- 2: 12025 observações;
- 3: 4284 observações.

e) *Denominação* - Variável nominal que representa a denominação da variável anterior:

- 0 → Incapaz;
- 1 → Dependente;
- 2 → Autónomo;
- 3 → Independente.

f) *Data IAI* - Esta variável diz respeito à data em que o utente foi avaliado:

- Data mais antiga: 01\01\2015;
- Data mais recente: 31\12\2015;
- Amplitude em dias: 365 (um ano).

g) *Data do Episódio* - Data em que o utente deu entrada na instituição:

- Data mais antiga: 01\01\2015;
- Data mais recente: 31\12\2015;
- Amplitude em dias: 365 (um ano).

6. 12 *_LOCOMOCAO.csv* - este ficheiro tem 6 variáveis e 179 061 observações:

a) *ID Utente* - Número do processo clínico;

b) *ID Episódio* - Número do episódio do doente;

c) *Data IAI* - Data em que foi avaliado o utente:

- Data mais antiga: 01\01\2015;
- Data mais recente: 31\12\2015;
- Amplitude em dias: 365 (um ano).

d) Existem 3 variáveis para a avaliação da locomoção dos indivíduos. O número de observações é possível ver na tabela seguinte (5.4):

Tabela 5.4: Número de observações das variáveis de locomoção

Variáveis	**Incapaz**	**Dependente_3s**	**Meios**	**Independente**
Casa	72 098	50 406	39 962	16 595
Rua	98 231	47 878	23 509	9 443
Escadas	124 502	30 488	14 548	9 523
Nº. Utentes		40 016		

7. 13_*Obitos_na_Rede .csv* - este ficheiro tem 8 variáveis e 12 249 observações:

a) *ID Utente* - Número do processo clínico;

b) *ID Episódio* - Número do episódio do doente;

c) *Data do Episódio* - Data do episódio anterior à morte:

- Data mais antiga: 01\01\2015;
- Data mais recente: 30\12\2015;
- Amplitude em dias: 363 (um ano).

d) *Data do Óbito* - Data em que o utente morreu:

- Data mais antiga: 02\01\2015;
- Data mais recente: 30\01\2016;
- Amplitude em dias: 393 (um ano e um mês).

e) *Data Internamento* - Data em que o utente foi internado:

- Data mais antiga: 17\12\2014;
- Data mais recente: 11\01\2016;
- Amplitude em dias: 390 (um ano e um mês).

f) *Internamento* - Dias desde que o indivíduo foi internado até à data de óbito:

- Mínimo: 0 dias (não chegou a ser internado);
- Máximo: 382 dias;
- Média: 48 dias.

g) *Episódio até ao óbito* - Número de dias desde que o utente deu entrada na rede até à data de óbito:

- Mínimo: 0 dias (não chegou a ser avaliado);

- Máximo: 369 dias;
- Média: 37 dias.

h) *Idade* - Idade do indivíduo à data de óbito:

- Mínimo: 60 anos;
- Máximo: 105 anos;
- Média: 80 anos aproximadamente.

5.2 Base de Dados Obtida por Junção

Para a análise da base de dados é necessário que cada utente tenha a avaliação das três áreas: Locomoção, Atividades da Vida Diária e Cognitivo. Uma vez que os ficheiros *Excel* têm, cada um, uma dimensão diferente, foi necessário agrupá-los. Para o mesmo utente deverá ter-se os vários episódios com datas e avaliações respetivas. Os ficheiros que se juntaram com as respetivas variáveis são:

- 09 *_evoluo_do_estado_cognitivo .csv*:
  - *ID Utente*;
  - *ID Episódio*;
  - *Data IAI*;
  - As 10 variáveis da avaliação: *Ano*, *Mês*, *Dia*, *Estação*, *Dia da Semana*, *País*, *Distrito*, *Terra*, *Casa* e *Andar*;
  - *Resultado*.
- 10*_evoluo_da_autonomia_fisica .csv*:
  - *ID Utente*;
  - *ID Episódio*;
  - *Data IAI*;
  - As 8 variáveis da avaliação: *Lavar*, *Vestir*, *Sanita*, *Deitar*, *Sentar*, *Continência Urinária*, *Continência das fezes*, *Alimentar*;
  - *Resultado*.
- 12 *_LOCOMOÇÃO.csv*:
  - *ID Utente*;
  - *ID Episódio*;
  - *Data IAI*;
  - As 3 variáveis da avaliação: *Casa*, *Rua* e *Escadas*.

- 02 _*caracterizaoscio-demografica .csv*:
  - *ID Utente*;
  - *ID Episódio*;
  - *Idade*;
  - *Concelho*;
  - *Género*;
  - *Estado Civil*.

Para juntar utilizou-se uma ferramenta do *software R*, a função *join_all*() do *package plyr*. Esta função junta bases de dados, considerando uma coluna comum, como é possível ver no código seguinte (5.1)

Listagem 5.1: Juntar Base de dados

```
a<-read.csv("09_evoluo_do_estado_cognitivo1.csv",header=TRUE, sep=";",
            stringsAsFactors=FALSE)

b<-read.csv("10_evoluo_da_autonomia_fisica.csv",header=TRUE, sep=";",
            stringsAsFactors=FALSE)

c<-read.csv("12_LOCOMOCAO.csv",header=TRUE, sep=";",
            stringsAsFactors=FALSE)

d<-read.csv("02_caracterizao_scio-demogrfica.csv",header=TRUE, sep=";",
            stringsAsFactors=FALSE)

library("plyr")

DadosAB <- join_all(list(a,b),"ID")
names(DadosAB)
nrow(DadosAB)

Dados_AB<-na.omit(DadosAB)
nrow(Dados_AB)

DadosABC <- join_all(list(Dados_AB,c),"ID")
Dados_ABC<-na.omit(DadosABC)
nrow(Dados_ABC)

DadosABCIdades <- join_all(list(Dados_ABC,d),"ID1")
Dados_ABCIdades<-na.omit(DadosABCIdades)
nrow(Dados_ABCIdades)
```

Obtendo-se, assim, uma base de dados nova, com diferentes dimensões; um único ficheiro com 93 145 observações e 30 variáveis.

5.2.1 Informação Descritiva da Base de Dados

Aspetos importantes da nova base de dados:

- Existem 23 894 utentes diferentes;
- A figura seguinte (5.1) apresenta o número de avaliações por indivíduo e a tabela 5.5 apresenta algumas medidas descritivas desta variável;

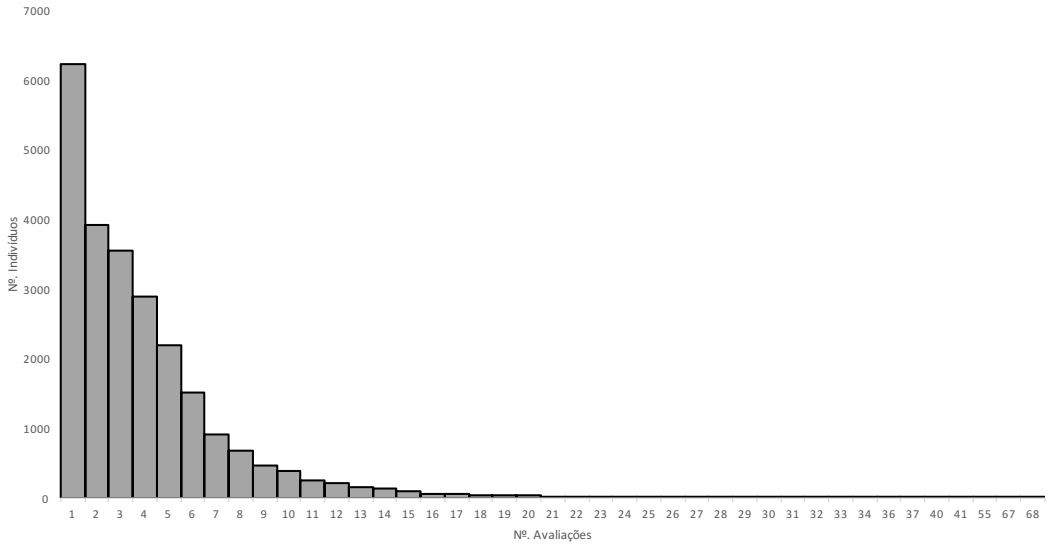


Figura 5.1: Distribuição do número de indivíduos por cada número de avaliações

Tabela 5.5: Medidas descritivas de avaliações por utente

Mínimo	**1ºQuartil**	**Mediana**	**Média**	**3ºQuartil**	**Máximo**
1.00	1.00	3.00	3.90	5.00	68.00

- As frequências das idades por utentes podem ver-se no histograma da Figura 5.2, bem como na Tabela 5.6 com algumas medidas de descrição:

Tabela 5.6: Medidas descritivas das idades

Mínimo	**1ºQuartil**	**Mediana**	**Média**	**3ºQuartil**	**Máximo**
60.00	73.00	80.00	79.03	85.00	107.00

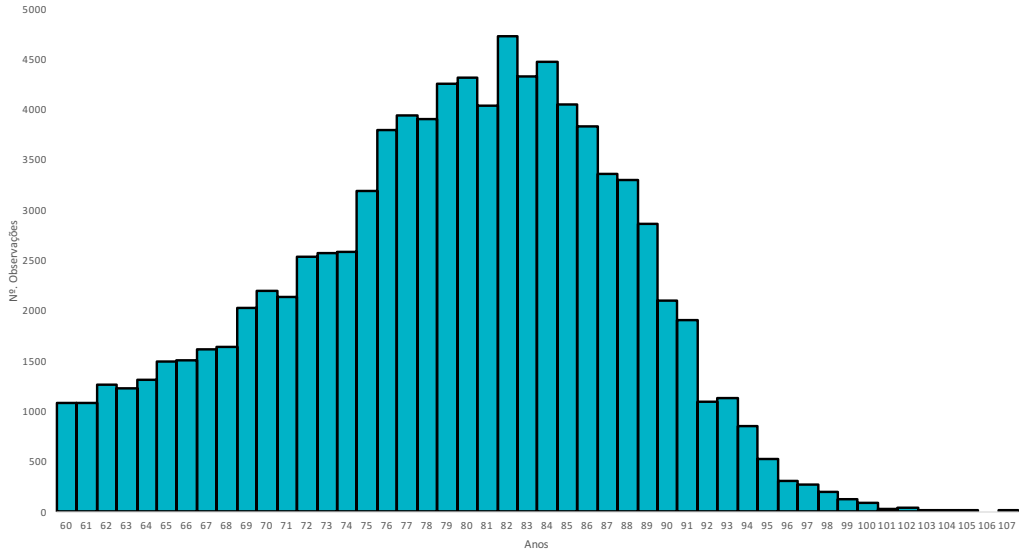


Figura 5.2: Histograma das idades

- Existem mais mulheres do que homens, tabela 5.7 e gráfico 5.3;

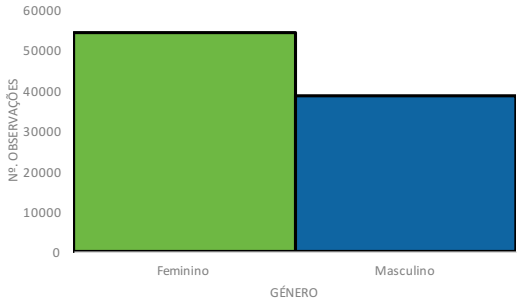


Figura 5.3: Histograma do género

Tabela 5.7: Número de observações por género

Feminino	**Masculino**
54 408	38 737

- O número de observações das avaliações do estado cognitivo pode ver-se na tabela seguinte (5.8);
- O número de observações das avaliações das atividades da vida diária pode ver-se na seguinte tabela (5.9);
- O número de observações das avaliações da locomoção dos utentes tem-se na seguinte tabela (5.10);
- Nesta base de dados, as avaliações decorrem desde o primeiro dia de 2015 (01\01\2015) ao último (31\12\2015), tendo assim uma amplitude de 364 dias de avaliações.

Tabela 5.8: Número de observações das avaliações do estado cognitivo

Variáveis	Errado	Certo
Ano	43 014	50 131
Mês	42 022	51 123
Dia	51 493	41 652
Estação	40 317	52 828
Dia Semana	47 622	45 523
Pais	30 187	62 958
Distrito	34 527	58 618
Terra	30 987	62 158
Casa	41 515	51 630
Andar	48 558	44 587
Nº. Utentes	23 894	

Tabela 5.9: Número de observações das avaliações das atividades da vida diária

Variáveis	Incapaz	Dependente_3s	Meios	Independente
Lavar	21 876	61 508	5 090	4 671
Vestir	21 233	56 593	9 061	6 258
Sanita	29 094	44 032	8 239	11 780
Deitar	19 507	54 237	8 559	10 842
Sentar	18 824	52 816	9 873	11 632
Cont. Urina	33 409	12 640	23 705	23 391
Cont. Fezes	33 392	14 090	16 439	29 224
Alimentar	12 616	42 539	13 104	24 886
Nº. Utentes	23 894			

Tabela 5.10: Número de observações das avaliações da locomoção

Variáveis	Incapaz	Dependente_3s	Meios	Independente
Casa	34 896	28 842	22 097	7 310
Rua	50 255	26 336	12 547	4 007
Escadas	63 529	17 474	8 156	3 986
Nº. Utentes	23 894			

CAPÍTULO 6

Tratamento e análise da Base de Dados

As variáveis da base de dados, apresentada na secção 5.2, são qualitativas e ordinais. Para serem utilizadas para o cálculo dos *clusters* é necessário que sejam quantitativas. Com esta transformação já é possível calcular medidas descritivas e fazer uma breve análise da população em estudo.

6.1 Variáveis Qualitativas

Uma vez que são ordinais, esta tarefa torna-se mais fácil.

Sendo assim, as variáveis referentes à avaliação do estado cognitivo do utente, que admitem o valor *Errado* ou *Certo*, podem, então, assumir o valor 0 e 3, respetivamente:

$$AvCog_{ij} = \{Ano_i; Mes_i; Dia_i; Estacao_i; DiaSemana_i; Pais_i; Distrito_i; Terra_i; Casa_i; Andar_i\}$$

Onde,

- i é referente à observação;
- j é referente à variável, $j \in \{1, 2, ..., 10\}$.

$$AvCog^*_{ij} = \begin{cases} 3, & se \quad AvCog_{ij} = Certo \\ 0, & se \quad AvCog_{ij} = Errado \end{cases}$$

As concretizações das variáveis da avaliação das atividades da vida diária são: *Incapaz*, *Dependente_3s* (Dependente de terceiros), *Meios* (necessita de meios para o fazer) e *Independente*, assim para se passarem a quantitativas tomam os valores: 0, 1, 2 e 3, respetivamente.

$$AvAVD_{ij} = \{Lavar_i; Vestir_i; Sanita_i; Deitar_i; Sentar_i; ContUrina_i; ContFezes_i; Alimentar_i\}$$

Onde,

- i é referente à observação;
- j é referente à variável, $j \leq 8$.

$$AvAVD^*_{ij} = \begin{cases} 3, & se \quad AvAVD_{ij} = Independente \\ 2, & se \quad AvAVD_{ij} = Meios \\ 1, & se \quad AvAVD_{ij} = Dependente_3s \\ 0, & se \quad AvAVD_{ij} = Incapaz \end{cases}$$

Por fim, têm-se as variáveis da avaliação da locomoção. Estas podem ser: *Incapaz*, *Dependente_3s* (Dependente de terceiros), *Meios* (necessita de meios para o fazer) e *Independente*. Ao transformá-las em quantitativas, estas podem assumir 0, 1, 2 e 3, respetivamente.

$$AvLoc_{ij} = \{Casa_i; Rua_i; Escadas_i\}$$

Onde,

- i é referente à observação;
- j é referente à variável, $j \leq 3$.

$$AvLoc^*_{ij} = \begin{cases} 3, & se \quad AvLoc_{ij} = Independente \\ 2, & se \quad AvLoc_{ij} = Meios \\ 1, & se \quad AvALoc_{ij} = Dependente_3s \\ 0, & se \quad AvLoc_{ij} = Incapaz \end{cases}$$

6.2 Análise Descritiva

Uma vez que as variáveis já são qualitativas já é possível calcular algumas medidas descritivas, necessárias para uma análise da base de dados.

Quanto às variáveis relativas à avaliação do estado cognitivo do utente, veja-se a tabela seguinte (6.1).

Tabela 6.1: Medidas descritivas das variáveis da avaliação cognitiva

	Ano	**Mês**	**Dia**	**Estação**	**Dia Semana**	**País**	**Distrito**	**Terra**	**Casa**	**Andar**
Mínimo	0	0	0	0	0	0	0	0	0	0
1º Quartil	0	0	0	0	0	0	0	0	0	0
Mediana	3	3	0	3	0	3	3	3	3	0
Média	1.61	1.65	1.34	1.70	1.47	2.03	1.89	2.00	1.66	1.44
3º Quartil	3	3	3	3	3	3	3	3	3	3
Máximo	3	3	3	3	3	3	3	3	3	3

As médias destas variáveis são próximas de 1.5, isto é, o valor intermédio entre 0(*Errado*) e 3 (*Certo*), ou seja, referente à avaliação cognitiva, está-se perante uma população moderadamente dependente. Veja-se que as variáveis referentes ao tempo são mais dependentes do que as do espaço.

Em relação à avaliação das atividades da vida diária, tem-se a tabela 6.2.

Tabela 6.2: Medidas descritivas das variáveis da avaliação das atividades da vida diária

	Lavar	**Vestir**	**Sanita**	**Deitar**	**Sentar**	**Cont. Urina**	**Cont. Fezes**	**Alimentar**
Mínimo	0	0	0	0	0	0	0	0
1º Quartil	1	1	0	1	1	0	0	1
Mediana	1	1	1	1	1	2	1	1
Média	0.92	1.00	1.03	1.12	1.15	1.40	1.45	1.54
3º Quartil	1	1	1	1	1	3	3	3
Máximo	3	3	3	3	3	3	3	3

Neste caso, as médias são muito próximas de 1(*Dependente_3s*), a população é mais dependente do que observando as variáveis da parte cognitiva.

Por fim, quanto à locomoção, apenas 3 variáveis, tem-se a seguinte tabela (6.3).

Tabela 6.3: Medidas descritivas das variáveis da avaliação da locomoção

	Casa	**Rua**	**Escadas**
Mínimo	0	0	0
1º Quartil	0	0	0
Mediana	1	0	0
Média	1.02	0.68	0.49
3º Quartil	2	1	1
Máximo	3	3	3

As médias das variáveis estão entre 0(*Incapaz*) e 1(*Dependente_3s*), ou seja, tem-se uma população severamente dependente.

Concluindo, a população em estudo é muito dependente, o que é de esperar uma vez que todos os utentes avaliados têm idade superior a 60 anos (5.6).

6.3 Normalização das Observações

Adaptando-se [13], para simplificação do cálculo dos *clusters* procedeu-se a uma normalização dos dados. Tendo-se uma variável x, com n observações, a observação x_i é normalizada a partir da seguinte expressão:

$$x_{i,normalizada} \equiv y_i = \frac{x_i - \min\limits_{0 \leq i \leq n} x_i}{\max\limits_{0 \leq i \leq n} x_i - \min\limits_{0 \leq i \leq n} x_i}$$

Ficando-se assim com todas as variáveis com valores entre 0 e 1.

Por fim, veja-se a matriz de correlação (tabela 6.4), as variáveis de cada tipo de avaliação são mais correlacionadas, uma vez que, por exemplo, se um utente não sabe a cidade onde vive pode implicar não saber o distrito, ou, um utente que não consegue andar em casa, muito provavelmente não irá conseguir subir nem descer escadas, e assim, para outras variáveis.

Tabela 6.4: Matriz de correlação das variáveis da avaliação

	Ano	Mês	Dia	Estação	Dia Semana	País	Distrito	Terra	Casa	Andar	Lavar	Vestir	Sanita	Deitar	Sentar	Cont. Urina	Cont. Fezes	Alimentar	Casa	Rua	Escadas
Ano	1.00	0.86	0.79	0.84	0.79	0.73	0.75	0.71	0.73	0.69	0.42	0.44	0.49	0.43	0.44	0.54	0.56	0.48	0.46	0.38	0.32
Mes	0.86	1.00	0.79	0.87	0.83	0.74	0.76	0.74	0.76	0.71	0.42	0.45	0.49	0.44	0.45	0.55	0.57	0.49	0.46	0.39	0.33
Dia	0.79	0.79	1.00	0.74	0.84	0.61	0.66	0.61	0.72	0.75	0.42	0.44	0.49	0.44	0.44	0.53	0.54	0.46	0.45	0.39	0.34
Estação	0.84	0.87	0.74	1.00	0.78	0.77	0.77	0.75	0.76	0.70	0.42	0.44	0.49	0.44	0.44	0.54	0.56	0.48	0.46	0.38	0.32
Dia Semana	0.79	0.83	0.84	0.78	1.00	0.66	0.69	0.67	0.75	0.75	0.42	0.45	0.50	0.45	0.45	0.55	0.57	0.48	0.46	0.39	0.34
País	0.73	0.74	0.61	0.77	0.66	1.00	0.87	0.89	0.72	0.63	0.40	0.42	0.45	0.42	0.42	0.50	0.53	0.47	0.44	0.35	0.29
Distrito	0.75	0.76	0.66	0.77	0.69	0.87	1.00	0.86	0.74	0.68	0.40	0.43	0.46	0.42	0.43	0.51	0.54	0.47	0.44	0.36	0.30
Terra	0.71	0.74	0.61	0.75	0.67	0.89	0.86	1.00	0.76	0.66	0.39	0.42	0.45	0.42	0.42	0.50	0.54	0.47	0.44	0.35	0.29
Casa	0.73	0.76	0.72	0.76	0.75	0.72	0.74	0.76	1.00	0.81	0.39	0.42	0.47	0.42	0.43	0.53	0.55	0.45	0.44	0.37	0.32
Andar	0.69	0.71	0.75	0.70	0.75	0.63	0.68	0.66	0.81	1.00	0.39	0.42	0.48	0.43	0.43	0.52	0.54	0.44	0.44	0.38	0.33
Lavar	0.42	0.42	0.42	0.42	0.42	0.40	0.40	0.39	0.39	0.39	1.00	0.88	0.72	0.74	0.72	0.54	0.54	0.57	0.62	0.59	0.55
Vestir	0.44	0.45	0.44	0.44	0.45	0.42	0.43	0.42	0.42	0.42	0.88	1.00	0.78	0.80	0.78	0.59	0.59	0.60	0.66	0.62	0.59
Sanita	0.49	0.49	0.49	0.49	0.50	0.45	0.46	0.45	0.47	0.48	0.72	0.78	1.00	0.84	0.84	0.69	0.69	0.61	0.73	0.65	0.63
Deitar	0.43	0.44	0.44	0.44	0.45	0.42	0.42	0.42	0.42	0.43	0.74	0.80	0.84	1.00	0.94	0.62	0.63	0.60	0.73	0.63	0.62
Sentar	0.44	0.45	0.44	0.44	0.45	0.42	0.43	0.42	0.43	0.43	0.72	0.78	0.84	0.94	1.00	0.63	0.63	0.61	0.74	0.63	0.62
Cont. Urina	0.54	0.55	0.53	0.54	0.55	0.50	0.51	0.50	0.53	0.52	0.54	0.59	0.69	0.62	0.63	1.00	0.88	0.60	0.61	0.51	0.48
Cont. Fezes	0.56	0.57	0.54	0.56	0.57	0.53	0.54	0.54	0.55	0.54	0.54	0.59	0.69	0.63	0.63	0.88	1.00	0.62	0.62	0.51	0.47
Alimentar	0.48	0.49	0.46	0.48	0.48	0.47	0.47	0.47	0.45	0.44	0.57	0.60	0.61	0.60	0.61	0.60	0.62	1.00	0.56	0.47	0.41
Casa	0.46	0.46	0.45	0.46	0.46	0.44	0.44	0.44	0.44	0.44	0.62	0.66	0.73	0.73	0.74	0.61	0.62	0.56	1.00	0.76	0.67
Rua	0.38	0.39	0.39	0.38	0.39	0.35	0.36	0.35	0.37	0.38	0.59	0.62	0.65	0.63	0.63	0.51	0.51	0.47	0.76	1.00	0.76
Escadas	0.32	0.33	0.34	0.32	0.34	0.29	0.30	0.29	0.32	0.33	0.55	0.59	0.63	0.62	0.62	0.48	0.47	0.41	0.67	0.76	1.00

Assim, já é possível encontrar os diversos *clusters* da base de dados.

CAPÍTULO 7

Cálculo de *Clusters*

Um dos objetivos desta dissertação consiste em encontrar conjuntos de utentes com o mesmo grau de dependência e, a partir daí, encontrar um representante de cada grupo e poder estabelecer caraterísticas dos diferentes graus de dependência. Para isso, foi utilizada a técnica de análise de *clusters* (Capítulo 4). Esta técnica, como referido anteriormente, divide uma base de dados em conjuntos com elementos homogéneos entre si.

7.1 Análise de *Clusters* - *Software R*

Pretende-se dividir a base de dados em subconjuntos disjuntos entre si e com o mesmo nível hierárquico. Para isso, utilizaram-se os métodos por partição (4.4.1). Dentro destes, como se tem o objetivo de obter os utentes mais representativos dos subgrupos, usou-se um método para o cálculo dos *K-medoids*. Dado que se está perante uma base de dados de elevadas dimensões, o método *CLARA* (4.4.1) foi o utilizado.

Assim, tirando partido do *software R*, usando a função *clara*() do *package cluster*, é possível calcular, então, subconjuntos. Esta função necessita de alguns argumentos, que aqui se expõem:

- a base de dados, cujas linhas correspondem às observações e as colunas às variáveis;
- o número de *clusters* no qual a base de dados é dividida, este tem de ser menor que o número de observações;
- a medida de dissimilaridade a ser utilizada para calcular os subconjuntos;
- se pretende que a base de dados seja estandardizada antes do cálculo das medidas de dissimilaridade, esta normalização consiste na subtração da média da variável e na divisão pelo desvio padrão;

- o número de amostras ao qual é aplicado o método para obter os *clusters*;
- o número de observações em cada amostra;
- se os *medoids* devem ser apresentados no *output*;
- por fim, se deve ser utilizada a mesma função objetivo que o método *PAM* utiliza.

A partir do código 7.1, foi possível obter, então, os *clusters* da base de dados.

Listagem 7.1: *Clustering* dos dados

```
clusters_Medoids_Dados<-clara(Medoids_Dados, k , metric = "euclidean",
stand=FALSE,  samples=10000,  sampsize=1000, medoids.x = TRUE,
pamLike = TRUE)
```

Tem-se que:

- *Medoids_Dados* é a base de dados normalizada (Secção 6.3);
- k é o número de *clusters* que se pretende obter;
- Foi utilizada como medida de dissimilaridade a distância *Euclidiana*;
- Não se pretende que a base de dados seja estandardizada;
- O método *CLARA* usa amostras da base de dados para calcular os *clusters* e verifica a que tem a melhor função objetivo, sendo assim, neste caso foram feitas 10 000 amostras de tamanho 100;
- Pretendeu-se que os *medoids* fossem apresentados no *output*;
- Por fim, a função objetivo utilizada tem de ser a mesma que no método *PAM*.

De seguida, serão apresentados os *medoids* não normalizados de cada *cluster*, que correspondem aos utentes representativos de cada grau de dependência.

Considerando que são graus de dependência, menos que 3 seria pouco, uma vez que ter-se-ia apenas um estado de dependência e o estado saudável. Assim, iniciou-se com 3 *clusters* (tabela 7.1).

Tabela 7.1: *Medoids* de 3 *clusters*

Cognitivo										AVD								Locomoção		
Ano	Mês	Dia	Estação	Dia Semana	Pais	Distrito	Terra	Casa	Andar	Lavar	Vestir	Sanita	Deitar	Sentar	Cont. Urina	Cont. Fezes	Alimentar	Casa	Rua	Escadas
3	3	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	2	2	2
3	3	3	3	3	3	3	3	3	3	1	1	1	1	1	2	2	2	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Neste caso (tabela 7.1), estava-se perante os seguintes representantes:

1. Um utente saudável:
    - A avaliação cognitiva é a melhor possível (3), ou seja, o utente sabe onde está e em que momento do tempo;

- A avaliação das atividades da vida diária, maior parte das variáveis têm o valor máximo (3), só duas apresentam o valor antes do máximo (2), *lavar* e *vestir*, que significa que precisam de meios;
- Por fim, as variáveis referentes à locomoção, apesar de não serem máximas, os valores que apresentam (2), representam que necessitam de meios.

2. Um utente fracamente dependente:

   - As variáveis da avaliação cognitiva é a melhor possível (3), ou seja, sabe onde está e em que momento do tempo;
   - Na maior parte das atividades da vida diária, o utente necessita de ajuda de terceiros (1) e nas outras precisa apenas de meios para as fazer (2);
   - Por fim, as variáveis referentes à locomoção, o utente necessita de ajuda de terceiros (1) para andar em *Casa* e na *Rua*, e é incapaz (0) de subir e descer *Escadas*.

3. Um utente severamente dependente:

   - As variáveis da avaliação cognitiva têm o valor mais baixo (0), ou seja, o utente nem sabe onde está nem em que momento do tempo vive;
   - O indivíduo é incapaz de fazer todas as atividades da vida diária (0);
   - Por último, o utente é incapaz de se mover em qualquer lugar (0).

Seguidamente, dividiu-se a base de dados em 4 (tabela 7.2).

Tabela 7.2: *Medoids* de 4 *clusters*

Cognitivo										AVD								Locomoção		
Ano	Mês	Dia	Estação	Dia Semana	Pais	Distrito	Terra	Casa	Andar	Lavar	Vestir	Sanita	Deitar	Sentar	Cont. Urina	Cont. Fezes	Alimentar	Casa	Rua	Escadas
3	3	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	2	2	2
3	3	3	3	3	3	3	3	3	3	1	1	1	1	1	2	2	2	1	1	0
0	0	0	0	0	3	3	3	0	0	1	1	1	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Neste caso, obtiveram-se os seguintes representantes:

1. Um utente saudável - igual ao caso de 3 *clusters*;
2. Um utente fracamente dependente - o mesmo que o caso anterior;
3. Um utente moderadamente dependente:

   - O indivíduo só sabe o *País*, o *Distrito* e a *Terra* (3);
   - Em todas as atividades da vida diária, o utente necessita de ajuda de terceiros (1);
   - Por último, o utente necessita de ajuda de terceiros (1) para se mover em *Casa*, nos outros sítios é incapaz (0).

4. Um utente severamente dependente - os mesmos resultados que no caso com 3 *medoids*.

Por fim, foi dividida a base de dados em 5 subconjuntos, tabela 7.3:

Tabela 7.3: *Medoids* de 5 *clusters*

Cognitivo										AVD								Locomoção		
Ano	Mês	Dia	Estação	Dia Semana	Pais	Distrito	Terra	Casa	Andar	Lavar	Vestir	Sanita	Deitar	Sentar	Cont. Urina	Cont. Fezes	Alimentar	Casa	Rua	Escadas
3	3	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	2	2	2
3	3	3	3	3	3	3	3	3	3	1	1	1	1	1	2	2	2	1	1	0
0	0	0	0	0	3	3	3	0	0	1	1	1	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Comparando com os utentes do caso de 4 *clusters*, a única diferença é o novo utente com as seguintes caraterísticas:

1. Referente às variáveis da avaliação das atividades da vida diária e da locomoção têm o mesmo valor que o utente com dependência severa;

2. A única diferença está nas três variáveis referentes à avaliação cognitiva que o utente, com dependência severa, tinha positivas (*País*, *Distrito* e *Terra*), neste caso este novo utente não sabe onde está de todo (0).

Posto isto, foi necessário decidir qual o número de *clusters* ótimo. Com 5 *clusters* não há uma grande desigualdade dos utentes com maior grau de dependência, por isso, foi logo posto de lado o casos dos 5 *clusters*.

O utente representativo adicionado nos 4 *clusters* tem diferenças significativas, uma vez que faz diferença um indivíduo precisar de meios para se alimentar ou precisar de terceiros. Sendo assim foi escolhido dividir a base de dados em 4 *clusters*. Adicionalmente, comparam-se as médias dos coeficientes de *Silhouette*, calculados a partir das expressões dadas na secção 4.5.1, tabela 7.4.

Tabela 7.4: Médias dos coeficientes de *Silhouette* para cada número de *clusters*

k	$\bar{s}$
3	0.39
4	0.35
5	0.26

Apesar de o coeficiente para 4 *clusters* não ser o mais elevado, para além de não estar longe do melhor, é preferível ao de 5 *clusters*. Por isso, continuou a considerar-se que a base de dados seria dividida em 4.

Como explicado em [11], foi feita uma análise de estabilidade dos *medoids*, isto é, se dividindo a base de dados, os representantes dos *clusters* se mantêm os mesmos, o que se verificou, pode levar a concluir que, se adicionarmos novos utentes, os elementos representativos dos graus de dependência se mantêm.

7.2 Caraterização dos Graus de Dependência

As observações de cada *cluster* são semelhantes entre si, de acordo com o critério de semelhança escolhido, e representam cada grau de dependência.

7.2.1 Estado Saudável

O estado saudável tem como elemento representativo o apresentando na tabela 7.5.

Tabela 7.5: Elemento representativo do estado saudável

Cognitivo										AVD								Locomoção		
Ano	Mês	Dia	Estação	Dia Semana	Pais	Distrito	Terra	Casa	Andar	Lavar	Vestir	Sanita	Deitar	Sentar	Cont. Urina	Cont. Fezes	Alimentar	Casa	Rua	Escadas
3	3	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	2	2	2

Ao calcularmos a média das variáveis para cada observação, obtiveram-se as seguintes medidas descritivas para essa média, tabela 7.6.

Tabela 7.6: Medidas descritivas para a média das variáveis no estado saudável

Estado Saudável	
Mínimo	1.53
1º Quartil	2.49
Mediana	2.63
Média	2.63
3º Quartil	2.81
Máximo	3.00
Desvio Padrão	0.23
N.º Observações	14 865

A partir destas medidas e com base no Teorema do Limite Central, pode obter-se um intervalo de confiança a 99% para a média das avaliações. Usando a fórmula $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$, com $z_{\frac{\alpha}{2}} = \mathbb{P}[z < 1 - \frac{\alpha}{2}]$ tem-se:

$$IC_{99\%} = (2.63\,;\, 2.64)$$

Isto significa que é altamente provável que, para um indivíduo saudável, a média das avaliações esteja neste intervalo particular. Adicionalmente, também é possível fazer este intervalo de confiança a 99% para as médias de cada tipo de avaliação (cognitivo, atividades da vida diária e locomoção), conforme se pode verificar na Tabela 7.7.

Tabela 7.7: Medidas descritivas para a média das variáveis dos diferentes tipos de avaliação no estado saudável

Estado Saudável			
	Avaliação Cognitiva	**Avaliação AVD**	**Avaliação Locomoção**
Mínimo	0.30	1.38	0.00
1º Quartil	3.00	2.25	1.67
Mediana	3.00	2.50	2.00
Média	2.89	2.54	1.98
3º Quartil	3.00	2.88	2.33
Máximo	3.00	3.00	3.00
Desvio Padrão	0.30	0.34	0.64
$IC_{99\%}$	**(2.88 ; 2.90)**	**(2.53 ; 2.55)**	**(1.97 ; 1.99)**

Em conclusão, o indivíduo no estado saudável deve ter todas as suas variáveis no valor mais elevado.

7.2.2 Estado Dependência Fraca

O elemento representativo do estado de dependência fraca pode ser apresentado na tabela seguinte (7.8).

Tabela 7.8: Elemento representativo do estado dependência fraca

Cognitivo										AVD								Locomoção		
Ano	Mês	Dia	Estação	Dia Semana	Pais	Distrito	Terra	Casa	Andar	Lavar	Vestir	Sanita	Deitar	Sentar	Cont. Urina	Cont. Fezes	Alimentar	Casa	Rua	Escadas
3	3	3	3	3	3	3	3	3	3	1	1	1	1	1	2	2	2	1	1	0

Considerando a média das variáveis, como feito no estado saudável, obteve-se a tabela 7.9.

Tabela 7.9: Medidas descritivas para a média das variáveis no estado dependência fraca

Estado Dependência Fraca	
Mínimo	0.90
1º Quartil	1.76
Mediana	1.95
Média	1.92
3º Quartil	2.14
Máximo	2.57
Desvio Padrão	0.29
N.º Observações	33 863

Do mesmo modo que foi feito para o estado saudável, também é possível fazer um intervalo de confiança a 99% das médias das avaliações:

$$IC_{99\%} = (1.92\,;\,1.93)$$

O que significa que um indivíduo com dependência fraca tem alta probabilidade da média das suas avaliações estar contida neste intervalo. Fazendo o mesmo para as médias de cada tipo de avaliação, obtêm-se os diferentes intervalos de confiança a 99 %, tabela 7.10.

Tabela 7.10: Medidas descritivas para a média das variáveis dos diferentes tipos de avaliação no estado dependência fraca

Estado Dependência Fraca			
	Avaliação Cognitiva	**Avaliação AVD**	**Avaliação Locomoção**
Mínimo	1.20	0.00	0.00
1º Quartil	2.70	1.00	0.33
Mediana	3.00	1.38	0.67
Média	2.80	1.29	0.70
3º Quartil	3.00	1.63	1.00
Máximo	3.00	2.50	3.00
Desvio Padrão	0.33	0.47	0.56
$IC_{99\%}$	**(2.79 ; 2.80)**	**(1.28 ; 1.30)**	**(0.69 ; 0.70)**

7.2.3 Estado Dependência Moderada

No estado de dependência moderada, o elemento representativo é apresentado na tabela 7.11.

Tabela 7.11: Elemento representativo do estado dependência moderada

Cognitivo										**AVD**								**Locomoção**		
Ano	**Mês**	**Dia**	**Estação**	**Dia Semana**	**Pais**	**Distrito**	**Terra**	**Casa**	**Andar**	**Lavar**	**Vestir**	**Sanita**	**Deitar**	**Sentar**	**Cont. Urina**	**Cont. Fezes**	**Alimentar**	**Casa**	**Rua**	**Escadas**
0	0	0	0	0	3	3	3	0	0	1	1	1	1	1	1	1	1	1	0	0

Considerando a média das variáveis, tem-se a tabela seguinte (7.12):

Tabela 7.12: Medidas descritivas para a média das variáveis no estado dependência moderada

Estado Dependência Moderada	
Mínimo	0.29
1º Quartil	0.81
Mediana	1.00
Média	1.03
3º Quartil	1.24
Máximo	2.05
Desvio Padrão	0.33
N.º Observações	14 933

O intervalo de confiança a 99% das médias das avaliações é dado por:

$$IC_{99\%} = (1.02\,;\,1.04)$$

O que significa que um indivíduo con dependência moderada tem alta probabilidade da média das avaliações estar neste intervalo. A tabela 7.13 apresenta os intervalos de confiança a 99% para as médias de cada tipo de avaliação.

Tabela 7.13: Medidas descritivas para a média das variáveis dos diferentes tipos de avaliação no estado dependência moderada

Estado Dependência Moderada			
	Avaliação Cognitiva	**Avaliação AVD**	**Avaliação Locomoção**
Mínimo	0.00	0.00	0.00
1º Quartil	0.90	0.63	0.00
Mediana	1.20	1.00	0.33
Média	1.17	1.03	0.57
3º Quartil	1.50	1.38	1.00
Máximo	2.10	3.00	3.00
Desvio Padrão	0.46	0.59	0.64
$IC_{99\%}$	**(1.16 ; 1.78)**	**(1.02 ; 1.05)**	**(0.56 ; 0.59)**

7.2.4 Estado Dependência Severa

Por último, na tabela 7.14, é apresentado o elemento representativo do estado de dependência severa.

Tabela 7.14: Elemento representativo do estado dependência severa

Cognitivo										AVD								Locomoção		
Ano	Mês	Dia	Estação	Dia Semana	Pais	Distrito	Terra	Casa	Andar	Lavar	Vestir	Sanita	Deitar	Sentar	Cont. Urina	Cont. Fezes	Alimentar	Casa	Rua	Escadas
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Na tabela 7.15, é possível observar algumas medidas descritivas para a média das variáveis.

Tabela 7.15: Medidas descritivas para a média das variáveis no estado dependência severa

Estado Dependência Severa	
Mínimo	0.00
1º Quartil	0.00
Mediana	0.24
Média	0.24
3º Quartil	0.38
Máximo	1.43
Desvio Padrão	0.23
N.º Observações	29 484

Obteve-se um intervalo de confiança a 99% das médias das avaliações:

$$IC_{99\%} = (0.24\,;\,0.25)$$

Isto é, um indivíduo dependente severo tem alta probabilidade da média das avaliações estar no intervalo anterior. Para as médias de cada tipo de avaliação têm-se a tabela 7.16.

Tabela 7.16: Medidas descritivas para a média das variáveis dos diferentes tipos de avaliação no estado dependência severa

Estado Dependência Severa			
	Avaliação Cognitiva	**Avaliação AVD**	**Avaliação Locomoção**
Mínimo	0.00	0.00	0.00
1º Quartil	0.00	0.00	0.00
Mediana	0.00	0.50	0.00
Média	0.04	0.51	0.22
3º Quartil	0.00	0.75	0.33
Máximo	1.80	2.13	3.00
Desvio Padrão	0.16	0.45	0.42
$IC_{99\%}$	**(0.039 ; 0.043)**	**(0.50 ; 0.51)**	**(0.21 ; 0.23)**

Com estas tabelas, é possível determinar qual o estado em que um indivíduo se encontra. Veja-se um exemplo de um indivíduo com as seguintes avaliações (tabela 7.17):

Tabela 7.17: Exemplo

Cognitivo										**AVD**								**Locomoção**		
Ano	**Mês**	**Dia**	**Estação**	**Dia Semana**	**Pais**	**Distrito**	**Terra**	**Casa**	**Andar**	**Lavar**	**Vestir**	**Sanita**	**Deitar**	**Sentar**	**Cont. Urina**	**Cont. Fezes**	**Alimentar**	**Casa**	**Rua**	**Escadas**
3	3	3	3	3	0	0	0	0	0	2	2	2	2	2	2	2	2	1	1	1

Neste caso, tem-se que a média das variáveis é 1.62, a média da avaliação cognitiva, das atividades da vida diária e da locomoção são 1.5 , 2 e 1, respetivamente. Quanto à parte cognitiva, estaríamos no caso de dependência moderada, no caso das atividades da vida diária, estaríamos no caso saudável, no caso de locomoção, no estado de dependência fraca, observando as medias das variáveis. Apesar das médias serem próximas de diferentes estados, iremos para o estado intermédio desses, ou seja, o estado de dependência fraca.

Outra maneira de obter este resultado é calcular a distância *Euclidiana* desta observação a cada *medoid*:

$$D(\text{Saudavel}Medoid,\ \text{Obs}) = \sqrt{\sum_{i=1}^{21}(\text{Saudável}Medoid_i - Obs_i)^2} = 7.35$$

$$D(\text{DependenciaFraca}Medoid,\ \text{Obs}) = \sqrt{\sum_{i=1}^{21}(\text{DependenciaFraca}Medoid_i - Obs_i)^2} =$$

7.14

$$D(\text{DependenciaModerada}Medoid,\ \text{Obs}) = \sqrt{\sum_{i=1}^{21}(\text{DependenciaModerada}Medoid_i - Obs_i)^2} = 9.06$$

$$D(\text{DependenciaSevera}Medoid,\ \text{Obs}) = \sqrt{\sum_{i=1}^{21}(\text{DependenciaSevera}Medoid_i - Obs_i)^2} = 8.94$$

A distância mínima é obtida com o *medoid* do estado da dependência fraca, ou seja, neste caso o utente está no estado de dependência fraca.

Repare-se na tabela 7.18, esta indica as distâncias entre os diferentes *medoids*.

Tabela 7.18: Distância entre os *medoids*

	Saudável	**Dependência Fraca**	**Dependência Moderada**	**Dependência Severa**
Saudável	0.00	4.80	9.90	12.81
Dependência Fraca	4.80	0.00	8.19	10.44
Dependência Moderada	9.90	8.19	0.00	6.00
Dependência Severa	12.81	10.44	6.00	0.00

Note-se que, quanto maiores as diferenças dos graus de dependência, maiores são as distâncias entre os *medoids* de cada *cluster*.

Com o objetivo de dividir a base de dados em subconjuntos que têm elementos com o mesmo grau de dependência, utilizou-se, então, a análise de *clusters*. Daí resultaram quatro grupos: saudável, dependência fraca, dependência moderada e dependência severa. Estes são bastante distintos, como é possível observar nos intervalos de confiança das várias avaliações, pois são muito distantes.

Capítulo 8

Estimação de Matrizes de Transições de Probabilidades

Uma vez que já foram obtidos os *clusters*, cada registo está associado a um estado: saudável, dependência fraca, dependência moderada e dependência severa. Por isso, criou-se uma matriz que tem apenas 4 colunas. São estas, pela ordem apresentada:

1. *ID Utente* - que corresponde ao nome do utente;
2. *Data IAI* - data em que foi feita a avaliação;
3. *Idade* - a idade do indivíduo a que foi feita a avaliação;
4. *Estado* - estado em que cada utente se encontra à data de avaliação.

No entanto, ainda existe informação sobre o óbito dos utentes, por isso, tem-se um novo estado, que corresponde à morte do utente. Assim, à matriz anterior acrescentaram-se todas as observações dos indivíduos, pertencentes à base de dados, na qual o óbito foi registado e a data correspondente.

Com os óbitos, a base de dados aumentou, tendo-se 105 394 observações, em que os estados se distribuem segundo o seguinte histograma (8.1).

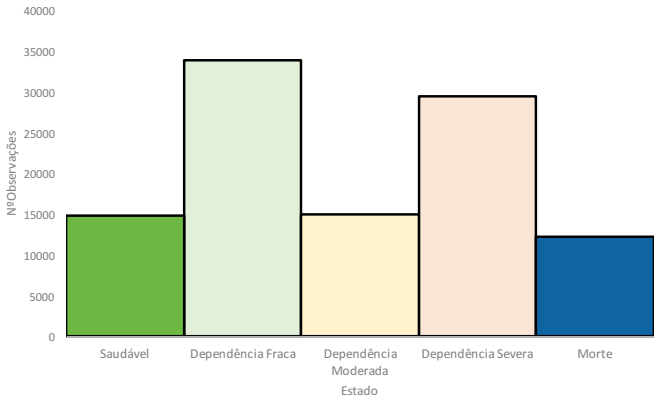


Figura 8.1: Distribuição das observações nos diferentes estados

Uma vez que se pode admitir, em primeira aproximação, que um indivíduo estar num determinado estado não depende do passado, considerou-se estar-se perante uma cadeia de *Markov*, ou mais especificamente, um modelo de estados múltiplos, com cinco estados, cujas transições são possíveis de observar na seguinte figura (8.2).

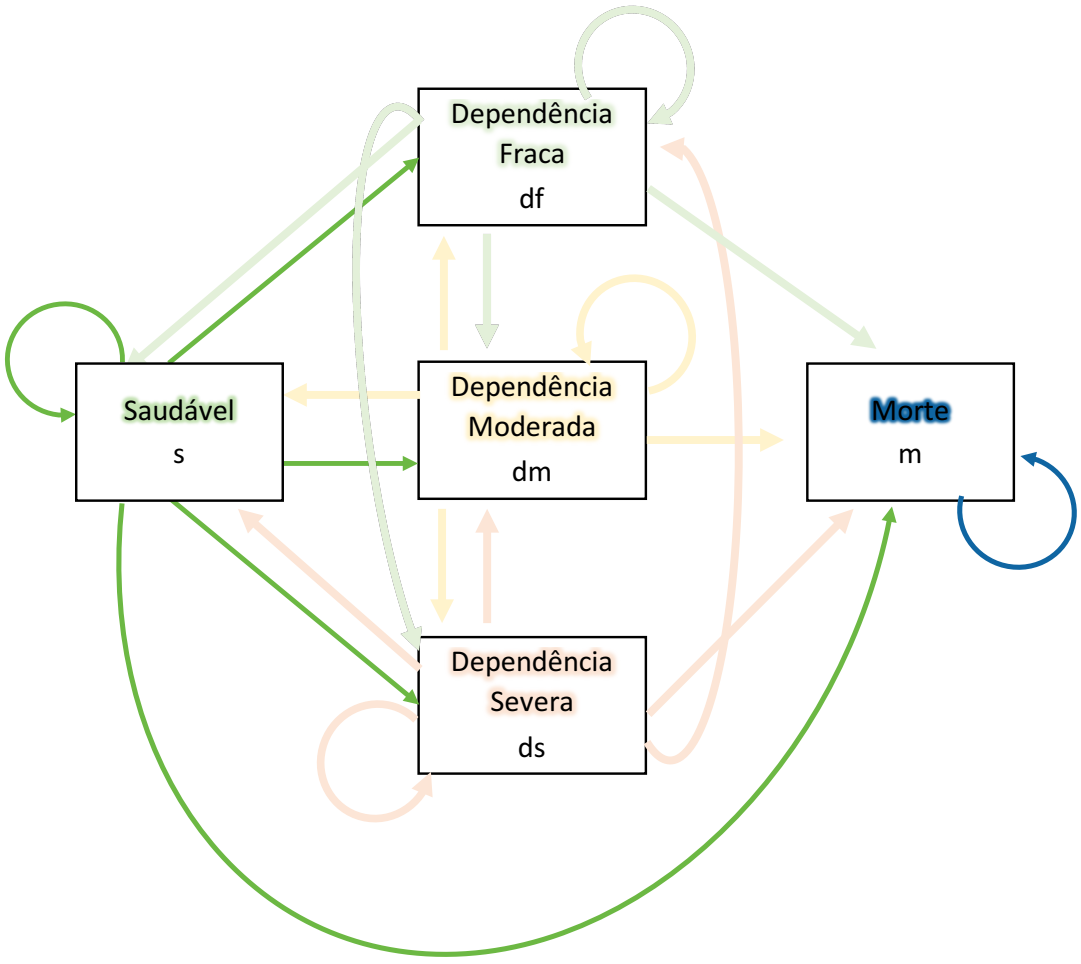


Figura 8.2: Representação do Modelo de Estados Múltiplos

O *software R* apresenta um *package* para o tratamento destes modelos, *msm*. Para o cálculo da matriz de transições, foi utilizada uma função do pacote referido anteriormente, chamada de *statetable.msm*. Esta necessita de apenas 3 argumentos, pela ordem apresentada:

1. As observações dos estados, que se assume estarem por ordem temporal;

2. Os nomes dos indivíduos a que correspondem os estados, cada indivíduo deve ter vários estados;

3. Nome da base de dados.

Para as observações dos estados estarem por ordem temporal, recorreu-se ao *Excel* e à sua ferramenta de ordenamento. Em primeiro lugar, aplicou-se ao nome dos utentes, ou seja, à variável *ID Utente*, de seguida à coluna das datas da avaliação (*Data IAI*).

Com as observações ordenadas e usando a função *statetable.msm* têm-se a seguinte matriz de transições (8.1):

Tabela 8.1: Matriz de contagem de transições

	Saudável	Dep. Fraca	Dep. Mod.	Dep. Sev.	Morte
Saudável	7929	1313	291	83	210
Dep. Fraca	4017	18794	2141	1067	1038
Dep. Mod.	494	2215	6507	2308	655
Dep. Sev.	168	1176	2581	18037	2553
Morte	0	0	0	0	0

No entanto, uma vez que se tem informação sobre as idades dos utentes, foram feitas matrizes por conjunto de idades. Estes conjuntos foram determinados com o objetivo de se ter os mesmos números de observações em cada matriz. Para isso, as idades foram divididas da seguinte forma (tabela 8.2 e figura 8.3)

Tabela 8.2: Número de observações por cada conjunto de idades

Idades	**Nº.Observações**
[60 ; 71]	20 752
[72 ; 77]	20 613
[78 ; 81]	18 424
[82 ; 86]	24 231
[87 ; 107]	21 374
Total	**105 394**

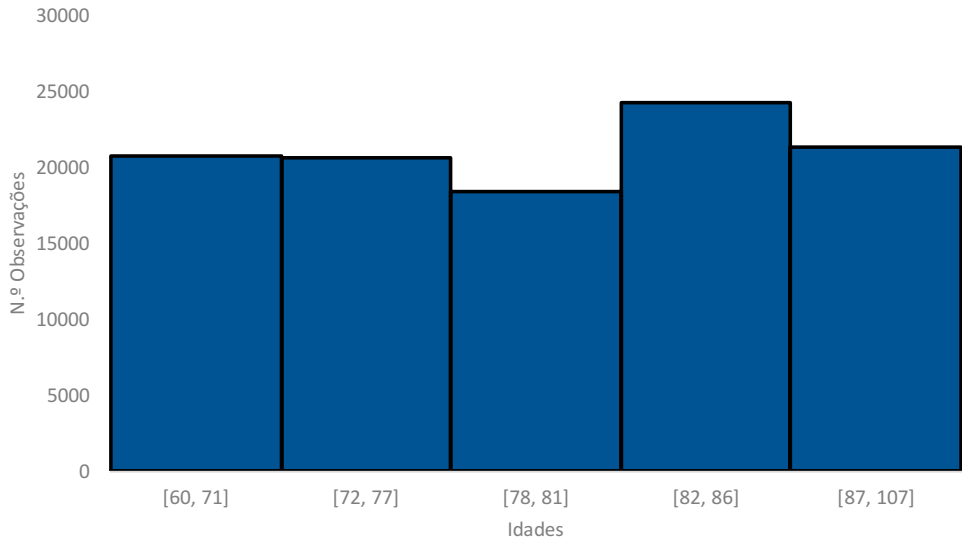


Figura 8.3: Histograma do número de observações por cada conjunto de idades

8.1 Matrizes de Transições de Probabilidades

Segundo o capítulo 3, sendo $S(t)$ a variável que representa o estado em que um indivíduo se encontra à idade t, esta, neste caso, pode assumir um de cinco estados: $S(t) = \{$saudável, dependência fraca, dependência moderada, dependência severa, morte$\}$, tem-se que a matriz de probabilidades de transição associada é dada por:

$$ {}_tP_x = \begin{bmatrix} {}_tp_x^{s\,s} & {}_tp_x^{s\,df} & {}_tp_x^{s\,dm} & {}_tp_x^{s\,ds} & {}_tp_x^{s\,m} \\ {}_tp_x^{df\,s} & {}_tp_x^{df\,df} & {}_tp_x^{df\,dm} & {}_tp_x^{df\,ds} & {}_tp_x^{df\,m} \\ {}_tp_x^{dm\,s} & {}_tp_x^{dm\,df} & {}_tp_x^{dm\,dm} & {}_tp_x^{dm\,ds} & {}_tp_x^{dm\,m} \\ {}_tp_x^{ds\,s} & {}_tp_x^{ds\,df} & {}_tp_x^{ds\,dm} & {}_tp_x^{ds\,ds} & {}_tp_x^{ds\,m} \\ {}_tp_x^{m\,s} & {}_tp_x^{m\,df} & {}_tp_x^{m\,dm} & {}_tp_x^{m\,ds} & {}_tp_x^{m\,m} \end{bmatrix} = \begin{bmatrix} {}_tp_x^{s\,s} & {}_tp_x^{s\,df} & {}_tp_x^{s\,dm} & {}_tp_x^{s\,ds} & {}_tp_x^{s\,m} \\ {}_tp_x^{df\,s} & {}_tp_x^{df\,df} & {}_tp_x^{df\,dm} & {}_tp_x^{df\,ds} & {}_tp_x^{df\,m} \\ {}_tp_x^{dm\,s} & {}_tp_x^{dm\,df} & {}_tp_x^{dm\,dm} & {}_tp_x^{dm\,ds} & {}_tp_x^{dm\,m} \\ {}_tp_x^{ds\,s} & {}_tp_x^{ds\,df} & {}_tp_x^{ds\,dm} & {}_tp_x^{ds\,ds} & {}_tp_x^{ds\,m} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} $$

Sendo esta matriz estocástica, isto é:

$$ \sum_{j \in S} {}_tp_x^{ij} = 1 \quad \forall i,t $$

Uma vez que a matriz de probabilidades é estocástica, a estimativa de cada probabilidade de transição é feita pelo estimador da proporção, isto é:

$$ {}_tp_x^{i\,j} = \frac{N_x^{ij}}{\sum\limits_{j \in S} N_x^{ij}} \quad \forall i,j $$

Sendo N^{ij} o número de transições de estado i para o estado j, calculado no *Software R*. Assim, é possível obter as matrizes de transição para cada conjuntos de idades.

8.1.1 Conjunto de Idades: 60 aos 71 anos

Para as idades dos 60 aos 71 anos obteve-se a seguinte matriz:

$$ {}_tP_{x\in[60\,;\,71]} = \begin{bmatrix} 84.43\% & 11.17\% & 1.44\% & 0.66\% & 2.31\% \\ 18.70\% & 69.45\% & 4.66\% & 3.00\% & 4.19\% \\ 6.98\% & 21.24\% & 47.56\% & 18.14\% & 6.09\% \\ 1.27\% & 6.82\% & 10.66\% & 72.14\% & 9.10\% \\ 0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\% \end{bmatrix} $$

8.1.2 Conjunto de Idades: 72 aos 77 anos

Para os indivíduos com idades entre os 72 e os 77 anos, pode-se observar a matriz seguinte:

$$ {}_tP_{x\in[72\,;\,77]} = \begin{bmatrix} 83.71\% & 12.16\% & 1.29\% & 0.57\% & 1.80\% \\ 16.98\% & 69.42\% & 6.41\% & 3.38\% & 3.81\% \\ 3.81\% & 19.99\% & 54.49\% & 16.46\% & 5.25\% \\ 1.03\% & 6.21\% & 11.47\% & 72.91\% & 8.38\% \\ 0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\% \end{bmatrix} $$

8.1.3 Conjunto de Idades: 78 aos 81 anos

Para o conjunto de indivíduos com idades entre os 78 e 81 anos, tem-se a seguinte matriz:

$$ {}_tP_{x\in[78\,;\,81]} = \begin{bmatrix} 80.02\% & 14.40\% & 2.82\% & 0.84\% & 1.92\% \\ 14.45\% & 69.48\% & 8.23\% & 4.11\% & 3.74\% \\ 3.78\% & 18.06\% & 52.88\% & 20.23\% & 5.05\% \\ 0.64\% & 4.73\% & 11.25\% & 74.25\% & 9.13\% \\ 0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\% \end{bmatrix} $$

8.1.4 Conjunto de Idades: 82 aos 86 anos

De seguida está apresentada a matriz dos utentes cujas idades estão entre os 82 e 86 anos.

$$ {}_tP_{x\in[82\,;\,86]} = \begin{bmatrix} 75.00\% & 17.19\% & 5.30\% & 1.09\% & 1.43\% \\ 12.54\% & 70.48\% & 9.72\% & 4.34\% & 2.92\% \\ 4.03\% & 17.82\% & 54.63\% & 19.08\% & 4.45\% \\ 0.55\% & 4.31\% & 10.98\% & 74.39\% & 9.77\% \\ 0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\% \end{bmatrix} $$

8.1.5 Conjunto de Idades: 87 aos 107 anos

Por fim, os utentes com mais idade, 87 aos 107, têm uma matriz de transição apresentada de seguida.

$$ {}_tP_{x\in[87\,;\,107]} = \begin{bmatrix} 70.53\% & 16.38\% & 8.42\% & 1.82\% & 2.84\% \\ 9.80\% & 68.48\% & 12.14\% & 5.43\% & 4.15\% \\ 2.92\% & 15.78\% & 55.37\% & 20.23\% & 5.70\% \\ 0.35\% & 3.46\% & 9.15\% & 74.49\% & 12.55\% \\ 0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\% \end{bmatrix} $$

8.2 Matrizes com Taxas de Mortalidade

Nas matrizes anteriormente calculadas, é possível observar que as probabilidades de morte são baixas, por isso, após uma análise aprofundada das informações relativas aos óbitos, concluiu-se que, quando um utente morre, nem sempre há o registo. Para resolver esta situação, usou-se uma taxa de mortalidade para assim calcular as transições, que possivelmente, tinham ocorrido para o estado de morte. Esta taxa de mortalidade (representada por q_x), ou seja, a probabilidade de um indivíduo estar vivo a uma certa idade mas no ano seguinte já ter falecido, foi obtida por uma tábua de mortalidade, isto é, uma tabela que apresenta um conjunto de dados demográficos que possibilitam a avaliação da mortalidade numa dada população. Normalmente, estes são apresentados para todas as idades, segundo [25]. Uma vez que os dados do trabalho correspondem à população portuguesa, ao ano de 2015, a tábua usada foi a de Portugal, de 2013-2015, calculada pelo Instituto Nacional de Estatística ([23]). Tendo em conta, que as observações não correspondem todas ao mesmo estado de saúde, a probabilidade de morte não deve ser igual para todos. Por exemplo, uma pessoa num estado de dependência severa tem maior probabilidade de morrer que uma no estado saudável. Assim, foram utilizados coeficientes de agravamento (α) para os estados de dependência de maior grau e tem-se que: $q_x^* = 1 - (1 - q_x)^\alpha$.

Como as matrizes anteriores foram calculadas para um conjunto de idades, as taxas de mortalidade utilizadas são uma combinação de várias taxas, ponderadas pelo número de indivíduos de cada idade. Assim, para a primeira matriz (idades entre 60 e 71), o cálculo das taxas foi feito do seguinte modo:

- Taxas de mortalidade para o estado saudável e de dependência fraca (primeiras 2 linhas da matriz):

$$ q_{x\in[60\,;\,71]} = \frac{\sum_{x=60}^{71} N_x * q_x}{\sum_{x=60}^{71} N_x} $$

  Sendo q_x dada pela tábua de mortalidade e N_x o número de indivíduos da base de dados com x anos.

- Taxa de mortalidade para o estado de dependência moderada (terceira linha da

matriz):

$$q^{*1}_{x\in[60\,;\,71]} = 1 - \left(1 - \frac{\sum_{x=60}^{71} N_x * q_x}{\sum_{x=60}^{71} N_x}\right)^{\alpha_1}$$

Sendo α_1 o coeficiente de agravamento do estado de dependência moderada.

- Por fim, taxa de mortalidade para o estado de dependência severa:

$$q^{*2}_{x\in[60\,;\,71]} = 1 - \left(1 - \frac{\sum_{x=60}^{71} N_x * q_x}{\sum_{x=60}^{71} N_x}\right)^{\alpha_2}$$

Os coeficientes de agravamento, α_1 e α_2, foram calculados com base numa matriz de probabilidades de transição já existente. Esta ($_tP_x^{SCM}$) foi obtida a partir de dados da Santa Casa da Misericórdia de Almada durante os anos 2008 a 2011:

$$_tP_x^{SCM} = \begin{bmatrix} 79.15\% & 9.27\% & 5.79\% & 1.93\% & 3.86\% \\ 6.19\% & 46.90\% & 24.78\% & 3.54\% & 18.58\% \\ 5.58\% & 8.37\% & 46.22\% & 14.34\% & 25.50\% \\ 2.29\% & 3.82\% & 6.87\% & 45.80\% & 41.22\% \\ 0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\% \end{bmatrix}$$

Os coeficientes de agravamento correspondem a proporções das taxas de mortalidade dos diferentes estados em relação ao estado de morte, foram divididas por dois, uma vez que se estão a utilizar dados referentes a utentes muito restritos, apenas à zona de Almada. Por isso, só foi considerada metade da proporção.

$$\alpha_1 = \frac{1}{2}\frac{p_x^{dm\,m\,SCM}}{p_x^{s\,m\,SCM}}$$

$$\alpha_2 = \frac{1}{2}\frac{p_x^{ds\,m\,SCM}}{p_x^{s\,m\,SCM}}$$

Com isto, obtiveram-se os coeficientes de agravamento apresentados na tabela 8.3.

Tabela 8.3: Coeficientes de agravamento

α_1	α_2
3.30	5.34

Seguindo o mesmo raciocínio para as outras idades, considerando os mesmos coeficientes de agravamento, obtiveram-se as taxas de mortalidade apresentadas na tabela 8.4.

Tabela 8.4: Taxas de mortalidade

Idades	q_x	q_x^{*1}	q_x^{*2}
[60 ; 71]	0.01	0.04	0.06
[72 ; 77]	0.03	0.08	0.13
[78 ; 81]	0.04	0.14	0.21
[82 ; 86]	0.09	0.26	0.39
[87 ; 107]	0.23	0.58	0.76

Com isto, têm-se novas matrizes de transição de probabilidades:

$$P_{x\in[60\,;\,71]} = \begin{bmatrix} 83.47\% & 11.05\% & 1.44\% & 0.66\% & 3.38\% \\ 18.49\% & 68.66\% & 4.61\% & 2.97\% & 5.28\% \\ 6.72\% & 20.48\% & 45.85\% & 17.50\% & 9.45\% \\ 1.21\% & 6.44\% & 10.04\% & 67.86\% & 14.45\% \\ 0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\% \end{bmatrix}$$

$$P_{x\in[72\,;\,77]} = \begin{bmatrix} 81.63\% & 11.88\% & 1.71\% & 0.57\% & 4.20\% \\ 16.57\% & 67.68\% & 6.26\% & 3.30\% & 6.20\% \\ 3.53\% & 18.41\% & 50.12\% & 15.16\% & 12.78\% \\ 0.91\% & 5.43\% & 10.01\% & 63.65\% & 20.00\% \\ 0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\% \end{bmatrix}$$

$$P_{x\in[78\,;\,81]} = \begin{bmatrix} 76.60\% & 13.80\% & 2.70\% & 0.84\% & 6.06\% \\ 13.82\% & 66.47\% & 7.87\% & 3.95\% & 7.89\% \\ 3.29\% & 15.63\% & 45.72\% & 17.48\% & 17.88\% \\ 0.52\% & 3.74\% & 8.90\% & 58.60\% & 28.24\% \\ 0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\% \end{bmatrix}$$

$$P_{x\in[82\,;\,86]} = \begin{bmatrix} 68.41\% & 15.69\% & 4.89\% & 1.02\% & 9.99\% \\ 11.43\% & 64.26\% & 8.87\% & 3.97\% & 11.47\% \\ 2.97\% & 13.13\% & 40.26\% & 14.06\% & 29.57\% \\ 0.34\% & 2.64\% & 6.70\% & 45.39\% & 44.93\% \\ 0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\% \end{bmatrix}$$

$$P_{x\in[87\,;\,107]} = \begin{bmatrix} 54.15\% & 12.63\% & 6.48\% & 1.48\% & 25.26\% \\ 7.53\% & 52.54\% & 9.33\% & 4.17\% & 26.43\% \\ 1.22\% & 6.57\% & 23.08\% & 8.45\% & 60.69\% \\ 0.09\% & 0.85\% & 2.23\% & 18.06\% & 78.77\% \\ 0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\% \end{bmatrix}$$

8.3 Observações sobre as Matrizes de Probabilidades

Existem algumas considerações a ter em conta sobre as matrizes anteriores.

À medida que a idade aumenta, as probabilidades do estado saudável e de menor dependência vão diminuindo. Em contrapartida, os estados com maior grau de dependência vão tendo maiores probabilidades.

As probabilidades de retorno para estados de dependência de menor gravidade são baixas e também diminuem com o aumento da idade.

Em cada matriz, as maiores probabilidades são as de permanência no mesmo estado.

Por fim, as probabilidades de morte são maiores nos estado de maior dependência.

Para o cálculo das intensidades de transição foi utilizada uma combinação destas matrizes, ponderada pela percentagem de indivíduos dos vários conjuntos de idades, isto é:

$$
\begin{aligned}
{}_tP_x &= {}_tP_{x\in[60\,;\,71]}\frac{20752}{105394} + {}_tP_{x\in[72\,;\,77]}\frac{20613}{105394} + \\
&+ {}_tP_{x\in[78\,;\,81]}\frac{18424}{105394} + {}_tP_{x\in[82\,;\,86]}\frac{24231}{105394} \\
&+ {}_tP_{x\in[87\,;\,107]}\frac{21374}{105394} \\
&= \begin{bmatrix}
72.50\% & 13.08\% & 3.53\% & 0.92\% & 9.97\% \\
13.45\% & 63.80\% & 7.44\% & 3.68\% & 11.63\% \\
3.52\% & 14.72\% & 40.76\% & 14.41\% & 26.59\% \\
0.60\% & 3.76\% & 7.48\% & 50.15\% & 38.00\% \\
0.00\% & 0.00\% & 0.00\% & 0.00\% & 100.00\%
\end{bmatrix}
\end{aligned}
$$

Note-se que esta matriz não depende de t, por isso, pode considerar-se ${}_tP_x = P_x$

As matrizes ${}_tP_x$ e ${}_tP_x^{SCM}$ são semelhantes, apesar dos valores serem claramente disjuntos apresentam o mesmo tipo de variação tanto por linha como por coluna. Em [10], é possível comparar os resultados obtidos pelas diferentes matrizes.

Foi possível, a partir dos estados dos utentes, obterem-se matrizes de transições de probabilidades. No entanto, estas são discretas e o objetivo é terem-se probabilidades contínuas no tempo, por isso, ainda têm de ser calculadas probabilidades contínuas.

CAPÍTULO 9

Cálculo por Calibração de Intensidades de Transição e Tempos de Permanência

No capítulo anterior foi calculada uma matriz de probabilidades de transição de uma cadeia de *Markov*, não dependente de t. No entanto, esta deve fazer-se variar com o tempo, por isso, a partir das equações diferenciais de *Chapman-Kolmogorov* (secção 3.3) é possível dadas certas intensidades dependentes da idade do indivíduo, obter probabilidades dependentes de t, ou seja, obter uma cadeia de *Markov* a tempo contínuo, tendo-se a seguinte matriz de intensidades:

$$M_x = \begin{bmatrix} \mu_x^{s\,s} & \mu_x^{s\,df} & \mu_x^{s\,dm} & \mu_x^{s\,ds} & \mu_x^{s\,m} \\ \mu_x^{df\,s} & \mu_x^{df\,df} & \mu_x^{df\,dm} & \mu_x^{df\,ds} & \mu_x^{df\,m} \\ \mu_x^{dm\,s} & \mu_x^{dm\,df} & \mu_x^{dm\,dm} & \mu_x^{dm\,ds} & \mu_x^{dm\,m} \\ \mu_x^{ds\,s} & \mu_x^{ds\,df} & \mu_x^{ds\,dm} & \mu_x^{ds\,ds} & \mu_x^{ds\,m} \\ \mu_x^{m\,s} & \mu_x^{m\,df} & \mu_x^{m\,dm} & \mu_x^{m\,ds} & \mu_x^{m\,m} \end{bmatrix}$$

Estas intensidades foram escritas, em função de x e de outros parâmetros, da seguinte forma:

$$\mu_{ij}^x = \gamma_{ij} + 10^{\alpha_{ij}x+\beta_{ij}}, \quad \gamma_{ij}, \quad \alpha_{ij}, \quad \beta_{ij} \quad \in \mathbb{R} \quad i,j \in \{s,\, df,\, dm,\, ds,\, m\}$$

Os parâmetros (γ_{ij}, α_{ij}, β_{ij}), usados nas transições do estado saudável para o estado de dependência fraca e para o estado de morte foram os mesmo utilizados em [12]:

- $(\gamma_{s\,df},\, \alpha_{s\,df},\, \beta_{s\,df}) = (0{,}0004;\, 0{,}06;\, -5{,}46)$;
- $(\gamma_{s\,m},\, \alpha_{s\,m},\, \beta_{s\,m}) = (0{,}0005;\, 0{,}038;\, -4{,}12)$.

Todas as outras intensidades foram calculadas a partir destes, segundo as seguintes expressões de calibração:

1. $\gamma_{ij} = \begin{cases} \gamma_{s\,df} + \dfrac{{}_tp_x^{s\,df} - {}_t\,p_x^{i\,j}}{{}_tp_x^{s\,df} - {}_t\,p_x^{s\,m}}.\rho_\gamma \quad se \quad j \neq m \\ \\ \gamma_{s\,m} + \dfrac{{}_tp_x^{s\,df} - {}_t\,p_x^{i\,m}}{{}_tp_x^{s\,df} - {}_t\,p_x^{s\,m}}.\rho_\gamma \quad se \quad j = m \end{cases}$

2. $\alpha_{ij} = \begin{cases} \alpha_{s\,df} - \dfrac{{}_tp_x^{s\,df} - {}_t\,p_x^{i\,j}}{{}_tp_x^{s\,df} - {}_t\,p_x^{s\,m}}.\rho_\alpha \quad se \quad j \neq m \\ \\ \alpha_{s\,m} - \dfrac{{}_tp_x^{s\,df} - {}_t\,p_x^{i\,m}}{{}_tp_x^{s\,df} - {}_t\,p_x^{s\,m}}.\rho_\alpha \quad se \quad j = m \end{cases}$

3. $\beta_{ij} = \begin{cases} -\left(\beta_{s\,df} - \dfrac{{}_tp_x^{s\,df} - {}_t\,p_x^{i\,j}}{{}_tp_x^{s\,df} - {}_t\,p_x^{s\,m}}.\rho_\beta\right) \quad se \quad j \neq m \\ \\ -\left(\beta_{s\,m} - \dfrac{{}_tp_x^{s\,df} - {}_t\,p_x^{i\,m}}{{}_tp_x^{s\,df} - {}_t\,p_x^{s\,m}}.\rho_\beta\right) \quad se \quad j = m \end{cases}$

Sendo ρ_γ, ρ_α e ρ_β números reais positivos.

De seguida, são apresentadas as equações diferenciais de *Chapman-Kolmogorov*, sendo $i, j = \{s, df, dm, ds, m\}$:

- Condições iniciais:

  $\rightarrow\ {}_0p_x^{ii} = 1,\ \forall i;$

  $\rightarrow\ {}_0p_x^{ij} = 0,$ com $i \neq j.$

- Estado saudável:

  $$\rightarrow\ \frac{d_t p_x^{s\,j}}{dt} = \sum_{k \neq j} {}_tp_x^{s\,k}\,\mu_{x+t}^{kj} - {}_tp_x^{s\,j} \sum_{k \neq j} \mu_{x+t}^{jk} \quad \forall j$$

- Estado dependência fraca:

  $$\rightarrow\ \frac{d_t p_x^{df\,j}}{dt} = \sum_{k \neq j} {}_tp_x^{df\,k}\,\mu_{x+t}^{kj} - {}_tp_x^{df\,j} \sum_{k \neq j} \mu_{x+t}^{jk} \quad \forall j$$

- Estado dependência moderada:

  $$\rightarrow\ \frac{d_t p_x^{dm\,j}}{dt} = \sum_{k \neq j} {}_tp_x^{dm\,k}\,\mu_{x+t}^{kj} - {}_tp_x^{dm\,j} \sum_{k \neq j} \mu_{x+t}^{jk} \quad \forall j$$

- Estado dependência severa:

  $$\rightarrow\ \frac{d_t p_x^{ds\,j}}{dt} = \sum_{k \neq j} {}_tp_x^{ds\,k}\,\mu_{x+t}^{kj} - {}_tp_x^{ds\,j} \sum_{k \neq j} \mu_{x+t}^{jk} \quad \forall j$$

Com a resolução das equações diferenciais de *Chapman-Kolmogorov* foram obtidas novas probabilidades de transição, representadas por ${}_tP_x^{\mu^\rho}$, em todas as resoluções posteriores admitiu-se que $x = 65$.

9.1 Função Perda para as Probabilidades

Considere-se a função perda,

$$L({}_tP_x^{\mu}, P_x) \quad = \quad \sum_{i,j} \sum_{t=1}^{40} \left({}_tp_x^{\mu_{ij}^{\rho}} - \left(p_x^{ij}\right)^t \right)^2 \quad , t \in \mathbb{N}$$

Veja-se [12], tem-se que ${}_nP_x =_1 P_x^n$, se a cadeia for homogénea.

O objetivo é que esta função seja mínima, isto é, que as probabilidades de transição não homogéneas contínuas se aproximem às probabilidades homogéneas discretas, fazendo variar-se $\rho = (\rho_\gamma, \rho_\alpha, \rho_\beta)$, ou seja, pretende-se determinar μ_{ij}^{ρ}, tal que:

$$\min L({}_tP_x^{\mu}, P_x) = \min_{(\rho_\gamma, \rho_\alpha, \rho_\beta)} \sum_{i,j} \sum_{t} \left({}_tp_x^{\mu_{ij}^{\rho}} - \left(p_x^{ij}\right)^t \right)^2 , t \in \mathbb{N}$$

Segundo [9], este problema de optimização tem solução única, sob hipóteses simples. Fazendo variar ρ podem ter-se vários resultados, veja-se a tabela 9.1.

Tabela 9.1: Valores da função perda média por probabilidade

ρ_γ	ρ_α	ρ_β	$\sqrt{\frac{L({}_tP_x^{\mu^\rho}, P_x)}{800}}$
0,0004	0,00004	0,0002	**16,288%**
0,00001	0,01	0,001	**15,252%**
0,0000001	0,004	0,0001	**11,761%**
0,00001	0,002	0,000001	**10,504%**

As percentagens da última coluna da tabela anterior (9.1) correspondem à função perda média por probabilidade. A função perda é dividida por 800 que corresponde ao número de probabilidades calculadas.

O melhor resultado para a função perda é utilizando os valores da última linha da tabela 9.1. Com estes, e resolvendo numericamente as equações de *Chapman-Kolmogorov* com recurso ao *Mathematica*, obtiveram-se as probabilidades de transição representadas pelos gráficos que se seguem.

- Para as probabilidades do estado saudável, têm-se a figura 9.1;

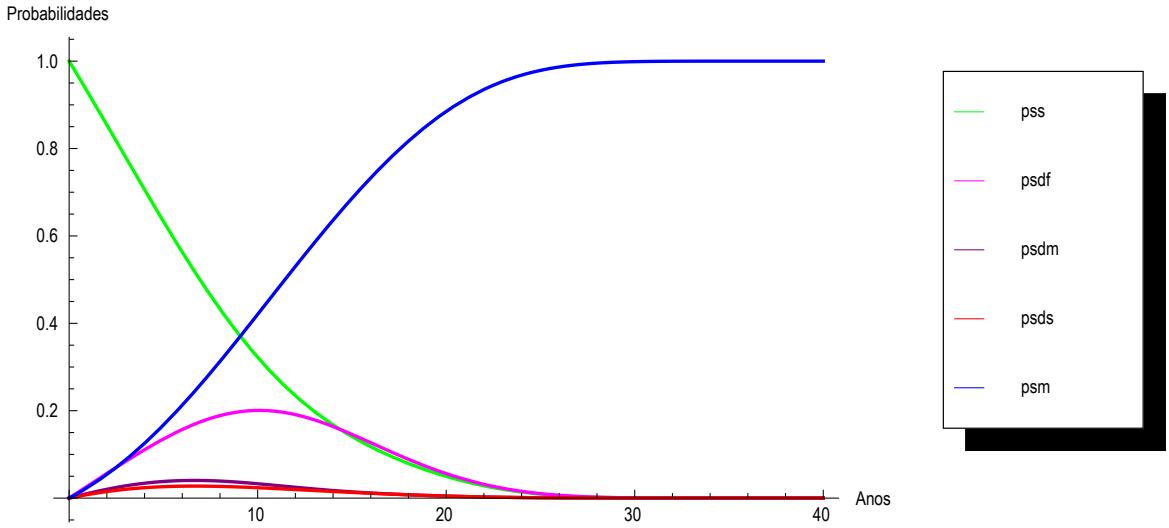


Figura 9.1: Probabilidades de transição a partir do estado saudável

- As probabilidades do estado de dependência fraca são apresentadas na figura 9.2;

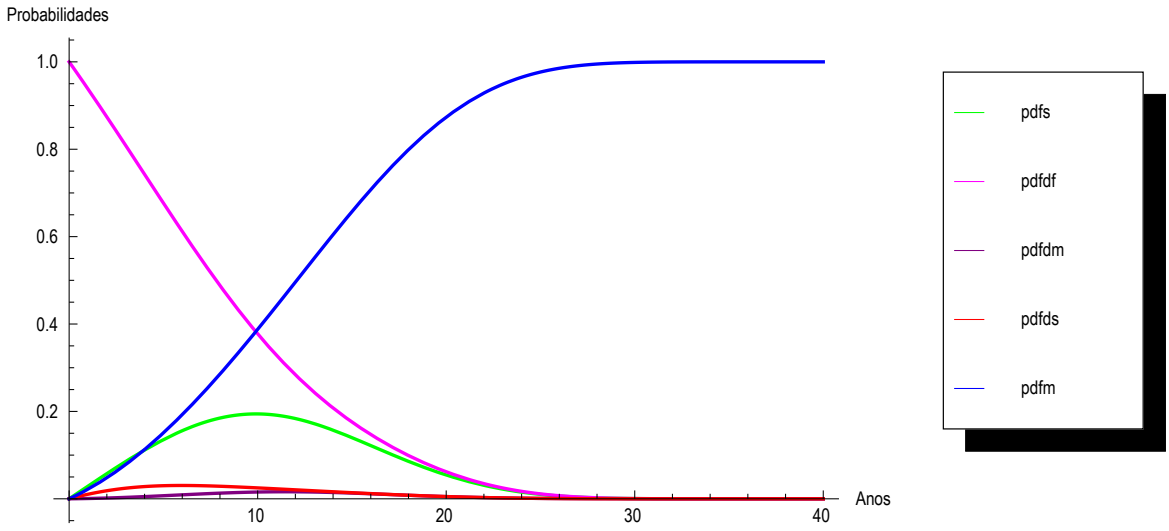


Figura 9.2: Probabilidades de transição a partir do estado de dependência fraca

- Na figura 9.3 têm-se as probabilidades do estado de dependência moderada;

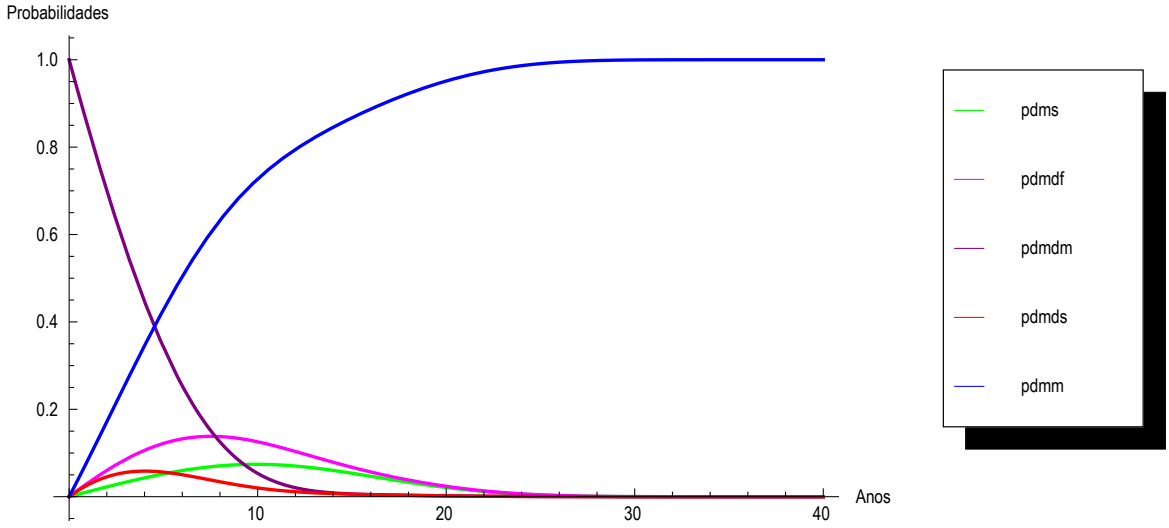


Figura 9.3: Probabilidades de transição a partir do estado de dependência moderada

- Por fim, para as probabilidades do estado de dependência severa têm-se a figura 9.4.

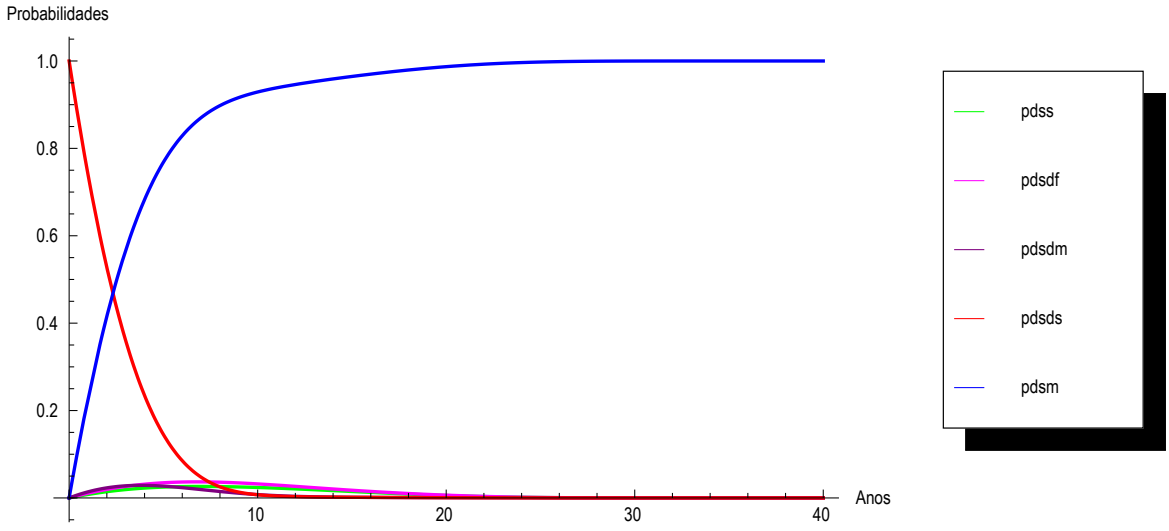


Figura 9.4: Probabilidades de transição a partir do estado de dependência severa

Também foi calculada a distribuição do tempo de permanência nos diversos estados. Seja T_x^{ii} a variável aleatória que representa o tempo de permanência interrupta no estado i, a sua função distribuição é dada por:

$$\begin{aligned} F_{T_x^{ii}}(t) &= \mathbb{P}[T_x^{ii} \leq t] = \mathbb{P}[S(x+t) \neq i | S(x+u) = i, 0 \leq u < t] = 1 - \mathbb{P}[S(x+t) = i | S(x+u) = i, 0 \leq u < t] \\ &= 1 -_t p_x^{ii} = 1 - exp(-\int_0^t \mu_{x+s}^i ds) \end{aligned}$$

Com recurso ao *Mathematica*, obteve-se o seguinte gráfico (9.5), para as distribuições dos tempos de permanência.

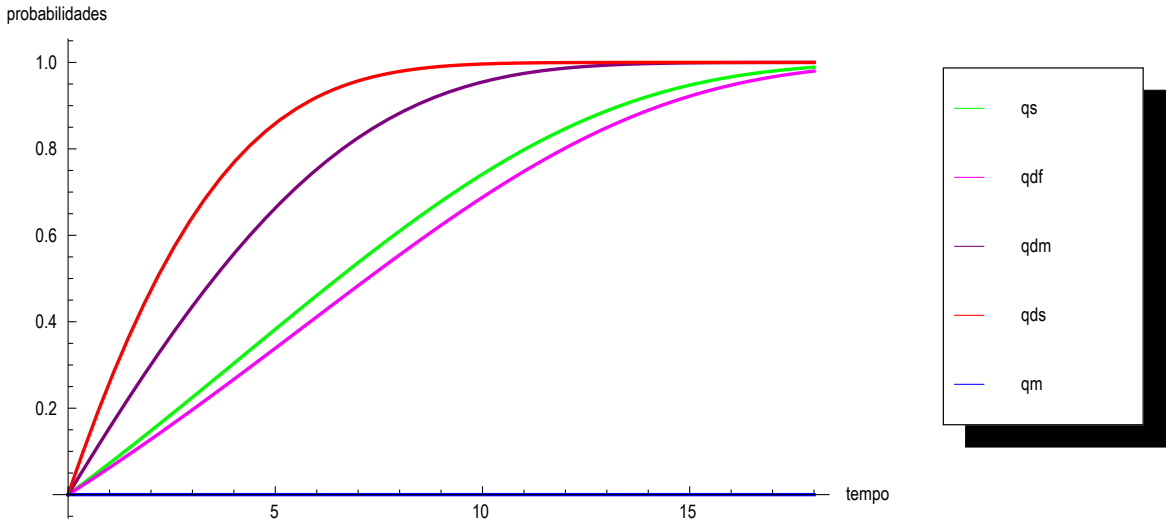


Figura 9.5: Distribuição do tempo de permanência nos diferentes estados

Do gráfico anterior, pode observar-se, por exemplo, que a probabilidade de permanência do estado de dependência fraca é pequena e até inferior ao estado saudável, em contrapartida com a probabilidade do estado de dependência severa que é bastante elevada.

Em suma, conseguiu obter-se probabilidades contínuas no tempo, e assim podem ser simuladas trajetórias com qualquer tempo de permanência nos diferentes estados.

CAPÍTULO 10

Simulação de Custos

A cada estado de dependência está associado um custo já que os indivíduos são dependentes e necessitam de apoio para realizar as suas tarefas. Com o objetivo de garantir recursos para financiar estes custos, tem-se o seguro *Long-Term Care*, secção 2.3.

O valor do prémio deste seguro tem de cobrir, em média, os gastos feitos devido à dependência. Para esse cálculo, fizeram-se diversas simulações dos custos para se obter uma amostra destes, ou seja, simulações dos estados em que o indivíduo pode estar e respetivos tempos de permanência nestes, podendo calcular, assim, os custos que um segurado acarreta por estar dependente.

Considerem-se os seguintes requisitos do seguro que cobrirá os gastos de dependência:

- Os prémios do seguro são pagos mensalmente a partir de uma certa idade, representada por x;
- Os prémios são pagos durante n anos;
- Existe um período de carência de n anos;
- O indivíduo só começa a receber uma mensalidade caso esteja dependente a partir dos $x+n$ anos, o valor dessa mensalidade varia consoante o grau de dependência;
- A taxa de juro é constante ao longo do tempo;
- Admite-se que os custos feitos devido a dependência são constantes ao longo do tempo, ou seja, não são inflacionados; para contrabalançar esta hipótese, os custos não foram atualizados.

Os custos referentes à dependência podem ser equivalentes a um prémio mensal, **P**, dado pela seguinte expressão:

$$\mathbf{P}\sum_{k=1}^{12n}(1+\frac{i}{12})^k = \mathbb{E}[\text{Custos}] \tag{10.1}$$

O pagamento dos prémios corresponde graficamente à seguinte figura (10.1).

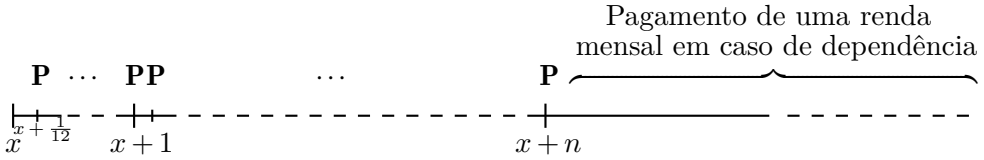


Figura 10.1: Representação gráfica do Seguro de Dependência

Os custos variam conforme o grau de dependência do utente, admitindo que uma hora de assistência domiciliária tem um custo de 10 €, consideraram-se os seguintes custos que correspondem ao valor mensal a receber, caso o indivíduo se encontre dependente:

- Se um segurado estiver no grau de dependência fraca irá receber uma renda mensal de 500 €, para garantir:

  1. Cuidados ao domicilio, em média, três vezes por semana, quatro horas por dia, perfazendo um custo total médio mensal de 480 €;
  2. Todos os medicamentos a necessitar, valor médio mensal de 20 €.

- Se um segurado for dependente moderado irá receber uma mensalidade de 1 500 €, para garantir:

  1. Cuidados ao domicilio, em média, seis vezes por semana, seis horas por dia, fazendo um custo total médio mensal de 1 440 €;
  2. Todos os medicamentos a necessitar, valor médio mensal de 60 €.

- Por fim, se o segurado estiver no grau de dependência severa irá receber uma renda mensal de 3 000 €, para garantir:

  1. O pagamento da estadia e todos os recursos que necessitar num estabelecimento próprio de cuidados (Hospital, Lar), o que perfaz um custo, médio, por mês de 3 000 €.

Para obter o custo total de um segurado procedeu-se, sequencialmente, a três simulações, tal como detalhado em [17]:

1. A primeira tem o objetivo de simular o estado em que o indivíduo se encontra quando começa a ser segurado, se estiver vivo (ou seja, à idade $x+n$), a partir da função de

probabilidade de $S^*(x+n)$, isto é,

$$\begin{aligned}
\mathbb{P}[S(x+n)=j|S(x+n)\neq m] &= \frac{\mathbb{P}[S(x+n)=j\cap S(x+n)\neq m]}{\mathbb{P}[S(x+n)\neq m]}, \quad j=s,\, df,\, dm,\, ds \\
&= \frac{\mathbb{P}[S(x+n)=j\cap S(x+n)\neq m]}{\mathbb{P}[S(x+n)\neq m]} = \\
&= \frac{\frac{N^j_{x+n}}{N_{x+n}}}{1-\frac{N^m_{x+n}}{N_{x+n}}} = \\
&= \frac{N^j_{x+n}}{\sum\limits_j N^k_{x+n}} = f(S^*(x+n)) \quad ,j=s,\, df,\, dm,\, ds
\end{aligned}$$

Tendo-se para cada estado: $f(S^*(x+n)) = \begin{cases} \frac{N^s_{x+n}}{N^*_{x+n}}, & \text{se} \quad S(x+n)=s \\ \frac{N^{df}_{x+n}}{N^*_{x+n}}, & \text{se} \quad S(x+n)=df \\ \frac{N^{dm}_{x+n}}{N^*_{x+n}}, & \text{se} \quad S(x+n)=dm \\ \frac{N^{ds}_{x+n}}{N^*_{x+n}}, & \text{se} \quad S(x+n)=ds \end{cases}$

Sendo, neste caso, $N^*_{x+n}=\sum\limits_k N^j_{x+n}$ com $j=s,\, df,\, dm,\, ds$, ou seja, o número de todos os utentes que estejam vivos à idade $x+n$.

2. De seguida, é simulado o tempo que o utente está em cada estado, com a função de distribuição das variáveis aleatórias $F_{T^{ii}_x}(t) \quad \forall i$, dadas pelo gráfico da figura 9.5.

3. Por fim, é simulado para qual estado irá o utente ao fim do tempo obtido no ponto anterior. Estes estados são obtidos utilizando as probabilidades da resolução numérica das equações diferenciais de *Chapman-Kolmogorov* (figuras 9.1, 9.2, 9.3 e 9.4).

Veja-se o seguinte exemplo de uma simulação de custos para uma pessoa. Considere-se:

- $x=30$;
- $n=35$
- $i=3\%$

Para a simulação do estado inicial do indivíduo, é necessário ter-se a distribuição da

variável $f(S^*(65))$:

$$f(S^*(65)) = \begin{cases} \frac{366}{1487}, & \text{se} \quad S(x+n) = s \\ \frac{655}{1487}, & \text{se} \quad S(x+n) = df \\ \frac{145}{1487}, & \text{se} \quad S(x+n) = dm \\ \frac{321}{1487}, & \text{se} \quad S(x+n) = ds \end{cases}$$

Assim, simulando com o *Mathematica* tem-se que:

1. O estado inicial é o estado saudável, de seguida foi simulado o tempo de permanência neste:
    - $T_{65}^{s\,s} = 10.6864$, ou seja, o indivíduo desde os 65 anos está durante, aproximadamente, 10 anos e 251 dias no estado saudável;
2. Ao fim, desse tempo, é simulado o novo estado do indivíduo, neste caso, obteve-se dependente moderado. É necessário calcular o tempo que o indivíduo permanece neste:
    - Simulando, o indivíduo está durante 5 anos e 26 dias neste estado;
3. Terminado esse tempo, simulando, o indivíduo irá para o estado de dependência severa:
    - Permanecendo nesse estado durante 6 anos e 298 dias;
4. Por fim, a simulação terminou, o que significa que o indivíduo foi para o último estado (morte).

Assim, desde os 65 anos até morrer, passaram 22 anos e 209 dias, ou seja, o utente morreu aos 87 anos. Este esteve aproximadamente 11 anos no estado saudável, e por isso, a seguradora não tem qualquer custo neste período. No entanto, esteve cerca de 5 anos no estado de dependência moderada, o que corresponde a um custo mensal de 1 500 €e assim um custo total de 91 264.19 €deste estado, nos últimos 6 anos. Uma vez que esteve no estado de dependência severa, acarreta um custo mensal de 3 000 €, ou seja, um custo total de 245 362.50 €. Em suma, tem um custo total de 336 626.70 €.

Fazendo as simulações anteriores, mas para um maior número de indivíduos, é possível fazer uma amostra significativa de custos. Na tabela 10.1, têm-se medidas descritivas para os custos, variando o tamanho da amostra.

Tabela 10.1: Medidas descritivas de custos para diferentes tamanhos de amostra em €

Tamanho da amostra	100	1 000	10 000	100 000
Mínimo	0.00	0.00	0.00	0.00
1ºQuartil	58 687.45	53421.88	50 992.85	53 616.25
Mediana	116 895.68	119 823.42	116 813.20	118 230.73
Média	**155 790.41**	**168 525.44**	**163 077.31**	**165 767.23**
3ª Quartil	202 553.16	119 823.42	238 228.48	240 892.36
Máximo	662 986.65	864 837.43	1 041 349.88	1 234 218.43
Desvio Padrão	143 414.64	159 106.95	155 479.34	157 396.78

Veja-se que os custos referentes à dependência, em média, por um utente variam entre os 150 000 €e os 175 000 €, no entanto, o desvio padrão é muito elevado, o que significa que pode haver uma grande diferença de custos entre diferentes indivíduos. Também é possível ter-se um histograma de custos, figura 10.2.

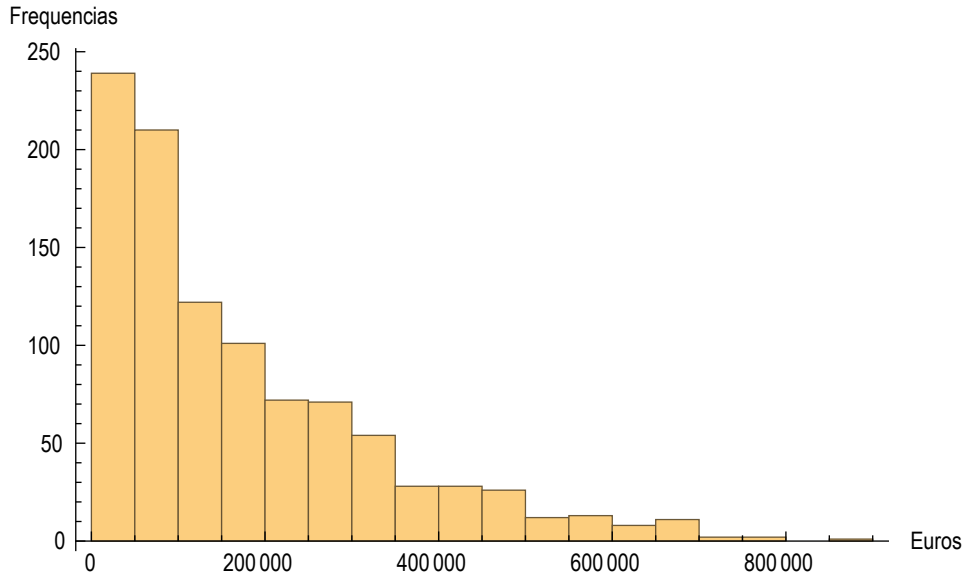


Figura 10.2: Histograma de custos por segurado

Calculados os custos, é possível obter o prémio do seguro, a partir da equação 10, tendo-se os seguintes valores, tabela 10.2.

Tabela 10.2: Prémios com diferentes tamanhos de amostras

Tamanho da amostra	100	1 000	10 000	100 000
Prémio	**209.28 €**	**226.39 €**	**219.07 €**	**222.68 €**

Veja-se que não existem grandes variações de prémios, cerca de 15 €, a partir de uma amostra de tamanho 1000 os resultados obtidos começam a ser estáveis. Assim, sempre que se fizeram simulações, fez-se com uma amostra de tamanho 1 000. No entanto, para uma melhor precisão do prémio, foram feitas 20 simulações diferentes para se terem diferentes

prémios. Obtendo-se as seguintes medidas descritivas desta amostra de prémios, tabela 10.3:

Tabela 10.3: Medidas descritivas para uma amostra de simulação de Prémios

Prémios	
Mínimo	205.08 €
1ºQuartil	216.56 €
Média	**220.81 €**
Mediana	223.03 €
3ª Quartil	226.53 €
Máximo	231.91 €
Desvio Padrão	7.32 €

Conclui-se com a tabela anterior que um prémio mensal, a cobrar a um indivíduo com 30 anos que queira ser segurado daqui a 35 anos, pode ser de 220.81 €, uma vez que é a média de uma amostra de prémios.

Por fim, também é possível obter uma amostra dos totais dos tempos de permanência, isto é, do tempo até o indivíduo chegar ao último estado (morte), ou seja, a esperança de vida, neste caso, aos 65 anos, uma vez que todos os valores calculados são com essa idade. Tendo-se as medidas descritivas na tabela 10.4.

Tabela 10.4: Medidas descritivas para os tempos de permanência

Tamanho da amostra	**100**	**1 000**	**10 000**	**100 000**
Mínimo	0.78	1.03	0.21	0.07
1ºQuartil	7.88	9.55	9.49	9.55
Mediana	13.57	15.72	15.16	15.30
Média	**17.90**	**19.09**	**19.22**	**19.32**
3º Quartil	26.08	26.34	26.48	26.53
Máximo	52.03	57.25	62.93	64.28
Desvio Padrão	13.28	12.46	12.97	12.95

Assim, foram feitas simulações a partir das probabilidades de transição calculadas com a base de dados fornecida, obtendo-se prémios para um seguro de dependência, mas também o tempo de permanência total nos estados, ou seja, o tempo que os indivíduos estão vivos.

CAPÍTULO 11

Tabelas de Prémios e Análises de Resultados

Neste capítulo, será feita uma análise de alguns resultados obtidos.

Note-se na média dos tempos totais de permanência, na tabela seguinte (11.1), que corresponde ao valor médio da tabela 10.4.

Tabela 11.1: Média da soma os tempos totais de permanência

Tamanho da amostra	**100**	**1 000**	**10 000**	**100 000**
$\sum \bar{T}^x_{ii}$	17.90	19.09	19.22	19.32

Em Portugal, segundo o Instituto Nacional de Estatística, em $2014-2016$, a esperança média de vida aos 65 anos, ou seja, a idade à qual foram calculados os prémios anteriores e os tempos totais de permanência, é de 19.31 (2.2), o que corresponde, por diferenças de milésimas aos valores obtidos na tabela anterior (11.1).

Note-se que existe uma estabilidade nos valores, são muito semelhantes, apenas variam nas décimas, pelo que se justifica a dimensão do estudo de simulação com 1000 trajetórias, tal como se referiu anteriormente.

11.1 Tabelas de Prémios

Todos os resultados obtidos até então, presumiam que o indivíduo começava a pagar os prémios mensalmente, a partir dos 30 anos, e um período de carência de 35 anos. Adicionalmente, foram calculados prémios fazendo-se variar estas idades e, assim, construiu-se uma tabela de prémios, tabela 11.2.

Note-se que, na última linha da tabela 11.2 que corresponde à média dos tempos totais de permanência nas diversas idades, estes podem ser comparados aos da esperança média

Tabela 11.2: Tabela de prémios

		$x+n$				
		60	**65**	**70**	**75**	**80**
x	**30**	344.01 €	226.39 €	134.33 €	74.44 €	44.63 €
	35	449.23 €	287.98 €	167.70 €	91.64 €	54.37 €
	40	609.81 €	376.07 €	213.33 €	114.41 €	66.93 €
	45	880.86 €	510.49 €	278.59 €	145.54 €	83.57 €
	50	1 426.82 €	737.40 €	378.17 €	190.06 €	106.30 €
	55	3 059.06 €	1 194.45 €	546.26 €	258.00 €	138.82 €
	60	201 311.36 €	2 560.85 €	884.83 €	372.68 €	188.44 €
$\sum \bar{T}_{ii}^{x}$		25.22	19.09	13.89	9.21	5.69

de vida a essas idades em Portugal (tabela 11.3). Veja-se que não são muito diferentes. A maior diferença, que corresponde à idade de 70 anos, não é superior a 30%, isto pode acontecer, devido ao facto da calibração ter sido feita para uma matriz média especifica e para uma idade de 65 anos. Para ultrapassar esta limitação o algoritmo de simulação poderá ser reformulado usando-se, em cada trajetória, as probabilidades de transição calculadas a partir da calibração das intensidades em cada classe de idades.

Tabela 11.3: Comparações da esperança média de vida (EMV) a diferentes idades, em Portugal, 2013-2015 (Fonte: INE) com média dos tempos totais de permanência

Idade	EMV	$\sum \bar{T}_{ii}^{x}$	$\Delta = \sum \bar{T}_{ii}^{x} - EMV$	$\frac{\Delta}{EMV}$
60	25.22	23.37	1.85	7.92%
65	19.09	19.19	-0.1	-0.52%
70	13.89	15.18	-1.29	-8.50%
75	9.21	11.4	-2.19	-19.21%
80	5.69	7.95	-2.26	-28.43%

Com o processo de simulação do presente trabalho, obtiveram-se prémios e tempos totais de permanência. Por uma lado, comparando estes tempos com valores da população portuguesa, já calculados pelo INE, pode notar-se que são semelhantes, isto é, os valores dos tempos simulados não fogem muito à realidade, ou seja, é possível concluir que os resultados obtidos na simulação não são desadequados, tanto para os tempos, bem como para os custos e respetivos prémios.

CAPÍTULO 12

Conclusão

Portugal tem sofrido um envelhecimento da população e, por consequência, um aumento do número de pessoas dependentes. Por isto, torna-se de extrema importância a existência de um seguro de dependência no mercado português. Assim, o objetivo deste trabalho foi calcular um prémio para um seguro desta natureza.

Para uma avaliação do risco de dependência da população portuguesa, foi ajustado ao seguro de *Long-term Care* um modelo de estados múltiplos e, com isto, foram feitas simulações para o cálculo do prémio.

Os resultados obtidos são esclarecedores pois permitem obter, por simulação, as distribuições da duração de vida até à morte e os correspondentes custos. Estas distribuições podem ser utilizadas para verificações, a *posteriori*, da qualidade do ajustamento e para outras formas de cálculo dos prémios.

Em Portugal, este seguro ainda não foi comercializado, porque há pouca informação consistente sobre a população, por isso, as seguradoras não têm a possibilidade de calcular os prémios.

Existem alguns aspetos que devem ser considerados em trabalhos futuros, um deles é o estudo da influência de taxas de juro variáveis no valor dos prémios, por exemplo, através de um modelo que analise a variação temporal destas.

Concluindo, este seguro é muito importante, pois as pessoas mais idosas têm menos capacidades monetárias para custos, cada vez mais caros, referentes à saúde, devendo assim pagar quando podem, ou seja, durante a sua vida ativa.

Bibliografia

[1] Instituto Nacional de Estatística.

[2] R. Carrujo. "Seguro de Dependência - Proposta de um modelo de avaliação financeiro-actuarial". Tese de mestrado. Faculdade de Ciências e Tecnologia, 2008.

[3] I. M. F. Cordeiro. "A multiple state model for the analysis of permanent health insurance claims by cause of disability". Em: *Insurance Mathematics e Economics* 30 (2002), pp. 167 –186.

[4] U. de Missão para os Cuidados Continuados Integrados. *Glossário Rede Nacional de Cuidados Continuados Integrados*. 2009.

[5] U. de Missão para os Cuidados Continuados Integrados. *Rede Nacional de Cuidados Continuados Integrados - Manual do Prestador - Recomendações para a Melhoria Contínua*. 2011.

[6] *Decreto Lei nº 101/2006, I Série – A,* Diário da Rebública, 6 de Junho de 2006.

[7] *Decreto-Lei n.º 136/2015, I Série – A,* Diário da Rebública, 28 de julho de 2015.

[8] E. S. Delgado e J. Castelblanque. "El seguro de Dependencia (II) Experiencia internacional y reaseguro". Em: *trébol* 34 (2005), pp. 9–14.

[9] M. L. Esquível, M. C. de Oliveira, G. R. Guerreiro e C. Nobre. *A Five State Non-Homogeneous Continous Time Markov Chain Model for Long Term Care: Calibration, Simulation and Premium Computations*. 2016.

[10] M. L. Esquível, M. C. de Oliveira, G. R. Guerreiro e C. Nobre. "Calibration and Simulation of a Continous Time Markov Chain Model for Long Term Care". Em: 2^{nd} *International Conference on Computacional Finance*. 2017, pp. 136–141.

[11] M. L. Esquível, M. C. de Oliveira, G. R. Guerreiro, S. Nascimento e H. Lopes. "Estimation of Markov Transition Probabilities via Clustering". Symposium on Big Data in Finance, Retail and Commerce: Statistical and Computational Challenges. 2017.

[12] S. Haberman e E. Pitacco. *Actuarial Models for Disability Insurance*. Taylor & Francis, 1998.

[13] J. Han, J. Pei e M. Kamber. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011.

[14] T. Hastie, R. Tibshirani e J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York, 2013.

[15] http://www.populationpyramid.net/.

[16] L. Kaufman e P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley, 2009.

[17] M. C. de Oliveira. "Determinação de Prémios de um Produto de Long Term Care por Simulação de Cadeias de Markov não Homogéneas a Tempo Contínuo". Seminário em Atuariado, Estatística e Investigação Operacional. 2016.

[18] E. Pitacco. "Actuarial models for pricing disability benefits: Towards a unifying approach". Em: *Insurance Mathematics and Economics* 16 (1995), pp. 39–62.

[19] *Portaria n.º 184/2015 Anexo, 1.ª série - N.º 120*. Diário da Rebública, 23 de junho de 2015.

[20] R. Portuguesa. *Plano de Desenvolvimento da RNCCI 2016-2019*. 2016.

[21] A. D. Santos. "Modelação atuarial de um Seguro Long Term Care". Tese de mestrado. Faculdade de Ciências e Tecnologia, 2016.

[22] I. da Segurança Social. *Guia Prático - Rede Nacional de Cuidados Continuados Integrados*. 2017.

[23] *Tábuas de Mortalidade para Portugal*. Rel. téc. Instituto Nacional de Estatística, 2016.

[24] H. R. Waters, M.A., D. Phil. e F.I.A. *An Approach to the study of Multiple State Models*. 1984.

[25] H. Waters, D. Dickson e M. Hardy. *Actuarial Mathematics for Life Contingent Risks*. International Series on Actuarial Science. Cambridge University Press, 2013.