Pedro André Arraia Gomes

Licenciado em Matemática

Fiabilidade da Imputação de valores omissos através de métodos dedutivos

Relatório de Estágio Profissional no Instituto Nacional de Estatística
para Obtenção do Grau de Mestre em
Matemática e Aplicações
Ramo Atuariado Estatística e Investigação Operacional

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

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

março, 2019

Fiabilidade da Imputação de valores omissos através de métodos dedutivos

Agradecimentos

Ao meu Orientador, Professor Doutor Manuel Leote Esquível, agradeço a motivação e a partilha de conhecimento.

Aos colegas de trabalho do Instituto Nacional de Estatística, que me receberam com todo o carinho e disponibilidade, em particular ao pessoal do Gabinete Censos, e em especial à Professora Doutora Sandra Lagarto, agradeço a dedicação no acompanhamento deste projeto, sem a sua intervenção seria impossível concluí-lo.

Agradeço à minha família por estarem presentes e pelas palavras de apoio, com um agradecimento especial para a minha parceira Vânia Furtado pela paciência e disponibilidade nesta fase importante da minha vida.

Aproveito para agradecer a todos os colegas e professores da Universidade Nova de Lisboa, que me acompanharam ao longo destes anos em torno da Matemática. Foi uma jornada emocionante que me permitiu desenvolver capacidades críticas para o meu futuro. Entre estas pessoas saliento os meus ex colegas de curso Ângelo Martins, André Ferreira, Susana Neves e Vanessa Silva bem como os colegas de faculdade Daniel Franco, José Esteves e Marco Silva.

Por fim um agradecimento aos meus amigos Artur Vasconcelos, Ruben Peixoto e Wilson Ferreira pela força e motivação, como a todos os outros que se preocupam comigo, desejando o melhor para o meu futuro.

Resumo

Neste relatório encontra-se descrita uma metodologia desenvolvida no projeto de estágio profissional, integrado no curso de formação específico para ingresso na carreira de Técnico Superior Especialista em Estatística, realizado no Instituto Nacional de Estatística. Esta metodologia é aplicada a dados provenientes de registos administrativos integrados na Base de População Residente (*BPR*), projeto em desenvolvimento no Gabinete Censos. Este projeto enquadra-se no plano de estudos do Mestrado em Matemática e Aplicações, para obter o grau de mestre e é orientado pelo Professor Doutor Manuel Esquível da Universidade Nova de Lisboa.

A metodologia tem como objetivo avaliar a qualidade dos registos da *BPR*, inicialmente com dados omissos, aos quais foram imputados valores recorrendo a métodos dedutivos. A necessidade desta avaliação prende-se com a existência de variáveis que não estão preenchidas a 100%. Estas derivam de dados administrativos[1], provenientes da Administração Pública. Existe informação omissa sempre que esta não é recolhida para o ano de referência. Para estes registos são usados métodos de imputação[2] dedutivos, através da informação disponível para o mesmo indivíduo, relativa a anos anteriores.

Foram estudadas duas variáveis, que caracterizam o estado civil e o nível de ensino completo de um indivíduo, com taxa de preenchimento de 99,5% e de 27,5%, respetivamente. Utilizam-se os dados de 2011 e 2015 para imputar a cada registo omisso em 2016, o último valor observado.

Estima-se a probabilidade dos valores imputados permanecerem iguais aos do ano de 2016, através da Lei dos Grandes Números (*LGN*) e obtêm-se as matrizes de transição a um e cinco anos. Não existindo informação para calcular estas probabilidades para outros períodos de tempo assumiu-se que os dados destas duas variáveis seguem cadeias de Markov[3] a tempo discreto com passos anuais, estimando-se as matrizes de transição a cinco anos aplicando as propriedades destas cadeias. Com o objetivo de verificar se os pressupostos assumidos são verificados são comparadas as matrizes a 5 passos para ambas as variáveis em estudo.

Os resultados relativos à variável *estado civil* apoiam a hipótese de que os dados cumprem os pressupostos e que aplicar as propriedades referidas traduz uma boa aproximação ao estimado pela *LGN*. Já para a variável *nível de ensino completo* os resultados apontam para a necessidade de aprofundar o estudo. Para incorporar os resultados na *BPR*, como prova do conceito, foi construída uma tabela que categoriza os dados nos *clusters* obtidos para posterior cruzamento com a tabela de probabilidades de permanência[4] calculada. Esta tabela considera todas as combinações possíveis de estados entre as duas variáveis em estudo, permitindo atribuir a probabilidade dos atuais estados serem os mesmos que os últimos observados. Os procedimentos efetuados serão detalhados neste relatório bem como uma análise crítica ao trabalho desenvolvido.

Palavras-Chave: dados administrativos; imputação; Markov; permanência

Abstract

This report describes a methodology developed within the professional internship, integrated in the study plan of the masters in Maths and Applications, at the INE (Instituto Nacional de Estatística) headquarters in Lisboa. It is applied to administrative data[1] from BPR, a project under development at the Census Unit.

The objective of this methodology is to measure the quality of the records from the BPR that hold missing data, which have been targeted for imputation[2] by deductive methods. This measuring is a necessity due to the existence of variables that are incomplete. These missing values were treated using deductive imputation[2] methods that use previous information, when they match the same individual. Two variables were studied, the civil state and the academic level of an individual, displaying 99,5% and 27,5% fill rate respectively.

To obtain the probabilities of imputed data matching with the real data in the year 2016, the Law of Large Numbers (*LLN*) was used, as well as the calculation of the one and the five steps transition matrices. Without information to obtain the matrices at n steps, it was assumed that the data of the two variables follows Discrete-Time Markov[3] Chains with annual steps. Both one and five steps transition matrices of probabilities were compared to find out if the assumptions made could be verified, in order to calculate the permanence[4] probabilities.

The civil state results support the use of Markov[3] Proprieties and that using them provide a good approximation to the *LLN* estimation. On the other hand, the results of the academic level point to a need of further study. To incorporate the estimated probabilities in the BPR, as a proof of concept, a table was built which categorizes the date in clusters so it can be merged with the calculated probabilities table. This last table considers all the combinations between the last states seen for each variable, allowing the association of the probability of these last seen states to remain the same with the passing years. The procedures are detailed in this report as well as the critical analysis.

Keywords: administrative data; imputation;Markov ; permanence

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Enquadramento

Um dos problemas usuais encontrados na análise de dados consiste na existência de valores omissos nos dados. Existem metodologias que colmatam este problema como a construção de variáveis derivadas[1] ou a aplicação de métodos de imputação. Quando os dados são observados ao longo do tempo para os mesmos indivíduos é possível utilizar métodos dedutivos de imputação. Estes apoiam-se em informação prévia relativa à mesma unidade estatística (no caso de estudo ao mesmo indivíduo) e atribuem aos dados omissos o último valor observado desde que válido.

Nos casos em que os valores observados são passíveis de alteração é importante avaliar a qualidade da informação imputada ou qual a probabilidade de continuar atualizada sabendo quanto tempo passou desde a última observação. Para estimar estas probabilidades, para populações na ordem dos milhões de indivíduos, pode considerar-se a LGN para obter uma aproximação ao seu valor esperado. Caso não exista informação para o período pretendido é possível assumir que os dados seguem uma cadeia de Markov e extrapolar as probabilidades a n passos a partir da potência das matrizes de transição a 1 passo.

O problema exposto constitui a principal motivação do trabalho que se apresenta de seguida. A metodologia desenvolvida aplica-se a uma base de dados da população residente em Portugal (BPR) construída pelo Instituto Nacional de Estatística (INE).

Ao longo do desenvolvimento do programa de ação do Gabinete Censos foram perspetivadas diferentes opções metodológicas na transformação do modelo censitário e devidamente ponderadas as vantagens, os riscos e as condições prevalecentes na introdução de um modelo baseado em dados administrativos. O objetivo de repensar o modelo censitário prende-se com a necessidade de melhorar a eficiência do processo (reduzir custos e a sobrecarga sobre o respondente) bem como de divulgar estatísticas da população anuais, conforme legislação em preparação pelo EUROSTAT, a ser aplicada aos Estados-Membros.

É no âmbito do projeto *BPR* que foi estudada a aplicação da metodologia exploratória apresentada neste relatório. Para o efeito foram analisadas em particular e com o objetivo de melhorar a qualidade dos dados, as variáveis categóricas que caracterizam o estado civil e o nível de ensino completo de um indivíduo.

[1] Variável obtida a partir de outras variáveis através da transformação lógica, matemática ou de outro tipo [10]

1.2 Organização do relatório

Este relatório é constituído por 6 capítulos. No capítulo 2 é descrito o projeto em desenvolvimento no INE, base de todo o trabalho, aprofundando-se o conceito de fonte administrativa e apresentando-se os métodos de imputação mais utilizados.

Para a obtenção dos resultados foram ajustadas cadeias de Markov a tempo discreto a duas bases de dados distintas, uma referente ao estado civil e a outra ao nível de ensino completo. Com o objetivo de construir matrizes de transição mais específicas efetuou-se uma análise de *clusters*. Assim, nos capítulos 3 e 4, é feita uma breve abordagem a estes temas resumindo-se as definições e conceitos, os métodos usados e os dois casos em estudo como modelos de estados múltiplos.

No capítulo 5 são apresentados os resultados da exploração da base de dados extraída da *BPR*, qual o tratamento que sofreu e os pressupostos assumidos. Depois de construídos os dois universos, um para cada variável em estudo, são calculadas as matrizes de transição a partir dos dados tratados. Neste capítulo é estudada a aplicação de propriedades das cadeias de Markov a cada variável, comparando as matrizes a 5 passos estimadas com recurso a métodos diferentes. Apresentam-se ainda os *clusters* obtidos que permitem a construções de matrizes de transição para a população portuguesa, que reduzem a generalização da população como um só grupo. Finaliza-se com a apresentação da tabela de probabilidades de permanência, já preparada para ser incorporada na metodologia de construção da *BPR*.

Por fim, no capítulo 6, são apresentadas as conclusões deste estudo, assim como as perspectivas de trabalho futuro. Em anexo apresentam-se as matrizes de transição calculadas, por *cluster*, para cada base de dados e o código utilizado nos 3 *softwares*: *R*, *Stata* e *Wolfram Mathematica*.

Capítulo 2

Base de População Residente em Portugal

Pela Lei do SEN, em [1], dados administrativos são dados recolhidos por entidades do sector público sobre pessoas singulares ou colectivas, incluindo os dados individuais, com base em procedimentos administrativos que têm normalmente um fim primário que não é estatístico.

A base de população residente em Portugal é uma base de dados, construída no INE, que resulta da utilização de dados administrativos. A substituição da recolha via inquérito por informação administrativa foi a principal linha condutora do projeto, ao longo do qual se investigaram aprofundadamente 12 ficheiros, oriundos de diferentes serviços da administração pública, nomeadamente: Autoridade Tributária (AT); Instituto da Segurança Social (ISS); Instituto dos Registos e Notariado (BDIC); Direção Geral da Educação (EDUC); Instituto do Emprego e Formação Profissional (IEFP); Quadros de Pessoal (QP); Caixa Geral de Aposentações (CGA); Direção Geral de Saúde (SAUDE) e Serviço de Estrangeiros e Fronteiras (SEF). Os resultados da investigação permitiram construir, pela primeira vez, uma Base de População Residente em Portugal (*BPR*), a partir de informação de carácter administrativo. Em 2013, o INE iniciou o levantamento de requisitos para que esta base de dados fosse construída anualmente, identificando como fases do processo as seguintes:

- Criação das condições legais adequadas para acesso aos dados administrativos;
- Análise das variáveis e fontes administrativas de interesse censitário;
- Carregamento, limpeza e harmonização dos dados;
- Aplicação de técnicas de *record-linkage* com os diferentes ficheiros administrativos;
- Aplicação de regras de indícios de residência;
- Construção de variáveis socioeconómicas derivadas de variáveis administrativas;
- Imputação de valores omissos.

A fase de interligação dos ficheiros é uma etapa fundamental no processo de construção da BPR. Após esta fase e da sinalização da existência dos registos nas diferentes fontes administrativas, são aplicadas regras que traduzem os indícios de residência, fazendo convergir os ficheiros da BDIC e do SEF para a *BPR*.

A figura seguinte ilustra as principais etapas para a construção da BPR.

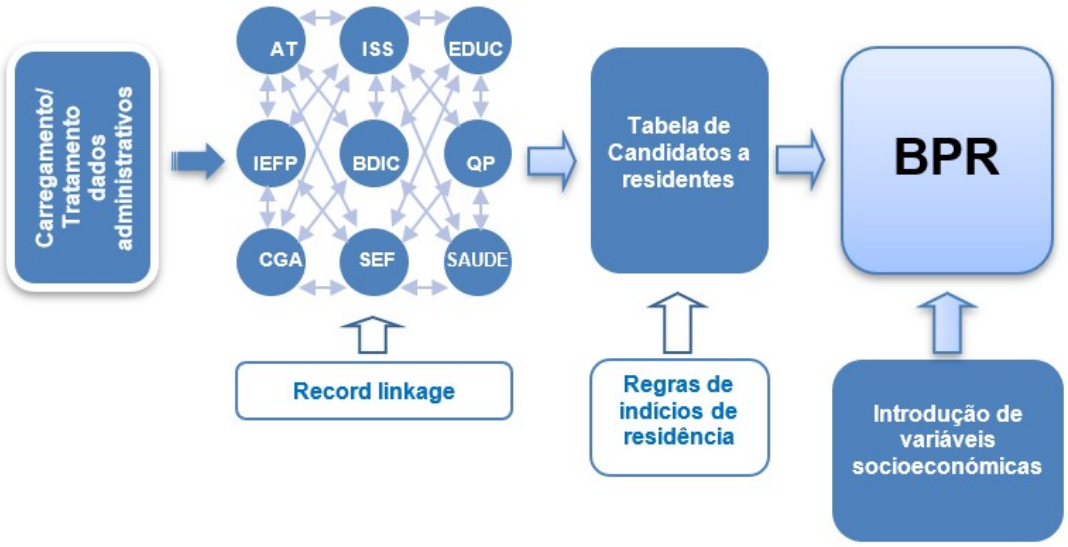


Figura 2.1: Etapas para construção da BPR

Fonte: INE [12]

A Deliberação da Comissão Nacional de Proteção de Dados (CNPD) nº 929/2014, datada de 11 de junho e a nº 163/2017, datada de 31 de janeiro, conjugadas com a Lei do SEN, em [1] conferem a moldura legal necessária e indispensável, para esta fase do projeto, de modo a manter a anonimidade dos indivíduos, possibilita o acesso aos dados individuais de acordo com as seguintes restrições:

- Identificadores numéricos encriptados na fonte que permitem a ligação por chave única;
- Nome do indivíduo limitado às 3 primeiras letras do primeiro nome e 3 últimas do último nome;
- Morada do indivíduo limitada à localidade e código postal.

Já em 2018, a proposta de lei nº 152/XIII (Lei dos Censos 2021) confere o enquadramento legal de acesso a dados administrativos para fins estatísticos, salvaguardando a proteção dos dados pessoais.

Até à data foram produzidas 3 edições da *BPR* para os anos de referência de 2011, 2015 e 2016. Está em construção a edição de 2017. Os resultados alcançados até ao momento mostram que esta Base de População Residente tem potencial para vir a constituir um repositório único de informação oficial estatística em Portugal, de cariz demográfico e socioeconómico, atualizável, por via administrativa.

2.1 Fontes de dados administrativos

Coube ao INE articular e definir com as diferentes entidades da Administração Pública o conteúdo informacional necessário e as condições de tratamento e transmissão da informação. De forma a permitir a cada entidade transmitir os dados ao INE, de forma segura e de acordo com os requisitos estabelecidos pela CNPD, foi utilizada uma aplicação de codificação de dados chamada CDA. Esta encripta os identificadores numéricos e abrevia os campos relativos ao nome do indivíduo.

Carregados os ficheiros das diversas fontes, cada ficheiro foi analisado separadamente, sendo produzido um relatório com as características de cada um. No sentido de se harmonizar a informação das diversas fontes foi definido um conjunto de regras, com objetivo de reduzir as diferenças nas codificações e converter as variáveis para os formatos pretendidos. Foram utilizadas todas as fontes de informação para o preenchimento das variáveis em observação. Existem no entanto dados omissos que têm de ser tratados através de métodos de imputação.

2.2 Imputação de dados omissos

Como já foi mencionado, existe o compromisso de disponibilizar um conjunto de estatísticas da população ao EUROSTAT. A divulgação dessa informação estatística é obrigatória para todos os Estados Membros. Para obter uma caracterização exaustiva da população residente, é necessário que exista informação nas várias fontes para o maior número possível de registos. Para obter um preenchimento de 100% é necessário recorrer a métodos de imputação que atribuem valores aos dados omissos em função da base de dados existente ou de bases prévias.

Existem diferentes técnicas de imputação, que serão mais ou menos adequadas em função das características dos dados em tratamento. De seguida apresenta-se uma explicação dos tipos de métodos usados habitualmente.

- **Métodos Dedutivos**: atribui-se um valor que é deduzido a partir de informação conhecida, referente ao mesmo registo, no caso em estudo ao mesmo indivíduo. Esta informação estará disponível através de registos prévios, pelo que terá um erro associado caso esta informação seja passível de se alterar com o passar do tempo;

- **Métodos Determinísticos**: atribui-se um valor igual para todos os registos, baseando-se na informação completa da base de dados em tratamento. Exemplos de métodos deste tipo são a imputação pela média ou mediana e as que recorrem a métodos regressivos;

- **Métodos Estocásticos**: atribuem-se valores diferentes consoante as características do registo a imputar. Apenas podem ser usados para o tratamento de omissões parciais nos dados, visto que são necessárias outras características do indivíduo para saber que valor atribuir. Os mais usados são os métodos de Hot-Deck e a imputação por associação flexível.

Na última operação censitária em 2011 aplicaram-se correções automáticas que segundo a respetiva metodologia, em [11], recorreu-se ao método de imputação por *Hot-deck*, em que, para cada resposta

omissa a determinada variável do indivíduo 1, o sistema recorreu a outro indivíduo 2, geograficamente próximo, com duas ou mais características idênticas e com resposta à variável em causa (INE, 2013).

O estudo cujos resultados agora se apresentam visa no âmbito dos trabalhos da *BPR_2017* (em preparação), propõem uma variável que calcula o erro associado à informação atribuída a registos com valores omissos, através de métodos dedutivos de imputação. A metodologia dedutiva em estudo consiste na atribuição do último valor observado quando certificado que se referem ao mesmo indivíduo.

Capítulo 3

Cadeias de Markov

Considerando o objetivo de avaliar a qualidade da informação imputada na *BPR* através de métodos dedutivos, calcula-se a probabilidade das variáveis permanecerem no mesmo estado em dois períodos de tempo (passo). Para estimar a probabilidade de permanência no estado i a n passos são definidas cadeias de Markov a tempo discreto. Segundo Manning, em [7], estas cadeias são um processo estocástico a tempo discreto e a distribuição de probabilidades dos estados seguintes depende apenas do atual e não dos estados que conduziram ao presente estado.

De seguida resume-se muito brevemente o conceito de cadeia de Markov, apresentam-se as matrizes de transição e as propriedades utilizadas neste estudo. Exemplificam-se as duas variáveis analisadas como modelos de estados múltiplos. É ainda definida a probabilidade de permanência que será um dos produtos finais deste trabalho.

3.1 Definição

Seja n pertencente aos números naturais e X o espaço dos estados possíveis, o processo X_n é uma cadeia de Markov a tempo discreto se:

- X_n é um processo estocástico;
- $\forall_{i,j \in X}$, $\Pr[X_{n+1} = j \mid X_n = i \, , \, X_u \, = \, k \, , \, 0 \leq u < n] = Pr[X_{n+1} = j \mid X_n = i \,]$.

3.2 Matriz de Probabilidade de Transição

Considere-se a probabilidade de transição para o estado j, sabendo que o indivíduo se encontra no estado i, representando-se por p_{ij}.

Seja N o número de estados diferentes em que um indivíduo pode estar, a probabilidade em função dos estados representa-se pela seguinte matriz de Transição:

$$\mathbf{P} = \begin{bmatrix} p_{11} & \dots & p_{1N} \\ \dots & \dots & \dots \\ p_{N1} & \dots & p_{NN} \end{bmatrix}$$

A matriz apresentada indica a probabilidade de todas as transições possíveis e satisfaz a seguinte propriedade:

$$\sum_{j \in X} p_{ij} = 1 \ \forall_i$$

Ou seja é estocástica, as probabilidades de transição de cada estado para todos os existentes no conjunto dos estados somadas dão 1. Quando é necessário estimar estas probabilidades a mais do que um passo, segundo Grinstead, em [4], a matriz a n passos pode ser obtida através da matriz de probabilidade a 1 passo.

$$_n p_{ij} \ = \ \mathbf{P}^{(n)} \ = \ \mathbf{P}^n$$

3.3 Distribuição de X_n

Considere-se X_n uma cadeia de Markov com N estados, pode escrever-se X_n como variável aleatória tal que:

$$\pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ ... \\ \pi_N \end{pmatrix} = \begin{pmatrix} \Pr(X_0 = 1) \\ \Pr(X_0 = 2) \\ ... \\ \Pr(X_0 = N) \end{pmatrix}$$

Para calcular esta distribuição passados n passos, usando as propriedades de Markov e a lei da probabilidade total é demonstrado em [9] o seguinte teorema:

Teorema 3.1: Seja X_n uma cadeia de Markov com N estados e $\mathbf{P}$ a matriz de Transição tem-se:

$$X_0 \sim \pi^T \implies X_n \sim \pi^T \mathbf{P}^n$$

Ou seja, sabendo a distribuição atual da variável e construindo a matriz de transição a 1 passo é possível estimar a distribuição a n passos.

3.4 Modelo de Estados Múltiplos

Considere-se a variável aleatória X_n, que representa o estado em que um indivíduo se encontra no momento n. Estamos perante uma cadeia de Markov a tempo discreto se se verificarem as propriedades enunciadas em 3.1.

3.4.1 Exemplos em estudo

De seguida apresentam-se dois exemplos de aplicação de modelos de estados múltiplos utilizando variáveis existentes na *BPR*.

3.4.1.1 Estado civil

Considerem-se os seguintes estados para a variável *Estado Civil*:

- Solteiro = 1;
- Casado = 2;
- Divorciado/Viúvo = 3

Na figura 3.1 observa-se as transições possíveis entre estados de estado civil.

Figura 3.1: Transições do Estado Civil

Assim a matriz de transição para o estado civil é dada por:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ 0 & p_{22} & p_{23} \\ 0 & p_{32} & p_{33} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ 0 & p_{22} & 1-p_{22} \\ 0 & 1-p_{33} & p_{33} \end{bmatrix}$$

3.4.1.2 Nível de Ensino Completo

Para a variável *Nível de Ensino Completo* temos mais estados relativamente ao exemplo anterior as transições são apenas num sentido. Neste caso, definem-se os estados:

- Ensino Básico 1º Ciclo = 1;
- Ensino Básico 2º Ciclo = 2;
- Ensino Básico 3º Ciclo = 3;

- Ensino Secundário/Profissional = 4;
- Ensino Superior = 5.

Na figura 3.2 observa-se o diagrama de transições.

Figura 3.2: Transições do Nível de Ensino Completo

A matriz de transição para o nível de ensino completo é dada por:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ 0 & p_{22} & p_{23} & p_{24} & p_{25} \\ 0 & 0 & p_{33} & p_{34} & p_{35} \\ 0 & 0 & 0 & p_{44} & p_{45} \\ 0 & 0 & 0 & 0 & p_{55} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ 0 & p_{22} & p_{23} & p_{24} & p_{25} \\ 0 & 0 & p_{33} & p_{34} & p_{35} \\ 0 & 0 & 0 & p_{44} & 1 - p_{44} \\ 0 & 0 & 0 & 0 & p_{55} \end{bmatrix}$$

3.5 Probabilidade de permanência

Seja A o acontecimento em que a variável em observação permanece no mesmo estado i entre dois períodos de tempo. Considerando o período de 1 ano a probabilidade do acontecimento A ocorrer pode escrever-se da seguinte forma:

$$Pr(A) = Pr(X_1 = i \mid X_0 = i) = p_{ii}\ , \forall_{i \in X}$$

Tendo em conta as matrizes de transição, esta probabilidade é dada em função do estado atual do indivíduo em observação e corresponde aos valores da diagonal da matriz de transição respetiva. No presente caso de estudo é assumido que as duas variáveis seguem cadeias de Markov independentes e que a probabilidade da interseção destes acontecimentos calcula-se multiplicando as duas probabilidades.

Capítulo 4

Análise de *Clusters*

Tendo em conta a dimensão das bases de dados populacionais, na ordem dos milhões de registos, para as analisar eficazmente é aconselhável tratar os dados antes de retirar conclusões estatísticas. Existem várias ferramentas que ajudam na análise de grandes bases de dados como complemento às técnicas de tratamento. A análise de *clusters*, também designada por *clustering* é utilizada neste estudo de forma a otimizar o algoritmo de construção das matrizes de transição de probabilidade, caracterizando também a população em segmentos que permitem uma aproximação mais acertada para cada grupo de indivíduos.

Neste capítulo são apresentados alguns métodos de análise de *clusters* e também um método que apoia a decisão sobre quantas partições aplicar a uma população, de forma a reduzir as distâncias entre elementos da mesma partição.

4.1 Definição

O *clustering* é um processo que tem como objetivo dividir um conjunto de dados em subconjuntos. Consiste na partição em k grupos distintos $C = \{C_1, C_2, ..., C_k\}$, sendo k o número de *clusters* tal que:

Seja X um conjunto com n elementos, $X = \{X_1, X_2, ...X_n\}$, em que $X_i \in IR^p$ é um vector com p variáveis, então:

- $C_1 \cup C_2 \cup ... \cup C_k = X$;
- $C_i \neq \emptyset$, $\forall i$, $1 \leq i \leq k$;
- $C_i \cap C_j = \emptyset$, $\forall i \neq j$, $1 \leq i \leq k$ *e* $1 \leq j \leq k$.

4.2 Métodos de *Clustering*

Os métodos usados na construção de *clusters* analisam a semelhança entre objetos, agrupando-os em função da distância entre certos aspectos de objetos diferentes. Para ser possível a sua aplicação é necessário ter alguns cuidados na construção do algoritmo de *clustering*. Segundo Han e Kamber, em [6], relativamente ao algoritmo a implementar deve ter-se em conta:

- **escalabilidade:** o método deve estar preparado para o aumento dos dados, independentemente da ordem de grandeza atual;
- **Versatilidade:** deve comportar diversos tipos de formatos no que diz respeitos às variáveis de *input* e *output*, ao tamanho dos *clusters* e à existência de ”ruído”nos dados;
- **Interpretabilidade e usabilidade:** ser possível interpretar o código e altera-lo consoante novas restrições sem ter que redesenhar todo o código é importante bem como a sua aplicabilidade;
- **Critério de particionamento:** existem tipos de partições hierárquicas e não (todos os *clusters* estão no mesmo nível), dependendo dos tipo de dados deve ser definido qual o critério a utilizar.

É assim possível separar em conjuntos os diversos algoritmos que se apresentam com maior detalhe de seguida.

4.2.1 Métodos hierárquicos

Estes métodos baseiam-se numa decomposição de forma hierárquica, existindo dois tipos, os divisivos e os aglomerativos. No caso de uma aproximação aglomerativa, começa-se com um número de *clusters* igual ao de objetos, formando de seguida grupos em função da distância entre os objetos até que se chega a um só grupo. Nos divisivos tem-se o contrário, começando-se com um único *cluster* e, nas seguintes iterações vai-se dividindo em subgrupos até atingir a condição de paragem.

4.2.2 Métodos por partições

Geralmente os métodos por partições ”encontram a melhor partição, de acordo com uma medida de similaridade”[8]. Inicia-se o algoritmo com uma partição inicial seguido de técnicas de realocação iterativas que melhorar a qualidade da partição movendo objetos de uns *clusters* para os outros.
Exemplos de métodos por partições são o *k-means* e o *k-medoids*, estes usam métodos heurísticos[1] para iniciar o algoritmo e vão iterativamente melhorando a qualidade da partição. Estes métodos definem para cada *cluster* um elemento representativo relativamente ao qual é calculado a medida de similaridade.

4.2.2.1 Algoritmo K-medoids e CLARA

Este método define como centros de cada *cluster* o elemento mais representativo e não a média como acontece no k-means, mais adequado para dados contínuos. Segundo Oliveira, em [8], o algoritmo inicia com elementos aleatórios como os representantes de cada um dos *clusters*, denominados de *medoids*,

[1] Métodos que simplificam um problema complexo, transformando-o em problemas mais simples cuja resolução conjunta permite chegar à solução do problema principal.

seguindo-se de forma iterativa de todas as substituições possíveis destes *medoids* até não existir nenhuma melhoria à qualidade dos *clusters.*

Quando se está perante bases de dados com elevadas dimensões é mencionado em [6] que antes de se aplicar o algoritmo aos dados deve retirar-se uma amostra significativa aplicando o método CLARA. Este algoritmo retira múltiplas amostras aleatórias e posteriormente aplica o algoritmo K-medoids a cada, devolvendo a melhor partição [5].

4.2.3 Outros métodos

Como alternativas aos métodos acima descritos tem-se os métodos baseados em densidades e os baseados em grelhas/secções. Estes métodos são menos usados mas úteis para certos tipos de dados. Os baseados em densidades particionam os dados em função do número de objetos por *cluster*, sendo útil para excluir *outliers*. Quanto aos baseados em grelhas tem como principal vantagem a sua rapidez pois apenas se preocupa com a posição na grelha dos objetos. A ideia é dividir o espaço quantitativo em pequenas secções e colocar em *clusters* os objetos que se encontram em secções diferentes. Existem ainda algoritmos mais específicos dada a importância desta ferramenta no estudo de Big Data[2].

[2]Big data é um termo que descreve grandes volumes de alta velocidade, dados complexos e variáveis que requerem técnicas avançadas e tecnologias para possibilitar a sua captura, armazenamento, distribuição, gestão e análise. [3]

4.3 Método para definição do número de *Clusters "Elbow"*

Considerando um conjunto de dados por agrupar existe k número de *clusters* a partir do qual o modelo não beneficia com o aumento deste. A partir desta ideia este método permite definir um k para ser usado nos métodos que necessitam de um número inicial de partições para efetuarem o *clustering*, como o k-means e o k-medoids. Segundo Bholowalia, em [2], este método baseia-se na observação do gráfico da *within-cluster sum of square* (WCSS) ou soma das distâncias entre objetos do mesmo *cluster* em função do número de *clusters*. Assim o número de partições a definir para ser utilizado no método é o primeiro que produzir no gráfico um ângulo significativamente menor que os anteriores como é exemplificado na figura 4.1.

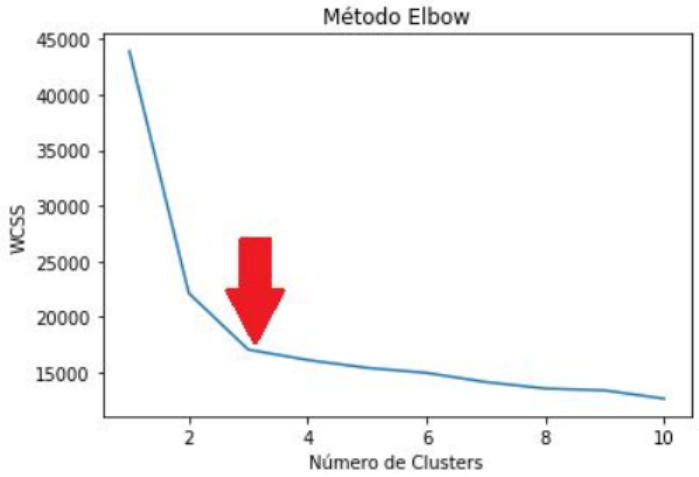


Figura 4.1: Exemplo da decisão do método Elbow

Neste exemplo a variância dos *clusters* quando são formados 3 ou 4 é muito semelhante constatando-se que a construção de mais do que 3 *clusters* não será significativa. No estudo desenvolvido foi aplicado o algoritmo CLARA, referido na secção 4.2.2.1, em duas bases de dados distintas depois de tomada a decisão de quantas partições dividir os dados com apoio do método acima descrito.

Capítulo 5

Resultados

Neste capítulo são apresentados os resultados da aplicação da metodologia proposta às duas variáveis da *BPR*, *estado civil* e *nível de ensino completo*, abreviadas respetivamente com as siglas *EC* e *NEC*. Partindo da descrição da base de dados, qual o tratamento que sofreu, aplicação dos conceitos inerentes a cadeias de Markov. Finaliza-se com a análise de *clusters* e a construção de tabelas de probabilidades de permanência no mesmo estado, passado um certo período de tempo.

5.1 Os Dados

Para este estudo foi utilizada uma base de dados obtida através do cruzamento entre a Base de dados dos Censos 2011, a *BPR_2015* e a *BPR_2016*, usando como chave a variável *IND* (identificador numérico que liga *BPR* com Censos), garantindo que as observações de diferentes anos são referentes ao mesmo universo de indivíduos. A base de dados tem 8.506.481 observações e é composta pelas 10 variáveis seguintes:

- *ind_bpr2016* – Identificador numérico único do registo na BPR_2016;
- *est_civ_2011* – Estado civil do indivíduo registado nos Censos 2011;
- *est_civ_2015* – Estado civil do indivíduo registado na BPR_2015;
- *est_civ_2016* – Estado civil do indivíduo registado na BPR_2016;
- *nec_2011* – Nível de Ensino Completo do indivíduo registado nos Censos 2011;
- *nec_2015* – Nível de Ensino Completo do indivíduo registado na BPR_2015;
- *nec_2016* – Nível de Ensino Completo do indivíduo registado na BPR_2016;
- *sexo_censos* – Sexo do indivíduo registado nos Censos 2011;
- *nac_censos* – Nacionalidade do indivíduo registado nos Censos 2011;
- *idade_2016* – Idade do indivíduo registado nos Censos 2011 com um incremento de 5 anos.

Na figura 5.1 apresenta-se um quadro resumo, extraído do software **Stata**, com o total de observações, número de categorias únicas, média, máximo e mínimo de cada variável. Observa-se que existem registos com dados omissos que são excluídos em função da variável em tratamento.

Variable	Obs	Unique	Mean	Min	Max	Label
ind_bpr2016	8506481	8506481	5184496	2	1.03e+07	Identificador
est_civ_2011	8506481	5	1.888211	1	5	Estado Civil 2011
est_civ_2015	8488391	7	1.979978	1	9	Estado Civil 2015
est_civ_2016	8506481	6	2.016165	1	9	Estado Civil 2016
nec_2011	8506481	10	3.500122	1	10	Nível de Ensino 2011
nec_2015	2310470	10	4.408079	1	10	Nível de Ensino 2015
nec_2016	2339585	10	4.464612	1	10	Nível de Ensino 2016
sexo_censos	8506481	2	1.521805	1	2	Sexo
nac_censos	8506481	194	14.58476	1	998	Nacionalidade
idade_2016	8506481	107	45.50902	5	111	Idade

Figura 5.1: Medidas descritivas da base de dados inicial

5.1.1 Pre-Tratamento dos Dados

Tendo em vista a análise das variáveis EC e NEC separadamente, dividiu-se a base de dados em duas partes, excluindo todas as observações com dados omissos para algum dos anos observados. Foram também excluídas as variáveis que não são relevantes para cada caso de estudo. Para todo o tratamento de dados e exportação de resultados foi utilizado o software **Stata**, encontrando-se em anexo o código comentado. Para otimizar a análise de *clusters* as variáveis correspondentes à idade e nacionalidade do indivíduo foram transformadas da seguinte forma:

$$Idade = \begin{cases} 1, \ se \ idade\_2016 < 16 \\ 2, \ se \ 16 \leq \ idade\_2016 < 36 \\ 3, \ se \ 36 \leq \ idade\_2016 < 56 \\ 4, \ se \ 56 \leq \ idade\_2016 \end{cases}$$

$$Nacionalidade = \begin{cases} 1, \ se \ nac\_censos \ = \ `010' \\ 2, c.c. \end{cases}$$

Para obter um universo de indivíduos sem dados omissos é necessário restringir a base de dados aos registos que têm informação para os 3 anos em observação. Nas secções seguintes apresentam-se as categorias e número de observações de cada variável depois de tratadas.

Nas figuras 5.2 e 5.3 apresentam-se as duas bases de dados tratadas, que serão designadas daqui em diante como *BD_EC* e *BD_NEC*, sendo a primeira relativa ao estudo da variável *EC* e a segunda da variável *NEC*.

Variable	Obs	Unique	Mean	Min	Max	Label
est_civ_2011	8463554	5	1.889396	1	5	Estado Civil 2011
est_civ_2015	8463554	5	1.964049	1	5	Estado Civil 2015
est_civ_2016	8463554	5	1.987253	1	5	Estado Civil 2016
sexo_censos	8463554	2	1.521915	1	2	Sexo
Nacionalid~e	8463554	2	1.0126	1	2	Nacionalidade
Idade_2016~t	8463554	4	2.901546	1	4	Idade

Figura 5.2: Medidas descritivas *BD_EC*

Variable	Obs	Unique	Mean	Min	Max	Label
nec_2011	1953784	10	4.329918	1	10	Nível de Ensino 2011
nec_2015	1953784	10	4.397583	1	10	Nível de Ensino 2015
nec_2016	1953784	10	4.408999	1	10	Nível de Ensino 2016
sexo_censos	1953784	2	1.47471	1	2	Sexo
Nacionalid~e	1953784	2	1.023012	1	2	Nacionalidade
Idade_2016~t	1953784	4	2.880155	1	4	Idade

Figura 5.3: Medidas descritivas *BD_NEC*

5.1.1.1 Tratamento da *BD_EC*

Para a análise da variável *EC* foram ignoradas as variáveis relativas ao nível de ensino e todas as categorias caracterizadas como desconhecidas perfazendo um total de 8.463.554 observações. Adicionalmente foram corrigidas cerca de 40 mil situações de regresso ao estado de solteiro aplicando-se a seguinte regra:

$$\begin{cases} est\_civ\_2015 = est\_civ\_2011 \text{ , } se\ est\_civ\_2015 = 1 \ \wedge \ nec\_2011 \ \neq \ 1 \\ est\_civ\_2016 = est\_civ\_2015 \text{ , } se\ est\_civ\_2016 = 1 \ \wedge \ nec\_2015 \ \neq \ 1 \end{cases}$$

Esta correção que altera as variáveis *estado civil 2015* e *estado civil 2016* reduz o número de solteiros, não permitindo o regresso a este estado a partir de qualquer outro estado. Por exemplo, se um indivíduo é ”viúvo” ($EC = 5$), num dado momento, passado n passos, não pode voltar ao estado ”solteiro” ($EC = 1$). As figuras seguintes apresentam as frequências para cada variável.

Estado Civil 2011	Freq.	Percent	Cum.
Solteiro	3,405,096	40.23	40.23
Casado	4,032,608	47.65	87.88
Separado	67,732	0.80	88.68
Divorciado	473,095	5.59	94.27
Viúvo	485,023	5.73	100.00
Total	8,463,554	100.00	

Figura 5.4: Frequências *Estado Civil* 2011

Estado Civil 2015	Freq.	Percent	Cum.
Solteiro	3,190,643	37.70	37.70
Casado	4,083,171	48.24	85.94
Separado	6,247	0.07	86.02
Divorciado	605,296	7.15	93.17
Viúvo	578,197	6.83	100.00
Total	8,463,554	100.00	

Figura 5.5: Frequências *Estado Civil* 2015

Estado Civil 2016	Freq.	Percent	Cum.
Solteiro	3,153,784	37.26	37.26
Casado	4,056,648	47.93	85.19
Separado	6,604	0.08	85.27
Divorciado	634,879	7.50	92.77
Viúvo	611,639	7.23	100.00
Total	8,463,554	100.00	

Figura 5.6: Frequências *Estado Civil* 2016

Da análise das frequências anteriores observa-se que existe uma redução perto dos 90% do número de indivíduos no estado ”Separado”entre 2011 e 2015. Este fato pode ser causado pelo método de recolha e conceitos serem diferentes. É necessário optar por um único conceito de forma a ultrapassar a limitação enunciada. Quanto à variável *idade* verifica-se que mais de 35% dos registos têm idade superior a 55 anos de idade, característica conhecida da população portuguesa (ver figura 5.9).

Sexo	Freq.	Percent	Cum.
Masculino	4,046,301	47.81	47.81
Feminino	4,417,253	52.19	100.00
Total	8,463,554	100.00	

Figura 5.7: Frequências *Sexo*

Nacionalidade	Freq.	Percent	Cum.
Portuguesa	8,356,909	98.74	98.74
Estrangeira	106,645	1.26	100.00
Total	8,463,554	100.00	

Figura 5.8: Frequências *Nacionalidade*

Idade	Freq.	Percent	Cum.
Menos de 16 anos	942,565	11.14	11.14
Entre 16 e 35 anos	1,923,057	22.72	33.86
Entre 36 e 55 anos	2,623,016	30.99	64.85
Mais de 55 anos	2,974,916	35.15	100.00
Total	8,463,554	100.00	

Figura 5.9: Frequências Idade

Este tipo de análise ajuda a interpretação dos futuros *clusters* e dos resultados obtidos. Segue-se um tratamento semelhante para a variável *NEC*.

5.1.1.2 Tratamento da *BD_NEC*

Para a análise da variável *NEC* foram ignoradas as variáveis relativas ao *EC* e todas as categorias caracterizadas como desconhecidas restando 1.953.784 observações. De seguida apresentam-se as frequências da variável *NEC* para os 3 anos observados.

Nível de Ensino 2011	Freq.	Percent	Cum.
Nenhum	55,469	2.84	2.84
1º Ciclo	332,951	17.04	19.88
2º Ciclo	382,696	19.59	39.47
3º Ciclo	495,305	25.35	64.82
Ensino Secundário/Profissional	282,561	14.46	79.28
Ensino Pós Secundário	21,010	1.08	80.36
Bacharelato	38,980	2.00	82.35
Licenciatura	308,914	15.81	98.16
Mestrado	32,971	1.69	99.85
Doutoramento	2,927	0.15	100.00
Total	1,953,784	100.00	

Figura 5.10: Frequências *Nível de Ensino* 2011

Nível de Ensino 2015	Freq.	Percent	Cum.
Nenhum	31,140	1.59	1.59
1º Ciclo	342,644	17.54	19.13
2º Ciclo	379,463	19.42	38.55
3º Ciclo	595,387	30.47	69.03
Ensino Secundário/Profissional	155,228	7.94	76.97
Ensino Pós Secundário	12,331	0.63	77.60
Bacharelato	41,403	2.12	79.72
Licenciatura	354,014	18.12	97.84
Mestrado	38,236	1.96	99.80
Doutoramento	3,938	0.20	100.00
Total	1,953,784	100.00	

Figura 5.11: Frequências *Nível de Ensino* 2015

Nível de Ensino 2016	Freq.	Percent	Cum.
Nenhum	30,992	1.59	1.59
1º Ciclo	339,880	17.40	18.98
2º Ciclo	376,663	19.28	38.26
3º Ciclo	596,062	30.51	68.77
Ensino Secundário/Profissional	158,229	8.10	76.87
Ensino Pós Secundário	12,482	0.64	77.51
Bacharelato	40,469	2.07	79.58
Licenciatura	355,183	18.18	97.76
Mestrado	39,781	2.04	99.79
Doutoramento	4,043	0.21	100.00
Total	1,953,784	100.00	

Figura 5.12: Frequências *Nível de Ensino* 2016

Depois de analisadas as transições a 5 anos entre 2011 e 2016 e a 1 ano entre 2015 e 2016 foram encontradas transições não compatíveis com a variável em questão, *NEC*. Essas observações foram tratadas da seguinte forma:

$$\begin{cases} nec\_2015 = nec\_2011 \text{ , } se\ nec\_2015 < nec\_2011 \\ nec\_2016 = nec\_2015 \text{ , } se\ nec\_2016 < nec\_2015 \end{cases}$$

Quanto às variáveis que serão alvo de uma análise de *clusters* (*sexo, nacionalidade e idade*), observa-se que a percentagem de indivíduos com menos de 16 anos é bastante inferior ao normal em qualquer amostra da nossa população residente em Portugal. Este fato deve-se à exclusão de todos os registos sem *NEC* registado em qualquer dos anos. Como nos inquéritos censitários, indivíduos com menos de 15 anos não necessitam de responder à questão que retorna o *NEC*, são poucos os registos com essa característica registada nos 3 anos observados.

Sexo	Freq.	Percent	Cum.
Masculino	1,026,303	52.53	52.53
Feminino	927,481	47.47	100.00
Total	1,953,784	100.00	

Figura 5.13: Frequências *Sexo*

Nacionalidade	Freq.	Percent	Cum.
Portuguesa	1,908,823	97.70	97.70
Estrangeira	44,961	2.30	100.00
Total	1,953,784	100.00	

Figura 5.14: Frequências *Nacionalidade*

Idade	Freq.	Percent	Cum.
Menos de 16 anos	2,638	0.14	0.14
Entre 16 e 35 anos	543,549	27.82	27.96
Entre 36 e 55 anos	1,092,923	55.94	83.89
Mais de 55 anos	314,674	16.11	100.00
Total	1,953,784	100.00	

Figura 5.15: Frequências Idade

5.2 Distribuição das variáveis EC e NEC

Para obter as distribuições das variáveis em estudo referentes a cada população nos diferentes anos foram calculadas as frequências das variáveis *EC* e *NEC* para os anos de 2011 e 2015, com o objetivo de caracterizar os estados iniciais para as transições a 5 e a 1 passo respetivamente, até 2016. Estas distribuições são utilizadas para definir os processos como cadeias de Markov no *software Wolfram Mathematica*.

5.2.1 Matriz Pi para o *EC* em 2011

São identificados na variável *EC* 5 estados: Solteiro, Casado, Separado, Divorciado e Viúvo, representando cada linha da matriz que caracteriza a distribuição da variável em 2011 apresentada de seguida.

$$\pi_{2011} = \begin{bmatrix} 40.23\% \\ 47.65\% \\ 0.80\% \\ 5.59\% \\ 5.73\% \end{bmatrix}$$

Esta matriz corresponde à distribuição de partida, considerando que os dados da variável EC seguem uma cadeia de Markov a tempo discreto com passos quinquenais, ou seja, transições entre 2011 e 2016.

5.2.2 Matriz Pi para o *EC* em 2015

Já em 2015 e com dados recolhidos administrativamente obtém-se a seguinte matriz, considerando que a variável *NEC* também segue uma cadeia de Markov a tempo discreto mas com passos anuais, entre 2015 e 2016.

$$\pi_{2015} = \begin{bmatrix} 37.70\% \\ 48.24\% \\ 0.07\% \\ 7.15\% \\ 6.83\% \end{bmatrix}$$

Chama-se a atenção para o estado ”separado”que conta com menos de 1% de registos em ambas distribuições.

5.2.3 Matriz Pi para a *BD_NEC* em 2011

São identificados na variável *NEC* 10 estados: Nenhum, 1º Ciclo, 2º Ciclo, 3º Ciclo, Ensino Secundário/Profissional, Ensino Pós-Secundário, Bacharelato, Licenciatura, Mestrado e Doutoramento. Cada estado é representado por cada linha da matriz que caracteriza a distribuição da variável em 2011 apresentada de seguida na respetiva ordem de enunciação.

$$\pi_{2011} = \begin{bmatrix} 2.84\% \\ 17.04\% \\ 19.59\% \\ 25.35\% \\ 14.46\% \\ 1.08\% \\ 2.00\% \\ 15.81\% \\ 1.69\% \\ 0.15\% \end{bmatrix}$$

5.2.4 Matriz Pi para a *BD_NEC* em 2015

Para o ano de 2015 foi calculada a mesma matriz mas com as frequências do respetivo ano.

$$\pi_{2015} = \begin{bmatrix} 1.59\% \\ 17.54\% \\ 19.42\% \\ 30.47\% \\ 7.94\% \\ 0.63\% \\ 2.12\% \\ 18.12\% \\ 1.96\% \\ 0.20\% \end{bmatrix}$$

É de notar que os últimos níveis de ensino completo, equivalentes aos níveis de ”mestre”e ”doutorado”são pouco populados com seria de esperar apesar de se observar um crescimento entre 2011 e 2015.

5.3 Matrizes de Transição de Probabilidades

Considerando, respectivamente, os processos relativos ao *EC* e *NEC* de um indivíduo cadeias de Markov, com matrizes de transição estimadas através da *LGN*. Neste estudo são obtidas as matrizes de transição a 1 e a 5 passos, sendo que as matrizes a 5 passos são estimas pela *LGN* e também pela potenciação da matriz a 1 passo.

Antes da obtenção dos *clusters* calcularam-se as matrizes de transição para cada base de dados com o objetivo de testar a homogeneidade dos dados. Para tal, as matrizes a 5 passos estimadas pela *LGN* e pela propriedade de Markov são comparadas.

5.3.1 Matrizes de Transição a 1 Passo

Para o cálculo das matrizes de transição, usando o software **Stata**, gerou-se uma variável de transição entre 2015 e 2016 tornando possível construir uma tabela de frequências para cada estado inicial. Através das tabelas obtidas constroem-se as matrizes de transição linha a linha.

5.3.1.1 Matriz a 1 passo: *BD_EC*

A matriz seguinte, estimada pela LGN a partir dos dados observados em 2015 e 2016, representa as probabilidades de transição a 1 ano da população da *BD_EC*. (Os valores apresentados estão em percentagem mas para as matrizes continuarem perceptíveis optou-se por não colocar o símbolo de %, o que acontecerá com as restantes matrizes).

$$\mathbf{P}_{2015} = \begin{bmatrix} 98.84 & 1.15 & 0.00 & 0.01 & 0.00 \\ 0.00 & 98.21 & 0.02 & 0.94 & 0.83 \\ 0.00 & 3.58 & 94.12 & 1.40 & 0.51 \\ 0.00 & 1.46 & 0.00 & 98.52 & 0.02 \\ 0.00 & 0.11 & 0.00 & 0.00 & 99.89 \end{bmatrix}$$

5.3.1.2 Matriz a 1 passo: *BD_NEC*

A matriz seguinte, construída de forma semelhante mas com dados do *nivel de ensino*, representa as probabilidades de transição a 1 ano da população da *BD_NEC*.

$$\mathbf{P}_{2015} = \begin{bmatrix} 96.12 & 2.61 & 0.68 & 0.54 & 0.04 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 96.91 & 1.73 & 1.27 & 0.07 & 0.01 & 0.00 & 0.01 & 0.00 & 0.00 \\ 0.00 & 0.00 & 96.22 & 3.55 & 0.19 & 0.01 & 0.01 & 0.02 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 98.64 & 1.24 & 0.03 & 0.02 & 0.07 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 98.64 & 0.28 & 0.09 & 0.91 & 0.08 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 96.60 & 0.47 & 2.73 & 0.20 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 95.89 & 3.79 & 0.31 & 0.01 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.52 & 1.42 & 0.06 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.68 & 0.32 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00 \end{bmatrix}$$

5.3.2 Matrizes de Transição a 5 Passos

De forma semelhante ao cálculo anterior, a matriz a 5 passos pode ser estimada, desta vez criando uma variável de transição de 2011 para 2016 para estimar as matrizes pela LGN. Recorre-se às matrizes calculadas a 1 passo (apresentadas na secção anterior) para estimar as matrizes pela propriedade de Markov enunciada na secção 4.2.

5.3.2.1 Matrizes a 5 passos pela LGN : *BD_EC*

A matriz seguinte representa as probabilidades de transição a 5 anos da população da *BD_EC*.

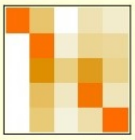
$$\mathbf{P}^{(5)} = \mathbf{P}_{2011} = \begin{bmatrix} 92.62 & 6.72 & 0.00 & 0.58 & 0.08 \\ 0.00 & 92.26 & 0.05 & 3.88 & 3.81 \\ 0.00 & 51.05 & 4.67 & 39.91 & 4.37 \\ 0.00 & 8.87 & 0.27 & 90.39 & 0.47 \\ 0.00 & 6.40 & 0.03 & 0.79 & 92.78 \end{bmatrix}$$

5.3.2.2 Matrizes a 5 passos pela propriedade de Markov: *BD_EC*

Assumindo que os dados são homogêneos aplicam-se as propriedades de Markov a tempo discreto e obtém-se também a matriz de transição a 5 passos para a mesma população.

$$\mathbf{P}^{(5)} = (\mathbf{P}_{2015})^5 = \begin{bmatrix} 94.36 & 5.42 & 0.00 & 0.12 & 0.10 \\ 0.00 & 91.53 & 0.07 & 4.40 & 4.00 \\ 0.00 & 15.60 & 73.89 & 7.96 & 2.55 \\ 0.00 & 6.88 & 0.00 & 92.91 & 0.21 \\ 0.00 & 0.50 & 0.00 & 0.03 & 99.47 \end{bmatrix}$$

Considerando as matrizes a 5 passos obtidas por métodos diferentes para a *BD_EC* observa-se que existe uma diferença significativa na 3ª linha, correspondente às probabilidades de transição partindo do estado "separado". Este já tinha sido mencionado anteriormente como aquele onde mais se evidenciou a diferença entre os dados de 2011 e os recolhidos administrativamente em 2015 e 2016, ou seja, esta diferença nas matrizes de transição a 5 passos era expectável. De seguida apresentam-se as propriedades das cadeias de Markov para cada uma das matrizes. Para obtenção destes resultados recorreu-se ao *software Wolfram Mathematica*(ver código em anexo).

Basic Properties	
TransitionMatrix	
HoldingTimeMean	{12.5501, 11.9032, 0.0489877, 9.40583, 12.8313}
HoldingTimeVariance	{170.056, 153.59, 0.0513875, 97.8754, 177.472}
Structural Properties	
CommunicatingClasses	{2, 3, 4, 5}, {1}
RecurrentClasses	{2, 3, 4, 5}
TransientClasses	{1}
AbsorbingClasses	None
PeriodicClasses	None
Periods	{}
Irreducible	False
Primitive	False
Aperiodic	True

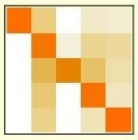
Basic Properties	
TransitionMatrix	
HoldingTimeMean	{16.6991, 10.7925, 2.82995, 13.1044, 17.8679}
HoldingTimeVariance	{295.56, 127.269, 10.8386, 184.829, 337.131}
Structural Properties	
CommunicatingClasses	{2, 3, 4, 5}, {1}
RecurrentClasses	{2, 3, 4, 5}
TransientClasses	{1}
AbsorbingClasses	None
PeriodicClasses	None
Periods	{}
Irreducible	False
Primitive	False
Aperiodic	True

Figura 5.16: Propriedades básicas de $\mathbf{P}_{2011}$ *e* $(\mathbf{P}_{2015})^5$

Os resultados são animadores apesar de não se poder concluir quanto à equivalência entre as duas matrizes ou que as discrepâncias entre elas decorram de erros de estimação. Observa-se que as propriedades das duas matrizes são diferentes excepto as estruturais, concluindo-se pela figura anterior que estruturalmente são iguais. As diferenças podem ser explicadas pela limitação referida na secção 5.1.1.1 relativa ao estado "separado". As probabilidades de cada estado alguma vez ser atingido são apresentadas figura 5.17. Para o estado "solteiro"($EC = 1$) na matriz A corresponde a 37,3% e na B a 35,6%, valores muito próximos.

Transient Properties	
TransientVisitMean	{5.45122}
TransientVisitVariance	{68.4135}
TransientTotalVisitMean	0.4023
Limiting Properties	
ReachabilityProbability	{0.37261, 1., 1., 1., 1.}
LimitTransitionMatrix	
Reversible	False

Transient Properties	
TransientVisitMean	{6.67257}
TransientVisitVariance	{111.426}
TransientTotalVisitMean	0.377
Limiting Properties	
ReachabilityProbability	{0.3557, 1., 1., 1., 1.}
LimitTransitionMatrix	
Reversible	False

Figura 5.17: Propriedades transientes de $\mathbf{P}_{2011}$ *e* $(\mathbf{P}_{2015})^5$

Por fim obtêm-se os grafos com as transições entre estados. O conjunto composto pelos estados 2, 3, 4 e 5 é em ambos os grafos um conjunto fechado. O estado "solteiro"($EC = 1$) é um estado de partida e o único ao qual não é possível regressar.

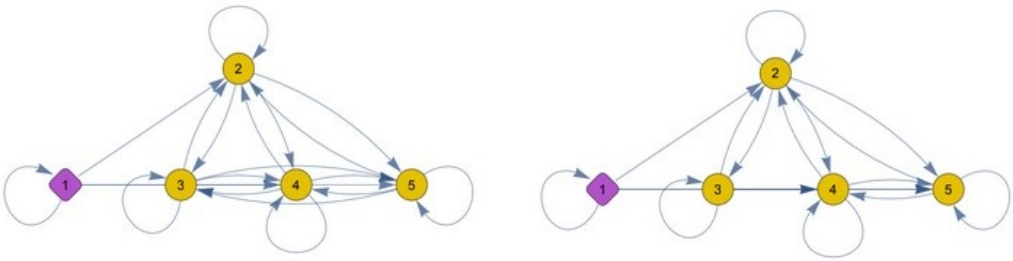


Figura 5.18: Diagrama de Transições de $\mathbf{P}_{2011}$ *e* $(\mathbf{P}_{2015})^5$

5.3.2.3 Matrizes a 5 passos pela LGN: *BD_NEC*

Repetindo o exercício anterior, apresentado nas subsecções anteriores, calculam-se as matrizes de transição, desta vez para os dados referentes ao nível de ensino.

$$\mathbf{P}^{(5)} = \mathbf{P}_{2011} = \begin{bmatrix} 25.92 & 38.59 & 16.34 & 15.17 & 2.05 & 0.16 & 0.26 & 1.39 & 0.12 & 0.01 \\ 0.00 & 67.22 & 19.21 & 12.48 & 0.78 & 0.06 & 0.04 & 0.18 & 0.02 & 0.00 \\ 0.00 & 0.00 & 65.48 & 30.48 & 3.56 & 0.14 & 0.07 & 0.24 & 0.02 & 0.00 \\ 0.00 & 0.00 & 0.00 & 88.36 & 9.07 & 0.46 & 0.19 & 1.79 & 0.12 & 0.01 \\ 0.00 & 0.00 & 0.00 & 0.00 & 75.78 & 1.45 & 1.66 & 17.20 & 3.84 & 0.07 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 71.72 & 4.03 & 21.43 & 2.74 & 0.08 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 64.75 & 32.96 & 2.05 & 0.24 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 93.10 & 6.35 & 0.56 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.81 & 2.19 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00 \end{bmatrix}$$

5.3.2.4 Matrizes a 5 passos pela propriedade de Markov: *BD_NEC*

$$\mathbf{P}^{(5)} = (\mathbf{P}_{2015})^5 = \begin{bmatrix} 82.03 & 11.31 & 3.32 & 3.00 & 0.28 & 0.03 & 0.00 & 0.01 & 0.00 & 0.00 \\ 0.00 & 85.48 & 7.52 & 6.35 & 0.51 & 0.05 & 0.02 & 0.07 & 0.00 & 0.00 \\ 0.00 & 0.00 & 82.46 & 16.02 & 1.28 & 0.06 & 0.04 & 0.13 & 0.01 & 0.00 \\ 0.00 & 0.00 & 0.00 & 93.40 & 5.87 & 0.17 & 0.09 & 0.44 & 0.03 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 93.40 & 1.26 & 0.44 & 4.39 & 0.50 & 0.01 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 84.11 & 2.02 & 12.51 & 1.31 & 0.04 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 81.06 & 16.92 & 1.93 & 0.09 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 92.82 & 6.87 & 0.32 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.39 & 1.61 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00 \end{bmatrix}$$

As matrizes $\mathbf{P}_{2011}$ *e* $(\mathbf{P}_{2015})^5$, estimadas por métodos diferentes mas correspondentes à mesma matriz de transição, a matriz a 5 passos, ao contrário do caso de estudo anterior, são estruturalmente diferentes. Note-se que a partir da comparação da diagonal de cada matriz existe uma diferença significativa entre as respetivas probabilidades. Assumindo que os dados da *BPR* estão corretos, que os indivíduos observados nos diferentes anos são os mesmos e que os cálculos das probabilidades de transição estão certos a matriz calculada pela *LGN* será a que mais se aproxima da realidade. Resta-nos a matriz calculada a partir das propriedades de uma cadeia de Markov, onde se pressupõe a homogeneidade dos dados. Este pressuposto pode não ser verificado implicando um erro de estimação quando a matriz a 5 passos é obtida pela matriz a 1.

É de notar a limitação que existe na variável *NEC* no que toca ao período mínimo que um indivíduo necessita para transitar de estado. Por exemplo, um indivíduo que tenha acabado o ensino secundário necessita no mínimo de 3 anos para concluir uma licenciatura ao contrário do que acontece com o estado civil. Pelo exposto, o estudo detalhado das duas matrizes não se considera necessário. Por outro lado terá interesse fazer um estudo mais aprofundado de como calcular as probabilidades de permanência de um indivíduo no mesmo nível de ensino em função das características observadas.

5.4 Análise de *Clusters*

Considerando o objetivo deste estudo, de criar uma variável que caracteriza a probabilidade de um registo alvo de imputação estar correto é necessário segmentar os registos em *clusters* de forma a ser possível calcular as matrizes de transição para cada variável, *EC* e *NEC*. Utilizam-se como variáveis explicativas o *sexo* a *nacionalidade* e a *idade*.

5.4.1 Método *Elbow*

O método *k-medoid*s tem como *input*, além dos dados, o número de *clusters* que o algoritmo tem que construir. Nas figuras seguintes são apresentados os gráficos de apoio à aplicação do método *"Elbow"*.

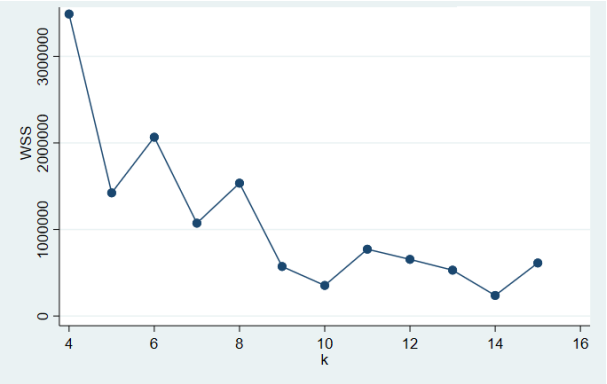


Figura 5.19: Método de Elbow *BD_EC*

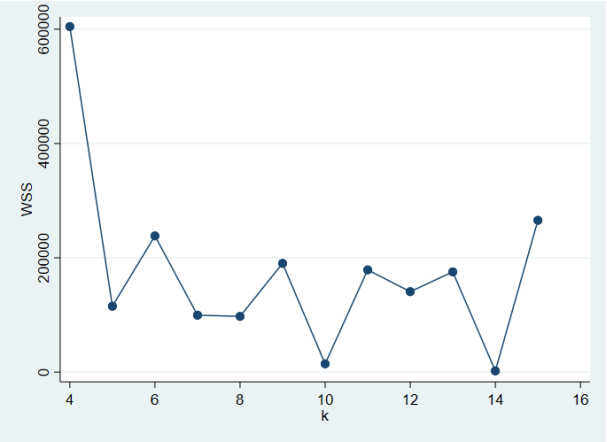


Figura 5.20: Método de Elbow *BD_NEC*

Em ambos os casos o valor de k a definir como *input* do *k-medoids*, a partir do qual aumentar o número de *clusters* não reduz significativamente as distâncias entre objetos do mesmo grupo, é o 10. Na subsecção seguinte será aplicado o algoritmo CLARA, caso particular do *k-medoids*.

5.4.2 Cálculo de *Clusters*

De seguida são apresentadas as frequências das variáveis *cluster* criadas pelo algoritmo *CLARA* com um k inicial igual a 10. Utilizou-se a função clara do *package cluster* do *software* **R** para programar o algoritmo, encontrando-se o código em anexo.

Cluster_EC	Freq.	Percent	Cum.
1	35,708	0.44	0.44
2	45,730	0.57	1.01
3	1,493,143	18.46	19.47
4	1,332,170	16.47	35.94
5	1,207,665	14.93	50.87
6	456,188	5.64	56.51
7	901,135	11.14	67.65
8	890,563	11.01	78.66
9	1,245,339	15.40	94.06
10	480,767	5.94	100.00

Figura 5.21: Frequências *Cluster_EC*

Cluster_NEC	Freq.	Percent	Cum.
1	8,544	0.44	0.44
2	9,065	0.46	0.90
3	14,504	0.74	1.64
4	12,848	0.66	2.30
5	267,588	13.70	16.00
6	506,483	25.92	41.92
7	265,965	13.61	55.53
8	559,088	28.62	84.15
9	178,235	9.12	93.27
10	131,464	6.73	100.00

Figura 5.22: Frequências *Cluster_NEC*

5.4.3 Descrição dos *Clusters*

O algoritmo utilizado segmentou os dados a partir das variáveis *Sexo*, *Nacionalidade* e *Idade*. Para cada base de dados, referentes às duas variáveis em estudo, constroem-se os respetivos *clusters*. Seguidamente são caracterizados os *clusters* aplicados a cada base de dados.

5.4.3.1 Base de dados *EC*

Para a *BD_EC* obtiveram-se 10 *Clusters*. O algoritmo agrupou com características elementares os indivíduos com menos de 16 anos e os indivíduos com mais de 55 anos por sexo (*clusters* 6, 10, 3 e 4).

- ***Cluster 1:*** Indivíduos do sexo masculino, nacionalidade estrangeira, com idade entre os 16 e os 55 anos;
- ***Cluster 2:*** Indivíduos do sexo feminino, nacionalidade estrangeira, com idade entre os 16 e os 55 anos;
- ***Cluster 3:*** Indivíduos do sexo feminino, com idade superior a 55 anos;
- ***Cluster 4:*** Indivíduos do sexo masculino, com idade superior a 55 anos;
- ***Cluster 5:*** Indivíduos do sexo masculino, nacionalidade portuguesa, com idade entre os 36 e os 55 anos;
- ***Cluster 6:*** Indivíduos do sexo feminino, com idade inferior a 16 anos;
- ***Cluster 7:*** Indivíduos do sexo feminino, nacionalidade portuguesa, com idade entre os 16 e os 35 anos;
- ***Cluster 8:*** Indivíduos do sexo masculino, nacionalidade portuguesa, com idade entre os 16 e os 35 anos;
- ***Cluster 9:*** Indivíduos do sexo feminino, nacionalidade portuguesa, com idade entre os 36 e os 55 anos;
- ***Cluster 10:*** Indivíduos do sexo masculino, com idade inferior a 16 anos.

A tabela 5.1 apresenta uma síntese das características de cada *cluster* da *BD_EC* por sexo nacionalidade e idade.

Cluster_EC	*Sexo*	*Nacionalidade*	*Idade*
1	Masculino	Estrangeira	Entre 16 e os 55 anos
2	Feminino	Estrangeira	Entre 16 e os 55 anos
3	Feminino	Portuguesa ou Estrangeira	Mais de 55 anos
4	Masculino	Portuguesa ou Estrangeira	Mais de 55 anos
5	Masculino	Portuguesa	Entre 36 e 55 anos
6	Feminino	Portuguesa ou Estrangeira	Menos de 16 anos
7	Feminino	Portuguesa	Entre 16 e 35 anos
8	Masculino	Portuguesa	Entre 16 e 35 anos
9	Feminino	Portuguesa	Entre 36 e 55 anos
10	Masculino	Portuguesa ou Estrangeira	Menos de 16 anos

Tabela 5.1: Tabela descritiva de *Clusters* da *BD_EC*

5.4.3.2 Base de Dados *NEC*

Já para a *BD_NEC* o algoritmo agrupou os indivíduos por nacionalidade, sexo e classe etária, juntando no mesmo *cluster* indivíduos de categorias diferentes apenas para a variável *Idade*.

- ***Cluster 1:*** Indivíduos do sexo masculino, nacionalidade estrangeira, menos de 36 anos ou mais de 55 anos.

- ***Cluster 2:*** Indivíduos do sexo feminino, nacionalidade estrangeira, menos de 36 anos ou mais de 55 anos.

- ***Cluster 3:*** Indivíduos do sexo feminino, nacionalidade estrangeira, com idade entre os 36 e os 55 anos.

- ***Cluster 4:*** Indivíduos do sexo masculino, nacionalidade estrangeira, com idade entre os 36 e os 55 anos.

- ***Cluster 5:*** Indivíduos do sexo masculino, nacionalidade portuguesa, com idade inferior a 36 anos.

- ***Cluster 6:*** Indivíduos do sexo feminino, nacionalidade portuguesa, com idade entre os 36 e os 55 anos.

- ***Cluster 7:*** Indivíduos do sexo feminino, nacionalidade portuguesa, com idade inferior a 36 anos.

- ***Cluster 8:*** Indivíduos do sexo masculino, nacionalidade portuguesa, com idade entre os 36 e os 55 anos.

- ***Cluster 9:*** Indivíduos do sexo masculino, nacionalidade portuguesa, com idade superior a 55 anos.

- ***Cluster 10:*** Indivíduos do sexo feminino, nacionalidade portuguesa, com idade superior a 55 anos.

A tabela 5.2 apresenta uma síntese das características de cada *cluster* da *BD_EC* por sexo nacionalidade e idade.

Cluster_NEC	*Sexo*	*Nacionalidade*	*Idade*
1	Masculino	Estrangeira	Inferior a 36 anos ou Mais de 55 anos
2	Feminino	Estrangeira	Inferior a 36 anos ou Mais de 55 ano
3	Feminino	Estrangeira	Entre 36 e 55 anos
4	Masculino	Estrangeira	Entre 36 e 55 anos
5	Masculino	Portuguesa	Inferior a 36 anos
6	Feminino	Portuguesa	Entre 36 e 55 anos
7	Feminino	Portuguesa	Inferior a 36 anos
8	Masculino	Portuguesa	Entre 36 e 55 anos
9	Masculino	Portuguesa	Mais de 55 anos
10	Feminino	Portuguesa	Mais de 55 anos

Tabela 5.2: Tabela descritiva de *Clusters* da *BD_NEC*

5.4.4 Tabela de Descodificação em *Clusters*

Para associar a cada registo da *BPR* o respetivo *cluster* foi construí-se uma tabela de descodificação, que a partir das 3 variáveis observáveis: *Sexo*, *Nacionalidade* e *Idade* atribuem o número do *cluster* para cada base de dados, *BD_EC* e *BD_NEC*.

Sexo	*Nacionalidade*	*Idade mínima*	*Idade máxima*	*Cluster_EC*	*Cluster_NEC*
1	1	0	15	10	5
1	1	16	35	8	5
1	1	36	55	5	8
1	1	55	110	4	9
1	2	0	15	10	1
1	2	16	35	1	1
...	...	...	...	...	...
2	2	16	35	8	2
2	2	36	55	5	3
2	2	55	125	4	2

Tabela 5.3: Tabela de descodificação de *Clusters*

5.5 Tabela de Probabilidades de permanência

As probabilidades de permanência no mesmo estado para dados imputados na *BPR* são obtidas, quando o último valor observado é referente ao ano anterior, através do cálculo das matrizes de transição a 1 passo para cada variável em estudo e para cada *cluster*. Quando o período desde a última observação é maior aplicam-se as propriedades de Markov para estimar as matrizes de transição através da potência da matriz a um passo, elevando-a à diferença entre o ano atual, ou de referência, e o ano da última observação. Estas matrizes representam-se, por exemplo para o *cluster* 1 da *BD_EC*, pela seguinte notação:

$$_{EC\_1}\mathbf{P}^{Ano\_EC-2011} = {}_{EC\_1}\mathbf{P}^{2017-2011} = {}_{EC\_1}\mathbf{P}^{6}$$

De forma a integrar probabilidade de permanência na *BPR* construiu-se uma tabela com todas as combinações possíveis para as variáveis que indicam o número do *cluster* ao qual o indivíduo pertence, que indicam o último estado observado para as cada variável em estudo e para o ano em que essas observações foram registadas (ver tabelas 5.4 e 5.5).

Clust_EC	*Clust_NEC*	*EC*	*NEC*	Ano_EC	Ano_NEC	Probabilidade de permanência
1	1	1	1	2011	2011	$_{EC\_1}p_{11}^{6} \times_{NEC\_1} p_{11}^{6}$
...	...	...	...	...	...	...
1	1	1	1	2016	2016	$_{EC\_1}p_{11} \times_{NEC\_1} p_{11}$
1	1	2	1	2011	2011	$_{EC\_1}p_{22}^{6} \times_{NEC\_1} p_{11}^{6}$
...	...	...	...	...	...	...
10	10	5	9	2016	2016	$_{EC\_10}p_{55} \times_{NEC\_10} p_{99}$
...	...	...	...	...	...	...
10	10	5	10	2015	2015	$_{EC\_10}p_{55}^{2} \times_{NEC\_10} p_{1010}^{2}$
10	10	5	10	2016	2016	$_{EC\_10}p_{55} \times_{NEC\_10} p_{1010}$

Tabela 5.4: Tabela de probabilidades de permanência

Da tabela anterior obtem-se a tabela seguinte, substituindo as probabilidades obtidas a partir das matrizes de transição calculadas.

Clust_EC	*Clust_NEC*	*EC*	*NEC*	Ano_EC	Ano_NEC	Probabilidade de permanência
1	1	1	1	2011	2011	$0.9377 \times 0.5109 = 47.91\%$
...	...	...	...	...	...	...
1	1	1	1	2016	2016	$0.9893 \times 0.8941 = 88.45\%$
1	1	2	1	2011	2011	$0.9038 \times 0.9920 = 89.66\%$
...	...	...	...	...	...	...
10	10	5	9	2016	2016	$0.9990 \times 0.9920 = 99.10\%$
...	...	...	...	...	...	...
10	10	5	10	2015	2015	$0.9980 \times 1 = 99.8\%$
10	10	5	10	2016	2016	$0.9990 \times 1 = 99.9\%$

Tabela 5.5: Tabela de probabilidades de permanência calculadas

Note-se que a cada ano que passa são acrescentadas combinações com o novo ano observado, ou seja, para cada combinação das primeiras 4 variáveis da tabela são criadas linhas com todas as combinações possíveis entre os anos observados. Cada vez que for realizada uma operação censitária, esta tabela deixa de ser válida para calcular probabilidades a mais do que um passo, visto termos dados exaustivos da população.

Na subsecção seguinte ilustra-se a aplicação das tabelas anteriores para os *clusters* 1 e 10.

5.5.1 *Clusters* da *BD_EC*

- *Cluster* 1 - Indivíduo do sexo masculino, de nacionalidade estrangeira e com idade entre 16 e 55 anos inclusive;
- *Cluster* 10 - Indivíduo do sexo masculino, de nacionalidade portuguesa ou estrangeira e com idade inferior a 16 anos;

Considere-se a probabilidade de permanência no estado solteiro ($EC = 1$) para um homem estrangeiro em Portugal (*cluster* 1). Esta probabilidade é elevada ($> 93\%$), para 1 ou 6 anos, apesar de ser menor em função do número de anos passados desde a última observação. Se calcularmos a probabilidade para o *cluster* 2, semelhante ao 1 mas com apenas mulheres, a mesma probabilidade, dada por ${}_{EC\_2}p^{6}_{11}$, também é superior a 93%.

5.5.2 *Clusters* da *BD_NEC*

- *Cluster* 1 - Indivíduo do sexo masculino, de nacionalidade estrangeira e com idade inferior a 36 anos ou com mais de 55 anos;
- *Cluster* 10 - Indivíduo do sexo feminino, de nacionalidade portuguesa e com idade superior a 55 anos;

Se observarmos a probabilidade de um indivíduo do *cluster* 1 continuar no estado "Nenhum"(*NEC* = 1) esta situa-se perto dos 50%. Esta estimativa pode não ser representativa pois deriva de uma estimativa feita com menos de 100 pessoas, como podemos observar nas frequências relativas à variável de transição entre 2015 e 2016 no estado em questão.

TNiv_Ens_Comp_15_16	Freq.	Percent	Cum.
1-1	76	89.41	89.41
1-2	4	4.71	94.12
1-3	3	3.53	97.65
1-4	2	2.35	100.00
Total	85	100.00	

Figura 5.23: Frequências da variável de transição 2015/2016

Quanto às probabilidades de uma mestre (*NEC* = 9) portuguesa, com mais de 55 anos não registar em bases administrativas um diploma que confira o grau de doutor é estimada em 99.2%. Esta probabilidade é obtida a partir do elemento da linha 9 e coluna 9 da matriz de transição a um passo para o décimo *cluster*. Recorrendo a microdados constata-se que este acontecimento ocorreu 743 vezes em 749 observações.

Aplicando a metodologia aos dados do estado civil, considerando o ano de referência o de 2016 e a distribuição dos indivíduos sem estado civil pelos *clusters*, calcula-se a probabilidade de permanência para cada estado a partir das matrizes de transição a 5 passos. No caso de usarmos as probabilidades estimadas pela LGN obtém-se uma probabilidade ponderada de 90,5% dos dados imputados estarem corretos, contra uma probabilidade de 93% se calculada utilizando as propriedades de Markov.

Capítulo 6

Conclusões e Trabalho Futuro

Os resultados obtidos transmitem confiança de qual o caminho a percorrer nos próximos desenvolvimentos de melhoria da *BPR*. O estudo agora feito deve ser aprofundado, recomendando-se a revisão de todos os resultados e código utilizado, de forma a aferir a sua qualidade.

A semelhança entre as matrizes de transição a cinco passos para o estado civil apoiam a decisão de estimar as probabilidades de permanência assumindo que os dados cumprem os pressupostos de uma cadeia de Markov a tempo discreto, com períodos de um ano. A partir desta hipótese é possível calcular as probabilidades de permanência a n passos a partir da potência da matriz a um passo. Ou seja, este método possibilita avaliar a qualidade dos dados imputados independentemente do período que passou desde a última observação. Quanto ao nível de ensino verifica-se que a estimação destas probabilidades pelo mesmo método, não traduz uma boa aproximação ao estimado pela LGN.

A metodologia proposta, depois de implementada, será a origem da variável que avalia a qualidade dos dados omissos, aos quais foram imputados valores através de métodos dedutivos. É essencial que se repita o estudo para ambas as variáveis com dados de 2017 e 2018, aproveitando as melhorias já propostas na *BPR*. Para a análise da variável *nível de ensino* pode ser interessante incluir na obtenção dos *clusters* a variável que caracteriza a frequência de ensino de um indivíduo quando esta se encontrar estabilizada. Estimar as probabilidades de permanência a n passos a partir da matriz de transição a 2 ou a 3 passos pode também ser uma boa abordagem, visto que as transições entre estados no nível de ensino geralmente têm uma frequência superior a 1 ano.

Um dos desenvolvimentos possíveis será otimizar o código de forma a ser flexível no que toca às variáveis que são utilizadas para o *clustering*, versátil quanto às variáveis para as quais serão calculadas as probabilidades de permanência e escalável, comportando um número de observações maior na expectativa desta metodologia ser aplicada a populações de ordem superior à portuguesa. Depois de testada esta metodologia pode ser adaptada a variáveis de interesse censitário como a *profissão*, *CAE*, *distrito de residência* e *condição perante a atividade económica*.

Por último, a aproximação aos Censos 2021, como a oportunidade para aferir a qualidade da *BPR*. Neste sentido, todos os processos que possam contribuir para melhorar a qualidade desta base de dados, como é o exemplo da metodologia proposta, o cálculo das probabilidades de permanência que caracterizam o erro associado à imputação, podem ser relevantes neste contexto.

Bibliografia

[1] Lei n.°22/2008. *Diário da República n.°92/2008, Série I*, I(92/2008), 2008.

[2] Purnima Bholowalia and Arvind Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9), 2014.

[3] TechAmerica Foundation's Federal Big Data Commission and others. Demystifying big data: a practical guide to transforming the business of government. *Washington, DC*, 2012.

[4] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 2012.

[5] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of intelligent information systems*, 17(2):107–145, 2001.

[6] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier Science, 2011.

[7] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[8] Matilde Oliveira. Calibração e simulação de um modelo de cadeias de markov para um seguro long-term care. Master's thesis, 2017.

[9] Department of Statistics. *Stochastic Processes*. Auckland, University of, 2018.

[10] Instituto Nacional de Estatística. Censos 2001 - Antecedentes, Metodologia e Conceitos. 2003.

[11] Instituto Nacional de Estatística. Censos 2011 - Preparação, Metodologia e Conceitos. 2013.

[12] Instituto Nacional de Estatística. Documentação interna. 2018.

Anexos

Anexo A

Matrizes de Transição do *estado civil* por *Cluster*

A.1 Cluster 1

Indivíduos do sexo masculino, nacionalidade estrangeira, com idade entre os 16 e os 55 anos.

A.1.1 Matriz a 1 Passo

$$\begin{bmatrix} 98.93 & 1.07 & 0.00 & 0.00 & 0.00 \\ 0.00 & 98.29 & 0.02 & 0.91 & 0.79 \\ 0.00 & 4.76 & 95.24 & 0.00 & 0.00 \\ 0.00 & 1.48 & 0.00 & 98.52 & 0.00 \\ 0.00 & 0.20 & 0.00 & 0.00 & 99.80 \end{bmatrix}$$

A.1.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix} 92.64 & 6.72 & 0.00 & 0.51 & 0.12 \\ 0.00 & 92.23 & 0.04 & 3.93 & 3.80 \\ 0.00 & 54.55 & 4.90 & 36.36 & 4.20 \\ 0.00 & 8.72 & 0.15 & 90.56 & 0.57 \\ 0.00 & 6.87 & 0.00 & 0.59 & 92.54 \end{bmatrix}$$

A.1.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix} 94.78 & 5.04 & 0.00 & 0.09 & 0.08 \\ 0.00 & 91.88 & 0.08 & 4.25 & 3.80 \\ 0.00 & 20.89 & 78.36 & 0.40 & 0.35 \\ 0.00 & 6.95 & 0.00 & 92.93 & 0.11 \\ 0.00 & 0.98 & 0.00 & 0.02 & 99.00 \end{bmatrix}$$

A.2 Cluster 2

Indivíduos do sexo feminino, nacionalidade estrangeira, com idade entre os 16 e os 55 anos.

A.2.1 Matriz a 1 Passo

$$\begin{bmatrix} 98.79 & 1.20 & 0.00 & 0.01 & 0.00 \\ 0.00 & 98.27 & 0.02 & 0.89 & 0.82 \\ 0.00 & 0.00 & 94.12 & 5.88 & 0.00 \\ 0.00 & 1.77 & 0.00 & 98.20 & 0.03 \\ 0.00 & 0.13 & 0.00 & 0.00 & 99.87 \end{bmatrix}$$

A.2.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix} 92.65 & 6.70 & 0.00 & 0.57 & 0.07 \\ 0.00 & 92.38 & 0.04 & 3.80 & 3.78 \\ 0.00 & 48.14 & 5.16 & 42.69 & 4.01 \\ 0.00 & 9.18 & 0.44 & 90.02 & 0.36 \\ 0.00 & 6.21 & 0.00 & 0.77 & 93.02 \end{bmatrix}$$

A.2.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix} 94.11 & 5.66 & 0.00 & 0.13 & 0.10 \\ 0.00 & 91.79 & 0.10 & 4.18 & 3.94 \\ 0.00 & 0.95 & 73.85 & 25.18 & 0.02 \\ 0.00 & 8.23 & 0.00 & 91.48 & 0.29 \\ 0.00 & 0.62 & 0.00 & 0.01 & 99.37 \end{bmatrix}$$

A.3 Cluster 3

Indivíduos do sexo feminino, com idade superior a 55 anos.

A.3.1 Matriz a 1 Passo

$$\begin{bmatrix} 98.84 & 1.15 & 0.00 & 0.00 & 0.00 \\ 0.00 & 98.21 & 0.02 & 0.93 & 0.83 \\ 0.00 & 3.32 & 93.53 & 2.34 & 0.81 \\ 0.00 & 1.45 & 0.00 & 98.52 & 0.03 \\ 0.00 & 0.10 & 0.00 & 0.01 & 99.89 \end{bmatrix}$$

A.3.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix} 92.60 & 6.75 & 0.00 & 0.56 & 0.08 \\ 0.00 & 92.26 & 0.05 & 3.88 & 3.81 \\ 0.00 & 50.77 & 4.50 & 40.44 & 4.29 \\ 0.00 & 8.80 & 0.29 & 90.41 & 0.51 \\ 0.00 & 6.11 & 0.03 & 0.76 & 93.11 \end{bmatrix}$$

A.3.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix} 94.35 & 5.43 & 0.00 & 0.12 & 0.10 \\ 0.00 & 91.53 & 0.08 & 4.38 & 4.01 \\ 0.00 & 14.38 & 71.58 & 10.23 & 3.81 \\ 0.00 & 6.81 & 0.00 & 92.95 & 0.24 \\ 0.00 & 0.49 & 0.00 & 0.04 & 99.47 \end{bmatrix}$$

A.4 Cluster 4

Indivíduos do sexo masculino, com idade superior a 55 anos.

A.4.1 Matriz a 1 Passo

$$\begin{bmatrix} 98.85 & 1.14 & 0.00 & 0.00 & 0.00 \\ 0.00 & 98.21 & 0.02 & 0.93 & 0.84 \\ 0.00 & 3.03 & 94.85 & 1.92 & 0.20 \\ 0.00 & 1.44 & 0.00 & 98.54 & 0.02 \\ 0.00 & 0.10 & 0.00 & 0.00 & 99.89 \end{bmatrix}$$

A.4.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix} 92.58 & 6.74 & 0.00 & 0.59 & 0.08 \\ 0.00 & 92.21 & 0.05 & 3.90 & 3.84 \\ 0.00 & 51.41 & 5.12 & 39.13 & 4.34 \\ 0.00 & 8.89 & 0.24 & 90.39 & 0.49 \\ 0.00 & 6.48 & 0.03 & 0.77 & 92.72 \end{bmatrix}$$

A.4.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix} 94.38 & 5.39 & 0.00 & 0.12 & 0.10 \\ 0.00 & 91.52 & 0.07 & 4.38 & 4.04 \\ 0.00 & 13.41 & 76.79 & 8.65 & 1.15 \\ 0.00 & 6.74 & 0.00 & 93.04 & 0.22 \\ 0.00 & 0.50 & 0.00 & 0.03 & 99.47 \end{bmatrix}$$

A.5 Cluster 5

Indivíduos do sexo masculino, nacionalidade portuguesa, com idade entre os 36 e os 55 anos.

A.5.1 Matriz a 1 Passo

$$\begin{bmatrix} 98.83 & 1.17 & 0.00 & 0.00 & 0.00 \\ 0.00 & 98.23 & 0.02 & 0.94 & 0.82 \\ 0.00 & 3.49 & 94.48 & 1.46 & 0.56 \\ 0.00 & 1.50 & 0.00 & 98.48 & 0.01 \\ 0.00 & 0.11 & 0.00 & 0.00 & 99.89 \end{bmatrix}$$

A.5.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix} 92.61 & 6.73 & 0.00 & 0.58 & 0.08 \\ 0.00 & 92.22 & 0.05 & 3.91 & 3.83 \\ 0.00 & 50.89 & 4.74 & 40.03 & 4.33 \\ 0.00 & 8.91 & 0.27 & 90.41 & 0.41 \\ 0.00 & 6.55 & 0.02 & 0.81 & 92.62 \end{bmatrix}$$

A.5.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix} 94.27 & 5.50 & 0.00 & 0.12 & 0.10 \\ 0.00 & 91.58 & 0.07 & 4.39 & 3.95 \\ 0.00 & 15.27 & 75.30 & 6.65 & 2.78 \\ 0.00 & 7.04 & 0.01 & 92.77 & 0.18 \\ 0.00 & 0.54 & 0.00 & 0.02 & 99.44 \end{bmatrix}$$

A.6 Cluster 6

Indivíduos do sexo feminino, com idade inferior a 16 anos.

A.6.1 Matriz a 1 Passo

$$\begin{bmatrix} 98.87 & 1.12 & 0.00 & 0.01 & 0.00 \\ 0.00 & 98.19 & 0.02 & 0.96 & 0.83 \\ 0.00 & 4.57 & 93.29 & 1.83 & 0.30 \\ 0.00 & 1.44 & 0.00 & 98.54 & 0.02 \\ 0.00 & 0.12 & 0.00 & 0.00 & 99.87 \end{bmatrix}$$

A.6.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix} 92.65 & 6.69 & 0.00 & 0.58 & 0.08 \\ 0.00 & 92.26 & 0.05 & 3.89 & 3.81 \\ 0.00 & 50.26 & 4.48 & 41.08 & 4.18 \\ 0.00 & 8.70 & 0.25 & 90.54 & 0.51 \\ 0.00 & 6.64 & 0.02 & 0.77 & 92.56 \end{bmatrix}$$

A.6.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix} 94.48 & 5.28 & 0.00 & 0.14 & 0.10 \\ 0.00 & 91.42 & 0.08 & 4.49 & 4.02 \\ 0.00 & 19.49 & 70.68 & 8.15 & 1.69 \\ 0.00 & 6.74 & 0.00 & 93.02 & 0.23 \\ 0.00 & 0.59 & 0.00 & 0.03 & 99.39 \end{bmatrix}$$

A.7 Cluster 7

Indivíduos do sexo feminino, nacionalidade portuguesa, com idade entre os 16 e os 35 anos.

A.7.1 Matriz a 1 Passo

$$\begin{bmatrix} 98.86 & 1.13 & 0.00 & 0.00 & 0.00 \\ 0.00 & 98.20 & 0.02 & 0.95 & 0.82 \\ 0.00 & 3.62 & 93.48 & 1.74 & 1.16 \\ 0.00 & 1.47 & 0.00 & 98.52 & 0.02 \\ 0.00 & 0.10 & 0.00 & 0.00 & 99.90 \end{bmatrix}$$

A.7.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix} 92.70 & 6.65 & 0.00 & 0.57 & 0.08 \\ 0.00 & 92.29 & 0.05 & 3.88 & 3.79 \\ 0.00 & 51.79 & 4.80 & 39.08 & 4.33 \\ 0.00 & 9.00 & 0.29 & 90.23 & 0.47 \\ 0.00 & 6.35 & 0.03 & 0.83 & 92.79 \end{bmatrix}$$

A.7.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix} 94.44 & 5.35 & 0.00 & 0.12 & 0.10 \\ 0.00 & 91.49 & 0.08 & 4.46 & 3.97 \\ 0.00 & 15.55 & 71.38 & 7.71 & 5.36 \\ 0.00 & 6.86 & 0.00 & 92.94 & 0.19 \\ 0.00 & 0.47 & 0.00 & 0.03 & 99.50 \end{bmatrix}$$

A.8 Cluster 8

Indivíduos do sexo masculino, nacionalidade portuguesa, com idade entre os 16 e os 35 anos.

A.8.1 Matriz a 1 Passo

$$\begin{bmatrix} 98.83 & 1.16 & 0.00 & 0.00 & 0.00 \\ 0.00 & 98.22 & 0.02 & 0.93 & 0.83 \\ 0.00 & 3.24 & 95.07 & 1.39 & 0.31 \\ 0.00 & 1.50 & 0.00 & 98.47 & 0.03 \\ 0.00 & 0.09 & 0.00 & 0.01 & 99.91 \end{bmatrix}$$

A.8.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix} 92.61 & 6.72 & 0.00 & 0.58 & 0.09 \\ 0.00 & 92.31 & 0.04 & 3.87 & 3.78 \\ 0.00 & 50.58 & 4.76 & 40.28 & 4.38 \\ 0.00 & 8.95 & 0.30 & 90.29 & 0.46 \\ 0.00 & 6.33 & 0.04 & 0.83 & 92.79 \end{bmatrix}$$

A.8.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix} 94.29 & 5.48 & 0.00 & 0.12 & 0.10 \\ 0.00 & 91.56 & 0.08 & 4.35 & 4.00 \\ 0.00 & 14.32 & 77.67 & 6.36 & 1.65 \\ 0.00 & 7.01 & 0.01 & 92.73 & 0.26 \\ 0.00 & 0.42 & 0.00 & 0.04 & 99.54 \end{bmatrix}$$

A.9 Cluster 9

Indivíduos do sexo feminino, nacionalidade portuguesa, com idade entre os 36 e os 55 anos.

A.9.1 Matriz a 1 Passo

$$\begin{bmatrix} 98.86 & 1.14 & 0.00 & 0.01 & 0.00 \\ 0.00 & 98.20 & 0.02 & 0.95 & 0.83 \\ 0.00 & 3.23 & 94.77 & 1.67 & 0.33 \\ 0.00 & 1.47 & 0.00 & 98.52 & 0.01 \\ 0.00 & 0.11 & 0.00 & 0.01 & 99.89 \end{bmatrix}$$

A.9.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix} 92.60 & 6.72 & 0.00 & 0.59 & 0.09 \\ 0.00 & 92.24 & 0.05 & 3.90 & 3.81 \\ 0.00 & 51.39 & 4.40 & 39.93 & 4.29 \\ 0.00 & 8.89 & 0.26 & 90.38 & 0.47 \\ 0.00 & 6.53 & 0.03 & 0.79 & 92.64 \end{bmatrix}$$

A.9.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix} 94.42 & 5.35 & 0.00 & 0.13 & 0.09 \\ 0.00 & 91.45 & 0.08 & 4.46 & 4.01 \\ 0.00 & 14.22 & 76.46 & 7.57 & 1.75 \\ 0.00 & 6.88 & 0.00 & 92.94 & 0.18 \\ 0.00 & 0.52 & 0.00 & 0.04 & 99.44 \end{bmatrix}$$

A.10 Cluster 10

Indivíduos do sexo masculino, com idade inferior a 16 anos.

A.10.1 Matriz a 1 Passo

$$\begin{bmatrix} 98.81 & 1.18 & 0.00 & 0.01 & 0.00 \\ 0.00 & 98.24 & 0.01 & 0.92 & 0.83 \\ 0.00 & 5.81 & 92.35 & 1.22 & 0.61 \\ 0.00 & 1.46 & 0.00 & 98.51 & 0.03 \\ 0.00 & 0.10 & 0.00 & 0.00 & 99.90 \end{bmatrix}$$

A.10.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix} 92.61 & 6.71 & 0.00 & 0.59 & 0.08 \\ 0.00 & 92.32 & 0.04 & 3.89 & 3.76 \\ 0.00 & 50.48 & 4.11 & 40.32 & 5.09 \\ 0.00 & 8.78 & 0.23 & 90.52 & 0.47 \\ 0.00 & 6.50 & 0.02 & 0.75 & 92.74 \end{bmatrix}$$

A.10.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix} 94.20 & 5.55 & 0.00 & 0.14 & 0.11 \\ 0.00 & 91.63 & 0.06 & 4.33 & 3.98 \\ 0.00 & 24.18 & 67.20 & 5.57 & 3.06 \\ 0.00 & 6.84 & 0.00 & 92.92 & 0.24 \\ 0.00 & 0.47 & 0.00 & 0.02 & 99.51 \end{bmatrix}$$

Anexo B

Matrizes de Transição do *nível de ensino* por *Cluster*

B.1 Cluster 1

Indivíduos do sexo masculino, nacionalidade estrangeira, menos de 36 anos ou mais de 55 anos.

B.1.1 Matriz a 1 Passo

$$\begin{bmatrix}
89.41 & 4.71 & 3.53 & 2.35 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 92.84 & 3.77 & 3.20 & 0.19 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 89.74 & 8.83 & 1.42 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 97.54 & 2.14 & 0.10 & 0.06 & 0.13 & 0.03 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 98.16 & 0.56 & 0.20 & 0.96 & 0.12 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.20 & 0.00 & 0.80 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.33 & 2.67 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.87 & 1.13 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.49 & 0.51 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.1.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix}
13.57 & 22.14 & 16.43 & 36.43 & 6.96 & 1.07 & 0.54 & 2.68 & 0.18 & 0.00 \\
0.00 & 53.13 & 16.10 & 24.93 & 3.56 & 0.00 & 0.85 & 1.28 & 0.14 & 0.00 \\
0.00 & 0.00 & 38.72 & 48.40 & 11.50 & 0.35 & 0.09 & 0.86 & 0.09 & 0.00 \\
0.00 & 0.00 & 0.00 & 83.21 & 12.92 & 1.32 & 0.35 & 2.05 & 0.15 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 86.70 & 1.87 & 1.30 & 8.60 & 1.48 & 0.04 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 87.25 & 1.47 & 9.80 & 1.47 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 77.84 & 19.89 & 1.70 & 0.57 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 95.35 & 4.36 & 0.29 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.83 & 2.17 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.1.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix}
57.14 & 16.24 & 12.68 & 12.82 & 1.01 & 0.03 & 0.02 & 0.04 & 0.01 & 0.00 \\
0.00 & 68.99 & 13.09 & 15.87 & 1.90 & 0.05 & 0.03 & 0.07 & 0.01 & 0.00 \\
0.00 & 0.00 & 58.21 & 34.07 & 7.20 & 0.15 & 0.08 & 0.24 & 0.05 & 0.00 \\
0.00 & 0.00 & 0.00 & 88.27 & 9.80 & 0.57 & 0.33 & 0.82 & 0.19 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 91.12 & 2.66 & 0.91 & 4.62 & 0.68 & 0.01 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 96.06 & 0.00 & 3.85 & 0.09 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 87.33 & 12.38 & 0.29 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 94.47 & 5.48 & 0.06 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.50 & 2.50 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.2 Cluster 2

Indivíduos do sexo feminino, nacionalidade estrangeira, menos de 36 anos ou mais de 55 anos.

B.2.1 Matriz a 1 Passo

$$\begin{bmatrix}
84.83 & 7.87 & 2.81 & 3.37 & 0.56 & 0.56 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 93.02 & 2.82 & 3.49 & 0.67 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 89.91 & 9.59 & 0.50 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 95.94 & 3.67 & 0.11 & 0.00 & 0.28 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 98.54 & 0.24 & 0.12 & 0.98 & 0.12 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.16 & 0.00 & 2.84 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.32 & 1.79 & 0.89 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.45 & 1.55 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.2.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix}
19.28 & 36.53 & 10.60 & 23.75 & 5.62 & 0.64 & 0.51 & 2.68 & 0.38 & 0.00 \\
0.00 & 58.88 & 17.48 & 19.44 & 2.66 & 0.14 & 0.14 & 1.12 & 0.14 & 0.00 \\
0.00 & 0.00 & 45.04 & 44.37 & 8.17 & 0.50 & 0.25 & 1.67 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 79.33 & 14.92 & 0.82 & 0.37 & 4.24 & 0.20 & 0.12 \\
0.00 & 0.00 & 0.00 & 0.00 & 84.37 & 1.41 & 1.01 & 10.70 & 2.38 & 0.12 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 82.22 & 3.33 & 12.78 & 1.11 & 0.56 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 76.55 & 22.57 & 0.88 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 94.26 & 5.40 & 0.34 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.29 & 0.71 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.2.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix}
43.93 & 24.70 & 9.78 & 15.58 & 3.71 & 1.99 & 0.01 & 0.29 & 0.01 & 0.00 \\
0.00 & 69.64 & 9.87 & 16.08 & 4.16 & 0.05 & 0.01 & 0.17 & 0.01 & 0.00 \\
0.00 & 0.00 & 58.76 & 35.83 & 4.97 & 0.10 & 0.01 & 0.31 & 0.01 & 0.00 \\
0.00 & 0.00 & 0.00 & 81.28 & 16.41 & 0.54 & 0.04 & 1.64 & 0.09 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 92.93 & 1.08 & 0.54 & 4.72 & 0.73 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 86.57 & 0.00 & 13.01 & 0.42 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 87.31 & 8.20 & 4.50 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 92.48 & 7.52 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.3 Cluster 3

Indivíduos do sexo feminino, nacionalidade estrangeira, com idade entre os 36 e os 55 anos.

B.3.1 Matriz a 1 Passo

$$\begin{bmatrix}
82.96 & 8.89 & 5.19 & 2.96 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 93.63 & 3.15 & 2.86 & 0.21 & 0.07 & 0.07 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 93.70 & 5.65 & 0.54 & 0.05 & 0.00 & 0.05 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 98.20 & 1.72 & 0.02 & 0.02 & 0.02 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 99.21 & 0.14 & 0.14 & 0.52 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.71 & 0.69 & 1.38 & 0.00 & 0.23 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.45 & 0.55 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.57 & 0.43 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.07 & 0.93 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.3.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix}
11.50 & 39.63 & 15.30 & 23.20 & 5.44 & 0.31 & 1.13 & 3.29 & 0.21 & 0.00 \\
0.00 & 58.63 & 19.33 & 19.15 & 1.57 & 0.25 & 0.31 & 0.75 & 0.00 & 0.00 \\
0.00 & 0.00 & 60.83 & 33.52 & 4.01 & 0.46 & 0.23 & 0.87 & 0.09 & 0.00 \\
0.00 & 0.00 & 0.00 & 88.08 & 8.95 & 0.38 & 0.44 & 1.94 & 0.19 & 0.03 \\
0.00 & 0.00 & 0.00 & 0.00 & 92.15 & 1.25 & 1.73 & 4.68 & 0.20 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 86.99 & 3.13 & 9.64 & 0.24 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 76.03 & 22.24 & 1.55 & 0.17 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 96.47 & 3.12 & 0.41 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.03 & 1.97 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.3.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix}
39.30 & 27.21 & 17.95 & 14.48 & 0.88 & 0.08 & 0.06 & 0.04 & 0.00 & 0.00 \\
0.00 & 71.95 & 12.12 & 13.67 & 1.56 & 0.33 & 0.33 & 0.05 & 0.00 & 0.00 \\
0.00 & 0.00 & 72.22 & 23.99 & 3.23 & 0.25 & 0.02 & 0.29 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 91.32 & 8.18 & 0.14 & 0.14 & 0.21 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 96.09 & 0.64 & 0.68 & 2.57 & 0.02 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 89.05 & 3.25 & 6.55 & 0.06 & 1.10 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.27 & 2.71 & 0.02 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.85 & 2.11 & 0.04 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 95.44 & 4.56 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.4 Cluster 4

Indivíduos do sexo masculino, nacionalidade estrangeira, com idade entre os 36 e os 55 anos.

B.4.1 Matriz a 1 Passo

$$\begin{bmatrix}
86.67 & 5.00 & 5.00 & 1.67 & 1.67 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 90.54 & 3.95 & 4.66 & 0.56 & 0.00 & 0.00 & 0.28 & 0.00 & 0.00 \\
0.00 & 0.00 & 93.62 & 5.98 & 0.27 & 0.07 & 0.00 & 0.07 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 98.63 & 1.20 & 0.05 & 0.00 & 0.12 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 99.51 & 0.03 & 0.16 & 0.30 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.19 & 0.41 & 0.41 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.99 & 1.57 & 0.45 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.40 & 0.53 & 0.07 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.4.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix}
7.87 & 23.00 & 20.73 & 37.67 & 6.66 & 0.76 & 0.61 & 2.57 & 0.15 & 0.00 \\
0.00 & 43.16 & 24.56 & 28.86 & 1.93 & 0.61 & 0.18 & 0.44 & 0.09 & 0.18 \\
0.00 & 0.00 & 56.12 & 39.55 & 3.02 & 0.49 & 0.16 & 0.60 & 0.05 & 0.00 \\
0.00 & 0.00 & 0.00 & 90.71 & 7.23 & 0.41 & 0.25 & 1.30 & 0.06 & 0.03 \\
0.00 & 0.00 & 0.00 & 0.00 & 94.17 & 0.96 & 1.25 & 3.48 & 0.14 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 90.38 & 2.78 & 6.41 & 0.43 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 77.50 & 20.42 & 1.88 & 0.21 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 95.54 & 3.90 & 0.56 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.34 & 1.66 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.4.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix}
48.89 & 15.42 & 18.01 & 10.52 & 6.88 & 0.04 & 0.02 & 0.21 & 0.00 & 0.00 \\
0.00 & 60.83 & 14.22 & 20.70 & 2.93 & 0.05 & 0.01 & 1.25 & 0.01 & 0.00 \\
0.00 & 0.00 & 71.92 & 25.56 & 1.82 & 0.32 & 0.01 & 0.37 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 93.34 & 5.77 & 0.24 & 0.02 & 0.63 & 0.01 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 97.55 & 0.13 & 0.79 & 1.51 & 0.02 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 95.99 & 1.92 & 2.04 & 0.04 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 90.33 & 7.43 & 2.23 & 0.01 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.05 & 2.62 & 0.33 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.5 Cluster 5

Indivíduos do sexo masculino, nacionalidade portuguesa, com idade inferior a 36 anos.

B.5.1 Matriz a 1 Passo

$$\begin{bmatrix}
95.92 & 2.89 & 0.79 & 0.26 & 0.13 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 91.40 & 4.58 & 3.88 & 0.13 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 91.57 & 7.90 & 0.49 & 0.01 & 0.01 & 0.02 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 97.92 & 1.91 & 0.06 & 0.03 & 0.08 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 97.62 & 0.63 & 0.16 & 1.46 & 0.13 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 96.16 & 0.62 & 2.95 & 0.26 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 89.13 & 9.49 & 1.38 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 96.96 & 2.98 & 0.07 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.81 & 0.19 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.5.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix}
19.34 & 17.78 & 19.66 & 30.41 & 5.47 & 0.34 & 0.61 & 5.70 & 0.58 & 0.11 \\
0.00 & 34.54 & 30.78 & 30.65 & 2.96 & 0.14 & 0.06 & 0.69 & 0.16 & 0.01 \\
0.00 & 0.00 & 40.67 & 48.75 & 9.81 & 0.32 & 0.09 & 0.31 & 0.04 & 0.01 \\
0.00 & 0.00 & 0.00 & 84.35 & 12.12 & 0.90 & 0.24 & 2.22 & 0.17 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 61.82 & 1.98 & 1.47 & 27.20 & 7.43 & 0.09 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 68.86 & 3.00 & 24.80 & 3.27 & 0.07 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 40.92 & 53.66 & 5.17 & 0.25 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 86.67 & 12.86 & 0.46 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.55 & 1.45 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.5.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix}
81.20 & 11.15 & 4.12 & 2.76 & 0.73 & 0.01 & 0.00 & 0.02 & 0.00 & 0.00 \\
0.00 & 63.78 & 16.06 & 18.60 & 1.43 & 0.04 & 0.02 & 0.07 & 0.01 & 0.00 \\
0.00 & 0.00 & 64.37 & 31.90 & 3.28 & 0.13 & 0.07 & 0.23 & 0.02 & 0.00 \\
0.00 & 0.00 & 0.00 & 90.02 & 8.72 & 0.37 & 0.14 & 0.67 & 0.07 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 88.66 & 2.77 & 0.65 & 6.83 & 1.06 & 0.03 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 82.24 & 2.31 & 13.32 & 2.11 & 0.02 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 56.24 & 35.68 & 7.99 & 0.08 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 85.68 & 13.96 & 0.36 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.06 & 0.94 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.6 Cluster 6

Indivíduos do sexo feminino, nacionalidade portuguesa, com idade entre os 36 e os 55 anos.

B.6.1 Matriz a 1 Passo

$$\begin{bmatrix}
96.44 & 2.57 & 0.49 & 0.44 & 0.05 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 96.25 & 2.09 & 1.54 & 0.09 & 0.01 & 0.00 & 0.02 & 0.00 & 0.00 \\
0.00 & 0.00 & 96.60 & 3.20 & 0.16 & 0.01 & 0.01 & 0.02 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 98.94 & 0.98 & 0.02 & 0.02 & 0.04 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 99.47 & 0.11 & 0.05 & 0.35 & 0.02 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.78 & 0.31 & 1.80 & 0.11 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.13 & 2.70 & 0.16 & 0.01 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.38 & 0.58 & 0.04 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.42 & 0.58 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.6.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix}
29.68 & 35.73 & 17.66 & 13.87 & 1.84 & 0.08 & 0.24 & 0.83 & 0.07 & 0.00 \\
0.00 & 63.59 & 21.93 & 13.49 & 0.73 & 0.05 & 0.04 & 0.16 & 0.01 & 0.00 \\
0.00 & 0.00 & 71.22 & 26.73 & 1.72 & 0.09 & 0.05 & 0.18 & 0.01 & 0.00 \\
0.00 & 0.00 & 0.00 & 92.12 & 6.99 & 0.21 & 0.13 & 0.52 & 0.03 & 0.01 \\
0.00 & 0.00 & 0.00 & 0.00 & 90.69 & 1.12 & 1.69 & 6.16 & 0.32 & 0.03 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 78.91 & 4.91 & 15.23 & 0.91 & 0.03 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 69.27 & 29.05 & 1.46 & 0.21 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 95.69 & 3.77 & 0.55 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.06 & 2.94 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.6.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix}
83.43 & 11.08 & 2.63 & 2.54 & 0.31 & 0.00 & 0.00 & 0.01 & 0.00 & 0.00 \\
0.00 & 82.61 & 9.05 & 7.60 & 0.59 & 0.05 & 0.02 & 0.09 & 0.00 & 0.00 \\
0.00 & 0.00 & 84.13 & 14.62 & 1.02 & 0.07 & 0.04 & 0.12 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 94.80 & 4.75 & 0.11 & 0.08 & 0.26 & 0.01 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 97.37 & 0.54 & 0.23 & 1.73 & 0.13 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 89.40 & 1.39 & 8.56 & 0.64 & 0.02 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 86.44 & 12.61 & 0.90 & 0.06 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 96.92 & 2.83 & 0.25 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.12 & 2.88 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.7 Cluster 7

Indivíduos do sexo feminino, nacionalidade portuguesa, com idade inferior a 36 anos.

B.7.1 Matriz a 1 Passo

$$\begin{bmatrix}
97.15 & 1.92 & 0.31 & 0.62 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 94.45 & 3.05 & 2.23 & 0.18 & 0.06 & 0.03 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 92.87 & 6.77 & 0.28 & 0.03 & 0.01 & 0.04 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 97.16 & 2.61 & 0.04 & 0.02 & 0.16 & 0.01 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 97.07 & 0.44 & 0.13 & 2.15 & 0.20 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 94.53 & 0.49 & 4.60 & 0.38 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 84.76 & 14.00 & 1.24 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.57 & 2.38 & 0.06 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.79 & 0.21 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.7.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix}
32.86 & 14.86 & 13.48 & 24.72 & 5.62 & 0.34 & 0.96 & 6.64 & 0.47 & 0.05 \\
0.00 & 46.84 & 25.39 & 22.97 & 2.77 & 0.07 & 0.14 & 1.52 & 0.28 & 0.02 \\
0.00 & 0.00 & 41.77 & 42.99 & 14.19 & 0.30 & 0.08 & 0.61 & 0.04 & 0.00 \\
0.00 & 0.00 & 0.00 & 76.89 & 15.32 & 0.73 & 0.22 & 6.40 & 0.43 & 0.01 \\
0.00 & 0.00 & 0.00 & 0.00 & 53.88 & 1.31 & 1.10 & 34.44 & 9.17 & 0.10 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 63.69 & 3.01 & 28.00 & 5.17 & 0.12 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 35.51 & 59.10 & 5.15 & 0.25 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 90.69 & 8.89 & 0.42 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.57 & 1.43 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.7.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix}
86.56 & 8.10 & 1.76 & 3.34 & 0.20 & 0.01 & 0.01 & 0.02 & 0.00 & 0.00 \\
0.00 & 75.18 & 11.75 & 11.16 & 1.38 & 0.26 & 0.11 & 0.15 & 0.02 & 0.00 \\
0.00 & 0.00 & 69.10 & 27.63 & 2.70 & 0.15 & 0.04 & 0.35 & 0.03 & 0.00 \\
0.00 & 0.00 & 0.00 & 86.57 & 11.61 & 0.28 & 0.10 & 1.28 & 0.16 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 86.20 & 1.85 & 0.48 & 9.98 & 1.46 & 0.03 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 75.48 & 1.59 & 20.12 & 2.78 & 0.03 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 43.75 & 48.84 & 7.32 & 0.09 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 88.42 & 11.26 & 0.32 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.93 & 1.07 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.8 Cluster 8

Indivíduos do sexo masculino, nacionalidade portuguesa, com idade entre os 36 e os 55 anos.

B.8.1 Matriz a 1 Passo

$$\begin{bmatrix}
95.56 & 2.78 & 0.92 & 0.74 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 95.98 & 2.40 & 1.51 & 0.08 & 0.01 & 0.00 & 0.01 & 0.00 & 0.00 \\
0.00 & 0.00 & 96.53 & 3.25 & 0.18 & 0.01 & 0.01 & 0.02 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 99.22 & 0.70 & 0.02 & 0.01 & 0.05 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 99.30 & 0.15 & 0.07 & 0.46 & 0.03 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.53 & 0.41 & 1.97 & 0.08 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 96.52 & 3.28 & 0.17 & 0.03 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.36 & 0.59 & 0.05 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.60 & 0.40 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.8.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix}
24.61 & 37.25 & 21.19 & 15.13 & 1.09 & 0.10 & 0.15 & 0.45 & 0.03 & 0.00 \\
0.00 & 59.62 & 24.64 & 14.75 & 0.76 & 0.07 & 0.03 & 0.12 & 0.01 & 0.00 \\
0.00 & 0.00 & 69.97 & 28.06 & 1.63 & 0.10 & 0.06 & 0.16 & 0.01 & 0.00 \\
0.00 & 0.00 & 0.00 & 93.06 & 5.91 & 0.28 & 0.16 & 0.55 & 0.03 & 0.01 \\
0.00 & 0.00 & 0.00 & 0.00 & 86.89 & 1.59 & 2.34 & 8.61 & 0.50 & 0.06 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 75.96 & 6.30 & 16.64 & 1.03 & 0.07 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 62.24 & 35.32 & 2.21 & 0.23 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 94.88 & 4.52 & 0.61 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.32 & 2.68 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.8.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix}
79.70 & 11.68 & 4.49 & 4.02 & 0.09 & 0.01 & 0.00 & 0.01 & 0.00 & 0.00 \\
0.00 & 81.46 & 10.32 & 7.59 & 0.50 & 0.06 & 0.02 & 0.05 & 0.00 & 0.00 \\
0.00 & 0.00 & 83.83 & 14.92 & 1.04 & 0.04 & 0.05 & 0.11 & 0.01 & 0.00 \\
0.00 & 0.00 & 0.00 & 96.15 & 3.39 & 0.10 & 0.06 & 0.29 & 0.01 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 96.53 & 0.68 & 0.33 & 2.30 & 0.15 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 88.27 & 1.82 & 9.39 & 0.51 & 0.02 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 83.77 & 15.10 & 0.98 & 0.16 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 96.83 & 2.88 & 0.29 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.00 & 2.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.9 Cluster 9

Indivíduos do sexo masculino, nacionalidade portuguesa, com idade superior a 55 anos.

B.9.1 Matriz a 1 Passo

$$\begin{bmatrix}
97.46 & 1.92 & 0.42 & 0.17 & 0.04 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 98.17 & 1.01 & 0.77 & 0.03 & 0.00 & 0.00 & 0.01 & 0.00 & 0.00 \\
0.00 & 0.00 & 97.83 & 2.02 & 0.14 & 0.00 & 0.01 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 99.40 & 0.53 & 0.01 & 0.02 & 0.04 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 99.73 & 0.07 & 0.03 & 0.17 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.83 & 0.23 & 0.94 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.87 & 1.97 & 0.16 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.56 & 0.36 & 0.09 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.72 & 0.28 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.9.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix}
25.98 & 53.17 & 11.91 & 7.90 & 0.72 & 0.07 & 0.04 & 0.19 & 0.01 & 0.00 \\
0.00 & 75.96 & 14.14 & 9.13 & 0.59 & 0.05 & 0.04 & 0.08 & 0.01 & 0.00 \\
0.00 & 0.00 & 75.42 & 22.10 & 2.06 & 0.08 & 0.12 & 0.19 & 0.02 & 0.00 \\
0.00 & 0.00 & 0.00 & 91.32 & 7.59 & 0.25 & 0.26 & 0.51 & 0.04 & 0.03 \\
0.00 & 0.00 & 0.00 & 0.00 & 92.49 & 1.34 & 2.07 & 3.79 & 0.27 & 0.04 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 65.31 & 12.24 & 19.05 & 3.40 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 69.01 & 28.94 & 1.74 & 0.32 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 94.84 & 3.81 & 1.34 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.22 & 2.78 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.9.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix}
87.93 & 8.77 & 2.08 & 1.01 & 0.22 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 91.17 & 4.68 & 3.88 & 0.21 & 0.02 & 0.01 & 0.04 & 0.00 & 0.00 \\
0.00 & 0.00 & 89.61 & 9.55 & 0.76 & 0.02 & 0.03 & 0.03 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 97.03 & 2.63 & 0.05 & 0.09 & 0.20 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 98.66 & 0.34 & 0.14 & 0.85 & 0.01 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 94.28 & 1.10 & 4.58 & 0.04 & 0.01 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 89.78 & 9.37 & 0.83 & 0.02 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.80 & 1.75 & 0.45 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.60 & 1.40 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.10 Cluster 10

Indivíduos do sexo feminino, nacionalidade portuguesa, com idade superior a 55 anos.

B.10.1 Matriz a 1 Passo

$$\begin{bmatrix}
96.80 & 2.46 & 0.37 & 0.37 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 97.88 & 1.09 & 0.94 & 0.06 & 0.01 & 0.00 & 0.01 & 0.00 & 0.00 \\
0.00 & 0.00 & 97.74 & 2.15 & 0.09 & 0.00 & 0.00 & 0.01 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 99.30 & 0.66 & 0.01 & 0.00 & 0.02 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 99.71 & 0.05 & 0.05 & 0.17 & 0.01 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.68 & 0.44 & 0.88 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 98.52 & 1.34 & 0.14 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.58 & 0.29 & 0.14 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 99.20 & 0.80 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.10.2 Matriz a 5 Passos estimada pela LGN

$$\begin{bmatrix}
27.11 & 53.54 & 9.64 & 8.23 & 0.88 & 0.12 & 0.06 & 0.34 & 0.05 & 0.03 \\
0.00 & 78.34 & 12.45 & 8.38 & 0.56 & 0.05 & 0.04 & 0.16 & 0.01 & 0.01 \\
0.00 & 0.00 & 76.96 & 20.85 & 1.71 & 0.10 & 0.09 & 0.27 & 0.01 & 0.01 \\
0.00 & 0.00 & 0.00 & 92.44 & 6.59 & 0.17 & 0.21 & 0.56 & 0.02 & 0.01 \\
0.00 & 0.00 & 0.00 & 0.00 & 93.80 & 0.97 & 1.82 & 3.19 & 0.17 & 0.05 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 71.93 & 5.26 & 22.81 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 77.83 & 20.99 & 0.80 & 0.38 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 95.53 & 3.28 & 1.19 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 95.99 & 4.01 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

B.10.3 Matriz a 5 Passos estimada pela propriedade de Markov

$$\begin{bmatrix}
84.99 & 11.04 & 1.91 & 2.01 & 0.04 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 89.84 & 5.01 & 4.68 & 0.37 & 0.03 & 0.01 & 0.07 & 0.00 & 0.00 \\
0.00 & 0.00 & 89.22 & 10.12 & 0.57 & 0.02 & 0.02 & 0.05 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 96.55 & 3.26 & 0.05 & 0.02 & 0.12 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 98.58 & 0.25 & 0.25 & 0.87 & 0.05 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 93.56 & 2.08 & 4.31 & 0.03 & 0.01 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 92.81 & 6.46 & 0.70 & 0.03 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 97.92 & 1.39 & 0.69 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 96.06 & 3.94 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 100.00
\end{bmatrix}$$

Anexo C

Código Stata

```
// Limpar workspace

cap clear all

cap log close

// definir worksplace

cd "C:\Users\******\******"

// abrir ficheiro de registo

log using "Estagio_log.smcl", replace


// desliga a necessidade de pressionar numa tecla para o display continuar

set more off

// mostrar tempo que demora a correr comandos

set rmsg on

// importar dados

import delimited DADOS_ESTAGIO.csv, clear

// Sumário das variáveis

codebook, compact

// Fase 1 - Estado Civil

Preserve // Comando que guarda base de dados atual até ponto de restauro
```

```
// retirar NULLS e Desconhecidos

drop if mi(est_civ_2015)

drop if est_civ_2015 == 6
drop if est_civ_2015 == 9
drop if est_civ_2016 == 9


// Apagar variaveis não relevantes para a analise do EC

drop ind_bpr2016
drop nec_2011
drop nec_2015
drop nec_2016

// construir variavel de NACIONALIDADE

gen Nacionalidade = cond(nac_censos == 10 , 1, 2)

drop nac_censos


// Clusterizar Idade

gen Idade_2016_Clust = cond(idade_2016 < 16, 1,
cond( idade_2016 < 36, 2,
cond(idade_2016 < 56, 3, 4 )))

drop idade_2016

// Labelizar variáveis criadas

label values Nacionalidade Label_Nacionalidade
label variable Nacionalidade "Nacionalidade"
label values Idade_2016_Clust Label_Idade
label variable Idade_2016_Clust "Idade"


// Sumário e contagem de todas as variaveis

codebook, compact


// Encontrar número otimo de clusters (ELBOW METHOD)

local list1 "sexo_censos Idade_2016_Clust Nacionalidade"

foreach v of varlist ‘list1’ {
egen z_‘v’= std(‘v’)
}
```

```
// clusterizar para 4,5,...,15 clusters

local list2 "sexo_censos Idade_2016_Clust Nacionalidade"

forvalues k = 4(1)15 {
cluster kmeans `list2', k(`k') start(random(123)) name(cs`k')
}



matrix WSS = J(15,5,.)
matrix colnames WSS = k WSS log(WSS) eta-squared PRE

// WSS para cada cluster

local list2 "sexo_censos Idade_2016_Clust Nacionalidade"

forvalues k = 4(1)15 {
scalar ws`k' = 0
foreach v of varlist `list2' {
quietly anova `v' cs`k'
scalar ws`k' = ws`k' + e(rss)
}
matrix WSS[`k', 1] = `k'
matrix WSS[`k', 2] = ws`k'
matrix WSS[`k', 3] = log(ws`k')
matrix WSS[`k', 4] = 1 - ws`k'/WSS[4,2]
matrix WSS[`k', 5] = (WSS[`k'-1,2] - ws`k')/WSS[`k'-1,2]
}


matrix list WSS


local squared = char(178)

_matplot WSS, columns(2 1) connect(l) xlabel(#10) name(plot1, replace) nodraw noname



gr di plot1

graph export "ELBOWEC.png", as(png) replace




// guardar dados para clusterizar em R
```

```
export delimited sexo_censos Nacionalidade Idade_2016_Clust using "DadosEstagioEC.csv
    ", replace


// importar clusters do R

gen ID = [_n]

merge 1:1 ID using "DadosECCluster.dta", keepusing(cluster_clara_EC) nogenerate

label variable cluster_clara_EC "Cluster_EC"



// corrigir regresso ao estado solteiro

replace est_civ_2015 = est_civ_2011 if est_civ_2015 == 1 & est_civ_2011 != 1

replace est_civ_2016 = est_civ_2015 if est_civ_2016 == 1 & est_civ_2015 != 1


// Construir variaveis de transição

egen TEst_Civ_11_15 = concat(est_civ_2011 est_civ_2015), punct(-)
egen TEst_Civ_11_16 = concat(est_civ_2011 est_civ_2016), punct(-)
egen TEst_Civ_15_16 = concat(est_civ_2015 est_civ_2016), punct(-)


// CONSTRUIR MATRIZES


// MATRIZ PI_2011

tab est_civ_2011, matcell(Pi_2011)

matrix Pi_2011 = Pi_2011 / cond(r(N)==0,1,r(N))

matrix list Pi_2011

// MATRIZ PI_2015

tab est_civ_2015,matcell(Pi_2015)

matrix Pi_2015 = Pi_2015 / cond(r(N)==0,1,r(N))

matrix list Pi_2015
```

```
// CONSTRUÇÃO DAS MATRIZES DE TRANSIÇAO


// MATRIZ DE TRANSIÇAO 2011_2016

local t1 = "1-1 1-2 1-3 1-4 1-5"
local t2 = "2-1 2-2 2-3 2-4 2-5"
local t3 = "3-1 3-2 3-3 3-4 3-5"
local t4 = "4-1 4-2 4-3 4-4 4-5"
local t5 = "5-1 5-2 5-3 5-4 5-5"

tabcount TEst_Civ_11_16 , c1(`t1') zero matrix(x1_2011 )
tab TEst_Civ_11_16 if est_civ_2011 == 1
matrix P_2011 = x1_2011'/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_11_16 , c1(`t2') zero matrix(x2_2011 )
tab TEst_Civ_11_16 if est_civ_2011 == 2
matrix P_2011 = P_2011 \ x2_2011'/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_11_16 , c1(`t3') zero matrix(x3_2011 )
tab TEst_Civ_11_16 if est_civ_2011 == 3
matrix P_2011 = P_2011 \ x3_2011'/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_11_16 , c1(`t4') zero matrix(x4_2011 )
tab TEst_Civ_11_16 if est_civ_2011 == 4
matrix P_2011 = P_2011 \ x4_2011'/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_11_16 , c1(`t5') zero matrix(x5_2011 )
tab TEst_Civ_11_16 if est_civ_2011 == 5
matrix P_2011 = P_2011 \ x5_2011'/ cond(r(N)==0,1,r(N))


// 2015 2016


tabcount TEst_Civ_15_16 , c1(`t1') zero matrix(x1_2015 )
tab TEst_Civ_15_16 if est_civ_2015 == 1
matrix P_2015 = x1_2015'/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_15_16 , c1(`t2') zero matrix(x2_2015 )
tab TEst_Civ_15_16 if est_civ_2015 == 2
matrix P_2015 = P_2015 \ x2_2015'/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_15_16 , c1(`t3') zero matrix(x3_2015 )
tab TEst_Civ_15_16 if est_civ_2015 == 3
matrix P_2015 = P_2015 \ x3_2015'/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_15_16 , c1(`t4') zero matrix(x4_2015 )
tab TEst_Civ_15_16 if est_civ_2015 == 4
matrix P_2015 = P_2015 \ x4_2015'/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_15_16 , c1(`t5') zero matrix(x5_2015 )
```

```
tab TEst_Civ_15_16 if est_civ_2015 == 5
matrix P_2015 = P_2015 \ x5_2015'/ cond(r(N)==0,1,r(N))



// LISTAR MATRIZES CONSTRUIDAS

// MATRIZ PI_2011 E pi_2015

matrix results = Pi_2011 * 100

matrix list Pi_2011

outtable using "MatrizPI2011EC.xlsx", mat(results) replace format(%9.2f)

matrix results = Pi_2015 * 100

matrix list Pi_2015

outtable using "MatrizPI2015EC.xlsx", mat(results) replace format(%9.2f)

// MATRIZ DE TRANSIÇAO A 1 PASSO

matrix results = P_2015 * 100

matrix list P_2015

outtable using "MatrizP2015EC.xlsx", mat(results) replace format(%9.2f)



// PARA CADA CLUSTER

forvalues i = 1(1)10{

display "Cluster `i'"

// MATRIZ PI_2011


tabcount est_civ_2011 if cluster_clara_EC == `i', v1(1/5) zero matrix(Pi_2011_`i')
tab est_civ_2011 if cluster_clara_EC == `i'
matrix Pi_2011_`i' = Pi_2011_`i' / cond(r(N)==0,1,r(N))

// MATRIZ PI_2011


tabcount est_civ_2015 if cluster_clara_EC == `i', v1(1/5) zero matrix(Pi_2015_`i')
tab est_civ_2015 if cluster_clara_EC == `i'
matrix Pi_2015_`i' = Pi_2015_`i' / cond(r(N)==0,1,r(N))
```

```
// PARA CADA CATEGORIA

// MATRIZ DE TRANSIÇAO 2011_2016

local t1 = "1-1 1-2 1-3 1-4 1-5"
local t2 = "2-1 2-2 2-3 2-4 2-5"
local t3 = "3-1 3-2 3-3 3-4 3-5"
local t4 = "4-1 4-2 4-3 4-4 4-5"
local t5 = "5-1 5-2 5-3 5-4 5-5"

tabcount TEst_Civ_11_16 if cluster_clara_EC == `i', c1(`t1') zero matrix(x1_2011_`i')
tab TEst_Civ_11_16 if est_civ_2011 == 1 & cluster_clara_EC == `i'
matrix P_2011_`i' = x1_2011_`i''/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_11_16 if cluster_clara_EC == `i', c1(`t2') zero matrix(x2_2011_`i')
tab TEst_Civ_11_16 if est_civ_2011 == 2 & cluster_clara_EC == `i'
matrix P_2011_`i' = P_2011_`i' \ x2_2011_`i''/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_11_16 if cluster_clara_EC == `i', c1(`t3') zero matrix(x3_2011_`i')
tab TEst_Civ_11_16 if est_civ_2011 == 3 & cluster_clara_EC == `i'
matrix P_2011_`i' = P_2011_`i' \ x3_2011_`i''/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_11_16 if cluster_clara_EC == `i', c1(`t4') zero matrix(x4_2011_`i')
tab TEst_Civ_11_16 if est_civ_2011 == 4 & cluster_clara_EC == `i'
matrix P_2011_`i' = P_2011_`i' \ x4_2011_`i''/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_11_16 if cluster_clara_EC == `i', c1(`t5') zero matrix(x5_2011_`i')
tab TEst_Civ_11_16 if est_civ_2011 == 5 & cluster_clara_EC == `i'
matrix P_2011_`i' = P_2011_`i' \ x5_2011_`i''/ cond(r(N)==0,1,r(N))


// 2015 2016


tabcount TEst_Civ_15_16 if cluster_clara_EC == `i', c1(`t1') zero matrix(x1_2015_`i')
tab TEst_Civ_15_16 if est_civ_2015 == 1 & cluster_clara_EC == `i'
matrix P_2015_`i' = x1_2015_`i''/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_15_16 if cluster_clara_EC == `i', c1(`t2') zero matrix(x2_2015_`i')
tab TEst_Civ_15_16 if est_civ_2015 == 2 & cluster_clara_EC == `i'
matrix P_2015_`i' = P_2015_`i' \ x2_2015_`i''/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_15_16 if cluster_clara_EC == `i', c1(`t3') zero matrix(x3_2015_`i')
tab TEst_Civ_15_16 if est_civ_2015 == 3 & cluster_clara_EC == `i'
matrix P_2015_`i' = P_2015_`i' \ x3_2015_`i''/ cond(r(N)==0,1,r(N))

tabcount TEst_Civ_15_16 if cluster_clara_EC == `i', c1(`t4') zero matrix(x4_2015_`i')
tab TEst_Civ_15_16 if est_civ_2015 == 4 & cluster_clara_EC == `i'
matrix P_2015_`i' = P_2015_`i' \ x4_2015_`i''/ cond(r(N)==0,1,r(N))
```

```
tabcount TEst_Civ_15_16 if cluster_clara_EC == `i', c1(`t5') zero matrix(x5_2015_`i')
tab TEst_Civ_15_16 if est_civ_2015 == 5 & cluster_clara_EC == `i'
matrix P_2015_`i' = P_2015_`i' \ x5_2015_`i''/ cond(r(N)==0,1,r(N))


}



// LISTAR MATRIZES DE TRANSIÇAO CLUSTERS

forvalues i = 1/10 {
matrix results = P_2015_`i' * 100
outtable using "Matriz2015EC1passo_`i'.xlsx", mat(results) replace format(%9.2f)
}



// MATRIZES DE TRANSIÇAO A 5 PASSOS

// Definir programa para elevar matrizes

cap program drop Elevar_Matriz

program define Elevar_Matriz
args n Q
matrix `Q'`n' = `Q'
forvalues i = 2/`n' {
matrix `Q'`n' = `Q'`n' * `Q'
}
end


// MATRIZ A 5 PASSOS PELA PROP DE MARKOV

Elevar_Matriz 5 P_2015

matrix results = P_20155*100

outtable using "Matriz2015EC5passos", mat(results) replace format(%9.2f)



// PARA CADA CLUSTER

forvalues i = 1/10 {
Elevar_Matriz 5 P_2015_`i'
matrix results = P_2015_`i'5 * 100
outtable using "Matriz2015NEC5passo_`i'.xlsx", mat(results) replace format(%9.2f)
}
```

```
// MATRIZ A 5 PASSOS ESTIMADA

matrix list P_2011

matrix results = P_2011*100

outtable using "Matriz2011EC5passosL", mat(results) replace format(%9.2f)


// PARA CADA CLUSTER

forvalues i = 1/10 {
matrix results = p_2011_‘i’ * 100
outtable using "Matriz2015EC5passoL_‘i’.xlsx", mat(results) replace format(%9.2f)
}




// CONSTRUÇAO INTERVALOS DE CONFIANÇA

// definir programa para calculo da raiz da matriz

cap program drop Raiz_Matriz


program define Raiz_Matriz
args n Q
forvalues i = 1(1)‘n’ {
forvalues j = 1(1)‘n’ {
matrix ‘Q’_raiz[‘i’,‘j’] = sqrt( 2 * ‘Q’[‘i’,‘j’] /_N )
}
}
end


// CHAMAR FUNÇAO RAIZ * 2 / N

matrix P_2011_raiz = P_2011

Raiz_Matriz 5 P_2011


matrix P_2011_inf = P_2011 - 1.96 * P_2011_raiz

matrix P_2011_sup = P_2011 + 1.96 * P_2011_raiz


// MATRIZ COM LIMITES INFERIORES

matrix list P_2011_inf
```

```
matrix results = P_2011_inf*100

outtable using "Matriz2011EC_IC_I", mat(results) replace format(%9.2f)

// MATRIZ COM LIMITES SUPERIORES

matrix list P_2011_sup


matrix results = P_2011_sup*100

outtable using "Matriz2011EC_IC_S", mat(results) replace format(%9.2f)









// restaurar base de dados para processar Nivel de ensino

restore


// retirar NULLS e Desconhecidos

drop if mi(nec_2011)
drop if mi(nec_2015)
drop if mi(nec_2016)

// Apagar variáveis não relevantes para o NEC

drop ind_bpr2016
drop est_civ_2011 est_civ_2015 est_civ_2016


// construir variavel de NACIONALIDADE


gen Nacionalidade = cond(nac_censos == 10 , 1, 2)

drop nac_censos

// Clusterizar Idade

gen Idade_2016_Clust = cond(idade_2016 < 16, 1,
cond( idade_2016 < 36, 2,
cond(idade_2016 < 56, 3, 4 )))
```

```
// Apagar variavel transformada

drop idade_2016

// Labelizar variáveis criadas

label values Nacionalidade Label_Nacionalidade
label variable Nacionalidade "Nacionalidade"
label values Idade_2016_Clust Label_Idade
label variable Idade_2016_Clust "Idade"

// Sumário e contagem de todas as variaveis

codebook, compact



// Encontrar número otimo de clusters (ELBOW METHOD)

local list1 "sexo_censos Idade_2016_Clust Nacionalidade"

foreach v of varlist `list1' {
egen z_`v'= std(`v')
}

// clusterizar para 4,5,...,15 clusters

local list2 "sexo_censos Idade_2016_Clust Nacionalidade"

forvalues k = 4(1)15 {
cluster kmeans `list2', k(`k') start(random(123)) name(cs`k')
}

matrix WSS = J(15,5,.)
matrix colnames WSS = k WSS log(WSS) eta-squared PRE

// WSS para cada cluster

local list2 "sexo_censos Idade_2016_Clust Nacionalidade"

forvalues k = 4(1)15 {
scalar ws`k' = 0
foreach v of varlist `list2' {
quietly anova `v' cs`k'
scalar ws`k' = ws`k' + e(rss)
}
matrix WSS[`k', 1] = `k'
matrix WSS[`k', 2] = ws`k'
matrix WSS[`k', 3] = log(ws`k')
matrix WSS[`k', 4] = 1 - ws`k'/WSS[4,2]
matrix WSS[`k', 5] = (WSS[`k'-1,2] - ws`k')/WSS[`k'-1,2]
```

```
}


matrix list WSS


local squared = char(178)

_matplot WSS, columns(2 1) connect(l) xlabel(#10) name(plot1, replace) nodraw noname

gr di plot1

// Exportar gráfico

graph export "ELBOWNEC.png", as(png) replace

// define-se um número de 10 clusters



// guardar dados para clusterizar em R


export delimited sexo_censos Nacionalidade Idade_2016_Clust using "DadosEstagioNEC.csv
    ", replace


// importar clusters do R

gen ID = [_n] // gerar variável para combinar tabelas

merge 1:1 ID using "DadosNECCluster.dta", keepusing(cluster_clara_NEC) nogenerate

// Labelizar variável

label variable cluster_clara_NEC "Cluster_NEC"


// corrigir descidas de graduação

replace nec_2015 = nec_2011 if nec_2015 < nec_2011
replace nec_2016 = nec_2015 if nec_2016 < nec_2015

// Construir variaveis de transição

egen TNiv_Ens_Comp_11_15 = concat(nec_2011 nec_2015), punct(-)
egen TNiv_Ens_Comp_11_16 = concat(nec_2011 nec_2016), punct(-)
egen TNiv_Ens_Comp_15_16 = concat(nec_2015 nec_2016), punct(-)
```

```
// CONSTRUIR MATRIZES de DISTRIBUIÇÃO

// MATRIZ PI_2011


tabcount nec_2011, v1(1/10) zero matrix(Pi_2011)
tab nec_2011
matrix Pi_2011 = Pi_2011 / cond(r(N)==0,1,r(N))

// MATRIZ PI_2015


tabcount nec_2015 , v1(1/10) zero matrix(Pi_2015)
tab nec_2015
matrix Pi_2015 = Pi_2015 / cond(r(N)==0,1,r(N))




// CONSTRUÇÃO DAS MATRIZES DE TRANSIÇAO


// MATRIZ DE TRANSIÇAO 2011_2016

local t1 = "1-1 1-2 1-3 1-4 1-5 1-6 1-7 1-8 1-9 1-10"
local t2 = "2-1 2-2 2-3 2-4 2-5 2-6 2-7 2-8 2-9 2-10"
local t3 = "3-1 3-2 3-3 3-4 3-5 3-6 3-7 3-8 3-9 3-10"
local t4 = "4-1 4-2 4-3 4-4 4-5 4-6 4-7 4-8 4-9 4-10"
local t5 = "5-1 5-2 5-3 5-4 5-5 5-6 5-7 5-8 5-9 5-10"
local t6 = "6-1 6-2 6-3 6-4 6-5 6-6 6-7 6-8 6-9 6-10"
local t7 = "7-1 7-2 7-3 7-4 7-5 7-6 7-7 7-8 7-9 7-10"
local t8 = "8-1 8-2 8-3 8-4 8-5 8-6 8-7 8-8 8-9 8-10"
local t9 = "9-1 9-2 9-3 9-4 9-5 9-6 9-7 9-8 9-9 9-10"
local t10 = "10-1 10-2 10-3 10-4 10-5 10-6 10-7 10-8 10-9 10-10"



tabcount TNiv_Ens_Comp_11_16 , c1(‘t1’) zero matrix(x1_2011 )
tab TNiv_Ens_Comp_11_16 if nec_2011 == 1
matrix P_2011 = x1_2011 ’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 , c1(‘t2’) zero matrix(x2_2011 )
tab TNiv_Ens_Comp_11_16 if nec_2011 == 2
matrix P_2011 = P_2011 \ x2_2011’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 , c1(‘t3’) zero matrix(x3_2011 )
tab TNiv_Ens_Comp_11_16 if nec_2011 == 3
matrix P_2011 = P_2011 \ x3_2011’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 , c1(‘t4’) zero matrix(x4_2011 )
tab TNiv_Ens_Comp_11_16 if nec_2011 == 4
matrix P_2011 = P_2011 \ x4_2011’/ cond(r(N)==0,1,r(N))
```

```
tabcount TNiv_Ens_Comp_11_16 , c1(‘t5’) zero matrix(x5_2011 )
tab TNiv_Ens_Comp_11_16 if nec_2011 == 5
matrix P_2011 = P_2011 \ x5_2011’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 , c1(‘t6’) zero matrix(x6_2011 )
tab TNiv_Ens_Comp_11_16 if nec_2011 == 6
matrix P_2011 = P_2011 \ x6_2011’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 , c1(‘t7’) zero matrix(x7_2011 )
tab TNiv_Ens_Comp_11_16 if nec_2011 == 7
matrix P_2011 = P_2011 \ x7_2011’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 , c1(‘t8’) zero matrix(x8_2011 )
tab TNiv_Ens_Comp_11_16 if nec_2011 == 8
matrix P_2011 = P_2011 \ x8_2011’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 , c1(‘t9’) zero matrix(x9_2011 )
tab TNiv_Ens_Comp_11_16 if nec_2011 == 9
matrix P_2011 = P_2011 \ x9_2011’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 , c1(‘t10’) zero matrix(x10_2011 )
tab TNiv_Ens_Comp_11_16 if nec_2011 == 10
matrix P_2011 = P_2011 \ x10_2011’/ cond(r(N)==0,1,r(N))

// 2015 2016


tabcount TNiv_Ens_Comp_15_16 , c1(‘t1’) zero matrix(x1_2015 )
tab TNiv_Ens_Comp_15_16 if nec_2015 == 1
matrix P_2015 = x1_2015’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 , c1(‘t2’) zero matrix(x2_2015 )
tab TNiv_Ens_Comp_15_16 if nec_2015 == 2
matrix P_2015 = P_2015 \ x2_2015’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 , c1(‘t3’) zero matrix(x3_2015 )
tab TNiv_Ens_Comp_15_16 if nec_2015 == 3
matrix P_2015 = P_2015 \ x3_2015’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 , c1(‘t4’) zero matrix(x4_2015 )
tab TNiv_Ens_Comp_15_16 if nec_2015 == 4
matrix P_2015 = P_2015 \ x4_2015’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 , c1(‘t5’) zero matrix(x5_2015 )
tab TNiv_Ens_Comp_15_16 if nec_2015 == 5
matrix P_2015 = P_2015 \ x5_2015’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 , c1(‘t6’) zero matrix(x6_2015 )
tab TNiv_Ens_Comp_15_16 if nec_2015 == 6
matrix P_2015 = P_2015 \ x6_2015’/ cond(r(N)==0,1,r(N))
```

```
tabcount TNiv_Ens_Comp_15_16 , c1(‘t7’) zero matrix(x7_2015 )
tab TNiv_Ens_Comp_15_16 if nec_2015 == 7
matrix P_2015 = P_2015 \ x7_2015’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 , c1(‘t8’) zero matrix(x8_2015 )
tab TNiv_Ens_Comp_15_16 if nec_2015 == 8
matrix P_2015 = P_2015 \ x8_2015’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 , c1(‘t9’) zero matrix(x9_2015 )
tab TNiv_Ens_Comp_15_16 if nec_2015 == 9
matrix P_2015 = P_2015 \ x9_2015’/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 , c1(‘t10’) zero matrix(x10_2015 )
tab TNiv_Ens_Comp_15_16 if nec_2015 == 10
matrix P_2015 = P_2015 \ x10_2015’/ cond(r(N)==0,1,r(N))


// LISTAR MATRIZES CONSTRUIDAS

// MATRIZ PI_2011 E pi_2015

matrix results = Pi_2011 * 100

matrix list Pi_2011

outtable using "MatrizPI2011NEC.xlsx", mat(results) replace format(%9.2f)

matrix results = Pi_2015 * 100

matrix list Pi_2015

outtable using "MatrizPI2015NEC.xlsx", mat(results) replace format(%9.2f)

// MATRIZ DE TRANSIÇAO A 1 PASSO

matrix results = P_2015 * 100

matrix list P_2015

outtable using "MatrizP2015NEC.xlsx", mat(results) replace format(%9.2f)



// PARA CADA CLUSTER

forvalues i = 1(1)10{

display "Cluster ‘i’"

// MATRIZ PI_2011
```

```
tabcount nec_2011 if cluster_clara_NEC == `i', v1(1/10) zero matrix(Pi_2011_`i')
tab nec_2011 if cluster_clara_NEC == `i'
matrix Pi_2011_`i' = Pi_2011_`i' / cond(r(N)==0,1,r(N))

// MATRIZ PI_2011


tabcount nec_2015 if cluster_clara_NEC == `i', v1(1/10) zero matrix(Pi_2015_`i')
tab nec_2015 if cluster_clara_NEC == `i'
matrix Pi_2015_`i' = Pi_2015_`i' / cond(r(N)==0,1,r(N))


// PARA CADA CATEGORIA

// MATRIZ DE TRANSIÇAO 2011_2016

local t1 = "1-1 1-2 1-3 1-4 1-5 1-6 1-7 1-8 1-9 1-10"
local t2 = "2-1 2-2 2-3 2-4 2-5 2-6 2-7 2-8 2-9 2-10"
local t3 = "3-1 3-2 3-3 3-4 3-5 3-6 3-7 3-8 3-9 3-10"
local t4 = "4-1 4-2 4-3 4-4 4-5 4-6 4-7 4-8 4-9 4-10"
local t5 = "5-1 5-2 5-3 5-4 5-5 5-6 5-7 5-8 5-9 5-10"
local t6 = "6-1 6-2 6-3 6-4 6-5 6-6 6-7 6-8 6-9 6-10"
local t7 = "7-1 7-2 7-3 7-4 7-5 7-6 7-7 7-8 7-9 7-10"
local t8 = "8-1 8-2 8-3 8-4 8-5 8-6 8-7 8-8 8-9 8-10"
local t9 = "9-1 9-2 9-3 9-4 9-5 9-6 9-7 9-8 9-9 9-10"
local t10 = "10-1 10-2 10-3 10-4 10-5 10-6 10-7 10-8 10-9 10-10"



tabcount TNiv_Ens_Comp_11_16 if cluster_clara_NEC == `i', c1(`t1') zero matrix(
    x1_2011_`i')
tab TNiv_Ens_Comp_11_16 if nec_2011 == 1 & cluster_clara_NEC == `i'
matrix P_2011_`i' = x1_2011_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 if cluster_clara_NEC == `i', c1(`t2') zero matrix(
    x2_2011_`i')
tab TNiv_Ens_Comp_11_16 if nec_2011 == 2 & cluster_clara_NEC == `i'
matrix P_2011_`i' = P_2011_`i' \ x2_2011_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 if cluster_clara_NEC == `i', c1(`t3') zero matrix(
    x3_2011_`i')
tab TNiv_Ens_Comp_11_16 if nec_2011 == 3 & cluster_clara_NEC == `i'
matrix P_2011_`i' = P_2011_`i' \ x3_2011_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 if cluster_clara_NEC == `i', c1(`t4') zero matrix(
    x4_2011_`i')
tab TNiv_Ens_Comp_11_16 if nec_2011 == 4 & cluster_clara_NEC == `i'
matrix P_2011_`i' = P_2011_`i' \ x4_2011_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 if cluster_clara_NEC == `i', c1(`t5') zero matrix(
    x5_2011_`i')
```

```
tab TNiv_Ens_Comp_11_16 if nec_2011 == 5 & cluster_clara_NEC == `i'
matrix P_2011_`i' = P_2011_`i' \ x5_2011_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 if cluster_clara_NEC == `i', c1(`t6') zero matrix(
    x6_2011_`i')
tab TNiv_Ens_Comp_11_16 if nec_2011 == 6 & cluster_clara_NEC == `i'
matrix P_2011_`i' = P_2011_`i' \ x6_2011_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 if cluster_clara_NEC == `i', c1(`t7') zero matrix(
    x7_2011_`i')
tab TNiv_Ens_Comp_11_16 if nec_2011 == 7 & cluster_clara_NEC == `i'
matrix P_2011_`i' = P_2011_`i' \ x7_2011_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 if cluster_clara_NEC == `i', c1(`t8') zero matrix(
    x8_2011_`i')
tab TNiv_Ens_Comp_11_16 if nec_2011 == 8 & cluster_clara_NEC == `i'
matrix P_2011_`i' = P_2011_`i' \ x8_2011_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 if cluster_clara_NEC == `i', c1(`t9') zero matrix(
    x9_2011_`i')
tab TNiv_Ens_Comp_11_16 if nec_2011 == 9 & cluster_clara_NEC == `i'
matrix P_2011_`i' = P_2011_`i' \ x9_2011_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_11_16 if cluster_clara_NEC == `i', c1(`t10') zero matrix(
    x10_2011_`i')
tab TNiv_Ens_Comp_11_16 if nec_2011 == 10 & cluster_clara_NEC == `i'
matrix P_2011_`i' = P_2011_`i' \ x10_2011_`i''/ cond(r(N)==0,1,r(N))

// 2015 2016


tabcount TNiv_Ens_Comp_15_16 if cluster_clara_NEC == `i', c1(`t1') zero matrix(
    x1_2015_`i')
tab TNiv_Ens_Comp_15_16 if nec_2015 == 1 & cluster_clara_NEC == `i'
matrix P_2015_`i' = x1_2015_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 if cluster_clara_NEC == `i', c1(`t2') zero matrix(
    x2_2015_`i')
tab TNiv_Ens_Comp_15_16 if nec_2015 == 2 & cluster_clara_NEC == `i'
matrix P_2015_`i' = P_2015_`i' \ x2_2015_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 if cluster_clara_NEC == `i', c1(`t3') zero matrix(
    x3_2015_`i')
tab TNiv_Ens_Comp_15_16 if nec_2015 == 3 & cluster_clara_NEC == `i'
matrix P_2015_`i' = P_2015_`i' \ x3_2015_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 if cluster_clara_NEC == `i', c1(`t4') zero matrix(
    x4_2015_`i')
tab TNiv_Ens_Comp_15_16 if nec_2015 == 4 & cluster_clara_NEC == `i'
matrix P_2015_`i' = P_2015_`i' \ x4_2015_`i''/ cond(r(N)==0,1,r(N))
```

```
tabcount TNiv_Ens_Comp_15_16 if cluster_clara_NEC == `i', c1(`t5') zero matrix(
    x5_2015_`i')
tab TNiv_Ens_Comp_15_16 if nec_2015 == 5 & cluster_clara_NEC == `i'
matrix P_2015_`i' = P_2015_`i' \ x5_2015_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 if cluster_clara_NEC == `i', c1(`t6') zero matrix(
    x6_2015_`i')
tab TNiv_Ens_Comp_15_16 if nec_2015 == 6 & cluster_clara_NEC == `i'
matrix P_2015_`i' = P_2015_`i' \ x6_2015_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 if cluster_clara_NEC == `i', c1(`t7') zero matrix(
    x7_2015_`i')
tab TNiv_Ens_Comp_15_16 if nec_2015 == 7 & cluster_clara_NEC == `i'
matrix P_2015_`i' = P_2015_`i' \ x7_2015_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 if cluster_clara_NEC == `i', c1(`t8') zero matrix(
    x8_2015_`i')
tab TNiv_Ens_Comp_15_16 if nec_2015 == 8 & cluster_clara_NEC == `i'
matrix P_2015_`i' = P_2015_`i' \ x8_2015_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 if cluster_clara_NEC == `i', c1(`t9') zero matrix(
    x9_2015_`i')
tab TNiv_Ens_Comp_15_16 if nec_2015 == 9 & cluster_clara_NEC == `i'
matrix P_2015_`i' = P_2015_`i' \ x9_2015_`i''/ cond(r(N)==0,1,r(N))

tabcount TNiv_Ens_Comp_15_16 if cluster_clara_NEC == `i', c1(`t10') zero matrix(
    x10_2015_`i')
tab TNiv_Ens_Comp_15_16 if nec_2015 == 10 & cluster_clara_NEC == `i'
matrix P_2015_`i' = P_2015_`i' \ x10_2015_`i''/ cond(r(N)==0,1,r(N))




}



// LISTAR MATRIZES DE TRANSIÇAO PARA CLUSTERS

forvalues i = 1/10 {
matrix results = P_2015_`n' * 100
outtable using "Matriz2015NEC1passo_`n'.xlsx", mat(results) replace format(%9.2f)
}




// MATRIZES DE TRANSIÇAO A N PASSOS


// MATRIZ A 5 PASSOS PELA PROP DE MARKOV
```

```
Elevar_Matriz 5 P_2015

matrix results = P_20155 * 100

matrix list P_20155

outtable using "Matriz2015NEC5passos.xlsx", mat(results) replace format(%9.2f)


// PARA CADA CLUSTER

forvalues i = 1/10 {
Elevar_Matriz 5 P_2015_‘i’
matrix results = P_2015_‘i’5 * 100
outtable using "Matriz2015NEC5passo_‘i’.xlsx", mat(results) replace format(%9.2f)
}




// MATRIZ A 5 PASSOS ESTIMADA

matrix list P_2011

matrix results = P_2011 * 100

outtable using "Matriz2015NEC5passosL.xlsx", mat(results) replace format(%9.2f)


// PARA CADA CLUSTER

forvalues i = 1/10 {
matrix results = P_2011_‘i’ * 100
outtable using "Matriz2015NEC5passoL_‘i’.xlsx", mat(results) replace format(%9.2f)
}
```

Anexo D

Código R

```
library(readr)
library(cluster)

### Carregar dados do Estado Civil e Nível de Ensino e usar função CLARA

x <- model.matrix( ~ ., DadosEstagioNEC)

cluster_clara<-clara(x, 10 , metric="euclidean", stand=FALSE,
samples=10000, sampsize= 1000, medoids.x=TRUE,pamLike=TRUE)


x2 <- model.matrix( ~ ., DadosEstagioEC)

cluster_clara2<-clara(x2, 10 , metric="euclidean", stand=FALSE,
samples=10000, sampsize= 1000, medoids.x=TRUE,pamLike=TRUE)


### Acrescentar aos dados uma coluna com os resultados

y <- cbind(DadosEstagioNEC,cluster_clara$clustering)

y2 <- cbind(DadosEstagioEC,cluster_clara2$clustering)

### Exportar Resultados

write.csv(y, file = "DadosNECCluster.csv")

write.csv(y2, file = "DadosECCluster.csv")
```

Anexo E

Código Mathematica

```
(* Definir processo markov utilizando distribuição em 2011 e P_2011 *)

CadeiaComplA = DiscreteMarkovProcess[initA, MatrizA];

(* Apresentar propriedades *)

MarkovProcessProperties[CadeiaComplA]

(* Grafos de transições *)

grA = Graph[CadeiaComplA, GraphLayout ? "LayeredDrawing"]



(* Definir processo markov utilizando distribuição em 2015 e P_2015^5 *)

CadeiaComplB = DiscreteMarkovProcess[initB, MatrizB];

(* Apresentar propriedades *)

MarkovProcessProperties[CadeiaComplB]

(* Grafos de transições *)

grB = Graph[CadeiaComplB, GraphLayout ? "LayeredDrawing"]
```