

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

CONHECIMENTO SOBRE O MERCADO PORTUGUÊS

SEGMENTAÇÃO GEODEMOGRÁFICA PARA DEFINIÇÃO
DE ESTRATÉGIAS COMERCIAIS NO MERCADO DO
CONSUMO DOMÉSTICO DE ELECTRICIDADE

Sandra Cristina Pereira Margalha Galvão Pisco

Trabalho de Projeto apresentado como requisito parcial para
obtenção do grau de Mestre em Estatística e Gestão de
Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

CONHECIMENTO SOBRE O MERCADO PORTUGUÊS

SEGMENTAÇÃO GEODEMOGRÁFICA PARA DEFINIÇÃO DE ESTRATÉGIAS COMERCIAIS NO MERCADO DO CONSUMO DOMÉSTICO DE ELECTRICIDADE

por

Sandra Cristina Pereira Margalha Galvão Pisco

Trabalho de Projeto apresentado como requisito parcial para a obtenção do grau de Mestre em Estatística e Gestão de Informação, Especialização em Marketing Research e CRM

Orientador/Coorientador: Professor Doutor Fernando Bação

Novembro 2017

DEDICATÓRIA

“Do not judge me by my successes, judge me by how many times I fell down and got back up again.”

“It always seems impossible until it's done.”

Nelson Mandela

Durante o período letivo deste mestrado vi a minha mãe partir, o que me fez perder algum alento. Consegui terminar a parte curricular mas o trabalho de projeto ficou em *stand by* por período indeterminado.

Mas sei que ela queria que eu cumprisse os objetivos a que me proponho e que conseguisse chegar à meta, apesar de todos os obstáculos.

Portanto dedico este trabalho à minha mãe, pois é em memória dela que me esforço por fazer mais e melhor.

AGRADECIMENTOS

Ao meu marido: Pelo apoio.

Às minhas filhas: Pela paciência e tempo que lhe roubei.

Restante família e amigos: Pela força.

Ao Professor Fernando Bação: pela ajuda no desenvolvimento e melhoramento do trabalho.

RESUMO

Com a liberalização do mercado energético em Portugal, tendo ficado concluída em 2012, várias foram as empresas estrangeiras que decidiram investir em Portugal neste sector.

Sendo o mercado empresarial o mais apetecível ao nível de volume de negócio e sobre o qual recaiu o investimento inicial, o mercado residencial apresenta-se também com um forte potencial de negócio. O problema neste segmento de negócio (ao contrário do mercado empresarial) coloca-se devido à falta de informação pública, acessível ou disponível sobre o mesmo, de forma a dotar estes novos *players* com mecanismos de avaliação adequados para poderem tomar as decisões mais acertadas e definir estratégias mais assertivas.

Dado este cenário, o grande objetivo deste Trabalho de Projeto é, com recurso a informação pública disponível (nomeadamente os Censos), aplicar conceitos da geodemografia e contruir uma segmentação geodemográfica da população portuguesa, que consiga responder a problemas de negócio concretos e assim auxiliar estas empresas no dia-a-dia do seu negócio.

Para obter a segmentação pretendida, técnicas de Análise de *Clusters* foram utilizadas, bem como o recurso a sistemas de informação geográfica para análise e demonstração dos resultados.

Os resultados obtidos foram suficientemente relevantes para passarem a ser considerados na tomada de decisões desta empresa de ora em diante.

PALAVRAS-CHAVE

Data Mining; Segmentação; Geodemografia ; Análise de Clusters; Censos

ABSTRACT

With the liberalization of the Energy Market in Portugal, which has been completed in 2012, several foreign companies decided to enter in Portugal and invest in this industry.

Being the corporate segment the most attractive in terms of business volume and on which had occur the initial investment, the residential segment also presents a strong business potential. The problem in this business segment (as opposed to corporate segment) arises due to the lack of public, accessible or available information on it, in order to equip these new players with evaluation mechanisms that enables them to take the right decisions and define more assertive strategies.

Given this scenario, the major objective of this Project is, using the available public information (namely the census), too use geodemography concepts and build a geographic segmentation of the Portuguese population, which can respond to specific business problems and thus help these companies their on a daily basis activity.

To obtain the desired segmentation, clusters analysis techniques were used, as well as geographic information systems for analysis and demonstration of results.

The results obtained were sufficiently relevant to be considered in the decision-making process of this company from now on.

KEYWORDS

Data Mining; Segmentation; Geodemography; Cluster Analysis; Census

ÍNDICE

1. Introdução	1
1.1. Enquadramento e definição do problema	1
1.1.1. Enquadramento.....	1
1.1.2. Definição do Problema	3
1.2. Objetivo do Estudo	2
1.3. Importância e Relevância do Estudo	4
2. Revisão da Literatura	5
2.1. O <i>Data Mining</i> e a Perspetiva Empresarial	5
2.2. Segmentação do mercado e a sua aplicabilidade	5
2.3. A Geodemografia e Segmentação geodemográfica.....	6
2.4. Os censos e a Segmentação Geodemográfica	7
2.5. Técnicas de Análise de Clusters e Segmentação Geodemográfica	8
3. Metodologia	11
3.1. Definição do problema	12
3.1.1. Objetivo da segmentação.....	12
3.1.2. Recolha de Dados e Seleção das Variáveis.....	12
3.2. Análise exploratória de dados (AED)	14
3.2.1. Estandardização dos dados	15
3.2.2. Identificação de problemas nos dados.....	20
3.2.3. Análise de correlações e seleção das variáveis	22
3.3. Seleção do método de Análise de <i>Clusters</i> e validação dos resultados.....	25
3.3.1. Segmentação de Nível de Risco.....	25
3.3.2. Segmentação de Nível de Consumo.....	27
3.3.3. Segmentação de Nível de Resposta	28
4. Resultados e Discussão.....	31
4.1. Segmentação Risco.....	31
4.2. Segmentação Consumo	33
4.3. Segmentação Resposta	35
4.4. Segmentação Final.....	37
5. Conclusões.....	39
6. Limitações e Recomendações para Trabalhos Futuros	40
7. Bibliografia.....	41
8. Anexos	43

8.1. Anexo 1 – Histograma Distribuição Variáveis Input	43
8.2. Anexo 2 – Código SAS para gráfico cotovelo.....	44

ÍNDICE DE FIGURAS

Figura 1 – Ilustração de exemplo de segmentação geográfica de Portugal	3
Figura 2 - Tipos de segmentação e categorias de informação utilizada	8
Figura 3 – Caracterizações das segmentações por nível de análise	13
Figura 4 – Fluxo de AED com recurso ao EM	20
Figura 5 - Fluxo de AED e segmentação com recurso ao EM.....	25
Figura 6 - Fluxo de AED, Segmentação e Caraterização dos segmentos com recurso ao EM .	31

ÍNDICE DE TABELAS

Tabela 1 - Variáveis iniciais de <i>input</i> – categoria “População”	16
Tabela 2 - Variáveis iniciais de <i>input</i> – categoria “Condições de Vida”	16
Tabela 3 - Variáveis iniciais de <i>input</i> – categoria “Emprego e Mercado de Trabalho”	17
Tabela 4 - Variáveis iniciais de <i>input</i> – categoria “Escolaridade”	17
Tabela 5 - Variáveis iniciais de <i>input</i> – categoria “Rendimento e Despesas Familiares”	17
Tabela 6 - Variáveis iniciais de <i>input</i> – categoria “Habitação e Conforto”	18
Tabela 7 – Variáveis de <i>input</i> após transformações	19
Tabela 8 – Análise univariada das variáveis	21
Tabela 9 – Matriz de correlações variáveis segmentação risco	22
Tabela 10 – Seleção final variáveis segmentação risco	23
Tabela 11 - Matriz de correlações variáveis segmentação consumo	23
Tabela 12 - Seleção final variáveis segmentação consumo	24
Tabela 13 - Matriz de correlações variáveis segmentação resposta	24
Tabela 14 - Seleção final variáveis segmentação resposta	24
Tabela 15 - Tabela de distribuição registos para 4 segmentos risco	26
Tabela 16 - Tabela de distribuição registos para 3 segmentos risco	26
Tabela 17 – Valores médios obtidos para os segmentos de risco	27
Tabela 18 - Tabela de distribuição registos para 4 segmentos consumo	27
Tabela 19 - Tabela de distribuição registos para 3 segmentos consumo	28
Tabela 20 - Valores médios obtidos para os segmentos de consumo	28
Tabela 21 - Tabela de distribuição registos para 4 segmentos resposta	29
Tabela 22 - Tabela de distribuição registos para 3 segmentos resposta	29
Tabela 23 - Valores médios obtidos para os segmentos de resposta	30
Tabela 24 – Pontuação Segmentação Final	38

ÍNDICE DE MAPAS

Mapa 1 – Caracterização da Região Metropolitana de Lisboa por Total de População	15
Mapa 2 – Caracterização da Região Metropolitana de Lisboa por Densidade Populacional ..	15
Mapa 3 – Caracterização geográfica da segmentação de risco	33
Mapa 4 - Caracterização geográfica da segmentação de consumo.....	35
Mapa 5 - Caracterização geográfica da segmentação de resposta	36
Mapa 6 - Caracterização geográfica da segmentação final	38

ÍNDICE DE GRÁFICOS

Gráfico 1 – Caracterização do mercado de eletricidade por consumo anualizado	3
Gráfico 2 – Caracterização do mercado de eletricidade por número de cliente.....	3
Gráfico 3 – Caracterização do mercado de eletricidade por número de clientes e tipos de consumidores	1
Gráfico 4 – Caracterização do mercado de eletricidade por consumo anualizado e tipo de consumidores	1
Gráfico 5 – Quota de mercado em ML por número de clientes	1
Gráfico 6 – Quota de mercado em ML por consumo anualizado	1
Gráfico 7 – Histograma da variável CV1	22
Gráfico 8 – Histograma da variável CV2	22
Gráfico 9 – Gráfico da variabilidade intra-cluster segmentação risco	25
Gráfico 10 – Gráfico de distribuição registos para 4 segmentos risco.....	26
Gráfico 11 - Gráfico de distribuição registos para 3 segmentos risco	26
Gráfico 12 – Importância variáveis segmento 1 risco	26
Gráfico 13 - Importância variáveis segmento 2 risco	26
Gráfico 14 - Importância variáveis segmento 3 risco	26
Gráfico 15 – Gráfico da variabilidade intra-cluster segmentação consumo.....	27
Gráfico 16 - Gráfico de distribuição registos para 4 segmentos consumo	27
Gráfico 17 - Gráfico de distribuição registos para 3 segmentos consumo	28
Gráfico 21 - Gráfico da variabilidade intra-cluster segmentação resposta	29
Gráfico 22 - Gráfico de distribuição registos para 4 segmentos resposta	29
Gráfico 23 - Gráfico de distribuição registos para 3 segmentos resposta	29
Gráfico 24 – Gráfico de médias normalizadas para segmentação risco	32
Gráfico 25 - Gráfico de médias normalizadas para segmentação consumo.....	34
Gráfico 26 - Gráfico de médias normalizadas para segmentação resposta	35

LISTA DE SIGLAS E ABREVIATURAS

ERSE Entidade Reguladora dos Serviços Energéticos

ML Mercado Liberalizado

MR Mercado Regulado

DM Data Mining

AED Análise Exploratória dos Dados

EM SAS Enterprise Miner

QGIS Quantum GIS

RML Região Metropolitana de Lisboa

1. INTRODUÇÃO

O ideia deste Trabalho de Projeto foi delineado durante o ano de 2013 e derivado de um projeto profissional que na altura me foi proposto. O grande objetivo era o de auxiliar uma entidade comercializadora de energia elétrica estrangeira a obter um maior conhecimento sobre o mercado português e particularmente sobre os comportamentos, capacidades e hábitos de consumo no que se refere ao consumo de eletricidade num segmento específico.

Dado que a necessidade demonstrada pela empresa em causa remonta ao ano de 2013 e todo o enquadramento faria sentido para esse período, toda a abordagem desenvolvida neste Trabalho de Projeto se fixa nesse período, embora só agora se apresentem os resultados.

Detendo já uma quota de mercado no que refere ao mercado de empresas, pretendia nesse altura entrar no mercado residencial e expandir a sua carteira de clientes também neste nicho de mercado.

Não tendo uma base interna para trabalhar, pretendia o desenvolvimento de uma segmentação geográfica com recurso a dados estatísticos disponíveis em *sites* públicos, distribuídos por zona geográfica, como os Censos, para alcançar um conhecimento e planejar estratégias de atuação comercial e efetuar decisões de negócio mais ajustadas e com maior confiança.

1.1. ENQUADRAMENTO E DEFINIÇÃO DO PROBLEMA

1.1.1. Enquadramento

A liberalização do mercado¹ de eletricidade em Portugal está em curso desde 2000 e já entrou na sua fase plena, com a extinção total das tarifas reguladas² para todos os clientes.

Na maior parte dos países europeus, o processo de liberalização dos sectores elétricos foi comum e efetuado de forma faseada, tendo começado por incluir os clientes de maiores consumos e níveis de tensão mais elevados. Em Portugal seguiu-se uma metodologia idêntica, tendo a abertura de mercado sido efetuada de forma progressiva entre 1995 e 2006.

Desde 4 de Setembro de 2006 todos os consumidores em Portugal continental podem escolher o seu fornecedor de energia elétrica, antecipando a Diretiva 2003/54/CE do Parlamento Europeu, emitida 26 de Junho de 2003, que definia como prazo limite 1 de Julho de 2007, e depois da publicação do Decreto-Lei nº 192/2004, de 17 de Agosto, em que o direito de elegibilidade foi alargado a todos os clientes, e da publicação do decreto-Lei nº 297/2006, de 15 de Fevereiro estabeleceu os princípios de organização e funcionamento do sistema nacional energético (transpondo os princípios constantes na Diretiva atrás mencionada).

¹ Mercado deixa de ser um monopólio e passa a haver a possibilidade de entrada de novos comercializadores de energia.

² Preços fixados por uma entidade reguladora do mercado.

A 26 de Março de 2012, os decretos-lei nº75/2012 e nº74/2012 vieram apresentar a extinção gradual das tarifas reguladas de venda de eletricidade e gás, definindo um período transitório máximo de 3 anos, para incentivar os consumidores a mudar de comercializador de energia.

Assim, a partir de 1 de Julho de 2012, deixou de ser possível realizar novos contratos em mercado regulado para os consumidores de eletricidade com potências contratadas iguais ou superiores a 10,35 kVA e gás 500m³. A partir de 1 de Janeiro de 2013, esta impossibilidade estendeu-se aos restantes consumidores.

A diferença entre o mercado regulado e o mercado livre é que no mercado regulado, os preços de venda da energia são fixados anualmente pela Entidade Reguladora dos Serviços Energéticos (ERSE)³, sendo essa a tarifa praticada pela EDP Serviço Universal⁴. No mercado livre, os preços da energia são estabelecidos por cada comercializador, respeitando as regras da concorrência e o Regulamento das Relações Comerciais.

O mercado liberalizado (ML) veio permitir a livre concorrência nos mercados de eletricidade e gás, possibilitando o aparecimento de vários comercializadores de energia e uma maior escolha por parte dos consumidores. A eliminação das tarifas reguladas veio obrigar os consumidores a optar por uma comercializadora do mercado liberalizado e desta forma tornar visível as diferentes tarifas existentes no mercado de forma a gerar uma concorrência saudável e benéfica para os consumidores.

Desta forma, associada à liberalização e à construção do mercado interno de eletricidade está um esperado aumento da concorrência, com reflexos ao nível dos preços e da melhoria da qualidade de serviço, a que deverá corresponder uma maior satisfação dos consumidores de energia elétrica.

Neste momento existem já diferentes comercializadores de eletricidade presentes no mercado liberalizado em Portugal. A designar:

- EDP Comercial - Comercialização de Energia, SA;
- Enat – Energias Naturais, Lda.;
- Endesa – Endesa Energia Sucursal Portugal;
- Galp Power S.A;
- Iberdrola Generación – Energia e Serviços Portugal, Unipessoal, Lda.;
- Union Fenosa Comercial, S.L. – Suc. Em Portugal.
- AUDAX ENERGIA:
- Axpo Iberia, S.L.;

³ Entidade que controlava e definia as tarifas em mercado regulado e atualmente controla as tarifas em mercado liberalizado de clientes de último recurso, ou seja, clientes de condições especiais.

⁴ Empresa monopolista pertencente ao Estado português, que detinha 100% de quota de mercado antes da liberalização.

- ENFORCESCO, SA;
- Nexus Energía, S.A.

1.1.2. Definição do Problema

De acordo com o Relatório do Mercado Liberalizado de Eletricidade, referente a Novembro 2013 disponibilizado pela ERSE (ERSE, 2013), a situação caracteriza-se por:

O consumo no mercado liberalizado tem vindo a crescer, representando já 71,5% do consumo global de eletricidade no país;

A fatia dos 28,5% de consumo ainda presente em mercado regulado representa 64,64% do total dos consumidores.

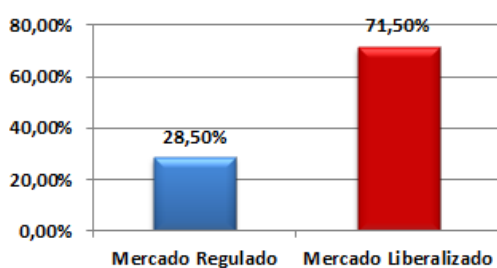


Gráfico 1 – Caracterização do mercado de eletricidade por consumo anualizado

Fonte: ERSE

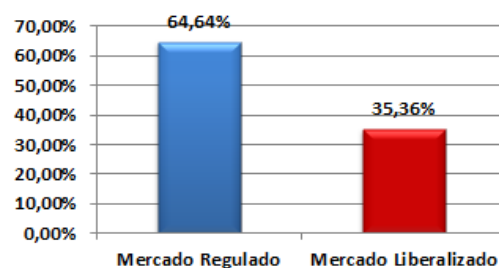


Gráfico 2 – Caracterização do mercado de eletricidade por número de cliente

Fonte: ERSE

O número de consumidores domésticos em mercado regulado baixou da fasquia dos 4 milhões quando há um ano era ainda de 5,2 milhões de clientes, representando ainda mais de 60% do total de consumidores;

A maior fatia de consumidores que resta fazer a passagem para mercado liberalizado são consumidores residenciais.

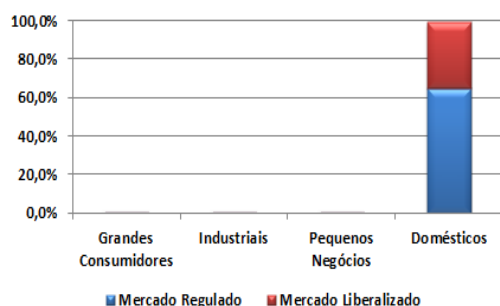


Gráfico 3 – Caracterização do mercado de eletricidade por número de clientes e tipos de consumidores
Fonte: ERSE

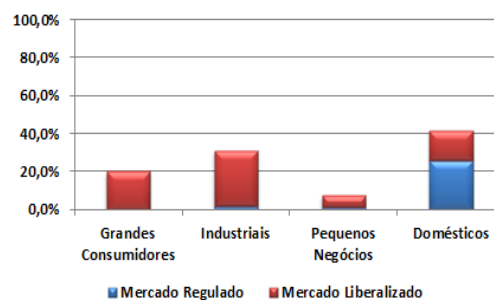


Gráfico 4 – Caracterização do mercado de eletricidade por consumo anualizado e tipo de consumidores
Fonte: ERSE

A EDP Comercial continua como o principal operador no mercado livre, tendo registado um ligeiro aumento da sua quota de clientes para 83,7%, e quota de consumo de 43,6%. A Endesa ocupa a segunda posição com uma quota de consumo abastecido de 21,3% e quota por clientes de 7,7%. Quer a Iberdrola, quer a Galp ganharam em Novembro quota em termos de consumos abastecidos com 21% e 5,9%, respetivamente.

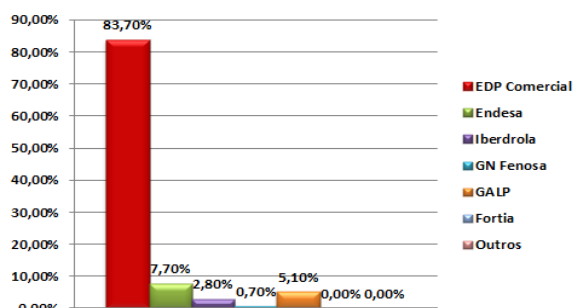


Gráfico 5 – Quota de mercado em ML por número de clientes
Fonte: ERSE

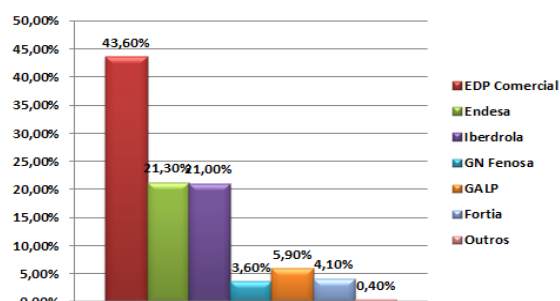


Gráfico 6 – Quota de mercado em ML por consumo anualizado
Fonte: ERSE

De acordo com este relatório, embora mais de 70% do total de consumo de eletricidade esteja já a ser comercializada no mercado liberalizado, analisando o número de consumidores está em falta a passagem de uma grande quota de consumidores domésticos para este mercado sendo que têm até final de 2015 para o fazer. Adicionalmente se salienta que para a parte de consumidores domésticos já implantados no mercado liberalizado, a quota de mercado da EDP Comercial é substancialmente superior aos das outras entidades concorrentes.

Face a esta conjuntura de fatores, que se resumem em oportunidade de negócio vs falta de informação sobre o mercado alvo, esta entidade expressou a necessidade em obter um conhecimento prévio do mercado português no que diz respeito à capacidade de pagamento dos consumidores do mercado doméstico de eletricidade e hábitos de consumo para poder sustentar as suas estratégias comerciais a aplicar neste novo segmento e decisões de negócio em factos concretos e reais e fazer com que a sua estratégia de investimento em Portugal seja sustentada e fundamentada em conhecimento.

1.2. OBJETIVO DO ESTUDO

Para alcançar o objetivo proposto e conseguir captar perfis comportamentais mediante várias perspetivas, o projeto baseia-se assim no desenvolvimento de uma segmentação geodemográfica do mercado que deverá devolver informação mediante três vertentes de análise:

- **Risco – Nível de risco**

Esta segmentação pretende aferir sobre a “qualidade” da população residente em cada zona geográfica. Irá recorrer a informação sobre características da população residente, nomeadamente idade, género, nacionalidade, bem como informação sobre condições de vida da população e como a mesma se enquadra no mercado de trabalho (taxa de desemprego, taxa de atividade, etc).

- **Consumo – Nível de consumo**

Pretende-se a captação de informação sobre o nível de consumo, em termos de energia, tentando produzindo informação sobre as zonas onde os consumidores detêm um maior ou menor consumo médio neste tipo de serviços. Para o objetivo, esta segmentação recorre a informações sobre as características das habitações existentes em cada zona, nomeadamente se existem sistemas de aquecimento incluídos ou mesmo ar condicionados incorporados nas mesmas, desta forma concluindo sobre as necessidades existentes.

- **Resposta – Nível de capacidade financeira**

Informação sobre quais as zonas de concentração de pessoas com um maior ou menor nível qualificações e com níveis de rendimentos mais elevados ou mais baixos. Pretende-se obter uma segmentação que especifique características da população ligadas ao seu poder de compra.

Através da utilização de técnicas de *Data Mining* (DM) adequadas e da análise da informação recolhida, pretende-se:

- Identificação das variáveis de segmentação e segmentar o mercado;
- Avaliação da atratividade de cada segmento;
- Identificação dos possíveis conceitos de posicionamento para cada segmento-alvo.

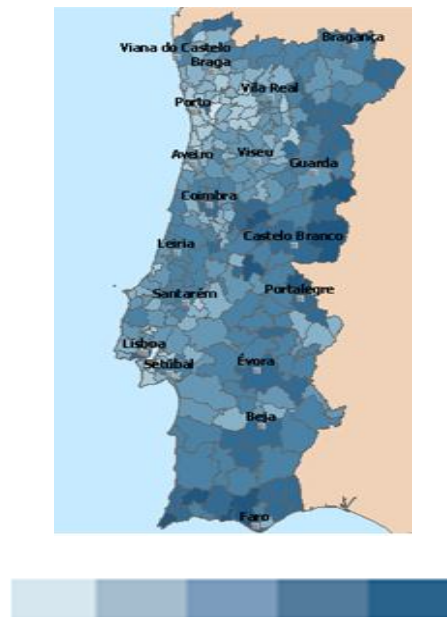


Figura 1 – Ilustração de exemplo de segmentação geográfica de Portugal
Fonte: Wikipédia

Os objetivos específicos para alcançar os resultados pretendidos podem-se descrever como:

- Investigar sobre informação disponível para o desenvolvimento proposto;
- Selecionar informação relevante para o desenvolvimento e para a caracterização geográfica pretendida;
- Aplicar técnicas de DM ao tipo de problema em análise:
 - Trabalho de preparação dos dados;
 - Efetuar o pré-processamento e executar a Análise Exploratória dos Dados;
 - Selecionar algoritmos de segmentação adequados.
- Efetuar a caracterização dos segmentos;
- Avaliar a aplicabilidade do desenvolvimento ao sector em análise e ao mercado alvo, através da definição de estratégias de negócio mediante as caracterizações obtidas.

1.3. IMPORTÂNCIA E RELEVÂNCIA DO ESTUDO

A realidade para o segmento de atual do país caracteriza-se por:

- Forte volatilidade da Economia;
- Deterioração poder de compra;
- Aumento desemprego;
- Aumento do incumprimento;
- Incremento da competitividade.

É imperativo conhecer e compreender o comportamento dos clientes, e assim conseguir reduzir o risco e aumentar a rentabilidade.

É com base nesta premissa que surge este projeto, que vem propor o desenvolvimento de uma solução que permita obter conhecimento completo sobre o mercado de atuação e que se pautar pela avaliação sobre várias perspectivas fulcrais para o sucesso da entidade.

Várias serão as mais-valias obtidas pela segmentação de mercado proposta. Através da utilização desse mesmo conhecimento obtido e transformado em estratégia de negócio mais ajustadas e adaptadas, esta entidade conseguirá alcançar:

- Redução de custos de investimento e de tempo em clientes sem potencial de negócio;
- Melhora na eficiência através de maior taxa de resposta/concretização, realizando uma pré-seleção dos potenciais clientes com os perfis pretendidos;
- Prevenção de futuros incumprimentos minimizando a taxa de morosidade;
- Maior e melhor conhecimento sobre o mercado português.

Ficará assim com a possibilidade de otimizar os seus recursos e alocar os mesmos de uma forma eficiente e assertiva, produzindo mais-valias concretas para a empresa e prevenindo situações indesejadas.

2. REVISÃO DA LITERATURA

Neste capítulo serão abordadas as metodologias de DM para dar resposta ao problema identificado e desta forma obter a caracterização de segmentos pretendidos. Começa-se por fazer uma breve descrição do DM, em seguida evidenciar a aplicabilidade das segmentações de mercado na vertente empresarial, evidenciar a opção da abordagem da segmentação geodemográfica no caso em análise e por último enumerar as técnicas estatísticas e algoritmos disponíveis para o desenvolvimento deste projeto.

2.1. O DATA MINING E A PERSPETIVA EMPRESARIAL

Muitos são os autores que possuem distintas definições para o conceito de DM. Para Bação, uma descrição suficientemente genérica é “a extração de informação escondida e de carácter eminentemente preditivo de grandes bases de dados” (Fernando Bação, 2012).

Assim, o grande objetivo do DM prende-se com a captação de comportamentos não conhecidos em grandes volumes de dados e que possam vir a servir de mais-valia de alguma forma, seja para prever comportamentos futuros, seja apenas para análise e classificação de comportamentos ou extração de ligações existentes mas desconhecidas.

Estando hoje a viver desafios sem antecedentes em termos de Gestão do Conhecimento (NASCIMENTO, 2004) (criação, armazenamento e distribuição de enormes volumes de dados, rápida alteração da informação e constante mudança das condições de trabalho), a gestão do conhecimento passa a ser uma das competências essenciais para a competitividade das organizações.

O tema central da performance das empresas no ambiente atual, que se caracteriza desafiador e em rápida mutação, resulta da sua capacidade de obter vantagem competitiva face aos seus concorrentes (Bal, Bal, & Demirhan, 2011). Uma organização que pretende ter vantagem competitiva tem que criar e alavancar suas capacidades. Uma das bases centrais para alcançar a vantagem competitiva é a capacidade organizacional para criar novo conhecimento e transferi-lo através de vários níveis da organização.

Assim, o DM tornou-se uma área de pesquisa com importância crescente em organizações de todos os tamanhos, que começaram a desenvolver e implementar novas tecnologias para tirar maior proveito dos dados disponíveis e para melhorar a sua capacidade de tomada de decisão.

2.2. SEGMENTAÇÃO DO MERCADO E A SUA APLICABILIDADE

A segmentação é um processo de identificação de grupos homogéneos de elementos. Assim, para uma empresa, segmentar o seu mercado consiste em efetuar a sua divisão num certo número de subconjuntos, tão homogéneos quanto possível, para permitir uma adaptação das políticas e definição de estratégias a cada um desses subconjuntos ou a alguns deles.

Aos subconjuntos dá-se o nome de segmentos, e pretende-se que cada um seja constituído por elementos com as mesmas necessidades, comportamentos, valores (McCarthy & Perreault, 1991). No entanto, os segmentos devem ser significativamente diferentes entre si, de forma a justificar a aplicação de diferentes políticas e estratégias.

Dado que a elevada quantidade, multidimensionalidade e variância dos dados é frequente neste tipo de problemas, a sua análise não é trivial. É, portanto, uma situação que justifica uma abordagem típica de *DM* (Fenando Bação & Loureiro, 1995).

Existem vários critérios de segmentação, sendo que os mais frequentes são (McCarthy & Perreault, 1991):

- a) Segmentação geográfica - divide o mercado a partir de áreas geográficas tais como: país, região, cidade, bairro;
- b) Segmentação demográfica - divide o mercado a partir de variáveis populacionais tais como: escalão etário, género, nível de rendimentos, ocupação, formação;
- c) Segmentação psicográfica - divide o mercado a partir de variáveis como a classe social, estilo de vida, convicções, personalidade;
- d) Segmentação por comportamento - divide o mercado com base no conhecimento, uso, e comportamento.

Sendo que é possível a combinação de critérios, como é o exemplo da segmentação geodemográfica que segmenta mediante informação geográfica e demográfica (Tanner Jr. & Raymond, 2011).

Técnicas de segmentação são utilizadas por empresas de todos os setores da indústria com aplicabilidade no marketing e planeamento estratégico da empresa. Para muitos, a segmentação do mercado é até considerada a panaceia do marketing moderno.

Tanto a finalidade de utilização como as mais-valias associadas à segmentação estão bem estabelecidas na literatura. As necessidades dos clientes apresentam-se cada vez mais díspares e essas necessidades não conseguem ser satisfeitas por recurso a uma abordagem de *Mass Marketing*. É possível para as empresas então lidar com essa diversidade agrupando clientes com requisitos similares e comportamento de compra em segmentos. As escolhas sobre quais segmentos são os mais apropriados para servir podem então ser feitas, fazendo assim uma melhor gestão dos recursos finitos que existem (Dibb, 1998).

2.3. A GEODEMOGRAFIA E SEGMENTAÇÃO GEODEMOGRÁFICA

Sobre uma perspetiva de planeamento e de negócio é fundamental conhecer e compreender a estratificação espacial da sociedade (SCHWABE & O'DONOVAN, 1997). Através da identificação

de comunidades com perfis similares, e consequentemente preocupações e necessidades básicas similares, as empresas poderão enfocar melhor as suas estratégias e alocar melhor os recursos.

A Geodemografia inclui a aplicação de classificações geodemográficas na atividade empresarial. Pode-se resumir como a área que estuda as pessoas com base no sítio onde elas vivem. Os sistemas geodemográficos pretendem estimar características e comportamentos das pessoas com base no perfil agrupado de todas as pessoas que vivem numa determinada área. Funciona através da recolha de dados sobre os consumidores com referência espacial, construindo modelos estatísticos que identificam e mapeiam a distribuição dessas características ou tipos de consumidores (Goss, 1995)

Muita tem sido a evolução da geodemografia como uma ferramenta de gestão para as empresas (Webber, n.d.). A Geodemografia começou a ser utilizada dado que permitia descrever tipologias de áreas que se mostraram como poderosas discriminadoras do comportamento do consumidor e desta forma na ajuda da 'análise de mercado' (Longley & Clarke, 1995).

Sistemas geodemográficos, com capacidades avançadas de visualização e exploração tem sido desenvolvidos por empresas para ajudar as outras empresas nessa gestão. Já existem no mercado produtos desenvolvidos nesta área, nomeadamente o *ACORN* e *MOSAIC* no Reino Unido e *LIFESTYLE*, *ClusterPlus* e *VISION* na América do Norte. Em Portugal ainda nenhuma empresa desenvolveu este tipo de análise, como produto integrado dentro do seu portfolio.

De particular importância para o aparecimento da indústria geodemográfica foi o desenvolvimento de técnicas de agrupamento para agrupar zonas estatisticamente similares em classes. Este tipo de segmentação é primariamente usada na descoberta e exploração de novos mercados (Weinstein, 1994)

A segmentação geodemográfica é baseada em dois princípios simples (Sivadas, 1997):

- As pessoas que vivem na mesma zona são mais propensas a ter características semelhantes do que duas pessoas que são escolhidas aleatoriamente.
- As zonas podem ser segmentadas em termos das características da população que eles contêm. Quaisquer duas zonas podem ser colocados na mesma categoria, ou seja, contêm pessoas com características semelhantes, embora estejam amplamente separados.

2.4. OS CENSOS E A SEGMENTAÇÃO GEODEMOGRÁFICA

Sistemas geodemográficos podem usar vários tipos de informação mas de longe a mais importante e relevante são os dados recolhidos nos Censos (Birkin, 1995).

A ideia de que os resultados do recenseamento poderiam servir para identificar e caracterizar as geografias das cidades ganhou força com a maior disponibilidade de dados nacionais do censo e a capacidade computacional de procurar padrões em grandes volumes de dados.

Diferentes estudos já evidenciaram que características demográficas, como idade, estado civil, nível social, educação, etc, são fatores discriminantes para explicar, por exemplo, comportamentos de consumo por zona geográfica (Webber, 1813). Já amplamente explorado que vizinhos tendem a ter comportamentos similares.

A figura seguinte mostra formas de segmentação e informação utilizada em cada tipo de segmentação (Tanner Jr. & Raymond, 2011).

By Behavior	By Demographics	By Geography	By Psychographics
■ Benefits sought from the product ■ How often the product is used (usage rate) ■ Usage situation (daily use, holiday use, etc.) ■ Buyer's status and loyalty to product (nonuser, potential user, first-time users, regular user)	■ Age/ generation ■ Income ■ Gender ■ Family life cycle ■ Ethnicity ■ Family size ■ Occupation ■ Education ■ Nationality ■ Religion ■ Social class	■ Region (continent, country, state, neighborhood) ■ Size of city or town ■ Population density ■ Climate	■ Activities ■ Interests ■ Opinions ■ Values ■ Attitudes ■ Lifestyles

Figura 2 - Tipos de segmentação e categorias de informação utilizada
Fonte: Livro "Market Segmentation, Targeting and Positioning"

Outros estudos mostram também a aplicação de informação de alojamentos e famílias como dados relevantes para explicar os comportamentos do consume de energia, como é referido em "Drivers of variability in energy consumption" (Albert, Gebru, & Ku, 2013)

2.5. TÉCNICAS DE ANÁLISE DE CLUSTERS E SEGMENTAÇÃO GEODEMOGRÁFICA

A Análise de *Clusters* é uma técnica multivariada amplamente utilizada em segmentações tanto em ambiente profissional como académico (Aldenderfer and Blashfield, 1984). Funciona através da utilização se uma medida de similaridade ou distância para a criação de grupos homogéneos ou *clusters* (Dibb, 1998).

O uso de diferentes algoritmos leva a resultados diferentes, mas não existe uma abordagem para selecionar o melhor algoritmo, assim como nenhum algoritmo oferece qualquer prova teórica de sua certeza (Grekousis & Thomas, 2012). Interessa conhecer as propriedades associadas a cada um deles para poder decidir qual se deverá utilizar.

Uma das técnicas de *Clustering* mais utilizadas na segmentação geodemográfica é o amplamente conhecido algoritmo de agrupamento k-Means. Na verdade, a maioria dos sistemas geodemográficos comerciais atuais são baseados em um algoritmo de k-Means. No

entanto, a utilização de técnicas de segmentação como redes neurais artificiais, algoritmos genéticos ou lógica difusa tem sido validados como mais eficientes quando existem necessidades computacionais acrescidas e mais complexas, seja pela razão da existência de grandes volumes de dados, nível elevado de complexidades ou multidimensionalidade dos dados (Brimicombe, 2007).

As redes neurais podem lidar com relações não-lineares, são robustas ao ruído e exibem um alto grau de automação. Não assumem hipóteses quanto à natureza ou distribuição dos dados e fornecem assistência valiosa para lidar com problemas de natureza geográfica que, até à data, foram impossíveis de resolver. Um dos métodos de rede neural mais conhecidos e mais eficientes para alcançar agrupamentos não supervisionados é o Self-Organizing Map (SOM). O SOM foi proposto como uma melhoria em relação ao método k-means, pois fornece uma abordagem mais flexível para o agrupamento de dados do recenseamento. O método SOM foi usado recentemente por Spielman e Thill (Spielman & Thill, 2008) para desenvolver o agrupamento geodemográfico de um conjunto de dados do recenseamento sobre a cidade de Nova York.

O algoritmo K-Means, proposto por J. MacQueen em 1967, é um dos mais conhecidos e utilizados, além de ser o que possui o maior número de variações (Macqueen, 1967). O algoritmo inicia com a escolha de k elementos, que formaram as sementes iniciais. A escolha pode ser feita mediante vários critérios, nomeadamente:

- Através da seleção das k primeiras observações;
- Através da seleção das observações aleatoriamente;
- Através da seleção das k observações de forma que seus valores sejam bastante diferentes. Por exemplo, para segmentar uma população em três grupos de acordo com a altura dos indivíduos, poderíamos optar por escolher um indivíduo de baixa estatura, um de estatura mediana e um alto.

Escolhidas as sementes, a distância às mesmas é calculada para cada elemento, agrupando esse elemento ao grupo para o qual a menor distância foi encontrada (mais similar) e recalculando o centróide do mesmo. O processo é repetido até que todos os elementos façam parte de um dos grupos.

Após agrupar todos os elementos, procura-se encontrar uma partição melhor do que a gerada arbitrariamente. Desta forma, é calculado o nível de homogeneidade interna dos grupos através da Soma de Quadrados dos Resíduos (SQRes), medida utilizada para avaliar a qualidade da segmentação.

Após o cálculo, o processo passa por mover os objetos para outros grupos e verificar se existem ganhos no valor da SQRes, ou seja, se ocorre uma diminuição desse valor. Caso se verifique, o objeto é movido para o grupo que produzir o maior ganho, a SQRes dos grupos é recalculada e passa-se ao objeto seguinte. Depois de um determinado número de iterações ou inexistência de alterações, o processo é terminado.

O algoritmo *K-Means* foi o algoritmo utilizado para o estudo objeto deste projeto.

3. METODOLOGIA

A metodologia adotada neste projeto, dado que se trata de uma segmentação de mercado com recurso a atributos demográficos distribuídos geograficamente (Portugal Continental), e obtida através da utilização de técnicas de Análise de *Clusters*, pauta-se por se dividir em várias fases:

1. Definição do problema, que engloba:
 - a. Identificar o objetivo da segmentação;
 - b. Seleção das variáveis a incorporar na segmentação, que possam descrever melhor as semelhanças/disseminhanças entre os indivíduos da população alvo e sejam consideradas relevantes para o objetivo do estudo.
2. Análise Exploratória dos Dados (AED), que engloba:
 - a. Análise de estatísticas descritivas para obter sensibilidade relativamente aos mesmos e avançar para as análises subsequentes com maior capacidade de deteção de problemas;
 - b. Identificação de *missings* e *outliers* (observações sem valor ou com valores atípicos) e avaliação da remoção dos mesmos;
 - c. Estandarização dos dados para transformação das variáveis numa mesma escala de medida (dependente do método de análise de clusters escolhido);
 - d. Análise de correlações para reduzir as variáveis de *input* e tratar o problema da multidimensionalidade e seleção final das variáveis a entrar na segmentação.
3. Segmentação com recurso ao método de análise de *clusters* – K-MEANS.
4. Análise da atratividade dos segmentos produzidos (interpretação e caracterização).
5. Aplicação prática dos desenvolvimentos, ou seja, conversão em estratégias de negócio.

O trabalho centra-se na segmentação da Região Metropolitana de Lisboa (RML) pois, por um lado, a metodologia apresentada é replicável na segmentação das restantes áreas em que possa haver interesse e por outro simplifica-se a carga de trabalho repetitivo neste projeto sem penalizar a riqueza do detalhe da metodologia desenvolvida.

O nível de detalhe de segmentação desce até à identificação da freguesia.

As plataformas informáticas utilizadas para os processos de preparação, pré-processamento e segmentação dos dados foram o *SAS Enterprise Miner (EM)*⁵, como ferramenta de análise

⁵ Ferramenta tecnológica, desenvolvida por uma empresa privada, SAS, que permite e auxilia no desenvolver modelos de *Data Mining*, tal como modelos comportamentais ou preditivos. Possui um

estatística e segmentação e o QGIS⁶, como o sistema de informação geográfica para análise dos dados em mapas de localização. A importação dos dados foi efetuada e a base de dados de trabalho para efetuar as segmentações foi criada.

3.1. DEFINIÇÃO DO PROBLEMA

3.1.1. Objetivo da segmentação

O objetivo da segmentação é encontrar grupos homogêneos de freguesias, da RML, que possam ser caracterizados mediante as várias perspetivas de análise mencionadas: Risco, Consumo e Resposta.

Através da identificação destes grupos, é possível caracterizá-los e obter conhecimento para posterior definição de estratégia a executar dentro de cada grupo.

3.1.2. Recolha de Dados e Seleção das Variáveis

As mais-valias de utilização de dados provenientes dos Censos já foi detalhada no capítulo anterior. Desta forma, foi recolhida e compilada informação do portal do INE⁷, relativa aos Censos efetuados no ano de 2011 em Portugal. Apesar do período de tempo transcorrido, não existe informação mais recente, onde se possa obter este nível de riqueza de informação estatística.

- INE – informação consultada em www.ine.pt;

Através da análise à informação disponível, foram selecionadas determinadas categorias de dados, as quais, de acordo com a literatura analisada, são relevantes no estudo em causa. Desta forma, o projeto focar-se-á na caracterização das zonas (freguesias) com recurso às seguintes categorias de informação demográfica:

- População;
- Condições de Vida;
- Emprego e Mercado de Trabalho;
- Educação;

conjunto de funcionalidades que permitem conhecer e explorar os dados, bem como desenvolver segmentações recorrendo a diversas técnicas.

⁶ QGIS (anteriormente conhecido como "Quantum GIS") é um software livre/open source multiplataforma de sistema de georreferenciamento (GIS) que provê visualização, edição e análise de dados georreferenciados. (Fonte: Wikipédia, <<https://pt.wikipedia.org/wiki/QGIS>>, consultado em 01/11/2017).

⁷ O Instituto Nacional de Estatística (também conhecido pela sua sigla INE, pronunciada como "iné") é o organismo oficial de Portugal responsável por produzir e divulgar informação estatística oficial de qualidade, promovendo a coordenação, o desenvolvimento e a divulgação da actividade estatística nacional. O INE foi criado em 1935 por transformação da Direcção-Geral de Estatística. (Fonte: Wikipédia, <[https://pt.wikipedia.org/wiki/Instituto_Nacional_de_Estat%C3%ADstica_\(Portugal\)](https://pt.wikipedia.org/wiki/Instituto_Nacional_de_Estat%C3%ADstica_(Portugal))>, consultado em 01/11/2017).

- Rendimento e Despesas Familiares;
- Habitação e Conforto.

Dentro de cada categoria foram selecionadas várias variáveis, consideradas relevantes para o estudo, e distribuídas de acordo com a sua aplicação, para cada tipo de segmentação que se pretende desenvolver. A divisão encontra-se descrita na tabela seguinte.

Risco	Consumo	Resposta
•População •Densidade populacional (N.º/ km²) •População residente - género masculino •População residente - Faixa Etária Menos de 10 anos •... •Condições de Vida •Índice de dependência total •Índice de sustentabilidade potencial •... •Emprego e Mercado de Trabalho •Taxa de desemprego •...	•Habitação e Conforto •Alojamentos com Escalão de divisões - 2 ou menos •Alojamentos com Dimensão (famílias no alojamento) - 1 família •Alojamentos com Escalão de área útil por ocupante - 60 m² ou mais •Alojamentos com Ar Condicionado •Alojamentos com Aquecimento •...	•Educação •População residente com ensino superior completo •Taxa de abandono escolar •... •Rendimento e Despesas familiares •Encargos médios mensais por aquisição de habitação própria •Valor médio mensal das rendas dos alojamentos familiares clássicos arrendados •...

Figura 3 – Caracterizações das segmentações por nível de análise

Fonte: A autora

Desta forma, a tabela final utilizada para o projeto, inclui os seguintes tipos de variáveis, e distribuídas mediante o tipo de segmentação:

Segmentação Risco

- 1. População**
 - a. Densidade populacional (N.º/ km²)
 - b. População residente - género
 - c. População residente - faixa Etária
 - d. População residente - nacionalidade
- 2. Condições de Vida**
 - a. Índice de dependência de idosos
 - b. Índice de dependência de jovens
 - c. Índice de dependência total
 - d. Índice de sustentabilidade potencial
- 3. Emprego e Mercado de Trabalho**
 - a. Taxa de atividade
 - b. Taxa de desemprego
 - c. Profissionais socialmente mais valorizados

Segmentação Consumo

4. Habitação e Conforto

- a. Densidade de alojamentos (N.º/ km²)
- b. Alojamentos - Escalão de divisões
- c. Alojamentos com Dimensão (famílias no alojamento)
- d. Alojamentos com Escalão de área útil por ocupante
- e. Alojamentos – Existência de Ar Condicionado
- f. Alojamentos - Existência de Aquecimento
- g. Idade média dos edifícios (Ano)

Segmentação Resposta

5. Educação

- a. População residente com ensino superior completo
- b. População residente com 15 e mais anos de idade sem nenhum nível de escolaridade completo
- c. Taxa de abandono escolar

6. Rendimento e Despesas Familiares

- a. Encargos médios mensais por aquisição de habitação própria
- b. Valor médio mensal das rendas dos alojamentos familiares clássicos arrendados

3.2. ANÁLISE EXPLORATÓRIA DE DADOS (AED)

Tal como já descrito foi efetuada a recolha de informação para as várias freguesias da RML, à data dos Censos de 2011.

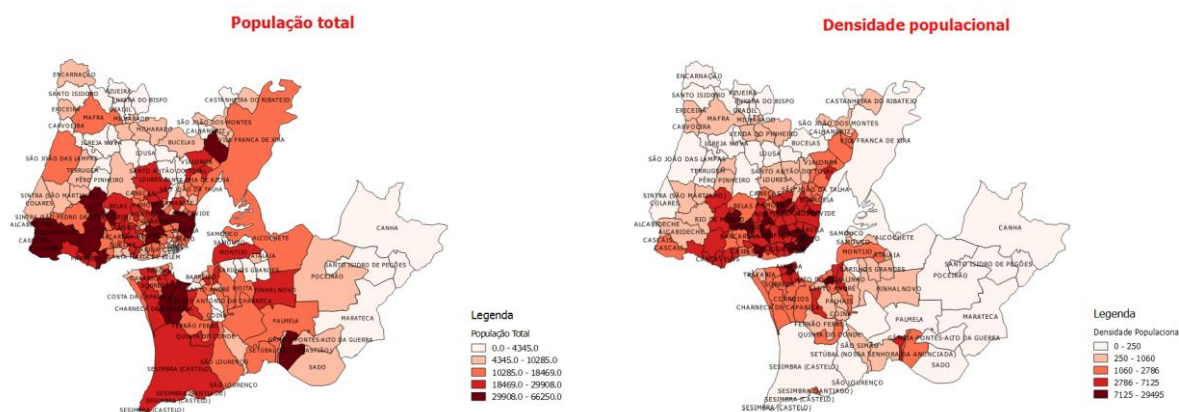
A RML nessa altura caracterizava-se por:

- População Total: 2.821.876 habitantes
- Densidade Populacional: 940 Número de pessoas por Km²
- Número de Freguesias: 211 freguesias

É sobre esta informação que vai ser executado e estudo e analisados os resultados.

Analisando mais em detalhe a região em causa, podemos identificar que os municípios que compõem a Área Metropolitana de Lisboa são 18, agrupados em duas sub-regiões, Grande Lisboa e Península de Setúbal. A Região da Grande Lisboa caracteriza-se por deter 9 municípios, e 153 freguesias. Relativamente à Península de Setúbal, caracteriza-se por deter também 9 municípios, mas apenas 58 freguesias.

Os mapas seguintes descrevem, ao nível da freguesia, a distribuição da população total por freguesia, bem como o nível de densidade populacional para o mesmo nível de detalhe geográfico.



Mapa 1 – Caracterização da Região Metropolitana de Lisboa por Total de População
Fonte: INE

Mapa 2 – Caracterização da Região Metropolitana de Lisboa por Densidade Populacional
Fonte: INE

Esta caracterização geográfica reflete a ordenação do território e a Carta Administrativa Oficial de Portugal (CAOP)⁸ em vigo na altura dos Censos 2011 (CAOP 2011), a qual já sofreu alterações, sendo a ultima versão data de março de 2016 (CAOP 2016).

O trabalho de AED nesta fase será efetuado para garantir a fiabilidade do estudo e do trabalho final.

3.2.1. Estandardização dos dados

O processo de transformação de dados é muito importante, dado que poderá trazer mais-valias substanciais aos desenvolvimentos das segmentações devido à natureza dos algoritmos utilizados. Nesta fase será analisada a necessidade de estandardização de dados para que os mesmos estejam todos posicionados numa mesma escala e a diferença inicial existente não interfira no algoritmo de segmentação.

⁸ CAOP significa Carta Administrativa Oficial de Portugal. A Carta Administrativa Oficial de Portugal regista o estado da delimitação e demarcação das circunscrições administrativas do País. A Assembleia da República é o organismo com competência, reconhecida por lei, para alterar e fixar limites administrativos. A Direção-Geral do Território (DGT) é responsável pela execução e manutenção da CAOP, de acordo com a alínea l) do n.º 2 do artigo 2.º do Decreto Regulamentar n.º 30/2012, de 13 de março. As competências da DGT, em matéria de delimitação administrativa estão circunscritas à determinação de limites para fins cadastrais e cartográficos. Os limites administrativos constantes na CAOP têm origem em diversas fontes de dados. Partindo da base de limites com origem nos Censos 2001, a CAOP tem vindo a ser atualizada com limites mais precisos, nomeadamente limites definidos nos diplomas de criação, extinção ou modificação de freguesias, limites constantes nas Secções de Cadastro Geométrico da Propriedade Rústica ou limites obtidos no âmbito dos Procedimentos de Delimitação Administrativa (PDA), através de acordo expresso por parte de todos os órgãos autárquicos envolvidos, dando origem às várias versões da CAOP publicadas anualmente desde 2001. (Fonte: Direção Geral do território, <<http://www.dgterritorio.pt>>, consultado em 15/10/2017).

Por outro lado, e derivado da interpretação também ser mais fácil, também várias transformações foram produzidas aos dados iniciais, nomeadamente transformações de variáveis de valor para percentagens e proporções, para que a sua leitura e interpretação pudesse ser mais fácil a correta.

Apresentam-se de seguida as tabelas com a base inicial de trabalho, ou seja, todas as variáveis que foram recolhidas para análise, distribuídas pelas categorias de informação:

População	Densidade populacional (N.º/ km²) por Local de residência (à data dos Censos 2011); Decenal (1)			N.º/ km²
	População residente (N.º) por Local de residência (à data dos Censos 2011),	Genero	HM	N.º
			H	N.º
População	População residente (N.º) por Local de residência (à data dos Censos 2011),	Grupo Etário	M	N.º
			Total	N.º
			Menos de 10 anos	N.º
			10 anos	N.º
			11 anos	N.º
			12 anos	N.º
			13 anos	N.º
			14 anos	N.º
			15 anos	N.º
			16 anos	N.º
			17 anos	N.º
			18 anos	N.º
			19 anos	N.º
			20 anos	N.º
			21 anos	N.º
			22 anos	N.º
			23 anos	N.º
			24 anos	N.º
			25 - 29 anos	N.º
			30 - 34 anos	N.º
			35 - 39 anos	N.º
			40 - 44 anos	N.º
			45 - 49 anos	N.º
			50 - 54 anos	N.º
			55 - 59 anos	N.º
			60 - 64 anos	N.º
			65 - 69 anos	N.º
			70 - 74 anos	N.º
			75 ou mais anos	N.º
População	População residente (N.º) por Local de residência (à data dos Censos 2011),	Nacionalidade	Total	N.º
			Portugal	N.º
			Estrangeira	N.º
			Europa	N.º
			África	N.º
			América	N.º
			Brasil	N.º
			Ásia	N.º
			Oceânia	N.º

Tabela 1 - Variáveis iniciais de *input* – categoria “População”

Condições de Vida	Índice de dependência de idosos (N.º) por Local de residência (à data dos Censos 2011); Decenal (1)			N.º
	Índice de dependência de jovens (N.º) por Local de residência (à data dos Censos 2011); Decenal (1)			N.º
	Índice de dependência total (N.º) por Local de residência (à data dos Censos 2011); Decenal (1)			N.º
	Índice de sustentabilidade potencial (N.º) por Local de residência (à data dos Censos 2011); Decenal (1)			N.º

Tabela 2 - Variáveis iniciais de *input* – categoria “Condições de Vida”

Emprego e Mercado de Trabalho	Taxa de actividade (%) da população residente por Local de residência (à data dos Censos 2011) e Sexo; Decenal		%
	Taxa de desemprego (%) por Local de residência (à data dos Censos 2011) e Sexo; Decenal		%
	Proporção de profissionais socialmente mais valorizados (%) por Local de residência (à data dos Censos 2011); Decenal		%
	Famílias clássicas (N.º) por Local de residência (à data dos Censos 2011)	Dimensão (indivíduos desempregados)	Total
			Famílias clássicas sem desempregados
			Famílias clássicas com desempregados
	Famílias clássicas (N.º) por Local de residência (à data dos Censos 2011)	Condição perante o trabalho (representante da família clássica)	População ativa
			Empregados
			Desempregados
			População inativa
	População residente (N.º) por Local de residência (à data dos Censos 2011), Sexo, Grupo etário e Grupo socioeconómico; Decenal	Grupo socioeconómico	Total
			Empresários com profissões intelectuais, científicas e técnicas
			Empresários da indústria comércio e serviços
			Empresários do sector primário
			Pequenos patrões com profissões intelectuais e científicas
			Pequenos patrões com profissões técnicas intermédias
			Pequenos patrões da indústria
			Pequenos patrões do comércio e serviços
			Pequenos patrões do sector primário
			Profissionais intelectuais e científicos independentes
			Profissionais técnicos intermédios independentes
			Trabalhadores industriais e artesanais independentes
			Prestadores de serviços e comerciantes independentes
			Trabalhadores independentes do sector primário
			Directores e quadros dirigentes do Estado e empresas
			Dirigentes de pequenas empresas e organizações
			Quadros intelectuais e científicos
			Quadros técnicos intermédios
			Quadros administrativos intermédios
			Empregados administrativos do comércio e serviços
			Operários qualificados e semi-qualificados
			Assalariados do sector primário
			Trabalhadores administrativos do comércio e serviços não qualificados
			Operários não qualificados
			Trabalhadores não qualificados do sector primário
			Pessoal das forças armadas
			Outras pessoas activas n.e.
			Inactivos

Tabela 3 - Variáveis iniciais de *input* – categoria “Emprego e Mercado de Trabalho”

Escolaridade	Proporção da população residente com ensino superior completo		%
	Proporção da população residente com 15 e mais anos de idade sem nenhum nível de escolaridade completo		%
	Taxa de abandono escolar		%
	Famílias clássicas (N.º) por Local de residência (à data dos Censos 2011)	Nível de escolaridade (representante da família clássica)	Total
			Nenhum nível de escolaridade
			Ensino básico
			Ensino secundário
			Ensino pós-secundário
			Ensino superior

Tabela 4 - Variáveis iniciais de *input* – categoria “Escolaridade”

Rendimento e Despesas Familiares	Encargos médios mensais por aquisição de habitação própria (€) por Localização geográfica (à data dos Censos 2011);	€
	Valor médio mensal das rendas dos alojamentos familiares clássicos arrendados (€) por Localização geográfica (à data	€

Tabela 5 - Variáveis iniciais de *input* – categoria “Rendimento e Despesas Familiares”

Habitação e Conforto	Densidade de alojamentos (N.º/ km²) por Localização geográfica (à data dos Censos 2011); Decenal			N.º/ km²
	Alojamentos familiares clássicos de residência habitual (N.º) por Localização geográfica (à data dos Censos 2011) e Regime de ocupação; Decenal	Regime de ocupação	Total	N.º
			Proprietário ou coproprietário	N.º
			Proprietário em regime de propriedade coletiva de cooperativa de habitação	N.º
			Arrendatário ou subarrendatário	N.º
			Outra situação	N.º
			Alojamentos familiares clássicos de residência habitual (N.º) por Localização geográfica (à data dos Censos 2011)	Escalaão de divisões
	Com 1 divisão	N.º		
	Com 2 divisões	N.º		
	Com 3 divisões	N.º		
	Com 4 divisões	N.º		
	Com 5 divisões	N.º		
	Com 6 divisões	N.º		
	Com 7 divisões	N.º		
	Com 8 divisões	N.º		
	Com 9 divisões	N.º		
	Com 10 ou mais divisões	N.º		
	Alojamentos familiares clássicos de residência habitual (N.º) por Localização geográfica (à data dos Censos 2011)	Dimensão (famílias no alojamento)	Total	N.º
			Alojamentos com 1 família	N.º
			Alojamentos com 2 famílias	N.º
			Alojamentos com 3 ou mais famílias	N.º
	Alojamentos familiares clássicos de residência habitual (N.º) por Localização geográfica (à data dos Censos 2011)	Escalaão de área útil por ocupante	Total	N.º
			Menos de 10 m²	N.º
			10 m² a menos de 15 m²	N.º
			15 m² a menos de 20 m²	N.º
			20 m² a menos de 30 m²	N.º
			30 m² a menos de 40 m²	N.º
			40 m² a menos de 60 m²	N.º
			60 m² a menos de 80 m²	N.º
			80 m² ou mais	N.º
	Alojamentos familiares clássicos de residência habitual (N.º) por Localização geográfica (à data dos Censos 2011)	Existência de ar condicionado	Total	N.º
			Com ar condicionado	N.º
			Sem ar condicionado	N.º
	Alojamentos familiares de residência habitual (N.º) por Localização geográfica (à data dos Censos 2011)	Existência de sistema de aquecimento e Fonte principal de energia utilizada para aquecimento	Total	N.º
			Com aquecimento	N.º
			Electricidade	N.º
Madeira, carvão ou outros combustíveis sólidos			N.º	
Petróleo, gasóleo ou outros combustíveis líquidos			N.º	
Gás natural, propano, butano ou outros combustíveis gasosos			N.º	
Outra (energia solar, geotérmica, ...)			N.º	
Idade média dos edifícios (Ano) por Localização geográfica (à data dos Censos 2011); Decenal			Ano	

Tabela 6 - Variáveis iniciais de *input* – categoria “Habitação e Conforto”

Sobre estas variáveis iniciais, várias transformações foram efetuadas, como por exemplo a transformação do conjunto de variáveis

- População residente (N.º) por Local de residência (à data dos Censos 2011)
 - Nacionalidade
 - Total
 - Portugal
 - Estrangeira
 - Europa
 - África
 - América
 - Brasil
 - Ásia
 - Oceânia

Em apenas duas variáveis que resultam em:

- $\text{Proporção População residente portuguesa} = (\text{População Residente - Nacionalidade - Portugal}) / (\text{População Residente - Nacionalidade - Total}) * 100$
- $\text{Proporção População residente estrangeira} = (\text{População Residente - Nacionalidade - Estrangeira}) / (\text{População Residente - Nacionalidade - Total}) * 100$

Todas as transformações efetuadas apresentam-se semelhantes à anteriormente descrita. O resultado final está caracterizado na tabela apresentada em seguida:

Segmentação	Categoria	Nome	Descrição	Tipo	Unidades
Segmentação Risco	População	POP1	Densidade populacional (N.º/ km²)	Nominal	Número
		POP2	Proporção População residente - genero masculino	Nominal	%
		POP3	Proporção População residente - genero feminino	Nominal	%
		POP4	Proporção População residente - Faixa Etária Menos de 10 anos	Nominal	%
		POP5	Proporção População residente - Faixa Etária - Entre 10 e 18 anos	Nominal	%
		POP6	Proporção População residente - Faixa Etária - Entre 18 e 30 anos	Nominal	%
		POP7	Proporção População residente - Faixa Etária - Entre 30 e 50 anos	Nominal	%
		POP8	Proporção População residente - Faixa Etária - Mais de 50 anos	Nominal	%
		POP9	Proporção População residente portuguesa	Nominal	%
		POP10	Proporção População residente estrangeira	Nominal	%
	Condições de Vida	CV1	Índice de dependência de idosos	Nominal	%
		CV2	Índice de dependência de jovens	Nominal	%
		CV3	Índice de dependência total	Nominal	%
		CV4	Índice de sustentabilidade potencial	Nominal	%
	Emprego e Mercado de Trabalho	EMT1	Taxa de atividade	Nominal	%
		EMT2	Taxa de desemprego	Nominal	%
		EMT3	Proporção de profissionais socialmente mais valorizados	Nominal	%
Segmentação Consumo	Habitação e Conforto	HC1	Densidade de alojamentos (N.º/ km²)	Nominal	Número
		HC2	Proporção Alojamentos com Escalão de divisões - 2 ou menos	Nominal	%
		HC3	Proporção Alojamentos com Escalão de divisões - 3 ou 4	Nominal	%
		HC4	Proporção Alojamentos com Escalão de divisões - 5 ou mais	Nominal	%
		HC5	Proporção Alojamentos com Dimensão (famílias no alojamento) - 1 família	Nominal	%
		HC6	Proporção Alojamentos com Dimensão (famílias no alojamento) - 2 ou mais famílias	Nominal	%
		HC7	Proporção Alojamentos com Escalão de área útil por ocupante - Menos de 20 m²	Nominal	%
		HC8	Proporção Alojamentos com Escalão de área útil por ocupante - 20 m² a menos de 60 m²	Nominal	%
		HC9	Proporção Alojamentos com Escalão de área útil por ocupante - 60 m² ou mais	Nominal	%
		HC10	Proporção Alojamentos com Ar Condicionado	Nominal	%
		HC11	Proporção Alojamentos sem Ar Condicionado	Nominal	%
		HC12	Proporção Alojamentos com Aquecimento	Nominal	%
		HC13	Proporção Alojamentos sem Aquecimento	Nominal	%
		HC14	Idade média dos edifícios (Ano)	Nominal	Anos
Segmentação Resposta	Educação	EDU1	Proporção da população residente com ensino superior completo	Nominal	%
		EDU2	Proporção da população residente com 15 e mais anos de idade sem nenhum nível de escolaridade completo	Nominal	%
		EDU3	Taxa de abandono escolar	Nominal	%
	Rendimento e Despesas	RD1	Encargos médios mensais por aquisição de habitação própria	Nominal	€
		RD2	Valor médio mensal das rendas dos alojamentos familiares clássicos arrendados	Nominal	€

Tabela 7 – Variáveis de *input* após transformações

Embora muitas transformações tenham sido efetuadas, ainda algumas variáveis não se encontram na mesma ordem de valores que as restantes.

Mas, dado que o próprio processo de segmentação no EM já produz automaticamente a normalização dos dados, não será necessário fazer transformações adicionais nesta fase.

3.2.2. Identificação de problemas nos dados

O trabalho de identificação de variáveis com valores *missing*, *outliers*, nulos, dados incongruentes ou incoerentes é fundamental para que todos estes problemas sejam tratados numa fase inicial não influenciam os resultados do estudo.

Foram utilizados os nós de *Input*⁹, *Multiplot*¹⁰ e *StatExplorer*¹¹ do EM para efetuar esta análise.

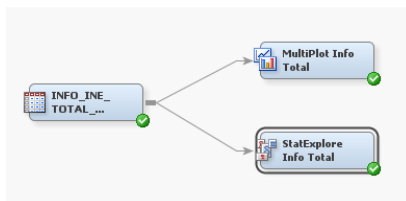


Figura 4 – Fluxo de AED com recurso ao EM

Através do nó de *Input* foram definidos o tipo e intervenção de cada variável no processo. Também foi identificada a descrição de cada variável bem como o tipo de variável (categórica ou nominal). A tabela 1 já apresentada, descreve a caracterização das 36 variáveis em estudo.

Em seguida aplicaram-se os nós *Multiplot* e *StatExplorer* de onde se obtiveram as estatísticas e visualizações necessárias para a análise das variáveis.

Começou-se por analisar estatísticas descritivas para cada uma das variáveis em estudo, como número de registos, valores médios, máximos, mínimos e desvio padrão, para poder ganhar sensibilidade sobre os dados e desde já perceber ou detetar algum problema nos dados, como valores em falta.

⁹ O nó *Input* pode ser alterado ao longo do processo, permitindo desta forma um maior controle sobre as mudanças nas variáveis entre as execuções sucessivas de um diagrama de fluxo do processo. Essa habilidade é importante quando as variáveis são adicionadas ou removidas dos dados de entrada. (Fonte: SAS Enterprise Miner Documentation, <<https://support.sas.com/documentation>>, consultado em 15/09/2017).

¹⁰ O nó *MultiPlot* é utilizado para produzir visualização de dados sobre várias perspetivas. O nó *MultiPlot* cria automaticamente gráficos de barras e gráficos de dispersão para as variáveis de entrada e objetivo, sem exigir a seleções específica dos itens. (Fonte: SAS Enterprise Miner Documentation, <<https://support.sas.com/documentation>>, consultado em 15/09/2017).

¹¹ O nó *StatExplore* serve para examinar as propriedades estatísticas de um conjunto de dados de entrada. (Fonte: SAS Enterprise Miner Documentation, <<https://support.sas.com/documentation>>, consultado em 15/09/2017).

Segmentação	Categoria	Nome	N	Mean	Minimum	Maximum	Std Dev
Segmentação Risco	População	POP1	211	4770.00	8.00	29495.00	5240.00
		POP2	211	47.56	42.16	58.08	1.93
		POP3	211	52.44	41.92	57.84	1.93
		POP4	211	9.86	3.07	16.52	2.25
		POP5	211	8.81	3.33	12.89	1.70
		POP6	211	12.72	8.09	20.23	1.71
		POP7	211	29.96	21.36	43.75	3.57
		POP8	211	38.64	12.92	56.15	6.65
		POP9	211	87.06	59.93	96.78	5.79
		POP10	211	12.94	3.22	40.07	5.79
	Condições de Vida	CV1	211	30.76	5.50	58.90	10.23
		CV2	211	22.65	9.50	34.80	4.41
		CV3	211	53.41	32.20	76.70	8.17
		CV4	211	3.70	1.70	18.10	1.62
	Emprego e Mercado de Trabalho	EMT1	211	49.27	37.23	64.91	3.58
		EMT2	211	12.62	6.54	27.60	3.13
		EMT3	211	27.25	7.62	62.83	13.30
Segmentação Consumo	Habitação e Conforto	HC1	211	2901.00	5.68	21943.00	3556.00
		HC2	211	3.73	0.45	22.99	3.43
		HC3	211	48.79	12.07	78.53	10.63
		HC4	211	47.48	11.99	86.45	11.82
		HC5	211	98.73	93.21	100.00	1.16
		HC6	211	1.27	0.00	6.79	1.16
		HC7	211	14.84	3.13	47.44	6.98
		HC8	211	60.91	43.29	73.56	5.10
		HC9	211	24.25	6.52	51.25	6.64
		HC10	211	11.67	0.98	42.70	6.88
		HC11	211	88.33	57.30	99.02	6.88
		HC12	211	84.13	57.40	95.20	5.97
		HC13	211	15.87	4.80	42.60	5.97
		HC14	211	45.01	18.50	107.08	18.44
Segmentação Resposta	Educação	EDU1	211	20.48	4.00	56.49	11.90
		EDU2	211	8.10	2.35	21.78	3.29
		EDU3	211	1.82	0.00	9.00	1.19
	Rendimento e Despesas	RD1	211	445.98	283.00	643.00	76.25
		RD2	211	268.01	103.00	440.00	67.03

Tabela 8 – Análise univariada das variáveis

Desta primeira análise pode-se salientar, de forma geral, que temos 211 registos, os quais representam as 211 freguesias incluídas na análise, e que a média registada para a densidade populacional são 4770 habitantes por Km². Também que a média da proporção de pessoas do género feminino é ligeiramente superior que o masculino; a média de proporção de pessoas na faixa etária “Mais de 50” é superior às restantes faixas etárias, o índice de dependência de idosos é superior ao índice de dependência de jovens, e a taxa de desemprego ronda, em média, os 12.62%, sendo que o nível de profissionais mais valorizados está em média no valor de 27.25% para a Região em análise

De seguida recorreu-se à análise visual de gráficos, como os histogramas, para deteção de valores extremos que pudessem deturpar o estudo.

Apresenta-se em seguida a análise da distribuição de duas das variáveis, sendo que os restantes se encontram em anexo.

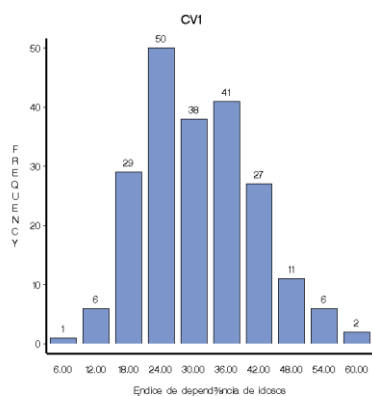


Gráfico 7 – Histograma da variável CV1

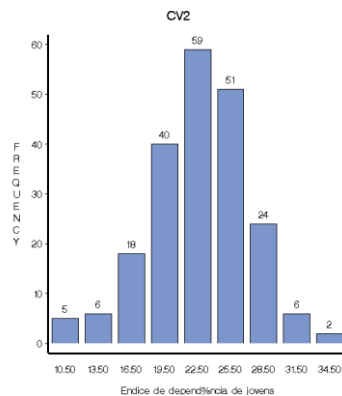


Gráfico 8 – Histograma da variável CV2

Como resultado da Análise Exploratória dos Dados podemos concluir que os mesmos não apresentam problemas de valores inexistentes ou muito fora do normal. Desta forma não foi necessário efetuar nenhum filtro nos dados e todos os registos iniciais foram mantidos para a segmentação.

3.2.3. Análise de correlações e seleção das variáveis

Também são analisadas as relações das variáveis entre si para deteção de correlações que possam existir com o objetivo de reduzir as variáveis de *input* e tratar o problema da multidimensionalidade.

Apresenta-se a análise de correlações entre as variáveis, divididas por cada tipo de segmentação.

Variáveis Segmentação Risco

Apresenta-se na tabela 3 a matriz de correlações para as variáveis selecionadas a entrar na segmentação de risco.

	POP1	POP2	POP3	POP4	POP5	POP6	POP7	POP8	POP9	POP10	CV1	CV2	CV3	CV4	EMT1	EMT2	EMT3
POP1	1.000	-0.065	0.065	-0.302	-0.211	0.012	-0.030	0.184	-0.225	0.225	0.192	-0.250	0.070	-0.245	-0.139	0.217	-0.257
POP2	-0.065	1.000	-1.000	-0.091	-0.274	0.622	0.778	-0.640	-0.830	0.830	-0.625	-0.341	-0.757	0.560	0.671	0.199	-0.410
POP3	0.065	-1.000	1.000	0.091	0.274	-0.622	-0.778	0.640	0.830	-0.830	0.625	0.341	0.757	-0.560	-0.671	-0.199	0.410
POP4	-0.302	-0.091	0.091	1.000	0.844	-0.213	0.020	-0.536	0.217	-0.217	-0.466	0.919	-0.030	0.526	-0.082	-0.034	0.217
POP5	-0.211	-0.274	0.274	0.844	1.000	-0.351	-0.248	-0.296	0.389	-0.389	-0.298	0.915	0.131	0.367	-0.321	-0.080	0.305
POP6	0.012	0.622	-0.622	-0.213	-0.351	1.000	0.498	-0.520	-0.545	0.545	-0.528	-0.454	-0.716	0.501	0.657	0.198	-0.409
POP7	-0.030	0.778	-0.778	0.020	-0.248	0.498	1.000	-0.792	-0.686	0.686	-0.733	-0.281	-0.834	0.670	0.877	0.005	-0.112
POP8	0.184	-0.640	0.640	-0.536	-0.296	-0.520	-0.792	1.000	0.471	-0.471	0.939	-0.238	0.795	-0.928	-0.708	-0.040	0.058
POP9	-0.225	-0.830	0.830	0.217	0.389	-0.545	-0.686	0.471	1.000	-1.000	0.457	0.428	0.636	-0.382	-0.602	-0.150	0.333
POP10	0.225	0.830	-0.830	-0.217	-0.389	0.545	0.686	-0.471	-1.000	1.000	-0.457	-0.428	-0.636	0.382	0.602	0.150	-0.333
CV1	0.192	-0.625	0.625	-0.466	-0.298	-0.528	-0.733	0.939	0.457	-0.457	1.000	-0.156	0.892	-0.958	-0.690	-0.124	0.090
CV2	-0.250	-0.341	0.341	0.919	0.915	-0.454	-0.281	-0.238	0.428	-0.428	-0.156	1.000	0.307	0.231	-0.367	-0.166	0.338
CV3	0.070	-0.757	0.757	-0.030	0.131	-0.716	-0.834	0.795	0.636	-0.636	0.892	0.307	1.000	-0.818	-0.833	-0.195	0.240
CV4	-0.245	0.560	-0.560	0.526	0.367	0.501	0.670	-0.928	-0.382	0.382	-0.958	0.231	-0.818	1.000	0.608	0.146	-0.094
EMT1	-0.139	0.671	-0.671	-0.082	-0.321	0.657	0.877	-0.708	-0.602	0.602	-0.690	-0.367	-0.833	0.608	1.000	-0.191	0.050
EMT2	0.217	0.199	-0.199	-0.034	-0.080	0.198	0.005	-0.040	-0.150	0.150	-0.124	-0.166	-0.195	0.146	-0.191	1.000	-0.817
EMT3	-0.257	-0.410	0.410	0.217	0.305	-0.409	-0.112	0.058	0.333	-0.333	0.090	0.338	0.240	-0.094	0.050	-0.817	1.000

Tabela 9 – Matriz de correlações variáveis segmentação risco

Da análise à matriz de correlações, e tendo salientado os valores de correlação mais elevados, ou seja, superiores a 0.65 e inferiores a -0.65, podemos observar que existem bastantes variáveis que possuem forte correlações entre si.

De notar que a variável POP1 não apresenta forte correlação com nenhuma outra variável e que as variáveis EMT1 e EMT2 só apresentam forte correlação uma com a outra, mas com nenhuma outra variável incluída no estudo. Nesta fase decidimos não retirar estas variáveis da análise.

Por outro lado, as restantes variáveis da categoria População não só apresentam fortes correlações entre si como com as variáveis da categoria Condições de Vida e com a variável EMT1. Dado que a interpretação dos segmentos pode ser mais complicada com um número elevado de variáveis, decidiu-se retirar variáveis que pudessem explicar o mesmo fenómeno.

Também, derivado aos problemas causados por variáveis que apresentam correlações perfeitas contrárias (-1), várias variáveis foram retiradas da análise derivado deste comportamento. Por exemplo, para as variáveis POP9 e POP10, percentagem de população residente portuguesa e percentagem de população residente estrangeira, que no seu total o somatório resulta em 100%, a primeira foi retirada da análise.

A seleção final de variáveis para a segmentação de risco esta representada na tabela 4.

Categoria	Nome	Descrição
População	POP1	Densidade populacional (N.º/ km²)
	POP10	Proporção População residente estrangeira
Condições de Vida	CV1	Índice de dependência de idosos
	CV2	Índice de dependência de jovens
Emprego e Mercado de Trabalho	EMT1	Taxa de actividade
	EMT2	Taxa de desemprego
	EMT3	Proporção de profissionais socialmente mais valorizados

Tabela 10 – Seleção final variáveis segmentação risco

Variáveis Segmentação Consumo

Apresenta-se na tabela 3 a matriz de correlações para as variáveis selecionadas a entrar na segmentação de risco.

	HC1	HC2	HC3	HC4	HC5	HC6	HC7	HC8	HC9	HC10	HC11	HC12	HC13	HC14
HC1	1.000	0.586	0.257	-0.416	-0.216	0.216	0.504	-0.270	-0.430	-0.434	0.434	-0.267	0.267	0.637
HC2	0.586	1.000	0.458	-0.725	0.010	-0.010	0.871	-0.539	-0.715	-0.606	0.606	-0.605	0.605	0.590
HC3	0.257	0.458	1.000	-0.945	0.306	-0.306	0.674	0.182	-0.802	-0.597	0.597	-0.449	0.449	0.114
HC4	-0.416	-0.725	-0.945	1.000	-0.241	0.241	-0.844	0.058	0.886	0.687	-0.687	0.571	-0.571	-0.306
HC5	-0.216	0.010	0.306	-0.241	1.000	-1.000	0.019	0.332	-0.159	0.105	-0.105	0.318	-0.318	-0.276
HC6	0.216	-0.010	-0.306	0.241	-1.000	1.000	-0.019	-0.332	0.159	-0.105	0.105	-0.318	0.318	0.276
HC7	0.504	0.871	0.674	-0.844	0.019	-0.019	1.000	-0.374	-0.922	-0.766	0.766	-0.766	0.766	0.426
HC8	-0.270	-0.539	0.182	0.058	0.332	-0.332	-0.374	1.000	-0.014	0.059	-0.059	0.295	-0.295	-0.293
HC9	-0.430	-0.715	-0.802	0.886	-0.159	0.159	-0.922	-0.014	1.000	0.802	-0.802	0.703	-0.703	-0.337
HC10	-0.434	-0.606	-0.597	0.687	0.105	-0.105	-0.766	0.059	0.802	1.000	-1.000	0.775	-0.775	-0.501
HC11	0.434	0.606	0.597	-0.687	-0.105	0.105	0.766	-0.059	-0.802	-1.000	1.000	-0.775	0.775	0.501
HC12	-0.267	-0.605	-0.449	0.571	0.318	-0.318	-0.766	0.295	0.703	0.775	-0.775	1.000	-1.000	-0.321
HC13	0.267	0.605	0.449	-0.571	-0.318	0.318	0.766	-0.295	-0.703	-0.775	0.775	-1.000	1.000	0.321
HC14	0.637	0.590	0.114	-0.306	-0.276	0.276	0.426	-0.293	-0.337	-0.501	0.501	-0.321	0.321	1.000

Tabela 11 - Matriz de correlações variáveis segmentação consumo

Neste caso, HC1 E HC14 não apresentam correlação com nenhuma outra variável em análise, bem como as variáveis HC5 e HC6 só apresentam correlação entre si mas com nenhuma outra variável.

Pelos motivos já anteriormente mencionados, a tabela 6 apresenta a lista de variáveis finais a entrar na segmentação de consumo.

Categoria	Nome	Descrição
	HC2	Proporção Alojamentos com Escalão de divisões - 2 ou menos
	HC4	Proporção Alojamentos com Escalão de divisões - 5 ou mais
	HC7	Proporção Alojamentos com Escalão de área útil por ocupante - Menos de 20 m²
	HC9	Proporção Alojamentos com Escalão de área útil por ocupante - 60 m² ou mais
	HC10	Proporção Alojamentos com Ar Condicionado
	HC12	Proporção Alojamentos com Aquecimento
	HC14	Idade média dos edifícios (Ano)

Tabela 12 - Seleção final variáveis segmentação consumo

Variáveis Segmentação Resposta

Por último, a tabela 7 apresenta a matriz de correlações para as variáveis selecionadas a entrar na segmentação de consumo.

	EDU1	EDU2	EDU3	RD1	RD2
EDU1	1.000	-0.878	-0.586	0.791	0.759
EDU2	-0.878	1.000	0.669	-0.670	-0.700
EDU3	-0.586	0.669	1.000	-0.497	-0.405
RD1	0.791	-0.670	-0.497	1.000	0.714
RD2	0.759	-0.700	-0.405	0.714	1.000

Tabela 13 - Matriz de correlações variáveis segmentação resposta

Neste caso todas as variáveis apresentam forte correlações entre si. A lista final manteve as mesmas variáveis, já que neste caso temos poucas variáveis iniciais e nenhuma delas apresenta correlações perfeitas entre elas.

A tabela 8 apresenta a seleção referida.

Categoria	Nome	Descrição
Educação	EDU1	Proporção da população residente com ensino superior completo
	EDU2	Proporção da população residente com 15 e mais anos de idade sem nenhum nível de escolaridade completo
	EDU3	Taxa de abandono escolar
Rendimento e Despesas	RD1	Encargos médios mensais por aquisição de habitação própria
	RD2	Valor médio mensal das rendas dos alojamentos familiares clássicos arrendados

Tabela 14 - Seleção final variáveis segmentação resposta

3.3. SELEÇÃO DO MÉTODO DE ANÁLISE DE *CLUSTERS* E VALIDAÇÃO DOS RESULTADOS

Nesta fase, de forma a conseguir captar as semelhanças/dissemelhanças entre as várias regiões em análise e mediante as variáveis de *input* selecionadas, algoritmos de Análise de *Clusters* serão utilizados para segmentação.

A aplicação *SAS Enterprise Miner* disponibiliza um conjunto vasto de algoritmos de Análise de *Clusters* que se poderá utilizar para comprovar os resultados obtidos. Para a segmentação iremos utilizar o método de *K-Means*.

O nó de **Cluster** foi utilizado para fazer a segmentação pretendida e o nó **SAS Code** foi utilizado para efetuar a computação do gráfico de cotovelo, o qual indica a evolução da variabilidade intra-segmentos à medida que o número de segmentos aumenta, ou seja, analisa a soma total das distâncias de cada ponto ao centroide do segmento alocado.

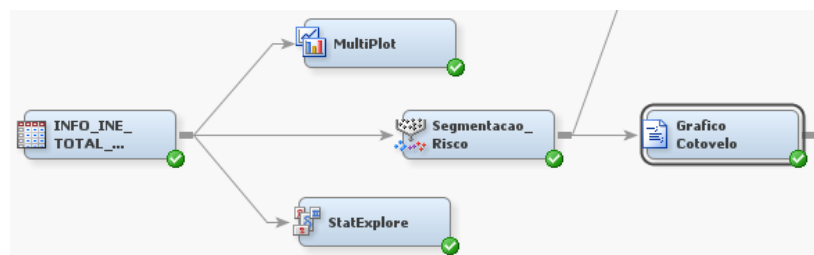


Figura 5 - Fluxo de AED e segmentação com recurso ao EM

3.3.1. Segmentação de Nível de Risco

Após seleção das variáveis, analisou-se o gráfico cotovelo.

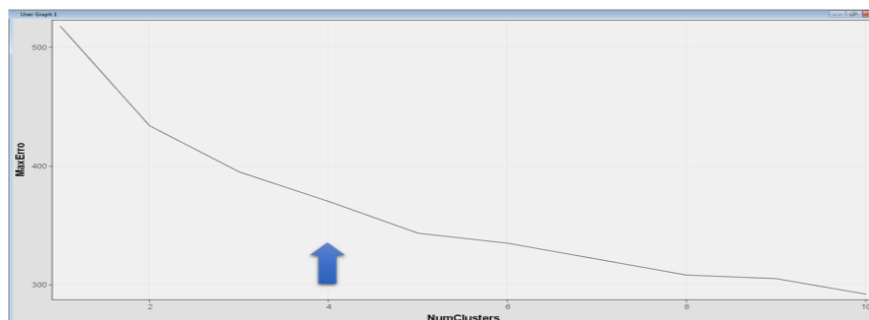


Gráfico 9 – Gráfico da variabilidade intra-cluster segmentação risco

Através do gráfico pode-se concluir que para esta segmentação se devem considerar quatro segmentos distintos, pois verifica-se neste ponto que o declive da curva diminui nesse ponto, ou seja, o ganho em termos de coesão dos segmentos não justifica a criação de um segmento adicional. Isto significa que a complexidade vai aumentando mas o nível de erro a partir do quarto segmentos não decresce substancialmente, de forma a justificar a introdução de maior complexidade na análise.

Após a escolha inicial do número de segmentos a utilizar procede-se à análise do tamanho de cada segmento, para verificar se todos os segmentos contruídos se caracterizam por deter observações suficientemente significativas. Através da análise ao gráfico 10 e tabela 9 podemos verificar que o segmento 3 apresenta um número mais reduzido de observações, em comparação com os outros segmentos.

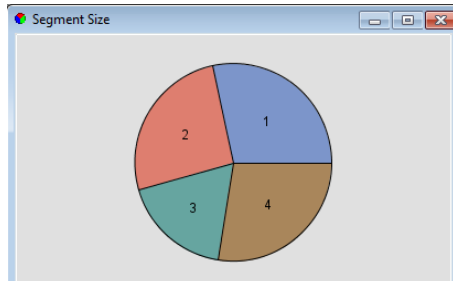


Gráfico 10 – Gráfico de distribuição registos para 4 segmentos risco

Frequencies: _SEGMENT_

Segment Variable	Segment Value	Frequency Count	Percent of Total Frequency
SEGMENT	1	60	28.4360
SEGMENT	4	58	27.4882
SEGMENT	2	55	26.0664
SEGMENT	3	38	18.0095

Tabela 15 - Tabela de distribuição registos para 4 segmentos risco

Desta forma, opta-se por reduzir o número de segmentos para 3. O gráfico 11 e tabela 10 apresentam essa distribuição, que se regista ser mais equilibrada.

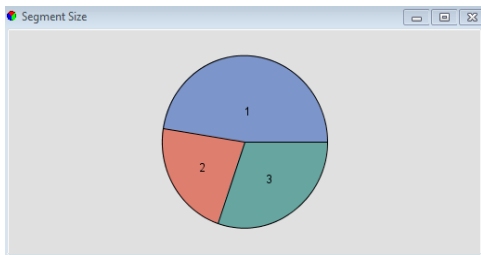


Gráfico 11 - Gráfico de distribuição registos para 3 segmentos risco

Frequencies: _SEGMENT_

Segment Variable	Segment Value	Frequency Count	Percent of Total Frequency
SEGMENT	1	100	47.3934
SEGMENT	3	64	30.3318
SEGMENT	2	47	22.2749

Tabela 16 - Tabela de distribuição registos para 3 segmentos risco

Pela análise da importância das variáveis selecionadas em cada segmento (gráficos 12, 13 e 14), podemos comprovar que nos segmentos mais significativos (segmentos 1 e 2) todas as variáveis selecionadas apresentam relevância nos mesmos.

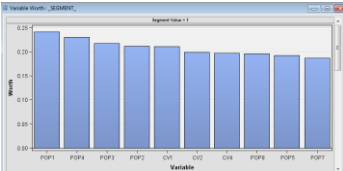


Gráfico 12 – Importância variáveis segmento 1 risco

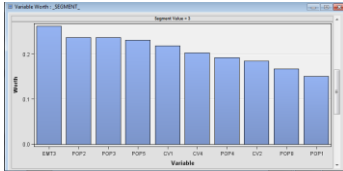


Gráfico 13 - Importância variáveis segmento 2 risco



Gráfico 14 - Importância variáveis segmento 3 risco

Como resultado para a segmentação de risco obtivemos 3 segmentos. A tabela seguinte apresenta a caracterização dos segmentos através dos valores médios obtidos em cada um e por cada variável em análise.

Segmento	Índice de dependência de idosos	Índice de dependência de jovens	Taxa de actividade	Taxa de desemprego	Proporção de profissionais socialmente mais valorizados	Densidade populacional (N.º/km²)	Proporção População residente estrangeira
1	24.4790	25.3810	50.8250	11.4295	22.6794	1734.5600	10.2100
2	31.0702	22.1745	47.3021	16.7743	16.9489	5924.2553	16.9626
3	40.3594	18.7344	48.2753	11.4255	41.9616	8663.5781	14.2551
Mean	30.7640	22.6507	49.2669	12.6188	27.2516	4769.5024	12.9411

Tabela 17 – Valores médios obtidos para os segmentos de risco

3.3.2. Segmentação de Nível de Consumo

Igualmente neste caso, foi efetuada a análise do gráfico do cotovelo.

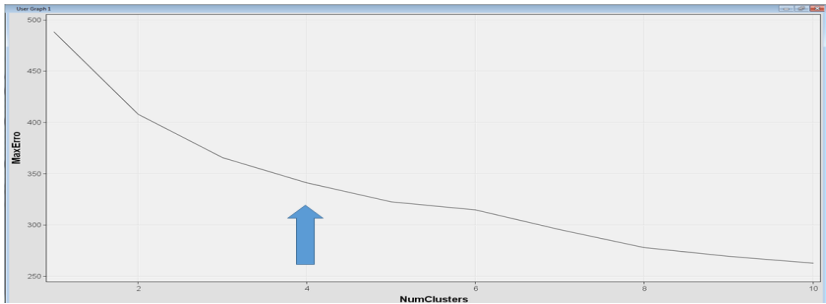


Gráfico 15 – Gráfico da variabilidade intra-cluster segmentação consumo

Também aqui, se valida que para esta segmentação se devem considerar quatro segmentos distintos, pelas mesmas razões anteriormente expostas.

Pela análise da distribuição de registos em cada segmento, também aqui se conclui que os segmentos 1 e 4 apresentam um número de registos inferiores aos segmentos 2 e 3. Pelo que se opta por reduzir o número de segmentos para 3.

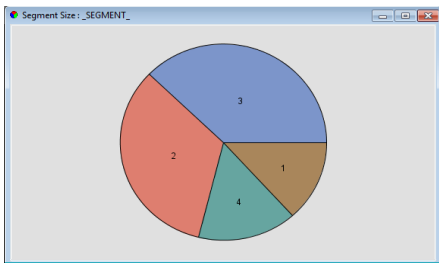


Gráfico 16 - Gráfico de distribuição registos para 4 segmentos consumo

Frequencies: `_SEGMENT_`

Segment Variable	Segment Value	Frequency Count	Percent of Total Frequency
<code>_SEGMENT_</code>	3	80	37.9147
<code>_SEGMENT_</code>	2	70	33.1754
<code>_SEGMENT_</code>	4	33	15.6398
<code>_SEGMENT_</code>	1	28	13.2701

Tabela 18 - Tabela de distribuição registos para 4 segmentos consumo

Pelo que se opta por reduzir o número de segmentos para 3.

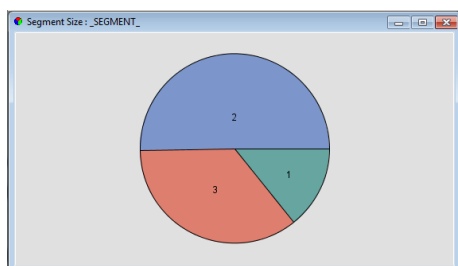


Gráfico 17 - Gráfico de distribuição registros para 3 segmentos consumo

Frequencies: _SEGMENT_

Segment Variable	Segment Value	Frequency Count	Percent of Total Frequency
SEGMENT	2	106	50.2370
SEGMENT	3	75	35.5450
SEGMENT	1	30	14.2180

Tabela 19 - Tabela de distribuição registros para 3 segmentos consumo

Embora persista um segmento com um número reduzido de registros, pode-se verificar que pela análise da importância das variáveis em cada segmento (gráficos 18, 19 e 20), todas as variáveis apresentam relevância em todos os segmentos.



Gráfico 18 - Importância variáveis segmento 1 consumo

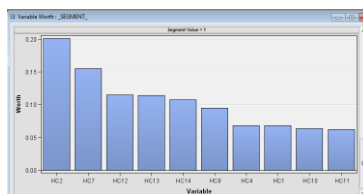


Gráfico 19 - Importância variáveis segmento 2 consumo

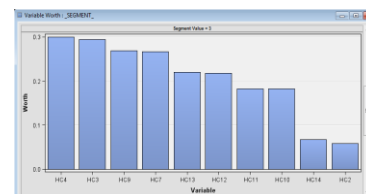


Gráfico 20 - Importância variáveis segmento 3 consumo

Desta forma, também na segmentação de consumo obtivemos 3 segmentos. A tabela seguinte apresenta a caracterização dos segmentos através dos valores médios obtidos em cada um e por cada variável em análise.

Segmento	Proporção Alojamentos com Ar Condicionado	Proporção Alojamentos com Aquecimento	Idade média dos edifícios (Ano)	Proporção Alojamentos com Escalão de divisões 2 ou menos	Proporção Alojamentos com Escalão de divisões 5 ou mais	Proporção Alojamentos com Escalão de área útil por ocupante - Menos de 20 m²	Proporção Alojamentos com Escalão de área útil por ocupante - 60 m² ou mais
1	6.0273	75.6980	70.1917	10.1673	34.4702	26.7155	18.9722
2	9.4894	83.4986	40.2036	2.7965	43.6578	15.3389	21.4151
3	17.0145	88.3875	41.7240	2.4648	58.0926	9.3982	30.3595
Mean	11.6720	84.1272	45.0077	3.7266	47.4824	14.8448	24.2470

Tabela 20 - Valores médios obtidos para os segmentos de consumo

3.3.3. Segmentação de Nível de Resposta

A mesma análise se efetuou para a segmentação de caracterização do consumo. Novamente, o gráfico do cotovelo salientou a escolha de quatro segmentos neste caso. Mas também aqui se

verificou a existência de dois segmentos com número menor de registros, pelo que a segmentação final resultou em três segmentos.

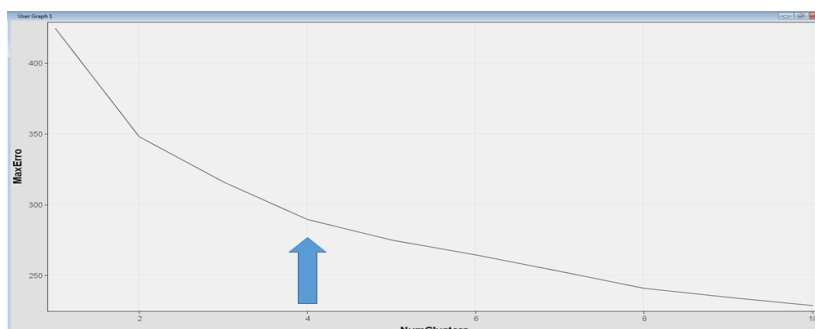


Gráfico 18 - Gráfico da variabilidade intra-cluster segmentação resposta

O Gráfico 22 e Tabela 15 apresentam a análise inicial de escolha de quatro segmentos.

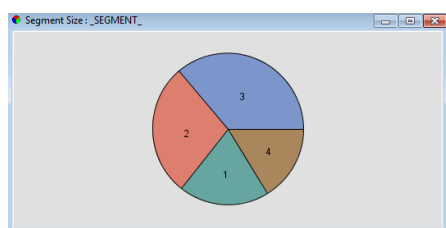


Gráfico 19 - Gráfico de distribuição registros para 4 segmentos resposta

Frequencies: _SEGMENT_

Segment Variable	Segment Value	Frequency Count	Percent of Total Frequency
SEGMENT	3	76	36.0190
SEGMENT	2	60	28.4360
SEGMENT	1	41	19.4313
SEGMENT	4	34	16.1137

Tabela 21 - Tabela de distribuição registros para 4 segmentos resposta

Os Gráfico 23 e Tabela 16 apresentam a distribuição de registros final pelos três segmentos obtidos.

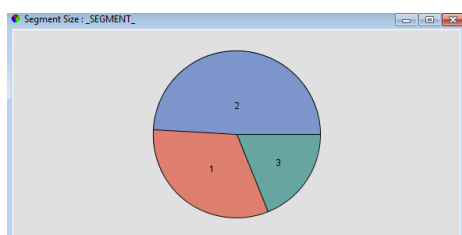


Gráfico 20 - Gráfico de distribuição registros para 3 segmentos resposta

Frequencies: _SEGMENT_

Segment Variable	Segment Value	Frequency Count	Percent of Total Frequency
SEGMENT	2	104	49.2891
SEGMENT	1	67	31.7536
SEGMENT	3	40	18.9573

Tabela 22 - Tabela de distribuição registros para 3 segmentos resposta

A Tabela 17 apresenta a caracterização dos segmentos através dos valores médios obtidos em cada um e por cada variável em análise

Segmento	Proporção da população residente com ensino superior completo	Proporção da população residente com 15 e mais anos de idade sem nenhum nível de escolaridade completo	Taxa de abandono escolar	Encargos médios mensais por aquisição de habitação própria	Valor médio mensal das rendas dos alojamentos familiares clássicos arrendados
1	11.0960	11.1279	2.6501	395.5918	209.2025
2	19.2427	7.4203	1.5447	450.0558	273.7914
3	39.4280	4.7780	1.1405	519.7540	351.4298
Mean	20.4824	8.0967	1.8191	445.9745	268.0003

Tabela 23 - Valores médios obtidos para os segmentos de resposta

4. RESULTADOS E DISCUSSÃO

O objetivo final prende-se com a transformação das segmentações em caracterizações que permitam auxiliar a entidade no planeamento de ações e definição da estratégia comercial a implementar. Essa caracterização e interpretação torna-se crucial neste projeto dado que dela depende a aplicabilidade empresarial do projeto.

Esta caraterização foi efetuada com auxílio ao nó *Segment Profile* do EM.

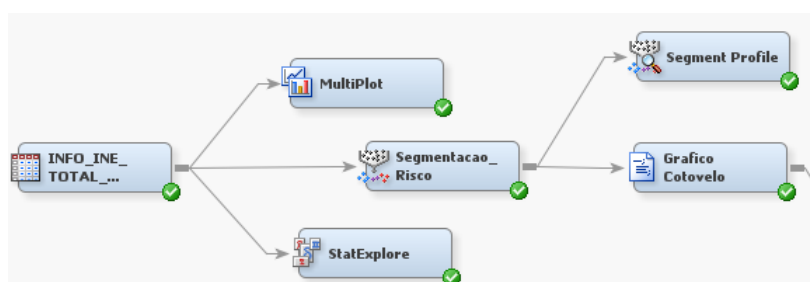


Figura 6 - Fluxo de AED, Segmentação e Caraterização dos segmentos com recurso ao EM

A finalidade será a produção de um conjunto de segmentos que se caracterizem por apresentar melhores condições ao nível de:

- Nível de risco
- Nível de consumo
- Nível de resposta

Através da obtenção destes segmentos a entidade poderá abordar zonas que se apresentem mais atrativas, ou seja, zonas onde os residentes apresentam características onde se possa deduzir que as condições de vida são melhores, com características que indiretamente se reflitam em maiores níveis de consumo e apresentem níveis de rendimento mais elevados.

4.1. SEGMENTAÇÃO RISCO

Para analisar e caracterizar os segmentos obtidos, apresenta-se o gráfico com a análise comparativa entre as médias normalizadas de cada variável por segmento e para o total da amostra. Através da análise do gráfico produziu-se a caraterização dos segmentos por nível de risco.

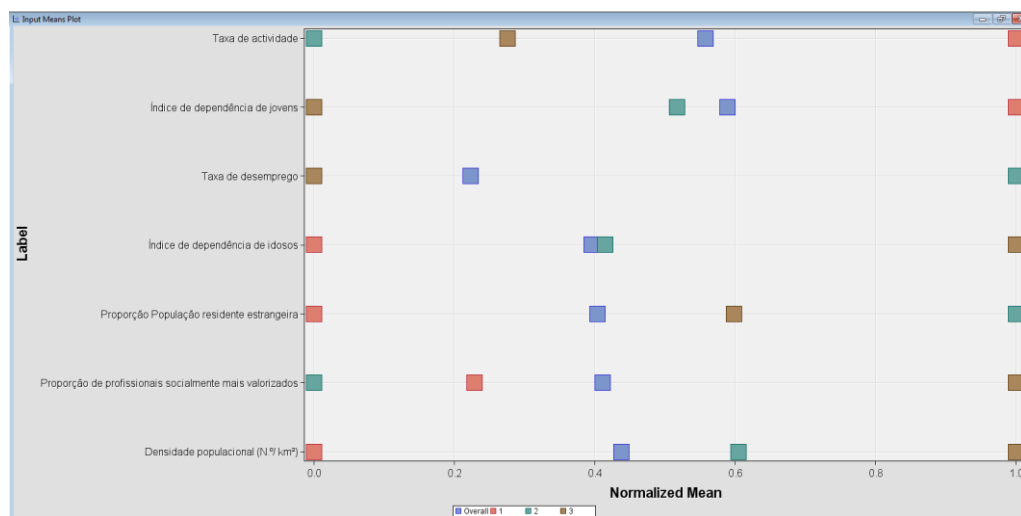


Gráfico 21 – Gráfico de médias normalizadas para segmentação risco

Da análise do gráfico, chegamos a uma caracterização dos segmentos que se apresenta em seguida:

Segmento 1 - Médio Risco

Zonas de menor densidade populacional mas com elevada concentração de jovens o que a torna uma zona de atividade elevada. Taxa de desemprego abaixo da média e existência de profissionais mais valorizados em linha com a média da população em análise.

Segmento 2 - Alto Risco

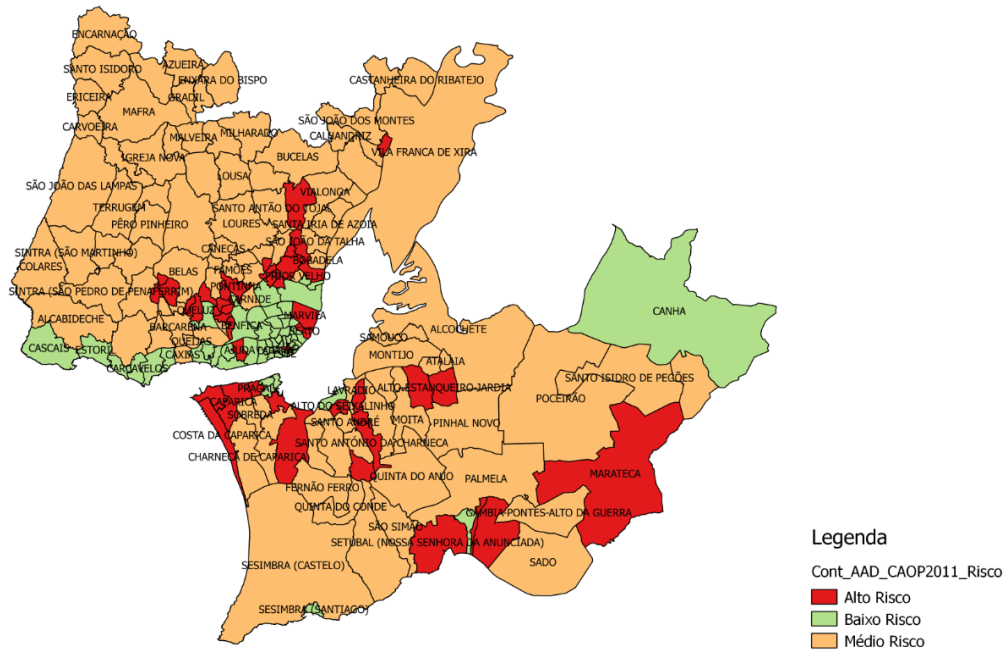
Zonas da densidade populacional acima da média mas forte concentração de estrangeiros. Taxa de desemprego com os valores mais elevados e taxa de atividade baixa. Com valores centrados na média no que respeita á concentração de jovens e idosos.

Segmento 3 - Baixo Risco

Zonas de maior densidade populacional, com indicação e existência de população estrangeira mas com alta concentração de segmentos de idade mais elevados e de profissionais socialmente mais valorizados. Taxa de desemprego nos valores mais baixos mas também taxa de atividade em valores abaixo da média com indicação de concentração de jovens em valores baixos.

O Mapa 3 apresenta a distribuição dos segmentos acima descritos pelas distintas freguesias da RML.

Segmentação Risco



Mapa 3 – Caracterização geográfica da segmentação de risco

4.2. SEGMENTAÇÃO CONSUMO

Para analisar e caracterizar os segmentos obtidos, apresenta-se o gráfico com a análise comparativa entre as médias normalizadas de cada variável por segmento e para o total da amostra. Através da análise do gráfico produziu-se a caracterização dos segmentos por nível de consumo.

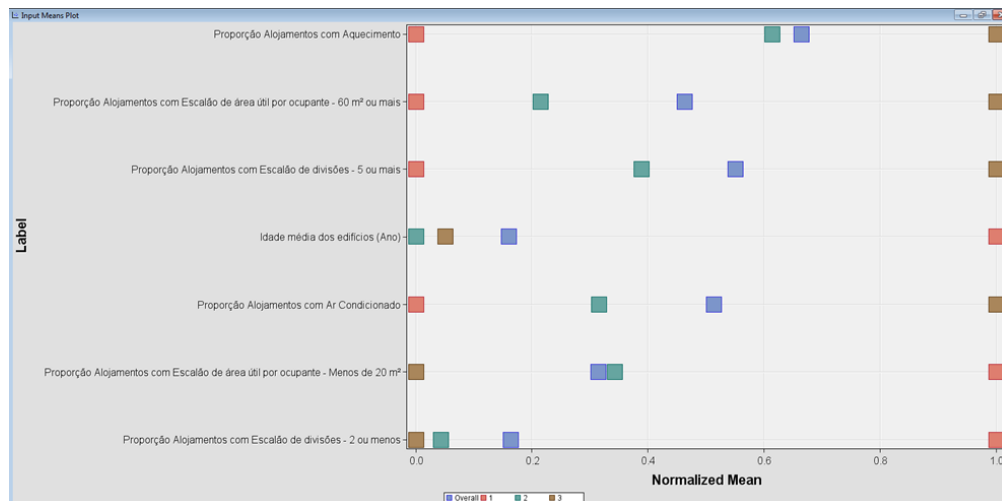


Gráfico 22 - Gráfico de médias normalizadas para segmentação consumo

Da análise do gráfico, chegamos a uma caracterização dos segmentos que se apresenta em seguida:

Segmento 1 - Baixo Consumo

Zonas onde existe uma maior concentração de alojamentos mais pequenos e com existência de menos sistemas de aquecimento ou ar condicionado. A idade média dos edifícios neste segmento também se apresenta superior à média da amostra.

Segmento 2 - Médio Consumo

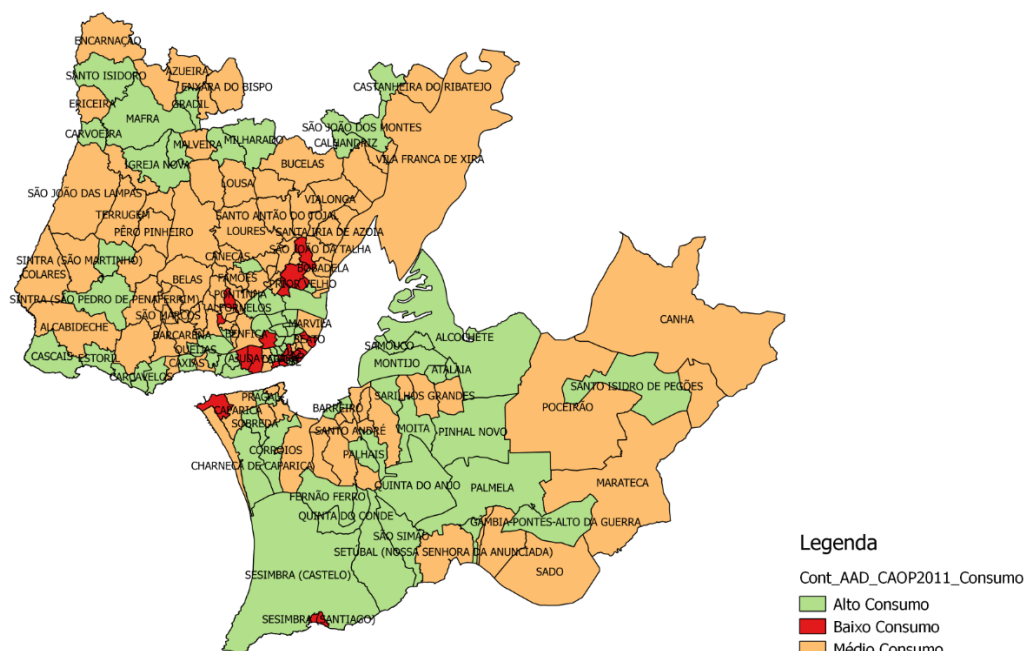
Zonas com alojamentos com menor antiguidade face à média da amostra e existência alojamentos com sistemas de aquecimento e ar condicionado de acordo com os valores médios da amostra.

Segmento 3 - Alto Consumo

Zonas com alojamentos de dimensão e espaço por ocupante em valores superiores à média e onde se verifica a existência, acima da média, de sistemas de aquecimento e ar condicionado. Também se verifica que a média obtida para a idade dos edifícios se encontra nos valores mais baixos.

A segmentação de consumo obtida apresenta-se distribuída pela RML, de acordo com o Mapa apresentado em seguida.

Segmentação Consumo



Mapa 4 - Caracterização geográfica da segmentação de consumo

4.3. SEGMENTAÇÃO RESPOSTA

Para analisar e caracterizar os segmentos obtidos, apresenta-se o gráfico com a análise comparativa entre as médias normalizadas de cada variável por segmento e para o total da amostra. Através da análise do gráfico produziu-se a caraterização dos segmentos por nível de resposta.

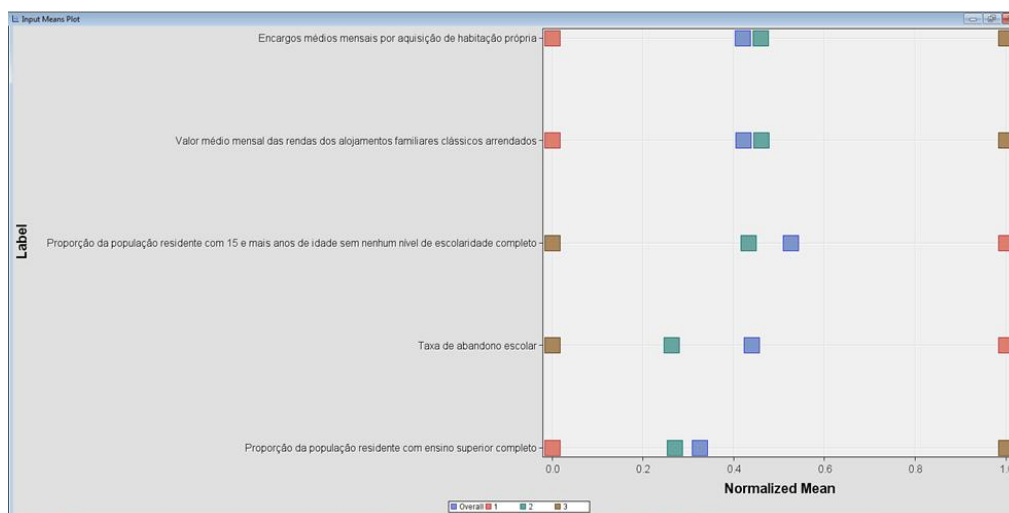


Gráfico 23 - Gráfico de médias normalizadas para segmentação resposta

Da análise do gráfico, chegamos a uma caracterização dos segmentos que se apresenta em seguida:

Segmento 1 - Baixa Resposta

Zonas onde existe uma maior concentração de alojamentos mais pequenos e com existência de menos sistemas de aquecimento ou ar condicionado. A idade média dos edifícios também se apresenta superior à média da amostra.

Segmento 2 - Médio Resposta

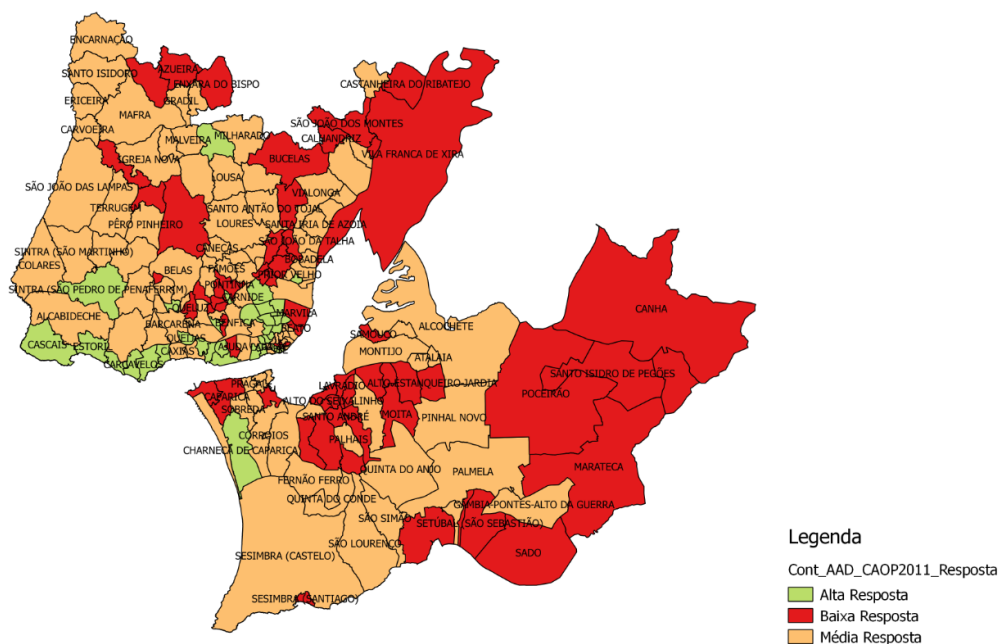
Zonas com alojamentos com menor antiguidade e existência alojamentos com sistemas de aquecimento e ar condicionado de acordo com os valores médios da amostra.

Segmento 3 - Alto Resposta

Zonas com alojamentos de dimensão e espaço por ocupante superiores e onde se verifica a existência, acima da média, de sistemas de aquecimento e ar condicionado. Também se verifica que a média obtida para a idade dos edifícios se encontra nos valores mais baixos.

A distribuição desta caracterização pela RML apresenta-se no Mapa 5.

Segmentação Resposta



Mapa 5 - Caracterização geográfica da segmentação de resposta

4.4. SEGMENTAÇÃO FINAL

A segmentação final junta as três vertentes de análise e tem como objetivo caracterizar as distintas freguesias de forma a potenciar de conhecimento à entidade que está a entrar no mercado e pretende definir as suas estratégias de atuação no mercado residencial. O foco poderão ser freguesias que apresentem níveis de baixo risco, níveis de alto consumo elevado e alta capacidade de resposta mas também é importante ter a capacidade de definir diferentes abordagens mediante os níveis de segmentação obtidos.

O critério de agregação final teve em conta os três segmentos obtidos para cada segmentação e foi definido da seguinte forma:

1. Definição de todas as possíveis combinações sobre as três vertentes de segmentação;
2. Atribuição de pontuação a cada nível de segmentação e para cada uma das vertentes de análise - classificação em valores compreendidos entre 1 e 3, com saltos de um ponto entre eles. A categoria 1 representava a melhor pontuação nas três vertentes de análise (menor nível de risco, maior nível de consumo, maior nível de resposta) e a categoria 3 a pior pontuação (maior nível de risco, menor nível de consumo, menor nível de resposta).
3. Cálculo de uma pontuação auxiliar, que representa uma média ponderada das três classificações anteriormente obtidas, sendo que para a pontuação da segmentação de risco se atribuiu uma ponderação ligeiramente superior que as restantes por se considerar esta categorização de maior importância que as restantes na caracterização final.

Calculo Auxiliar = 0,4 * Pontuação Risco + 0,3 * Pontuação Consumo + 0,3 * Pontuação Resposta

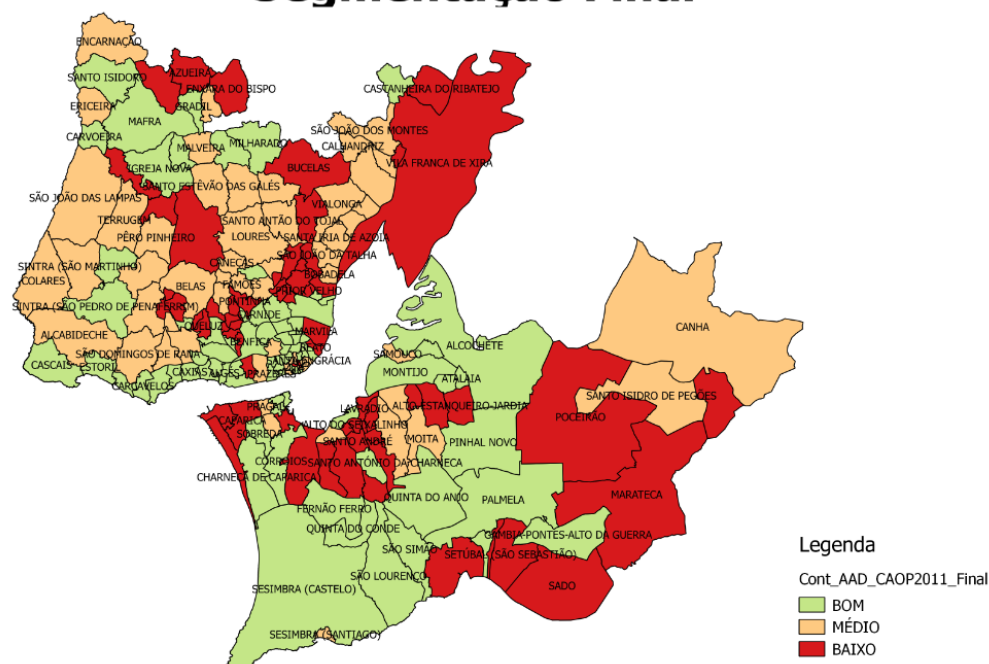
4. A pontuação final obtida para todas as combinações foi ordenada e dividida em três categorias – BOM, MÉDIO e MAU.
5. BOM reflete os resultados de pontuações mais baixas obtidas na média ponderada e MAU pontuações mais altas obtidas na média ponderada. MÉDIO reflete os valores intermédios de classificação final obtidos. A tabela seguinte mostra a forma de obtenção das três categorias descritas.

Segmento Risco	Segmento Consumo	Segmento Resposta	Calculo Auxiliar	Segmento Final	Descrição
1	1	1	1	1	1 BOM
1	1	2	1.3	1	1 BOM
1	2	1	1.3	1	1 BOM
2	1	1	1.4	1	1 BOM
1	1	3	1.6	1	1 BOM
1	3	1	1.6	1	1 BOM
1	2	2	1.6	1	1 BOM
2	2	1	1.7	1	1 BOM
2	1	2	1.7	1	1 BOM
3	1	1	1.8	2	2 MÉDIO
1	2	3	1.9	2	2 MÉDIO
1	3	2	1.9	2	2 MÉDIO
2	1	3	2	2	2 MÉDIO
2	2	2	2	2	2 MÉDIO
2	3	1	2	2	2 MÉDIO
3	1	2	2.1	2	2 MÉDIO
3	2	1	2.1	2	2 MÉDIO
1	3	3	2.2	2	2 MÉDIO
2	2	3	2.3	3	3 MAU
2	3	2	2.3	3	3 MAU
3	3	1	2.4	3	3 MAU
3	1	3	2.4	3	3 MAU
3	2	2	2.4	3	3 MAU
2	3	3	2.6	3	3 MAU
3	2	3	2.7	3	3 MAU
3	3	2	2.7	3	3 MAU
3	3	3	3	3	3 MAU

Tabela 24 – Pontuação Segmentação Final

1. O passo seguinte foi refletir esta classificação para todos os registos incluídos na segmentação. A distribuição geográfica da segmentação final obtida mediante o método descrito apresenta-se no Mapa 6.

Segmentação Final



Mapa 6 - Caracterização geográfica da segmentação final

5. CONCLUSÕES

Derivado da Liberalização do mercado e da conjuntura económica vivida no ano de 2013, uma entidade estrangeira mostrou a necessidade de obter um maior conhecimento sobre o mercado português. Pretendia definir as suas estratégias de forma mais informada e consciente e tomar as decisões mais acertadas com base em informação sobre o seu mercado alvo.

Desta forma, o objetivo deste trabalho derivou no desenvolvimento de segmentações geodemográficas, que por um lado pudessem dividir zonas em vários grupos com hábitos e comportamentos distintos de forma a justificar estratégias diferentes.

Decidiu-se dividir a segmentação em três categorias distintas: risco, consumo e resposta. Por um lado pretendia-se obter a caracterização qualitativa da população, com base em atributos que pudessem identificar características positivas e negativas para cada segmento, por outro obter uma divisão de grupos por caracterização de níveis de consumo associados a energia, e por fim uma divisão de grupos que pudessem separar aqueles com maiores capacidades financeiras daqueles com menores capacidades financeiras.

Como resultado final, através da junção das três caracterizações foi efetuada uma classificação final que passa a segmentar pelas três vertentes de análise anteriormente descritas.

Ao devolver estes resultados à entidade, esta poderá analisar os mesmos e definir as suas estratégias de acordo com os segmentos obtidos. Poderá alocar melhor os seus recursos comerciais em zonas que apresentem melhores resultados e identificar estratégias diferenciadas e de precaução para as zonas identificadas com pior classificação.

Este trabalho pretende demonstrar que o recurso a informação pública, sem custos, e o tratamento adequado sobre essa informação podem trazer mais-valias ao negócio de qualquer entidade, principalmente quando não existem dados internos para trabalhar e explorar.

Estando numa era de extração de conhecimento sobre grandes volumes de dados e numa era em que muita informação está disponível, sendo a tendência de crescimento, ter capacidades e ferramentas de tratamentos de grandes volumes são essenciais para o sucesso de qualquer entidade.

6. LIMITAÇÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Tendo o trabalho sido iniciado na época de 2013/2014, e ter sido interrompido por causalidades da vida, a maior dificuldade foi a capacidade de retomar o trabalho após essa interrupção. Voltar a rever toda a temática e validar o trabalho já desenvolvido foi complicado derivado do tempo transcorrido. Validar que a formulação do problema e a necessidade identificada continuariam validas, sem prejudicar o trabalho que havia sido desenvolvido foi um desafio adicional na hora de conjugar os dois períodos de execução do trabalho

Por outro lado, como recomendação deixo alguns apontamentos sobre o que poderia ser adicionalmente desenvolvido no âmbito deste trabalho. Os desenvolvimentos e investigações mais recentes referentes a novos algoritmos de segmentação, com maiores capacidades de ligação com o próprio comportamento humano, como é o caso das redes neuronais. Algoritmos como o SOM ou HSOM poderiam ter sido validados como possíveis alternativas ao famoso e amplamente utilizado algoritmo de Análise de Clusters K-Means, evidenciando mais valias na interpretação dos resultados obtidos e quando estão em tratamento grandes volumes de dados.

7. BIBLIOGRAFIA

- Albert, A., Gebru, T., & Ku, J. (2013). Drivers of Variability in Energy Consumption. *Proceedings of the 1st ECML/PKDD International Workshop on Data Analytics for Renewable Energy Integration (DARE '13)*, 1–12. Retrieved from http://www.stanford.edu/~adalbert/papers/drivers_variability.pdf
- Baçaõ, F. (2012). DATA MINING.
- Baçaõ, F., & Loureiro, M. (1995). O SELF-ORGANIZING MAP COMO FERRAMENTA NA ANÁLISE GEO-DEMOGRÁFICA.
- Bal, M., Bal, Y., & Demirhan, A. (2011). Creating Competitive Advantage by Using Data Mining Technique as an Innovative Method for Decision Making Process in Business. *International Journal of Online Marketing*, 1(3), 38–45. <https://doi.org/10.4018/ijom.2011070104>
- Birkin, M. (1995). Customer targeting, geodemographics and lifestyle approaches.
- Brimicombe, A. J. (2007). A dual approach to cluster discovery in point event data sets. *Computers, Environment and Urban Systems*, 31(1), 4–18. <https://doi.org/10.1016/j.compenvurbsys.2005.07.004>
- Dibb, S. (1998). Market segmentation: strategies for success. *Marketing Intelligence & Planning*, 16(7), 394–406. <https://doi.org/10.1108/02634509810244390>
- ERSE. (2013). Relatório Mercado Liberalizado de Eletricidade - Novembro 2013.
- Goss, J. (1995). “ We Know Who You Are and We Know Where You Live”: The Instrumental Rationality of Geodemographic Systems. *Economic Geography*. Retrieved from <http://www.jstor.org/stable/10.2307/144357>
- Grekousis, G., & Thomas, H. (2012). Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The fuzzy C-means and Gustafson-Kessel methods. *Applied Geography*, 34, 125–136. <https://doi.org/10.1016/j.apgeog.2011.11.004>
- Longley, P., & Clarke, G. (1995). *GIS for business and service planning*. Retrieved from <http://www.lavoisier.fr/livre/notice.asp?ouvrage=1343459>
- Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(233), 281–297. <https://doi.org/citeulike-article-id:6083430>
- McCarthy, E., & Perreault, W. (1991). *Essentials of marketing*. Retrieved from <http://www.getcited.org/pub/102647486>
- NASCIMENTO, J. DO. (2004). Ambiente Integrado de Colaboração Corporativa: Estratégia de Apoio a Gestão do Conhecimento. ... *E GESTÃO DO CONHECIMENTO NA ...*, 1–14. Retrieved from <http://www.redciencia.cu/empres/Intempres2004/Sitio/Ponencias/2.pdf>
- SCHWABE, C., & O'DONOVAN, M. (1997). THE USE OF CENSUS DATA TO DEVELOP A GEODEMOGRAPHIC OR LIFESTYLE SEGMENTATION SYSTEM FOR SOUTH AFRICA. *Southern African Journal of Demography*. Retrieved from <http://www.jstor.org/stable/10.2307/20853245>
- Sivadas, E. (1997). A preliminary examination of the continuing significance of social class to marketing: a geodemographic replication. *Journal of Consumer Marketing*, 14(6), 463–479.

<https://doi.org/10.1108/07363769710186097>

Spielman, S. E., & Thill, J.-C. (2008). Social area analysis, data mining, and GIS. *Computers, Environment and Urban Systems*, 32(2), 110–122.

<https://doi.org/10.1016/j.compenvurbsys.2007.11.004>

Tanner Jr., J. F., & Raymond, M. A. (2011). Market Segmenting , Targeting , and Positioning. In *PRINCIPLES OF MARKETING AND BUSINESS COMMUNICATION* (Flat World).

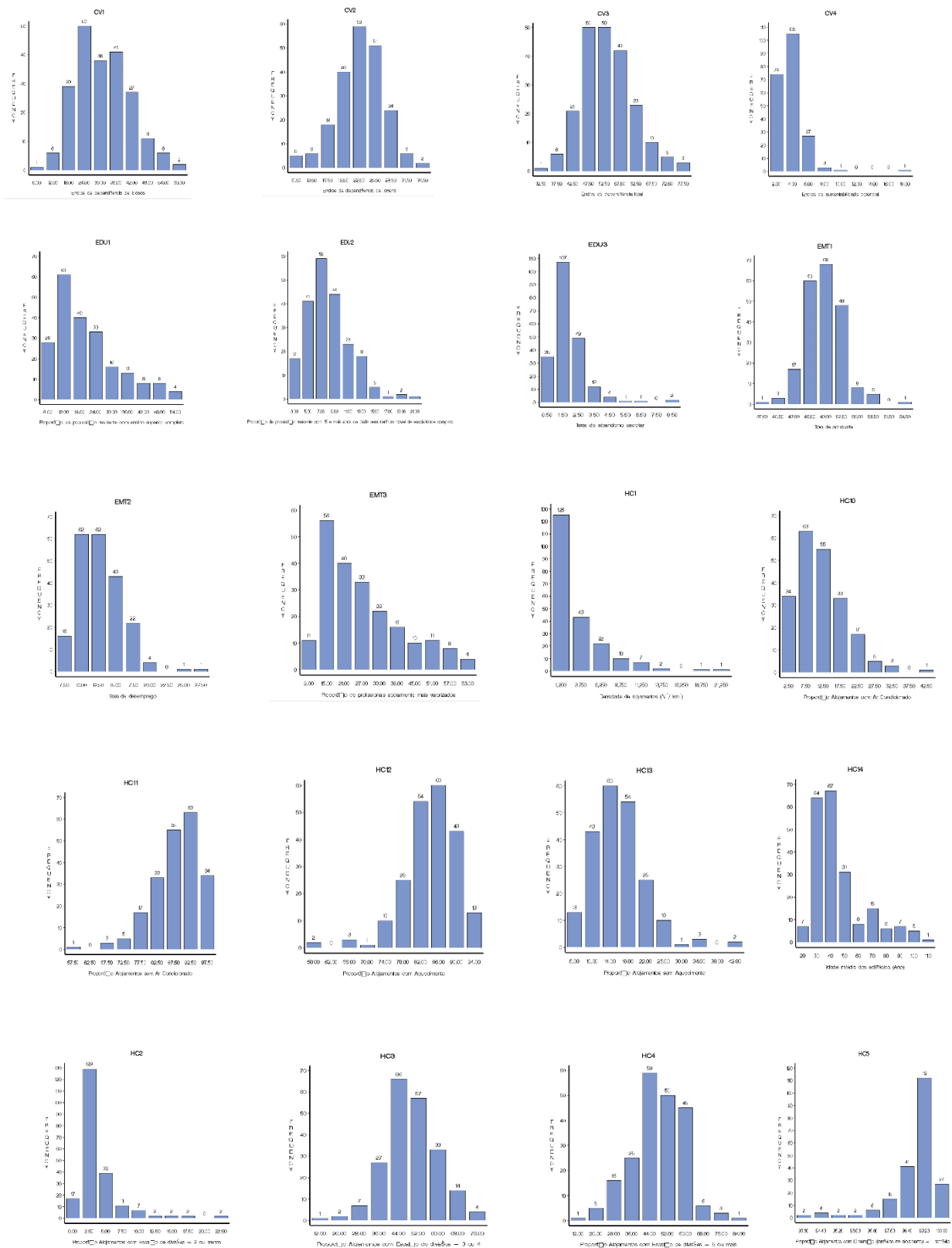
Webber, R. (n.d.). How to use geodemographic and lifestyle data, 1–20.

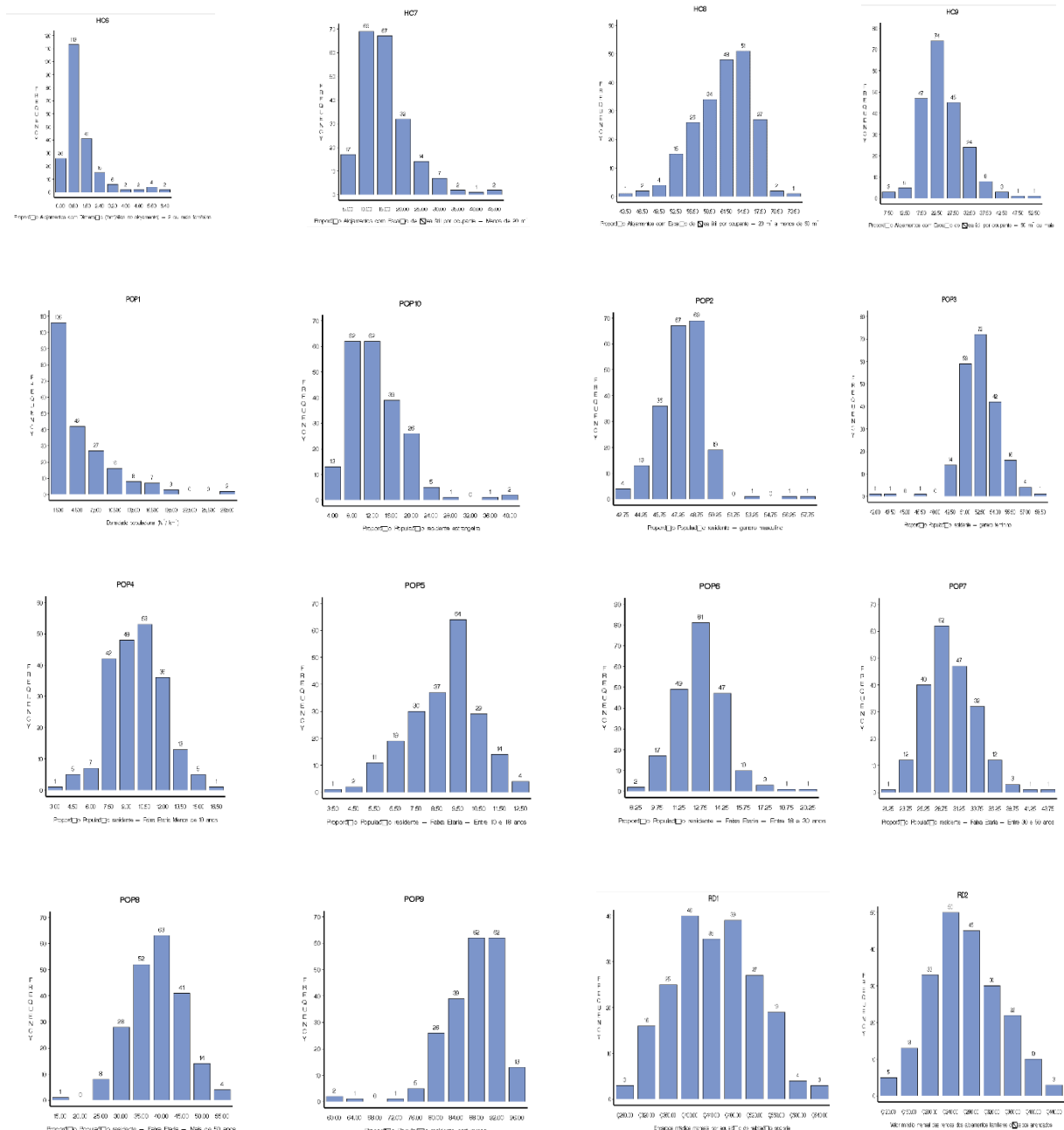
Webber, R. (1813). THE RELATIVE POWER OF HOUSEHOLD LEVEL DEMOGRAPHIC VARIABLES AS DISCRIMINATORS OF CONSUMER BEHAVIOUR consumer behaviour, 44(0).

Weinstein, A. (1994). Market Segmentation: Using Demographics, Psychographics and Other Niche Marketing Techniques to Predict and Model Customer Behavior. Retrieved from <http://www.getcited.org/pub/103228075>

8. ANEXOS

8.1. ANEXO 1 – HISTOGRAMA DISTRIBUIÇÃO VARIÁVEIS INPUT





8.2. ANEXO 2 – CÓDIGO SAS PARA GRÁFICO COTOVELO

```
%Let Segmentacao=Segmentacao_Consumo;

Data &EM_EXPORT_VALIDATE.;

Set &EM_IMPORT_DATA.(Keep=%EM_ID_Segment_Label_);

Rename _Segment_Label_=Seg_&Segmentacao_;
```

```

Run;

%Macro Ficheiro;

Proc Sql;

Select Sum(Distance),

Max(_SEGMENT_) into:Distance,

:NCluster From &EM_Import_data.;

Quit;

%Let Distance=&Distance.;

%Let NCluster=&NCluster.;

%Put =&NCluster.=&Distance.=;

%If %sysfunc(exist(&Em_Lib..&Segmentacao.,DATA)) ne %str(1) or &NCluster. eq 1 %Then %Do;

Data &Em_Lib..&Segmentacao.;

NumClusters=&NCluster.;

ErroTotal=&Distance.;

Output;

Run;

%End;

%Else %Do;

Data New;

NumClusters=&NCluster.;

ErroTotal=&Distance.;

Output;

Run;

Data &Em_Lib..&Segmentacao.;

Set &Em_Lib..&Segmentacao. New;

Run;

```

```
%End;

%EM_Register(Key=SEG,Type=DATA);

Proc Sql;

  Create Table &EM_USER_SEG As

  Select NumClusters,

  Avg(ErroTotal) As AvgErro,

  Min(ErroTotal) As MinErro,

  Max(ErroTotal) As MaxErro

  From &Em_Lib..&Segmentacao.

  Group By NumClusters;

Quit;

%EM_Report(Key=SEG,

Viewtype=DATA,

Autodisplay=Y,

Block=Cotovelo,

Description=Erro Total);

%Mend Ficheiro;

%Ficheiro;
```