



ANA MARGARIDA PINTO BORBINHA

Bachelor in Computer Science

CLIENTE 360 – SEGMENTAÇÃO E CARACTERIZAÇÃO DA BASE DE CLIENTES DA NOS

MESTRADO EM ANÁLISE E ENGENHARIA DE BIG DATA

Universidade NOVA de Lisboa
Março, 2023



CLIENTE 360 – SEGMENTAÇÃO E CARACTERIZAÇÃO DA BASE DE CLIENTES DA NOS

ANA MARGARIDA PINTO BORBINHA

Bachelor in Computer Science

Orientadora: Cláudia Soares
Assistant Professor, NOVA University Lisbon

Coorientador: José Pedro Moreno
Lead Data Scientist, NOS

Cliente 360 – Segmentação e Caracterização da base de clientes da NOS

Copyright © Ana Margarida Pinto Borbinha, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa.

A Faculdade de Ciências e Tecnologia e a Universidade NOVA de Lisboa têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objetivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

À minha família.

AGRADECIMENTOS

Gostaria de agradecer aos orientadores da minha tese de mestrado, a Professora Cláudia Soares e o Doutor José Pedro Moreno por todo o suporte, *feedback* e incansável otimismo.

Gostaria de deixar um especial agradecimento à minha primeira orientadora da NOS SGPS, Susana Brandão, pela orientação e por fomentar discussões das quais resultaram sugestões valiosas para esta dissertação. Acima de tudo, um obrigado por ter sido um exemplo de um profissional que eu aspiro ser no futuro.

Gostaria de agradecer à Professora Susana Nascimento pela sua orientação na fase inicial desta dissertação e por me ter introduzido à temática de aprendizagem multi vista, que se revelou um ponto central desta dissertação.

Queria agradecer a toda a NOS SGPS e, em especial, à equipa do Center of Excellence of Advanced Analytics. É um prazer fazer parte desta incrível comunidade, aprendi e cresci imenso convosco! Há, definitivamente, mais em NOS!

Queria agradecer à NOVA School of Science and Technology por me ter acompanhado durante o meu mestrado e a todos os docentes do Mestrado e Engenharia de Big Data pelas bases do conhecimento que me transmitiram e pela excelente relação cultivada entre docentes e alunos nesta instituição.

Gostaria de agradecer à minha família e amigos por todo o suporte e amor incondicional durante este percurso e por me lembrarem que o caminho se faz caminhando.

Por último, mas, igualmente, necessário, ao caminhante: “as tuas pegadas são o caminho, caminhante. Não há caminho, o caminho faz-se a andar.

“Diversity and independence are important because the best collective decisions are the product of disagreement and contest, not consensus or compromise.” (James Surowiecki)

ÍNDICE

Índice de Figuras	ix
Índice de Tabelas	xv
Siglas	xvii
Símbolos	xix
1 Introdução	1
1.1 Contexto e Motivação	1
1.2 Problema	2
1.3 Principais Objetivos	3
1.4 Estrutura do Documento	3
2 Análise de segmentação de cliente	5
2.1 Definição de análise de segmentação de cliente	5
2.1.1 Fatores de atratividade de um segmento	5
2.2 Abordagens quanto à segmentação de cliente	6
2.2.1 Abordagens quanto à emergência de segmentos	6
2.2.2 Abordagens quanto à escolha de variáveis de segmentação	6
2.2.3 Abordagens quanto à estrutura dos dados em segmentações <i>a poste-</i> <i>riori</i>	7
2.2.4 Abordagens quanto aos pilares de segmentação	8
2.3 Revisão de estado da arte	9
3 Modelação e exploração de dados	11
3.1 Modelação	11
3.2 Recolha de dados	14
3.2.1 Dados da PORDATA	14
3.2.2 Recolha de dados psicográficos	18

3.2.3	Dados comportamentais	20
3.3	Criação da amostra	20
3.4	Pré-processamento	21
3.5	Clusterização com k-means	21
3.6	Métricas de avaliação	22
3.6.1	Índices de validação internos	22
3.7	Resultados experimentais utilizando k-means	25
3.7.1	Experiências considerando apenas uma vista	25
3.7.2	Apenas vista psicográfica	29
3.7.3	Apenas vista comportamental	32
3.7.4	Multi-vista	37
4	Algoritmos de aprendizagem multi-vista	44
4.1	Motivação para modelos <i>multi-view</i>	44
4.2	Princípios para a aprendizagem multi-vista	46
4.2.1	Princípios de complementaridade e de consenso	46
4.2.2	Construção das vistas	46
4.2.3	Combinação das vistas	46
4.3	Grupos de algoritmos de aprendizagem multi-vista	47
4.3.1	Co-training style	47
4.3.2	Co-EM	48
4.3.3	Co-regularization	48
4.3.4	Co-clustering	48
4.3.5	Multi-kernel learning	49
4.3.6	Multi-view subspace clustering	49
4.3.7	Aplicabilidade de Multi-view subspace clustering em tempo linear a conjuntos de dados com uma única vista	55
4.3.8	Multi-task multi-view clustering	56
4.4	Aplicações de Multi-view Learning	57
4.5	Discussão de Resultados utilizando conjuntos de dados populares na lite- ratura	57
4.5.1	Resultados com o conjunto de dados Multiple Features handwritten digit	60
4.5.2	Resultados com o conjunto de dados Caltech101-20	67
4.5.3	Resultados sem considerar a vista HOG	73
4.5.4	Resultados com o conjunto de dados Yale Face	76
4.6	Conclusões	81
5	Segmentação de cliente com aprendizagem multi vista	83
5.1	Resultados considerando uma única vista	83
5.1.1	Vista Demográfica	83

5.1.2	Vista Psicográfica	88
5.1.3	Vista Comportamental	93
5.2	Experiências multi vista	99
5.2.1	Apenas vistas psicográfica e comportamental	108
6	Conclusões e Trabalho Futuro	115
	Bibliografia	117
	Apêndices	
A	Dados da PORDATA	121
A.1	Dados recolhidos da PORDATA	121
A.2	Dados normalizados da PORDATA	123
B	Experiências com conjuntos de dados <i>multi-view</i>	125
B.1	Handwritten	125
B.1.1	Vistas originais	126
B.1.2	Resultados de clusterização por vista	127
B.2	Caltech101-20	130
B.2.1	Vistas originais	130
B.2.2	Resultados de clusterização por vista	131
B.3	Yale Faces	134
B.3.1	Resultados de clusterização por vista	134
C	Variáveis utilizadas	136

ÍNDICE DE FIGURAS

3.1	Pilares de segmentação e informações contempladas por cada um dos pilares de segmentação. [33].	12
3.2	Composição da população residente por faixa etária, em 2011.	16
3.3	Índice de envelhecimento da população, por município, em 2011.	16
3.4	Composição do nível de educação, por município, em 2011.	17
3.5	Comparação entre o ganho médio mensal, em euros, com o poder de compra de cada município, em 2011.	17
3.6	Composição das famílias, agrupando as famílias unipessoais com as famílias com dois elementos, por município, em 2011.	18
3.7	Comparação da prevalência de clientes dos diferentes concelhos considerados entre a população e a amostra estratificada extraída.	21
3.8	Comparação das métricas de validação internas, considerando o algoritmo <i>k-means</i> para todos os clientes da amostra, considerando, apenas, dados demográficos.	26
3.9	Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelo algoritmo <i>k-means</i> para todos os clientes da amostra, utilizando, apenas, dados demográficos.	27
3.10	Comparação da prevalência de clientes com participações técnicas entre a população e a amostra estratificada extraída, considerando, apenas, dados demográficos.	28
3.11	Comparação das métricas de validação internas, considerando o algoritmo <i>k-means</i> para todos os clientes da amostra.	29
3.12	Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelo algoritmo <i>k-means</i> para todos os clientes da amostra, utilizando, apenas, dados psicográficos.	31
3.13	Comparação da prevalência de clientes com participações técnicas entre a população e a amostra estratificada extraída, considerando, apenas, dados psicográficos.	31

3.14	Comparação das métricas de validação internas, considerando o algoritmo <i>k-means</i>	33
3.15	Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelo algoritmo <i>k-means</i> para todos os clientes da amostra, utilizando, apenas, dados comportamentais.	34
3.16	Comparação das métricas de validação internas, considerando o algoritmo <i>k-means</i> para todos os clientes que evidenciam utilização do router, considerando, apenas, dados comportamentais.	36
3.17	Comparação da prevalência de clientes com participações técnicas entre a população e a amostra estratificada extraída.	38
3.18	Comparação da prevalência de clientes com participações técnicas entre a população e a amostra estratificada extraída.	39
3.19	Comparação da prevalência de clientes com participações técnicas entre a população e a amostra estratificada extraída.	42
3.20	Comparação da distribuição de Average Revenue Per Unit (ARPU) e idade, bem como da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando, apenas, dados psicográficos e comportamentais.	43
4.1	Ilustração dos princípios de consenso e de complementaridade [40].	46
4.2	Procedimento de algoritmos de co-training [40].	47
4.3	Procedimento de algoritmos de multi-kernel learning [40].	49
4.4	Ilustração do procedimento dos algoritmos de <i>subspace clustering</i> [40].	49
4.5	Esquematização do funcionamento do método <i>Large-Scale Multi-view Subspace Clustering in Linear Time</i> (LMVSC) [34].	53
4.6	Esquematização do funcionamento do método <i>Scalable Multi-view Subspace Clustering with Unified Anchors</i> (SMVSC) [34].	55
4.7	Ilustração do procedimento dos algoritmos de multi-task multi-view clustering [40].	56
4.8	Visualização dos resultados de clusterização para o conjunto de dados Multiple Features handwritten digit, utilizando a técnica de redução de dimensionalidade t-SNE, utilizando todas as vistas.	60
4.9	Matriz de confusão para o conjunto de dados Multiple Features handwritten digit, utilizando todas as vistas, normalizada ao longo das linhas.	60
4.10	Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Multiple Features handwritten digit, na vista Fourier coefficients of the character shapes.	64
4.11	Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Multiple Features handwritten digit, na vista Karhunen-Love coefficients.	64

4.12	Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Multiple Features handwritten digit, na vista Morphological features.	65
4.13	Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Multiple Features handwritten digit, na vista Pixel averages in 2 x 3 windows.	65
4.14	Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Multiple Features handwritten digit, na vista Zernike moments.	66
4.15	Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE, utilizando todas as vistas.	68
4.16	Matriz de confusão para o conjunto de dados Caltech101-20, utilizando todas as vistas, normalizada ao longo das linhas.	68
4.17	Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Caltech101-20, na vista Gabor.	70
4.18	Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Caltech101-20, na vista Wavelet moments.	70
4.19	Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Caltech101-20, na vista CENTRIST.	71
4.20	Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Caltech101-20, na vista HOG.	71
4.21	Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Caltech101-20, na vista GIST.	72
4.22	Diferença entre as matrizes de confusão entre modelo <i>single-view</i> e o modelo <i>multi-view</i> para o conjunto de dados Caltech101-20, na vista LBP.	72
4.23	Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE, utilizando todas as vistas exceto a 4.	73
4.24	Matriz de confusão para o conjunto de dados Caltech101-20, utilizando todas as vistas exceto a vista HOG, normalizada ao longo das linhas.	74
4.25	Visualização dos resultados de clusterização para o conjunto de dados Yale Face, utilizando a técnica de redução de dimensionalidade t-SNE, utilizando todas as vistas.	76
4.26	Matriz de confusão para o conjunto de dados Yale Face, utilizando todas as vistas, normalizada ao longo das linhas.	77
4.27	Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Yale Face, na vista Intensity.	77
4.28	Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Yale Face, na vista LBP.	78

4.29	Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Yale Face, na vista Gabor.	78
5.1	Comparação das métricas de validação internas, considerando os algoritmo LMVSC e SMVSC para todos os clientes, considerando, apenas, dados demográficos.	84
5.2	Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelos algoritmos LMVSC e SMVSC para todos os clientes da amostra, utilizando, apenas, dados demográficos.	85
5.3	Comparação da distribuição de ARPU e idade na amostra estratificada extraída, considerando, apenas, dados demográficos.	86
5.4	Comparação da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando, apenas, dados demográficos.	87
5.5	Comparação das métricas de validação internas, considerando os algoritmo LMVSC e SMVSC para todos os clientes, considerando, apenas, dados psicográficos.	88
5.6	Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelos algoritmos LMVSC e LMVSC para todos os clientes da amostra, utilizando, apenas, dados psicográficos.	89
5.7	Comparação da distribuição de ARPU e idade na amostra estratificada extraída, considerando, apenas, dados psicográficos.	91
5.8	Comparação da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando, apenas, dados psicográficos.	92
5.9	Comparação das métricas de validação internas, considerando os algoritmos LMVSC e SMVSC para todos os clientes, utilizando, apenas, dados comportamentais.	93
5.10	Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelos algoritmos LMVSC e LMVSC para todos os clientes da amostra, utilizando, apenas, dados comportamentais.	94
5.11	Comparação da distribuição de ARPU e idade na amostra estratificada extraída, considerando, apenas, dados comportamentais.	97
5.12	Comparação da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando, apenas, dados comportamentais.	98
5.13	Comparação das métricas de validação internas, considerando os algoritmo LMVSC e SMVSC para todos os clientes, considerando dados demográficos, psicográficos e comportamentais.	99
5.14	Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelos algoritmos LMVSC e LMVSC para todos os clientes da amostra, utilizando dados demográficos, psicográficos e comportamentais.	100

5.15	Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelo algoritmo LMVSC para todos os clientes da amostra, utilizando dados demográficos, psicográficos e comportamentais.	104
5.16	Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelo algoritmo SMVSC para todos os clientes da amostra, utilizando dados demográficos, psicográficos e comportamentais.	105
5.17	Comparação da distribuição de ARPU e idade, bem como da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando dados demográficos, psicográficos e comportamentais.	106
5.18	Comparação da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando dados demográficos, psicográficos e comportamentais.	107
5.19	Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelos algoritmos LMVSC e SMVSC para todos os clientes da amostra, utilizando, apenas, dados psicográficos e comportamentais.	108
5.20	Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelos algoritmos LMVSC e SMVSC para todos os clientes da amostra, utilizando dados psicográficos e comportamentais.	111
5.21	Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelo algoritmo LMVSC para todos os clientes da amostra, utilizando dados psicográficos e comportamentais.	112
5.22	Comparação da distribuição de ARPU e idade na amostra estratificada extraída, considerando, apenas, dados psicográficos e comportamentais.	113
5.23	Comparação da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando dados psicográficos e comportamentais.	114
B.1	Visualização das vistas do conjunto de dados Handwritten, utilizando a técnica de redução de dimensionalidade t-SNE.	126
B.2	Visualização dos resultados de clusterização para o conjunto de dados Multiple Features handwritten digit, utilizando a técnica de redução de dimensionalidade t-SNE.	127
B.2	Visualização dos resultados de clusterização para o conjunto de dados Multiple Features handwritten digit, utilizando a técnica de redução de dimensionalidade t-SNE.	128
B.2	Visualização dos resultados de clusterização para o conjunto de dados Multiple Features handwritten digit, utilizando a técnica de redução de dimensionalidade t-SNE.	129
B.3	Visualização das vistas do conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE.	130
B.4	Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE.	131

B.4	Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE.	132
B.4	Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE.	133
B.5	Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE.	134
B.5	Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE.	135

ÍNDICE DE TABELAS

3.1	Excerto da tabela de utilização de <i>internet</i> pelo discretizado em períodos de 15 minutos, por cada endereço MAC.	19
3.2	Protocolos de <i>internet</i> possíveis na variável <i>App Protocol</i> presente na tabela 3.1	19
3.3	Centróides dos <i>clusters</i> pelo algoritmo <i>k-means</i> , considerando todos os clientes da amostra e utilizando, apenas, variáveis comportamentais.	35
3.4	Centróides dos <i>clusters</i> , utilizando dados demográficos, psicográficos e comportamentais.	41
4.1	Complexidade dos métodos comparados, em que $h = \sum_{i=1}^v d_i$ e c é a dimensão comum	51
4.2	Performance da clusterização no dataset Multiple Features handwritten digit, utilizando o algoritmo LMVSC.	61
4.3	Performance da clusterização no dataset Multiple Features handwritten digit, utilizando o algoritmo SMVSC.	62
4.4	Performance da clusterização no dataset Caltech101-20, utilizando o algoritmo LMVSC.	69
4.5	Performance da clusterização no dataset Caltech101-20, utilizando o algoritmo SMVSC.	69
4.6	Performance da clusterização no dataset Caltech101-20, utilizando o algoritmo LMVSC.	74
4.7	Performance da clusterização no dataset Caltech101-20, utilizando o algoritmo SMVSC.	75
4.8	Performance da clusterização no dataset Yale Face, utilizando o algoritmo LMVSC.	79
4.9	Performance da clusterização no dataset Yale Face, utilizando o algoritmo SMVSC.	79
5.1	Âncoras utilizadas pelo algoritmo LMVSC e SMVSC, utilizando dados demográficos.	85

5.2	Âncoras utilizadas pelo algoritmo LMVSC e SMVSC, utilizando dados psicográficos.	90
5.3	Âncoras utilizadas pelo algoritmo LMVSC, utilizando dados psicográficos.	96
5.4	Âncoras utilizadas pelo algoritmo LMVSC e SMVSC, utilizando dados demográficos, psicográficos e comportamentais.	103
5.5	Âncoras utilizadas pelo algoritmo LMVSC e SMVSC, utilizando dados psicográficos e comportamentais.	111
A.1	População residente por município e em Portugal Continental, em 2011	121
A.2	Nível de educação, por município e em Portugal Continental, em 2011	121
A.3	Poder de compra e ganho médio mensal, por município e em Portugal Continental, em 2011	122
A.4	Dimensão das famílias, por município e em Portugal Continental, em 2011	122
A.5	Composição da população residente por grandes grupos etários, por município e em Portugal Continental, em 2011	123
A.6	Composição da população residente por nível de educação, por município e em Portugal Continental, em 2011	123
A.7	Composição da população residente por dimensões do agregado familiar, por município e em Portugal Continental, em 2011	124

SIGLAS

ACC	Accuracy 58
ARPU	Average Revenue Per Unit x, xii, xiii, 2, 8, 15, 20, 21, 32, 36, 37, 42, 43, 81, 86, 91, 97, 106, 109, 113
FN	Número de falsos negativos 59
FP	Número de falsos positivos 59
IE	Índice de Envelhecimento 16, 17
KPI	Indicador-chave de Desempenho, do inglês <i>Key Performance Indicator</i> 2
LMVSC	<i>Large-Scale Multi-view Subspace Clustering in Linear Time</i> x, xii, xiii, xv, xvi, 3, 51, 52, 53, 57, 69, 73, 79, 83, 84, 85, 87, 88, 89, 90, 92, 93, 94, 96, 98, 99, 100, 103, 104, 107, 108, 109, 111, 112, 116
LTV	Lifetime Value 8, 9
MARPU	Marginal Average Revenue Per Unit 8, 9
ML	Machine Learning 2
MvC	Multi-view Clustering 45, 49, 81, 82
MVSC	Multi-View Subspace Clustering 50, 81
NIF	Número de Identificação Fiscal 32
NMI	Normalized Mutual Information 58
SA	Identificador do contrato, do inglês <i>Service Account</i> 32
SMVSC	<i>Scalable Multi-view Subspace Clustering with Unified Anchors</i> x, xii, xiii, xv, xvi, 3, 52, 55, 57, 73, 81, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 97, 98, 99, 100, 103, 105, 106, 107, 108, 109, 111, 113, 114

SVD	Decomposição em valores singulares, do inglês, <i>Singular Value Decomposition</i> 53, 54, 55
TP	Número de verdadeiros positivos 59
VC	Vista Comportamental 37, 39, 40, 101, 108, 109
VD	Vista Demográfica 25, 37, 39, 83, 100, 108
VoD	Video on Demand 13, 20, 32
VP	Vista Psicográfica 37, 39, 101, 108, 109

SÍMBOLOS

A^i	Matriz com os protótipos de <i>k-means</i> na vista i 53
a_j	Protótipo j obtido por aplicação de <i>k-means</i> 53
d_i	Dimensionalidade da vista i 54
k	número de clusters 25, 32, 37, 83, 88, 93
n	Número de observações 54, 58
X_i	Matriz de dados na vista i 54

INTRODUÇÃO

1.1 Contexto e Motivação

Esta segmentação é motivada pela necessidade das empresas de telecomunicações fazerem uma monitorização cuidadosa e regular da qualidade de serviço entregue e a forma como é percecionada pelo cliente. Enquanto a qualidade do serviço (QoS, do Inglês *Quality of Service*) é mensurável, objetiva e observável, a qualidade da experiência (QoE, do Inglês *Quality of Experience*) é dependente do consumidor final e expressa a sua satisfação tanto objetiva como subjetivamente [6]. A qualidade de experiência é definida pelo grau de satisfação ou incómodo sentido pelo utilizador de um serviço ou aplicação, resultando da satisfação das suas expectativas face à utilidade ou à satisfação em relação à aplicação ou serviço utilizado, à luz da sua personalidade e do seu estado atual, no que diz respeito a fatores como sentimentos, pensamentos ou comportamentos [3].

A qualidade da experiência diferencia-se da qualidade de serviço quanto ao âmbito de aplicação, foco e métodos. Tem um domínio de aplicação mais abrangente que, apenas, as telecomunicações, focando-se na qualidade percecionada pelo consumidor final, a qual pode ser influenciada por fatores além dos fatores técnicos de entrega de serviço, como as expectativas do consumidor face à performance do serviço, perfis psicológicos e fatores socioculturais. Nesta ótica, surge a necessidade de uma segmentação a nível geográfico, demográfico, psicográfico e comportamental, tentando criar um artefacto que caracterize o cliente, como uma ferramenta que permita a obtenção de conhecimento acerca dos seus consumidores e das suas necessidades, além de forçar à reflexão da empresa quanto aos seus pontos fortes face à competição [11].

Utilizando a qualidade da experiência como ponto de partida para uma caracterização do cliente, a motivação estende-se à ambição de um modelo que seja tão geral quanto possível, cobrindo todos os ângulos do cliente, uma vez que poderá ser uma mais-valia em diversos projetos na área das telecomunicações, que acompanham o cliente nas várias etapas do seu ciclo de vida, desde a sua aquisição, passando pela recomendação de novos serviços, não esquecendo a resolução, de forma proativa, de problemas do cliente antes de uma possível reclamação, até à mitigação do *churn* de clientes.

1.2 Problema

Este trabalho vai ser desenvolvido na empresa de telecomunicações NOS. Os clientes desta empresa são heterogêneos nos motivos e necessidades pelos quais celebram contratos, nas expectativas que têm quanto ao serviço prestado e na sua sensibilidade a problemas na entrega de serviço. Esta heterogeneidade obriga a que as abordagens tanto de marketing como de assistência devam ser customizadas ao cliente em questão.

Neste momento, já são produzidos bons resultados através da utilização de modelos de [Machine Learning \(ML\)](#), nas várias fases do ciclo de vida de um cliente, desde tentar crescer o seu [ARPU](#), ao longo do tempo, como tentar recomendar serviços em que, provavelmente, esteja interessado, bem como tentar detetar clientes com elevada probabilidade de pretenderem rescindir contrato, não esquecendo tentar, proativamente, resolver problemas de serviço. Estas abordagens baseadas em [Indicador-chave de Desempenho, do inglês Key Performance Indicator \(KPI\)](#) técnicos, ainda que contenham informação que permita uma tomada de decisão que incorpore alguma informação sobre quem é o cliente, mostram-se insuficiente para distinguir dois clientes que, apesar de terem, por exemplo, as mesmas características de entrega de serviço, possuem um comportamento diferente, tanto quanto à sua propensão a apresentar queixas quanto à rescisão voluntária de contrato.

Cada cliente pode ser visto de acordo com as suas características geográficas, demográficas, psicográficas e comportamentais, sendo que cada uma destas características constitui uma vista diferente sobre o cliente. Uma das preocupações que esta tese procura abordar é o equilíbrio do peso relativo de cada vista no resultado final da clusterização, de modo a que seja o mais relevante possível para as tarefas a que nos endereçamos. Propomos, então, a utilização de algoritmos de *multi-view clustering*, frequentemente, utilizadas para outros problemas de multimodalidade dos dados.

É, assim, necessário realizar uma segmentação de cliente que permita agrupar consumidores em segmentos, que podem existir naturalmente ou podem ser artificialmente criados, partilhando preferências ou características semelhantes [11]. Contudo, a visão clássica do marketing que propõe a segmentação baseada em apenas uma variável de segmentação revela-se insuficiente para o objetivo da criação de um modelo que permita uma multiplicidade de utilizações e que englobe múltiplas perspetivas sobre o cliente. Com vista a uma modelação e segmentação que considere o cliente de forma multiangular, na tentativa de compreender melhor as suas necessidades e motivações, iremos contemplar os seguintes tipos de segmentação: geográfica, demográfica, psicográfica e comportamental.

Uma segmentação geográfica consiste em segmentar a base de clientes de acordo com a sua localização geográfica, ajustando a oferta de serviço às expectativas locais. Apesar de ser uma informação facilmente obténivel e segmentável, pertencer à mesma área geográfica não é um fator que, por si só, indique que clientes no mesmo *cluster* partilhem preferências semelhantes [11].

No caso da segmentação demográfica, tenta-se segmentar com base na idade, dimensão do agregado familiar, género, estado civil, nível de educação e rendimento familiar. Contudo, estima-se que a segmentação demográfica só explique 5% da variância no comportamento do consumidor [15].

Quanto à segmentação psicográfica, pretende-se caracterizar os traços de personalidade, interesses, crenças, valores e estilo de vida [37]. Apesar de ser mais informativa sobre as razões que motivam o comportamento dos consumidores, adiciona complexidade à segmentação.

Por fim, a segmentação comportamental estuda as semelhanças entre o comportamento real e o comportamento reportado, não sendo aplicável a potenciais clientes [11].

Assim, uma boa segmentação da base de clientes contempla estes quatro pilares, sendo que a incorporação da segmentação feita para cada pilar numa mesma segmentação poderá melhorar a *performance* da clusterização final, uma vez que cada pilar caracteriza o cliente de forma diferente e, parcialmente, independente [14].

O problema em causa é proceder a uma segmentação de cliente que permita uma multiplicidade de utilizações desta mesma segmentação. Deste modo, pretendemos ir ao encontro da ambição expressa no título desta dissertação, a produção de uma visão 360º sobre o cliente.

1.3 Principais Objetivos

Esta dissertação tem como objetivos:

1. compreensão da composição dos clientes; compreensão das características dos clientes valiosos; compreensão das características dos clientes perdidos e compreensão das suas características de consumo.
2. produção de um artefacto que permita a segmentação do parque de clientes da empresa com relativa homogeneidade. Os segmentos irão ser selecionados de acordo com os fatores de atratividade descritos na Secção 2.1.1 e apenas os promissores serão descritos consoante as suas necessidades e motivações.
3. realização de um estudo teórico da aplicabilidade dos algoritmos [LMVSC](#) e [SMVSC](#) a conjuntos de dados com uma única vista, utilizando conjuntos de dados populares da literatura.

1.4 Estrutura do Documento

O documento é organizado da seguinte forma:

No capítulo 2, apresenta-se uma breve introdução ao processo de segmentação de cliente, de acordo com as suas possíveis abordagens. No fim deste capítulo, é feito um estudo quanto às abordagens seguidas na literatura desta área.

No capítulo 3 em que se apresenta o processo de modelação, recolha de dados e criação da amostra e pré-processamento. Em seguida, é apresentado o algoritmo *k-means* e métricas de avaliação. No final, é apresentado os resultados da primeira abordagem de segmentação de cliente, utilizando o algoritmo *k-means*.

No capítulo 4, é apresentada a motivação para a consideração dos algoritmos multi vista, os princípios para a aprendizagem multi vista, a taxonomia dos diferentes tipos de algoritmos de aprendizagem multi vista. as aplicações deste tipo de algoritmos, a realização de experiências que estudam a aplicabilidade dos métodos de *subspace clustering* em tempo linear a conjuntos de dados com uma única vista e a sua comparação com o cenário multi vista.

No capítulo 5 são apresentados os resultados da segmentação de cliente, considerando, primeiramente, cada pilar de forma individual, e em seguida dois cenários multi vista.

As conclusões e trabalho futuro foram apresentados no capítulo 6.

ANÁLISE DE SEGMENTAÇÃO DE CLIENTE

Neste capítulo, é definido o conceito de análise de segmentação de cliente na secção 2.1. Na secção 2.2, são apresentadas abordagens de acordo com questões de emergência de segmentos quanto à escolha de variáveis de segmentação, quanto à estrutura dos dados em segmentações *a posteriori* e quanto aos pilares de segmentação considerados. Na secção 2.3, é apresentada uma breve revisão do estado da arte.

2.1 Definição de análise de segmentação de cliente

A intensa competição na indústria de telecomunicações fez com que as estratégias baseadas no preço não se apresentem capazes de satisfazer as necessidades do cliente, o que forçou as empresas a repensarem a sua estratégia focada na oferta tecnológica, orientando-se para uma estratégia focada no mercado e no consumidor. Assim, nasce a importância de cargos de gestão da relação com o cliente (do inglês *Customer relationship management*), com o objetivo de cobrir todos os processos relacionados com a aquisição de um cliente, o fomento de uma relação entre o cliente e a marca e ainda a retenção do cliente. Para garantir esta boa relação, é necessária uma constante avaliação do seu valor, motivações e necessidades.

A segmentação de cliente é definida como o processo de agrupar clientes que partilham características ou preferências semelhantes em segmentos, garantindo que os clientes, num mesmo grupo, têm a maior similaridade entre si e que os clientes, em grupos distintos, têm uma maior diferença.

2.1.1 Fatores de atratividade de um segmento

Os autores de [11] consideram a necessidade de especificar *a priori* os fatores que medem a atratividade de um segmento. Assim, consideramos a necessidade do segmento ser homogéneo, ser distinto dos outros segmentos, ser identificável no parque de clientes e ser possível contactar esses clientes para aplicar estratégias customizadas, ter dimensão suficiente para justificar o investimento económico no desenvolvimento de estratégias e,

por fim, ser compatível com os pontos fortes da empresa, ou seja, a empresa consegue satisfazer as necessidades do cliente.

2.2 Abordagens quanto à segmentação de cliente

2.2.1 Abordagens quanto à emergência de segmentos

Existem três abordagens para a segmentação de cliente, baseadas em limitações empresariais:

A abordagem quantitativa baseada em inquéritos é a visão tradicional de segmentação de cliente. O processo é começado a partir do zero, logo todas as visões que a empresa já possuía sobre os seus clientes não são tidas em conta na criação dos novos segmentos, podendo implicar uma revolução dos segmentos existentes. Das três abordagens existentes, apresenta a mais elevada probabilidade de colher todos os benefícios do processo de segmentação de cliente, embora nem sempre seja viável na prática, uma vez que seria necessário que a empresa estivesse disposta a começar do zero.

A criação de segmentos a partir de classificações de clientes já existentes representa uma evolução dos segmentos já existentes, através da sua refinação ou subdivisão, auxiliada tanto por dados que a empresa possui sobre os seus clientes como por novas pesquisas de mercado.

A emergência de segmentos a partir de investigação qualitativa encontra os segmentos como parte de um processo de investigação exploratória, fruto de *data mining* de várias fontes de dados, o que faz sentido na indústria de telecomunicações, dado o seu uso intensivo de dados.

2.2.2 Abordagens quanto à escolha de variáveis de segmentação

Segmentação *a priori* é considerada quando se sabe, à partida, qual é a única característica que permite uma segmentação dos clientes em dois ou mais segmentos, sem que seja necessário qualquer tipo de análise de dados, tradicionalmente, executada pelo departamento de marketing. Por exemplo, na indústria de telecomunicações, sempre se procedeu à segmentação dos clientes relativamente à sua contribuição para as receitas. Nesta perspectiva, o ganho deste tipo de segmentação limita-se ao ganho de conhecimento sobre um grupo de clientes, em vez de procurar as características-chave que definem um grupo relativamente homogêneo de clientes, de acordo com as suas motivações e necessidades.

Segmentação *a posteriori* é considerada quando não se conhecem as variáveis de segmentação que definem os *clusters*, procedendo-se a técnicas de prospeção de dados que

permitem descobrir novos segmentos de clientes, contudo é necessário que a organização tenha algumas suposições sobre as características relevantes para a identificação de um segmento. A organização deve ter certas suposições quanto às características-chave de identificação de um segmento, tendo dois objetivos: primeiramente, a exploração dos diferentes segmentos que podem ser criados através das variáveis de segmentação escolhidas e, em seguida, o desenvolvimento de perfis e descrições dos segmentos encontrados.

2.2.3 Abordagens quanto à estrutura dos dados em segmentações *a posteriori*

Quando se idealiza um estudo de análise de segmentação de cliente, é assumido que estes segmentos existem naturalmente nos dados. Contudo, segundo os autores [11], ao utilizar-se dados reais, este caso raramente se verifica. Uma vez que, no início de um processo de segmentação, não existe conhecimento se os clientes se irão agrupar em segmentos de mercado distintos e bem separados, este caso apresenta-se problemático, pois as definições de segmento de cliente e de *cluster* deixam de coincidir. Segundo [12] existem três abordagens quanto à estrutura dos dados:

A segmentação natural considera que os segmentos de clientes existem naturalmente nos dados, ou seja, que os segmentos de cliente são consistentes com a definição natural de *cluster*, existindo grupos de clientes que são muito semelhantes entre si e muito diferentes de clientes que não pertencem ao mesmo grupo. Assim, existe uma visão tradicional de que os segmentos de mercado existem naturalmente nos dados, sendo que esta análise visa descobri-los.

A segmentação reproduzível considera o caso em que os segmentos de clientes não existem naturalmente nos dados, ou seja, os clientes não se agrupam de acordo com uma estrutura consistente com a definição de *cluster*, mas também não são desprovidos de qualquer estrutura. Deste modo, é possível gerar a mesma segmentação repetidamente, o que permite uma análise fiável.

A segmentação construtiva considera o caso em que os clientes nem se agrupam de acordo com uma estrutura consistente com a definição de *cluster*, nem existe uma estrutura que permita gerar a mesma segmentação repetidamente como é o caso da segmentação reproduzível. Ainda assim, faz sentido prosseguir com uma segmentação de cliente, uma vez que a seleção de subconjuntos de clientes permite o desenvolvimento de estratégias com resultados mais promissores do que enveredar por estratégias indiferenciadas, que pretendem satisfazer uma grande base de clientes com a mesma estratégia, ignorando as suas diferenças.

2.2.4 Abordagens quanto aos pilares de segmentação

A segmentação geográfica permite segmentar os clientes, exclusivamente, de acordo com a região onde vivem. Esta segmentação é de fácil implementação, sendo que cada cliente pode ser atribuído a uma região geográfica, independentemente, da granularidade com que é feita a segmentação. Os autores de [11] apresentam como desvantagem para este tipo de segmentação o facto de não haver garantias que dois clientes partilhem características relevantes por viverem no mesmo segmento geográfico, como é o caso dos benefícios que procuram quando compram ou contratualizam serviços/produtos.

A segmentação demográfica permite segmentar os clientes de acordo com a sua informação sociodemográfica como é o caso da idade, género, rendimentos, educação, ocupação, entre outros. Os autores de [11] consideram uma desvantagem para este tipo de segmentação, a baixa explicabilidade acerca da variância do comportamento do cliente (cerca de 5%).

A segmentação psicográfica permite segmentar os clientes de acordo com características psicológicas como é o caso das crenças, interesses, preferências, aspirações ou benefícios que um cliente procura quando compra ou contratualiza serviços/produtos, segundo os autores de [11]. As agências de marketing defendem que um dos melhores exemplos deste tipo de segmentação é a promoção de campanhas focadas em clientes frugais, utilizando frases como "Preços incríveis" ou "Promoções imperdíveis" [35].

A segmentação comportamental permite a segmentação de clientes com base em características como a utilização dos seus serviços ou produtos, a intensidade e frequência da sua utilização e a sua lealdade à marca. Este tipo de segmentação é considerado como o método de segmentação mais eficaz por 91% dos profissionais de marketing, sendo considerado por 73% dos inquiridos como o meio de segmentação que se tem tornado mais eficaz nos últimos cinco anos, segundo um estudo realizado pela *Marketing Week*, em [30]. Uma desvantagem apresentada pelos autores de [11] é que esta segmentação só é possível para clientes que já tiveram oportunidade de comprar ou experimentar o produto, limitando esta análise a atuais clientes.

A segmentação baseada em valor permite a segmentação de clientes com base no valor que um cliente acrescenta à empresa em termos de [ARPU](#), [Marginal Average Revenue Per Unit \(MARPU\)](#) ou [Lifetime Value \(LTV\)](#), sendo calculado sem recurso a modelos de prospeção de dados, uma vez que os segmentos de valor podem ser calculados pela discretização do valor em segmentos, fazendo parte desta análise a seleção dos valores que os delimitam.

2.3 Revisão de estado da arte

Os autores em [36] consideram os critérios de segmentação baseados em comportamento, demografia, valor, lealdade e ainda em necessidades e atitudes. Estes autores focam-se na segmentação comportamental e baseada em valor. O processo de segmentação é visto como tendo cinco principais etapas: a compreensão do negócio e desenho do processo de segmentação, a compreensão dos dados, a preparação e enriquecimento, a identificação dos segmentos através do treino de modelos de clusterização, a avaliação e perfilagem dos segmentos e a implementação da solução de segmentação. É apresentado um processo de segmentação de cliente aplicado às telecomunicações, nomeadamente, aos tarifários móveis, ao acesso à internet e à TV, sendo considerado o **MARPU**, como variável de segmentação baseada em valor, o que permite uma priorização dos clientes de maior valor. Como variáveis de segmentação por comportamento, foram consideradas informações sobre a frequência e intensidade de utilização para cada tipo de chamada, para a internet, para a banda larga, o dia e a hora de utilização de serviço, utilização de *roaming*, direção das chamadas e o período de tempo enquanto cliente. Neste estudo, foram considerados, apenas, os clientes residenciais pós-pagos e os clientes atuais ou em incumprimento, excluindo-se os clientes que já abandonaram o serviço. Aplicou-se um processo de *binning* ao **MARPU**, privilegiando uma segmentação detalhada, sobretudo nos grupos de topo para discriminar os clientes mais valiosos. Para clusterizar os clientes com base no seu perfil, utilizou-se o algoritmo de *TwoStep Clustering* e, em seguida, procedeu-se à caracterização do seu perfil comportamental, do seu perfil demográfico e consequentes descobertas de estudo de mercado, como os seus *drivers* de satisfação, os seus desejos e necessidades bem como a abordagem de marketing mais adequada. Procedeu-se também à quantificação da percentagem de clientes neste segmento bem como da sua contribuição para o **MARPU** total. A implementação da segmentação é feita através do treino de uma árvore de decisão com os mesmos dados que foram utilizados na clusterização como *input* e os resultados da segmentação como *target*. Este método supervisionado permite que os vários perfis sejam divididos através de regras intuitivas, facilitando a compreensão das características que os definem.

Os autores em [16] consideram uma segmentação de cliente baseada num modelo de **LTV** que contempla as contribuições passadas, o valor potencial e a lealdade do cliente. O modelo proposto contempla o valor dos clientes ao longo do seu ciclo de vida e a probabilidade de um cliente churnar, o que é um tópico popular em indústrias muito competitivas como é o caso da indústria de telecomunicações. Este modelo considera que o **LTV** resulta da soma do lucro passado, ao que acresce o fluxo de caixa esperado no futuro, constituído por dois fatores: a soma dos lucros expectados com os benefícios potenciais durante o período em que se espera servir o cliente. O valor atual do cliente foi definido como o lucro que este gerou durante um certo período, ou seja, a diferença entre o valor pago e as dívidas do cliente, a dividir pelo período de serviço. O valor potencial é definido como o produtório do produto da probabilidade de um cliente utilizar um

serviço pelo lucro associado a esta utilização. Este artigo define a lealdade de um cliente como uma medida de retenção do mesmo, calculando a probabilidade de saída de cada cliente para obter a taxa de *churn*. Estes autores consideram uma segmentação sobre as variáveis de valor atual, valor potencial e lealdade do cliente, em que cada uma destas variáveis, normalizada entre 0 e 1, é um eixo de um espaço tridimensional. Neste artigo, não foi considerado nenhum algoritmo de *clustering*, uma vez que foi tido em conta que cada segmento era um octante do espaço resultante da translação da origem das coordenadas para o ponto (0.5, 0.5, 0.5), criando estratégias de marketing para cada um destes segmentos.

Os autores em [41] consideram uma análise de *clusters* para resolver o problema de agrupamento de consumidores, utilizando informação básica sobre o cliente como a sua identificação, os produtos contratualizados, o tempo de acesso à rede, os serviços contratados, os benefícios, as reclamações, o valor do cliente como as mensalidades pagas e outros serviços de valor acrescentado e os custos associados ao cliente. Foi também considerada informação comportamental como é o caso da duração das chamadas, as horas a que foram feitas e a quantidade de números de telefone distintos contactados. Neste artigo, foi utilizado o algoritmo de *k-means clustering*, apresentando as seguintes vantagens que justificam a sua escolha: capacidade de chegar a uma boa solução em que, frequentemente, se consegue alcançar um mínimo, ainda que local; escalável e eficiente em *datasets* de elevadas dimensões; não é sensível à ordem do *input*; assume que a incapacidade de lidar com dados com ruído não impacta a performance do algoritmo, uma vez que o *dataset* está relativamente completo e que o pré-processamento de dados pode compensar este problema. Para fazer o pré-processamento dos dados, foi utilizada a ferramenta de prospeção de dados *KXEN*, por ser orientada para os resultados, em vez de ser orientada para os procedimentos, conseguindo um processamento automático dos *missing values* e *outliers*, não precisando de espaço adicional para armazenar os dados, uma vez que são armazenados na própria *data warehouse*, não sendo também necessário a escolha do melhor modelo e respetivos parâmetros, visto que são determinados automaticamente, ao utilizar a teoria de minimização de risco estrutural de Vapnik.

MODELAÇÃO E EXPLORAÇÃO DE DADOS

Neste capítulo foi descrito o processo de modelação, aquisição e pré-processamento dos dados recolhidos. É realizada uma primeira abordagem ao desafio de clusterização aplicada à segmentação de cliente, com o algoritmo *k-means*, primeiramente, a cada pilar de segmentação, e em seguida, considerando todos os pilares em simultâneo.

3.1 Modelação

Nesta dissertação procurou-se segmentar a base de clientes da NOS, dos clientes em Portugal Continental que possuem serviço de Internet e televisão fixa e, com esse fim em mente, foram utilizados dados recolhidos pela empresa, referentes à utilização diária de internet e televisão, à descrição do contrato celebrado com o cliente e de informação sobre os contactos entre a empresa e o cliente. Estes dados representam o comportamento atual dos clientes, não existindo problemas comuns de recolha de dados por inquéritos como, por exemplo, os inquiridos serem afetados por memória imperfeita ou outros estilos de resposta, ainda que à custa de assunções sobre as razões que motivam o comportamento do cliente, uma vez que só conhecemos os efeitos e não as suas causas.

Neste projeto, foram considerados os dados sobre os clientes da empresa, o que totaliza mais de um milhão de contratos. Consideramos que, apesar de um mesmo cliente poder celebrar mais do que um contrato com a empresa, as suas necessidades, expetativas e utilização do serviço podem não ser partilhadas entre os vários contratos e, por essa razão irão ser consideradas entidades separadas.

O conjunto de dados a ser utilizado reflete as necessidades de caracterização do cliente, segundo o descrito em 2.2.4 e sumarizado na Figura 3.1, tendo os atributos apresentados no *dataset* final resultado da transformação de outras atributos presentes na base de dados da NOS ou através de agrupamentos de variáveis, gerando novos conceitos e subsequentemente novo conhecimento.

Durante este processo, foram materializadas variáveis de forma a proceder a uma caracterização geográfica, demográfica, psicográfica e comportamental, de acordo com os pilares de segmentação considerados relevantes para a modelação do cliente, de acordo

com o especificado na Subsecção 2.2.4.

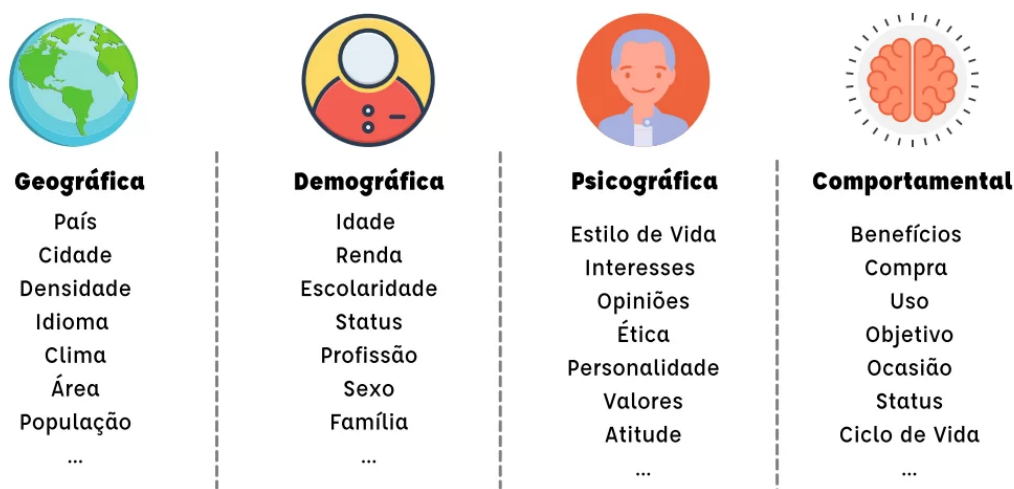


Figura 3.1: Pilares de segmentação e informações contempladas por cada um dos pilares de segmentação. [33].

3.1.0.1 Segmentação Geográfica

Localização O cliente, quanto à sua localização geográfica, é caracterizado segundo o seu distrito, concelho, freguesia e código postal registados no contrato.

3.1.0.2 Segmentação Demográfica

Não existem variáveis recolhidas pela empresa que permitam fazer uma caracterização demográfica, além da variável idade e, mesmo assim, é, apenas, uma estimativa. Assim, decidimos imputar esses dados à base de dados PORDATA como detalhado em 3.2.1.

3.1.0.3 Segmentação Psicográfica

Tipo de utilização de Internet caracterização do tipo de tráfico realizado, com uma baixa granularidade, relativamente à utilização de internet em serviços de streaming, redes sociais ou consulta de notícias. Com esta informação, tentamos caracterizar os interesses do agregado familiar quanto à utilização da internet e, como essas utilizações não têm todas a mesma intensidade de utilização nem a mesma sensibilidade a degradações de serviço, conseguimos caracterizar também a sensibilidade do cliente dada a utilização.

Estilo de vida Uma vez que estamos a caracterizar o cliente de acordo com consumos de internet fixa e televisão, assumimos que a utilização destes serviços é indicador se o cliente está ou não em casa. Gostaríamos de caracterizar a sua utilização durante os dias da semana, fins de semana e feriados.

Extensores de sinal Os extensores de sinal garantem Internet em toda a casa sem falhas e com máxima velocidade. A presença destes dispositivos revelam quer sofisticação tecnológica quer má cobertura da rede Wi-fi tanto problemas de estabilidade ou baixas velocidades, quer de *download* como de *upload* do sinal de Wi-Fi.

3.1.0.4 Segmentação comportamental

Número de equipamentos Após um levantamento dos nomes dos dispositivos que estão visíveis na rede local através do seu endereço MAC, decidimos agrupá-los de acordo com as seguintes categorias: telefones, televisões, *routers*.

Quanto à contabilização das televisões, considerou-se os endereços MAC das *set top boxes*. Quanto à contabilização dos *routers*, considerou-se os endereços MAC, salvaguardando que estes *routers* não tenham um número de telefone associado, o que sinaliza que este endereço observado se pode tratar de um telefone que iria ser mal classificado como *router*. Utilizando o nome com o qual estes dispositivos se registam na rede local, classificamos os dispositivos como *routers*, telefones, HUB, *plumes*, *powerboxes*, *boxes* de TV.

Intensidade de utilização O cliente foi modelado quanto à sua intensidade de serviço de internet fixa e televisão. Quanto à internet fixa, o cliente foi modelado de acordo com a velocidade contratada, hipotizada como uma característica do cliente reveladora de sofisticação tecnológica, especialmente quando opta pela maior velocidade de internet, o que reflete uma maior predisposição a sentir quebras de qualidade de serviço; número de *routers*, uma característica de maior intensidade de serviço e sensibilidade a degradações da qualidade de serviço entregue, quer seja por ser uma maior área do local onde o serviço foi contratado como uma necessidade de cobertura maior; total de bytes de upload e de download. Quanto à utilização de TV, o cliente foi modelado de acordo com o número de sessões em [Video on Demand \(VoD\)](#).

Utilização ocasional Procuramos caracterizar os clientes que têm mais do que um contrato em que um deles não regista atividade durante vários dias seguidos ou, até mesmo, vários meses seguidos sem qualquer atividade, sendo esses contratos caracterizados como "segundas" habitações ou casas de férias.

Lealdade à marca A lealdade à marca é definida como a associação positiva entre os clientes e a empresa de telecomunicações. A compra repetida de produtos ou serviços, independentemente dos esforços da concorrência em tentar criar melhores ofertas tanto em relação ao preço como em relação às características da oferta. Segundo [18], uma das maiores táticas para criar lealdade à marca passa por oferecer um serviço de apoio excepcional. Assim, o cliente é caracterizado de acordo com o seu tempo enquanto cliente, período até ao final da fidelização, número de vezes que o cliente é colocado num programa de retenção e os seus motivos.

3.2 Recolha de dados

Nesta dissertação procurou-se segmentar a base de clientes em Portugal Continental que possuem serviço de Internet e televisão fixa, apenas nos municípios de maior interesse estratégico da empresa, sendo estes Almada, Amadora, Cascais, Lisboa, Loures, Porto, Sintra e Vila Nova de Gaia.

A informação recolhida foi, posteriormente, reagrupada, segundo os pilares de segmentação, variáveis geográficas, demográficas, psicográficas e comportamentais e, em seguida, pré-processada.

3.2.1 Dados da PORDATA

A falta de informação demográfica sobre os clientes na base de dados mantida pela empresa, que apenas tem informação sobre a idade, não tendo qualquer informação sobre outros atributos demográficos tais como o género, rendimentos, educação, ocupação, tamanho do agregado familiar, entre outros, pode ser colmata através da inclusão de dados provenientes da PORDATA.

A PORDATA, a Base de Dados de Portugal Contemporâneo, é organizada e desenvolvida pela Fundação Francisco Manuel dos Santos. A PORDATA dá corpo a uma das prioridades da Fundação: a recolha, organização, sistematização e divulgação da informação sobre múltiplas áreas da sociedade para Portugal, municípios e países europeus. As estatísticas divulgadas são provenientes de fontes oficiais e certificadas, com competências de produção de informação nas áreas respetivas. A Fundação produz o esforço de recolher e organizar a informação disponível, tornando-a mais clara e acessível, juntando-se a este trabalho a adição de informação contextualizadora ou "metainformação". A Fundação disponibiliza esta informação a título gratuito.

A PORDATA disponibiliza informações sobre diversas temáticas, embora, para a caracterização da demografia destes clientes se tenha considerado que os mais relevantes são os temas de: **População**, que contém séries sobre residentes, famílias, nascimentos, óbitos, migrações, casamentos e outros; **Educação**, que contém séries sobre estabelecimentos, alunos, professores, grau de escolaridade e outros, **Habitação, Conforto e Condições de Vida**, que contém séries sobre edifícios, alojamentos, poder de compra e outros; e, por fim, sobre **Emprego e Mercado de Trabalho**, que contém séries sobre a população empregada, desempregada e inativa, salários, greves e outros.

Cruzando a necessidade destes atributos para explicar a demografia dos clientes, com os dados que a PORDATA disponibiliza foram considerados dados, municipais, sobre a **população residente por grandes grupos etários** [28], sendo a população dividida em 3 grupos, jovens, dos 0 aos 14 anos, pessoas em idade ativa, dos 14 aos 64 anos, e idosos, com 65 ou mais anos, estes dados são relevantes uma vez que a idade dos membros do agregado familiar traduz-se em diferentes níveis de sofisticação, intensidade e expectativas em relação ao serviço oferecido; sobre as famílias segundo os censos: total e

por número de indivíduos [13], sendo considerados os seguintes dimensões familiares, unipessoais, bipessoais, com 3 a 5 elementos, inclusive e famílias numerosas, com 6 ou mais elementos, a relevância da inclusão destes dados deve-se a que famílias de maior de dimensão tenham necessidades de utilização superiores e com necessidades mais diversificadas do que famílias de menor dimensão, uma vez que estes agregados familiares podem conter elementos com idades dispares; sobre a **População residente com 15 e mais anos segundo os Censos: total e por nível de escolaridade completo mais elevado** [27], sendo a população dividida pelos seguintes níveis de educação: sem nível de educação, com o 1º ciclo, com o 2º ciclo, com o 3º ciclo, com o ensino secundário, médio e superior, uma informação relevante para entender a sofisticação tecnológica e a sua capacidade de auto-resolução de problemas técnicos ao invés de optar pela participação técnica deste problema; sobre **Remuneração base média mensal dos trabalhadores por conta de outrem: total e por sexo** [29], para o qual só consideramos a coluna correspondente ao total da população, uma vez que não é possível discriminar os clientes pelo género, uma vez que rendimentos mais elevados podem levar à aquisição de mais serviços *premium* e a pagamentos de ARPU mais elevados; sobre a **Poder de compra per capita** [25], tentando complementar a informação relativa aos rendimentos auferidos pelos clientes com uma quantificação do bem-estar material, tendo em conta o custo de vida naquele município. Os dados recolhidos foram sistematizados, para os municípios de interesse estratégico, no Apêndice A nas tabelas A.1, A.2, A.3 e A.4.

Dado que as tabelas A.1, A.2 e A.4 contêm dados relativamente à composição da população quer em termos de idade, quer de nível de educação, quer de dimensão dos agregados familiares, em número de cidadãos de cada município em cada uma das categorias destas variáveis, tornando-as fortemente correlacionadas, uma vez que que municípios de maiores dimensões terão, naturalmente, uma maior quantidade de cidadãos em cada uma das categorias. Torna-se, deste modo, relevante a normalização destes dados, dado que a relevância da incorporação destes dados é a compreensão da composição, percentual, de cada um dos níveis de cada variável.

Uma vez que o procedimento é semelhante para cada uma das tabelas, A.1, A.2 e A.4, pelo que o processo vai ser descrito, apenas, para a tabela A.1. Para cada um dos grandes grupos etários, foi criada uma nova variável com o mesmo nome seguida de percentagem que resulta da fração do número de cidadãos nesse grupo etário e nesse município pelo total de cidadãos no município. O resultado das normalizações resultou nas tabelas A.5, A.6 e A.7.

Observando a distribuição das percentagem da população em cada grande grupo etário 3.2, conclui-se que a população com idades compreendidas entre os 15 e os 64 anos mantêm-se relativamente constantes ao longo dos municípios. O fator interessante torna-se as percentagens de população jovem (entre os 0-14 anos) e a percentagem de população idosa (acima dos 65 anos), sobretudo quando a percentagem de jovens é superior à percentagem de idosos, uma tendência contrária a Portugal Continental. Esta relação foi

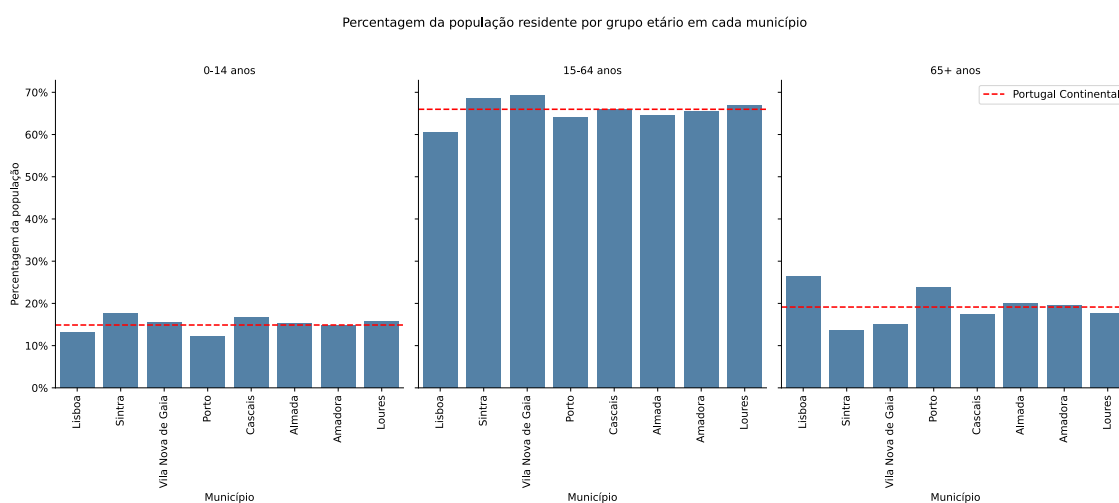


Figura 3.2: Composição da população residente por faixa etária, em 2011.

sistematizada utilizando o conceito de **Índice de Envelhecimento (IE)** que procura responder à pergunta de onde há mais e menos idosos por 100 jovens, sendo calculado a partir da equação e incluída na tabela A.5. Esta métrica torna evidente que o envelhecimento destes municípios é distinto 3.3, os que têm praticamente duas vezes mais idosos que população jovem, dois concelhos que se alinham com o IE em Portugal Continental, e os restantes encontram-se abaixo, ressaltando-se o caso de Sintra que tem mais população jovem que idosa.

$$IE = \frac{\# \text{ de idosos}}{\# \text{ de jovens}} \times 100 \quad (3.1)$$

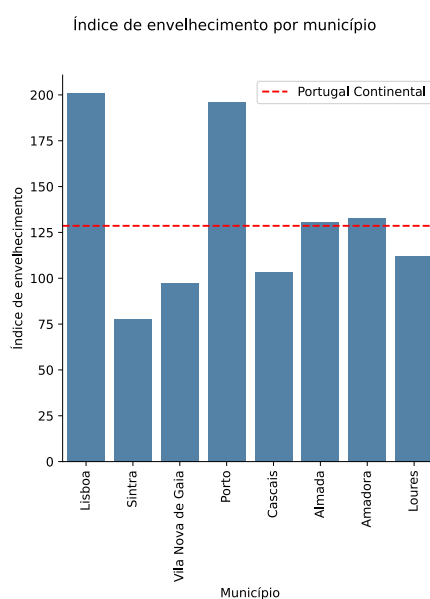


Figura 3.3: Índice de envelhecimento da população, por município, em 2011.

Assim, esta falta de informação demográfica é colmatada, atribuindo, a cada cliente, o poder de compra e ganho médio mensal daquele município. Para as variáveis que foram

normalizadas nas tabelas A.5, A.6 e A.7, foi associado, a cada cliente, o IE, a percentagem de pessoas com cada nível de educação, e a percentagem de famílias com cada dimensão, do município onde residem como métrica de envelhecimento do agregado familiar e a sua propensão a ter cada nível de educação e dimensão, respetivamente.

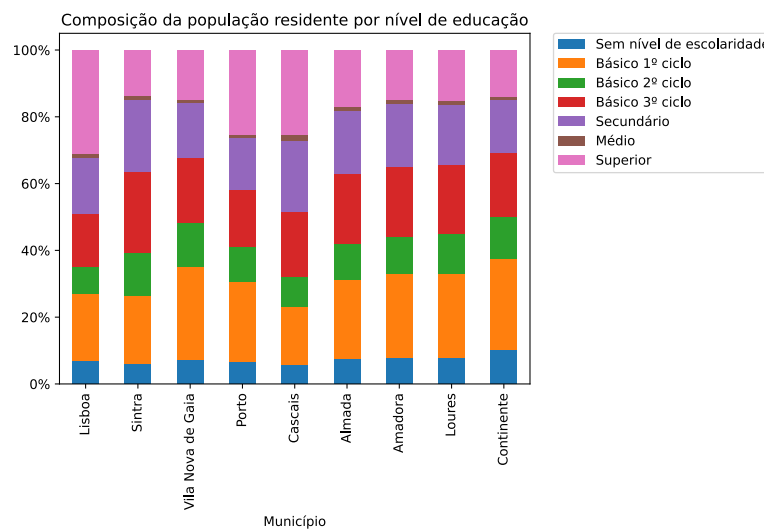


Figura 3.4: Composição do nível de educação, por município, em 2011.

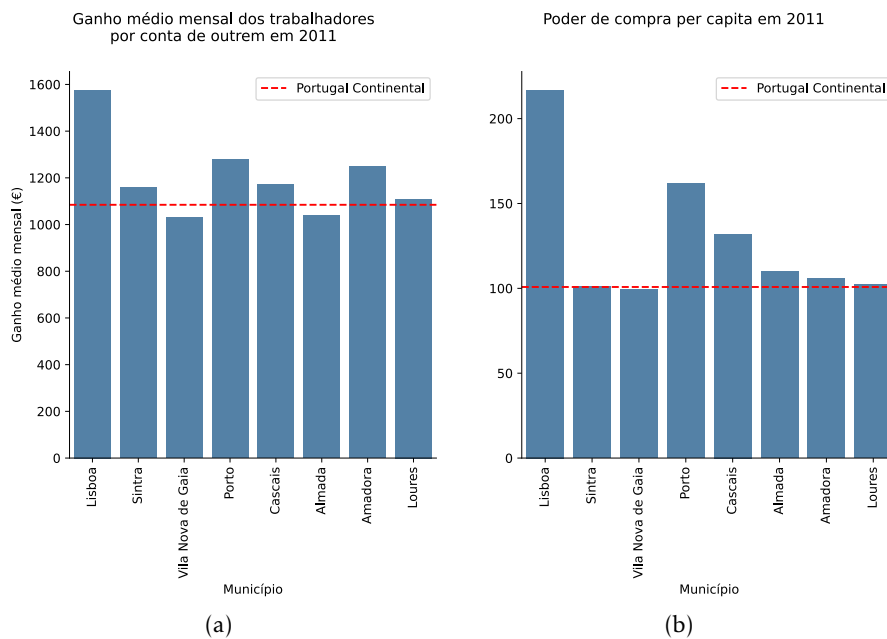


Figura 3.5: Comparação entre o ganho médio mensal, em euros, com o poder de compra de cada município, em 2011.

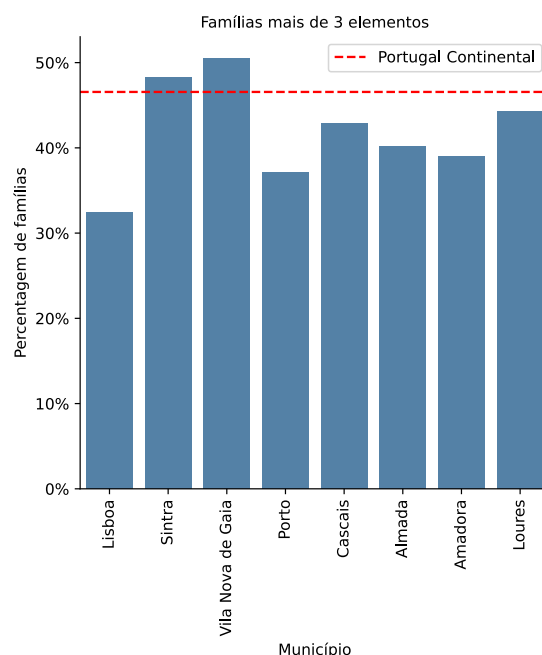


Figura 3.6: Composição das famílias, agrupando as famílias unipessoais com as famílias com dois elementos, por município, em 2011.

3.2.2 Recolha de dados psicográficos

Subscrição de canais *premium* Os canais *premium* relevantes para a modelação do cliente foram a Sport TV, Benfica TV, TV CINE e os restantes foram agrupados numa categoria de outros.

A utilização de internet por protocolos foi considerada a fonte para a caracterização psicográfica dos clientes relativamente ao serviço de *internet*.

Estes dados inicialmente encontravam-se no formato apresentado na tabela 3.1, em que para cada endereço MAC é registado o protocolo dominante num intervalo de quinze minutos, bem como o número de bytes recebidos e transmitidos.

Os protocolos possíveis no campo "App Protocol" da tabela estão descritos na tabela 3.2 em que cada protocolo foi caracterizado de acordo com a sofisticação tecnológica e tolerância a falhas por parte dos clientes que a utilizam, e por fim qual a percentagem dos *routers* dos clientes que utilizam este protocolo.

Considerando a tabela 3.2, concluiu-se que as variáveis necessárias para caracterizar os clientes quanto as suas expectativas relativamente ao seu serviço de *internet* são *flags* relativamente à sua utilização dos protocolos *remote connectivity*, *network admin*, *online games*, *database*, *tunneling*, *file access* e *VoIP*.

mac	App Protocol	Period Start	Period End	Total Received Bytes	Total Transmitted Bytes
xx-xx-xx-xx-xx-xx	Network Storage	2021-12-30 15:30:00	2021-12-30 15:45:00	22,857	10,817
xx-xx-xx-xx-xx-xx	Instant Messaging	2021-12-30 15:30:00	2021-12-30 15:45:00	4,706,324	429,226
xx-xx-xx-xx-xx-xx	Instant Messaging	2021-12-30 15:30:00	2021-12-30 15:45:00	5,113	4,295
xx-xx-xx-xx-xx-xx	Web Video	2021-12-30 15:30:00	2021-12-30 15:45:00	297,314	61,439
xx-xx-xx-xx-xx-xx	Software Updates	2021-12-30 15:30:00	2021-12-30 15:45:00	28,301	15,942

Tabela 3.1: Excerto da tabela de utilização de *internet* pelo discretizado em períodos de 15 minutos, por cada endereço MAC.

App Protocol	Sofisticação tecnológica	Tolerância a falhas	% de macs que utilizam
Anonymity	média	média	0.00
News Groups	baixa	alta	0.03
Peercasting	média	baixa	0.04
Remote Connectivity	alta	média	0.24
Database	média	média	0.82
Email	baixa	alta	1.29
Tunneling	alta	média	1.39
Network Admin	média	alta	1.44
File Access	média	média	1.84
Online Games	média	baixa	3.03
VoIP	baixa	baixa	28.05
Network Storage	média	média	70.13
Web video	baixa	média	74.09
Instant Messaging	baixa	alta	75.07
Software Updates	baixa	alta	75.71
Peer to Peer	média	média	84.46
Streaming Protocols	baixa	baixa	88.39
Web Browsing	baixa	alta	88.92

Tabela 3.2: Protocolos de *internet* possíveis na variável *App Protocol* presente na tabela 3.1

3.2.3 Dados comportamentais

Dados sobre o tipo de equipamentos em casa do cliente, tais como a quantidade de equipamentos, a quantidade de números de telefone fixo, número de routers, quantidade de telefones, quantidade de HUBs, de extensores de sinal Wi-Fi, boxes de TV, e quantidade de TV. Quanto ao comportamento passado relativamente ao churn, destaca-se as flags de churn parcial voluntário, churn voluntário, churn involuntário, churn voluntário total. Informação sobre quantos contratos estão associados ao NIF registado nesta conta.

Quanto à fidelização, o cliente pode não estar fidelizado, fidelizado durante um período inferior, igual ou superior a dois anos. Foi recolhido, também, o número de dias até ao período de fidelização acabar e o número de dias desde a última ativação do serviço.

A intensidade de utilização da internet é caracterizada com a quantidade de tráfego de download e upload médio ao longo de uma semana. A intensidade de utilização da televisão, pela quantidade de sessões em VoD, tanto em dias de semana como em fins-de-semana.

3.3 Criação da amostra

Os clientes representados na amostra são provenientes dos concelhos considerados mais relevantes pela empresa, sendo eles, Lisboa, Sintra, Vila Nova de Gaia, Porto, Cascais, Almada, Amadora, Loures.

Foi extraída uma amostra de 15,000 clientes, estratificada nas seguintes variáveis, concelho, idade, ARPU e *pts_flg*.

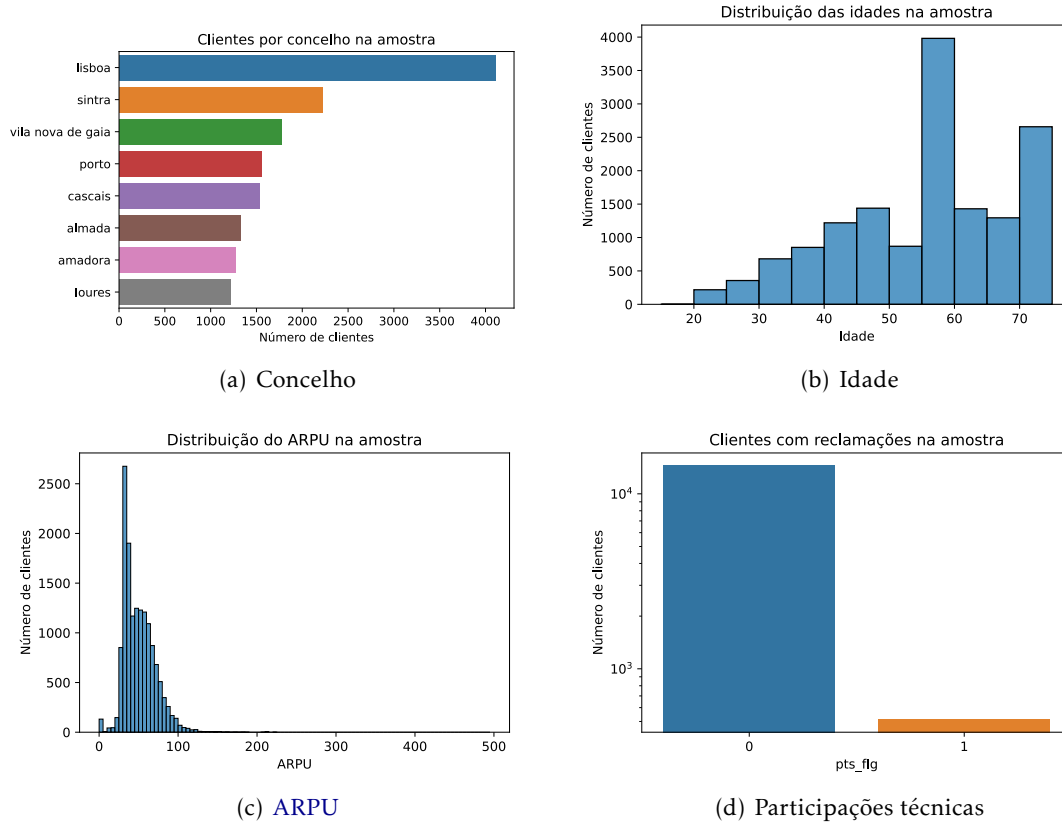


Figura 3.7: Comparação da prevalência de clientes dos diferentes concelhos considerados entre a população e a amostra estratificada extraída.

3.4 Pré-processamento

Normalização do conjunto de dados Para cada atributo no conjunto de dados X , foi efetuado uma normalização *min-max*, que coloca cada atributo numa escala entre $[0, 1]$

$$x = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3.2)$$

One-hot encoding Para cada atributo categórico, codifica cada categoria usando one-hot, o que transforma cada categoria numa coluna binária.

3.5 Clusterização com k-means

A clusterização surgiu como uma ferramenta para explorar a estrutura dos dados, e tem como base a ideia de particionar os dados de acordo com o critério de os objetos semelhantes serem colocados num mesmo *cluster* e objectos dissemelhantes serem colocados em clusters diferentes [40]. Este tipo de técnica é especialmente relevante para aplicações de aprendizagem automática não supervisionadas, para o qual não sabemos a verdadeira

classe de cada observação ou até mesmo em casos em que a sua verdadeira classe não é consensual entre os especialistas.

A primeira abordagem a este problema foi feita recorrendo ao método de *k-means clustering*.

K-means clustering [22] Um dos mais populares métodos de análise de dados que produz uma partição das entidades em *clusters*, sem sobreposição, devolvendo os centróides que representam esses *clusters*. Este método introduz um conjunto de vetores μ_k , em que μ_k é o protótipo associado ao k^{th} *cluster*. O objetivo deste algoritmo é encontrar um conjunto de vetores μ_k tal que a distância de cada ponto ao μ_k mais próximo seja mínima. A função objetivo, que representa a soma dos quadrados das distâncias de cada ponto até ao respetivo vetor μ_k , é dada por:

$$J = \sum_{n=1}^K \sum_{k=1}^K r_{nk} \|x_n - \mu_k\|^2 \quad (3.3)$$

Os valores ótimos para r_{nk} e μ_k podem ser descobertos através de um método iterativo que envolve dois passos em cada iteração, um passo de expectativa, minimização da função J em relação a r_{nk} , mantendo μ_k constante, e outro de maximização, minimizando a função J em relação a μ_k , mantendo r_{nk} constante. r_{nk} é definido de tal forma que r_{nk} é 1 para o valor de k cujo $\|x_n - \mu_k\|^2$ é mínimo, o que intuitivamente é o mesmo que atribuir o ponto n ao centroide mais próximo:

$$r_{nk} = \begin{cases} 1 & \text{se } k = \operatorname{argmin}_j \|x_n - \mu_j\| \\ 0 & \text{caso contrário} \end{cases} \quad (3.4)$$

3.6 Métricas de avaliação

3.6.1 Índices de validação internos

Silhouette Score Proposto por Peter Rousseeuw em [31] com o intuito de criar uma avaliação da validade de um exercício de clusterização e a escolha de um número apropriado de *clusters*.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3.5)$$

em que $a(i)$ representa a dissimilaridade média entre a observação i e todas as observações do seu cluster C_I .

$$a(i) = \frac{1}{\|C_I\| - 1} \sum_{j \in C_I, i \neq j} d(i, j) \quad (3.6)$$

e que $b(i)$ representa a dissimilaridade média entre a observação i e todas as observações do seu *cluster* mais próximo.

$$b(i) = \min_{J \neq I} \frac{1}{\|C_J\|} \sum_{j \in C_J} d(i, j) \quad (3.7)$$

O *silhouette score* para uma amostra foi calculado

$$s = \frac{b - a}{\max\{a, b\}} \quad (3.8)$$

a = distância média entre todas as observações de uma amostra e todas as observações da mesma classe; b = distância média entre todas as observações de uma amostra e todas as observações do *cluster* mais próximo;

Este índice de validação, que varia entre -1 e 1, deve ser maximizado, uma vez que valores mais elevados indicam *clusters* mais densos e melhor separados, o que coincide com a definição natural de *cluster*. Valores próximos de 0 indicam *clusters* que se sobrepõem, e valores próximos de -1 indicam uma clusterização incorreta.

Índice de Calinski-Harabasz Também conhecido pelo *Variance Ratio Criterion* [4], é um método de avaliação de modelos de clusterização quando não são conhecidos as verdadeiras *labels* das observações. É definido como o rácio entre a dispersão entre *clusters*, representado pela matriz B_k e a dispersão dentro do próprio *cluster*, representado pela matriz W_k .

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1} \quad (3.9)$$

$$W_k = \sum_{q=1}^k \sum_{x \in C_q} (x - c_q)(x - c_q)^T \quad (3.10)$$

$$B_k = \sum_{q=1}^k n_q (c_q - c_E)(c_q - c_E)^T \quad (3.11)$$

em que E é o conjunto de dados, C_q é o conjunto de observações do cluster q , c_q o centro do cluster q , c_E é o centro de E e n_q , o número de observações no cluster q .

Índice de Davies-Bouldin Proposto por David L. Davies e Donald W. Bouldin em 1979 [9] é um índice de validação interna que mede a similaridade média entre cada *cluster* e o seu cluster mais próximo. Este índice é definido como:

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} R_{ij} \quad (3.12)$$

em que R_{ij} é definido como:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (3.13)$$

sendo s_i a distância média de cada ponto do *cluster* i e o centróide desse *cluster*; d_{ij} a distância entre os centróides dos *clusters* i e j .

Assim, este índice é tanto melhor quanto mais próximo de 0, uma vez que o índice considera o *cluster* com um maior R_{ij} , ou seja, o *cluster* mais próximo, e a partição é tão melhor quanto mais distantes os centróides dos dois *clusters* mais próximos e menor for a soma das distâncias médias dos pontos *clusters* i e j aos seus respectivos centróides.

Bayesian Information Criterion (BIC) Proposta por Gideon Schwarz em 1978 [32], como um critério de seleção da dimensionalidade de um modelo dado um conjunto de dados, utilizando um critério com base em teoria da informação. O critério utilizado para escolher o melhor modelo foi maximizar a seguinte métrica:

$$\log M_j(X_i, \dots, X_n) - \frac{1}{2} k_j \log(n) \quad (3.14)$$

onde M_j é o modelo j , k_j é o número de parâmetros estimados para o modelo, n é a dimensão do conjunto de dados X e X_i é a observação i do conjunto de dados X .

3.7 Resultados experimentais utilizando k-means

3.7.1 Experiências considerando apenas uma vista

3.7.1.1 Apenas vista demográfica

Considerando, apenas, a **Vista Demográfica (VD)**, aplicou-se o algoritmo de *k-means* a fim de obter uma segmentação demográfica da base de clientes.

Observando a Figura 3.8 que contém as métricas internas para cada *k* considerado conclui-se que o melhor *k* é 18.

Escolhendo, apenas, os três segmentos com maior percentagem de clientes com participações técnicas tanto de TV como de NET. Os segmentos 2, 7 e 15 são aqueles que têm uma maior quantidade de participações técnicas de TV e são, também, onde a idade do titular é mais avançada. Os segmentos 4, 6 e 10 são os segmentos com maior percentagem de clientes com participações técnicas de NET. O segmento 6 é um segmento em que o índice de envelhecimento é muito alto e com uma idade do titular mais baixa, enquanto os segmentos 4 e 10 são segmentos com uma idade do titular média, com as maiores percentagens de famílias com três ou mais elementos e população com o 2º ciclo. Contudo, o segmento 4 tem uma muito maior percentagem da população sem nível de escolaridade que o segmento 10 e, por isso, o segmento 4 tem uma maior percentagem de clientes participações técnicas de TV face ao segmento 10.

Segmento	age	índice envelhecimento	Famílias com mais de 3 elementos (%)	Poder de Compra	Sem Nível Escolaridade (%)	2 ciclo (%)
2	0.883	0.158	1.000	0.000	0.778	1.000
7	0.927	1.000	0.000	1.000	0.520	0.000
15	0.911	0.208	0.577	0.279	0.000	0.147
4	0.611	0.158	1.000	0.000	0.778	1.000
6	0.370	0.960	0.258	0.531	0.436	0.398
10	0.635	0.000	0.873	0.019	0.123	0.876

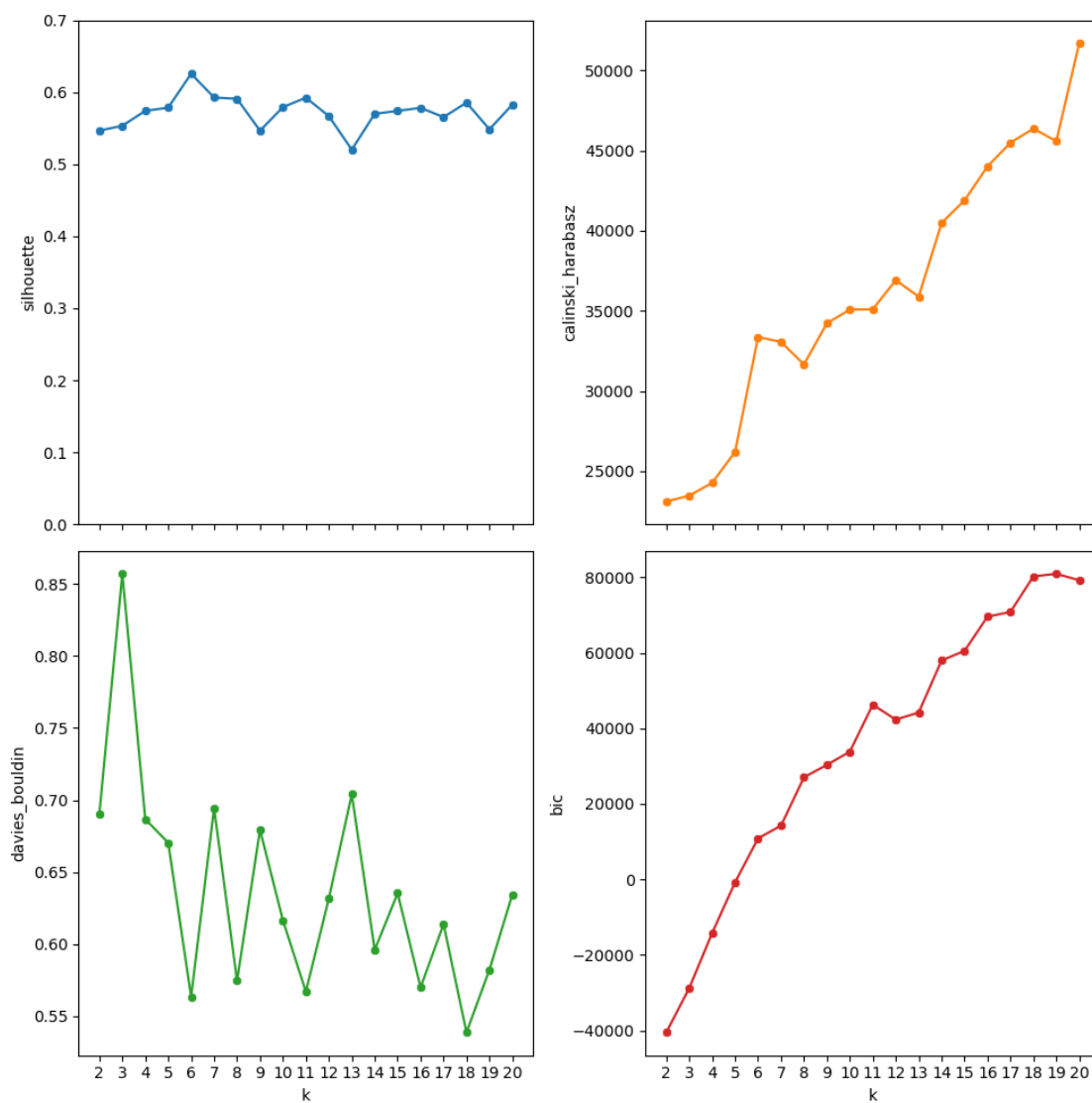


Figura 3.8: Comparação das métricas de validação internas, considerando o algoritmo *k-means* para todos os clientes da amostra, considerando, apenas, dados demográficos.

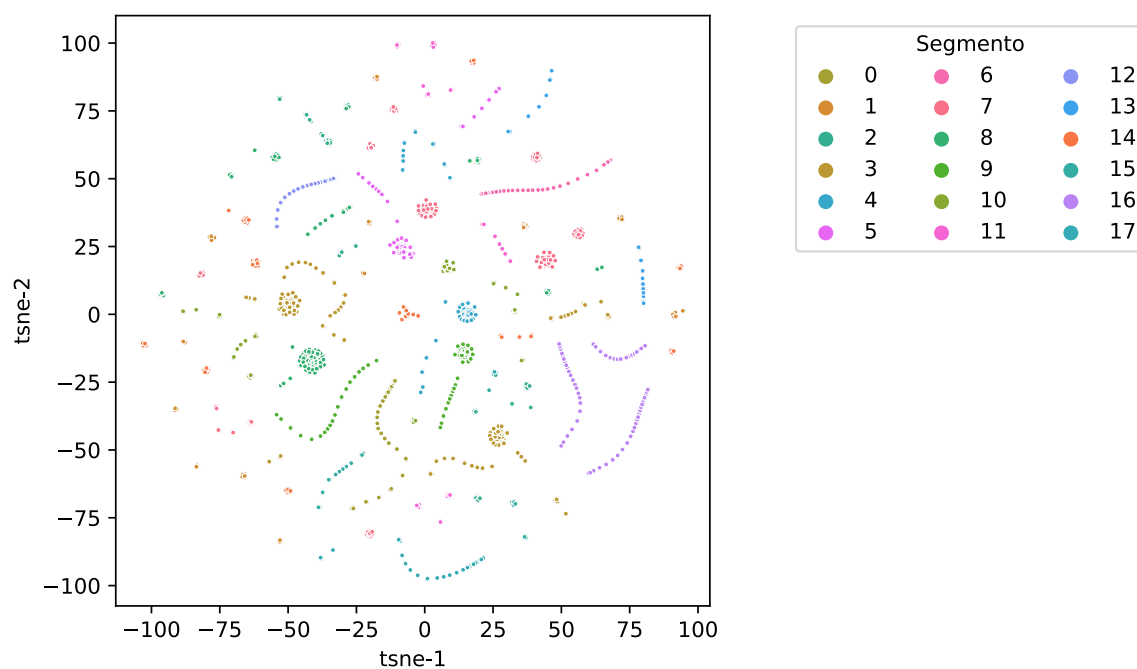


Figura 3.9: Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelo algoritmo *k-means* para todos os clientes da amostra, utilizando, apenas, dados demográficos.

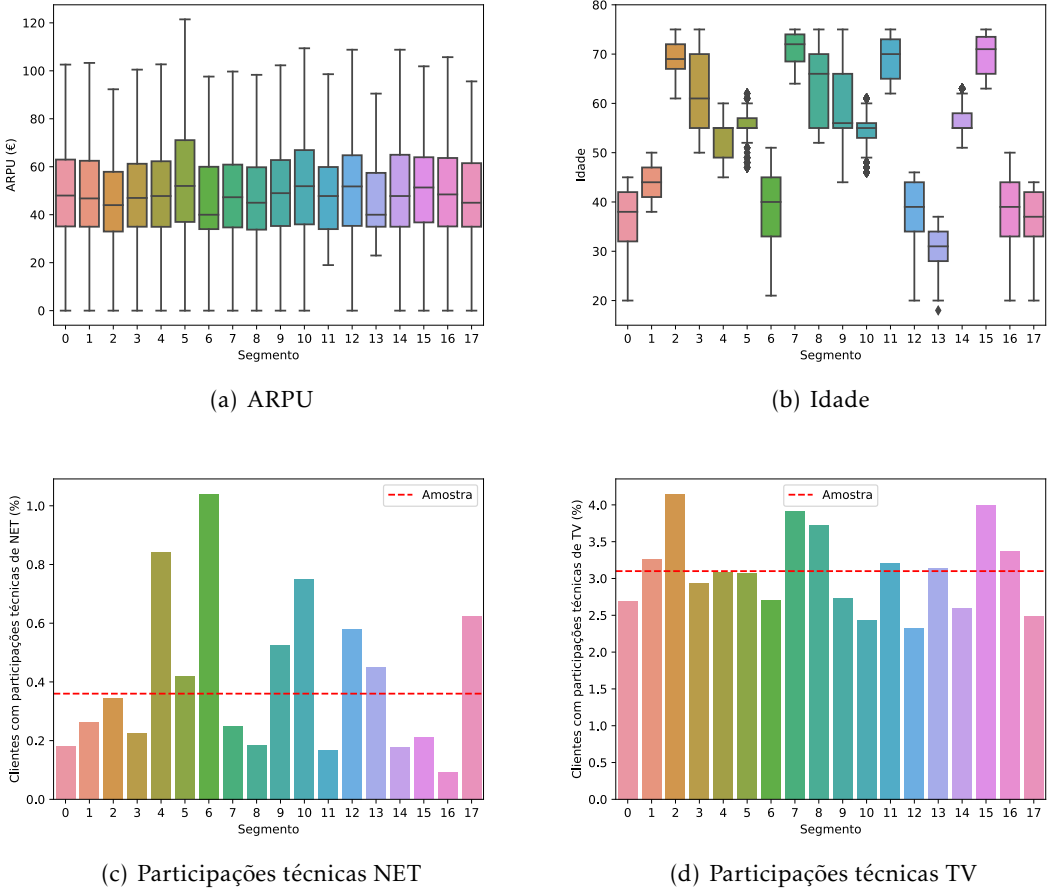


Figura 3.10: Comparação da prevalência de clientes com participações técnicas entre a população e a amostra estratificada extraída, considerando, apenas, dados demográficos.

3.7.2 Apenas vista psicográfica

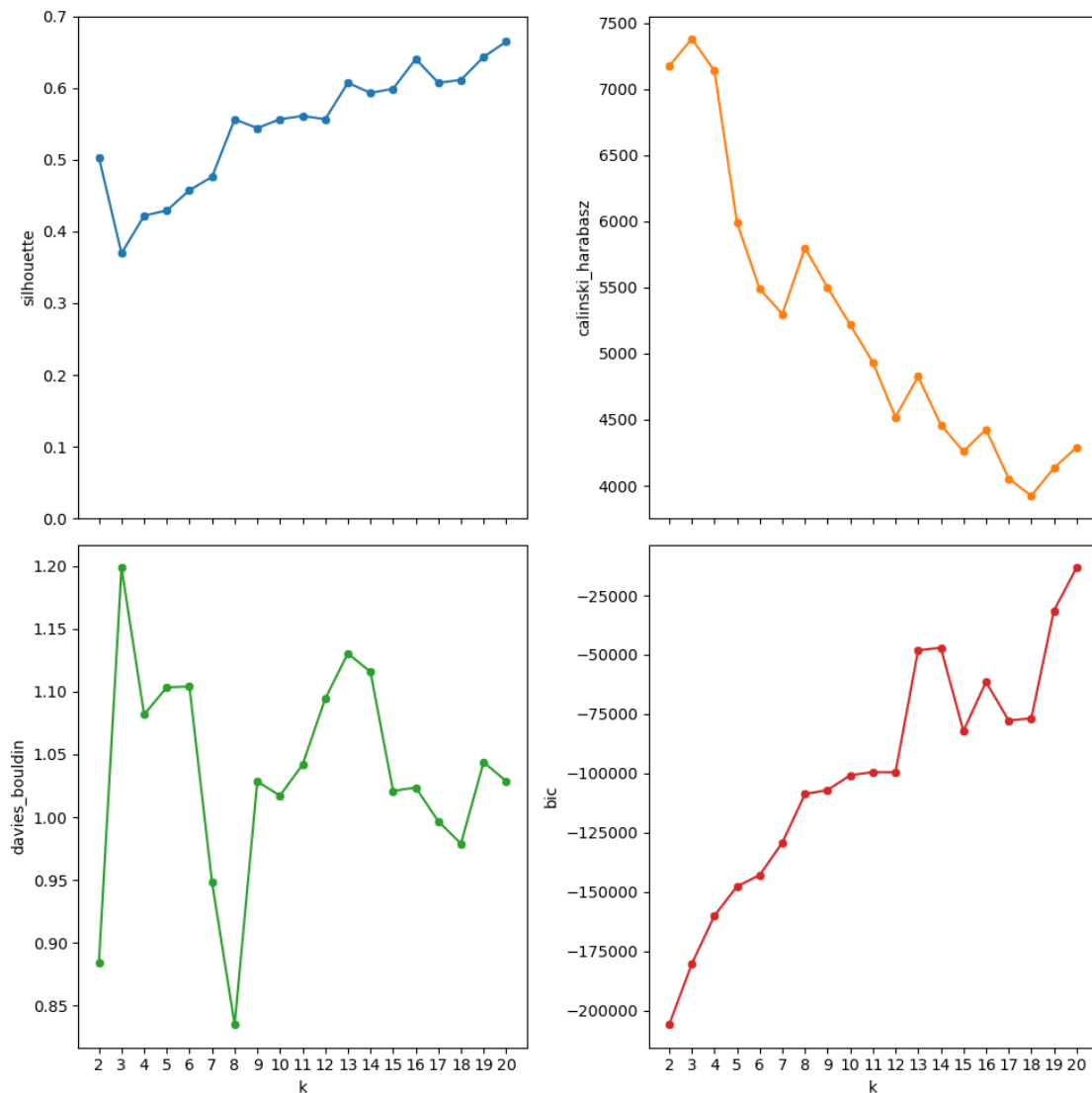


Figura 3.11: Comparação das métricas de validação internas, considerando o algoritmo *k-means* para todos os clientes da amostra.

Variáveis	c_4	c_{10}	c_{14}	c_{15}	c_{16}
digital_invoice_flg	0.852	0.978	0.986	1.000	0.000
direct_payment_flg	0.000	1.000	0.391	0.508	1.000
no_marketing_flg	0.038	0.038	0.015	0.000	0.083
tv_cine_qty	0.000	0.162	1.000	0.142	0.000
benfica_tv_qty	0.007	0.010	0.016	0.004	0.003
sport_tv_qty	0.008	0.011	0.020	0.008	0.002
other_premium_qty	0.004	0.008	0.021	0.006	0.001
remote_connectivity_flg	0.000	0.184	0.000	1.000	0.000

network_admin_flg	0.000	1.000	1.000	1.000	0.000
online_games_flg	1.000	0.908	0.949	0.958	0.090
database_flg	0.000	0.000	1.000	1.000	0.000
tunneling_flg	0.000	0.995	1.000	1.000	0.000
file_access_flg	0.031	1.000	0.986	1.000	0.021
voip_flg	1.000	1.000	1.000	1.000	1.000

Os segmentos com uma maior percentagem de clientes com participações de *internet* são os segmentos 14, 15 e 4, respetivamente. O segmento 14 é caracterizado pelo segmento que exhibe uma utilização sofisticada de *internet*, apenas não utiliza *remote connectivity*, e é o segundo segmento com mais participações técnicas de TV, uma vez que é um segmento que contrata TVCINE, e dos segmentos analisados com mais detalhe que tem maior contratação de serviços *premium*. O segmento 15 é o segundo segmento que tem o maior número de participações técnicas utiliza todos os protocolos de *internet* sofisticados e, o terceiro segmento com um maior número de participações técnicas de NET é o segmento dos *gammers* sem utilização de outros protocolos sofisticados de internet.

O segmento com um maior percentagem de clientes com participações técnicas de TV, o segmento 16, não tem expressão de utilização de protocolos de *internet* sofisticados ou contratação de canais *premium*, mas que possuem débito direto, características de uma população mais envelhecida. O terceiro segmento com maior percentagem de participações técnicas de TV é o segmento 10, é o cliente digital, com fatura digital e débito automático, alguma expressão de subscrição de canais premium e utilização sofisticada de internet.

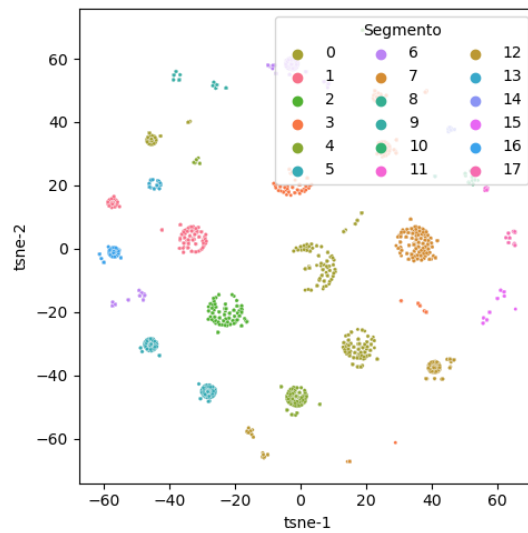


Figura 3.12: Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelo algoritmo *k-means* para todos os clientes da amostra, utilizando, apenas, dados psicográficos.

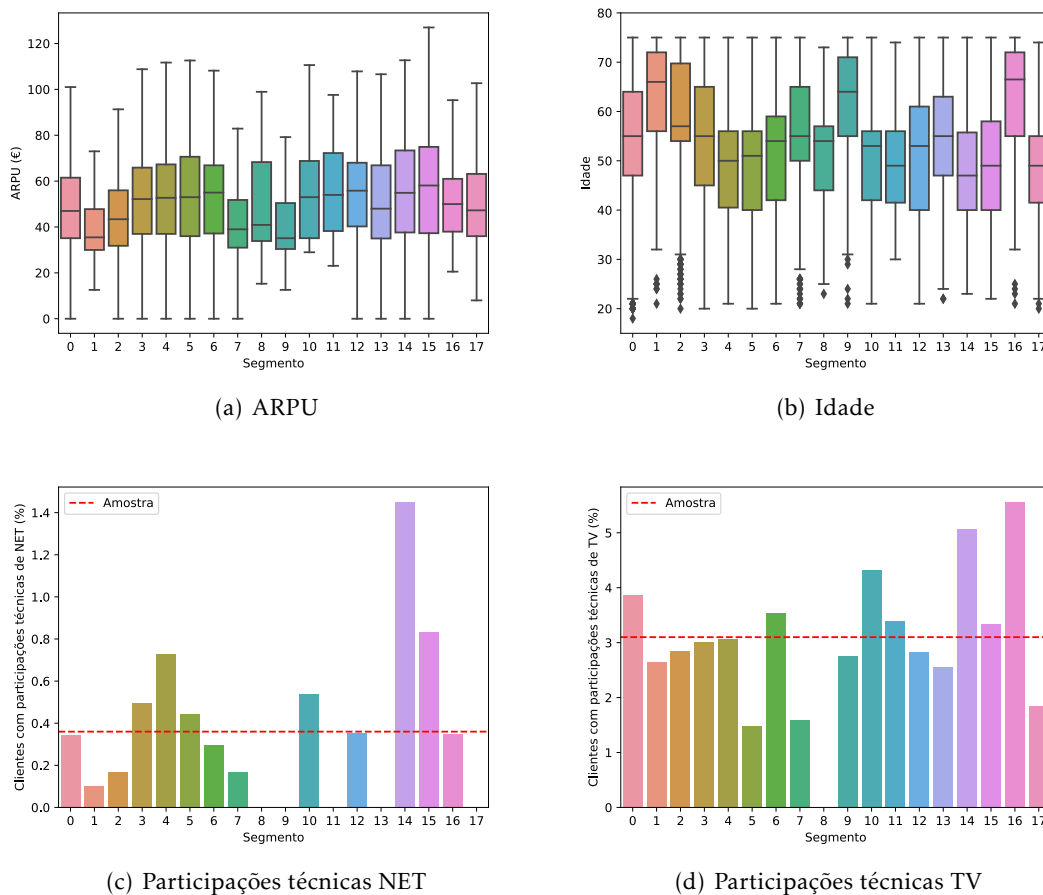


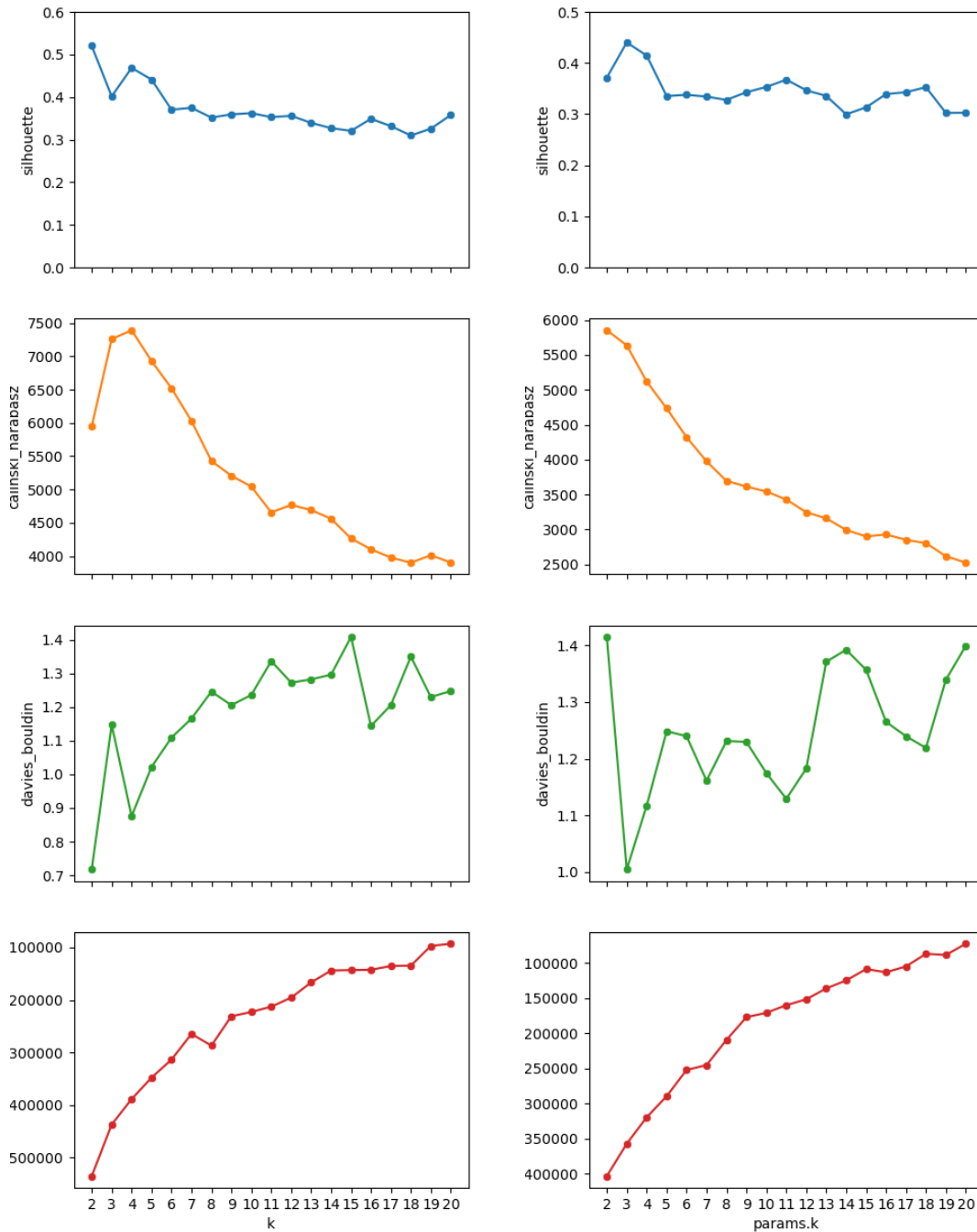
Figura 3.13: Comparação da prevalência de clientes com participações técnicas entre a população e a amostra estratificada extraída, considerando, apenas, dados psicográficos.

3.7.3 Apenas vista comportamental

Quando consideramos toda a amostra de cliente o melhor k é 2, em que se separam os clientes que utilizam *internet* dos que não utilizam internet. Ao remover os clientes que não utilizam, simultaneamente, a rede 2.4Gz, nem a rede 5Gz, nem se ligam por cabo *ethernet*, quer em dias de semana quer em feriados, quer em fins de semana (fortemente correlacionada com a utilização nos feriados), realizou-se uma segunda segmentação comportamental.

Na clusterização final do pilar comportamental, o segmento 0 é aquele que contém os clientes sem utilização de *internet* que foram removidos, mas os restantes segmentos isolam, apenas, os clientes com base na *box* de TV que contratualizam, os clientes com uma box IRIS isolam-se no segmento 1, os clientes sem serviço de televisão (NÃO APLICÁVEL) isolam-se no segmento 2 e os clientes com uma box UMA isolam-se no Segmento 3. Deste modo, observa-se que os clientes com maior propensão a participações técnicas de *NET* são aqueles que não contratualizam serviço de com box de TV, mas que mesmo assim pagam *ARPU* praticamente tão altos como os clientes que contratualizam a box mais avançada da empresa, a UMA. O segmento 3, o que tem os clientes que subscrevem a box UMA, é o segmento mais propenso a participações técnicas de TV, uma vez que, como é a box mais avançada da empresa, é contratualizada por clientes mais sofisticados e, desta forma, mais exigentes com o serviço de TV prestado. Já o segmento 0, sem utilização de *NET* e *VoD*, é o segmento menos propenso a reclamações quer de *NET* como de TV, mas também é o com maior propensão a ter mais do que uma habitação, uma vez que é o segmento com maior (*sa_same_fiscal_num_qty* número se *Identificador do contrato, do inglês Service Account (SA)* que partilham o mesmo *Número de Identificação Fiscal (NIF)*).

3.7. RESULTADOS EXPERIMENTAIS UTILIZANDO K-MEANS



(a) Todos os clientes

(b) Com utilização de internet

Figura 3.14: Comparação das métricas de validação internas, considerando o algoritmo *k-means*

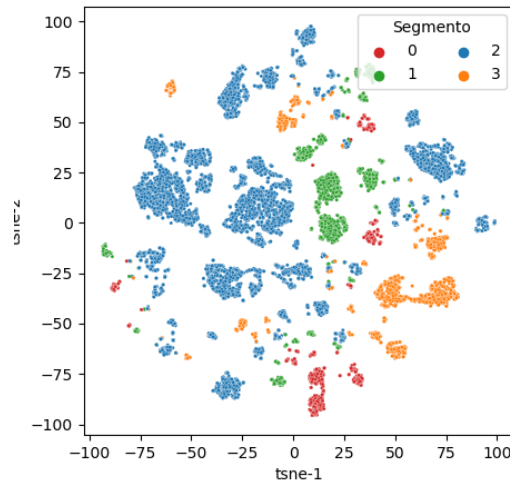


Figura 3.15: Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelo algoritmo *k-means* para todos os clientes da amostra, utilizando, apenas, dados comportamentais.

Variáveis	c_0	c_1	c_2	c_3
devices_qty_median	0.026	0.099	0.115	0.109
phones_number_qty_median	0.457	0.471	0.500	0.500
router_qty_median	0.094	0.109	0.160	0.216
phone_qty_median	0.000	0.002	0.003	0.001
hub_qty_median	0.235	0.309	0.246	0.067
plumes_qty_median	0.002	0.005	0.009	0.014
powerbox_qty_median	0.008	0.022	0.021	0.011
tv_qty_median	0.012	0.081	0.087	0.089
down_traffic_1w_avg_median	0.000	0.015	0.016	0.024
up_traffic_1w_avg_median	0.000	0.003	0.003	0.004
tv_sessions_qty	0.000	0.023	0.027	0.031
weekend_tv_sessions_perc	0.000	0.191	0.207	0.233
weekday_tv_sessions_perc	0.000	0.428	0.477	0.528
holiday_uses_5g_flg	0.000	0.000	0.000	0.000
holiday_uses_24g_flg	0.000	0.000	0.000	0.000
holiday_uses_eth_flg	0.000	0.000	0.000	0.000
weekday_uses_5g_flg	0.000	0.920	0.877	0.835
weekday_uses_24g_flg	0.000	0.976	0.960	0.936
weekday_uses_eth_flg	0.000	0.930	0.935	0.944
sa_same_fiscal_num_qty	0.019	0.003	0.003	0.002
product_revolving_churn_qty	1.000	1.000	1.000	1.000
product_effective_churn_qty	0.500	0.501	0.500	0.500

past_involuntary_churn_flg	0.000	0.000	0.000	0.000
past_partial_voluntary_churn_flg	0.001	0.000	0.000	0.000
past_total_voluntary_churn_flg	0.000	0.000	0.000	0.000
days_until_loyalty_ends	0.795	0.790	0.811	0.844
days_since_last_activation	0.126	0.216	0.258	0.132
tv_type_dsc_IRIS	0.001	0.982	0.000	0.000
tv_type_dsc_NÃO APLICÁVEL	0.655	0.000	1.000	0.000
tv_type_dsc_OUTROS	0.314	0.018	0.000	0.002
tv_type_dsc_UMA	0.028	0.000	0.000	0.998
loyalty_period 2 YEARS	0.926	0.910	0.955	0.954
loyalty_period LESS THAN 2 YEARS	0.020	0.037	0.019	0.005
loyalty_period MORE THAN 2 YEARS	0.012	0.005	0.005	0.006
loyalty_period NO LOYALTY	0.007	0.011	0.005	0.009

Tabela 3.3: Centróides dos *clusters* pelo algoritmo *k-means*, considerando todos os clientes da amostra e utilizando, apenas, variáveis comportamentais.

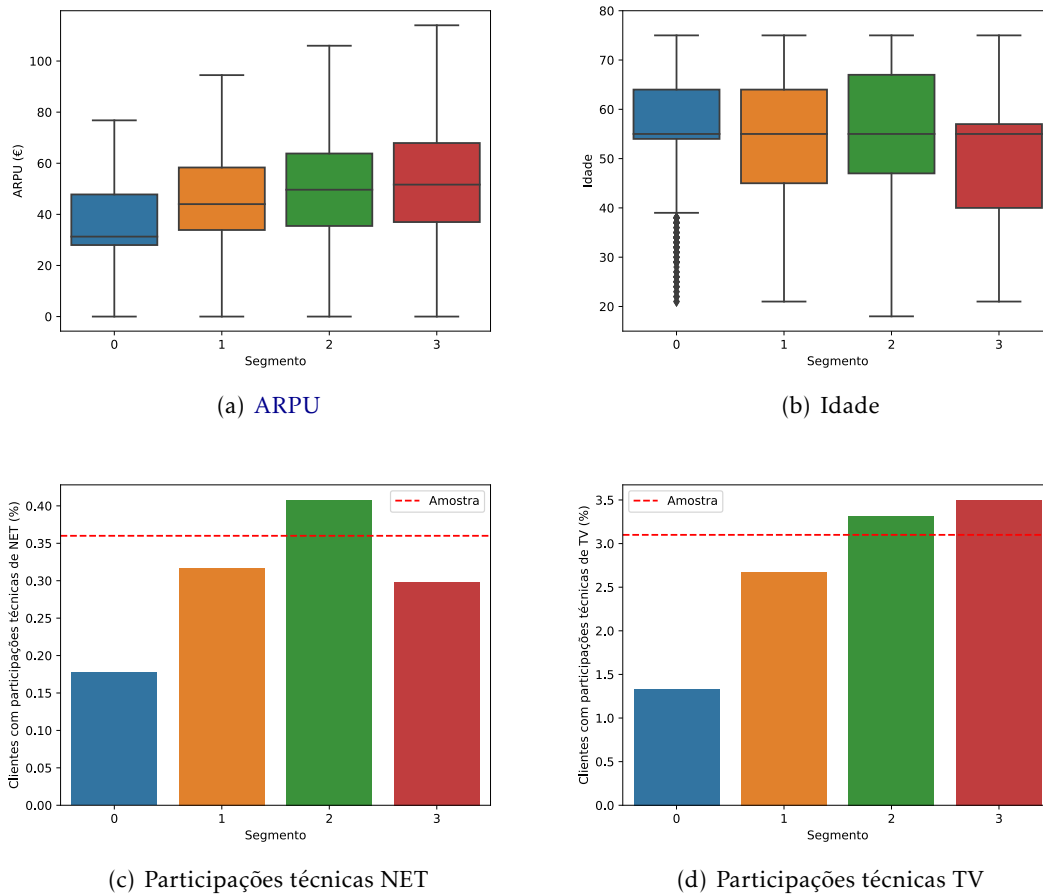


Figura 3.16: Comparação das métricas de validação internas, considerando o algoritmo *k-means* para todos os clientes que evidenciam utilização do router, considerando, apenas, dados comportamentais.

3.7.4 Multi-vista

A primeira tentativa de obter uma visão unificada de cliente consiste na aplicação do algoritmo de *k-means* considerando um conjunto de dados com atributos de mais do que uma vista.

As combinações de vistas a ser exploradas serão a combinação das **VD**, **Vista Psicográfica (VP)** e **Vista Comportamental (VC)**, mas também a combinação de, apenas, as vistas **VP** e **VC**, uma vez que a **VD** é a vista menos preditiva, explica somente 5 % da variância do comportamento do cliente (2.2.4) e é complementar às outras duas vistas, uma vez que adiciona contexto sociocultural a cada cliente, enquanto a **VP** e **VC** tendem a ter um maior consenso entre si.

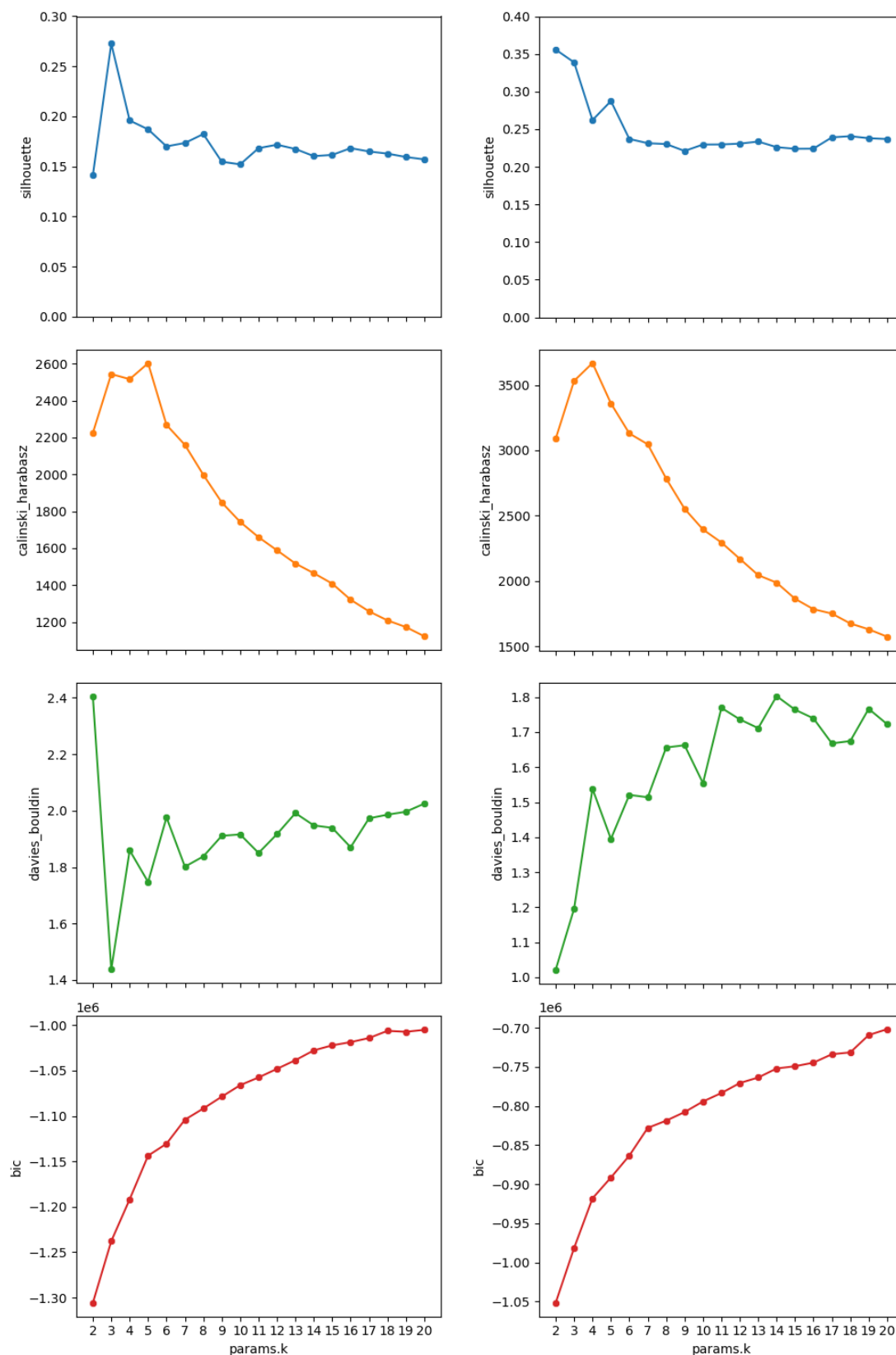
Observando a Figura 3.17 que contém as métricas internas para cada *k* considerado, conclui-se que no caso 3.17(a) o melhor *k* é 3 e no caso 3.17(b) o melhor *k* é 4. Verifica-se, também, que as métricas internas pioram quando se considera mais do que uma vista em simultâneo, o que põe em evidência que as vistas, quando consideradas de forma individual, os clientes agrupam-se em grupos consistentes com a definição de *cluster*, mas que se tendem a ficar cada vez mais equidistantes e perdem a sua estrutura estatística.

Tanto na experiência em que se considerou a combinação **VD+VP+VC** como a experiência com **VP+VC**, isolaram um grupo de clientes a que chamados **Clientes que não utilizam o serviço**, o segmento 0 e o segmento 2 em cada uma das experiências, respetivamente. Estes clientes são o segmento que paga **ARPU** menores e têm uma idade estimada menor, e com uma forte redução do número de participações técnicas tanto de NET como de TV.

O segundo grupo que se isolou tanto na primeira como na segunda experiência, foi o grupo a que chamamos *Cliente Sofisticado*, o segmento 1 e o segmento 3 em cada uma das experiências respetivamente. Estes clientes são aqueles que têm o maior volume de tráfego tanto de upload como de download e apresentam uma utilização esmagadoramente superior de protocolos de internet que evidenciam sofisticação tecnológica, como *remote connectivity*, *network admin*, *online games*, *database*, *tunneling* e *file access*, e também os clientes com a maior quantidade de *routers* e extensores de sinal.

Na primeira experiência (**VD+VP+VC**), os restantes clientes isolaram-se num único grupo. Na segunda experiência (**VP+VC**), destacou-se um terceiro grupo, o segmento 1, clientes sem subscrição de canais premium ou serviço de TV, sem utilização sofisticada de *internet*, que tem um número de participações técnicas ligeiramente superiores aos da amostra como um todo, que constituem, um grupo tendencialmente mais envelhecido.

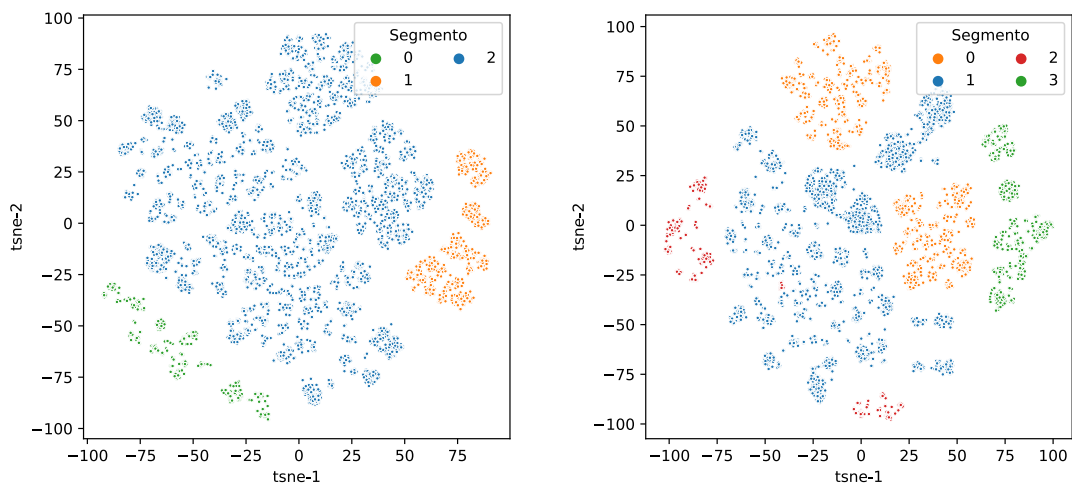
Apesar das duas experiências multi-vista produzirem resultados que permitem uma atuação diferenciada é insuficiente para uma segmentação que contemple os diversos pilares, contemplando toda a riqueza da informação das várias vistas sobre o cliente em que a complementariedade desta informação é explorada, bem como a importância que cada vista tem na visão unificada desejada.



(a) Demográfica Psicográfica Comportamental

(b) Psicográfica Comportamental

Figura 3.17: Comparação da prevalência de clientes com participações técnicas entre a população e a amostra estratificada extraída.



(a) Demográfica + Psicográfica + Comportamental

(b) Psicográfica + Comportamental

Figura 3.18: Comparação da prevalência de clientes com participações técnicas entre a população e a amostra estratificada extraída.

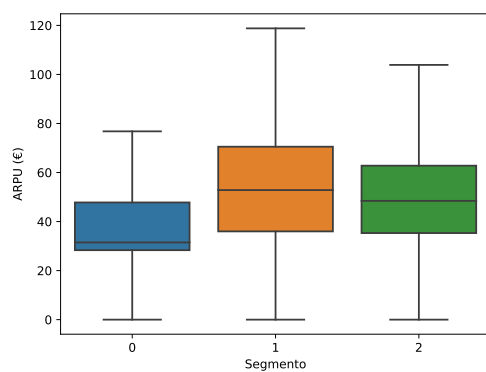
Pilar	Variáveis	VD+VP+VC			VP+VC			
		c_0	c_1	c_2	c_0	c_1	c_2	c_3
VD	age	0.671	0.557	0.657	-	-	-	-
	índice envelhecimento	0.592	0.487	0.507	-	-	-	-
	families_3+ percentage	0.387	0.482	0.458	-	-	-	-
	poder_compra	0.462	0.349	0.370	-	-	-	-
	sem_nivel escolaridade percentage	0.532	0.533	0.535	-	-	-	-
	2_ciclo_percentage	0.398	0.489	0.464	-	-	-	-
VP	digital_invoice_flg	0.745	0.902	0.801	0.844	0.780	0.748	0.902
	direct_payment_flg	0.386	0.451	0.432	0.423	0.437	0.383	0.451
	no_marketing_flg	0.068	0.051	0.071	0.062	0.076	0.069	0.051
	premium_qty	0.002	0.017	0.013	0.040	0.000	0.002	0.017
	tv_cine_qty	0.012	0.148	0.105	0.323	0.000	0.010	0.148
	benfica_tv_qty	0.000	0.005	0.005	0.014	0.000	0.000	0.005
	sport_tv_qty	0.001	0.007	0.007	0.022	0.000	0.001	0.007
	other_premium_qty	0.002	0.006	0.004	0.012	0.000	0.002	0.006
	remote_connectivity_flg	0.000	0.209	0.000	0.000	0.000	0.000	0.209
	network_admin_flg	0.000	1.000	0.000	0.000	0.000	0.000	1.000

	online_games_flg	0.000	0.950	0.116	0.133	0.108	0.000	0.950
	database_flg	0.000	0.758	0.000	0.000	0.000	0.000	0.758
	tunneling_flg	0.000	0.998	0.000	0.000	0.000	0.000	0.998
	file_access_flg	0.000	0.959	0.021	0.026	0.019	0.000	0.959
	voip_flg	0.000	1.000	0.799	0.824	0.784	0.000	1.000
VC	devices_qty	0.033	0.113	0.111	0.104	0.114	0.031	0.113
	phones_number qty	0.458	0.498	0.495	0.484	0.500	0.458	0.498
	router_qty	0.108	0.192	0.156	0.160	0.154	0.103	0.192
	phone_qty	0.000	0.005	0.002	0.002	0.003	0.000	0.005
	hub_qty	0.223	0.089	0.160	0.130	0.173	0.227	0.089
	plumes_qty	0.002	0.014	0.008	0.009	0.008	0.002	0.014
	powerbox_qty	0.008	0.019	0.019	0.016	0.021	0.008	0.019
	tv_qty	0.016	0.09	0.086	0.085	0.086	0.015	0.090
	down_traffic 1w_avg	0.000	0.046	0.013	0.015	0.012	0.000	0.046
	up_traffic 1w_avg	0.000	0.010	0.002	0.002	0.002	0.000	0.010
	tv_sessions_qty	0.000	0.030	0.027	0.027	0.027	0.000	0.030
	weekend_tv sessions_perc	0.004	0.235	0.206	0.209	0.205	0.001	0.235
	weekday_tv sessions_perc	0.006	0.521	0.474	0.472	0.474	0.001	0.521
	holiday_uses 5g_flg	0.055	1.000	1.000	1.000	1.000	0.038	1.000
	holiday_uses 24g_flg	0.055	1.000	1.000	1.000	1.000	0.038	1.000
	holiday_uses eth_flg	0.055	1.000	1.000	1.000	1.000	0.038	1.000
	weekday_uses 5g_flg	0.000	0.892	0.879	0.872	0.880	0.000	0.892
	weekday_uses 24g_flg	0.000	0.950	0.965	0.954	0.967	0.000	0.950
	weekday_uses eth_flg	0.000	0.965	0.937	0.933	0.936	0.000	0.965
	sa_same_fiscal num_qty	0.018	0.003	0.003	0.003	0.003	0.018	0.003
	product_revolving churn_qty	1.000	1.000	1.000	1.000	1.000	1.000	1.000

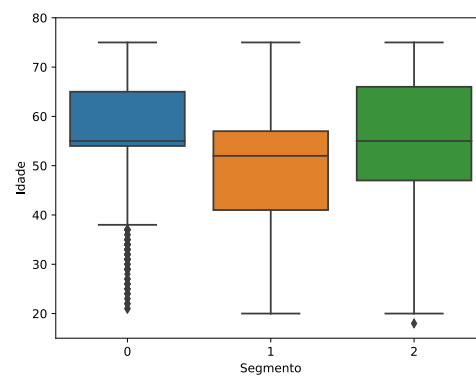
3.7. RESULTADOS EXPERIMENTAIS UTILIZANDO K-MEANS

product_effective churn_qty	0.500	0.500	0.500	0.500	0.500	0.500	0.500
past_involuntary churn_flg	0.000	0.000	0.000	0.000	0.000	0.000	0.000
past_partial_voluntary churn_flg	0.001	0.001	0.000	0.000	0.000	0.001	0.001
past_total_voluntary churn_flg	0.000	0.000	0.000	0.000	0.000	0.000	0.000
days_until loyalty_ends	0.795	0.828	0.811	0.816	0.809	0.795	0.828
days_since last_activation	0.129	0.222	0.231	0.174	0.259	0.129	0.222
tv_type_dsc IRIS	0.006	0.126	0.161	0.495	0.000	0.001	0.126
tv_type_dsc NÃO APLICÁVEL	0.663	0.642	0.675	0.000	1.000	0.668	0.642
tv_type_dsc OUTROS	0.297	0.000	0.004	0.011	0.000	0.302	0.000
tv_type_dsc TDC	0.002	0.000	0.000	0.000	0.000	0.002	0.000
tv_type_dsc UMA	0.033	0.232	0.160	0.494	0.000	0.027	0.232
loyalty_period 2_YEARS	0.926	0.959	0.946	0.930	0.954	0.926	0.959
loyalty_period LESS_THAN 2_YEARS	0.020	0.012	0.021	0.022	0.020	0.021	0.012
loyalty_period MORE_THAN 2_YEARS	0.013	0.001	0.005	0.006	0.005	0.012	0.001
loyalty_period NO_LOYALTY	0.007	0.007	0.006	0.009	0.005	0.007	0.007

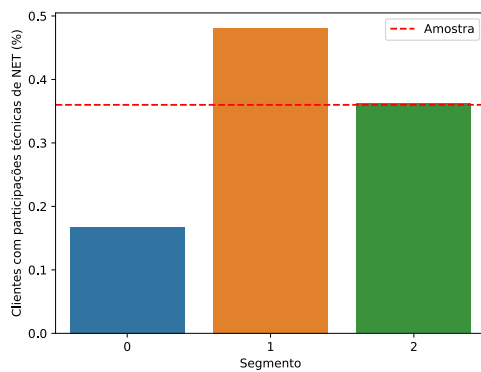
Tabela 3.4: Centróides dos *clusters*, utilizando dados demográficos, psicográficos e comportamentais.



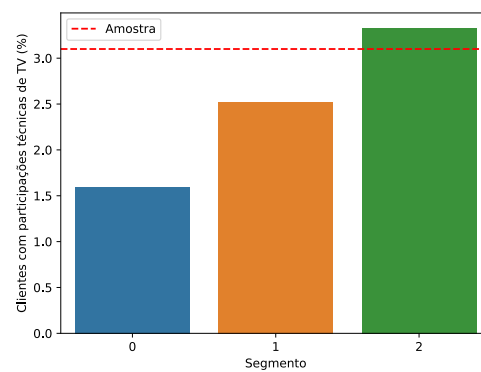
(a) ARPU



(b) Idade



(c) Participações técnicas NET



(d) Participações técnicas TV

Figura 3.19: Comparação da prevalência de clientes com participações técnicas entre a população e a amostra estratificada extraída.

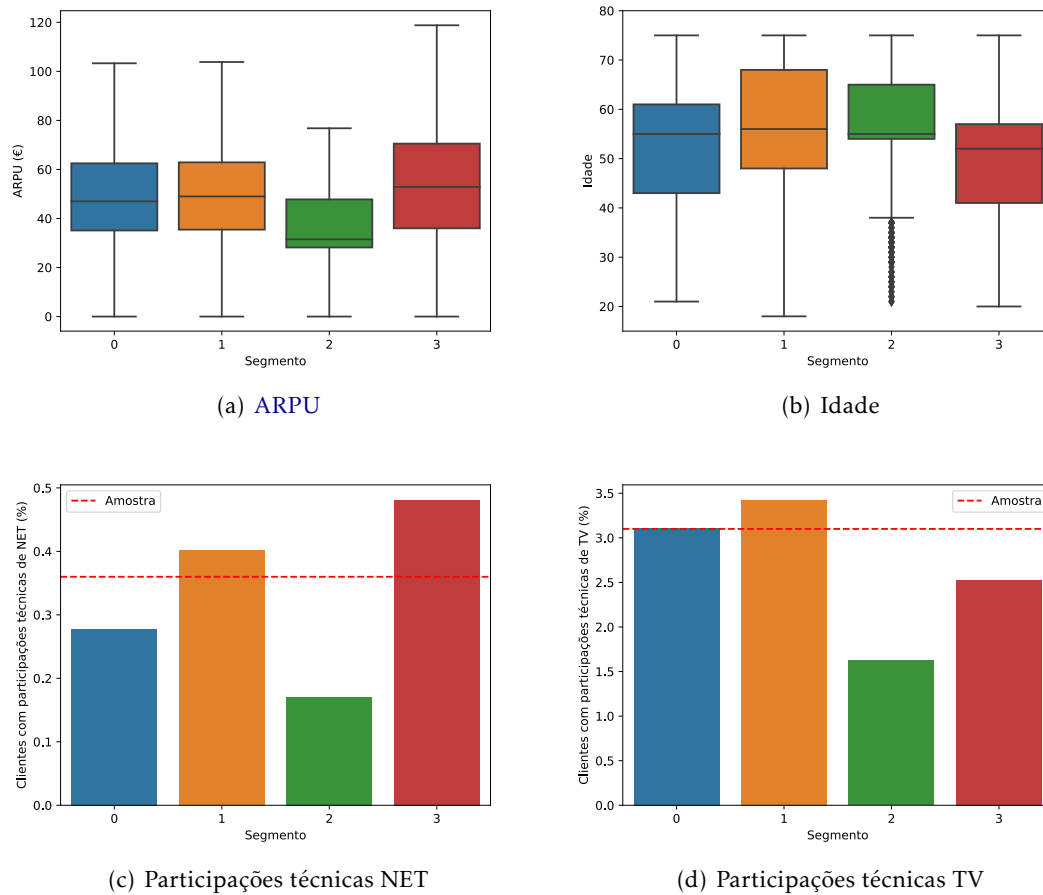


Figura 3.20: Comparação da distribuição de **ARPU** e idade, bem como da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando, apenas, dados psicográficos e comportamentais.

ALGORITMOS DE APRENDIZAGEM MULTI-VISTA

Neste capítulo são apresentados os algoritmos de aprendizagem multi-vista. Na primeira secção é apresentada uma motivação para a aprendizagem multi-vista, seguida dos princípios necessários para a aprendizagem multi-vista. Na terceira secção é apresentada a taxonomia dos métodos de aprendizagem multi vista, com especial destaque para *subspace clustering* e as abordagens escaláveis baseadas em âncoras.

4.1 Motivação para modelos *multi-view*

Na era de grandes volumes de dados, do inglês *big data*, os dados podem ser gerados de diferentes fontes, domínios, diferentes objetos que recolhem novos atributos ou observados de diferentes vistas, sendo, naturalmente, heterogêneos, mas têm uma conexão potencial, o que faz com que este tipo de dados seja considerado multi vista [40]. Alguns exemplos da proliferação deste tipo de dados encontra-se em aplicações de: caracterização de documentos, que pode ser realizada de várias óticas, tais como imagens, vídeos e áudio; a caracterização de atividades humanas, que podem ser caracterizadas por câmaras de vídeo, de profundidade e sensores corporais; imagens em *websites*, que têm uma vista dedicada às etiquetas textuais presentes na imagem e outra vista dedicada às suas descrições [40]; uma mesma notícia pode ser reportada por diferentes agências, em que cada uma destas agência contempla uma nova vista sobre os mesmos acontecimentos [40]; os sinais de sensores podem ser decompostos em duas vistas diferentes, uma vista relativa ao tempo e outra em relação à frequência [40]; o mesmo sentido semântico pode ser representado em diferentes linguagens, sendo cada uma delas uma vista sobre os dados; as imagens também podem ser descritas por várias vistas, como por exemplo, uma vista relativamente a atributos de cor e outra vista relativa a atributos de textura; na identificação de pessoas, uma vez que uma pessoa pode ser identificada utilizando métodos de reconhecimento diferentes, tais como reconhecimento facial, impressão digital, assinatura ou pela íris, sendo esses dados recolhidos por instrumentos e sensores distintos, constituindo, naturalmente, vistas distintas sobre o mesmo objeto.

Assim, a extração de conhecimento deste tipo de dados é bastante importante em análise de dados. Contudo, o desafio prende-se com a incorporação de diferentes vistas, para *clusterizar* grandes volumes de dados. Um problema comum em análise de dados é a integração destas mesmas perspetivas e descoberta de estruturas subjacentes, uma vez que cada vista, separadamente, pode ser insuficiente ou incompleta porque descreve, apenas, uma perspetiva dos dados.

Os algoritmos de clusterização habituais, tais como *k-means*, *spectral clustering* ou *subspace clustering*, consideram clusterização numa única vista, o que significa que a vista com os atributos mais importante e a vista com os atributos menos importantes são tratadas de igual forma quando consideramos que o nosso espaço de atributos é dado pela união dos atributos de cada vista, produzindo resultados insatisfatórios. Esta união pode, também, fazer com que incorramos num problema chamado *curse of dimensionality*, que ocorre quando as métricas de distâncias perdem o seu significado à medida que o número de dimensões do nosso espaço de atributos aumenta. Este problema é causado pelo facto de cada dimensão adicional dispersar cada vez mais os nossos pontos, chegando a um ponto em que estes estão equidistantes uns dos outros, invalidando a definição de *cluster* e impedindo o propósito de obter uma boa segmentação de clientes. Além disso, concatenar todas as vistas numa única pode não aumentar a performance dos algoritmos, porque cada vista tem a sua própria estrutura estatística, que pode ser perdida com esta fusão [40]. Para este objetivo, é desejável produzir uma clusterização para cada vista e unificar os resultados consoante a sua importância, resolvendo dois problemas desafiantes em simultâneo: combinar os vários resultados de clusterização e aprender a importância das várias vistas para a clusterização final.

Deste modo, **Multi-view Clustering (MvC)** atraiu atenção nos últimos anos ao considerar a complementaridade e diversidade da informação ao longo das vistas e alguns modelos, mais avançados, consideram, ainda, a diversidade de diferentes vistas aquando da sua fusão [40].

O objectivo dos algoritmos de multi-view clustering é como maximizar a qualidade da clusterização em cada vista, mantendo a consistência da clusterização ao longo das diversas vistas [40].

Os autores de [39] propõem a seguinte taxonomia para algoritmos de *multi-view clustering*: *co-training*, *multi-kernel learning* e *subspace learning*, às quais os autores de [40] adicionam *multi-task multi-view subspace clustering* e estas subdividem-se em aprendizagem supervisionada e não supervisionada, mas o foco deste trabalho, é a aprendizagem não supervisionada, dada a natureza da aplicação de segmentação de cliente, para o qual não são conhecidos os verdadeiros segmentos aos quais um cliente pertence, nem estes são consensuais.

4.2 Princípios para a aprendizagem multi-vista

4.2.1 Princípios de complementaridade e de consenso

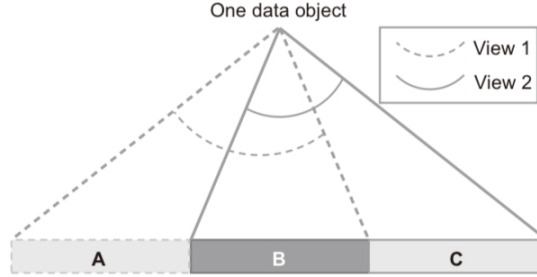


Figura 4.1: Ilustração dos princípios de consenso e de complementaridade [40].

Princípio de consenso procura maximizar a concordância entre as várias vistas [39]. Recorrendo ao resultado obtido por [8] que demonstra que

$$P(f^1 \neq f^2) \geq \max\{P_{err}(f^1), P_{err}(f^2)\}$$

Esta desigualdade permite nos concluir que a probabilidade de discórdia entre duas hipóteses independentes majora o erro de cada uma das hipóteses. Assim, a minimização da discordância entre as duas hipóteses, o erro de cada uma também é minimizado.

Princípio de complementaridade várias vistas devem ser utilizadas para descrever os objetos de forma a descrever as observações com mais detalhe e precisão [40] e indica que algumas vistas possam conter informação que outras não têm [39].

4.2.2 Construção das vistas

Segundo os autores de [39] a prioridade de *multi-view learning* é a aquisição de vistas redundantes, o que é uma diferença para a aprendizagem *single-view*.

Objetos podem ser frequentemente descritos de diferentes perspetivas. Em muitos casos, esta multiplicidade de vistas não existe, porque apenas uma vista foi apresentada para representar os dados. Uma vez que é difícil de aplicar um algoritmo de *multi-view* em com dados com uma única vista, um dos trabalhos preliminares é a construção de múltiplas vistas, particionando o conjunto de atributos, criando múltiplos subconjuntos disjuntos. Esta não é uma tarefa trivial, uma vez que esta partição está dependente tanto do algoritmo como do domínio dos dados.

4.2.3 Combinação das vistas

Combinação *a posteriori* A combinação *a posteriori* das vistas consiste na consideração das vistas de forma independente, enquanto se treina um modelo para cada vista. No

fim, os vários modelos passam por uma etapa de colaboração em que se estes modelos são forçados a serem consistentes. Este tipo de combinação é característico de algoritmos de multi-view que utilizam um mecanismo de co-training, esta fusão tardia pode ser observada na Figura 4.2.

Combinação intermédia A combinação intermédia de múltiplas vistas é feita considerando as vistas de forma independente, mas a sua combinação é feita imediatamente antes ou durante o treino do modelo *multi-view*. Os algoritmos de multiple kernel learning são exemplo deste tipo de combinação de vistas, uma vez que calculam, primeiro, um *kernel*, em separado para cada vista, e são combinados quer de forma linear ou não linear, de modo a melhorar a performance da clusterização [40], esta fusão intermédia pode ser observada na Figura 4.3.

Combinação a priori A combinação a priori considera desde o início as várias vistas de forma conjunta para obter a representação unificada. Assume-se que as vistas de input são geradas por um espaço latente, como pode ser observado na Figura 4.4.

4.3 Grupos de algoritmos de aprendizagem multi-vista

4.3.1 Co-training style

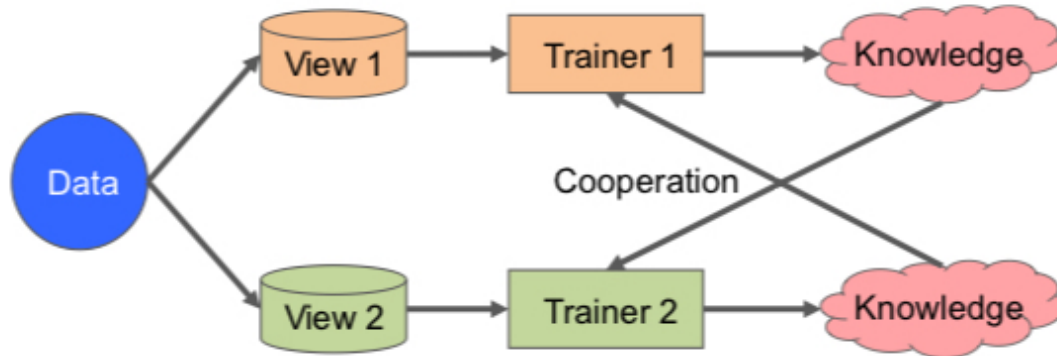


Figura 4.2: Procedimento de algoritmos de co-training [40].

Os algoritmos baseados em *co-training* treinam alternadamente de modo a maximizar a concordância entre duas vistas [39]. Este tipo de abordagens, segundo [40] consistem de duas fases: uma local e uma fase de colaboração. Na fase local cada vista é treinada com um modelo de clusterização e é produzido conhecimento, como é ilustrado na Figura 4.2, na fase de colaboração é utilizado o conhecimento e resultados de clusterização.

O sucesso desta estratégia depende de três suposições:

1. **Suficiência** - Cada vista é suficiente para a tarefa de aprendizagem, por si só.

2. **Compatibilidade** - As funções objetivo devem produzir a mesma predição para atributos que co- ocorram em duas vistas, com elevada probabilidade
3. **Independência Condicional** - As vistas produzem *labels* com independência condicional. Esta última assunção é difícil de satisfazer na prática e foram propostas alternativas mais fracas, como a assunção de uma dependência condicional fraca [1].

Foram propostas várias extensões de co-training, tais como o co-EM, co-regularization e co-clustering.

4.3.2 Co-EM

Este tipo de algoritmos procuram utilizar um classificador h_1 que adiciona exemplos a um conjunto de dados para o qual existe uma variável alvo, que um classificador h_2 tenta aprender. A ideia por detrás de co-EM é correr EM em cada vista individualmente, atribuindo probabilidades de cada label a cada exemplo não supervisionado. Nesta versão de co-training é essencial a assunção de independência condicional [39].

4.3.3 Co-regularization

Assumindo um caso de um conjunto de dados X com duas vistas, $X = \{X^1, X^2\}$ e dois espaços de hipóteses H^1 e H^2 , que aproximam a função objetivo e são treinadas de forma alternada para maximizar a concordância entre exemplos para os quais não existe um label. Esta maximização da concordância não é ideal para uma aplicação de segmentação de cliente, uma vez que as vistas naturalmente heterogêneas e não concordantes, logo terão, de certo, consenso 4.2.1, mas a sua complementaridade também terá de ser explorada para tirar partido de uma abordagem *multi-view* [39].

4.3.4 Co-clustering

Esta abordagem foi originalmente desenhada para ambientes semi-supervisionados. Trabalha sobre a assunção de que a verdadeira clusterização vai colocar as observações correspondentes em cada vista no mesmo *cluster*.

Considerando o estudo feito por [2] que explora versões multi-view de abordagens de abordagens single-view como k-means, k-meloids e EM, como apresentado em [39]. No caso em que se utiliza k-means, para cada iteração corre-se o algoritmo de k-means em cada vista, em seguida partilham a informação sobre a partição obtida noutra vista e volta-se a correr o algoritmo de *k-means* na segunda vista. Após o algoritmo terminar é calculada a média do consenso entre cada *cluster*. No final cada observação é atribuída a um *cluster* distinto de acordo com o vetor fechado de conceitos. Uma das características da utilização da estratégia das vistas utilizarem conhecimento umas das outras é a promoção de um consenso geral entre as diversas vistas [40].

4.3.5 Multi-kernel learning

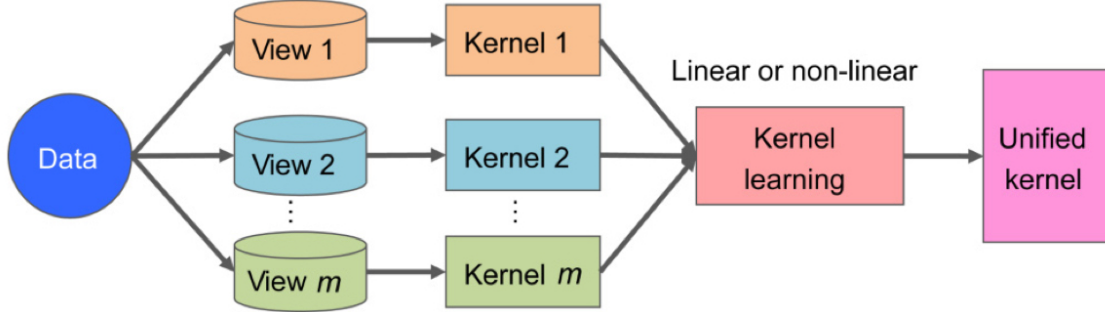


Figura 4.3: Procedimento de algoritmos de multi-kernel learning [40].

Multi-kernel learning foi, originalmente, desenvolvido para melhorar o espaço de procura de possíveis funções de *kernel*, tanto *kernels* lineares, como polinomiais, mas também gaussianos. [40]

Desta forma, *multi-kernel learning* tem sido aplicado para lidar com dados *multi-view*, uma vez que estes *kernels* são combinados de forma linear ou não linear para obter um *kernel* unificado, como ilustrado na Figura 4.3. Os *kernels* utilizados em *multi-kernel learning* correspondem, naturalmente, a vistas, contudo encontrar essas funções de *kernel*, bem como a sua combinação continua a ser um desafio [40]

4.3.6 Multi-view subspace clustering

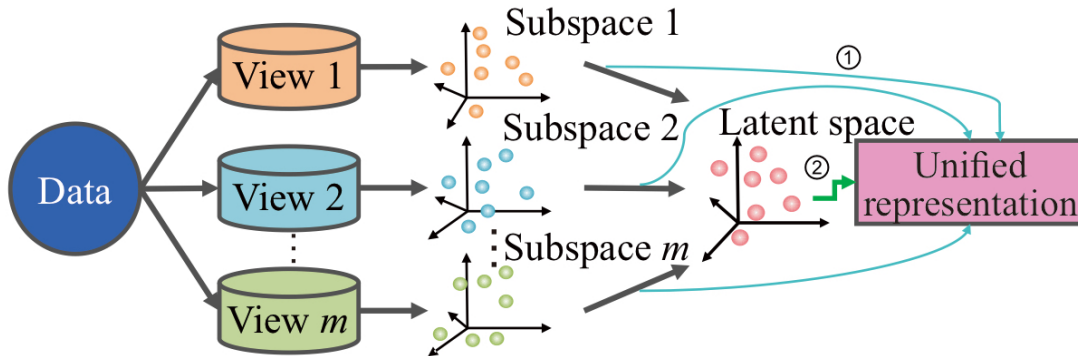


Figura 4.4: Ilustração do procedimento dos algoritmos de *subspace clustering* [40].

Multi-view subspace clustering é considerado um tópico popular na área de *MvC* que tem como ideia base a aprendizagem de uma representação unificada dos dados provenientes de várias vistas [40], procurando obter um espaço latente, sobre a assunção de que as vistas são geradas a partir deste mesmo espaço [39]. Esta representação unificada pode ser obtida de uma de duas formas, ambas ilustradas na Figura 4.1, sendo a primeira, (1), conseguida diretamente a partir dos subespaços, a (2), passa por primeiro aprender um espaço latente e depois chegar à representação unificada. Esta representação unificada

obtida pode ser, posteriormente, utilizada por um algoritmo de clusterização, tal como o algoritmo *k-means* para obter a clusterização desejada.

Subspace clustering tem como objetivo expressar cada observação como uma combinação linear das outras observações, procurando minimizar o erro de reconstrução para obter os coeficientes de combinação, que representa a similaridade entre dois pontos. O problema de otimização resolvido por subspace clustering, para uma matriz de dados $X \in \mathbb{R}^{d \times n}$ é o seguinte:

$$\min_S \|X - XS\|_F^2 + \alpha f(S) \text{ s.t. } S \geq 0, S\mathbf{1} = \mathbf{1}$$

$f(\cdot)$ é uma função regularizadora, $\alpha > 0$, é um meta-parâmetro.

A ideia principal de Guo [40] é a detecção de um subespaço partilhado entre várias vistas e, em seguida, procede-se à aplicação de um algoritmo standard de clusterização como o *k-means*.

Multi-view subspace clustering procura um subespaço unificado a partir da fusão de dados com várias vistas [34]. A clusterização é feita em duas fases: a primeira constrói o grafo, uma estrutura de baixa dimensionalidade dos dados, a segunda fase consiste na aplicação de um algoritmo espectral ao grafo construído na primeira fase.

Apesar de algoritmos de MVSC atingirem grande sucesso na melhoria da performance de clusterização, existem desafios computacionais quando se trabalha com clusterização de dados de larga escala, dado que sofrem de uma complexidade cúbica em relação ao tamanho da amostra [34]. A primeira fase destes algoritmos, a construção do grafo, tem uma complexidade temporal de pelo menos $O(n^3)$. A segunda fase, a aplicação de um algoritmo de clusterização espectral necessita de $O(n^3)$ para a decomposição dos valores singulares.

Esta elevada complexidade dos algoritmos levou à proposta de algoritmos de [Multi-View Subspace Clustering \(MVSC\)](#) escaláveis, a fim de reduzir o seu custo computacional, como é o caso das abordagens baseadas em âncoras.

4.3.6.1 Abordagens baseadas em âncoras

Apesar de toda a visibilidade que *multi-view subspace clustering* recebeu, este tipo de abordagens consideram uma abordagem clássica, uma vez que utilizam todas as observações para a estratégia de auto representação. Apesar da relação global ser bem explorada, o tempo e memória necessários para a construção de um grafo restringe a escalabilidade da aplicação destes algoritmos, dado a sua complexidade temporal cúbica [34]. Esta dificuldade levou alguns investigadores a considerarem abordagens em que são amostrados alguns pontos para servir como âncoras.

Representar cada observação utilizando todas as observações é desnecessário e redundante [34]. Deste modo, motivada-se a consideração de um pequeno subconjunto de âncoras para reconstruir o subespaço e capturar a estrutura o que permite a substituição dos grafos originais com uma dimensão $n \times n$ por grafos cuja dimensão é de apenas, $n \times k$

[34]. Estas âncoras podem ser obtidas por amostragem aleatória, distribuição uniforme ou pela utilização dos centróides obtidos pela aplicação de um *k-means*.

A ideia base por de trás desta abordagem é a escolha, independente, de m pontos de referência, igualmente fundidos no grafo produzido no primeiro passo deste tipo de algoritmos, e de seguida feita aplicado a habitual fase de clusterização espectral de modo a obter o resultado final.

Com a estratégia da utilização de âncoras, a complexidade computacional é reduzida para $O(n)$ por iteração, na fase de construção do grafo, um dos pontos que restringia a aplicação destes algoritmos a conjuntos de dados de larga escala [34].

Método	Complexidade Espacial	Complexidade Temporal	Número Máximo de instâncias
LMVSC [17]	$vm(n+h)$	$O(n)$	30,000
SMVSC [34]	$mn+(h+m)c$	$O(n)$	101,499

Tabela 4.1: Complexidade dos métodos comparados, em que $h = \sum_{i=1}^v d_i$ e c é a dimensão comum

4.3.6.2 Large-Scale Multi-View Subspace Clustering in Linear Time

O método **LMVSC** é proposto no artigo [17] e é a primeira tentativa de aplicar multi-view subspace clustering a dados de larga escala, pela sua complexidade computacional linear, em contraste com o tempo quadrático ou cúbico que métodos existentes à data necessitavam. Este método começa por tirar partido da redundância dos dados, utilizando apenas um subconjunto de observações para reconstruir os subespaços subjacentes, utilizando uma abordagem baseada em âncoras 4.3.6.1, aproxima a matriz de similaridades $S^i \in \mathbf{R}^{n \times n}$, com dimensão $n \times n$ com uma matriz $Z^i \in \mathbf{R}^{n \times m}$, em que m é o número de âncoras.

Estas m âncoras escolhidas são os centróides calculados pela aplicação do algoritmo *k-means* aos dados de cada vista X^i , constituindo a matriz $A^i = \{a_j\}_{j \in [1,m]}$, em que Z^i é aprendido através dos dados, utilizando a matriz de âncoras A^i e a matriz com os dados da vista i , X^i , resolvendo o problema, por programação quadrática convexa:

$$\begin{aligned} \min_{Z_i} \sum_{i=1}^v \|X_i - A_i Z_i^T\|_F^2 + \alpha \|Z_i\|_F^2 \\ \text{s.t. } S_0 \leq Z_i, Z_i^T \mathbf{1} = \mathbf{1} \end{aligned} \quad (4.1)$$

A utilização da estratégia de âncoras permite considerar apenas as similaridades entre mn pontos, em vez de n^2 . No final deste passo são produzidas v grafos $Z^i_{i \in [1,v]}$. Em seguida, é necessário fazer um spectral clustering utilizando as matrizes Z^i , o que já não é possível, uma vez que não são de dimensão $n \times n$.

Para resolver este problema são possíveis duas abordagens: a primeira, envolve a aplicação da equação

$$S = \hat{Z}\hat{Z}^T \quad (4.2)$$

a fim de restaurar i matrizes S^i com dimensão $n \times n$, aplicando a média das várias matrizes

$$\bar{S} = \frac{\sum_i S_i}{v} \quad (4.3)$$

Em que a solução é os k vetores próprios que correspondem aos maiores k valores próprios. Infelizmente, esta abordagem demora, pelo menos, $O(n^2k)$, o que impede a ambição de obter um algoritmo em tempo linear. Assim, a segunda abordagem utiliza a preposição 1 de (Chen e Cai 2011), em que se propõe uma alternativa para calcular os k vetores singulares esquerdos de $\bar{Z} \in \mathbb{R}^{n \times mv}$,

$$\bar{Z} = \frac{1}{\sqrt{v}} [\hat{Z}^1, \dots, \hat{Z}^i, \dots, \hat{Z}^v] \quad (4.4)$$

e uma vez que $mn \ll n$, fazendo com que o uso de $\bar{Z}$ seja computacionalmente mais eficaz do que utilizar $\bar{S}$. É provado que os vetores singulares esquerdos de $\bar{Z}$ são os mesmos que os vetores próprios de $\bar{S}$.

O algoritmo do método de [LMVSC](#) é sistematizado na tabela [4.3.6.2](#). Por questão de uniformização do input com o método [SMVSC](#), foi considerado que as matrizes $\{A_i\}_{i \in [1, v]}$ eram produzidas como primeiro passo do método, ao contrário do que foi proposto no artigo onde foi proposto [\[17\]](#), em que estas matrizes também eram parte do input deste método.

Input Matrizes de dados $X = \{X^1, \dots, X^v\}$, número de <i>clusters</i> k , e o hiperparâmetro α
Passo 1: Obtenção das matrizes $A^i_{i \in [1, v]} = \{a_j\}_{j \in [1, m]}$ por aplicação do método <i>k-means</i> a cada vista X^i
Passo 2: Resolver o problema de otimização (4.1)
Passo 3: Construir a matriz $\tilde{Z}$ como na equação 4.4
Passo 4: Calcular Q , utilizando <i>Decomposição em valores singulares, do inglês, Singular Value Decomposition (SVD)</i> na matriz $\tilde{Z}$
Resultado Final: Correr o algoritmo <i>k-means</i> em Q para obter a clusterização final

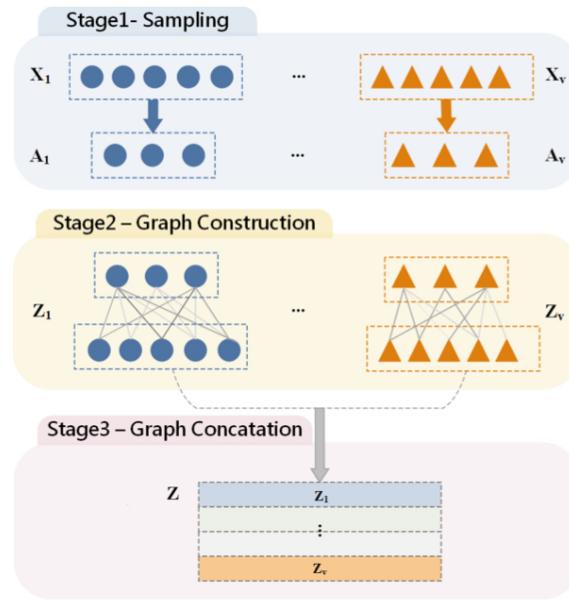


Figura 4.5: Esquematização do funcionamento do método LMVSC [34].

Embora este método venha permitir a aplicação de *multi-view subspace clustering* em tempo linear sofre de problemas, uma vez que, em primeiro lugar, a escolha das âncoras que está isolada do processo de clusterização, em segundo lugar, é feita de forma isolada para cada vista, não permitindo a captura de informação complementar entre vista nem considera a sua importância para a clusterização e em terceiro lugar, as âncoras não participam no processo de otimização, mantendo-se estáticas uma vez inicializadas. Esta separação do processo de amostragem das âncoras escolhidas e o processo de clusterização leva à escolha de âncoras com fraco poder discriminatório.

4.3.6.3 Scalable Multi-view Subspace Clustering with Unified Anchors

Este algoritmo procura colmatar as dificuldades de uma abordagem tradicional apresentadas em [34], em que o processo de escolha das âncoras está isolada do processo de clusterização, e é feita de forma isolada para cada vista, não permitindo a captura de informação complementar entre vistas nem a sua importância para a clusterização com várias vistas e as âncoras não participam na otimização. Deste modo, o método proposto em [34], procura-se combinar a aprendizagem das âncoras e a construção do grafo num único framework, a fim de obter um subespaço mais discriminante, mantendo uma complexidade linear, permitindo a sua aplicação a dados em larga escala.

$$\min_{\alpha, W_i, A, Z} \sum_{i=1}^v \alpha_i^2 \|X_i - W_i A Z\|_F^2 + \|Z\|_F^2$$

Em que $X_i \in \mathbf{R}^{d_i \times n}$

A otimização da matriz W_i é feita mantendo A , Z e α_i fixos, a função objetivo em relação a W_i pode ser formulada como:

$$\min_{W_i} \sum_{i=1}^v \alpha_i^2 \|X_i A Z\|_F^2, \text{ s.t. } W_i^T W_i = I_d \quad (4.5)$$

Uma vez que cada W_i está separada das outras vistas, a equação 4.5 pode ser transformada no problema 4.6, pela expansão da norma de Frobenius pelo traço e removendo todos os fatores que não estão relacionados com W_i

$$\max_{W_i} \text{Tr}(W_i^T B_i), \text{ s.t. } W_i^T W_i = I_d \quad (4.6)$$

Em que $B_i = X_i Z^T A^T$, supondo que o resultado de aplicar SVD a B_i é $U_b \Sigma_b V_b^T$ e o W_i ótimo é $U_b V_b^T$

A atualização da matriz A é feita mantendo W_i , Z e α_i constantes,

$$\min_A \sum_{i=1}^v \alpha_i^2 \|X_i A Z\|_F^2, \text{ s.t. } A^T A = I_m \quad (4.7)$$

De forma análoga a W_i , o problema de otimização de A pode ser reescrito da seguinte forma:

$$\max_A \text{Tr}(A^T C), \text{ s.t. } A^T A = I_m \quad (4.8)$$

Onde $C = \sum_{i=1}^v \alpha_i^2 W_i^T X_i Z^T$. Assim, a solução ótima para $A = U_c V_c^T$ em que $C = U_c \Sigma_c V_c^T$

A atualização da matriz Z é feita, mantendo W_i , A e α_i constantes e rescrevendo o problema de otimização par atualizar a variável Z :

$$\min_Z \sum_{i=1}^v \alpha_i^2 \|X_i A Z\|_F^2 + \|Z\|_F^2, \text{ s.t. } Z \geq 0, Z^T \mathbf{1} = \mathbf{1} \quad (4.9)$$

O problema de otimização em função de Z pode ser reformulado como um problema de programação quadrática :

$$\max \frac{1}{2} Z_{:,j}^T H Z_{:,j} + f^T Z_{:,j}, \text{ s.t. } Z_{:,j}^T \mathbf{1} = 1, Z \geq 0 \quad (4.10)$$

Onde $H = 2(\sum_{i=1}^v \alpha_i^2 + 1)\mathbf{I}$, $f^T = -2\sum_{i=1}^m X_{i[:,j]}^T W_i A$ A otimização pode ser feita resolvendo o problema de programação quadrática para cada coluna de Z .

A atualização de α_i envolve de fixar as variáveis irrelevantes obtemos o seguinte problema de otimização:

$$\min_{\alpha_i} \sum_{i=1}^v \alpha_i^2 M_i^2 \text{ s.t. } \alpha^T \mathbf{1} = 1 \quad (4.11)$$

Input Matrizes de dados $X = \{X^1, \dots, X^v\}$, número de <i>clusters</i> k
Passo 1: Inicialização das matrizes A , Z e W como uma matriz de zeros. Inicialização de $\alpha^i = \frac{1}{v}$ while não converge: Passo 1: Atualizar W_i, resolvendo o problema de otimização 4.6 Passo 2: Atualizar A, resolvendo o problema de otimização 4.8 Passo 3: Atualizar Z, resolvendo o problema de otimização 4.10 Passo 4: Atualizar α^i, resolvendo o problema de otimização 4.11 end while Obter os valores singulares esquerdos de U aplicando SVD em Z
Resultado Final: Correr o algoritmo <i>k-means</i> em U para obter a clusterização final height

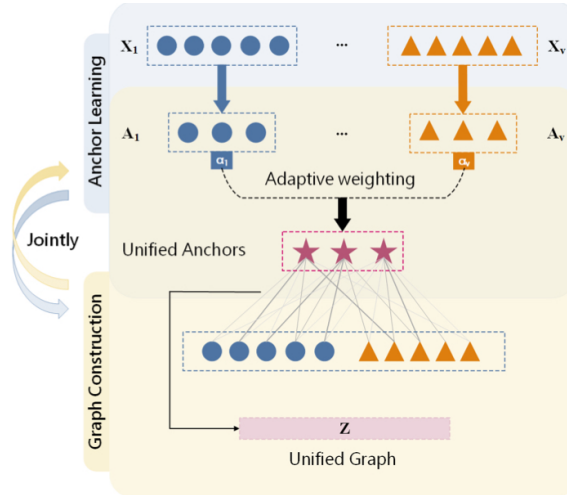


Figura 4.6: Esquemática do funcionamento do método SMVSC [34].

4.3.7 Aplicabilidade de Multi-view subspace clustering em tempo linear a conjuntos de dados com uma única vista

Remark. Os autores de [17] demonstram a aplicabilidade de LMVSC a dados com uma única vista.

Demonstração. Utilizando a equação 4.4, demonstra-se que $v = 1$, $\bar{Z} = \hat{Z}$ □

4.3.8 Multi-task multi-view clustering

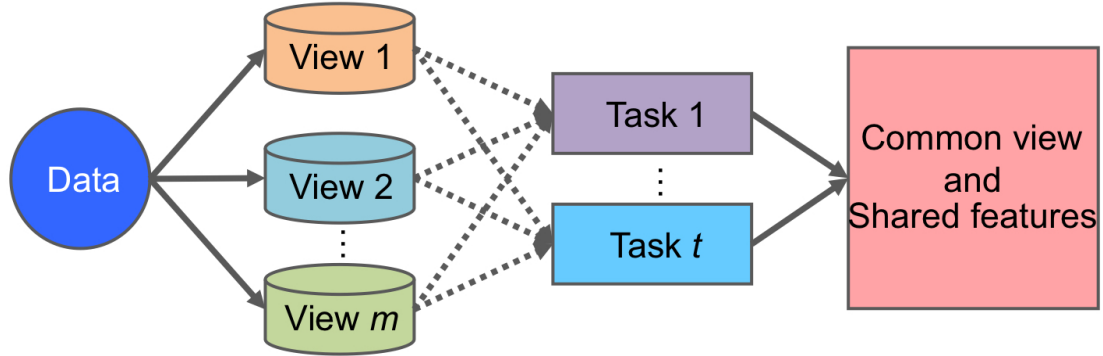


Figura 4.7: Ilustração do procedimento dos algoritmos de multi-task multi-view clustering [40].

O conceito de *multi-task learning* [40] consiste na utilização de diversas tarefas e utilizar a relação entre as mesmas para melhorar a performance de algoritmos com uma única vista. O conceito de multi-view clustering aliado ao de multi-task learning dá origem ao conceito de multi-task multi-view clustering trata cada vista individual com uma ou mais tarefas.

Abordagem proposta por Zhang et al. [42] é uma framework com três componentes, clusterização para cada vista-tarefa, aprendizagem da relação entre as diversas vistas e a aprendizagem da relação entre as tarefas. Durante a fase de clusterização para cada vista-tarefa é utilizado uma abordagem de co-clustering para cada vista em cada tarefa. Esta abordagem garante uma das partes essenciais deste algoritmo, a preservação do conhecimento que se encontra localmente em cada vista. A segunda fase desta *framework*, consiste na aprendizagem da relação entre as vistas, explorando a abordagem de minimizar a discórdia entre cada par vistas que são utilizadas para a mesma tarefa. Na última fase em que se procede à aprendizagem da relação entre as diversas tarefas, é utilizado, novamente, uma abordagem de co-clustering de modo a obter uma representação do subespaço comum sobre as tarefas consideradas.

4.4 Aplicações de Multi-view Learning

Um dos conjuntos de dados mais utilizados por investigadores na área de aprendizagem multi-vista é o WebKB [39], que contém 8282 páginas web recolhidas em departamentos de engenharia informática de quatro universidades americanas. É um conjunto de dados que possui duas vistas, a primeira contém o texto da página e a segunda o texto associado ao link para essa página.

Outro conjunto de dados amplamente utilizado é o Citeseer [39], uma coleção de 3312 publicações científicas distribuídas por 6 categorias. As vistas consideradas para este dataset são, naturalmente, uma vista com o texto, quer o título quer o *abstract*, e as outras duas vistas contemplam as citações, uma para os artigos que citam o artigo em questão e outra para os artigos citados.

Uma aplicação de multi-view learning consiste na anotação, classificação e recuperação de imagens, onde é comum utilizar o conjuntos de dados TRECVID2003 e o Caltech256 [39], em que diferentes atributos visuais são utilizados como vistas, tais como histogramas de cor, histogramas da direção das arestas e texturas de *wavelet*.

Nesta tese pretende-se validar que a segmentação de cliente também constitui uma aplicação promissora de aprendizagem multi vista. Deste modo, são consideradas quatro vistas consideradas para esta aplicação são os pilares de segmentação do marketing, sendo que a primeira vista contempla atributos geográficos, a segunda vista atributos demográficos, a terceira atributos psicográficos e a quarta atributos comportamentais.

4.5 Discussão de Resultados utilizando conjuntos de dados populares na literatura

Inspirados na experiência realizada pelos autores de [17], realizamos, também, experiências para comprovar a eficácia do algoritmo LMVSC em dados com uma única vista, considerando cada vista do conjunto de dados multi-view como um conjunto de dados com uma única vista. Analogamente, consideramos a aplicação do algoritmo SMVSC [34] em dados com uma única vista, de modo a comprovar a vantagem proposta por este método, a aprendizagem das âncoras e a construção do grafo serem mutuamente otimizadas. Deste modo, pretende-se comparar os dois algoritmos tanto no caso com uma única vista como no caso multi vista a fim de compreender as suas vantagens e limitações consoante a estrutura dos conjuntos de dados utilizados.

Os conjuntos de dados utilizados nesta experiência são: o *handwritten digits* com múltiplas vistas, em que se pretende clusterizar os diferentes dígitos; o *Caltech101-20*, em que se pretende clusterizar imagens de 20 objetos diferentes e, por último, o *Yale Faces*, em que se pretende clusterizar imagens de 15 pessoas distintas.

Todos os conjuntos de dados considerados contêm a verdadeira classificação de cada observação e, assim, além das métricas de avaliação internas 3.6.1, os resultados de clusterização produzidos com estes conjuntos de dados serão avaliados, também, com as

seguintes métricas de avaliação externas:

Accuracy (ACC) representa a fração entre o número de observações corretamente classificadas e o número total de observações.

$$accuracy(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1(\hat{y}_i = y_i) \quad (4.12)$$

em que $1(x)$ é a função indicadora. Varia entre 0 e 1, e é tão melhor quanto mais próximo de 1.

Normalized Mutual Information (NMI) Assumindo duas classificações U e V para as mesmas n observações:

A entropia é a quantidade de incerteza da partição é definida:

$$H(U) = - \sum_{i=0}^{|U|} P(i) \log(P(i)) \quad (4.13)$$

onde $P(i) = \frac{|U_i|}{n}$ é a probabilidade de uma observação escolhida de forma aleatória de U pertencer a classe U_i .

$$H(V) = - \sum_{j=0}^{|V|} P'(j) \log(P'(j)) \quad (4.14)$$

onde $P'(j) = \frac{|V_j|}{n}$

$$MI(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} P(i, j) \log \left(\frac{P(i, j)}{P(i)P'(j)} \right) \quad (4.15)$$

$$MI(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{n} \log \left(\frac{n|U_i \cap V_j|}{|U_i||V_j|} \right) \quad (4.16)$$

onde $P(i, j) = \frac{|U_i \cap V_j|}{n}$ é a probabilidade de uma observação escolhida ao acaso pertencer às classes U_i e V_j .

$$NMI(U, V) = \frac{MI(U, V)}{\text{mean}\{H(U), H(V)\}} \quad (4.17)$$

O **NMI** tem um limite superior de 1, que indica que as duas classificações U e V são iguais, com ou sem permutações, valores perto de 0 indicam que as classificações são independentes.

Purity é uma medida de se um cluster possui, apenas, uma classe. Para cada *cluster* contabiliza-se o número de observações da classe maioritária desse *cluster*

$$\frac{1}{n} \sum_{m \in M} \max_{d \in D} |m \cap d| \quad (4.18)$$

em que M é o conjunto de clusters e

F-score é a média harmónica entre a precisão e a sensibilidade. Varia entre 0, o seu pior valor, e 1, o seu melhor valor.

$$precisão = \frac{TP}{TP + FP} \quad sensibilidade = \frac{TP}{TP + FN} \quad (4.19)$$

onde Número de verdadeiros positivos (TP), Número de falsos positivos (FP) e Número de falsos negativos (FN).

$$F_\beta = (1 + \beta^2) \frac{precisão \times sensibilidade}{\beta^2 precisão + sensibilidade} \quad (4.20)$$

4.5.1 Resultados com o conjunto de dados Multiple Features handwritten digit

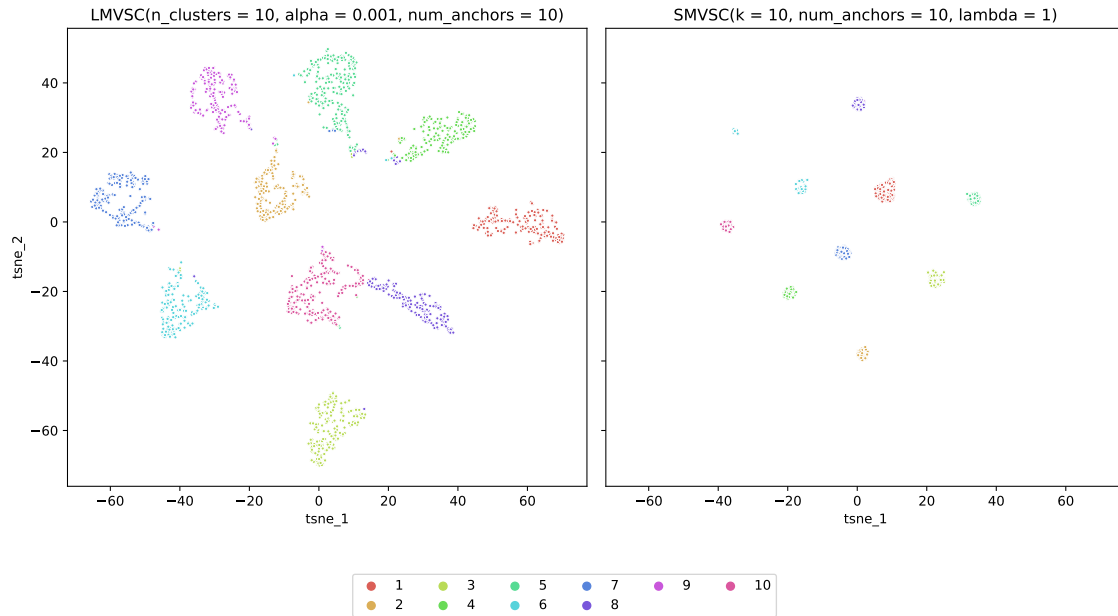


Figura 4.8: Visualização dos resultados de clusterização para o conjunto de dados Multiple Features handwritten digit, utilizando a técnica de redução de dimensionalidade t-SNE, utilizando todas as vistas.

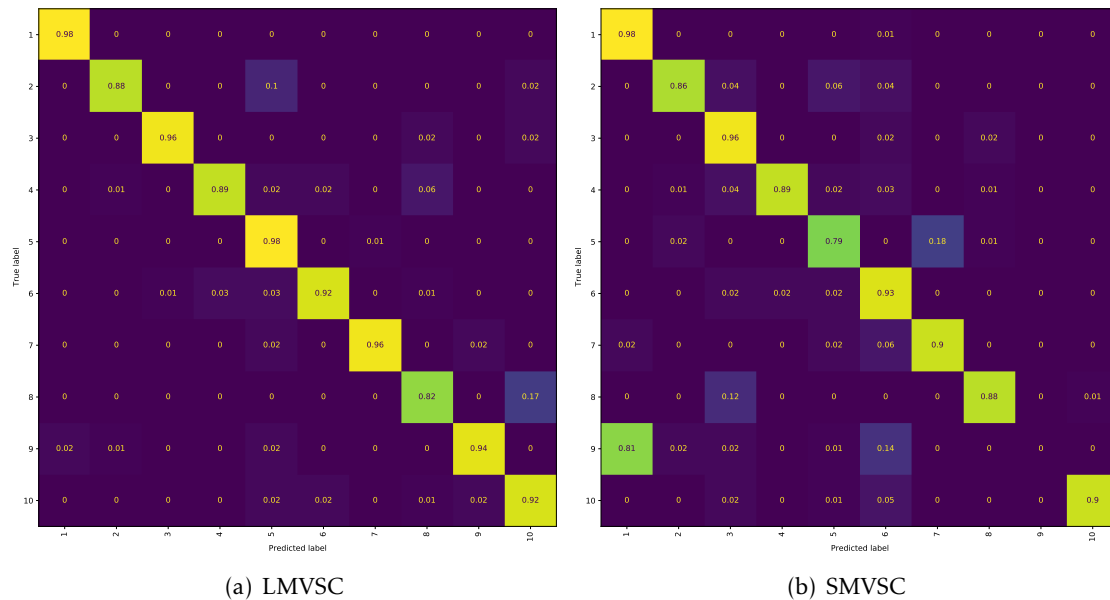


Figura 4.9: Matriz de confusão para o conjunto de dados Multiple Features handwritten digit, utilizando todas as vistas, normalizada ao longo das linhas.

4.5. DISCUSSÃO DE RESULTADOS UTILIZANDO CONJUNTOS DE DADOS POPULARES NA LITERATURA

	Multi Vista	Vista 1	Vista 2	Vista 3	Vista 4	Vista 5	Vista 6
ACC ($\uparrow$)	0.926	0.663	0.733	0.875	0.449	0.740	0.548
NMI ($\uparrow$)	0.861	0.657	0.690	0.791	0.400	0.736	0.512
Purity ($\uparrow$)	0.926	0.699	0.733	0.875	0.473	0.782	0.566
F-score ($\uparrow$)	0.859	0.583	0.624	0.776	0.317	0.687	0.454
SC ($\uparrow$)	0.592	0.548	0.538	0.517	0.440	0.602	0.502
CH ($\uparrow$)	1217.328	1066.051	814.382	810.952	601.367	1323.541	900.192
DB ($\downarrow$)	0.609	0.638	0.944	0.996	1.029	0.558	0.700

Tabela 4.2: Performance da clusterização no dataset Multiple Features handwritten digit, utilizando o algoritmo LMVSC.

Os resultados obtidos, utilizando o algoritmo LMVSC, são superiores aos produzidos pelo algoritmo SMVSC, no caso multi-view e em todas as experiências, considerando cada vista como um conjunto de dados com uma única vista, à exceção da experiência realizada com a vista Morphological features. Era expectável uma melhor performance do algoritmo SMVSC face ao algoritmo LMVSC, uma vez que os autores [34] propuseram o algoritmo SMVSC de forma a colmatar as duas deficiências do algoritmo LMVSC, identificadas pelos autores, nomeadamente, a fixação das âncoras obtidas assim que são inicializadas, o que causa um isolamento entre a aprendizagem das âncoras e a construção do grafo, e as âncoras serem geradas de forma independente para as diferentes vistas, o que causa a não exploração da complementaridade entre as mesmas e dificulta a fusão dos grafos. Dado que as expectativas teóricas e os resultados práticos não coincidiram, procedemos a uma análise mais detalhada. Para tal, visualizou-se, utilizando a técnica de redução de dimensionalidade t-SNE, os dados originais das várias vistas e, assim, demonstrou-se que a separação dos vários clusters já existia, nas mesmas vistas em que a performance deste algoritmo era superior, como se pode observar nas figuras B.1(a), B.1(b), B.1(c), B.1(e) e B.1(f). Assim, a abordagem de escolher as âncoras através da aplicação do algoritmo de k-means a cada vista individualmente, prova-se eficaz, por cada vista segmentar, naturalmente, os vários dígitos, permitindo que as âncoras obtidas sejam discriminativas. Dado que cada vista é suficiente para produzir uma boa clusterização, a matriz Z que resulta da concatenação dos vários grafos construídos, utilizando as âncoras obtidas, melhora os resultados de clusterização.

Os resultados obtidos pelo algoritmo SMVSC são inferiores aos obtidos no algoritmo LMVSC, como se pode observar na figura 4.9. Embora este algoritmo mantenha uma capacidade de classificar corretamente a maioria das classes, perde *accuracy* por classificar a grande maioria dos exemplares da classe 9 como exemplares da classe 1. É importante ressaltar que, na ausência de separação entre os *clusters*, como é o caso da vista 4, este algoritmo consegue obter melhores resultados na experiência considerando, apenas, esta vista, o que confirma a superioridade de uma abordagem de aprendizagem de âncoras, face à abordagem tradicional da sua seleção através do algoritmo de *k-means*, isolada do processo de construção do grafo. Apesar da performance da abordagem multi-view

	Multi Vista	Vista 1	Vista 2	Vista 3	Vista 4	Vista 5	Vista 6
ACC (↑)	0.821	0.591	0.497	0.706	0.607	0.676	0.537
NMI (↑)	0.789	0.614	0.469	0.617	0.614	0.643	0.474
Purity (↑)	0.821	0.614	0.501	0.706	0.622	0.676	0.537
F-score (↑)	0.753	0.539	0.417	0.558	0.555	0.575	0.446
SC (↑)	0.944	0.807	0.652	0.658	0.442	0.794	0.686
CH (↑)	3720.899	1709.140	942.480	952.322	494.051	1420.186	1622.385
DB (↓)	0.425	0.702	0.808	0.818	1.208	0.691	0.495

Tabela 4.3: Performance da clusterização no dataset Multiple Features handwritten digit, utilizando o algoritmo SMVSC.

neste algoritmo ter sido inferior nas métricas externas, é de salientar a coesão, separabilidade e distinguibilidade dos segmentos obtidos, como se pode observar na figura 4.8, verificando-se métricas internas muito superiores às obtidas pelo algoritmo LMVSC. O seu silhouette score é bastante alto, cerca de 0.94, sendo comprovativo dos *clusters* obtidos, pela aplicação do algoritmo *k-means* na matriz U , constituída pelos vetores singulares esquerdos resultantes da decomposição em valores singulares do grafo unificado Z . O valor do índice de Davies-Bouldin para a aplicação do algoritmo SMVSC na experiência multi-view é o melhor comparando com todas as outras experiências realizadas com este conjunto de dados, o que permite afirmar que a aplicação deste algoritmo, num contexto multi-view, produz a melhor partição no grafo unificado aprendido, ainda que à custa da perda de *accuracy*, causada pela aglomeração, inadvertida, de dois *clusters*. Em suma, o estudo realizado, utilizando o conjunto de dados *Multiple Features handwritten digit*, demonstra a superioridade da abordagem tradicional seguida pelo algoritmo LMVSC, ao nível dos índices externos, quando já existe separabilidade entre os vários *clusters* em cada vista dos dados originais, o que torna este algoritmo um melhor classificador dos diversos dígitos presentes neste conjunto de dados. Relativamente aos índices internos, estes apresentam valores muito mais baixos, o que pode ser explicado pela fraca separabilidade entre os extitclusters obtidos, considerando a matriz Z , o grafo ancorado unificado. A utilização deste algoritmo evidencia-se, assim, menos adequada para a aplicação de *multi-view* presente neste trabalho, a segmentação de cliente, uma vez que não nos encontramos num caso em que esta segmentação já exista nos dados originais, o que questiona a possibilidade de obtenção de bons resultados com esta abordagem. Como tal, consideramos que a utilização do algoritmo SMVSC permite a obtenção de *clusters* com uma melhor coesão, separabilidade e distinguibilidade, o que é coincidente com o conceito de segmentação natural 2.2.3. Embora este tipo de abordagem não seja isenta de defeitos, tal como a aglomeração de duas classes distintas, numa só, consideramos que, à luz desta aplicação, *clusters* que, após exploradas as características dos clientes que os constituem, se chegue à conclusão que podemos ter cometido o referido erro, faz sentido explorar duas alternativas: a primeira consiste em recorrer o algoritmo, mas considerando, apenas, os clientes contidos no *cluster* problemático a fim de que estes, ao serem isolados,

se revelem diferentes o suficiente para produzir uma segmentação adequada; a segunda, quando verificamos que a primeira estratégia falha, concluimos que só é possível obter segmentações mais finas recorrendo a segmentações construtivas 2.2.3, uma vez que não se observa mais separabilidade por via algorítmica, deixando de existir a possibilidade de segmentações consistentes com a definição de *cluster* e, igualmente, deixando de existir estrutura suficiente para gerar a mesma segmentação de forma consistente. Assim, a única alternativa consiste na sua criação de forma artificial, através de conhecimento do negócio, visto que a sua subsegmentação pode ser mais promissora do que a aplicação de estratégias indiferenciadas.

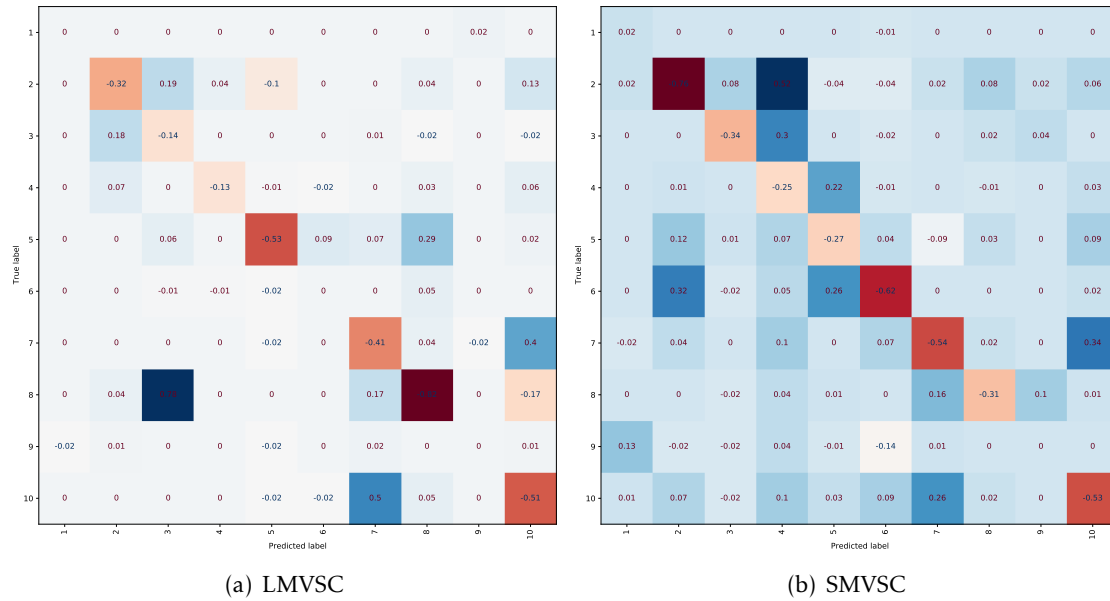


Figura 4.10: Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Multiple Features handwritten digit, na vista Fourier coefficients of the character shapes.

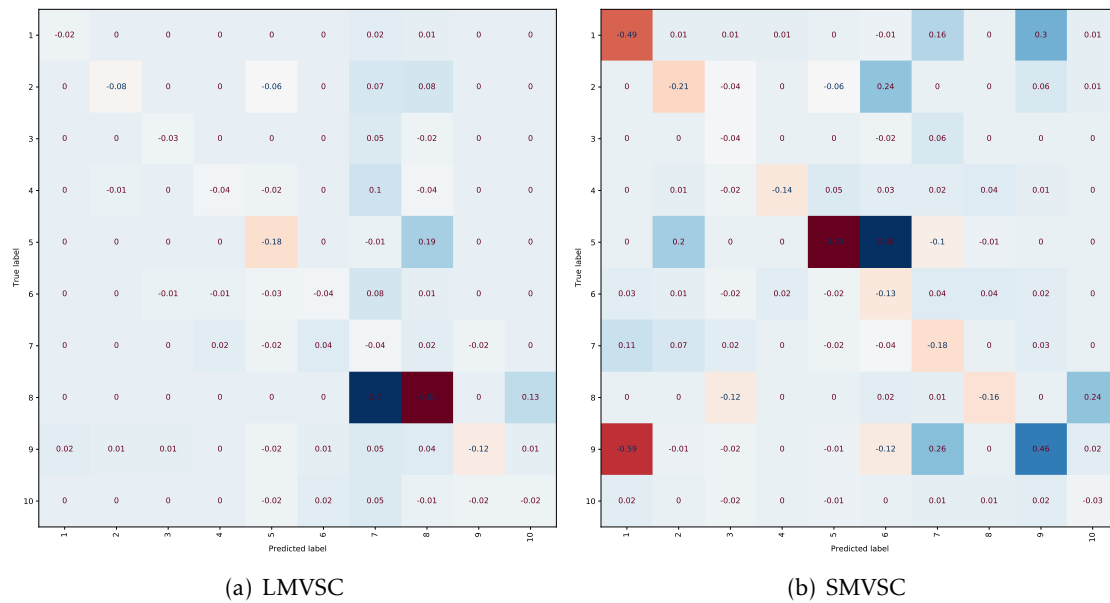


Figura 4.11: Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Multiple Features handwritten digit, na vista Karhunen-Love coefficients.

4.5. DISCUSSÃO DE RESULTADOS UTILIZANDO CONJUNTOS DE DADOS POPULARES NA LITERATURA

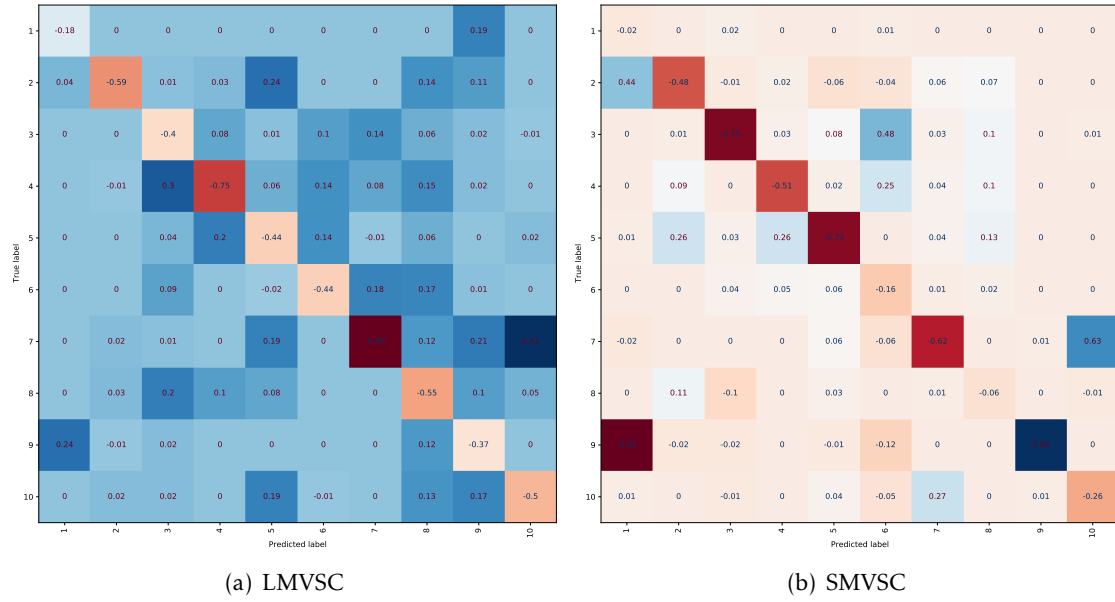


Figura 4.12: Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Multiple Features handwritten digit, na vista Morphological features.

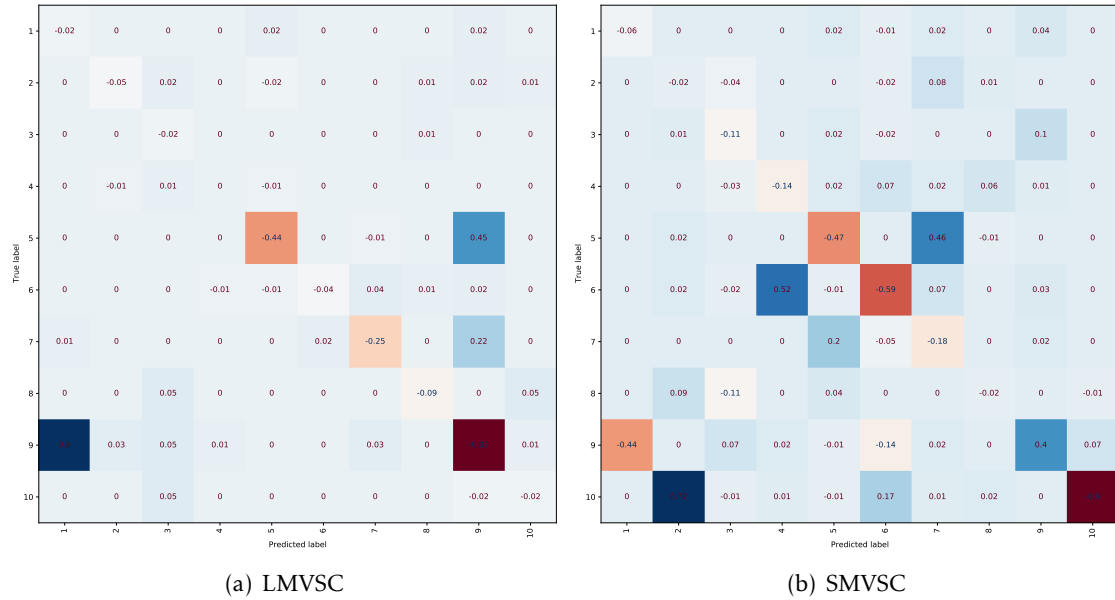


Figura 4.13: Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Multiple Features handwritten digit, na vista Pixel averages in 2 x 3 windows.

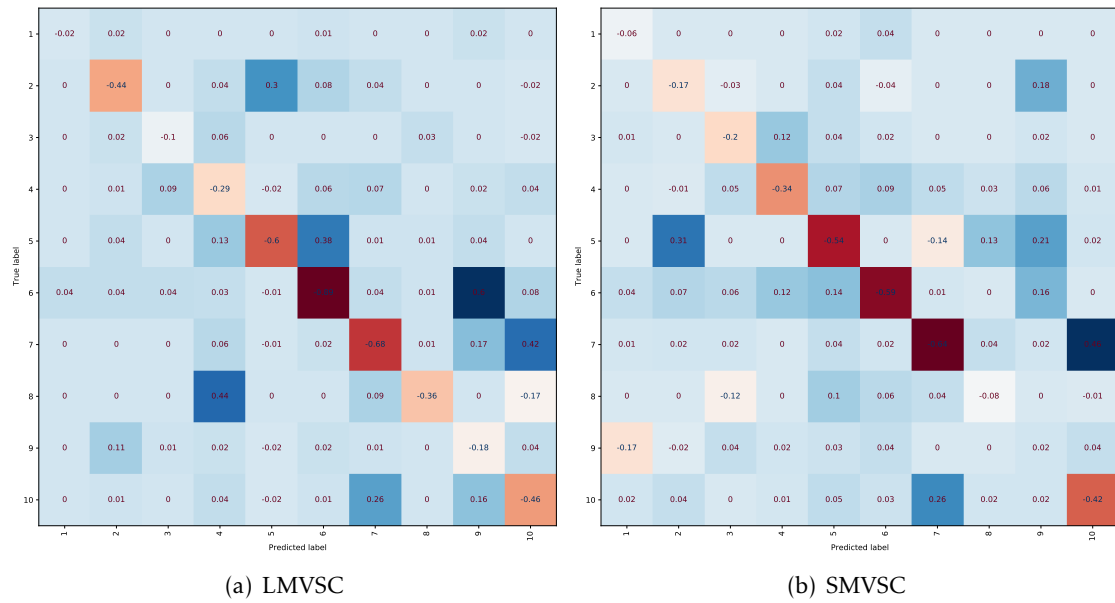


Figura 4.14: Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Multiple Features handwritten digit, na vista Zernike moments.

4.5.2 Resultados com o conjunto de dados Caltech101-20

Procedeu-se à validação dos resultados obtidos foi utilizado o conjunto de dados Caltech101-20, com 2386 observações 20 *clusters* e 6 vistas:

- Gabor [19], com 48 atributos.
- Wavelet moments [10], com 40 atributos.
- CENTRIST [38], com 254 atributos.
- HOG [7], com 1984 atributos.
- GIST [24], com 512 atributos.
- LBP [23], com 928 atributos.

Este conjunto de dados contém imagens de 20 categorias diferentes, mas a sua distribuição não é equitativa, uma vez que as quatro categorias com mais imagens, Motorbikes com 798 exemplares, Faces com 435 exemplares, Leopards com 200 exemplares e Car side com 123 exemplares, contêm 1556 das 2386 imagens totais, o que equivale a 65.21%. Os autores de [26] apresentam limitações do conjunto de dados Caltech101. Apesar de este ser um dos conjuntos de dados mais diversos em termos de variabilidade entre classes, tem uma fraca variabilidade dentro da mesma classe, uma vez que muitos dos objetos têm um tamanho e orientação uniforme com fundos pouco ricos. O autor defende, também, que as imagens neste conjunto de dados contêm pouco ou nenhum ruído, os objetos representados estão centrados nas imagens e a maioria dos objetos representados estão numa posição estereotípica dessa classe. Apesar das limitações levantadas, este conjunto de dados continua a ser um standard para a avaliação de algoritmos de classificação multi-classe.

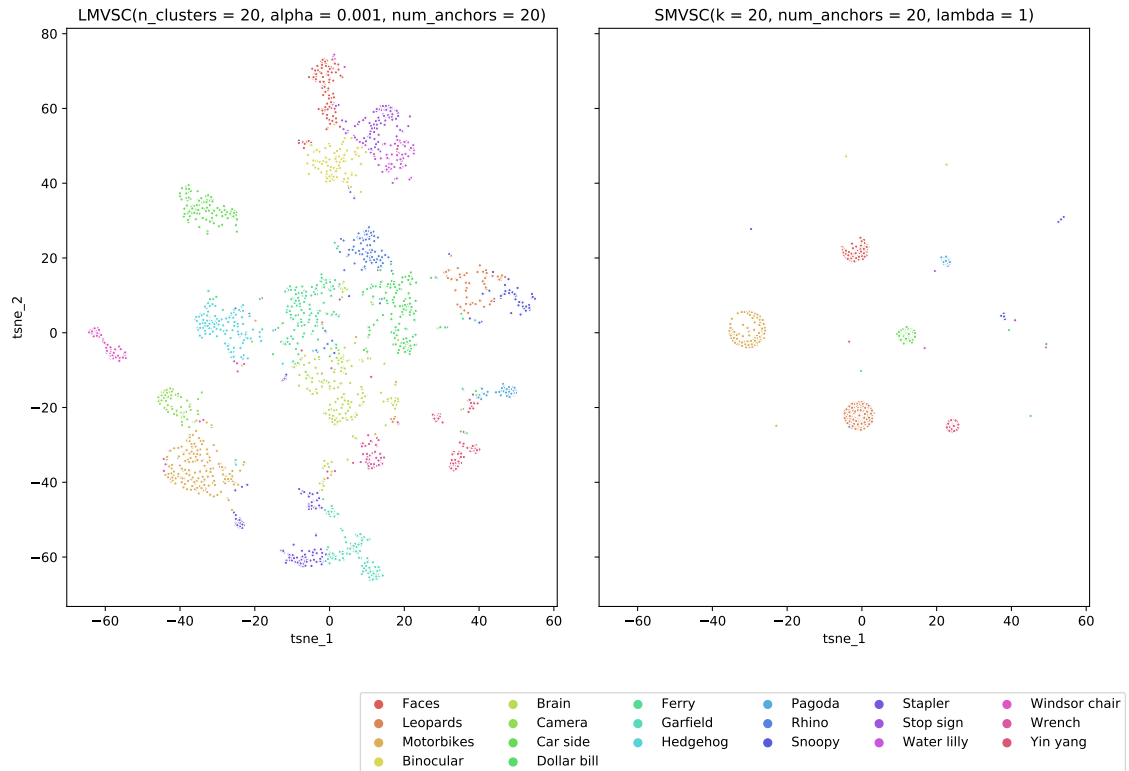


Figura 4.15: Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE, utilizando todas as vistas.

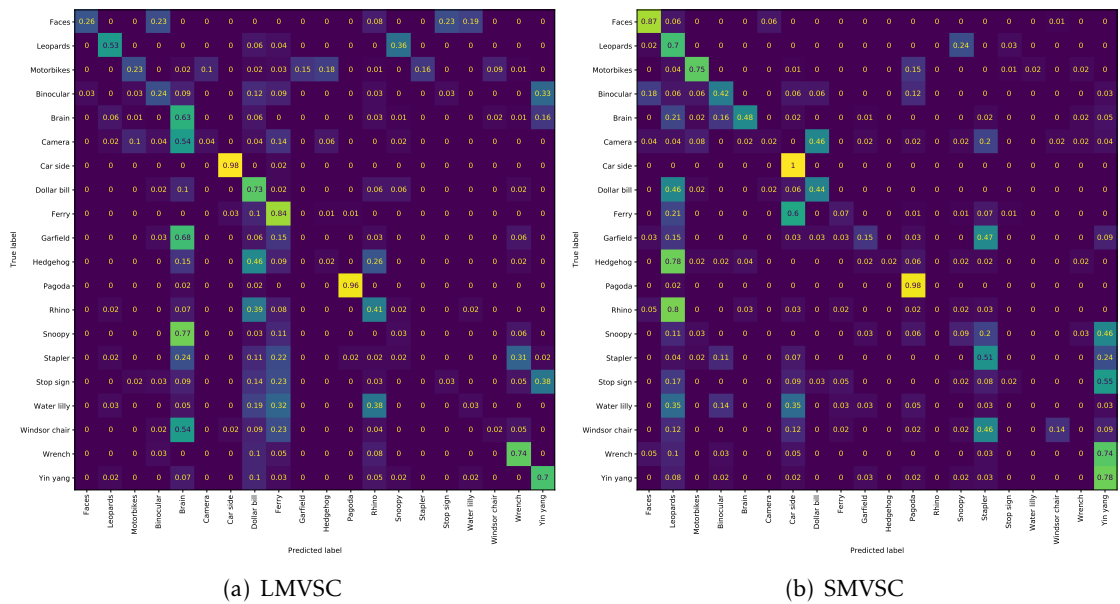


Figura 4.16: Matriz de confusão para o conjunto de dados Caltech101-20, utilizando todas as vistas, normalizada ao longo das linhas.

4.5. DISCUSSÃO DE RESULTADOS UTILIZANDO CONJUNTOS DE DADOS POPULARES NA LITERATURA

	Multi Vista	Vista 1	Vista 2	Vista 3	Vista 4	Vista 5	Vista 6
ACC (↑)	0.441	0.247	0.334	0.331	0.534	0.467	0.428
NMI (↑)	0.566	0.244	0.368	0.389	0.650	0.560	0.499
Purity (↑)	0.730	0.480	0.612	0.596	0.780	0.717	0.671
F-score (↑)	0.415	0.202	0.294	0.278	0.526	0.552	0.413
SC (↑)	0.289	0.278	0.373	0.319	0.526	0.345	0.410
CH (↑)	219.973	235.477	292.935	253.103	533.529	337.422	285.337
DB (↓)	1.271	1.271	1.241	1.373	1.055	1.004	1.246

Tabela 4.4: Performance da clusterização no dataset Caltech101-20, utilizando o algoritmo LMVSC.

	Multi Vista	Vista 1	Vista 2	Vista 3	Vista 4	Vista 5	Vista 6
ACC (↑)	0.615	0.249	0.384	0.493	0.555	0.539	0.489
NMI (↑)	0.587	0.269	0.348	0.399	0.614	0.541	0.519
Purity (↑)	0.690	0.469	0.574	0.584	0.748	0.729	0.674
F-score (↑)	0.661	0.195	0.354	0.456	0.524	0.498	0.456
SC (↑)	0.93	0.599	0.576	0.813	0.907	0.805	0.827
CH (↑)	1874.637	582.165	765.726	1442.759	1783.391	1123.323	1138.102
DB (↓)	0.579	1.326	1.273	0.996	0.762	0.608	1.0311

Tabela 4.5: Performance da clusterização no dataset Caltech101-20, utilizando o algoritmo SMVSC.

CAPÍTULO 4. ALGORITMOS DE APRENDIZAGEM MULTI-VISTA

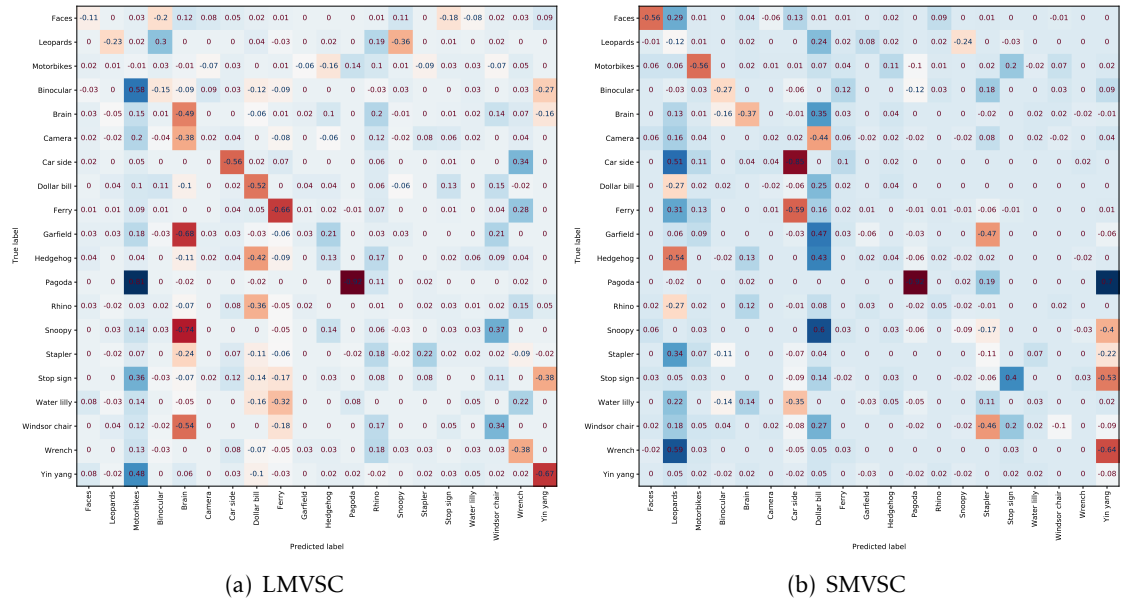


Figura 4.17: Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Caltech101-20, na vista Gabor.

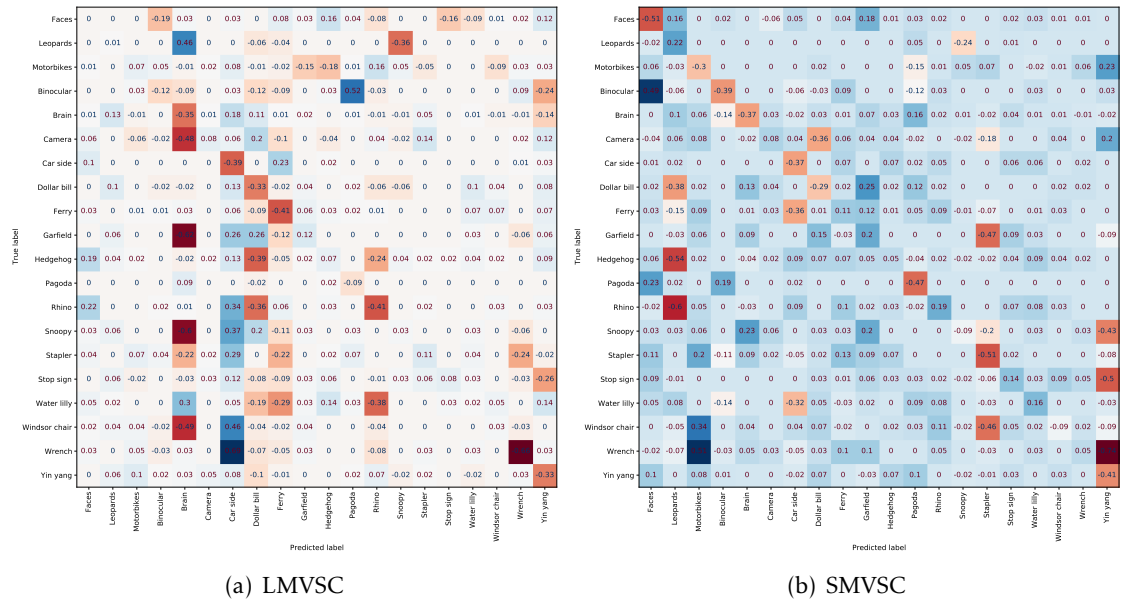


Figura 4.18: Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Caltech101-20, na vista Wavelet moments.

4.5. DISCUSSÃO DE RESULTADOS UTILIZANDO CONJUNTOS DE DADOS POPULARES NA LITERATURA

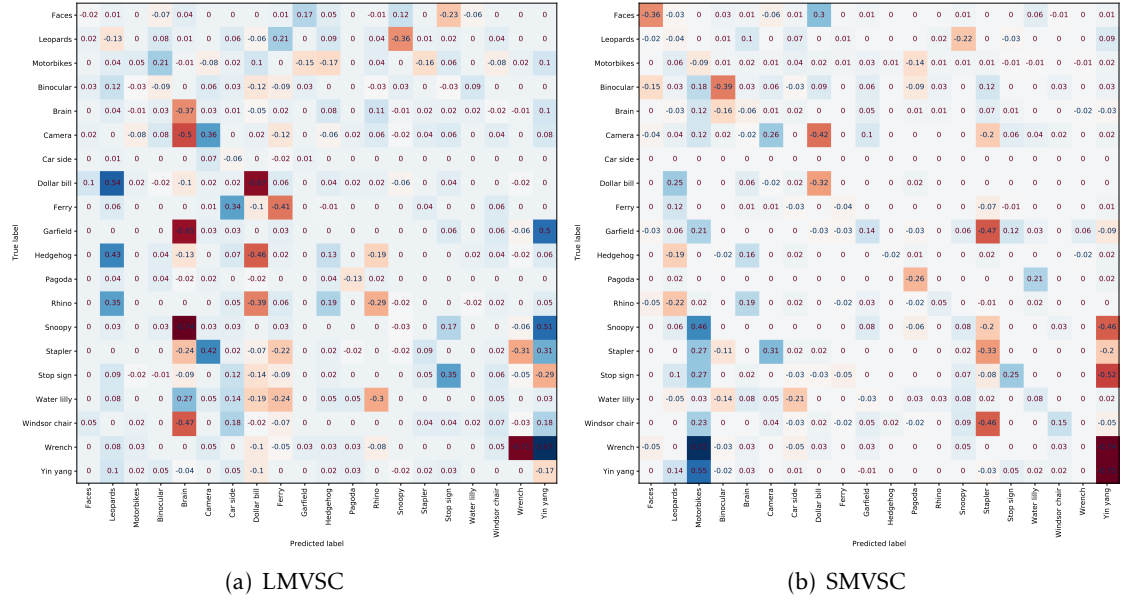


Figura 4.19: Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Caltech101-20, na vista CENTRIST.

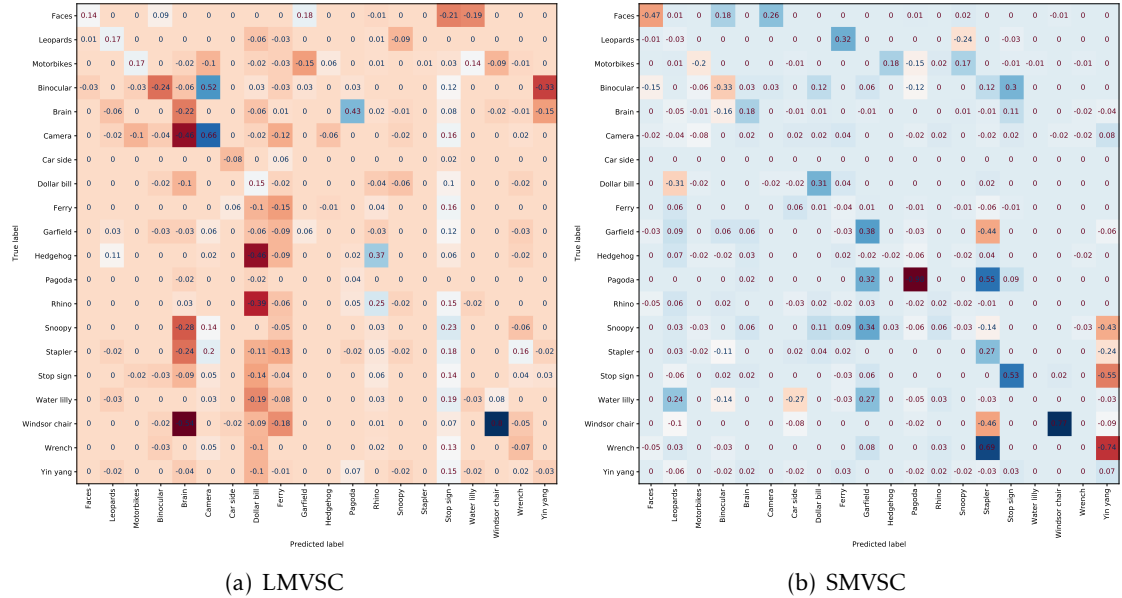


Figura 4.20: Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Caltech101-20, na vista HOG.

CAPÍTULO 4. ALGORITMOS DE APRENDIZAGEM MULTI-VISTA

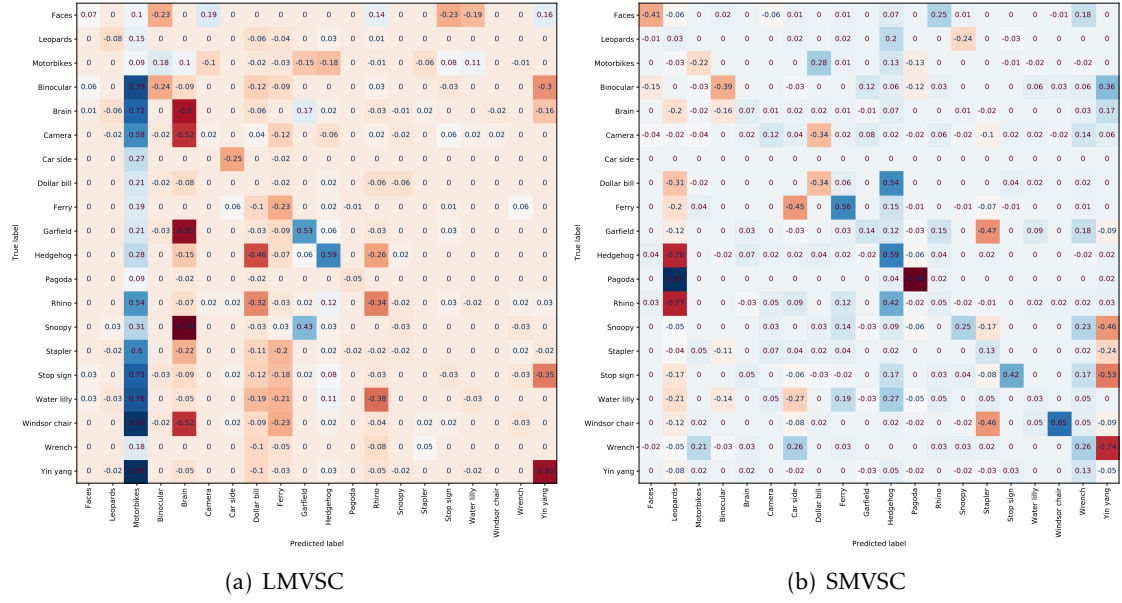


Figura 4.21: Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Caltech101-20, na vista GIST.

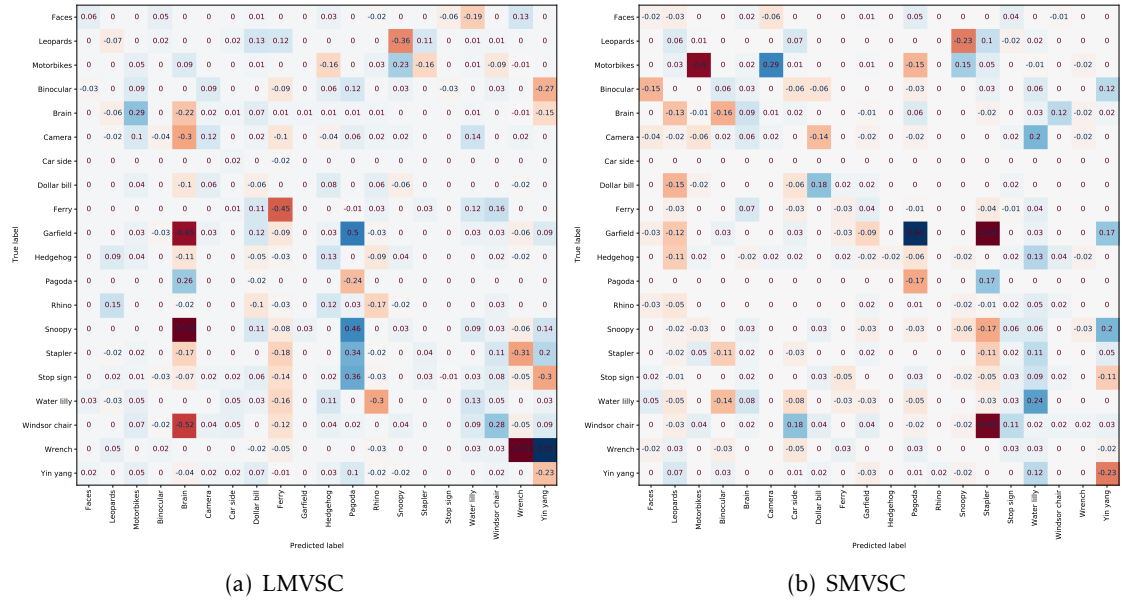


Figura 4.22: Diferença entre as matrizes de confusão entre modelo *single-view* e o modelo *multi-view* para o conjunto de dados Caltech101-20, na vista LBP.

4.5.3 Resultados sem considerar a vista HOG

Através da experiência *single-view* realizada, considerando a quarta vista, a vista HOG, foi posto em evidência que, quando aplicado o algoritmo **LMVSC**, tanto as métricas internas como as métricas externas revelam que esta vista, sozinha, produz não só uma melhor classificação do que qualquer outra vista, de forma isolada, mas também uma melhor classificação, cerca de 10% superior, que no caso *multi-view*.

Quando aplicado o algoritmo **SMVSC**, a melhor performance é obtida pelo caso *multi-view*, quer nas métricas internas quer nas métricas externas, ainda assim a vista HOG apresenta uma performance, ainda que inferior, bastante próxima do caso *multi-view*. Deste modo, considera-se relevante a criação de uma experiência em que os algoritmos **LMVSC** e **SMVSC** sejam aplicados a um cenário *multi-view*, removendo-se a vista que se considera mais discriminativa na classificação dos diversos objetos.

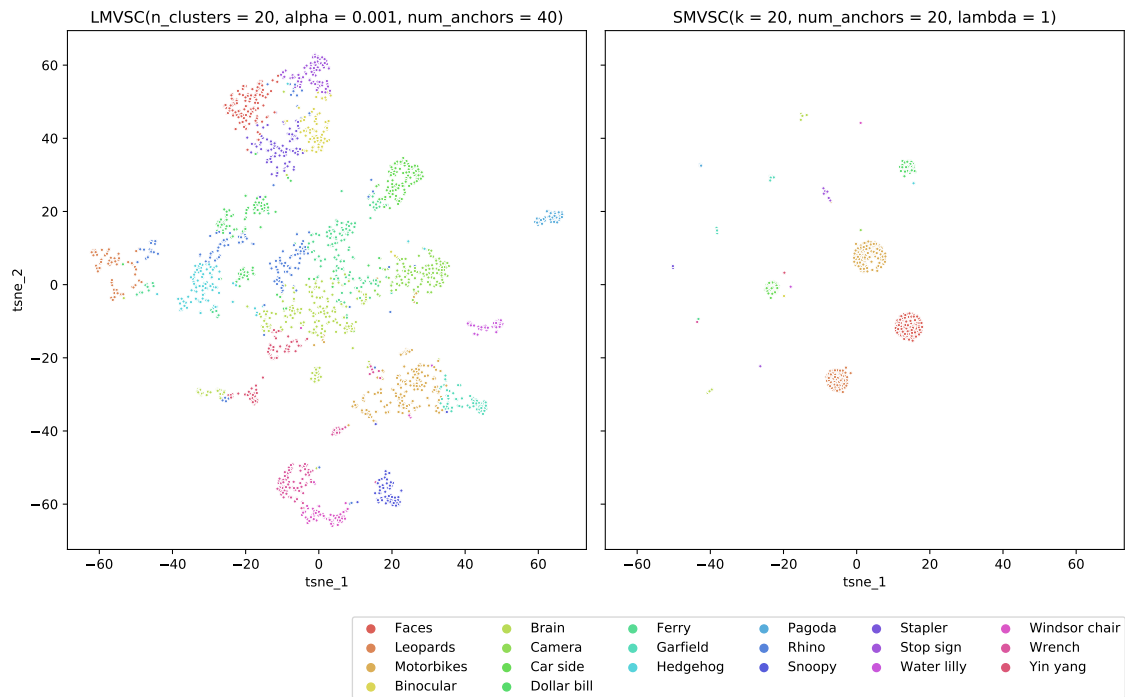


Figura 4.23: Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE, utilizando todas as vistas exceto a 4.

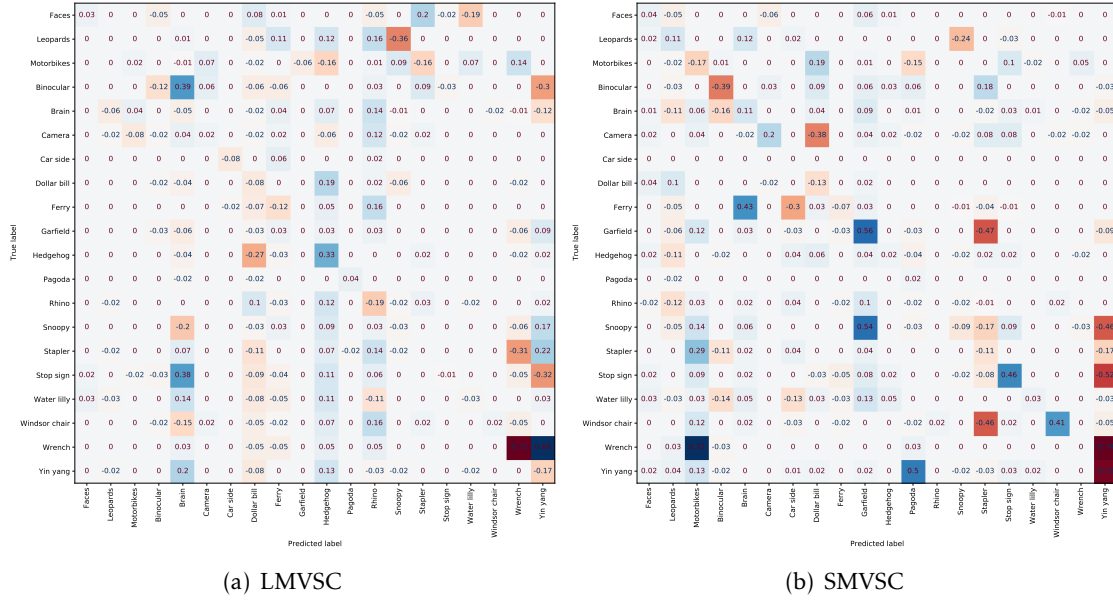


Figura 4.24: Matriz de confusão para o conjunto de dados Caltech101-20, utilizando todas as vistas exceto a vista HOG, normalizada ao longo das linhas.

	Multi Vista	Multi Vista (sem 4)	Vista 1	Vista 2	Vista 3	Vista 5	Vista 6
ACC ($\uparrow$)	0.441	0.387	0.247	0.334	0.331	0.467	0.428
NMI ($\uparrow$)	0.566	0.534	0.244	0.368	0.389	0.560	0.499
Purity ($\uparrow$)	0.733	0.695	0.482	0.612	0.596	0.717	0.671
F-score ($\uparrow$)	0.415	0.355	0.202	0.294	0.278	0.552	0.413
SC ($\uparrow$)	0.289	0.275	0.278	0.373	0.319	0.345	0.410
CH ($\uparrow$)	219.973	227.638	235.477	292.935	253.103	337.422	285.337
DB ($\downarrow$)	1.271	1.301	1.271	1.241	1.373	1.004	1.246

Tabela 4.6: Performance da clusterização no dataset Caltech101-20, utilizando o algoritmo LMVSC.

A realização desta experiência com o algoritmo LMVSC demonstra a perda significativa de *accuracy* causada pela remoção da vista mais discriminativa, o que demonstra que a estratégia de fundir os grafos ancorados, utilizando pesos iguais para produzir o grafo de consenso, produz resultados que deteoram a performance obtida pelas vistas individuais, uma vez que o algoritmo ignora o facto de as vistas consideradas poderem não ter um igual contributo para a clusterização final e ignora, também a informação complementar entre as várias vistas, dado que as três fases deste algoritmo (a escolha das âncoras, a construção dos grafos âncorados e a criação do grafo unificado) são independentes umas das outras e não partilham informação entre as vistas. No caso particular desta experiência, ao nível da área de processamento de imagem, as imagens podem ser descritas pelos filtros que lhes são aplicados. Os filtros utilizados foram considerados como vistas separadas, sendo tal procedimento dado como exemplo pelos autores de [40]

4.5. DISCUSSÃO DE RESULTADOS UTILIZANDO CONJUNTOS DE DADOS POPULARES NA LITERATURA

os quais fundamentam a exploração do princípio da complementariedade como forma de garantir a eficácia de multi-view clustering. Alinhando o argumento dos autores de [40] com os atributos deste conjunto de dados, temos uma vista com LBP [23], um atributo de textura, e outra vista com HOG, sensível a informação marginal. Embora cada uma destas vistas seja suficiente para uma determinada tarefa de descoberta de conhecimento, é necessária a exploração da informação complementar entre as várias vistas para descrever os objetos e aprofundar conhecimento sobre o agrupamento produzido.

	Multi Vista	Multi Vista (sem 4)	Vista 1	Vista 2	Vista 3	Vista 5	Vista 6
ACC (↑)	0.615	0.597	0.249	0.384	0.493	0.539	0.489
NMI (↑)	0.587	0.562	0.269	0.348	0.399	0.541	0.519
Purity (↑)	0.690	0.725	0.469	0.574	0.584	0.729	0.674
F-score (↑)	0.661	0.566	0.195	0.354	0.456	0.498	0.456
SC (↑)	0.930	0.911	0.599	0.576	0.813	0.805	0.827
CH (↑)	1874.637	1317.215	582.165	765.726	1442.759	1123.323	1138.102
DB (↓)	0.579	0.707	1.326	1.273	0.996	0.608	1.031

Tabela 4.7: Performance da clusterização no dataset Caltech101-20, utilizando o algoritmo SMVSC.

A realização da mesma experiência com o algoritmo SMVSC, demonstra a sua superioridade, uma vez que, apesar de também lhe ser retirada a vista com a melhor performance individual tanto a nível de métricas internas como externas, este algoritmo, ao contrário do LMVSC, é capaz de uma performance comparável com a experiência multi-view que contém esta vista e superior à performance de qualquer vista quando considerada de forma individual. Ainda assim, a performance desta experiência é inferior, em quase todas as métricas, à experiência multi-view que contém a vista HOG, o que confirma a importância desta informação para a clusterização final. Este facto, juntamente com a constatação da perda de performance pela abordagem tradicional, demonstra a vantagem da abordagem do algoritmo SMVSC, especialmente no caso multi-view, nomeadamente a superioridade da aprendizagem das âncoras e a construção dos grafos serem otimizados de forma conjunta, de modo a melhorar a estrutura da clusterização, como sugerido pelos autores deste algoritmo [34].

4.5.4 Resultados com o conjunto de dados Yale Face

Procedeu-se à validação dos resultados obtidos foi utilizado o conjunto de dados Yale Faces, com 165 observações, de 15 pessoas, em que cada pessoa constitui um *clusters* e 3 vistas:

- Intensity, com 4096 atributos.
- LBP [23], com 3304 atributos.
- Gabor [19], com 6750 atributos.

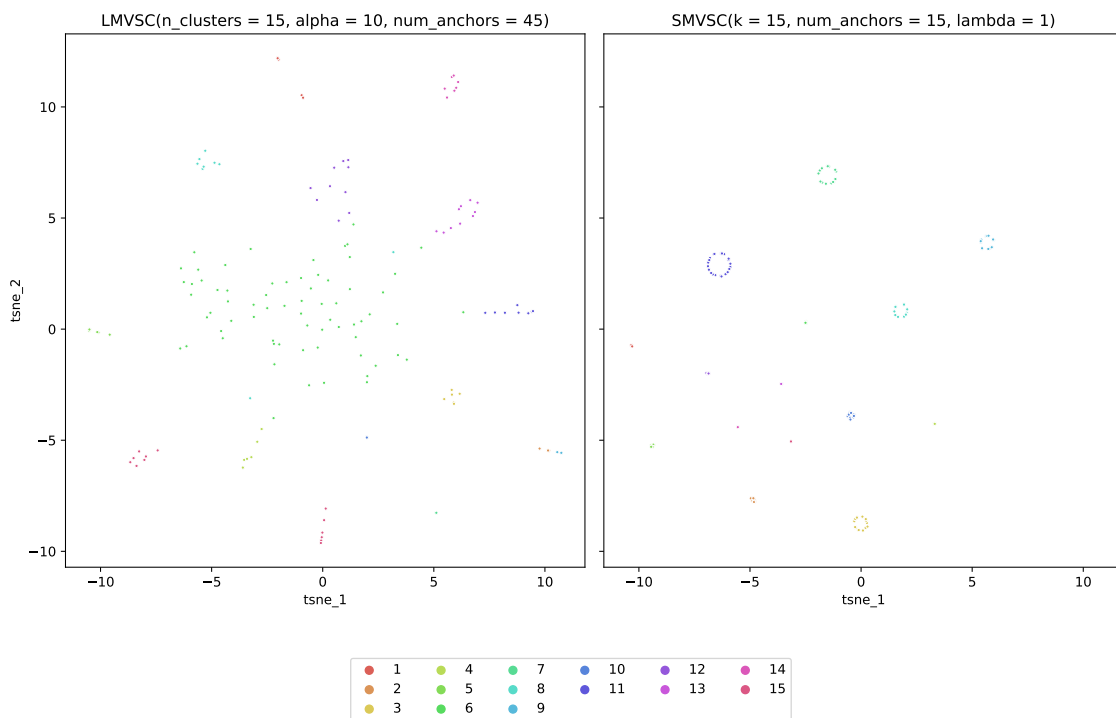


Figura 4.25: Visualização dos resultados de clusterização para o conjunto de dados Yale Face, utilizando a técnica de redução de dimensionalidade t-SNE, utilizando todas as vistas.

4.5. DISCUSSÃO DE RESULTADOS UTILIZANDO CONJUNTOS DE DADOS POPULARES NA LITERATURA

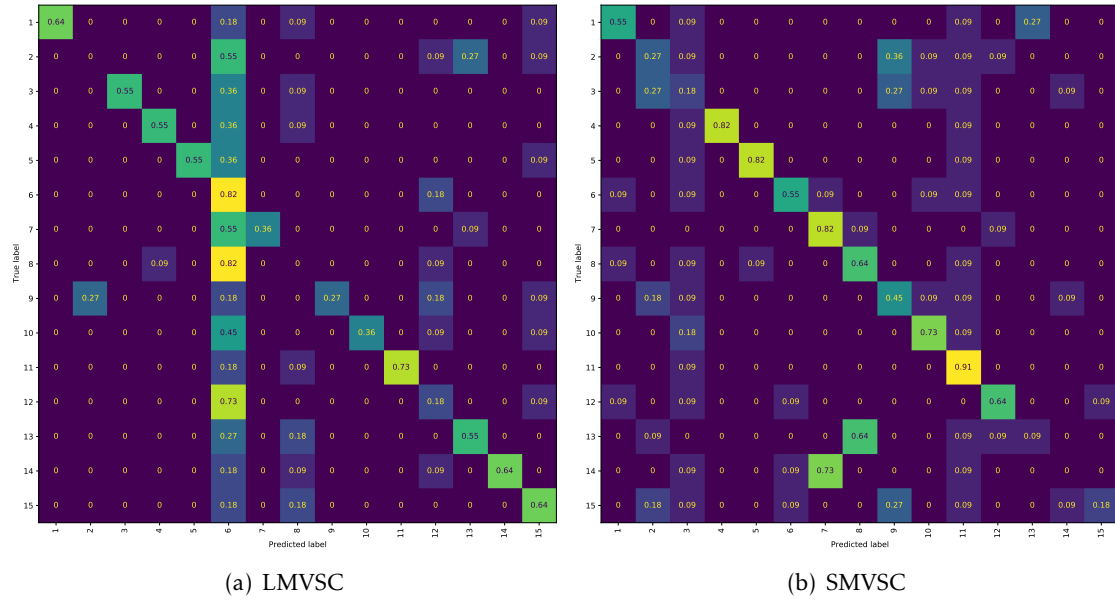


Figura 4.26: Matriz de confusão para o conjunto de dados Yale Face, utilizando todas as vistas, normalizada ao longo das linhas.

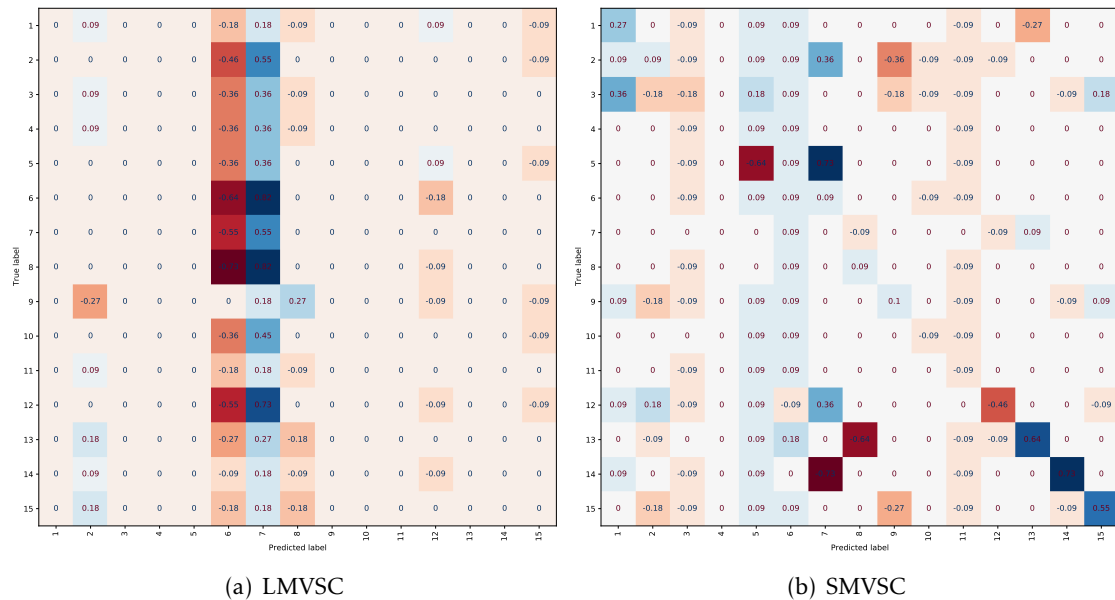


Figura 4.27: Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Yale Face, na vista Intensity.

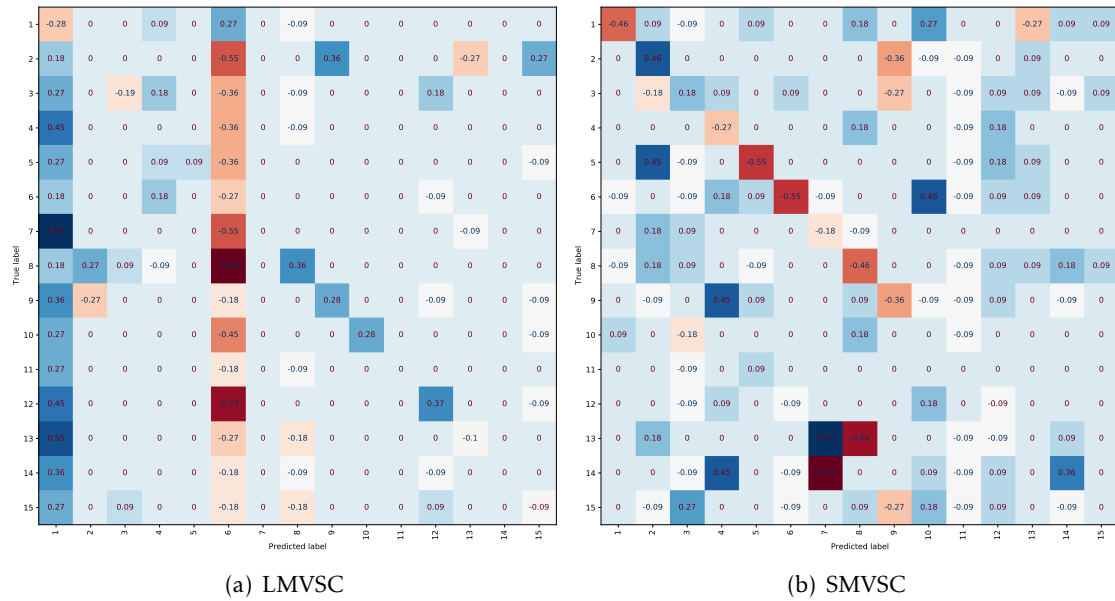


Figura 4.28: Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Yale Face, na vista LBP.

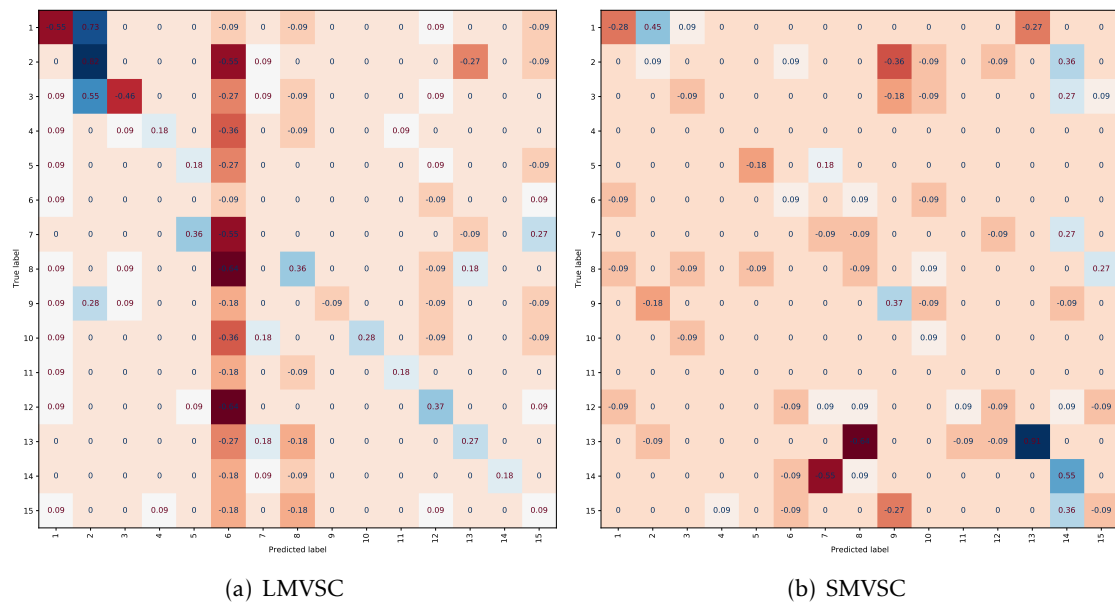


Figura 4.29: Diferença entre as matrizes de confusão entre modelo single-view e o modelo multi-view para o conjunto de dados Yale Face, na vista Gabor.

4.5. DISCUSSÃO DE RESULTADOS UTILIZANDO CONJUNTOS DE DADOS POPULARES NA LITERATURA

	MV	Vista 1 Intensity	Vista 2 LBP	Vista 3 Gabor
ACC (↑)	0.521	0.521	0.509	0.582
NMI (↑)	0.568	0.566	0.563	0.624
Purity (↑)	0.527	0.527	0.533	0.588
F-score (↑)	0.309	0.288	0.263	0.410
SC (↑)	0.502	0.498	0.645	0.527
CH (↑)	42.823	42.094	48.734	44.210
DB (↓)	0.873	0.6555	0.864	0.907

Tabela 4.8: Performance da clusterização no dataset Yale Face, utilizando o algoritmo LMVSC.

Como podemos observar na tabela 4.8, a performance do algoritmo LMVSC, considerando todas as vistas, é inferior à verificada quando se considera apenas a terceira vista, com os atributos do filtro de Gabor. A abordagem multi-view, isto é, considerando todas as vistas, provoca a classificação incorreta da classe 6 que contém 68 observações quando, idealmente, cada classe devia conter apenas 15. Assim, este algoritmo, apesar de conseguir clusterizar corretamente algumas imagens, classifica incorretamente imagens de todas as classes, identificando-as como sendo exemplares de imagens da classe 6. Através da experiência de considerar cada vista como um conjunto de dados single-view, confirma-se que esta perda de performance é resultante do facto de duas das três vistas, a vista Intensity e LBP, produzirem clusterizações em que uma classe contém, pelo menos, um terço das observações, não conseguindo distinguir os diferentes indivíduos dentro desta classe.

	Multi Vista	Vista 1 Intensity	Vista 2 LBP	Vista 3 Gabor
ACC (↑)	0.509	0.582	0.424	0.588
NMI (↑)	0.564	0.594	0.498	0.614
Purity (↑)	0.527	0.594	0.430	0.612
F-score (↑)	0.364	0.374	0.281	0.403
SC (↑)	1.0	0.997	0.801	0.997
CH (↑)	1.096e+32	19436.120	58.952	39839.524
DB (↓)	0.0	0.018	0.689	0.020

Tabela 4.9: Performance da clusterização no dataset Yale Face, utilizando o algoritmo SMVSC.

Como podemos observar na tabela 4.9, a performance do algoritmo SMVSC, considerando todas as vistas, obteve resultados semelhantes ao algoritmo LMVSC para as métricas *accuracy*, *nmi*, *purity* e *f-score*, contudo a diferença está na composição dos *clusters*. Enquanto, o algoritmo LMVSC produz um *cluster* com cerca de 40% da amostra, este algoritmo produz 15 segmentos, com um *silhouette score* máximo, igual a 1. Este valor permite-nos afirmar que a segmentação obtida pela aplicação do algoritmo *k-means* à matriz *U*, que provém da decomposição de valores singulares da matriz *Z*, é perfeita, ou seja, os extitclusters obtidos são tão compactos que a distância média entre dois pontos no

mesmo *cluster* é 0. Apesar de este algoritmo ter uma performance superior à do algoritmo LMVSC, principalmente nas classes 4 a 12, como se pode observar na figura B.5(a), este algoritmo, em vez de perder performance, criando um cluster com uma grande parte da amostra, perde performance, aglomerando duas classes distintas numa só, como é o caso das classes 13 e 14, que são confundidas com as classes 8 e 7, respetivamente, bem como a classe 3, que contém imagens de todas as classes. Igualmente, no caso do algoritmo SMVSC, conclui-se, através da experiência realizada, considerando cada vista como um conjunto de dados com uma única vista, que a melhor performance deste algoritmo ocorre quando considerada, apenas, a vista 3, com os atributos do filtro de Gabor, para clusterizar as caras presentes neste conjunto de dados. Em suma, o estudo realizado, utilizando o conjunto de dados *Yale Faces*, corrobora a consideração feita por [40] de garantir a corretude das vistas, uma vez que os algoritmos de clusterização multi-view, ao utilizarem todas as vistas disponíveis para produzir a clusterização final, torna-se essencial endereçar ou mitigar este problema, de modo a garantir a qualidade da clusterização obtida.

4.6 Conclusões

Com a proliferação de dados com múltiplas vistas torna-se necessário a criação e utilização de algoritmos de clusterização que consigam extrair conhecimento destes conjuntos de dados [40]. Estes algoritmos, segundo os autores de [39] podem ser categorizados em três categorias: co-training, multi-kernel learning, multi-view subspace clustering. Os autores de [40] adicionam a categoria de algoritmos de clusterização multi-task multi-view. A categoria de multi-view graph clustering propostas por [40] não foi contemplada.

Os principais desafios de multi-view clustering são: **o garante da correção das vistas** [40] uma vez que dado que todas as vistas são utilizadas para aumentar a performance da clusterização, conter vistas incorretas pode ser prejudicial. Alguns métodos utilizam pesos para cada vista, o que pode colmatar o problema, como é o caso do **SMVSC**. Ainda assim, os erros podem ser propagados para outras vistas, como no caso de algoritmos baseados em estratégias de *co-training* 4.3.1.

Momento oportuno para combinar as vistas [40] a estratégia utilizada pode ser uma de três tipos, combinação a priori, combinação intermédia, e fusão a posteriori. Para isso, vai ser estudada uma combinação à priori, bem como experiências utilizando os algoritmos de **MvC** aplicados a cada vista individualmente, para comparar os resultados considerando cada vista individualmente com os resultados considerando o caso multi-view, demonstrando a superioridade de considerar as diversas vistas com um algoritmo de **MvC**.

MvC em larga escala [43], nomeadamente **MVSC**, uma vez que a complexidade pelo menos cúbica restringe a aplicação destes algoritmos a conjuntos de dados reais, pelo que se procurou utilizar algoritmos de **MVSC** em tempo linear, utilizando estratégias baseadas em âncoras.

MvC incompleto [40] quando existem falhas de dados o que é frequente em aplicações reais, ainda assim a investigação nesta área não tem sido extensa. Este desafio coloca-se quando se considera clientes com um de dois tipos de serviço, internet ou TV, ou os dois simultâneo, sendo que alguns clientes não tem informação para algumas variáveis quer por não contratualizarem um destes serviços quer por potenciais falhas de recolhas destes mesmos dados. Por não serem estudadas este tipo de abordagens, considera-se que nestes casos que o cliente não teve utilização do serviço cujas variáveis estão em falta, verificando-se na análise dos segmentos de foi uma decisão apropriada.

Multi-task multi-view clustering [40] apresenta desafios como explorar a relação entre as diferentes tarefas e as diferentes vistas e como transferir conhecimento entre as outras vistas. A sua relevância para um problema de segmentação de cliente prende-se com o facto de procurarmos explicar vários fatores com uma mesma segmentação, a propensão a reclamações sobre o serviço de TV, a propensão a reclamações sobre o serviço de internet e a propensão ao pagamento de determinados **ARPU**. Cada um destes fatores poderia ser uma tarefa considerando estes modelos, mas dada a complexidade temporal e o escalamento de algoritmos de **MvC** procurou-se não o fazer, reservando estas

tarefas para caracterização dos segmentos criados resultantes de **MvC**.

Modelos sequenciais MvC [43]. Os dados sequenciais são muito comuns e estes também podem ter dados com múltiplas vistas e o desenvolvimento de algoritmos que suportem estes mesmos dados ainda é um problema em aberto. Apesar desta abordagem ser interessante para estudar a evolução dos segmentos, uma vez que o comportamento do cliente e a seu consequente segmento não é estático, por isso demonstra-se interessante não só saber o segmento atual como os segmentos seguintes e a propensão que os clientes de um certo segmento têm em transitar para outro segmento.

SEGMENTAÇÃO DE CLIENTE COM APRENDIZAGEM MULTI VISTA

Neste capítulo, são apresentados os resultados da aplicação de algoritmos de aprendizagem multi-vista, nomeadamente os algoritmos de *subspace clustering* baseadas em âncoras [LMVSC](#) e [SMVSC](#). Analogamente, as experiências realizadas com conjuntos de dados populares da literatura com algoritmos multi vista e as experiências realizadas com o algoritmo *k-means*, serão realizadas experiências para a aplicação de segmentação de cliente, considerando, primeiramente cada vista de forma individual e, em seguida, comparar os resultados obtidos considerando as várias vistas em simultâneo.

5.1 Resultados considerando uma única vista

5.1.1 Vista Demográfica

Observando a figura [5.1](#), conclui-se que o melhor k para a experiência que considera o algoritmo [LMVSC](#) é 4 e que o melhor k para a experiência com o algoritmo [SMVSC](#) é 18, considerando dados demográficos.

Considerando o algoritmo [LMVSC](#), observa-se que o algoritmo coloca as âncoras baseado, exclusivamente, em fatores demográficos ignorando a idade do cliente titular. O primeiro segmento é caracterizado pelo um baixo índice de envelhecimento e a menor percentagem de clientes sem nível de escolaridade, o segundo segmento é caracterizado por um elevado índice de envelhecimento e poder de compra, o terceiro segmento por uma elevada percentagem de clientes sem nível de escolaridade e baixo poder de compra, o quarto segmento tem um elevado índice de envelhecimento, poder de compra médio, mas uma grande quantidade da população sem nível de educação ou com o segundo ciclo, o quinto segmento representa aquele que tem a maior percentagem de famílias com três ou mais elementos. Mesmo assim, o algoritmo [LMVSC](#) produz resultados satisfatórios o que corrobora a afirmação feita em [4.5.1](#) de que este algoritmo é capaz de produzir resultados adequados quando existe, naturalmente, separabilidade entre os vários segmentos. No caso de uma segmentação considerando, unicamente, a [VD](#), esta segmentação existe

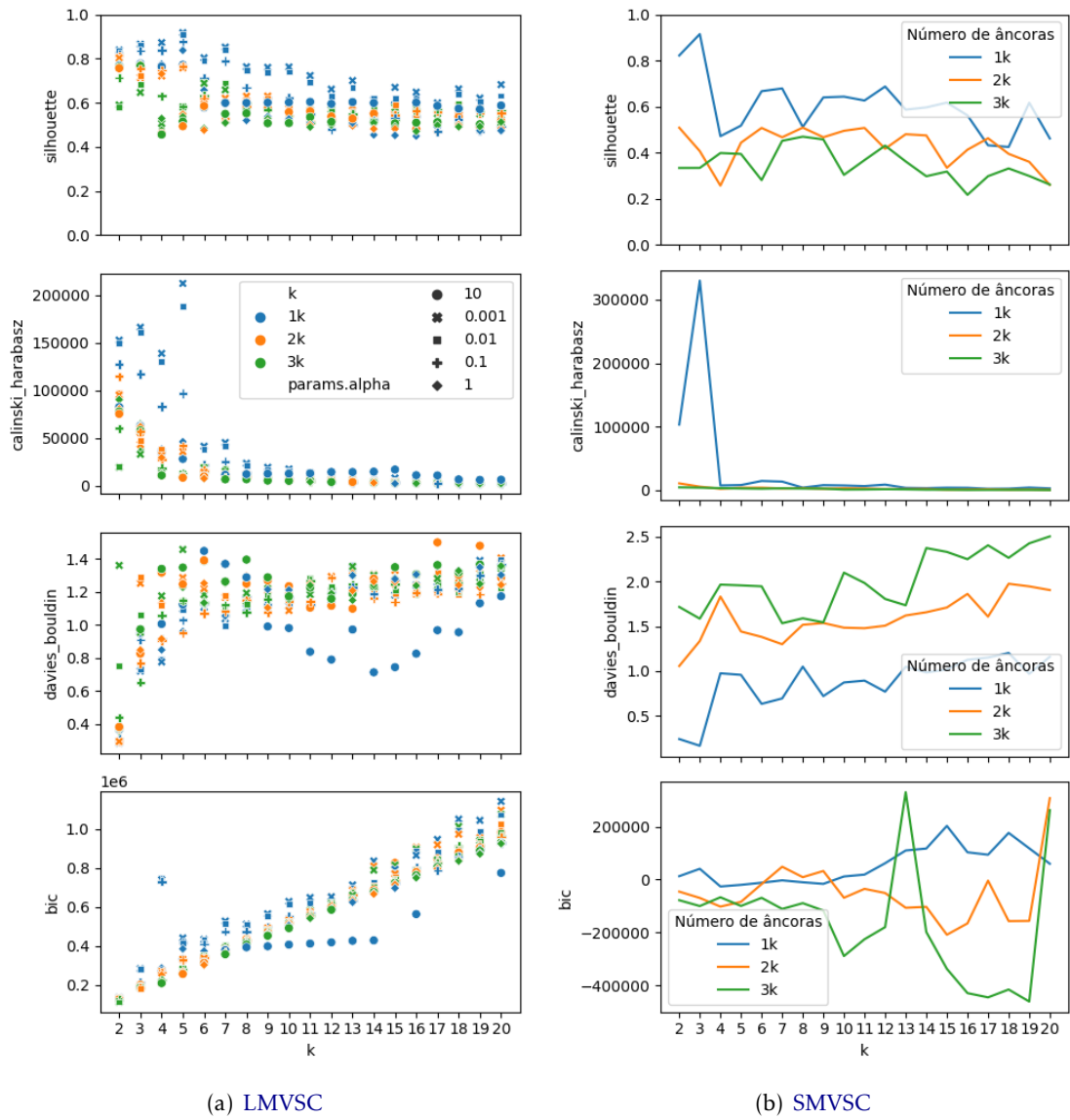


Figura 5.1: Comparação das métricas de validação internas, considerando os algoritmo LMVSC e SMVSC para todos os clientes, considerando, apenas, dados demográficos.

uma vez que todas as variáveis, com exceção da idade são iguais entre clientes do mesmo município.

Considerando o algoritmo **SMVSC**, o segmento que tem o maior número de participações técnicas de TV e NET é o segmento 14, e é caracterizado por não ser o segmento mais jovem, com uma propensão média a ter uma família com três ou mais elementos e uma alta percentagem da população sem nível de educação.

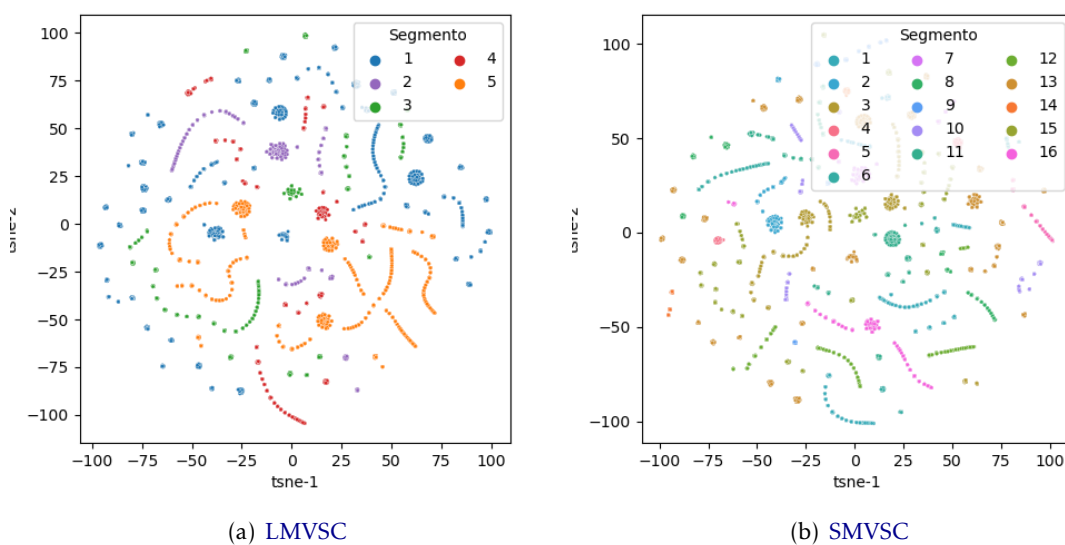
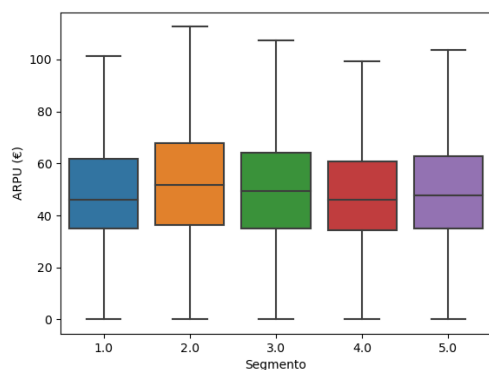


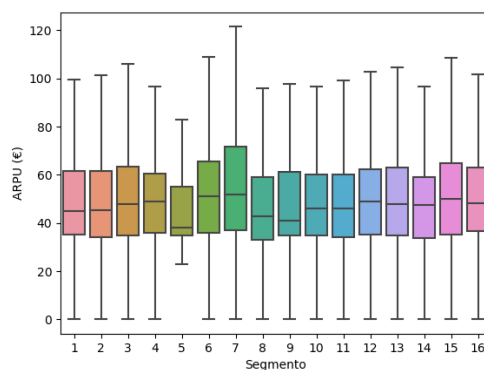
Figura 5.2: Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelos algoritmos **LMVSC** e **SMVSC** para todos os clientes da amostra, utilizando, apenas, dados demográficos.

	LMVSC					SMVSC		
Variáveis	c_1	c_2	c_3	c_4	c_5	c_4	c_8	c_{14}
age	0.663	0.658	0.644	0.665	0.625	0.012	0.002	0.331
indice _envelhecimento	0.208	1.000	0.387	0.960	0.070	0.066	0.007	-0.130
families_3+ _percentage	0.577	0.000	0.479	0.258	0.930	-0.001	-0.012	0.494
poder_compra	0.279	1.000	0.059	0.531	0.010	-0.051	-0.017	0.050
sem_nivel _escolaridade _percentage	0.000	0.520	0.933	0.436	0.414	-0.018	-0.000	0.716
2_ciclo _percentage	0.147	0.000	0.620	0.398	0.931	0.010	0.001	0.005

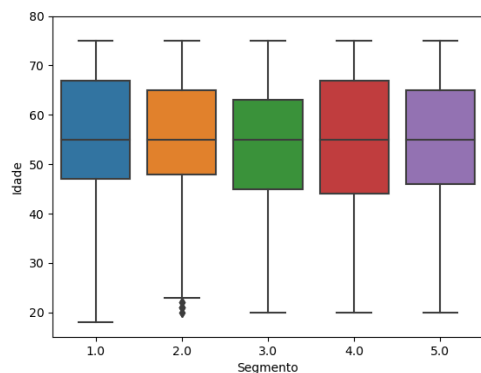
Tabela 5.1: Âncoras utilizadas pelo algoritmo **LMVSC** e **SMVSC**, utilizando dados demográficos.



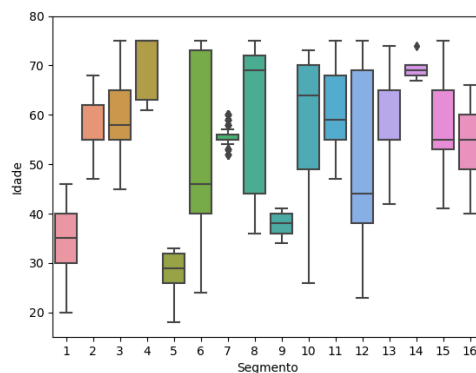
(a) LMVSC - ARPU



(b) SMVSC - ARPU



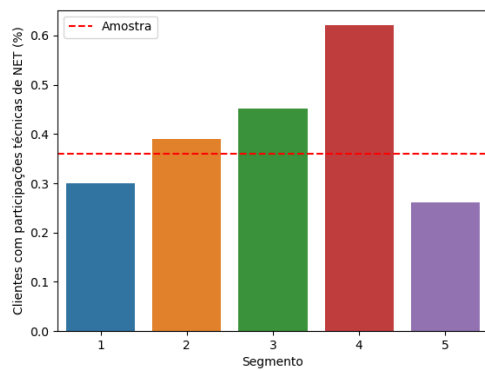
(c) LMVSC - Idade



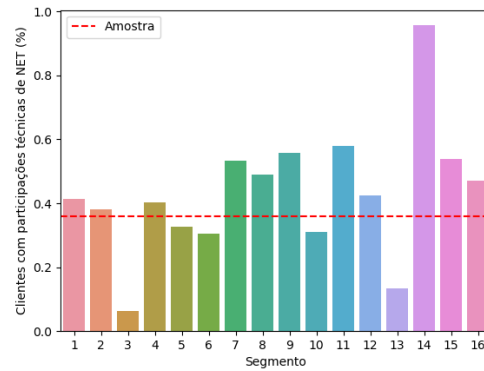
(d) SMVSC - Idade

Figura 5.3: Comparação da distribuição de **ARPU** e idade na amostra estratificada extraída, considerando, apenas, dados demográficos.

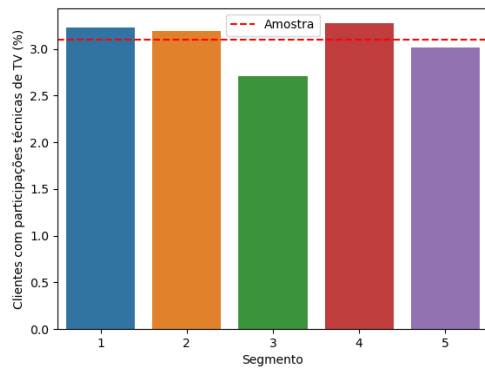
5.1. RESULTADOS CONSIDERANDO UMA ÚNICA VISTA



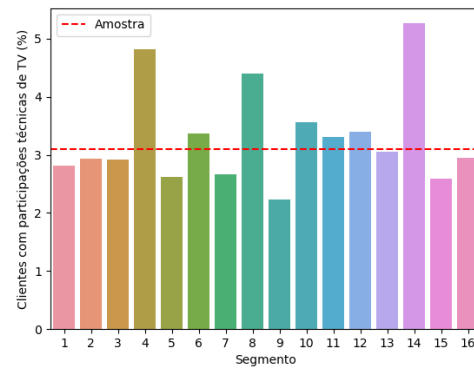
(a) **LMVSC** - Participações técnicas de NET



(b) **SMVSC** - Participações técnicas de NET



(c) **LMVSC** - Participações técnicas de TV



(d) **SMVSC** - Participações técnicas de TV

Figura 5.4: Comparação da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando, apenas, dados demográficos.

5.1.2 Vista Psicográfica

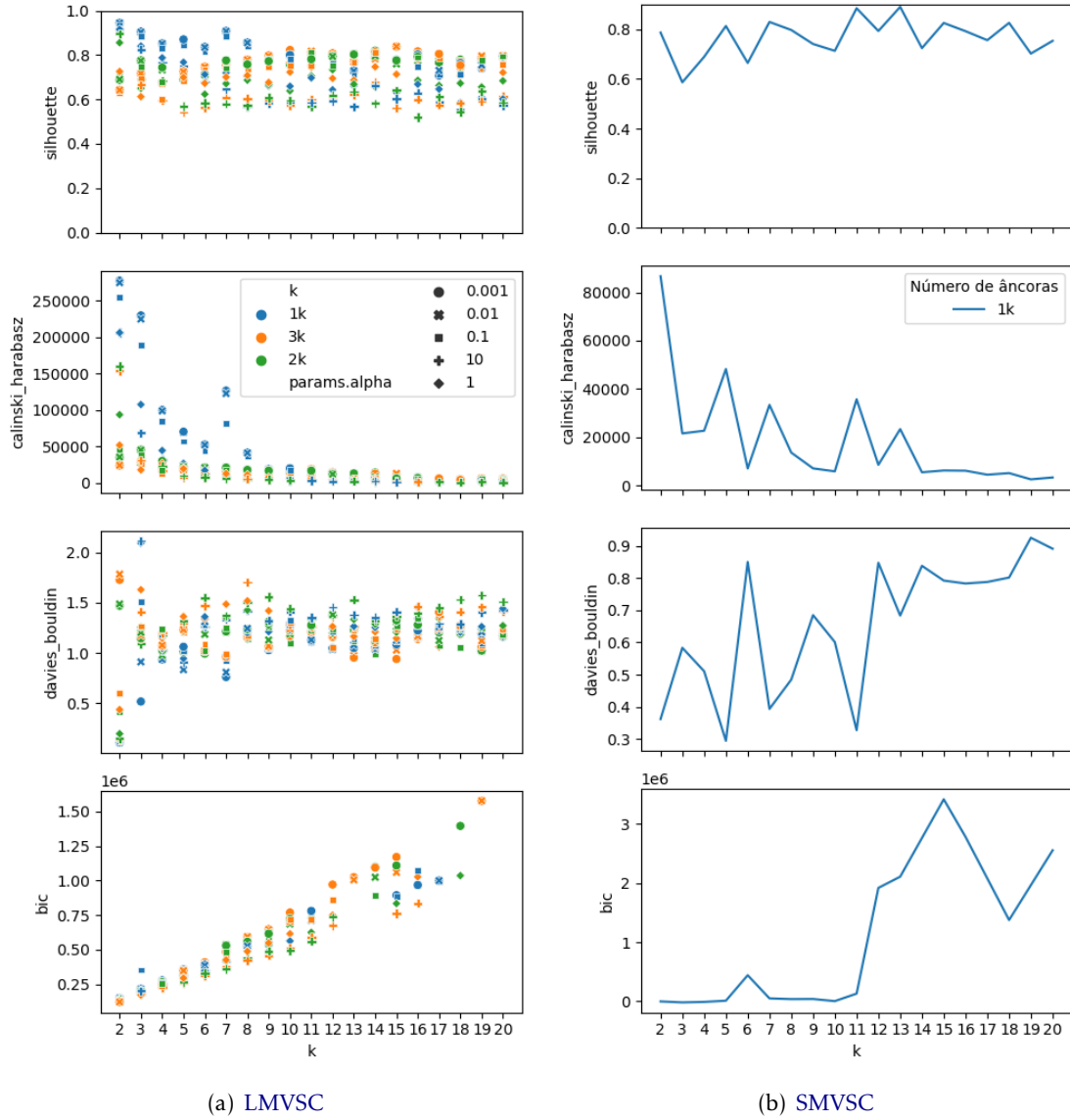


Figura 5.5: Comparação das métricas de validação internas, considerando os algoritmo LMVSC e SMVSC para todos os clientes, considerando, apenas, dados psicográficos.

Observando a figura 5.5, conclui-se que o melhor k para a experiência que considera o algoritmo LMVSC é 7 e que o melhor k para a experiência com o algoritmo SMVSC é 15, considerando dados psicográficos.

Considerando o algoritmo LMVSC observa-se que os segmentos 4 é os segmentos com mais participações técnicas de NET, enquanto o segmento é o segmento com mais participações técnicas de TV. O segmento 4 é caracterizado por uma utilização sofisticada de *internet* e contratação de TVCINE, e ainda uma parte dos clientes com débito direto (*direct_payment_flg*). O segmento 6 apesar de também ter contratação de TVCINE, não exibe uma utilização sofisticada de *internet*.

Quando consideramos o algoritmo **SMVSC**, os segmentos 5 e 15 são aqueles que tem um maior número de participações técnicas de NET, em que o segmento 5 é caracterizado por clientes que contratam Sport TV e o segmento 15 é caracterizado por clientes que utilizam protocolos sofisticados de internet como *network_admin*, *database*, *tunneling* e *file_access*. O segmento com maior percentagem de clientes com participações técnicas de TV é o segmento 14 e é caracterizado por clientes que subscrevem Benfica TV.

Estes dois segmentos reforçam que a utilização sofisticada de internet é torna um cliente mais propenso a participações técnicas de NET, enquanto a ausência de uma utilização sofisticada de internet com subscrição de canais premium torna um cliente mais propenso a participações técnicas de TV.

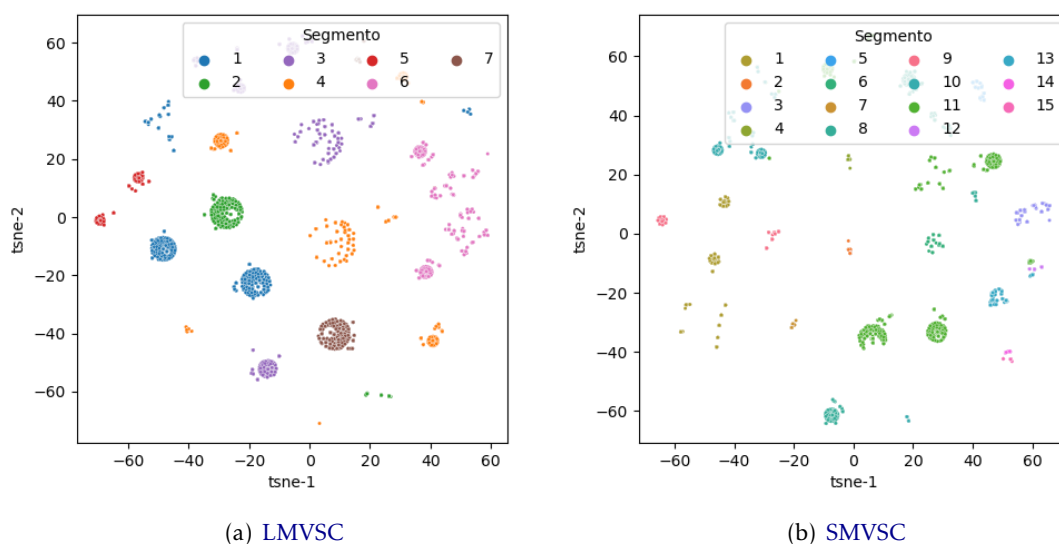
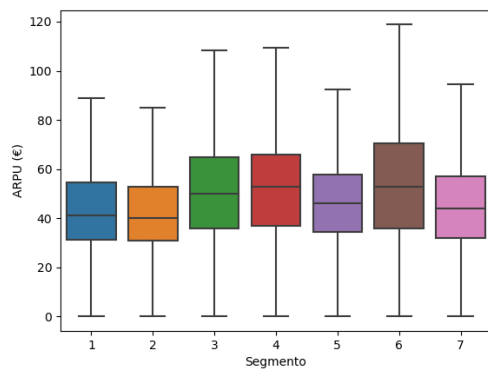


Figura 5.6: Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelos algoritmos **LMVSC** e **SMVSC** para todos os clientes da amostra, utilizando, apenas, dados psicográficos.

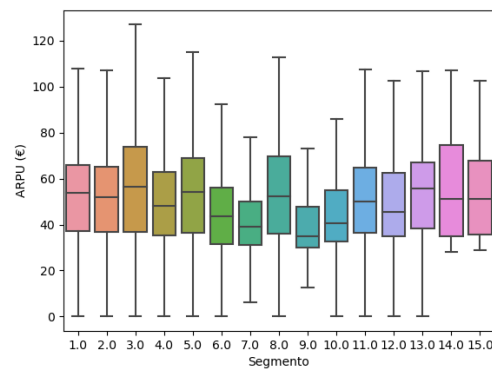
Variáveis	LMVSC		SMVSC		
	c_4	c_5	c_5	c_{14}	c_{15}
digital_invoice_flg	0.902	1.000	-0.002	-0.001	0.009
direct_payment_flg	0.451	0.000	0.016	0.011	0.027
no_marketing_flg	0.051	0.062	-0.018	-0.025	-0.045
tv_cine_qty	0.148	0.123	-0.098	-0.128	-0.100
benfica_tv_qty	0.005	0.006	-0.356	0.872	-0.085
sport_tv_qty	0.007	0.008	0.832	0.137	-0.142
other_premium_qty	0.006	0.004	-0.134	-0.070	-0.189
remote_connectivity_flg	0.209	0.000	-0.048	-0.035	-0.277
network_admin_flg	1.000	0.000	0.002	0.011	0.482

online_games_flg	0.950	0.166	-0.017	-0.018	0.200
database_flg	0.758	0.000	0.049	0.057	0.496
tunneling_flg	0.998	0.000	-0.039	-0.029	0.440
file_access_flg	0.959	0.026	-0.048	-0.054	0.333
voip_flg	1.000	1.000	0.026	0.032	0.092

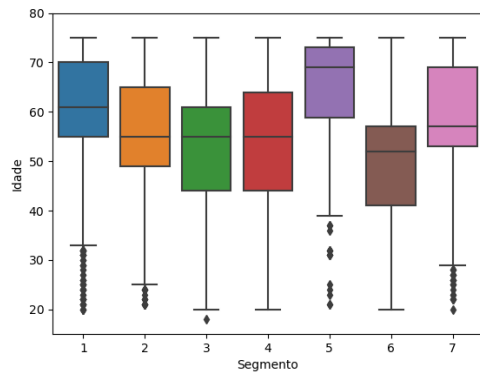
Tabela 5.2: Âncoras utilizadas pelo algoritmo **LMVSC** e **SMVSC**, utilizando dados psicográficos.



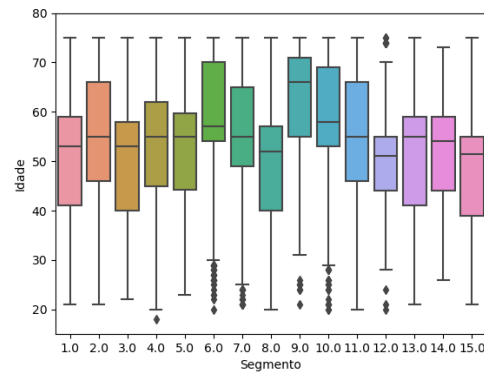
(a) LMVSC - ARPU



(b) SMVSC - ARPU

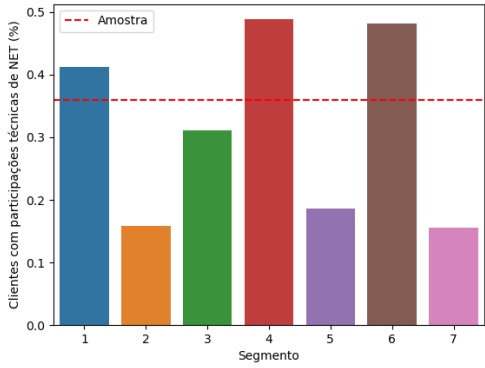


(c) LMVSC - Idade

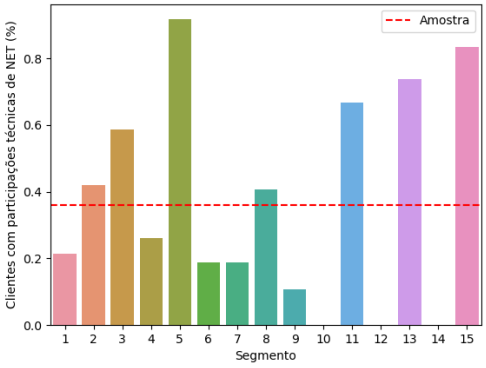


(d) SMVSC - Idade

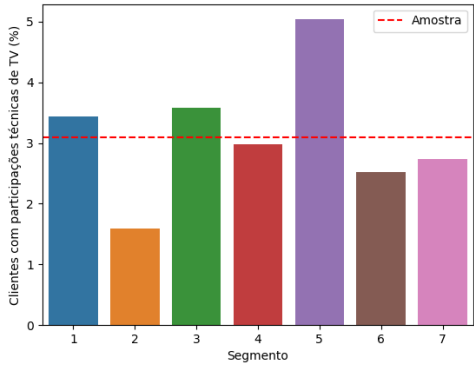
Figura 5.7: Comparação da distribuição de **ARPU** e idade na amostra estratificada extraída, considerando, apenas, dados psicográficos.



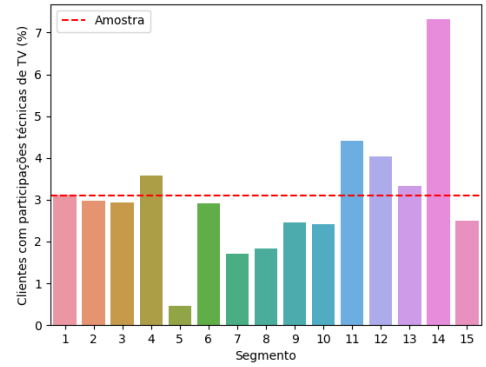
(a) *LMVSC* - Participações técnicas de NET



(b) *SMVSC* - Participações técnicas de NET



(c) *LMVSC* - Participações técnicas de TV



(d) *SMVSC* - Participações técnicas de TV

Figura 5.8: Comparação da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando, apenas, dados psicográficos.

5.1.3 Vista Comportamental

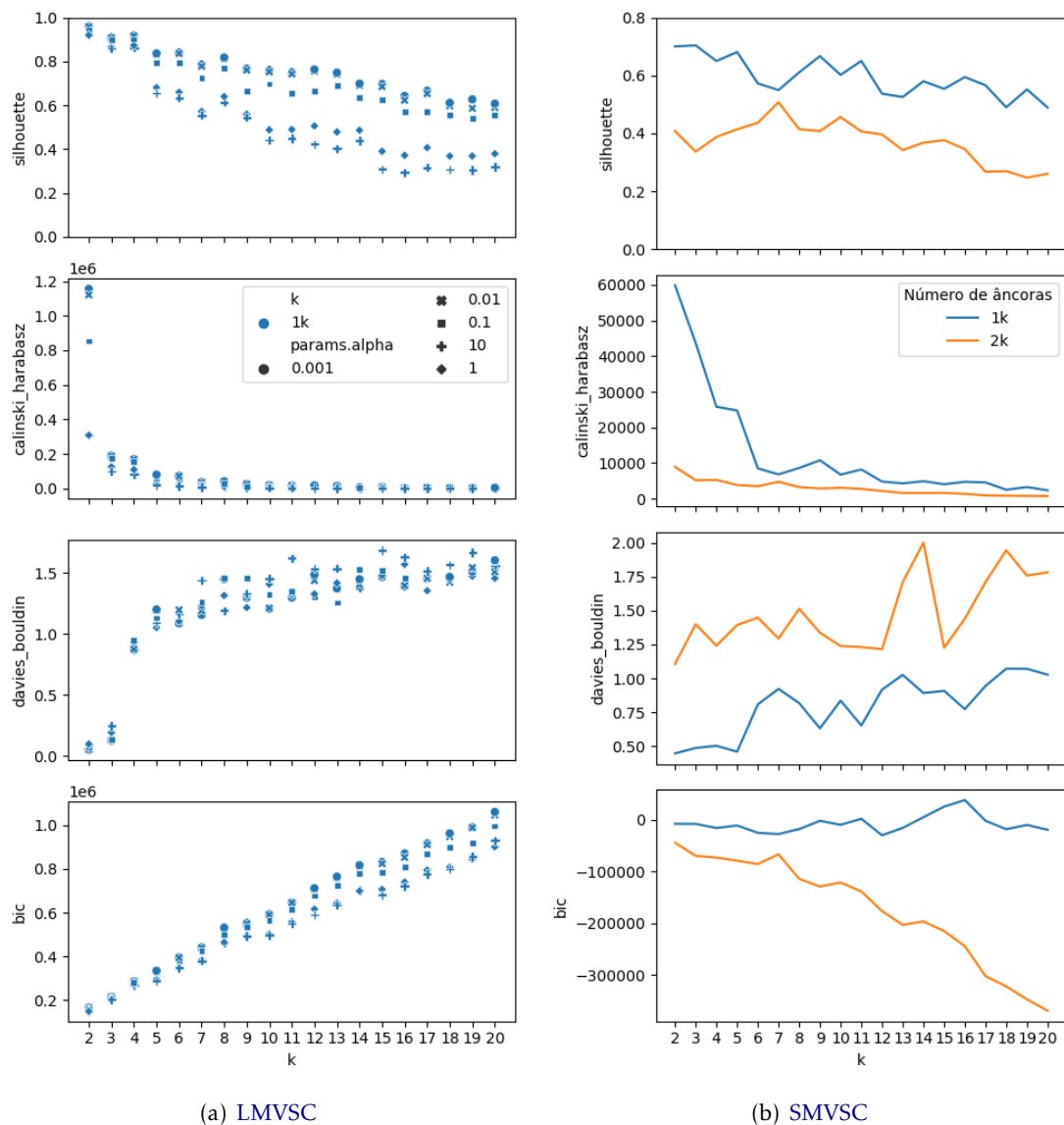


Figura 5.9: Comparação das métricas de validação internas, considerando os algoritmos LMVSC e SMVSC para todos os clientes, utilizando, apenas, dados comportamentais.

Observando a figura 5.9, conclui-se que o melhor k para a experiência que considera o algoritmo LMVSC é 4 e que o melhor k para a experiência com o algoritmo SMVSC é 18, considerando dados demográficos, psicográficos e comportamentais.

Considerando o algoritmo LMVSC, observa-se que mais uma vez, o algoritmo coloca as âncoras com base na contratação ou não de uma *box* de TV e da utilização de NET. O segmento 3 isola os clientes que não têm utilização de NET, enquanto o segmento 1 isola os clientes que não contratualizam serviço de TV, o segmento 2 contém os restantes clientes.

CAPÍTULO 5. SEGMENTAÇÃO DE CLIENTE COM APRENDIZAGEM MULTI VISTA

Considerando o algoritmo **SMVSC**, observa-se que os segmentos com maior percentagem de clientes com participações técnicas de NET são os segmentos 2 e 3, e caracterizam-se por terem clientes com muito tempo desde que celebraram o seu contrato e clientes com um período de fidelização inferior ao período usual de 2 anos. Os segmentos com maior percentagem de clientes com participações técnicas de TV são os segmentos 4 e 15 e são caracterizados por terem uma grande quantidade de *routers* e uma elevada quantidade de *powerboxes*, respetivamente.

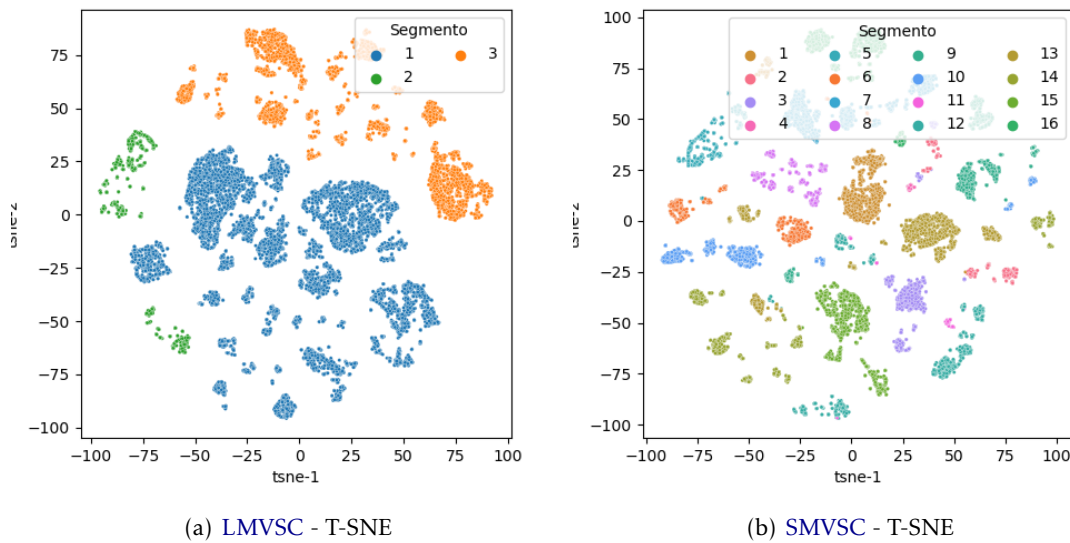


Figura 5.10: Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelos algoritmos **LMVSC** e **SMVSC** para todos os clientes da amostra, utilizando, apenas, dados comportamentais.

	LMVSC			SMVSC			
Variáveis	c_1	c_2	c_3	c_2	c_3	c_4	c_{16}
devices_qty	0.115	0.104	0.026	-0.209	-0.186	-0.049	0.098
phones_number_qty	0.500	0.486	0.457	0.001	-0.061	0.024	0.039
router_qty	0.160	0.164	0.094	0.271	-0.015	0.447	-0.029
phone_qty	0.003	0.002	0.000	-0.062	-0.043	-0.050	-0.056
hub_qty	0.164	0.123	0.235	-0.204	-0.034	-0.353	0.045
plumes_qty	0.009	0.010	0.002	-0.097	-0.047	-0.116	-0.058
powerbox_qty	0.021	0.017	0.008	-0.096	0.047	-0.000	0.973
tv_qty	0.087	0.085	0.012	-0.054	0.128	0.018	0.045
down_traffic_1w_avg	0.016	0.019	0.000	-0.065	-0.038	-0.199	-0.009

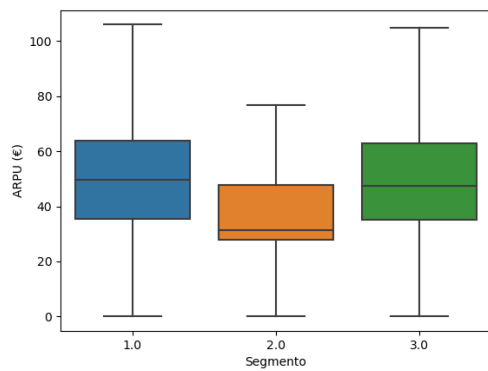
5.1. RESULTADOS CONSIDERANDO UNA ÚNICA VISTA

up_traffic _1w_avg	0.003	0.003	0.000	-0.040	-0.013	-0.089	-0.009
tv_sessions_qty	0.027	0.027	0.000	-0.102	-0.029	-0.075	-0.023
weekend_tv _sessions_perc	0.207	0.213	0.000	-0.047	-0.024	-0.084	-0.012
weekday_tv _sessions_perc	0.477	0.479	0.000	0.120	-0.013	0.050	0.010
holiday_uses _5g_flg	1.000	1.000	0.000	0.008	0.060	0.187	0.039
holiday_uses _24g_flg	1.000	1.000	0.000	0.008	0.060	0.187	0.039
holiday_uses _eth_flg	1.000	1.000	0.000	0.008	0.060	0.187	0.039
weekday_uses _5g_flg	0.877	0.876	0.000	0.073	-0.004	-0.348	-0.004
weekday_uses _24g_flg	0.960	0.955	0.000	0.032	0.026	0.062	0.024
weekday_uses _eth_flg	0.935	0.937	0.000	0.094	-0.026	-0.421	-0.028
sa_same_fiscal _num_qty	0.003	0.003	0.019	-0.008	-0.013	-0.038	-0.003
product_revolving _churn_qty	1.000	1.000	1.000	0.005	0.002	0.001	0.001
product_effective _churn_qty	0.500	0.500	0.500	-0.000	0.001	0.001	0.018
past_involuntary _churn_flg	0.000	0.000	0.000	-0.007	-0.002	-0.002	0.027
past_partial _voluntary _churn_flg	0.000	0.000	0.001	-0.017	0.005	-0.007	-0.007
past_total _voluntary _churn_flg	0.000	0.000	0.000	0.000	0.000	0.000	0.000
days_until _loyalty_ends	0.811	0.818	0.795	0.089	-0.167	0.166	0.061
days_since _last_activation	0.258	0.173	0.126	0.864	0.006	-0.266	0.129
tv_type_dsc_IRIS	0.000	0.476	0.001	-0.057	0.022	0.017	-0.003

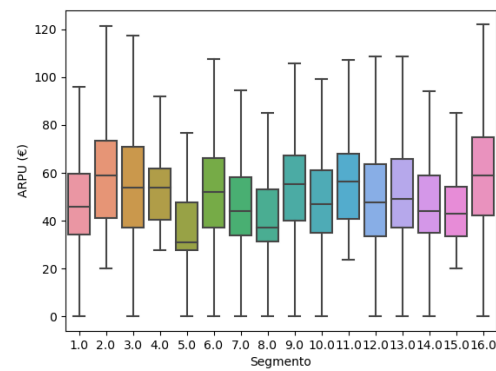
CAPÍTULO 5. SEGMENTAÇÃO DE CLIENTE COM APRENDIZAGEM MULTI VISTA

tv_type_dsc _NÃO APLICÁ- VEL	1.000	0.000	0.655	0.077	-0.012	0.176	0.008
tv_type_dsc _OUTROS	0.000	0.010	0.314	-0.040	-0.041	-0.045	-0.028
tv_type_dsc _TDC	0.000	0.000	0.002	-0.003	-0.004	-0.003	-0.002
tv_type_dsc _UMA	0.000	0.514	0.028	-0.026	0.013	-0.222	0.006
loyalty_period 2 YEARS	0.955	0.932	0.926	0.024	-0.673	0.038	0.027
loyalty_period LESS THAN 2 YE- ARS	0.019	0.020	0.020	-0.014	0.421	-0.020	-0.019
loyalty_period MORE THAN 2 YEARS	0.005	0.005	0.012	-0.004	0.189	-0.007	-0.001
loyalty_period NO LOYALTY	0.005	0.010	0.007	-0.008	0.222	-0.014	-0.008

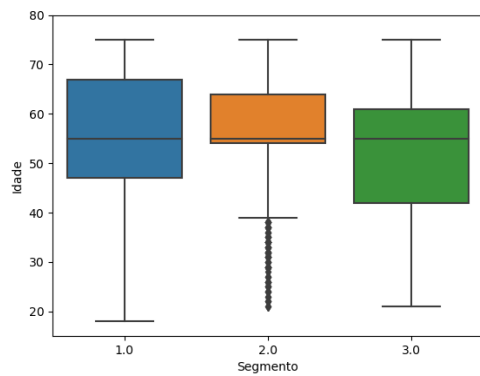
Tabela 5.3: Âncoras utilizadas pelo algoritmo **LMVSC**, utilizando dados psicográficos.



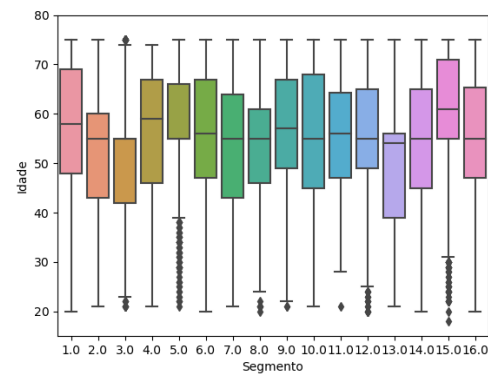
(a) LMVSC - ARPU



(b) SMVSC - ARPU

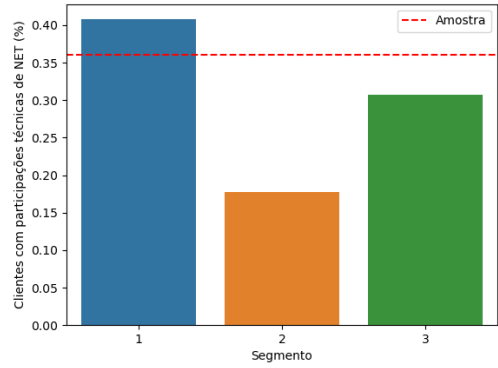


(c) LMVSC - Idade

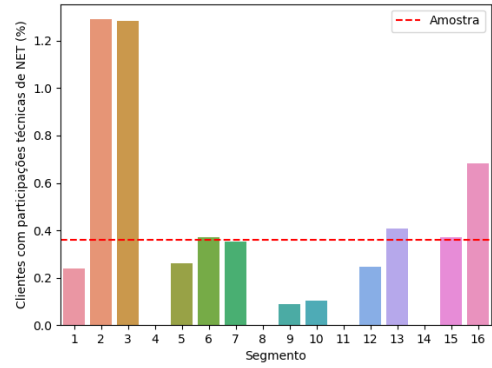


(d) SMVSC - Idade

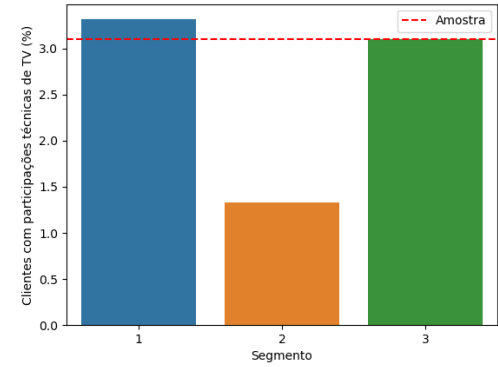
Figura 5.11: Comparação da distribuição de **ARPU** e idade na amostra estratificada extraída, considerando, apenas, dados comportamentais.



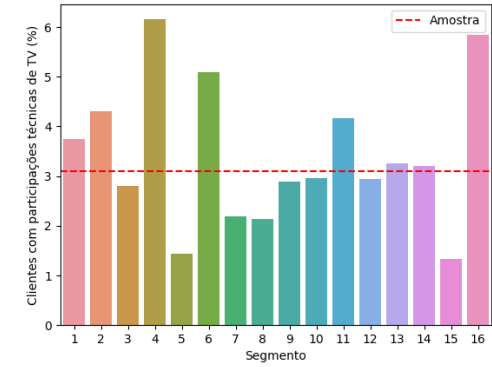
(a) *LMVSC* - Participações técnicas de NET



(b) *SMVSC* - Participações técnicas de NET



(c) *LMVSC* - Participações técnicas de TV



(d) *SMVSC* - Participações técnicas de TV

Figura 5.12: Comparação da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando, apenas, dados comportamentais.

5.2 Experiências multi vista

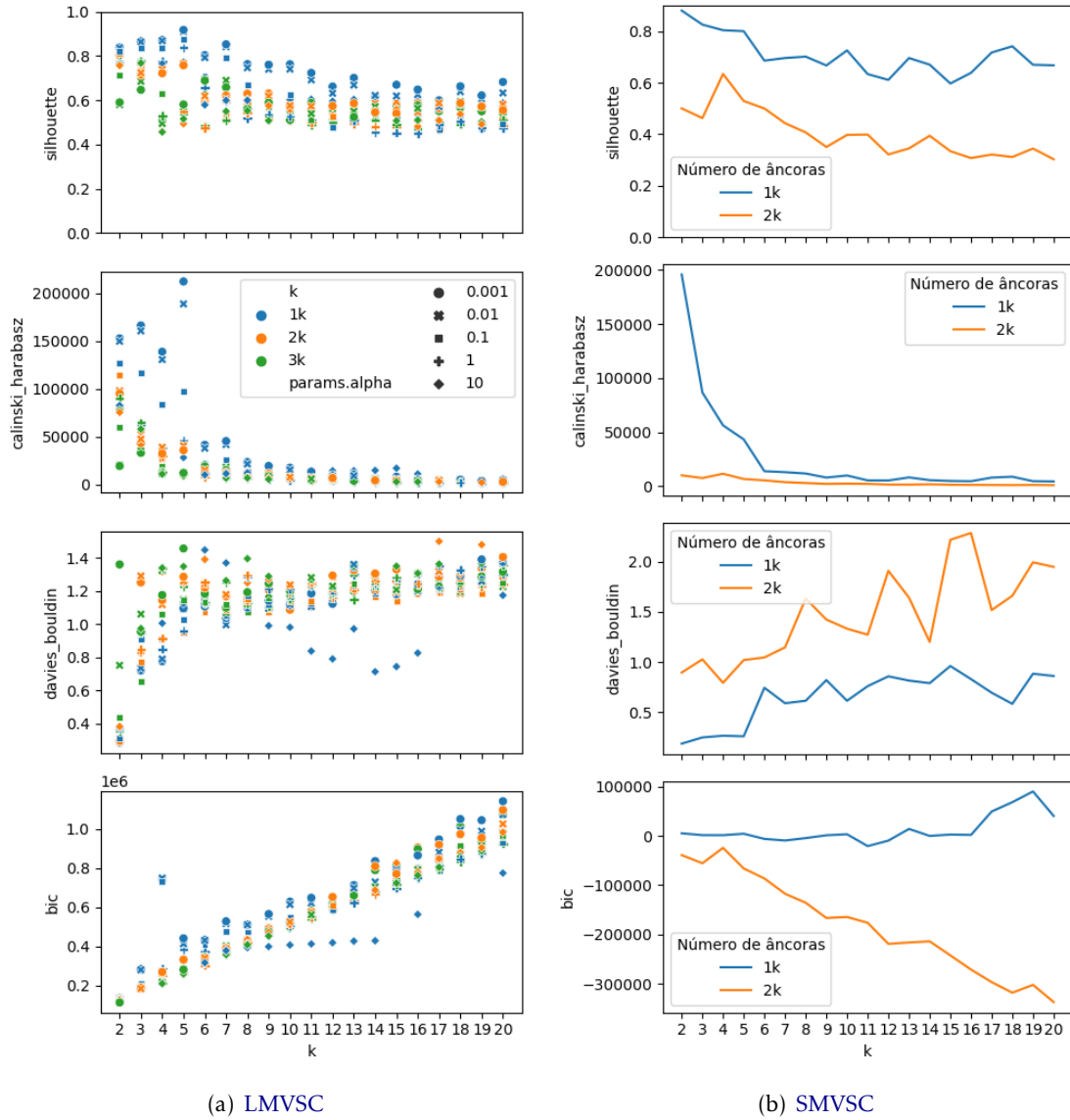


Figura 5.13: Comparação das métricas de validação internas, considerando os algoritmo LMVSC e SMVSC para todos os clientes, considerando dados demográficos, psicográficos e comportamentais.

Observando a figura 5.13, conclui-se que o melhor k para a experiência que considera o algoritmo LMVSC é 4 e que o melhor k para a experiência com o algoritmo SMVSC é 18, considerando dados demográficos, psicográficos e comportamentais.

Considerando o algoritmo LMVSC observa-se que mais uma vez, o algoritmo coloca as âncoras com base no equipamento e utilização. O segmento 3 isola os clientes que têm uma baixa utilização de *internet*, com um elevado índice de envelhecimento e poder de compra, enquanto o segmento 2 isola os clientes que não contratualizam serviço de TV, o segmento 4 isola clientes com a box UMA, um segmento com maior contratualização

CAPÍTULO 5. SEGMENTAÇÃO DE CLIENTE COM APRENDIZAGEM MULTI VISTA

de canais premium com o TVCINE e a utilização sofisticada de internet, e o o segmento 1 isola clientes com a box IRIS ou OUTROS, com baixa sofisticação da internet, excepto para o protocolo de *online games*. Este método quando considerado com vistas com informação dissonante cria uma segmentação trivial, possível de ser obtida considerando apenas uma vista, a comportamental, e não consegue explicar as variáveis alvo, idade, arput, participações técnicas de NET e TV.

Quando considerado o algoritmo **SMVSC**, os segmentos com mais participações técnicas de TV são o segmento 2 e o segmento 13, em que o segmento 2 é caracterizado por utilização de *file_access* e períodos de fidelização superior a dois anos; o segmento 13 é caracterizado por uma grande quantidade de Benfica TV e extensores de sinal de internet (plumes). O segmento 4 tem o dobro dos clientes com participações técnicas de NET e uma maior percentagem de clientes com participações técnicas TV é caracterizado por uma elevada quantidade de tráfego de *upload* (*up_traffic_1w_avg*) e de (*tunneling_flg*); segmento 11 com uma grande percentagem de participações técnicas de NET e TV é caracterizado pela utilização de do router nos feriados, que está correlacionada com a utilização de *internet* aos fins de semana.

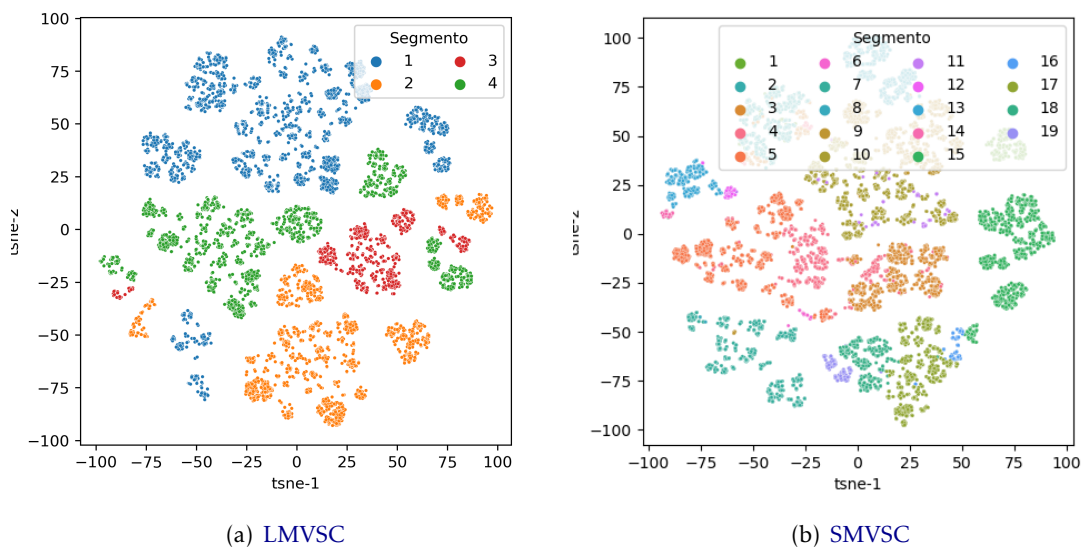


Figura 5.14: Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelos algoritmos **LMVSC** e **LMVSC** para todos os clientes da amostra, utilizando dados demográficos, psicográficos e comportamentais.

		LMVSC				SMVSC			
Pilar	Variáveis	c_1	c_2	c_3	c_4	c_2	c_4	c_{11}	c_{13}
VD	age	0.663	0.644	0.660	0.625	-0.000	0.001	0.015	-0.034
	idade _envelhecimento	0.208	0.387	0.989	0.070	0.046	-0.009	-0.039	0.065

5.2. EXPERIÊNCIAS MULTI VISTA

	families_3+_percentage	0.577	0.479	0.071	0.930	0.134	0.001	0.005	0.140
	poder_compra	0.279	0.059	0.871	0.010	-0.046	-0.005	-0.022	-0.109
	sem_nivel_escolaridade_percentage	0.000	0.933	0.497	0.414	0.033	0.003	0.035	0.013
	2_ciclo_percentage	0.147	0.620	0.109	0.931	-0.164	-0.013	-0.072	-0.223
VP	digital_invoice_flg	1.000	0.948	0.000	0.902	0.025	-0.000	-0.002	0.006
	direct_payment_flg	0.000	1.000	0.103	0.451	-0.021	-0.001	0.007	0.032
	no_marketing_flg	0.065	0.058	0.113	0.051	-0.029	-0.002	-0.011	-0.012
	tv_cine_qty	0.107	0.104	0.056	0.148	-0.026	0.025	0.030	0.056
	benfica_tv_qty	0.005	0.004	0.003	0.005	0.042	-0.001	0.005	0.878
	sport_tv_qty	0.007	0.007	0.005	0.007	0.053	0.038	0.038	-0.302
	other_premium_qty	0.004	0.004	0.002	0.006	-0.036	0.022	0.013	-0.231
	remote_connectivity_flg	0.000	0.000	0.000	0.209	0.001	0.016	-0.013	0.021
	network_admin_flg	0.000	0.001	0.000	1.000	-0.243	-0.659	-0.027	0.046
	online_games_flg	0.131	0.102	0.059	0.950	-0.065	0.006	-0.005	0.064
	database_flg	0.000	0.000	0.000	0.758	-0.162	-0.063	0.009	-0.017
	tunneling_flg	0.000	0.000	0.000	0.998	-0.347	0.704	-0.018	-0.020
	file_access_flg	0.022	0.020	0.012	0.959	0.847	0.013	0.032	-0.096
	voip_flg	0.769	0.765	0.549	1.000	-0.002	0.000	-0.022	0.046
VC	devices_qty	0.099	0.115	0.026	0.109	-0.159	-0.146	-0.157	-0.170
	phones_number_qty	0.471	0.500	0.457	0.500	-0.059	0.008	0.030	0.036
	router_qty	0.109	0.160	0.094	0.217	0.015	0.014	0.015	0.027
	phone_qty	0.002	0.003	0.000	0.001	-0.022	0.005	-0.001	-0.040
	hub_qty	0.206	0.164	0.235	0.045	-0.026	0.030	-0.062	-0.057
	plumes_qty	0.005	0.009	0.002	0.014	-0.015	-0.004	-0.161	0.955
	powerbox_qty	0.022	0.021	0.008	0.011	0.042	0.038	0.046	0.024
	tv_qty	0.081	0.087	0.012	0.089	0.099	0.078	0.119	0.101
	down_traffic_1w_avg	0.015	0.016	0.000	0.024	0.001	-0.277	-0.025	-0.031

CAPÍTULO 5. SEGMENTAÇÃO DE CLIENTE COM APRENDIZAGEM MULTI
VISTA

up_traffic _1w_avg	0.003	0.003	0.000	0.004	-0.045	0.933	-0.036	-0.057
tv_sessions_qty	0.023	0.027	0.000	0.031	-0.032	-0.004	0.004	-0.002
weekend_tv _sessions_perc	0.191	0.207	0.000	0.233	-0.037	-0.011	-0.056	-0.017
weekday_tv _sessions_perc	0.428	0.477	0.000	0.528	-0.005	0.014	-0.092	0.043
holiday_uses _5g_flg	1.000	1.000	0.000	1.000	0.004	0.010	0.249	0.030
holiday_uses _24g_flg	1.000	1.000	0.000	1.000	0.004	0.010	0.249	0.030
holiday_uses _eth_flg	1.000	1.000	0.000	1.000	0.004	0.010	0.249	0.030
weekday_uses _5g_flg	0.920	0.877	0.000	0.835	0.039	-0.019	-0.092	-0.039
weekday_uses _24g_flg	0.976	0.960	0.000	0.936	0.026	0.041	-0.825	-0.080
weekday_uses _eth_flg	0.930	0.935	0.000	0.944	-0.075	-0.089	-0.096	-0.013
sa_same _fiscal_num _qty	0.003	0.003	0.019	0.002	-0.015	-0.015	-0.040	-0.012
product_revolving _churn_qty	1.000	1.000	1.000	1.000	-0.003	-0.005	-0.004	-0.003
product_effective _churn_qty	0.501	0.500	0.500	0.500	0.047	0.003	-0.000	-0.001
past_involuntary _churn_flg	0.000	0.000	0.000	0.000	0.001	0.004	0.002	0.000
past_partial _voluntary _churn_flg	0.000	0.000	0.001	0.000	0.222	-0.004	-0.010	-0.012
past_total _voluntary _churn_flg	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
days_until _loyalty_ends	0.790	0.811	0.795	0.844	0.136	0.065	0.073	0.095
days_since _last_activation	0.216	0.258	0.126	0.132	0.100	0.081	0.036	0.110
tv_type_dsc _IRIS	0.981	0.000	0.001	0.000	-0.014	-0.004	0.017	-0.002

tv_type_dsc NÃO APLICÁVEL	0.000	1.000	0.655	0.000	-0.001	-0.010	0.033	0.001
tv_type_dsc OUTROS	0.019	0.000	0.314	0.001	0.011	-0.005	-0.133	-0.018
tv_type_dsc TDC	0.000	0.000	0.002	0.000	-0.002	-0.000	-0.010	-0.001
tv_type_dsc _UMA	0.000	0.000	0.028	0.999	0.010	0.019	0.000	0.009
loyalty_period 2 YEARS	0.910	0.955	0.926	0.954	-0.204	-0.011	0.007	0.021
loyalty_period LESS THAN 2 YEARS	0.037	0.019	0.020	0.005	-0.050	0.013	0.017	-0.003
loyalty_period MORE THAN 2 YEARS	0.005	0.005	0.012	0.006	0.903	-0.016	-0.042	-0.047
loyalty_period NO LOYALTY	0.011	0.005	0.007	0.009	-0.033	-0.012	0.002	0.010

Tabela 5.4: Âncoras utilizadas pelo algoritmo **LMVSC** e **SMVSC**, utilizando dados demográficos, psicográficos e comportamentais.

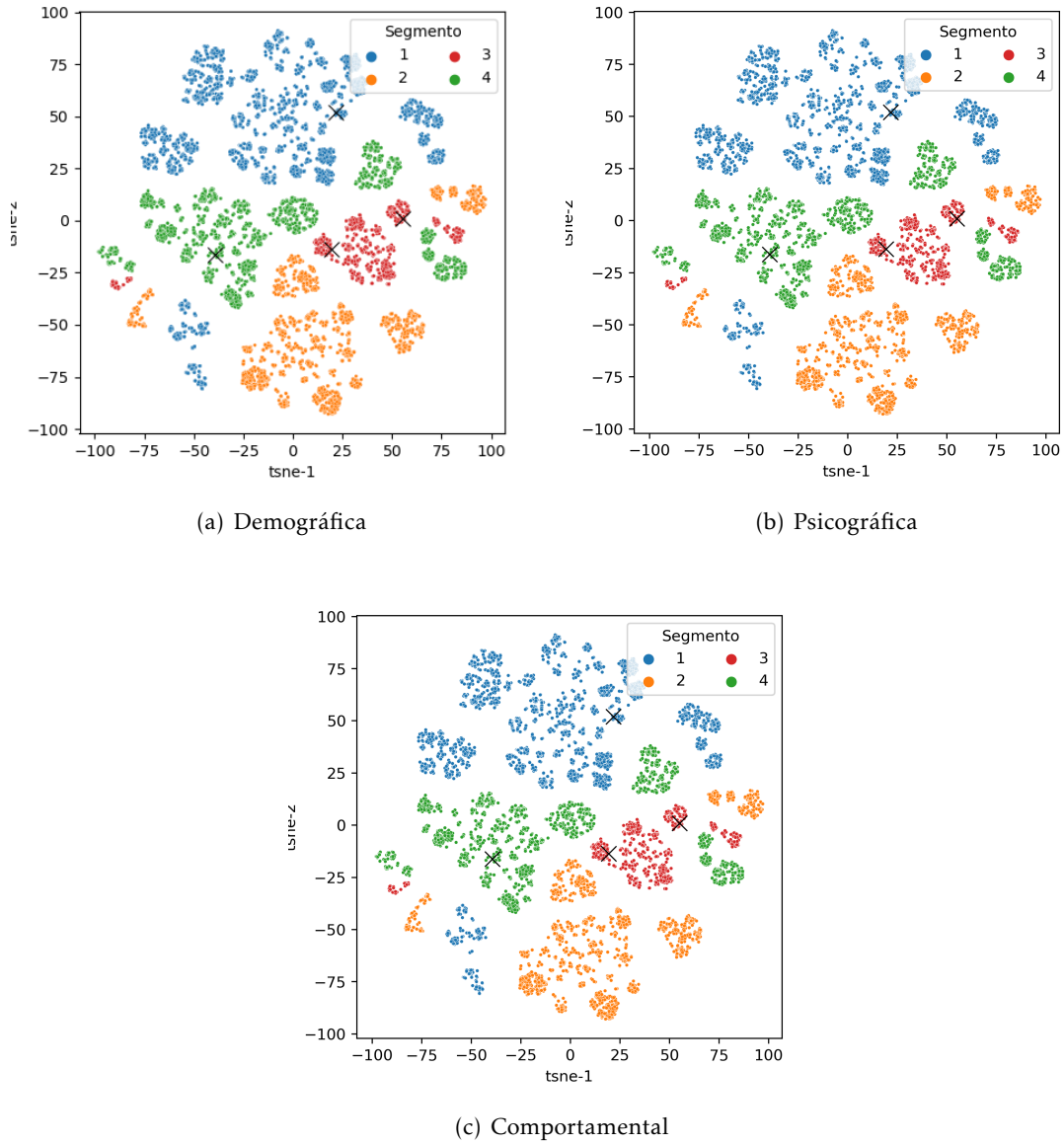


Figura 5.15: Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelo algoritmo [LMVSC](#) para todos os clientes da amostra, utilizando dados demográficos, psicográficos e comportamentais.

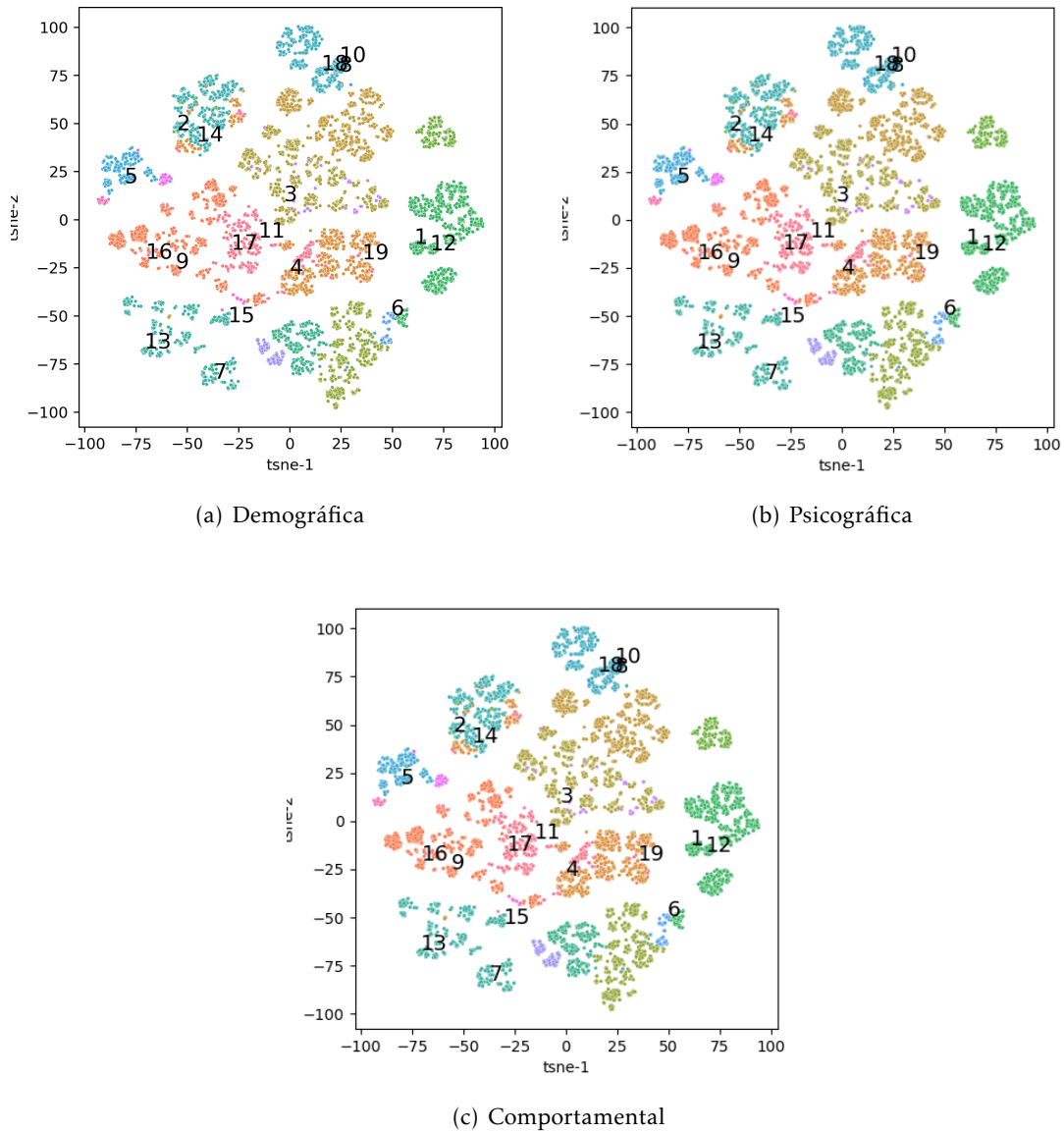
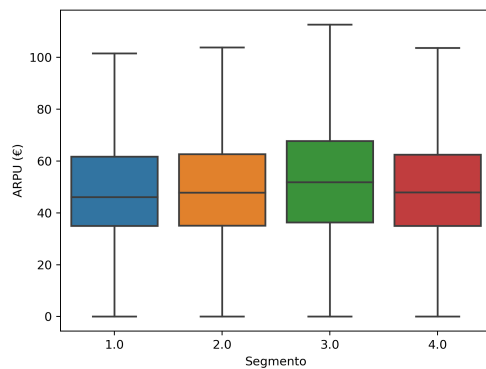
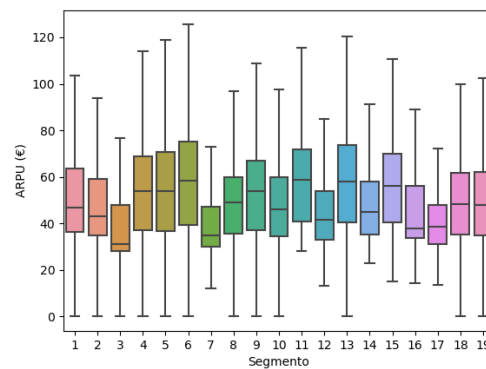


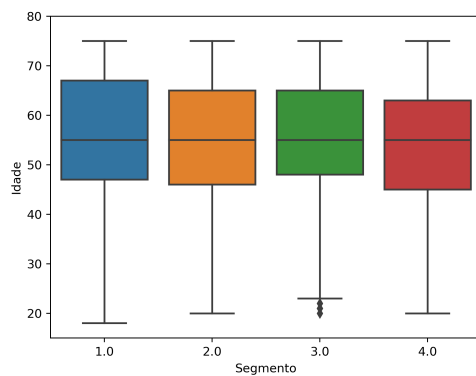
Figura 5.16: Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelo algoritmo [SMVSC](#) para todos os clientes da amostra, utilizando dados demográficos, psicográficos e comportamentais.



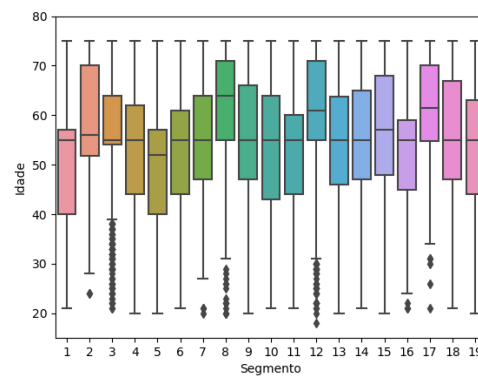
(a) LMVSC - ARPU



(b) SMVSC - ARPU

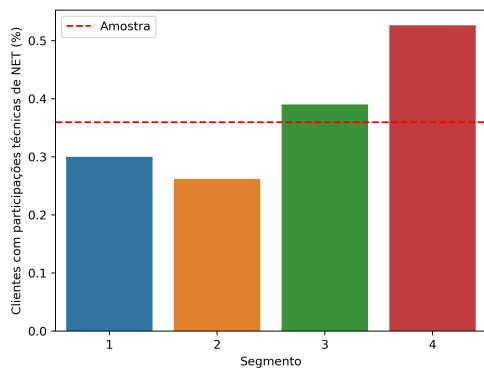


(c) LMVSC - Idade

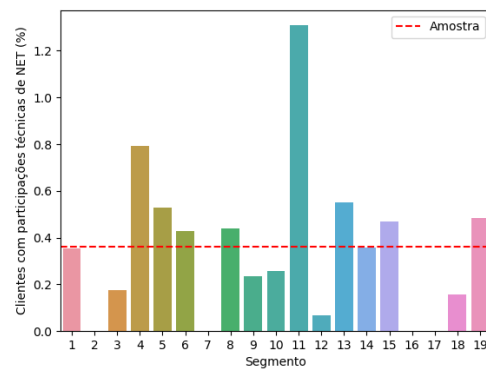


(d) SMVSC - Idade

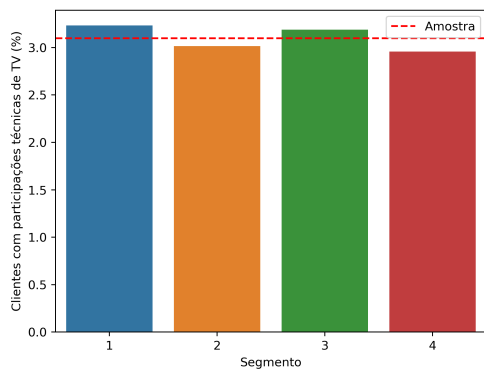
Figura 5.17: Comparação da distribuição de **ARPU** e idade, bem como da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando dados demográficos, psicográficos e comportamentais.



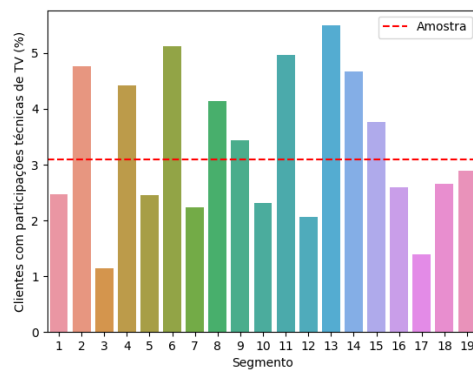
(a) LMVSC - Participações técnicas de NET



(b) SMVSC - Participações técnicas de NET



(c) LMVSC - Participações técnicas de TV



(d) SMVSC - Participações técnicas de TV

Figura 5.18: Comparação da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando dados demográficos, psicográficos e comportamentais.

5.2.1 Apenas vistas psicográfica e comportamental

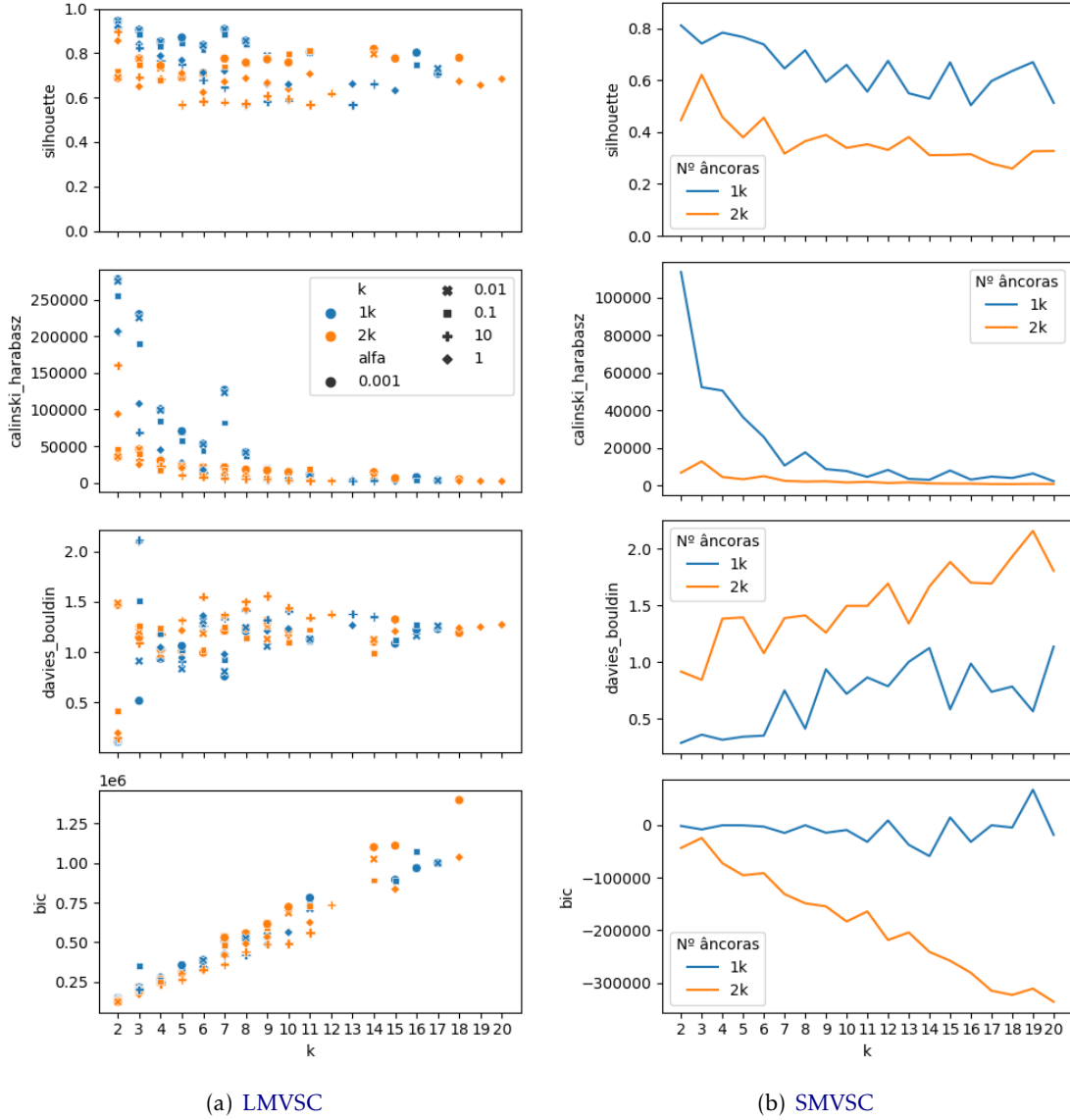


Figura 5.19: Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelos algoritmos **LMVSC** e **SMVSC** para todos os clientes da amostra, utilizando, apenas, dados psicográficos e comportamentais.

Observando a figura 5.13, conclui-se que o melhor k para a experiência que considera o algoritmo **LMVSC** é 4 e que o melhor k para a experiência com o algoritmo **SMVSC** é 18, considerando dados demográficos, psicográficos e comportamentais.

Considerando o algoritmo **LMVSC**, observa-se os mesmos resultados que quando se considera as **VD + VP + VC**, apenas explica melhor as variáveis, uma vez que a matriz de âncoras Z deste algoritmo não considera as âncoras obtidas de forma independente da **VD**, dando-lhes a mesma importância do que vistas mais discriminativas. Contudo, mais uma vez incorre numa segmentação baseada na utilização ou não utilização e por

tipo de box de TV. Este método quando utilizado vistas com informação dissonante cria uma segmentação trivial, possível de ser obtida considerando apenas uma vista, a comportamental, e não consegue explicar as variáveis alvo, idade, [ARPU](#), participações técnicas de NET e TV.

Considerando o algoritmo [SMVSC](#), observa-se que os segmentos 4 e 11 são os segmentos que tem o maior número de participações técnicas, em que o segmento 4 é caracterizado por contratação de TVCINE e uma elevada quantidade de tráfego de upload (*up_traffic_1w_avg*) e o segmento 11 é caracterizado por débito direto e utilização da rede 2.4 Gz no dias de semana. O segmento 13 é dos segmentos com maior número de participações técnicas de técnicas de TV e é caracterizada por ter uma elevada utilização do protocolo de *file_access* e boxes de TV UMA.

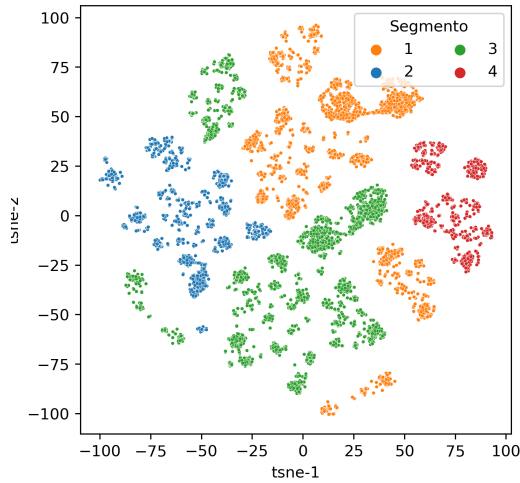
		LMVSC				SMVSC			
Pilar	Variáveis	c_1	c_2	c_3	c_4	c_2	c_4	c_{11}	c_{13}
VP	digital_invoice_flg	1.000	0.948	0.000	0.902	-0.008	0.013	0.358	0.036
	direct_payment_flg	0.000	1.000	0.103	0.451	0.008	-0.017	0.741	0.005
	no_marketing_flg	0.065	0.058	0.113	0.051	-0.016	-0.013	-0.016	-0.057
	tv_cine_qty	0.107	0.104	0.056	0.148	0.026	0.850	-0.154	-0.077
	benfica_tv_qty	0.005	0.004	0.003	0.005	0.007	-0.309	-0.168	-0.042
	sport_tv_qty	0.007	0.007	0.005	0.007	0.015	-0.029	0.049	-0.006
	other_premium_qty	0.004	0.004	0.002	0.006	0.028	-0.097	-0.103	-0.042
	remote_connectivity_flg	0.000	0.000	0.000	0.209	0.011	0.127	0.142	0.063
	network_admin_flg	0.000	0.001	0.000	1.000	-0.021	-0.050	-0.001	-0.356
	online_games_flg	0.131	0.102	0.059	0.950	-0.006	-0.078	-0.238	-0.100
	database_flg	0.000	0.000	0.000	0.758	0.003	-0.047	0.066	-0.204
	tunneling_flg	0.000	0.000	0.000	0.998	-0.023	-0.084	-0.094	-0.261
	file_access_flg	0.022	0.020	0.012	0.959	0.020	-0.086	-0.209	0.810
	voip_flg	0.769	0.765	0.549	1.000	-0.018	0.025	0.182	0.054
VC	devices_qty	0.099	0.115	0.026	0.109	-0.167	-0.148	-0.221	-0.243
	phones_number_qty	0.471	0.500	0.457	0.500	-0.005	0.014	0.029	0.067
	router_qty	0.109	0.160	0.094	0.217	0.021	0.008	0.029	0.122
	phone_qty	0.002	0.003	0.000	0.001	-0.034	0.012	0.005	-0.160
	hub_qty	0.206	0.164	0.235	0.045	-0.021	0.016	-0.159	-0.149
	plumes_qty	0.005	0.009	0.002	0.014	-0.020	-0.002	-0.099	-0.082
	powerbox_qty	0.022	0.021	0.008	0.011	0.032	0.043	0.035	0.023
	tv_qty	0.081	0.087	0.012	0.089	0.085	0.085	0.103	-0.022
	down_traffic_1w_avg	0.015	0.016	0.000	0.024	0.011	-0.285	-0.045	-0.122
	up_traffic_1w_avg	0.003	0.003	0.000	0.004	-0.054	0.931	-0.072	-0.094

CAPÍTULO 5. SEGMENTAÇÃO DE CLIENTE COM APRENDIZAGEM MULTI
VISTA

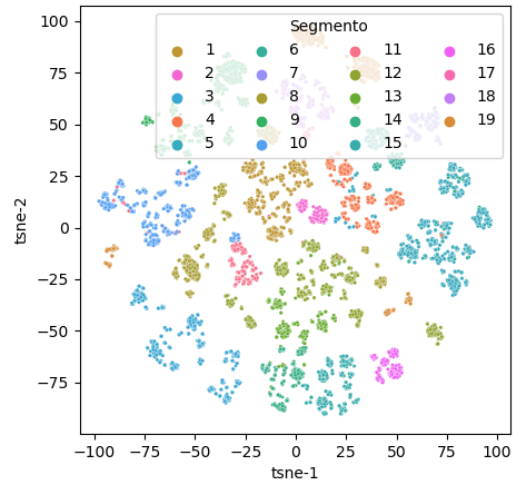
tv_sessions_qty	0.023	0.027	0.000	0.031	-0.029	-0.002	-0.055	-0.021
weekend_tv_sessions_perc	0.191	0.207	0.000	0.233	-0.032	-0.020	-0.043	-0.054
weekday_tv_sessions_perc	0.428	0.477	0.000	0.528	-0.020	-0.008	0.022	0.017
holiday_uses_5g_flg	1.000	1.000	0.000	1.000	0.026	0.009	0.091	0.027
holiday_uses_24g_flg	1.000	1.000	0.000	1.000	0.026	0.009	0.091	0.027
holiday_uses_eth_flg	1.000	1.000	0.000	1.000	0.026	0.009	0.091	0.027
weekday_uses_5g_flg	0.920	0.877	0.000	0.835	0.052	-0.001	-0.876	0.087
weekday_uses_24g_flg	0.976	0.960	0.000	0.936	0.055	0.042	0.257	0.117
weekday_uses_eth_flg	0.930	0.935	0.000	0.944	-0.061	-0.086	0.107	-0.025
sa_same_fiscal_num_qty	0.003	0.003	0.019	0.002	-0.013	-0.016	-0.018	-0.015
product_revolving_churn_qty	1.000	1.000	1.000	1.000	0.008	-0.004	0.019	0.010
product_effective_churn_qty	0.501	0.500	0.500	0.500	0.040	-0.001	0.001	-0.004
past_involuntary_churn_flg	0.000	0.000	0.000	0.000	-0.014	0.002	-0.029	-0.017
past_partial_voluntary_churn_flg	0.000	0.000	0.001	0.000	0.058	-0.006	-0.004	-0.006
past_total_voluntary_churn_flg	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
days_until_loyalty_ends	0.790	0.811	0.795	0.844	0.075	0.081	0.084	0.103
days_since_last_activation	0.216	0.258	0.126	0.132	-0.004	0.054	0.020	-0.097
tv_type_dsc_IRIS	0.981	0.000	0.001	0.000	0.019	-0.003	0.008	-0.289
tv_type_dsc_NÃO APLICÁVEL	0.000	1.000	0.655	0.000	-0.006	-0.002	0.062	-0.363
tv_type_dsc_OUTROS	0.019	0.000	0.314	0.001	-0.019	-0.006	-0.047	-0.021

tv_type_dsc _TDC	0.000	0.000	0.002	0.000	-0.001	-0.001	-0.004	-0.001
tv_type_dsc _UMA	0.000	0.000	0.028	0.999	-0.001	0.009	-0.066	0.756
loyalty_period _2 YEARS	0.910	0.955	0.926	0.954	-0.489	-0.012	0.020	0.053
loyalty_period _LESS THAN 2 YEARS	0.037	0.019	0.020	0.005	-0.139	0.017	-0.001	-0.010
loyalty_period _MORE THAN 2 YEARS	0.005	0.005	0.012	0.006	0.305	-0.011	-0.009	-0.034
loyalty_period _NO LOYALTY	0.011	0.005	0.007	0.009	-0.065	0.001	0.006	-0.000

Tabela 5.5: Âncoras utilizadas pelo algoritmo **LMVSC** e **SMVSC**, utilizando dados psicográficos e comportamentais.



(a) **LMVSC**



(b) **SMVSC**

Figura 5.20: Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelos algoritmos **LMVSC** e **SMVSC** para todos os clientes da amostra, utilizando dados psicográficos e comportamentais.

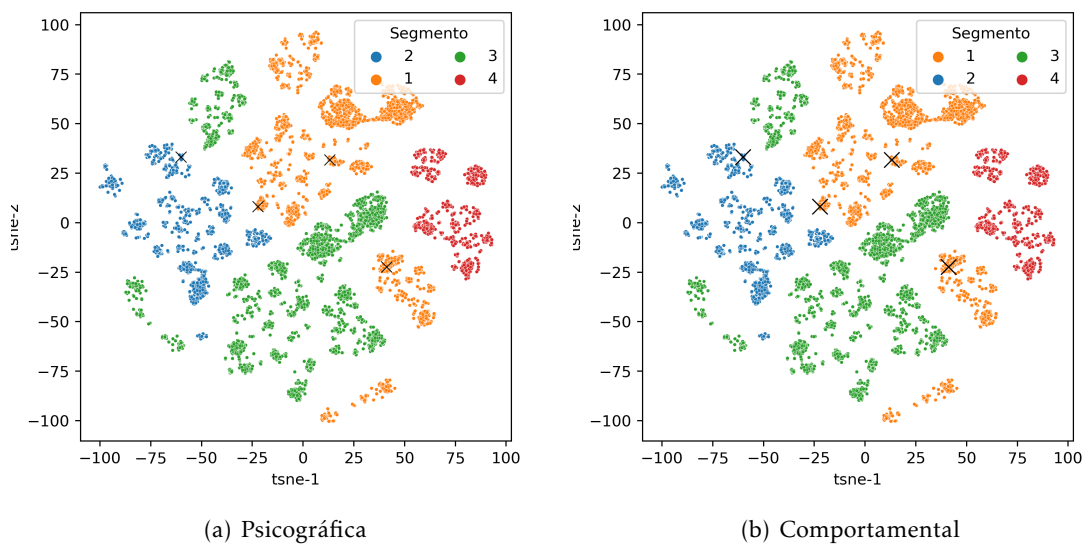
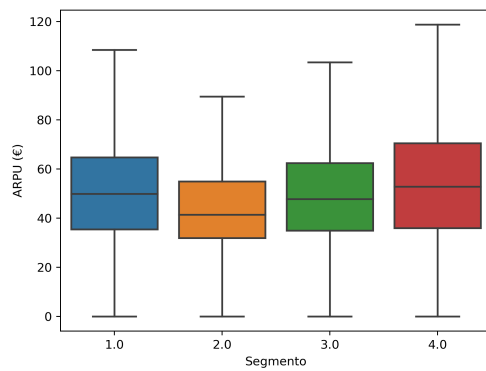
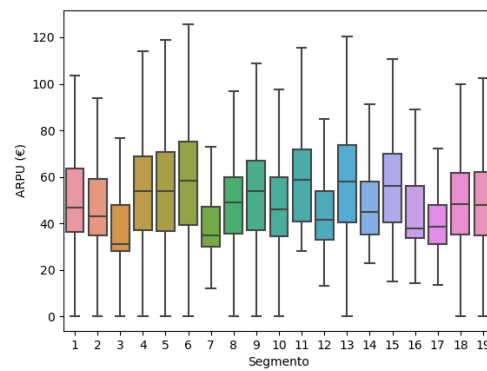


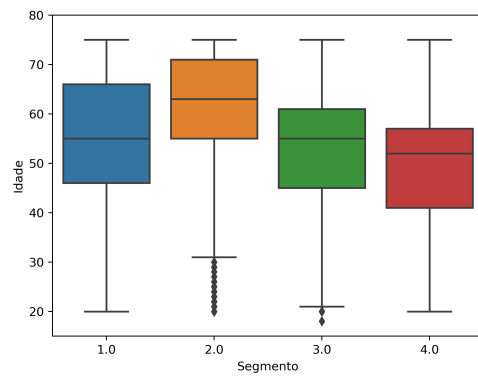
Figura 5.21: Visualização do conjunto de dados utilizando t-SNE, considerando os segmentos produzidos pelo algoritmo [LMVSC](#) para todos os clientes da amostra, utilizando dados psicográficos e comportamentais.



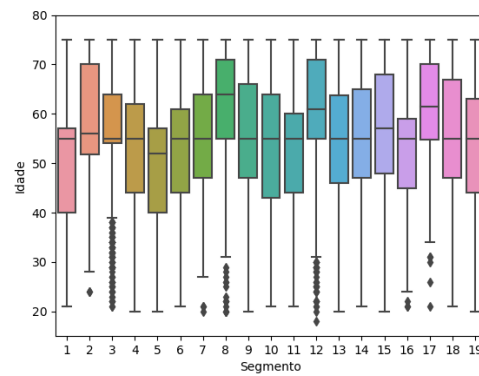
(a) LMVSC - ARPU



(b) SMVSC - ARPU

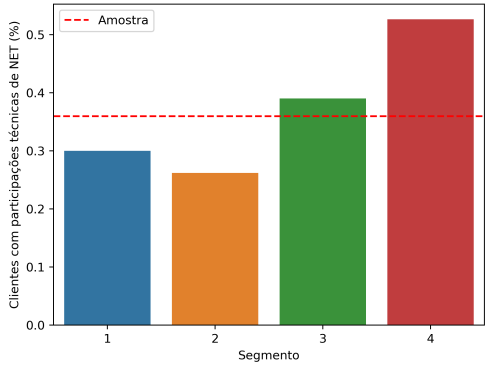


(c) LMVSC - Idade

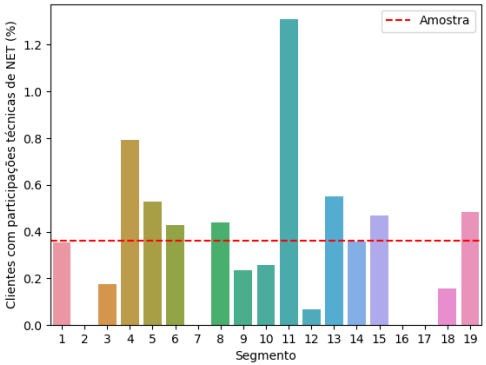


(d) SMVSC - Idade

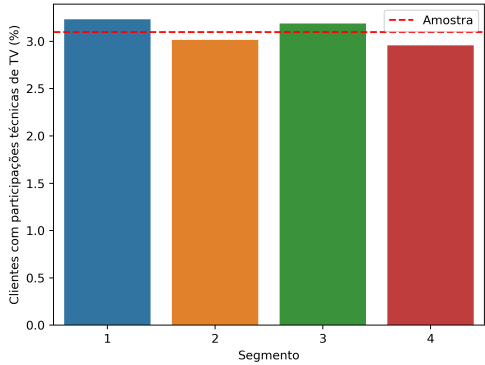
Figura 5.22: Comparação da distribuição de **ARPU** e idade na amostra estratificada extraída, considerando, apenas, dados psicográficos e comportamentais.



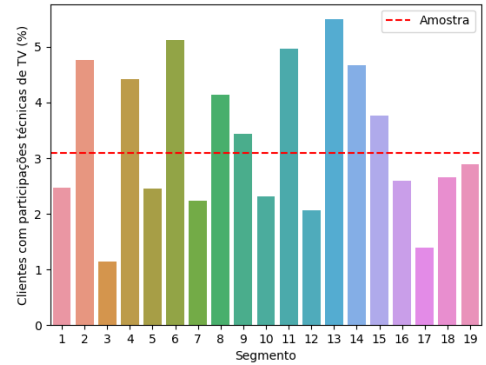
(a) LMVSC - Participações técnicas de NET



(b) SMVSC - Participações técnicas de NET



(c) LMVSC - Participações técnicas de TV



(d) SMVSC - Participações técnicas de TV

Figura 5.23: Comparação da prevalência de clientes com participações técnicas na amostra estratificada extraída, considerando dados psicográficos e comportamentais.

CONCLUSÕES E TRABALHO FUTURO

A segmentação de cliente permite dividir a base de clientes de uma empresa em segmentos que partilhem características comuns para permitir melhorar a sua experiência enquanto clientes de uma empresa de telecomunicações, personalizando o tratamento efetuado a cada segmento.

Tradicionalmente, existem quatro principais formas de efetuar uma segmentação de cliente, a segmentação geográfica, demográfica, psicográfica e comportamental e cada uma destas segmentações pode ser feita de forma individual, mas cada uma delas é apenas uma visão sobre o cliente, sendo o objetivo desta dissertação uma visão unificada sobre estas segmentações unificadas.

Para efetuar esta segmentação foi considerado uma abordagem padrão de segmentação, utilizando uma abordagem escalável e amplamente utilizada em várias aplicações em diversos domínios. Apesar de produzir uma segmentação consistente com a definição de *cluster*, são, sempre, dependentes de um conjunto reduzido de variáveis, sendo possível obter tal segmentação realizando uma segmentação à priori. A segmentação feita juntando os dados demográficos, psicográficos e comportamentais

Com este trabalho mostrou-se a importância e relevância de considerar o cliente como uma entidade que pode ser vista à luz das suas características geográficas, demográficas, psicográficas e comportamentais, sendo que estas características podem ser consideradas como diferentes vistas do mesmo objeto. Assim, a exploração de abordagens de multi-view learning são capazes de detetar segmentos bem definidos num subespaço latente.

Quando considerada uma abordagem de clusterização com um único conjunto de atributos, como no caso com *k-means*, conclui-se que, apesar, de produzirem segmentos com algum significado de negócio quando se segmenta o parque de clientes, utilizando um pilar de segmentação de cada vez, mas quando se considera mais do que um pilar de segmentação perde a capacidade de obter uma segmentação com significado.

A consideração de abordagens de aprendizagem multi vista, nomeadamente de *multi-view subspace clustering* baseadas em âncoras revela-se uma abordagem interessante para detetar segmentos de mercado com uma abordagem *data-driven*. Enquanto uma primeira

abordagem de obter as âncoras de forma independente falha em conseguir obter segmentos mais interessantes do que uma abordagem como o *k-means* (LMVSC), a abordagem em que se combina a aprendizagem das âncoras com a construção do grafo unificado, o que permite a obtenção de um subespaço mais discriminante.

Estudando os segmentos mais interessantes de acordo com uma variável alvo, conseguimos compreender que características do cliente é que as motivam.

Deste modo, demonstra-se que a segmentação de cliente também constitui uma aplicação de aprendizagem multi-vista não supervisionada.

Como trabalho futuro é proposta: a exploração de outras funções regularizadoras como [5] ou [21], a fim de melhor explorar a diversidade ou a consistência e especificidade, respetivamente; a exploração da evolução dos segmentos ao longo do tempo, uma vez que esta abordagem considera o comportamento dos clientes num mês em concreto, não tendo em consideração o comportamento do cliente ao longo do tempo e do possível significado de este poder mudar de segmento tanto de forma sazonal como definitiva e essa transição ter um significado de negócio e *multi-view clustering* com dados incompletos, uma vez que se revelou um desafio lidar com a utilização ou não utilização do serviço de internet ou de TV.

BIBLIOGRAFIA

- [1] S. P. Abney. “Bootstrapping”. Em: *Annual Meeting of the Association for Computational Linguistics*. 2002 (ver p. 48).
- [2] S. Bickel e T. Scheffer. “Multi-View Clustering”. Em: dez. de 2004, pp. 19–26. ISBN: 0-7695-2142-8. DOI: [10.1109/ICDM.2004.10095](https://doi.org/10.1109/ICDM.2004.10095) (ver p. 48).
- [3] K. Brunnström et al. *Qualinet White Paper on Definitions of Quality of Experience*. Mar. de 2013, pp. - (ver p. 1).
- [4] T. Caliński e H. JA. “A Dendrite Method for Cluster Analysis”. Em: *Communications in Statistics - Theory and Methods* 3 (jan. de 1974), pp. 1–27. DOI: [10.1080/03610927408827101](https://doi.org/10.1080/03610927408827101) (ver p. 23).
- [5] X. Cao et al. “Diversity-induced Multi-view Subspace Clustering”. Em: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 586–594. DOI: [10.1109/CVPR.2015.7298657](https://doi.org/10.1109/CVPR.2015.7298657) (ver p. 116).
- [6] T. Contributor. *What is quality of experience (QoE or Qox)? - definition from whatis.com*. Abr. de 2008. URL: <https://searchcustomerexperience.techtarget.com/definition/Quality-of-Experience-QoE-or-QoX> (ver p. 1).
- [7] N. Dalal e B. Triggs. “Histograms of oriented gradients for human detection”. Em: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Vol. 1. 2005, 886–893 vol. 1. DOI: [10.1109/CVPR.2005.177](https://doi.org/10.1109/CVPR.2005.177) (ver p. 67).
- [8] S. Dasgupta, M. Littman e D. Mcallester. “PAC Generalization Bounds for Co-training.” Em: jan. de 2001, pp. 375–382 (ver p. 46).
- [9] D. L. Davies e D. W. Bouldin. “A Cluster Separation Measure”. Em: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979), pp. 224–227. DOI: [10.1109/TPAMI.1979.4766909](https://doi.org/10.1109/TPAMI.1979.4766909) (ver p. 23).
- [10] Y. Ding et al. “Fall detection based on depth images via wavelet moment”. Em: *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. 2017, pp. 1–5. DOI: [10.1109/CISP-BMEI.2017.8302057](https://doi.org/10.1109/CISP-BMEI.2017.8302057) (ver p. 67).

- [11] S. Dolnicar, B. Grün e F. Leisch. “Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful”. Em: jul. de 2018, pp. 11–22. ISBN: 978-981-10-8817-9. DOI: [10.1007/978-981-10-8817-9_2](https://doi.org/10.1007/978-981-10-8817-9_2) (ver pp. 1–3, 5, 7, 8).
- [12] S. Dolnicar e F. Leisch. “Evaluation of structure and reproducibility of cluster solutions using the bootstrap”. Em: *Marketing Letters* 21.1 (2010), pp. 83–101. ISSN: 09230645, 1573059X. URL: <http://www.jstor.org/stable/40604700> (ver p. 7).
- [13] *Famílias Segundo OS censos: Total e por número de indivíduos*. URL: <https://www.pordata.pt/municipios/familias+segundo+os+censos+total+e+por+numero+de+indivíduos-33> (ver p. 15).
- [14] H. Gao et al. “Multi-view Subspace Clustering”. Em: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 4238–4246. DOI: [10.1109/ICCV.2015.482](https://doi.org/10.1109/ICCV.2015.482) (ver p. 3).
- [15] R. I. Haley. *Developing effective communications strategy: A benefit segmentation approach*. Wiley, 1985 (ver p. 3).
- [16] H. Hwang, T. Jung e E. Suh. “An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry”. Em: *Expert Systems with Applications* 26.2 (2004), pp. 181–188. DOI: [10.1016/s0957-4174\(03\)00133-7](https://doi.org/10.1016/s0957-4174(03)00133-7) (ver p. 9).
- [17] Z. Kang et al. “Large-Scale Multi-View Subspace Clustering in Linear Time”. Em: *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (abr. de 2020), pp. 4412–4419. DOI: [10.1609/aaai.v34i04.5867](https://doi.org/10.1609/aaai.v34i04.5867) (ver pp. 51, 52, 55, 57).
- [18] C. M. Kopp. *Brand loyalty: What you need to know*. Dez. de 2021. URL: <https://www.investopedia.com/terms/b/brand-loyalty.asp> (ver p. 13).
- [19] C. Liu e H. Wechsler. “Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition”. Em: *IEEE Transactions on Image Processing* 11.4 (2002), pp. 467–476. DOI: [10.1109/TIP.2002.999679](https://doi.org/10.1109/TIP.2002.999679) (ver pp. 67, 76).
- [20] J. M. Lourenço. *The NOVAthesis L^AT_EX Template User’s Manual*. NOVA University Lisbon. 2021. URL: <https://github.com/joaomlourenco/novathesis/raw/master/template.pdf> (ver p. ii).
- [21] S. Luo et al. “Consistent and Specific Multi-View Subspace Clustering”. Em: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (abr. de 2018). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/11617> (ver p. 116).
- [22] B. Mirkin. *Core Concepts In Data Analysis: Summarization, correlation and visualization*. Springer, 2019 (ver p. 22).

- [23] T. Ojala, M. Pietikainen e T. Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". Em: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.7 (2002), pp. 971–987. DOI: [10.1109/TPAMI.2002.1017623](https://doi.org/10.1109/TPAMI.2002.1017623) (ver pp. 67, 75, 76).
- [24] A. Oliva e A. Torralba. "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope". Em: *International Journal of Computer Vision* 42 (mai. de 2001), pp. 145–175. DOI: [10.1023/A:1011139631724](https://doi.org/10.1023/A:1011139631724) (ver p. 67).
- [25] *Poder de compra per capita*. URL: <https://www.pordata.pt/municipios/poder+de+compra+per+capita-118> (ver p. 15).
- [26] J. Ponce et al. "Dataset Issues in Object Recognition". Em: *Toward Category-Level Object Recognition*. Ed. por J. Ponce et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 29–48. ISBN: 978-3-540-68795-5. DOI: [10.1007/11957959_2](https://doi.org/10.1007/11957959_2). URL: https://doi.org/10.1007/11957959_2 (ver p. 67).
- [27] *População Residente com 15 e mais anos segundo os censos: Total e por nível de escolaridade completo mais elevado*. URL: <https://www.pordata.pt/municipios/populacao+residente+com+15+e+mais+anos+segundo+os+censos+total+e+por+nivel+de+escolaridade+completo+mais+elevado-69> (ver p. 15).
- [28] *População Residente: Total E por grandes grupos etários*. URL: <https://www.pordata.pt/municipios/populacao+residente+total+e+por+grandes+grupos+etarios-390> (ver p. 14).
- [29] *Remuneração Base Média mensal dos trabalhadores por conta de outrem: Total e Por Sexo*. URL: <https://www.pordata.pt/municipios/remuneracao+base+media+mensal+dos+trabalhadores+por+conta+de+outrem+total+e+por+sexo-235> (ver p. 15).
- [30] C. Rogers. *Why behaviour beats demographics when it comes to segmentation*. Abr. de 2019. URL: <https://www.marketingweek.com/behaviour-demographics-segmentation/> (ver p. 8).
- [31] P. Rousseeuw. "Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Comput. Appl. Math.* 20, 53-65". Em: *Journal of Computational and Applied Mathematics* 20 (nov. de 1987), pp. 53–65. DOI: [10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7) (ver p. 22).
- [32] G. Schwarz. "Estimating the Dimension of a Model". Em: *The Annals of Statistics* 6.2 (1978), pp. 461–464. DOI: [10.1214/aos/1176344136](https://doi.org/10.1214/aos/1176344136). URL: <https://doi.org/10.1214/aos/1176344136> (ver p. 24).
- [33] *Segmentação de Mercado - marketing Júnior USP*. Abr. de 2022. URL: <https://www.marketingjr.com.br/segmentacao-de-mercado/> (ver p. 12).
- [34] M. Sun et al. "Scalable Multi-view Subspace Clustering with Unified Anchors". Em: ago. de 2021. DOI: [10.1145/3474085.3475516](https://doi.org/10.1145/3474085.3475516) (ver pp. 50, 51, 53–55, 57, 61, 75).

- [35] *the 4 types of customer segmentation and how to apply them - profilename*. URL: <https://profiletree.com/customer-segmentation/> (ver p. 8).
- [36] K. Tsipis e A. Chorianopoulos. *Data mining techniques in CRM: Inside customer segmentation*. Wiley-Blackwell (an imprint of John Wiley & Sons), 2010 (ver p. 9).
- [37] *What is market segmentation? 4 types and 5 benefits*. Dez. de 2021. URL: <https://www.lotame.com/what-is-market-segmentation/> (ver p. 3).
- [38] J. Wu e J. M. Rehg. "CENTRIST: A Visual Descriptor for Scene Categorization". Em: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.8 (2011), pp. 1489–1501. DOI: [10.1109/TPAMI.2010.224](https://doi.org/10.1109/TPAMI.2010.224) (ver p. 67).
- [39] C. Xu, D. Tao e C. Xu. "A Survey on Multi-view Learning". Em: *ArXiv abs/1304.5634* (2013) (ver pp. 45–49, 57, 81).
- [40] Y. Yang e H. Wang. "Multi-view clustering: A survey". Em: *Big Data Mining and Analytics* 1 (jun. de 2018), pp. 83–107. DOI: [10.26599/BDMA.2018.9020003](https://doi.org/10.26599/BDMA.2018.9020003) (ver pp. 21, 44–50, 56, 74, 75, 80, 81).
- [41] L. Ye et al. "Telecom customer segmentation with K-means clustering". Em: *2012 7th International Conference on Computer Science Education (ICCSE)*. 2012, pp. 648–651. DOI: [10.1109/ICCSE.2012.6295158](https://doi.org/10.1109/ICCSE.2012.6295158) (ver p. 10).
- [42] X. Zhang, H. Liu e X. Liu. "Multi-Task Multi-View Clustering". Em: *IEEE Transactions on Knowledge and Data Engineering* 28 (dez. de 2016). DOI: [10.1109/TKDE.2016.2603983](https://doi.org/10.1109/TKDE.2016.2603983) (ver p. 56).
- [43] J. Zhao et al. "Multi-view Learning Overview: Recent Progress and New Challenges". Em: *Information Fusion* 38 (fev. de 2017). DOI: [10.1016/j.inffus.2017.02.007](https://doi.org/10.1016/j.inffus.2017.02.007) (ver pp. 81, 82).

DADOS DA PORDATA

A.1 Dados recolhidos da PORDATA

	0-14	15-64	65+	Total
Lisboa	71,277	328,537	143,106	542,917
Sintra	66,866	259,850	52,022	378,733
Vila Nova De Gaia	47,058	210,173	45,761	302,986
Porto	28,556	151,061	55,943	235,554
Cascais	34,775	136,461	35,957	207,187
Almada	26,690	112,351	34,869	173,906
Amadora	25,951	115,065	34,428	175,440
Loures	32,216	137,658	36,157	206,025
Portugal Continental	1,494,871	6,627,338	1,922,281	10,044,484

Tabela A.1: População residente por município e em Portugal Continental, em 2011

	Sem Nível de Escolaridade	1 ° Ciclo	2 ° Ciclo	3 ° Ciclo	Secundário	Médio	Superior	Total
Lisboa	32,413	96,850	38,897	74,655	80,869	5,142	148,413	477,239
Sintra	18,351	64,044	39,362	76,335	66,819	4,091	42,200	311,202
Vila Nova de Gaia	18,851	70,853	33,970	49,801	41,476	2,752	37,951	255,654
Porto	13,814	50,642	21,333	35,585	32,936	1,917	52,985	209,212
Cascais	9,767	30,426	15,484	34,083	37,171	2,709	44,184	173,824
Almada	11,069	35,248	16,036	31,068	28,068	1,701	25,257	148,447
Amadora	11,752	37,290	16,934	30,929	28,390	1,677	22,261	149,233
Loures	13,596	43,559	20,600	35,867	31,177	1,942	26,257	172,998
Portugal Continental	884,705	2,322,454	1,082,748	1,636,402	1,355,254	83,485	1,198,453	8,563,501

Tabela A.2: Nível de educação, por município e em Portugal Continental, em 2011

Zona	Poder de Compra	Ganho Médio Mensal (€)
Lisboa	216.9	1576.4
Sintra	101.3	1159.6
Vila Nova de Gaia	99.1	1029.5
Porto	161.7	1279.9
Cascais	132.0	1171.9
Almada	109.8	1040.3
Amadora	105.7	1249.4
Loures	102.4	1109.3
Portugal Continental	100.8	1084.6

Tabela A.3: Poder de compra e ganho médio mensal, por município e em Portugal Continental, em 2011

	1	2	3-5	6+	Total
Lisboa	85,244	79,452	75,126	4,070	243,892
Sintra	30,283	44,361	66,585	2,931	144,160
Vila Nova de Gaia	21,242	35,309	55,725	1,987	114,263
Porto	30,736	32,663	35,713	1,714	100,826
Cascais	20,210	26,681	33,715	1,487	82,093
Almada	18,544	24,487	27,763	1,107	71,901
Amadora	20,119	24,660	27,245	1,409	73,433
Loures	17,492	27,289	34,012	1,671	80,464
Portugal Continental	834,680	1,232,982	1,729,796	71,730	3,869,188

Tabela A.4: Dimensão das famílias, por município e em Portugal Continental, em 2011

A.2 Dados normalizados da PORDATA

Zona	0-14 anos (%)	15-64 anos (%)	65+ anos (%)	Índice Envelhecimento
Lisboa	13.1	60.5	26.36	200.8
Sintra	17.7	68.6	13.74	77.8
Vila Nova De Gaia	15.5	69.4	15.1	97.2
Porto	12.1	64.1	23.75	195.9
Cascais	16.8	65.9	17.4	103.4
Almada	15.4	64.6	20.1	130.6
Amadora	14.8	65.6	19.6	132.7
Loures	15.6	66.8	17.6	112.2
Continente	14.9	66.0	19.1	128.6

Tabela A.5: Composição da população residente por grandes grupos etários, por município e em Portugal Continental, em 2011

	Sem Nível Escolaridade (%)	Básico 1º ciclo (%)	Básico 2º ciclo (%)	Básico 3º ciclo (%)	Ensino Secundário (%)	Ensino Médio (%)	Ensino Superior (%)
Lisboa	6.8	20.3	8.2	15.6	17.0	1.1	31.1
Sintra	5.9	20.6	12.7	24.5	21.5	1.3	13.6
Vila Nova de Gaia	7.4	27.7	13.3	19.5	16.2	1.1	14.8
Porto	6.6	24.2	10.2	17.0	15.7	0.9	25.3
Cascais	5.6	17.5	8.9	19.6	21.4	1.6	25.4
Almada	7.5	23.7	10.8	20.9	18.9	1.2	17.0
Amadora	7.9	25.0	11.4	20.7	19.0	1.1	14.9
Loures	7.9	25.2	11.9	20.7	18.0	1.1	15.2
Continente	10.3	27.1	12.6	19.1	15.8	1.0	14.0

Tabela A.6: Composição da população residente por nível de educação, por município e em Portugal Continental, em 2011

	1 elemento (%)	2 elementos (%)	3 a 5 elementos (%)	Mais de 6 elementos (%)
Lisboa	35.0	32.6	30.8	1.7
Sintra	21.0	30.8	46.2	2.0
Vila Nova de Gaia	18.6	30.9	48.8	1.7
Porto	30.5	32.4	35.4	1.7
Cascais	24.6	32.5	41.1	1.8
Almada	25.8	34.1	38.6	1.5
Amadora	27.4	33.6	37.1	1.9
Loures	21.7	33.9	42.3	2.1
Continente	21.6	31.9	44.7	1.9

Tabela A.7: Composição da população residente por dimensões do agregado familiar, por município e em Portugal Continental, em 2011

EXPERIÊNCIAS COM CONJUNTOS DE DADOS *MULTI-VIEW*

B.1 Handwritten

B.1.1 Vistas originais

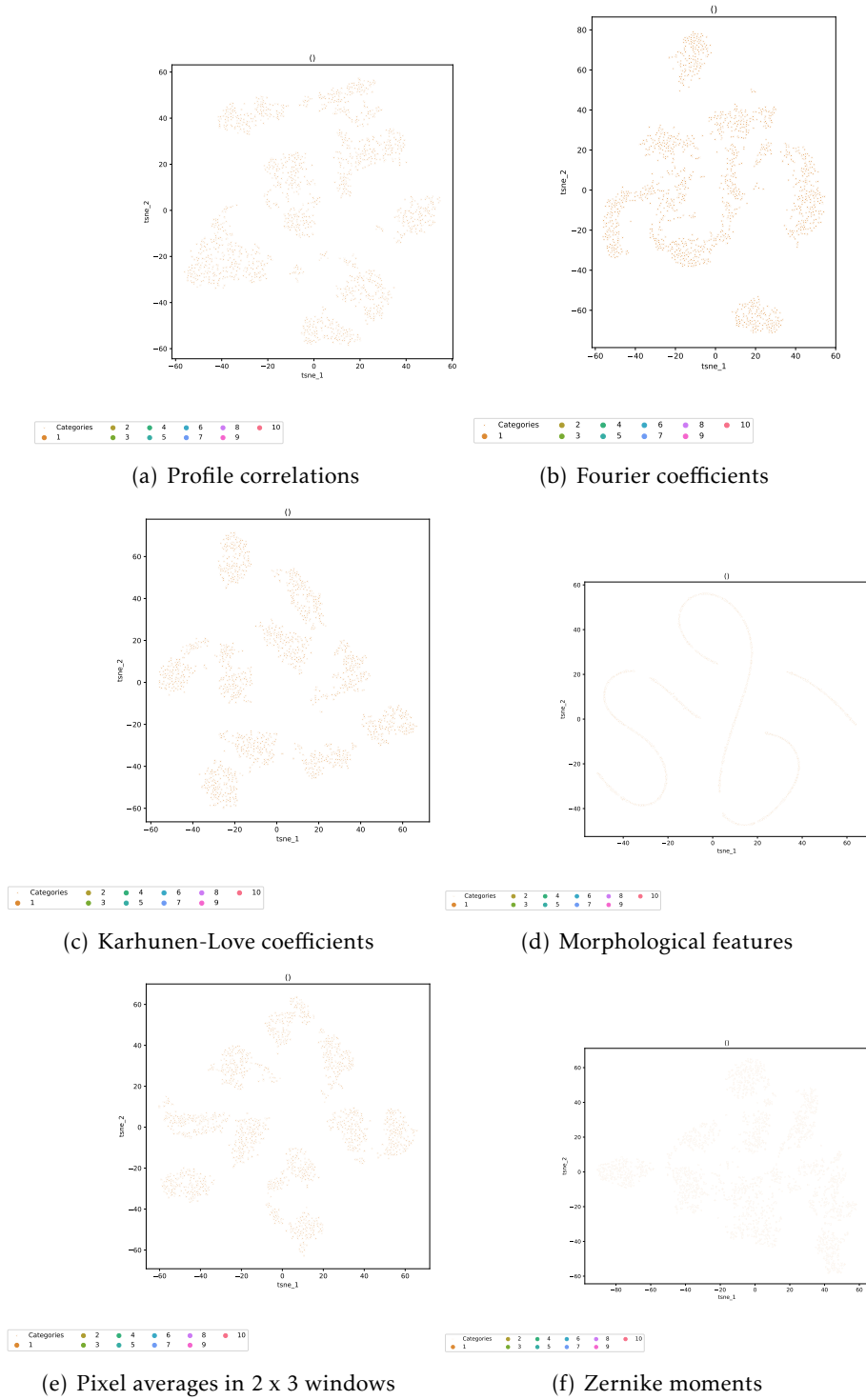
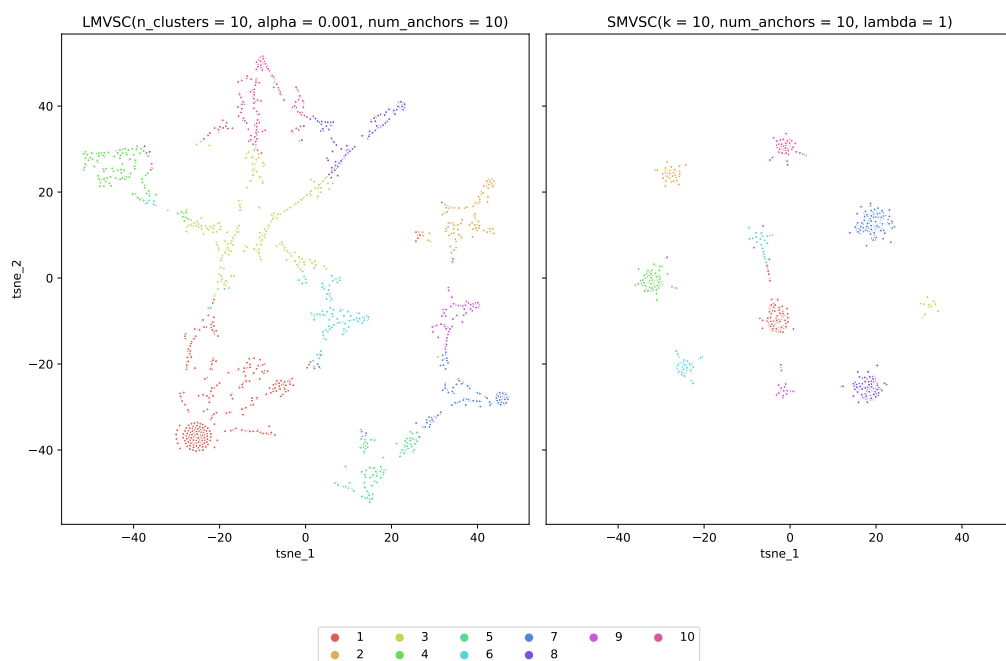
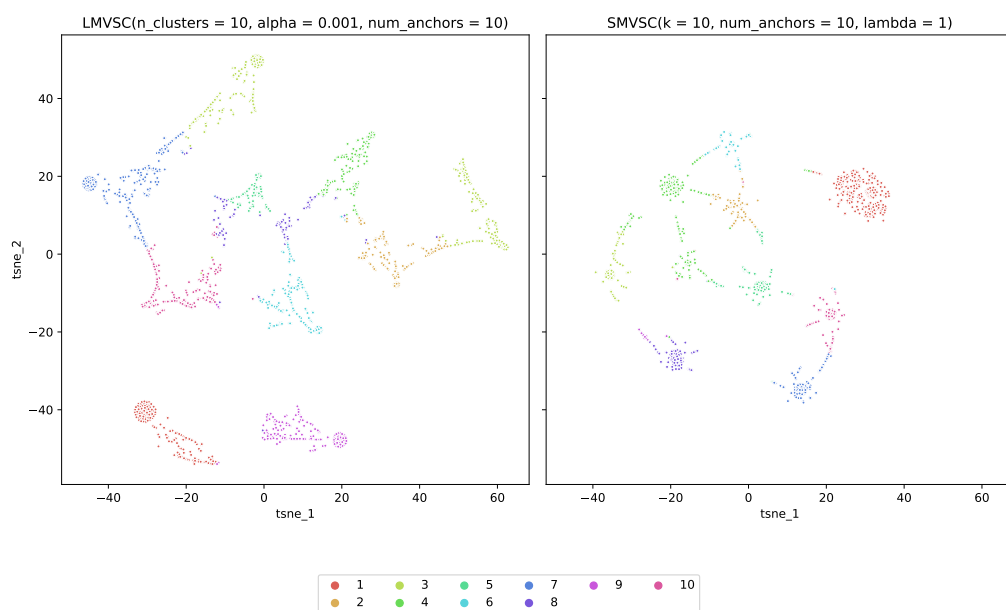


Figura B.1: Visualização das vistas do conjunto de dados Handwritten, utilizando a técnica de redução de dimensionalidade t-SNE.

B.1.2 Resultados de clusterização por vista

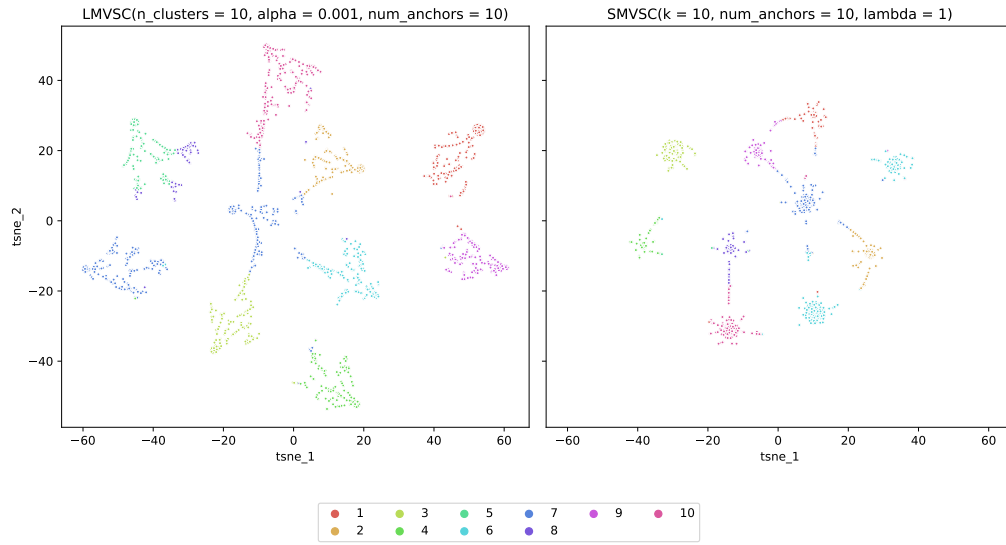


(a) Vista Profile Correlations

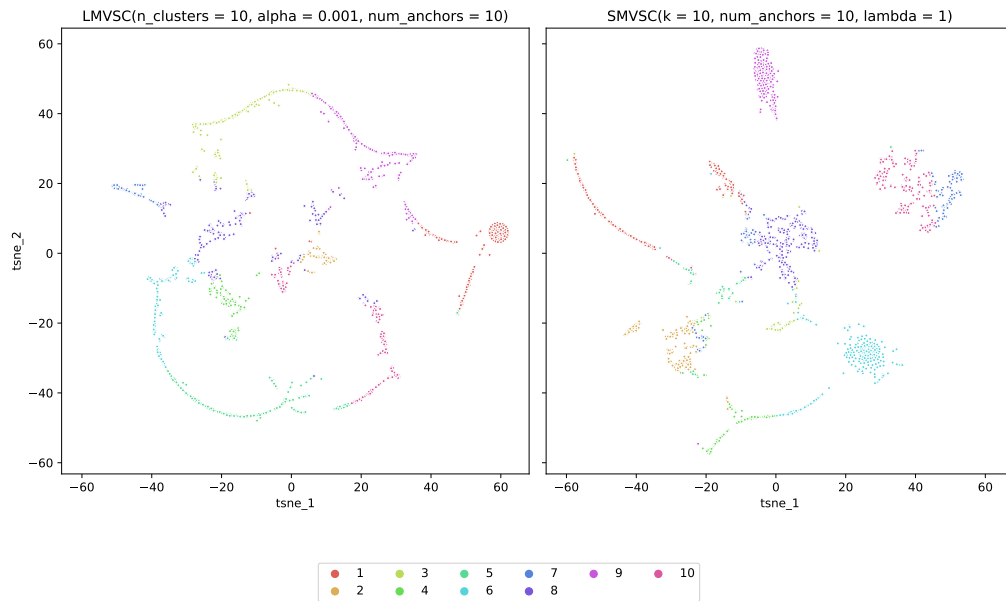


(b) Vista Fourier coefficients of the character

Figura B.2: Visualização dos resultados de clusterização para o conjunto de dados Multiple Features handwritten digit, utilizando a técnica de redução de dimensionalidade t-SNE.

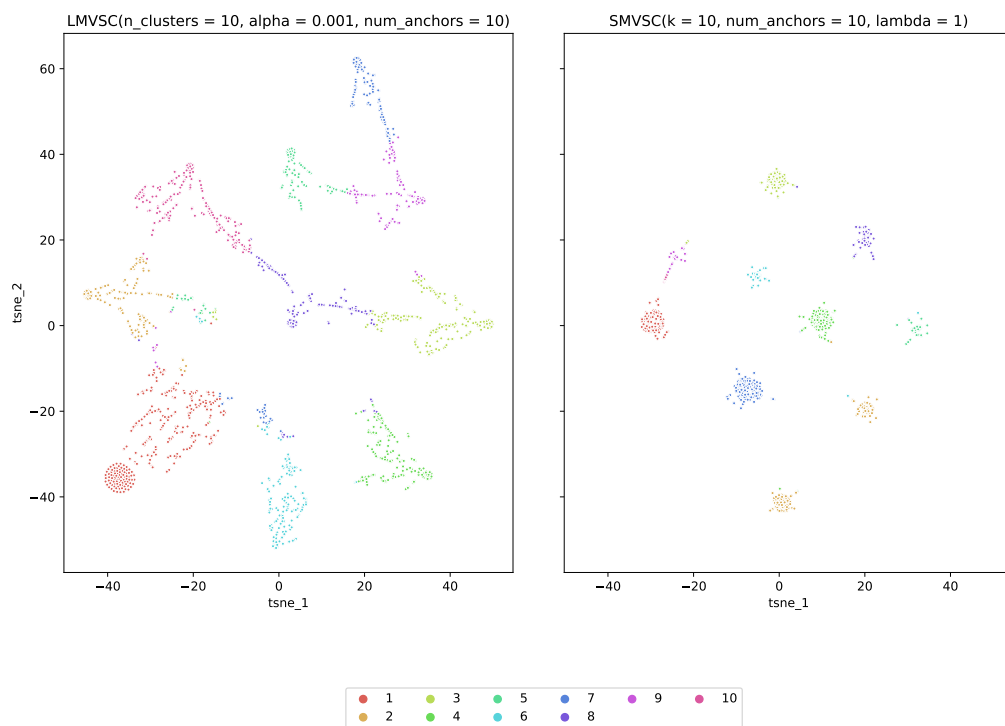


(a) Vista Karhunen-Love coefficients

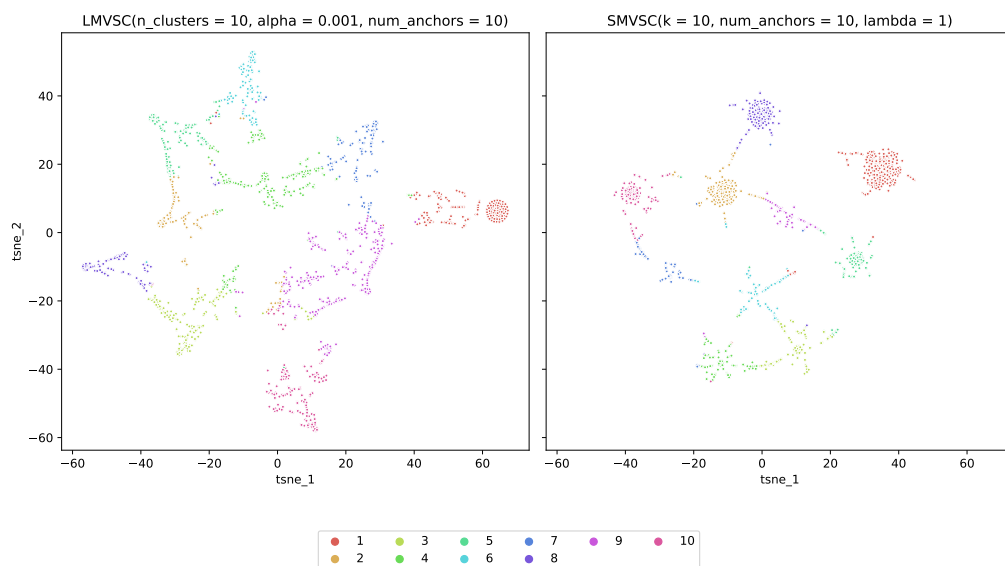


(b) Vista Morphological features

Figura B.2: Visualização dos resultados de clusterização para o conjunto de dados Multiple Features handwritten digit, utilizando a técnica de redução de dimensionalidade t-SNE.



(a) Vista Pixel averages in 2 x 3 windows



(b) Vista Zernike moments

Figura B.2: Visualização dos resultados de clusterização para o conjunto de dados Multiple Features handwritten digit, utilizando a técnica de redução de dimensionalidade t-SNE.

B.2 Caltech101-20

B.2.1 Vistas originais

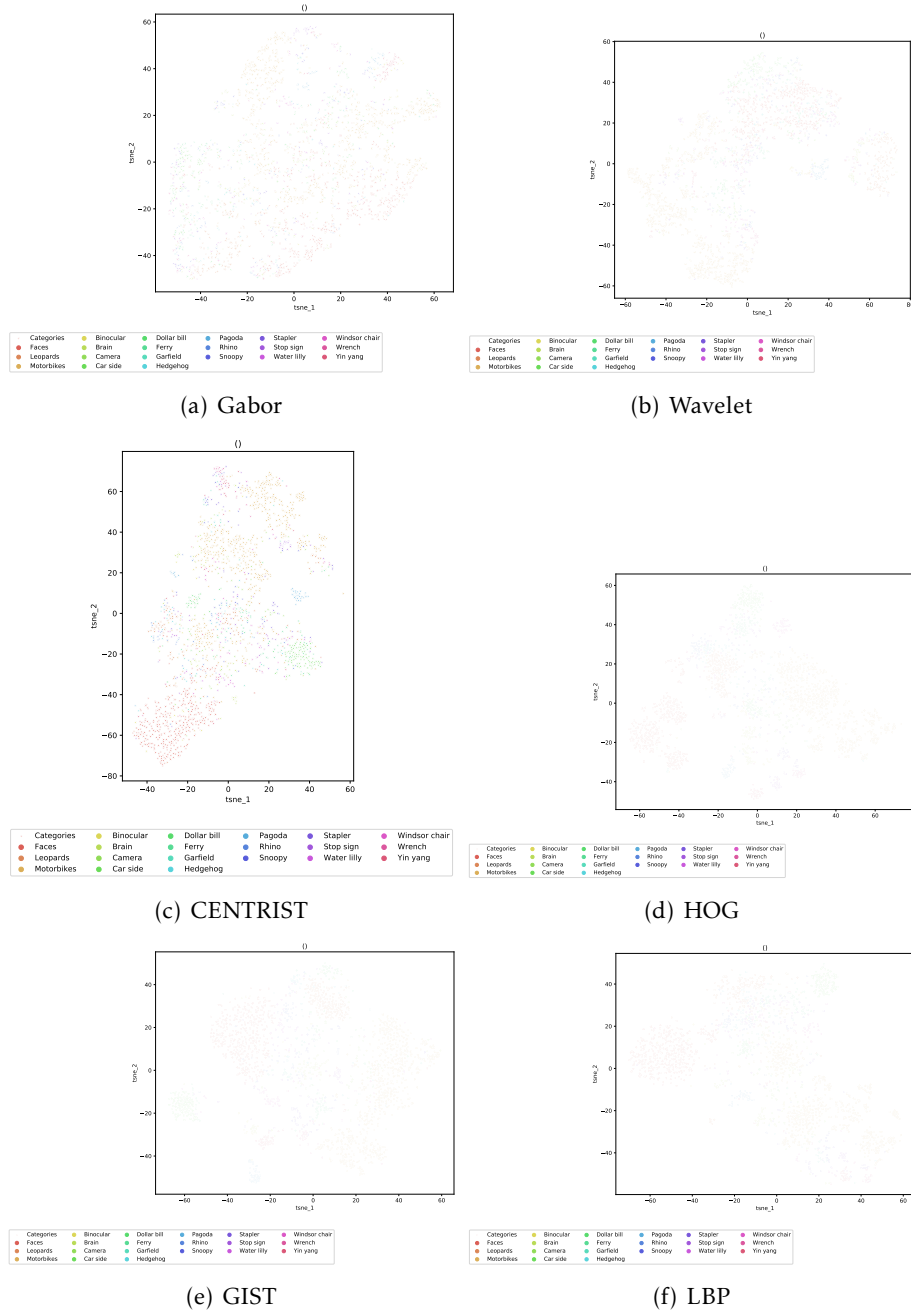
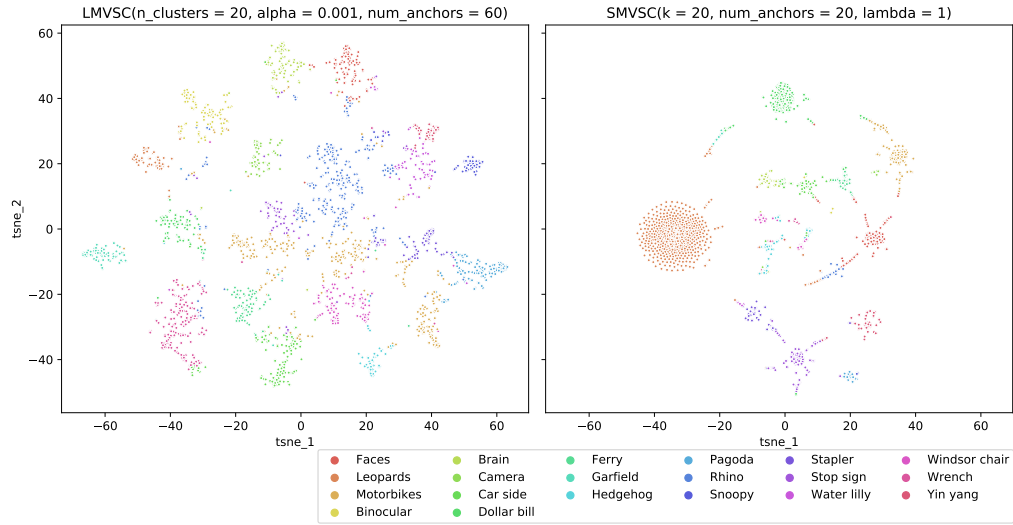
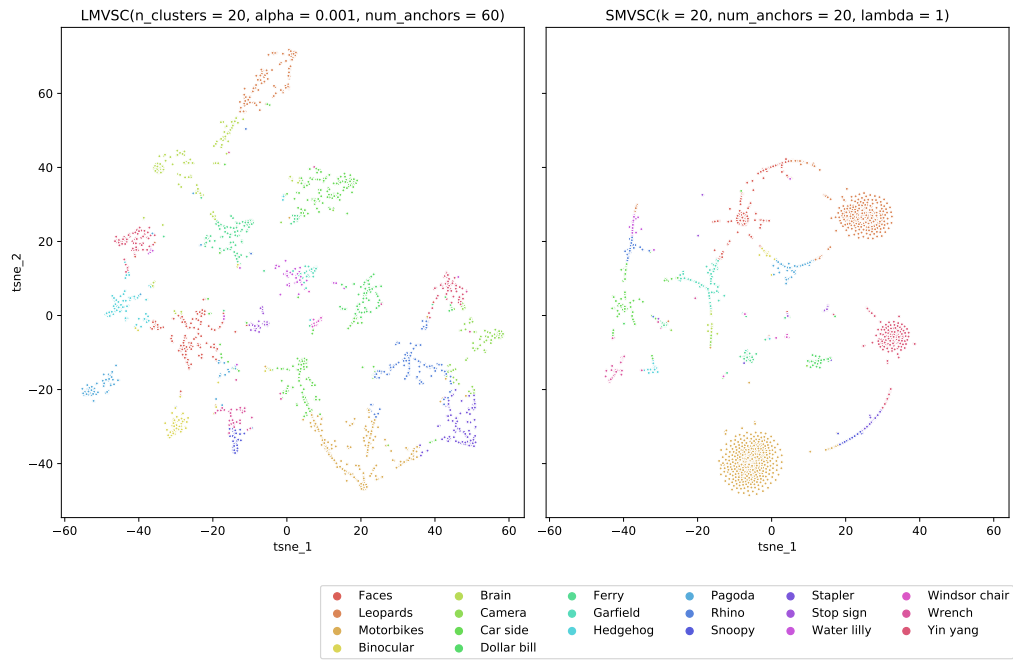


Figura B.3: Visualização das vistas do conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE.

B.2.2 Resultados de clusterização por vista

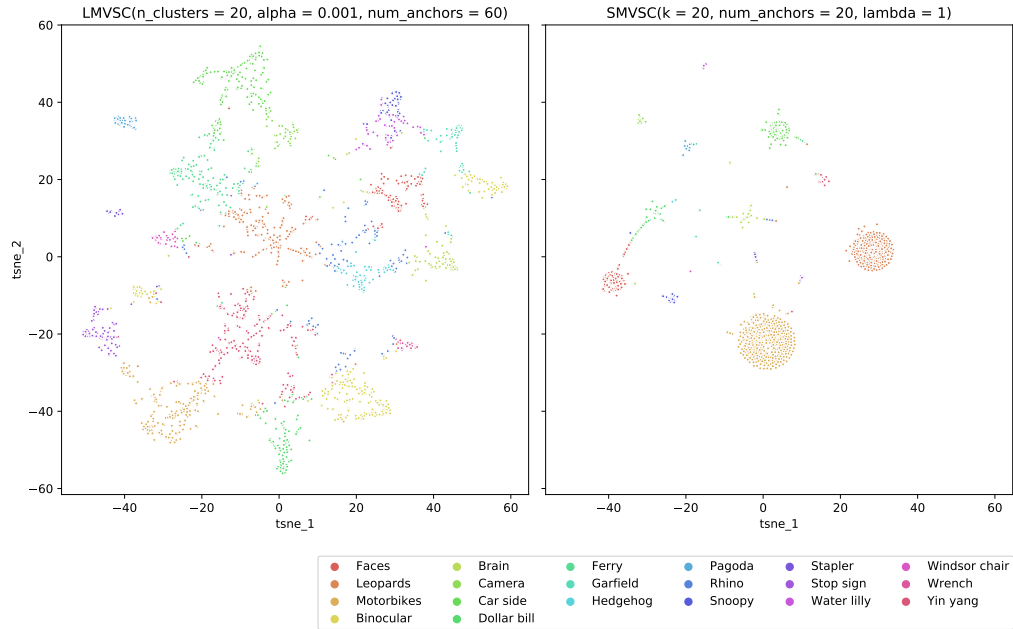


(a) Gabor

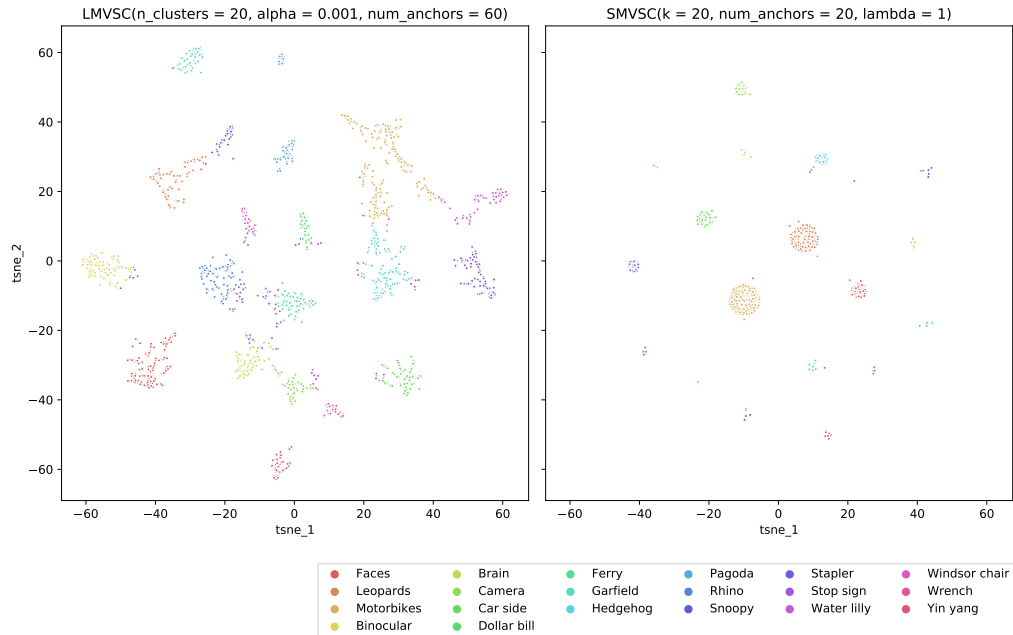


(b) Wavelet moments

Figura B.4: Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE.

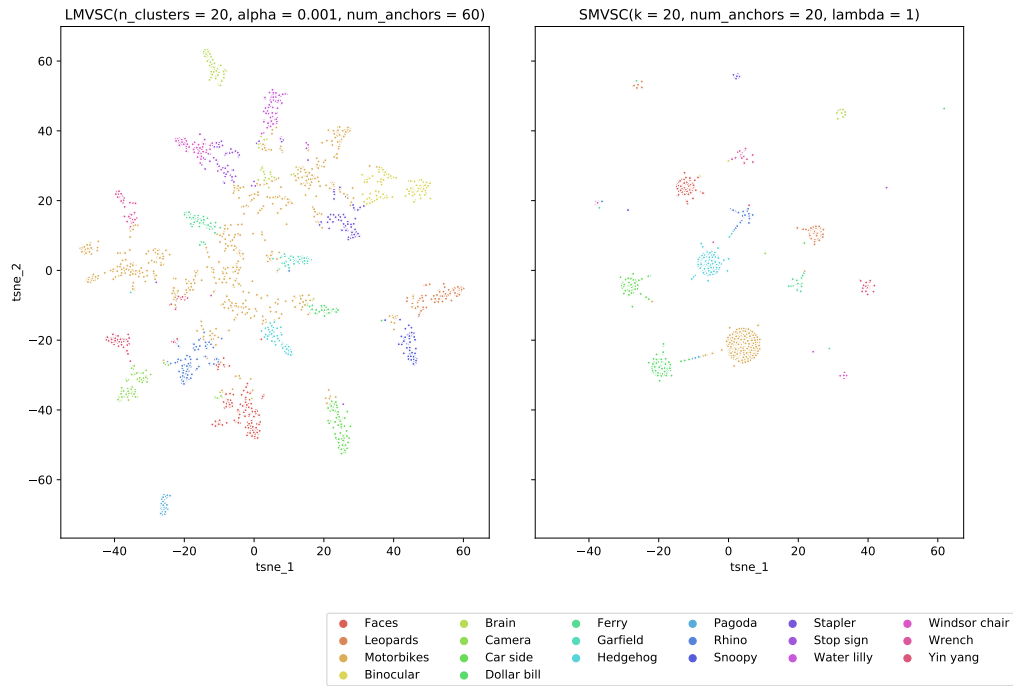


(a) CENTRIST

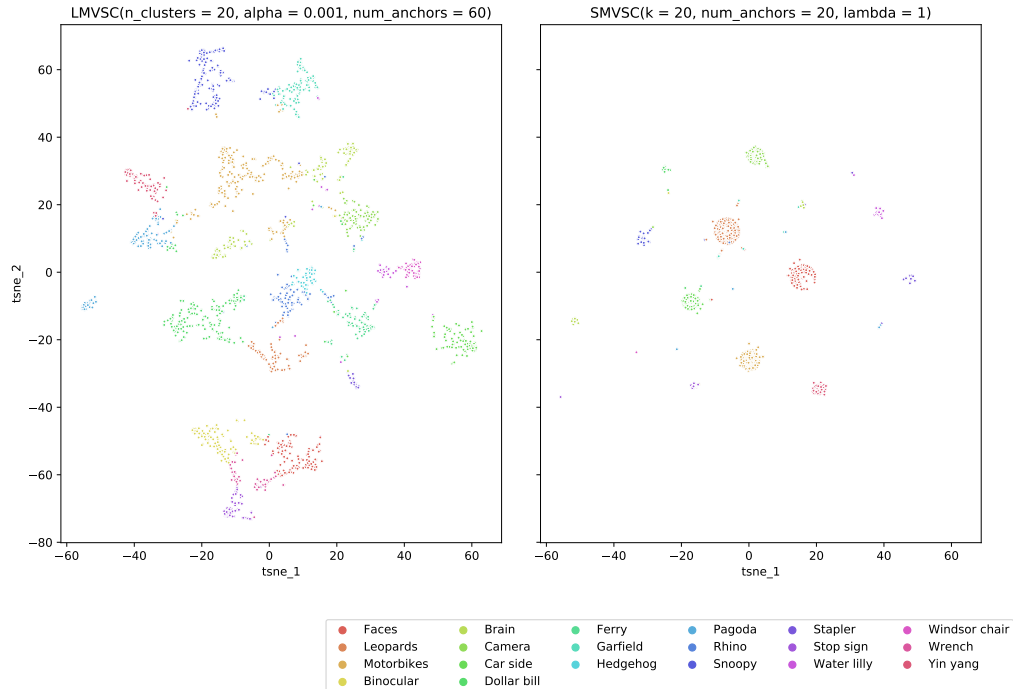


(b) HOG

Figura B.4: Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE.



(a) GIST

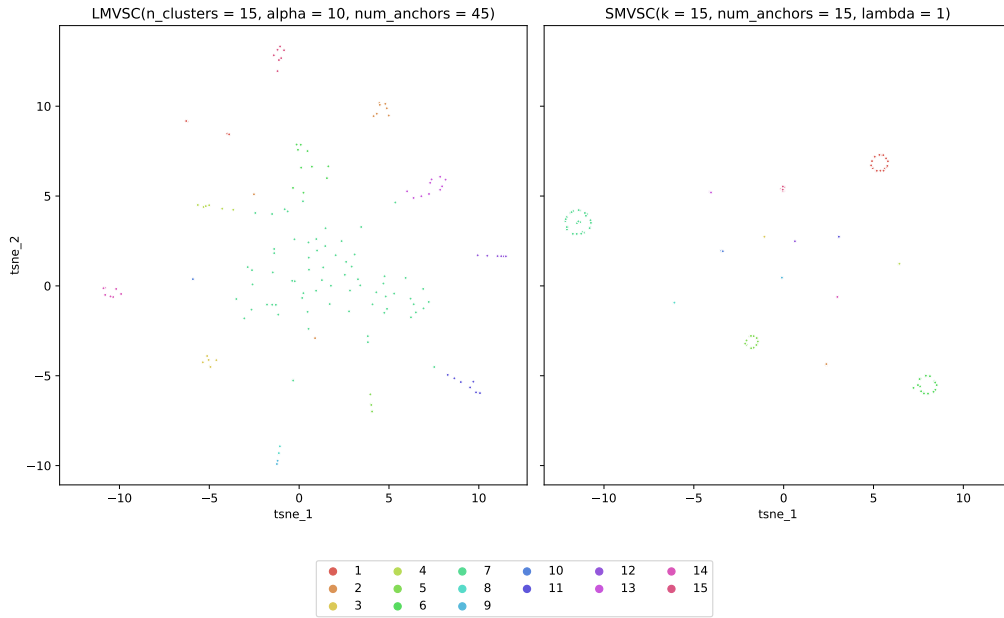


(b) LBP

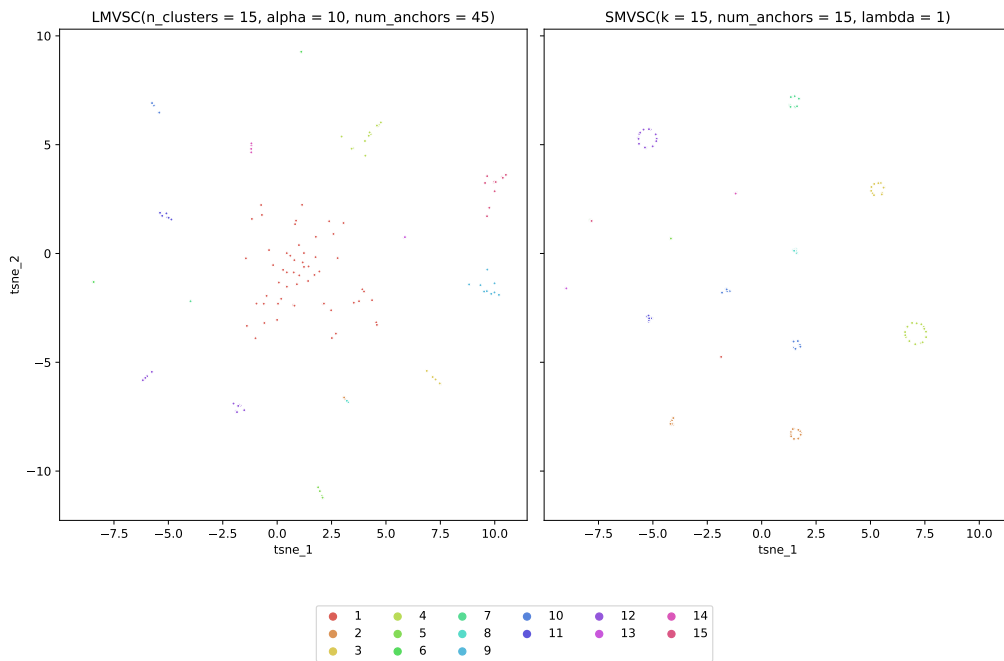
Figura B.4: Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE.

B.3 Yale Faces

B.3.1 Resultados de clusterização por vista

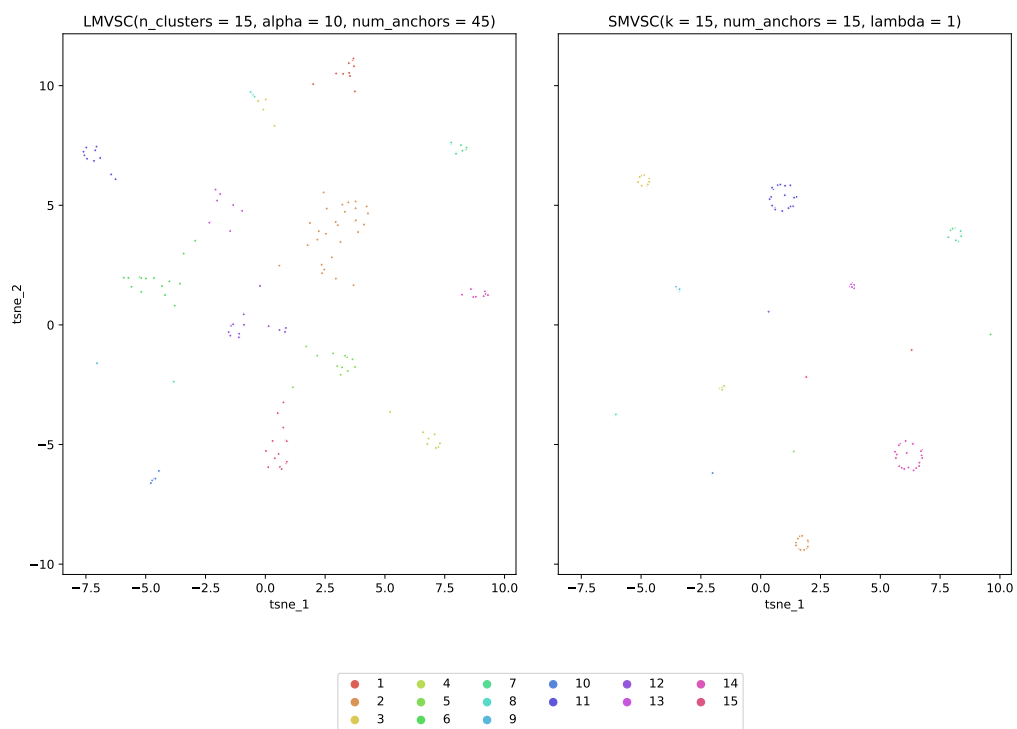


(a) Intensity



(b) LBP

Figura B.5: Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE.



(a) Gabor

Figura B.5: Visualização dos resultados de clusterização para o conjunto de dados Caltech101-20, utilizando a técnica de redução de dimensionalidade t-SNE.

VARIÁVEIS UTILIZADAS

Pilar	Conceitos	Variável	Tipo
Demográfico	Idade	age	Numérica
		indice_envelhecimento	
	Educação	sem_nivel_escolaridade_percentage	
	Rendimentos	poder_compra	
	Famílias	families_3+_percentage	
Psicográfico	Pagamentos	digital_invoice_flg direct_payment_flg	Binária
		no_marketing_flg	
	Premium	premium_qty tv_cine_qty benfica_tv_qty sport_tv_qty other_premium_qty	Númerica
	Protocolos sofisticados utilizados	remote_connectivity_flg network_admin_flg online_games_flg tunneling_flg database_flg tunneling_flg file_access_flg voip_flg	Binária
Comportamental	Quantidade de Dispositivos	devices_qty phones_number_qty router_qty phone_qty hub_qty plumes_qty powerbox_qty	Numérica

	tv_qty	
Tipo de box TV	tv_type_dsc	Categórica
Intensidade de Utilização	down_traffic_1w_avg_median up_traffic_1w_avg_median holiday_tv_sessions_qty weekend_tv_sessions_qty weekday_tv_sessions_qty	Númerica
Utilização do Router	holiday_uses_5g_flg holiday_uses_24g_flg holiday_uses_eth_flg weekday_uses_5g_flg weekday_uses_24g_flg weekday_uses_eth_flg	Binária
Churn passado	product_revolving_churn_qty product_effective_churn_qty	Numérica
	past_involuntary_churn_flg past_voluntary_churn_flg past_partial_voluntary_churn_flg past_total_voluntary_churn_flg	Binária
Lealdade à marca	loyalty_period days_until_loyalty_ends days_since_last_activation sa_same_fiscal_num_qty	Categórica Numérica Numérica Numérica

