NOTA DE ANÁLISE - APÊNDICE METODOLÓGICO

PLANAPP
PLANEAMENTO | POLÍTICAS | PROSPETIVA

Organização do tempo de trabalho na Administração Pública Central

Apêndice metodológico aos resultados do inquérito

Fevereiro 2024

FICHA TÉCNICA

Título
Organização do tempo de trabalho na Administração Pública Central: Apêndice metodológico aos resultados do inquérito

Data
19 de fevereiro de 2024

Coordenação
Pedro Mazeda Gil (Assessoria Estratégica e Projetos Especiais – PlanAPP)

Autoria
Henrique Pereira (Assessoria Estratégica e Projetos Especiais – PlanAPP)

Revisão e *layout*
Equipa Multidisciplinar de Comunicação e Gestão do Conhecimento (EMCGC) – PlanAPP

PlanAPP – Centro de Competências de Planeamento, de Políticas e de Prospetiva da Administração Pública
Rua Filipe Folque, 44
1069-123 Lisboa
planapp@planapp.gov.pt
www.planapp.gov.pt

Índice

1. Introdução

Serve o presente apêndice para expor, esclarecer e justificar as opções metodológicas consideradas nas análises de inferência e de associação cujos resultados são apresentados na Nota de Análise do PlanAPP e DGAEP intitulada “Organização do Tempo de Trabalho na Administração Pública: Inquérito e Análise de Resultados”, no que diz respeito às Secção 2 e 3 do referido documento.

Quanto ao remanescente deste apêndice, a Secção 2 descreve, *grosso modo*, o processo de extrapolação da amostra considerada para o universo alvo, esclarecendo sobre a sua representatividade, a delimitação final do universo e o processo inferencial.

De seguida, a Secção 3 do apêndice detalha a forma como foram conduzidas as análises de associação apresentadas na Nota de Análise, esclarecendo quais os métodos de regressão utilizados e os desenhos dos mesmos.

Por fim, Secção 4 do apêndice apresenta a estratégia seguida para a análise dos campos de resposta aberta do inquérito aos trabalhadores da AC respeitantes a duas perguntas que visaram capturar as perceções dos trabalhadores face a outras necessidades de investimento, por um lado, e outras necessidades de ajustamento, por outro, tendo em vista a redução do tempo de trabalho.

2. Amostra, universo representado e inferência

A amostra abrangeu uma multiplicidade de tipos de entidades das diferentes áreas governativas, refletindo a diversidade inerente à AC. Foi possível apurar 14 232 respostas, correspondentes a 2,7% do total de trabalhadores registados no Sistema de Informação da Organização do Estado (SIOE) com referência a 31 de março de 2023 e respeitantes a um total de 649 entidades da AC.[1]

Em traços gerais, relativamente à amostra apurada, tem-se que:

- A amostra cobre as diferentes áreas ministeriais, embora não respeite, em geral, o peso que cada uma delas tem no universo da AC. Em especial, nota-se uma sub-representação nos ministérios da Administração Interna, da Defesa Nacional e da Saúde. Ao nível das idades, a amostra sobre-representa a população de 45 e mais anos e a população de mulheres. Em determinados estratos, que resultam do cruzamento das dimensões área ministerial da entidade do respondente, carreira, escalão etário e sexo dos inquiridos, existiu apenas um trabalhador inquirido. Para estes casos, essa informação não foi tida em conta.
- Existem carreiras para as quais a amostra final não permite obter informação suficiente, nomeadamente: Bombeiros, Guardas Prisionais, Magistrados, Oficiais de Justiça, Representantes do poder legislativo e de órgãos executivos, Diplomatas, antigo Serviço de Estrangeiros e Fronteiras, Polícia Judiciária, Guarda Nacional Republicana e Forças Armadas.
- A amostra final, resultante do tratamento de algumas incongruências e da eliminação de estratos com insuficiente representação, conta com 13 573 respostas.

Através de uma estratificação feita *a posteriori*, tendo em conta as dimensões de área ministerial da entidade do inquirido, carreira, escalão etário e sexo dos inquiridos, foi analisada a capacidade da amostra replicar apropriadamente as distribuições populacionais de cada uma destas dimensões.

Para as carreiras cuja amostra apresenta estratos sem representação e esses estratos correspondem a mais de metade dos indivíduos da carreira associada, a carreira em causa foi descartada. A exceção a esta regra são as carreiras dos Oficiais de Justiça, e dos Militares das Forças Armadas. Estas exceções justificam-se pelo seguinte:

- Relativamente a Oficiais de Justiça, só existe representação para individuos com mais de 55 anos e menos de 25.
- Quanto aos Militares das Forças Armadas, os estratos sem representação foram considerados especialmente relevantes, não fazendo sentido representar a carreira sem estes (embora o seu peso seja de apenas 25% do universo da carreira).

As diferenças que se sinalizam entre a composição da amostra e a composição do universo, nomeadamente no que diz respeito a categorias sobre e sub-representadas, não representadas e insuficientemente representadas, traduzem-se num efeito de seleção com potêncial de enviesar a

[1] O inquérito foi respondido pelos trabalhadores a partir de uma ligação eletrónica para o questionário, sendo o controlo da pertença a determinada entidade aferido unicamente pela resposta dada pelo trabalhador quanto ao organismo da AC em que desempenha a maior parte da sua atividade profissional.

análise sempre que se procure inferir parâmetros populacionais (ou seja, extrapolar para o total da AC).

Na medida em que as respostas dependam de especificidades da área ministerial ou da carreira, ou que o sexo e idade dos respondentes influencie o sentido de resposta dos inquiridos, existirá um enviesamento decorrente do facto de estas características não estarem igualmente distribuídas entre a amostra e o universo que se pretende considerar.

De forma a mitigar o enviesamento do efeito de seleção decorrente do desequilíbrio das características atrás mencionadas e que compõe a base do sistema de estratificação, foi calculado um ponderador individual que permite ajustar as distribuições dessas características às distribuições do universo passível de ser representado pela amostra (cerca de 89%).

Tabela A1 - Representação pela amostra e peso na AC de carreiras insuficientemente representadas

Carreira	Representação pela amostra (%)	Peso na AC (%)
Bombeiro	0%	<1%
Diplomata	32%	<1%
Guarda Nacional Republicana	2%	4%
Guarda Prisional	0%	1%
Magistrado	0%	<1%
Militar das Forças Armadas	75%	5%
Oficial de Justiça	54%	<1%
Outro Pessoal de Segurança	0%	<1%
Polícia Judiciária	13%	<1%
Representantes do poder legislativo e de órgãos executivos	0%	<1%
Antigo Serviço Estrangeiros Fronteiras	0%	<1%

Tabela A2 - Militares das forças armadas sem qualquer representação

Até aos 34	Feminino	Oficial
Até aos 34	Masculino	Praça
Até aos 34	Masculino	Sargento
35 aos 44	Feminino	Oficial

Tabela A3 - Representatividade da amostra por ministério

Ministério	Peso no Universo	Peso na Amostra	% Amostragem	Efeito de seleção
MAA	1%	4%	15%	4 pp
MAAC	1%	1%	6%	1 pp
MAI	9%	7%	2%	-2 pp
MC	0%	1%	5%	1 pp
MCT	0%	1%	11%	1 pp
MCTES	9%	10%	3%	1 pp
MDN	6%	1%	1%	-4 pp
MEdu	36%	39%	3%	3 pp
MEM	1%	3%	14%	2 pp
MF	2%	9%	12%	7 pp
MIH	0%	1%	5%	0 pp
MJ	3%	3%	3%	0 pp
MNE	0%	1%	11%	1 pp
MS	28%	8%	1%	-21 pp
MTSSS	3%	8%	7%	5 pp
PCM	1%	1%	3%	0 pp

Nota: pp – pontos percentuais (valores arredondados à unidade).

Tabela A4 – Representatividade da amostra por escalão etário

Escalão Etário	Peso no Universo	Peso na Amostra	% Amostragem	Efeito de seleção
Até aos 34 anos	15%	7%	1%	-8 pp
35 aos 44	22%	18%	2%	-4 pp
45 e mais anos	63%	75%	3%	12 pp

Nota: pp – pontos percentuais (valores arredondados à unidade).

Tabela A5 – Representatividade da amostra por sexo

Sexo	Peso no Universo	Peso na Amostra	% Amostragem	Efeito de seleção
Feminino	66%	69%	3%	3 pp
Masculino	34%	31%	2%	-3 pp
Todos	100%	100%	3%	0 pp

Nota: pp – pontos percentuais (valores arredondados à unidade).

Tabela A6 – Representatividade da amostra por carreira

Carreira	Peso no Universo	Peso na Amostra	% Amostragem	Efeito de seleção
Assistente Operacional, Operário/a, Auxiliar	16%	9%	2%	-7 pp
Assistente Técnico/a, Técnico/a Nível Intermédio, Administrativo/a	9%	12%	4%	3 pp
Bombeiro	0%	0%	0%	0 pp
Chefia Tributária	0%	1%	8%	0 pp
Conservador e Notário	0%	0%	11%	0 pp
Diplomata	0%	0%	11%	0 pp
Dirigente 1º grau	0%	0%	21%	0 pp
Dirigente 2º grau	0%	1%	17%	1 pp
Dirigente Intermédio	1%	3%	6%	2 pp
Docente do Ensino Superior Politécnico	2%	2%	2%	0 pp
Docente do Ensino Superior Universitário	3%	2%	1%	-2 pp
Educ.Infância e Doc. do Ens. Básico e Secundário	25%	24%	3%	-1 pp
Enfermeiro/a	10%	2%	1%	-8 pp
Guarda Nacional Republicana	4%	0%	0%	-4 pp
Guarda Prisional	1%	0%	0%	-1 pp
Informático	1%	1%	7%	1 pp
Magistrado	0%	0%	0%	0 pp
Médico/a	6%	1%	0%	-5 pp
Militar das Forças Armadas	5%	1%	0%	-4 pp
Oficial de Justiça	0%	0%	9%	0 pp
Oficial dos Registos e do Notariado	1%	2%	6%	1 pp
Outro Pessoal de Segurança	0%	0%	0%	0 pp
Pessoal Aduaneiro	0%	1%	13%	1 pp
Pessoal de Administração Tributária	1%	5%	10%	3 pp
Pessoal de Inspecção	0%	2%	15%	2 pp
Pessoal de Investigação Científica	1%	1%	2%	0 pp
Polícia de Segurança Pública	4%	6%	4%	2 pp
Polícia Judiciária	0%	0%	0%	0 pp
Representantes do poder legislativo e de órgãos executivos	0%	0%	0%	0 pp
Serviço Estrangeiros Fronteiras	0%	0%	0%	0 pp
Técnico Superior de Diagnóstico e Terapêutica	2%	1%	1%	-1 pp
Técnico Superior de Saúde	0%	1%	10%	1 pp
Técnico/a Superior	7%	24%	9%	16 pp

Nota: pp – pontos percentuais (valores arredondados à unidade).

3. Análises de associação entre características dos trabalhadores e as suas perceções

A análise de fatores estatisticamente associados às respostas relativas a perceções dos trabalhadores (reportadas nas Tabelas 23, 27 e 30), foi elaborada tendo por base a noção de "riscos relativos" (apresentados como rácios de probabilidades). Esta opção justifica-se pela complicada interpretação dos rácios de *odds* (a opção mais comum na literatura), que muitas vezes é erroneamente interpretada como rácios de probabilidades. Desta feita, evitando uma interpretação errónea e facilitando a intuição das associações identificadas, foi escolhida como métrica de associação os rácios de probabilidades. Em termos simples, a métrica apresentada reflete a probabilidade de determinada resposta (e.g., discordar da S4D) de um determinado grupo (e.g., pertencer a carreiras gerais) relativa (dividida) à mesma probabilidade para o grupo contrário (e.g., não pertencer às carreiras gerais).

Zou (2004) sugere o uso de regressões de "Poisson modificadas" para estimar os riscos relativos (rácios de probabilidades) e respetivos intervalos de confiança usando métodos de estimação de variâncias robustas. Os erros padrão apresentados neste relatório seguem a metodologia de estimação robusta do tipo Horvitz-Thompson. Note-se que, de acordo com a análise feita em todo o documento, procurou-se extrapolar as diversas associações para todo o universo considerado, recorrendo-se para isso a regressões ponderadas (respeitando os pesos apurados por estrato).

De forma a isolar o efeito de determinadas características dos trabalhadores (que podem ser confundidos por efeitos de outras características), as especificações das regressões aplicadas aos dados em causa consideram as várias variáveis suspeitas de introduzir esta confundibilidade. Assuma-se o seguinte exemplo hipotético: as carreiras gerais são predominantemente compostas por mulheres. Neste caso, importa distinguir o efeito que se identifica relativamente a um individuo estar inserido nas carreiras gerais do efeito de se ser mulher. A não inclusão da variável 'mulher' captaria no coeficiente de regressão associado à variável carreira os efeitos de se ser mulher (por estas estarem sobre-representadas neste grupo).

Por sua vez, de forma a que os grupos de comparação para cada característica correspondam ao seu contrário, as várias características foram binarizadas (definidas com o valor 1 e 0 caso o trabalhador possua, ou não, determinada característica, respetivamente) e foram estimadas regressões separadas para cada característica (controlando para os vários efeitos que podiam confundir as associações que se pretendem identificar). Caso contrário, a inclusão de, por exemplo, várias carreiras exceto uma (evitando colinearidade) levaria a que a base de comparação das probabilidades não fosse o contrário da característica em causa (no exemplo, uma carreira específica), mas sim o grupo que fica de fora. Tal, embora válido estatisticamente, tornaria os rácios de probabilidade estimados mais difíceis de interpretar. Assim, a opção metodológica seguida estabelece que, para os diferentes rácios de probabilidade apresentados, a interpretação deva ter como base de comparação o contrário da característica à qual o rácio é relativo, mantendo tudo o resto constante (ou seja, isolando o melhor possível o efeito dessa característica).

As tabelas apresentadas de seguida dizem respeito aos relatórios das várias regressões estimadas para cada uma das análises apresentadas (totalizando 65 regressões). Mais uma vez, embora não sejam apresentados todos os coeficientes estimados, foram sempre incluídas na regressão todas as variáveis passíveis de confundir os efeitos que se pretendiam isolar. Note-se que, de forma a obter os rácios de probabilidades apresentados no relatório, deve-se proceder à exponenciação dos coeficientes estimados e respetivos intervalos de confiança, uma vez que se usaram regressões de Poisson no procedimento de estimação.

3.1. Análise dos fatores associados à discordância da redução da jornada semanal de trabalho

Tabela A7 – Análise às carreiras dos respondentes

	Discorda da redução da jornada semanal de trabalho (binária)									
Carreiras Segurança	0,334									
	(0,263)									
Carreiras Saúde		-0,371								
		(0,308)								
Carreiras na Justiça e Notariado			**-1,693****							
			(0,706)							
Carreiras Inspeção				-0,097						
				(0,197)						
Carreiras Educação (excepto Ensino Superior)					0,007					
					(0,155)					
Carreiras Dirigentes						**0,976*****				
						(0,207)				
Carreiras Gerais							-0,117			
							(0,124)			
Carreiras Informática								**-1,743*****		
								(0,544)		
Carreiras Ensino Superior e Investigação Científica									**0,668*****	
									(0,242)	

Notas: *p<0.1; **p<0.05; ***p<0.01
As colunas da tabela correspondem a diferentes regressões. Foram omitidos da tabela os coeficientes associados a características demográficas, de composição dos agregados e de perceção face à conciliação da vida pessoal e profissional, apesar destas terem sido incluídas na regressão

Tabela A8 – Análise relativa ao sexo, características do agregado e outras perceções dos respondentes

	Discorda da redução da jornada semanal de trabalho (binária)
Sexo Masculino	**0,372****
	(0,162)
Tem Crianças a Cargo	0,102
	(0,152)
Tem Adultos a Cargo	0,165
	(0,141)
Faz Teletrabalho	-0,144
	(0,141)
Consegue Conciliar a Vida Profissional Pessoal Familiar	**0,502****
	(0,227)
Considerar ter tempo para si e para os seus hobbies	**0,844*****
	(0,225)
Considera ter flexibilidade para gerir o trabalho	0,068
	(0,167)

Notas: *p<0,1; **p<0,05; ***p<0,01
Foram omitidos da tabela os coeficientes associados às diferentes carreiras, idade e educação, apesar destas características terem sido incluídas na regressão

Tabela A9 – Análise relativa à escolarização dos respondentes

	Discorda da redução da jornada semanal de trabalho (binária)					
Tem apenas Ensino Básico	-0,256					
	(0,276)					
Tem apenas Ensino Primário		1,064*				
		(0,553)				
Tem apenas Ensino Secundário			-0,222			
			(0,178)			
Tem apenas Ensino Pós-Secundário				-0,073		
				(0,384)		
Tem apenas Licenciatura					-0,013	
					(0,168)	
Tem apenas Mestrado						0,205
						(0,172)
Doutoramento						0,139
						(0,246)

Notas: *p<0,1; **p<0,05; ***p<0,01
As colunas da tabela correspondem a diferentes regressões. Foram omitidos da tabela os coeficientes associados às diferentes carreiras, características demográficas e perceções da análise de concordância, apesar destas características terem sido incluídas na regressão.

Tabela A10 – Análise relativa às idades dos respondentes

	Discorda da da redução da jornada semanal de trabalho (binária)				
Idade inferior a 25 anos	-1,003				
	(0,763)				
Idade entre 25 e 34 anos		-0,458			
		(0,335)			
Idade entre 35 e 44 anos			-0,096		
			(0,213)		
Idade entre 45 e 54 anos				-0,118	
				(0,124)	
Idade superior ou igual a 55 anos					**0,456*** **
					(0,144)

Notas: *p<0,1; **p<0,05; ***p<0,01
Foram omitidos da tabela os coeficientes associados às diferentes carreiras, características demográficas e perceções da análise de concordância, apesar destas características terem sido incluídas na regressão.

3.2. Análise dos fatores associados à perceção de necessidades de investimento

Tabela A11 – Análise relativa ao sexo e características do agregado dos respondentes

	Identifica necessidades de investimento (binária)
Sexo Masculino	**-0,043*** **
	(0,011)
Tem Crianças a Cargo	-0,005
	(0,009)
Tem Adultos a Cargo	0,010
	(0,009)

Notas: *p<0,1; **p<0,05; ***p<0,01
Foram omitidos da tabela os coeficientes associados às diferentes carreiras, idade e educação, apesar destas características terem sido incluídas na regressão

Tabela A12 – Análise relativa às carreiras dos respondentes

	Identifica necessidades de investimento (binária)								
Carreiras Segurança	**0,066*** **								
	(0,018)								
Carreiras Saúde		**0,056*** **							
		(0,012)							
Carreiras na Justiça e Notariado			**0,066*** **						
			(0,019)						
Carreiras Inspecao				**0,043*** **					
				(0,012)					
Carreiras Educação (exceto Ensino Superior)					**0,033*** **				
					(0,009)				
Carreiras Dirigentes						-0,023			
						(0,023)			
Carreiras Gerais							**-0,117*** **		
							(0,010)		
Carreiras Informática								**-0,118*** **	
								(0,044)	
Carreiras Ensino Superior e Investigação Científica									-0,019
									(0,024)

Notas: *p<0,1; **p<0,05; ***p<0,01
As colunas da tabela correspondem a diferentes regressões. Foram omitidos da tabela os coeficientes associados às características demográficas e de agregado, apesar destas características terem sido incluídas na regressão

Tabela A13 – Análise relativa à escolarização dos respondentes

	Identifica necessidades de investimento (binária)					
Tem apenas Ensino Básico	-0,022					
	(0,030)					
Tem apenas Ensino Primário		0,039				
		(0,053)				
Tem apenas Ensino Secundário			-0.006			
			(0.017)			
Tem apenas Ensino Pós-Secundário				**0,039****		
				(0,019)		
Tem apenas Licenciatura					0,006	
					(0,009)	
Tem apenas Mestrado						-0,008
						(0,011)
Doutoramento						-0,003
						(0,024)
Notas:	*p<0,1; **p<0,05; ***p<0,01 As colunas da tabela correspondem a diferentes regressões. Foram omitidos da tabela os coeficientes associados às características demográficas, de caracterização do agregado familiar, e carreira dos respondentes, apesar destas características terem sido incluídas na regressão					

Tabela A14 – Análise relativa à idade dos respondentes

	Identifica necessidades de investimento (binária)					
Idade inferior a 25 anos	0,035					
	(0,038)					
Idade entre 25 e 34 anos		**0,044*****				
		(0,011)				
Idade entre 35 e 44 anos			0,017			
			(0,011)			
Idade entre 45 e 54 anos				-0,004		
				(0,009)		
Idade superior ou igual a 55 anos						**-0,040*****
						(0,010)
Notas:	*p<0,1; **p<0,05; ***p<0,01 As colunas da tabela correspondem a diferentes regressões. Foram omitidos da tabela os coeficientes associados às diferentes carreiras, características demográficas e perceções da análise de concordância, apesar destas características terem sido incluídas na regressão.					

3.3. Análise dos fatores associados à perceção de necessidades de ajustamento

Tabela A15 – Análise relativa ao sexo e características do agregado familiar dos respondentes

	Identifica necessidades de ajustamento (binária)
Sexo Masculino	**-0,034****
	(0,017)
Tem Crianças a Cargo	0,011
	(0,018)
Tem Adultos a Cargo	-0,018
	(0,017)
Notas:	*p<0,1; **p<0,05; ***p<0.01 Foram omitidos da tabela os coeficientes associados às diferentes carreiras, idade e educação, apesar destas características terem sido incluídas na regressão

Tabela A16 – Análise relativa às carreiras dos respondentes

	Identifica necessidades de ajustamento (binária)
Carreiras Segurança	**0,101*****
	(0,038)
Carreiras Saúde	0,023
	(0,021)
Carreiras na Justiça e Notariado	**0,124*****
	(0,031)
Carreiras Inspeção	**0,046*****
	(0,017)
Carreiras Educação (exceto Ensino Superior)	**0,138*****
	(0,014)
Carreiras Dirigentes	**-0,062****
	(0,027)
Carreiras Gerais	**-0,200*****
	(0,020)
Carreiras Informática	**-0,116****
	(0,050)
Carreiras Ensino Superior e Investigação Científica	**-0,132*****
	(0,032)
Notas:	*p<0,1; **p<0,05; ***p<0,01 As colunas da tabela correspondem a diferentes regressões. Foram omitidos da tabela os coeficientes associados às características demográficas e de agregado, apesar destas características terem sido incluídas na regressão

Tabela A17 – Análise relativa à escolarização dos respondentes

	Identifica necessidades de ajustamento (binária)					
Tem apenas Ensino Básico	-0,124					
	(0,087)					
Tem apenas Ensino Primário		-0,065				
		(0,127)				
Tem apenas Ensino Secundário			0,012			
			(0,036)			
Tem apenas Ensino Pós-Secundário				-0,002		
				(0,064)		
Tem apenas Licenciatura					0,001	
					(0,016)	
Tem apenas Mestrado						0,009
						(0,018)
Doutoramento						**0,064****
						(0,031)

Notas: *p<0.1; **p<0.05; ***p<0.01
As colunas da tabela correspondem a diferentes regressões. Foram omitidos da tabela os coeficientes associados às diferentes carreiras, características demográficas e perceções da análise de concordância, apesar destas características terem sido incluídas na regressão.

Tabela A18 – Análise relativa às idades dos respondentes

	Identifica necessidades de ajustamento (binária)				
Idade inferior a 25 anos	**0,172****				
	(0,067)				
Idade entre 25 e 34 anos		-0,058			
		(0,044)			
Idade entre 35 e 44 anos			-0,009		
			(0,018)		
Idade entre 45 e 54 anos				**0,044*****	
				(0,015)	
Idade superior ou igual a 55 anos					-0,033*
					(0,019)

Notas: *p<0,1; **p<0,05; ***p<0,01
As colunas da tabela correspondem a diferentes regressões. Foram omitidos da tabela os coeficientes associados às diferentes carreiras, características demográficas e perceções da análise de concordância, apesar destas características terem sido incluídas na regressão.

4. Análise de tópicos em respostas em campos de resposta aberta

Tendo em conta a natureza aberta das questões em causa e o número de respostas, foi utilizada uma abordagem de análise de texto por via computacional, conhecida como análise de tópicos, de forma a identificar grupos de respostas que fossem semelhantes entre si. Esta secção detalha o procedimento seguido, apresentando breves justificações e explicações do método.

Primeiramente, é realizada uma etapa de pré-processamento dos dados. Isso envolve a remoção de caracteres estranhos, o que ajuda a limpar o texto e a garantir consistência nos dados. A substituição do termo "desburocratização" e "desburocratizar" por "menos burocracia" foi imposta manualmente de forma a padronizar os termos utilizados nos dados, evitando variações que prejudicavam a análise. A tradução do texto das respostas para inglês, com recurso ao tradutor da Baidu, foi realizada de forma a permitir a utilização de modelos de linguagem natural mais complexos, que se encontram pre-treinados com dados em inglês.

Em seguida, os textos são transformados em *embeddings*, ou seja, representações vetoriais que capturam o significado semântico dos documentos. Para isso, é utilizada a biblioteca "sentence_transformers" juntamente com o modelo de linguagem natural "all-mpnet-base-v2" disponível na plataforma HuggingFace. Esses modelos pré-treinados são capazes de mapear os textos para um espaço vetorial de alta dimensionalidade, preservando as relações semânticas entre as frases respondidas. Em termos simples, as respostas foram codificadas em vetores, sendo que respostas semanticamente semelhantes estão mais próximas no espaço vetorial do que respostas menos relacionadas.

Para reduzir a dimensionalidade dos *embeddings* – isto é, para reduzir o número de dimensões dos vetores que representam cada resposta – é aplicado o algoritmo "UMAP". Este algoritmo realiza uma redução não linear da dimensionalidade, preservando as estruturas de similaridade entre as respostas (identificadas pela proximidade entre vetores). A pertinência de se reduzir a dimensionalidade prende-se com os conhecidos problemas de "maldição da dimensionalidade" que afetam especialmente os algoritmos de *clustering*. Os parâmetros deste algoritmo são ajustados para otimizar a qualidade da redução dimensional. Este processo é manual e dependente do utilizador, que ajusta os parâmetros iterativamente até se aproximar de uma solução adequada.

Na etapa de redução de dimensionalidade dos *embeddings* utilizando o algoritmo "UMAP", foram otimizados os seguintes parâmetros (os restantes foram empregues conforme pré-definidos pela biblioteca utilizada):

- **n_neighbors**: Este parâmetro determina o número de vizinhos mais próximos considerados ao construir o grafo de vizinhança utilizado pelo "UMAP". Um valor maior pode capturar estruturas de dados mais complexas, mas também pode introduzir ruído. No caso específico, foi definido como 20, o que pode ser considerado moderado, permitindo uma boa captura das relações locais entre os documentos.

- **n_components:** Indica a dimensionalidade da representação reduzida. Neste caso, foi definido como 5, o que significa que os *embeddings* originais serão reduzidos para um espaço vetorial de 5 dimensões. Escolher um número adequado de dimensões é crucial para garantir que a estrutura dos dados seja preservada, ao mesmo tempo em que se reduz a dimensionalidade para facilitar a visualização e a interpretação.
- **min_dist:** Este parâmetro controla a distância mínima entre pontos no espaço de representação reduzido. Um valor menor indica que pontos podem estar mais próximos uns dos outros, enquanto um valor maior impõe uma separação mais rígida. Foi definido como 0,0, o que significa que não há restrição para a distância mínima entre pontos, facilitando a tarefa de clustering a realizar posterior.
- **metric:** Determina a métrica de distância utilizada para calcular a proximidade entre os pontos no espaço de representação reduzido. Neste caso, foi escolhida a métrica *"cosine simmilarity"*, que mede a similaridade entre os vetores através da sua direção (ângulo entre eles), independentemente da sua magnitude.

De seguida, agruparam-se os documentos utilizando o algoritmo "HDBSCAN". Este algoritmo é especialmente adequado para identificar clusters de diferentes densidades e formas nos dados. Baseado no algoritmo "DBSCAN", as principais características desta estratégia de *clustering* são o recurso às noções de densidade de pontos no espaço e vizinhança entre pontos. Adicionalmente, este algoritmo identifica "ruído" nos dados, pontos em regiões do espaço com relativamente pouca densidade e com vizinhos muito distantes. Os parâmetros deste algoritmo são ajustados também iterativamente, de forma manual e dependente do utilizador, com base na adequabilidade dos grupos resultantes. O critério seguido foi a definição de *clusters* bastante puros como base da estratégia aglomerativa. Ou seja, primeiro definem-se grupos de respostas muito semelhantes e, posteriormente, agrupam-se em grupos de grupos com base na semelhança entre estes.

Na etapa de *clustering* com "HDBSCAN", foram otimizados os seguintes parâmetros (os restantes foram utilizados conforme pre-definidos pela biblioteca utilizada):

- **min_cluster_size:** Define o tamanho mínimo que um *cluster* pode ter. Este parâmetro influencia diretamente na sensibilidade do algoritmo em identificar *clusters* pequenos. Foi definido como 10, o que indica que *clusters* com menos de 10 pontos serão considerados como ruído e não serão agrupados.
- **cluster_selection_method:** Este parâmetro determina o método utilizado para selecionar os *clusters*. Foi escolhido o método 'leaf', que conduz à construção de *clusters* mais puros que outros métodos.

Após o *clustering*, é realizada uma representação dos tópicos para análise manual. A biblioteca "scikit-learn" é utilizada, especialmente a função "CountVectorizer", para realizar essa representação. Neste

processo, são removidas as *stopwords* (palavras com pouca relevância semântica) e são criados n-gramas – palavras-chave multitermo entre 1 e 3 termos – para sumarizar cada tópico.

Por fim, é feita a aglomeração manual de tópicos. Isso envolve a análise dos *clusters* identificados (com recurso à representação feita no passo anterior) e o seu agrupamento ou separação com base na semelhança do conteúdo (tendo em conta análise feita pelo utilizador). O resultado desta etapa deu origem aos tópicos e subtópicos reportados nas Tabelas 32 e 33 da Nota de Análise.

5. Referências

L. McInnes, J. Healy, S. Astels, 2017, “hdbscan: Hierarchical density based clustering”, Journal of Open Source Software, The Open Journal, volume 2, number 11.

L. McInnes, J. Healy, J. Melville. 2018, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," ArXiv e-prints.

N. Reimers, I. Gurevych, 2019, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.

Z. He, 2015, "Baidu Translate: Research and Products," Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra), pp. 61-62.

G. Zou, 2004, "A Modified Poisson Regression Approach to Prospective Studies with Binary Data" American Journal of Epidemiology, vol. 159, no. 7, pp. 702-706.

www.planapp.gov.pt

PlanAPP

@planapp_

Newsletter

PLANAPP
PLANEAMENTO | POLÍTICAS | PROSPETIVA