Análise de Componentes Principais e Análise de Agrupamento dos Municípios do Brasil

Weidmam Milagres Leles Nº 95410

Trabalho Individual da Unidade Curricular de Reconhecimento de Padrões

Docente: Prof. Doutor José G Dias

19 de dezembro de 2020

Índice

Introdução

Apoiado na visão filosófica desenvolvida por Barros, Henriques e Mendonça (2001) de que o Brasil não pode ser classificado como um país pobre, mas com muitos pobres, este trabalho pretende identificar, por meio da Análise de Componentes Principais e Análise de Agrupamento de Dados, os perfis dos municípios brasileiros no sentido de estimular o debate sobre as prioridades das ações governamentais que enfatizam o papel de políticas redistributivas à luz dos aspetos distintivos da sociedade brasileira.

Põe-se ênfase no facto de que o Brasil, embora seja a 9º maior economia mundial, é um país desigual, com notória objeção imposta pelo contexto de forte injustiça social, económica e política. É esta grande discrepância vista no Brasil que motiva a realização deste trabalhado, uma vez que é importante que sejam desenvolvidas políticas públicas adequadas e que levem em consideração toda a diversidade brasileira.

Isto posto, com o intuito de conhecer melhor os municípios brasileiros para que políticas públicas sejam adequadamente desenhadas e aplicadas, utiliza-se a Análise de Componentes Principais, com o objetivo de reduzir a dimensionalidade da base de dados e consequentemente aumentar a sua interpretabilidade, além de Identificar novas variáveis subjacentes significativas.

Posteriormente, faz-se a Análise de Agrupamento de Dados, técnica que auxiliará na perceção de como os municípios brasileiros podem ser caracterizados por meio de um conjunto de padrões representativos e intrínsecos com o intuito de agrupá-los.

Dados

O conjunto de dados a ser analisado é uma coleção retirada do site Kaggle.com composta por 79 atributos para cada um dos 5 070 municípios do Brasil e do Distrito Federal – Brasília [1]. Esta base de dados é uma compilação de diversos indicadores, como população, gastos municipais Índice de Desenvolvimento Humano, Produto Interno Bruto, área, valor acrescentado bruto e vários outros dados socioeconómicos.

Neste sentindo, utiliza-se as seguintes variáveis ativas (INPUT) para fazer a análise dos dados:

PER_POP_REGULAR: Percentagem da população a viver em áreas com ordenamento urbano regular;
IDHM: Índice de Desenvolvimento Humano
COMP_CAPITA: Quantidade de empresas por habitantes
GAS_MUN_CAPITA: Gastos municipais por habitantes
TAXES_CAPITA: Impostos, líquidos de subsídios, sobre produtos, a preços correntes por habitantes
GVA_CAPITA: Valor adicionado bruto total, a preços correntes por habitantes
GDP_CAPITA: Produto interno bruto per capita

Uma vez determinadas as variáveis de INPUT, fez-se a limpeza dos dados excluindo as entradas em que havia valores não atribuídos para alguma das variáveis ativas (INPUT). Deste modo, 4 075 municípios brasileiros serão analisados, ou seja, 80,37% dos municípios.

Posteriormente, em termos de PROFILE, estes grupos serão analisados tendo em consideração as dimensões do IDH, densidade populacional, divisão administrativa dos estados (se é capital ou não), valor acrescentado bruto, e setores de atividade empresarial.

Identificação das dimensões da análise

A Análise de Componentes Principais (PCA – Principal Component Analysis) é uma técnica estatística que resume os dados, por meio da ortogonalização de vetores, encontrando as principais correlações em combinações lineares das observações, transformando um conjunto de variáveis correlacionadas em um conjunto menor de variáveis não correlacionadas.

No sentido de conduzir tal análise, serão realizados cinco passos: (1) Verificar adequação para conduzir a Análise de Componentes Principais; (2) Extração das Componentes Principsi; (3) Rotação e interpretação das Componentes Principais; (4) Tomada de decisão final sobre o número de Componentes Principais; (5) Criar pontuações.

Passo 1: adequação da Análise de Componentes Principais

[1] Conforme o Art. 32 da Constituição Federal Brasileira de 1988, Ao Distrito Federal são atribuídas as competências legislativas reservadas aos Estados e Municípios, pelo que não podemos classificá-lo como um município.

A análise da **matriz dos gráficos de dispersão** é uma boa maneira de iniciar este processo de verificar se a realização da Análise de Componentes Principais é adequada ao conjunto de dados em questão, uma vez que por meio desta matriz, é possível perceber melhor as variáveis e como elas se corelacionam, conforme abaixo:

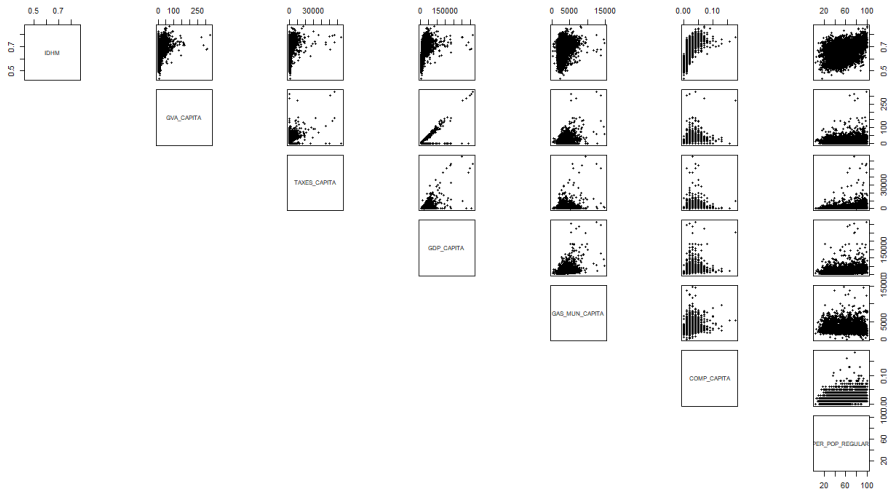

Figura 1 - Matriz dos Gráficos de Dispersão

Quando as correlações entre as variáveis são fracas, realizar a Análise de Componentes Principais pode não ser muito útil, já que não é possível fazer uma redução significativa da nossa base de dados sem perder uma quantidade considerável de informações. Como podemos verificar na Figura 1, as variáveis *GVA_CAPITA* e *GPD_CAPITA* são as variáveis mais correlacionadas, seguida da variável *IDH* e *COMP_CAPITA*.

Complementarmente, outro gráfico que é extremamente útil antes de se iniciar a Análise de Componentes Principais propriamente dita é o **Gráfico de Correlação** (Figura 2), pois através dos Coeficientes de Correlação, conseguimos ver com mais clareza o quão correlacionadas estão as variáveis[2]. No caso deste conjunto de dados, as variáveis ativas em sua grande maioria, estão correlacionadas de forma positiva, com exceção da *PER_POP_REGULAR* e *GAS_MUN_CAPITA.* Isto é, a relação entre as despesas municipais e a percentagem de habitantes que residem em áreas regulamente planeadas, embora seja muito próxima de nula, tem uma ligeira relação negativa, o que significa que quanto maior for esta percentagem, menor será o gasto municipal.

As demais variáveis estão correlacionadas positivamente, seja em menor (azul claro e arredondado) ou maior grau (azul escuro e forma de elipse). É importante notar que os coeficientes de correlação maiores que 0,3 em valor absoluto são bons indicadores para se preformar a PCA.

[2] Veja ANEXO I para detalhes da matriz de correlação entre as variáveis Ativas.

Figura 2 - Gráfico de Correlação

Uma vez identificado que a maioria das variáveis desta base de dados apresentam uma correlação superior a 0.3, para dar sequência a Analise de Componentes Principais realizou-se o **Teste de Esfericidade de Bartett**[3] e verificou que a PCA seria um método eficaz para a redução da nossa base de dados. Adicionalmente, fez-se a avaliação de uma outra premissa para dar sequência na Análise de Componentes Principais, o **teste de Kaiser-Meyer Olkin (KMO)**[4]. Este teste foi importante, pois foi possível verificar que não seria adequado prosseguir com a variável *PER_POP_REGULAR* com uma variável ativa nesta análise, pelo que foi removida.

Após a análise da matriz dos gráficos de dispersão, do Gráfico de Correlação e de performar e verificar resultados satisfatórios por meio do Teste de Esfericidade de Bartett e do teste de Kaiser-Meyer Olkin (KMO) realizou os ajustes necessários para se obter um melhor aproveitamento da Análise de Componentes Principais, a qual é adequada ao conjunto de dados em questão.

Com este primeiro passo, preparamos o nosso conjunto de variáveis ativas para que a Análise de Componentes Principais fosse realizado, o que permite o avanço para o passo seguinte, o de extração de componentes principais.

Passo 2: extração de Componentes Principais[5]

É neste segundo passo que efetivamente começa a análise das Componentes Principais, ou seja, das novas variáveis que serão criadas para explicarmos a nossa base de dados. Embora a decisão final sobre quantidade de componentes principais que vão ser criadas seja feita apenas no passo

[3] O de **Teste de Esfericidade de Bartett** é usado para testar a hipótese nula de que o as variáveis não estão correlacionadas na população. Isto é, se as amostras são de populações com variâncias iguais, não podemos rejeitar a hipótese nula do teste e, portanto, a PCA é inadequada. No caso em questão, o P-Value é menor que 0.05, pelo que a PCA é uma técnica adequada ao nosso conjunto de dados.

[4] Teste realizado para validar se os dados das variáveis ativas são adequados no sentido de se realizar a Análise de Componentes principais. Veja Anexo II para detalhes.

[5] Veja Anexo III para detalhes técnicos dos procedimentos realizados neste passo.

4, é no passo 2 em que se inicia as análises e colhe-se bons indicadores de quantas componentes serão necessárias para poder explicar o conjunto de dados. Com isto, o objetivo neste passo reside em criar variáveis, que são chamadas de componentes principais, que sejam capazes de explicar uma boa variância da nossa base de dados. É desejável que seja gerado poucas componentes, uma vez que o objetivo da Análise de Componentes principais é reduzir a dimensão da base de dados.

Com efeito, as componentes principais serão as novas variáveis construídas por meio de combinações lineares das variáveis iniciais. Essas combinações serão feitas de tal forma que as componentes principais não estarão correlacionadas e a maioria das informações dentro das variáveis iniciais será comprimida nos primeiros componentes. Em outras palavras, a ideia é que dado as 6 variáveis ativas obtém-se 6 componentes principais. Porém, o máximo de informações possíveis é inserido logo no primeiro componente, o máximo de informações restantes no segundo e assim por diante.

É importante destacar que ao analisar as quantidades totais de variação que uma variável partilha com todas as outras variáveis da análise, foi necessário remover a variável *GAS_MUN_CAPITA.* Após a remoção da variável *GAS_MUN_CAPITA*, 5 variáveis serão utilizadas como variáveis de INPUT:

IDHM: Índice de Desenvolvimento Humano
COMP_CAPITA: Quantidade de empresas por habitantes
TAXES_CAPITA: Impostos, líquidos de subsídios, sobre produtos, a preços correntes por habitantes
GVA_CAPITA: Valor adicionado bruto total, a preços correntes por habitantes
GDP_CAPITA: Produto interno bruto per capita

Com estas cinco variáveis ativas, os testes indicam que será possível reduzi-las em dois componentes, os quais conseguirão explicar 82% da variabilidade geral dos dados originais. Esta é uma boa percentagem, pois recomenda-se que as componentes extraídas levem em conta pelo menos 60% da variância.

Passo 3 e 4 : Rotação e Interpretação das Componentes Principais e Tomada de decisão final sobre o número de Componentes Principais[6]

Nesta etapa, no sentido de se confirmar que serão retidas duas Componentes Principais, é importante mais algumas interações com os nossos dados. Assim sendo, girar estas Componentes Principais será o próximo passo. A rotação das Componentes Principais tem o intuito de torná-las mais significativas e fáceis de interpretar, já que cada variável está associada a um número mínimo de PCs. Em outras palavras, as rotações são feitas para fins de interpretação dos componentes extraídos na PCA em busca de alguma estrutura da matriz de

[6] Veja anexo IV para mais detalhes sobre os procedimentos técnicos realizados no passo 3 e 4.

loadings, que pode ser chamada de estrutura simples: quando diferentes Componentes Principais tendem a carregar diferentes variáveis.

Utilizou-se o método Varimax para fazer esta rotação e finalmente foi possível confirmar que duas componentes principais serão criadas:

- **Componente Principal 1** – Componente com alto *loadings* de *GVA_CAPITA*, *TAXES_CAPITA* e *GDP_CAPITA*, variáveis intimamente ligadas à Produção Económica de cada Município do Brasil. Deste modo, a Componente Principal 1 chamará **Produção;**
- **Componente Principal 2** – Componente com alto *loadings* de *IDHM* e *COMP_CAPITA,* variáveis que estão intimamente ligadas ao Desenvolvimento Humano, como expectativa de vida ao nascer, educação e PIB per capita, e a quantidade de empresas no município, as quais são capazes não só aumentarem a oferta de emprego, mas também de oferecem trabalhos dignos. Embora com diversas limitações em termos de definição, esta Componente Principal, de alguma maneira, é capaz de quantificar o **Bem-Estar**[7] dos Munícipes.

Passo 5: Criar pontuações

Ao adentrar no último passo da Análise de Componentes Principais, é necessário criar as pontuações para as variáveis. Ou seja, cria-se variáveis, neste caso as variáveis Produção e Bem-Estar, que são simplesmente a soma ponderada das pontuações das variáveis, onde os pesos são os autovalores.

A representação gráfica das pontuações de cada um dos municípios brasileiros, pode ser visualizada por meio da Figura 3. Deste modo, nota-se há uma grande concentração à esquerda ao redor do nível 0 de bem-estar e de produção. No entanto, é possível também notar que há uma tendência negativa na medida em que produção aumenta o bem-estar diminui, porém não é uma situação que se aplica a todos municípios brasileiros, conforme pode-se observar a seguir:

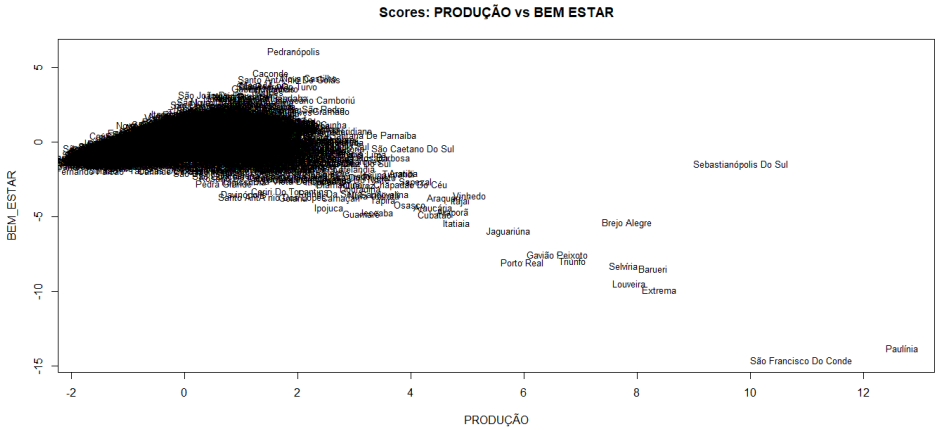

Figura 3 - Pontuações: PRODUÇÃO VS BEM-ESTAR

[7] Nota-se que o conceito de Bem-Estar é um conceito subjetivo e de difícil quantificação. Este trabalho não tem intenção e nem a pretensão de trazer uma nova definição do que é Bem-Estar para a sociedade brasileira. Deste modo, este trabalho apenas considera que o emprego e o IDH têm influências no Bem-Estar dos indivíduos.

Uma vez que as componentes principais estão criadas tona-se muito mais simples identificar padrões e perceber melhor se é possível dividir os municípios brasileiros em grupos e se for possível, em quantos grupos eles podem ser agrupamentos.

Deste modo, o que pretende a partir de agora é utilizar a produção e o bem-estar de cada um dos municípios brasileiros para verificar a heterogeneidade do Brasil.

Identificação da heterogeneidade dos municípios brasileiros: Ecléticos, Administradores e Transformadores

O maior e mais populoso país da América do Sul, o Brasil continua a buscar o crescimento industrial e agrícola e o desenvolvimento de seu interior. Tendo enfrentado com sucesso um período de dificuldade financeira global no final do século 20, o Brasil é visto como um dos mercados emergentes mais fortes do mundo. No entanto, a disparidade entre ricos e pobres continua a ser um dos grandes problemas do Brasil. Deste modo, tendo realizado a Análise de Componentes Principais, é interessante prosseguir com análise adicional dos Municípios do Brasil por meio da Identificação das diferentes realidades existentes no país, a qual objetiva auxiliar a compreensão dos municípios brasileiros no sentido de orientar as prioridades das ações governamentais que enfatizem o papel de políticas redistributivas à luz dos aspetos distintivos da sociedade brasileira.

Utilizou-se, complementarmente, a abordagem hierárquica e de particionamento[8] para que fosse possível dividir os municípios brasileiros em três diferentes grupos, conforme é possível ver na Figura 4, os quais receberam os nomes **Ecléticos, Administradores e Transformadores**, tendo como referência a atividade que mais contribui para o VAB - Valor Acrescentado Bruto de cada um dos grupos.

O grupo **Ecléticos** é formado por 2027 municípios, o que representa 49,74% dos municípios analisados, e é o grupo que tem os Demais Serviços[9] como a atividade que mais contribui para o VAB;

O grupo **Administradores** é formado por 2000 municípios, o que representa 49,07% dos municípios analisados, e é o grupo que tem o setor da Administração, defesa, educação e saúde públicas e seguridade social como as atividades que mais contribuem para o VAB;

O grupo **Transformadores** é composto por 48 municípios, o que representa 1,17% dos municípios analisados e é o grupo que tem as Indústrias de transformação como a atividade que mais contribui para o VAB. Embora o grupo Transformadores represente apenas 1,17% dos municípios, pelo seria possível concluir que seriam municípios outliers, estes 48 municípios de têm um impacto de **7,25%** no PIB brasileiro, pelo que representam um nicho de municípios com altíssima produção.

[8] A análise de agrupamento de dados é essencialmente uma abordagem exploratória. Veja ANEXO V – AGRUPAMENTO DOS DADOS para mais informações.

[9] A Classe Demais Serviços compreende a agregação dos setores: Transporte, armazenagem e correio; Alojamento e alimentação; Informação e comunicação; Atividades financeiras, de seguros e serviços relacionados; Atividades imobiliárias; Atividades profissionais, científicas e técnicas, administrativas e serviços complementares; Educação e saúde privadas; Artes, cultura, desporto e recreação e outras atividades de serviços e serviços domésticos.

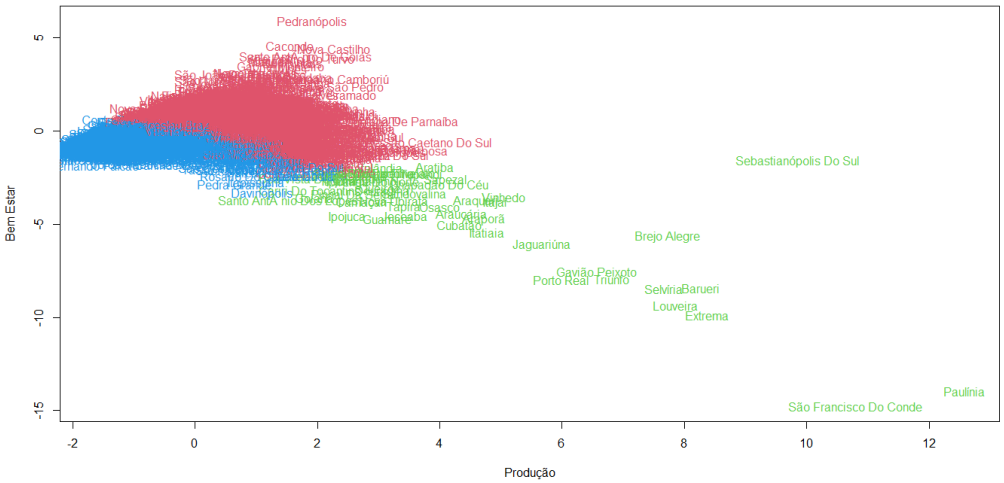

Figura 4 - Bem-Estar vs Produção dos municípios brasileiros divididos em três grupos

Em termos geográficos, é possível identificar por meio da análise da Figura 5, que os municípios Ecléticos estão concentrado mais ao sul e sudeste do Brasil, com algumas aglomerações também no centro-oeste brasileiro; o grupo Administradores tem uma concentração mais proeminente no nordeste e com alguns destaques no norte do país; O grupo Transformadores é um grupo menor comparativamente com os outros dois, no entanto, é o grupo em que os municípios estão mais dispersos, conforme é possível analisar a seguir.

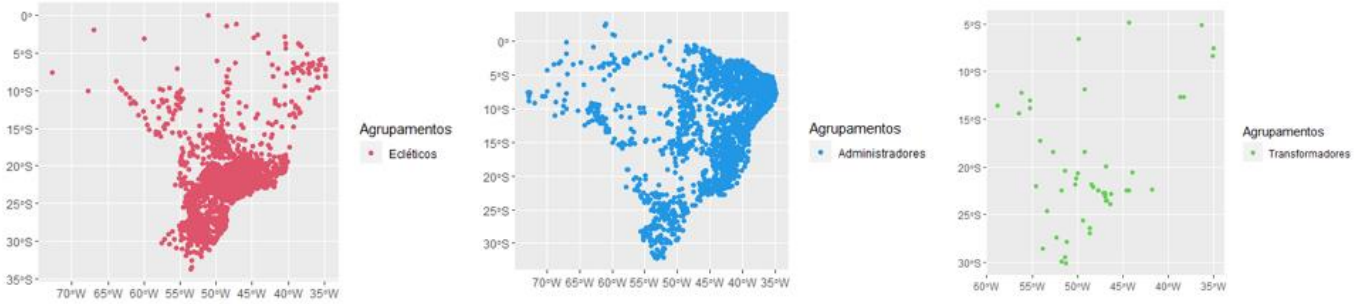

Figura 5 - Distribuição geográfica dos agrupamentos dos municípios brasileiros

Assim sendo, é possível constatar que mesmo que a localização dos municípios não tenha sido objeto de análise deste estudo, verifica-se que o padrão de Bem-Estar e de Produção está também associada à região em que este município se encontra. Deste modo, pode-se inferir que a região sul e sudeste do Brasil, tem um melhor equilíbrio entre o nível de produção e de bem-estar comparativamente com os outros dois grupos, o que vai ao encontro com o que podemos observar na Tabela 1.

	Valor Acrescentado Bruto (VAB)[10]	Capitais	Renda (IDH)	Educação (IDH)	Longevidade (IDH)	DENSIDADE POPULACIONAL	DIVISÃO DA POPULAÇÃO
ECLÉTICOS 2027 municípios 49,74% dos municípios	**51%** dos municípios tem os Demais Serviços como a atividade que mais contribui para o VAB.	**95,46%** das capitais dos estados estão neste grupo	**60%** dos municípios têm rendimento **alto**, **2 %** tem rendimento **muito alto** e **36%** tem rendimento **médio**	**77%** dos municípios tem um nível de educação **médio** e **13%** tem um nível **alto**	**93%** dos municípios tem um nível de longevidade **muito alto**	**50,71%** dos municípios têm densidade populacional **média**	**71,5 %** da população do Brasil vive em um município deste grupo
ADMINISTRADORES 2000 municípios 49,07% dos municípios	**76%** dos municípios tem a Administração, defesa, educação e saúde públicas e seguridade social como as atividades que mais contribuem para o VAB.	**4,54%** das capitais dos estados estão neste grupo	**69%** dos municípios têm rendimento **médio** e **25%** tem rendimento **baixo**	**78%** dos municípios tem um nível educacional **baixo** e **20,8%** tem um nível educacional **médio**	**69,4%** dos municípios tem um nível de longevidade **muito alto**	**50,35%** dos municípios têm densidade populacional **baixa**	**26,8%** da população vive em um município deste grupo
TRANSFORMADORES 48 municípios 1,17% dos municípios	**36%** dos municípios tem as Indústrias de transformação como a atividades que mais contribuem para o VAB. A Agricultura, inclusive apoio à agricultura e a pós colheita e os Demais Serviços contribuem **16%** cada.	Não tem nenhuma capital de estado neste grupo	**54%** dos municípios têm rendimento **alto** e **39%** tem rendimento **médio**	**70%** dos municípios tem um nível de educação **médio**, **14,58%** tem um nível alto e **14,58%** tem um nível **baixo**	**91,48%** dos municípios tem um nível de longevidade **muito alto**	**41,66%** dos municípios têm densidade **baixa** e **37,5%** tem densidade **alta**	**1.7%** da população do Brasil vive em um município deste grupo

Tabela 1 - Caracterização dos Agrupamentos

Escala da Renda, Educação e Longevidade:
Até 0.555 – Baixo
De 0.556 até 0.700 – Médio
De 0.701 até 0.800 – Alto
Maior que 0.800 – Muito alto

Escala Densidade Populacional:
Até 20 hab/km ^{2} – Baixo
De 21 hab/km ^{2} até 100 hab/km ^{2} – Médio
Maior que 100 hab/km ^{2} – Alto

[10] O VAB é uma medida muito importante, pois é usado para determinar o produto interno bruto (PIB). Além disso, o VAB é um bom indicador do bem-estar económico da população, uma vez que inclui todas as rendas primárias e justamente por isso foi utilizado como fator diferenciador dos grupos.

Ao observar a Tabela 1, é possível notar que o grupo **Administradores**, embora tenha uma quantidade de municípios muito próxima da do grupo **Ecléticos**, representa apenas **26,8%** da população total do Brasil, o que confirma a alta concentração não só populacional na Região Sudeste, mas também altas taxas de renda, educação e longevidade. O que por sua vez acaba por ser a causa e o efeito do "esvaziamento" de investimento privado na região nordeste e norte do Brasil.

Assim sendo, o grupo **Eclético** deve ser visto como o grupo que oferece melhores oportunidades de trabalho aos habitantes, uma vez que **92%** dos seus municípios oferecem uma muito boa ou boa relação entre quantidade de empresas por número de habitantes. Em contrapartida, no grupo **Administradores, 93,1%** dos municípios têm uma má ou média relação entre a quantidade de empresas por número de habitantes, o que significa que são menores as chances de se conseguir empregos formais e com boas condições de trabalho nos municípios do grupo **Administradores,** além de ser o grupo com o pior nível educacional.

Já o grupo **Transformadores** possui um bom nível de renda, de educação e de longevidade; em sua maioria, são municípios com baixa e média densidade populacional, porém apresentam os maiores níveis de produção e os menores níveis de bem-estar do país. Deste modo, podemos inferir que as cidades no canto inferior direito da Figura 4, têm alta dependência de algumas empresas para a sua alta produção e geração de empregos, o que acaba por prejudicar o bem-estar da população. Isto é, o alto nível de produção que estes municípios apresentam é oriundo de poucas empresas em tendo em vista a relação quantidade de empresas por habitantes.

Assim sendo, é importante percebermos como as empresas estão distribuídas (por setor) no território brasileiro:

SETOR DE ATIVIDADE	ECLÉTICOS	ADMINISTRADORES	TRANSFORMADORES
Agriculture, livestock, forestry, fishing and aquaculture	20,87	54,45	16,67
Extractive industries	63,64	82,50	56,25
Industries of transformation	3,16	23,55	6,25
Electricity and gas	93,04	98,20	89,58
Water, sewage, waste management	62,90	89,15	50,00
Construction	12,23	38,85	4,17
Trade; repair of motor vehicles and motorcycles	0,00	0,00	0,00
Transport, storage and mail	4,29	40,25	6,25
Accommodation and food	5,87	36,35	0,00
Information and communication	35,47	68,20	43,75
Financial, insurance and related services activities	49,58	87,80	45,83
Real estate activities	45,78	83,90	41,67
Professional, scientific and technical	13,42	46,30	18,75
Administrative activities and complementary services	11,35	36,25	10,42
Public administration, defense and social security	3,75	3,45	2,08
Education	12,48	27,00	12,50
Human health and social services	18,45	48,65	22,92
Arts, culture, sport and recreation	24,67	65,40	27,08
Other service activities	3,65	8,00	4,17
International and other extraterritorial institutions	99,51	99,95	100,00

Tabela 2- COMPARAÇÃO DAS PERCENTAGENS DE MUNICÍPIOS QUE NÃO POSSUI NENHUMA EMPRESA DE UM DETERMINADO SETOR DE ATIVIDADE

Conforme é possível observar na Tabela 2, o grupo **Administradores** é um grupo onde a iniciativa privada tem ainda um grande caminho por percorrer em comparação com os outros dois grupos, pois é alta a percentagem de municípios que não possuem nenhuma empresa de um determinado segmento. Tal facto tem ligação com o já havia sido constatado na Tabela 1, onde foi possível observar que **69%** dos municípios têm rendimento médio e **25%** dos municípios tem rendimento baixo, uma vez que o número de empresas por habitantes tem uma correlação positiva com o PIB per capita, variável com maior peso na produção dos municípios. Em contrapartida, a Administração pública, defesa e segurança social está bem presente nos municípios deste grupo, o que mostra uma boa presença dos governos de forma transversal neste grupo.

No que toca o grupo **Transformadores** e **Ecléticos**, nota-se que estes grupos são bem compostos pelos vários segmentos dos sectores de atividade, o que traduz as suas características de serem os grupos mais industrializados e onde se concentra a produção de riqueza do Brasil, o que por sua vez acaba por influenciar outros índices como educação, por exemplo.

Complementarmente, ao analisarmos comparativamente a proporção da população e o PIB de cada um dos agrupamentos, é possível identificar que há uma discrepância entre os grupos, uma vez que o **26,8%** da população brasileira vive nos municípios do grupo **Administradores**, porém apenas **13,2%** do Produto Interno Bruto brasileiro é gerado nesta região. Esta é uma realidade muito própria do grupo **Administradores**, pois a realidade tanto dos municípios Ecléticos e Transformadores é oposta. Em outras palavras, **71,5%** da população brasileira vive nos municípios **Ecléticos** e contribuem com **79.53%** do PIB. Já os municípios do grupo **Transformadores** contribuem com **7.25%** PIB e apenas 1,7% da população vive em um dos municípios deste grupo. Tal discrepância pode ser melhor percebida através do seguinte gráfico:

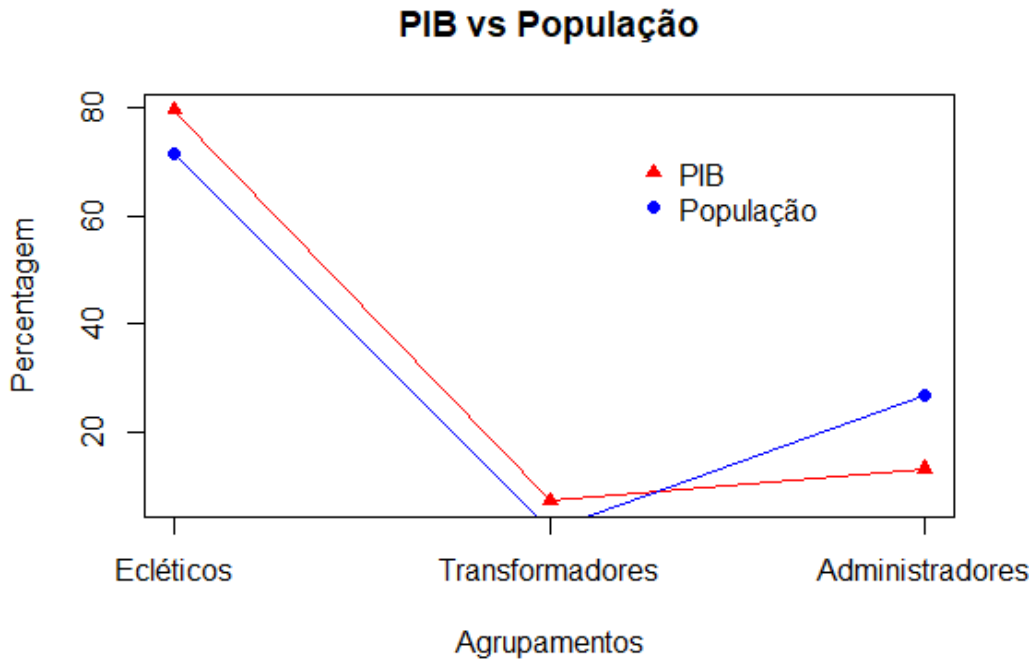

Figura 6 - Comparação entre o PIB e População em cada um dos agrupamentos

Conclusão

Neste trabalho foi realizada a Análise de Componentes Principais e Análise de Agrupamento de Dados com a intenção de identificar os perfis dos municípios brasileiros no sentido de incentivar a discussão sobre as prioridades das ações governamentais que enfatizam o papel de políticas redistributivas à luz dos aspetos distintivos da sociedade brasileira.

Como foi possível observar, o Brasil é um país com realidades bastantes distintas, mas ainda assim, foi possível agrupar os municípios brasileiros em três grupos: **Ecléticos, Administradores e Transformadores**. grupo **Eclético** é o grupo com maior produção de riqueza e com maior concentração da população, e por isso tem um grande setor de serviços já desenvolvido; o grupo **Transformadores** é um grupo pequeno, porém com municípios com alta produção de riqueza tendo em vista a sua população, e por isso é possível identificar alguns desequilíbrios na relação população e quantidade de empresas; o grupo **Administradores** é o grupo onde se apresenta o menor rendimento per capita, o pior nível de escolaridade e também a pior diversificação dos setores empresariais, mas ainda assim, representa uma boa parcela da população.

Levando em consideração os aspetos apontados no decorrer neste trabalho à luz da necessidade de firmar um debate sobre políticas redistributivas no Brasil, uma vez que o Brasil não pode ser considerado um país pobre, mas com muitos pobres, em termos de recomendações ao governo brasileiro, é possível destacar a necessidade de se desenvolver políticas no sentido de atrair empresas privadas dos mais diversos setores para os municípios do grupo administradores com o intuito de se quebrar o ciclo vicioso da pobreza nesta região.

No entanto, só atrair empresas não seria a solução para o problema, pois é possível notar no caso da educação, por exemplo, que há uma acentuada diferença em sua qualidade do grupo administradores para os outros dois grupos. Deste modo, é possível também concluir que as políticas publicas a serem desenvolvidas devem levar em consideração o papel que uma educação de qualidade pode desempenhar para tonar o Brasil um país mais justo para a sua população.

No que tange os municípios do grupo Transformadores, é preciso tecer políticas públicas em torno das atividades empresariais visando aumentar o bem-estar da população, uma vez que é interessante é que haja um balanço entre estas duas questões.

Por fim, tal análise não quer dizer que todos os municípios do grupo Eclético ou Transformadores são municípios ricos e que não há problemas internos dentro de cada grupo, pelo que tal recomendação também é válida para que haja uma redistribuição transversal a todos os municípios dos grupos, mesmo que embora esta análise micro não faça parte do escopo deste trabalho.

Referências Bibliográficas

[1] BARROS, R. P.; HENRIQUES, R.; MENDONÇA, R. A Estabilidade inaceitável: desigualdade e Pobreza no Brasil. IPEA, Textos para discussão n. 800, jun. 2001

[2] Abascal, Elena y Grande, Ildefonso. M´etodos multivariantes para la investigaci´on comercial, ArielEconomía, Barcelona, 1989.

ANEXOS

ANEXO I – MATRIZ DE CORRELAÇÃO DAS VARIÁVEIS ATIVAS

	IDHM	GVA_CAPITA	TAXES_CAPITA	GDP_CAPITA	GAS_MUN_CAPITA	COMP_CAPITA	PER_POP_REGULAR
IDHM	1.000	0.406	0.340	0.496	0.325	0.748	0.515
GVA_CAPITA	0.406	1.000	0.495	0.784	0.426	0.368	0.183
TAXES_CAPITA	0.340	0.495	1.000	0.684	0.283	0.282	0.232
GDP_CAPITA	0.496	0.784	0.684	1.000	0.516	0.442	0.209
GAS_MUN_CAPITA	0.325	0.426	0.283	0.516	1.000	0.301	-0.069
COMP_CAPITA	0.748	0.368	0.282	0.442	0.301	1.000	0.330
PER_POP_REGULAR	0.515	0.183	0.232	0.209	-0.069	0.330	1.000

ANEXO II – TESTE DE KAISER MEYER OLKIN (KMO)

Com base no resultado do **teste de Kaiser-Meyer Olkin (KMO)** o valor de KMO sugere que os dados das variáveis ativas são adequados[11] para a aplicação da PCA, exceto o da variável *PER_POP_REGLAR*. Quanto mais próxima a medida KMO estiver de 1, maior será a adequação de amostragem. Isto é, se o índice KMO for alto (>0.8), o PCA pode atuar de forma eficiente; se KMO é baixo (<0.5), o PCA não é relevante. Neste sentido, a variável *PER_POP_REGULAR* será removida da análise.

[11] Referência para avaliação Kaiser-Meyer Olkin (KMO): 0,8 e superior são ótimos; 0,7 é aceitável; 0,6 é medíocre; menos de 0,5 é inaceitável

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = correlation)
Overall MSA =  0.73
MSA for each item =
      IDHM GVA_CAPITA TAXES_CAPITA GDP_CAPITA GAS_MUN_CAPITA COMP_CAPITA PER_POP_REGULAR
      0.69        0.77         0.77       0.70          0.80        0.74           0.63
```

Figura 7 - Critério de Kaiser–Meyer–Olkin

ANEXO III – ESCOLHA DA QUANTIDADE DE COMPONENTES

Extraiu-se, as informações para 6 componentes no intuito de começar a análise de quantas componentes principais serão retidas. Chama-se a atenção para a variância das Componentes Principais, conforme marcação na Figura 8.

```
> pc6 <- principal(SCALED_DATA, nfactors=6, rotate="none", scores=TRUE)
> pc6
Principal Components Analysis
Call: principal(r = SCALED_DATA, nfactors = 6, rotate = "none",
    scores = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
                 PC1   PC2   PC3   PC4   PC5   PC6 h2       u2 com
IDHM            0.74  0.56 -0.07  0.04 -0.36  0.03  1 3.3e-16 2.4
GVA_CAPITA      0.80 -0.30 -0.02 -0.49  0.01  0.19  1 1.0e-15 2.1
TAXES_CAPITA    0.70 -0.37 -0.42  0.42  0.03  0.12  1 2.2e-16 3.1
GDP_CAPITA      0.90 -0.29 -0.07 -0.09 -0.01 -0.32  1 7.8e-16 1.5
GAS_MUN_CAPITA 0.62 -0.16  0.73  0.23  0.01  0.05  1 -4.4e-16 2.3
COMP_CAPITA     0.70  0.63 -0.04  0.00  0.34  0.00  1 -4.4e-16 2.5

                       PC1  PC2  PC3  PC4  PC5  PC6
SS loadings           3.34 1.05 0.73 0.48 0.25 0.16
Proportion Var        0.56 0.18 0.12 0.08 0.04 0.03
Cumulative Var        0.56 0.73 0.85 0.93 0.97 1.00
Proportion Explained  0.56 0.18 0.12 0.08 0.04 0.03
Cumulative Proportion 0.56 0.73 0.85 0.93 0.97 1.00

Mean item complexity =  2.3
Test of the hypothesis that 6 components are sufficient.

The root mean square of the residuals (RMSR) is  0
 with the empirical chi square  0  with prob <  NA

Fit based upon off diagonal values = 1
> |
```

Apenas fatores com autovalores maiores que 1 serão retidos

Os dois componentes extraídos (regra de autovalor > 1) explicam 73% da variância geral dos dados originais.

Figura 8 - PCA para 6 Componentes

No entanto, como decisão ainda não é final, é importante que se faça a análise do Scree Plot com o intuito obter melhor compreensão sobre quantas componentes deverão ser retidas. O Scree Plot é uma ferramenta para verificar se a PCA é adequada ao conjunto de dados dos Municípios do Brasil ou não. As Componentes Principais são criadas na ordem da quantidade de variação que cobrem; PC1 captura a maior variação, já a PC2, a segunda maior e assim por diante, de forma que cada um delas contribui com algumas informações dos dados. Por isso, há um *trade-off* na escolha de quantas Componentes Principais serão geradas, pois embora quanto menos componentes, melhor, ao deixar algumas componentes de fora perde-se informações.

Em relação às variáveis ativas do conjunto de dados dos Municípios do Brasil, o Scree Plot, assim como o critério de Kaiser, também sugere que duas Componentes Principais devem ser geradas (autovalores > 1), conforme gráfico abaixo:

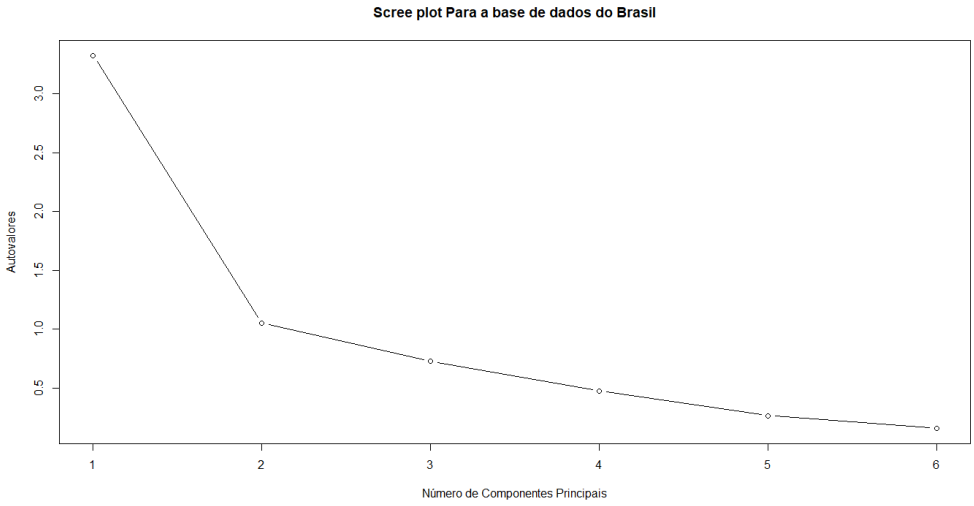

Figura 9 - Scree Plot

À luz das análises realizadas, cria-se uma solução para extrair duas Componentes Principais, as quais explicam 73% da variância geral.

```
> pc2 <- principal(SCALED_DATA, nfactors=2, rotate="none")
> pc2
Principal Components Analysis
Call: principal(r = SCALED_DATA, nfactors = 2, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
                 PC1   PC2   h2   u2 com
IDHM            0.74  0.56 0.86 0.14 1.9
GVA_CAPITA      0.80 -0.30 0.73 0.27 1.3
TAXES_CAPITA    0.70 -0.37 0.63 0.37 1.5
GDP_CAPITA      0.90 -0.29 0.88 0.12 1.2
GAS_MUN_CAPITA 0.62 -0.16 0.41 0.59 1.1
COMP_CAPITA     0.70  0.63 0.88 0.12 2.0

                       PC1  PC2
SS loadings           3.34 1.05
Proportion Var        0.56 0.18
Cumulative Var        0.56 0.73
Proportion Explained  0.76 0.24
Cumulative Proportion 0.76 1.00

Mean item complexity =  1.5
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.09
 with the empirical chi square  951.05  with prob <  1.4e-204

Fit based upon off diagonal values = 0.97
```

Figura 10 - Solução com Duas Componentes Principais

Para além disso, na figura a seguir é possível verificar as quantidades totais de variação que uma variável original partilha com todas as outras variáveis da análise em questão. Isto é, quanto mais próximo de 1 for o valor, maior o nível de explicação. Por isso, exclui-se a variável *GAS_MUN_CAPITA* da análise, uma vez que possui um valor baixo, conforme é possível observar abaixo marcado de vermelho:

```
> round(pc2$communality,2)
          IDHM     GVA_CAPITA    TAXES_CAPITA     GDP_CAPITA GAS_MUN_CAPITA    COMP_CAPITA
          0.86           0.73           0.63           0.88           0.41           0.88
>
```

Figura 11 - Communalities da Solução com Duas Componentes Principais

Após a remoção da variável GAS_MUN_CAPITA, tendo em vista o seu Communality de 0,41, cria-se uma solução com duas componentes principais, a qual explica 82% da variância, conforme é possível ver abaixo, na Figura 12:

```
> pc2 <- principal(AJUSTED_SCALED_DATA, nfactors=2, rotate="none")
> pc2
Principal Components Analysis
Call: principal(r = AJUSTED_SCALED_DATA, nfactors = 2, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
              PC1   PC2   h2   u2 com
IDHM         0.76  0.54 0.87 0.13 1.8
GVA_CAPITA   0.80 -0.32 0.74 0.26 1.3
TAXES_CAPITA 0.72 -0.43 0.70 0.30 1.6
GDP_CAPITA   0.89 -0.31 0.89 0.11 1.2
COMP_CAPITA  0.72  0.61 0.88 0.12 1.9

                       PC1  PC2
SS loadings           3.04 1.04
Proportion Var        0.61 0.21
Cumulative Var        0.61 0.82
Proportion Explained  0.74 0.26
Cumulative Proportion 0.74 1.00

Mean item complexity =  1.6
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.09
 with the empirical chi square  597.91  with prob <  4.8e-132

Fit based upon off diagonal values = 0.97
> |
```

Figura 12 - Solução com Duas Componentes Principais sem a variável GAS_MUN_CAPITA

Neste sentido, as communalities são verificadas e conforme na Figura 13, todos os valores das novas communalities estão mais próximos de 1, o que significa que podemos dar sequência com os atributos selecionados até aqui.

```
> round(pc2$communality,2)
      IDHM   GVA_CAPITA TAXES_CAPITA   GDP_CAPITA  COMP_CAPITA
      0.87          0.74          0.70          0.89          0.88
> |
```

Figura 13 – Novas communalities da Solução com Duas Componentes Principais

ANEXO IV – ROTAÇÃO E INTERPRETAÇÃO DAS COMPONENTES PRINCIPAIS E TOMADA DE DECISÃO FINAL SOBRE O NÚMERO DE COMPONENTES PRINCIPAIS

Conforme é possível observar a seguir, na Figura 14, ao realizar a rotação, todas as variáveis são diferentes de zero e as cargas significativas de cada uma das variáveis acabaram por se concentrar em apenas uma das Componentes Principais (assinalado de verde na figura abaixo), o que era justamente o que se esperava, já que esta rotação maximiza as variações de carregamento dentro de cada uma das variáveis.

```
> #Roda a solução para 2 componentes e faz interpretação dos dados
> pc2r <- principal(AJUSTED_SCALED_DATA, nfactors=2, rotate="varimax")
> pc2r$loadings

Loadings:
             RC1   RC2
IDHM         0.266 0.892
GVA_CAPITA   0.828 0.237
TAXES_CAPITA 0.831 0.105
GDP_CAPITA   0.891 0.305
COMP_CAPITA  0.191 0.919

                 RC1   RC2
SS loadings    2.278 1.799
Proportion Var 0.456 0.360
Cumulative Var 0.456 0.815
```

Figura 14 - Solução com Duas Componentes com Rotação Varimax

Utiliza-se o método *Varimax* para fazer esta rotação; os *loadings* maiores ficam ainda maiores e os menores, ficam ainda menores, e ainda, as componentes continuam não relacionadas. Além disso, através desta rotação e por meio do *loadings* de cada variável em cada uma das Componentes Principais, é possível identificar o cada uma das cada PCs, de modo geral, acaba por representar.

ANEXO V – AGRUPAMENTO DOS DADOS

No intuito de definir a quantidade de grupos em que os municípios brasileiros serão agrupados, realizou-se a análise hierárquica dos dados e posteriormente a análise de Particionamento. A Análise hierárquica produz um conjunto de clusters aninhados e organizados como uma árvore hierárquica (dendrograma), conforme é possível observar na imagem a seguir:

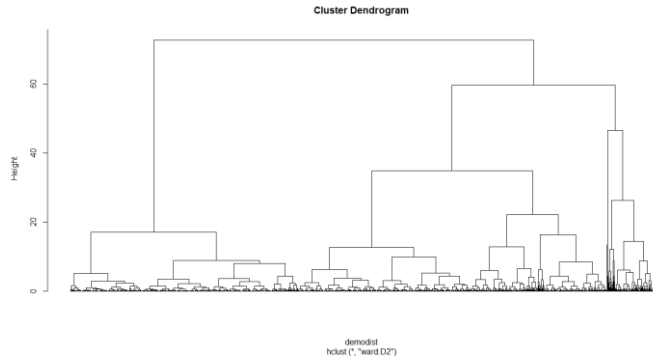

Figura 15 - Dendrograma - Árvore Hierárquica Euclidiana formada por meio do Método de variação mínima de Ward

No entanto, o agrupamento hierárquico não dita quais são os grupos existentes no conjunto de dados em questão. Deste modo, após a elaboração do dendrograma foi necessário cortá-lo; o método utilizado para definir onde realizar o referido corte no dendrograma foi a perceção simples do local onde os galhos da árvore fossem mais longos, o que originou três agrupamentos, conforme pode-se observar a seguir:

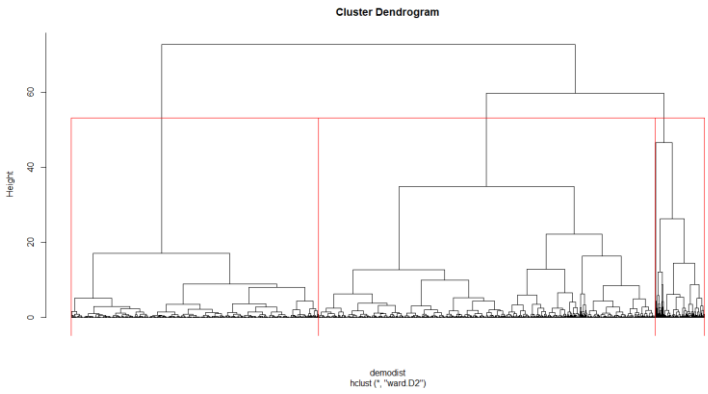

Figura 16 - Corte no Dendrograma K = 3 (Tamanho dos grupos: 2000, 2027 e 48)

Após a realização da análise hierárquica dos dados é preciso realizar o particionamento dos dados. Isto é, importa-se o número de K originado no método anterior e faz-se a partição dos dados em um número (K) pré-especificado de grupos. Neste caso em questão, o valor de K é 3,

o que corresponde ao número de grupos originados após o corte da árvore. Para tal particionamento, neste trabalho, utiliza-se o algoritmo k-means de MacQueen.

É importante notar que o valor de K originado por meio do método de variação mínima de Ward é um valor indicativo, pelo que testes com outros valores de K foram realizados (Figura 17 e Figura 18) e posteriormente, realizou-se uma análise das médias das silhuetas para valores de K de 1 a 10, tendo em vista que por meio desta abordagem, o valor ótimo de K é aquele em que a silhueta apresentar a média mais elevada, conforme é possível observar na Figura 19.

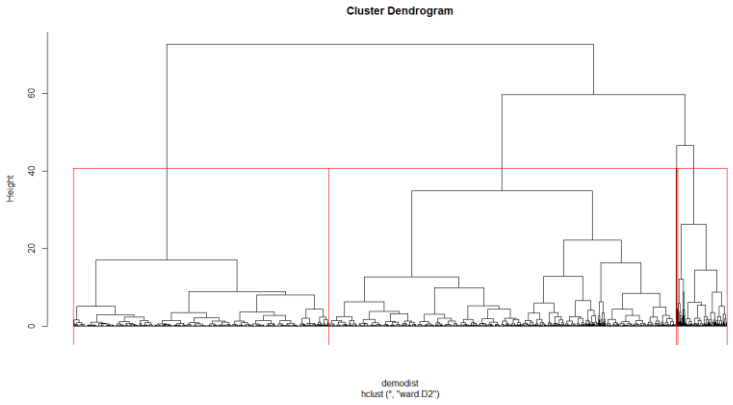

Figura 17 - Teste K = 4 (tamanho dos grupos: 1836, 287, 15, 1937)

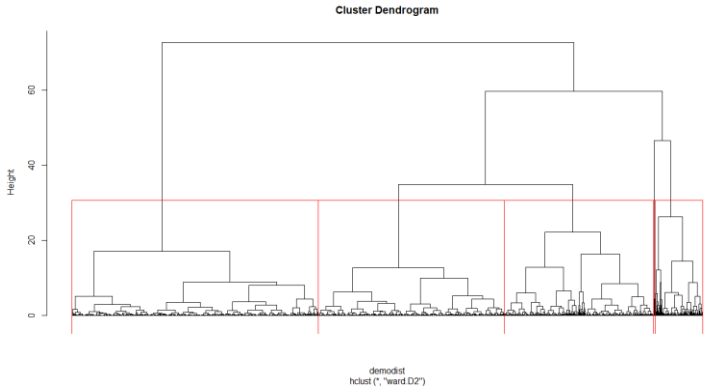

Figura 18 - Teste K =5 (tamanho dos grupos: 1485, 15, 267, 1539 e 769)

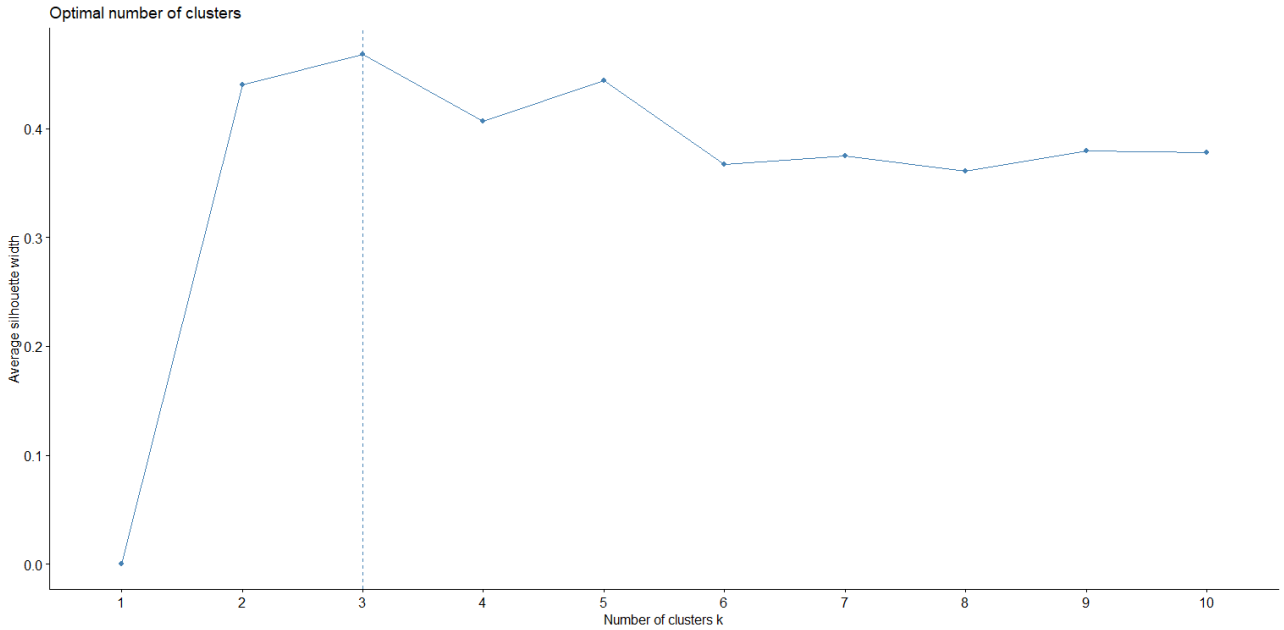

Figura 19 - Método Média das silhuetas para definir o valor de K

Neste sentido, definiu-se que o valor de K a ser utilizado neste trabalho é 3. Ou seja, os municípios brasileiros serão divididos em três grupos.

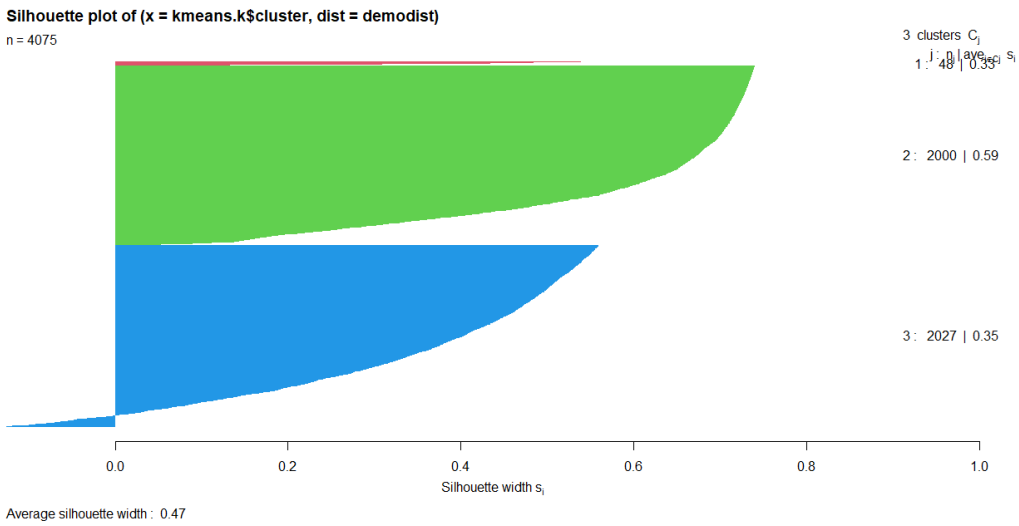

Figura 20 - Silhueta para k = 3

ANEXO VI – ATIVIDADE COM MAIOR CONTRIBUIÇÃO DE VALOR ACRESCENTADO BRUTO POR MUNICÍPIO E POR AGRUPAMENTO

Atividade com maior contribuição de Valor Acrescentado Bruto (VAB)	*CONTAGEM DE MUNICÍPIOS POR AGRUPAMENTOS E VAB*		
	1	2	3
Administração, defesa, educação e saúde públicas e seguridade social	286	1515	1
Agricultura, inclusive apoio à agricultura e a pós colheita	386	193	8
Comércio e reparação de veículos automotores e motocicletas	27	4	5
Construção	5	2	0
Demais Serviços	1028	161	8
Eletricidade e gás, água, esgoto, atividades de gestão de resíduos e descontaminação	38	20	6
Indústrias de transformação	162	36	18
Indústrias extrativas	14	9	2
Pecuária, inclusive apoio à pecuária	75	52	0
Produção florestal, pesca e aquicultura	6	8	0

Tabela 3- ATIVIDADE COM MAIOR CONTRIBUIÇÃO DE VALOR ACRESCENTADO BRUTO POR MUNICÍPIO E POR AGRUPAMENTO