Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

ESTUDO SOBRE IMIGRANTES NA ÁREA METROPOLITANA DE BRASÍLIA UTILIZANDO A TÉCNICA DE ANÁLISE DE AGRUPAMENTO

Juliana Pacheco de Almeida

Brasília
2013

Juliana Pacheco de Almeida

ESTUDO SOBRE IMIGRANTES NA ÁREA METROPOLITANA DE BRASÍLIA UTILIZANDO A TÉCNICA DE ANÁLISE DE AGRUPAMENTO

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientadora: Ana Maria Nogales Vasconcelos

Brasília
2013

Resumo

O trabalho tem como objetivo analisar o perfil dos imigrantes na Área Metropolitana de Brasília (AMB), relacionando com seu perfil socioeconômico. Os dados utilizados foram a amostra de pessoas do Censo Demográfico de 2010 e o método estatístico adotado foi a Análise de Agrupamento (ou Análise de Cluster).

Os imigrantes considerados para a pesquisa foram os de data fixa, ou seja, aqueles que não residiam na AMB no dia 31/07/2005.

Após a seleção dos imigrantes de data fixa, foi realizado um corte, onde foram selecionados apenas os imigrantes que tinham idade maior ou igual a 20 anos em 2010, pois entende-se, como migrante de data fixa, que os que tinham idade menor ou igual a 15 anos no momento da migração (há 5 anos antes) não tiveram escolhas sobre a mudança, pois vieram para acompanhar os pais, por isso não é interessante para nossa pesquisa traçar o perfil dessas pessoas.

Os imigrantes foram separados em 4 regiões de acordo com as características econômicas, sendo a região 1, a mais rica, a 2, a intermediária, e a 3, a mais pobre dentro do DF, a região 4 é composta com os municípios goianos que fazem parte da AMB. Os perfis dos imigrantes foram traçados a partir dessa classificação e essas regiões não foram utilizadas na análise de agrupamento, pois o objetivo era classificar os indivíduos sem a influência do local onde eles residiam.

A análise foi feita utilizando as variáveis Idade, Estado Civil, Grau de Instrução, UF de Residência Anterior (em 31/07/2005) e Sexo e o Método de Agrupamento de Duas Etapas, que é feito em duas partes, na primeira foi utilizado o Método de Agrupamento Não-Hierárquico e na segunda, o Método de Agrupamento Hierárquico, com tratamento de ruído no nível de 20%.

A análise feita resultou em 8 grupos, onde todas as variáveis são de extrema importância para a montagem dos grupos. No grupo 1, predominam os indivíduos das Regiões 1 e 2 e nos grupos 3, 5 e 6, os indivíduos das Regiões 2 e 4, os demais grupos foram separados homogeneamente entre as regiões.

Índice

1. Introdução

> *A busca por melhor qualificação, por emprego, renda e melhores condições de vida, por residir junto de familiares e a busca de suporte são alguns exemplos de estímulo, estreitamente relacionados às etapas do ciclo de vida, que muitas vezes culminam em migração.* (De Jong e Gardner, 1981).

Levando em consideração a citação acima que explica o que leva a pessoa a migrar de um local para outro e o porquê das pessoas estarem sempre em movimento, sendo assim a migração faz parte das populações de sua formação.

> *A zona de expansão teve um crescimento populacional de 1970 a 1991, em consequência da rápida ampliação da agricultura e do grande incremento das cidades-satélite de Brasília. A oferta de empregos do setor de serviços dessas cidades é a principal explicação do crescimento dessa área no período analisado.* (Amaral; Rodrigues; Fígoli, 2002)

Segundo a citação acima, o fluxo de migração não se dá apenas para o Distrito Federal em si, mas para toda a Área Metropolitana de Brasília (AMB), que é constituída de Brasília e suas cidades-satelites, localizadas no Distrito Federal e no estado de Goiás.

Segundo Amaral, Rodrigues e Fígoli, o local que o indivíduo migra na AMB, corresponde diretamente a sua renda e ocupação, porém queremos saber se há uma tendência nesse fluxo migratório.

Nessa pesquisa, tem-se como objetivo saber se podemos estudar os migrantes que vêm para a Área Metropolitana de Brasília como grupos em vez de estudar cada pessoa individualmente.

O objetivo geral dessa pesquisa é separar os imigrantes em grupos homogêneos utilizando a análise de agrupamento (ou cluster) de acordo com a técnica escolhida e os objetivos específicos são conhecer os aspectos, poder classificar e agrupar as pessoas que migram para a AMB, assim como conhecer um pouco mais sobre os métodos e técnicas de agrupamento existentes.

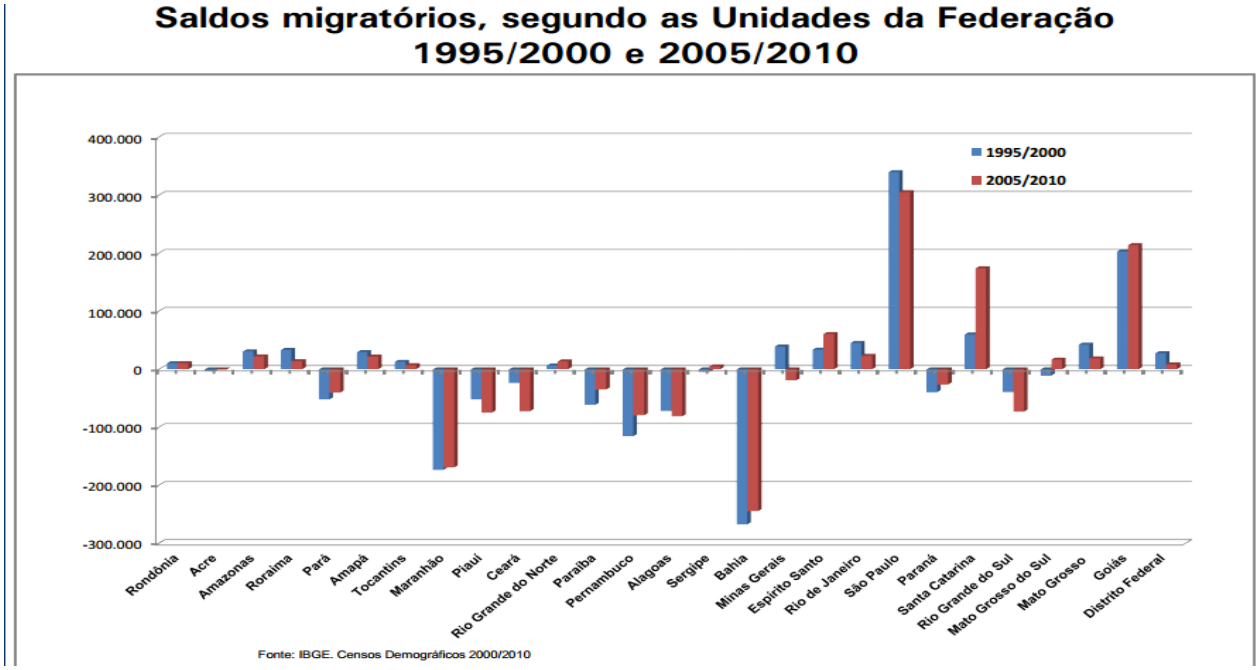
Uma das importâncias do trabalho é identificar os fluxos de migração, ou seja, se há uma tendência de pessoas vindas de certo lugar para outro dentro da AMB, que se encaixam no mesmo perfil socioeconômico.

Para que esses objetivos sejam alcançados devemos especificar quem são essas pessoas, sabendo sua idade, sexo, nível de escolaridade, condição de emprego, quanto ganha e qual sua ocupação, assim como saber de onde elas vieram e onde se instalaram na AMB.

> *O caso do Distrito Federal é peculiar em dois sentidos: em primeiro lugar, é uma cidade nova, criada há cerca de 50 anos para ser sede do governo federal. Por esse motivo, houve forte incentivo migratório.* (IPEA, 2011)

2. Descrição do problema

Para que se possa perceber o problema que vamos analisar, segue um estudo detalhado sobre o mesmo. Segundo o gráfico abaixo (Fonte: IBGE, Censos Demográficos, 2000/2010), há um saldo migratório positivo (ou seja, maior número de imigrantes[1] do que emigrantes[2]) em 16 estados brasileiros dos anos 2005 a 2010 e, entre eles, está o estado de Goiás (como segundo maior saldo migratório) e o Distrito Federal.



[1] Imigrante é o indivíduo que muda do seu local de origem pela perspectiva do local que o acolhe.
[2] Emigrante é o indivíduo que muda do seu local de origem pela perspectiva do seu local de origem.

> *Todavia, se a migração na década de 1990 não se destaca como a principal componente na evolução da população do Distrito Federal, ela assume essa importância na evolução da população do chamado Entorno Imediato, onde a proporção de migrantes recentes aumentou de 22% em 1970 para 55,3% em 2000.* (De Vasconcelos et al., 2006).

Ao observar o gráfico e citação acima, vemos que a migração para a Área Metropolitana de Brasília é uma prática em crescimento, haja vista que o entorno de Brasília é um local em desenvolvimento populacional.

3. Metodologia

Nessa seção será descrita a forma que foram feitas as análises, assim como a seleção dos dados e das variáveis utilizadas. A mesma foi separada em duas partes, sendo elas: 3.1. Fonte de dados: Censo 2010: onde será explicitado de onde foram retirados os dados utilizados no trabalho e como é feito o recolhimento deles; 3.2. Análise de agrupamento: como foi feito a análise selecionada a partir dos dados escolhidos na seção 3.1.

3.1. Fonte de dados: CENSO 2010

A base de dados do Censo 2010 será utilizada como nosso banco de dados, pois é um retrato de corpo inteiro do país com o perfil da população e as características de seus domicílios, ou seja, ele nos diz como somos, onde estamos e como vivemos, foi realizada uma análise descritiva acerca dos imigrantes da AMB para que possamos retratar quem eles são, onde e como vivem, assim como uma análise multivariada, com o auxilio da técnica de análise de agrupamento (ou análise de conglomerados, classificação ou cluster).

O Censo utiliza duas formas de coleta de dados: a primeira é utilizada um questionário grande, para ser aplicado em uma amostra de domicílios (e cada um de seus moradores), chamado de questionário da amostra; e a segunda, é utilizado um questionário pequeno, para ser aplicado aos domicílios (e cada um de seus moradores) não selecionados para a amostra, chamado de questionário básico. As variáveis do questionário básico estão contidas no questionário da amostra e todos os domicílios (e seus indivíduos) em território nacional participam do censo.

O plano amostral utilizado no Censo de 2010 equivale a um plano de amostragem complexa. A amostra foi selecionada em três estágios (setores, domicílios e adulto). No primeiro estágio foram selecionados 250 setores censitários com probabilidade proporcional do seu número de domicílios particulares permanentes (DPP), sendo excluídos da população amostrada os setores censitários especiais (quartéis, bases militares, alojamentos, asilos, orfanatos, etc.), no segundo estágio, foram selecionados 20 DPPs com equiprobabilidade em cada setor censitário escolhido. E, no último estágio, foi selecionado um adulto em cada domicílio para responder a entrevista individual.

As unidades primárias de seleção (setores censitários) foram estratificadas segundo o tamanho da população do município e sua situação (rural ou urbana). O tamanho da amostra de setores foi alocada entre os estratos de forma proporcional à população de cada estrato.

Para cada indivíduo selecionado na amostra, foi gerado um peso amostral (ou fator de expansão) que são utilizados para obtenção de estimativas dos parâmetros relacionados com as características investigadas para a população, a partir da amostra.

A base de dados foi selecionada com base nos indivíduos que moram na Área Metropolitana de Brasília (AMB), formada pelo Distrito Federal mais 10 municípios do Goiás, sendo eles Águas Lindas de Goiás, Cristalina, Cidade Ocidental, Formosa, Luziânia, Novo Gama, Padre Bernardo, Planaltina, Santo Antônio do Descoberto e Valparaíso de Goiás.

Essa base de dados foi utilizada para classificar os imigrantes, empregando o conceito de migrante de data fixa, ou seja, são aqueles que residem na AMB, com cinco anos completos ou mais de idade, e que nesta data não residia nessa área há cinco anos (ou seja, em 31/07/2005) onde foi realizada a entrevista.

Para fazer a análise utilizamos as seguintes variáveis do Censo 2010:

- V0001 – Unidade de Federação que reside (Distrito Federal ou Goiás);
- V1006 – Situação do domicílio (Urbano ou Rural);
- V0601 – Sexo (Homem ou Mulher);
- V6036 – Idade (calculada em anos);
- V0606 – Cor ou Raça (Branca, Parta, Preta ou Outros);

- V6222 – Unidade de Federação de nascimento (26 estados brasileiros mais o DF ou Estrangeiro);
- V6252 – Unidade de Federação que residia anteriormente (26 estados brasileiros mais o DF ou Estrangeiro);
- V6400 – Grau de Instrução (Sem Instrução/Fundamental Incompleto, Fundamental Completo/Médio Incompleto, Médio Completo/Superior Incompleto ou Superior Completo);
- V0640 – Estado Civil (Solteiro, Divorciado, Casado ou Viúvo);
- V6643 – Total de filhos vivos.

Os imigrantes foram separados, segundo Vasconcelos, César e Costa, em quatro regiões, sendo a região 1 a mais rica (Plano Piloto, Sudoeste/Octogonal, Lago Norte, Lago Sul) a 2, intermediária (Cruzeiro, Candangolândia, Núcleo Bandeirante, Guará, Gama, Taguatinga, Águas Claras, Vicente Pires, Riacho Fundo I e São Sebastião) e a 3, a mais pobre (Brazlândia, Ceilândia, Itapoã, Planaltina, Santa Maria, Recanto das Emas, Riacho Fundo II, Samambaia e áreas rurais). A quarta região agrupa os municípios goianos que fazem parte da AMB que, em geral, tem renda ainda menor do que a 3ª região do DF.

3.2. Análise de Agrupamento

Na análise de agrupamentos a população estudada é dividida em grupos homogêneos, ou seja, similares entre si com respeito às características que neles foram medidas, e onde elementos com características distintas fiquem em grupos diferentes.

Primeiramente, o banco de dados é selecionado utilizando o conceito de imigrante de data fixa, após é feito um corte no nosso banco para indivíduos com idade maior ou igual há 20 anos, pois entende-se que os que tinham idade menor ou igual a 15 anos no momento da migração (há 5 anos antes) não tiveram escolhas sobre a mudança, pois vieram para acompanhar os pais, por isso não é interessante para nossa pesquisa traçar o perfil dessas pessoas.

Como as variáveis estudadas são categóricas com um número grande de categorias, foi feita uma transformação nas variáveis utilizadas em variáveis binárias para que seja mais fácil trabalhar com as mesmas, conforme descrito abaixo:

- Sexo:
  - 0 - Homem
  - 1 - Mulher;
- Raça/Cor:
  - 0 - Branca (Branca ou Amarela)
  - 1 - Outros (Parda, Preta, Indígena, outros);
- Unidade de Federação que residia em 31/07/2005:
  - 0 - Regiões Sul e Sudeste
  - 1 - Regiões Norte, Nordeste e Centro-Oeste;
- Grau de Instrução:
  - 0 - Sem Instrução a Ensino Fundamental Completo
  - 1 - Ensino Médio Completo e Ensino Superior Completo;
- Estado Civil:
  - 0 - Casado;
  - 1 - Outros (Viúvo, Solteiro e Separado).

A variável idade será a única utilizada na sua forma inicial, ou seja, ela será utilizada como uma variável contínua para que seja feita a análise.

Após fazer essa nova representação, é necessário decidir a medida de similaridade ou dissimilaridade que será empregada, cada tipo de medida produz um tipo de agrupamento. Depois de calcular essa medida, é necessário decidir, utilizando métodos e análises estatísticas, o número de grupos em que os indivíduos farão parte e quais as variáveis de maior importância para esse agrupamento.

As análises são feitas utilizando o software estatístico SPSS e não será feito uso do peso amostral na hora da análise, já que o objetivo é avaliar a amostra, e também não será utilizada a variável região na análise de agrupamento, pois o objetivo proposto é classificar o imigrante independente do local onde ele reside, essa variável somente será utilizada para traçar o perfil do indivíduo.

4. Análise de Agrupamento

Nessa seção será explicada a técnica utilizada (Análise de agrupamento), assim como seus diferentes métodos que podem ser utilizados.

4.1. Conceitos Básicos:

Vetor aleatório: Seja (Ω, Ғ,Ƥ) um espaço de probabilidade. Definimos como vetor aleatório a função X de Ω, tal que $X' = [X_1\ X_2\ ...\ X_p]$, para todo p = 1, 2, 3, ...

Vetor de médias: Seja X um vetor aleatório, o vetor μ é igual a esperança do vetor X, assim:

$$\mu' = [E(x_1), ... , E(X_p)] = [\mu_1, ... , \mu_p]$$

Combinações Lineares: Seja X um vetor aleatório, composto por p variáveis, com vetor de médias μ e matriz de covariâncias ∑ (pxp). Seja a (px1) um vetor de constantes conhecidas, isto é, $a = (a_1\ a_2\ ...\ a_p)$. Seja Z a variável definida por:

$$Z_i = a_1X_1 + ... + a_pX_p$$

A esperança de Z é igual a $\mu_z = a_1\ \mu_1 + ... + a_p\ \mu_p$ e a variança de Z é:

$$var\ (Z) = a'\sum a$$

Autovalor e autovetor: Seja $T : V \rightarrow V$ um operador linear. Um vetor $v \in V$, $v \neq 0$, é dito autovetor do operador T, se existir $\lambda \in R$ tal que $T(v) = v\lambda$.

O escalar λ é denominado autovalor do operador linear T associado ao autovetor v. Seja v é um autovetor do operador linear T associado ao autovalor λ então $kv \in V$ também é um autovetor de T associado ao autovalor λ , para todo $k \in R$, $k \neq 0$.

Teorema da decomposição espectral: Existe uma matriz ortogonal O (pxp), isto é, O'O = OO' = I, tal que O'∑O será uma matriz com a diagonal igual aos autovalores e os demais números iguais a 0. Sendo ∑ a matriz de covariâncias e $\lambda_1 \geq ... \geq \lambda_p$:

Det (∑) = |∑| = multiplicação dos λi;

Traço (∑) = $\lambda_1 + ... + \lambda_p$

A matriz O é dada por [e_1 ... e_p] e, pelo teorema da decomposição espectral, temos:

$$\sum = \Sigma (\lambda_i e_i e'_i)$$

4.2. Análise de Agrupamento:

A análise de agrupamento é uma técnica em que não são feitas suposições sobre o número de grupos ou a estrutura do grupo.

O agrupamento é feito com base nas similaridades ou nas distâncias (dissimilaridades).

O objetivo básico da análise de agrupamento é descobrir grupos naturais de variáveis. Por outro lado, precisamos criar uma escala quantitativa para medir a similaridade entre os objetos, no nosso caso, os indivíduos.

A análise de agrupamento compreende cinco etapas:

1. A seleção de indivíduos ou de uma amostra para o estudo;
2. A escolha de um conjunto de variáveis de onde serão retiradas as informações sobre os indivíduos ou a amostra;
3. A definição de uma medida de semelhança ou distância entre os indivíduos;
4. A escolha de um algoritmo de classificação;
5. Validação dos resultados encontrados.

a) Medidas de Similaridade:

Quanto maior a medida de similaridade, maior a semelhança entre os indivíduos. Algumas das medidas de similaridade que são normalmente usadas são:

1. Distância Euclidiana: A distância euclidiana dada entre dois indivíduos a e b é dada por:

$$d_{ab} = \left[\sum_{j=1}^{p} \left(X_{aj} - X_{bj}\right)^2\right]^{1/2}$$

Onde p=1,2,...,j;

X_{aj} = valor da variável j para o indivíduo a;

X_{bj} = valor da variável j para o indivíduo b;

É recomendável a padronização das variáveis antes de se obter o valor da distância euclidiana, devido que normalmente todos os dados não estão no mesmo padrão de medidas.

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j} \quad , \quad Z_{ij} \sim (0\,,1_j)$$

2. Distância Euclidiana Média: A distância euclidiana cresce à medida que cresce o número de variáveis. Uma maneira de eliminar o efeito do número de variáveis é dividir o valor da distância euclidiana pela raiz quadrada do número de variáveis.

$$\bar{d}_{ab} = \frac{1}{\sqrt{p}} \cdot d_{ab}$$

Onde p = número de variáveis;

d_{ab} = distância euclidiana entre a e b.

3. Distância de Mahalanobis: A distância de Mahalanobis entre os indivíduos a e b é dada por:

$$D^2_{ab} = [X_a - X_b]' \cdot S^{-1} \cdot [X_a - X_b]$$

Onde D^2_{ab} = distância de Mahalanobis entre os indivíduos a e b;

X_a = vetor de características do indivíduo a;

X_b = vetor de características do indivíduo b;

S = matriz de variância amostral da população.

4. Distância de log-verossimilhança: A distância entre 2 grupos i e s é definida por:

$$d(i,s) = \xi i + \xi s + \xi(i,s) \quad (1)$$

Onde:

$$\xi_i = -n_i \left(\sum_{j=1}^{p} \frac{1}{2} \log(\hat{\sigma}_{ij}^2 + \hat{\sigma}_j^2) - \sum_{j=1}^{q} \sum_{l=1}^{m_j} \hat{\pi}_{ijl} \log(\hat{\pi}_{ijl}) \right) \quad (2)$$

$$\xi_s = -n_s \left(\sum_{j=1}^{p} \frac{1}{2} \log(\hat{\sigma}_{sj}^2 + \hat{\sigma}_j^2) - \sum_{j=1}^{q} \sum_{l=1}^{m_j} \hat{\pi}_{sjl} \log(\hat{\pi}_{sjl}) \right) \quad (3)$$

$$\xi_{\langle i,s \rangle} = -n_{\langle i,s \rangle} \left(\sum_{j=1}^{p} \frac{1}{2} \log(\hat{\sigma}_{\langle i,s \rangle j}^2 + \hat{\sigma}_j^2) - \sum_{j=1}^{q} \sum_{l=1}^{m_j} \hat{\pi}_{\langle i,s \rangle jl} \log(\hat{\pi}_{\langle i,s \rangle jl}) \right) \quad (4)$$

ξ_v pode ser interpretado como a dispersão dentro do grupo v (v = i,s,<i,s>). ξ_v consiste em duas partes, sendo a primeira $-n_v\Sigma(1/2)\log(\sigma^2_{vj}+\sigma^2_j)$ mede a dispersão das variáveis contínuas x_j dentro do grupo v e a segunda, a entropia $-n_v\Sigma\Sigma\pi_{vjl}\log(\pi_{vjl})$ para medir a dispersão das variáveis categóricas.

b) Métodos Hierárquicos:

Técnicas de agrupamentos hierárquicos são feitas por séries de fusões sucessivas ou divisões sucessivas.

Nos Métodos Hierárquicos de Agrupamento começamos com os objetos individuais (no nosso caso, cada um dos indivíduos). Assim, os indivíduos mais semelhantes são agrupados inicialmente e assim por diante. Eventualmente, as similaridades vão diminuindo até todos os subgrupos serem fundidos em um único grupo.

Nos Métodos hierárquicos de Repartição começamos com um único grupo inicial de indivíduos que é dividido em dois subgrupos, de tal forma que os indivíduos em um subgrupo estão "longe“ (são mais “diferentes”) dos indivíduos no outro e assim sucessivamente, o processo continua até que cada objeto forma um grupo.

Temos 2 métodos mais utilizados:

- Vizinho mais próximo (ou single linkage method): calcula-se a matriz de distâncias entre os ‘n’ indivíduos da população, em seguida os indivíduos mais próximos são agrupados.

- Vizinho mais distante (ou complete linkage method): é o inverso do vizinho mais próximo. Calcula-se a matriz de distâncias entre os ‘n’ indivíduos da população, em seguida os indivíduos mais distantes são agrupados.

c) Métodos não-hierárquicos: Método das K Médias:

Técnicas de agrupamento não-hierárquica são projetados para os itens do grupo (ou indivíduos), ao invés de variáveis, em uma coleção de K grupos. O número de grupos, K, pode ser especificado antecipadamente ou como parte do processo de agrupamento. Como uma matriz de distâncias (semelhanças) não tem de ser determinada e os dados de base não precisam que ser armazenados durante o funcionamento do computador, esses métodos pode ser aplicado a conjuntos de dados muito maiores do que os métodos hierárquicos.

Esse método associa cada individuo para o grupo com o centróide (ou média) mais próximo. Na versão mais simples, o processo é composto por três passos:

Passo 1: Partição de k grupos iniciais;

Passo 2: Atribua um item ao grupo cujo centróide (média) é mais próximo (A distância é geralmente calculada usando distância euclidiana com observações padronizadas ou não padronizadas). Recalcular o centróide do grupo que recebeu o novo item e para o grupo que perdeu o item.

Passo 3: Repita o passo 2 até que não haja mais deslocamentos para acontecerem.

Número de grupos: Determinar o número de grupos da análise é uma das tarefas mais difíceis de realizar.

Para Barroso & Artes, o número de grupos pode ser definido a priori, através de algum conhecimento que se tenha sobre os dados, pela conveniência do pesquisador, por simplicidade, ou ainda pode ser definido a posteriori com base nos resultados da análise.

Qualquer que seja a abordagem empregada, geralmente é aconselhável observar o padrão total de agrupamentos. Isto pode proporcionar uma medida da qualidade do processo de agrupamento e do número de agrupamentos que nos emerge vários níveis do critério de agrupamento.

d) Dendograma

Dendograma é a representação matemática e gráfica da análise de agrupamento através de uma estrutura de árvore.

Os nós apresentados no dendograma representam os agrupamentos feitos entre um grupo já formado e outro ou um indivíduo e um grupo ou dois indivíduos, formando um novo grupo. Se cortarmos o dendograma num nível qualquer, teremos a classificação dos grupos nesse nível.

e) Método de agrupamento de duas etapas:

Quando se tem um banco de dados muito grande ou precisa-se de um procedimento de agrupamento que pode rapidamente formar grupos com bases em dados categóricos ou contínuos, nenhum dos outros dois procedimentos demonstrados acima convém.

O método de Agrupamento hierárquico requer uma matriz de distância entre todos os pares de casos e o método de K médias requer remanejar os casos para dentro e fora dos clusters e saber o número de clusters antecipadamente.

A análise de agrupamento de duas etapas foi designada para essas aplicações. Ela requer somente uma passagem pelos dados (o que é importante para grandes bancos de dados) e pode ter soluções baseados em variáveis contínuas e categóricas e com um número variante de grupos. O número de grupos formados pode ser especificado ou pode-se selecionar um algoritmo que seleciona um número “ótimo” (Critério de Informação de Akaike – AIC – ou Critério de Informação de Bayes - BIC).

O algoritmo do agrupamento é baseado em uma medida de distância que proporciona o melhor resultado se todas as variáveis são independentes, se as variáveis contínuas tem uma distribuição normal e as variáveis categóricas tem uma distribuição multinominal. Na prática, essas suposições serão raramente verdade, mas o algoritmo foi desenvolvido para se comportar razoavelmente bem quando as suposições não são cumpridas.

Como análise de agrupamento não envolve teste de hipótese e cálculo de níveis de significância é perfeitamente aceitável dados de agrupamento que não atendem os pressupostos ter um bom desempenho. Somente o pesquisador que

utiliza os métodos pode determinar se a solução é satisfatória para a sua pesquisa. O procedimento é separado em dois passos:

1. Passo 1: Pré-agrupamento: Fazendo pequenos grupos:

O primeiro passo do procedimento de duas etapas é a formação de pré-grupos. O objetivo do pré-agrupamento é de reduzir o tamanho da matriz que contém as distâncias entre todas as possibilidades de pares de indivíduos. Pré-grupos são somente aglomerados dos indivíduos originais que são usados no lugar dados brutos no agrupamento hierárquico. Quando um indivíduo é estudado, o algoritmo decide, baseado na medida de distância, se o indivíduo em questão deve ser mesclado com um pré-grupo já formado ou se deve iniciar um novo pré-grupo. Quando o pré-agrupamento está completo, todos os casos no mesmo pré-grupo são tratados como um único indivíduo. O tamanho da matriz de distâncias não depende mais do número de indivíduos, mas do número de pré-grupos.

2. Passo 2: Agrupamento hierárquico dos pré-grupos:

No segundo passo, é utilizado o algoritmo padrão da análise de agrupamento hierárquico nos pré-grupos. Similar ao modelo hierárquico, os grupos com as menores distancias d(i,s) são juntados a cada etapa até que formem um cluster só. A função de log-verossimilhança para o passo com k grupos é dado por:

$$l_k = \sum_{v=1}^{k} \xi_v.$$

A função l_k pode ser interpretada como a dispersão dentro do grupo. Se utilizarmos somente variáveis categóricas, l_k é a entropia dentro dos k grupos.

Número de grupos: O número de grupos pode ser previamente determinado ou calculado pelos estimadores de Critério de Informação de Akaike (AIC):

$$AIC_k = -2l_k + 2r_k$$

Onde r_k é o número de parâmetros independentes ou pelo Critério de Informação Bayesiano:

$$BIC_k = -2l_k + r_k \log n$$

O estimador escolhido é calculado no primeiro passo. BIC ou AIC resultam em uma boa estimação inicial do número máximo de clusters. O número máximo de

grupos é determinado ao número do grupo onde a proporção BIC_k/BIC_1 é menor do que c_1 (valor calculado em simulação pelo programa estatístico utilizado) pela primeira vez.

No segundo passo, usa-se a proporção R(k) na distância do k grupo, definido como:

$$R(k) = d_{k-1}/d_k,$$

Onde d_{k-1} é a distância se o grupo K é adicionado ao grupo k-1. O números de grupos é obtido quando ocorre um grande salto na razão das mudanças, definido por:

$$R(k_1)/R(k_2)$$

5. Resultados

Os resultados obtidos a partir dos dados e variáveis selecionados seguem abaixo. A seção foi dividida em duas partes: 5.1. Análise exploratória dos dados, onde é apresentada as características gerais dos imigrantes; 5.2.Análise de agrupamento, onde será apresentado os resultados referente ao método escolhido.

5.1. Análise Exploratória dos Dados

Tabela 1 – Imigrante, segundo algumas características, da Área Metropolitana de Brasília - 2010.

	Sem Ponderação		**Ponderado**	
Características	n	%	N	%
TOTAL	**13.875**		**240.012**	
UF de Residência				
Goiás	6.056	43,6%	63.615	26,5%
Distrito Federal	7.819	56,4%	176.397	73,5%
Situação do domicílio				
Rural	854	6,2%	12.968	5,4%
Urbano	13.021	93,8%	227.044	94,6%
Sexo				
Homem	6.846	49,3%	116.898	48,7%
Mulher	7.029	50,7%	123.114	51,3%

Continuação Tabela 1 – Imigrante, segundo algumas características, da Área Metropolitana de Brasília - 2010.

Cor/Raça				
Branca	5.380	38,8%	93.565	39,0%
Preta	1.476	10,6%	22.602	9,4%
Parda	7.079	51,0%	118.511	49,4%
Outros	40	0,3%	945	0,4%
Nível de Instrução				
Sem instrução/Fund. Incompleto	6.135	44,2%	96.632	40,3%
Fund. Completo/Médio Incompleto	2.339	16,9%	39.816	16,6%
Médio Completo/Sup. Incompleto	3.546	25,6%	64.213	26,8%
Superior Completo	1.803	13,0%	38.457	16,0%
Não determinado	52	0,4%	894	0,4%
Estado Civil				
Casado	3.450	24,9%	61.430	25,6%
Divorciado	493	3,6%	9.081	3,8%
Viúvo	282	2,0%	4.607	1,9%
Solteiro	8.557	61,7%	147.286	61,4%
Não Informado	1.093	7,9%	17.608	7,3%
Ocupação				
Empregado com carteira de trabalho assinada	4.051	29,2%	72.245	30,1%
Militar/Funcionário Público	794	5,7%	17.001	7,1%
Empregado sem carteira de trabalho assinada	1.836	13,2%	30.858	12,9%
Conta Própria	1.040	7,5%	17.427	7,3%
Empregador	59	0,4%	995	0,4%
Não Remunerado	66	0,5%	889	0,4%
Trabalhador para consumo próprio	54	0,4%	665	0,3%
Não Informado	5.975	43,1%	99.931	41,6%
Região				
Região 1	1.519	10,9%	35.261	14,7%
Região 2	3.825	27,6%	85.574	35,7%
Região 3	2.475	17,8%	55.562	23,1%
Região 4	6.056	43,6%	63.615	26,5%

É possível observar que as características mais frequentes dos imigrantes da AMB são: vive na zona urbana, é mulher (apesar de que a diferença entre o número de homens e mulheres é tênue), cor da pele parda, sem nível de instrução ou nível fundamental incompleto, solteiro e habita a região 4, que é constituída pelo entorno, onde o nível econômico é o menor das 4 regiões apresentadas.

Tabela 2 – Descrição da Idade dos imigrantes da Área Metropolitana de Brasília - 2010.

Características	N					Quartil		
	Válido	Ausente	Média	Mínimo	Máximo	25%	50%	75%
Idade								
Sem Ponderação	13.875	0	28,10	5	103	19	25	35
Ponderado	240.012	0	28	5	103	19	26	35

Pela tabela acima, percebe-se que os imigrantes são jovens, onde a média de idade é de 28 anos, tendo seu 3º quartil de 35 anos, isto é, onde 75% da população está compreendida abaixo dessa idade.

Tabela 3 – Imigrante acima de 20 anos, segundo algumas características, da Área Metropolitana de Brasília - 2010.

Características	Sem Ponderação		Ponderado	
	N	%	N	%
TOTAL	**10.113**		**177.313**	
UF de Residência				
Goiás	4.206	41,60%	44.423	25,05%
Distrito Federal	5.907	58,40%	132.891	74,95%
Situação do domicílio				
Rural	599	94,08%	168.215	94,87%
Urbano	9514	5,92%	9.098	5,13%
Sexo				
Homem	5.049	49,93%	87.241	49,20%
Mulher	5.064	50,07%	90.072	50,80%
Cor/Raça				
Branca	4.086	40,4%	72.186	40,71%
Preta	1.042	10,30%	17.065	9,62%
Parda	4.952	48,97%	84.129	47,45%
Outros	33	0,33%	774	0,44%

Continuação Tabela 3 – Imigrante acima de 20 anos, segundo algumas características, da Área Metropolitana de Brasília - 2010.

Nível de Instrução				
Sem instrução/Fund. Incompleto	3.465	34,26%	53.475	30,16%
Fund. Completo/Médio Incompleto	1.654	16,36%	27.741	15,65%
Médio Completo/Sup. Incompleto	3.176	31,41%	57.360	32,35%
Superior Completo	1.795	17,75%	38.297	21,60%
Não determinado	23	0,23%	440	0,25%
Estado Civil				
Casado	3.376	33,38%	60.224	33,96%
Divorciado	488	4,83%	8.982	5,07%
Viúvo	282	2,79%	4.607	2,60%
Solteiro	5.967	59,00%	103.500	58,37%
Ocupação				
Empregado com carteira de trabalho assinada	3.794	37,52%	67.859	38,27%
Militar/Funcionário Público	770	7,61%	16.526	9,32%
Empregado sem carteira de trabalho assinada	1.534	15,17%	25.683	14,48%
Conta Própria	937	9,27%	15.758	8,89%
Empregador	59	0,58%	995	0,56%
Não Remunerado	34	0,34%	496	0,28%
Trabalhador para consumo próprio	45	0,44%	573	0,32%
Não Informado	2.940	29,07%	49.424	27,87%
Região				
Região 1	1.225	12,11%	28.162	15,88%
Região 2	2.945	29,12%	65.714	37,06%
Região 3	1.737	17,18%	39.015	22,00%
Região 4	4.206	41,59%	44.423	25,05%

Quando é feito o corte para imigrantes maiores de 20 anos, há algumas mudanças nas características, a variável Grau de Instrução diminui a quantidade de imigrantes sem instrução/Ensino Fundamental Incompleto e há um aumento nos imigrantes com Ensino Médio ou/e Superior Completos, o número de casados também aumenta, assim como o número de trabalhadores com carteira assinada. Há um aumento na quantidade de migrantes na Região 1 e 2.

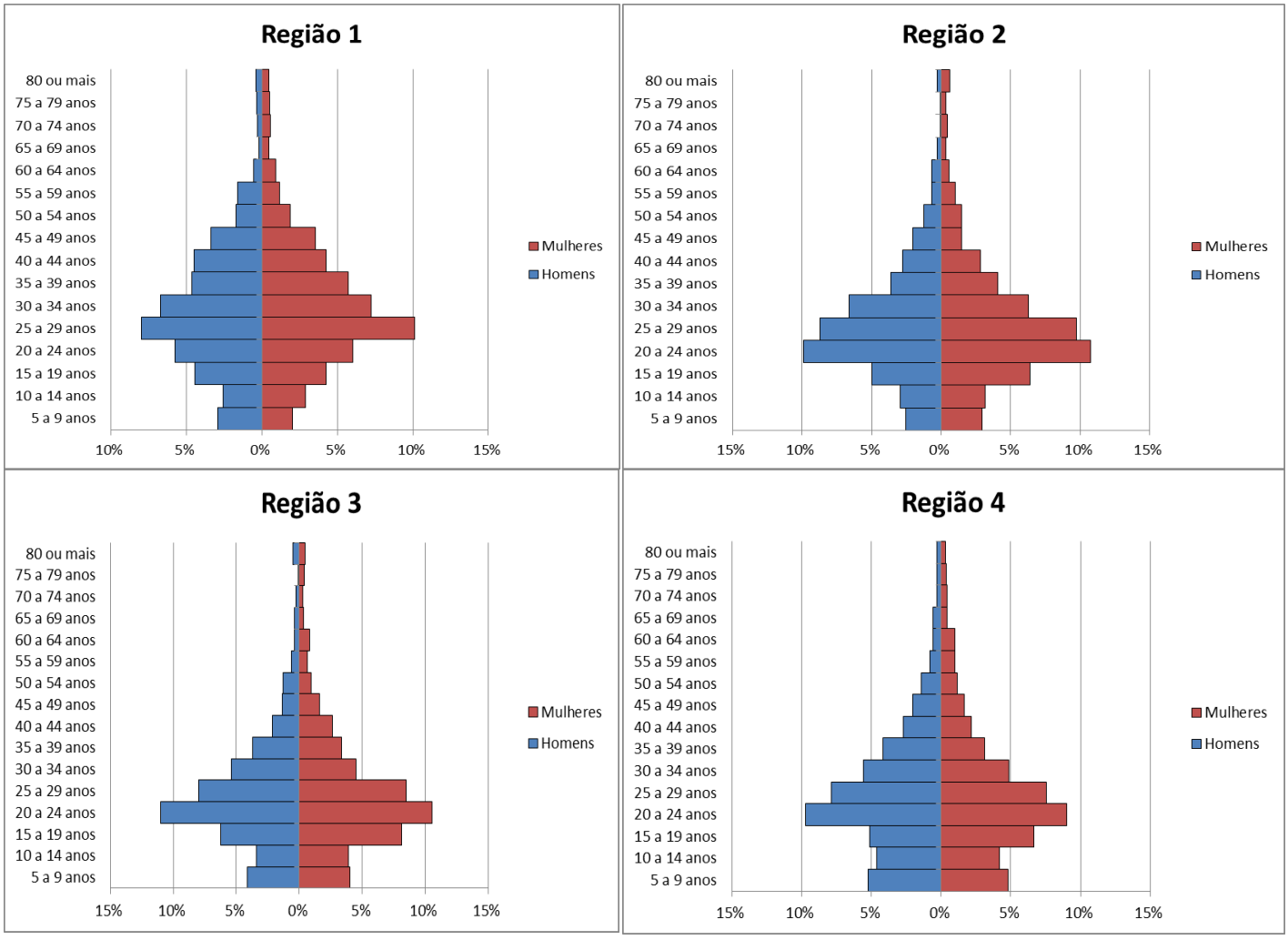
Tabela 4 – Imigrante acima de 20 anos, segundo Idade, da Área Metropolitana de Brasília - 2010.

Características		N					Quartil		
		Válido	Ausente	Média	Mínimo	Máximo	25%	50%	75%
Idade	Sem Ponderação	10.113	0	33,78	20	103	24	30	39
	Ponderado	177.313	0	33,65	20	103	24	30	39

Como já era esperada, a média da idade aumentou, já que foram retirados os imigrantes mais novos, de 28 para 33,8 anos.

5.2. Gráficos

Gráfico 1 – Pirâmide etária dos imigrantes da AMB segundo região de residência – 2010.

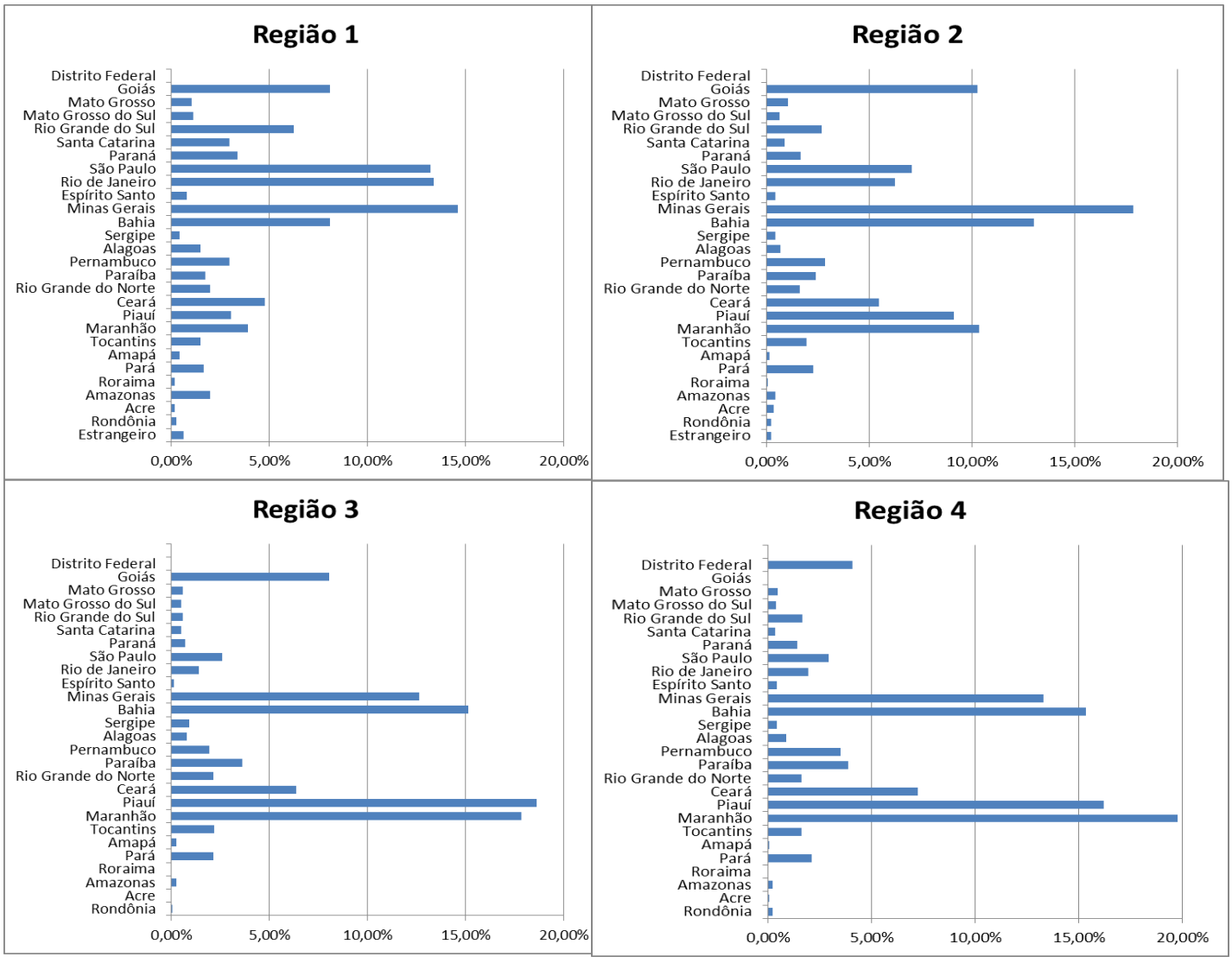


É possível observar pelos gráficos acima que os imigrantes jovens são maioria em todas as quatro regiões estudadas. Podemos observar na Região 4, uma base mais larga, ou seja, uma quantidade proporcional maior de crianças que as outras regiões.

É também visível, que entre os jovens, que são demonstrados no centro da pirâmide, um número maior de mulheres na região 1, enquanto as outras regiões tem uma distribuição equilibrada entre os dois sexos.

Também é necessário analisar que o topo da pirâmide, onde estão evidenciados os idosos, é mais evidente nas regiões 1 e 2.

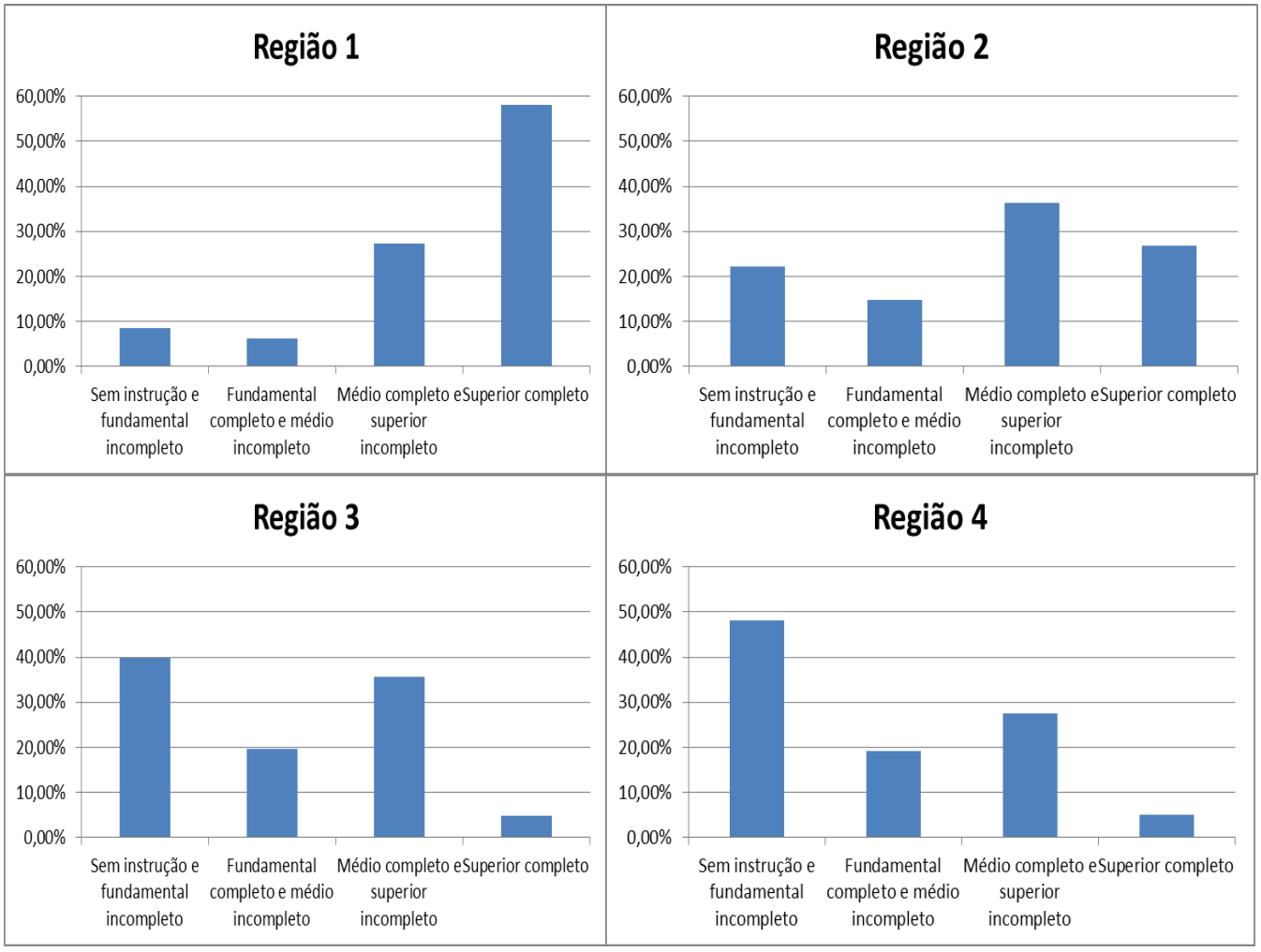
Gráfico 2 – UF de Nascimentos dos imigrantes maiores de 20 anos da AMB – 2010.



Pode-se observar que na região 1 temos maioria dos imigrantes nascidos na região Sudeste do país e dos estados da Bahia e do Goiás; na região 2 temos maioria nascido na região Sudeste e parte do Nordeste e no estado do Goiás; na região 3 temos maioria nascido na região Nordeste e nos estados de Minas Gerais; e na região 4 temos maioria vindos da região Nordeste e do estado de Minas Gerais.

Vale ressaltar que os estados que mais aparecem como estado de nascimento dos imigrantes da AMB, independente da região, são os estados de Minas Gerais e da Bahia.

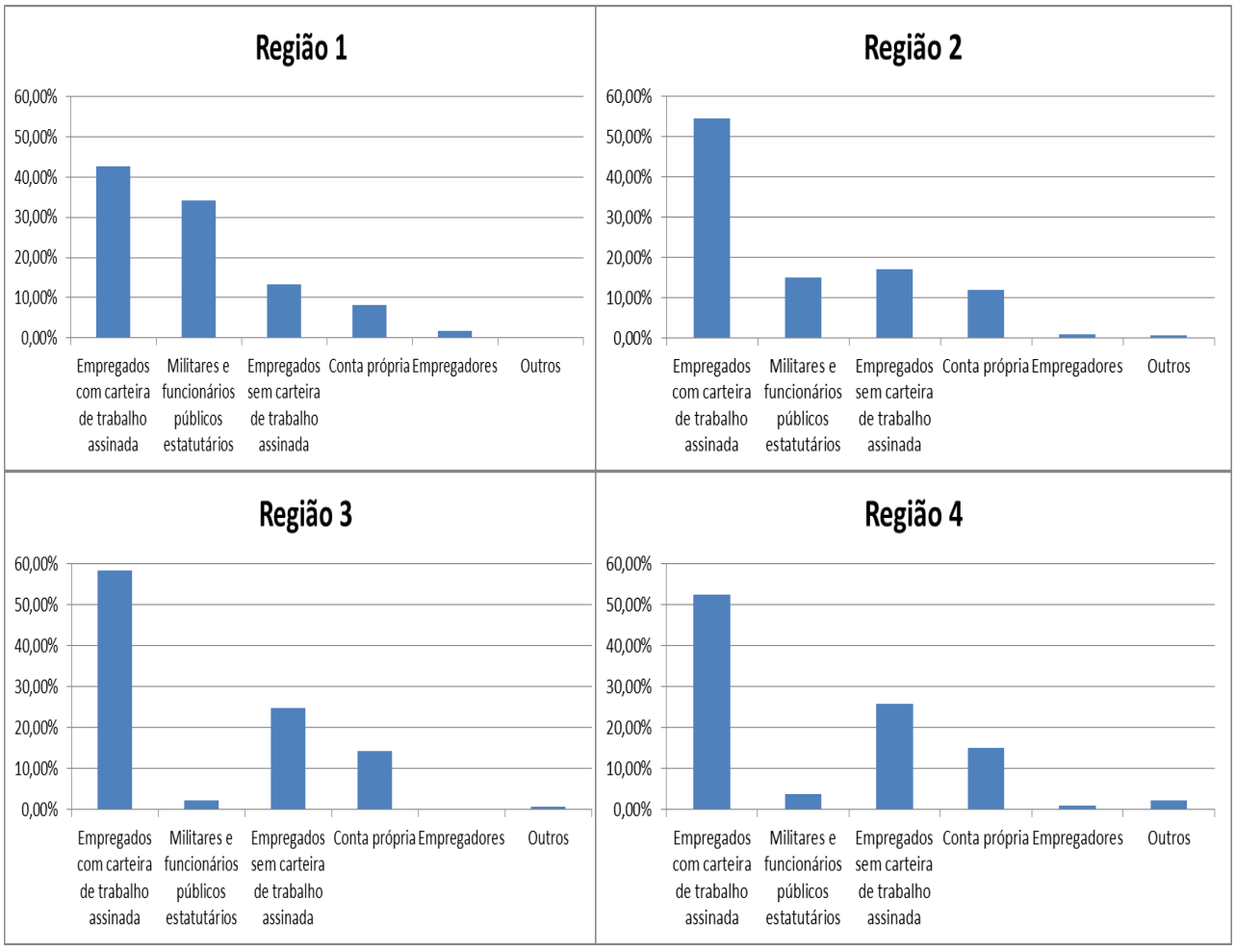
Gráfico 3 – Imigrantes maiores de 20 anos da AMB segundo níveis educacionais – 2010.



A região 1 tem um contraste acerca dos níveis de escolaridade da sua população de imigrantes em comparação com as regiões 3 e 4, dá para ver

claramente que na região 1 há um predomínio do nível superior completo, enquanto nas outras duas regiões citadas há uma dominância dos imigrantes sem instrução ou, no máximo, completo. Na região 2, é possível observar que não há muita variação entre os níveis escolares, ou seja, há pequenas diferenças entre um nível e outro, sendo os imigrantes com Ensino Médio Completo/Superior Incompleto e com Superior Completo a maioria nessa região.

Gráfico 4 – Imigrantes maiores de 20 anos da AMB segundo categorias ocupacionais – 2010.



Em todas as regiões a maioria dos imigrantes é empregada com carteira de trabalho assinada (acima de 40% em todas as regiões), então é preciso levar em consideração a segunda maior ocupação trabalhista. Na região 1 são os Militares e Funcionários públicos; enquanto na região 2, 3 e 4 são aqueles que estão empregados sem carteira de trabalho assinada, gerando uma grande contraste.

5.3. Análise de Agrupamento:

a) Análises de duas etapas.

Tabela 5 – Critérios de BIC.

Número de Grupos	Critério Bayesiano de Schwarz (BIC)	Mudança de BIC	Razão das mudanças de BIC	Razão das Medidas de Distância
1	74768,156			
2	62510,938	-12257,218	1,000	1,819
3	55800,795	-6710,144	,547	1,009
4	49149,668	-6651,126	,543	1,034
5	42717,267	-6432,402	,525	1,543
6	38571,178	-4146,089	,338	1,024
7	34522,033	-4049,145	,330	1,229
8	31239,289	-3282,744	,268	1,221
9	28562,358	-2676,931	,218	1,363
10	26615,193	-1947,165	,159	1,057

Pelo Critério de Informação de Bayes, a maior diferença entre as razões das medidas das distâncias calculada é dada no grupo 2, sendo assim, vamos utilizar a Análise de Duas Etapas com dois grupos.

Tabela 6 – Distribuição dos Grupos.

		N	% de Combinados	% do Total
Cluster	1	5063	50,2%	50,1%
	2	5027	49,8%	49,7%
	Combinado	10090	100,0%	99,8%
Casos excluídos		23		0,2%
Total		10113		100,0%

Os grupos são de tamanhos homogêneos. Os 23 casos excluídos são aqueles que apresentam a variável Instrução sem valor.

Tabela 7 – Centróides da variável Idade.

		Idade		
		Média	Desvio padrão	Mediana
Cluster	1	32,05	10,769	29,04
	2	35,53	14,941	31
	Combinado	33,79	13,131	30,01

Os dois grupos apresentam valores pertos das médias. O desvio padrão é alto, pois a idade varia até 103 anos, e esses valores foram distribuídos entre os grupos de acordo com as medidas de similaridades já estudadas que analisam todas as variáveis.

Tabela 8 – Distribuição dos Grupos.

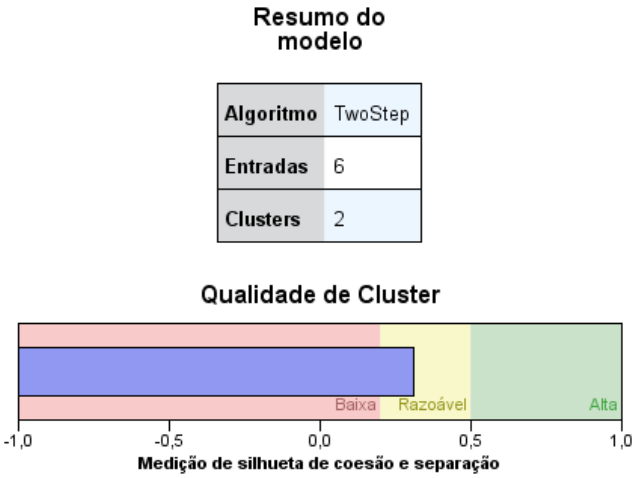
Variáveis		Grupos	
		1	2
Sexo	Homem	46,8%	53,2%
	Mulher	53,5%	46,5%
Cor/Raça	Branca	65,5%	34,5%
	Outros	39,8%	60,2%
Região de Residência Anterior	N/NE/CO	41,6%	69,4%
	S/SE	58,4%	30,6%
Estado Civil	Solteiro	46,4%	53,6%
	Outros	55,6%	44,4%
Grau de Instrução	Sem Instrução a Ensino Fundamental Completo	1,8%	98,2%
	Ens. Médio Completo ou Ens. Superior Completo	100,0%	0,0%

As variáveis sexo e estado civil estão distribuídas homogeneamente dentro dos grupos, ou seja, cada grupo apresenta cerca de 50% de cada opção da variável (com pequenas variações).

A variável Cor/Raça tem cada uma de suas características predominando em cada grupo, no grupo 1 temos 65,5% do grupo com cor branca, enquanto no grupo 2 temos 60,2% do grupo em outra cor/raça. O mesmo é visto na variável de Região de Residência Anterior onde há a predominância das regiões Norte/Nordeste/Centro-Oeste no grupo 2 e no grupo 1, da região Sul/Sudeste.

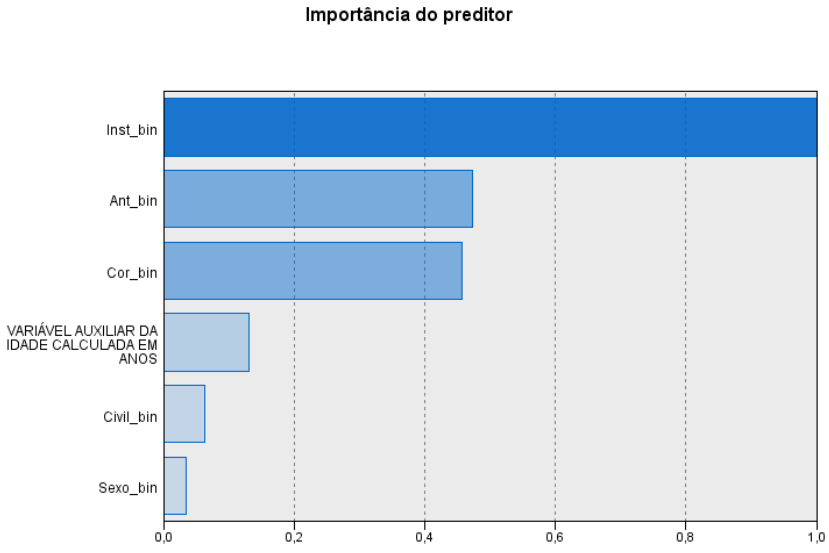
É possível ver claramente que o Grau de Instrução foi uma variável muito importante na divisão dos grupos, o grupo 1 tem praticamente só indivíduos Sem Instrução/Ensino Fundamental Incompleto e Ensino Fundamental Completo/Ensino Médio Incompleto, enquanto no grupo 2, 100% dos indivíduos possuem Ensino Médio Completo ou Ensino Superior Completo.

Figura 1 – Resumo e qualidade do modelo



A figura acima mostra a informação de quantas variáveis foram utilizadas no modelo, no caso foram 6 variáveis, quantos grupos foram formados (2 grupos) e qual a qualidade do nosso modelo. O modelo que está sendo trabalhado foi classificado em Razoável, mais abaixo iremos saber o porquê do modelo não ter sido classificado com qualidade Alta.

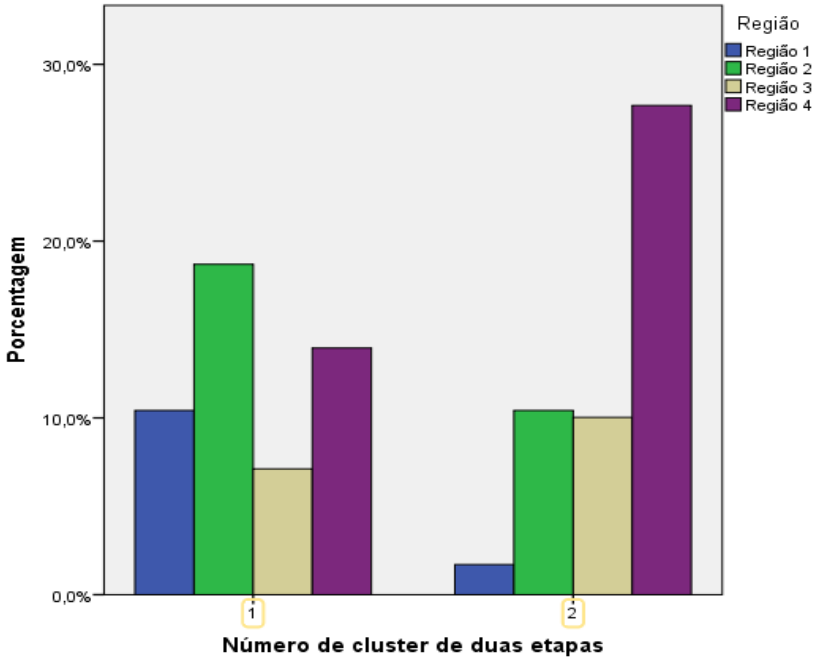
Gráfico 5 – Importância do Preditor.



Nesse gráfico é apresentado a importância de cada variável na formação dos grupos, é perceptível que a variável Grau de Instrução é a mais importante, o que já havia sido percebido, já que na ‘Tabela 6 – Distribuição dos Grupos’ foi visto que em cada grupo só temos uma opção da variável, ou seja, cada grupo é formado com

100% (ou aproximadamente) de uma das opções dessa variável. A forte relação entre a escolha dos grupos e a variável Grau de Instrução é uma das razões da qualidade dos grupos não ser muito alta.

Gráfico 6 – Distribuições das Regiões dentro dos Grupos.



Apesar de não ter utilizado a variável Região na Análise de Agrupamento feita, já que era uma variável altamente correlacionada com as outras variáveis, o que iria interferir na montagem dos grupos, é importante ver como os grupos se comportam dentro dessa variável.

No Grupo 1 não temos nenhum grupo que tenha muito destaque, ele é formado pela junção de todas as regiões, principalmente da região 2 2. No grupo 2, temos uma presença significativa da região 4 e quase não temos a presença da Região 1.

b) Análise de duas etapas (Sem a variável Grau de Instrução).

Como a variável Grau de Instrução estava muito correlacionada com a montagem dos grupos, vamos retirar essa variável e ver se a qualidade dos grupos aumenta.

Tabela 9 – Critérios de BIC.

Número de Clusters	Critério Bayesiano de Schwarz (BIC)	Mudança de BIC	Razão das mudanças de BIC	Razão das Medidas de Distância
1	60919,835			
2	45798,849	-15120,987	1,000	2,088
3	38586,209	-7212,640	,477	1,225
4	32707,198	-5879,010	,389	1,157
5	27633,657	-5073,541	,336	1,184
6	23356,301	-4277,356	,283	1,553
7	20621,201	-2735,100	,181	1,094
8	18124,841	-2496,360	,165	1,199
9	16051,162	-2073,679	,137	1,007
10	13992,815	-2058,347	,136	1,157

O salto na Razão de Medidas de Distância continua sendo no 2º Grupo, sendo assim, o número de grupos continua igual.

Tabela 10 – Distribuição dos Grupos.

		N	% de Combinados
Cluster	1	5967	59,0%
	2	4146	41,0%
	Combinado	10113	100,0%

O grupo 1 é um pouco maior do que o grupo 2, porém se a Razão de Tamanho entre o maior e o menor grupo for menor do que 3, sendo assim, é aceito que os grupos não tem uma diferença tão significativa entre eles. A Razão de tamanho entre o Grupo 1 e o Grupo 2 é de 1,45. Não existem mais casos excluídos da análise, já que eles estavam ligados à ausência de resposta na variável que foi retirada.

Tabela 11 – Centróides da variável Idade.

		Idade		
		Média	Desvio padrão	Mediana
Cluster	1	29,28	9,533	26,00
	2	40,25	14,781	36,85
	Combinado	33,78	13,127	30,00

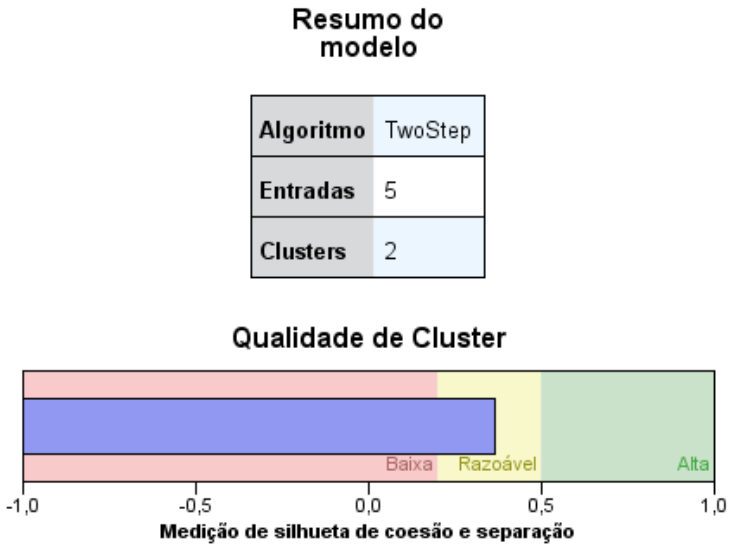
Em comparação à análise anterior, o grupo 1 diminuiu a sua média de idade e o grupo 2 aumentou. O desvio padrão diminui nos dois grupos e a mediana diminui no primeiro e aumentou no segundo.

Tabela 12 – Distribuição dos Grupos.

Variáveis		Grupos	
		1	2
Sexo	Homem	61,1%	38,9%
	Mulher	56,9%	43,1%
Cor/Raça	Branca	50,7%	49,3%
	Outros	64,6%	35,4%
Região de Residência Anterior	N/NE/CO	63,5%	36,5%
	S/SE	49%	51%
Estado Civil	Solteiro	100%	0%
	Outros	0%	100%

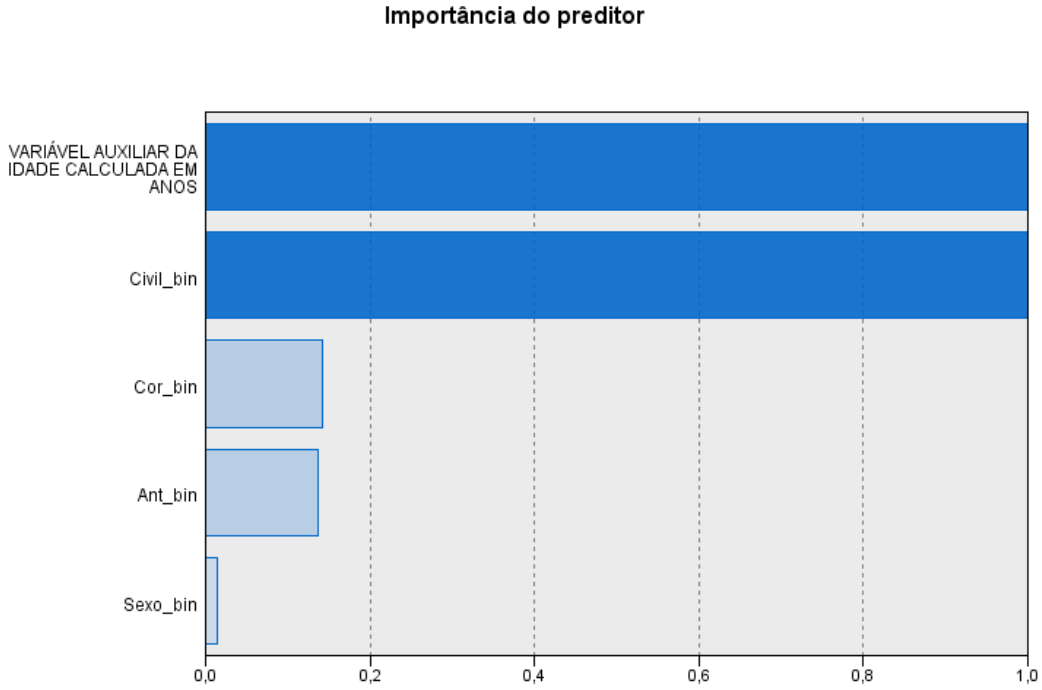
Nessa análise, é possível ver que quase todas as variáveis estão distribuídas homogeneamente, exceto a variável Estado Civil que apresenta somente uma opção em cada um dos grupos.

Figura 2 – Resumo e qualidade do modelo



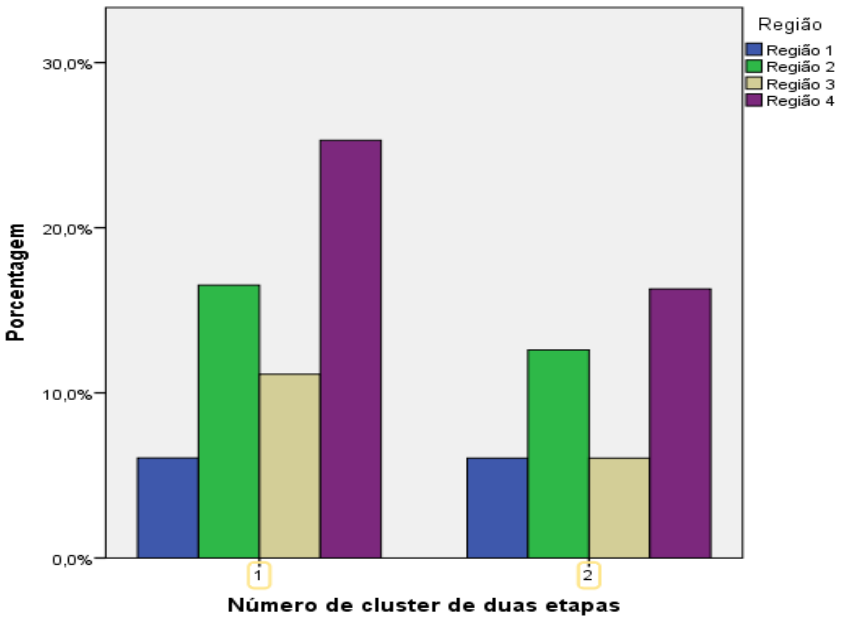
No novo caso, foram utilizadas 5 variáveis que foram divididas em dois clusters. A qualidade dos grupos continua como Razoável.

Gráfico 7 – Importância do Preditor.



Nessa nova análise, as variáveis Idade e Estado Civil que estão fortemente relacionadas a criação dos grupos, sendo isso uma das causas da nova análise também ter a qualidade Razoável.

Gráfico 8 – Distribuições das Regiões dentro dos Grupos.



Uma distribuição mais homogênea das regiões segundo os grupos é obtida, há uma predominância da Região 4 e da Região 2 nos dois grupos, mas essas são as Região com maior número de pessoas, sendo assim, é esperado que elas tenham maior quantidade de indivíduos em cada grupo.

c) Análise de duas etapas usando tratamento de ruído de 20%.

Irá se utilizar um novo recurso chamado Tratamento de Ruído na análise de agrupamento já feito. O Tratamento de Ruído é utilizado para retirar dos grupos os valores outliers que não se encaixam em nenhum grupo e coloca-los como valores discrepantes. Utilizaremos o Tratamento de Ruído com nível de 20%.

Tabela 13 – Critérios de BIC.

Número de Clusters	Critério Bayesiano de Schwarz (BIC)	Mudança de BIC	Razão das mudanças de BIC	Razão das Medidas de Distância
1	47756,47			
2	36874,62	-10881,9	1	2,028
3	31540,92	-5333,7	0,49	1,171
4	26993,34	-4547,58	0,418	1,074
5	22761,53	-4231,81	0,389	1,702
6	20300,46	-2461,08	0,226	1,015
7	17876,38	-2424,08	0,223	1,139
8	15754,79	-2121,59	0,195	1,248
9	14066,85	-1687,94	0,155	1,165
10	12626,19	-1440,67	0,132	1,003

O salto na Razão de Medidas de Distância continua sendo no 2º Grupo, sendo assim, utiliza-se o mesmo número de grupos.

Tabela 14 – Distribuição dos Grupos.

		N	% de Combinados	% do Total
Cluster	1	5599	55,5%	55,4%
	2	3833	38,0%	37,9%
	Valor Discrepante	658	6,5%	6,5%
	Combinado	10090	100,0%	99,8%
Casos excluídos		23		0.2%
Total		10113		100%

O segundo grupo é menor que o primeiro, porém a Razão de Tamanho entre o Grupo 1 e o Grupo 2 é igual a 1,46, sendo assim, o tamanho dos grupos estão bons. Vemos que 6,5% dos casos não se encaixaram em nenhum grupo, por isso foram retirados da análise de agrupamento.

Tabela 15 – Centróides da variável Idade.

		Idade		
		Média	Desvio padrão	Mediana
Cluster	1	30,54	9,169	27,98
	2	32,83	9,976	30,04
	Valor Discrepante	67,00	12,278	
	Combinado	33,79	13,131	29,05

A média das idades dos grupos 1 e 2 continuam perto do valor de 30 anos e o desvio padrão entre eles diminuiu consideravelmente com a retirada dos valores discrepantes.

Os valores discrepantes têm como média de 67 anos, o que é esperado, já que os jovens foram alocados nos outros grupos (que diminuíram o desvio padrão). A mediana está mais perto da média.

Tabela 16 – Distribuição dos Grupos.

Variáveis		Grupos		
		1	2	Valor Discrepante
Sexo	Homem	57,2%	37,1%	5,7%
	Mulher	53,8%	38,9%	7,3%
Cor/Raça	Branca	0%	94%	6%
	Outros	93,1%	0%	6,9%
Região de Residência Anterior	N/NE/CO	63,8%	31,7%	4,4%
	S/SE	36,7%	52%	11,2%
Estado Civil	Solteiro	64,5%	34,2%	1,3%
	Outros	42,5%	43,5%	14%
Grau de Instrução	Sem Instrução a Ensino Fundamental Completo	65,1%	25,6%	9,2%
	Ens. Médio Completo ou Ens. Superior Completo	45,5%	50,7%	3,7%

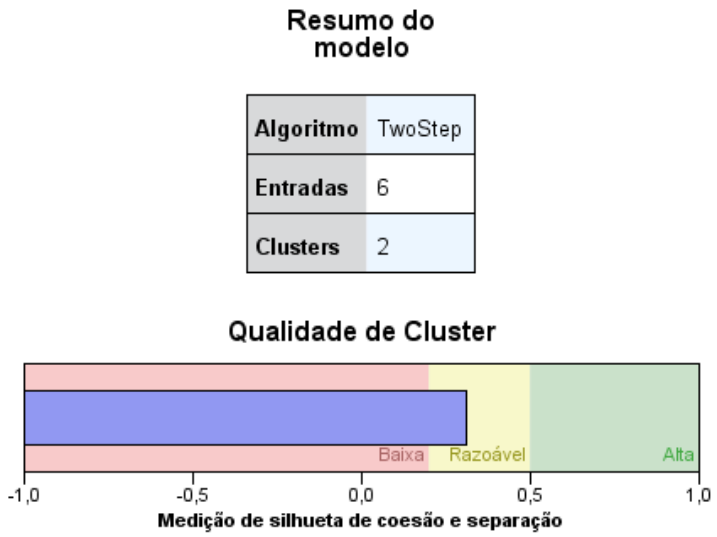
No 1º Grupo tem uma predominância de indivíduos com Cor/Raça Parda/Preta/Outros nos dois sexos, que residiam anteriormente nas Regiões

Norte/Nordeste/Centro-Oeste, Solteiro e Sem Instrução ou Ensino Fundamental Completo.

No Grupo 2, predominam os indivíduos de Cor/Raça branca de ambos os sexos, que residiam anteriormente nas Regiões Sul/Sudeste, não são solteiros e tem Ensino Médio ou Ensino Superior Completo.

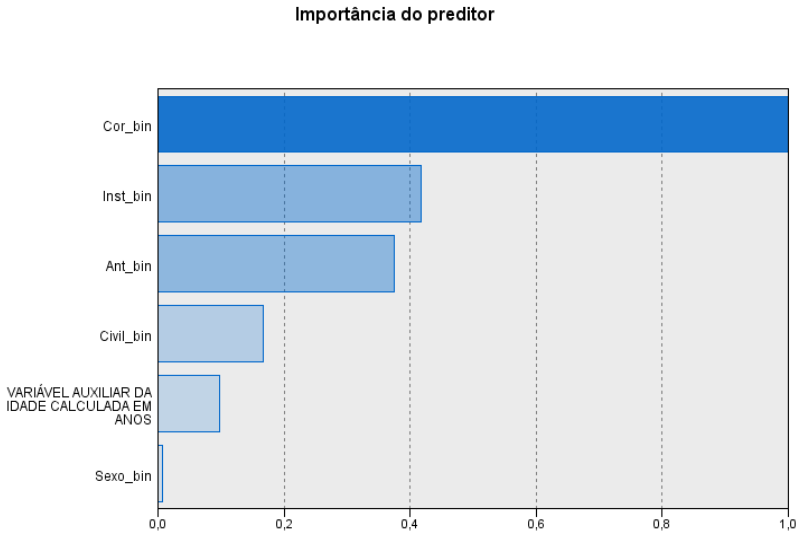
Os valores discrepantes apresentam indivíduos nos dois sexos, com ambas as opções de Cor/Raça, na sua maioria que residiam anteriormente nas Regiões Sul/Sudeste, não solteiros e sem instrução ou somente Ensino Fundamental Completo.

Figura 3 – Resumo e qualidade do modelo



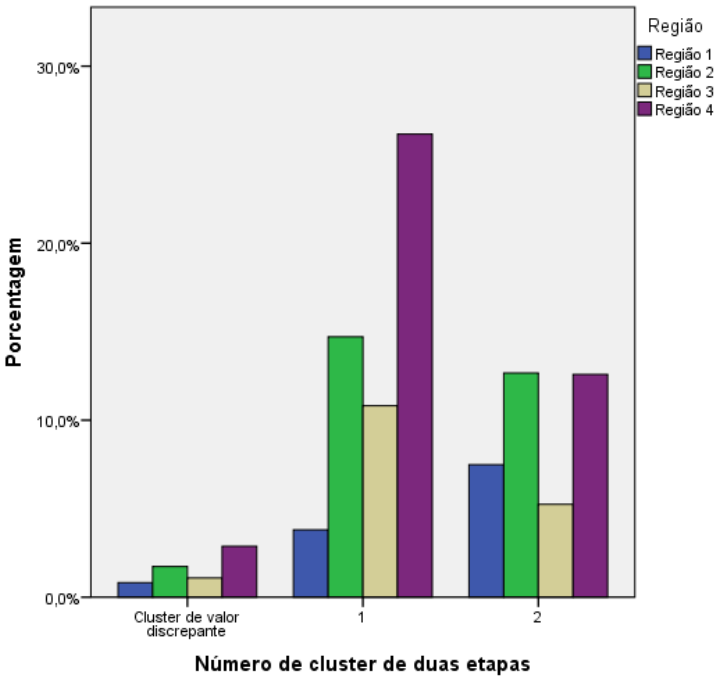
O modelo utilizado é constituído de 6 variáveis e os indivíduos divididos e, 2 grupos. A qualidade é também considerada Razoável.

Gráfico 9 – Importância do Preditor.



A variável Cor/Raça é fortemente relacionada com a construção dos grupos, enquanto que as outras variáveis têm menos de 45% na importância da montagem grupos, por essa razão, a qualidade dos grupos é considerada somente como razoável,

Gráfico 10 – Distribuições das Regiões dentro dos Grupos.



Os indivíduos com valores discrepantes que foram retirados da análise são homogêneos em relação a região, há uma pequena diferença em relação as Regiões 2 e 4, mas elas são as regiões que mais aparecem na população de imigrantes estudados como um todo.

No Grupo 1, há baixo aparecimento de indivíduos da Região 1. No Grupo 2, há igualmente a presença de indivíduos da Região 2 e 4 e há um número alto (proporcional a sua quantidade total) de indivíduos da Região 1.

d) Análise de duas etapas usando tratamento de ruído de 20% (Sem a Variável Raça/Cor)

Como a variável Cor/Raça estava muito correlacionada com a montagem dos grupos, retira-se essa variável para ver se há algum impacto na qualidade dos grupos.

Tabela 17 – Critérios de BIC.

Número de Clusters	Critério Bayesiano de Schwarz (BIC)	Mudança de BIC	Razão das mudanças de BIC	Razão das Medidas de Distância
1	41099,438			
2	31275,880	-9823,557	1,000	1,350
3	24012,256	-7263,624	,739	1,364
4	18701,104	-5311,152	,541	1,641
5	15485,055	-3216,049	,327	1,229
6	12877,395	-2607,660	,265	1,135
7	10587,192	-2290,203	,233	1,096
8	8502,747	-2084,444	,212	1,604
9	7223,396	-1279,351	,130	1,030
10	5982,904	-1240,492	,126	1,131

Nesse caso, diferente com as análises já feitas, a maior variação da Razão das Medidas de Distâncias ocorre quando a quantidade de grupos é igual a 8, por isso, será utilizado esse como nosso número de grupos.

Tabela 18 – Distribuição dos Grupos.

		N	% de Combinados	% do Total
Cluster	1	1079	10,7%	10,7%
	2	1342	13,3%	13,3%
	3	994	9,9%	9,8%
	4	1147	11,4%	11,3%
	5	1500	14,9%	14,8%
	6	791	7,8%	7,8%
	7	1475	14,6%	14,6%
	8	1144	11,3%	11,3%
	Valor Discrepante	618	6,1%	6,1%
	Combinado	10090	100,0%	99,8%
	Casos excluídos	23		0,2%
	Total	10113		100%

O maior grupo abrange 14,9% do total e o menor 7,8%, o que leva a uma Razão de Tamanho entre o Grupo 5 e o Grupo 6 igual a 1,90, sendo assim, o tamanho dos grupos estão bons. Temos 618 indivíduos no grupo discrepante (6,1% dos estudados) e 23 indivíduos que foram excluídos da análise, pois possuem alguma informação em branco.

Tabela 19 – Centróides da variável Idade.

		Idade		
		Média	Desvio padrão	Mediana
Cluster	1	37,85	10,820	35,13
	2	38,23	14,077	35,07
	3	27,53	7,367	25,08
	4	39,07	12,464	36,98
	5	29,88	8,460	27,96
	6	26,36	6,161	25,00
	7	29,08	8,778	26,92
	8	30,39	10,359	26,99
	Valor Discrepante	53,83	20,775	
	Combinado	33,79	13,131	29,10

As médias das idades continuam por volta de 30 anos, assim como as medianas. O desvio padrão está entre 6,161 (Grupo 6) e 14,077 (Grupo 2) entre os grupos.

O grupo com a maior média de idade é o Grupo 4 com média de 39,07 anos e o com menor média é o Grupo 6, com 26,36 anos. O grupo com a maior mediana é o Grupo 4 com mediana de 36,98 anos e o com menor mediana é o grupo 6, com 25,00 anos e que são justamente os grupos com maiores e menores médias.

Os valores discrepantes tem média de 53,83 anos e um desvio padrão de 20,775, o que já é esperado, já que ele absorve os indivíduos com maiores e menores idades que não entraram em nenhum grupo.

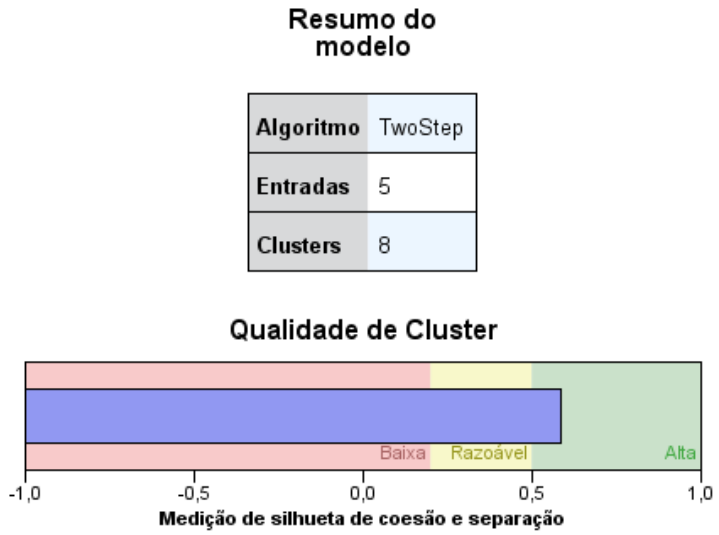
Tabela 20 – Distribuição dos Grupos.

Variáveis		Grupos								
		1	2	3	4	5	6	7	8	Valores Discrepantes
Sexo	Homem	10,70%	0,00%	0,00%	22,80%	15,80%	15,70%	29,30%	0,00%	2,70%
	Mulher	10,70%	26,60%	19,70%	0,00%	13,90%	0,00%	0,00%	22,60%	6,50%
Região de Residência Anterior	N/NE/CO	0,00%	19,20%	14,20%	16,40%	0,00%	11,30%	21,10%	16,40%	1,20%
	S/SE	34,70%	0,00%	0,00%	0,00%	48,20%	0,00%	0,00%	0,00%	17,10%
Estado Civil	Solteiro	0,00%	0,00%	16,70%	0,00%	25,20%	13,30%	24,80%	19,20%	0,00%
	Outros	26,10%	32,50%	0,00%	27,70%	0,00%	0,00%	0,00%	0,00%	13,70%
Grau de Instrução	Sem Instrução a Ensino Fundamental Completo	0,00%	14,20%	0,00%	12,70%	10,50%	0,00%	28,80%	22,30%	11,40%
	Ens. Médio Completo ou Ens. Superior Completo	21,70%	12,40%	20,00%	10,00%	19,40%	15,90%	0,00%	0,00%	0,60%

Ao observar o gráfico acima, é possível perceber que os grupos 2, 3 e 8 não apresentam indivíduos homens em sua composição e os grupos 4, 6 e 7 não tem mulheres. Os grupos 1 e 5 apresentam somente indivíduos que não residiam nas Regiões Norte, Nordeste e Centro-Oeste e os demais grupos, nas Regiões Sul e Sudeste. Os grupos 1, 2 e 4 só possuem indivíduos que não são solteiros e nos demais grupos somente solteiros. Os grupos 1, 3 e 6 só possuem indivíduos com Ensino Médio ou Superior Completos, nos grupos 7 e 8 só possuem indivíduos sem instrução e Ensino Fundamental Completo.

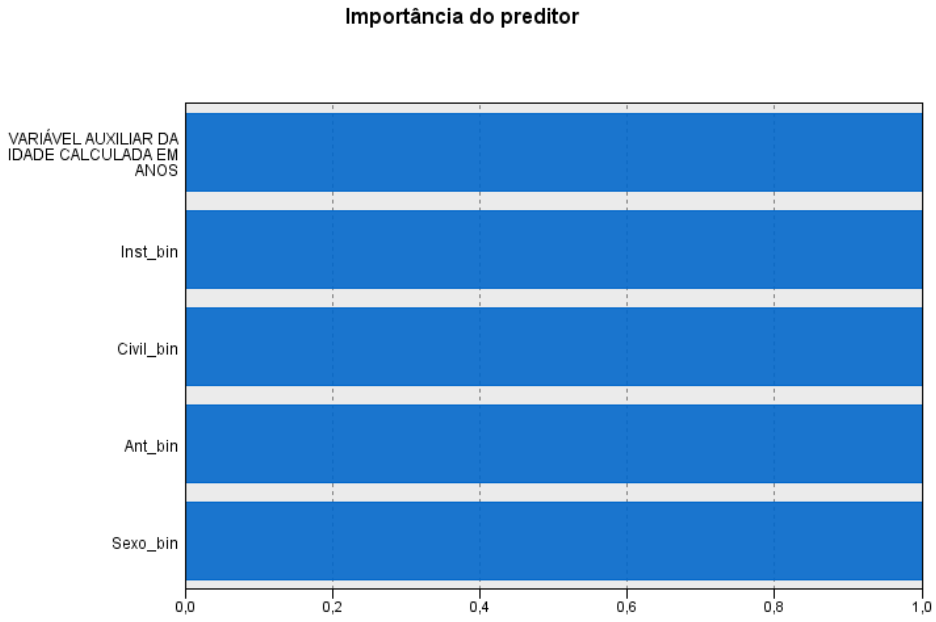
Nos valores discrepantes não possuem indivíduos solteiros e quase nenhum com Ensino Médio ou Superior Completos.

Figura 4 – Resumo e qualidade do modelo



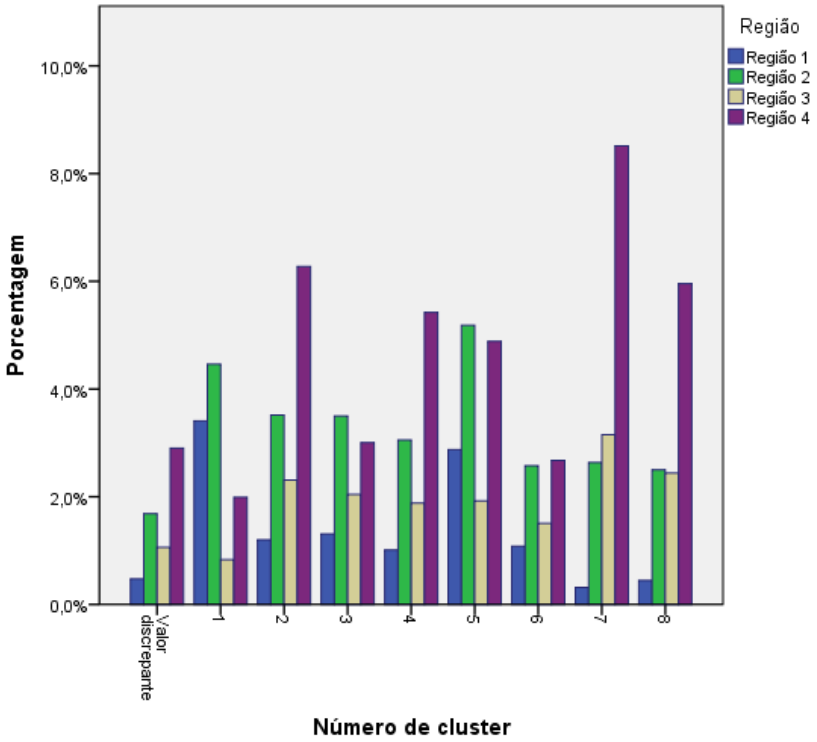
Foi obtido um modelo com o modelo inicial com 5 variáveis e 8 grupos. A qualidade é considerada Alta e abaixo será demonstrada a diferença entre esse modelo e os outros já feitos anteriormente e o porquê dele ter considerado com qualidade alta.

Gráfico 11 – Importância do Preditor.



Diferente das outras análises já feitas, todas as variáveis tem muita importância para a montagem dos grupos, o que faz a diferença na qualidade da análise.

Gráfico 12 – Distribuições das Regiões dentro dos Grupos.



Diferente das outras análises, agora há grupos onde a maioria dos indivíduos faz parte da Região 4.

No grupo1, predomina os indivíduos das Regiões 1 e 2 e nos grupos 3, 5 e 6, predominam os indivíduos das Regiões 2 e 4.

6. Discussão e Conclusões

A Análise de Duas Etapas foi o método escolhido para a nossa análise, pois ela pode ser utilizada em dados com variáveis categóricas e variáveis contínuas, pode rapidamente formar grupos com banco de dados grande como o utilizado, que, apesar de todos os cortes feitos, continuava um banco muito grande com 10.113 observações. A Análise de Agrupamento de Duas Etapas foi desenvolvida para lidar com o problema de bancos de dados com variáveis mistas, ou seja, onde haja variáveis numéricas e categóricas.

O Método de Duas Etapas combina os princípios da Análise Hierárquica e da Análise não-Hierárquica e diferente do Método de K-médias não é necessário indicar o número de grupos que se quer usar, ele utiliza o Critério de Informação de Akaike ou o Critério de Informação de Bayes pra estimar um bom inicial de máximo de número de grupos.

Após ter sido feita a análise com quatro diferentes modelos (com todas as variáveis selecionadas e sem tratamento de ruído, sem a variável 'Grau de Escolaridade' e sem tratamento de ruído, com todas as variáveis selecionadas e com tratamento de ruído e sem a variável 'Raça/Cor' e com tratamento de ruído), foi concluído que a melhor o forma de agrupar os dados foi dado no último modelo feito, onde havia somente 5 variáveis (Idade, Grau de Instrução, Estado Civil, Local de Residência Anterior e Sexo) e foi utilizado o tratamento de ruído num nível de 20%.

Oito grupos foram formados, sendo ele:

- Grupo 1: Separado igualmente entre ambos os sexos com idade média de 37 anos que moraram anteriormente na Região Sul/Sudeste e onde todos os indivíduos do grupo tem pelo menos o Ensino Médio Completo, casados, separados ou viúvos;
- Grupo 2: Mulheres com idade média de 38 anos que moraram anteriormente na Região Norte/Nordeste/Centro-Oeste, 54,2% do grupo não tem Instrução ou somente Ensino Fundamental Completo, todos os indivíduos são casado, separado ou viúvo;
- Grupo 3: Mulheres com idade média de 28 anos que moraram anteriormente na Região Norte/Nordeste/Centro-Oeste com pelo menos Ensino Médio Completo, todos os indivíduos são solteiros;
- Grupo 4: Homens com idade média de 39 anos que moraram anteriormente na Região Norte/Nordeste/Centro-Oeste, 56,8% do grupo não tem Instrução ou somente Ensino Fundamental Completo, todos os indivíduos são casado, separado ou viúvo;
- Grupo 5: Separado igualmente entre ambos os sexos com idade média de 30 anos que moraram anteriormente na Região Sul/Sudeste, 64,3% do grupo pelo menos Ensino Médio Completo, todos os indivíduos são solteiros;

- Grupo 6: Homens com idade média de 26 anos que moraram anteriormente na Região Norte/Nordeste/Centro-Oeste, com pelo menos Ensino Médio Completo, onde todos os indivíduos são solteiros;
- Grupo 7: Homens com idade média de 29 anos que moraram anteriormente na Região Norte/Nordeste/Centro-Oeste, sem nenhuma Instrução ou somente Ensino Fundamental Completo, onde todos os indivíduos são solteiros;
- Grupo 8: Mulheres com idade média de 30 anos que moraram anteriormente na Região Norte/Nordeste/Centro-Oeste, sem nenhuma Instrução ou somente Ensino Fundamental Completo, onde todos os indivíduos são solteiros.

Tabela 21 – Distribuição dos Grupos dentro das Regiões.

	Grupos								
	Valores discrepante	1	2	3	4	5	6	7	8
Região 1	3,92%	28,13%	9,89%	10,79%	8,34%	23,71%	8,91%	2,62%	3,68%
Região 2	5,79%	15,32%	12,09%	12,02%	10,49%	17,81%	8,85%	9,06%	8,58%
Região 3	6,18%	4,86%	13,47%	11,91%	10,98%	11,21%	8,79%	18,38%	14,22%
Região 4	6,98%	4,79%	15,07%	7,21%	13,02%	11,74%	6,43%	20,45%	14,31%

A tabela acima mostra a distribuição dos indivíduos segundo a Região de moradia dentro de cada grupo em que foi alocado. Os indivíduos da Região 1 estão distribuída em sua maioria nos Grupos 1 e 5; os da Região 2, nos Grupos 1, 2, 3 e 5; os da Região 3, nos Grupos 2, 7 e 8; e os da Região 4, nos Grupos 2, 7 e 8. Os grupos 4 e 6 tem distribuições homogêneas nas regiões.

Os resultados acima já eram esperados já que os Grupos 1 e 5 tiveram predominância de indivíduos que residiam anteriormente da Região Sul e Sudeste com pelo menos o Ensino Superior completo, características mais observadas nas Regiões 1 e 2. E os Grupos 2, 7 e 8 contém indivíduos que residiam anteriormente da Região Norte, Nordeste e Centro-Oeste e em sua maioria sem Instrução ou somente Ensino Fundamental Completo, características mais observadas nas Regiões 3 e 4.

7. Referências Bibliográficas

ALBIERI, SÔNIA. Trabalhando com amostragem complexa: a produção de informações estatísticas da área sociodemográfica da Diretoria de Pesquisas do IBGE. I POSDEM – Encontro Nacional de Pós-graduação em Demografia e Áreas Afins –Campinas, 24.02.2010.

AMARAL, ERNESTO F. L.; RODRIGUES, ROBERTO N.; FÍGOLI, MOEMA G. B. Síntese da migração em Goiás e no Distrito Federal nas últimas décadas. Sociedade e Cultura, julho-dezembro, ano/vol. 5, número 002, página 127-136. UFG, 2002.
approaches to microlevel studies in developed and developing countries. New York:

BACHER, JOHANN; WENZIG, KNUT; VOGLER, MELANIE. SPSS twostep cluster. A first evaluation. Universität Erlangen-Nürnberg. 2004.

BARROSO, L.P.; ARTES, R. Análise multivariada. Lavras: UFLA, 2003.

CENSO 2010. Disponível em: http://www.ibge.gov.br

DE JONG, G.F.; GARDNER, R.W (Eds.). Migration decision making: multidisciplinary. Pergamon Press, 1981.

GOLGHER, ANDRÉ B. *Fundamentos da migração,* Belo Horizonte: UFMG/CEDEPLAR, 2004.

IBGE. Notas Técnicas – Histórico da investigação sobre cor ou raça nas pesquisas domiciliares do IBGE. 2008.

IPEA – Instituto de Pesquisa Econômica Aplicada. Comunicados do IPEA: Perfil dos migrantes em São Paulo. 2011.

JOHNSON, RICHARD A. e WICHERN, DEAN W. *Applied Multivariate Statistical Analysis,* Ed. Pearson, 2007.

MINGOTI, SUELI A. *Análise de dados através de métodos de estatística multivariada,* ed. UFMG, 2007.

MOOI, E.; SARSTEDT, M. The process, Data, and Methods Using IBM SPSS Statistics. 2011.

PAVIANI, Aldo [et al]. Brasilia 50 anos, da capital a metrópole. Brasília: Editora UNB, 2010.

SZWARCWALD, CÉLIA L.; DAMACENA, GISELI N. Amostras complexas em inquéritos populacionais: planejamento e implicações na análise estatística dos dados. Revista Bras Epidemiol, páginas 38-45. 2008.

VASCONCELOS, A.M.N ;CESAR, L.J.T.; COSTA, M.T.L. A Universidade de Brasília e o acesso ao ensino superior na Área Metropolitana de Brasília. A publicar.

VASCONCELOS, ANA MARIA N. [et al]. Da utopia à realidade: uma análise dos fluxos migratórios para o aglomerado urbano de Brasília. Trabalho apresentado no XV Encontro Nacional de Estudos Populacionais, ABEP. Caxambu, 2006.