PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO

PUC-SP

CLASSIFICAÇÃO DO BRASIL

Focando principalmente indicadores relacionados a habitação, educação, saúde e muito particularmente TRABALHO

MÉTODOS QUANTITATIVOS NA PESQUISA EMPÍRICA

Diego de Melo Conti

CAP I ANÁLISE DE CONGLOMERADOS

1. INTRODUÇÃO.

O presente trabalho tem por objetivo efetuar uma análise exploratória dos dados relacionados ao Indicador Social de Desenvolvimento dos Municípios (ISDM), mais profundamente de suas variáveis analíticas relacionadas ao 'trabalho'. Além disso, será feita uma análise do Índice FIRJAN de Desenvolvimento Municipal (IFDM). De tal forma, iniciaremos este trabalho com a definição e discussão de ambos indicadores, apresentando um breve histórico, o seu funcionamento metodológico e as suas variáveis. Na sequencia será realizada uma análise e interpretação dos dados que serão manipulados, utilizando as ferramentas do software estatístico **MINITAB.**

2. ENTENDENDO OS INDICADORES.

2.1. Indicador Social de Desenvolvimento dos Municípios – ISDM.

O Indicador Social de Desenvolvimento Municipal (ISDM) foi desenvolvido pelo Centro de Microeconomia Aplicada da Fundação Getúlio Vargas (FGV). O ISDM é um indicador sintético, que tem como objetivo reunir num único indicador vários aspectos referentes ao desenvolvimento social de um município, tornando possível a comparação do desempenho de todas as cidades brasileiras, de forma transversal ou longitudinal, medindo a performance dos entes federativos nas dimensões analisadas.

Em termos metodológicos, o indicador é calculado a partir de dados secundários, tendo como fontes o Instituto Brasileiro de Geografia e Estatística (IBGE), o Ministério da Saúde e o Ministério da Educação. Ele abrange cinco dimensões: Habitação, Renda, Trabalho, Saúde e Segurança e Educação. Essas dimensões e as variáveis que as compõem foram escolhidas de maneira a englobar algumas das questões mais prementes nas políticas públicas direcionadas para o município.

2.2. Índice Firjan de Desenvolvimento Municipal – IFDM.

O Índice FIRJAN de Desenvolvimento Municipal (IFDM) é um estudo anual do Sistema FIRJAN que acompanha o desenvolvimento de todos os municípios

brasileiro. O índice é baseado em três pilares: Emprego & Renda, Educação e Saúde. Ele é feito, exclusivamente, com base em estatísticas públicas oficiais, disponibilizadas pelos ministérios do Trabalho, Educação e Saúde.

De leitura simples, o índice varia de 0 a 1. Quanto mais próximo de 1, maior o desenvolvimento da localidade. Além disso, sua metodologia possibilita determinar, com precisão, se a melhora relativa ocorrida em determinado município decorre da adoção de políticas específicas ou se o resultado obtido é apenas reflexo da queda dos demais municípios.

2.3. As variáveis.

São 6 (seis) as variáveis desta pesquisa, incluindo o nome dos indicadores. A seguir uma breve explanação através da tabela 1.

TABELA 1 - AS VARIÁVEIS

VARIÁVEL	SIGNIFICADO	TIPO	UNIDADE DE MEDIDA
ISDM	Indicador Social de Desenvolvimento dos Municípios (ISDM). Trata-se de uma média ponderada de diferentes indicadores analíticos: Habitação (H), Renda (R), Trabalho (T), Saúde & Segurança (S) e Educação (E), padronizada pela média do Brasil.	Variável Quantitativa. UF / Município.	Numérico
T	Indicador da dimensão Trabalho. Trata-se da média ponderada dos indicadores da dimensão Trabalho (T1_1, T1_2 e T2_1) padronizada pela média do Brasil.	Variável Quantitativa. UF / Município.	Numérico
T1_1	Taxa de ocupação. Percentual da população economicamente ativa (PEA) que esteja ocupada na semana de referência. Pessoas ocupadas podem ser empregados, empregadores, conta própria e não remunerados. Define-se como PEA a população entre 15 e 60 anos, que esteja ocupada (incluindo pessoas que estavam de férias) ou procurando emprego, exceto os deficientes físicos. Foram consideradas deficiências físicas a Tetraplegia (paralisia permanente total de ambos os braços e pernas), Paraplegia (paralisia permanente das pernas), Hemiplegia (paralisia permanente de um dos lados do corpo) ou Falta de membro ou de parte dele (falta de perna, braço, mão, pé ou do dedo polegar ou a falta de parte da perna ou braço).	Variável Quantitativa. UF / Município.	Numérico
T1_2	Taxa de formalização entre os empregados. Percentual dos empregados ocupados na semana de referência	Variável Quantitativ	

	no setor formal, dentre o total de empregados da PEA. Define-se como empregados ocupados no setor formal aqueles que possuíam carteira de trabalho assinada. Define-se como PEA a população entre 15 e 60 anos, que esteja ocupada (incluindo pessoas que estavam de férias) ou procurando emprego, exceto os deficientes físicos. Foram consideradas deficiências físicas a Tetraplegia (paralisia permanente total de ambos os braços e pernas), Paraplegia (paralisia permanente das pernas), Hemiplegia (paralisia permanente de um dos lados do corpo) ou Falta de membro ou de parte dele (falta de perna, braço, mão, pé ou do dedo polegar ou a falta de parte da perna ou braço).	a. UF / Município.	Numérico
T2_1	Taxa de trabalho infantil. Percentual das crianças de 10 a 14 anos que se encontram trabalhando ou procurando emprego na semana de referência em relação a população total residente dessa mesma faixa etária.	Variável Quantitativa. UF / Município.	Numérico
IFDM	Índice Firjan de Desenvolvimento Municipal – IFDM. O índice é resultado da média ponderada de diferentes indicadores: Emprego & Renda, Educação e Saúde.	Variável Quantitativa. UF / Município.	Numérico

2.4. A Tabela de Dados.

Os dados utilizados nesta pesquisa foram extraídos da planilha “ISDM por município 2000 e 2010 FGV Firjan IF GF IFDM”, tendo considerado os seguintes elementos:

UF2	Município	IFDM	ISDM	T	T1_1	T1_2	T2_1

3. ANÁLISE DAS VARIÁVEIS

3.1 VARIÁVEIS CATEGÓRICAS

Este tipo de variável indica que o foco de concentração deve ser a análise de gráficos do tipo *pie chart* e barras.

3.1.1 Variável: “Estado”

Fazem parte desta pesquisa os 27 estados brasileiros e suas cidades. O gráfico abaixo exibe o número de cidades por estado.

A variação no número de cidades por estado é acentuada. Considerando que o Distrito Federal é um estado brasileiro, é o estado com o menor número de cidades (1), enquanto o Mato Grosso possui mais de 852 cidades.

3.1.2 Variável: “REGIÃO”

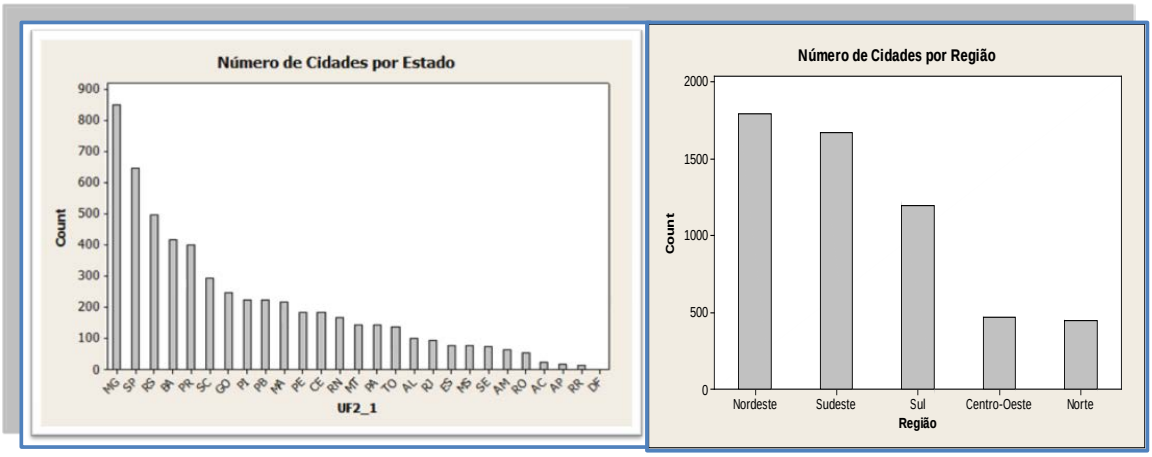


Figura 1. Número de Cidades por Estado e Região do Brasil

Podemos verificar no gráfico acima que a Região Nordeste é a que possui o maior número de cidades do Brasil (1790) e seguido pela Região Sudeste (1669). A Região que possui o menor número de cidades é a Norte, com 447 cidades, muito próxima da Região Centro-Oeste (468). A Região Sul possui 1191 cidades.

3.2 VARIÁVEIS QUANTITATIVAS

A análise deste tipo de variável permite a utilização de uma maior gama de ferramentas de análise como histogramas, curvas de densidade, gráfico de ramos, box-plot e dot-plot, além de informações numéricas como média, desvio-padrão, mediana, quartis, 5 números, intervalo de confiança e teste de normalidade de Anderson-Darling.

3.2.1. DENDOGRAMA DE TRABALHO POR ESTADO (-DF)

O Dendograma permite uma análise do grau de similaridade dos dados para uma determinada variável. Em seguida geramos o Dendograma de Trabalho por Estado

STAT >> MULTIVARIATE >> CLUSTER OBSERVATION

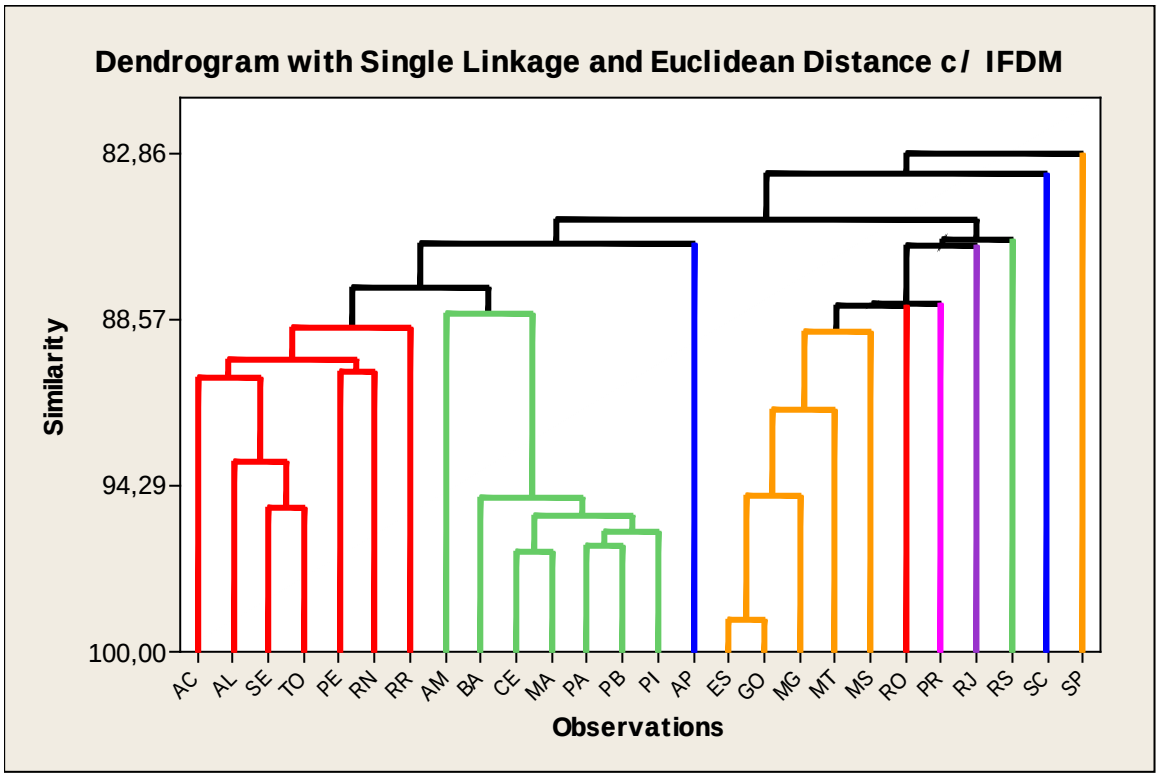


Figura 2. Dendograma da variável Educação por estados do Brasil (classificação não supervisionada)

Na figura acima podemos verificar dtrês grandes grupos de variáveis, agrupadas pela similaridade dos dados. Além disso, uma série de outras variáveis não tão similares (cada uma de uma cor). Os estados que possuem maior similaridade são ES e GO no grupo laranja e CE e MA no grupo verde. O nível de similaridade dos dados destes estados está por volta de 98%, conforme indicado na escala apresentada no exio Y do gráfico.

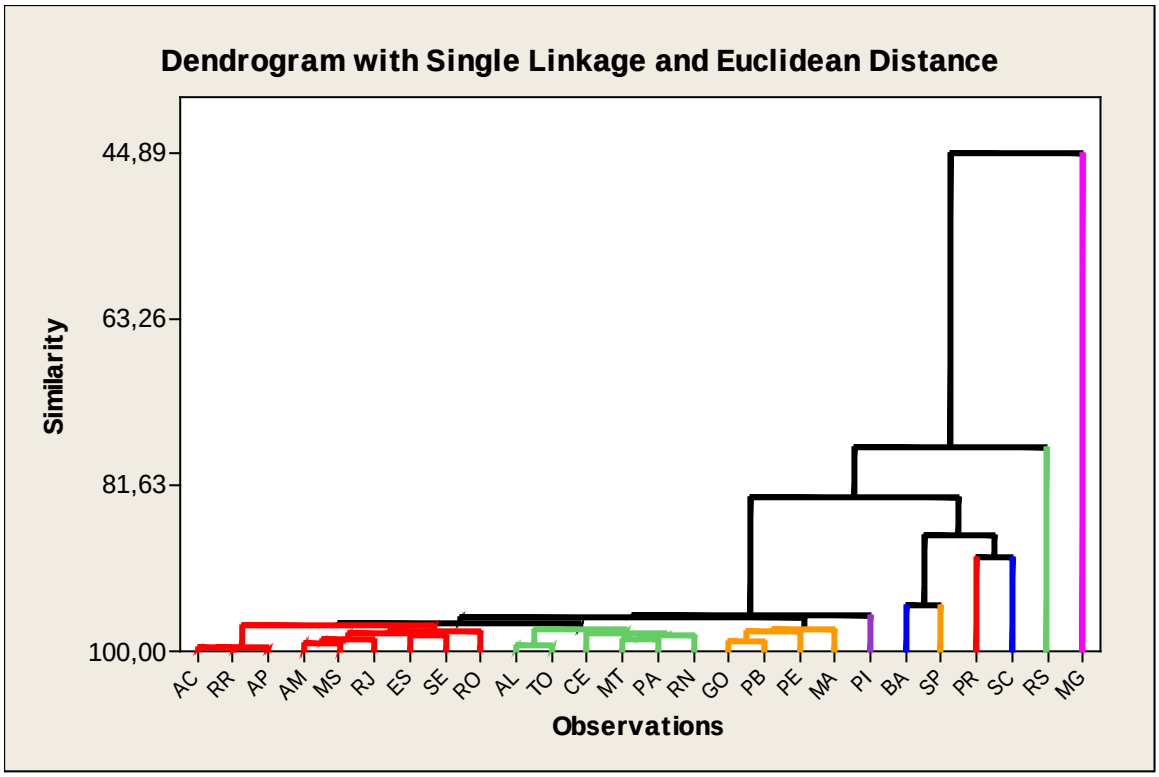


Figura 3. Dendrograma “Desiglalômetro” de Trabalho dos municípios por estado.

No gráfico acima, podemos verificar 3 grandes agrupamento de dados, compostos pelos estados do Brasil, elem de seis estados que ficaram isolados por não ter seus dados em similaridade com os outros estados.

Na classificação não supervisionada não se tem informações prévias sobre estes grupos. Não se tem informações sobre os por quês ou os critérios de agrupamento utilizados neste agrupamento.

Podemos observar estados com alto nível de similaridade o que significa que a desigualdade é baixa. O menor nível de desigualdade se encontra nos estados mais próximos do eixo X, por exemplo AC e RR, que tem um nível de similaridade próximo de 98%.

Quando o nível de desigualdade é baixo poderíamos erroneamente dizer que a situação é boa. Isso não é verdade. Baixa desigualdade não significa que as coisas vão bem, e sim que existe um padrão nos municípios do estado em termos de educação, uma maior similaridade entre este municípios, e não é possível responder se esta similaridade é boa ou não.

3.2.2. DENDOGRAMA DOS DADOS AGRUPADOS PELO RESULTADO DAS MÉDIAS

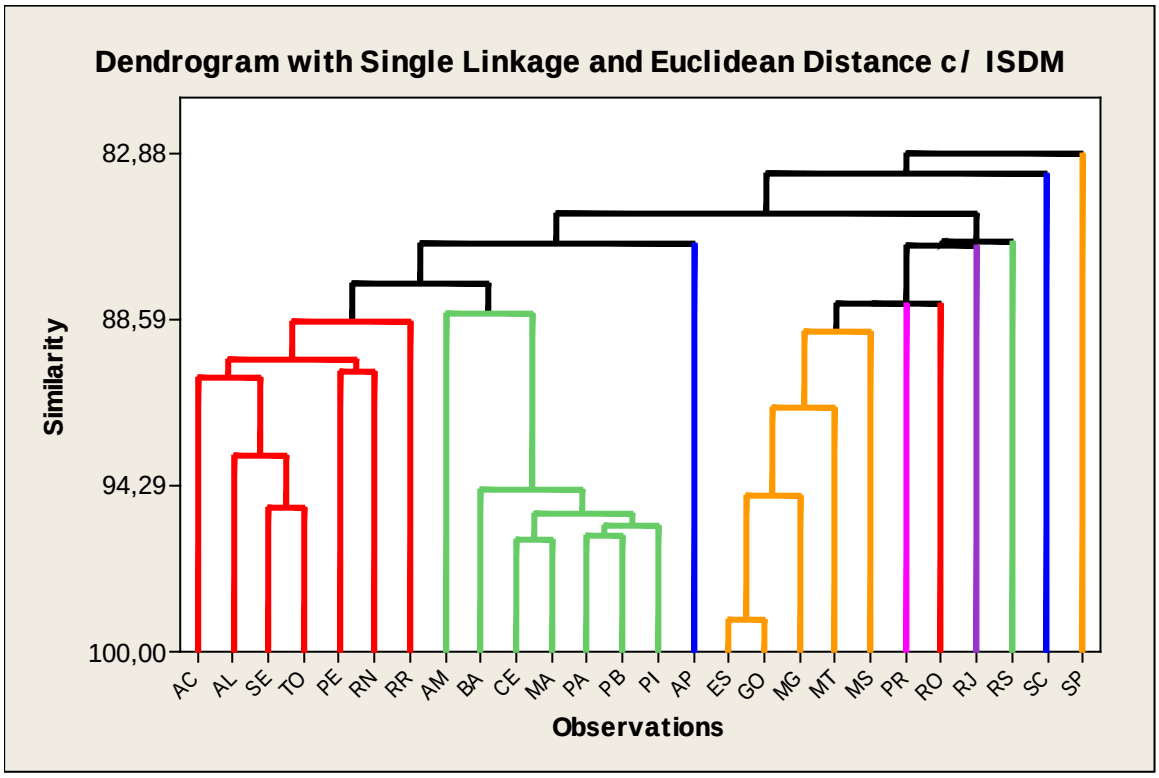


Figura 8. Dendograma dos dados das médias de ISDM e Trabalho por Estado

Podemos observar que existem três grandes grupos de similaridade e podemos considerar um quarto no canto direito com cores variadas. Estes estado tem baixo nível de similaridade com os outros mais para efeito de análise iremos agrupá-los para maior entendimento da situação da educção nos municípios do Brasil.

3.2.4. CONSIDERAÇÕES FINAIS

As análise comparativas dos dados nos permitem um resumo dos dados através de cálculos específicos como médias e desvios padrões, tornando a análise dos dados mais fácil e simples. Os gráficos de Dendograma são excelentes figuras visuais para podermos analisar e interpretar os diferentes comportamentos dos dados. No dendograma podemos analisar as similaridades dos dados e no Boxplot podemos ver as relações entre as médias e as variâncias dos agrupamentos analisados. Trata-se de ferramentas úteis para análise de grandes volumes de dados.

CAP II ANALISE DISCRIMINANTE

1. INTRODUÇÃO.

O presente trabalho tem por objetivo efetuar uma análise exploratória dos dados relacionados ao Indicador Social de Desenvolvimento dos Municípios (ISDM), mais profundamente de suas variáveis analíticas relacionadas ao 'trabalho'. Além disso, será feita uma análise do Índice FIRJAN de Desenvolvimento Municipal (IFDM). De tal forma, iniciaremos este trabalho com a definição e discussão de ambos indicadores, apresentando um breve histórico, o seu funcionamento metodológico e as suas variáveis. Na sequencia será realizada uma análise e interpretação dos dados que serão manipulados, utilizando as ferramentas do software estatístico **MINITAB.**

2. ENTENDENDO OS INDICADORES.

2.1. Indicador Social de Desenvolvimento dos Municípios – ISDM.

O Indicador Social de Desenvolvimento Municipal (ISDM) foi desenvolvido pelo Centro de Microeconomia Aplicada da Fundação Getúlio Vargas (FGV). O ISDM é um indicador sintético, que tem como objetivo reunir num único indicador vários aspectos referentes ao desenvolvimento social de um município, tornando possível a comparação do desempenho de todas as cidades brasileiras, de forma transversal ou longitudinal, medindo a performance dos entes federativos nas dimensões analisadas.

Em termos metodológicos, o indicador é calculado a partir de dados secundários, tendo como fontes o Instituto Brasileiro de Geografia e Estatística (IBGE), o Ministério da Saúde e o Ministério da Educação. Ele abrange cinco dimensões: Habitação, Renda, Trabalho, Saúde e Segurança e Educação. Essas dimensões e as variáveis que as compõem foram escolhidas de maneira a englobar algumas das questões mais prementes nas políticas públicas direcionadas para o município.

2.2. Índice Firjan de Desenvolvimento Municipal – IFDM.

O Índice FIRJAN de Desenvolvimento Municipal (IFDM) é um estudo anual do Sistema FIRJAN que acompanha o desenvolvimento de todos os municípios brasileiro. O índice é baseado em três pilares: Emprego & Renda, Educação e Saúde. Ele é feito, exclusivamente, com base em estatísticas públicas oficiais, disponibilizadas pelos ministérios do Trabalho, Educação e Saúde.

De leitura simples, o índice varia de 0 a 1. Quanto mais próximo de 1, maior o desenvolvimento da localidade. Além disso, sua metodologia possibilita determinar, com precisão, se a melhora relativa ocorrida em determinado município decorre da adoção de políticas específicas ou se o resultado obtido é apenas reflexo da queda dos demais municípios.

2.3. As variáveis.

São 6 (seis) as variáveis desta pesquisa, incluindo o nome dos indicadores. A seguir uma breve explanação através da tabela 1.

TABELA 1 - AS VARIÁVEIS

VARIÁVEL	SIGNIFICADO	TIPO	UNIDADE DE MEDIDA
ISDM	Indicador Social de Desenvolvimento dos Municípios (ISDM). Trata-se de uma média ponderada de diferentes indicadores analíticos: Habitação (H), Renda (R), Trabalho (T), Saúde & Segurança (S) e Educação (E), padronizada pela média do Brasil.	Variável Quantitativa. UF / Município.	Numérico
T	Indicador da dimensão Trabalho. Trata-se da média ponderada dos indicadores da dimensão Trabalho (T1_1, T1_2 e T2_1) padronizada pela média do Brasil.	Variável Quantitativa. UF / Município.	Numérico
T1_1	Taxa de ocupação. Percentual da população economicamente ativa (PEA) que esteja ocupada na semana de referência. Pessoas ocupadas podem ser empregados, empregadores, conta própria e não remunerados. Define-se como PEA a população entre 15 e 60 anos, que esteja ocupada (incluindo pessoas que estavam de férias) ou procurando emprego, exceto os deficientes físicos. Foram consideradas deficiências físicas a Tetraplegia (paralisia permanente total de ambos os braços e pernas), Paraplegia (paralisia permanente das pernas), Hemiplegia (paralisia permanente de um	Variável Quantitativa. UF / Município.	Numérico

	dos lados do corpo) ou Falta de membro ou de parte dele (falta de perna, braço, mão, pé ou do dedo polegar ou a falta de parte da perna ou braço).		
T1_2	Taxa de formalização entre os empregados. Percentual dos empregados ocupados na semana de referência no setor formal, dentre o total de empregados da PEA. Define-se como empregados ocupados no setor formal aqueles que possuíam carteira de trabalho assinada. Define-se como PEA a população entre 15 e 60 anos, que esteja ocupada (incluindo pessoas que estavam de férias) ou procurando emprego, exceto os deficientes físicos. Foram consideradas deficiências físicas a Tetraplegia (paralisia permanente total de ambos os braços e pernas), Paraplegia (paralisia permanente das pernas), Hemiplegia (paralisia permanente de um dos lados do corpo) ou Falta de membro ou de parte dele (falta de perna, braço, mão, pé ou do dedo polegar ou a falta de parte da perna ou braço).	Variável Quantitativa. UF / Município.	Numérico
T2_1	Taxa de trabalho infantil. Percentual das crianças de 10 a 14 anos que se encontram trabalhando ou procurando emprego na semana de referência em relação a população total residente dessa mesma faixa etária.	Variável Quantitativa. UF / Município.	Numérico
IFDM	Índice Firjan de Desenvolvimento Municipal – IFDM. O índice é resultado da média ponderada de diferentes indicadores: Emprego & Renda, Educação e Saúde.	Variável Quantitativa. UF / Município.	Numérico

2.4. A Tabela de Dados.

Os dados utilizados nesta pesquisa foram extraídos da planilha “ISDM por município 2000 e 2010 FGV Firjan IF GF IFDM”, tendo considerado os seguintes elementos:

UF2	Município	IFDM	ISDM	T	T1_1	T1_2	T2_1

3. ANÁLISE DAS VARIÁVEIS

3.1 VARIÁVEIS CATEGÓRICAS

Este tipo de variável indica que o foco de concentração deve ser a análise de gráficos do tipo *pie chart* e barras.

3.1.1 Variável: “Estado”

Fazem parte desta pesquisa os 27 estados brasileiros e suas cidades. O gráfico abaixo exibe o número de cidades por estado.

A variação no número de cidades por estado é acentuada. Considerando que o Distrito Federal é um estado brasileiro, é o estado com o menor número de cidades (1), enquanto o Mato Grosso possui mais de 852 cidades.

3.1.2 Variável: “REGIÃO”

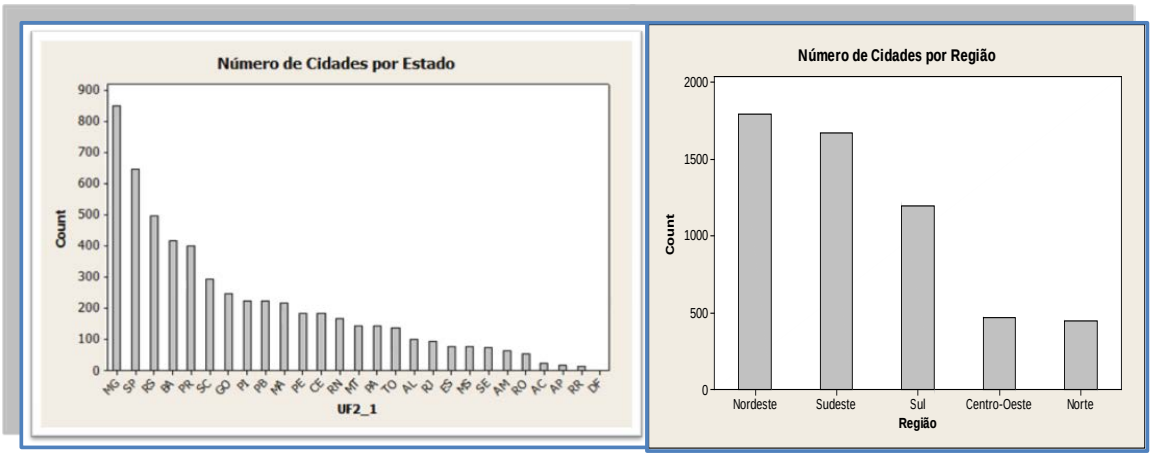


Figura 1. Número de Cidades por Estado e Região do Brasil

Podemos verificar no gráfico acima que a Região Nordeste é a que possui o maior número de cidades do Brasil (1790) e seguido pela Região Sudeste (1669). A Região que possui o menor número de cidades é a Norte, com 447 cidades, muito próxima da Região Centro-Oeste (468). A Região Sul possui 1191 cidades.

3.2 VARIÁVEIS QUANTITATIVAS

A análise deste tipo de variável permite a utilização de uma maior gama de ferramentas de análise como histogramas, curvas de densidade, gráfico de ramos, box-plot e dot-plot, além de informações numéricas como média, desvio-padrão, mediana, quartis, 5 números, intervalo de confiança e teste de normalidade de Anderson-Darling. Também podemos fazer classificações supervisionadas das variáveis quantitativas, através da análise discriminante.

3.2.1. ANÁLISE DISCRIMINANTE LINEAR DOS MUNICIPIOS POR REGIÃO

A análise discriminante é uma técnica da estatística multivariada utilizada para discriminar e classificar objetos, e estuda a separação de objetos de uma população em duas ou mais classes. Neste caso queremos discriminar os valores de educação dos municípios do Brasil, e utilizaremos inicialmente a variável categórica Região. Para geração de análise discriminante utilizaremos o comando do Minitab:

STAT >> MULTIVARIATE >> DISCRIMINANT ANALISYS

Discriminant Analysis: Região versus ISDM; T; T1_1; T1_2

```
Linear Method for Response: Região

Predictors: ISDM; T; T1_1; T1_2

Group  Centro-Oeste  Nordeste  Norte  Sudeste   Sul
Count           467      1790    447     1669  1191

Summary of classification

                                 True Group
Put into Group  Centro-Oeste  Nordeste  Norte  Sudeste    Sul
Centro-Oeste            225       153     84      219    148
Nordeste                 23      1015    126      116      2
Norte                    27       579    217       23      4
Sudeste                 118        42      7     1155    238
Sul                      74         1     13      156    799
Total N                 467      1790    447     1669   1191
N correct               225      1015    217     1155    799
Proportion            0,482     0,567  0,485    0,692  0,671

N = 5564          N Correct = 3411          Proportion Correct = 0,613
```

Figura 2. Resultado do comando STAT >> MULTIVARIATE >> DISCRIMINANT ANALISYS

A região que acertou mais é Sudeste (0,692) e a que errou mais é a Centro-Oeste (0,482). O gráfico exibe o cruzamento de dados entre as regiões. Por exemplo, a região Sudeste possui 1663 municípios e apenas 1155 correspondem a região, sendo que 238 são semelhantes aos dados da região Sul. O nome desta matriz é *confusion matrix* ou matriz de confusão. Podemos concluir que o agrupamento por região não é uma boa escolha segundo esta avaliação.

3.2.2. ANÁLISE DISCRIMINANTE LINEAR DOS MUNICIPIOS POR “3 BRASIS”

Esta segunda análise está interessada em verificar os possíveis agrupamento de dados utilizando a variável 3 Brasis, calculada no exercício anterior, e demonstra os agrupamentos do Brasil segundo sua proximidade de dados de educação.

Discriminant Analysis: 3 Brasis versus ISDM; T; T1_1; T1_2

```
Linear Method for Response: 3 Brasis


Predictors: ISDM; T; T1_1; T1_2


Group  Centro-Oeste   Nor    Su
Count           467  2237  2860


Summary of classification

                        True Group
Put into Group  Centro-Oeste    Nor     Su
Centro-Oeste            282    325    638
Nor                      42   1898    117
Su                      143     14   2105
Total N                 467   2237   2860
N correct               282   1898   2105
Proportion            0,604  0,848  0,736

N = 5564          N Correct = 4285          Proportion Correct = 0,770
```

A % de acertos melhorou de 61%(5 Regiões) a 77% (3 Brasis).

Existem duas possibilidades análise discriminante que são a linear e a quadrática. Dependendo da variável deve-se dar mais peso e mais atenção a um método que outro. Neste caso utilizamos a linear. Podemos observar que alguns estados e municípios da região centro-oeste tem características das regiões Sul, visto pelo número 638 municípios foram encontrados na intersecção entre sul e centro-oeste.

3.2.3. ANÁLISE DISCRIMINANTE QUADRÁTICA DOS MUNICPIOS POR “3 BRASIS”

Uma boa classificação deve resultar em pequenos erros, isto é, deve haver pouca probabilidade de má classificação, e para que isso ocorra a regra de classificação deve considerar as probabilidades a priori e os custos de má classificação. Outro fator que uma regra de classificação deve considerar é se as variâncias das populações são iguais ou não. Quando a regra de classificação assume que as variâncias das populações são iguais, as funções discriminantes são ditas lineares e quando não são funções discriminantes quadráticas. Vamos agora verificar a função quadrática para 3 Brasis.

Discriminant Analysis: 3 Brasis versus ISDM; T; T1_1; T1_2

Quadratic Method for Response: 3 Brasis

Predictors: ISDM; T; T1_1; T1_2

Group	Centro-Oeste	Nor	Su
Count	467	2237	2860

Summary of classification

	True Group		
Put into Group	Centro-Oeste	Nor	Su
Centro-Oeste	329	276	726
Nor	50	1924	173
Su	88	37	1961
Total N	467	2237	2860
N correct	329	1924	1961
Proportion	0,704	0,860	0,686

N = 5564 N Correct = 4214 **Proportion Correct = 0,757**

No modelo quadrático a proporção foi alterada em apenas 2% (de 0,77 para 0,75). Seguindo o pensamento da simplicidade, vamos escolher o método linear pois é o mais simples e mais precisso.

Em ciência, parcimônia é a preferência pela explicação mais simples para uma observação. Esta geralmente é considerada a melhor maneira de julgar as hipóteses. Parcimônia também é um conceito utilizado na sistemática moderna que estabelece que ao construir e selecionar árvores filogenéticas, ou seja, os dados, o melhor critério é baseado em seus princípios: normalmente é correto o relacionamento mais simples encontrado entre dois indivíduos, aquele que

apresente o menor número de passos intermediários ou mudanças evolucionárias. Portanto a diferença entre o método linear e o quadrático é pequena e não justifica a utilização do método quadrático.

3.2.4. ANÁLISE DISCRIMINANTE LINEAR DOS MUNICIPIOS PARA DADOS AGRUPADOS em 5 BRASIS

Neste exemplo abaixo vamos através do dendograma pesquisar o grau de similaridade das variáveis de desvio padrão do trabalho nos municípios do Brasil. Com base na similaridade poderemos definimos agrupamento de dados e após utilizamos a análise discriminante para verificar a proporção correta dos agrupamentos.

Discriminant Analysis: 5 Brasis versus ISDM; T; T1_1; T1_2

```
Linear Method for Response: 5 Brasis


Predictors: ISDM; T; T1_1; T1_2


Group        B1        B2        B3        B4        B5
Count       703      1466      1397       467      1531


Summary of classification

                              True Group
Put into Group     B1     B2     B3     B4     B5
B1                341    367    131     23      5
B2                272    993     72      5      3
B3                 70     90    694     94    262
B4                 17     14    181    175    193
B5                  3      2    319    170   1068
Total N           703   1466   1397    467   1531
N correct         341    993    694    175   1068
Proportion      0,485  0,677  0,497  0,375  0,698

N = 5564           N Correct = 3271          Proportion Correct = 0,588
```

Neste caso a proporção correta um pouco melhor, ou seja, os agrupamentos gerados anteriormente pelo agrupamento em 5 Brasis gerou a mesma proporção do método linear utilizado na análise discriminante.

4. CONSIDERAÇÕES FINAIS

A tarefa da análise discriminante é encontrar a melhor função discriminante linear de um conjunto de variáveis que reproduza, tanto quanto possível, um agrupamento a priori de casos considerados.

Um procedimento em passos é utilizado nesse programa, e em cada passo a variável mais poderosa é introduzida na função discriminante. A função critério para selecionar a próxima variável depende do número de grupos especificados.

Quando o número de variáveis é maior do que dois, então o critério de seleção de variáveis é o traço do produto da matriz de covariância para as variáveis envolvidas e a matriz de covariância interclasse em um passo particular.

Os cálculos podem ser realizados em toda a população ou em amostra de dados ou mesmo em dados previamente agrupados.

Focando nas variáveis referentes a trabalho, utilizamos a análise discriminante linear e conseguimos um resultado de 77% de proporção correta quando consideramos 3 Brasis e 61% para 5 Brasis..

CAP III REGRESSÃO LOGÍSTICA

1. INTRODUÇÃO.

O presente trabalho tem por objetivo efetuar uma análise exploratória dos dados relacionados ao Indicador Social de Desenvolvimento dos Municípios (ISDM), mais profundamente de suas variáveis analíticas relacionadas ao 'trabalho'. Além disso, será feita uma análise do Índice FIRJAN de Desenvolvimento Municipal (IFDM). De tal forma, iniciaremos este trabalho com a definição e discussão de ambos indicadores, apresentando um breve histórico, o seu funcionamento metodológico e as suas variáveis. Na sequencia será realizada uma análise e interpretação dos dados que serão manipulados, utilizando as ferramentas do software estatístico **MINITAB.**

A ferramenta utilizada neste trabalho será a 'regressão logística'. A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias1 2. A regressão logística é amplamente usada em ciências médicas e sociais, e tem outras denominações, como modelo logístico e classificador de máxima entropia.

2. ENTENDENDO OS INDICADORES.

2.1. Indicador Social de Desenvolvimento dos Municípios – ISDM.

O Indicador Social de Desenvolvimento Municipal (ISDM) foi desenvolvido pelo Centro de Microeconomia Aplicada da Fundação Getúlio Vargas (FGV). O ISDM é um indicador sintético, que tem como objetivo reunir num único indicador vários aspectos referentes ao desenvolvimento social de um município, tornando possível a comparação do desempenho de todas as cidades brasileiras, de forma transversal ou longitudinal, medindo a performance dos entes federativos nas dimensões analisadas.

Em termos metodológicos, o indicador é calculado a partir de dados secundários, tendo como fontes o Instituto Brasileiro de Geografia e Estatística

(IBGE), o Ministério da Saúde e o Ministério da Educação. Ele abrange cinco dimensões: Habitação, Renda, Trabalho, Saúde e Segurança e Educação. Essas dimensões e as variáveis que as compõem foram escolhidas de maneira a englobar algumas das questões mais prementes nas políticas públicas direcionadas para o município.

2.2. Índice Firjan de Desenvolvimento Municipal – IFDM.

O Índice FIRJAN de Desenvolvimento Municipal (IFDM) é um estudo anual do Sistema FIRJAN que acompanha o desenvolvimento de todos os municípios brasileiro. O índice é baseado em três pilares: Emprego & Renda, Educação e Saúde. Ele é feito, exclusivamente, com base em estatísticas públicas oficiais, disponibilizadas pelos ministérios do Trabalho, Educação e Saúde.

De leitura simples, o índice varia de 0 a 1. Quanto mais próximo de 1, maior o desenvolvimento da localidade. Além disso, sua metodologia possibilita determinar, com precisão, se a melhora relativa ocorrida em determinado município decorre da adoção de políticas específicas ou se o resultado obtido é apenas reflexo da queda dos demais municípios.

2.3. As variáveis.

São 6 (seis) as variáveis desta pesquisa, incluindo o nome dos indicadores. A seguir uma breve explanação através da tabela 1.

TABELA 1 - AS VARIÁVEIS

VARIÁVEL	SIGNIFICADO	TIPO	UNIDADE DE MEDIDA
ISDM	Indicador Social de Desenvolvimento dos Municípios (ISDM). Trata-se de uma média ponderada de diferentes indicadores analíticos: Habitação (H), Renda (R), Trabalho (T), Saúde & Segurança (S) e Educação (E), padronizada pela média do Brasil.	Variável Quantitativa. UF / Município.	Numérico
T	Indicador da dimensão Trabalho. Trata-se da média ponderada dos indicadores da dimensão Trabalho (T1_1, T1_2 e T2_1) padronizada pela média do Brasil.	Variável Quantitativa. UF / Município.	Numérico
T1_1	Taxa de ocupação. Percentual da população economicamente ativa (PEA) que esteja ocupada na	Variável Quantitativ	Numérico

	semana de referência. Pessoas ocupadas podem ser empregados, empregadores, conta própria e não remunerados. Define-se como PEA a população entre 15 e 60 anos, que esteja ocupada (incluindo pessoas que estavam de férias) ou procurando emprego, exceto os deficientes físicos. Foram consideradas deficiências físicas a Tetraplegia (paralisia permanente total de ambos os braços e pernas), Paraplegia (paralisia permanente das pernas), Hemiplegia (paralisia permanente de um dos lados do corpo) ou Falta de membro ou de parte dele (falta de perna, braço, mão, pé ou do dedo polegar ou a falta de parte da perna ou braço).	a. UF / Município.	
T1_2	Taxa de formalização entre os empregados. Percentual dos empregados ocupados na semana de referência no setor formal, dentre o total de empregados da PEA. Define-se como empregados ocupados no setor formal aqueles que possuíam carteira de trabalho assinada. Define-se como PEA a população entre 15 e 60 anos, que esteja ocupada (incluindo pessoas que estavam de férias) ou procurando emprego, exceto os deficientes físicos. Foram consideradas deficiências físicas a Tetraplegia (paralisia permanente total de ambos os braços e pernas), Paraplegia (paralisia permanente das pernas), Hemiplegia (paralisia permanente de um dos lados do corpo) ou Falta de membro ou de parte dele (falta de perna, braço, mão, pé ou do dedo polegar ou a falta de parte da perna ou braço).	Variável Quantitativa. UF / Município.	Numérico
T2_1	Taxa de trabalho infantil. Percentual das crianças de 10 a 14 anos que se encontram trabalhando ou procurando emprego na semana de referência em relação a população total residente dessa mesma faixa etária.	Variável Quantitativa. UF / Município.	Numérico
IFDM	Índice Firjan de Desenvolvimento Municipal – IFDM. O índice é resultado da média ponderada de diferentes indicadores: Emprego & Renda, Educação e Saúde.	Variável Quantitativa. UF / Município.	Numérico

2.4. A Tabela de Dados.

Os dados utilizados nesta pesquisa foram extraídos da planilha “ISDM por município 2000 e 2010 FGV Firjan IF GF IFDM”, tendo considerado os seguintes elementos:

UF2	Município	IFDM	ISDM	T	T1_1	T1_2	T2_1

3. ANÁLISE DAS VARIÁVEIS

3.1 VARIÁVEIS CATEGÓRICAS

Este tipo de variável indica que o foco de concentração deve ser a análise de gráficos do tipo *pie chart* e barras.

3.1.1 Variável: “Estado”

Fazem parte desta pesquisa os 27 estados brasileiros e suas cidades. O gráfico abaixo exibe o número de cidades por estado.

A variação no número de cidades por estado é acentuada. Considerando que o Distrito Federal é um estado brasileiro, é o estado com o menor número de cidades (1), enquanto o Mato Grosso possui mais de 852 cidades.

3.1.2 Variável: “REGIÃO”

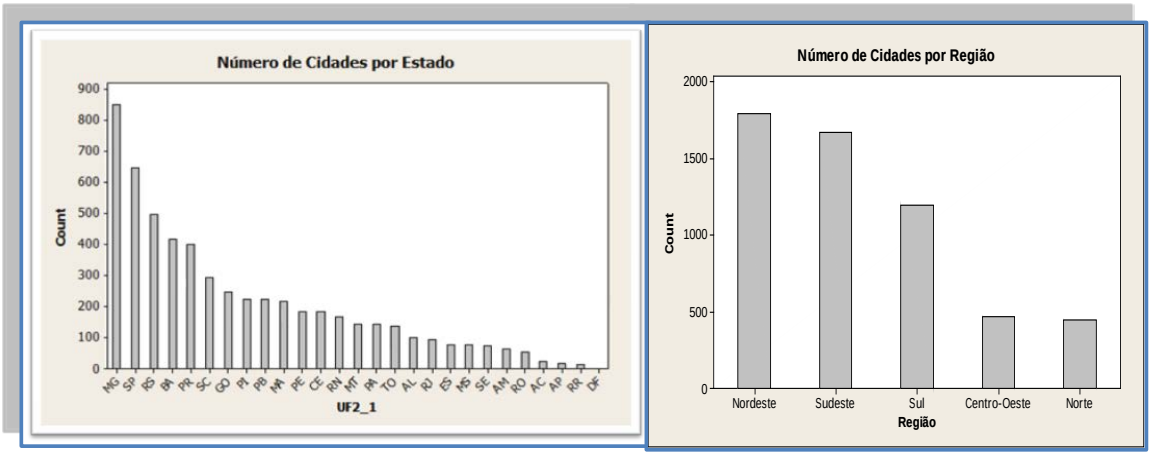


Figura 1. Número de Cidades por Estado e Região do Brasil

Podemos verificar no gráfico acima que a Região Nordeste é a que possui o maior número de cidades do Brasil (1790) e seguido pela Região Sudeste (1669). A Região que possui o menor número de cidades é a Norte, com 447 cidades, muito próxima da Região Centro-Oeste (468). A Região Sul possui 1191 cidades.

3.2 VARIÁVEIS QUANTITATIVAS

A análise deste tipo de variável permite a utilização de uma maior gama de ferramentas de análise como histogramas, curvas de densidade, gráfico de ramos, box-plot e dot-plot, além de informações numéricas como média, desvio-padrão, mediana, quartis, 5 números, intervalo de confiança e teste de normalidade de Anderson-Darling. Também podemos fazer classificações supervisionadas das variáveis quantitativas, através da análise discriminante.

Trata-se de um modelo de regressão para variáveis dependentes ou de resposta binomialmente distribuídas. É útil para modelar a probabilidade de um evento ocorrer como função de outros factores. Os dados são originários da pesquisa da FGV / FIRJAM sobre o desenvolvimento dos municípios do Brasil. Neste trabalho abordaremos as variáveis referentes à educação dos municípios.

3.2.1. REGRESSÃO LOGÍSTICA

Stat >> Regression >> Ordinal Logistical Regression

Ordinal Logistic Regression: Região versus ISDM; T; T1_1; T1_2; T2_1

```
Link Function: Logit


Response Information

Variable  Value         Count
Região    Centro-Oeste    467
          Nordeste       1790
          Norte           447
          Sudeste        1669
          Sul            1191
          Total          5564


Logistic Regression Table

                                                      Odds         95% CI
Predictor        Coef    SE Coef       Z      P  Ratio  Lower        Upper
Const(1)      15,2236    9,63302    1,58  0,114
Const(2)      17,7410    9,63365    1,84  0,066
Const(3)      18,3331    9,63387    1,90  0,057
Const(4)      20,5789    9,63476    2,14  0,033
ISDM        -0,716523  0,0434651  -16,49  0,000   0,49   0,45         0,53
T             1,00124    8,82646    0,11  0,910   2,72   0,00  88729842,90
T1_1        -0,162923   0,366727   -0,44  0,657   0,85   0,41         1,74
T1_2        -0,0799999   0,366608   -0,22  0,827   0,92   0,45         1,89
T2_1        -0,0041186   0,733213   -0,01  0,996   1,00   0,24         4,19
```

```
Log-Likelihood = -6612,888
Test that all slopes are zero: G = 3093,924, DF = 5, P-Value = 0,000


Goodness-of-Fit Tests

Method    Chi-Square     DF      P
Pearson      39012,0  22247  0,000
Deviance     13225,8  22247  1,000


Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs         Number  Percent  Summary Measures
Concordant   9185141     79,4  Somers' D              0,59
Discordant   2313175     20,0  Goodman-Kruskal Gamma  0,60
Ties           67712      0,6  Kendall's Tau-a        0,44
Total       11566028    100,0
```

4. CONSIDERAÇÕES FINAIS

Enquanto método de predição para variáveis categóricas, a regressão logística é comparável às técnicas supervisionadas propostas em aprendizagem automática (árvores de decisão, redes neuronais, etc.), ou ainda a análise discriminante preditiva em estatística exploratória. É possível de as colocar em concorrência para escolha do modelo mais adaptado para um certo problema preditivo a resolver.

A Regressão Logística aumentou os acertos a 79% para 5 Brasis; no entanto nenhum dos coeficientes, salvo o referente a ISDM é confiável !

CAP IV ÁRVORES DE CLASSIFICAÇÃO

1. INTRODUÇÃO.

O presente trabalho tem por objetivo efetuar uma análise exploratória dos dados relacionados ao Indicador Social de Desenvolvimento dos Municípios (ISDM), mais profundamente de suas variáveis analíticas relacionadas ao 'trabalho'. Além disso, será feita uma análise do Índice FIRJAN de Desenvolvimento Municipal (IFDM). De tal forma, iniciaremos este trabalho com a definição e discussão de ambos indicadores, apresentando um breve histórico, o seu funcionamento metodológico e as suas variáveis. Na sequencia será realizada uma análise e interpretação dos dados que serão manipulados, utilizando as ferramentas dos softwares estatísticos **MINITB e SPSS.**

A ferramenta utilizada neste trabalho será a 'análise de correspondência'. Análise de correspondência é uma técnica de análise exploratória de dados adequada para analisar tabelas de duas entradas ou tabelas de múltiplas entradas, levando em conta algumas medidas de correspondência entre linhas e colunas. Consiste na conversão de uma matriz de dados não negativos em um tipo particular de representação gráfica em que as linhas e colunas da matriz são simultaneamente representadas em dimensão reduzida, isto é, por pontos no gráfico. Este método permite estudar as relações e semelhanças existentes entre as categorias de linhas e entre as categorias de colunas de uma tabela de contingência ou o conjunto de categorias de linhas e o conjunto categorias de colunas.

2. ENTENDENDO OS INDICADORES.

2.1. Indicador Social de Desenvolvimento dos Municípios – ISDM.

O Indicador Social de Desenvolvimento Municipal (ISDM) foi desenvolvido pelo Centro de Microeconomia Aplicada da Fundação Getúlio Vargas (FGV). O ISDM é um indicador sintético, que tem como objetivo reunir num único indicador vários aspectos referentes ao desenvolvimento social de um município, tornando possível a comparação do desempenho de todas as cidades brasileiras, de forma

transversal ou longitudinal, medindo a performance dos entes federativos nas dimensões analisadas.

Em termos metodológicos, o indicador é calculado a partir de dados secundários, tendo como fontes o Instituto Brasileiro de Geografia e Estatística (IBGE), o Ministério da Saúde e o Ministério da Educação. Ele abrange cinco dimensões: Habitação, Renda, Trabalho, Saúde e Segurança e Educação. Essas dimensões e as variáveis que as compõem foram escolhidas de maneira a englobar algumas das questões mais prementes nas políticas públicas direcionadas para o município.

2.2. Índice Firjan de Desenvolvimento Municipal – IFDM.

O Índice FIRJAN de Desenvolvimento Municipal (IFDM) é um estudo anual do Sistema FIRJAN que acompanha o desenvolvimento de todos os municípios brasileiro. O índice é baseado em três pilares: Emprego & Renda, Educação e Saúde. Ele é feito, exclusivamente, com base em estatísticas públicas oficiais, disponibilizadas pelos ministérios do Trabalho, Educação e Saúde.

De leitura simples, o índice varia de 0 a 1. Quanto mais próximo de 1, maior o desenvolvimento da localidade. Além disso, sua metodologia possibilita determinar, com precisão, se a melhora relativa ocorrida em determinado município decorre da adoção de políticas específicas ou se o resultado obtido é apenas reflexo da queda dos demais municípios.

2.3. As variáveis.

São 6 (seis) as variáveis desta pesquisa, incluindo o nome dos indicadores. A seguir uma breve explanação através da tabela 1.

TABELA 1 - AS VARIÁVEIS

VARIÁVEL	SIGNIFICADO	TIPO	UNIDADE DE MEDIDA
ISDM	Indicador Social de Desenvolvimento dos Municípios (ISDM). Trata-se de uma média ponderada de diferentes indicadores analíticos: Habitação (H), Renda (R), Trabalho (T), Saúde & Segurança (S) e Educação (E), padronizada pela média do Brasil.	Variável Quantitativa. UF / Município.	Numérico
T	Indicador da dimensão Trabalho. Trata-se da média ponderada dos indicadores da dimensão Trabalho	Variável Quantitativ	

	(T1_1, T1_2 e T2_1) padronizada pela média do Brasil.	a. UF / Município.	Numérico
T1_1	Taxa de ocupação. Percentual da população economicamente ativa (PEA) que esteja ocupada na semana de referência. Pessoas ocupadas podem ser empregados, empregadores, conta própria e não remunerados. Define-se como PEA a população entre 15 e 60 anos, que esteja ocupada (incluindo pessoas que estavam de férias) ou procurando emprego, exceto os deficientes físicos. Foram consideradas deficiências físicas a Tetraplegia (paralisia permanente total de ambos os braços e pernas), Paraplegia (paralisia permanente das pernas), Hemiplegia (paralisia permanente de um dos lados do corpo) ou Falta de membro ou de parte dele (falta de perna, braço, mão, pé ou do dedo polegar ou a falta de parte da perna ou braço).	Variável Quantitativa. UF / Município.	Numérico
T1_2	Taxa de formalização entre os empregados. Percentual dos empregados ocupados na semana de referência no setor formal, dentre o total de empregados da PEA. Define-se como empregados ocupados no setor formal aqueles que possuíam carteira de trabalho assinada. Define-se como PEA a população entre 15 e 60 anos, que esteja ocupada (incluindo pessoas que estavam de férias) ou procurando emprego, exceto os deficientes físicos. Foram consideradas deficiências físicas a Tetraplegia (paralisia permanente total de ambos os braços e pernas), Paraplegia (paralisia permanente das pernas), Hemiplegia (paralisia permanente de um dos lados do corpo) ou Falta de membro ou de parte dele (falta de perna, braço, mão, pé ou do dedo polegar ou a falta de parte da perna ou braço).	Variável Quantitativa. UF / Município.	Numérico
T2_1	Taxa de trabalho infantil. Percentual das crianças de 10 a 14 anos que se encontram trabalhando ou procurando emprego na semana de referência em relação a população total residente dessa mesma faixa etária.	Variável Quantitativa. UF / Município.	Numérico
IFDM	Índice Firjan de Desenvolvimento Municipal – IFDM. O índice é resultado da média ponderada de diferentes indicadores: Emprego & Renda, Educação e Saúde.	Variável Quantitativa. UF / Município.	Numérico

2.4. A Tabela de Dados.

Os dados utilizados nesta pesquisa foram extraídos da planilha “ISDM por município 2000 e 2010 FGV Firjan IF GF IFDM”, tendo considerado os seguintes elementos:

UF2	Município	IFDM	ISDM	T	T1_1	T1_2	T2_1

3. ANÁLISE DAS VARIÁVEIS

3.1 VARIÁVEIS CATEGÓRICAS

Este tipo de variável indica que o foco de concentração deve ser a análise de gráficos do tipo *pie chart* e barras.

3.1.1 Variável: “Estado”

Fazem parte desta pesquisa os 27 estados brasileiros e suas cidades. O gráfico abaixo exibe o número de cidades por estado.

A variação no número de cidades por estado é acentuada. Considerando que o Distrito Federal é um estado brasileiro, é o estado com o menor número de cidades (1), enquanto o Mato Grosso possui mais de 852 cidades.

3.1.2 Variável: “REGIÃO”

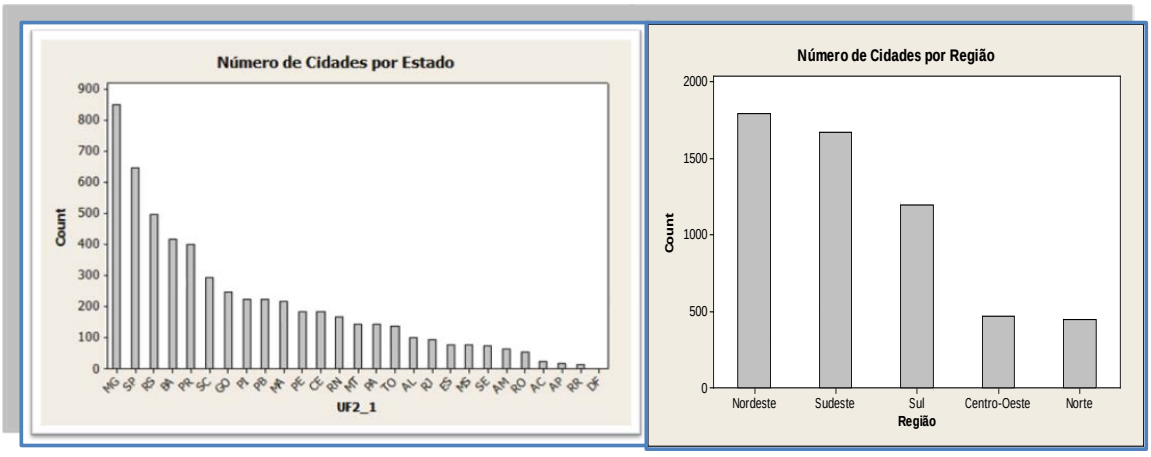


Figura 1. Número de Cidades por Estado e Região do Brasil

Podemos verificar no gráfico acima que a Região Nordeste é a que possui o maior número de cidades do Brasil (1790) e seguido pela Região Sudeste (1669). A Região que possui o menor número de cidades é a Norte, com 447 cidades, muito próxima da Região Centro-Oeste (468). A Região Sul possui 1191 cidades.

3.2 VARIÁVEIS QUANTITATIVAS

A análise deste tipo de variável permite a utilização de uma maior gama de ferramentas de análise como histogramas, curvas de densidade, gráfico de ramos, box-plot e dot-plot, além de informações numéricas como média, desvio-padrão, mediana, quartis, 5 números, intervalo de confiança e teste de normalidade de Anderson-Darling. Também podemos fazer classificações supervisionadas das variáveis quantitativas, através da análise discriminante.

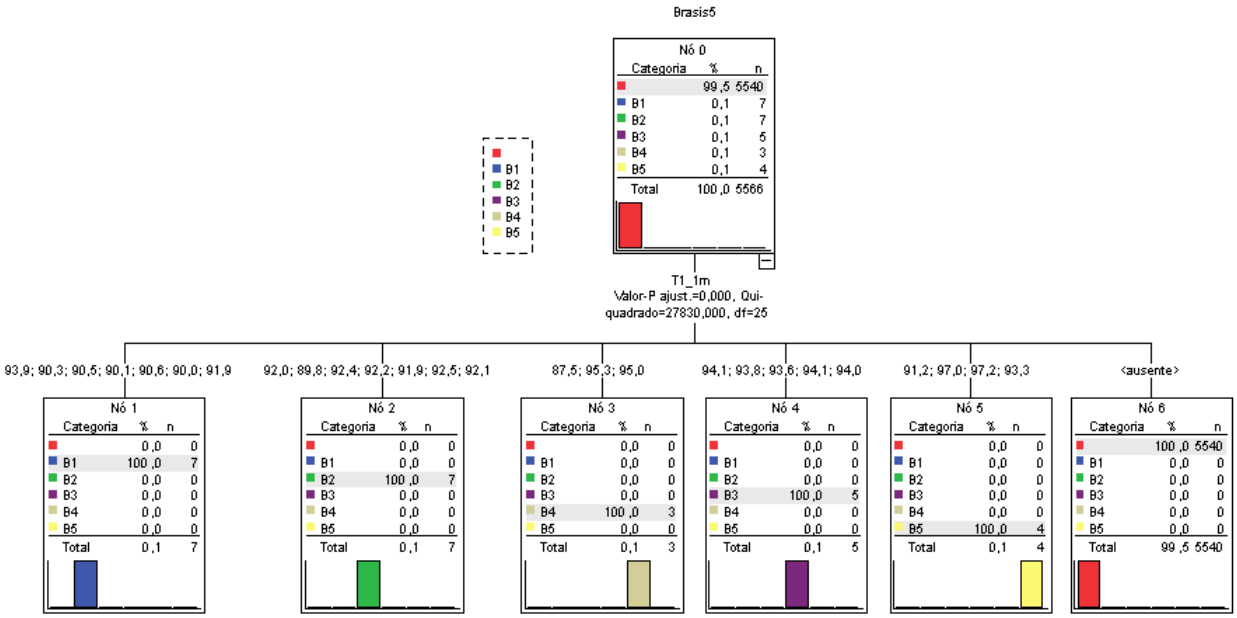
ÁRVORES DE CLASSIFICAÇÃO.

Arvore agrupada por região (5 brasis) com as médias de desenvolvimento por estado focando a variável trabalho.

Árvore classificatória

Resumo do modelo

Especificações	Método de crescimento	CHAID
	Variável dependente	Brasis5
	Variáveis independentes	TM, T1_1m, T1_2m, T2_1m, ISDMm, E&R-M, LIQ-M, H6-M, R1-M, S1_1-M, E2_4-M
	Validação	Nenhum
	Profundidade de árvore máxima	3
	Casos mínimos em nó pai	2
	Casos mínimos em nó filho	1
	Variáveis independentes incluídas	T1_1m
Resultados	Número de nós	7
	Número de nós de terminal	6
	Profundidade	1



Risco

Estimativas	Modelo padrão
,000	,000

Método de crescimento: CHAID

Variável dependente: Brasis5

Posto

Observado	Previsto						
		B1	B2	B3	B4	B5	Porcentagem Correta
	5540	0	0	0	0	0	100,0%
B1	0	7	0	0	0	0	100,0%
B2	0	0	7	0	0	0	100,0%
B3	0	0	0	5	0	0	100,0%
B4	0	0	0	0	3	0	100,0%
B5	0	0	0	0	0	4	100,0%
Porcentagem global	99,5%	0,1%	0,1%	0,1%	0,1%	0,1%	100,0%

Método de crescimento: CHAID

Variável dependente: Brasis5

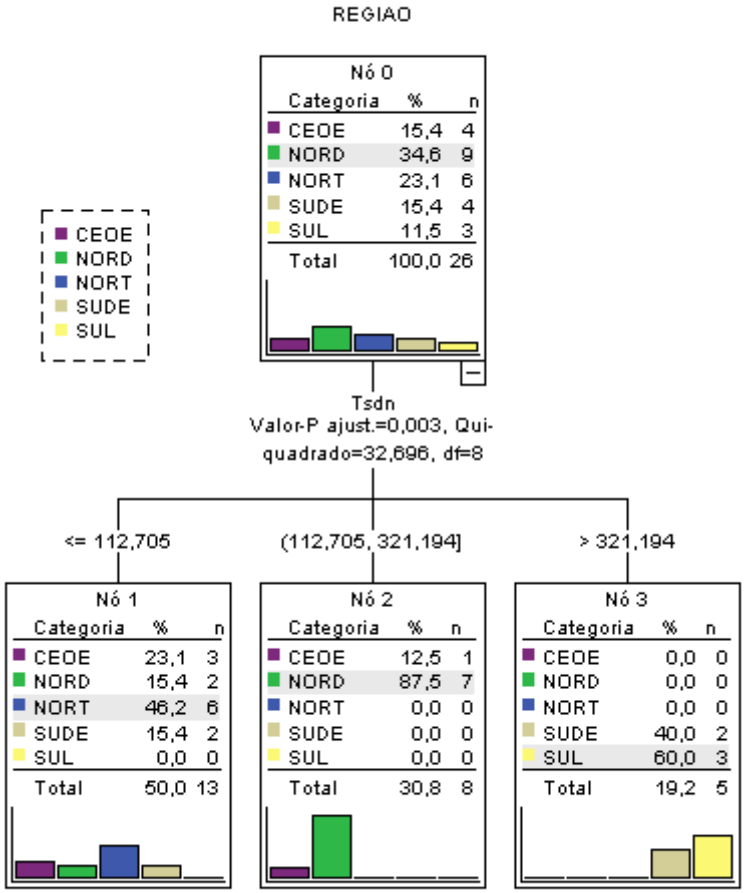
100% de acertos sendo que a variável que entra é da taxa de ocupação !

Arvore agrupada por região (5 brasis) considerando os desvios padrões (disparidades) no desenvolvimento por estado e focando a variável trabalho.

A árvore de classificação indica qual a variável que melhor separa os grupos e classifica as variáveis por ordem de importância na separação dos grupos. A seguir é demonstrado o teste desse modelo.

Resumo do modelo

Especificações	Método de crescimento	CHAID
	Variável dependente	REGIAO
	Variáveis independentes	ISDMsdn, Tsdn, T11sdn, T12sdn, T21sdn
	Validação	Nenhum
	Profundidade de árvore máxima	3
	Casos mínimos em nó pai	2
	Casos mínimos em nó filho	1
	Variáveis independentes incluídas	Tsdn
Resultados	Número de nós	4
	Número de nós de terminal	3
	Profundidade	1



Risco

Estimativas	Modelo padrão
,385	,095

Método de crescimento: CHAID

Variável dependente: REGIAO

Posto

Observado	Previsto					
	CEOE	NORD	NORT	SUDE	SUL	Porcentagem Correta
CEOE	0	1	3	0	0	0,0%
NORD	0	7	2	0	0	77,8%
NORT	0	0	6	0	0	100,0%
SUDE	0	0	2	0	2	0,0%
SUL	0	0	0	0	3	100,0%
Porcentagem global	0,0%	30,8%	50,0%	0,0%	19,2%	61,5%

Método de crescimento: CHAID

Variável dependente: REGIAO

4. CONSIDERAÇÕES FINAIS

A análise de correspondência pode ser considerada como um caso especial da análise de componentes principais, porém dirigida a dados categóricos organizados em tabelas de contingência e não a dados contínuos. O problema é análogo a encontrar o maior componente principal de um conjunto de I observações e J variáveis, com modificações devido à ponderação das observações e à métrica ponderada.

Já a árvore de decisão representa um instrumento de decisão sob a forma de uma árvore, porém podem haver outras aplicações. Tem a mesma utilidade da tabela de decisão. Trata-se de uma maneira alternativa de expressar as mesmas regras que são obtidas quando se constrói a tabela. **Tomando em consideração as medias por estado a variável ocupação permite classificar totalmente(100 %) as cinco regiões do Brasil; entanto que considerando agora níveis de disparidades (sd) dos 5 Brasis a % de acertos da arvore foi similar aquela da Análise Discriminante 61%.**