Classificação de dados geográficos e representação cartográfica: discussões metodológicas

Paulo Fernando Braga Carvalho
PUC Minas

Resumo

A classificação de dados é um passo importante na elaboração de mapas coropléticos, pois cores serão atribuídas às unidades espaciais básicas de acordo com as classes definidas. Outra representação cartográfica para dados quantitativos é feita com o uso de símbolos proporcionais, sendo apropriada para destacar a dimensão da unidade espacial básica em relação ao fenômeno em estudo. Em algum momento o responsável pela representação cartográfica precisa organizar os dados e, para isso, é preciso definir o número de classes a adotar, qual o esquema de cores ou símbolos mais apropriados e, dentre vários critérios de classificação, qual o mais adequado para o fenômeno em estudo. Neste trabalho são apresentadas contribuições para melhor compreensão e uso com embasamento metodológico dos critérios mais comumente adotados na classificação de dados geográficos para produção de gráficos, mapas coropléticos, mapas de símbolos proporcionais ou, ainda, explorar a relação existente entre eles.

Palavras-Chave: Dado Gegráfico; Classificação de Dados; Representação Cartográfica

Abstract

The classification of data is an important step towards the elaboration of choropleth maps because the use of colors will be attributed to the basic spatial units in accordance with the classes defined. Another representation for quantitative data analysis is made through the usage of proportional symbols, which are appropriate to the data related to the dimension of the Basic Spatial Unit in relation to the phenomenon under consideration. At a given moment the responsible for the cartographic representation, which will be led to organize the presented data, and for that to happen, it is necessary to define the number of classes to be adopted, which colors or symbols scheme is the most appropriate and, among several classification criteria, which is the best appropriate for the phenomenon under study. In this work, contributions for better understanding and use with methodological basis of the criteria more commonly adopted in the classification and visual representation of geographic data to produce graphs, choropleth maps, maps of proportional symbols or, in addition, to explore the existing relationship between them.

Keywords: Geographical Data; Data Classification; Cartographic Representation

paulofernando@pucminas.br

Considerações iniciais

Estudos geográficos que envolvem análise espacial vêm aplicando grande volume de dados de diferentes tipos e ordens de grandezas. Esses dados precisam ser organizados, tratados e apresentados de forma a mostrar a informação, inicialmente oculta ou difusa, e destacar a existência de algum padrão espacial no comportamento da distribuição do fenômeno ou mesmo identificar unidades espaciais que fogem de algum padrão espacial de concentração ou dispersão.

Com a evolução tecnológica dos Sistemas de Informação Geográfica, os mapas coropléticos vêm se tornando cada vez mais populares na representação de dados geográficos, tendo em vista sua aparente simplicidade e são usados para os mais diversos fins: acadêmico, gestão pública ou privada, relatórios, etc.

> Os mapas coropléticos são elaborados com dados quantitativos e apresentam sua legenda ordenada em classes conforme as regras próprias de utilização da variável visual valor por meio de tonalidades de cores, ou ainda, por uma sequência ordenada de cores que aumentam de intensidade conforme a sequência de valores apresentados nas classes estabelecidas... são indicados para representar distribuições espaciais de dados que se refiram as áreas. (ARCHELA ; THERY, 2008, p.9)

A classificação dos dados é um passo importante na elaboração de mapas coropléticos pois cores e/ou símbolos serão atribuídos às unidades espaciais básicas de acordo com as classes definidas.

Outra representação cartográfica para dados quantitativos é feita com o uso de símbolos proporcionais, sendo apropriada para dados que estão relacionados à dimensão da unidade espacial básica em relação ao fenômeno em estudo como população, Produto Interno Bruto, número de casos de doenças confirmados, tamanhos de rebanhos, etc.

> A variação do tamanho do signo depende diretamente da proporção das quantidades que se pretende representar. Geralmente, o número de classes com utilização do tamanho, deve atingir no máximo cinco classes. (ARCHELA; THERY, 2008, p.10)

Em algum momento o responsável pela representação cartográfica precisa organizar os dados e, para isso, é preciso definir o número de classes a adotar, qual o esquema de cores ou símbolos mais apropriados e, dentre vários critérios de classificação, qual o mais adequado para o fenômeno em estudo.

A escolha do critério de classificação de dados depende de alguns fatores como, por exemplo, a amplitude relativa do intervalo de variação dos dados, existência ou não de dados discrepantes, presença de simetria ou assimetria e, até mesmo, o público alvo, tendo em vista que em estudos de determinadas áreas do conhecimento observa-se a preferência por parâmetros baseados nas medidas de posição, como os quantis, mas em outras, nota-se o uso mais frequente de parâmetros baseados nas medidas de tendência central e de dispersão como, por exemplo, média e desvio padrão.

Neste trabalho são apresentadas contribuições para melhor compreensão e uso com embasamento metodológico dos critérios mais comumente adotados na classificação de dados geográficos para uso em gráficos, mapas coropléticos, mapas de símbolos proporcionais ou, ainda, para

explorar a relação existente entre eles. Destaca-se que, apesar de sua relevância, a semiologia gráfica não é discutida neste estudo, mas discussões relevantes sobre o tema podem ser vistas em Martinelli (2008), Martinelli (2014), Archela (2001), Archela e Théry (2008).

Para tanto, este trabalho é divido em 5 seções. Nesta primeira seção, são feitas algumas considerações iniciais para contextualizar e apresentar a motivação para produção do texto. Em seguida, são apresentados estudos relacionados ao tema proposto que fundamentam algumas escolhas metodológicas e apresentam formas diferenciadas na abordagem da questão proposta. Na terceira seção, são apresentados alguns conceitos fundamentais e duas formas para tratamento preliminar dos dados em busca de valores discrepantes. Na penúltima seção, a quarta, são apresentados critérios para classificação de dados, tomando como referência dados sobre a Taxa de Mortalidade Infantil e, finalmente, na última seção são traçadas algumas considerações finais e indicação de temas a serem aprofundados.

Estudos relacionados

Tendo em vista a proposta de discutir o tema da classificação de dados aplicada à representação cartográfica de dados espaciais, nesta seção são apresentados alguns estudos relacionados ao tema.

Coulson (1987) trata o tema da definição de intervalos de classes do ponto de vista do objetivo geográfico dando particular atenção às abordagens sugeridas nos trabalhos de George F. Jenks. Para Jenks e Coulson (JENKS; COULSON *apud* COULSON, 1987, p. 16), uma análise de mapas preparados por autores em várias disciplinas acadêmicas falham em mostrar algum procedimento racional ou padronizado para a seleção dos intervalos de classes.

> O limite superior do número de classes é determinado, não apenas pela natureza dos dados, mas pelo fato de que o olho humano não pode distinguir entre pequenas diferenças nos valores de tonalidade. O limiar de diferenciação para a tonalidade não é conhecido, tanto para cores como para tons de cinza, mas acredita-se que sete ou oito tonalidades em uma sequência espectral se aproximam do limite para um leitor de mapas habitual. (JENKS *apud* COULSON, 1987, p.17)

São discutidos aspectos como padrões espaciais simples, os impactos das unidades espaciais que podem variar em tamanho e forma, séries de mapas e número de classes. As variâncias observadas são comparadas ao aplicar métodos de determinação de intervalos de classes como: Jenks Ótimo, desvio padrão, quantis, passos iguais, aritmético modificado e geométrico.

Ramos e Sanches (2000), fazem análise de diversas técnicas de classificação de dados utilizados em cartografia para estabelecimento de intervalos de classe e destacam três questões a serem consideradas com este fim: o número de classes, o intervalo de classe ideal e a possibilidade de estabelecer o nível de generalização ou perda de detalhe. Os autores destacam, ainda, que dados espaciais apresentados sob forma de tabelas apresentam grande potencial analítico, porém não comunicam a dinâmica espacial. Fica evidenciada nessa fala a relevância deste tipo de estudo para a produção de mapas coropléticos.

Brewer e Pickle (2002) abordam a evolução de métodos para

classificação de dados e buscam determinar quais são mais adequados para representação de indicadores epidemiológicos. Os autores detectaram que os epidemiologistas adotam, com maior frequência, critérios baseados em quantis. Sete metodologias foram avaliadas partindo-se das respostas de cinquenta e seis pessoas sobre mapas individuais e sobre coleções de mapas.

Os autores destacam que o responsável pela produção cartográfica precisa saber a finalidade do mapa a ser elaborado para identificar o método de classificação mais adequado, na tentativa de evidenciar as questões específicas que o leitor verá no mapa e, ainda, que quando estiver trabalhando em um ambiente computacional, várias opções devem ser testadas para identificar o quanto os padrões espaciais são sensíveis aos critérios de classificação.

Para Brewer e Pickle (2002) os critérios baseados em quantis parecem ser os melhores métodos para auxiliar a comparação, além de facilitar a leitura geral de mapas.

Kumar (2004) discute o uso do histograma de frequência como legenda em substituição às tradicionais legendas observadas em mapas coropléticos. Para o autor, grandes variações de áreas e formas das unidades espaciais básicas podem dificultar a habilidade dos leitores na compreensão das distribuições estatísticas em mapas coropléticos. Acredita-se que o uso do histograma junto ao mapa coroplético pode auxiliar o leitor a interpretar, simultaneamente, a distribuição espacial e a distribuição estatística dos dados em estudo.

O uso do histograma como legenda foi testado em diferentes tipos de distribuições estatísticas e uma aplicação customizada foi desenvolvida por Kumar, usando o ArcInfo 9.0. O autor destaca, ainda, que o conhecimento estatístico prévio ajuda efetivamente o uso e a interpretação das legendas com histograma de frequência nos mapas coropléticos.

Archela e Théry (2008) tratam de conceitos fundamentais do processo cartográfico que devem ser observados na construção de mapas úteis para a análise e compreensão do espaço geográfico apresentando uma orientação metodológica para construção e leitura de mapas.

A semiologia gráfica desenvolvida por Bertin (1967) dá suporte à discussão de Archeda e Théry para análise das vantagens e limites da percepção empregada na simbologia permitindo formulação de regras para uso racional da linguagem cartográfica. Os fenômenos são classificados em qualitativos, ordenados ou quantitativos e, assim, são apresentadas sugestões mais adequadas para produção de mapas capazes de transmitir a informação de modo eficaz.

Conceitos fundamentais

Ao classificar um conjunto de dados, o primeiro passo a tomar é a definição do número de classes ou categorias que será adotado. O número de classes (k) não deve ser muito pequeno, pois provocará a perda de detalhes na análise dos dados, tendo em vista que cada classe deverá comportar muitos dados e que, provavelmente, não devem ser homogêneos. Um número excessivo de classes poderá ter o efeito oposto, ou seja, detalhes em demasia, o que poderá dificultar a análise da distribuição dos dados, além de dificultar a distinção de cores ou tonalidades na produção de mapas coropléticos.

Assim, sugere-se que o número de classes varie de 5 a 8 e, em casos excepcionais, números que se afastem pouco deste intervalo.

Uma sugestão para definição do número de classes é dada pela fórmula de Sturges $k = 1 + \log_2 n$ ou, usando logaritmos decimais, $k = 1+\log n/\log 2$. Por exemplo, para um conjunto com n = 30 valores, temos:

$$k = 1 + \frac{\log 30}{\log 2} = 1 + \frac{1,477}{0,301} \cong 6 \text{ classes.}$$

A média aritmética simples, ou média, dada pela expressão $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ é uma medida estatística de tendência central, também conhecida como centro de gravidade da distribuição, sugerindo que os dados circulam em seu entorno.

Para avaliar a dispersão dos dados em relação à média, ou seja, como os dados gravitam em seu entorno, faz-se o uso de medidas de dispersão como o desvio padrão (s), dado pela expressão $s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}$. Quanto maior o desvio padrão em relação à média maior a variabilidade do conjunto de dados e quanto menor o desvio padrão em relação à média mais concentrados em torno da média devem estar os dados.

. Um uso simples da conjugação das medidas média e desvio padrão é a determinação do intervalo de tendência central. Para dados com distribuição próxima da normal, espera-se que aproximadamente 64% dos dados estejam no intervalo que vai de $\left(\bar{x} - s\right)$ a $\left(\bar{x} + s\right)$. Por exemplo, um conjunto de dados normalmente distribuído com média igual a 9 e desvio padrão igual a 2, terá como intervalo de tendência central, 7 a 11, ou seja, uma parcela significativa dos dados deverá estar inserida entre esses valores.

A Mediana (Md) é uma medida estatística de tendência central e também uma medida de posição. Sua principal característica é dividir um conjunto ordenado de valores em dois grupos iguais. Sendo assim, a primeira metade deve conter valores menores ou iguais à Mediana e a outra metade deve ser composta de valores maiores ou iguais à Mediana.

Como exemplo, considere o conjunto de dados previamente ordenados: 2, 7, 8, 11 e 13. Para encontrar a posição da Mediana usamos a expressão $p = \frac{n+1}{2}$, em que n é a quantidade de dados. Ou seja, $p = \frac{n+1}{2} = \frac{5+1}{2} = 3$ e, portanto, a Mediana é o dado que ocupa a terceira posição do conjunto ordenado de dados, logo, Md = 8.

Outra medida estatística de posição importante é o Percentil. Enquanto a Mediana é calculada com a divisão do conjunto de dados em duas partes iguais, os percentis usam a divisão do conjunto de dados ordenados em 100 partes iguais. Ou seja, o Percentil 80, P_{80} , separa os dados deixando 80% das informações nele, ou abaixo dele e, obviamente, os 20% restantes nele ou acima dele.

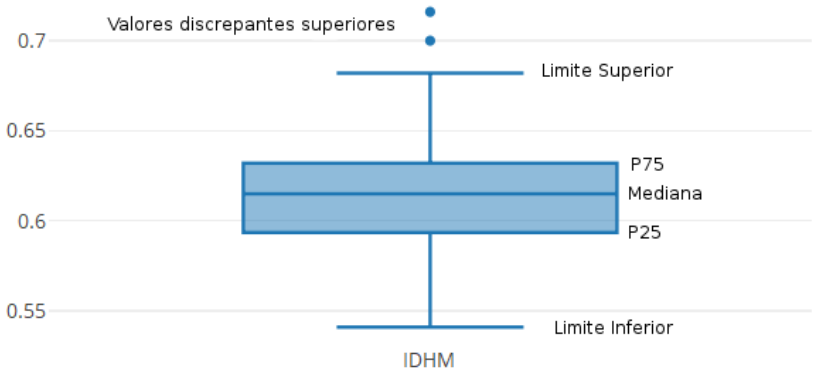
Um dado discrepante ou *outlier* é um valor atípico dentro do conjunto de dados, podendo ser um valor discrepante superior, quando é muito maior que os demais valores, ou discrepante inferior, quando é muito menor. A detecção de valores discrepantes é importante, pois eles podem influenciar significativamente os resultados das medidas estatísticas, como, por exemplo, a média, fazendo com que essa medida deixe de representar uma tendência central do conjunto.

Quanto à classificação dos dados, valores discrepantes podem prejudicar a visualização da distribuição dos dados, como será discutido mais adiante.

O Boxplot, ou gráfico de caixa, é um gráfico utilizado para avaliar a distribuição dos dados e a presença de valores discrepantes, sendo construído tomando como base cinco medidas estatísticas: Mínimo, Percentil 25, Mediana, Percentil 75 e Máximo, conforme Figura 1.

O Limite Superior (LS) é obtido com a operação $LS = P_{75} + 1{,}5 \cdot (P_{75} - P_{25})$ e o Limite Inferior (LI) é dado por $LI = P_{25} - 1{,}5 \cdot (P_{75} - P_{25})$. O termo $(P_{75} - P_{25})$ é denominado diferença interquartis (DIQ), já que o percentil 25 é equivalente ao quartil 1 e o percentil 75 é equivalente ao quartil 3, medidas obtidas ao dividir o conjunto de dados ordenados em quatro partes iguais.

O deslocamento do box (caixa) ou da Mediana para cima ou para baixo sugere distribuição heterogênea dos dados. Quanto mais homogênea a distribuição dos dados mais centralizado fica o box e mais igualmente espaçadas devem estar as linhas dos percentis.



Elaboração: CARVALHO, P.F.B.

Figura 1: Elementos fundamentais de um boxplot (Gráfico de caixa)

Metodologia

Nesta seção metodológica são discutidos os princípios da classificação de dados e apresentados alguns dos critérios mais comumente encontrados na literatura, ou seja, a divisão dos dados em classes usando: intervalos de amplitudes iguais; média e desvio padrão; quantis; quebras naturais.

No início do processo de classificação de dados, a existência ou não de valores discrepantes deve ser avaliada e, caso sejam observados, é tomada a decisão de como trata-los. Para mostrar a importância desta etapa, considere, como exemplo, o resultado apresentado na Figura 2 obtido com a classificação dos dados de densidade demográfica nos 853 municípios de Minas Gerais no ano de 2010.

Fonte: IBGE (2017) – Elaboração: CARVALHO, P.F.B.

Figura 2 Mapa coroplético da Densidade Demográfica em Minas Gerais/2010

Observe que 846 municípios, dos 853 municípios existentes, estão alocados na classe que varia de 1 a 1420 habitantes/km^2, provocando perda de detalhes na discriminação e análise dos dados, pois foram agrupadas unidades espaciais muito distintas. Isso ocorre pela ocorrência de dados discrepantes, em particular, Belo Horizonte com 7093,42 habitantes/km^2. Mas este é o único valor atípico para o estado de Minas Gerais?

Para identificação de dados discrepantes são apresentadas duas técnicas empíricas: pelo boxplot e com uso da média e desvio padrão.

Triola (2013, p.54) sugere uma técnica baseada no boxplot para identificação de valores discrepantes suaves e valores discrepantes extremos. São considerados dados discrepantes suaves aqueles que superam o percentil 75 em 1,5·DIQ a 3,0·DIQ (DIQ = $P_{75} - P_{25}$ é a diferença interquartis, definida anteriormente) ou estão 1,5·DIQ a 3,0·DIQ abaixo do percentil 25. E os valores que excedem o percentil 75 ou são inferiores ao percentil 25 por mais de 3,0·DIQ são considerados valores discrepantes extremos.

Outra técnica adotada para identificação de valores discrepantes,

em especial para conjunto de dados normalmente distribuídos, toma como referências a média e o desvio padrão. Define-se o limite inferior (li) pela expressão e o limite superior (ls) por , onde é a média e s o desvio padrão. Valores abaixo de li ou acima de ls podem ser considerados discrepantes. Alguns autores adotam, alternativamente, e , ou consideram estes pontos como limites para valores discrepantes extremos.

Como visto, a identificação de valores atípicos é muito importante na fase inicial do tratamento de dados, pois pode revelar dados que estão incorretos desde seu registro ou que merecem uma abordagem especial.

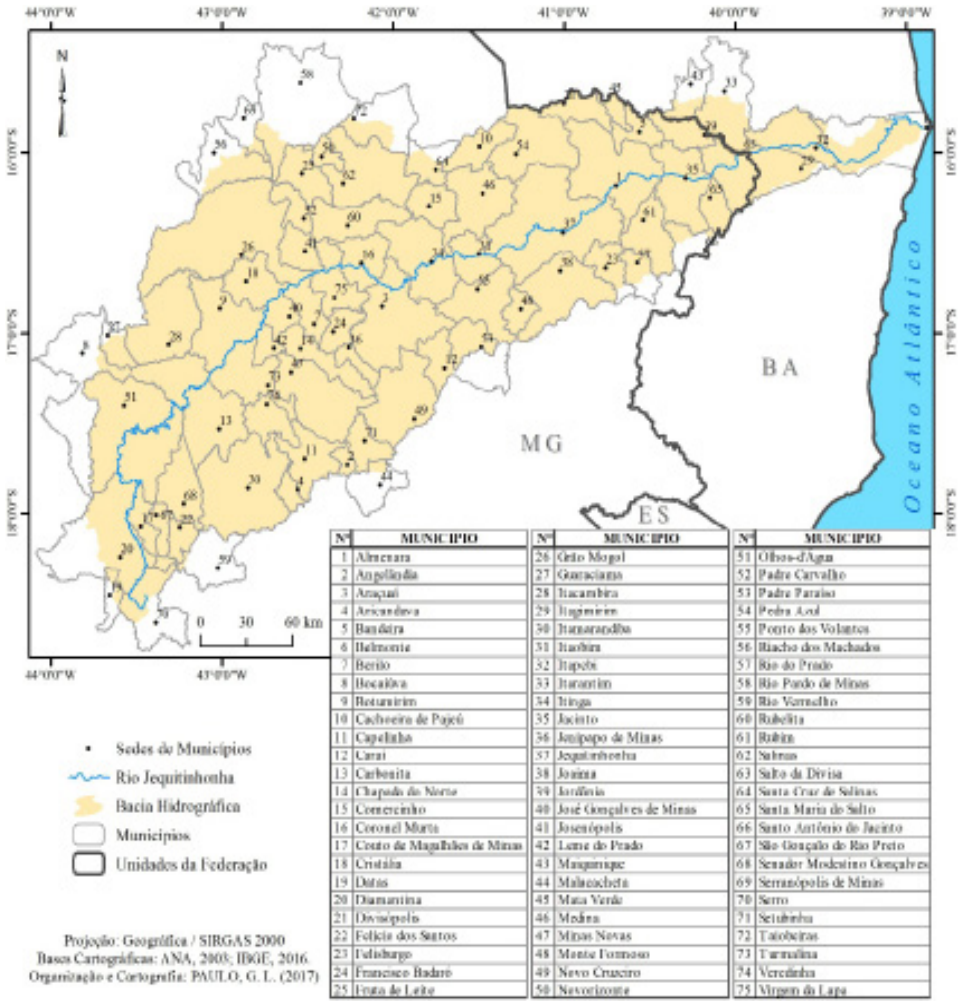
Sobre essa abordagem especial, o pesquisador pode optar por excluir os dados discrepantes do conjunto, criar classes específicas para esses dados ou mesmo não os considerar como dados atípicos classificá-los de acordo com o critério escolhido.

Ao classificar os dados alguns princípios devem ser respeitados:

1. As classes devem englobar todos os dados;
2. Cada dado deve pertencer a uma única classe;
3. Se possível, adotar o mesmo princípio lógico matemático para todas as classes;
4. Caso considere relevante, respeitar divisões de classes convencionadas pela comunidade como, por exemplo, para o Índice de Desenvolvimento Municipal-IDHM que, em sua metodologia[1], sugere 0 a 0,499 como IDHM muito baixo, 0,500 a 0,599 baixo, 0,600 a 0,699 médio, 0,700 a 0,799 alto e 0,800 e acima como muito alto.

A seguir, são apresentados os critérios de classificação usados com maior frequência. Serão discutidos as características fundamentais de cada critério, o processo de elaboração e aspectos relevantes a serem considerados na análise.

Para permitir comparações entre características e resultados, serão explorados os dados referentes às taxas de mortalidade infantil (número de óbitos de menores de um ano de idade por mil nascidos vivos) dos 75 municípios pertencentes à Bacia Hidrográfica do Rio Jequitinhonha para o ano de 2010.



Nº	MUNICÍPIO	Nº	MUNICÍPIO	Nº	MUNICÍPIO
1	Almenara	26	Grão Mogol	51	Olhos-d'Água
2	Angelândia	27	Guaraciama	52	Padre Carvalho
3	Araçuaí	28	Itacambira	53	Padre Paraíso
4	Araçuaíba	29	Itaipé	54	Pedra Azul
5	Bandeira	30	Itamarandiba	55	Ponto dos Volantes
6	Belmonte	31	Itaobim	56	Riacho dos Machados
7	Berilo	32	Itapebi	57	Rio do Prado
8	Bocaiúva	33	Itaranhin	58	Rio Pardo de Minas
9	Botumirim	34	Itinga	59	Rio Vermelho
10	Cachoeira de Pajeú	35	Jacinto	60	Rubelita
11	Capelinha	36	Jequitaí	61	Rubim
12	Caraí	37	Jequitinhonha	62	Salinas
13	Carbonita	38	Jenipapo de Minas	63	Salto da Divisa
14	Chapada do Norte	39	Jordânia	64	Santa Cruz de Salinas
15	Comercinho	40	José Gonçalves de Minas	65	Santa Maria do Salto
16	Coronel Murta	41	Josenópolis	66	Santo Antônio do Jacinto
17	Couto de Magalhães de Minas	42	Lelme do Prado	67	São Gonçalo do Rio Preto
18	Crisólita	43	Mangaratuba	68	Senador Modestino Gonçalves
19	Datas	44	Malacacheta	69	Serranópolis de Minas
20	Diamantina	45	Mata Verde	70	Serro
21	Divisópolis	46	Medina	71	Setubinha
22	Felício dos Santos	47	Minas Novas	72	Taiobeiras
23	Felisburgo	48	Monte Formoso	73	Turmalina
24	Francisco Badaró	49	Novo Cruzeiro	74	Veredinha
25	Fruta de Leite	50	Novorizonte	75	Virgem da Lapa

Fonte: IBGE (2017) – Elaboração: CARVALHO, P.F.B.

Figura 3: Mapa de localização da Bacia Hidrográfica do Rio Jequitinhonha

De acordo com o exposto nas fichas de qualificação disponibilizadas pelo DATASUS(2012) , essa medida estima o risco de um nascido vivo morrer em seu primeiro ano de vida e, ainda, altas taxas de mortalidade infantil refletem, de maneira geral, baixos níveis de saúde, de desenvolvimento socioeconômico e de condições de vida.

Assim, a Taxa de Mortalidade Infantil contribui na avaliação dos níveis de saúde e de desenvolvimento socioeconômico da população. Faz-se o alerta de que taxas reduzidas também podem encobrir más condições de vida em segmentos sociais específicos, logo é necessária prudência na análise desse indicador.

> As taxas de mortalidade infantil são geralmente classificadas em *altas* (50 ou mais), *médias* (20-49) e *baixas* (menos de 20), em função da proximidade ou distância de valores já alcançados em sociedades mais desenvolvidas. Esses parâmetros devem ser periodicamente ajustados às mudanças verificadas no perfil epidemiológico. (DATASUS, 2012).

Apesar da classificação apresentada: altas, médias ou baixas taxas de mortalidade, pode-se buscar discriminar mais detalhadamente os municípios da área de estudo.

Critério de classes com amplitudes iguais

Neste primeiro critério, os dados serão classificados em classes com amplitudes iguais, ou seja, a diferença entre o valor final e o valor inicial de cada classe se mantém constante.

Para definição dos limites de cada classe, primeiro deve-se determinar a amplitude total dos dados, dada por . Em seguida, é calculada a amplitude de classe (h), que será igual para todas as classes. Para tanto, basta dividir a amplitude (A) pelo número de classes (k): .

Os limites das classes são definidos partindo do menor valor e adicionando a amplitude de classe h sucessivamente até alcançar o maior valor. Para o caso da Mortalidade Infantil nos municípios da região da Bacia Hidrográfica do Rio Jequitinhonha com cinco classes, temos

Menor valor: 14,2

Maior valor: 39,8

$$A = 39{,}8 - 14{,}2 = 25{,}6$$

$$h = \frac{25{,}6}{5} = 5{,}12$$

E, assim, os intervalos das cinco classes ficam definidos como:

$$14{,}2 \rightarrow 19{,}3$$

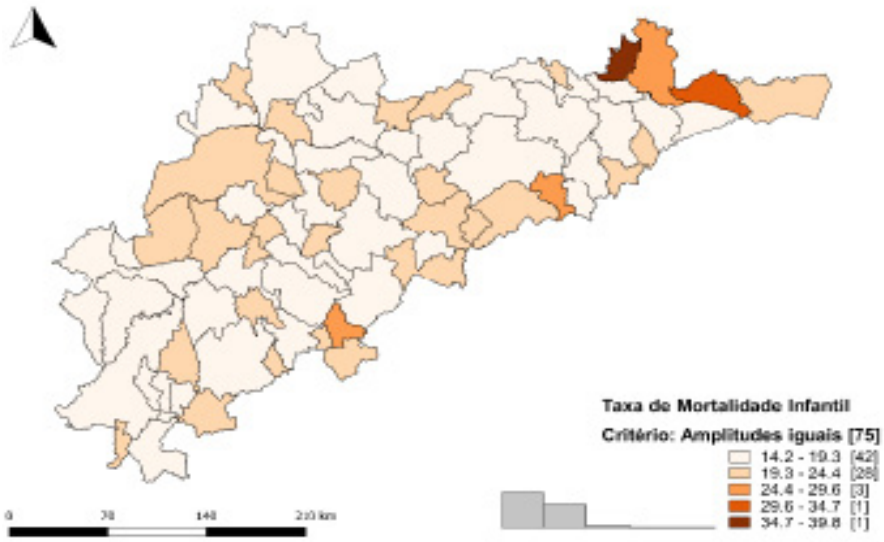
$$19{,}3 \rightarrow 24{,}4$$

$$24{,}4 \rightarrow 29{,}6$$

$$29{,}6 \rightarrow 34{,}7$$

$$34{,}7 \rightarrow 39{,}8$$

Na sequência, é feita a contagem de municípios com taxas de mortalidades dentro de cada classe definida e são elaborados uma tabela distribuição de frequências, o gráfico de frequência e o mapa coroplético, como na Figura 3.



Fonte: Atlas do Desenvolvimento Humano no Brasil (2017) – Elaboração: CARVALHO, P.F.B.

Figura 4: Taxa de Mortalidade Infantil-Bacia Hidrográfica do Rio Jequitinhonha/2010 – Critério das amplitudes iguais.

A Figura 4, sugere que os municípios com piores taxas de mortalidade, no ano de 2010, estão na porção Nordeste da região de estudo – Maiquinique, Itarantim e Itapebi – mais os municípios de Felisburgo e Setubinha, sendo que 56% dos municípios (42) estão na classe de 14,2 a 19,3, considerada baixa taxa de mortalidade infantil, de acordo com as fichas de qualificação do DATASUS. Porém, 44% dos municípios (33) apresentam taxa de mortalidade considerada média.

Sobre o critério adotado, deve-se questionar a grande concentração de municípios nas classes iniciais, com menores taxas de mortalidade, evidenciada pelo gráfico de frequência que apresenta duas colunas muito mais altas que as das demais classes e pela contagem registrada junto à legenda. Essa característica é apenas reflexo dos dados e deve ser mantida assim? A alta concentração não está provocando perda de detalhes na diferenciação dos municípios em análise? É provocada por valores discrepantes superiores? Essa última questão pode ser respondida com uso do boxplot, como discutido anteriormente.

Uma estratégia comum na aplicação deste critério é a geração de outras distribuições de frequências, e respectivas representações cartográficas, com número de classes distintos. Por exemplo, na tentativa de discriminar melhor os dados pertencentes às classes iniciais, poderíamos tentar distribuições com 6 ou 7 classes.

Critério de classes definidas pela média e pelo desvio padrão

O próximo critério a ser discutido está baseado em duas medidas estatísticas descritivas muito comuns, a média e o desvio padrão, e é mais eficiente para dados normalmente distribuídos, ou que apresentam distribuição aproximadamente simétrica.

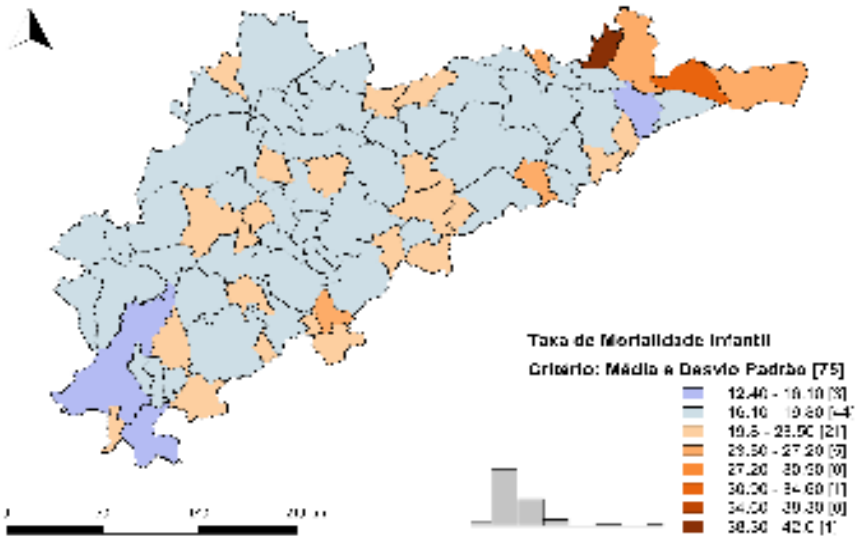
Neste critério, o pesquisador não tem controle sobre o número de classes,

pois a amplitude de cada classe será o desvio padrão. Assim, inicia-se o processo no ponto central, a média, , e faz-se acréscimos e decréscimos sucessivos do desvio padrão, , até alcançar o maior e o menor valor:

$$\begin{aligned} &\ldots \\ \bar{x} - 3s &\rightarrow \bar{x} - 2s \\ \bar{x} - 2s &\rightarrow \bar{x} - s \\ \bar{x} - s &\rightarrow \bar{x} \\ \bar{x} &\rightarrow \bar{x} + s \\ \bar{x} + s &\rightarrow \bar{x} + 2s \\ \bar{x} + 2s &\rightarrow \bar{x} + 3s \end{aligned}$$

A Figura 5, apresenta os resultados obtidos para os dados de Taxa de Mortalidade, com média e desvio padrão . Observe que apenas duas classes apresentam valores menores que a média e considerados com baixa mortalidade infantil, com um total de 47 municípios. Em contrapartida foram criadas seis classes para valores maiores ou iguais à média.

Com o intuito de exemplificar o processo de construção das classes, foram mantidos, na legenda, os valores 12,4 e 42,0. Tais valores podem ser trocados, respectivamente, pela menor taxa de mortalidade observada, 14,2, e pela maior taxa de mortalidade, 39,8 tendo em vista que os extremos adotados na legenda inexistem no conjunto de dados.



Fonte: Atlas do Desenvolvimento Humano no Brasil (2017) – Elaboração: CARVALHO, P.F.B.

Figura 5: Taxa de Mortalidade Infantil-Bacia Hidrográfica do Rio Jequitinhonha/2010 – Critério da Média e Desvio Padrão.

Mais uma vez, pelo gráfico de frequência, observa-se concentração maior de valores nas classes iniciais, com menores taxas de mortalidade, mas, diferentemente do resultado obtido no critério anterior, as classes com maiores frequências são a segunda e a terceira, ou seja, com valores mais próximos da média. Este critério também aponta a porção no extremo nordeste da bacia como aquela com piores indicadores.

A presença de duas classes com frequências nulas, a quinta e a sétima, sugere a presença de algum dado discrepante superior, neste caso, a taxa de 39,8 registrada em Maquinique, no estado da Bahia. Cabe aí uma

discussão sobre a exclusão, ou análise em separado, deste dado na elaboração dos gráficos, tabelas e mapas. Uma alternativa é a criação de uma classe exclusiva para Maquinique, e recálculo das medidas de referência para criação das classes, neste caso, calcular média e desvio padrão, sem o valor de 39,8.

Longley e Goodchild (2013, p.313) sugerem o uso de uma rampa de duas cores na legenda para enfatizar valores acima e abaixo da média, como feito na Figura 5.

Critério dos quartis

Nos critérios discutidos anteriormente, as classes apresentavam mesma amplitude, diferença entre o limite inferior e o limite superior da classe, e frequências que variavam de uma classe para outra.

Os dois próximos critérios, dos quartis e dos percentis, poderiam ser tratados em uma única sessão, tendo em vista que partem de uma mesma lógica matemática mas, pela popularidade observada para o primeiro, faz-se a opção de trabalhar o critério dos quartis com um pouco mais de profundidade.

Para classificação dos dados serão usados os quartis, equivalentes aos percentis 25, 50 e 75, ou seja, medidas que dividem os dados em quatro classes com mesmas frequências:

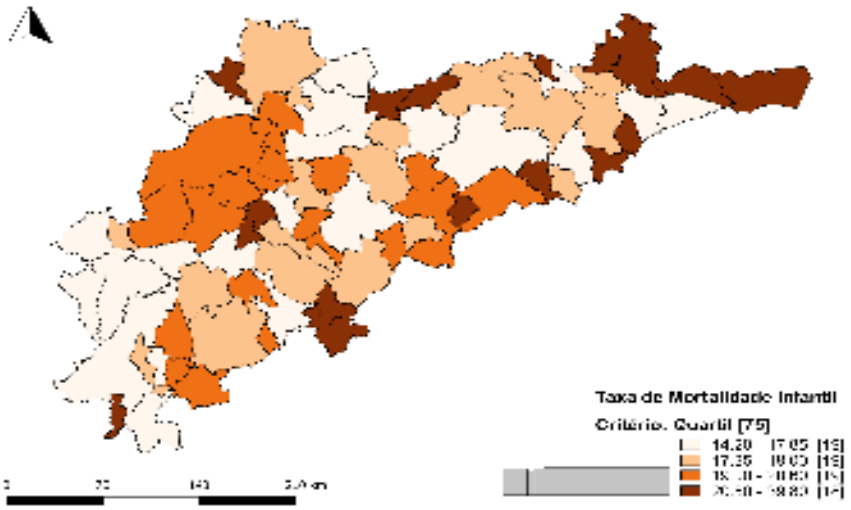
$$\text{Menor valor} \rightarrow Q_1$$

$$Q_1 \rightarrow Q_2$$

$$Q_2 \rightarrow Q_3$$

$$Q_3 \rightarrow \text{maior valor}$$

A Figura 6 apresenta o resultado obtido com a aplicação do critério dos quartis. Diferentemente dos critérios anteriores, no gráfico de frequência, deve-se observar as larguras das colunas, que indicam as amplitudes de cada classe, e não as alturas, que indicam as frequências. Isso porque todas as classes apresentam a mesma frequência, igual a 19, exceto a quarta, tendo em vista que 75, o número de municípios, não é divisível por 4.



Fonte: Atlas do Desenvolvimento Humano no Brasil (2017) – Elaboração: CARVALHO, P.F.B.

Figura 6: Taxa de Mortalidade Infantil-Bacia Hidrográfica do Rio Jequitinhonha/2010 – Critério dos quartis.

Observe que as três classes iniciais apresentam colunas bastante estreitas, ou seja, no pequeno intervalo de 14,20 a 20,60 estão 75% dos dados (57 municípios), sugerindo concentração na distribuição do indicador, enquanto na quarta e última classe, a que varia de 20,60 a 39,80 estão 25% dos dados. Observe que essa última classe tem amplitude de 19,20, sendo igual a três vezes a amplitude que engloba as três classes iniciais, sugerindo, outra vez, a presença de valores discrepantes superiores.

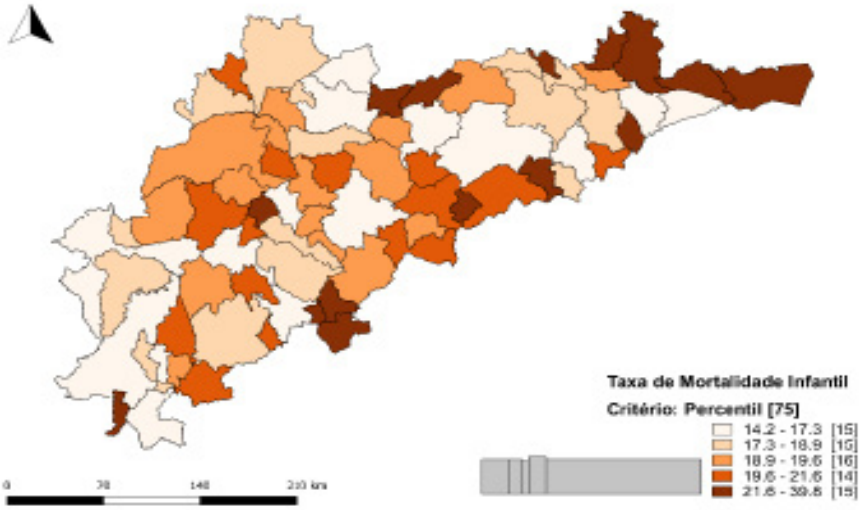
A adoção do critério dos quartis deixa evidente a heterogeneidade na distribuição espacial da taxa de mortalidade nos municípios que compõem a Bacia Hidrográfica do Rio Jequitinhonha. Também reforça que os municípios do extremo nordeste da região, pertencentes ao estado da Bahia, apresentam altas taxas de mortalidade infantil.

É importante destacar, mais uma vez, a mudança no foco de análise. Enquanto nos dois primeiros critérios a frequência era ponto central, ao usar o critério dos quartis as amplitudes das classes é que trazem maior contribuição ao estudo de variabilidade ou homogeneidade do fenômeno.

Critério dos percentis (ou quantis)

O critério dos percentis pode ser visto como uma generalização do critério anterior, pois, ao adotar os quartis como referência, foram criadas quatro classes com frequências iguais, mas com os percentis podemos criar quantas classes quisermos.

Por exemplo, ao adotar o percentil 20 como referência serão criadas 5 classes com mesmas frequências. Adotando o percentil 12,5 serão oito classes , etc.



Fonte: Atlas do Desenvolvimento Humano no Brasil (2017) – Elaboração: CARVALHO, P.F.B.

Figura 7: Taxa de Mortalidade Infantil-Bacia Hidrográfica do Rio Jequitinhonha/2010 – Critério dos percentis.

Na Figura 7, temos o caso dos dados de Mortalidade Infantil com cinco classes, ou seja, cada classe deve incorporar 20% dos dados, exceto por motivo de arredondamentos ou dados iguais, como nas terceira e quarta classes.

A leitura mantém o foco nas amplitudes e não nas frequências das classes. Assim, mais uma vez, observamos quatro classes com pequenas amplitudes, no intervalo de 14,2 a 21,6 e uma classe com grande amplitude, de 21,6 a 39,8 sugerindo a presença de dados discrepantes superiores e/ou alta concentração em dados intermediários ou menores.

Uma desvantagem dos critérios dos quartis e percentis é o fato permitir agrupamento de unidades espaciais associadas a valores muito diferentes.

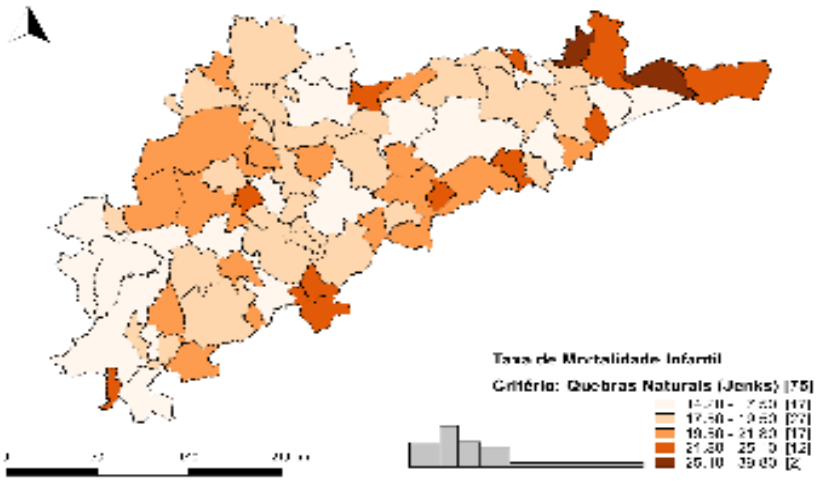
Critério das quebras naturais

O último critério apresentado, também conhecido como Critério de Jenks, explora o comportamento inerente à distribuição dos dados para determinação dos limites das classes. Além de, em geral, serem obtidas classes com amplitudes distintas, como nos quartis e percentis, também podem ser registradas classes com frequências também distintas, como nos dois primeiros critérios explorados neste texto. Assim, na análise dos resultados, deve-se ficar atento tanto às amplitudes das classes quanto às frequências de cada classe.

Tendo em vista a complexidade matemática em que esse critério está embasado, faz-se a opção por explora-lo intuitivamente.

O critério das quebras naturais busca identificar limites que provoquem maior homogeneidade dos dados dentro de cada classe e maior heterogeneidade entre as classes, ou seja, busca-se reduzir a variabilidade dentro das classes e maximizar as diferenças entre as classes.

Dependendo da irregularidade na distribuição dos dados esse critério pode gerar classes com frequências muito distintas, mas, ainda assim, é gerada uma classificação adequada para representação cartográfica de diversos fenômenos, sendo, inclusive, adotada como critério padrão em vários sistemas de informação geográfica.



Fonte: Atlas do Desenvolvimento Humano no Brasil (2017) – Elaboração: CARVALHO, P.F.B.

Figura 8: Taxa de Mortalidade Infantil-Bacia Hidrográfica do Rio Jequitinhonha/2010 – Critério das quebras naturais (Jenks).

A Figura 8 apresenta o resultado da classificação para os dados de Mortalidade Infantil usando o método das quebras naturais, com cinco classes. Observa-se discriminação dos municípios em classes com frequências e amplitudes distintas, cujos limites foram determinados pela natureza da distribuição dos dados.

Fica registrada maior frequência de municípios (51) com taxas de mortalidade nas três classes iniciais (14,20 → 21,80), sugerindo maior concentração de dados em um pequeno intervalo de dados e relacionados com taxas mais baixas. A quinta classe (25,10 → 39,80) apresenta grande amplitude e baixa frequência, sugerindo a presença de dados discrepantes superiores, mas a quarta classe (21,80 → 25,10), com valores intermediários superiores, se descola mais suavemente das classes predecessoras, ocorrendo melhor discriminação das taxas de mortalidade infantil.

Representação com uso de escala de valor visual e figuras proporcionais

Archela e Théry (2008, p. 10) recomendam mapas de símbolos proporcionais para representar dados absolutos tais como população em número de habitantes, produção, renda, mas que o uso de mais de um símbolo em um mesmo mapa deve ser evitado, pois dificultará a comunicação cartográfica.

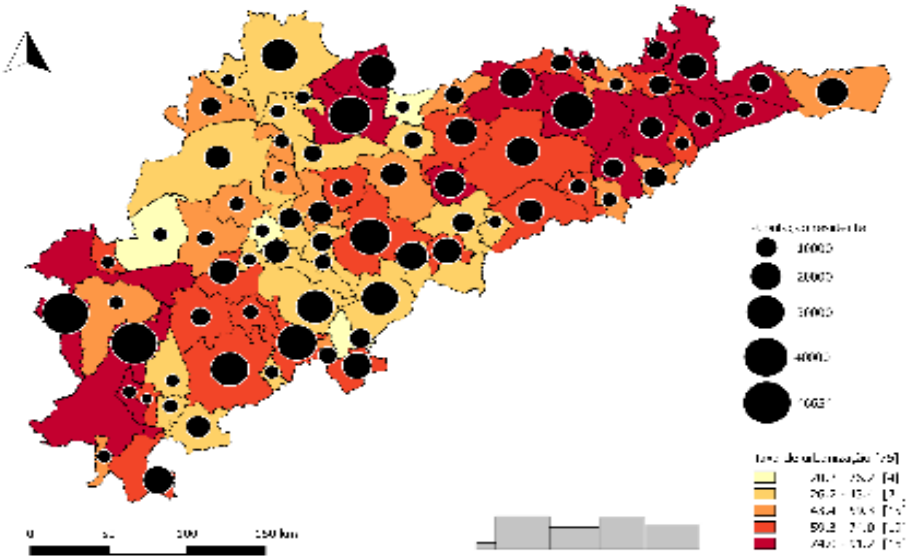
Neste caso, a definição de limites de classes não é adequada pois o tamanho do símbolo deve ser proporcional a cada valor registrado e, ainda, deve ser posicionado em um ponto dentro da unidade espacial básica como, por exemplo, no centroide do polígono ou na sede de um município.

Ao representar cartograficamente duas variáveis, o uso de símbolos proporcionais e de uma escala de valor visual são complementares. As variáveis população residente e taxa de urbanização dos municípios da Bacia

Hidrográfica do Rio Jequitinhonha em 2010 permitem análise complementar e estão representadas na Figura 9.

A taxa de urbanização foi obtida com a divisão do número de residentes em área urbana pela população total do município e o resultado foi multiplicado por 100. Assim, quanto maior a população urbana do município, maior a taxa de urbanização que, teoricamente, poderia variar de 0 a 100.

O mapa da Figura 9 mostra que os municípios com maiores populações não são, necessariamente, aqueles com maiores taxas de urbanização e, ainda, que um aglomerado de municípios com maiores taxas de urbanização se encontra na porção Nordeste da região. Outro aglomerado significativo, com baixas taxas de urbanização e poucos habitantes, se encontra na porção Centro-Norte.



Fonte: Atlas do Desenvolvimento Humano no Brasil (2017) – Elaboração: CARVALHO, P.F.B.

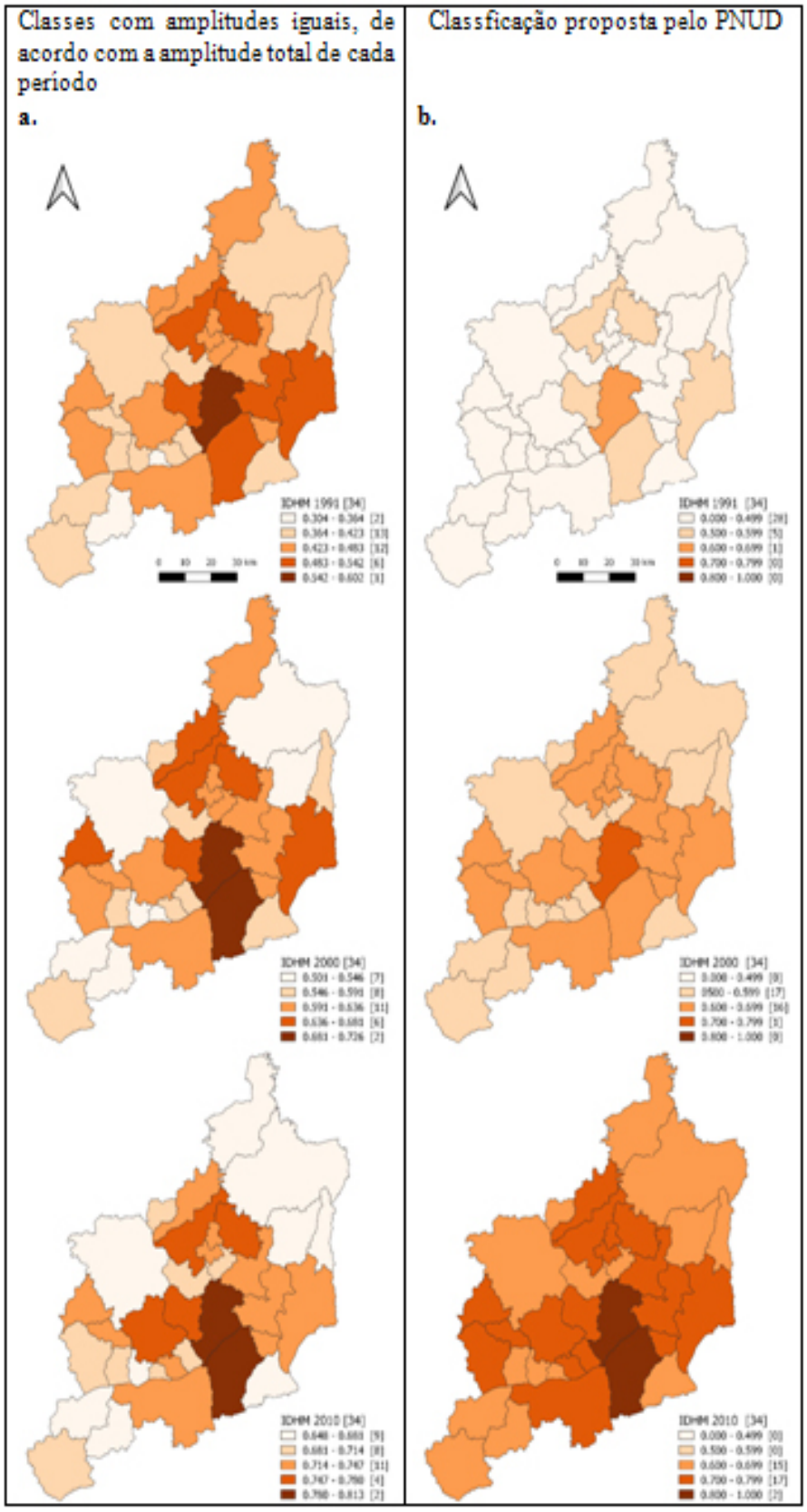
Figura 9: População residente e Taxa de urbanização-Bacia Hidrográfica do Rio Jequitinhonha/2010 – Critério das quebras naturais (Jenks).

Representação de série espaço-temporal

Uma coleção de mapas pode ser elaborada como apoio à análise comparativa para séries de dados geográficos observados em períodos diferentes. Neste caso, é importante que seja selecionado um critério de classificação único e com os mesmos limites de classes em todos os mapas.

Para que os dados de qualquer um dos períodos em estudo se enquadrem em alguma das classes definidas, o limite inferior da primeira classe deve ser igual ao menor valor da variável no conjunto de todos os períodos de estudo e o limite superior da última classe deve ser igual ao maior valor da variável no conjunto de todos os períodos de estudo.

Na Figura 10 estão representados, para os anos de 1991, 2000 e 2010, o Índice de Desenvolvimento Humano Municipal-IDHM para os 34 municípios da Região Metropolitana de Belo Horizonte-RMBH, usando o critério de classes com amplitudes iguais, de acordo com a amplitude total dos dados em cada período (Figura 10a), e pela classificação sugerida pelo PNUD (Figura 10b).



Fonte: Atlas do Desenvolvimento Humano no Brasil (2017) – Elaboração: CARVALHO, P.F.B.

Figura 10: Índice de Desenvolvimento Humano Municipal-Região Metropolitana de Belo Horizonte/1991-2000-2010

Ao atribuir os mesmos limites de classes para os três períodos (Figura 10b), a evolução do Índice de Desenvolvimento Humano na RMBH no período de 1991 a 2010 ficou evidenciada e comparável, o que não ocorre na classificação com amplitudes iguais, e classes definidas pelos dados de cada período (Figura 10a).

Como exemplo, na Figura 10a, o município de Jaboticatubas, ao Norte da RMBH, pertence à primeira classe na escala visual de valor para os anos de 2000 e 2010, apesar de essa classe representar valores diferentes de IDHM, 0,501 a 0,546 em 2000 e 0,648 a 0,681 em 2010, com isso, a evolução no indicador desse município não ficou tão simples de ser notada.

De modo análogo, ao produzir uma coleção de mapas para analisar a distribuição de um determinado fenômeno geográfico, mas em regiões diferentes, os mesmos limites de classes também devem ser usados em todos os mapas.

Considerações Finais

As Figuras 4 a 8 mostram cinco diferentes representações cartográficas para um mesmo fenômeno: a Taxa de Mortalidade Infantil nos 75 municípios que compõem a Bacia Hidrográfica do Rio Jequitinhonha no ano de 2010.

Espera-se que, com este trabalho, tenha ficado clara a importância de avaliar os critérios de classificação dos dados antes de sua representação cartográfica. É claro que outros aspectos, além dos tratados aqui, devem ser discutidos, como escala do mapa, cores, símbolos, objetivo, público alvo, dentre outros.

Na busca pela melhor representação, vários critérios de classificação de dados devem ser testados assim como variações no número de classes adotadas. O objetivo da pesquisa e o público alvo também devem ser levados em consideração na classificação dos dados.

Cada critério está focado em uma ou mais características na classificação dos dados. Por exemplo, ao analisar dados classificados pelo critério dos quantis (quartis ou percentis), deve-se dedicar especial atenção às amplitudes de classes, já que todas as classes apresentam mesmas frequências ou pequenas diferenças por questões de arredondamento ou divisibilidade. As variações das amplitudes podem sugerir concentrações na distribuição dos dados ou mesmo possível presença de valores discrepantes.

Ao adotar os critérios do desvio padrão e média ou de classes com amplitudes iguais a análise é preciso levar em conta as frequências e a distribuição dos dados, em busca de algum padrão na distribuição espacial ou da simetria/assimetria no espalhamento dos dados entre as classes.

Para tanto, a inclusão do gráfico de frequências pode se tornar uma importante ferramenta complementar de análise, pois, visualmente, as frequências e as amplitudes de classes ficam evidenciadas.

Sobre as taxas de mortalidade infantil, merece destaque o fato de o Brasil ter atendido, já em 2015, um dos Objetivos do Milênio proposto:

> O Brasil atingiu a meta assumida no quarto Objetivo de Desenvolvimento do Milênio (ODM 4) das Nações Unidas, de reduzir em dois terços os indicadores de mortalidade de crianças com até cinco anos. (Portal Federativo,2018)

Referências bibliográficas

ARCHELA, R. S. Contribuições da Semiologia Gráfica para a Cartografia Brasileira. *Geografia*, v. 10, n. 1, p. 45-50. Londrina, 2001.

ARCHELA, Rosely Sampaio e THÉRY, Hervé. Orientação metodológica para construção e leitura de mapas temáticos. *Confins*, v.3, 2008. Disponível em <http://journals.openedition.org/confins/3483> DOI: 10.4000/confins.3483. Acesso: 08 Mar. 2018.

Atlas do Desenvolvimento Humano no Brasil. Disponível em: <http://www.atlasbrasil.org.br/2013/pt/download/>. Acesso: 08 dez. 2017.

BREWER, Cynthia A.; PICKLE, Linda. Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series. *Annals Of The Association Of American Geographers,* [s.l.], v. 92, n. 4, p.662-681, dez. 2002. Informa UK Limited. Disponível em: <http://dx.doi.org/10.1111/1467-8306.00310>. Acesso: dez. 2017.

COULSON, Michael R C. In The Matter Of Class Intervals For Choropleth Maps: With Particular Reference To The Work Of George F Jenks. *Cartographica*: The International Journal for Geographic Information and Geovisualization, [s.l.], v. 24, n. 2, p.16-39, jun. 1987. University of Toronto Press Inc. (UTPress). http://dx.doi.org/10.3138/u7x0-1836-5715-3546.

IBGE - Instituto Brasileiro de Geografia e Estatística. [online] Disponível em: <http://www.ibge.gov.br>. Acesso: dez. 2017.

KUMAR, Naresh. Frequency Histogram Legend in the Choropleth Map: A Substitute to Traditional Legends. *Cartography And Geographic Information Science*, [s.l.], v. 31, n. 4, p.217-236, jan. 2004. Informa UK Limited. Disponível em: <http://dx.doi.org/10.1559/1523040042742411>. Acesso: dez. 2017.

LONGLEY, Paul A. et al. *Sistemas e ciência da informação geográfica*. 3. ed. Porto Alegre: Bookman, 2013. Disponível em: <http://integrada.minhabiblioteca.com.br/books/9788565837651>. Acesso: 09 fev. 2018.

MARTINELLI, M. *Mapas da geografia e cartografia temática. [registro eletrônico]*. [s.l.] : São Paulo : Contexto, 2008. Disponível em: <http://search.ebscohost.com/login.aspx?direct=true&db=cat03476a&AN=pon.8572442189&lang=pt-br&site=eds-live>. Acesso em: 4 jul. 2019.

MARTINELLI, M. *Mapas, gráficos e redes: elabore você mesmo.* São Paulo: Oficina de Textos, 120 p., 2014.

Ministério da Saúde. DATASUS. Disponível em <http://fichas.ripsa.org.br/2012/c-1/?l=pt_BR>. Acesso: 09/02/2018.

PAULO, G.L. *Mapa de Localização dos municípios da Bacia Hidrográfica do Rio Jequitinhonha.* PUC Minas, 2017.

PNUD – Programa das Nações Unidas para o Desenvolvimento. Brasília: Organização das Nações Unidas. *Atlas do desenvolvimento humano no Brasil.* Organização das Nações Unidas. Disponível em: <http://www.atlasbrasil.org.br/2013/pt/home/>. Acesso: dez. 2017.

PORTAL FEDERATIVO. Secretaria de Governo. Disponível em: <http://www.portalfederativo.gov.br/noticias/destaques/meta-da-onu-de-reduzir-a-mortalidade-infantil-e-superada-em-niveis-nacional-e-municipais>. Acesso: 28/02/2018.

RAMOS, Cristhiane da Silva; SANCHEZ, Miguel Cezar. Estudo metodológico de classificação de dados para cartografia temática. *Geografia*, Rio Claro, v. 2, n. 25, p.23-52, ago. 2000.

RAMOS, Ana Paula Marques et al. Avaliação qualitativa e quantitativa de métodos de classificação de dados para o mapeamento coroplético. *Revista Brasileira de Cartografia*, Rio de Janeiro, v. 3, n. 68, p.609-629, mar. 2016.

TRIOLA, Mário F. *Introdução à estatística*. 11. ed. Rio de Janeiro: Ltc, 2013. 707 p.

STERN, Boris; YSAKOWSKI, Yvone; HURNI, Lorenz. *Statistics for Thematic Cartography*. 2011. Disponível em: <http://www.gitta.info/Statistics/en/text/Statistics.pdf>. Acesso: 15 fev. 2018.