Revista Brasileira de Cartografia (2016), Nº 68/3: 609-629
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto
ISSN: 1808-0936

AVALIAÇÃO QUALITATIVA E QUANTITATIVA DE MÉTODOS DE CLASSIFICAÇÃO DE DADOS PARA O MAPEAMENTO COROPLÉTICO

Qualitative and Quantitative Evaluation of Data Classification Methods for the Choropleth Mapping

Ana Paula Marques Ramos[1], José Marcato Junior[2], Mônica Modesta Santos Decanini[3], Edmur Azevedo Pugliesi[4], Renan Furlan de Oliveira[5] & Antonio Conceição Paranhos Filho[6]

[1]Universidade do Oeste Paulista – UNOESTE
Departamento de Geografia – Faculdade de Artes, Ciências, Letras e Educação de Pres. Prudente - FACLEPP
Rua José Bongiovani, 700, Cidade Universitária. CEP 19050-920 - Presidente Prudente – São Paulo, Brasil
anaramos@unoeste.br

[2,6]Universidade Federal de Mato Grosso do Sul - UFMS
Departamento de Geografia – Faculdade de Engenharia, Arquitetura e Urbanismo e Geografia – FAENG
Campus Universitário, S/N. CEP 79070-900, Campo Grande – Mato Grosso do Sul, Brasil
jrmarcato@gmail.com, antonio.paranhos@pq.cnpq.br

[3,4,5] Universidade Estadual Paulista - UNESP
Departamento de Cartografia - Faculdade de Ciências e Tecnologia – FCT
Rua Roberto Simonsen, 305, Jardim das Rosas. CEP 19060-900, Presidente Prudente - São Paulo, Brasil
{monca, edmur}@fct.unesp.br, renanfurlanoliveira@hotmail.com

Recebido em 11 de Setembro, 2015/ Aceito em 22 de Janeiro, 2016
Received on September 11, 2015/ Accepted on January 22, 2016

RESUMO

Os mapas coropléticos são amplamente aplicados em estudos geográficos e cartográficos. Este trabalho avalia qualitativa e quantitativamente diferentes métodos de classificação de dados, para a produção de mapa coroplético. A principal contribuição do trabalho é apresentar uma estratégia que auxilie na escolha de método apropriado para classificar dados quantitativos. Realizaram-se dois estudos de caso com dados de densidade demográfica do Brasil. Fez-se a classificação dos dados pelos métodos de Intervalo Igual, Quantil, Desvio-padrão e Otimização de *Jenks*, com cinco classes de representação. Os resultados apontam que métodos distintos propõem diferentes padrões de representação para os dados. Nos casos estudados, conclui-se que a Otimização de *Jenks* foi a regra de classificação que melhor otimizou a distribuição dos dados nas cinco classes de representação. Conclui-se pela análise qualitativa que nenhum dos métodos atende, simultaneamente, a todos os critérios avaliados. A análise pelo melhor ajuste de variância mostrou-se uma ferramenta importante para desenvolver a análise qualitativa, pois viabilizou quantificar numericamente a eficácia de cada método de classificação. Recomenda-se uma análise detalhada da distribuição dos dados antes da seleção do método. Outros resultados são apresentados e suas implicações discutidas.

Palavras chaves: Cartografia Temática, Mapa Coroplético, Métodos de Classificação de Dados.

ABSTRACT

Choropleth maps are largely applied in geographical studies. This work evaluates qualitatively and quantitatively different data classifications methods to produce choropleth maps. The main contribution of this study is propose a strategy, which supports a proper choice of data classification method. Two study of case were performed to population density of Brazil. It It was applied the following data classification schemes: Equal Intervals, Quantile, Standard Deviation and Jenks Optimization for five number of classes. Results show that different data classification schemes provide distinct spatial data pattern distribution. For study cases, it concludes that the Jenks Optimization consists of the classification method that better has distributed the data into the five representation classes. It concludes from the qualitative analysis that none of the four evaluated data classification schemes is able to meeting all the criteria simultaneously. It also concludes that the Goodness of Variance Fit consists of an important tool to the development a quantitative analysis because it offers support to mathematically quantify the effectiveness of each data classification method applied. It is recommended to develop a detailed analysis about data distribution before the classification scheme be started. Other results are showed and their implications discussed.

Keywords: Thematic Cartography, Choropleth Map, Data Classification Methods.

1. INTRODUÇÃO

Mapas coropléticos são extensivamente empregados em aplicações envolvendo o estudo da distribuição espacial de distintos fenômenos (ANDRIENKO *et al*., 2001), com destaque para estudos de: geografia do trabalho (MAIA, 2009); da saúde (CARDIM *et al*., 2013; do crime (LUCENA *et al.*, 2012); geomarketing (PROCHNOW *et al*., 2011); geografia da população (BAPTISTA & RIGOTTI, 2014), dentre outros.

Um mapa coroplético é resultado da representação temática de dados de natureza quantitativa, na qual a dimensão espacial do fenômeno é associada a unidades de enumeração (ex.: países, estados, cidade, setor censitário) (DENT *et al.,* 2009; SLOCUM *et al.,* 2009; MARTINELLI, 2014). 'Coroplético' advém do grego '*choros'* que significa lugar, e '*plethos'* que significa quantidade, portanto, trata-se da descrição de quantidade por lugar (DENT *et al.,* 2009).

O termo 'coroplético' foi proposto pelo geógrafo americano John K. Wright, no início do século XX. Todavia, a primeira representação coroplética foi elaborada ainda no início do século XIX para demonstrar o nível de alfabetização na França. Adotou-se a variável visual valor para transcrever os padrões espaciais relacionados ao nível de alfabetização entre as divisões administrativas do país, em que os valores mais escuros no mapa foram atribuídos aos departamentos com menores índices de crianças na escola (MARTINELLI, 2014). A variação de valor é uma progressão contínua, a qual o olho percebe uma sequência de tons de cinza, por exemplo, do preto ao branco (BERTIN, 1983).

Os mapas coropléticos são apropriados para representar variações de uma unidade de enumeração para outra (CROMLEY, 2005; DENT *et al.* 2009; SLOCUM *et al.,* 2009; MARTINELLI, 2014). Portanto, não são adequados para expressar variações dentro de uma mesma unidade (CROMLEY, 2005; SLOCUM *et al.,* 2009), tampouco para apresentar dados absolutos, ou seja, dados que não podem ser expressos em termos de proporção, razão, taxa (DENT *et al.,* 2009). Sendo assim, no mapeamento coroplético, a padronização ou normalização dos dados é uma medida necessária para a correta representação da distribuição espacial de um fenômeno (PETERSON, 2009).

A elaboração de mapas coropléticos envolve tradicionalmente um processo de classificação de dados (ANDRIENKO *et al*., 2001), no qual distintas unidades, adjacentes ou não, são agrupadas em classes, e uma variável visual, tal como valor, cor-valor é associada às classes (MARTINELLI & MACHADO-HESS, 2014; SLOCUM *et al*., 2009). O objetivo da classificação é definir classes heterogêneas entre si, de modo que os constituintes das classe apresentem característica o mais similar possível com relação ao objeto de estudo.

A classificação ideal é aquela que reúne um único elemento amostral em cada classe de representação, sendo, porém, uma abordagem inviável por duas razões. Primeiro, por dificultar o estabelecimento de padrões espaciais entre

elementos mapeados; segundo, por ser limitada a capacidade do olho humano em discernir grandes variações acromáticas ou cromáticas (DENT *et al.,* 2009). Segundo Krygier & Wood (2005), a classificação é essencial, pois revela padrões que são difíceis de serem estabelecidos a partir de dados não agrupados. O processo de classificação deve ser compreendido como um facilitador da análise espacial, a qual pode ser desenvolvida em diferentes vertentes da Cartografia, tal como na Cartografia Temática.

Diversos métodos podem ser utilizados no processo de classificação de dados para a produção de mapas coropléticos, dentre os quais se destacam: Intervalo igual, Quantil, desvio-padrão e Otimização de *Jenks* (DENT *et al.,* 2009; SLOCUM *et al.,* 2009). Tais métodos podem ser divididos em duas categorias, de acordo com a estratégia empregada para gerar as classes de feições. Tem-se a categoria que leva em consideração a distribuição natural dos dados ao definir os intervalos das classes (desvio-padrão e Otimização de *Jenks*), e a categoria que não considera a semelhança entre as observações ao realizar a classificação do conjunto (Intervalo igual e Quantil) (SLOCUM *et al.,* 2009). Ambas as categorias de métodos estão implementados nas principais plataformas de Sistema de Informação Geográfica (SIG), tais como *ESRI ArcGIS*, *MapInfo*, *Quantum GIS* e *gvSIG*.

De acordo com Andrienko *et al*. (2001), a interpretação da informação em um mapa coroplético é fortemente influenciada tanto pelo método selecionado para classificar os dados quanto pelo número de classes definido para representar a variação do fenômeno. Nesse contexto, a problemática da produção de um mapa coroplético, eficiente e eficaz, isto é, que atenda à demanda do usuário, está em coordenar as diversas decisões que compõem o processo de agrupamento de dados. Isto corresponde em compreender a distribuição dos dados, o número de classes que deve ser representado, o método de classificação que irá estabelecer os intervalos das classes, os fatores a serem considerados ao utilizar as variáveis visuais na representação dessas classes. Em razão dessas complexas decisões, Dent *et al.* (2009) afirmam que os usuários de SIG, muitas vezes, confiam no padrão (*default*) oferecido pelos *softwares,* tanto ao que se refere à quantidade de classes e ao método para gerar essas classes, quanto nos esquemas de variáveis visuais para a representação. Outras vezes usam métodos inapropriados que expressam diferenças entre as classes, mesmo sem considerar a distribuição natural dos dados, simplesmente para dar ênfase ao resultado final. Isso pode ser um problema no processo de comunicação da informação espacial, pois cada método de classificação possui características distintas, as quais, dependendo das particularidades do conjunto amostral, podem consistir em vantagens ou desvantagens.

Assim, a questão central no processo de elaboração de mapas coropléticos está voltada para a seleção de um método de classificação que determine classes, de maneira a transcrever apropriadamente a distribuição dos dados que deveriam explicar o comportamento do fenômeno espacial. Por um lado esta questão pode ser tratada com base em análises subjetivas, as quais envolvem a consideração das características de cada método e as particularidades do conjunto amostral associadas à demanda do usuário (SLOCUM *et al.,* 2009). Por outro esta questão pode ser tratada com base em análise quantitativa, a qual consiste em estimar o Melhor Ajuste de Variância (*Goodness of Variance Fit* – GVF), para cada método aplicado, de maneira que o melhor agrupamento seja o que maximiza o valor do GVF, aproximando-o de 1,0, mas nunca sendo igual a este valor (DENT *et al.,* 2009).

O objetivo desse trabalho é apresentar as etapas do processo de produção de mapas coropléticos, por diferentes métodos de classificação, e realizar uma análise qualitativa e quantitativa da classificação visando auxiliar na elaboração de representações cartográficas que favoreçam uma análise dos padrões espaciais de forma eficiente e eficaz. As questões respondidas são: 'Quais as etapas que compõem o processo de classificação de dados?', 'Os diferentes métodos resultam em diferentes padrões de distribuição?', 'Como avaliar a classificação dos dados, em termos qualitativo e quantitativo?'. A contribuição desse trabalho está em apoiar o processo de comunicação cartográfica, ao que se refere a uma abordagem sistêmica de classificação de dados quantitativos, para a construção de mapas coropléticos que apoiem o desenvolvimento de análises espaciais de fenômenos geográficos.

2. MÉTODO

A abordagem empregada para a avaliação qualitativa e quantitativa dos métodos de classificação de dados para o mapeamento coroplético foi dividida em cinco etapas: análise da natureza do fenômeno; identificação das variáveis para a análise de estatística descritiva; seleção do número de classes e representação gráfica; aplicação dos métodos de classificação; e avaliação da classificação, em termos qualitativo e quantitativo. Para exemplificar a aplicação dessa abordagem, realizaram-se dois estudos de caso por meio do uso de dados de densidade demográfica (população total dividida pela área) de dois estados brasileiros, São Paulo (SP) e Mato Grosso do Sul (MS). Optou-se por realizar estes dois casos para ilustrar o processo de classificação de conjuntos amostrais com distintos números de observações (municípios), um com 79 municípios (MS) e outro com 645 (SP). Os dados foram obtidos junto ao Instituto Brasileiro de Geografia e Estatística (IBGE), no levantamento do Censo Demográfico de 2010 (IBGE, 2015).

2.1 Análise da natureza dos dados

A natureza dos dados tem sido traduzida por meio de três níveis de medida: qualitativo, ordinal ou quantitativo (RAMOS, 2005; MARTINELLI, 2014). Estes níveis de medida empregam uma regra importante na identificação do tipo de mapa temático a ser adotado. Os dados que se enquadram no nível qualitativos descrevem uma relação de similaridade ou diversidade e nenhuma operação matemática pode ser estabelecida entre essas classes. No nível ordinal, os dados mantêm uma relação de hierarquia entre as classes, pois as classes de feição se ordenam espontaneamente, porém, sem caracterizar qualquer manifestação de proporção, pois não há valores numéricos associados às classes. Assim, ao comparar duas classes A e B é possível apenas estabelecer que A é maior (ou menor) que B, sem afirmar o valor dessa magnitude (MARTINELLI, 2014).

Os dados de natureza quantitativa são oriundos de uma contagem, uma estatística (ex.: média, proporção) ou uma medição (ex.: temperatura, elevação). Por exemplo, os dados de população são provenientes de uma contagem, a temperatura em uma cidade advém de uma medição e a porcentagem de mulheres grávidas com mais de 30 anos uma relação/proporção/razão. Além possibilitarem o ordenamento das classes, os dados permitem a quantificação das classes, ou seja, a relação de proporção é imediata nesse tipo de dado (SLOCUM *et al.,* 2009). Neste caso, pode-se afirmar que a classe A é duas vezes maior que a B, ou que a classe C é 10% menor comparada à classe D, e assim sucessivamente.

Os dados quantitativos se subdividem em intervalar, quando o ponto zero da escala de mensuração é arbitrário, ou seja, quando a escala não tem um valor igual a zero absoluto como ponto de partida (ex.: 0º Celsius não significa ausência de temperatura), e de razão, quando o ponto zero é verdadeiro, ou seja, absoluto e indica ausência total (ex.: 0 habitantes em uma cidade significa ausência de habitantes) (SLOCUM *et al.,* 2009). Os dados de densidade demográfica são de natureza quantitativa do tipo razão, pois permite responder questões como: 'o quanto a densidade populacional de determinado município é maior (ou menor) comparada à densidade dos demais municípios?' ou 'qual é o município com maior densidade demográfica?'.

Ainda que um mapa coroplético seja construído a partir de dados quantitativos, a representação de suas classes ocorre por meio de variáveis visuais que expressam a percepção de ordem. Peterson (2009) afirma que a intensidade da cor associada a uma região deve denotar a grandeza do valor da variável nessa região. Portanto, um mapa coroplético é resultado da combinação de uma representação quantitativa, ao que se refere à criação dos intervalos das classes, e de ordem, ao que se refere à representação das classes que compõe sua legenda.

A caracterização da natureza dos dados é uma importante etapa para auxiliar no processo de representação cartográfica, realizando a transcrição gráfica de classes de feição no mapa. Deve-se considerar a natureza do dado para, adequadamente, selecionar variáveis visuais que possuem propriedades perceptivas (dissociativa, associativa, seletiva, ordenativa ou quantitativa) condizentes à relação fundamental (diversidade, ordem ou proporcionalidade) a ser estabelecida entre as feições (MARTINELLI, 2014). Quando

o intuito é sugerir diferenças qualitativas entre as classes, deve-se adotar uma variável visual com a propriedade de seletividade, como a variável visual cor, em sua dimensão de matiz. Isto porque esta dimensão da cor favorece o olho humano isolar instantaneamente as classes formadas, contribuindo para o processo de comunicação cartográfica (JOLY, 2011).

Quando o intuito é expressar diferenças quantitativas entre as classes, deve-se adotar variação de valor, ou cores, que denotem ordem na variação do dado, tal como a variável visual cor, em sua dimensão de saturação ou de valor (KRYGIER & WOOD, 2005). A ordem visual entre as feições pode ser transcrita utilizando tons, dos mais claros até os mais escuros, bem como uma ordem visual construída com texturas, que também vão das mais claras até as mais escuras (MARTINELLI & MACHADO-HESS, 2014; SLOCUM *et al*. 2009).

No mapeamento coroplético, as variáveis visuais utilizadas para transcrever uma percepção ordenada entre as classes de feição são a variável visual valor, proposta por Bertin (1963 apud MARTINELLI, 2014), e a variável visual cor, nas dimensões de valor e saturação (SLOCUM *et al*., 2009). A variável valor simboliza a variação dos tons de cinza, do claro ao escuro, enquanto o valor da cor e a saturação da cor simbolizam, respectivamente, a variação de brilho e de intensidade de um mesmo matiz (DENT *et al.,* 2009). Ao representar dados de densidade demográfica, pela variável visual valor ou valor ou saturação da cor, pode-se verificar claramente os municípios mais e menos populosos dentro de cada unidade mapeanda (MARTINELLI & MACHADO-HESS, 2014).

2.2 Estatística descritiva

Krygier e Wood (2005) afirmam que realizar a classificação de um conjunto amostral, sem antes examinar os dados que o constitui é um procedimento inapropriado. O processo de classificação dos dados se inicia com a análise de estatística descritiva, a qual provê as informações gerais do conjunto amostral (SIEGEL & JUNIOR, 2006). Na produção de mapas coropléticos, a finalidade da análise descritiva é apoiar a seleção do método de classificação que se ajuste às características dos dados, viabilizando o estudo do fenômeno de interesse.

A caracterização dos dados ocorre por meio da determinação de medidas de posição central (ex.: média, mediana) e de dispersão (ex.: desvio-padrão, variância). As medidas de dispersão revelam a variabilidade dos dados, e as medidas de posição a distribuição dos dados (MORETTIN, 2005). A distribuição de frequência pode ser avaliada por testes de normalidade nos dados, tal como o teste *Shapiro-Wilk* sugerido por Razali e Wah (2011).

No processo de classificação, conhecer a distribuição dos dados é importante, primeiro, porque alguns dos métodos de agrupamento definem os intervalos das classes com base na distribuição de probabilidade do conjunto. Segundo, pois, a partir da análise da distribuição de frequência da variável, pode-se estimar a probabilidade de ocorrência de um determinado valor de interesse (MORETTIN, 2005).

Nos casos em que os dados não têm distribuição normal, a distribuição de frequência é estudada por meio das propriedades de assimetria e curtose (MORETTIN, 2005). A assimetria indica a direção de deslocamento dos dados: acima ou abaixo da média amostral. Uma assimetria positiva indica valores concentrados abaixo da média, e a negativa indica valores distribuídos acima da média. Dados com distribuição normal apresentam assimetria nula. Curtose é a curva de achatamento da distribuição dos dados (MORETTIN, 2005). Dados com distribuição normal têm curtose nula, sendo a curva de distribuição nomeada de mesocúrtica. Isto significa que os elementos estão distribuídos uniformemente ao entorno da média. A curva é dita leptocúrtica, e a curtose é positiva, para a função de distribuição com pico superior ao pico da curva mesocúrtica, implicando que os dados concentram-se em um intervalo reduzido de valores. A curva é planicúrtica, e a curtose negativa, para a função com pico inferior ao da curva mesocúrtica (MORETTIN, 2005).

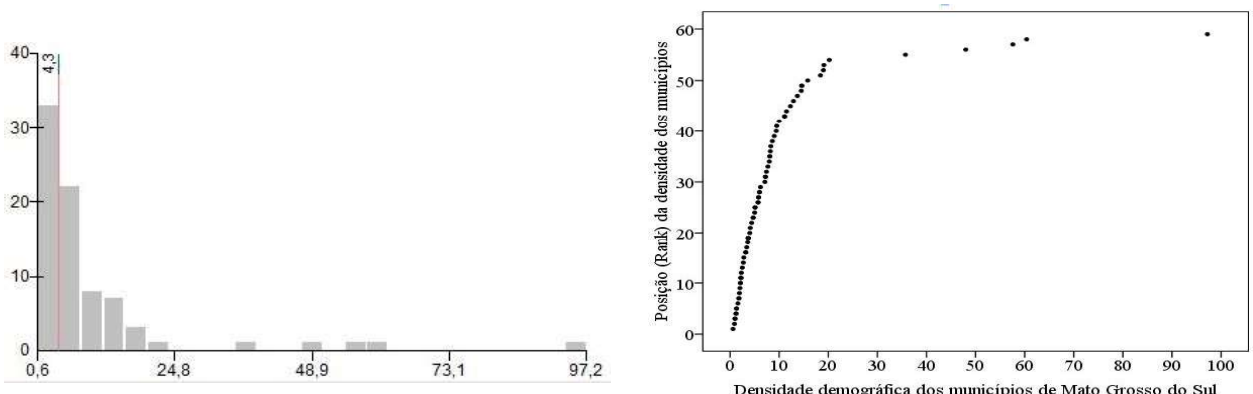
Para os dois estudos de caso, fez-se a análise de estatística descritiva e o teste de normalidade dos dados de densidade demográfica, e os resultados estão na Tabela 1. O histograma de frequência e o gráfico de dispersão desses dados encontram-se nas Figuras 1 e 2. O processamento dos dados ocorreu no *software* SPSS 16.0 (*Statistical Package for the Social Sciences*).

Os resultados da análise de estatística descritiva (Tabela 1) revelam que a densidade demográfica dos dois estados não segue uma distribuição de probabilidade normal, seja pela análise dos valores do teste de *Shapiro Wilk*, dos valores de assimetria e curtose, ou pela análise visual do histograma de frequências (Fig. 1 e 2). Os conjuntos têm assimetria e curtose positivas indicando, respectivamente, que maioria dos municípios possui densidade abaixo da média amostral e que curva de distribuição é leptocúrtica.

Tabela 1: Análise descritiva da densidade demográfica dos estados de SP e MS

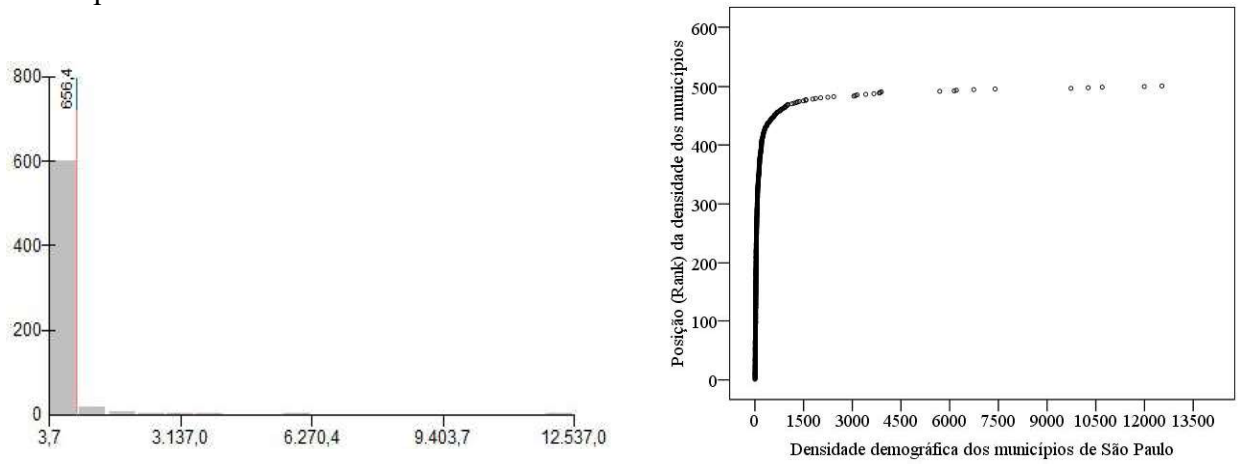
Estatísticas	São Paulo	Mato Grosso do Sul
Número de municípios (amostras)	645	79
Média amostral	301,98	9,83
Mediana	38,90	5,70
Soma	194.779,00	776,70
Valor mínimo	3,70	0,60
Valor máximo	12.537,00	97,20
Amplitude	12.533,30	96,60
Desvio-Padrão da média (σ)	1.196,74	14,86
Variância (σ^2)	1.432.197,10	220,82
Curtose	57,34	17,53
Desvio-Padrão da curtose	0,19	0,53
Assimetria	7,15	3,87
Desvio-Padrão da assimetria	0,10	0,27
Teste Shapiro-Wilk (p-valor)*	< 0,0001	< 0,0001

*p-valor maior que 0,05 assume-se distribuição de probabilidade normal para os dados, ao nível de confiabilidade de 95%.



(a) (b)

Fig. 1 – Histograma de frequência (a) e gráfico de dispersão (b) da densidade demográfica dos municípios do Estado de Mato Grosso do Sul.



(a) (b)

Fig. 2 – Histograma de frequência (a) e gráfico de dispersão (b) para a densidade demográfica dos municípios do Estado de São Paulo.

Conforme a Tabela 1, outras estatísticas também foram aplicadas nos dados de densidade demográfica, como a estimativa dos valores de mínimo e de máximo, a amplitude e a mediana. Os valores de mínimo e de máximo auxiliam na identificação de observações discrepantes na amostra, denominadas de *outliers*. Valores discrepantes geram uma translação da média, para mais ou para menos, de modo que a média não seja um indicador apropriado para os demais valores (MORETTIN, 2005).

A amplitude é a diferença entre o valor de máximo e de mínimo do conjunto. Para o caso de São Paulo (12.533,30), a amplitude dos dados se mostrou muito superior ao valor do Estado do Mato Grosso do Sul (96,60). Todavia, em ambos os estados, constata-se que são poucos os municípios que elevam o valor da amplitude do conjunto. O gráfico de dispersão (Fig. 2B) mostra que a maioria (97%) dos municípios paulistas tem menos de 2000 habitantes por quilômetro quadrado (hab./km^2). Pelo histograma de frequência (Fig. 2A), nota-se que aproximadamente 600 municípios têm densidade abaixo de 656 hab./km^2. No caso do Estado de MS, 74 dos 79 municípios têm menos de 20 hab./km^2 (Fig. 1B), sendo que mais de 30 dos municípios possui menos de cinco pessoas por quilômetro quadrado (Fig. 1A)

Discrepâncias entre valores de amostra resulta em inconsistências na classificação, tal como classes sem observações (SLOCUM *et al.* 2009). Para contornar este problema, Dent *et al.* (2009) sugerem trabalhar com o conceito de amplitude útil do conjunto, ao invés da amplitude total. O princípio é representar o(s) valor(es) mais discrepante(s) em uma única classe separada, e calcular uma nova amplitude do conjunto, sem considerar o(s) valor(es) disposto(s) nesta classe.

Nos casos de estudo, alguns municípios têm densidade populacional muito discrepante dos demais, e podem ser considerados *outliers* na amostra. Para o Estado de São Paulo, esses municípios, por exemplo, referem-se àqueles com mais de 2000 hab./km^2, enquanto no MS aqueles com densidade acima de 25 hab./ km^2. A amplitude útil desses conjuntos é de 1.869,40 hab./km^2 (média = 116,53; mediana = 37,00), para São Paulo, e de 19,60 hab./km^2 (média = 6,50; mediana = 5,10), para Mato Groso do Sul. Dessa forma, para o primeiro caso, uma única classe representaria os municípios paulistas com densidade acima de 2000 hab./km^2 e, no segundo caso, os municípios mato-grossenses com densidade superior a 25 hab./km^2.

A mediana consistiu em outra estatística calculada para os dados (Tab. 1). A mediana é um indicador que divide o conjunto amostral em duas partes iguais. Na ausência de *outliers*, tem-se que as observações com valores acima da média têm maior influência no cálculo da média comparado aos valores abaixo desta, e isso ocorre porque o atributo dos valores acima da média é maior que os valores abaixo. Os dados com distribuição normal têm valores iguais de média, moda e mediana (MORETTIN, 2005). Em um primeiro momento, a mediana da densidade demográfica era equivalente a 1/8 do valor da média, no caso do Estado de SP, e 1/2 no caso de MS (Tab. 1). Ao considerar a amplitude útil do conjunto, essa discrepância reduz para 1/3 e 1/1,28 para SP e MS, respectivamente. Esses resultados permitem afirmar que a densidade dos municípios de SP é mais discrepante comparada à densidade dos municípios sul-mato-grossenses. Nesse sentido, pela análise do histograma de frequência, gráfico de dispersão e valores de amplitude, média e mediana dos dados, constata-se que a densidade demográfica desses dois estados brasileiros não é uniformemente distribuída nos municípios.

Outras duas importantes estatísticas para um conjunto de dados são variância e desvio-padrão. A variância é o quanto os dados são similares entre si. Valor alto de variância indica observações discrepantes e valor baixo sugere similaridade (MORETTIN, 2005). Pela Tabela 1, nota-se um alto valor de variância para os dados do Estado de São Paulo (acima de 1.400.000) e isso confirma a discrepância no número de habitantes por km^2 nos municípios. O desvio-padrão é a raiz quadrada da variância. Um baixo valor de desvio-padrão indica que os dados se concentram ao entorno da média, e um alto valor implica valores distantes da média. Nos municípios de São Paulo, o desvio-padrão é 1.196,74 e, em MS, da ordem de 14,86. Nesse contexto, a análise do desvio-padrão e da variância é uma maneira de se confirmar as interpretações gráficas que se realiza pelo histograma de frequências, gráfico de dispersão, curva de distribuição dos dados.

2.3 Seleção do número de classes e representação gráfica

Para dados de natureza qualitativa ou ordinal, em geral, define-se o número de classes em função da ocorrência das categorias do fenômeno (ex.: tipos de solo, de vegetação). Contudo, nenhum critério estatístico determina o número exato de classes para o mapeamento de dados quantitativos (SLOCUM, *et al.,* 2009). Se por um lado quanto maior o número de classes, mais complexa é a leitura do mapa, por outro quanto menor o número de classes, maior é a generalização do fenômeno e, portanto, maior a omissão dos detalhes do conjunto (DENT *et al.,* 2009). Ainda que não haja regras estatísticas de apoio à seleção do número de classes para classificação de dados quantitativo, a demanda do usuário, os aspectos da percepção visual e a abordagem de Sturges*, de* 1926, são fatores que podem auxiliar nesse processo.

A fórmula de Sturges pode ser utilizada apenas como um ponto de partida para a definição do número de classes, afirmam Dent *et al.* (2009). Isso porque, à medida que o número de elementos amostrais aumenta, o número de classes cresce, e mais complexa é a classificação. *Sturges* relacionou o número de elementos com uma função logarítmica do tipo: $k = 1+ 3{,}33 \log_{10}(N)$, sendo k o número de classes e, N é a quantidade de observações não repetidas no conjunto (DENT *et al.,* 2009). Para amostras com 80 ou menos elementos, essa expressão é simplificada para: $k = \sqrt{N}$.

Pela fórmula de *Sturges,* o número de classes necessárias para representar os 645 municípios do Estado de SP, e os 79 municípios do Estado do MS equivale a, respectivamente, 10 (k_{SP} = 1 + 3,33 $\log_{10}(624)$ = 10,31 ≅ 10 e 9 ($k_{MS} = \sqrt{N} = \sqrt{75} = 8{,}66 \cong 9$) classes. Deve-se ressaltar que quatro municípios do Mato Grosso do Sul têm igual densidade populacional, portanto, foram desconsiderados do cálculo do número de classes, assim N = 75. Para São Paulo, 21 municípios foram desconsiderados da definição do número de classes (N = 624).

A percepção visual é outro fator que pode auxiliar na definição do número de classes em um mapeamento. A percepção se relaciona à apreensão de objetos, a partir do uso dos sentidos; a percepção visual é o uso da visão para a apreensão de objetos (DOWNS & STEA, 2005). O sistema perceptivo humano tem limite na capacidade de discernir variações acromáticas (do branco ao preto - tons de cinza) e cromáticas (valor, saturação da cor). Pelas leis da percepção visual, o olho humano é capaz de discriminar até oito variações de tons de cinza com exceção de treinamento prévio (DENT *et al.,* 2009). Para uma variação cromática ou acromática muito pequena entre as classes, a diferenciação se torna quase imperceptível aos olhos. Isso principalmente para as classes visualizadas no contexto do mapa, por estarem distantes umas das outras, ou simbolizadas na legenda, por estarem representadas próximas, sendo esta influência devido ao processo de contraste por indução (DENT, 1999). A literatura recomenda que o número de classes em um mapa não seja superior a 10, seja considerando a escala acromática ou cromática. Dent *et al.* (2009) sugerem o uso não mais que cinco classes para um mapa temático quantitativo. Outros trabalhos na literatura, tais como Slocum e Egbert (1993), Cromley e Ye (2006), também têm realizado a elaboração de mapa coroplético adotando-se 5 classes para agrupar dados.

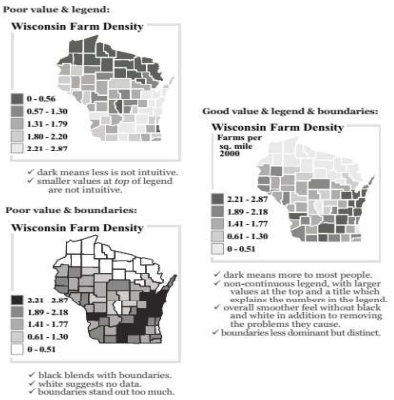
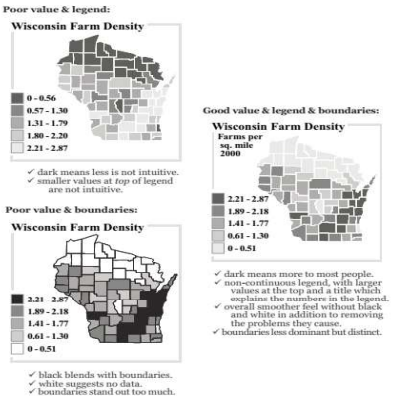
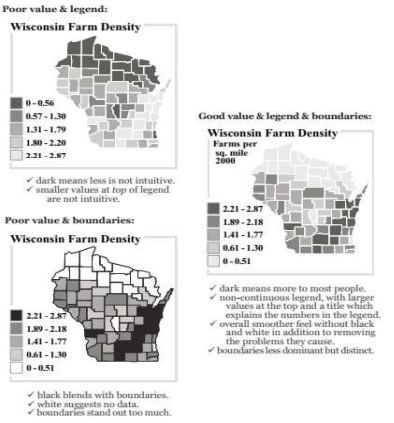
Segundo Krygier e Wood (2005), tão importante quanto a definição do número de classes é o projeto gráfico das representações. Para ilustrar essa afirmação, tais autores apresentam três exemplos de mapas coropléticos, elaborados com cinco classes de representação, a partir do uso da variável visual valor (Quadro 1). Krygier e Wood (2005) afirmam que dois desses três mapas possuem problemas de projeto, ao que se refere à representação da legenda, ou à simbolização das feições, e propõem solução para resolver esse problema de comunicação cartográfica.

2.4 Aplicação dos métodos de classificação

A classificação de dados pode ocorrer por diferentes métodos, e a maioria estão disponíveis nos Sistema de Informação Geográfica. Nesse trabalho, o *software* utilizado foi o da empresa *ESRI, ArcGIS* 10.0, e os métodos de classificação selecionados foram Intervalo igual, Quantil, desvio-padrão e Otimização de *Jenks*. Para cada método, produziram-se os mapas coropléticos com 5 classes.

O processo de classificação se inicia

Quadro 1: Considerações de projeto gráfico para mapa coroplético

<u>Legenda</u> Problemas: - Uso de valores escuros para as áreas com valores de atributos menores (ex.: 0 - 0,56) não é intuitivo para o usuário do mapa; -Valores menores no início da legenda podem não ser uma representação intuitiva para o usuário.	
<u>Representação das regiões</u> Problemas: - Regiões representadas com a borda em preto se destacam muito no mapa, resultado em uma representação visual densa; - A representação das áreas em branco sugere a ausência de dados, mas não é o que a legenda representa (0 – 0,51).	
<u>Solução</u> - Regiões mais escuras representam mais quantidades; - Valores maiores no início da legenda, torna a representação mais intuitiva para o usuário; - Mapa se mostra com aparência mais harmoniosa, sem o uso do preto e do branco. Além disso, sem o uso do branco, pode-se resolver o problema da ausência de dados; - Bordas das regiões se apresentam menos dominante no mapa, mais distintas.	

Fonte: Adaptado de Krygier e Wood (2005).

com a ordenação crescente dos dados, independentemente do método selecionado. A função de um método de classificação se resume em determinar a amplitude do intervalo de cada classe, sendo o que diferencia um método de outro é o critério usado para estimar essas amplitudes (JENKS & COULSON, 1963 apud DENT *et al.,* 2009). Nas seções seguintes, são descritas as características dos métodos, as etapas de classificação envolvidas em cada um e os mapas coropléticos resultantes da aplicação.

2.4.1 Método do Intervalo igual

A característica da classificação por Intervalo igual consiste em que as classes são definidas com amplitude constante, isto é, as classes têm faixa de valores iguais. Portanto, o que pode variar de uma classe para outra é o

número de feições dentro de cada uma. Este tipo de classificação se destaca pela simplicidade de cálculo, facilidade de interpretação dos resultados e legenda de fácil compreensão por ser formada por valores ordenados e contínuos (SLOCUM *et al.,* 2009). Uma desvantagem da classificação por Intervalo igual deve-se ao fato dos intervalos das classes não serem definidos com base na distribuição dos dados. Por isso, essa técnica não é aconselhada para agrupar dados com distribuição assimétrica, em que as observações são concentradas em uma ou duas classes apenas. Outra desvantagem é a possibilidade de se gerar classes vazias, sem observações (SLOCUM *et al.,* 2009). Classes vazias implicam que mapa e legenda não estão sincronizados, isto é, há classe na legenda que pode não estar simbolizada no mapa, e Dent *et al.* (2009) sugerem trabalhar com a amplitude útil para evitar esse problema.

As etapas que compõem a classificação por Intervalo igual se resumem na ordenação dos dados, no cálculo da amplitude da amostra, na estimativa do intervalo das classes (h) e na definição dos limites, inferior e superior, de cada classe. O intervalo das classes é determinado pela razão entre a amplitude do conjunto (A) e o número de classes (K) (DENT *et al.,* 2009; SLOCUM *et al.,* 2009). A partir dessas etapas, fez-se a classificação dos dois conjuntos dos estudos de caso, SP e MS, para um número de classes igual a 5. O intervalo de classes obtido foi igual a 2.506,66 hab./km^2 e 19,30 hab./km^2 para São Paulo e Mato Grosso do Sul, respectivamente. Os resultados da distribuição dos dados nessas cinco classes, considerando esses intervalos de classes, encontram-se na Figura 3.

Verifica-se (Fig. 3) que as observações são fortemente concentradas em uma única classe, tanto com relação ao conjunto de municípios paulistas quanto sul-mato-grossenses. No caso de SP (Fig. 3B), 97% dos municípios foram agrupados na primeira classe, a qual possui densidade de até 2.510,36 hab./km^2. No caso do MS (Fig. 3A), 92% dos dados se concentraram na classe com densidade até 19,93 hab./km^2. Esses resultados decorrem dos limites das classes serem definidos sem considerar a distribuição de frequência dos dados. Associado a essa característica do método, outro fator é o fato de ambos os estados (SP e MS) apresentam dados com distribuição em forma de curva leptocúrtica, como discutido na seção de análise de estatística descritiva. Ainda na Figura 3, nota-se que há duas quebras entre grupos similares, ao definir o limite superior da primeira (19,93) e da terceira (58,60) classe nos dados de MS (Fig. 3A), o que resulta na distribuição de dados similares em classes distintas.

A partir desses resultados, e com a finalidade de melhorar a distribuição das observações nas cinco classes de representação, propôs-se uma nova classificação para os dados considerando a amplitude útil dos conjuntos (Fig. 4). Agruparam-se em uma única classe os municípios com densidade acima de 20,2 hab./ km^2 e 1869,40 hab./ km^2, casos de MS e SP, respectivamente. Os mapas resultantes da aplicação do método de Intervalo igual, para a amplitude total e amplitude útil, encontram-se nas Figuras 5 e 6. Observa-se que ao utilizar a amplitude útil há uma melhor distribuição dos dados nas classes, principalmente para o caso do Estado de MS (Fig. 4A), favorecendo para a análise dos padrões de distribuição da densidade demográfica nos municípios.

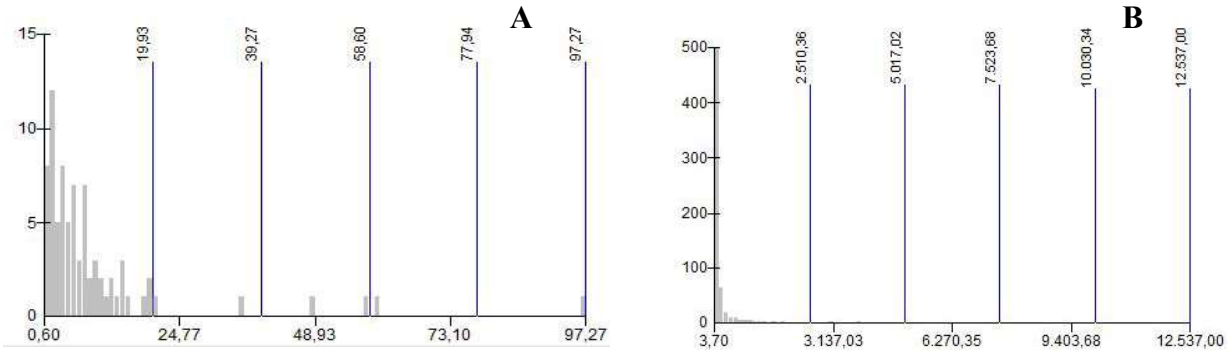
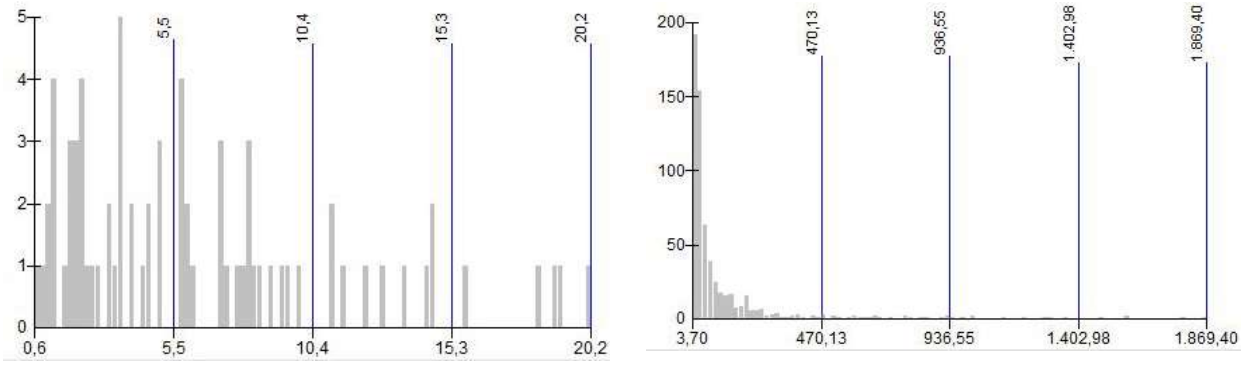
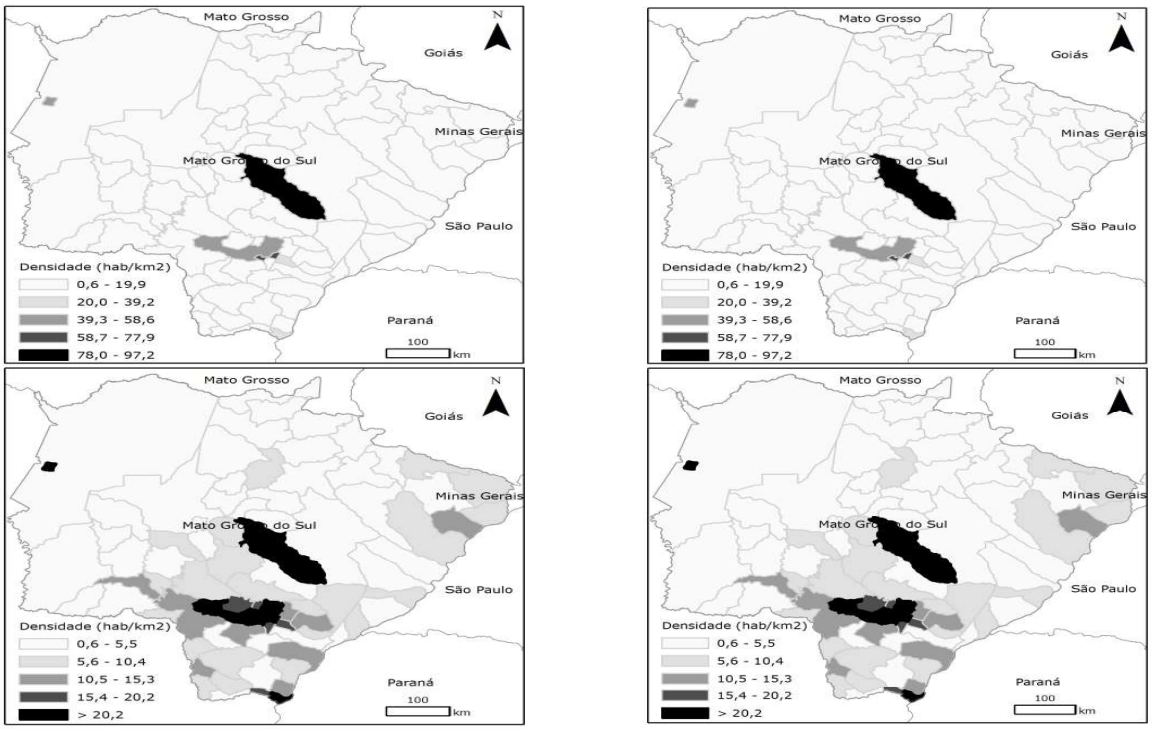


Fig. 3 – Distribuição das observações nas cinco classes de representação. Em (a) dados do Estado de Mato Grosso do Sul e, em (b), dados do Estado de São Paulo.



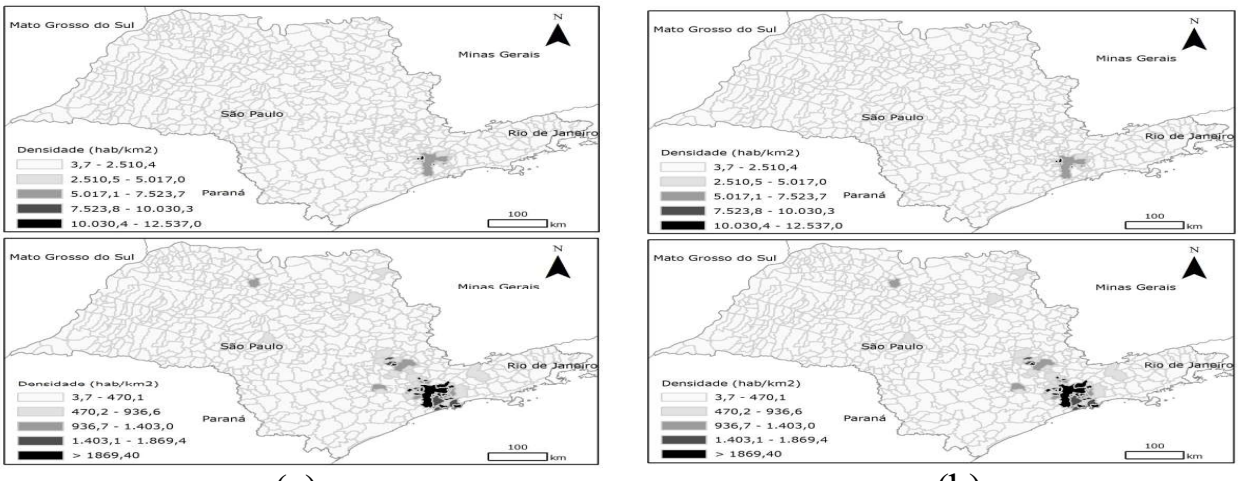
(a) (b)

Fig. 4 – Distribuição das observações nas cinco classes de representação considerando a amplitude útil do conjunto. Em (a) dados do Estado de Mato Grosso do Sul e, em (b), dados do Estado de São Paulo.



(a) (b)

Fig. 5 – Classificação da densidade demográfica do Estado de Mato Grosso do Sul usando o método de Intervalo igual. Em (a) classificação pela amplitude total do conjunto e, em (b), pela amplitude útil.



(a) (b)

Fig. 6 – Classificação da densidade demográfica do Estado de São Paulo usando o método de Intervalo igual. Em (a) classificação pela amplitude total do conjunto e, em (b), pela amplitude útil.

2.4.2 Método do Quantil

A classificação pelo método do Quantil, propõe que cada classe possua o mesmo número de elementos. Assim, o que pode variar de uma classe para outra é a amplitude (SLOCUM *et al.* 2009). As características positivas desse método são enfatizar diferenças no meio da faixa de valores e também não gerar classes vazias. Todavia, feições similares podem ser classificadas em classes diferentes, bem como feições dissimilares classificadas em uma mesma classe (CARVALHO, 2005). Denomina-se esse método de acordo com a quantidade de classes do mapa, tal como quartil é o mapa formados por quatro classes, quintil quando há cinco classes de representação, percentil para 100 classes no mapa, e assim sucessivamente (DENT *et al.,* 2009).

Na classificação por Quantil, inicia-se o processo pela definição do número de elementos (q) de cada classe. Para tanto, faz-se a ordenação crescente dos dados e a divisão do número de observações (N) pelo número de classes (k). Seja k = 5 (cinco classes de representação), nos estudos de caso, obteve-se: q_{ms} = (N/k = 79 / 5 = 15,8) e q_{sp} = (N/k = 645 / 5 = 129). Para divisões fracionadas, a primeira classe irá conter um elemento a menos comparada às demais classes (SLOCUM *et al.,* 2009). No caso do conjunto de MS, a primeira classe representará a densidade demográfica de 15 municípios, enquanto as quatro outras classes 16 municípios, o que garante que todos os 79 municípios sejam classificados em uma das 5 classes. Para o estado de SP, a divisão é exata, logo as 5 classes irão conter 129 elementos cada.

A segunda etapa da classificação por Quantil é a definição dos limites das classes. A primeira classe se inicia com o primeiro valor ordenado da amostra (valor mínimo) e termina com o último elemento sendo aquele que representar a quantidade máxima de elementos na classe. A segunda terá como limite inferior o próximo elemento ordenado da amostra e o limite superior feito como anteriormente, e assim sucessivamente. Slocum *et al.* (2009) ressaltam que o problema dessa abordagem é nos casos em que os elementos amostrais de igual valor são distribuídos em classes distintas, tal como dois ou mais municípios com a mesma densidade demográfica podem ser simbolizados em mais de uma classe. Isso gera uma classificação sem sentido, ambígua, e, para esses casos, Slocum *et al.* (2009) propõem a definição dos limites das classes pelo uso da média aritmética entre o último valor da classe anterior e o primeiro valor da próxima classe. Por exemplo, o limite inferior da classe 2 será a média aritmética do último valor da classe 1 e o primeiro valor da classe 2, e o limite superior dessa classe será a média aritmética entre seu último valor e o primeiro valor da classe 3, e assim sucessivamente.

Nas Figuras 7 e 8 estão apresentados os resultados da classificação da densidade demográfica dos estados de SP e MS, com 5 classes de representação. Fez-se a classificação também considerando a amplitude útil do conjunto (Fig. 7B e 8B). Deve-se ressaltar que para a classificação pela amplitude útil, a quantidade de elementos em cada classe foi calculada para o número de classes igual a 4, portanto, SP apresenta 645/4 = 156 elementos, e MS 79/4 = 19,75.

O método Quantil tem característica de desconsiderar a semelhança entre as observações ao agrupar os elementos, e de ser simples matematicamente, por requer uma operação de divisão apenas. Uma grande diferença entre esse método e dos Intervalos Iguais, deve-se ao fato do Quantil poder gerar descontinuidades (*gaps*) na legenda, ou seja, a possibilidade de ocorrer 'vazios' de uma classe para outra. Isso pode trazer dificuldades para o usuário interpretar a legenda, por esta não ser intuitiva. Slocum *et al.* (2009) ressaltam que se pode criar legendas sem descontinuidades para qualquer método de classificação, contudo ressaltam que essa abordagem não indicará a variação real dos valores em cada classe. O *software ArcGIS* adota por *default* esse esquema de representação da legenda, e por essa razão não se constatam descontinuidades nos limites das classes (ex. Fig. 7 e 8). A classificação pelo Quantil é recomendada para conjuntos que apresentem distribuição retangular, com valores não muito repetitivos na amostra (SLOCUM *et al.* 2009).

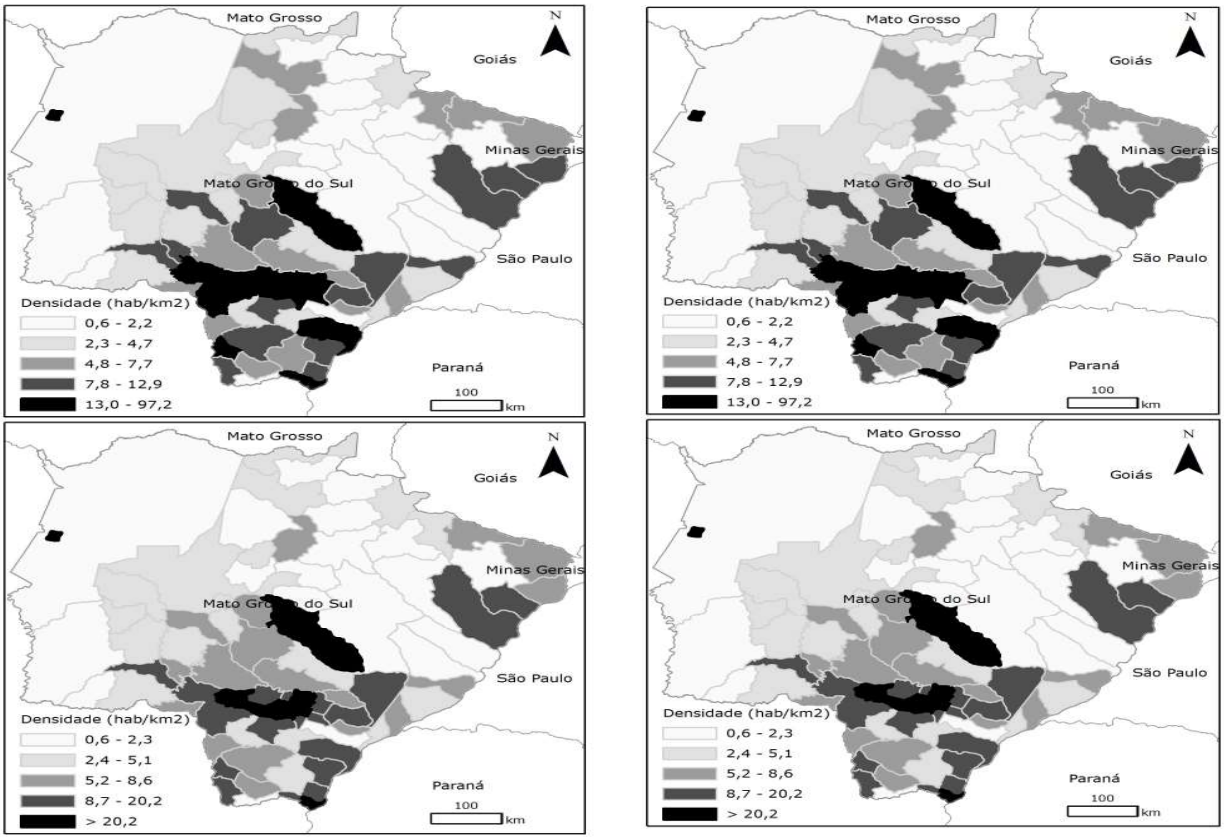


Fig. 7 – Classificação da densidade demográfica do Estado de Mato Grosso do Sul usando o método do Quantil. Em (a) classificação pela amplitude total do conjunto e, em (b), pela amplitude útil.

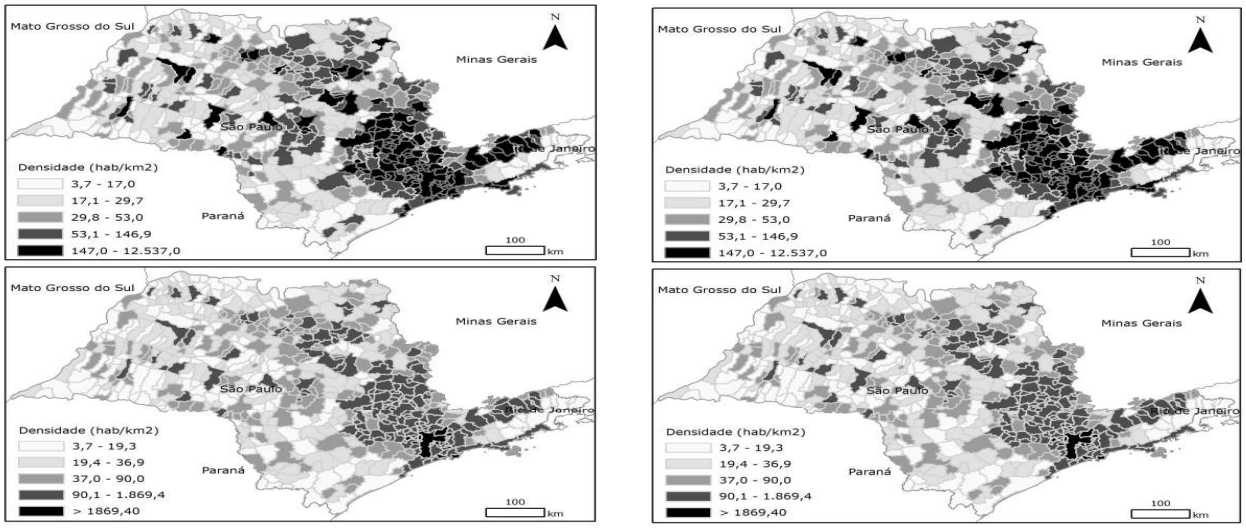


Fig. 8 – Classificação da densidade demográfica do Estado de São Paulo usando o método do Quantil. Em (a) classificação pela amplitude total do conjunto e, em (b), pela amplitude útil.

2.4.3 Método do Desvio-padrão

O princípio da classificação pelo método de Desvio-padrão consiste em apresentar a distribuição dos dados acima e abaixo da média amostral. Esse método é adequado na análise de dados, tais como precipitação, renda, temperatura, altitude, pois, nesses casos, a média é um estimador empregado para a comparação entre diferentes localidades (DENT *et al.,* 2009; SLOCUM *et al.,* 2009).

As etapas da classificação se constituem em determinar a média (Equação 1) e o desvio-

padrão das observações (Equação 2) e, em seguida, definir os limites do número de classes pré-determinado. Estabelece-se os limites das classes a partir da adição ou subtração repetitiva do Desvio-padrão (σ) (Equação 2) dos dados (χ_i) da média aritmética (μ) do conjunto (Equação 1). N é o número de elementos da amostra.

$$\mu = \frac{\sum_{i=1}^{N} \chi_i}{N} \quad (1)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (\chi_i - \mu)^2}{N-1}} \quad (2)$$

Uma vantagem do método de Desvio-padrão é não produzir *gaps* na legenda. Além disso, não mais que seis classes são necessárias para representar todo o conjunto de dados (DENT, *et al.,* 2009). Em contrapartida, sua desvantagem é a exigência de distribuição normal para os dados. Somente neste caso, a média representa um indicador para a amostra, pois representa um ponto de divisão que permite uma análise coerente de contraste entre os dados acima e abaixo desta (SLOCUM *et al.,* 2009). Uma alternativa para se obter a normalidade dos dados e desenvolver a classificação pelo Desvio-padrão é aplicar a função de Box-Cox, a qual transforma amostras com uma distribuição normal (JOHNSON & WICHERN, 2007). No entanto, o uso desse tipo de transformação nem sempre representa uma solução vantajosa, pois muitas vezes, o interesse é examinar os dados brutos. Outra desvantagem do método de Desvio-padrão é a exigência de se conhecer conceitos de estatística para que se possa interpretar a legenda (SLOCUM *et al.,* 2009).

Para os casos de estudo, a análise de estatística descritiva (Tabela 1) mostrou que os dados de São Paulo e Mato Grosso do Sul não apresentam distribuição normal. Portanto, pode-se assumir que o método de Desvio-padrão não é o mais apropriado para o agrupamento dos dados neste caso. As Figuras 9 e 10 apresentam o resultado da classificação. Para os dados de SP quatro classes foram suficientes para compreender todo o conjunto amostral.

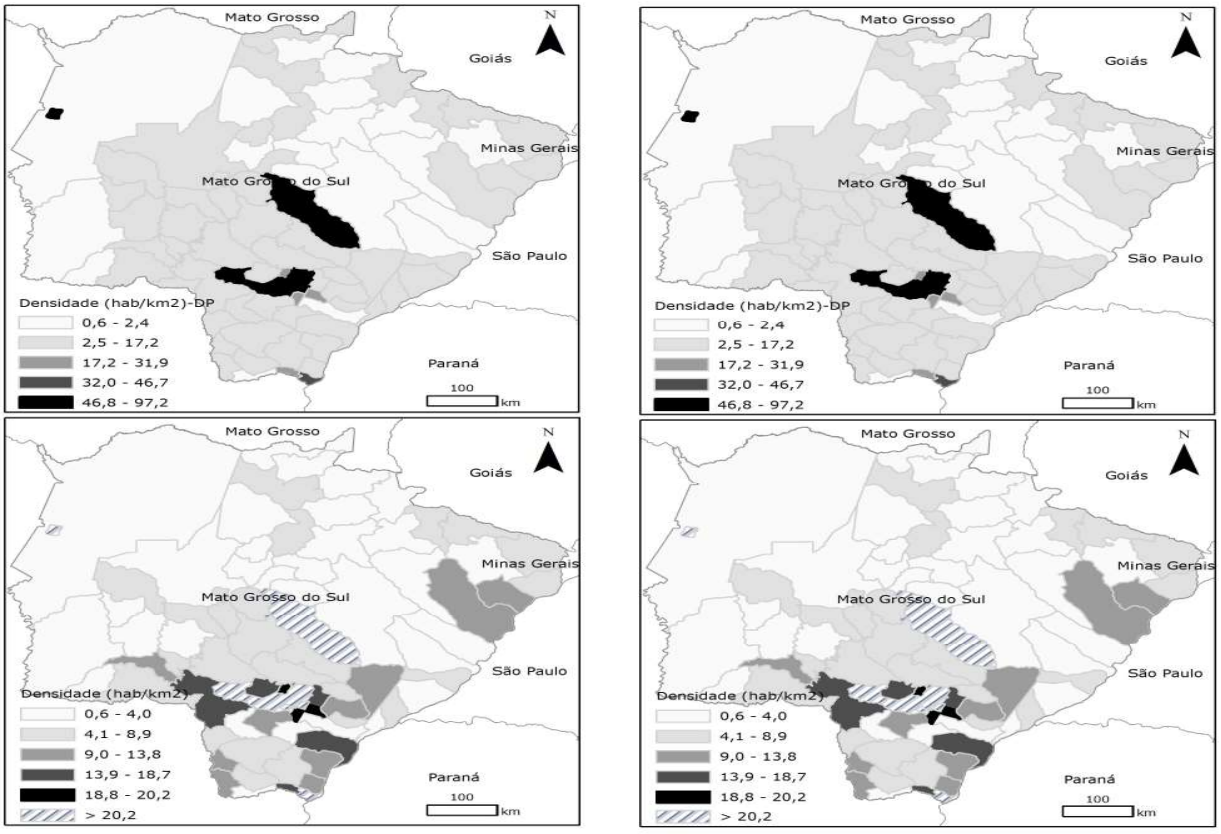


Fig. 9 – Classificação da densidade demográfica do Estado de Mato Grosso do Sul usando o método do Desvio-padrão. Em (a) classificação pela amplitude total do conjunto e, em (b), pela amplitude útil.

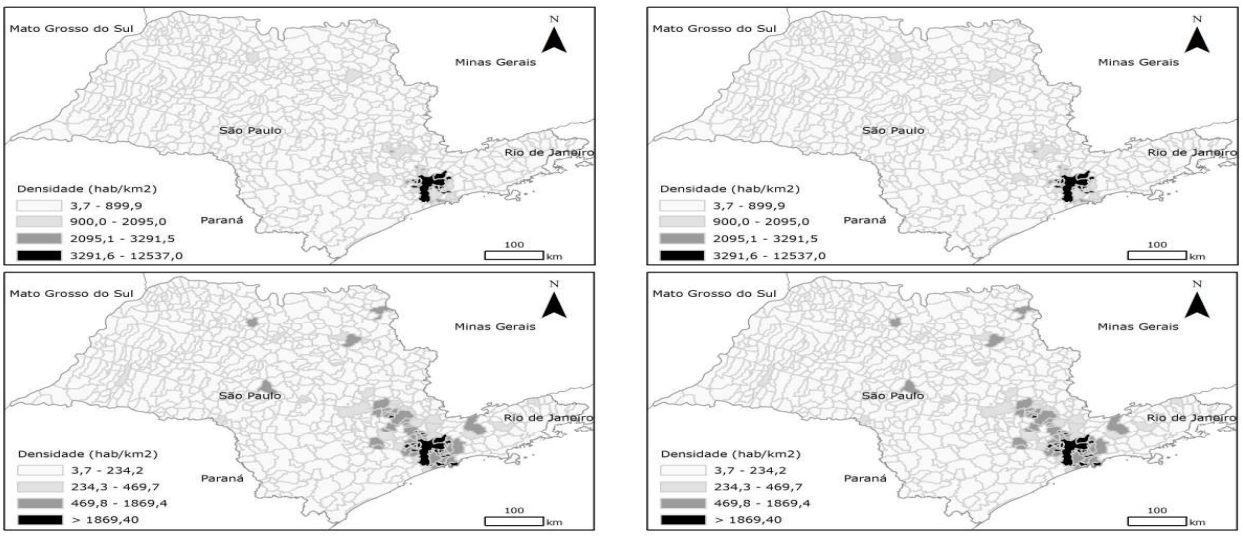


Fig. 10 – Classificação da densidade demográfica do Estado de São Paulo usando o método do Desvio-padrão. Em (a) classificação pela amplitude total do conjunto e, em (b), pela amplitude útil.

2.4.4 Método de Otimização de Jenks

Desenvolvido por *Walter Fisher* (1958) e implementado por *George Jenks* (1977), o método de Otimização de *Jenks* é também conhecido por 'quebras naturais' (DENT *et al*., 2009). O princípio dessa técnica de classificação é minimizar as diferenças entre os valores dispostos na mesma classe e maximizar as diferenças entre as classes (FINN *et al.,* 2006), isto é, formar classes homogêneas internamente, assegurando a heterogeneidade entre essas (DENT *et al*., 2009; SLOCUM *et al.,* 2009).

O processo de classificação se inicia com a ordenação crescente dos dados, seguida da construção do histograma de frequências, o qual auxilia na identificação dos possíveis agrupamentos. O cálculo dos limites das classes se baseia na estimativa de um índice denominado de Ajuste de Bondade ou Melhor Ajuste de Variância (GVF – *Goodness of Variance Fit)*. Este índice é utilizado para quantificar a qualidade da distribuição dos elementos nas classes de acordo com a similaridade entre as observações (SLOCUM *et al.* 2009).

A classificação por Otimização de *Jenks* consiste em um processo iterativo, e, por isso, na primeira interação adota-se o caso ideal, GVF = 1. Isso implica que cada observação do conjunto se encontra em uma classe distinta. Nas demais iterações realizam-se possibilidades de agrupamento entre os dados. À medida que as observações são agrupadas, cada iteração, determina-se um valor de GVF, o qual sempre será menor que 1 (um), sendo considerada a melhor classificação aquela que maximiza o valor do GVF. A classificação se resume em cinco etapas (DENT, 1999, p.148):

- Etapa 1: ordenação crescente dos dados e construção do histograma de frequências;
- Etapa 2: estimativa da média μ das observações (Equação 1);
- Etapa 3: determinação da 'somatória do desvio quadrático de cada observação em relação à média amostral' (*SDAM*) (Equação 3). N é o número de observações, x_i o valor da observação na posição *i* e μ é a média das observações:

$$SDAM = \sum_{i=1}^{N} \left(x_i - \mu\right)^2 \qquad (3)$$

- Etapa 4: definição do limite das classes. Para isso, calculam-se as médias de cada classe (Z_c). Deve-se lembrar que na primeira iteração as médias serão iguais ao número de observações, pois se considera cada observação como sendo uma classe. A partir do $\boldsymbol{Z}_c$, determina-se o somatório do desvio quadrático de cada observação em relação à média das observações contidas na sua respectiva classe' (*SDCM*), (Equação 4). *K* é o número de classes. Na primeira iteração, tem-se que as médias são iguais ao número de observações ($Z_c = x_i$), logo *SDCM* é igual a zero.

$$SDCM = \sum_{c=1}^{K} \sum_{i=1}^{N} \left(x_i - Z_c\right)^2 \quad (4)$$

Etapa 5: estimativa do GVF (Equação 5):

$$GVF = \frac{SDAM - SDCM}{SDAM} \quad (5)$$

As etapas 4 e 5 são repetidas até que o GVF seja maximizado e não haja melhora em seu valor.

Se por um lado a principal vantagem da classificação por Otimização de *Jenks* refere-se ao fato dos limites das classes se ajustarem aos padrões de agrupamento dos dados, por outro as várias etapas e a complexidade matemática desse método representam sua principal desvantagem. Slocum *et al.* (2009) ressaltam que a legenda pode ser descontínua e que podem ser geradas classes com números discrepantes de elementos, representando outros pontos negativos dessa técnica. Segundo esses autores, isso está relacionado à irregularidade da distribuição de frequência dos dados. Mesmo com tais desvantagens, a classificação por Otimização de *Jenks* é considerada apropriada para a maioria das situações de mapeamento coroplético, e tem sido usada como opção padrão nos sistemas de informação geográfica, tal como o *ArcGIS* (DENT *et al*., 2009; SLOCUM *et al.,* 2009). Os mapas da classificação da densidade demográfica de SP e MS, usando 5 classes, encontram-se nas Figuras 11 e 12.

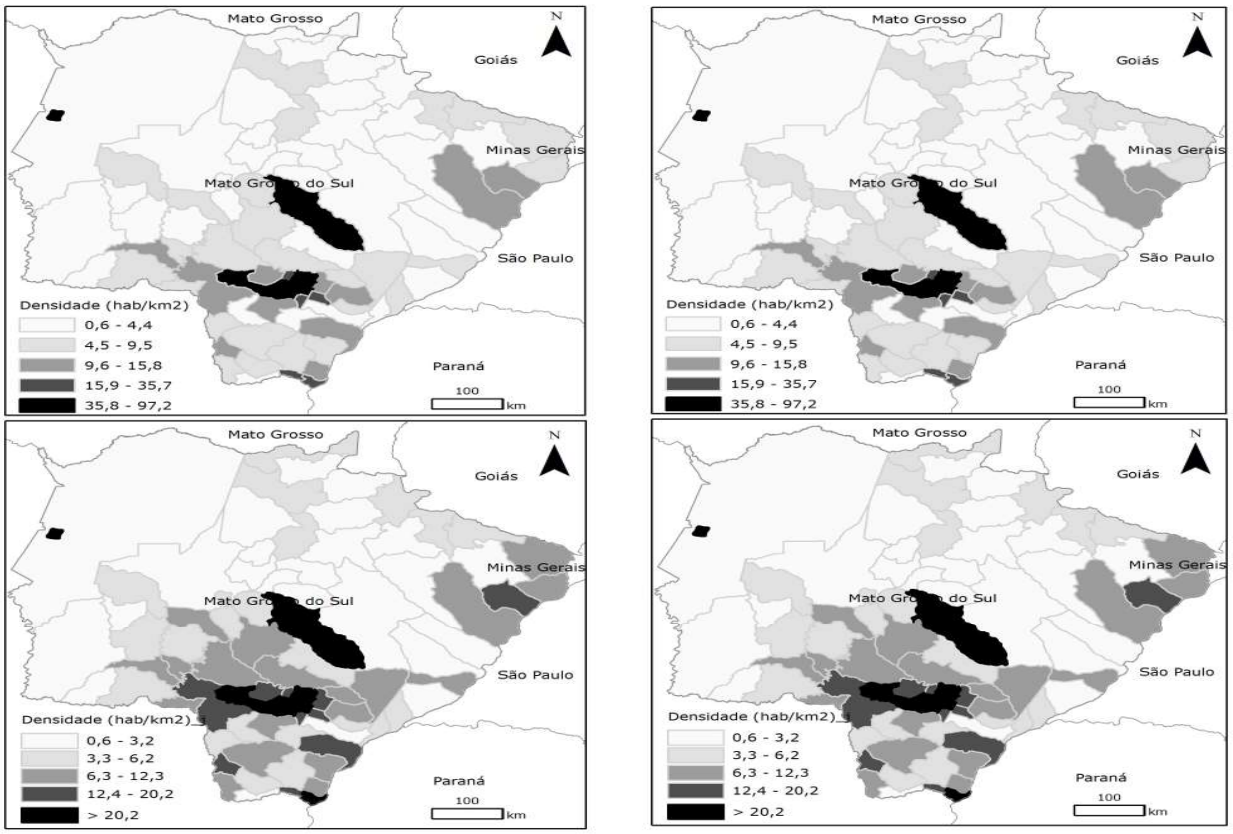


Fig. 11 – Classificação por Quebras naturais da densidade demográfica do Estado do Mato Grosso do Sul. Em (a) classificação pela amplitude total do conjunto e, em (b), pela amplitude útil.

2.5 Avaliação da classificação

O mapeamento coroplético é tema de estudo na Cartografia temática há mais de três décadas, e vários são os estudos encontrados na literatura, nos quais se constata o esforço científico para aprimorar esse método de classificação de dados quantitativos (CROMLEY,1984; LAURANCE & CARSTENSEN, 1984, 1986; SLOCUM & EGBERT, 1993; ANDRIENKO *et al.,* 2001; KRYGIER & WOOD, 2005; CROMLEY & YE, 2005, 2006; DENT et al., 2009; SLOCUM *et al.,* 2009).

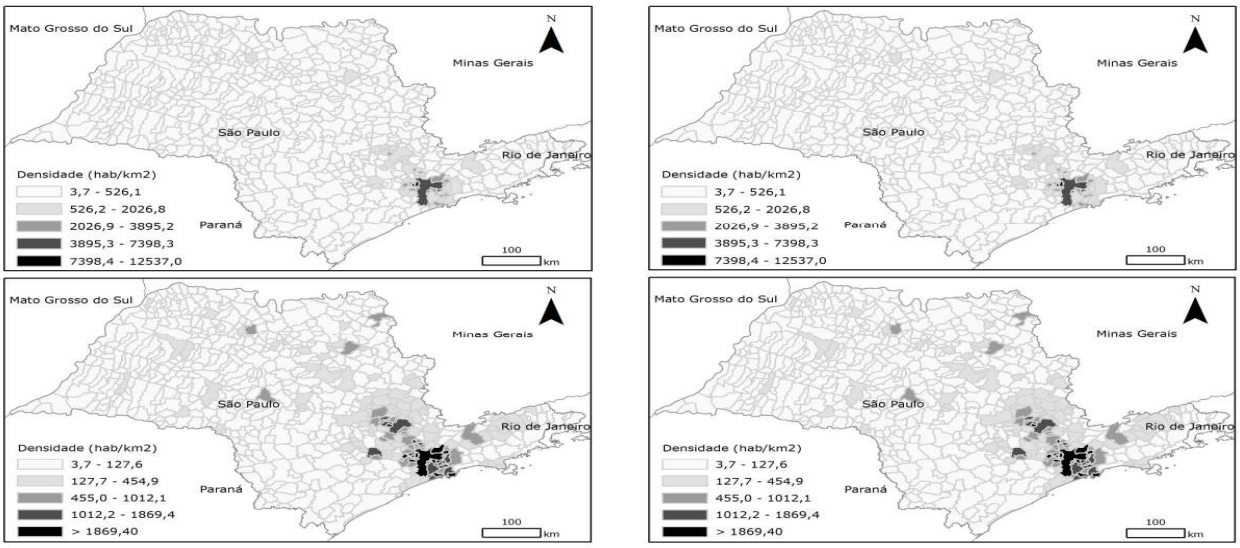


Fig. 12 – Classificação por quebras naturais da densidade demográfica do Estado de São Paulo. Em (a) classificação pela amplitude total do conjunto e, em (b), pela amplitude útil.

A avaliação qualitativa e quantitativa da classificação proposta nesse trabalho é baseada nas discussões apresentadas em Dent *et al*. (2009) e Slocum *et al*. (2009). A avaliação qualitativa consistiu na análise das características de cada método, em função de um conjunto de critérios que representam vantagens ou desvantagens para o método. De acordo com cada critério avaliado, foi atribuído o valor 'sim' ou 'não' ao método (Quadro 2).

Pela análise do Quadro 2, observa-se que nenhum método de classificação atende a todos os critérios simultaneamente, o que se nota são vantagens e desvantagens em cada método, e entende-se que essas devam ser ajustadas ao propósito do usuário e às particularidades do conjunto de dados. Se o propósito do usuário é observar a máxima heterogeneidade entre áreas que possuem valores distintos, entende-se que o método de Otimização de *Jenks* é a opção mais apropriada. Isso em razão dos dados serem agrupados com base em sua distribuição de frequência, de modo a maximizar as diferenças entre as classes e minimizar as diferenças entre os elementos de cada classe. Por outro lado, esse é o método de maior dificuldade de compreensão conceitual e matemática.

Quando o interesse se resume em evitar descontinuidades na legenda e gerar classes com intervalos constantes para facilitar a interpretação do mapa, a classificação por Intervalo igual se destaca como a melhor opção (Quadro 2). Outra possibilidade é o método do Desvio-padrão, contudo este requer a distribuição normal dos dados para que os resultados da classificação sejam coerentes com o fenômeno estudado. Deve-se ressaltar que não se constatou descontinuidades em cada uma das legendas criadas, por cada método de classificação (Fig. 5 a 12), pois o *software ArcGIS* constrói, por *default,* legendas contínuas, independentemente da variação real dos valores em cada classe.

A avaliação quantitativa consistiu da estimativa do melhor ajuste de variância para cada método da classificação aplicado: Intervalo igual, Quantil, Desvio-padrão e Otimização de *Jenks,* e a comparação desses valores. Os resultados encontram-se na Tabela 2. Para o cálculo do GVF considerou-se a amplitude total do conjunto, o número de classes igual a 5 (exceto para o Desvio-padrão) e divisão das classes proposta pelo *ArcGIS,* conforme apresentado na seção anterior.

Para ambos estudos de caso, SP e MS, nota-se (Tabela 2) que o pior GVF foi obtido para a classificação pelo método do Quantil (MS = 0,193; SP = 0,523). Uma das possíveis justificativas para esse resultado é o fato desse método formar classes com igual número de observações e, em razão da alta discrepância da densidade demográfica entre alguns municípios, ocorre que elementos distintos são mantidos dentro de uma mesma classe. Com relação ao método do Desvio-padrão, este apresentou o segundo pior resultado, em termos de maximização do índice GVF. Isto se deve ao fato de que os dados necessitam apresentar uma distribuição de probabilidade normal para serem

adequadamente classificados por essa técnica, o que não foi constatado na análise de estatística descritiva, tanto para os municípios paulistas quanto para os sul-mato-grossenses.

Na classificação do conjunto com menor número de elementos, Estado de Mato Grosso do Sul com 79 municípios, os índices de GVF foram exatamente iguais na classificação por Intervalo igual e Otimização de *Jenks*. Contudo, deve-se lembrar que, no método de Intervalo igual, a classificação ocorre de maneira adequada somente quando os dados descrevem um histograma retangular (SLOCUM *et al.,* 2009), comportamento este não observado na análise descritiva dos conjuntos (SP e MS). Na comparação do GVF para o conjunto mais denso, Estado de São Paulo com 645 municípios, a discrepância entre os índices de GVF aumentou. Neste caso, a melhor maximização do ajuste de variância ocorreu por Otimização de *Jenks* (GVF = 0,898).

Quadro 2: Análise qualitativa da classificação realizada por diferentes métodos

Critério	Método de classificação			
	Intervalos Iguais	Quantil	Desvio-padrão	Otimização de *Jenks*
Utiliza a distribuição dos dados para a definição dos limites das classes	Não	Não	*Sim	Sim
Fácil compreensão conceitual e matemática	Sim	Sim	Sim	Não
Maximiza as diferenças entre as classes e minimiza as diferenças entre os elementos de cada classe	Não	Não	Não	Sim
Produz classes de intervalos constantes	Sim	Não	Sim	Não
Pode produzir classes sem observações	Sim	Não	Não	Não
Pode produzir legenda com descontinuidades (*gaps*)	Não	Sim	Não	Sim
Legenda ordenada e contínua, de fácil compreensão	Sim	Não	Não	Não

*Apenas para dados que com distribuição normal.

Tabela 2: Comparação dos métodos de classificação

Método de classificação	Estado de São Paulo		Estado de Mato Grosso do Sul	
	Intervalos das classes	Melhor Ajuste de Variância - GVF	Intervalos das classes	Melhor Ajuste de Variância - GVF
Intervalo igual	1= [3,7; 2510,4] 2 = [2510,5; 5017,0] 3 = [5017,1; 7523,7] 4 = [7523,8; 10030,3] 5 = [10030,4; 12537,0]	0,898	1= [0,6; 19,9] 2 = [20,0; 39,2] 3 = [39,3; 58,6] 4 = [58,7; 77,9] 5 = [78,0; 97,2]	0,941
Quantil	1= [3,7; 17,0] 2 = [17,1; 29,7] 3 = [29,8; 53,0] 4 = [53,1; 146,9] 5 = [147,0; 12537,0]	0,523	1= [0,6; 2,2] 2 = [2,3; 4,7] 3 = [4,8; 7,7] 4 = [7,8; 12,9] 5 = [13,0; 97,2]	0,193
Desvio padrão	1= [3,7; 899,9] 2 = [900,0; 2095,0] 3 = [2095,1; 3291,5] 4 = [3291,6; 12537,0]	0,880	1= [0,6; 2,4] 2 = [2,5; 17,2] 3 = [17,3; 31,9] 4 = [32,0; 46,7] 5 = [46,8; 97,2]	0,829
Otimização de *Jenks*	1= [3,7; 526,1] 2 = [562,2; 2026,8] 3 = [2026,9; 3895,2] 4 = [3895,3; 7398,3] 5 = [7398,4; 12537,0]	0,898	1= [0,6; 4,4] 2 = [4,5; 9,5] 3 = [9,6; 15,8] 4 = [15,9; 35,7] 5 = [35,8; 97,2]	0,978

CONSIDERAÇÕES FINAIS

Esse trabalho discute as etapas de produção de mapa coroplético por diferentes métodos de classificação e apresenta uma abordagem, qualitativa e quantitativa, para a avaliação da classificação. Conclui-se que o processo de classificação é composto por várias etapas, as quais estão correlacionadas ao método empregado. A partir dos casos de estudo, observou-se que métodos distintos propõem diferentes representações para os dados, o que demonstra a necessidade de se conhecer o propósito do usuário e as características tanto do conjunto amostral quanto de cada método.

Entende-se que o grande desafio da classificação de dados quantitativos é estabelecer classes heterogêneas entre si, compostas por elementos homogêneos. A seleção de cinco classes se mostrou um número apropriado para classificar os dois conjuntos de densidade demográfica, independentemente da variação no número de elementos, pois essa quantidade de classes não realizou um excesso de generalização nos dados e se manteve de acordo com os limites da percepção visual.

Dentre os métodos de classificação estudados, conclui-se que Otimização de *Jenks* foi o método mais apropriado para elaborar os mapas coropléticos da distribuição de densidade demográfica dos estados de São Paulo e Mato Grosso do Sul. Além disso, entende-se que os dois estudos de caso apontam para a importância de se combinar a avaliação quantitativa e qualitativa na seleção do método de classificação, sobretudo, para conjuntos de dados com grande número de amostras.

A análise comparativa das características dos métodos se mostrou uma alternativa para a avaliação do resultado da classificação, em termos qualitativos. Em razão de cada técnica de agrupamento possuir vantagens e desvantagens, conclui-se que não há um método ideal, oportuno para todos os tipos de dados. Isto permite inferir que a seleção do método deve ser realizada com base nas suas características, mas também deve ser acompanhada de outros critérios que podem influenciar no resultado na classificação, tais como a natureza do conjunto amostral, o propósito do mapa e o nível de conhecimento do usuário do produto. Entende-se que a análise quantitativa, pelo melhor ajuste de variância, é uma ferramenta importante e de apoio à análise qualitativa, uma vez que possibilita quantificar matematicamente a eficácia de cada método de classificação. Sendo assim, assume-se que a combinação das análises qualitativas e quantitativas é uma estratégia que aumenta as chances da escolha apropriada do método para a classificação de dados de natureza quantitativa.

Recomenda-se que, para a classificação de conjuntos com disparidade de valores em alguns elementos, deve-se utilizar a amplitude útil do conjunto, evitando-se a amplitude total, sendo uma classe reservada para armazenar os elementos díspares. Nos casos estudados, o mapeamento que utilizou a amplitude útil tornou a classificação mais significativa, pois representou mais coerentemente a distribuição espacial das informações. Contudo, recomenda-se uma análise cuidadosa da distribuição dos dados antes da decisão pela amplitude útil. Para apoiar nessa decisão, entende-se que é de suma importância a etapa de análise de estatística descritiva dos dados, uma vez que quanto mais se conhece a distribuição dos dados, mais eficazes podem ser as decisões tomadas no processo de classificação.

Se por um lado se pode considerar simples a aplicação dos métodos de classificação para a produção de mapas coropléticos, sobretudo, quando apoiada por Sistemas de Informação Geográfica, por outro o resultado desse tipo de mapeamento pode apresentar problemas de acuracidade. Isso se deve ao fato da distribuição do fenômeno geográfico sobre a unidade de área ser considerada uniforme, enquanto, na realidade, pode consistir em uma distribuição desigual. Uma possibilidade para analisar a variação da distribuição de um fenômeno no interior de uma região administrativa é o uso do mapeamento dasimétrico. Sendo assim, sugere-se a elaboração de mapas dasimétricos com os dados de densidade demográfica, dos estados de SP e MS, e se recomenda a comparação desses produtos cartográficos. Com isso, pode-se responder às questões do tipo: 'para o mapeamento da densidade demográfica, qual é o mais eficaz, eficiente e de maior aceitação entre os usuários: mapa coroplético ou mapa dasimétrico?', ou 'para apoiar as ações de políticas públicas, em termos de controle da

expansão urbana, os gestores deveriam utilizar como ferramenta cartográfica o mapa coroplético ou mapa dasimétrico?'.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDRIENKO, G.; ANDRIENKO, N.; SAVINOV, A. Choropleth maps: classification revisited. In: *ICC,* Beijing, 2001. **Proceedings**. p.1209-1219, 2001.

BAPTISTA, E. A.; RIGOTTI, J. I. R. Minas Gerais e sua população de deficientes: um estudo a partir dos Censos Demográficos de 2000 e 2010. **Caderno de Geografia**, v. 24, p. 98-118, 2014.

BERTIN, J. **Semiology of Graphics: Diagrams, Networks, Maps**. Madison: University of Wisconsin, 1983. 415p.

CARDIM, M. F. M.; RODAS, L. A. C.; DIBO, M. R.; GUIRADO, M. M.; OLIVEIRA, A. M.; CHIARAVALLOTI NETO, F. Introdução e expansão da Leishmaniose visceral americana em humanos no estado de São Paulo, 1999-2011. **Revista de Saúde Pública**, v. 47, p. 691-700, 2013.

CROMLEY, R. G. A triangulation data structure for choropleth base maps. **The Cartographic Journal** v. 21, p. 19-22. 1984.

CROMLEY, R. G.; YE, Y. An Alternative to Maximum Contrast Symbolization for Classed Choropleth Mapping. **The Cartographic Journal** v. 42, n. 2, p. 137-144. 2005.

CROMLEY, R. G.; YE, Y. Ogive-based Legends for Choropleth Mapping. **Cartography and Geographic Information Systems,** v. 33, n. 4, p. 257-268. 2006. Publicado on-line em 14 Mar 2013.

DENT, B. D. **Cartography: Thematic Map Design**. 3 ed. Dubuque: Wm. C. Brown Publishers, 1999. 417p.

DENT, B. D.; TORGUSON, J.; HODLER, T. **Cartography: Thematic Map Design**. 6 ed. McGraw-Hill, Georgia, 2009. 368p.

DOWNS, R. M.; STEA, D. Cognitive Maps and Spatial Behaviour: Process and Products. *In:* Downs, R. M.; Stea, D. **Image and Environment: Cognitive Mapping and Spatial Behavior.** USA: Aldine Transaction, 2005.

FINN, M.; WILLIAMS, M.; URSEY, L. An Implementation of the Jenks-Caspall Algorithm for Optimal Classification of Data for Geographic Visualization. In: **American Society of Photogrametry and Remote Sensing**. Annual Conference: Reno, New York. 2006. Disponível em: <http://cegis.usgs.gov>. Acesso: 10 Abril 2015.

IBGE. Instituto Brasileiro de Geografia e Estatística. 2015. **Censo demográfico**, Rio de Janeiro, IBGE. Disponível em: < http://www.ibge.gov.br/home/estatistica/populacao/censo2010/default_atlas.shtm>. Acesso em: 09 Maio 2015.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis.** 6. ed. New Jersey: Prentice Hall, 2007. 800p.

JOLY, F. **A Cartografia.** 14 ed. São Paulo: PAPIRUS, 2011. 136p.

KRYGIER, J.; WOOD, D. **Making Maps: A visual guide to map design for GIS**. New York: Guilford Publications. 2005. 303p.

LAURANCE, W.; CARSTENSEN, JR. Perceptions of variable similarity on bivariate choroplethic maps. **The Cartographic Journal** v. 21, p. 23-29. 1984.

LAURANCE, W.; CARSTENSEN, JR. Hypothesis Testing Using Univariate and Bivariate Choropleth Maps. **The American Cartographer.** v. 13, v. 3, p. 231-251. 1986.

LUCENA, K. D. T.; SILVA, A. T. M. C.; MORAES, R. M.; SILVA, C. C.; BEZERRA, I. M. P. Análise espacial da violência doméstica contra a mulher entre os anos de 2002 e 2005 em João Pessoa, Paraíba, Brasil. **Cadernos de Saúde Pública**, v. 28, p. 1111-1121, 2012.

MAIA, A. G. Geografia do Trabalho no Brasil. Confins. **Revista franco-brasileira de geografia**, v. 6, p. 6, 2009.

MARTINELLI, M. **Mapas da Geografia e Cartografia Temática**. 6 ed, São Paulo, Contexto, 2014. 144p.

MARTINELLI, M. & MACHADO-HESS, E. S. Mapas estáticos e dinâmicos, tanto analíticos como de síntese, nos atlas geográficos escolares: a viabilidade metodológica. **Revista Brasileira**

de Cartografia, v. 66, n. 4, p. 899-920, 2014.

MORETTIN, L. G. **Estatística básica**. 2 ed. São Paulo: Pearson Makron Books, 2005. 185p.

PETERSON, G. N. **GIS cartography: A guide to effective map design.** New York: Taylor & Francis, 2009. 224p.

PROCHNOW, R. M.; OLIVEIRA, F. H. & OLIVEIRA, R. A. Potencial dos dados do setor censitário brasileiro aplicado ao marketing de um *fast food delivery*. **Revista Geográfica de América Central**, v. 2, p. 47, 2011.

RAMOS, C. S. **Visualização Cartográfica e Cartografia Multimídia: conceitos e tecnologias.** São Paulo, UNESP, 2005. 179p.

RAZALI, N. M.; WAH, Y. B. Power comparison of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson Darling tests. **Journal of Statistical Modeling and Analytics**, v. 2, n. 1, p. 21-33. 2011.

SIEGEL, S.; JUNIOR, N. J. C. **Estatística não-paramétrica para as ciências do comportamento.** Porto Alegre: Artmed, 2006. 448p.

SLOCUM, A. T., EGBERT, S. L. Knowledge Acquisition from Choropleth Maps. **Cartography and Geographic Information Systems,** v. 20, n. 2, p. 83-95. 1993. Publicado on-line em 14 Mar 2013.

SLOCUM, A. T., MCMASTER, R. B., KESSLER, F. C.; HOWARD, H. H. **Thematic Cartography and Geovisualization.** 3nd ed. Prentice Hall, 2009. 576p.