A propósito de uma técnica de seleção de intervalos de classe para fins de mapeamento

Barbara-Christine Nentwig Silva (*)

Na construção de um mapa estatístico, a seleção de intervalos de classe é de importância fundamental, porque podem ocorrer diversas interpretações dos mesmos dados originais, se diferentes intervalos de classe são utilizados, ou seja, o autor do mapa pode influenciar a interpretação de um fenômeno cuja perspectiva espacial é mostrada no mapa. Segundo Evans (1977 p. 98), o problema de seleção dos intervalos de classe é, freqüentemente, um ramo totalmente anárquico da Cartografia. Parece mesmo, como escreveu Jenks e Coulson (1963, p. 120), que numerosos autores acreditam que os mapas correspondem a uma forma artística, que possibilita muitas liberdades, inadmissíveis, por sua vez, na interpretação verbal ou tabular.

Recomendamos, no artigo "Técnica estatística para agrupamento e mapeamento de informações geográficas" (Bol. Ageteo, 1974) uma técnica que permite classificar de maneira científica e objetiva os dados geográficos para fins de tabulação ou de mapeamento. Para conseguir o agrupamento, cada passo é baseado em técnicas estatísticas, que envolvem necessariamente o cálculo dos quatro primeiros momentos, ou seja, da média da variância, da assimetria e da curtose, para determinar o tipo de distribuição de freqüência dos dados em questão. Comprovada a normalidade da distribuição (ou, por exemplo, a log-normalidade), podemos fazer uma classificação na base do desvio padrão com limites de classe, por exemplo, de $\bar{X}$, $\bar{X} \pm 1s$, $\bar{X} \pm 2s$. Assim, o intervalo de classe é uma função do desvio padrão.

Sabendo que existem muitas outras propostas de classificação de dados para fins de mapeamento (ou, até mesmo, discussões sobre nenhuma classificação), o nosso objetivo neste trabalho é o de avaliar a aplicação da técnica acima mencionada na Geografia e o de apresentar algumas considerações suplementares desta técnica, testada durante os últimos anos com vários tipos de dados.

(*) Instituto de Geociências da Universidade Federal da Bahia, Salvador.

Foi observado constantemente que os dados geográficos, em geral, mostram raramente uma distribuição de freqüência normal, que permita a aplicabilidade rápida e imediata da classificação baseada na média e no desvio padrão. Por outro lado, na Geografia, a distribuição simétrica dos dados é relativamente freqüente na Geografia Física e menos na Geografia Humana e Regional, onde encontramos, muitas vezes, a assimetria positiva dos dados. Assimetria negativa é rara na Geografia.

Nos exemplos onde dados têm uma distribuição normal, a classificação não apresenta nenhum problema. Para esses exemplos, a técnica é ideal. Para variáveis com assimetria, devemos tentar uma transformação para "normalizar" os dados, para depois classificá-los da mesma maneira, só que com os intervalos de classe baseados na média e no desvio padrão dos dados transformados. Os limites de classe dos dados transformados são, finalmente, retransformados em valores originais. Assim, as variáveis que mostram, nos dados reais, normalidade, têm intervalos de classe iguais, ainda que as variáveis transformadas tenham os intervalos de classe constantes só na classificação com dados transformados. A retransformação em unidades originais resulta em intervalos de classe irregulares. Devemos lembrar que os intervalos não precisam ser necessariamente de um desvio padrão, dependendo do objetivo da pesquisa.

É muito importante a escolha eficiente do número de classes, que é uma função do número de observações em consideração. Com poucos dados, menos classes são justificadas; com mais dados um maior número de classes é exigido. A fórmula de Sturges pode ser utilizada para calcular o número recomendável de classes. Segundo ela temos $K = 1+3{,}3 \log n$, onde K é o número de classes e n o número total de observações. Mas, devemos lembrar que experiências comprovaram que com mais de 10 classes o leitor tem dificuldades de distinguir as classes em um mapa. Segundo Evans (1977, p. 100), dentro de uma amplitude de 4 a 10 classes, a decisão sobre o número de classes deveria ser influenciada pelo público a que se destina a carta, pelas técnicas disponíveis e pelo padrão espacial de distribuição do fenômeno.

A tabela 1 mostra classificações com intervalos de classe de um desvio padrão. Juntamos na tabela a percentagem de observações que caem segundo a probabilidade dentro de cada classe, se a distribuição fosse totalmente normal. Os cálculos podem ser feitos na base das tabelas da distribuição normal (por exemplo Spiegel, 1974, p. 562).

Tabela 1. Classificações com intervalos de classe de um desvio padrão

4 CLASSES

Intervalos de classe	*% de ocorrências esperadas*
$\geq$ 1s	15,87
< 1s a $\bar{X}$	34,13
$\bar{X}$ a 1s	34,13
< — 1s	15,87

5 CLASSES

Intervalos de classe	*% de ocorrências esperadas*
< — 1,5 s	6,68
— 1,5 s a — 0,5 s	24,17
— 0,5 s a 0,5 s	38,30
— 0,5 s a 1,5 s	24,17
$\geq$ 1,5 s	6,68

6 CLASSES

Intervalos de classe	*% de ocorrências esperadas*
< — 2 s	2,28
— 2 s a — 1s	13,59
— 1s a $\bar{X}$	34,13
$\bar{X}$ a 1s	34,13
1s a 2s	13,59
$\geq$ 2s	2,28

7 CLASSES

Intervalos de classe	*% de ocorrências esperadas*
— 2,5s a — 1,5s	6,06
< — 2,5s	0,62
— 1,5s a — 0,5s	24,17
— 0,5s a 0,5s	38,30
0,5s a 1,5s	24,17
$\geq$ 2,5s	0,62

8 CLASSES

Intervalos de classe	*% de ocorrências esperadas*
< 3s	0,13
— 3s a — 2s	2,15
— 2s a — 1s	13,59
— 1s a $\bar{X}$	34,13
$\bar{X}$ a 1s	34,13
1s a 2s	13,59
2s a 3s	13,59
$\geq$ 3s	0,13

Destaca-se nesta proposição que a primeira e a última classe devem ser necessariamente abertas. A(s) classe(s) no meio, em torno da média, tem a maior freqüência, que permite ao leitor destacar imediatamente o "normal", o "típico", em relação à pesquisa em questtão. Segundo a tabela 1 observamos que, se temos um número par de classes, a média forma um limite de classe; se o número é impar, a média é um ponto médio. A primeira e a última classe com poucos dados mostram o "extremo", isto é, os valores extremos de um fenômeno. A tabela indica claramente que, segundo o número de classes, a percentagem de observações esperada em cada classe varia bastante.

A tabela 2, por sua vez, fornece uma classificação segundo a mesma técnica, mas na base de 0,5 desvio padrão para diferentes classes.

Tabela 2. Classificações com intervalos de classe de 0,5 desvio padrão

4 CLASSES

Intervalos de classe	*% de ocorrências esperadas*
$< -0{,}5s$	30,85
$-0{,}5s$ a $\bar{X}$	19,15
$\bar{X}$ a $0{,}5s$	19,15
$> 0{,}5s$	30,85

5 CLASSES

Intervalos de classe	*% de ocorrências esperadas*
$< -0{,}75s$	22,66
$-0{,}75s$ a $-0{,}25s$	17,47
$-0{,}25s$ a $0{,}25s$	19,74
$0{,}25s$ a $0{,}75s$	17,47
$> 0{,}75s$	22,66

6 CLASSES

Intervalos de classe	*% de ocorrências esperadas*
$< -1s$	15,87
$-1s$ a $0{,}5s$	14,98
$-0{,}5s$ a $\bar{X}$	19,15
$\bar{X}$ a $0{,}5s$	19,15
$0{,}5s$ a $1s$	14,98
$> 1s$	15,87

7 CLASSES

Intervalos de classe	*% de ocorrências esperadas*
$< -1{,}25s$	10,56
$-1{,}25s$ a $-0{,}75s$	12,10
$-0{,}75s$ a $-0{,}25s$	17,47
$-0{,}25s$ a $0{,}25s$	19,74
$0{,}25s$ a $0{,}75s$	17,47
$0{,}75s$ a $1{,}25s$	12,10
$> 1{,}25s$	10,56

8 CLASSES

Intervalos de classe	*% de ocorrências esperadas*
$> -1{,}5s$	6,68
$-1{,}5s$ a $-1s$	9,19
$-1s$ a $-0{,}5s$	14,98
$-0{,}5s$ a $\bar{X}$	19,15
$\bar{X}$ a $0{,}5s$	19,15
$0{,}5s$ a $1s$	14,98
$1s$ a $1{,}5s$	9,19
$> 1{,}5s$	6,68

Observa-se que no caso de 4 e 5 classes, a primeira e a última classe têm mais freqüência do que as classes intermediárias, eliminando assim a distribuição típica da curva normal, embora considerando que na classificação em 6 e 7 classes a freqüência tende a ser igual em cada classe. Em outras palavras, uma amplitude de 0,5 desvio padrão deixa uma grande amplitude de dados nas classes extremas abertas, suprimindo assim informações que são justamente importantes para a pesquisa geográfica. Por outro lado, classes maiores de um desvio padrão têm um efeito contrário em relação às classes abertas. Elas deixam muito poucas observações na primeira e última classe e a(s) classe(s) do meio é (são) sobrecarregada(s).

Tentativas similares podem ser feitas também com intervalos de classe de, por exemplo, 0,9, 0,8, 0,7, 0,6 desvio padrão, etc., com o objetivo de ver se esses valores dão melhor classificação para uma determinada pesquisa.

APLICAÇÃO

Escolhemos três exemplos do Estado da Bahia, com dados apresentando distribuições de freqüência originalmente diferentes. Aplicamos a técnica para os dados dos 336 municípios da Bahia utilisando, para os três casos, seis classes respectivamente. O intervalo de classe é, em todos os casos, um desvio padrão. Segundo as nossas tentativas, esta classificação deu o melhor resultado para a nossa pesquisa. O programa que calcula os quatro primeiros momentos foi usado em uma calculadora eletrônica Hewlett-Packard 91008.

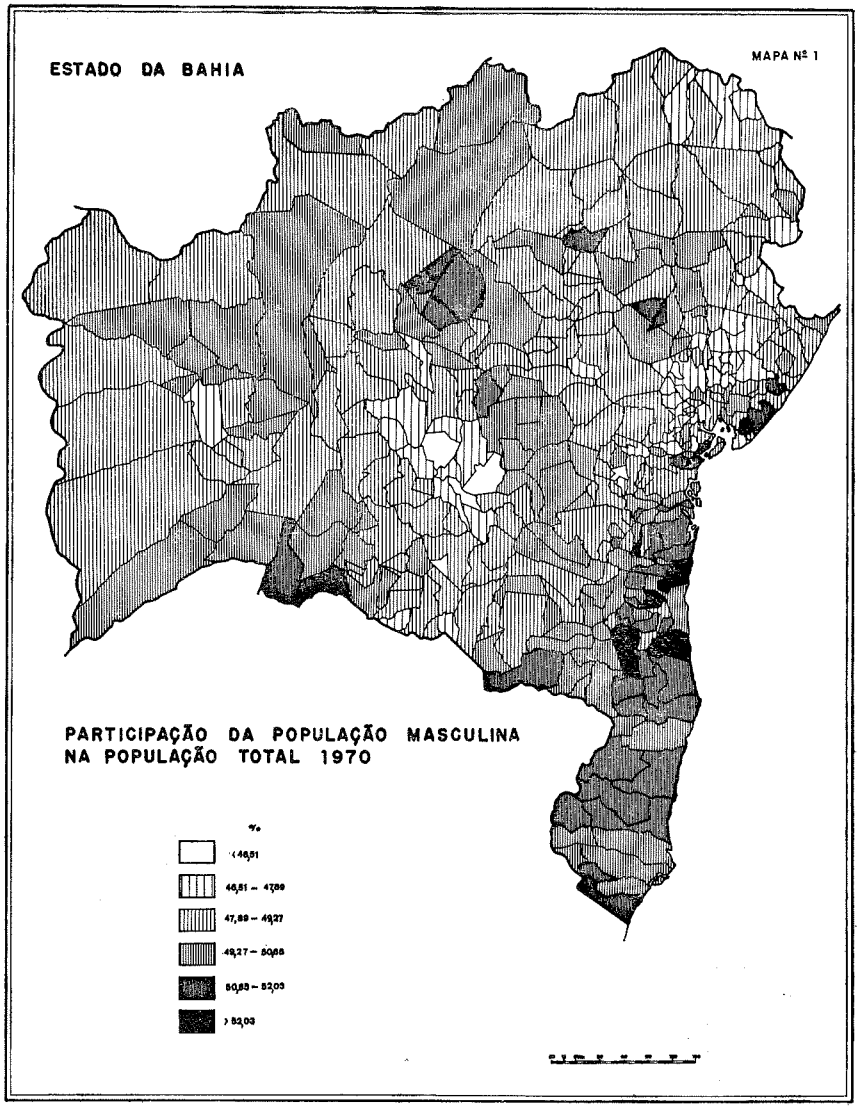
O nosso primeiro exemplo trata da participação da população masculina no Estado da Bahia. Os cálculos da relação de homens/população total (em%), nos municípios baianos, foi efetuado segundo o Censo Demográfico de 1970. A normalidade da distribuição dos dados é satisfatoriamente atingida, como verificamos nas tabelas para testar a assimetria e curtose (Pearson e Hartley, 1970). Considerando que a média é 49,27 e o desvio padrão 1,38, fizemos a seguinte classificação com os intervalos de classe correspondendo a um desvio padrão:

Intervalos de classe

$< 46{,}51$
$46{,}51 - 47{,}89$
$47{,}89 - 49{,}27$
$49{,}27 - 50{,}65$
$50{,}65 - 52{,}03$
$> 52{,}03$

Observa-se que neste caso a amplitude total das classes, com exceção das classes abertas, é constante, ou seja, 1,38. O mapa 1 é baseado nesta classificação.

O nosso segundo exemplo aborda a densidade demográfica (hab./km²) dos 336 municípios da Bahia, tomando também como fonte o Censo Demográfico de 1970. Os dados não têm uma distribuição normal segundo os testes de assimetria e curtose (v.



tab. 3). Assim, é preciso tentar uma transformação através da qual os dados são mudados matematicamente numa forma que se aproxima da curva normal. A transformação logarítmica dá, segundo o nosso controle, uma log-normalidade (v. tab. 3), com limites de confiança de 99%.

TABELA 3. Coeficientes do momento de assimetria e curtose para a densidade demográfica da Bahia — 1970

Transformação	*assimetria*	*curtose*
nenhuma	3,49988 0	19,13130
logarítmica.....	— 0,1100766	0,1172585

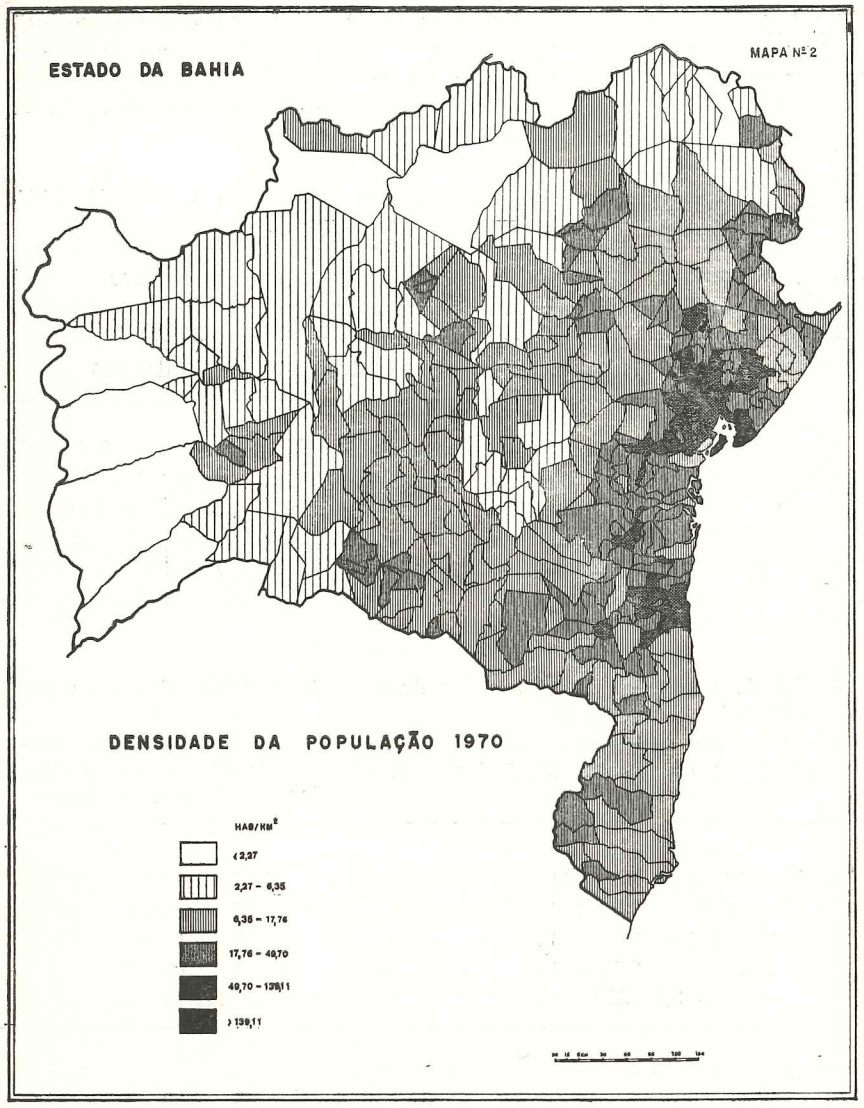
Sendo em valores logarítmicos X = 1,249426 e s = 0,4469622, a classificação em 6 classes, em analogia ao exemplo anterior e com um intervalo de classe de um desvio padrão, dá o resultado da tabela 4. Transformando os limites de classe em antilogarítmos, conseguimos a classificação utilizada no mapa 2. Destacamos que os intervalos de classe em unidades antilogarítmos (originais) não são constantes.

TABELA 4. Classificação dos dados da densidade demográfica da Bahia — 1970

Intervalos de classe em unidades logarítmicas	*Intervalos de classe em unidades antilogarítmicas (originais)*
< 0,355502	< 2,27
0,355502 — 0,802464	2,27 — 6,35
0,802464 — 1,249426	6,35 — 17,76
1,249426 — 1,696388	17,76 — 49,76
1,696388 — 2,143350	49,70 — 139,11
⩾ 2,143350	⩾ 139,11

Os efeitos da transformação logarítmica sobre distribuições com um alto valor de assimetria positiva são muitas vezes satisfatórios. Podemos dizer o mesmo para a curtose, particularmente se se trata de distribuição platicúrticas.

Como terceiro exemplo, escolhemos as taxas geométricas de crescimento anual na década 1960-1970 no Estado da Bahia, baseadas nos cálculos da SEPLANTEC (1976). A distribuição original não é normal como se vê na tabela 5. Foram feitas várias tentativas de transformação que mostramos abaixo com os respectivos valores da assimetria e curtose. Algumas tentativas de



transformação deram resultados insatisfatórios, inclusive a transformação logarítmica que, além disto, não podia ser aplicada imediatamente, como no exemplo anterior, porque vários municípios do Estado têm taxas de crescimento negativas. Neste caso precisou-se adicionar a cada valor de X uma constante para conseguir valores positivos. As transformações com resultados satisfatórios são indicadas na tabela 5.

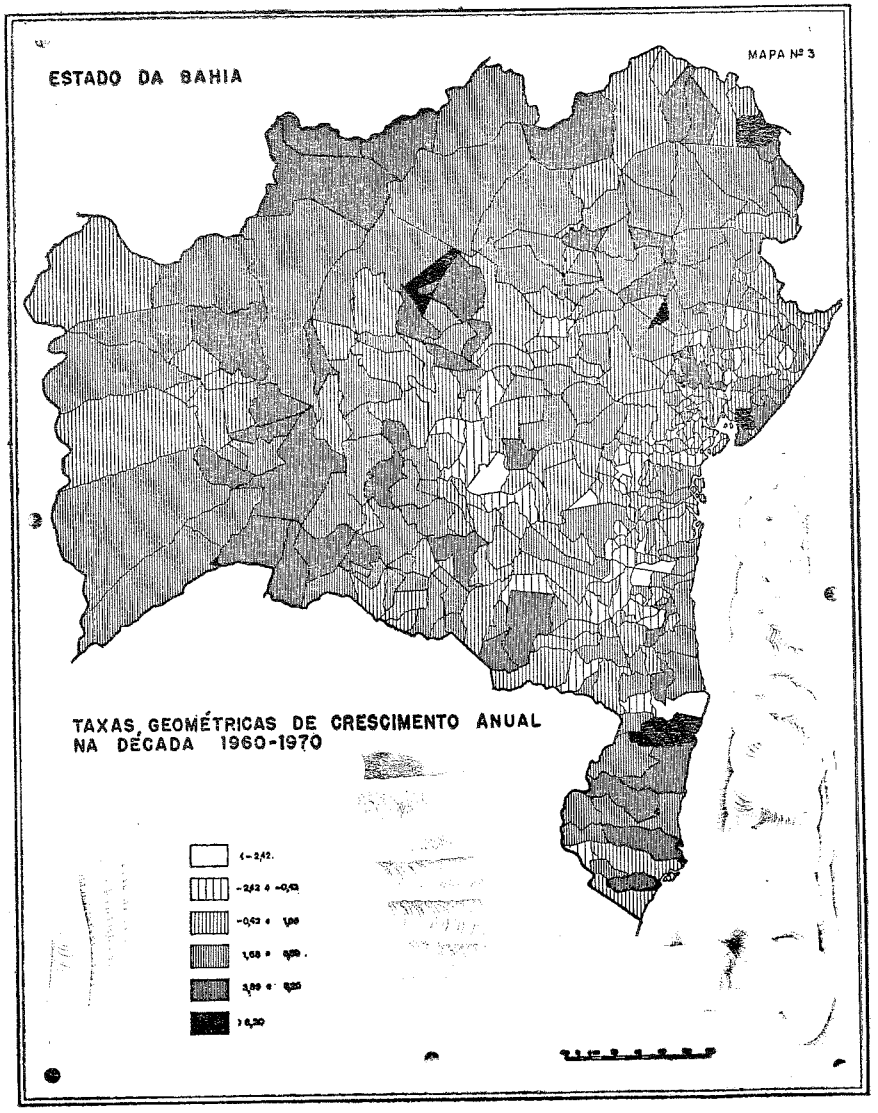
TABELA 5. Coeficientes do momento de assimetria e curtose para taxas geométricas de crescimento anual na Bahia. (1960-1970)

Transformação	*Momentos de assimetria*	*Momentos de curtose*
nenhuma	0,3342628	0,8150159
$\sqrt{X + 20}$	0,1401169	0,6319276
$\sqrt{X + 20}^{1}$	0,2451916	0,6407905
$\sqrt{X + 20}^{1}$	0,2451995	0,6407948

Escolhemos a transformação através de $\sqrt{X + 20}$ que se aproxima mais da normalidade. Com esta transformação o valor de $X = 4{,}656199$ e do desvio padrão é 0,2314052. A classificação com os dados em unidades transformadas ($\sqrt{X + 20}$) e dados transformados em unidades originais está na tabela 6. Aqui, como no exemplo anterior, os intervalos de classe em unidades transformadas têm intervalos de classe constantes, ou seja, de um desvio padrão. A retransformação em unidades originais resulta em amplitudes diferentes. A aplicação prática deste exemplo está no mapa 3.

TABELA 6. Classificação dos dados das taxas geométricas de crescimento anual na Bahia (1960-1970)

Intervalos de classe com dados transformados	*Intervalos de classe em unidades originais*
< 4,193389	< — 2,42
4,193389 a 4,424794	— 2,42 a — 0,42
4,424794 a 4,656199	— 0,42 a 1,68
4,656199 a 4,887604	1,68 a 3,89
4,887604 a 5,119009	3,89 a 6,20
⩾ 5,119009	⩾ 6,20



interessante é observar que os respectivos dados da taxa geométrica de crescimento anual da Bahia para as duas décadas 1940-50 e 1950-60 demonstram um mais alto valor de assimetria e curtose nos dados originais, sendo a década 1940-50 a mais acentuada. Para os dados destas duas décadas do Estado da Bahia tentamos fazer transformações com as mencionadas transformações, e ainda outras, e não conseguimos uma normalidade da distribuição. São os únicos dados geográficos encontrados em numerosas aplicações, que não conseguimos transformar numa distribuição normal, observando-se também que os efeitos da transformação para os dados do período 1950-60 são melhores do que para a década 1940-50. Assim, para estes dados não podemos recomendar a seleção dos intervalos de classe segundo a técnica proposta.

Concluindo, muitos testes mostraram que podemos transformar quase sempre os dados geográficos que não têm originalmente uma distribuição normal numa distribuição deste tipo, permitindo a aplicação da seleção de intervalos de classe segundo a técnica proposta. Normalmente, a transformação logarítmica ou o uso da raiz quadrada dá bom resultado para distribuições com assimetria positiva. São muito raras as variáveis geográficas que não conseguimos "normalizar". Com a grande difusão de pequenas calculadoras programáveis, o teste de assimetria e curtose pode ser feito rapidamente e com facilidade, dispensando o uso de grandes computadores. Comprovada a normalidade dos dados, os intervalos de classe, na base do desvio padrão, permitem uma classificação não só objetiva mas também muito favorável para a interpretação geográfica, porque destaca facilmente as regiões com altos ou baixos valores e as regiões com valores em torno da média. Assim, para a seleção dos intervalos de classe, provavelmente o passo mais importante, para a construção de um mapa estatístico — já que com este passo o autor do mapa controla a interpretação — a técnica em questão se mostrou muito propícia para os dados geográficos.

BIBLIOGRAFIA

Bahia. Seplantec. CPE. 1976. *Estudo do comportamento demográfico e divisão territorial do Estado da Bahia de 1940-1950.* Salvador, 6V.

Evans, I. S. 1977. The selection of class intervals. *Transations. The Institute of British Geographers.* New Series. 2: 98:124.

Jenks, G. F. e Coulson, M. R. 1963. Class intervals for statistical maps. *Int. Yearbook Cartography.* 3:119-134.

Nentwig Silva, B. C.; Galbraith, J. H.; Silva, S. C. Bandeira de Mello E. 1974. Técnica estatística para agrupamento e mapeamento de informações geográficas. *Bol. Geog. Teorética.* Rio Claro, *4* (7/8):29-42.

Nentwig Silva, B. C. 1978. Métodos quantitativos aplicados em Geografia: uma introdução. *Geografia*. Rio Claro, *3* (6):33-73.

Pearson, E. S. e Hartley, H. O. 1970. *Biometrika tables for statisticians*. Cambridge, The Biometrika Trustees at the University Press, V. 1.

Spiegel, M. R. 2974. *Estatística*. Rio de Janeiro, McGraw-Hill,