ESTUDO METODOLÓGICO DE CLASSIFICAÇÃO DE DADOS PARA CARTOGRAFIA TEMÁTICA

CRISTHIANE DA SILVA RAMOS[1]

MIGUEL CEZAR SANCHEZ[2]

Resumo

Dados espaciais apresentados sob forma de tabelas apresentam grande potencial analítico, porém não comunicam dinâmica espacial. Neste sentido é grande a contribuição da cartografia temática. Classificar dados para representação cartográfica resulta em generalização e perda de detalhe. Neste trabalho são analisados uma série de métodos estatísticos de classificação de dados para a construção de cartogramas coropléticos e expostos seus recursos e/ou restrições na construção cartográfica.

Palavras–chave: Classificação de dados; quantificação em geografia; cartografia temática; análise temporal em cartografia.

Abstract

Methodological Study of Classification of Data for Thematic Cartography

Spatial data presented as tables has great analytic potential, even so it don't communicate spatial dynamics. In this sense, the contribution of the thematic cartography is very important. The classification of data for cartographic representation results in generalization and loss of detail. In this work, a series of statistical methods of classification of data for the construction of coroplethic maps are analysed and its resources or restrictions in the cartographic construction are exposed.

Key words: Classification of data; quantification in geography; thematic cartography; temporal analysis in cartography.

[1] Pós-Graduanda em Geografia – IGCE – UNESP – Rio Claro. E-mail: csramos@rc.unesp.br
[2] Professor Doutor do Curso de Pós-Graduação em Geografia – IGCE – UNESP – Rio Claro
Caixa Postal 178 - 13.500-230 - Rio Claro-SP

INTRODUÇÃO

A questão de definição de intervalos de classe para fins de mapeamento é de fundamental importância para a cartografia temática. Entende-se por classificação a subdivisão de um conjunto de dados de acordo com um critério pré-estabelecido.

A classificação visa sempre uma simplificação dos dados objetivando sua melhor compreensão. Para fins de classificação pode-se utilizar desde critérios subjetivos, como o livre arbítrio do pesquisador até fórmulas que determinam o número de classes em uma série de dados e o de intervalo entre classes.

O presente trabalho, visa a análise de diversas técnicas de classificação de dados utilizados em cartografia para o estabelecimento de intervalos de classe. Esta pesquisa faz parte de um projeto mais amplo de desenvolvimento do Atlas de Industrialização do Estado de São Paulo, o qual pretende mapear, por regiões de governo e municípios, e para os anos de 1986, 1990 e 1995, a distribuição espacial da indústria paulista e conta com o apoio financeiro do CNPq.

Foram trabalhados aqui os dados relativos à distribuição espacial de estabelecimentos da Indústria de Material de Transporte, por regiões de governo, para o ano de 1995.

Para o estabelecimento de uma metodologia de classificação onde seja possível, ao mesmo tempo, o mínimo de perda de detalhe e a comparação entre os diferentes cartogramas, surgem, a princípio, três questões a serem consideradas: o número de classes a serem estabelecidas, o intervalo de classe ideal e a possibilidade de estabelecer o nível de generalização ou perda de detalhe.

Existem diversas técnicas para a definição dos intervalos de classe bem como do estabelecimento do número de classes a ser considerado. Estas técnicas variam desde as mais subjetivas, ou seja, que dependem do arbítrio do pesquisador, quanto as mais objetivas, que permitem inclusive a mensuração da perda de detalhe na definição dos intervalos de classe.

MATERIAL

Para o desenvolvimento desta pesquisa foram utilizados:

- Base cartográfica digital do Estado de São Paulo, desenvolvida pela equipe do Projeto Atlas da Industrialização do Estado de São Paulo, a partir da base cartográfica do IBGE, em 1/250.000, de 1976;

- CD-ROM do Relatório Anual de Informações Sociais (RAIS), Ministério do Trabalho, 1995, de onde foram extraídos os dados;
- *Software EXCEL*, versão 7.0, para tabulação dos dados e cálculos;
- *Software Idrisi*, versão 2.0, para processamento e análise dos cartogramas;
- *Software Corel D*raw, versão 8.0, para cartografia digital.

DEFININDO O NÚMERO DE CLASSES

De maneira geral a definição do número de classes a serem consideradas para fins de mapeamento depende do número de elementos da série que se pretende estudar. Segundo GERARDI & SILVA (1981) a existência de mais que 10 classes em um mapa dificulta a sua interpretação, neste sentido, Evans (1977, apud GERARDI & SILVA, 1981) propõe que seja considerado um intervalo entre quatro e dez classes, dependo do público alvo da representação, do comportamento espacial dos dados e dos recursos técnicos disponíveis.

Segundo FERREIRA & SIMÕES (1987) o número de classes em uma série não deve ser menor que cinco nem superior a vinte, por considerarem que um número inferior de classes dificulta comparações entre os valores, e um número superior de classes compromete a visualização do fenômeno mapeado.

MARTINS & DONAIRE (1987) propõem critérios para o estabelecimento de números de classes (k) que levam em consideração o número de elementos que compõem a série de dados (tabela 1).

Tabela 1. Número de classes segundo proposta de MARTINS & DONAIRE (1987, p.78)

Nº de elementos observados	Número de classes (k)	
	Mínimo	Máximo
Até 50	5	10
51 a 100	8	16
101 a 200	10	20
201 a 300	12	24
301 a 500	15	30
Mais de 500	20	40

Pode-se ainda utilizar a fórmula de Sturges, que é definida por:

$k=1+3,33 \log(N)$

Onde:

k é o número de classes

N é o número de elementos observados na série

Existem ainda métodos de classificação que pressupõem um número fixo de classes, como é o caso do quartil, que estabelece quatro classes e do método da Curva de Lorenz, que estabelece duas classes, a dos dados representativos e dos não representativos.

DETERMINANDO A AMPLITUDE DOS DADOS OU *"RANGE"*

O estabelecimento do número de classes em uma série depende basicamente de dois fatores: o número de elementos que compõem a série de dados e a amplitude da série.

A determinação da amplitude da série é feita através da subtração dos valores máximo e mínimo, no caso (tabela 2) 1.241 (valor máximo) e 2 (valor mínimo), o valor resultante, 1139 corresponde à amplitude total. No entanto vale destacar que, de acordo com o comportamento dos dados, nota-se grande diferença entre o primeiro e o segundo valor da série, sendo o primeiro, 1.241 relativo à região de governo de São Paulo quase dez vezes maior que o segundo 137, relativo à região de Campinas. Neste caso optou-se por trabalhar com a amplitude útil dos dados, ou seja, foi reservada uma classe para a região de governo de São Paulo, sendo considerada para fins de cálculo apenas a amplitude útil dos dados, 135, relativo à subtração dos valores das regiões de Campinas (137) e Guaratinguetá (2).

Em outras palavras, a série de dados apresentada na tabela 2 permite o estabelecimento da amplitude total (que é de 1.239) e da amplitude útil (135). Para fins de mapeamento, quando uma série apresenta disparidade de valores em alguns elementos recomenda-se trabalhar sempre com a amplitude útil, reservado-se uma classe especificamente para os elementos que se apresentam díspares, desta maneira pode-se representar satisfatoriamente a distribuição espacial das informações.

DEFININDO OS INTERVALOS DE CLASSE

Existem diversas maneiras de definir intervalos de classe, e muito já se discutiu a esse respeito. Buscou-se, no entanto, uma técnica que fosse ao mesmo tempo simples, objetiva e homogênea, facilitando assim o trabalho da equipe que compõe o Projeto Atlas, visto que o volume de dados a serem mapeados é muito grande.

Visando a aplicação de uma técnica de classificação de dados que permitisse melhor análise em cartogramas, optou-se por estudar algumas técnicas estatísticas mais utilizadas, tais como:

- Desvio Quartílico
- Diagrama de Dispersão
- Análise de Diagrama de Colunas
- Sturges
- Classificação Hierárquica por Pares Recíprocos
- Curva de Lorenz

Para exemplificar a aplicação destas técnicas foram analisados os dados relativos à distribuição dos estabelecimentos da indústria de material de transporte no Estado de São Paulo, por regiões de governo (figura 1) no ano de 1995 (tabela 2).

Desvio Quartílico

O Desvio Quartílico (Quartil) divide a série de dados em quatro grupos com igual número de ocorrências, cada um compreendendo 25% do total de valores apresentados na série (figura 2A). Como a série de dados aqui estudada apresenta grande diferença entre o valor de São Paulo e as demais regiões de governo, foi considerada apenas a amplitude útil para efeito de cálculo dos Quartis. Sendo assim, serão obtidas quatro classes dentro da amplitude útil, uma classe é reservada para São Paulo e uma para as regiões de governo com valores zero, resultando num total de seis classes.

A fórmula para o cálculo do quartil para séries ímpares é a seguinte:

1º Quartil: $(n + 1)/4$

2º Quartil: $(2 * (n+1))/4$

3º Quartil: $(3 * (n + 1))/4$

4º Quartil: $(4 * (n + 1))/4$

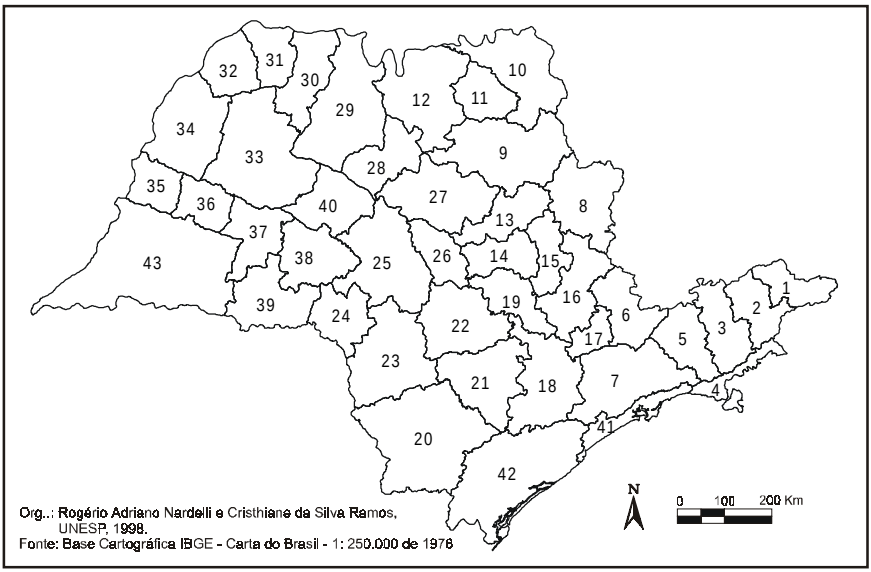
onde n corresponde ao número total de casos na série (excluídas as repetições). Em séries onde o número n é par, não é necessário somar 1.

Tabela 2. Distribuição dos Estabelecimentos da Indústria de Material de Transporte por Regiões de Governo, São Paulo, 1986/1990/1995.

ID	Região de Governo	1986	1990	1995
36	Adamantina	6	4	10
34	Andradina	5	4	7
33	Araçatuba	16	18	23
27	Araraquara	12	11	16
39	Assis	7	8	6
23	Avaré	6	6	6
12	Barretos	5	7	14
25	Bauru	8	12	30
22	Botucatu	6	6	14
6	Bragança Paulista	11	10	27
16	Campinas	46	52	137
4	Caraguatatuba	4	3	8
28	Catanduva	8	7	7
1	Cruzeiro	4	3	8
35	Dracena	5	4	6
31	Fernandópolis	2	2	8
10	Franca	2	2	11
2	Guaratinguetá	0	0	2
21	Itapetininga	2	3	13
20	Itapeva	1	0	7
32	Jales	4	6	6
26	Jaú	12	12	15
17	Jundiaí	7	10	34
15	Limeira	21	23	47
40	Lins	2	2	0
38	Marília	8	7	15
24	Ourinhos	3	4	9
19	Piracicaba	15	19	23
43	Presidente Prudente	24	20	26
42	Registro	1	2	4
9	Ribeirão Preto	25	32	61
14	Rio Claro	7	8	6
41	Santos	47	43	54
13	São Carlos	7	7	13
8	São João da Boa Vista	4	7	15
11	São Joaquim da Barra	4	4	6
29	São José do Rio Preto	18	26	32
5	São José dos Campos	20	20	33
7	São Paulo	873	904	1.241
18	Sorocaba	32	35	62
3	Taubaté	3	8	13
37	Tupã	5	6	5
30	Votuporanga	1	2	13

Fonte: RAIS/Ministério do Trabalho.

Figura 1 - Regiões de Governo do Estado de São Paulo



1	Cruzeiro	12	Barretos	23	Avaré	34	Andradina
2	Guaratinguetá	13	São Carlo	24	Ourinhos	35	Dracena
3	Taubaté	14	Rio Claro	25	Bauru	36	Adamantina
4	Caraguatatuba	15	Limeira	26	Jaú	37	Tupã
5	São José dos Campos	16	Campinas	27	Araraquara	38	Marília
6	Bragança Paulista	17	Jundiaí	28	Catanduva	39	Assis
7	São Paulo	18	Sorocaba	29	São José do Rio Preto	40	Lins
8	São João da Boa Vista	19	Piracicaba	30	Votuporanga	41	Santos
9	Ribeirão Preto	20	Itapeva	31	Fernandópolis	42	Registro
10	Franca	21	Itapetininga	32	Jales	43	Presidente Prudente
11	São Joaquim da Barra	22	Botucatu	33	Araçatuba		

Para a série analisada na tabela 2 chegou-se a um valor N de 25 ocorrências. Desta forma determinou-se a seguinte distribuição quartílica:

1º Quartil – Até o sexto elemento da série

2º Quartil – do sétimo ao décimo terceiro elemento da série

3º Quartil – do décimo terceiro ao décimo nono elemento da série

4º Quartil - do vigésimo ao vigésimo quinto elemento da série

A aplicação dos quartis determinou os seguintes intervalos de classes:

Tabela 3. Intervalos de Classe apontados pela técnica do Quartil.

Classes	Limite mínimo	Limite máximo
1º Quartil	2	8
2º Quartil	9	16
3º Quartil	23	33
4º Quartil	34	137
São Paulo	1241	1.241

A tabela em Excel com os valores da série colocados em ordem crescente e com os respectivos identificadores foi colada no editor de arquivos de valores (.val) do software Idrisi, recurso que diminui bastante o trabalho de digitação e reduz substancialmente a margem de erro. A associação entre o arquivo de valores (.val) e a imagem base região.img, permitiu a criação de um layer temático que foi posteriormente classificado de acordo com os critérios estabelecidos pelo cálculo dos Quartis, resultando no cartograma exposto na figura 2B.

O método de classificação de quartis oferece algumas vantagens, o cálculo é simples e rápido, principalmente considerando-se que o projeto em questão propõe-se a desenvolver um atlas e a quantidade de dados a serem classificados e de cartogramas a serem gerados é muito grande, todavia também oferece desvantagens como destaca SILVA (1978), pois este método não considera a magnitude dos dados, simplesmente considera a posição do dado na série.

Diagrama de Dispersão

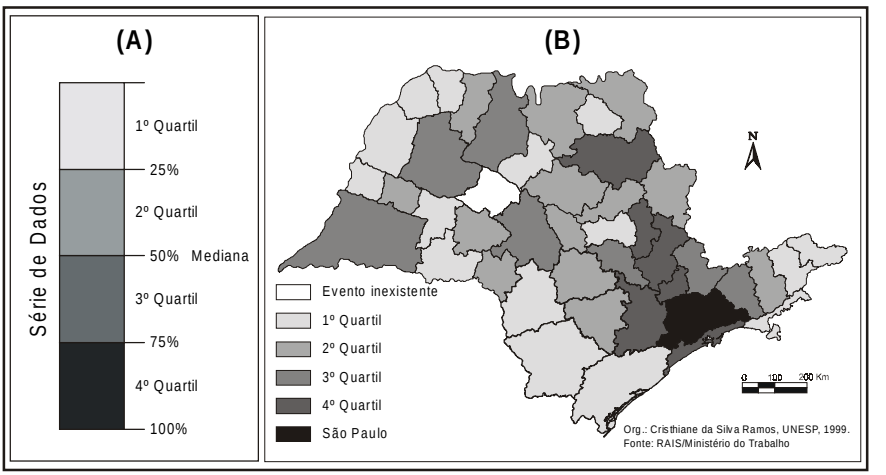
O diagrama de dispersão é um gráfico que permite a visualização do número de vezes em que os dados ocorrem na série. Constitui-se num eixo horizontal em escala, onde são lançados os valores da série, mesmo quando os valores se repetem, observa-se assim quais valores ocorrem com maior freqüência.

Para testar de que forma a análise deste tipo de diagrama pode ser influenciada pelo cuidado do pesquisador em sua confecção, foram elaborados diagramas de dispersão segundo a proposta de construção efetuada em 1978 por SILVA (figura 3A), e em 1991 por MARTINELLI, (figura 4A).

Na figura 3A o diagrama foi desenvolvido sem preocupação com a escala no eixo X, na figura 4A a escala foi considerada.

Observando-se a figura 3A nota-se que os valores mais baixos são os que mais se repetem na série. A classificação dos dados a partir do diagrama de dispersão pode se dar através da observação do pesquisador, delimitando-se as classes

Figura 2 - Representação Gráfica de uma série de dados em quartis (A) e Distribuição dos Estabelecimentos da Indústria de Material de Transporte por Regiões de Governo, São Paulo, 1995. Classificação pelo Método de Quartis (B)



pelas rupturas nos valores de freqüência. As rupturas aqui observadas (tabela 4) resultam no cartograma exposto na figura 3B.

Tabela 4. Intervalos de classe obtidos através da análise do Diagrama de Dispersão (figura 3A).

Classes	Limite mínimo	Limite máximo
1ª. classe	2	5
2ª. classe	6	8
3ª. classe	9	11
4ª. classe	13	23
5ª. classe	26	1241

Considerando-se a proposta de MARTINELLI (1991) foi elaborado o seguinte diagrama de dispersão (figura 4A).

Figura 3 - Diagrama de dispersão construído sem escala no eixo x (A) e Distribuição dos Estabelecimentos da Indústria de Material de Transporte por Regiões de Governo, São Paulo, 1995. Classificação pelo Método do Diagrama de Dispersão (B)

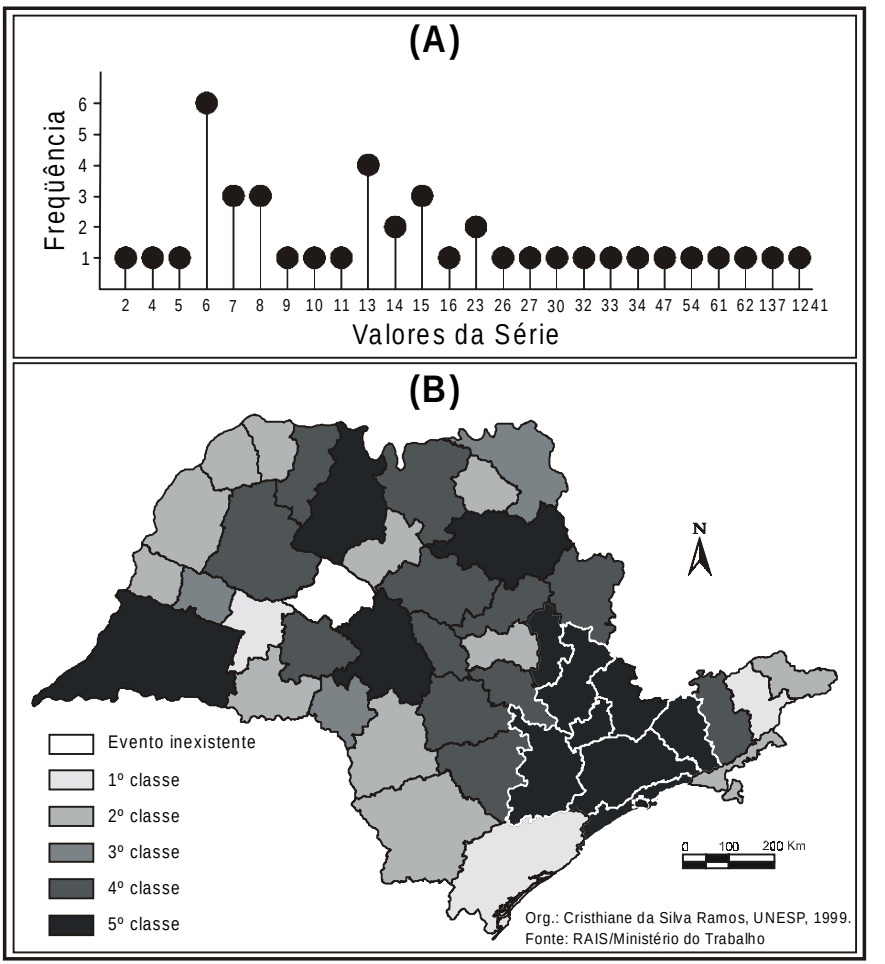
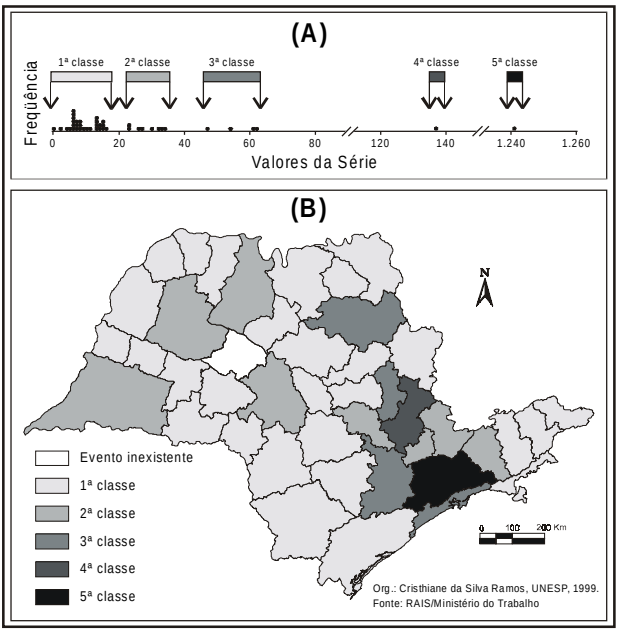


Figura 4 - Diagrama de dispersão construído considerando escala no eixo x (A) e Distribuição dos Estabelecimentos da Indústria de Material de Transporte por Regiões de Governo, São Paulo, 1995. - Classificação pelo Método do Diagrama de Dispersão (B)



A análise da figura 4A permite o estabelecimento dos intervalos de classe apontados na tabela 5, espacializados na figura 4B.

Tabela 5. Intervalos de Classe obtidos através da análise do Diagrama de Dispersão (figura 4A).

Classes	Limite mínimo	Limite Máximo
1ª. classe	2	16
2ª. classe	23	34
3ª. classe	47	62
4ª. classe	137	137
5ª. classe	1241	1241

A classificação através do diagrama de dispersão mostrou-se bastante subjetiva, pois depende da tendência visual de agrupamento dos dados de acordo com a freqüência com que ocorrem na série. Pela comparação dos resultados das análises dos diagramas das figuras 3A e 4A nota-se que o primeiro (figura 3B) foi pouco útil na análise da série apresentada visto que classificou com detalhe os baixos valores e generalizou os altos valores compondo uma classe de grande amplitude (a quinta classe, com amplitude de 1215), mascarando a extrema concentração de estabelecimentos industriais na região metropolitana de São Paulo.

Já a análise da figura 4B permite constatar o inverso, ou seja, fica mais explícita a concentração de estabelecimentos nas regiões de governo de São Paulo e Campinas.

Na construção de ambos os diagramas foram utilizados todos os valores da série, e não a amplitude útil como na técnica do quartil. Isto porque, ao se considerar a escala na construção do diagrama, os valores discrepantes afastam-se dos demais, permitindo assim a definição de uma classe específica (figura 4A).

Desta maneira demonstrou-se que o cuidado na confecção do diagrama de dispersão é fator determinante na análise, que já é complicada pelo fato de ser puramente visual. Também há que se considerar o fato de que a construção de diagramas de dispersão para muitas séries de dados, ou mesmo para séries com grande número de elementos, seria tarefa por demais trabalhosa, o que poderia inviabilizar a pesquisa.

Análise de Diagrama de Colunas

Este tipo de análise compreende a elaboração de um diagrama de colunas para a série de dados; através das rupturas de continuidade apresentadas no gráfico (figura 5A) são estabelecidos os intervalos de classe.

Ao contrário do diagrama de freqüência, este diagrama lança no eixo Y o valor de cada elemento da série e no eixo X o cada um dos elementos.

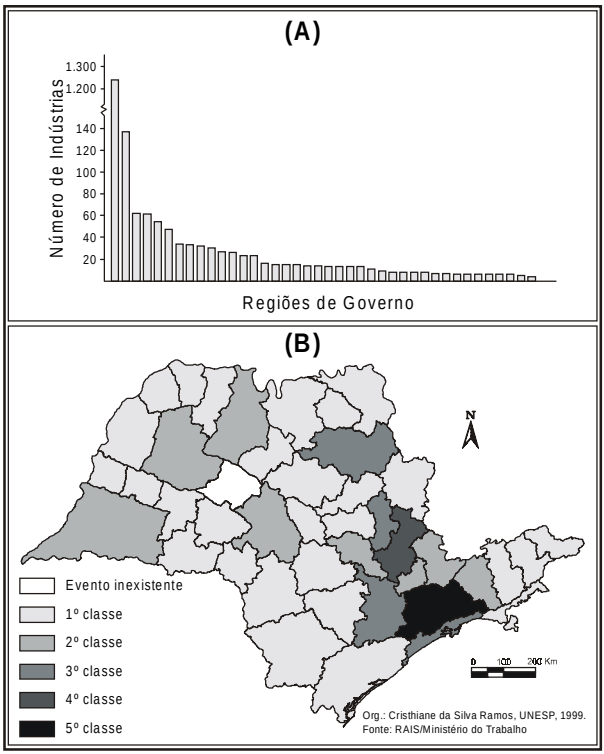
Pela análise das rupturas do diagrama de colunas (figura 5A) chegou-se à distribuição de intervalos de classe exposta na tabela 6.

Tabela 6. Intervalos de Classe obtidos através da análise das rupturas do diagrama de colunas.

Classes	Limite mínimo	Limite máximo
1ª. classe	2	16
2ª. classe	23	34
3ª. classe	47	62
4ª. classe	137	137
5ª. classe	1241	1241

A espacialização dos intervalos de classe obtidos através deste método de classificação resultou no cartograma exposto na figura 5B.

Figura 5 - Diagrama de Colunas (A) e Distribuição dos Estabelecimentos da Indústria de Material de Transporte por Regiões de Governo, São Paulo, 1995. - Classificação pelo Método do Diagrama de Colunas (B)



Este método de classificação é de fácil aplicação, especialmente se o pesquisador utilizar o software EXCEL para a manipulação dos dados, pois ele permite, com relativa facilidade a construção de diagramas de colunas.

Apesar de permitir melhor visualização da concentração dos valores na região metropolitana de São Paulo e adjacências, este método apresenta as mesmas desvantagens do diagrama de dispersão, pois implicaria na confecção de um gráfico para cada série de dados, para posterior análise e classificação, além disso, para cada série poderia haver um número de classes distinto, dependendo da distribuição dos valores, o que acabaria tornando impossível a comparação entre cartogramas.

Da mesma forma, a utilização de uma escala diferente no eixo Y permitiria a visualização de outras rupturas, gerando outros intervalos de classe. Portanto, cabe destacar a extrema subjetividade que este método implica, uma vez que a classificação é puramente visual e depende do arbítrio do observador; duas pessoas analisando o mesmo diagrama, ou mesmo construindo o diagrama de formas diferentes, poderiam estabelecer classificações distintas.

Sturges

Um método estatístico bastante utilizado para a definição do número de classes em uma série é o Método de Sturges, que propõe a aplicação da seguinte fórmula:

$$k = 1 + 3{,}33 * \text{Log } N$$

onde:

k corresponde ao número de classes

N corresponde ao número de elementos da série, da qual, para fins de mapeamento excluem-se os valores repetidos

Aplicando-se a fórmula à série analisada chegou-se ao número de 6 (seis) classes. O intervalo de cada classe é obtido pela divisão da amplitude dos dados pelo número de classes sugerido por Sturges. Considerando-se a extrema concentração de valores na região metropolitana de São Paulo, optou-se por trabalhar apenas com a amplitude útil, reservando-se uma classe para a região de São Paulo.

Desta maneira, dividiu-se a amplitude útil pelo número de classes:

$$h = \frac{amplitude}{k}$$

Onde:

h é o intervalo de classe

amplitude pode ser a amplitude total ou útil, dependendo do comportamento da série de dados

k é o número de classes sugerido por Sturges.

Desta forma chegou-se a um intervalo de classe de amplitude 27, que resultou na seguinte classificação:

Tabela 7. Intervalos de Classe obtidos através do método de Sturges.

Classe	Limite Mínimo		Limite Máximo	
	Valor Absoluto	Observação	Valor Absoluto	Observação
1ª classe	2	-	29 *	-
2ª classe	30	-	56*	-
3ª classe	57*	-	83*	-
4ª classe	84*	Classe vazia	110*	Classe vazia
5ª classe	111*	-	137	-
6ª classe	1241	-	1241	-

* valores inexistentes na série, utilizados apenas como limites de classe.

A aplicação destes intervalos de classe resultou no cartograma exposto na figura 6.

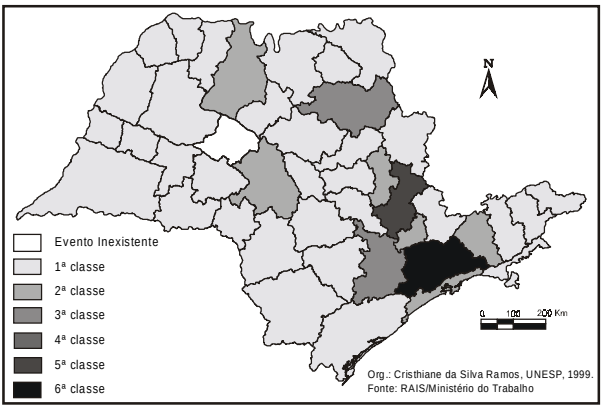
O cartograma resultante retratou satisfatoriamente a concentração de estabelecimentos na região de São Paulo e adjacências. Destaca-se como vantagem deste método a extrema facilidade com que ele é aplicado, no entanto pode haver número distinto de classes entre diferentes cartogramas da série pois o número de classes varia de acordo com o número de elementos de cada rol analisado, fato que preocupa quando o objetivo é comparar cartogramas representativos de diferentes períodos cronológicos.

Como desvantagem pode-se também destacar que por considerar intervalos de classe fixos, podem ocorrer classes vazias, como é o caso da classe 4, que apesar de existir na legenda inexiste no cartograma.

Classificação Hierárquica por Pares Recíprocos

A técnica da Classificação Hierárquica por Pares Recíprocos consiste na elaboração de matrizes de agrupamento, onde são analisadas as distâncias entre os

Figura 6 - Distribuição dos Estabelecimentos da Indústria de Material de Transporte por Regiões de Governo, São Paulo, 1995. - Classificação pelo Método de Sturges



diversos valores da série de dados, agrupadas crescentemente, chegando-se ao final em uma única classe que compreenderia todos os valores da série e corresponderia a 100% de generalização

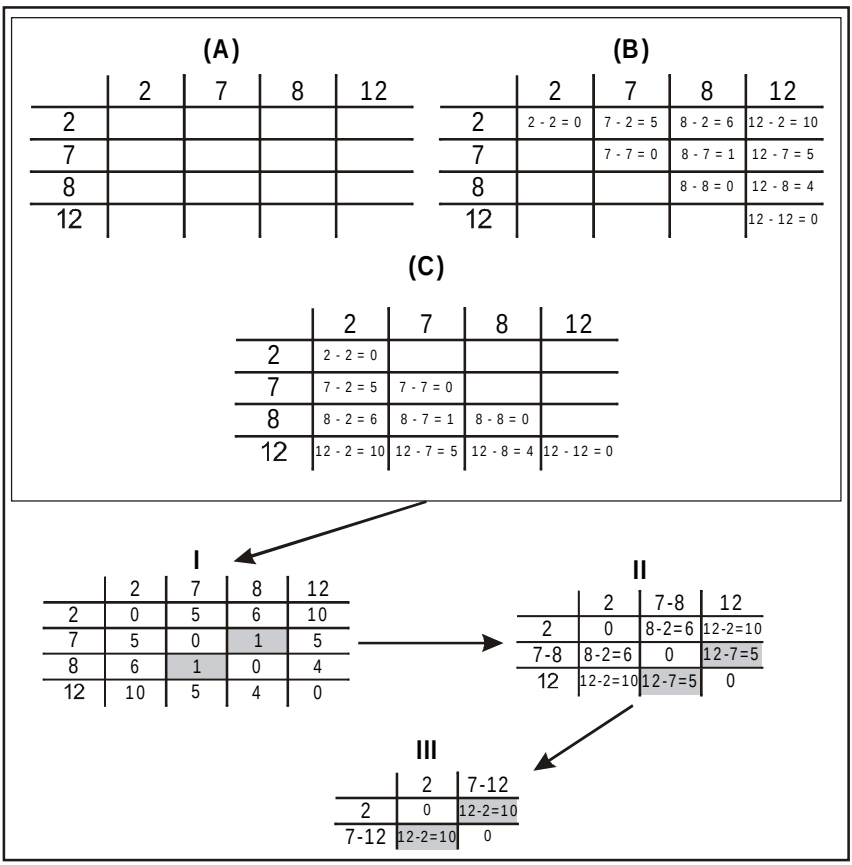
Para compor a primeira matriz de afastamento são lançados na primeira linha e primeira coluna os valores do rol colocados em ordem crescente, os demais valores da matriz são resultantes do cálculo da diferença existente entre os valores da primeira linha e primeira coluna da matriz. Existem três formas de determinar esta distância:

- Mínima distância
- Máxima distância
- Centróide

Neste trabalho optou-se pelo método da máxima distância que é determinada pela subtração dos valores da primeira linha e primeira coluna para cada posição na matriz.

Por exemplo, para a série 2, 7, 8 e 12 seria composta uma matriz de 5 colunas e cinco linhas (figura 7A), na primeira linha e primeira coluna seriam lançados os quatro valores da série.

Figura 7 - Procedimento para o cálculo das matrizes de agrupamento



Para a composição dos valores da matriz são subtraídos os valores das linhas e colunas, por exemplo para a posição 2,2 (segunda linha e segunda coluna) são subtraídos os valores respectivos na primeira linha e primeira coluna, ou seja 2 e 2 resultando em 0, como ilustra a figura 7B.

No sentido das colunas o cálculo segue a mesma lógica, porém serão invertidos os operadores, seguindo o procedimento: valores da primeira coluna menos valores da primeira linha, como mostra a figura 7C, ou seja, a matriz é simétrica.

Os procedimentos ilustrados nas figuras 7A, 7B e 7C permitem a elaboração da primeira matriz de afastamento (figura 7I). Através da observação da primeira

matriz, são agrupados os pares de elementos que apresentam as menores distâncias reciprocamente, neste caso 1, criando a classe 7-8 que será lançada na segunda matriz (figura 7II).

Observa-se que para estabelecimento de distância entre a 7-8 e os demais valores usa-se a técnica da maior distância, ou seja, se o valor a ser subtraído for menor que 8 usa-se 8 como valor na subtração, caso o valor seja maior que 8 usa-se o valor 7 na subtração. Na segunda matriz agrupam-se os valores 7-8 e 12, por apresentar menor distância (5), gerando a terceira e última matriz (figura 7III).

Desta forma chegou-se ao agrupamento final, que apresenta distância máxima igual a 10, que corresponde a 100% de generalização, ou seja, uma única classe. Calculados os percentuais relativos às distâncias apresentadas nos demais agrupamentos pode-se lançar estes valores em um diagrama, que dependendo de sua construção pode ser chamado árvore de ligação ou dendograma, este tipo de diagrama permite melhor análise dos agrupamentos.

Foram elaboradas as matrizes de afastamento para os dados aqui analisados. Ao todo foram criadas doze matrizes. Verificados os valores de agrupamento, estes foram lançados em uma tabela e convertidos em percentual, considerando-se 100% a maior distância verificada, ou seja 1.239.

Estas distâncias, convertidas em percentual, foram lançadas em um gráfico onde são visualizados os agrupamentos (figura 8A).

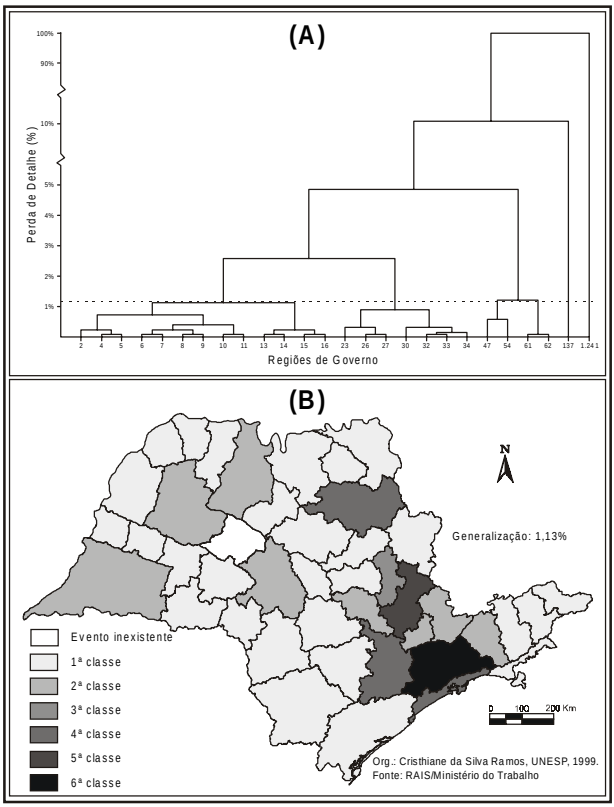
A linha pontilhada na figura 8A mostra o corte para o estabelecimento de seis classes (com percentual de perda de detalhe de 1,13%), seguindo sugestão obtida na aplicação do método de Sturges. São criados assim os seguintes intervalos de classe:

Tabela 8. Intervalos de Classe obtidos através da técnica da Classificação Hierárquica por Pares Recíprocos.

Classes	Limite mínimo	Limite máximo
1ª classe	2	16
2ª classe	23	34
3ª classe	47	54
4ª classe	61	62
5ª classe	137	137
6ª classe	1.241	1.241

Os intervalos de classe apontados na tabela 8 resultam no cartograma exposto na figura 8B.

Figura 8 - Dendograma (A) e Distribuição dos Estabelecimentos da Indústria de Material de Transporte por Regiões de Governo, São Paulo, 1995. - Classificação pelo Método da Classificação Hierárquica por Pares Recíprocos (B)



A classificação hierárquica por pares recíprocos resultou em um cartograma onde é explicitada a concentração dos estabelecimentos da indústria de material de transporte na região metropolitana de São Paulo e regiões próximas. SANCHEZ (1972) aponta como principal vantagem deste tipo de classificação o fato dela permitir a mensuração da perda de detalhe (neste caso de 1,13%) e também a liberdade que dá ao pesquisador para escolher o número de classes ideal de acordo com o objetivo de sua pesquisa.

Entretanto, aponta-se como desvantagem a extrema complexidade deste método, que no caso de séries de dados com muitos elementos se torna ainda mais

complexo. Buscando viabilizar a aplicação deste método no projeto Atlas foram pesquisados dois softwares que realizassem este tipo de classificação, o GEO-INF, desenvolvido por TEIXEIRA (1987) e Statistica, versão 4.3, desenvolvido pela StatSoft (1993). O primeiro foi desenvolvido para microcomputadores de 8 bits, este tipo de barramento de dados atualmente está bastante ultrapassado, portanto, softwares desenvolvidos para 8 bits requerem uma série de adaptações para funcionar satisfatoriamente. O software Statistica mostrou-se bastante interessante, pois possui um grande leque de recursos estatísticos, e realiza a classificação hierárquica por pares recíprocos (cluster analisys) apenas com séries multivariadas, o que não é o caso aqui estudado.

No entanto, optou-se por utilizar o software Excel para a realização das matrizes de afastamento. Apesar de não possuir funções prontas para a realização deste tipo de classificação, o EXCEL é um software bastante comum, pois integra o pacote Office da Microsoft. Este software mostrou-se útil na medida em que permite o estabelecimento de fórmulas que efetuam os cálculos diminuindo assim a margem de erro envolvida no processo manual, porém a utilização da planilha eletrônica mostrou-se bastante trabalhosa.

Curva de Lorenz

A Curva de Lorenz é "um recurso gráfico utilizado para medida da distribuição espacial que se baseia na plotagem de porcentagens cumulativas representadas em dois eixos ortogonais." (GERARDI & SILVA, 1981, p. 117)

Para elaborar a curva de Lorenz, deve-se seguir as seguintes etapas:

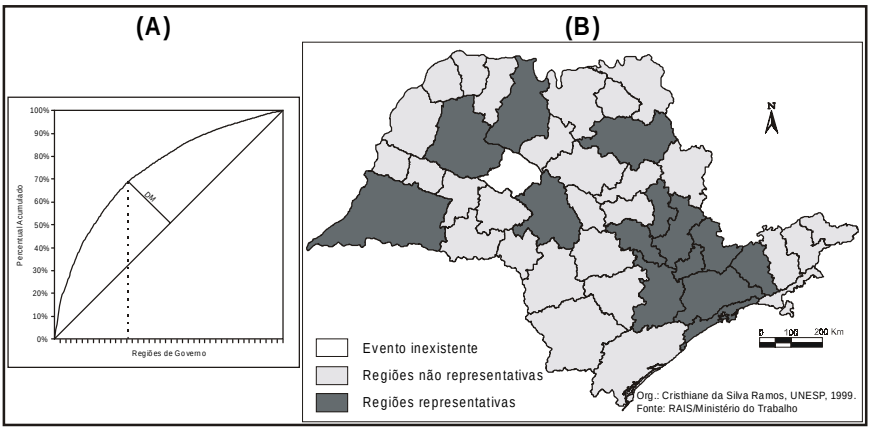
1. Ordenar a série de dados em ordem decrescente
2. Calcular o percentual de cada elemento na série
3. Calcular o percentual acumulativo da série

Os dados inerentes às quatro etapas mencionadas são organizados em tabela. De posse dos valores de percentual acumulado, cria-se um sistema de eixos cartesianos representando valores de 0 a 100%, é traçada a bissetriz que representa distribuição igualitária e, portanto, uma situação de equilíbrio, ou seja, uma situação hipotética, onde nenhum dos elementos da série pode ser considerado estatisticamente mais representativo que os demais.

Os valores em percentual são plotados no gráfico (figura 10A) e então medese a distância de cada ponto da curva em perpendicular à bissetriz, os elementos da série compreendidos entre o início da curva e a distância máxima (DM) são considerados representativos, os valores compreendidos depois da distância máxima são considerados não-representativos.

A análise da Curva de Lorenz estabelece como representativas as 13 primeiras regiões de governo da série, ou seja, aquelas com os valores compreendidos entre 23 e 137. Como a região de São Paulo não foi considerada no cálculo (considerou-se apenas a amplitude útil) ela foi incluída como representativa no mapeamento.

Figura 9 - Curva de Lorenz (A) e Distribuição dos Estabelecimentos da Indústria de Material de Transporte por Regiões de Governo, São Paulo, 1995. - Classificação pelo Método da Curva de Lorenz (B)



GERARDI & SILVA (1981) observam que quando a Curva de Lorenz não tornar possível determinar graficamente a divisão entre dados representativos e não representativos, pode-se recorrer à formula de Ayyar, que determina a distância máxima através da seguinte fórmula:

$$dm = Y \cos\phi - X \operatorname{sen}\phi$$

Onde:

dm é a distância máxima

f é a razão entre os comprimentos dos eixos X e Y (Y/X)

Y é o percentual acumulado dividido por 10

X corresponde a cada um dos elementos da série

Tanto a análise da Curva de Lorenz quanto o método de Ayyar pressupõem duas classes, dados representativos e não representativos. Embora este tipo de classificação possa ser bastante útil quando o que se deseja mostrar é a representatividade de uma determinada parcela dos elementos em relação ao conjunto, tal método não seria válido para o projeto Atlas da Industrialização do Estado de São Paulo, pois um número de classes tão pequeno empobreceria muito a análise dos dados.

TRABALHANDO COM UMA SÉRIE TEMPORAL

A aplicação dos métodos expostos anteriormente mostrou vantagens e desvantagens (tabela 9), embora estatisticamente corretos, cada um possui uma aplicação específica e representa melhor determinados tipos de situação, cabe ao pesquisador, partindo deste conhecimento, decidir pelo método que melhor se aplica em sua situação de trabalho.

No caso do Atlas da Industrialização do Estado de São Paulo optou-se por utilizar o método do Quartil que é de fácil aplicação, e resulta em um número satisfatório de classes: quatro dentro da amplitude útil, uma classe reservada para São Paulo, e outra classe para regiões desprovidas de estabelecimentos industriais, resultando num total de seis classes.

Desta forma, partiu-se para a análise da aplicação das várias técnicas em uma série temporal, utilizando os dados relativos à distribuição do total dos estabelecimentos da indústria de material de transporte para os anos de 1986, 1990 e 1995.

O quadro síntese das técnicas aplicadas ilustra o resultado final e permite a comparação entre todas as técnicas (figura 10). Nesta figura é possível notar que cada técnica gerou diferente número de classes com intervalos distintos, inclusive entre os diferentes anos. Aplicando-se os métodos de classificação para cada série de dados isoladamente (considerando cada ano separadamente) é praticamente impossível observar a dinâmica temporal do fenômeno mapeado.

Buscou-se, então, uma maneira de uniformizar os intervalos de classe que permitisse a análise temporal dos cartogramas. Utilizou-se duas técnicas de uniformização, considerando-se os três anos como uma única série de dados (unificação da série de dados) e através de taxas de variação.

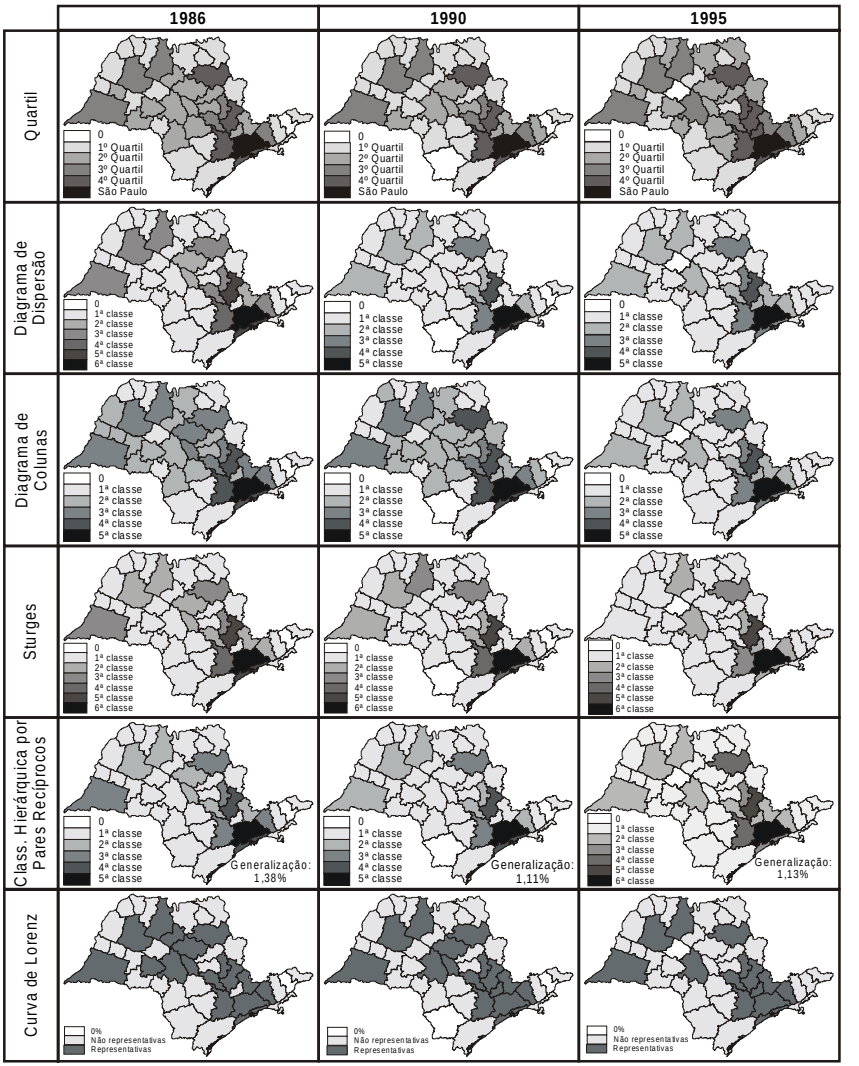
Todo o processo de classificação por si só compreende uma generalização dos dados, a partir do momento em que se torna necessária uma adaptação desses intervalos para que se chegue a um intervalo comum (como no caso de séries temporais), pressupõe-se que a perda de informação será ainda maior, portanto é

Tabela 9. Vantagens e desvantagens dos métodos estatísticos analisados.

Método de Classificação	Vantagens	Desvantagens
Quartil	Cálculo simples e rápido	Não considera a amplitude dos dados, apenas sua posição na série.
Diagrama de Dispersão	Método de fácil aplicação, pois a classificação acontece de acordo com o agrupamento "visual" dos dados.	Método subjetivo. O cuidado (ou falta dele) na elaboração do diagrama pode prejudicar a análise.
Diagrama de colunas	Método de fácil aplicação.	Subjetividade: duas pessoas analisando o mesmo diagrama poderiam estabelecer classificações distintas. O estabelecimento dos intervalos de classe está diretamente relacionado à escala adotada no eixo Y.
Sturges	Estabelece o número de classes ideal de acordo com o número de elementos da série. Pode ser integrado a outros métodos estatísticos, servindo de parâmetro para o estabelecimento do número de classes.	Por não considerar a amplitude dos dados, pode resultar em classes vazias.
Classificação Hierárquica por Pares Recíprocos	Permite a mensuração da perda de detalhe implícita no processo de classificação. Dá ao pesquisador a liberdade de escolher o número de classes que melhor lhe convier.	Método complexo, de difícil aplicação, salvo quando realizado através de *software* estatístico.
Curva de Lorenz	Diagrama de fácil construção. Apresenta os valores representativos da série de dados.	Por apresentar apenas duas classes, deve ser utilizado apenas quando se pretende descobrir quais elementos da série são representativos (ou não).

pertinente pesquisar o melhor método para realizar esta adaptação, de modo a reduzir a generalização.

Figura 10 - Quadro-síntese de todas as técnicas aplicadas



Unificação da Série de Dados

Uma alternativa para solucionar a questão do mapeamento de uma série temporal de dados seria a unificação da série. Ou seja, considerar todos os valores dos diferentes anos (independente da região de governo a qual pertençam) como uma série única.

Para tanto, foram dispostos os valores da tabela 2 em ordem crescente, excluídas as repetições. Obteve-se então a série:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 23, 24, 25, 26, 27, 30, 32, 33, 34, 35, 43, 46, 47, 52, 54, 61, 62, 137, 873, 904, 1.241

Da mesma forma que nas séries consideradas individualmente, a série unificada apresenta grande diferença entre os três últimos valores (873, 904 e 1.241) e os demais. Estes valores correspondem à região de governo de São Paulo. Optou-se por excluí-los da série, trabalhando-se apenas com a amplitude útil para fins de classificação, os valores discrepantes foram inseridos numa classe em separado.

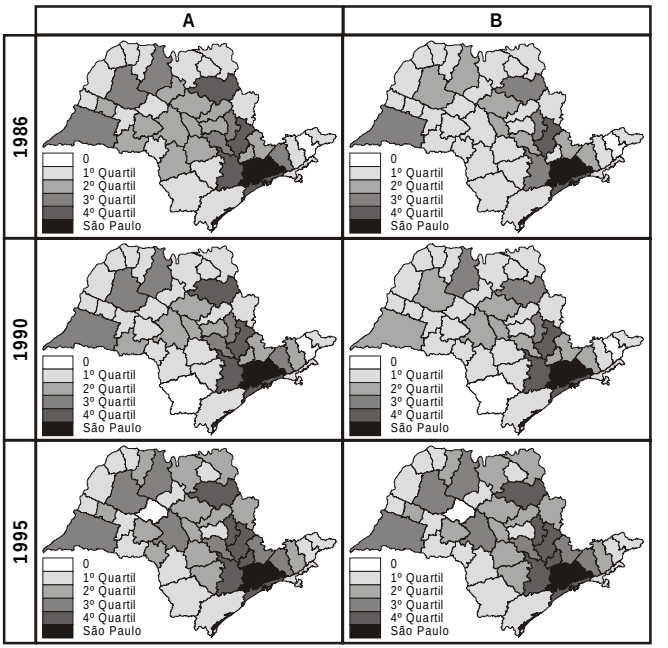
Desta maneira, chegou-se a uma série de 38 elementos, aplicando-se o método do Quartil, foram obtidos os intervalos de classe expostos na tabela 10.

Tabela 10. Intervalos de Classe obtidos através da técnica do Quartil, aplicada na série de dados unificada (1986/1990/1995).

Classe	Limite Mínimo	Limite Máximo
1º. quartil	1	9
2º. quartil	10	20
3º. quartil	21	33
4º. quartil	34	137
São Paulo	873	1241

A comparação entre o mapeamento obtido através da aplicação da técnica do Quartil para cada série de dados (figura 11A) e o obtido através da aplicação dos Quartis na série unificada (figura 11B) permite a identificação de pequenas diferenças. Cabe ao pesquisador avaliar qual a melhor forma de mapear os dados quando em série cronológica.

Figura 11 - Quadro comparativo entre os intervalos de classe apontados pela técnica do Quartis aplicada separadamente para cada série de dados (A) e a unificação da série (B)



Taxa de Crescimento

A taxa de crescimento compreende o cálculo da diferença percentual entre os diferentes anos. Permite desta forma, analisar quais regiões de governo cresceram em termos de número de estabelecimentos industriais e quais diminuíram.

A fórmula para o cálculo da taxa de crescimento entre os anos de 1986 e 1990 para a região de governo de Adamantina, por exemplo, é a seguinte:

$$\text{Taxa de Variação } 86\text{-}90 = \left(\frac{\text{indústrias em } 90 - \text{indústrias em } 86}{\text{indústrias em } 86} \right) * 100$$

$$\text{Taxa de Variação } 86\text{-}90 = \left(\frac{4 - 6}{6} \right) * 100$$

$$\text{Taxa de Variação } 86\text{-}90 = -33{,}33\%$$

A região de governo de Adamantina apresentava em 1986 seis estabelecimentos da indústria de material de transporte, em 1990 apresentava quatro estabelecimentos, ou seja, houve uma redução (crescimento negativo) de 33,33% no número de estabelecimentos da indústria de material de transporte no período.

O único problema verificado no cálculo, acontece quando a região não apresenta estabelecimentos industriais em um ano e no ano seguinte sim. Neste caso ocorre erro de divisão por zero. Como solução adotou-se taxa de variação positiva de 100% nesses casos.

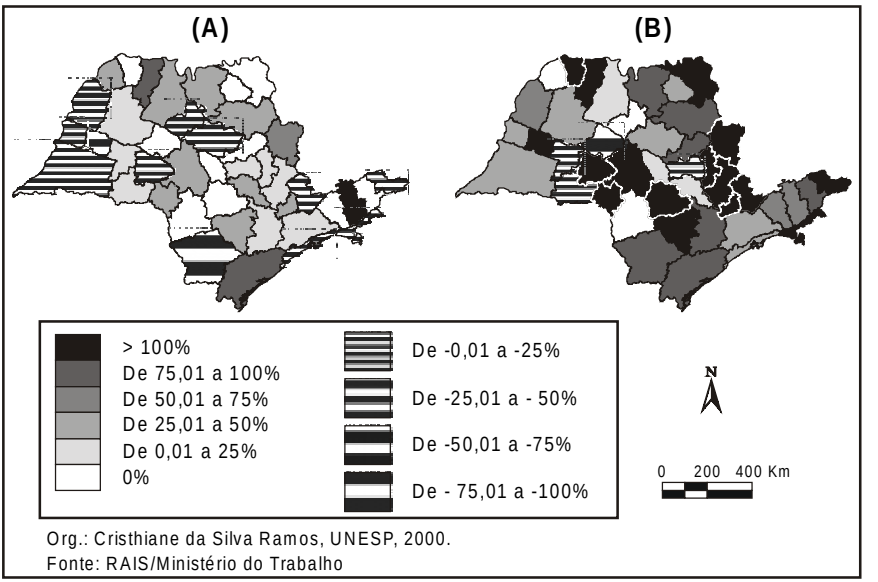
As figuras 12A e 12B mostram os cartogramas relativos às taxas de crescimento para os anos 86-90 e 90-95, para efeito de classificação foram adotados limites de 25%.

A observação das figuras 12A e 12B permite ao leitor a noção da dinâmica deste ramo industrial no período mapeado, pois mostra tons de cinza as regiões que apresentaram taxas positivas e em hachuras as regiões que apresentaram taxas de crescimento negativas.

A taxa de crescimento, no entanto deve ser analisada como tal e não como valor absoluto, caso contrário o pesquisador estaria cometendo um equívoco, por exemplo, no período 90-95 a região de Jundiaí apresentou uma taxa de variação positiva de 240% contra 37,28% apresentada pela região de São Paulo. Isto quer dizer então que a indústria de material de transporte de Jundiaí cresceu mais que a de São Paulo? Em termos percentuais sim, porém se considerarmos os valores absolutos não. Neste período a região de Jundiaí teve um acréscimo de 24 novos estabelecimentos industriais, São Paulo apresentou no mesmo período um acréscimo de 337 novos estabelecimentos.

Portanto, ao trabalhar com uma série de dados temporal, quando o que se pretende é analisar o comportamento dos dados em relação ao total, ou analisar os valores absolutos da série, deve-se optar por um método de classificação adotando a técnica de série unificada. Quando o que se pretende é analisar a variação de cada elemento da série no período, em relação a ele próprio, deve-se optar pela taxa de variação.

Figura 12 - Taxa de Crescimento do Número de Estabelecimentos da Indústria de Material de Transporte, por Regiões de Governo no Estado de São Paulo, para os anos 1986-1990 (A) e 1990-1995 (B)



CONSIDERAÇÕES FINAIS

Uma das principais vantagens deste estudo foi a utilização do *software* EXCEL para as análises estatísticas. Este programa, normalmente utilizado para a tabulação de dados ou para a confecção de diagramas, é na verdade uma poderosa planilha de cálculo, que quando manipulada com habilidade, torna-se bastante flexível.

A experiência de um projeto, onde texto, mapas, diagramas e tabelas de dados são interdependentes mostra a importância da cartografia no trabalho do geógrafo. Como destaca SILVA (1982, p. 57) "mapas e teorias têm propósitos similares. Podemos usar mapas para produzir informação, podemos usá-los para a predição e também para analisar relações. O mesmo ocorre com o uso da teoria.", portanto o documento cartográfico, quando corretamente construído, pode apontar tendências, hipóteses, fenômenos, que não seriam visíveis apenas em uma tabela de dados. Muitas vezes, um mapa bem construído pode comunicar de imediato ao

leitor um fenômeno descrito em várias páginas de texto. O mapa não se restringe apenas a uma ilustração, ele é de fato um instrumento de análise.

No âmbito da cartografia temática, o tema da classificação de dados está longe de se esgotar. Outras formas de representação e classificação devem ser testadas visando a melhor representação da dinâmica temporal e espacial de dados sócio-econômicos.

BIBLIOGRAFIA

CASTRO, J. F. M., GERARDI, L. H. O., BUFALO, A. C. Utilização de SIG na Integração de Dados dos Quadros Físico-Natural e Sócio-Econômico da Região Administrativa de Campinas: Uma proposta Metodológica. *Geografia*, v. 23, n. 3, p. 65-93, 1998.

CERON, A. O. Classificações Espaciais e Regionalização. *Boletim de Geografia Teorética*, v. 7, n. 14, p. 9-45, 1977.

FERREIRA, C. C., SIMÕES, N. N. *Tratamento Estatístico e Gráfico em Geografia*, Lisboa, Gradiva, 1987. 151p.

GERARDI, L. H. O., SILVA, B. C. N. *Quantificação em Geografia*, São Paulo, DIFEL, 1981. 161p.

MANTELLI, J., SANCHEZ, M. C. Técnicas Cartográficas em Geografia. *Geografia: Ensino & Pesquisa*, n. 4, p.7-68, 1990.

MARTINELLI, M. *Curso de Cartografia Temática*, São Paulo, Contexto, 1991. 178p.

MARTINS, G. A., DONAIRE, D. *Princípios de Estatística*, 3ª edição, São Paulo, Editora Atlas, 1987.255p.

MONKHOUSE, F. J., WILKINSON, H. R. *Mapas y Diagramas*, Barcelona, Oikostau, 1968. 533p.

SANCHEZ, M. C. A problemática dos intervalos de classe na elaboração de cartogramas, *Boletim de Geografia Teorética*, v. 3, n. 4, p. 53-65, 1972.

SILVA, B. C. N. Métodos quantitativos aplicados em Geografia: uma Introdução, *Geografia*, v. 3, n. 6, p. 33-73, 1978.

SILVA, S. C. B. M. Cartografia da Acessibilidade e da Interação no Estado da Bahia, *Geografia*, v. 7, n. 13/14, p. 51-73, 1982.

TEIXEIRA, A. L. A. *Sistemas de Informação Geográfica:* uma solução para microcomputadores de oito bits. Rio Claro, dez/1987. 242p. Tese (Doutorado em Geografia, Organização do Espaço) – Instituto de Geociências e Ciências Exatas, Universidade Estadual Paulista.

AGRADECIMENTOS

Ao Prof. Dr. João Afonso Zavatini, Prof. Dr. Marcos César Ferreira, Prof. José Flávio Morais Castro, Prof. Dr. Celso Dal Ré Carneiro, Profa. Dra. Iandara Alves Mendes e Profa. Dra. Lucia Helena de Oliveira Gerardi, por suas críticas e sugestões, sempre pertinentes.