

1 - Introdução

O Sistema de Informação Geográfica (SIG) é uma tecnologia do Geoprocessamento, que possibilita integrar dois tipos de operações realizadas com base de dados. As operações convencionais, como captura, armazenamento, manipulação, análise e apresentação de dados, conjuntamente com as operações que possibilitam a visualização e análise geográfica oferecidas pelos mapas (CARVALHO et al., 2000).

O SIG pode ser utilizado como suporte para análise espacial de dados resultantes de levantamentos de recursos naturais, tais como: mapas geológicos, topográficos, ecológicos, fitogeográficos e pedológicos; e recursos humanos, tais como: dados sócio-econômicos, demográficos, urbanos e de saúde (CÂMARA et al., 2001b).

A utilização do SIG, na Análise Espacial de dados da saúde pública, tem se destacado, principalmente, nas áreas de vigilância epidemiológica, avaliação dos serviços de saúde, urbanização e ambiente. O conhecimento dessas áreas é uma etapa indispensável no processo de planejamento da oferta de serviços de saúde e na avaliação do impacto das ações de saúde (CARVALHO et al., 2000). Uma das principais fontes desse tipo de informação são os Sistemas Nacionais de Informação sobre Saúde, sob responsabilidade do Departamento de Informática do Sistema Único de Saúde – DATASUS.

Este projeto trata-se de um estudo sobre Sistemas de Informação Geográfica e sua aplicação na Análise Espacial de dados da saúde pública, nos municípios paraibanos, entre os anos de 1998 e 2001.

No primeiro semestre do projeto, realizou-se um estudo sobre os princípios básicos de Geoprocessamento e Sistemas de Informação Geográfica. Enfatizou-se neste estudo a utilização do Sistema de Processamento de Informação Georeferenciada - SPRING (SPRING, 1996) como suporte na Análise Espacial de fenômenos geográficos. Foram estudados também, alguns métodos de análise aplicados em SIG como: Análise Exploratória de dados, Classificação Hierárquica, Consultas, Agrupamentos e a Análise de Autocorrelação Espacial de Moran.

No segundo semestre do projeto, esses conceitos e métodos de análise estudados foram utilizados para fazer um estudo com base em seis variáveis da saúde pública do estado da Paraíba, entre os anos de 1998 e 2001. Estas variáveis são: número de hospitais por município,

número de leitos por município, morbidade hospitalar por local (município) de internação, produção ambulatorial por local (município) de atendimento, recursos financeiros do SUS destinados aos municípios e população residente estimada por município.

O grupo de SIG do Departamento de Estatística já possuía um banco de dados composto por essas variáveis, correspondentes ao período de 1998 à 2000, que foram obtidas no DATASUS (DATASUS, 2000). Esses dados foram previamente utilizados por Borges e Moraes (2000) e Teles e Nascimento (2001). Infelizmente, a formatação desses bancos de dados não era adequada ao trabalho que deveria ser desenvolvido. Por esta razão, houve a necessidade de modificação de suas estruturas para posteriormente incluir os dados do ano 2001, que foram divulgados pelo DATASUS (DATASUS, 2003). Em seguida, estes dados foram implementados no sistema SPRING (SPRING, 1996), sob a forma de mapa cadastral, sendo um mapa cadastral para cada ano. Por mapa cadastral entende-se ser a associação de atributos não-gráficos, os dados; a um atributo gráfico, que neste caso é o mapa político da Paraíba, obtido também, no DATASUS (DATASUS, 2003).

2 - Objetivo

O objetivo principal deste projeto é o de criar e difundir uma cultura do uso de Sistemas de Informação Geográfica. Um sistema desse tipo, bem estruturado, poderá auxiliar pesquisas sobre o estado da Paraíba por pesquisadores da própria Paraíba, como de outros estados e até mesmo de outros países.

Para atingir esse objetivo, fez-se um estudo sobre os princípios básicos de Sistemas de Informação Geográfica, o sistema SPRING (SPRING, 1996) e a utilização, deste sistema, na análise espacial de fenômenos geográficos. Em seguida aplicou-se essa metodologia para analisar as variáveis da saúde pública do estado da Paraíba, entre os anos de 1998 e 2001, que foram implementadas no sistema SPRING. Visando com isso, atingir o objetivo de difusão dessa cultura e troca de experiência com outros grupos similares dentro e fora da UFPB, disponibilizando o relatório final deste projeto para acesso público via Internet.

3 - Revisão Bibliográfica

3.1 - Geoprocessamento e Sistemas de Informação Geográfica

Geoprocessamento é um conjunto de tecnologias de coleta, tratamento, manipulação e apresentação de dados espaciais. Dentre essas tecnologias, destacam-se: o sensoriamento remoto, a digitalização de dados, a automação de tarefas cartográficas, a utilização de Sistemas de Posicionamento Global – GPS e o Sistema de Informação Geográfica – SIG. Este último pode ser entendido como a mais completa das tecnologias do Geoprocessamento, uma vez que pode englobar todas as demais (CARVALHO et al., 2000).

Para entender melhor o que é Sistema de Informação Geográfica, apresentam-se, a seguir, algumas de suas definições encontradas na literatura.

“Um conjunto manual ou computacional de procedimentos utilizados para armazenar e manipular dados georreferenciados” (ARONOFF, 1989);

“Conjunto poderoso de ferramentas para coletar, armazenar, recuperar, transformar e visualizar dados sobre o mundo real” (BURROUGH, 1986);

“Um sistema de suporte à decisão que integra dados referenciados espacialmente num ambiente de respostas a problemas” (COWEN, 1988);

“Um banco de dados indexados espacialmente, sobre o qual opera um conjunto de procedimentos para responder a consultas sobre entidades espaciais” (SMITH et al., 1987).

Diante dessas definições, observa-se que cada autor expõe, a sua maneira, a diversidade de aplicações que essa tecnologia pode proporcionar.

3.2 - O Sistema SPRING

O Sistema de Processamento de Informação Georeferenciada (SPRING) é um programa (*software*) de SIG que opera como um banco de dados geográfico e suporta grande volume de dados. Esse programa está sendo desenvolvido pelo Instituto Nacional de Pesquisas Espaciais –

INPE, para ambiente UNIX e WINDOWS. Pode ser obtido pela *Internet*, gratuitamente, no endereço: <<http://www.dpi.inpe.br/spring>> (SPRING, 1996).

3.3 - Definições de Conceitos Básicos em SIG

O Sistema de Informação Geográfica tem como propósito inicial, representar em ambiente computacional os fenômenos geográficos, que se pretende estudar (CÂMARA et al., 2001a). Para isso é necessário compreender conceitos que envolvem essa realidade a ser representada.

Um desses conceitos é o de **espaço geográfico** que pode ser definido, como sendo uma coleção de localizações na superfície da Terra, onde ocorrem os fenômenos geográficos. O espaço geográfico define-se, portanto, em função de suas coordenadas, sua altitude e sua posição relativa. Sendo um espaço localizável, o espaço geográfico é possível de ser cartografado (DOLFUS, 1991).

Em Geoprocessamento, o espaço geográfico é modelado segundo duas visões complementares: os geo-campos e os geo-objetos (WORBOYS, 1995).

- Um **geo-campo** representa a distribuição espacial de uma variável, que possui valores em todos os pontos pertencentes a uma região do espaço geográfico, num dado tempo t. Por exemplo, pode-se ter um geo-campo de vegetação, onde cada um de seus pontos está associado a um tema, como floresta densa, aberta, cerrada, etc. (CÂMARA et al., 2001a).
- Os **geo-objetos** (ou **objetos geográficos**) são entidades distintas e localizáveis que compõem uma determinada região do espaço geográfico. Como exemplo, considere uma pequena cidade, na qual pode-se identificar componentes urbanos como praças, escolas, posto médico, rua principal, etc. Esses componentes urbanos são exemplos de geo-objetos (TELES; MORAES, 1999).

As informações obtidas de um espaço geográfico (Informação Espacial), que depende da localização dos geo-objetos (ou de pontos de um geo-campo) são denominadas de dados espaciais, ou seja, são dados que possuem uma localização geográfica definida. Os dados espaciais são compostos por duas componentes distintas: a parte **gráfica**, que representa

simbolicamente o geo-objeto ou o geo-campo no mapa (ou imagem) e a parte **não-gráfica**, que armazena as características qualitativas e quantitativas dos geo-objetos ou geo-campos (CARVALHO et al., 2000).

No SIG, os dados espaciais são organizados em Bancos de Dados Geográficos e esta organização varia de acordo com o tipo de programa (ou *software*) adotado.

No caso do **SPRING**, o Banco de Dados Geográfico possui uma organização baseada numa arquitetura dual, isto é, as componentes gráficas e as não-gráficas, dos dados espaciais, são armazenadas em ambientes computacionais diferentes. As duas componentes se associam através de identificadores, chamados de geocodificadores (ou GeoID). Desta forma ao selecionar um geo-objeto (ou um ponto de um geo-campo) no mapa, o sistema fornecerá as informações espaciais correspondentes (CÂMARA et al., 2001a).

O banco de dados do SPRING divide-se em partes, denominadas de Projetos. Cada Projeto define uma área física de trabalho e é composto por Planos de Informação - PIs. Um Plano de Informação representa a distribuição espacial de uma variável e é formado pela associação das componentes gráfica (mapas ou imagens) e não-gráficas (atributos). Como exemplo, pode-se ter Planos de Informação de rodovias, ferrovias, redes de drenagem, altimetria, geomorfologia, vegetação, relevo, etc. Cada PI representa a mesma área, mas com informações geográficas (variáveis) diferentes. Estes PIs quando superpostos formarão a cartografia básica da região de estudo (CÂMARA et al., 2001a).

3.4 - Escalas de Mensuração aplicadas em SIG

A escala (ou nível) de mensuração é formada pelo processo de atribuição de números a qualidades de um geo-objeto (ou pontos de um geo-campo), segundo regras definidas (GERARDI; SILVA, 1981).

De acordo com o trabalho desenvolvido por Stevens (1951), existem quatro escalas de mensuração básicas aplicadas em SIG: nominal, ordinal, intervalar e razão.

- No nível **nominal**, os elementos (objetos ou pontos) se diferenciam segundo classes distintas. Como exemplo de classes usadas em medidas nominais tem-se: classes de solo, classes de rocha, classes de cobertura vegetal.

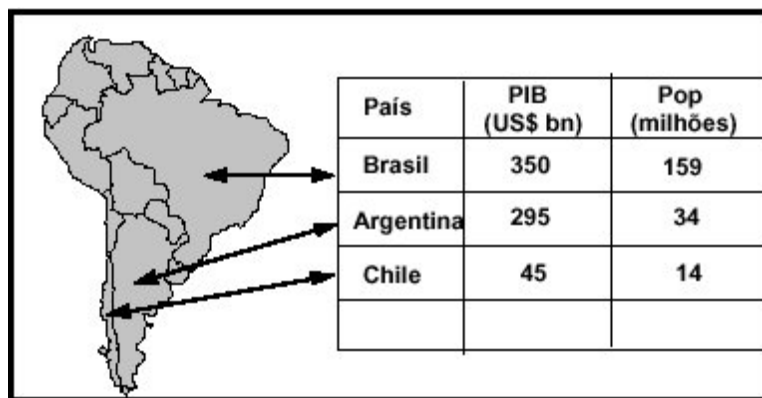
- No nível **ordinal**, os elementos se diferenciam segundo um conjunto ordenado de classes, baseado em critérios como tamanho (maior do que, menor do que), altura (baixo, médio, alto), etc.
- Nível **intervalar**, baseia-se em uma escala de números reais, permitindo atribuir aos elementos valores negativos e positivos. Neste nível de mensuração, a razão entre dois intervalos quaisquer é independente da unidade de medida e do ponto de referência zero, que são arbitrários. Como exemplo de medidas no nível de mensuração intervalar tem-se a temperatura (Centígrada ou Fahrenheit) e a localização geográfica através de latitude e longitude.
- O nível de medida **razão** possui todas as características do nível intervalar, sendo que o ponto de referência zero não é arbitrário, mas sim, determinado por alguma condição natural. Por exemplo, na descrição de atributos como peso, área, volume e distância, que tem o propósito de medição, não faz sentido físico valores negativos, sendo a ausência destes atributos o ponto de origem zero, na escala dos números reais.

3.5 - Tipos de dados em Geoprocessamento.

Como foi visto anteriormente, as feições geográficas do mundo real são modeladas por geo-campos ou por geo-objetos. A partir desta modelagem, os dados em Geoprocessamento, podem ser classificados em: mapa temático, mapa cadastral, redes, imagens e modelo numérico de terreno (CÂMARA et al., 2001a).

- Os **Mapas Temáticos** representam graficamente a distribuição espacial de uma variável. Esta variável associa a cada ponto, de uma região do espaço geográfico, uma classe, que pode ter nível de mensuração nominal ou ordinal. Desta forma, os mapas temáticos descrevem fenômenos modelados por geo-campos.
- Um **Mapa Cadastral** distingue-se de um temático, pois cada um de seus elementos é um objeto geográfico, que possui atributos e pode estar associado a várias representações

gráficas. Por exemplo, a Figura 3.1 mostra um mapa cadastral da América do Sul, onde os países são objetos geográficos que possuem atributos não-gráficos (PIB e população) e que podem ter representações gráficas diferentes em mapas de escalas distintas.



Fonte: Câmara et al. (2001a).

FIGURA 3.1 - Exemplo de um mapa cadastral da América do Sul

- **Redes** são estruturas lineares, formadas por geo-objetos conectados. Sua parte gráfica é armazenada no sistema, através de uma representação vetorial, denominada de topologia arco-nó. Cada arco é constituído de pontos, que estão conectados linearmente e o nó é o ponto de interseção entre dois ou mais arcos. À parte não-gráfica é armazenada no banco de dados geográfico, onde os atributos correspondentes ao arco incluem o sentido do fluxo.
- As **Imagens** representam formas de captura indireta de informação espacial e são fontes atualizadas de informação para produção de novos mapas. Podem ser obtidas por satélites, fotografias aéreas ou *scanners* aerotransportados. As imagens são armazenadas no sistema como matrizes, onde cada um de seus elementos (denominados de “*pixel*”), tem uma coordenada e um valor proporcional à energia eletromagnética refletida ou emitida pela área da superfície terrestre correspondente. Descrevem então, fenômenos modelados por geo-campo.

- **Modelo Numérico de Terreno** (ou **MNT**) é um mapa (em 3 dimensões) de um geo-campo. Representa a distribuição espacial de uma variável que associa a cada ponto (x,y) de uma região do espaço geográfico um número real (z). Os MNT são comumente associados a altimetria, mas também podem ser utilizados para modelar unidades geológicas, como teor de minério, ou propriedades do solo (ou subsolo), como aeromagnetismo.

3.6 - Análise Espacial em SIG.

Os Sistemas de Informação Geográfica são sistemas computacionais, usados para o entendimento de fatos e fenômenos que ocorrem no espaço geográfico. Permite reunir uma grande quantidade de dados espaciais estruturando-os e integrando-os adequadamente. Esses sistemas tornam-se ferramentas essenciais para manipulação e análise de informações geográficas, no que é chamada de análise espacial (PINA, 1994). A Tabela 3.1 ilustra alguns tipos de análise que podem ser feitas através de um SIG (CÂMARA; MEDEIROS, 1996).

Diversas outras análises podem ser realizadas com o SIG. Uma das operações mais utilizadas em SIG é a apresentação dos dados espaciais em um mapa coroplético (nome dado pelos geógrafos aos mapas coloridos). Este mapa mostra a distribuição espacial de uma variável, onde se pode visualizar o padrão espacial do fenômeno (CÂMARA et al., 2001b).

TABELA 3.1 - Exemplos de Análise em um Sistema de Informação Geográfica

<i>Análise</i>	<i>Pergunta</i>	<i>Exemplo</i>
Condição	"O que está...?"	"Qual é a população desta cidade?"
Localização	"Onde está...?"	"Quais as áreas com declividade acima de 20%?"
Tendência	"O que mudou...?"	"Esta terra era produtiva há 5 anos atrás?"
Roteamento	"Por onde ir...?"	"Qual o menor caminho para se passar um metrô?"
Padrões	"Qual é o padrão...?"	"Qual a distribuição da dengue na Paraíba?"
Modelo	"O que acontece se...?"	"Qual o impacto no clima se desmatarmos a Amazônia?"

Fonte: Câmara e Medeiros (1996).

Além de visualizar o padrão espacial, pode-se traduzi-lo em considerações objetivas, tais como: o padrão observado apresenta uma agregação definida? Esta distribuição pode ser associada a causas mensuráveis? Como exemplo, considere alguns problemas típicos (CÂMARA et al., 2001b):

- Epidemiologistas coletam dados sobre ocorrência de doenças. A distribuição dos casos de uma doença forma um padrão no espaço? Existe associação com alguma fonte de poluição? Evidência de contágio?
- A polícia deseja investigar se existe algum padrão espacial na distribuição de roubos. Roubos que ocorrem em determinadas áreas estão correlacionados com características sócio-econômicas dessas áreas?
- Geólogos desejam estimar a extensão de um depósito mineral em uma região a partir de amostras. Pode-se usar essas amostras para construir um mapa de contaminação?
- Queremos analisar uma região geográfica para fins de zoneamento agrícola. Como escolher as variáveis explicativas (p.ex., o solo, a vegetação e a geomorfologia) e determinar qual a contribuição de cada uma delas para a obtenção de um mapa resultante?

Esses problemas são exemplos de questões que podem ser resolvidas com a Análise Espacial. Em geral, esses problemas lidam com quatro tipos de dados, segundo a natureza de sua observação aleatória (MORAES et al., 2003):

- **Dados pontuais** _ são dados obtidos quando a ocorrência do fenômeno em estudo se dá em uma determinada coordenada geográfica, e esta localização é o fator de interesse no estudo, como por exemplo, a localização da ocorrência dos casos de dengue em uma cidade;
- **Dados de área** _ são dados obtidos quando a ocorrência de um fenômeno em estudo se dá em uma área geográfica aleatória, por exemplo, áreas de incidência de alguma endemia ou

epidemia. Na prática, estes dados são associados a levantamentos de recursos humanos, como censos e estatísticas da saúde. Referem-se então, a indivíduos localizados em pontos específicos do espaço, mas que por razões de confidencialidade ou de tratamento estatístico são agregados em unidades de análise (geo-objeto), usualmente delimitados por polígonos fechados (setores censitários, zonas de endereçamento postal, municípios). Como exemplo desses dados, temos: dados de saúde, sócio-econômico, demográficos;

- **Dados de superfície** _ são dados obtidos de levantamentos de recursos naturais, nos quais, a sua natureza aleatória é a própria superfície do fenômeno estudado. Estão disponíveis usualmente como um conjunto de valores, que podem estar regularmente ou irregularmente distribuídos e são modelados como uma amostra de uma superfície contínua (geo-campo). Por exemplo: a distribuição da temperatura em um estado ou a distribuição do *ph* em uma área de solo;
- **Dados de interação** _ são dados obtidos quando há um fluxo aleatório de informações entre dois ou mais pontos com localização geográfica fixa. Por exemplo, o fluxo migratório entre duas cidades, o fluxo de veículos e cargas, etc.

Neste estudo, serão utilizados apenas os dados de área, uma vez que os dados de saúde pública, divulgados pelo SUS, estão agregados por município (geo-objeto).

Os fenômenos que são representados por dados de área, podem ser descritos por variáveis aleatórias geo-referenciadas. Estas podem ser entendidas, como sendo uma função que associa a cada geo-objeto (que por definição é geo-referenciado) um valor numérico, que possui um determinado nível de mensuração.

No processo de análise espacial, procura-se descobrir se as variáveis aleatórias geo-referenciadas seguem algum padrão e se esse padrão depende da localização espacial, de modo a descobrir um modelo inferencial que melhor represente o relacionamento espacial presente no fenômeno.

4 – Análise Espacial de Dados de Área

As ferramentas de análise, para dados de área estudadas, foram as seguintes: Análise Exploratória de Dados, Classificação Hierárquica, Consultas, Agrupamentos e Análise de Autocorrelação Espacial de Moran.

4.1 - Análise Exploratória de Dados

Análise Exploratória de Dados é um conjunto de procedimentos estatísticos que visam descrever e identificar o tipo de distribuição das variáveis aleatórias. Apesar de não ser um método de Análise Espacial, a Análise Exploratória é uma etapa primordial para realização de análises mais complexas, já que vários procedimentos estatísticos, utilizados ou não na Análise espacial, são baseados na suposição de que os dados provêm de uma distribuição Normal (também denominada de Gaussiana) ou supõem conhecido o tipo de distribuição das variáveis aleatórias.

A descrição dos dados é realizada através da Estatística Descritiva, que se utiliza, de métodos numéricos para resumir as informações contidas nos dados, como por exemplo: Medidas de Tendência Central, Medidas de Dispersão e os Quantis. Utiliza-se também de métodos gráficos, que permitem visualizar padrões de comportamento nos dados (VIEIRA, 1999).

Medidas de Tendência Central

Medidas de Tendência Central são medidas que descrevem a tendência que os dados têm de se agrupar em torno de certos valores. Dessas medidas as mais usuais são: a média aritmética, a mediana e a moda.

A **Média Aritmética** - \bar{X} , representa o ponto de equilíbrio de um conjunto com n valores, x_1, x_2, \dots, x_n e é definido por (4.1):

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} \quad (4.1)$$

Dado um conjunto de valores ordenados em uma seqüência crescente (ou decrescente) x_1, x_2, \dots, x_n , a **Mediana** é o valor que ocupa a posição central (quando o número de valores for ímpar) ou a média dos dois valores que ocupam as posições centrais (quando o número de valores for par).

Os valores do conjunto de valores, de uma variável aleatória, que ocorre com maior freqüência são denominados de **Moda**. Se a moda for um único valor, a distribuição é denominada de unimodal. Caso exista mais de um valor, as denominações são: bimodal para dois valores, trimodal para três valores e etc.

Medidas de Dispersão

As informações fornecidas pelas medidas de Tendência Central, em geral, necessitam ser complementadas pelas medidas de Dispersão. Estas servem para indicar, o quanto os dados apresentam-se dispersos, em torno da região central (COSTA NETO, 1977). Dentre essas medidas destacam-se a variância e o desvio-padrão.

A **Variância** - S^2 , de um conjunto com n observações x_1, x_2, \dots, x_n , é dada por (4.2):

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (4.2)$$

A variância é uma medida expressa numa unidade igual ao quadrado da unidade dos dados, o que pode causar problemas de interpretação. Costuma-se usar então, o **Desvio Padrão** - S , que é uma medida definida como sendo a raiz quadrada positiva da variância. O desvio padrão indica qual será o “erro” (desvio) cometido, ao tentar substituir cada observação pela média aritmética.

Quantis

Apenas com as medidas de Tendência Central e Dispersão não dá pra se ter idéia da simetria (ou assimetria) da distribuição de uma variável aleatória X . Para contornar esse fato, pode-se utilizar a medida p-quantil.

Dado um conjunto de valores ordenados x_1, x_2, \dots, x_n , o **p-quantil** (ou quantil de ordem p) é uma medida, indicada por $q(p)$ e definida por (BUSSAB; MORETTIN, 2002):

$$q(p) = \begin{cases} x_1, & \text{se } p < p_1 \\ x_i, & \text{se } p = p_i \\ (1 - f_i)q(p_i) + f_i q(p_{i+1}), & \text{se } p_i < p < p_{i+1} \\ x_n, & \text{se } p > p_n \end{cases} \quad \text{com } i=1, 2, \dots, n \quad (4.3)$$

Onde:

x_i : é a observação de ordem i ;

p : é uma proporção qualquer, talque $0 < p < 1$;

$p_i = (i-0,5)/n$, com $i = 1, 2, \dots, n$;

$f_i = (p - p_i)/(p_{i+1} - p_i)$, com $i = 1, 2, \dots, n$.

Quando a distribuição de uma variável aleatória apresenta uma forma aproximadamente simétrica, verificam-se as seguintes relações, onde os valores $q(0,25)$, $q(0,5)$ e $q(0,75)$ são denominados de quartis (BUSSAB; MORETTIN, 2002):

$$(a) q(0,5) - x_1 \cong x_n - q(0,5);$$

$$(b) q(0,5) - q(0,25) \cong q(0,75) - q(0,5);$$

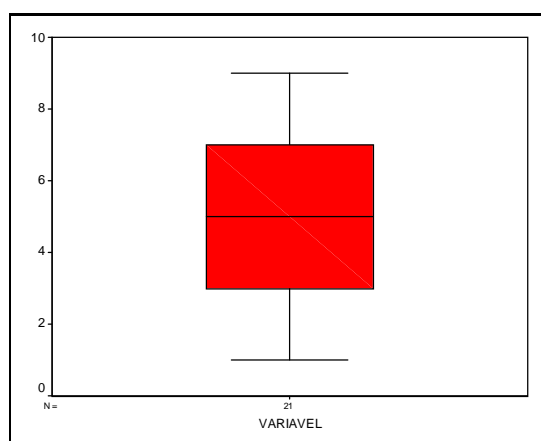
$$(c) q(0,25) - x_1 \cong x_n - q(0,75).$$

Histograma

Agrupando os dados em intervalos com amplitudes iguais, denominados de intervalos de classes, o **Histograma** é definido como sendo um gráfico de barras contíguas, com as bases proporcionais aos intervalos de classes e a área de cada retângulo proporcional à respectiva frequência (BUSSAB; MORETTIN, 2002). Permite visualizar a simetria (ou assimetria) da distribuição.

Boxplot

O **Boxplot** (ou desenho esquemático) é um diagrama que representa os quartis, como pode ser observado na Figura 4.1. Este diagrama dá uma idéia da dispersão, posição, assimetria e caudas da distribuição de uma variável aleatória X (BUSSAB; MORETTIN, 2002). A dispersão pode ser observada pela distância de $q(0,25)$ à $q(0,75)$, denominada de distância interquartílica e que é representada no *boxplot* pela altura do retângulo; a posição central é dada por $q(0,5)$ (a mediana), que no *boxplot* corresponde a linha que corta o retângulo; a posição relativa dos quartis, no *boxplot*, dão uma noção da assimetria da distribuição; os comprimentos das caudas são observados na Figura 4.1, pelas linhas que vão do retângulo aos valores extremos.



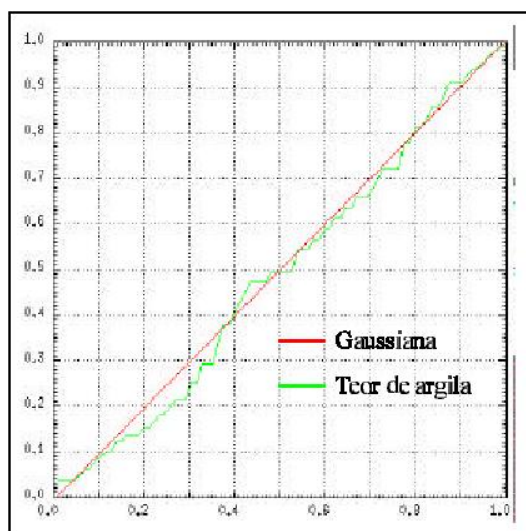
Fonte: Exemplo hipotético do "boxplot" gerado no SPSS (SPSS, 1997).

FIGURA 4.1 - Exemplo de um gráfico "Boxplot"

Qqplot

O **Qqplot** é um gráfico de comparação entre a distribuição dos valores observados da variável X e a distribuição dos quantis esperados, $q(p)$, se a variável X , tivesse distribuição Normal (ou Gaussiana). Esse gráfico pode ser usado para avaliar a hipótese de normalidade dos dados. Se a variável tem distribuição Normal, os pontos $(x_i, q(p_i))$ desenhados num gráfico

cartesiano devem se localizar sobre a reta $X = q(p)$ (JOHNSON; WICHERN, 1992). Na Figura 4.2, tem-se um exemplo de um gráfico *qqplot*.



Fonte: Camargo (1997).

FIGURA 4.2 – *Qqplot* da amostra de Teor de Argila da fazenda Canchim, no município São Carlos, estado de São Paulo.

Transformação de Variáveis

Se uma variável aleatória não tiver uma distribuição Normal, pode-se efetuar uma transformação nos dados obtendo uma distribuição mais simétrica e próxima da Normal. A transformação consiste em efetuar algum tipo de operação matemática a todos valores de uma variável aleatória X , obtendo uma nova variável aleatória Y . Uma família de transformações freqüentemente explorada é (BUSSAB; MORETTIN, 2002):

$$Y = \begin{cases} x_i^k, & \text{se } k > 0 \\ \ln(x_i), & \text{se } k = 0 \\ 1/x_i^k, & \text{se } k < 0 \end{cases} \text{ com } i=1, 2, \dots, n \quad (4.4)$$

Onde:

x_i : é a observação de ordem i da variável aleatória X , com $i = 1, 2, \dots, n$;

Y : é a variável obtida com a transformação;

k : é um número real que define a transformação Y .

Normalmente, escolhe-se valores de k na sequência (...,-3,-2,-1,-1/2,-1/3,-1/4, 0, 1/4, 1/3, 1/2, 1, 2, 3, ...), examina-se os gráficos histograma e *Qqplot*, de modo a obter uma variável transformada Y mais simétrica e aproximadamente Normal.

4.2 - Classificação Hierárquica

Classificação Hierárquica é um método de Análise Multivariada que tem o objetivo de identificar agrupamentos (*clustering*) naturais de objetos, itens ou variáveis. Esse método não requer a especificação do número de agrupamentos, que são feitos com base em critérios de distância (JOHNSON E WICHERN, 1992).

Para realizar a classificação nos dados é necessário que todas as variáveis tenham distribuição Normal e valores compatíveis quanto à unidade e grandeza. As variáveis que não tiverem distribuição Normal devem ser normalizadas através de transformações (ver seção 4.1), obtendo-se uma maior eficiência do método de agrupamento. A compatibilidade das variáveis pode ser obtida, padronizando-se cada variável X_j , para obter a distribuição Normal(0,1), através da transformação: $Z_j = (X_j - \bar{X})/S$, onde \bar{X} é a média e S é o desvio padrão da variável X_j . Na Classificação Hierárquica, os erros ou variações não são considerados, o que torna este método sensível a *outliers* (pontos extremos).

O critério de distância que será utilizado, para realizar os agrupamentos de objetos, é o da ligação média entre os grupos ou "*average linkage between groups*". Os passos para execução de

uma Classificação Hierárquica, utilizando esse critério e partindo de N grupos de objetos são os seguintes (JOHNSON E WICHERN, 1992):

- 1) Inicia-se com N agrupamentos, cada um contendo apenas um objeto. Calcula-se a matriz de distância (ou similaridade) $D = \{d_{ij}\}$, de ordem $N \times N$, onde d_{ij} é a distância Euclidiana ao quadrado, dada por:

$$d_{ij} = \sqrt{\sum (X_i - X_j)^2} \text{ para } i \neq j. \quad (4.5)$$

- 2) Busca-se na matriz de distância D , o mais próximo, ou mais similar, par de agrupamentos U e V . Une-se os agrupamentos U e V em novo agrupamento (UV) .

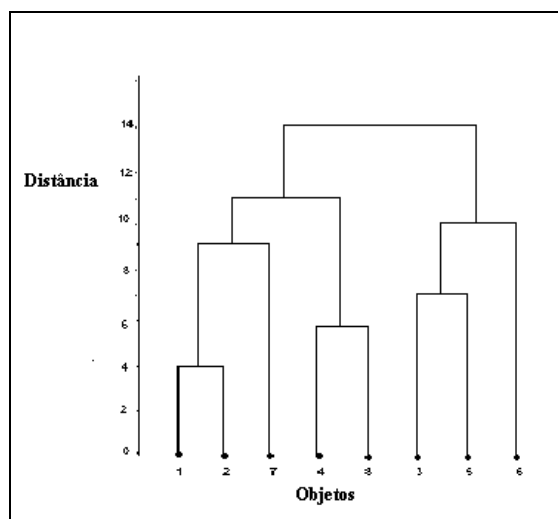
- 3) Recalcula-se a matriz de distância $D = \{d_{ij}\}$, eliminando-se as informações dos grupos U e V separadamente e substituindo-as pelas informações do novo grupo (UV) . A distância entre (UV) e qualquer outro agrupamento W , é dada por:

$$d_{(UV)W} = \frac{\sum_i \sum_j d_{ij}}{N_{(UV)}N_W} \quad (4.6)$$

Onde d_{ij} é a distância entre o objeto i no agrupamento (UV) e o objeto j no agrupamento W . Já $N_{(UV)}$ e N_W são os números de objetos nos agrupamentos (UV) e W , respectivamente.

- 4) Repetem-se os passos 2 e 3, até que todos os objetos estejam contidos em um único agrupamento.

O resultado da Classificação Hierárquica pode ser visualizado em um gráfico cartesiano, denominado de dendrograma, como mostra a Figura 4.3. Com o dendrograma pode-se escolher o número de classe com que se deseja trabalhar ou estabelecer a que distância os objetos devem ser classificados. Determinando as classes de agrupamentos, os resultados podem ser visualizados em um mapa coroplético, visto que estaremos agrupando objetos geográficos.



Fonte: Adaptado de Johnson e Wichern (1992).

FIGURA 4.3 - Exemplo de um Dendrograma.

4.3 - Consultas

A consulta consiste num processo de seleção de objetos geográficos que satisfaçam a determinadas condições. Estas condições são restrições feitas a uma ou mais variáveis do banco de dados do SIG, através de expressões lógicas. Cada expressão é formada por uma igualdade ou desigualdade entre duas variáveis, tais como: = (igual), < (menor que), > (maior que), <= (menor ou igual), >= (maior ou igual), <> (diferente). Duas ou mais expressões podem se combinar através dos operadores lógicos AND (interseção) e OR (união). Os objetos geográficos selecionados pela consulta formam uma “Coleção de Objetos”, que pode ser visualizada por meio de mapas coropléticos. Desta forma a Consulta é um método de análise lógica aplicada a um SIG.

4.4 - Agrupamentos

O Agrupamento consiste num processo de agrupar geo-objetos, com base nos valores de suas variáveis (SPRING, 1996). No SPRING encontram-se três métodos de Agrupamentos: Passo Igual, Quantis e Estatístico.

Passo Igual

O método Passo Igual permite agrupar os geo-objetos em até 15 grupos, que são formados dividindo os dados de uma determinada variável X , em intervalos com amplitudes iguais (intervalos de classe). Cada intervalo de classe corresponde um grupo e os geo-objetos que fazem parte deste grupo são coloridos no mapa, com uma cor específica, obtendo-se então um mapa coroplético.

Quantil

O método Quantil também permite agrupar os geo-objetos em até 15 grupos, sendo que cada grupo tem aproximadamente a mesma proporção ($1/k$) de geo-objetos, onde k é o número de grupos. No entanto, os intervalos de classe podem ter amplitudes desiguais e são dados em termos dos quantis da variável aleatória X . Estes intervalos são denominados de intervalos inter-quantis. Os geo-objetos que fazem parte de um intervalo inter-quantil são agrupados e coloridos no mapa com uma determinada cor, compondo um mapa coroplético.

Estatístico

O método Estatístico, por sua vez, permite agrupar geo-objetos com base na média aritmética \bar{X} e no desvio padrão S de uma variável aleatória X , isto é, a distribuição da variável X pode ser dividida em intervalos com amplitudes iguais a: $\pm S$, $\pm(1/2)S$ ou $\pm(1/4)S$ em relação à média. Determinada a amplitude, os geo-objetos, que fazem parte de cada um dos intervalos formados, são agrupados e coloridos com uma cor específica, obtendo-se também um mapa coroplético.

4.5 - Análise Espacial e Análise de Autocorrelação Espacial de Moran

A Análise Espacial é o estudo quantitativo de fenômenos que são localizados no espaço, ou seja, permite identificar padrões espaciais na distribuição dos fenômenos. Um conceito chave na

compreensão da análise espacial é a dependência espacial. Essa noção parte do que Waldo Tobler (CÂMARA et al., 2001b) chama de primeira lei da geografia: "todas as coisas são parecidas, mais coisas mais próximas se parecem mais que coisas mais distantes". Ou, como afirma Noel Cressie (1991), "a dependência [espacial] está presente em todas as direções e fica mais fraca à medida que aumenta a dispersão na localização dos dados" (CÂMARA et al., 2001b).

Para estudar quantitativamente, a dependência espacial existente entre os geo-objetos, utiliza-se a Autocorrelação Espacial. Este termo foi derivado do conceito Estatístico de correlação, utilizado para mensurar o relacionamento entre duas variáveis aleatórias. A preposição "*auto*" indica que a medida de correlação é realizada com a mesma variável aleatória, medida em locais distintos do espaço, ou seja, a autocorrelação espacial mede a correlação existente entre geo-objetos para um determinado fenômeno (CÂMARA et al., 2001b).

Matriz de Proximidade Espacial

Para estimar a Autocorrelação Espacial de um fenômeno geográfico, faz-se necessário especificar uma medida que relacione os geo-objetos. Uma ferramenta básica é a Matriz de Proximidade Espacial. Dada uma região do espaço geográfico, composta por n geo-objetos $\{G_1, G_2, \dots, G_n\}$, a matriz de proximidade W é tal que, cada um de seus elementos w_{ij} representa uma medida de proximidade entre os geo-objetos G_i e G_j , segundo um critério de proximidade (CÂMARA et al., 2001b). Um critério possível é: $w_{ij}=1$, se G_i compartilha um lado com G_j e caso contrário, $w_{ij}=0$.

Média Móvel Espacial

Para obter uma primeira aproximação da variabilidade espacial de uma variável aleatória Y , pode-se calcular a Média Móvel Espacial \bar{y}_i , para cada geo-objeto G_i . Considerando a matriz de proximidade W , a estimativa da média móvel espacial \bar{y}_i , é dada por (4.7). Esta operação consiste em substituir em cada geo-objeto G_i , o valor da variável y_i , pela média aritmética dos

valores de Y de seus vizinhos, obtendo-se uma superfície menos descontínuas nos dados originais.

$$\bar{y}_i = \frac{\sum_{j=1}^n w_{ij} y_j}{\sum_{j=1}^n w_{ij}} \quad (4.7)$$

Para visualizar a média móvel espacial pode-se realizar um agrupamento de geo-objetos, segundo as médias móveis \bar{y}_i (ver seção 4.4), obtendo-se um mapa coroplético.

Índice Autocorrelação Espacial

A média móvel espacial é vista como uma primeira aproximação da variabilidade espacial, mostrando alguns padrões e tendências. Porém, também é importante verificar a dependência espacial dos dados, a fim de verificar a maneira pela qual os dados estão correlacionados no espaço. A autocorrelação espacial procura medir essa dependência espacial de um geo-objeto G_i e seus vizinhos, segundo um atributo estudado. Então, o Índice Global de Autocorrelação de Moran (I) mede a interdependência geográfica em uma região segundo um atributo e é dado por:

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} z_i z_j}{\sum_{i=1}^n z_i^2} \quad (4.8)$$

Onde z_i é a diferença entre o valor do atributo na região de estudo e a sua média.

O índice de Moran I verifica se os geo-objetos conectados apresentam algum afastamento de um esperado padrão aleatório. Em geral, seu valor fica limitado ao intervalo $[-1, 1]$, embora em alguns casos possa assumir valores diferenciados. Assim, valores positivos do índice I quantificam correlação direta entre o geo-objeto em questão e os seus vizinhos, segundo o atributo estudado e os valores negativos quantificam correlação inversa (CARVALHO, 1997).

O Índice Local de Moran (I_i) é uma simplificação do índice global, computando a variação do atributo no geo-objeto G_i pela variação do mesmo atributo em seus vizinhos.

$$I_i = \frac{z_i \sum_{j=1}^n w_{ij} z_j}{\sum_{i=1}^n z_i^2} \quad (4.9)$$

Segundo Anselin (1995a), os indicadores locais de autocorrelação devem atender a dois objetivos:

- (a) permitir a identificação de padrões de associação espacial e,
- (b) ser uma decomposição do índice global de autocorrelação de Moran.

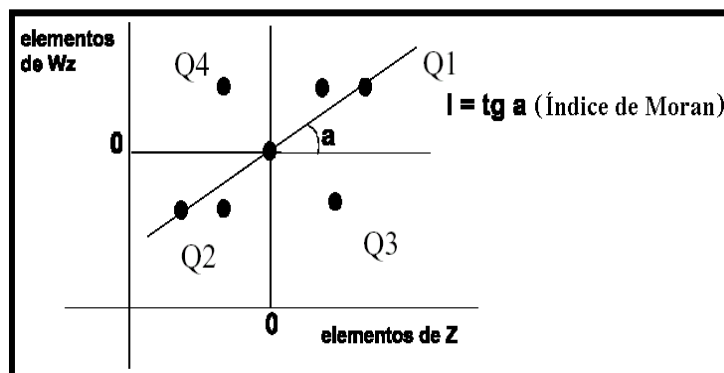
Diagrama de Espalhamento de Moran

Descrevendo a equação (4.8) em forma matricial, temos:

$$I = \frac{Z'WZ}{Z'Z} \quad (4.10)$$

O Diagrama de Espalhamento de Moran é uma forma gráfica adicional para visualizar o Índice de Moran. Procura-se visualizar espacialmente o relacionamento entre os valores observados Z e os valores das médias locais WZ .

O diagrama é mostrado na Figura 4.4. O gráfico é dividido intencionalmente em quatro quadrantes: $Q1$, $Q2$, $Q3$ e $Q4$. Os pontos colocados nos quadrantes $Q1$ e $Q2$ indicam regiões que seguem o mesmo processo de dependência espacial das demais observações. No caso dos pontos colocados nos quadrantes $Q3$ e $Q4$ indicam diferentes processos de dependência espacial, ou seja, podem indicar regiões de transição entre regimes espaciais. Esse gráfico pode ser apresentado na forma de um mapa coroplético, indicando-se os quatro quadrantes com cores diferenciadas e facilitando a interpretação.



Fonte: CÂMARA *et al.* (2001b)

FIGURA 4.4 - Diagrama de Espalhamento de Moran.

Indicadores Locais de Associação Espacial (LISA)

Uma vez admitida a existência da autocorrelação pelo índice local de Moran, pode-se identificar agrupamentos de regiões com atributos semelhantes e áreas anômalas, contendo mais de um regime espacial, com significância estatística.

Para a visualização destes índices, pode ser gerado um mapa coroplético, através do método de Agrupamento (ver seção 4.4), apresentando as regiões que possuem correlação local significativamente diferente do resto das outras regiões. Normalmente, este mapa contém apenas quatro categorias, apresentando a não significância, significância em 5%, significância em 1% e significância em 0,5%.

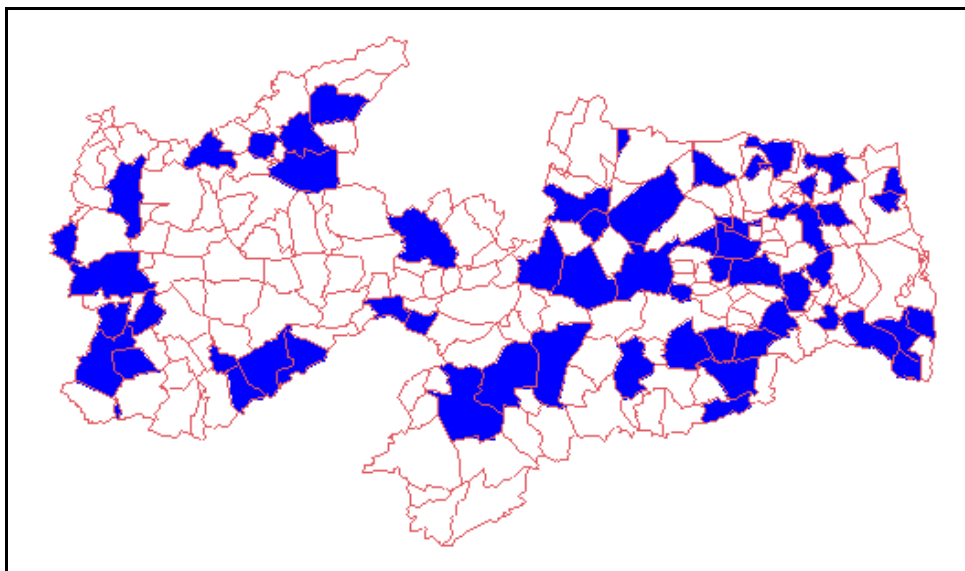
O valor da significância estatística no índice local de Moran pode ser determinado por simulação (pseudo-significância) ou por aproximação a distribuição normal utilizando critérios de convergência. Em ambos os casos, o grau de conhecimento estatístico exigido vai além do nível de graduação e por isso não foi estudado em maiores detalhes.

4.5 - Exemplos

4.5.1 - Exemplo 1:

Na sua tese de mestrado, Borges (BORGES; MORAES, 2000) propôs uma metodologia de Análise Espacial para dados de Saúde Pública, na presença da heterogeneidade nas observações. A metodologia consiste na utilização de Consultas Espaciais e principalmente Classificação Hierárquica, para analisar os recursos do SUS, destinados aos municípios paraibanos, através das seguintes variáveis: recursos financeiros destinados aos municípios paraibanos, número de leitos e internações hospitalares, produção ambulatorial e população total do município. Estas informações foram obtidas do DATASUS (DATASUS, 2000), referem-se ao período compreendido entre os anos de 1998 e 1999, e foram implementadas no sistema SPRING (SPRING, 1996), sob a forma de mapa cadastral.

A consulta teve como base o cruzamento da variável de interesse, os recursos, com as outras variáveis que representam as "despesas", por exemplo, considere o seguinte problema: Verificar se há municípios com baixa incidência de internações hospitalares recebendo valores relativamente altos de recursos, no mês de dezembro de 1999. Para isso, fez-se a seguinte consulta: Quais os municípios, onde a quantidade de internações hospitalares foi menor ou igual à 100 e os recursos foram maiores ou iguais à R\$ 20.000,00, no mês de dezembro de 1999?



Fonte: Adaptado de Borges e Moraes (2000)

FIGURA 4.5 - Municípios paraibanos com menos de 100 internações hospitalar e recursos maiores que R\$ 20.000,00

O resultado da consulta pode ser observado no mapa da Figura 4.5, no qual, verifica-se que os 60 municípios, que aparecem de cor azul no mapa, satisfazem a consulta.

Outras consultas deste tipo foram realizadas, evidenciando a existência de municípios paraibanos que recebem recursos relativamente altos do SUS, apesar de apresentarem poucas "despesas". Segundo Borges e Moraes (2000), esses municípios, provavelmente, utilizam uma parte dos recursos que recebem, com o deslocamento de pacientes para municípios vizinhos. O que levou a inclusão da variável distância de cada um dos 223 municípios do estado da Paraíba para o município mais próximo dentro do estado que possuísse uma melhor estrutura de saúde pública.

Realizou-se uma Classificação Hierárquica utilizando as cinco variáveis mencionadas anteriormente, agrupando os municípios em 7 classes que foram modeladas por Análise de Regressão Múltipla incluindo também a variável "distância".

4.5.2 - Exemplo 2:

No seu relatório PIBIC/CNPq, Teles (TELES; MORAES, 2000) fez um estudo sobre Sistemas de Informação Geográfica e sua aplicação na análise de alguns indicadores dos municípios paraibanos, obtidos do Censo Brasileiro de 1991 (TELES; MORAES, 2000). Os indicadores são: população residente; renda por faixa de salário; área do município; densidade populacional; população por: grau de escolaridade, sexo, cor ou raça, situação de urbanização, condição de moradia, faixa etária e religião. Esses dados foram agregados por municípios e foram armazenados no sistema SPRING (SPRING, 1996), sob a forma de mapa cadastral.

O método de análise utilizado foi a Classificação Hierárquica. Como esse método só é indicado para agrupar variáveis que possuem uma distribuição Normal, fez-se, inicialmente, um estudo sobre a distribuição das variáveis, através de métodos de Análise Exploratória de Dados. Os resultados de algumas variáveis estão apresentados na Tabela 4.1.

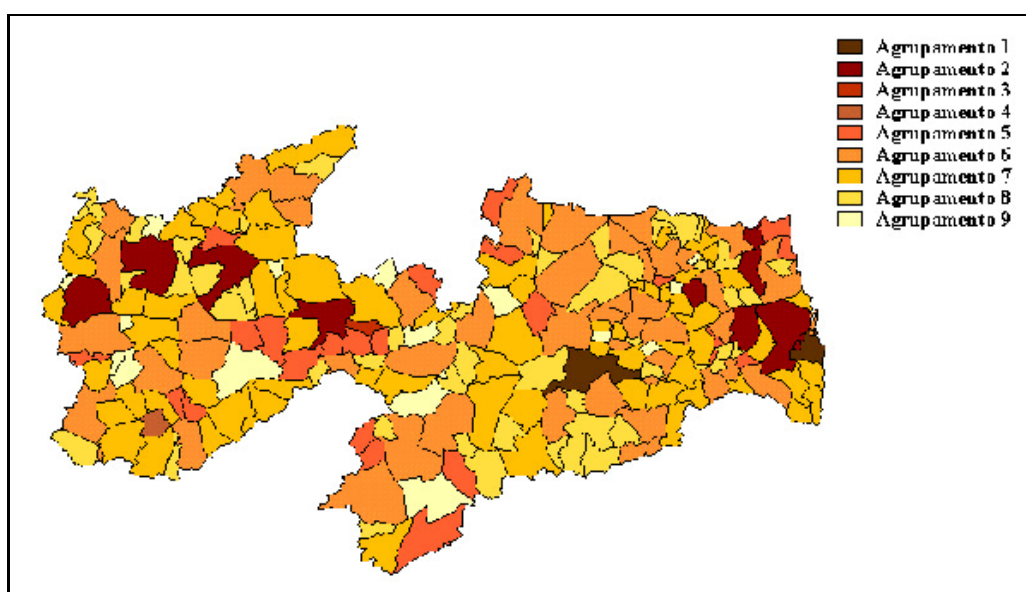
A partir da análise de gráficos, como o *qqplot*, verificou-se que todas as variáveis não possuíam uma distribuição Normal. Para a realização da Classificação Hierárquica, realizou-se a normalização dos dados, através da transformação logarítmica, além da padronização de todas as variáveis.

TABELA 4.1 - Estatística descritiva de algumas variáveis

<i>População</i>	<i>Máximo</i>	<i>Média</i>	<i>Mediana</i>	<i>Desvio padrão</i>
População Residente	497.600	14.354,77	7.257	41.246,99
Com 2º grau completo	19.918	294,40	75,5	1.576,57
Com Domicílio próprio	81.105	2.556,46	1241,5	7.124,22
Com idade de 25 a 29 anos	46.894	1.234,55	512,5	4.014,51
População Residente	497.600	14.354,77	7.257	41.246,99

Fonte: Adaptado de Teles e Moraes (2000).

Deste modo, realizou-se a classificação hierárquica no SPSS (SPSS, 1997), sendo obtidos sete agrupamentos, mais dois agrupamentos de municípios excluídos, totalizando nove agrupamentos. O mapa da distribuição dos agrupamentos, pode ser observado na Figura 4.6.



Fonte: Teles e Moraes (2000).

FIGURA 4.6 - Mapa dos nove agrupamentos no estado da Paraíba

- Grupo 1: neste agrupamento estão os 2 maiores municípios do estado, João Pessoa (Capital do Estado) e Campina Grande, que apresentam um desenvolvimento sócio-econômico semelhante.

- Grupo 2: é composto por municípios que estão localizados próximos às principais rodovias do estado, a BR 230 e a BR 101 e que se destacam por serem cidades que possuem uma boa condição sócio-econômica em relação a outros municípios com população inferior a deles.
- Grupos 3 e 4: são compostos por municípios que possuem as menores populações do estado, sendo Quixabá o 1º (grupo 3) e Curral Velho o 4º (grupo 4) menor em população no ano de 1991. Os dois municípios diferenciam-se dos demais, por apresentarem rendas semelhantes à cidades de porte médio.
- Grupo 5: possui características bem predominantes, com população entre 3.129 a 6.198 habitantes. Vários de seus municípios estão localizados bem próximo às divisas do estado, onde existem rodovias estaduais de acesso a outros estados.
- Grupo 6: é composto por 46 municípios, que apresentam uma população que varia de 13.773 a 33.255 habitantes com um número razoável de crianças cursando o primeiro grau (entre 2460 a 6.214).
- Grupo 7: possui 72 municípios e é o grupo com maior número de municípios. Seus municípios possuem uma população entre 4.448 a 15.007 habitantes com um número pequeno de crianças cursando o primeiro grau (entre 838 e 2.757).
- Grupo 8: é formado pelos municípios criados após 1991. Esses municípios foram excluídos da análise por apresentarem valores nulos em várias das suas variáveis.
- Grupo 9: temos neste grupo os municípios excluídos antes da execução do método de agrupamento hierárquico, por apresentarem valores indeterminados em algumas das variáveis transformadas pelo logaritmo Neperiano.

4.5.3 - Exemplo 3:

Para ilustrar uma aplicação dos indicadores de Autocorrelação de Moran, estudamos o artigo (MORAIS NETO et al., 2001), no qual investigou-se o padrão espacial da mortalidade Neonatal o Pós-neonatal, produzindo mapas que indicaram áreas de risco.

A população considerada foi a coorte de 101 mil nascidos vivos, residentes em Goiânia - GO, entre os anos de 1992 à 1996. As variáveis estudadas foram: nascidos vivos (informação obtida no Sistema de Informação sobre Nascidos Vivos - SINAS); óbitos infantis e probabilidades de morte para os períodos neonatal e pós-neonatal (informação obtida no Sistema de Informação sobre Mortalidade - SIM). Cada declaração de óbito foi emparelhada com a respectiva declaração de nascidos vivos e geocodificadas nos 65 distritos urbanos (geo-objeto) da cidade de Goiânia. Em seguida as informações obtidas por distrito foram implementadas no sistema de informação geográfica ArcView (ESRI, 1996), além dos aplicativos de análise espacial Info-Map (BAILEY; GATRELL, 1995) e Spacestat (ANSELIN, 1995b).

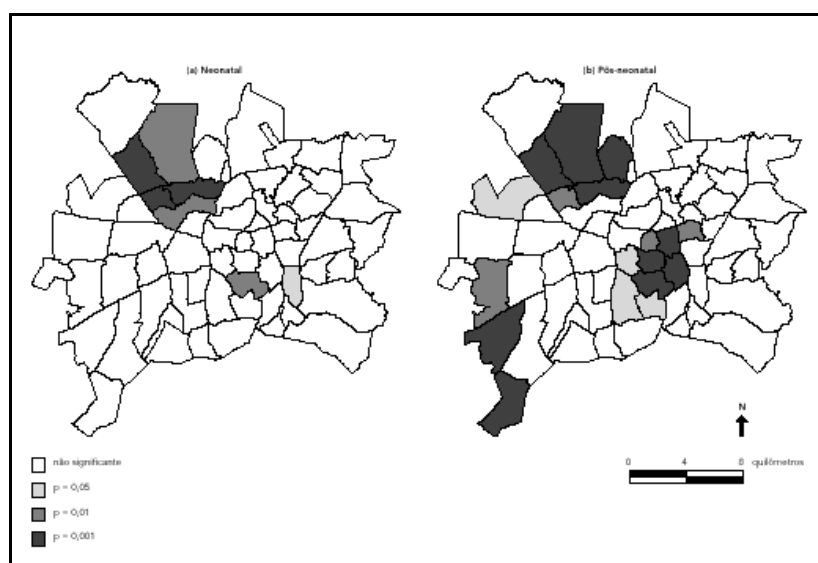
Com os dados das variáveis nascidos vivos e óbitos infantis, foram estimadas as probabilidades de morte neonatal e pós-neonatal.

As probabilidades de morte neonatal e pós-neonatal observadas e as estimadas foram analisadas por meio dos Índices de Autocorrelação Espacial de Moran (Global e Local). Os valores da significância, no índice local de Moran, foram obtidos segundo os métodos de simulação (pseudo-significância) e pela aproximação da distribuição normal. Os resultados obtidos com os dois métodos foram os mesmos.

No período neonatal, as probabilidades de morte observadas apresentaram autocorrelação global não-significativa, $I = 0,0230$. No entanto, as probabilidades de morte neonatal estimadas indicaram autocorrelação global significativa, com $I = 0,1699$.

No período Pós-neonatal, as probabilidades de morte observadas e estimadas apresentaram os seguintes índices de Moran, $I = 0,2324$ e $I = 0,7052$, indicando uma alta correlação global, significativa.

O exame da Autocorrelação Local de Moran mostra que existem, para o período neonatal, três áreas com índices significativos, ou seja, áreas de risco (Figura 4.7 (a)). No período pós-neonatal, observam-se também, três áreas de risco, que abrange um número maior de distritos (Figura 4.7 (b)).



Fonte: (MORAIS NETO et al., 2001)

FIGURA 4.7 - Índice de Moran Local utilizando-se as probabilidades de morte neonatal e pós-neonatal, dos Distritos urbanos de Goiânia, 1992 - 1996.

Os Índices de Autocorrelação de Moran possuem algumas limitações quanto ao uso e a principal delas é a sua utilização para analisar doenças raras e regiões com alta heterogeneidade populacional (MORAIS NETO et al., 2001).