ANÁLISE EXPLORATÓRIA ESPACIAL DE DADOS SÓCIO-ECONÔMICOS DE SÃO PAULO

MARCOS CORRÊA NEVES * marcos@dpi.inpe.br
FREDERICO ROMAN RAMOS ** fred@ltid.inpe.br,
EDUARDO CELSO GERBI CAMARGO ** eduardo@dpi.inpe.br,
GILBERTO CÂMARA ** gilberto@dpi.inpe.br,
ANTÔNIO MIGUEL MONTEIRO ** miguel@dpi.inpe.br,

* Embrapa Meio Ambiente
Rod. SP 340, km 127,5 , CEP 13820-000, Campinas(SP)
Tel.: (019) 3867- 8700 - Fax.: (019) 3867 8740

** Instituto Nacional de Pesquisa Espacial - INPE
Av. dos Astronautas, 1758, CEP:12201-027 - São José dos Campos(SP)
tel.: (012) 345 6444, fax.: (012) 345 6468

RESUMO

Este trabalho apresenta um conjunto de ferramentas de análise exploratória de dados espaciais (ESDA) existentes no SPRING. Estas ferramentas ampliam a capacidade do analista em extrair informações de seu conjunto de dados, permitindo-lhe uma melhor compreensão da dinâmica espacial existente no fenômeno estudado. Elas são baseadas no conceito de autocorrelação espacial, sendo aplicáveis aos objetos espaciais com área definida e atributos numéricos associados. Estas ferramentas geram como resultados, índices globais e locais que fornecem uma medida da associação espacial, além de gráficos e mapas auxiliares. Com estes dispositivos, o analista pode compreender melhor os padrões de associação espacial, visualizar, identificar e classificar agrupamentos de objetos com valores de atributos semelhantes, áreas de transição e situações atípicas. Como forma de demonstrar a utilidade destas técnicas, elas foram aplicadas a um conjunto de dados sócio-econômicos, referentes aos distritos da cidade de São Paulo. Os resultados obtidos, índices e dispositivos gráficos, são apresentados e discutidos.

ABSTRACT

This paper presents a set of tools of exploratory spatial data analysis (ESDA) existent in SPRING. These tools increase the capacity of extraction of additional information of the spatial data, making possible the analyst to understand the dynamics spatial present in your study. The tools are based on the concept of spatial autocorrelation, being applicable to area data, spatial objects with defined area and numeric attributes. These tools generate as results, global and local indexes, that supply measures of the space association, scatter plots and auxiliary maps. these devices help the analyst to identify patterns of space association, to visualize, to identify and to classify clusters of objects, transition areas, atypical locations and spatial outliers. As example of the usefulness of the ESDA techniques, the tools were applied to socioeconomic data, referring to the districts of the city of São Paulo. The indexes, graphs and maps resultants are presented and discussed**.**

1- INTRODUÇÃO

Recentes avanços em computação, referentes tanto à capacidade de processamento dos equipamentos, quanto à evolução dos sistemas de informação geográfica - SIG, aumentaram as possibilidades do emprego de técnicas mais elaboradas de análise espacial. Em paralelo a estes desenvolvimentos, algumas técnicas de estatística foram sendo definidas e adaptadas aos dados espaciais. Estas técnicas, combinadas com funções de visualização, formam, em alguns SIGs atuais, um conjunto de ferramentas que suporta a análise exploratória de dados espaciais *(Exploratory Spatial Data Analysis - ESDA).* Este conjunto de ferramentas é definido na literatura como sendo uma coleção de técnicas para descrever e visualizar distribuições espaciais, identificar situações atípicas, descobrir padrões de associação espacial, agrupamento de valores semelhantes (*clusters*) e sugerir regimes espaciais ou outras formas de heterogeneidade espacial (Anselin & Bao, 1997).

Anselin (1998) apresenta algumas formas e exemplos de integração de SIGs com ferramentas de análise exploratória, dando destaque ao acoplamento entre o SpaceStat e o ArcView. O SpaceStat é um programa especializado em análise estatística de dados espaciais, sem contudo, possuir funções de entrada e saída de dados.

Neste trabalho, utilizou-se um conjunto de ferramentas de análise exploratória existentes no SPRING (Sistema para Processamento de Informações Georeferenciadas). O SPRING é desenvolvido pelo INPE (Instituto Nacional de Pesquisas Espaciais) e está disponível para plataformas UNIX e Windows. As ferramentas, aqui utilizadas, são apresentadas e discutidas quanto a sua aplicação e utilização na análise. Elas atuam como forma de extração e visualização de informações não diretamente perceptíveis ao analista, quando este utiliza procedimentos comuns de classificação e visualização de dados espaciais. As técnicas são baseadas no conceito de autocorrelação espacial e são aplicáveis à abjetos-área. Este tipo de objeto espacial, possui um ou um conjunto de atributos numéricos associados e são representadas, espacialmente, por linhas poligonais fechadas.

Como forma de exemplificar as utilidades destas técnicas, elas foram aplicadas a alguns dados sócio-econômicos. Estes dados são um oriundos do mapeamento da exclusão social da cidade de São Paulo, conduzido por SPOSATI (1996) e referem-se aos 96 distritos do município de São Paulo.

Os resultados obtidos são apresentados em forma de índices que medem a associação espacial (índice de Moran), global e local, gráficos de espalhamento e mapas. Estes dispositivos auxiliaram na identificação de agrupamentos de objetos, de altos e baixos valores, áreas de transição e casos atípicos.

2 - MATERIAL E MÉTODOS

O trabalho conduzido por SPOSATI (1996) considerou uma grande quantidade de variáveis para o estabelecimento de um índice de exclusão/inclusão social para os distritos de São Paulo. Aqui, utilizamos um pequeno subconjunto destas informações, além do índice resultante, como forma de identificar e comparar os regimes espaciais presentes nestes atributos. Foram escolhidos, inicialmente, cinco atributos:

- percentagem de domicílios precariamente servido por água;
- percentagem de domicílios precariamente servido por esgoto;
- percentagem de chefes de famílias com instrução acima dos 15 anos;
- percentagem de chefes de família com renda inferior a 1,5 salário mínimo;
- índice de exclusão social.

Tanto as informações vetoriais, descrevendo os limites dos distritos, como seus atributos alfanuméricos, encontravam-se em formato digital e foram importadas para o SPRING. Cada polígono, representando graficamente um distrito, foi associado a um registro de uma tabela de dados, onde as colunas da tabela continham os atributos (dados sócio-econômicos) do distrito correspondente. Dentro do universo conceitual do SPRING, este tipo de dado é denominado de *dado cadastral* (Câmara & Medeiros, 1999).

Em todas as técnicas de ESDA empregadas neste trabalho, estão presentes três elementos básicos: a matriz de proximidade espacial (***W***), o vetor de desvios (***Z***) e o vetor de médias ponderadas (***Wz***).

A matriz de proximidade espacial é uma ferramenta geral e bastante útil para descrever o arranjo espacial dos objetos (Bailey & Gatrell, 1995). ***W***, é uma matriz quadrada, com n^2 elementos, onde cada elemento, w_{ij}, representa uma medida de proximidade espacial entre o polígono ***i*** e o polígono ***j***, sendo ***n,*** o número total de objetos. Neste trabalho, utilizou a seguinte medida de proximidade:

objetos com fronteira comum, $w_{ij} = 1$;

objetos sem fronteira comum, $w_{ij} = 0$.

Para o cálculo do vetor de desvios, Z, é calculada, primeiramente, a média (μ) dos valores dos atributos, considerando os ***n*** objetos. Cada elemento ***i*** de Z, z_i, é obtido subtraindo-se o valor da média, do valor do atributo correspondente ($z_i = y_i - \mu$).

O terceiro elemento básico, o vetor de médias ponderadas (***Wz***), é obtido pela multiplicação do vetor transposto dos desvios, pela matriz de proximidade espacial com linhas normalizadas, onde cada elemento de uma linha i qualquer, originariamente com valor 1, é dividido pelo número de elementos não nulos da mesma linha. Desta forma, como resultado, cada elemento wz_i, contém um valor correspondente à média dos desvios dos vizinhos ao objeto i.

Os três elementos, ***W***, ***Z*** e ***Wz***, são gerados automaticamente pelo SPRING, sendo necessário indicar apenas o atributo para o qual serão realizados os cálculos. Inicialmente, o SPRING constrói a matriz de proximidade espacial utilizando informações da topologia, extraída da representação gráfica dos objetos. O cálculo de ***Z***, é realizado utilizando os valores, contidos na coluna da tabela de dados correspondente ao atributo selecionado. E, por último, é calculado a média dos vizinhos, por uma operação de multiplicação matricial (Wz = WxZ).

Estes elementos básicos são usados para gerar os resultados (índices, e classificações) que serão utilizados em conjunto com as ferramentas de visualização do SPRING. A seguir, descrevemos as ferramentas de análise exploratória disponível.

Índice Global de associação espacial: índice de Moran: I

O Índice de Moran fornece uma medida geral da associação espacial existente no conjunto dos dados. Seu valor varia de –1 a 1. Valores próximos de zero, indicam a inexistência de autocorrelação espacial significativa entre os valores dos objetos e seus vizinhos. Valores positivos para o índice, indicam autocorrelação espacial positiva, ou seja, o valor do atributo de um objeto tende a ser semelhante aos valores dos seus vizinhos. Valores negativos para o índice, por sua vez, indicam autocorrelação negativa.

O índice de Moran pode ser descrito em função dos elementos básicos vistos anteriormente, é dado por:

$$I = \frac{\mathbf{Z}^t . \mathbf{W}_z}{\mathbf{Z}^t . \mathbf{Z}}$$

onde:

Z^t: é o vetor de desvios transposto.

Índice Local de Associação Espacial (LISA)

Enquanto os indicadores globais, como o índice de Moran, fornecem um único valor como medida da associação espacial para todo o conjunto de dados, os indicadores locais produzem um valor específico para cada objeto, permitindo assim, a identificação de agrupamentos de objetos com valores de atributos semelhantes (*clusters*), objetos anômalos (*outliers*) e de mais de um regime espacial. Segundo Anselin (1994), um LISA tem que atender à dois objetivos:

i) permitir a identificação de padrões de associação espacial significativos;

ii) ser uma decomposição do índice global de associação espacial.

O Lisa utilizado neste trabalho é o índice local de Moran. Uma das formas de representação deste índice, em função dos elementos básicos, é:

$$I_i = z_i . Wz_i / \sigma^2$$

onde:

I_i : índice local para o objeto i;

z_i : valor do desvio do objeto i;

Wz_i : valor médio dos desvios dos objetos vizinhos de i;

σ^2 : variância da distribuição dos valores dos desvios.

Existem variações possíveis para a formula acima. Quando apresentada desta forma, o valor do índice global de Moran, é a média aritmética dos índices locais.

Gráfico de Espalhamento de Moran

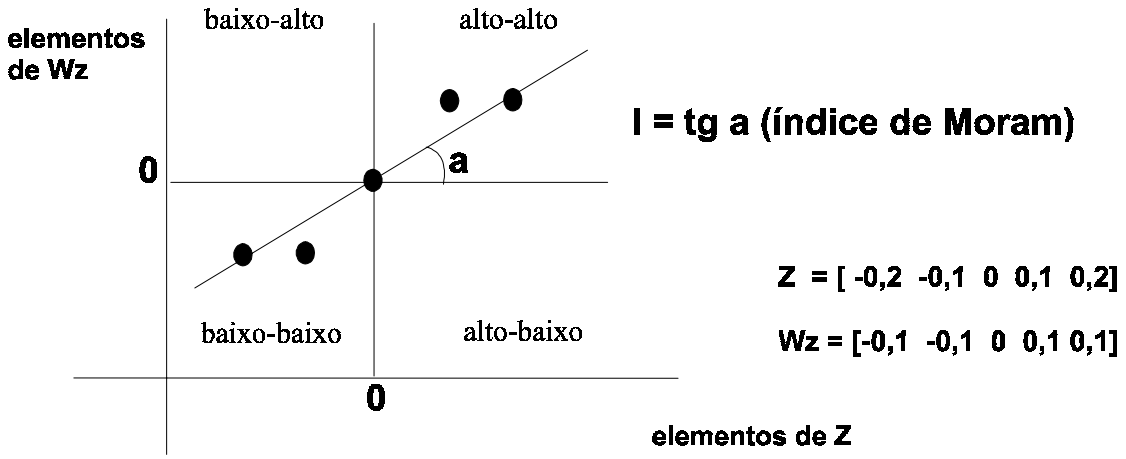
Como vimos anteriormente, o índice de Moran global, na forma matricial é dado por:

$$I = \frac{Z^t . W_Z}{Z^t . Z}$$

.

I é formalmente equivalente ao coeficiente de regressão linear. Este coeficiente indica a inclinação da reta de regressão (βo) de W_z em Z (Neter & Wasserman, 1974). A interpretação do índice de Moran como um coeficiente de regressão, indica o caminho para se construir um dispositivo gráfico para visualizar a associação espacial entre o valor do atributo de cada elemento (z_i) com a média dos valores dos atributos dos seus vizinhos (Wz_i). Este procedimento é denominado de gráfico de espalhamento de Moran (Anselin, 1996). A Figura 1, apresenta como o gráfico de espalhamento é construído.

Figura 1: Construção do gráfico de espalhamento de Moran



Mapa de barras (Z x W_z)

Este interessante mecanismo de visualização é apresentado em Anselin & Bao (1997) como uma forma de percepção da associação espacial. Este dispositivo permite a visualização simultânea do valor relacionado ao atributo do objeto e do valor correspondente ao valor médio dos atributos dos objetos vizinhos, com o uso de duas barras gráficas sobre a área correspondente ao objeto, no mapa. A altura das barras são proporcionais aos valores do atributo do objeto (ou o desvio) e à média dos vizinhos. Ambas informações podem ser obtidas facilmente dos elementos básicos, os vetores *Z* e W_z.

Box map, Lisa map e Moran map

Estes três dispositivos gráficos de visualização são baseadas nos resultados obtidos para os indicadores locais e do gráfico de espalhamento de Moran. No *box map*, cada objeto é classificado conforme sua posição em relação aos quadrantes do gráfico de espalhamento, recebendo uma cor correspondente no mapa gerado. Na geração do *LISA map*, é avaliada a significância dos valores do índice de Moran Local obtido para cada objeto, em relação à hipótese de não existência de autocorrelação espacial (hipótese nula). Na avaliação da significância é utilizada a abordagem de permutação dos atributos dos vizinhos, conforme descrito em Anselin (1995). Os objetos são classificados em quatro grupos: não siginificantes; com significânia entre 0,05 e 0,01; com significância entre 0,01 e 0,001; e maior que 0,001.

No *Moran map,* de forma semelhante ao *LISA map*, somente os objetos para os quais os valores de LISA foram considerados significantes ($p > 0,05$), são destacados, porém, aparecem classificados em quatro grupos, conforme sua localização no quadrante do gráfico de espalhamento. Os demais objetos, ficam classificados como *sem significância*.

3 - RESULTADOS

Aplicamos as técnicas descritas acima, ao conjunto de atributos sócio-econômicos, como forma de ilustrar a sua utilização prática. Primeiramente, calculamos o índice global de Moran, para os cinco atributos. A Tabela 1, mostra os resultados. Podemos observar que todos os índices são positivos, indicando existir, em todos os casos, uma autocorrelação espacial positiva. O maior índice foi obtido para a variável PERINST que representa a percentagem de chefes de família com mais de quinze anos de estudo, enquanto que o menor índice, foi obtido para o atributo PERAGUA, que representa a percentagem de domicílios precariamente abastecido de água.

Tabela 1: Índice global de autocorrelação espacial para algumas variáveis socio-econômicas.

Atributo	**Índice Moran Global**
índice de exclusão social	0,625
percentagem de chefe de família com mais de 15 anos de estudo	0,764
percentagem de chefes de família com renda inferior a 1,5 salário mínimo	0,640
percentagem de domicílios sem esgoto	0,625
percentagem de domicílios sem abastecimento de água	0,317

Na Figura 2, são apresentados os mapas, classificados em 6 quantis, para as variáveis PERAGUA e PERINST a fim de visualizarmos os diferentes regimes espaciais existentes para as duas variáveis, comportamentos estes, que foram refletidos em seus respectivos índices de associação espacial.

O gráfico de espalhamento de Moran, Figura 3, também apresenta comportamentos bens distintos para as duas varáveis. No caso de PERINST, há uma forte relação entre o valor do desvio, z_i, e os valores dos vizinhos, Wz_i. Já para o atributo, PERAGUA, existe uma forte concentração de pontos na região inferior-esquerda do gráfico, refletindo baixos valores de z e Wz. Três distritos apresentam comportamentos bem distintos dos demais. Estes três distritos, localizados geograficamente ao sul do município de São Paulo, são: Marsilac, Parelheiros e Grajaú. A situação discrepante destes distritos, apontada no gráfico de espalhamento, pode ser explicada pelo alto índice de moradias sem água tratada em Marsilac (99%) e Parelheiros (39%), destoando totalmente da situação dos demais distritos. A posição do distrito Grajaú, no gráfico de espalhamento (canto inferior esquerdo), é determinada pelos altos valores dos atributos dos seus vizinhos, Marsilac e Parelheiros.

A Figura 4, apresenta um mapa de barras, para a variável PERINST. Esta ferramenta apresenta duas barras sobre cada objeto (distrito), uma com altura proporcional ao valor do desvio do distrito (barra vermelha), e a outra barra (verde), proporcional ao valor médio dos desvios dos distritos vizinhos. Esta ferramenta permite visualizar facilmente os agrupamentos de altos e baixos valores, onde as barras aparecem com alturas semelhantes e, também as áreas de transição e situações atípicas (barras com alturas distintas). A Figura 4.b, mostra em detalhes, os agrupamentos de distritos das áreas central e oeste da cidade de São Paulo.

O *box map* para a variável PERIST, é apresentado na Figura 5. Os distritos da classe 1, possuem os atributos acima da média (desvios positivos) e distritos vizinhos com média de atributos também positivos (quadrante alto-alto no gráfico de espalhamento). Na classe 2, é o caso oposto, estão os distritos com atributo e média dos vizinhos abaixo da média global (quadrante baixo-baixo). Na classe 3, estão os distritos com valores acima da média e valor médio dos atributos dos vizinhos menor que a média (quadrante alto-baixo). Finalmente, na classe 4 estão os distritos com valores abaixo da média e valor médio dos atributos dos vizinhos acima da média (baixo-alto). A Figura 5, nos mostra que, para a variável PERIST, os distritos da classe 1 (alto-alto) aparecem absolutamente concentrados na região central da cidade. Os distritos da classe 2 (baixo-baixo) aparecem no sul, oeste e norte da cidade. Os distritos pertencentes às classes 3 e 4, estão localizados nas zonas de transição entre o agrupamento de altos valores e os agrupamentos de baixos valores.

Na Figura 6, é apresentado a LISA Map, também relacionado à variável PERIST, com os distritos classificados em função da significância dos valores de seus índices locais, em: não significantes; significância entre 0,05 e 0,01; entre 0,01 e 0,001; e maior que 0,001. Esta ferramenta permite identificar os agrupamentos de distritos com valores de Lisa

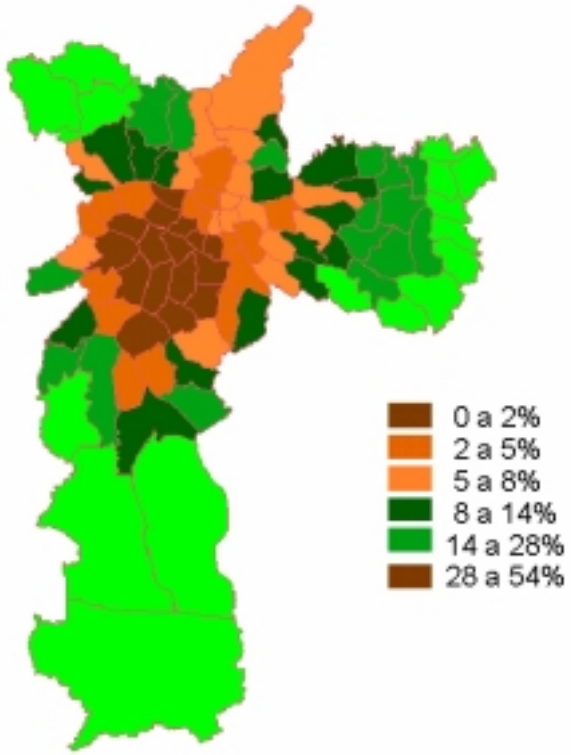


Figura 2.a: Percentagem de chefes de família com mais de 15 anos de instrução.

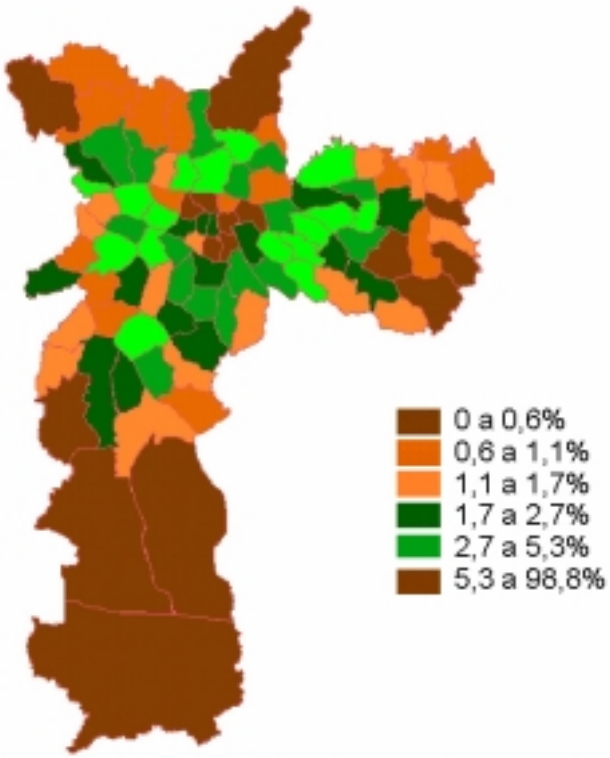


Figura 2.b: Percentagem de domicílios sem abastecimento água.

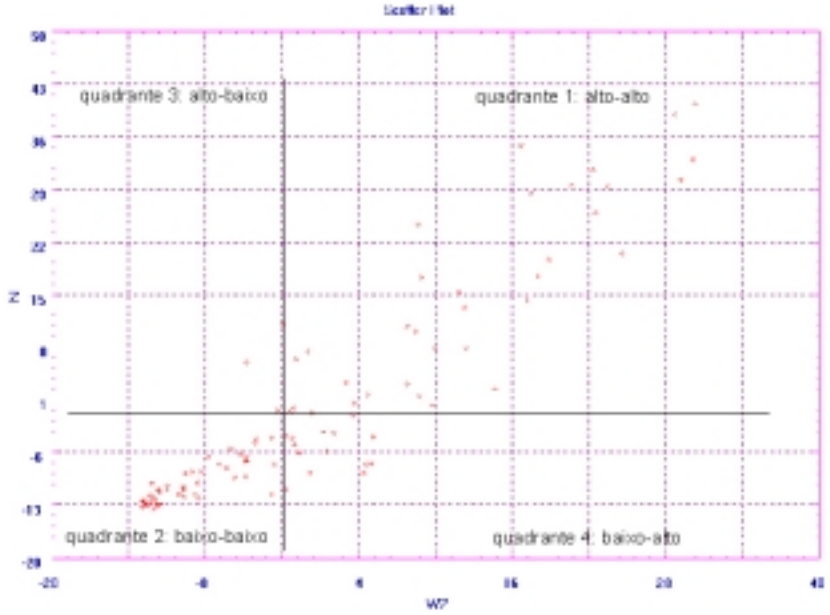


Figura 3.a: Gráfico de espalhamento de Moran (Z x Wz) para a variável PERINST.

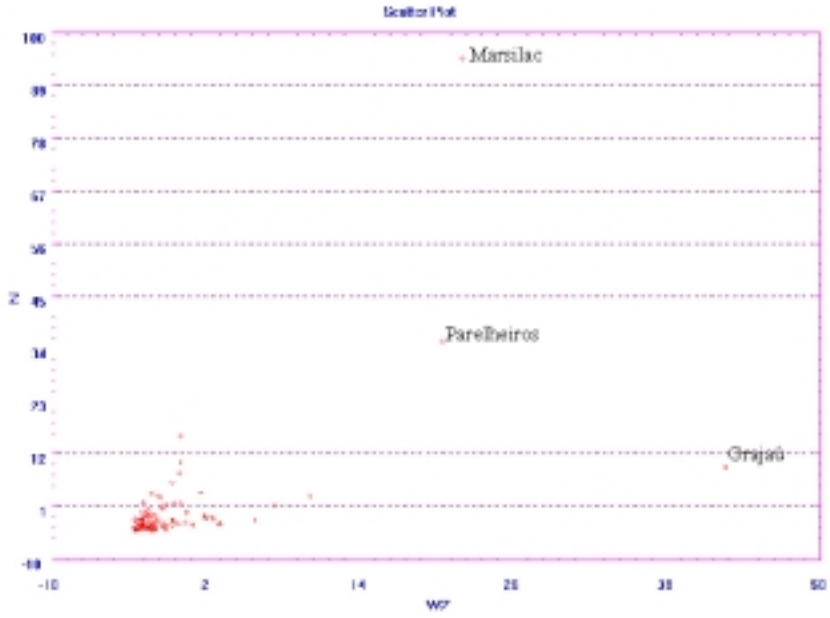


Figura 3.b: Gráfico de espalhamento de Moran, para a variável PERAGUA.

Figura 4.a: Mapa de barras (desvio x média dos desvios vizinhos).

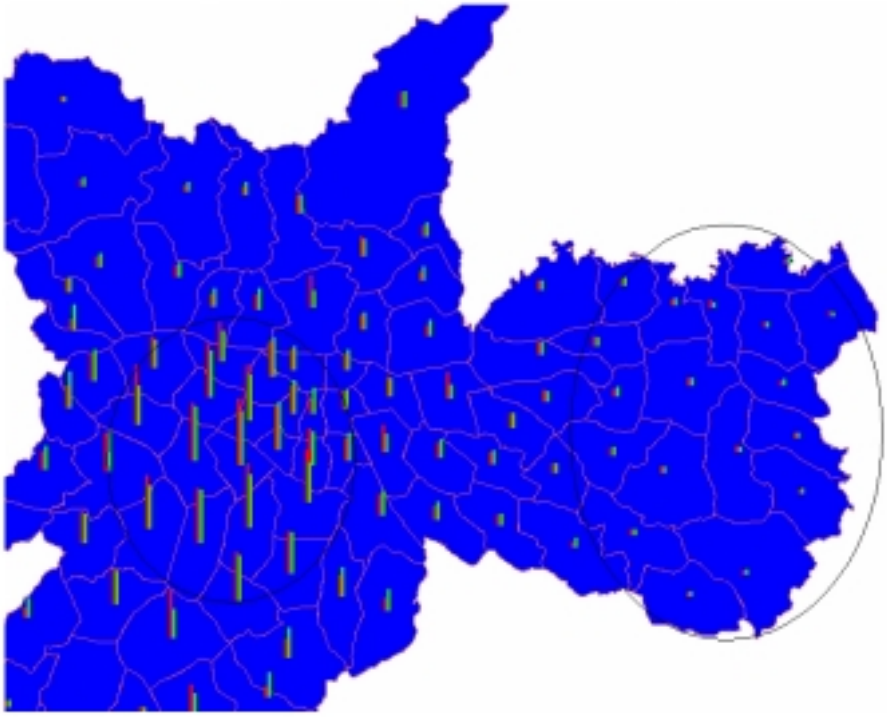


Figura 4.b: detalhe dos agupamento de altos valores (regial central) e baixos valores (oeste)

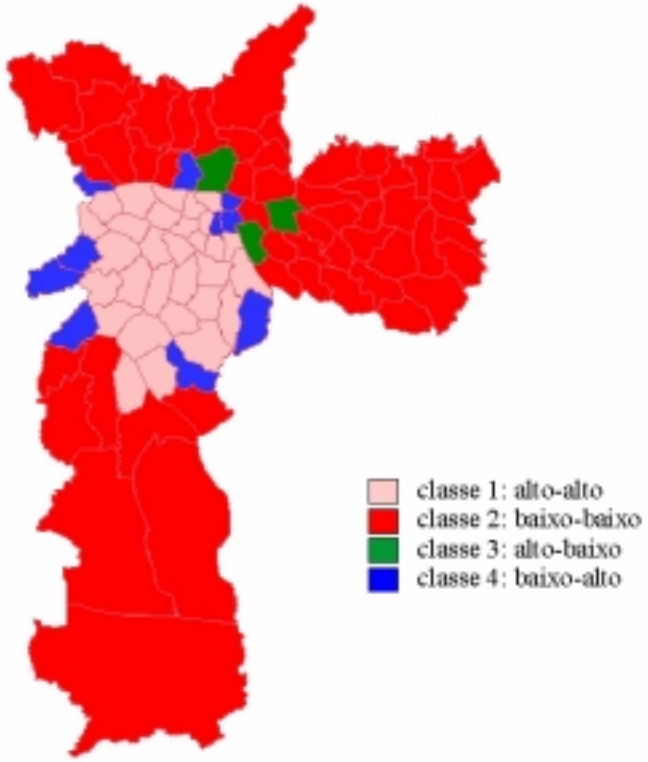


Figura 5: Box map - Distritos classificados conforme sua posição no gráfico de espalhamento de Moran.

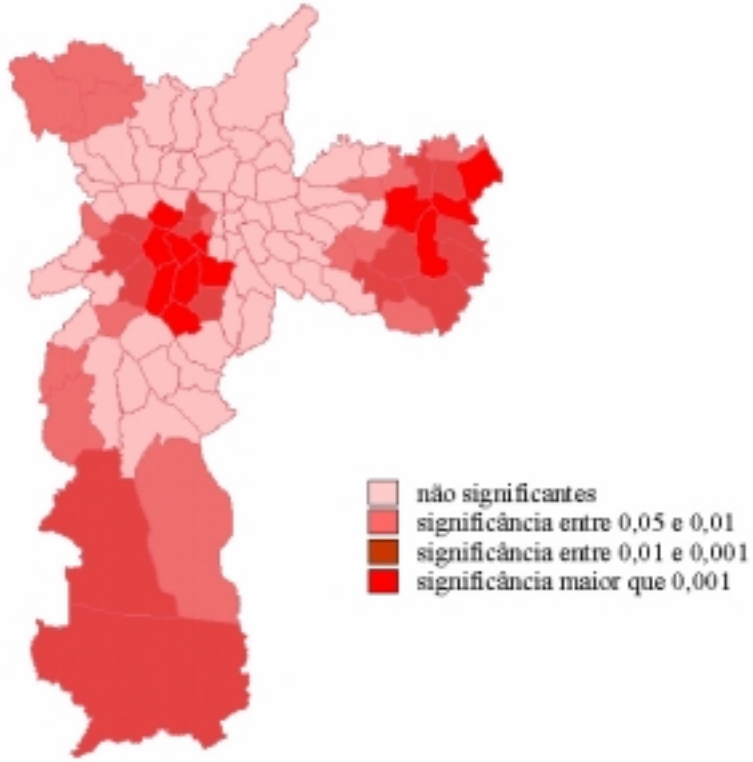


Figura 6: Lisa map - Distritos classificados conforme a significância do índice local de associação espacial.

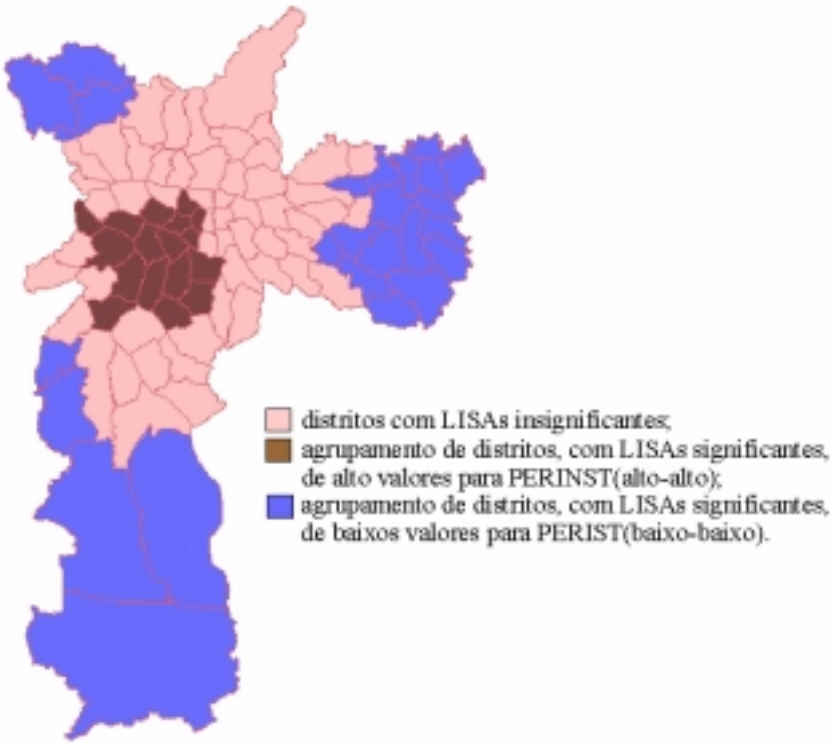


Figura 7.a: *Moran map*, atributo PERINST, indicando "clusters" de distritos de alto e baixos valores, com Lisa significantes.

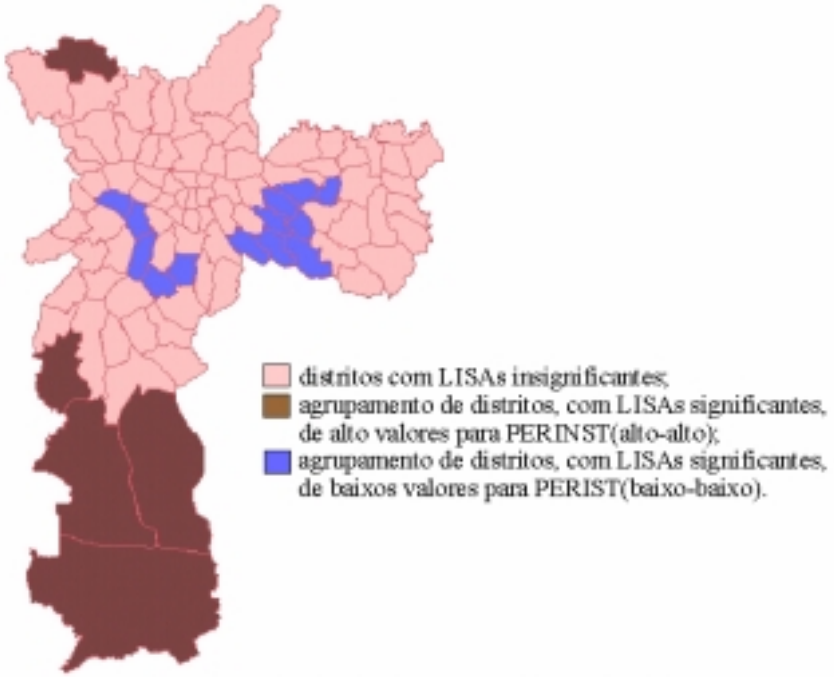


Figura 7.b: *Moran map*, atributo PERAGUA, indicando "clusters" de distritos de alto e baixos valores, com Lisa significantes.

significantes, sem contudo diferencia-los em agrupamentos de altos ou baixos valores.

O resultado obtido pela aplicação da última ferramenta, *Moran map*, é apresentado na Figura 7. Ele combina os dois resultados anteriores, apresentando todos os distritos com significância maior que 0,05, classificados em quatro classes, conforme sua posição no gráfico de espalhamento. Desta forma, podemos através desta ferramenta, destacarmos os distritos com significativo índice de autocorrelação espacial e ao mesmo tempo, sabermos que se trata de um agrupamento de alto ou baixo valor. A Figura 7.a, refere-se ao atributo PERINST, e a 7.b, ao atributo PERAGUA.

4 - CONCLUSÃO

Para os dados sócio-econômicos analisados, foram produzidos, como resultado, uma série de índices de associação espacial, gráficos e mapas. Embora, neste trabalho, o objetivo não fosse analisar os dados e sim as ferramentas disponíveis no SPRING, foi possível verificar que este conjunto de ferramentas aumenta a possibilidade de compreensão da dinâmica espacial dos dados e contribui para o embasamento de hipóteses que explicassem a distribuição e relação espacial dos dados.

Os cinco atributos analisados apresentarem autocorrelação positiva, indicando existir uma associação espacial, tendendo haver semelhança entre o valores dos atributos dos distritos fisicamente mais próximos.

As técnicas utilizadas mostraram-se úteis na identificação de agrupamentos contínuos (*clusters* de distritos), de altos e baixos valores, áreas de transição entre *clusters*. Também foi possível, avaliar os agrupamentos quanto ao nível de significância da associação espacial, selecionando os agrupamentos mais importantes.

O trabalho permitiu concluir que o conjunto de ferramentas avaliadas fornece, de fato, a possibilidade de explorar os dados espaciais, extraindo informações adicionais não diretamente perceptíveis, quando se utiliza as técnicas de classificação e visualização comuns.

BIBLIOGRAFIA

Anselin, L. Exploratory Spatial Data Analysis in a Geocomputational Environment. In: Longley, P. A.; Brooks; S. M.; MCDONNELL, R.; MACMILLIAN; B**.** ***Geocomputation a primer.*** Chichester: John Willey & Sons Ltd, 1998, p.77-94.

Anselin, L. & Bao, S. Exploratory Spatial Data Analysis Linking SpaceStat and ArcView. In: Fischer, M. M. & Getis, A. ***Recent developments in spatial analysis***. New York: Springer, 1997, p. 35-59.

Anselin, L. The Moran scatterplot as ESDA tool to assess local instability in spatial association. In: Fisher, M.; Scholten, H. J.; Unwin, D. ***Spatial Analytical Perspectives on GIS***. London: Taylor & Francis, 1996, p. 111-126.

Anselin, L. Local Indicators of Spacial Association – LISA. Geographical Analysis. v.27, n.2, p.93-115, 1995.

Bailey, T. C. & Gatrell, A. C. Interactive spatial data analisys. Harlow: Longman, 1995.

Câmara, G. & Medeiros, J. S. Modelagem de dados em geoprocessamento. In: Assad, E. D. & Sano, E. E**. Sistemas de Informações Geográficas. Aplicações na agricultura.** Brasília: Embrapa, 1999, p. 47-65.

Neter, J & Wasserman, W. Applied Linear Statistical Models - Regression, Analysis of Variance and Experimental Designs. Homewood: Richard Irwin, 1974.

Sposati, A. Mapa da Exclusão/Inclusão Social da Cidade de São Paulo. São Paulo: EDUC, 1996.