Análise Espacial Intra-Urbana em São José dos Campos com Mapas Auto-Organizáveis

Marcos A.S. da Silva[1,2] , Antônio M.V. Monteiro[1] , José S. de Medeiros[1]
[1]*Instituto Nacional de Pesquisas Espaciais*
Divisão de Processamento de Imagens
São José dos Campos, SP
{miguel,simeao} @dpi.inpe.br
[2]*Laboratório de Geotecnologias Aplicadas*
Embrapa Tabuleiros Costeiros, Aracaju, SE
aurelio@cpatc.embrapa.br

Resumo

Este trabalho aplicou a rede neural do tipo Mapa Auto-Organizável (SOM) na tarefa de análise exploratória de dados espaciais multivariados. O SOM foi usado sobre o problema de exclusão/inclusão social urbana em São José dos Campos-SP. Os resultados mostraram que o SOM é uma ferramenta eficiente e de fácil aplicação.

Abstract

This work applied the Self-Organizing Map (SOM) neural network in the task of exploratory multivariate spatial data analysis. The SOM was used in the problem of urban social exclusion/inclusion in São José dos Campos-SP. Through results we realized that SOM is a very efficient and easy to use tool for this kind of exploratory analysis.

1. Introdução

A capacidade para geração, armazenamento e recuperação de dados, com referência no espaço e no tempo, cresceu muito nos últimos anos. Contribuíram para isto: A ampliação da oferta de dados de satélites, em várias resoluções espaciais, espectrais e temporais e de Mapas Urbanos Básicos Digitais (MUB) para diversas cidades; a possibilidade de coleta direta de dados posicionais com o uso de sistemas GPS (*Global Positioning Systems*); a facilidade de acesso a um conjunto bem mais amplo de dados demográficos e ambientais, como é o caso do censo 2000, IBGE, com a malha de setores censitários disponível por municípios.

As tecnologias da informação para lidar com grandes bases de dados, em particular, a tecnologia dos SGBDs (Sistemas Gerenciadores de Bancos de Dados) e as de Sistemas de Informação Geográfica (SIG) permitiram acomodar parte desta capacidade geradora de dados posicionais, com a possibilidade de armazenamento duradouro e com sua recuperação simples, mais eficiente e facilitada. No entanto, a nossa capacidade de analisar este conjunto de dados, em várias escalas e com existência em unidades espaciais, por vezes distintas, é bem menor que a nossa capacidade de produzí-lo.

Neste trabalho os Mapas Auto-Organizáveis de Kohonen serão usados para a tarefa de análise exploratória de dados espaciais de área. A análise de exclusão/inclusão social urbana em São José dos Campos a partir de dados censitários foi usada como estudo de caso. O principal objetivo é verificar se os resultados obtidos pelo SOM são coerentes com resultados obtidos por técnicas estatísticas aplicadas anteriormente [8].

2. Análise Espacial de Área

O estudo de caso deste trabalho assim como a maior parte das aplicações do SOM na Análise Espacial trabalha com Análise Espacial de Área. Este tipo de análise considera a análise de dados associada com zonas espaciais ou áreas. Estas áreas podem estar dispostas de forma regular, como em imagens de sensores remotos, ou ser um conjunto de áreas irregulares, como áreas de distritos administrativos ou de setores censitários. Os atributos associados com estas áreas não variam continuamente em função do espaço. As áreas consideradas são a única posição espacial na qual os atributos podem ser medidos. [2]

O principal objetivo da Análise Espacial de Área é a detecção e possível exploração de padrões espaciais ou tendências nos valores dos atributos. Dada uma região de estudo *R*, particionada em subáreas $A_i,...,A_n$ com $A_1 \cup ... \cup A_n = R$, têm-se o vetor de características $x(A_i)$, $A_i \in \{A_1,...,A_n\}$. Neste trabalho este vetor de características será denotado por x_i, ver Figura 1.

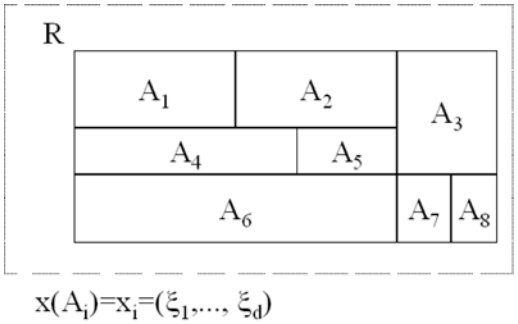

Figura 1. **Elementos da análise espacial de área**

Existem várias formas para visualização deste tipo de análise de dado geo-espacial. Neste trabalho usou-se padrão de cores para colorir os objetos geográficos de forma a realçar possíveis padrões.

3. Mapa Auto-Organizável

O Mapa Auto-Organizável de Kohonen é uma rede neural de aprendizagem competitiva organizada em duas camadas, [11]. A primeira camada representa o vetor dos dados de entrada, x_k , a segunda corresponde a uma grade de neurônios, geralmente bidimensional, totalmente conectada aos componentes do vetor de entrada. Cada neurônio possui um vetor de código associado, w_j.

O processo de aprendizagem consiste de três fases. Na primeira fase, competitiva, cada padrão de entrada é apresentado a todos os neurônios para que aquele mais próximo do padrão apresentado seja o vencedor. Na segunda fase, cooperativa, é definida a vizinhança relativa ao neurônio vencedor. Na terceira fase, adaptativa, os vetores de código do neurônio vencedor e dos seus vizinhos serão alterados segundo algum critério de atualização.

Após a aprendizagem os vetores de código do SOM corresponderam a uma aproximação não-linear dos padrões de entrada. O SOM também preserva a formação topológica dos padrões de entrada, ou seja, padrões próximos no conjunto amostral estarão relacionados a neurônios próximos na grade neural.

O SOM pode variar em algoritmos de aprendizagem, estrutura topológica da grade, função de vizinhança, parametrização inicial etc.

3.1 Matriz de Distância Unificada (U-matriz)

A matriz de distâncias unificada, U-matriz [17] tem o objetivo de permitir a detecção visual das relações topológicas entre os neurônios. Usa-se a mesma forma de cálculo de distância usada no treinamento para calcular a distância entre os vetores de código dos neurônios adjacentes.

O resultado gerado a partir da aplicação da U-matrix sobre o mapa é uma imagem *f(x,y)* onde o nível de intensidade de cada pixel corresponde a uma distância calculada. Um mapa 2-D *NXM* gera uma imagem *(2N-1)X(2M-1)*.

Dado um mapa bidimensional hexagonal encontra-se a U-matriz calculando-se as distâncias *dx, dy e dz*, ver Figura 2, para cada neurônio. O valor *du* da U-matriz é calculado em função dos valores dos elementos circunvizinhos do neurônio relativo ao *du*. O valor *du* pode ser a média, mediana, valor máximo ou mínimo destes valores. O processo é análogo para o caso de uma rede bidimensional retangular.

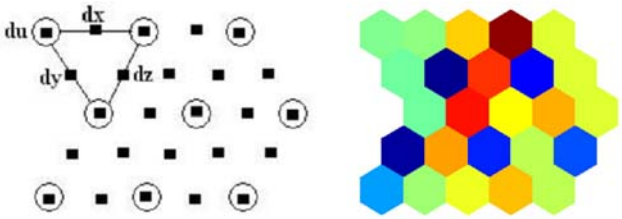

Figura 2. **Exemplo de geração da imagem relativa a U-matriz a partir de uma rede 3x3 hexagonal**

A matriz de distância unificada pode ser interpretada como uma imagem através da coloração dos pixels de acordo com a intensidade de cada componente da matriz. Valores altos correspondem a neurônios vizinhos dissimilares e valores baixos correspondem a neurônios vizinhos similares. Regiões com baixos valores do gradiente correspondem a vales que agrupam neurônios especializados em padrões similares. Regiões com valores altos correspondem a fronteiras entre agrupamentos.

3.2 Planos de Componentes

Para que se possa ter uma noção de como cada componente do vetor de característica x_k se organizou no Mapa treinado usa-se algum método de coloração do SOM baseado nos valores de cada componente. Para um dado componente *j*, de uma Mapa bidimensional *MxN*, gera-se uma imagem *f(x,y)* com dimensões iguais ao do Mapa, *MxN*, onde cada pixel corresponderá ao valor do componente *j* na posição *(x,y)*. Para imagens em escalas de cinza pode-se convencionar o branco para valores máximos, o preto para valores mínimos e tons de cinza

para valores intermediários. A ilustração da Figura 3 mostra uma plano de componente para uma rede hipotética de dimensões 3x3 hexagonal.

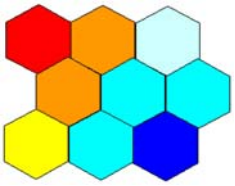

Figura 3. **Considerando uma rede 3x3 pode-se ter a seguinte configuração para um determinado componente do vetor de código. Aqui o neurônio mais acima e a esquerda indica que o valor deste componente neste neurônio é elevado. Para o neurônio mais abaixo e a direita tem-se que o valor do componente é o mais baixo do conjunto de neurônios.**

4. Análise Espacial com SOM

Os mapas auto-organizáveis têm se mostrado bastante útil na Análise Espacial. Haja vista o crescente número de publicações sobre o assunto presente na literatura, [2], [9], [3], [18]. Sua principal função é atuar como um mecanismo não-supervisionado de mapeamento de dados multivariados numa grade de dimensão menor, resguardando as propriedades dos dados originais. Sua simplicidade conceitual aliada a suas variantes estruturais e de aprendizagem tem proporcionado variedades de aplicações.

Porém, é a partir da propriedade de geração de mapas topologicamente ordenados que os trabalhos de uso da rede SOM na Análise Espacial têm sido desenvolvidos. [12], [13] considera que este tipo de rede é extremamente útil para análise de dados geográficos cujas propriedades impedem que sejam usados métodos estatísticos. Segundo Openshaw, problemas como análise de dados multivariados, dependência de incerteza sobre os dados, distribuições não normais das variáveis etc, podem ser convenientemente tratados com as RNA, em especial a rede SOM.

Observou-se que os trabalhos de aplicação dos mapas auto-organizáveis na Análise Espacial, [19], [5], [7], [12], [16], [10], [4], [6], [15], apresentam algumas características comuns, como o uso do algoritmo padrão de treinamento. Em geral são usados os modelos com topologias bidimensionais, pois permitem a visualização natural dos agrupamentos. O que difere um trabalho do outro é a forma de interpretação da formação topológica no mapa neural. O que aumenta a importância da necessidade do especialista na área de aplicação para entendimento semântico dos agrupamentos.

Foi possível observar que existi uma ampla variedade de formas de se explorar dados multivariados a partir de redes neurais do tipo SOM. Para o caso não-supervisionado pode-se: (a) usar a U-matriz para descobrir manualmente agrupamentos de dados; (b) usar os Planos de Componentes para descobrir relações e tendências entre as variáveis; (c) usar uma rede com poucos neurônios e considerar que cada neurônio corresponde a um agrupamento. Neste trabalho usou-se as técnicas (a) e (b).

5. Estudo de Caso

A análise de exclusão/inclusão social urbana em São José dos Campos-SP foi baseada nos estudos conduzidos por Genovez em sua dissertação de mestrado, [8].

A metodologia consiste na análise de atributos associados aos setores censitários da área urbana de São José dos Campos. Cada setor censitário possui um conjunto de atributos relativos a dados do IBGE que correspondem a questões relacionadas com o nível de vida daquela população.

Genovez propôs uma metodologia dividida em três fases: análise quantitativa dos dados para transformar percentuais do censo em índices; análise qualitativa para verificar a correlação e significância de componentes; e análise espacial. Este trabalho atuou sobre a segunda fase.

6. Material

Para proceder com as simulações usou-se os sistemas CΛS*Λ* e o pacote SOM ToolBox. Para geração dos gráficos usou-se o Excel 2000.

7. Seleção dos Dados e Pré-processamento

Para este estudo selecionou-se os índices de Distribuição de Renda dos Chefes de Família (ARENDR), Desenvolvimento Educacional (DESEDUCR), Estímulo Educacional (ESTEDUCR), Longevidade (LONGR), Qualidade Ambiental (QAMBR), Conforto Domiciliar (QDOMR) e Mulheres não Alfabetizadas (MANALFR), Concentração de Mulheres Chefes de Família (MCHFR).

Todo o conjunto de dados compreende um total de n=342 padrões de dimensionalidade iguala a d=8.

8. Configuração da rede SOM

Definiu-se um conjunto fixo de configurações de rede a serem testadas. Todas as redes eram bidimensionais, com disposição hexagonal dos neurônios, função de vizinhança gaussiana, aprendizagem em lote e em uma única fase, 1000 épocas de treinamento. Restando assim, as dimensões da rede e o raio inicial de vizinhança como parâmetros livres.

Para cada aplicação tem-se um processo distinto de escolha da melhor rede.

9. Identificando Dados Atípicos

A U-matriz, como visto na seção 3.1, permite que a estrutura geral do conjunto de dados amostrais seja avaliada de maneira visual. Inclusive permitindo que dados atípicos sejam facilmente identificados.

Avaliar cada configuração de rede não apresenta valor prático uma vez que a estrutura da U-matriz para vários Mapas são semelhantes. A Figura 4 mostra que para redes pequenas (3x3) a estrutura da U-matriz se apresenta complexa e não fornece subsídios para a análise dos dados, já para redes muito grandes (50x30) percebe-se uma superespecialização no Mapa representada pelos vários grupos de dados observados. Esta superespecialização foi ilustrada através da plotagem do histograma do nível de atividade dos neurônios (em branco).

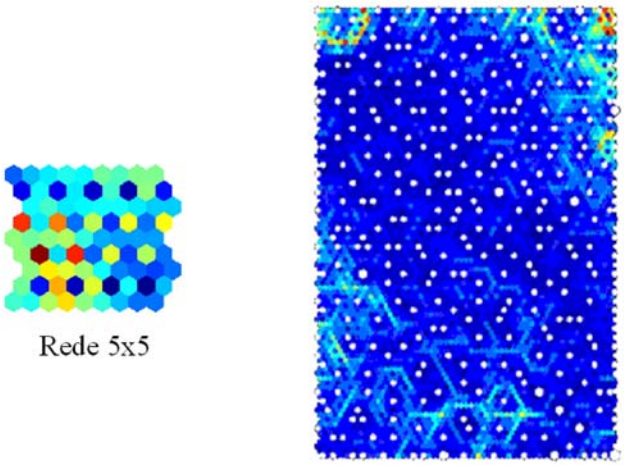

Figura 4. **U-matrizes geradas paras as redes 3x3 e 50x30.**

Analisando-se a curva dos erros de quantização e topológico observa-se que a curva do erro topológico é irregular, porém crescente para redes com m/n>1, a curva do erro de quantização decai suavemente até, aproximadamente, m/n=1, ver Figura 5. Logo, da análise visual da formação da U-matriz e dos gráficos do erro de qauntização e topológico optou-se pela configuração de rede com dimensão 20x15, pois equilibra o tamanho e formação da U-matriz com os valores do erro de quantização e topológico.

A U-matriz gerada pela rede 20x15 esta ilustrada na Figura 6. Através desta U-matriz pode ser observado dois agrupamentos de dados bem definidos nos cantos superiores da imagem. Na parte inferior central da imagem há uma região candidata a agrupamento, mas não muito bem definida. A região central forma, aparentemente uma região homogênea, ou seja, sem formação explícita de agrupamentos. Para o conjunto de setores censitários que se encontram relacionados com os neurônios do agrupamento do canto superior esquerdo denominou-se, Grupo1, e Grupo 2 para os setores relacionados com os neurônios do agrupamento do canto superior direito.

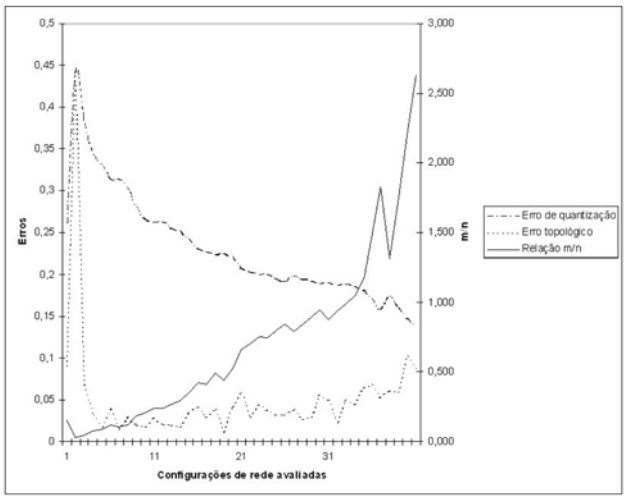

Figura 5. **Gráfico dos erros de quantização e topológico**

Usando o mapa dos setores censitários de São José dos Campos para mostrar quais são os setores dos grupos 1 e 2 identifica-se que correspondem a áreas de exclusão social. Estas mesmas áreas foram encontradas por [8] usando-se outros métodos de detecção de dados atípicos, o que evidencia e confirma a capacidade do SOM em descobrir facilmente padrões atípicos dentro do conjunto amostral. O mapa com os grupos 1 e 2 está ilustrado na Figura 7.

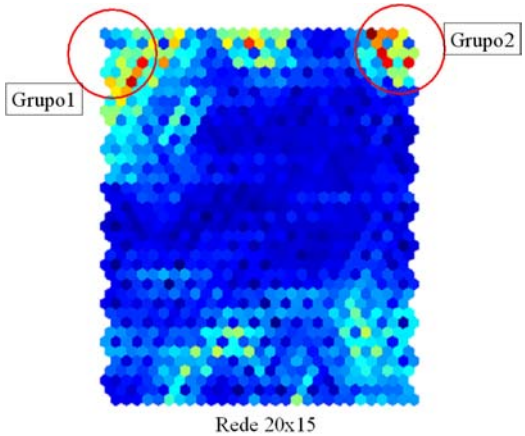

Figura 6. **U-matriz gerada para a rede 20x15**

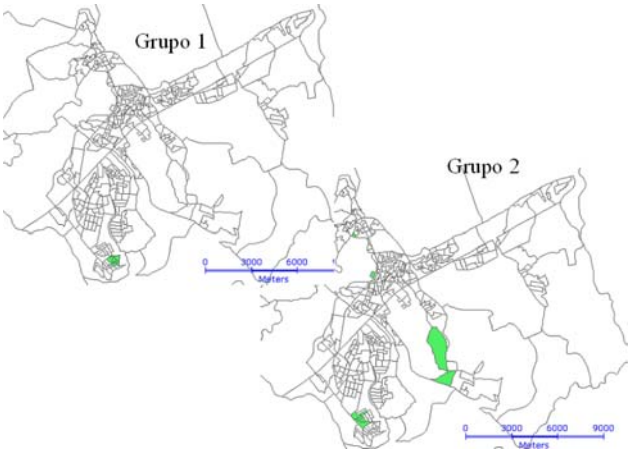

Figura 7. **Mapas dos setores censitários identificados como setores atípicos.**

10. Análise de Correlação e Significância de Componentes

Para a análise dos Planos de Componentes percebe-se que, para o conjunto de dados estudado, os planos de componentes gerados para a rede 5x5 são semelhantes àqueles gerados pela rede 20x15, Figura 8. Através da observação visual dos SOMs avaliados constata-se que o tamanho do Mapa não influencia significativamente a formação dos planos de componentes, embora mapas muito pequenos acabem escondendo determinados comportamentos dos componentes. Assim escolheu-se a rede 20x15 para a análise dos Planos de Componentes.

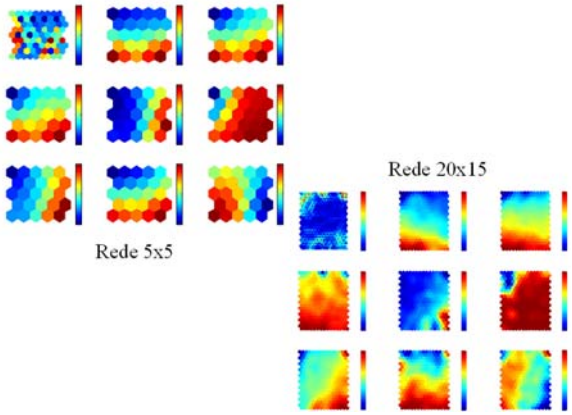

Figura 8. **Planos de Componentes. Tanto para redes pequenas (5x5) qaunto para redes maiores (20x15) os planos de componentes são semelhantes**

A Figura 9 mostra a estrutura dos Planos de Componentes para a rede 20x15. Como a cor vermelha indica valores altos e o azul escuro indica valores baixos pode-se fazer uma relação direta entre o padrão de cores dos Planos de Componentes com regiões de inclusão e exclusão social. Assim regiões em vermelho correspondem as áreas do Mapa especializadas em setores censitários com alta inclusão social (valores próximos de +1), analogamente tem-se regiões em azul especializadas em setores com alta exclusão social (valores próximos de –1).

A primeira conclusão da observação dos Planos de Componentes é que as variáveis ARENDR, DESEDUCR e MANALFR estão correlacionadas, pois apresentam o mesmo padrão de coloração. Com destaque para a alta correlação entre as variáveis ARENDR e DESEDUCR.

Observa-se que as variáveis LONGR e QAMBR contribuem muito pouco para formação da U-matriz uma vez que possuem grandes áreas bem homogêneas. Aqui destaca-se a variável LONGR que é variável que menos contribui no computo geral.

Estas mesmas conclusões foram alcançadas por [8], usando métodos distintos.

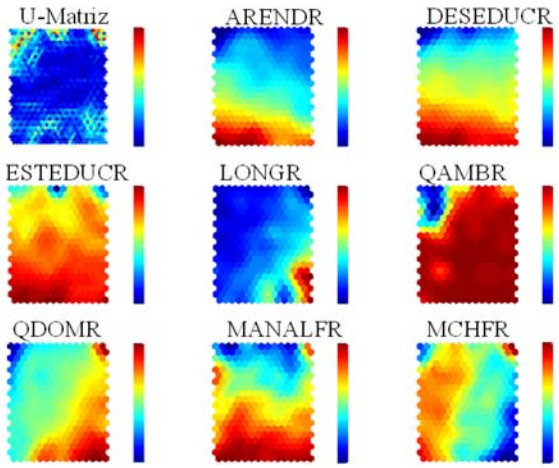

Figura 9. **Planos de Componentes para a rede 20x15**

11. Análise da Distribuição Espacial do Fenômeno

Da análise dos Planos de Componentes, rede 20x15, chega-se a conclusão que existe um sentido exclusão/inclusão na distribuição do Mapa e que este é vertical. Rotulando-se os neurônios no sentido vertical do mapa e mapeando esta rotulação para os setores censitários têm-se o mapa da Figura 10. Observa-se que as áreas de inclusão estão concentradas no centro do mapa enquanto que os setores com maior exclusão social concentram-se na periferia do mapa. Esta é uma das conclusões mais importantes do trabalho [8] e que foi confirmada através da análise dos Planos de Componentes, distribuição centro-periferia do fenômeno de exclusão/inclusão social urbana em São José dos Campos.

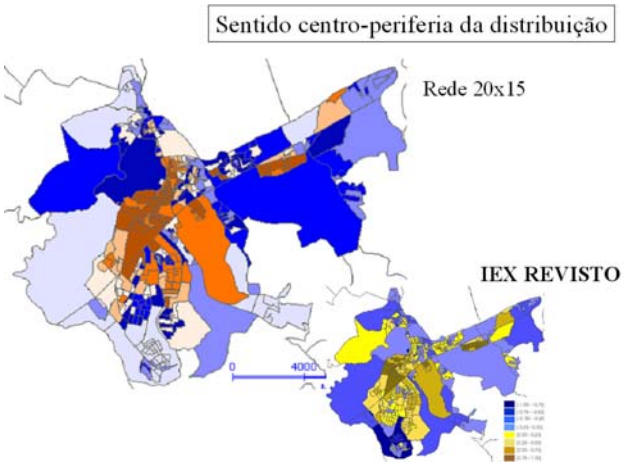

Figura 10. **Mapa gerado a partir da rotulação no sentido vertical da grade de neurônios baseada na distribuição dos Planos de Componentes.**

12. Conclusões

Observou-se que o SOM obteve os mesmos resultados obtidos por Genovez para a detecção de dados atípicos, análise de correlação de componentes, análise de significância dos componentes e distribuição espacial do problema de exclusão/inclusão social urbana.

A parametrização inicial da rede influencia a formação final do Mapa e nos resultados como um todo, porém o fator que mais influencia é o tamanho da rede. Para redes muito pequenas ou grandes o resultado não é satisfatório, todavia, para redes intermediárias os resultados tendem a ser bons e similares entre si.

Conclui-se que para a tarefa de análise exploratória de dados espaciais multivariados sobre o problema de exclusão/inclusão social urbana o SOM apresentou bons resultados para os quesitos estudados.

13. Referências

[1] Babu, G. P. Self-organizing neural networks for spatial data. **Pattern Recognition Letters**, v. 18, p. 133–142, 1997.

[2] Bailey, T. C.; Gatrell, A. C. **Interactive Spatial Data** Analysis. Longman, 1995.

[3] Cereghino, R.; Giraudel, J.; Compin, A. Spatial analysis of stream invertebrates distribution in the Adour-Garonne drainage basin (France), using Kohonen self organizing maps. **Ecological Modelling**, v. 146, n. 1-3, p. 167–180, 2001.

[4] Cottrell, M.; Gaubert, P.; Letremy, P.; Rousset, P. **Kohonen Maps**. Elsevier, 1999. Cap. Analyzing and representing multidimensional quantitative and qualitative data: Demographic study of the Rhˆone valley. The domestic consumption of the Canadian families, p. 1–14.

[5] Foody, G. Applications of the self-organising feature map neural network in community data analysis. **Ecological Modelling**, v. 120, p. 97–107, 1999.

[6] Franzini, L.; Bolchi, P.; Diappi, L. Self Organizing Maps,: A Clustering neural method for urban analysis. **Proceeding of the V Recontres de Th´eo Quant**. 2001. p. 1–15.

[7] Gahegan, M.; Takatsuka, M.; Wheeler, M.; Hardisty, H. Introducing GeoVISTA Studio: an integrated suite of visualization and computational methods for exploration and knowledge construction in geography. **Computers, Environment and Urban Systems**, v. 26, p. 267–292, 2002.

[8] Genovez, P. C. **Território e Desigualdades: Análise Espacial Intra-urbana no estudo da dinâmica de exclusão/inclusão social no espaçoo urbano em São José dos Campos-SP**. Dissertação – INPE, Dezembro 2002. Disponível em www.dpi.inpe.br/genovez/, acessado em 01/06/2003.

[9] Ji, C. Y. Land-use classification of remotely sensed data using self-organizing feature mapa neural networks. **Photogrammetric Engineering & Remote Sensing**, v. 66, n. 12, p. 1451–1460, 2000.

[10] Kaski, S.; Kohonen, T. Exploratory Data Analysis by The Self-Organizing Map:Structures of Welfare and Poverty in the World. **Proceeding of the Third International Conference on Neural Networks in the Capital Markets**. World Scientific, 1996. p. 498–507.

[11] Kohonen, T. ***Self-organizing maps***. Springer, 1995. Third Edition 2001.

[12] Kropp, J. A neural network approach to the analysis of city systems. **Applied Geography**, v. 18, n. 1, p. 83–96, 1998.

[13] Openshaw, S.; Blake, M.; Wymer, C. **Using neurocomputing methods to classify britain's residential areas**, 1995. www.geog.leeds.ac.uk/papers/95-1/.

[14] Openshaw, S.; Turton, I. **A parallel Kohonen algorithm for the classification of large spatial datasets**. Computers & Geosciences, v. 22, n. 9, p. 1019–1026, 1996.

[15] SMILEY. **SIMILEY: Signature miner & interface language for Earth-data, Yet another**, 2003. Disponível em midas.cs.ndsu.nodak.edu/˜smiley/, acessado em 04/06/2003.

[16] Takatsuka, M. An application of the self-organizing map and interactive 3-D visualization to geospatial data. **Proceedings of the 6th International Conference on GeoComputation University of Queensland**, 24 – 26 September 2001. Brisbane, Australia, 2001.

[17] Ultsch, A. **Information and Classification**. Springer, 1993. Cap. Knowledge extraction from self-organizing neural networks.

[18] Villmann, T.; Mer´enyi, E.; Hammer, B. **Neural maps in remote sensing image analysis**. Neural Networks, v. 16, p. 389–403, 2003.

[19] Winter, K.; Hewitson, B. **Neural nets: applications in geography**. Kluwer, 1994. Cap. Self organizing maps - applications to census data.