



MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

INPE-12434-TDI/996

**MAPAS AUTO-ORGANIZÁVEIS NA ANÁLISE EXPLORATÓRIA
DE DADOS GEOESPACIAIS MULTIVARIADOS**

Marcos Aurélio Santos da Silva

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada,
orientada pelos Drs. Antônio Miguel Vieira Monteiro e José Simeão de Medeiros,
aprovada em 08 de março de 2004.

681.3.019

SILVA, M. A. S.

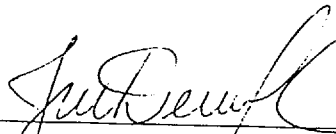
Mapas auto-organizáveis na análise exploratória de dados geoespaciais multivariados / M. A. S. Silva. – São José dos Campos: INPE, 2004.

117p. – (INPE-12434-TDI/996).

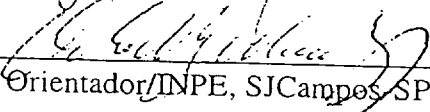
1.Redes Neurais. 2.Inteligência artificial. 3.Distribuição espacial. 4.Sistemas de Informação Geográfica (SIG). 5.Análise de agrupamentos. I.Título.

Aprovado(a) pela Banca Examinadora
em cumprimento a requisito exigido
para a obtenção do Título de Mestre
em Computação Aplicada.

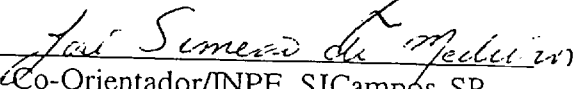
Dr. José Demísio Simões da Silva


Presidente/INPE, SJCampos-SP

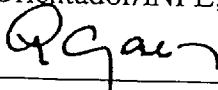
Dr. Antônio Miguel Vieira Monteiro


Orientador/INPE, SJCampos-SP

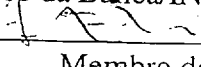
Dr. José Simeão de Medeiros


Co-Orientador/INPE, SJCampos-SP

Dr. Gilberto Câmara Neto


Membro da Banca/INPE, SJCampos-SP

Dr. Aluízio Fausto Ribeiro Araújo


Membro da Banca
Convidado - Univ. Federal de Pernambuco

Candidato: Marcos Aurélio Santos da Silva

São José dos Campos, 08 de março de 2004.

Ao meu amor,
Lilian Dias Dantas.

AGRADECIMENTOS

Agradeço às sincronicidades Divinas que permitiram a realização deste Mestrado.

Aos meus pais Pedro e Maria de Lourdes, meus irmãos Lílian, Márcio e Patrícia, meu sobrinho Pedro Paulo, aos meus sogros João Batista e Nora, ao meu tio José Teles e meus cunhados Michelangelo, Liliane e Lorena. Agradeço, principalmente, a minha noiva Lílian pela sua compreensão, paciência e tolerância.

Aos meus orientadores Antônio Miguel Vieira Monteiro e José Simeão de Medeiros, pela oportunidade de trabalhar com ambos, pessoas de imensa capacidade criativa e disposição, além de motivadores do trabalho baseado no consenso.

Aos Drs. Lafayette Franco Sobral, Ederlon Ribeiro de Oliveira, Antônio Carlos Barreto e Amaury Apolônio de Oliveira que, em esferas distintas, ajudaram decisivamente para a viabilidade deste projeto e depositaram total confiança em nosso trabalho.

Aos professores Henrique Nou Schneider e Leila Maciel de Almeida e Silva, Departamento de Ciência da Computação da Universidade Federal de Sergipe - UFS, pelo apoio irrestrito, motivação e exemplo pessoal de perseverança e competência.

À Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), pelo auxílio financeiro correspondente a dois anos de bolsa de mestrado e outras despesas.

Ao Instituto Nacional de Pesquisas Espaciais (INPE), pela disponibilização da ótima estrutura física e humana.

Aos professores do INPE pelo conhecimento compartilhado, em especial aos professores Gilberto Câmara e José Demisio Simões da Silva.

Ao Centro de Pesquisas Agropecuárias dos Tabuleiros Costeiros (EMBRAPA Tabuleiros Costeiros), através do Chefe Geral, Dr. Lafayette Franco Sobral, por acreditar em nosso potencial de trabalho, pela liberação total para a realização do curso e financiamento de despesas.

A Patrícia Genovez e Marcelo Alves, pelos seus trabalhos, que foram fundamentais para a formulação da idéia e elaboração desta dissertação.

Aos amigos Alex Pessoa, Jacques Politi, Rodrigo Rizzi, Élcio Shiguemori, Ana Paula Castro, Marcelino Silva, Dimitry Fedorov, Eliana Fonseca, Arley Souza, Tantravahi Adytia, Ana Paula Figueiredo, Marinaldo Gleriani, Emiliano Castejon, Isabela Drummond, Alexandre Oliveira, Fabrício Harter, Leonardo Chiwiacowsky, Lúcio Franco, que me receberam, ajudaram e tornaram meu trabalho o menos penoso possível.

Concluo, agradecendo a todos aqueles que não foram citados mas que ajudaram e contribuíram para a conclusão deste trabalho.

RESUMO

Os Mapas Auto-Organizáveis têm sido aplicados, com sucesso, em variados problemas de análise exploratória de dados multivariados, todavia, poucos são os trabalhos voltados para a análise de dados coletados para unidades espaciais definidas, como os dados produzidos por censos demográficos, aqui chamados genericamente por dados geoespaciais. Este trabalho apresenta uma metodologia para a análise destes dados que contempla desde a verificação de dados atípicos até a análise de dependência espacial, fazendo uso, como ferramentas, somente do Mapa Auto-Organizável e seus algoritmos relacionados. Pode-se dividir a metodologia em quatro partes: detecção de dados atípicos através da análise da matriz de distância unificada (U-matriz), análise de componentes através dos Planos de Componentes, detecção automática de agrupamentos através do algoritmo Costa-Netto e análise da distribuição e dependência espaciais a partir dos Planos de Componentes e do particionamento dos dados na etapa anterior, respectivamente. Para a aplicação da metodologia proposta foi projetado e implementado um ambiente computacional integrado para análise exploratória de dados, que faz extensivo uso de banco de dados geográfico através da biblioteca aberta **TerraLib** (terralib.dpi.inpe.br). Compõem este ambiente a biblioteca SOMLib e o sistema *CASA - Connectionist Approach for Spatial Analysis of Areal Data*. A biblioteca SOMLib compreende um conjunto de classes em C++ projetadas a partir do uso de padrões de projeto e outras técnicas modernas de programação, cujo objetivo é implementar os algoritmos relacionados com os Mapas Auto-Organizáveis, de maneira a considerar a escalabilidade, a capacidade de manutenção do projeto e a conectividade com a biblioteca **TerraLib**. O sistema *CASA* é um simulador neural SOM 2-D, com interface gráfica, desenvolvido sobre as bibliotecas SOMLib e QT para execução das tarefas de análise exploratória de dados geoespaciais. Para validação da metodologia e do ambiente computacional de análise geográfica com redes SOM a mesma foi aplicada no problema de análise de indicadores de exclusão/inclusão social no município de São José dos Campos, São Paulo. Os resultados desta aplicação foram comparados com resultados anteriores, para o mesmo conjunto de dados, obtidos através de técnicas estatísticas e mostraram que os Mapas Auto-Organizáveis e os algoritmos relacionados são ferramentas robustas para a tarefa de análise exploratória de dados geoespaciais.

SELF-ORGANIZING MAPS IN THE EXPLORATORY ANALYSIS OF MULTIVARIATE GEOSPATIAL DATA

ABSTRACT

Self-Organizing Maps (SOM) have been applied, successfully, in a variety of problems of exploratory analysis of multivariate data, however, few are the works related to the analysis of geospatial data. This work considers a methodology of analysis of geospatial data that contemplates from the verification of outliers to the analysis of space dependence using a Self-Organizing Map and its related algorithms as tools. The methodology is divided into four parts: detection of outliers through the analysis of the matrix of unified distance (U-matrix), analysis of components through the Component Planes, automatic clustering through the Costa-Netto's algorithm and analysis of the space distribution and dependence from the Component Planes and analysis of the partitioning of the data in the previous stage, respectively. The application of the methodology was carried out with direct access to geographic data through the integration with the **TerraLib** library (terralib.dpi.inpe.br) by a computational environment. The SOMLib library and the system *CASA - Connectionist Approach for Spatial Analysis of Areal Data* are components of that environment. The SOMLib library is a set of C++ classes designed by using Design Patterns and other modern programming techniques, whose objective is to implement the algorithms related with the Self-Organizing Maps in way that considers the scalability, maintainability of the project and the connectivity with the **TerraLib**. The *CASA* system is a neural 2D SOM simulator, with graphical interface, developed on the SOMLib and QT libraries. The methodology was applied to the problem of social exclusion/inclusion in the City of São José dos Campos, São Paulo. The results of this application were compared with previous statistical results and showed that a Self-Organizing Map and related algorithms are robust tools for the task of exploratory analysis of geospatial data.

SUMÁRIO

| | <u>Pág.</u> |
|--|-------------|
| LISTA DE FIGURAS | 15 |
| LISTA DE TABELAS | 19 |
| LISTA DE SÍMBOLOS | 21 |
| LISTA DE SIGLAS E ABREVIATURAS | 23 |
| CAPÍTULO 1 – INTRODUÇÃO | 25 |
| 1.1 – Introdução e Motivação | 25 |
| 1.2 – Objetivos | 27 |
| 1.3 – Contribuições | 28 |
| 1.4 – Organização da Dissertação | 28 |
| 1.5 – Sumário | 29 |
| CAPÍTULO 2 – MAPAS AUTO-ORGANIZÁVEIS | 31 |
| 2.1 – Redes Neurais Artificiais | 31 |
| 2.2 – Mapas Auto-Organizáveis | 32 |
| 2.2.1 – Aprendizagem padrão ou seqüencial | 34 |
| 2.2.2 – Aprendizagem em lote | 35 |
| 2.2.3 – Considerações sobre o treinamento do SOM | 36 |
| 2.2.4 – Avaliação da qualidade da aprendizagem | 37 |
| 2.2.5 – Propriedades | 37 |
| 2.3 – Visualização do Mapa Auto-Organizável | 38 |
| 2.3.1 – Representação dos vetores de código no espaço \mathbb{R}^d | 38 |
| 2.3.2 – Histograma | 39 |
| 2.3.3 – Planos de Componentes | 39 |
| 2.3.4 – Matriz de distância unificada (U-Matriz) | 40 |
| 2.4 – Análise Exploratória de Dados com Mapas Auto-Organizáveis | 41 |
| 2.4.1 – Seleção dos dados | 42 |
| 2.4.2 – Pré-processamento | 42 |
| 2.4.3 – Configuração da rede e seleção do Mapa | 42 |
| 2.4.4 – Interpretação do Mapa neural | 43 |
| 2.5 – Sumário | 45 |

| | |
|---|-----------|
| CAPÍTULO 3 – DETECÇÃO AUTOMÁTICA DE AGRUPAMENTOS COM MAPAS AUTO-ORGANIZÁVEIS | 47 |
| 3.1 – Introdução | 47 |
| 3.2 – Métodos Automáticos de Segmentação | 47 |
| 3.3 – Segmentação Baseada em Particionamento de Grafos | 49 |
| 3.4 – Índices de Validação dos Agrupamentos | 50 |
| 3.4.1 – Índice Davies-Bouldin | 50 |
| 3.4.2 – Índice CDbw | 52 |
| 3.4.3 – Usando os vetores de código como vetores de referência no CDbw | 53 |
| 3.5 – Validando o Particionamento do SOM | 54 |
| 3.6 – Sumário | 54 |
| | |
| CAPÍTULO 4 – ANÁLISE EXPLORATÓRIA DE DADOS GEOES- PACIAIS MULTIVARIADOS ATRAVÉS DE MAPAS AUTO-ORGANIZÁVEIS | 57 |
| 4.1 – Análise Espacial de Dados em Área | 57 |
| 4.2 – Análise Espacial com o SOM | 58 |
| 4.3 – Inclusão da Variável Espacial no Algoritmo SOM | 60 |
| 4.4 – Análise da Distribuição Espacial do Fenômeno | 60 |
| 4.5 – A Proposta de um Índice de Valoração da Dependência Espacial | 61 |
| 4.6 – Sumário | 64 |
| | |
| CAPÍTULO 5 – PROJETO E PROGRAMAÇÃO DO MAPA AUTO- ORGANIZÁVEL | 65 |
| 5.1 – Introdução | 65 |
| 5.2 – Projeto e Programação | 66 |
| 5.3 – Características | 72 |
| 5.4 – Avaliação da Biblioteca | 72 |
| 5.5 – Uso da Biblioteca SOMLib | 72 |
| 5.6 – Descrição do Sistema <i>CASA</i> | 74 |
| 5.7 – Sumário | 79 |
| | |
| CAPÍTULO 6 – ESTUDO DE CASO: ANÁLISE ESPACIAL INTRA- URBANA EM SÃO JOSÉ DOS CAMPOS-SP | 81 |
| 6.1 – Estudo de Caso | 81 |
| 6.2 – Seleção dos Dados e Pré-processamento | 82 |
| 6.3 – Configuração da Rede SOM | 82 |
| 6.4 – Identificando Dados Atípicos e Organização Geral da Estrutura dos Dados | 85 |

| | |
|---|------------|
| 6.5 – Análise de Componentes | 90 |
| 6.6 – Análise da Distribuição Espacial do Fenômeno | 93 |
| 6.7 – Avaliando a Inclusão das Coordenadas Planas (x, y) em x_k | 95 |
| 6.8 – Descoberta de Agrupamentos e Análise da Dependência Espacial | 97 |
| 6.9 – Sumário | 103 |
| CAPÍTULO 7 – CONCLUSÕES | 105 |
| 7.1 – Considerações Finais | 105 |
| 7.2 – Trabalhos Futuros | 107 |
| REFERÊNCIAS BIBLIOGRÁFICAS | 109 |
| APÊNDICE A – INICIAÇÃO LINEAR DOS VETORES DE CÓDIGO DO SOM | 117 |

LISTA DE FIGURAS

| | <u>Pág.</u> |
|--|-------------|
| 2.1 Modelo básico de um neurônio j com entrada x_k , pesos sinápticos w_j , nível de ativação J e saída $f(J)$ | 31 |
| 2.2 Exemplo de um Mapa Auto-Organizável bidimensional $N \times M$, com entrada x_k | 33 |
| 2.3 Diferentes topologias para estruturação dos mapas auto-organizáveis: a) unidimensional; b) bidimensional hexagonal; c) bidimensional retangular. | 34 |
| 2.4 Do lado esquerdo tem-se os dados, sendo dois conjuntos de dados formando um elo de corrente; do lado direito tem-se a estrutura final do SOM, após treinamento, usando os valores dos vetores de código, como coordenadas no espaço \mathbb{R}^3 | 39 |
| 2.5 Representação dos componentes da U-matriz dx, dy, dz e du para uma rede 3x3 hexagonal. Os círculos representam os neurônios e os quadrados representam os valores dos componentes. | 40 |
| 2.6 Exemplo de geração da imagem relativa a U-matriz, a partir de uma rede 3x3 hexagonal. | 41 |
| 2.7 Fluxograma do processo de análise exploratória de dados com SOM. | 42 |
| 3.1 Exemplo da aplicação do método de segmentação da U-matriz (<i>SL-SOM</i>): (a) SOM bidimensional 10x10; (b) U-matrix gerada a partir desta SOM treinada; (c) Rotulação dos neurônios da SOM com o auxílio da imagem (b). FONTE:(Costa, 1999). | 48 |
| 3.2 A estratégia de segmentação do SOM baseia-se na eliminação de conexões inconsistentes entre os neurônios. Neste exemplo, uma rede 3x3 foi particionada de forma que 3 grupos foram constituídos. | 49 |
| 4.1 Elementos da Análise Espacial de Dados em Área. | 57 |
| 4.2 Coordenadas do centróide de uma área qualquer relativo ao MBR do conjunto de áreas. | 61 |

| | | |
|------|--|----|
| 4.3 | Diferentes possíveis sentidos da distribuição dos componentes no Mapa neural. | 62 |
| 4.4 | Correspondência entre a relação de vizinhança no espaço de atributos e no espaço físico. | 62 |
| 4.5 | Conjunto de áreas rotuladas, 2 agrupamentos, para exemplificar o cálculo do índice IRVE. | 63 |
| 5.1 | Diagrama de Classes para representação das famílias de Mapas Auto-Organizáveis. | 66 |
| 5.2 | Diagrama de Classe: a) Representação da classe base e das classes de aprendizagem; b) Nova estrutura do diagrama -a- baseada no padrão <i>Strategy</i> . | 67 |
| 5.3 | Diagrama de Classes. Aqui observa-se o alto acoplamento entre as classes de topologia e de aprendizagem. | 68 |
| 5.4 | Através do padrão <i>Bridge</i> separou-se os detalhes de topologia e aprendizagem. | 69 |
| 5.5 | Diagrama de Classe final | 70 |
| 5.6 | Representação do uso do padrão <i>Abstract Factory</i> sobre o diagrama de classes da Figura 5.5. | 71 |
| 5.7 | Representação da estrutura de classes relativas aos dados e algoritmo de leitura e gravação dos dados de entrada da rede neural. | 72 |
| 5.8 | Tela inicial do sistema <i>CASA</i> . | 74 |
| 5.9 | Formulário de acesso ao banco de dados geográfico. | 76 |
| 5.10 | Formulário com informações sobre o processo de aprendizagem da rede, número de agrupamentos encontrados pelo algoritmo Costa-Netto e índices de validação deste particionamento. | 76 |
| 5.11 | Resultado do processo de segmentação do Mapa neural através do algoritmo Costa-Netto. O formulário <i>Data Label</i> informa, para cada neurônio, quais padrões de entrada estão relacionados com o mesmo, sua posição (<i>Neuron number</i>) e a qual agrupamento pertence (<i>cluster ID</i>). | 77 |
| 5.12 | Planos de Componentes gerados pelo sistema. | 78 |

| | | |
|------|--|----|
| 5.13 | U-matriz pelo sistema. | 78 |
| 6.1 | Gráfico do erro de quantização. | 85 |
| 6.2 | Gráfico do erro topológico. | 86 |
| 6.3 | Número de agrupamentos encontrados pelo algoritmo de segmentação Costa-Netto. | 86 |
| 6.4 | Índice de validação CDbw. | 87 |
| 6.5 | Índice de validação Davies-Bouldin ($p=2, q=1$). | 87 |
| 6.6 | Índice de validação Davies-Bouldin ($p=2, q=2$). | 88 |
| 6.7 | U-matrizes geradas para as redes 5×5 e 50×30 | 89 |
| 6.8 | Gráfico dos erros de quantização e topológico. | 90 |
| 6.9 | U-matriz gerada para a rede 20×15 | 91 |
| 6.10 | Mapas dos setores censitários identificados como setores atípicos. | 91 |
| 6.11 | Planos de Componentes. Tanto para redes pequenas (5×5), quanto para redes maiores (20×15), os planos de componentes são semelhantes. | 92 |
| 6.12 | Planos de Componentes para a rede 20×15 | 94 |
| 6.13 | Mapa gerado a partir da rotulação, no sentido vertical, da grade de neurônios, baseada na distribuição dos Planos de Componentes “a”. Mapa baseado no Iex revisto “b”. FONTE: (Genovez, 2002). | 94 |
| 6.14 | Efeito, na U-matriz, da inclusão das coordenadas planas. | 95 |
| 6.15 | Efeito, nos Planos de Componentes, da inclusão das coordenadas planas. | 96 |
| 6.16 | Fases do processo de particionamento dos dados em c agrupamentos. | 97 |
| 6.17 | Gráficos para o índice Davies-Bouldin. | 98 |
| 6.18 | Mapa neural particionado segundo o índice Davies-Bouldin. | 98 |
| 6.19 | Gráficos para o índice CDbw. | 99 |

| | | |
|------|---|-----|
| 6.20 | Mapa particionado segundo o índice Cdbw. | 100 |
| 6.21 | Mapa dos setores censitários gerados a partir do SOM particionado segundo o algoritmo Costa-Netto e validação do índice CDbw. Em destaque o setor sul da área urbana onde pode-se verificar que o algoritmo identificou dentro de uma área de exclusão sub-agrupamentos que podem ser caracterizados como fragmentos urbanos. | 101 |
| 6.22 | Relação entre os índices IRVE e CDbw. | 102 |

LISTA DE TABELAS

| | <u>Pág.</u> |
|---|-------------|
| 6.1 Configurações de rede avaliadas. | 83 |
| 6.2 Experimentos conduzidos para uma rede neural SOM bidimensional, hexagonal, com função de vizinhança gaussiana e aprendizagem em lote. | 84 |
| 6.3 Resultados para o índice IRVE do experimento 001, configuração de rede 26. | 100 |

LISTA DE SÍMBOLOS

| | |
|--------------------|--|
| n | – número de padrões amostrais |
| m | – quantidade de neurônios na rede neural |
| x_k | – vetor de características $k = 1, \dots, n$ |
| d | – dimensão do vetor x_k |
| d' | – dimensão da grade de neurônios |
| Ξ | – conjunto dos vetores x_k |
| ξ_{kj} | – j -ésimo componente do vetor x_k , $j = 1, \dots, d$ |
| N | – dimensão vertical da rede neural SOM |
| M | – dimensão horizontal da rede neural SOM |
| I | – espaço de entrada da rede neural |
| U | – espaço de saída da rede neural |
| w_j | – vetor de código ou pesos do neurônio j |
| d_{ij} | – distância entre os neurônios i e j |
| h_{ij} | – função de vizinhança aplicada aos neurônios i e j |
| $\delta(t)$ | – raio de abrangência da vizinhança no tempo discreto t |
| V_i | – região de Voronoi para o neurônio i |
| n_{V_i} | – número de padrões na região de Voronoi V_i |
| s_i | – somatório das amostras relativas a região de Voronoi i |
| p | – parâmetro para cálculo da dispersão intra-agrupamento do índice Davies-Bouldin |
| q | – parâmetro para cálculo da dispersão inter-agrupamento do índice Davies-Bouldin |
| dx | – distância entre o vetor de código de um neurônio e o seu vizinho à direita |
| dy | – distância entre o vetor de código de um neurônio e o seu vizinho abaixo |
| dz | – distância entre o vetor de código de um neurônio e o seu vizinho na diagonal |
| du | – distância calculada a partir dos valores dx , dy e dz |
| $H(i)$ | – nível de atividade do neurônio i |
| $d(w_i, w_j)$ | – distância entre os vetores de código dos neurônios i e j |
| H_{min} | – limiar para o nível de atividade do neurônio |
| c | – número de agrupamentos encontrados após partição dos dados |
| Q_k | – conjunto dos padrões relativos ao agrupamento k |
| $S_c(Q_k)$ | – dispersão interna do agrupamento Q_k |
| N_k | – número de amostras no agrupamento Q_k |
| $d_{ce}(Q_k, Q_l)$ | – distância entre os agrupamentos Q_k e Q_l |
| V_i' | – conjunto dos vetores representativos do agrupamento i |
| A_k | – área de estudo k , $k = 1, \dots, n$ |
| v_{ij} | – vetor representativo j do agrupamento i |
| \bar{x}_i | – média das amostras do i -ésimo agrupamento |
| R | – região de estudo $R = A_1 \cup \dots \cup A_n$ |
| W | – matriz de proximidade |
| w'_{ij} | – elementos da matriz de proximidade |
| (x, y) | – coordenadas planas relativas ao centróide das áreas de estudo A |
| p_i | – número de áreas (A) pertencentes ao agrupamento i |
| q_i | – número de grupos de áreas (A) distintas do agrupamento i |

E_q – medida do erro de quantização
 E_t – medida do erro topológico

LISTA DE SIGLAS E ABREVIATURAS

| | |
|----------|--|
| ART | – Teoria da Ressonância Adaptativa (<i>Adaptative Resonance Theory</i>) |
| BMU | – Neurônio vencedor (<i>Best Match Unit</i>) |
| CASA | – Abordagem Conexionista para Análise Espacial de Área (<i>Connectionist Approach for Spatial Analysis of Areal Data</i>) |
| CDbw | – Densidade composta inter e intra agrupamentos (<i>Compose Density between and within clusters</i>) |
| EECNI | – Eliminação do Efeito de Cadeia dos Neurônios Inativos |
| GeoVista | – Sistema visual escrito em Java para análise espacial |
| GPS | – Sistema de Posicionamento Global (<i>Global Positioning System</i>) |
| IBGE | – Instituto Brasileiro de Geografia e Estatística |
| INPE | – Instituto Nacional de Pesquisas Espaciais |
| IRVE | – Índice de Relação de Vizinhança Espacial |
| MBR | – Mínimo Retângulo Envolvente (<i>Minimum Bound Rectangle</i>) |
| MEDALUS | – Uso da Terra e Desertificação do Mediterrâneo (<i>Mediterranean Desertification and Land Use</i>) |
| MLP | – Perceptron de Múltiplas Camadas (<i>Multi-Layer Perceptron</i>) |
| MUB | – Mapas Urbanos Básicos |
| OSAMS | – Sistema Otago de Análise Espacial e Modelagem (<i>Otago Spatial Analysis and Modelling System</i>) |
| SL-SOM | – SOM auto-rotulável (<i>Self-Labeling SOM</i>) |
| SIG | – Sistema de Informação Geográfica |
| SGBD | – Sistema Gerenciador de Banco de Dados |
| SOM | – Mapa Auto-Organizável (<i>Self-Organizing Map</i>) |
| SOMPAK | – Pacote SOM (<i>SOM Package</i>) |
| SOMLib | – Biblioteca de classes SOM (<i>SOM library</i>) |
| SOMSD | – SOM para Dados Espaciais (<i>SOM for Spatial Data</i>) |

CAPÍTULO 1

INTRODUÇÃO

1.1 Introdução e Motivação

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; oferta 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, realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), com a malha de setores censitários disponível por município.

As tecnologias da informação que lidam com essas bases de dados, em particular a tecnologia dos SGBDs (Sistemas Gerenciadores de Bancos de Dados) e a 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 distintas, é bem menor que a nossa capacidade de produzi-lo.

Várias décadas depois de seu surgimento, os SIGs ampliaram suas capacidades funcionais para a análise desta massa de dados geograficamente referenciados, aqui chamados dados geoespaciais, mas ainda estão no início da efetiva integração dos métodos de exploração e análise destes dados. Estes métodos têm surgido dentro de áreas já tradicionais, como os métodos estatísticos, assim como em áreas emergentes, como aquelas ligadas à Inteligência Artificial e Sistemas Complexos, entre outras (Hewitson e Crane, 1994; Fischer e Getis, 1996; Openshaw e Openshaw, 1997; Couclelis, 1998; Openshaw e Abrahart, 2000; Câmara e Monteiro, 2001; Koga, 2003).

O termo “geocomputação” descreve o uso extensivo de técnicas computacionais no auxílio à solução de problemas em geografia (Gahegan, 1999; Openshaw e Abrahart, 2000). A ênfase original para o termo situava “geocomputação” como técnicas ou algoritmos computacionalmente intensivos, usados para buscar e revelar padrões escondidos em grandes conjuntos de dados derivados de problemas em geografia física ou humana, e, em particular, métodos não convencionais para análise de agrupamentos. O termo foi, mais tarde, aplicado em um contexto mais amplo, para incluir aquelas técnicas

matemático-computacionais que tratassem de análise estatística espacial, visualização de dados geoespaciais, modelos dinâmicos de interação espacial e modelos de dinâmicas espaço-temporais, entre outros (Longley et al., 1998). De maneira simplificada, porém definitiva, para o escopo deste trabalho, pode-se usar a definição de geocomputação definida por Openshaw e Abrahart (2000, p. 19)¹ como sendo “*o processo de aplicação da tecnologia computacional para a solução de problemas de análise geográfica*”.

É nesse contexto que se posiciona esta dissertação em Computação Aplicada. O desafio aqui foi procurar oferecer, em um ambiente de processamento de informação geográfica integrado, a possibilidade de uso de uma técnica computacional, a de Redes Neurais Artificiais (RNA), mais precisamente dos Mapas Auto-Organizáveis de Kohonen - SOM²(Kohonen, 2001), como uma ferramenta para análise exploratória de dados geoespaciais multivariados. O propósito é avaliar este tipo de rede como técnica possível para detecção de padrões, visualização e análise de componentes em bases de dados geográficas.

O uso de Redes Neurais Artificiais na Análise Espacial intensificou-se a partir dos anos 90 (Hewitson e Crane, 1994; Openshaw e Openshaw, 1997). Desde então, muitos trabalhos surgiram na literatura, tratando de temas diversos, como: classificação de imagens de sensores remotos (Ji, 2000; Tso e Mather, 2001; Villmann et al., 2003); predição dos níveis de pluviosidade (Bollivier et al., 1997; Lee et al., 1998); determinação dos índices de vulnerabilidade à erosão (Cai, 1995; Medeiros, 1999); produção de mapas potenciais para diversos bens minerais (Nobrega, 2001); e para avaliação de erosão em áreas agrícolas (Rosa et al., 1999). Como iniciativas no desenvolvimento de sistemas computacionais nesta direção destacam-se os projetos GeoVista (Gahegan et al., 2002) e OSAMS (Purvis et al., 2001).

A análise exploratória de dados geoespaciais multivariados é de extrema relevância para os especialistas interessados em modelar fenômenos geográficos (Openshaw e Turton, 1996). Métodos estatísticos vêm sendo aplicados, com sucesso, nesta tarefa, todavia, estes modelos, que exigem hipóteses *a priori*, atuam como filtros lineares, alterando as características dos dados, escondendo padrões e criando outros acidentalmente (Openshaw e Turton, 1996). As RNAs não são, necessariamente, mecanismos automáticos de análise exploratória de dados, mas funcionam sem que nenhuma hipótese *a priori* seja feita, além de serem métodos não lineares.

¹Geocomputation is the process of computational technology application to solve geographical problems.

²Do inglês Self-Organizing Map.

Uma RNA que tem sido usada com bastante êxito na análise de dados espaciais é o Mapa Auto-Organizável - SOM (Oja et al., 2003; Kohonen, 2001; Kaski et al., 1998). O SOM é um tipo de RNA competitiva, cuja função principal é mapear os dados de entrada de dimensão d numa grade de dimensão d' , $d' \ll d$, de forma que esta grade de neurônios, totalmente conectada ao vetor de entrada pelos vetores de código, ou pesos, represente, de maneira topologicamente ordenada, os dados de entrada (Kohonen, 2001). Ou seja, o SOM identifica, nos padrões de entrada, os agrupamentos e os correlaciona a regiões específicas da grade de neurônios. É justamente sobre esta propriedade de ordenação topológica dos agrupamentos gerados pelo SOM que surgem aplicações de análise exploratória de dados geoespaciais (Cereghino et al., 2001; Openshaw e Turton, 1996; Winter e Hewitson, 1994).

1.2 Objetivos

O objetivo deste trabalho foi propor e avaliar uma metodologia de análise exploratória de dados geoespaciais a partir do Mapa Auto-Organizável de Kohonen e produzir as ferramentas computacionais necessárias para sua implementação. Para testar e validar a metodologia e os métodos computacionais projetados e implementados foi avaliado um estudo de caso sobre o mapeamento da exclusão/inclusão social urbana em São José dos Campos-SP. Este problema já foi abordado através de técnicas estatísticas e de análise espacial (Genovez, 2002), estes resultados foram usados como parâmetros comparativos.

São objetivos específicos:

- a) Programar uma estrutura de software que integre os algoritmos relativos ao Mapa Auto-Organizável e a biblioteca **TerraLib**³ (Câmara et al., 2001), criando um meio para acesso direto a bancos de dados geográficos neste formato, obedecendo a lógica de desacoplamento entre estruturas de dados e algoritmos, fortemente utilizada no ambiente **TerraLib**;
- b) Especificar quais métodos de análise exploratória do SOM podem ser aplicados aos dados geoespaciais e como esta aplicação deve ser conduzida;
- c) Verificar a sensibilidade dos métodos às variações nos parâmetros livres da rede neural, principalmente quanto às dimensões da rede;
- d) Pesquisar métodos automáticos de análise de agrupamentos em dados geoespaciais a partir do SOM;

³TerraLib é uma API, de código aberto e gratuita, para acesso e manipulação de dados geográficos armazenados em bancos de dados, desenvolvida na Divisão de Processamento de Imagens do Instituto Nacional de Pesquisas Espaciais (DPI/INPE). A TerraLib está disponível em: terralib.dpi.inpe.br.

- e) Identificar a distribuição espacial e regimes específicos de dependência espacial global e local do fenômeno a partir do SOM.

1.3 Contribuições

A principal contribuição deste trabalho é a proposição de uma metodologia de análise exploratória de dados geospaciais baseada nos Mapas Auto-Organizáveis de Kohonen, metodologia esta suportada pelo sistema *CASA*, também desenvolvido no escopo desta dissertação. A metodologia contempla a análise de presença de dados atípicos, a análise de correlação e significância de componentes, distribuição espacial do fenômeno, detecção automática de agrupamentos e análise da dependência espacial.

Como contribuições conseqüentes da metodologia têm-se o uso dos Planos de Componentes como mecanismos automáticos de verificação da distribuição espacial do fenômeno estudado e a aplicação do algoritmo de segmentação automática dos vetores de código do SOM em conjunto com os índices de validação de partição de agrupamentos, Davies-Bouldin e CDbw, na tarefa de partição do conjunto de dados relativos às áreas estudadas.

Também como produto deste trabalho foi desenvolvida a biblioteca de classes SOMLib, cuja finalidade é prover os desenvolvedores de sistemas conexonistas com um projeto de código aberto, com alto grau de manutenibilidade e facilidade de uso. O SOMLib implementa o Mapa Auto-Organizável e os algoritmos relacionados, como a U-matriz, algoritmos de partição de dados, Planos de Componentes etc. O sistema *CASA* é um ambiente gráfico de interface com o usuário, desenvolvido em C++, com o auxílio da biblioteca de classes QT 3.2.0 e sobre a biblioteca SOMLib, cujo objetivo é facilitar o uso combinado das diversas ferramentas de visualização e de análise de agrupamentos implementadas neste trabalho.

Finalmente, considerando os poucos trabalhos na área, tem-se uma contribuição na aplicação dos Mapas Auto-Organizáveis na análise espacial de um problema urbano. Como colocado por Franzini et al. (2001, p. 2)⁴: “*As potencialidades das RNAs ainda estão inexploradas, especialmente quando aplicadas a estudos urbanos*”.

1.4 Organização da Dissertação

O Capítulo 1 faz uma breve introdução ao problema, apresenta os objetivos e expõe as contribuições deste trabalho.

O Capítulo 2 faz uma revisão bibliográfica sobre o Mapa Auto-Organizável de Kohonen

⁴Artificial Neural Network (ANN) possibilities are still largely unexplored, specially when applied to urban studies.

e como este pode ser usado como ferramenta de análise exploratória de dados multivariados. Neste capítulo são abordados os algoritmos básicos do SOM como algoritmos de aprendizagem e métodos de visualização.

O Capítulo 3 faz uma breve revisão dos métodos de partição automática de dados a partir da RNA do tipo SOM, enfatizando o algoritmo Costa-Netto. Este capítulo também trata dos índices de avaliação ou validação dos agrupamentos gerados pela rede neural. O Capítulo 4 faz uma breve revisão de literatura sobre o uso do SOM na Análise Espacial e descreve as contribuições deste trabalho no uso do SOM na Análise Espacial de Dados em Área. O Capítulo 5 trata do projeto e implementação do Mapa Auto-Organizável e do sistema *CASA*, desenvolvido para auxiliar o processo de Análise Espacial de Dados em Área. Em cada capítulo é realizada uma revisão bibliográfica e logo após é descrita a contribuição deste trabalho no tópico. No Capítulo 3 as contribuições são descritas a partir da Seção 3.4.3, no Capítulo 4 a partir da Seção 4.3 e no Capítulo 5 a partir da Seção 5.2.

No Capítulo 6 usam-se as técnicas, métodos e sistemas apresentados ou propostos nos capítulos anteriores no estudo de caso de exclusão/inclusão social intra-urbana em São José dos Campos-SP. As conclusões e discussões finais são apresentadas no Capítulo 7.

1.5 Sumário

Este capítulo teve o objetivo de introduzir o leitor no contexto dos temas abordados nesta dissertação através da exposição das motivações, objetivos e contribuições relevantes.

CAPÍTULO 2

MAPAS AUTO-ORGANIZÁVEIS

2.1 Redes Neurais Artificiais

As Redes Neurais Artificiais constituem-se em modelos computacionais paralelos baseados numa unidade atômica, o neurônio (Figura 2.1). Em geral, estes modelos possuem inspiração neurobiológica, porém, na prática, são algoritmos computacionais representando, de maneira bastante elementar, o mecanismo de funcionamento cerebral. Atualmente, existe uma extensa variedade de RNAs disponíveis.

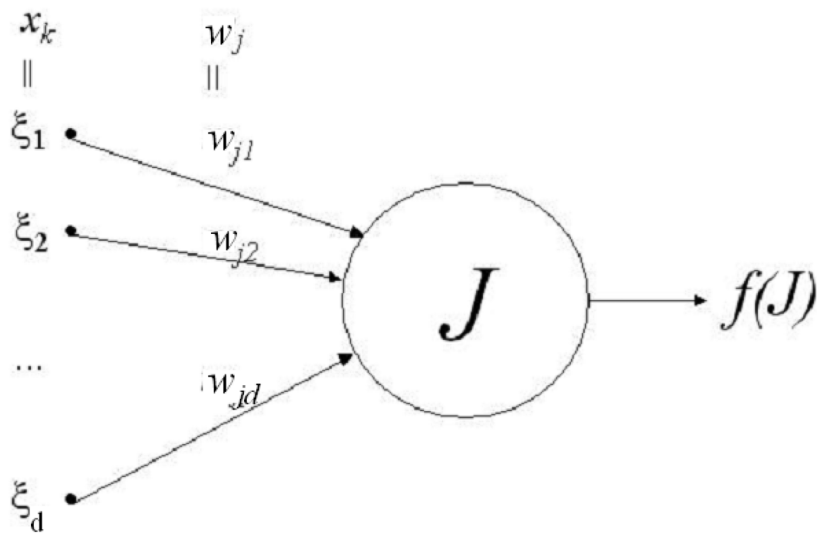


FIGURA 2.1 – Modelo básico de um neurônio j com entrada x_k , pesos sinápticos w_j , nível de ativação J e saída $f(J)$.

As RNA são caracterizadas pela arquitetura, pela característica dos neurônios que as compõem e pela regra de treinamento usada para absorção do conhecimento. Cada neurônio j possui um vetor de dados de entrada $x_k = [\xi_1, \xi_2, \dots, \xi_n]^T$, uma ativação interna J , uma função de ativação $f(J)$ e os pesos sinápticos $w_j = [w_{j1}, \dots, w_{jd}]^T$, que conectam os elementos de x_k ao neurônio j . Existem variações deste modelo básico, assim como várias funções de ativação. As RNAs são formadas pela combinação destas unidades básicas.

As RNAs apresentam como vantagens as características de adaptabilidade, generalização e tolerância a ruídos, dentre outras (Haykin, 2001). Estas características são extremamente importantes quando aplicadas a problemas geográficos, dada a natureza complexa

e ruidosa dos mesmos.

Kohonen (2001) propõe a divisão das redes neurais em três categorias: redes de transferência de sinal, redes de transferência de estado e redes competitivas.

Nas **redes de transferência de sinais** a saída da rede dependerá, única e exclusivamente, do valor de entrada. Estas redes são usadas para transformação de sinais. São exemplos deste tipo de rede aquelas “alimentadas adiante”, como os Perceptrons de Múltiplas Camadas - *Multi-Layer Perceptron (MLP)* (Rumelhart et al., 1986) e as redes de função de base radial - *Radial Basis Function (RBF)* (Bishop, 1995). Essas redes são usadas como identificadores e classificadores de padrões, controle, avaliação de dados de entrada etc.

As redes de transferência de estado têm como base os efeitos de relaxação. A retroalimentação e a não-linearidade são tal que garantem que o estado de atividade rapidamente convirja para um de seus valores estáveis. Os valores de entrada acionam o estado inicial de atividade, a rede então inicia o processamento até chegar no estado final. São exemplos deste tipo de rede, as redes de *Hopfield* (Fausett, 1994) e a máquina de *Boltzman* (Haykin, 2001). As principais aplicações destas redes são: como função de memória associativa e em problemas de otimização, embora também sejam usadas no reconhecimento de padrões.

As redes de aprendizagem competitiva estão baseadas no processo competitivo de aprendizagem entre suas unidades. Uma das principais aplicações destas redes é a descoberta de agrupamentos de dados. Estão incluídas nesta categoria as redes SOM (Kohonen, 2001) e ART - *Adaptive Resonance Theory* (Fausett, 1994). A aprendizagem competitiva é um processo adaptativo onde os neurônios, numa rede neural artificial, tornam-se gradualmente sensíveis a diferentes categorias de entrada e a conjuntos de amostras num domínio específico do espaço de entrada. Este trabalho de pesquisa concentrou-se nesta última categoria, mais especificamente no SOM. Nas seções seguintes serão descritos com mais detalhes, a arquitetura, o algoritmo de aprendizagem e as aplicações do SOM.

2.2 Mapas Auto-Organizáveis

O Mapa Auto-Organizável de Kohonen é uma RNA com duas camadas (Kohonen, 2001): a camada de entrada I e a de saída U . A entrada da rede corresponde a um vetor no espaço d -dimensional em \mathfrak{R}^d , representado por $x_k = [\xi_1, \dots, \xi_d]^T$, $k = 1, \dots, n$, sendo n o número de vetores de entrada. Cada neurônio j da camada de saída possui um vetor de código w , também no espaço \mathfrak{R}^d , associado ao vetor de entrada x_k , $w_j = [w_{j1}, \dots, w_{jd}]^T$.

Os neurônios da camada de saída estão interconectados por uma relação de vizinhança que descreve a estrutura do mapa. Por exemplo, na Figura 2.2 tem-se um mapa com a camada de saída U , bidimensional, retangular, de dimensões $N \times M$. Nesta figura somente estão representados os vetores de código w , conectados ao neurônio j .

O SOM foi idealizado a partir da analogia com a região do córtex cerebral humano. Descobriu-se que esta parte do cérebro aloca regiões específicas para atividades específicas e que, para uma determinada ativação cerebral, o grau de ativação dos neurônios diminuía à medida que se aumentava a distância da região de ativação inicial (Kohonen, 2001).

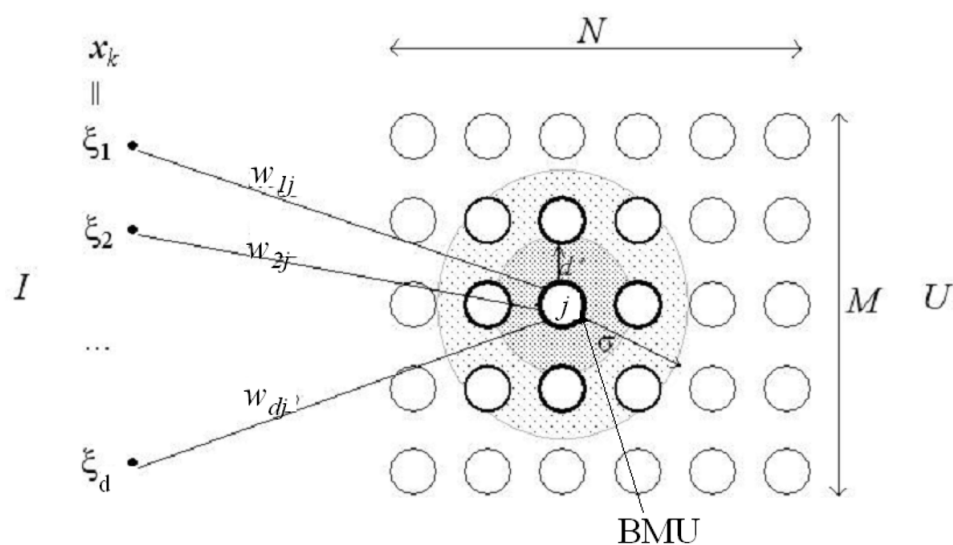


FIGURA 2.2 – Exemplo de um Mapa Auto-Organizável bidimensional $N \times M$, com entrada x_k .

Existem diferentes topologias para estruturação de um Mapa Auto-Organizável, sendo que a estrutura mais comum é a de duas dimensões. Na Figura 2.3 vê-se uma rede SOM unidimensional (a), uma rede bidimensional com organização hexagonal dos neurônios, com 6-vizinhos (b) e uma rede bidimensional com disposição retangular dos neurônios, com 4-vizinhos (c).

Desde o seu surgimento, em 1982, o SOM vem sendo aplicado numa ampla variedade de problemas de engenharia, medicina etc. Destacam-se as potencialidades de visualização de dados multivariados, análise de agrupamentos, mineração de dados, descoberta de conhecimento e compressão de dados (Kohonen, 2001).

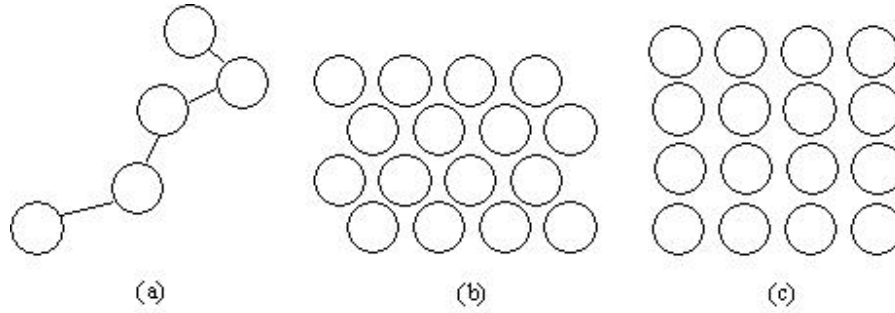


FIGURA 2.3 – Diferentes topologias para estruturação dos mapas auto-organizáveis: a) unidimensional; b) bidimensional hexagonal; c) bidimensional retangular.

2.2.1 Aprendizagem padrão ou seqüencial

O algoritmo básico de treinamento do SOM consiste de três fases. Na primeira fase, competitiva, os neurônios da camada de saída competem entre si, segundo algum critério, geralmente a distância Euclideana, para encontrar um único vencedor, também chamado de BMU (*Best Match Unit*). Na segunda fase, cooperativa, é definida a vizinhança deste neurônio. Na última fase, adaptativa, os vetores de código do neurônio vencedor e de sua vizinhança são ajustados.

A relação de vizinhança entre os neurônios é estabelecida segundo alguma função. O principal objetivo da função de vizinhança é controlar o nível de atuação dos neurônios em torno do neurônio vencedor do processo competitivo. Seguindo o modelo neurobiológico tem-se que o nível de atuação dos neurônios vizinhos decai à medida que o mesmo se distancia do BMU.

Seja $h_{j,i}$ a vizinhança topológica centrada no neurônio i e com um conjunto de neurônios cooperativos J , $j \in J$. Seja $d_{i,j}$ a distância lateral entre o neurônio vencedor i e o neurônio j . Para que $h_{j,i}$ atenda aos requisitos neurobiológicos, a mesma tem que ser simétrica em relação ao ponto de valor máximo ($d_{i,i} = 0$), e $h_{j,i}$ deve decair monotonicamente com o aumento da distância lateral ($d_{i,j}$), decaindo para próximo de 0 quando $d_{i,j} \rightarrow \infty$.

A função gaussiana $h_{j,i} = \exp(-d_{i,j}^2/2\delta^2)$ satisfaz estas exigências e é invariante à translação. δ representa o raio da vizinhança topológica e o grau que os neurônios vizinhos do BMU participam do processo de aprendizagem adaptativa. A rede SOM converge mais rapidamente com este tipo de função de vizinhança (Lo et al., 1991; Erwin et al., 1992; Lo et al., 1993). $\delta(t)$ é definido como uma função monotonicamente decrescente em função do tempo (épocas), $\delta(t) = \delta(0)\exp(-t/\tau_i)$, sendo τ_i uma constante.

Seja Ξ o conjunto dos padrões de entrada composto por $x_k, k = 1, \dots, n$, tem-se o algoritmo de aprendizagem padrão ou seqüencial, como segue:

a) Os vetores de código, $w_j = [w_{j1}, \dots, w_{jp}]^T$, são iniciados linearmente (ver apêndice A).

b) Para cada época t

1) Para todo $x_k \in \Xi, k = 1, \dots, n$, para o tempo discreto t , encontre o neurônio vencedor c segundo a distância Euclideana:

$$c = \operatorname{argmin}_j \{\|x_k - w_j\|\}, j = 1, \dots, m \quad (2.1)$$

onde m corresponde ao número de neurônios na rede. A ordem de apresentação dos padrões deve ser aleatória.

2) Os vetores de código w_j do neurônio vencedor e dos seus vizinhos são, então, atualizados segundo a equação:

$$w_{ji}(t+1) = w_{ji}(t) + \alpha(t) h(t) [x_{ik}(t) - w_{ij}(t)] \quad (2.2)$$

onde $\alpha(t)$ é uma função que determina a taxa de aprendizagem na iteração t e $h(t)$ é a função que determina a vizinhança entre o neurônio vencedor c e seus vizinhos.

2.2.2 Aprendizagem em lote

O algoritmo de aprendizagem em lote difere da aprendizagem seqüencial quanto à forma de atualização dos vetores de código, na ausência da taxa de aprendizagem $\alpha(t)$ e na não obrigatoriedade de apresentação aleatória dos padrões. Neste algoritmo, os vetores de código são atualizados ao final de cada época. Em cada passo (época) o conjunto de dados é particionado de acordo com as regiões de Voronoi dos vetores de código do Mapa neural, definido segundo o critério de proximidade do vetor de código com o conjunto de dados. Os vetores de código w podem ser atualizado a partir das equações que seguem (Vesanto e Ahola, 1999):

$$s_i(t) = \sum_j^{nv_i} x_j \quad (2.3)$$

onde s_i representa o somatório dos padrões relativos à região de Voronoi V_i e n_{V_i} corresponde ao número de amostras do conjunto de Voronoi do neurônio i .

$$w_i(t+1) = \frac{\sum_j^m h_{ji}(t) s_j(t)}{\sum_j^m n_{V_j} h_{ji}(t)} \quad (2.4)$$

Em virtude da ausência do parâmetro de aprendizagem $\alpha(t)$ e do bom desempenho do algoritmo em lote (Costa, 1999; Vesanto e Ahola, 1999) decidiu-se optar pelo mesmo no decorrer deste trabalho. A ausência deste parâmetro e a independência do resultado quanto à ordem de apresentação dos padrões facilitam o processo de análise por usuários não muito experientes na análise com SOM.

2.2.3 Considerações sobre o treinamento do SOM

Para o algoritmo de aprendizagem seqüencial as apresentações dos padrões Ξ têm de ser de forma aleatória, para que seja garantida a uniformidade de apresentação de todos os $x_k \in \Xi$. Para o algoritmo em lote não existe esta restrição.

Define-se uma época como a apresentação completa do conjunto de padrões Ξ à rede. A normalização dos dados de entrada e vetores de código não é obrigatória (Demartines e Blayo, 1992), mas pode ser feita segundo a Equação 2.5, seja ξ_i o componente i do padrão x_k , então:

$$\xi'_i = \xi_i / \|x_k\| \quad (2.5)$$

O processo competitivo é o mais custoso em processo de treinamento. Geralmente trata-se de uma busca seqüencial pelo neurônio vencedor. Este processo pode ser otimizado usando-se algum mecanismo heurístico para minimizar a busca (Costa, 1999), ou por meio da paralelização do algoritmo (Openshaw e Turton, 1996).

A determinação dos parâmetros de aprendizagem em geral é empírica, fortemente baseada na experiência do usuário e em métodos de tentativa e erro. Porém, algumas técnicas de determinação automática dos parâmetros de aprendizagem têm sido propostas, seja por meio de algoritmos genéticos (Silva e Rosa, 2002), ou métodos numéricos (Haese, 1998; Haese e Goodhill, 2001). A taxa de aprendizagem $\alpha(t)$, Equação 2.2, deve assumir um valor pré-fixado, $\alpha(0) < 1$, e deve decair com o tempo t até um valor próximo de zero. Analogamente, a função de vizinhança $h(t)$ também deve assumir um valor pré-fixado, $h(0)$, adequado de forma que maximize a qualidade da formação do mapa.

A dimensionalidade do mapa auto-organizável e seu tamanho (m) dependerão do tipo de problema e propósito. A literatura mostra que a determinação do tamanho do SOM é um processo empírico (Flexer, 2001; Kohonen, 2001). Em geral, o SOM bidimensional $N \times M$ é usado devido sua capacidade de projeção dos dados de dimensão p num Mapa bidimensional. Este trabalho está baseado única e exclusivamente neste tipo de Mapa. O tamanho da amostra de treinamento também auxilia o processo de decisão sobre o tamanho do Mapa. Para grandes volumes de dados, Mapas razoavelmente grandes são mais adequados. Todavia, grandes Mapas comprometem o desempenho do algoritmo e Mapas muito pequenos comprometem a integridade da formação topológica do SOM (Costa, 1999; Flexer, 2001; Kohonen, 2001; Park et al., 2003).

2.2.4 Avaliação da qualidade da aprendizagem

Existe um conjunto razoável de mecanismos de avaliação da qualidade do Mapa gerado após o processo de aprendizagem. Escolheu-se duas destas métricas, o erro da quantização vetorial e o erro topológico (Kohonen, 2001).

O erro de quantização (E_q) corresponde à média do erro correspondente à diferença entre o vetor de características x_k e o vetor de código w_{BMU} , vetor de código vencedor no processo competitivo para o padrão x_k :

$$E_q = \frac{\sum_{k=1}^n \|x_k - w_{BMU}\|}{n} \quad (2.6)$$

O erro topológico (E_t) procura avaliar o quanto a estrutura da grade aproxima padrões próximos no espaço de entrada. Considerando que, para cada padrão x_k tem-se o BMU como o primeiro neurônio na ordem de competição na grade, o BMU2 corresponderá ao segundo neurônio nesta escala. Assim, o erro topológico corresponderá ao percentual de padrões cujo BMU e BMU2 não são vizinhos na grade:

$$E_t = \frac{1}{n} \sum_{k=1}^n u(x_k) \quad (2.7)$$

onde $u(x_k)$ corresponde a 1, se o BMU e BMU2 não são vizinhos, e 0 caso contrário.

2.2.5 Propriedades

Uma vez concluído o processo de aprendizagem da rede SOM, o mapa de códigos gerado, representado pelos vetores w_j , mostrará propriedades importantes dos dados de entrada

(Haykin, 2001; Kohonen, 2001).

- *Propriedade 1.* Ordenação topológica. O mapa de características calculado pelo algoritmo SOM é ordenado topologicamente, no sentido de que a localização espacial de um neurônio na grade corresponde a um domínio particular ou características dos padrões de entrada. O inverso nem sempre é verdadeiro.
- *Propriedade 2.* Casamento de densidade. O mapa de características reflete variações na estatística da distribuição da entrada, embora a distribuição das unidades do SOM não seja exatamente a mesma da distribuição dos dados amostrais (para SOM 1D a densidade das unidades de saída é proporcional a $p(x_k)^{2/3}$ em torno do ponto x_k).
- *Propriedade 3.* Seleção de características. Pode-se afirmar que os Mapas Auto-Organizáveis fornecem uma aproximação discreta das assim chamadas curvas principais, e podem, portanto, ser vistos como uma generalização não-linear da análise de componentes principais.

Este trabalho baseou-se nessas propriedades para, através de métodos distintos, proceder à análise exploratória de dados geoespaciais multivariados.

2.3 Visualização do Mapa Auto-Organizável

Após o processo de aprendizagem do Mapa é necessário que se possa verificar visualmente o resultado da ordenação topológica. Destacam-se três formas de representação visual. A primeira forma usa os vetores de código como coordenadas no espaço d -dimensional. Este processo pode ser aplicado quando $d \leq 3$. A segunda forma é através da matriz de distância entre os vetores de código. Esta matriz, em especial a matriz de distância unificada (Ultsch, 1993), pode ser analisada como uma imagem, o que facilita o processo de análise. A terceira forma, os Planos de Componentes, usa os valores de cada componente dos vetores de código para colorir o Mapa Auto-Organizável. Este método permite que seja avaliada a distribuição do componente no Mapa, após a aprendizagem.

2.3.1 Representação dos vetores de código no espaço \mathfrak{R}^d

Para o caso onde os vetores de código possuem dimensão d , menor ou igual a 3, pode-se usar os seus valores como coordenadas no espaço \mathfrak{R}^d para visualização da organização dos neurônios. Dado o conjunto de dados da Figura 2.4 (à esquerda), onde $d = 3$, correspondente a dois toróides que formam um elo de corrente. Treinando-se uma rede 15×15 hexagonal com aprendizagem em lote, pode-se visualizar o resultado final do treinamento, usando os valores dos vetores de código como coordenadas no espaço \mathfrak{R}^3 , Figura 2.4

(à direita).

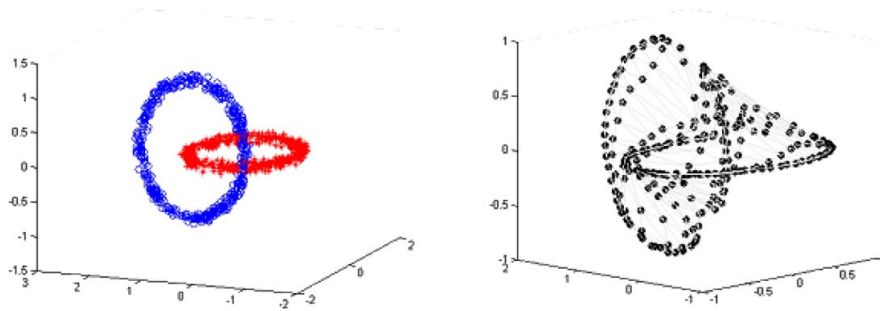


FIGURA 2.4 – Do lado esquerdo tem-se os dados, sendo dois conjuntos de dados formando um elo de corrente; do lado direito tem-se a estrutura final do SOM, após treinamento, usando os valores dos vetores de código, como coordenadas no espaço \mathcal{R}^3 .

Existem outras formas de projeção dos valores dos vetores de código no espaço \mathcal{R}^2 como através do mapa de Sammon ou através de outros métodos de projeção (Kaski et al., 1999; Kohonen, 2001). Todavia, estes métodos não foram tratados neste trabalho.

2.3.2 Histograma

Os dados podem ser projetados no Mapa pesquisando-se, para cada x_k , o seu respectivo BMU. A frequência de ocorrências de BMUs para um dado neurônio gera um histograma que refletirá o nível de atividade $H(i)$ deste neurônio. Destacam-se dois tipos de histogramas, aquele que registra o número de BMUs após a fase de aprendizagem e o que registra a frequência de ocorrências nesta fase. Ambas abordagens são úteis, todavia neste trabalho o nível de atividade $H(i)$ referir-se-á ao primeiro tipo de histograma.

2.3.3 Planos de Componentes

Para que se possa ter uma noção de como cada componente do vetor de característica x_k organizou-se no Mapa treinado, usa-se algum método de coloração do SOM baseado nos valores de cada componente. Para um dado componente j , de um Mapa bidimensional $M \times N$, gera-se uma imagem $f(x, y)$ com dimensões iguais às do Mapa $M \times N$, 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, preto para valores mínimos e tons de cinza para valores intermediários.

2.3.4 Matriz de distância unificada (U-Matriz)

A matriz de distâncias unificada, U-matriz (Ultsch, 1993), 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, distância Euclideana, 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 $N \times M$ gera uma imagem $(2N - 1) \times (2M - 1)$.

Dado um Mapa bidimensional hexagonal encontra-se a U-matriz calculando-se as distâncias dx , dy e dz (Figura 2.5), 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.

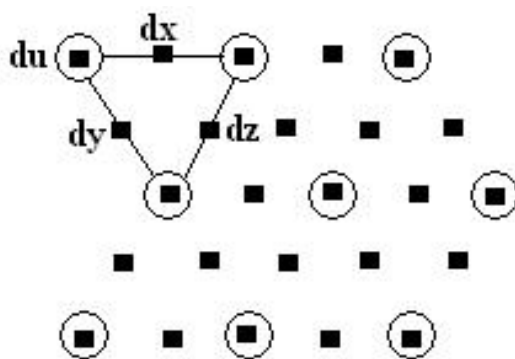


FIGURA 2.5 – Representação dos componentes da U-matriz dx , dy , dz e du para uma rede 3x3 hexagonal. Os círculos representam os neurônios e os quadrados representam os valores dos componentes.

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.

Pelo fato da U-matrix gerar uma imagem relativamente complexa (Figura 2.6), sua principal aplicação é a visualização do mapa para separação manual dos agrupamentos.

Porém, já existe alguma iniciativa para detecção automática dos agrupamentos por meio de técnicas de processamento desta imagem (Costa, 1999; Costa e Andrade Netto, 2001).

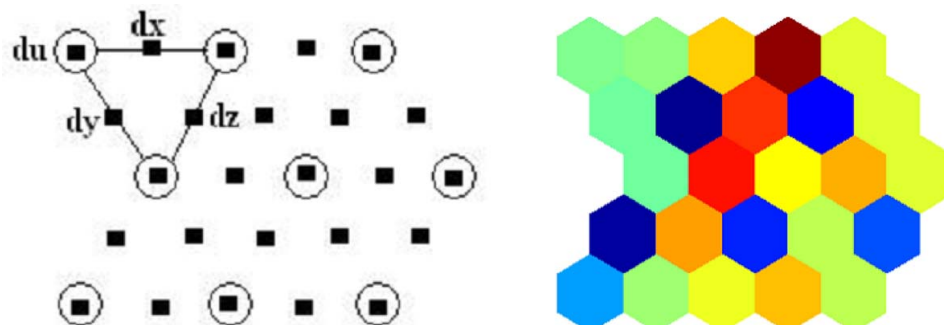


FIGURA 2.6 – Exemplo de geração da imagem relativa a U-matriz, a partir de uma rede 3x3 hexagonal.

A U-matriz é um método cujo objetivo é permitir a detecção visual das relações topológicas dos neurônios (Ultsch, 1993). Esta técnica é extremamente útil quando se tem os vetores de código com dimensão maior que 3. Para estes casos não se pode representar graficamente, ou pelo diagrama de Voronoi ou por superfícies de influência, a organização final dos neurônios.

2.4 Análise Exploratória de Dados com Mapas Auto-Organizáveis

A análise exploratória de dados consiste na busca por padrões em dados amostrais a partir de técnicas de visualização e de análise de agrupamentos, dentre outras. Para o caso de problemas estatisticamente bem conhecidos, quanto maior o volume de dados amostrais melhor a acurácia da análise. Para problemas complexos esta tarefa torna-se difícil e exige que vários métodos sejam avaliados até que se chegue a algum resultado conclusivo. Este trabalho baseou-se em trabalhos que aplicaram, de formas distintas, as propriedades dos Mapas Auto-Organizáveis na análise exploratória de dados (Kaski e Kohonen, 1996; Vesanto, 1997; Vesanto e Ahola, 1999; Vesanto, 1999; Kaski et al., 1999; Kohonen, 2001).

Os estágios da análise exploratória de dados com SOM compreendem a escolha do conjunto de dados, o pré-processamento dos dados, a parametrização da rede e escolha de “bons” Mapas neurais e a interpretação dos resultados (Kaski e Kohonen, 1996). Todas estas fases são críticas e relevantes para a geração de resultados confiáveis (Figura 2.7). Todavia, destaca-se aqui a tarefa de interpretação dos resultados como a mais difícil, em

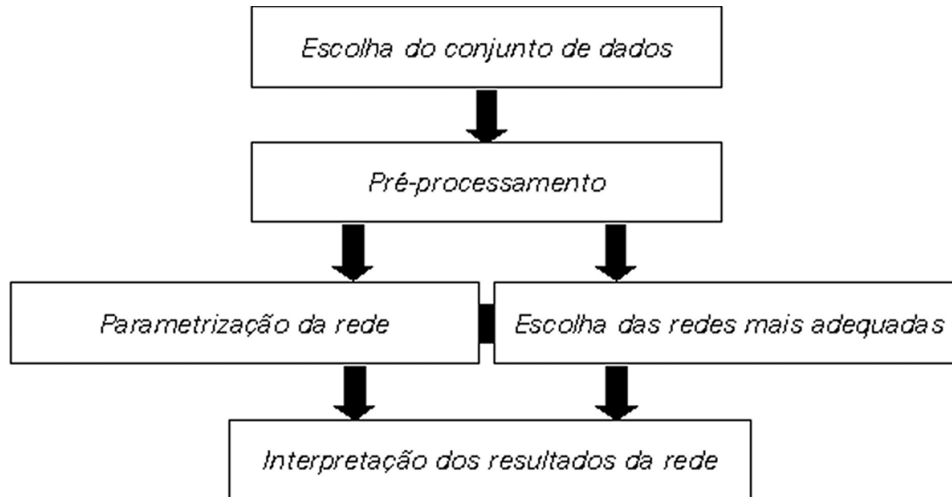


FIGURA 2.7 – Fluxograma do processo de análise exploratória de dados com SOM.

função das variadas formas de análise disponível.

2.4.1 Seleção dos dados

A seleção dos dados a serem analisados dependerá da natureza do problema. O cuidado nas fases de formulação do problema e seleção dos componentes é de extrema importância e segue os mesmos preceitos daqueles usados em qualquer tipo de análise exploratória.

2.4.2 Pré-processamento

Existem diversas técnicas para pré-processamento do conjunto amostral de dados. Cada uma delas possui objetivo distinto e depende do domínio do problema. Dentre os pré-processamentos mais usados em conjunto com o SOM destacam-se a normalização, a marcação de valores ausentes e a remoção de valores atípicos.

2.4.3 Configuração da rede e seleção do Mapa

Assim como nas etapas anteriores, toda a parametrização do Mapa Auto-Organizável dependerá do tipo de aplicação. Classificação de padrões, controle, visualização, detecção de agrupamento, cada tarefa exige que uma análise posterior seja conduzida até que se chegue à configuração ideal da rede neural. Como se aplicou somente o SOM bidimensional, hexagonal, com função de vizinhança gaussiana, com aprendizagem em lote, os comentários estarão restritos à parametrização relativa a este tipo de rede neural. Para este tipo de rede precisa-se definir as dimensões da rede $M \times N$, o raio inicial da função de vizinhança e o número de épocas do processo de aprendizagem.

Uma forma de se avaliar a qualidade do Mapa é através das medidas dos erros de quantização e topológico. Ou seja, para cada configuração de rede tem-se valores distintos destes erros. Assim, pode-se construir dois gráficos em \mathbb{R}^2 , ambos tendo nos eixos das ordenadas as configurações de rede avaliadas e nos eixos das abscissas os valores dos erros de quantização e topológico. As redes candidatas serão aquelas com os menores valores para ambos os erros. Embora este seja um processo válido, existem limitações práticas para o seu uso, como, por exemplo, a irregularidade da curva do erro topológico.

Uma outra forma para determinar os parâmetros iniciais da rede é observar a formação final do Mapa através de algum método de visualização. A U-matriz é um dos métodos mais completos para esta análise. Todavia, observa-se que, para problemas complexos, redes pequenas não conseguem exprimir, através da U-matriz, a estrutura do conjunto de dados amostrais (Costa e Andrade Netto, 2003). Porém, quanto maior a rede, melhor a U-matriz exprime a estrutura real do conjunto de dados (Ultsch, 1999). Pode-se, também, usar algum tipo de índice de avaliação para o Mapa final gerado. Para o caso de separação automática do conjunto de dados em agrupamentos distintos pode-se usar alguma métrica de validação de particionamento de dados para verificar a eficácia da rede neural. Neste trabalho usou-se estas três técnicas para avaliação dos Mapas nas diferentes fases do processo de análise exploratória dos dados geoespaciais.

2.4.4 Interpretação do Mapa neural

Neste trabalho o SOM é usado para tarefas específicas como: detecção de dados atípicos, análise de componentes, análise da distribuição espacial do fenômeno estudado, detecção automática de agrupamentos e análise da dependência espacial. As técnicas do uso do SOM para a Análise Espacial serão estudadas no Capítulo 4. Nesta seção será abordado somente o uso do SOM para detecção de dados atípicos, análise de correlação e significância de componentes e análise simples de agrupamentos.

2.4.4.1 Detecção de dados atípicos

Dados atípicos identificados pelo SOM são, em geral, os mesmos detectados por procedimentos estatísticos como análise de fatores e o k-médias (Morlini, 1998; Muñoz e Muruzábal, 1998). Isto deve-se ao fato de que os vetores de código correspondem a uma aproximação da densidade dos dados de entrada; logo, tem-se que o problema de detecção de dados atípicos no conjunto amostral de dados pode ser transferido para a detecção de dados atípicos no conjunto de vetores de código da grade de neurônios.

Vetores de código atípicos podem ser observados de diversas maneiras. Destacam-se os

métodos de Morlini (1998), que usa a distância média, para cada neurônio, do vetor de código analisado para os seus vizinhos, identificando o vetor atípico através dos maiores valores para esta média, e de Muñoz e Muruzábal (1998) que define um fluxo de passos que devem ser seguidos para se detectar dados atípicos num conjunto de dados, baseado em ferramentas auxiliares ao Mapa Auto-Organizável. Uma outra forma de análise é através do uso da U-matriz, onde os vetores atípicos são identificados por regiões pequenas e separadas das demais regiões por áreas de transição.

Justifica-se o uso do SOM para a detecção de dados atípicos devido ao fato de que este método é computacionalmente barato, de fácil interpretação e não está baseado em modelos específicos. O SOM é robusto o suficiente para gerar, a partir de configurações distintas, os mesmos resultados quanto à identificação de vetores de código atípicos (Morlini, 1998).

2.4.4.2 Análise de componentes

Durante o processo de aprendizagem os vetores de código se aproximam discretamente dos padrões de entrada, mantendo a ordenação topológica dos mesmos. Ou seja, estes vetores de código são uma aproximação da distribuição de densidade dos padrões de entrada. A visualização desses vetores de código podem auxiliar no entendimento e na contribuição de cada componente. Esta visualização está relacionada com a análise de componentes principais e está diretamente relacionada à discriminação da influência de cada componente no Mapa neural (Kohonen, 2001). Para analisar a contribuição de cada componente na formação dos agrupamentos do Mapa após a aprendizagem usa-se os Planos de Componentes. Os Planos de Componentes vêm sendo amplamente usados nesta tarefa (Kaski e Kohonen, 1996). Vesanto e Ahola (1999) propõem um método automático para busca de componentes correlacionados. Este método não foi usado neste trabalho devido ao mesmo só se aplicar para um número muito grande de componentes a serem analisados.

2.4.4.3 Análise simples de agrupamentos

Um expediente comum na análise de agrupamentos, usando o SOM, é definir o número de neurônios como o número de agrupamentos possíveis (Ultsch, 1993). Após o treinamento da rede, cada neurônio estará associado a um grupo de vetores de entrada. Embora seja um método válido, já que se trata de uma análise exploratória de dados, este procedimento impõe uma restrição sobre a estrutura dos agrupamentos, pois assume-se uma estrutura hiperesférica para cada grupo de dados. Este método é aplicado para o caso de redes pequenas, pois a separação manual de padrões nessas redes é mais fácil e menos

trabalhosa.

2.5 Sumário

Os Mapas Auto-Organizáveis são estruturas neurais artificiais formuladas sobre o conceito de auto-organização de unidades básicas (neurônios) segundo algum processo de aprendizagem competitiva. A escolha da topologia e do mecanismo de aprendizagem dependem de fatores como tipo do dado a ser analisado, grau de generalização desejado, tipo de aplicação etc. São propriedades do SOM a ordenação topológica, casamento de densidade e seleção de características.

Em função das razões expostas neste capítulo foi definida como rede de trabalho o Mapa neural bidimensional, com disposição hexagonal da grade de neurônios, função de vizinhança gaussiana e aprendizagem em lote. A avaliação dos Mapas será de acordo com as medidas do erro de quantização e topológico.

Visualização e análise de agrupamentos são as principais aplicações dos Mapas Auto-Organizáveis. Dentre as técnicas existentes de visualização foram usadas neste trabalho a U-matriz e os Planos de Componentes. Para análise de agrupamentos será usado um mecanismo de detecção automática de agrupamentos exposto no Capítulo 3.

Embora as ferramentas relacionadas com o SOM para análise exploratória de dados sejam numerosas, as mesmas não cobrem todo o espectro de Análise Espacial de Dados em Área. Um outro fator a ser analisado é o quão automático o processo pode ser para que qualquer usuário do SOM possa proceder com a análise, sem grandes esforços de entendimento e interpretação dos resultados.

Os três próximos capítulos abordam temas relativos a essas questões. O Capítulo 3 dedicou-se à pesquisa de métodos automáticos de detecção de agrupamentos. Adaptações e uso das técnicas de análise exploratória com o SOM foram extendidas no Capítulo 4, onde foram propostas técnicas para proceder a Análise Espacial de Dados em Área com o SOM. No Capítulo 5 deu-se especial atenção aos métodos de acesso à base de dados geográfica, de forma a tornar ainda mais fácil o acesso e posterior análise de dados geográficos.

CAPÍTULO 3

DETECÇÃO AUTOMÁTICA DE AGRUPAMENTOS COM MAPAS AUTO-ORGANIZÁVEIS

3.1 Introdução

Como observado no Capítulo 2, existem vários mecanismos de análise exploratória de dados através dos Mapas de Kohonen. A tarefa de descoberta de agrupamentos é uma delas e tem sido feita visualmente, através da projeção do Mapa por meio da U-matriz e dos Planos de Componentes. Todavia, existem casos onde a complexidade da U-matriz gerada inviabiliza ou dificulta a descoberta de agrupamentos pela verificação visual. Para estes casos seriam bastante úteis técnicas de detecção automática de agrupamentos baseadas nos vetores de código gerados pelo SOM.

O método de identificação visual de agrupamentos através da U-matriz apresenta algumas restrições. Para Mapas pequenos a U-matriz gerada tende a ser complexa e de difícil identificação visual dos agrupamentos (Figura 2.6), além do que, a U-matriz só pode ser gerada a partir de mapas com grade bi-dimensional. Para mapas com dimensões de grade maior que 2 o processo de visualização da matriz de distância torna-se complexo.

Este capítulo avaliou o método de segmentação automática do SOM proposto por Costa e Netto (2003). Este método foi aplicado em conjunto com os índices de validação de partição de dados, Davies-Bouldin (Davies e Bouldin, 1979) e o CDbw (Halkidi e Vazirgiannis, 2002). Neste capítulo também foi realizada uma breve revisão bibliográfica sobre outros métodos automáticos de segmentação do SOM.

3.2 Métodos Automáticos de Segmentação

Com o objetivo de particionar e rotular automaticamente um SOM treinado, baseando-se no gradiente dos componentes, cuja informação é apresentada na U-matriz, foi desenvolvido o algoritmo *SL-SOM Self-Labeling SOM* (Costa, 1999; Costa e Andrade Netto, 2001). O algoritmo *SL-SOM* usa o método de segmentação de imagens *watershed* para particionar a U-matriz em regiões conectadas. O algoritmo *SL-SOM* somente se aplica a Mapas com grade bidimensional. Esta restrição não chega a ser proibitiva devido ao fato de que a maioria das aplicações do SOM presentes na literatura usam este tipo de rede.

Embora o algoritmo *SL-SOM* tenha obtido bons resultados (Costa, 1999), como pode ser observado através do exemplo da Figura 3.1, a sua aplicação não é recomendada para Mapas com poucos neurônios ou problemas cujos possuam estrutura complexa, pois a U-

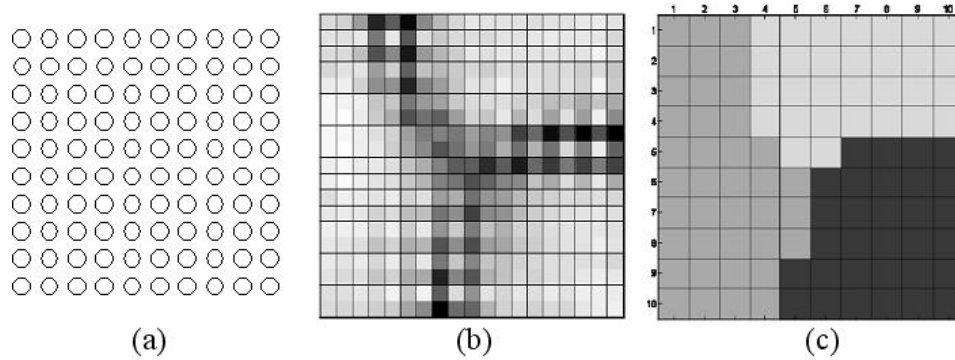


FIGURA 3.1 – Exemplo de segmentação da U-matriz (*SL-SOM*): (a) SOM bidimensional 10x10; (b) U-matrix gerada a partir desta SOM treinada; (c) Rotulação dos neurônios da SOM com o auxílio da imagem (b).
 FONTE:(Costa, 1999).

matriz para estes casos tende a ser de difícil interpretação, dificultando a separação dos padrões. Apesar das limitações do algoritmo com relação à complexidade da U-matriz pode-se afirmar que o SL-SOM oferece um bom mecanismo de investigação de dados multivariados.

A proposta de Vesanto e Alhoniemi (2000) é a de usar o SOM como um redutor do tamanho do conjunto de dados a ser analisado. O trabalho propõe a descoberta de agrupamentos em duas fases, primeiro os dados são usados para treinar uma rede SOM, os vetores de código da rede serão, então, usados para a descoberta de agrupamentos através de métodos tradicionais de descoberta de agrupamentos. O autor usou um método hierárquico aglomerativo e o algoritmo de k-médias. A principal função da rede SOM, neste método, é a de reduzir o volume de dados e, conseqüentemente, o custo computacional do processamento dos algoritmos tradicionais de agrupamento. Além de reduzir este custo computacional, o SOM também diminui o número de dados atípicos, reduzindo o seu impacto na quantização vetorial.

Como critério para fusão ou separação de grupos de dados, o autor usou o índice Davies-Bouldin (Davies e Bouldin, 1979). Este índice também foi usado na comparação entre os métodos de descoberta de agrupamentos. A validação comparou os resultados obtidos usando-se os vetores de código do SOM diretamente sobre o conjunto de dados. Os resultados foram semelhantes para ambos os casos. Observou-se que, neste processo de descoberta de agrupamentos exige-se bastante interação do usuário, não configurando, portanto, um método totalmente automático. Além de que, trata-se do uso de técnicas tradicionais para detecção de agrupamentos em um conjunto de dados menor do que o original. Este método não faz uso direto de informações agregadas aos neurônios como o

seu nível de atividade, formação topológica etc.

Em função da simplicidade e generalidade escolheu-se o método de segmentação do SOM baseado no particionamento de grafos, ou algoritmo Costa-Netto, detalhado na próxima seção e aplicado no estudo de caso do Capítulo 6.

3.3 Segmentação Baseada em Particionamento de Grafos

Costa e Netto (2003) propõem um método para segmentação do mapa baseado no particionamento de grafos. Neste caso, o algoritmo é independente da U-matriz e da dimensão da grade da rede SOM. O algoritmo proposto baseia-se em informações geométricas de distância entre os neurônios, no erro de quantização e no nível de atividade do neurônio. A estratégia é considerar que todos os neurônios fazem parte de um grafo não orientado, totalmente conectado e, a partir de regras heurísticas, eliminar conexões inconsistentes entre neurônios vizinhos, restando grupos conectados, representando agrupamentos distintos (Figura 3.2).

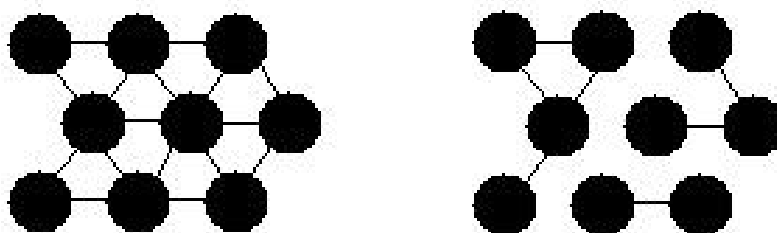


FIGURA 3.2 – A estratégia de segmentação do SOM baseia-se na eliminação de conexões inconsistentes entre os neurônios. Neste exemplo, uma rede 3x3 foi particionada de forma que 3 grupos foram constituídos.

Para um Mapa Auto-Organizável treinado tem-se o algoritmo Costa-Netto como segue:

- a) Obter as distâncias entre os pesos dos neurônios adjacentes i e j , $d(w_i, w_j)$; e a atividade de cada neurônio i , $H(i)$;
- b) Para cada par de neurônios adjacentes i e j , a aresta será considerada inconsistente:

Se a distância entre os pesos excede em 2 a distância média dos outros neurônios adjacentes a i ou a j ;

Se os dois neurônios adjacentes i e j possuem atividade (H) abaixo de 50% da mínima permitida (H_{min}), ou um dos neurônios for inativo ($H(i) = 0$); $H_{min} = \omega H_{med}$, sendo que $0.1 \leq \omega \leq 0.6$ e $H_{med} = n/m$.

Se a distância entre os centróides dos conjuntos de dados associados aos neurônios i e j exceder em 2 vezes a distância entre os pesos $d(w_i, w_j)$;

- c) Remoção dos ramos (arestas) inconsistentes. Para cada aresta (i, j) considerada inconsistente resultará em uma conexão nula no endereço (i, j) da matriz de adjacência A . Ramos consistentes recebem entrada 1 no endereço (i, j) de A ;
- d) Atribuir um código distinto para cada conjunto de neurônios conectados. Remover componentes conectados pequenos (com menos de 3 neurônios).

O que acontece com a aplicação do algoritmo é uma poda dos neurônios conectados adjacentes. Ou seja, ao final vários grupos de neurônios conectados estarão representando um agrupamento específico. O algoritmo é independente da U-matriz e da dimensionalidade da grade do Mapa, o que o torna mais genérico que a proposta de segmentação baseada na U-matriz (Costa, 1999). O algoritmo faz uso de alguns limiares empíricos definidos por meio de experimentações, porém, consegue particionar os dados usando somente as informações inerentes ao Mapa treinado, como a distância entre os neurônios, o erro de quantização e o nível de atividade.

3.4 Índices de Validação dos Agrupamentos

Para validar o particionamento dos dados gerados pelo algoritmo de segmentação baseado no particionamento de grafos usou-se dois índices, já aplicados aos Mapas Auto-Organizáveis. O índice Davies-Bouldin (Davies e Bouldin, 1979), usado para auxiliar o processo de definição do número de agrupamentos corretos (Vesanto e Alhoniemi, 2000; Park et al., 2003) e o índice CDbw (Halkidi e Vazirgiannis, 2002) usado numa aplicação semelhante à anterior (Wu e Chow, 2004).

3.4.1 Índice Davies-Bouldin

O índice Davies-Bouldin (Davies e Bouldin, 1979) é uma medida que indica a similaridade entre agrupamentos. Esta medida pode ser usada para a avaliação da partição dos dados e, conseqüentemente, para a comparação relativa entre diferentes divisões do conjunto de dados. O índice Davies-Bouldin é independente do número de agrupamentos e do método de partição dos dados, o que o torna indicado para a avaliação de algoritmos de partição de dados.

O índice Davies-Bouldin é dado por:

$$\frac{1}{c} \sum_{k=1}^c \max_{c \neq l} \left\{ \frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Q_l)} \right\} \quad (3.1)$$

onde: c é o número de agrupamentos, $S_c(Q_k)$ representa a distância intra-agrupamento (dispersão interna do agrupamento Q_k), baseado na distância para o centróide, $d_{ce}(Q_k, Q_l)$ representa a distância entre os agrupamentos Q_k e Q_l , também baseado na distância entre os centróides. $S_c(Q_k)$ é dado por:

$$S_c(Q_k) = \left\{ \frac{1}{N_k} \sum_{j=1}^{N_k} |x_j - c_k|^q \right\}^{1/q} \quad (3.2)$$

onde: $x_j \in Q_k$, N_k é o número de amostras no agrupamento Q_k e $c_k = 1/N_k \sum_{x_i \in Q_k} x_i$. d_{ce} é dado por:

$$d_{ce}(Q_k, Q_l) = \left\{ \sum_{k=1}^d |c_{ki} - c_{kj}|^p \right\}^{1/p} \quad (3.3)$$

onde: d corresponde à dimensão do vetor x_k .

Para $p = 2$ tem-se d_{ce} como a distância Euclideana entre os centróides dos dois agrupamentos. S_c representa a raiz q -ésima do q -ésimo momento dos pontos no agrupamento k em torno da média. Se $q = 1$, S_c torna-se a média Euclideana dos vetores no agrupamento k em relação ao centróide deste grupo. Se $q = 2$, S_c torna-se o desvio padrão da distância das amostras com relação ao centróide do grupo. Neste trabalho, fixou-se $p = 2$ e variou-se $q = 1, 2$.

Vesanto e Alhoniemi (2000) usaram o índice Davies-Bouldin com $p = 2$ e $q = 2$ para avaliação da partição dos dados feita através do Mapa Auto-Organizável. Neste caso, o SOM foi usado como um redutor do volume de dados a ser particionado. Após esta redução, dois métodos de partição tradicionais, o k -médias e o método hierárquico aglomerativo, foram aplicados, separadamente, para encontrar os agrupamentos. O índice Davies-Bouldin foi usado como critério de junção ou separação de agrupamentos nos algoritmos de partição.

3.4.2 Índice CDbw

O índice CDbw - *Compose Density between and within clusters* (Halkidi e Vazirgiannis, 2002; Wu e Chow, 2004) também baseia-se na medição das distâncias intra e inter-agrupamentos, porém, enfatiza as características geométricas de cada agrupamento, tratando eficientemente agrupamentos com formatos arbitrários. A característica geométrica do agrupamento é representada através do uso de vetores representativos de cada agrupamento. Ao invés de usar o centróide como referência, usa-se um conjunto de vetores. Isto permite que o índice avalie corretamente estruturas não hiperesféricas, o que não ocorre com o índice Davies-Bouldin.

Para um conjunto de dados particionados em c agrupamentos, define-se um conjunto de pontos representativos $V_i' = \{v_{i1}, v_{i2}, \dots, v_{ir_i}\}$ para o agrupamento i , onde: r_i representa o número de pontos de representação para o agrupamento i . Para cada componente ρ do agrupamento i tem-se que o desvio padrão $stdev(i)$ é dado por:

$$stdev^\rho(i) = \sqrt{\sum_{k=1}^{n_i} (x_k^\rho - \bar{x}_i^\rho)^2 / (n_i - 1)} \quad (3.4)$$

onde: n_i representa o número de amostras no agrupamento i , $x_k \in Q_i$, e \bar{x}_i é a média da amostra do i -ésimo agrupamento. A média do desvio padrão é dada por:

$$stdev = \frac{1}{c} \sqrt{\sum_{i=1}^c \|stdev(i)\|^2} \quad (3.5)$$

A densidade intra-agrupamento é definida como:

$$Intra_dens(c) = \frac{1}{c} \sum_{i=1}^c \frac{1}{r_i} \sum_{j=1, j \neq i}^{r_i} density(v_{ij}), c > 1 \quad (3.6)$$

O termo $density(v_{ij})$ é definido como $density(v_{ij}) = \sum_{l=1}^{n_i} f(x_l, v_{ij})$, onde $x_l \in Q_i$, v_{ij} é a j -ésima representação do i -ésimo agrupamento, e $f(x_l, v_{ij})$ é dado por 1, se $\|x_l - v_{ij}\| \leq stdev$, ou 0 caso contrário.

A densidade inter-agrupamento é dada por:

$$Inter_dens(c) = \sum_{i=1}^c \sum_{j=1, j \neq i}^c \frac{\|close_rep(i) - close_rep(j)\|}{\|stdev(i)\| + \|stdev(j)\|} density(v_{ij}), c > 1, c \neq n \quad (3.7)$$

onde: $close_rep(i)$ e $close_rep(j)$ representam o par de pontos de representação mais próximos entre o agrupamento i e o j , v_{ij} é o ponto médio entre este par de pontos. $density(v_{ij})$ é dado por $density(v_{ij}) = \sum_{k=1}^{n_i+n_j} f(x_k, v_{ij})$, onde $x_k \in Q_i$ ou $x_k \in Q_j$, e $f(x_k, v_{ij})$ é dado por 1 se $\|x_k - v_{ij}\| \leq (\|stdev(i)\| + \|stdev(j)\|)$, ou 0 caso contrário.

A separação entre os agrupamentos é dado por:

$$Sep(c) = \sum_{i=1}^c \sum_{j=1, j \neq i}^c \frac{\|close_rep(i) - close_rep(j)\|}{1 + Inter_dens(c)}, c > 1 \quad (3.8)$$

O índice CDbw é definido por:

$$CDbw(c) = Intra_dens(c) * Sep(c) \quad (3.9)$$

Uma boa partição dos dados é indicada para valores altos do índice. A complexidade $O(n)$ é favorável para dados geoespaciais (Halkidi e Vazirgiannis, 2002).

Uma questão importante a ser considerada neste algoritmo é a definição dos vetores de referência para cada agrupamento. Segundo Halkidi e Vazirgiannis (2002) este processo é iterativo. Primeiro escolhe-se o ponto mais distante da média do agrupamento, posteriormente o ponto mais distante do ponto anterior é escolhido e assim sucessivamente.

3.4.3 Usando os vetores de código como vetores de referência no CDbw

Os vetores de referência para o cálculo do CDbw podem ser encontrados de forma iterativa a partir do conjunto de dados particionado (Halkidi e Vazirgiannis, 2002). Todavia, estes autores não definem o critério de parada para o algoritmo de criação dos vetores de referência. Ou seja, o número de vetores de referência, para cada agrupamento, tem de ser definido empiricamente para servir como critério de parada, caso contrário todos os vetores poderiam ser escolhidos como vetores de referência.

Para o caso de partição dos dados através do SOM tem-se os vetores de código como uma aproximação da distribuição dos dados de entrada sendo, portanto, vetores representativos dos dados amostrais. Logo, pode-se usar os vetores de código do SOM particionado como vetores de referência dos seus respectivos agrupamentos. Isto simplifica o processo de cálculo do CDbw para o caso de partição dos dados com o SOM.

A adequação desta abordagem dependerá da relação entre o número n de padrões e o número m de neurônios. Para m/n muito pequeno pode-se ter uma deficiência em número de neurônios para representação de cada agrupamento. Para m/n muito grande tem-se o inverso.

3.5 Validando o Particionamento do SOM

O algoritmo de detecção automática de agrupamentos baseado na partição do SOM (Costa e Andrade Netto, 2003) separa os padrões, mas não garante que todos os vetores de entrada serão rotulados. Por exemplo, dados atípicos podem não ser rotulados devido a alguma restrição do Item “b” do algoritmo da Seção 3.3.

Este problema pode ser solucionado usando-se o critério do vizinho mais próximo para a rotulação de todos os neurônios especializados do Mapa. Este procedimento evitará que o cálculo dos índices de validação sejam comprometidos.

O processo de avaliação dos agrupamentos usado neste trabalho (Capítulo 6) será realizado da seguinte forma:

- a) Define-se um conjunto de redes que serão testadas e, para cada rede, deve-se:
 - efetuar o treinamento da rede;
 - aplicar o algoritmo Costa-Netto;
 - rotular todos os neurônios através do método do vizinho mais próximo;
 - calcular os índices Davies-Bouldin e CDbw;
- b) Escolhe-se as redes com os melhores valores dos índices.

3.6 Sumário

O processo de detecção automática de agrupamentos com Mapas Auto-Organizáveis pode ser realizado em três fases. Na primeira fase o Mapa neural é treinado, na segunda este mesmo mapa é segmentado, na terceira e última fase os dados são particionados de acordo com o neurônio ao qual está associado.

Foram expostas três formas de detecção automática de agrupamentos. Através da segmentação da imagem gerada pela U-matriz, através do uso de técnicas estatísticas para particionar os pesos do SOM ou através do particionamento de grafos. Este último método foi escolhido para trabalho considerando que é o único que se baseia unicamente nas informações contidas nos próprios neurônios, após o processo de aprendizagem, além de ser independente das dimensões da rede neural.

Para validar os partiucionamentos dos dados foram analisados duas métricas: o índice Davies-Bouldin e o CDbw. Ambos avaliam as densidades intra e inter agrupamentos, todavia o primeiro é baseado no centróide dos agrupamentos, enquanto que o segundo baseia-se em vetores de referência. O objetivo desses vetores é inserir a estrutura geométrica do agrupamento no computo do índice. Neste trabalho os vetores de referência serão determinados a partir dos vetores de código do Mapa treinado.

Portanto, o processo de avaliação dos agrupamentos passará primeiramente pela definição das configurações de redes a serem avaliadas; treinamento destas redes; aplicação do algoritmo Costa-Netto; rotulação de todos os neurônios pelo método do vizinho mais próximo; e cálculo dos índices de validação. As redes serão selecionadas de acordo com os valores dos índices.

CAPÍTULO 4

ANÁLISE EXPLORATÓRIA DE DADOS GEOESPACIAIS MULTIVARIADOS ATRAVÉS DE MAPAS AUTO-ORGANIZÁVEIS

4.1 Análise Espacial de Dados em Área

O estudo de caso deste trabalho, assim como boa parte das aplicações do SOM na Análise Espacial, trabalha com Análise Espacial de Dados em Área, que considera a análise de dados associados 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 (Bailey e Gatrell, 1995).

Na fase exploratória da Análise Espacial de Dados em Área, a detecção e possível exploração de padrões espaciais, ou tendências nos valores dos atributos, são as tarefas principais. Dada uma região de estudo R , particionada em subáreas (A_1, \dots, A_n) com $A_1 \cup \dots \cup A_n = R$, tem-se o vetor de características $x(A_k) = (\xi_1, \dots, \xi_d)$, $A_k \in \{A_1, \dots, A_n\}$. Neste trabalho, este vetor de características será denotado por x_k (Figura 4.1).

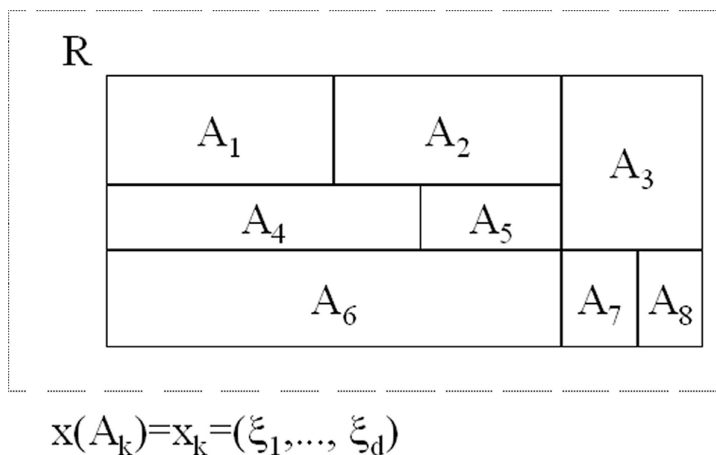


FIGURA 4.1 – Elementos da Análise Espacial de Dados em Área.

Existem várias formas para a visualização deste tipo de dado geoespacial (Bailey e Gatrell, 1995). Neste trabalho, usou-se mapas coropléticos para identificação e realce de possíveis padrões.

Para a análise exploratória de área uma questão deve ser respondida: como medir a proximidade das observações relativas às áreas A_k ? Pode-se usar o centróide das áreas e usar alguma medida de distância para avaliar a proximidade ou usar algum outro mecanismo, como uma matriz de proximidade W , definida sobre algum critério específico. A primeira opção é válida, porém limitada, uma vez que não contempla a estrutura dos objetos geográficos (Bailey e Gatrell, 1995). A segunda opção, w_{ij} , é mais genérica e será usada neste trabalho.

Seja W uma matriz de proximidade espacial, onde cada elemento, w_{ij} , representa uma medida de proximidade entre as áreas A_i e A_j . Existem vários critérios para o cálculo dos valores w_{ij} , desde baseado no centróide até aqueles baseados no compartilhamento de fronteiras entre A_i e A_j , ou uma combinação destes (Bailey e Gatrell, 1995). Para este trabalho W foi definido como 1, caso A_j compartilhe fronteira com A_i , ou 0 caso contrário.

Uma vez definido o critério de proximidade espacial pode-se determinar a dependência espacial do conjunto de dados. A dependência espacial avalia a variação dos atributos quanto à disposição espacial das áreas. Há várias técnicas para medir a dependência espacial (Bailey e Gatrell, 1995); aqui será destacado o índice de correlação espacial global de Moran. Para uma determinada matriz de proximidade W , o índice de Moran (I) calcula a correlação espacial para o atributo ξ_i da seguinte forma:

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (\xi_i - \bar{\xi}) (\xi_j - \bar{\xi})}{\left(\sum_{i=1}^n (\xi_i - \bar{\xi})^2 \right) \left(\sum_{i \neq j} w_{ij} \right)} \quad (4.1)$$

onde $\bar{\xi}$ representa a média do atributo ξ_i .

4.2 Análise Espacial com o SOM

Os Mapas Auto-Organizáveis têm-se mostrado bastante úteis na Análise Espacial, haja vista o crescente número de publicações presentes na literatura (Openshaw e Turton, 1996; Ji, 2000; Cereghino et al., 2001; Villmann et al., 2003). 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 gerado uma variedade 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. Este tipo de rede é extremamente útil para a análise de dados ge-

ográficos cujas propriedades impedem que sejam usados métodos estatísticos (Openshaw et al., 1994; Openshaw e Turton, 1996). Segundo Openshaw e Turton (1996), problemas como análise de dados multivariados, dependência de incerteza sobre os dados, distribuições não normais das variáveis etc., podem ser convenientemente tratados com as RNAs, em especial a rede SOM. Outros trabalhos exploram a propriedade de mapeamento e visualização da rede SOM para as tarefas de Análise Espacial (Winter e Hewitson, 1994; Foody, 1999; Cereghino et al., 2001; Gahegan et al., 2002).

Nenhum dos trabalhos disponíveis na literatura faz referência às metodologias de análise exploratória de dados geoespaciais que contemplem os diversos pontos de vista da análise, tais como: descoberta de dados atípicos, análise da distribuição espacial do fenômeno, análise de correlação e significância de componentes, análise de agrupamentos e dependência espacial. Também há poucas referências sobre o uso do SOM para particionamento dos dados geoespaciais (Franzini et al., 2001), ou para análise de dependência espacial (Kaski e Kohonen, 1996). Pôde-se observar que os trabalhos de aplicação dos Mapas Auto-Organizáveis na Análise Espacial apresentam algumas características comuns, como o uso do algoritmo padrão de treinamento (Winter e Hewitson, 1994; Kaski e Kohonen, 1996; Foody, 1999; Takatsuka, 2001; Cereghino et al., 2001; Franzini et al., 2001). Em geral são usados os modelos com topologias bidimensionais, pois permitem a visualização natural dos agrupamentos através da U-matriz (Kaski e Kohonen, 1996; Cereghino et al., 2001), e dos Planos de Componentes (Winter e Hewitson, 1994; Kaski e Kohonen, 1996; Franzini et al., 2001). 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 do Mapa gerado. Em geral, as aplicações estão concentradas na análise da dinâmica de cidades (Winter e Hewitson, 1994; Kropp, 1998; Takatsuka, 2001), ou na análise de dinâmicas ambientais (Foody, 1999; Cereghino et al., 2001). Franzini et al. (2001) foi o único trabalho encontrado sobre o estudo de problemas urbanos através dos Mapas Auto-Organizáveis.

Embora os trabalhos estejam relacionados a dados geográficos armazenados em Sistemas de Informação Geográfica, nenhum modelo aplicado insere em seus cálculos algo que identifique a posição espacial, por exemplo, as coordenadas planas entre os objetos geográficos estudados, sejam sítios de coleta de dados ou distritos censitários. Em (Babu, 1997) é proposta uma rede SOM modificada que considera a posição espacial entre os objetos espaciais, porém seu objetivo não é descobrir relações fenomenológicas entre os objetos, mas sim, facilitar a tarefa de indexação e visualização dos objetos geográficos. Também foi observado que nenhum método de detecção automática de agrupamentos foi aplicado no processo de análise do Mapa gerado pelas redes. Observou-se que o processo

de determinação do tamanho dos Mapas é totalmente empírico, baseado na experiência do usuário e no método de tentativa e erro.

Foi possível concluir, a partir desta revisão bibliográfica, que o Mapa Auto-Organizável tem despertado um crescente interesse por parte dos profissionais da geociência, haja vista o crescente número de ferramentas recentemente disponíveis (Takatsuka, 2001; Gahegan et al., 2002). Pôde-se, finalmente, observar que existe 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; d) para séries temporais, pode-se usar o SOM para análise de trajetória; e) usar redes com dimensões maiores que 2 para agrupamento de dados; f) usar o SOM para a análise de deslocamento entre grupos após alterações do vetor de característica de determinado objeto.

4.3 Inclusão da Variável Espacial no Algoritmo SOM

Na seção anterior apresentou-se uma breve revisão da aplicação do SOM na Análise Espacial, pela qual pôde-se observar que, em nenhum momento, as variáveis posicionais (x, y) são incluídas explicitamente no algoritmo. Babu (1997) propõe que a questão espacial de objetos geográficos seja incluída nos Mapas Auto-Organizáveis através da criação de uma medida de dissimilaridade D que contemple o objeto geográfico de qualquer dimensão, de maneira simples e representativa. Ou seja, o autor substitui a função de distância do SOM padrão, em geral a distância Euclideana, por uma outra. Esta variante é chamada de *SOM for Spatial Data* (SOMSD) e objetiva a visualização espacial e a indexação de objetos geográficos. Este trabalho mostrou que o SOM pode ser usado de forma combinada com as coordenadas espaciais, para fins de visualização e indexação.

A fim de avaliar o efeito da inclusão das variáveis posicionais, (x, y) , neste trabalho propõe-se incluí-las no vetor de características x_k . Através deste procedimento, espera-se verificar se isto afetará significativamente a formação final do Mapa Auto-Organizável, através da análise da U-matriz e dos Planos de Componentes.

4.4 Análise da Distribuição Espacial do Fenômeno

A partir dos Planos de Componentes é possível a identificação, no Mapa neural, do sentido da variação dos componentes. Em geral, esta análise é feita visualmente (Kaski e Kohonen, 1996; Winter e Hewitson, 1994; Franzini et al., 2001). Porém, é possível

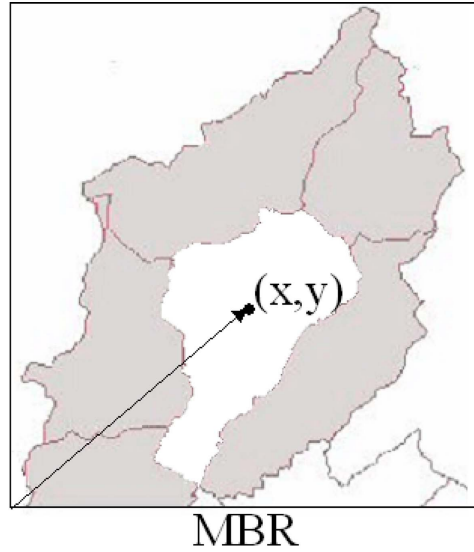


FIGURA 4.2 – Coordenadas do centróide de uma área qualquer relativo ao MBR do conjunto de áreas.

automatizar este processo através de um mecanismo bastante simples. Seja um Mapa neural bidimensional ($N \times M$), pode-se rotulá-lo de maneira que o sentido da distribuição predominante nos Planos de Componentes seja representado. Pode-se ter vários formatos para a distribuição da variação dos componentes no Mapa, porém definiu-se um conjunto fixo de distribuições, vertical, horizontal e diagonais (Figura 4.3).

4.5 A Proposta de um Índice de Valoração da Dependência Espacial

A propriedade de ordenação topológica dos dados amostrais de entrada, numa grade finita de neurônios, garante que padrões vizinhos no espaço de entrada I estejam próximos no espaço de saída U , porém o inverso nem sempre é verdadeiro. Todavia, pode-se afirmar que o Mapa Auto-Organizável representa uma relação de vizinhança no espaço de atributos. As variações nos valores de x_k são captadas pelo SOM através dos valores dos vetores de código e de sua posição na grade. Assim, a propriedade de ordenação topológica pode auxiliar no cômputo da dependência espacial. Tem-se, então, que o SOM determina a relação de vizinhança no espaço de atributos, exprimindo a ordenação da variação nos valores de x_k . Tem-se, também, a relação de vizinhança espacial expressa pela matriz de proximidade W . Portanto, a dependência espacial pode ser valorada definindo-se uma métrica que leve em consideração as relações de vizinhança no espaço de atributos e no espaço físico (Figura 4.4).

Uma forma simples de calcular esta dependência espacial, baseada no SOM e na matriz W , é medir a relação entre o número de padrões que estão simultaneamente próximos no

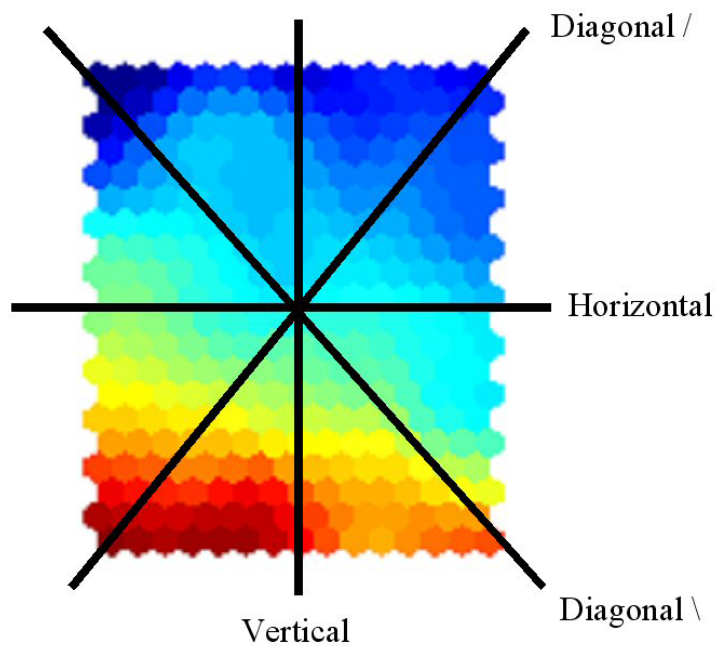


FIGURA 4.3 – Diferentes possíveis sentidos da distribuição dos componentes no Mapa neural.



FIGURA 4.4 – Correspondência entre a relação de vizinhança no espaço de atributos e no espaço físico.

espaço de atributos e no espaço físico. Todavia, como nem toda vizinhança no espaço de saída do SOM representa similaridade entre os padrões, a segmentação do SOM em zonas de similaridade é necessária. Para isto, usou-se o algoritmo Costa-Netto para partição do SOM, o qual dividiu o conjunto de dados em c agrupamentos e associou cada área A_i a seu respectivo agrupamento. Ou seja, as regiões A_i serão rotuladas de acordo com o grau de proximidade no espaço de atributos. Pode-se, então, avaliar a existência de regimes espaciais distintos medindo o grau de dispersão espacial intra-agrupamento.

Seja p_i o número de áreas A pertencentes ao agrupamento i . Seja q_i o número de grupos de áreas A distintas do agrupamento i . Tem-se que o índice de relação de vizinhança espacial ($IRVE_i$) para o agrupamento i será dado por:

$$IRVE_i = \begin{cases} 0, & \text{se } p_i = q_i \\ -\frac{1}{p_i}(q_i - 1) + 1, & \text{caso contrário.} \end{cases} \quad (4.2)$$

O índice global ($IRVE$) corresponderá à média ponderada dos índices $IRVE_i, i = 1, \dots, c$

$$IRVE = \frac{1}{n} \sum_{i=1}^c IRVE_i p_i \quad (4.3)$$

Por exemplo, dado um conjunto de áreas rotuladas, representadas pela Figura 4.5, tem-se que, aplicando a Equação (4.2) $IRVE_1 = -(1/11) * (2 - 1) + 1 = 0,90$ e $IRVE_2 = -(1/10) * (3 - 1) + 1 = 0,80$, aplicando a Equação (4.3) tem-se que $IRVE = (1/21) * (0,90 * 11 + 0,80 * 10) = 0,852$.

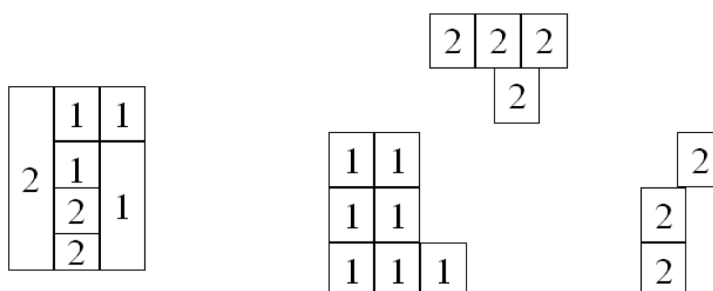


FIGURA 4.5 – Conjunto de áreas rotuladas, 2 agrupamentos, para exemplificar o cálculo do índice IRVE.

O índice $IRVE$ varia entre 0 e 1, sendo que valores próximos de zero significam alta

dispersão espacial dos agrupamentos e valores próximos de 1 significam alto nível de agregação dos agrupamentos espaciais. A rigor, este cálculo poderia ser feito para qualquer partição das áreas A , porém, o índice IRVE identifica que o nível de agregação dos agrupamentos foi alcançado a partir do SOM.

4.6 Sumário

A Análise Espacial de Dados em Área consiste na busca por informações em mapas de áreas cujos atributos associados não variam em função do espaço. Ou seja, cada área terá um único conjunto de atributos associado. Na análise exploratória desses dados o objetivo principal é verificar a existência de padrões espaciais. Antes de iniciar o processo de análise exploratória é necessário definir um critério de proximidades entre as áreas. Neste trabalho foi usado a matriz de proximidade baseada no compartilhamento de fronteiras entre as áreas.

Um conceito importante na Análise Espacial de Dados em Área é o de dependência espacial. Dependência espacial é o grau com que a variação nos atributos acompanham a variação na localização espacial. A Análise Espacial de Dados em Área compreende a análise da distribuição espacial do fenômeno, a análise de agregados espaciais e medida de dependência espacial.

Existem vários trabalhos que tratam do uso dos Mapas Auto-Organizáveis na Análise Espacial, todavia, nenhuma aplicação trata explicitamente da inserção da posição espacial, assim como também não há trabalhos que defina uma metodologia que cubra todo o escopo da Análise Espacial de Dados em Área.

Neste capítulo foram apresentadas três propostas de uso do SOM na Análise Espacial. A primeira sugere a inclusão das coordenadas planas no vetor de características x_k para verificar se isto afetaria a qualidade final do Mapa neural. A segunda propõe a automatização do processo de análise da distribuição espacial do fenômeno através dos Planos de Componentes. A terceira proposta trata da formulação de uma métrica para a dependência espacial a partir do mapa de áreas segmentado pelo algoritmo Costa-Netto.

CAPÍTULO 5

PROJETO E PROGRAMAÇÃO DO MAPA AUTO-ORGANIZÁVEL

5.1 Introdução

Para o caso de análise de dados geoespaciais multivariados é necessário que os resultados gerados a partir do Mapa Auto-Organizável possam ser visualizados graficamente por meio de mapas. Para que isto seja possível, sem a necessidade de importação/exportação de arquivos, uma solução possível, e aqui utilizada, é a conexão do algoritmo SOM à biblioteca de acesso ao banco de dados geográficos **TerraLib**, desenvolvida no INPE/DPI (Câmara et al., 2002). A TerraLib é uma biblioteca de classes voltada para o desenvolvimento de sistemas de informação geográfica customizados. A **TerraLib** foi desenvolvida na linguagem de programação C++, através da aplicação de modernas técnicas de programação, como padrões de projeto (Gamma et al., 1995), programação genérica (Stroustrup, 2000), STL (Musser e Saini, 1996) e programação multi-paradigma (Coplien, 1998).

Embora o algoritmo padrão de treinamento da rede SOM seja conceitualmente simples, sua implementação requer uma série de cuidados. Kohonen (2001), afirma que a maioria das implementações não se preocupa com os detalhes do processo de construção do algoritmo. Ciente deste problema, a equipe de pesquisas em Mapas Auto-Organizáveis da universidade da Finlândia desenvolveu dois pacotes de software que implementam a rede SOM. O SOM PAK, desenvolvido em C (Kohonen et al., 1995) e o SOM ToolBox, desenvolvido em MatLab (Vesanto et al., 1999). Ambos possuem código fonte aberto, são gratuitos e possuem características importantes para este projeto como confiabilidade, disponibilidade do código fonte e funcionalidade.

Porém, após a análise dos pacotes SOM PAK e SOM ToolBox, verificou-se que ambos demandariam um esforço muito grande de conexão com a biblioteca **TerraLib**, uma vez que estes pacotes foram desenvolvidos em linguagens distintas da C++ e não usam, extensivamente, conceitos de programação moderna, o que acarretaria sérias dificuldades de manutenção. Portanto, apesar das vantagens em termos de confiabilidade e funcionalidade, decidiu-se desenvolver um novo código para o algoritmo SOM. Outros pacotes foram analisados, mas não atendiam simultaneamente os requisitos de disponibilidade do código fonte, confiabilidade e manutenibilidade como o SNNS (Zell et al., 1992) e Nenet (Kohonen, 2001). O pacote SOM ToolBox foi usado neste projeto como mecanismo de comparação e teste do algoritmo SOM desenvolvido.

O desenvolvimento de qualquer simulador neural exige preocupações nas áreas de depuração do código, processamento de alto desempenho, com ou sem paralelização da implementação, e projeto (Lawrence et al., 1996). Este trabalho concentrou-se na elaboração do projeto de implementação baseado no paradigma de Orientação a Objetos (Gamma et al., 1995). Os pacotes SOM PAK e ToolBox auxiliaram na depuração do código projetado e implementado. O projeto consistiu no desenho e implementação de uma biblioteca de classes, SOMLib, que implementa algoritmos e encapsulam dados relativos ao uso da rede SOM para a análise exploratória de dados multivariados, geoespaciais ou não.

5.2 Projeto e Programação

Segundo Kohonen (2001), qualquer pacote SOM deve apresentar um conjunto mínimo de características, tais como: permitir que a grade da rede possa ter qualquer dimensão, definição automática das dimensões em função dos auto-valores da matriz de correlação dos padrões de entrada, disposição hexagonal e retangular, aprendizagem em lote e sequencial, função de vizinhança gaussiana e bolha, iniciação linear, tratamento de dados ausentes, algoritmos de visualização e cálculo dos erros de quantização e topológico.

Como observado no Capítulo 2, a rede neural SOM pode variar de diferentes formas. Pode-se ter redes de dimensões variadas, com formatos diferentes da grade de neurônios, funções de vizinhança distintas etc. Representando este conjunto de variações num diagrama de classes (Figura 5.1), pode-se observar a proliferação de classes. Observe-se que todas as características relativas à grade foram encapsuladas nas classes de topologia (2D, 3D ...).

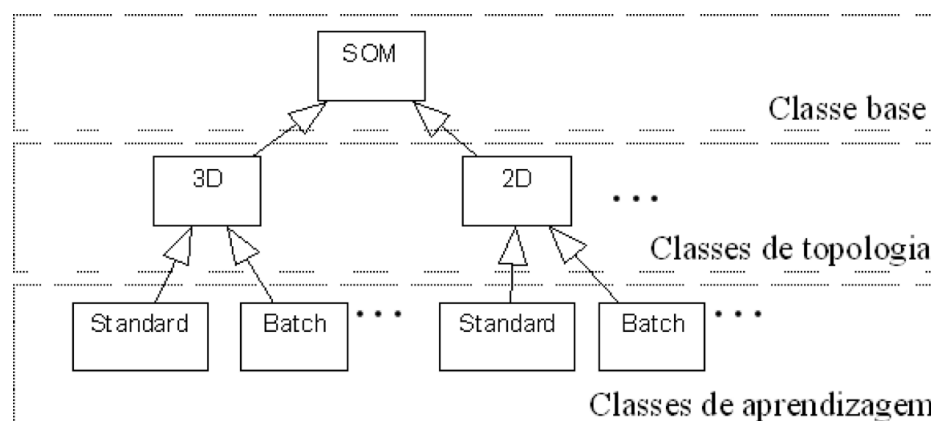


FIGURA 5.1 – Diagrama de Classes para representação das famílias de Mapas Auto-Organizáveis.

As classes foram agrupadas em três categorias: classe base (SOM), classes de topologia (2D, 3D ...) e classes de aprendizagem (Standard, Batch ...). Implementar a biblioteca com base nesta estrutura de classes não configura uma boa idéia, pois, além da duplicação de classes, observa-se um forte acoplamento entre as classes de topologia e de aprendizagem. Para resolver esta questão, dividiu-se o problema em dois: projeto e implementação das classes de aprendizagem e de topologia. Para o problema relativo às classes de aprendizagem tem-se que, a depender do contexto ou necessidade do usuário, deve ser possível variar entre os vários algoritmos de aprendizagem implementados. Pode-se, então, usar o padrão *Strategy* para resolver este problema. O padrão *Strategy* define uma família de algoritmos, encapsula cada um e os faz interoperáveis (Gamma et al., 1995). O diagrama da Figura 5.2 mostra que uma classe abstrata foi criada (*LearningAlgorithm*), pela qual as classes de aprendizagem serão derivadas. O trecho de código em seguida ilustra a implementação da classe base SOM, considerando a estrutura do diagrama de classes (Figura 5.2).

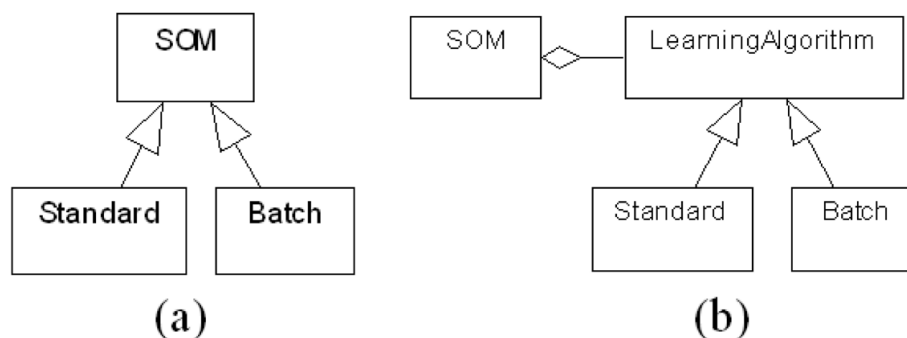


FIGURA 5.2 – Diagrama de Classe: a) Representação da classe base e das classes de aprendizagem; b) Nova estrutura do diagrama -a- baseada no padrão *Strategy*.

Como demonstrado através da Figura 5.3, as questões de topologia e aprendizagem estão “misturadas” de forma que a adição de mais uma classe de topologia implica na reconstrução das classes de aprendizagem relacionadas com a mesma. Este problema foi solucionado com o uso do padrão de projeto *Bridge*. Este padrão desacopla uma abstração de sua implementação, de forma que ambas possam variar independentemente (Gamma et al., 1995). Assim, criou-se mais uma classe abstrata, *TopologyImp*. Foi a partir dessa classe que se originaram as classes concretas de topologia. A Figura 5.4 mostra a nova estrutura de relacionamento entre as classes de topologia e as classes de aprendizagem. Com esta nova estrutura uma mesma implementação de uma classe de topologia pode servir a mais

```

class SOM {
public:
    SOM( Params par ) {          // Constructor
        switch (par.type) {
            case "batch" : _learningAlgorithm = new Batch;
            case "Standard" : _learningAlgorithm = new Standard; }
        }

    ~SOM();          // Destructor
protected:
    LearningAlgorithm * _learningAlgorithm;

    LearningAlgorithm * getLearningAlgorithm() { return _learningAlgorithm;
};
public:
    void Learning() {
        getLearningAlgorithm()->Learning( netParams );
    };
};

```

de uma classe de aprendizagem, sem a necessidade de duplicação de código. Em seguida, tem-se mais um trecho de código da classe abstrata *LearningAlgorithm*.

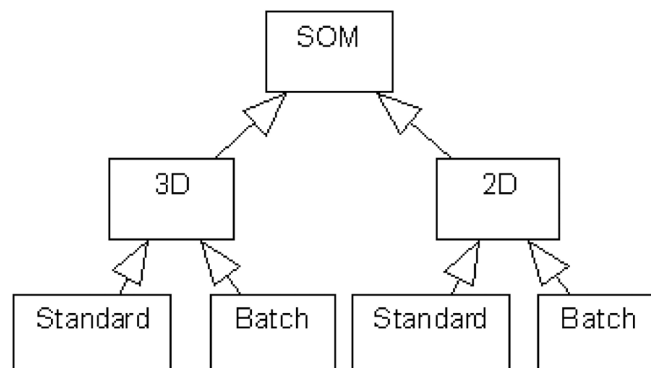


FIGURA 5.3 – Diagrama de Classes. Aqui observa-se o alto acoplamento entre as classes de topologia e de aprendizagem.

Optou-se pela mesma implementação para os dois padrões de projeto usados, mas observe-se que ambas foram motivadas por razões distintas. A Figura 5.5 mostra a configuração final do diagrama de classes após o uso dos padrões. Esta estrutura permitirá uma maior manutenibilidade e possibilidade de reuso de código para a biblioteca SOM-Lib.

Na implementação das classes base SOM e da classe abstrata *LearningAlgorithm* percebe-se que cada uma deve decidir qual objeto criar de acordo com os parâmetros passados no

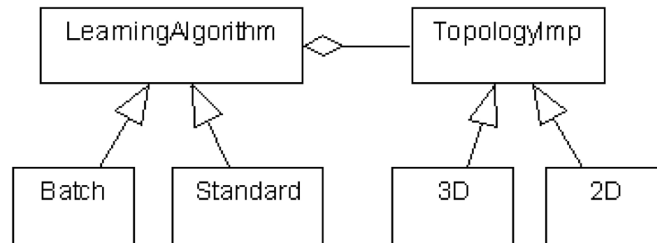


FIGURA 5.4 – Através do padrão *Bridge* separou-se os detalhes de topologia e aprendizagem.

```

class LearningAlgorithm {
public:
    virtual int Learning( Params& par ) = 0;
    TopologyImp * getTopology() { return _topology; };
    ~LearningAlgorithm();

protected:
    LearningAlgorithm(const TopolParams& par) {
        switch( par.type ) {
            case "two" : _topology = new TwoD;
            case "three" : _topology = new ThreeD; }
    };

private:
    TopologyImp * _topology;
};
  
```

construtor de cada classe. Após a passagem de parâmetro, a cláusula *switch* definirá qual objeto construir. Embora seja um método válido, cria a necessidade de se alterar todas as classes que contenham este tipo de cláusula, toda vez que uma nova classe de aprendizagem ou de topologia for implementada. Para este caso, usou-se o padrão de projeto *Abstract Factory*. Este padrão estrutural provê uma interface para a criação de famílias de objetos sem especificar as respectivas classes concretas. Usou-se uma implementação específica deste padrão (Câmara et al., 2001). Nesta implementação, os autores empregaram a programação genérica para definir um *Factory* genérico, cuja função é construir qualquer classe concreta, de um conjunto pré-definido, dispensando o uso de cláusulas do tipo *if..then* e *switch*.

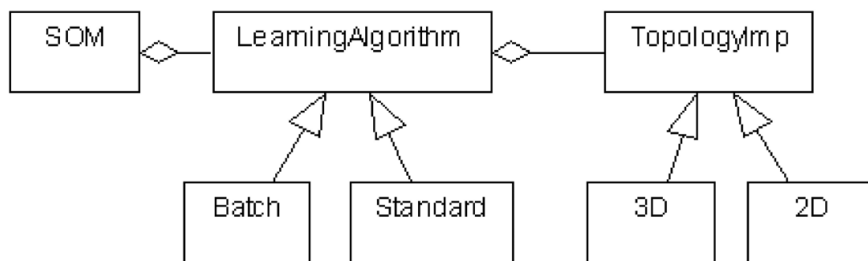


FIGURA 5.5 – Diagrama de Classe final

A Figura 5.6 mostra como ficou a estrutura de classes do diagrama da Figura 5.5, após o uso do padrão *Abstract Factory*. Note-se que, para cada classe de aprendizagem do diagrama da Figura 5.5, foi criada uma classe construtora, *LearningFactory*, *StandardFactory* e *BatchFactory*. A função das classe concretas *StandardFactory* e *BatchFactory* é a de implementar a função *build* da classe *Factory*. O mesmo método foi aplicado no diagrama de classes da Figura 5.4.

A Figura 5.7 mostra a estrutura de classes para a implementação das rotinas de leitura e gravação dos dados, *SOMData*, que alimentaram a rede neural. Optou-se por criar uma classe concreta, *SOMDataCadastro*, para isolar completamente os dados dos detalhes de armazenamento. Assim, a classe *SOMData* transfere todas as responsabilidades de gerenciamento dos dados para a classe *SOMDataCadastro*. Como há várias formas de armazenamento dos dados, usou-se o padrão *Strategy* de forma a facilitar o processo de implementação de novos algoritmos de acesso. Assim, surge a classe abstrata de interface, *ISOMDataRepository*, e as classes concretas derivadas desta e que implementam os métodos de acesso aos dados, *RepositorySOMDataFile*, sistemas de arquivos, e

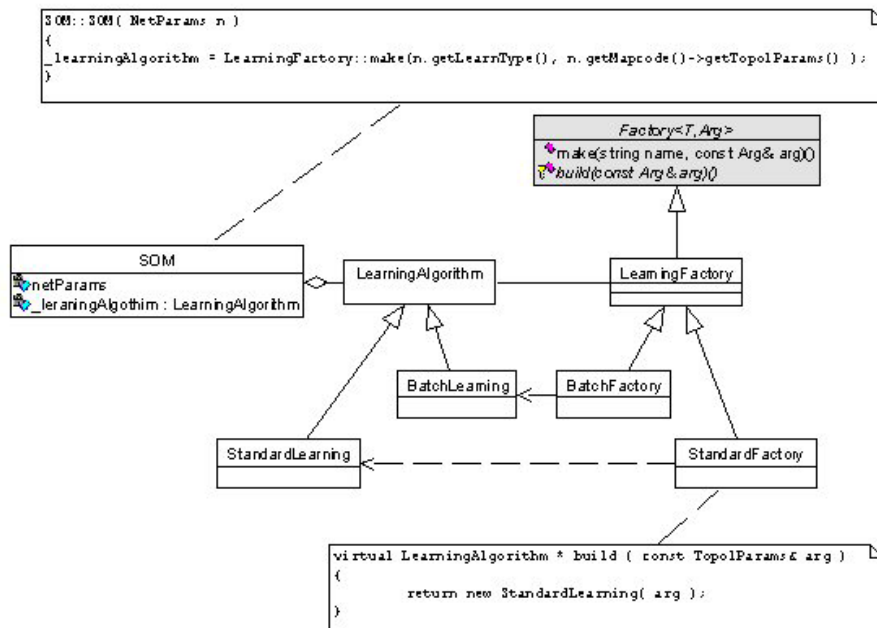


FIGURA 5.6 – Representação do uso do padrão *Abstract Factory* sobre o diagrama de classes da Figura 5.5.

```

class LearningFactory : public Factory<LearningAlgorithm,TopolParams> {
public:
  LearningFactory(const string& name): Factory < LearningAlgorithm, TopolParams> (name) {}
}

class BatchLearning : public LearningAlgorithm {
public:
  BatchLearning(const TopolParams& s):LearningAlgorithm(s) {};

  int Learning ( NetParams& net );
}

class BatchFactory : public LearningFactory {
public:
  BatchFactory( const string& name ) : LearningFactory( name ) {};

  virtual LearningAlgorithm * build ( const TopolParams& arg )
  { return new BatchLearning( arg ); }
}
  
```

RepositorySOMDataTerralib, banco de dados formato **TerraLib**.

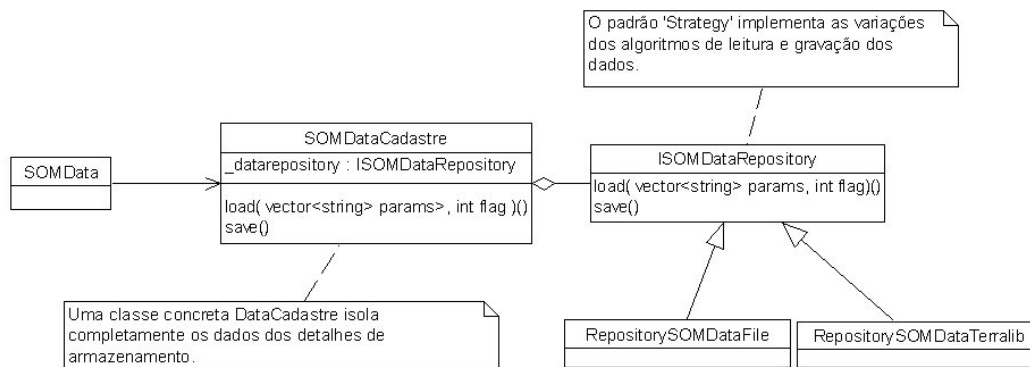


FIGURA 5.7 – Representação da estrutura de classes relativas aos dados e algoritmo de leitura e gravação dos dados de entrada da rede neural.

5.3 Características

No projeto SOMLib implementou-se os algoritmos de aprendizagem em lote e sequencial; as funções de vizinhança gaussiana, bolha e gaussiana cortada; a grade com arranjo hexagonal e retangular; o cálculo do erro de quantização e topológico; a iniciação por interpolação simples e linear; a grade bidimensional.

5.4 Avaliação da Biblioteca

Para a avaliação da SOMLib usou-se dois conjuntos de dados da base *UCI Repository of machine learning databases* (Blake e Merz, 1998): Íris e Wine. As análises de separabilidade das classes e comparação com os resultados gerados pelo SOM Toolbox validaram a biblioteca para estes casos.

5.5 Uso da Biblioteca SOMLib

A seguir, tem-se um exemplo, em C++, do uso da SOMLib. Neste exemplo, os padrões são lidos a partir de um arquivo de dados, 'dados.pat'. Após a leitura, um SOM com valores *default* é criado, bidimensional, com aprendizagem em lote. Em seguida, os parâmetros da rede são ajustados: dimensão 20x20, disposição hexagonal da grade de neurônios, função de vizinhança gaussiana, raio inicial igual a 15, iniciação linear, 2000 épocas de treinamento. As funções de iniciação, *InitMapcode()*, e de aprendizagem, *Learning()*, são então chamadas. Finalmente, os vetores de código da rede treinada serão gravados no arquivo "mapa_treinado.cod".

```

#include "..\Som.h"
#include "..\SOMDataCadastro.h"
#include "..\RepositorySOMDataFile.h"
#include "..\MapcodeCadastro.h"
#include "..\RepositoryMapcodeFile.h"

void
main() {

    vector<string> params;

    RepositorySOMDataFile repD;
    SOMDataCadastro cadD( repD );
    SOMData * data = new SOMData;
    params.push_back( "dados.pat" ); // Lê os dados do arquivo dados.pat

    SOM * mysom = new SOM; // Cria rede bidimensional com algoritmo
                          // de aprendizagem em lote

    mysom->setData( data );

    mysom->getMapcode()->setNumVar( mysom->getData()->getDimension() );
    mysom->getMapcode()->setDimensions( 0, 20 ); // Rede bidimensional 20x20
    mysom->getMapcode()->setDimensions( 1, 20 );
    mysom->getMapcode()->setLattice( "hexa" ); // Grade hexagonal
    mysom->getMapcode()->setNeighborType( "gaussian" ); // Função gaussiana
    mysom->getMapcode()->CreateCodebook( 20*20, mysom->getData()->getDimension() );

    mysom->setInitNeighbor( 15 ); // Vizinhaça inicial
    mysom->setInitType( LINEAR ); // Iniciação linear
    mysom->setNumIterations( 2000 ); // 2000 épocas

    mysom->InitMapcode(); // Inicia os vetores de código
    mysom->Learning(); // Executa algoritmo de aprendizagem

    RepositoryMapcodeFile repM;
    MapcodeCadastro cadM( repM ); // Salva mapa treinado em arquivo
    cadM.save( mysom->netParams.getMapcode(), "mapa_treinado.cod" );
};

```

5.6 Descrição do Sistema *CASA*

A fim de tornar possível a observação visual dos resultados obtidos pelo SOM quanto ao processamento de dados geográficos, foi desenvolvido o sistema *CASA* (*Connectionist Approach for Spatial Analysis of Areal Data*). O sistema *CASA* foi construído sobre as bibliotecas SOMLib e **TerraLib**. O sistema é um simulador neural que possibilita a avaliação de Mapas Auto-Organizáveis bidimensionais e implementa um conjunto de ferramentas de apoio à análise exploratória de dados geoespaciais armazenados em bancos de dados geográficos acessíveis via biblioteca **TerraLib**.

Na Figura 5.8, tem-se a tela principal do sistema. Por meio desta é possível fazer toda a parametrização do simulador. São elementos configuráveis a partir desta tela: os parâmetros de estrutura da rede e de aprendizagem, a análise de agrupamentos, a matriz de distância unificada e a conexão com banco de dados geográfico.

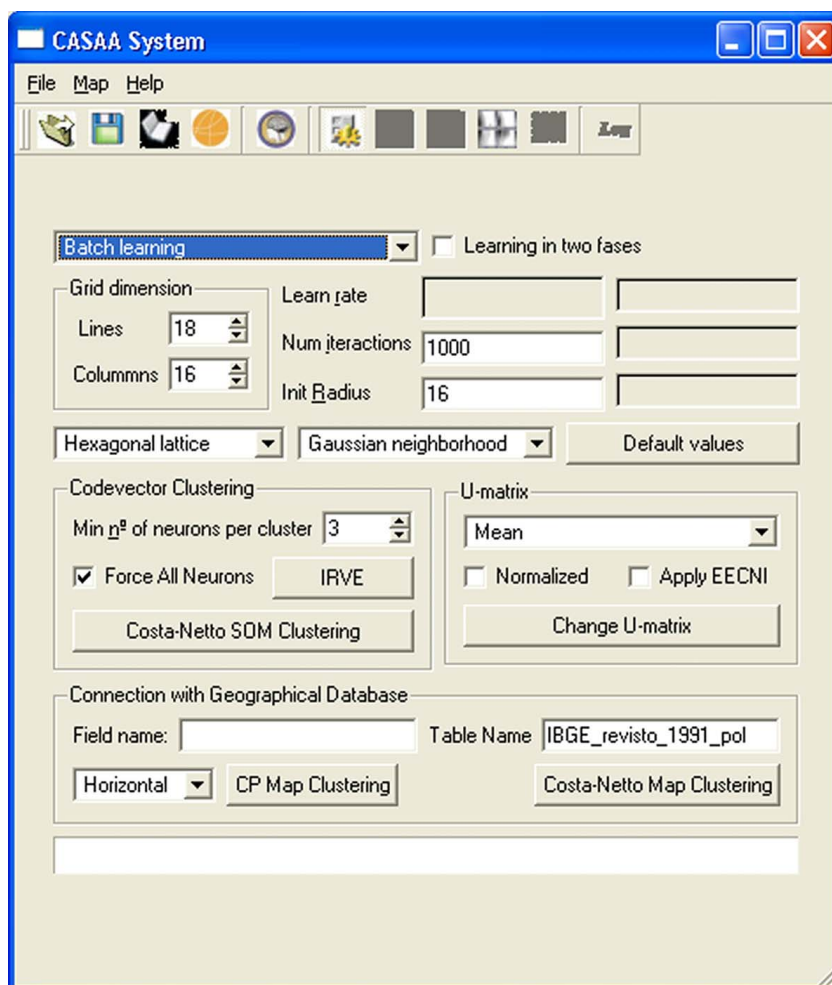


FIGURA 5.8 – Tela inicial do sistema *CASA*.

Para a definição da estrutura da rede o sistema permite a configuração das dimensões (*Grid dimension*) e formato da grade (*lattice*). São parâmetros de aprendizagem configuráveis a taxa de aprendizagem (*learning rate*), o algoritmo de aprendizagem, a função de vizinhança, o número de épocas de aprendizagem (*Num iterations*) e o número de fases de aprendizagem (uma ou duas). Ainda é possível optar por valores de parâmetros *default* (*Default values*) para as dimensões, a organização da grade, a função de vizinhança e o algoritmo de aprendizagem.

A análise de agrupamentos (*Codevector clustering*) está baseada no algoritmo Costa-Netto. Para esta análise, pode-se: optar pelo número de neurônios mínimos por agrupamento (*Min n° of neurons per cluster*), forçar que todos os neurônios especializados sejam rotulados segundo o critério do vizinho mais próximo e, através do botão IRVE, calcular este índice de avaliação da dependência espacial.

As modificações na matriz de distância unificada podem ser feitas através das opções do grupo '*U-matrix*'. Pode-se calcular a U-matriz pela média, mediana, valor máximo e valor mínimo; pode-se, ainda, normalizar os valores e aplicar o algoritmo de Eliminação do Efeito da Cadeia dos Neurônios Inativos - *Apply EECNI* (Costa, 1999).

A conexão com o banco de dados geográfico é feita na leitura e gravação dos dados. Para a leitura dos dados tem-se a tela representada pela Figura 5.9. Através desta tela é feita a conexão com o banco e a leitura das variáveis contidas numa tabela específica e sobre uma determinada restrição da cláusula *WHERE*. Também é neste momento que é lida a matriz de proximidade entre os objetos. Após a leitura dos dados e processamento (aprendizagem) da rede neural, os resultados podem ser gravados na base, através das opções do grupo *Connection with Geographical Database*. Esses dados de gravação estão relacionados com a análise de agrupamentos baseada nos Planos de Componentes (*CP Map Clustering*) ou no algoritmo Costa-Netto (*Costa-Netto Map Clustering*). A visualização destes resultados pode ser efetuada pelo sistema TerraView.

Após a fase de treinamento, o sistema gera uma tela (Figura 5.10) contendo informações sobre as opções do treinamento e resultados. São informações contidas na tela de informações (*Log Info*): arquivo de dados (*Data File*), tipo de aprendizagem (*Learning type*), número de épocas de treinamento (*Training epochs*), raio inicial (*Initial radius*), estrutura da grade de neurônios (*Lattice*), função de vizinhança (*Neighbourhood*), dimensões (*Dimensions*), erros de quantização (*Quantization error*) e topológico (*Topological error*), arquivo de dados do mapa neural (*Mapcode File*), número de agrupamentos encontrados pelo algoritmo Costa-Netto (*Number Cluster*) e dos índices de validação do particionamento dos dados Davies-Bouldin e CDbw.

Terralib Database Params

Host: localhost

Database File: D:/BancosDeDados/sjc.mdb [Load File...]

User: []

Password: []

Table: IBGE_revisto_1991_pol

WHERE statement: CATEGORIA < 0

Data Label Variable: object_id

Variable: [] [Add] [Delete Selected]

- ARENDR
- DESEDUCR
- ESTEDUCR
- LONGR
- QAMBR
- QDOMR
- MANALFR

[Cancel] [OK]

FIGURA 5.9 – Formulário de acesso ao banco de dados geográfico.

Log Info

| | Value |
|--------------------------------|-----------|
| Data File | [] |
| Learning type | Batch |
| Training epochs | 1000 |
| Initial radius | 16 |
| Lattice | Hexagonal |
| Neighbourhood | Gaussian |
| Dimension | 18x16 |
| Quantization error | 0.195594 |
| Topological error | 0.0497076 |
| Mapcode File | [] |
| Number Cluster | 20 |
| Davies-Bouldin (Data)(p=2,q=1) | 3.91791 |
| Davies-Bouldin(Data)(p=2,q=2) | 1.91794 |
| CDbw | 110.142 |

[OK]

FIGURA 5.10 – Formulário com informações sobre o processo de aprendizagem da rede, número de agrupamentos encontrados pelo algoritmo Costa-Netto e índices de validação deste particionamento.

O resultado do processo de segmentação do Mapa neural, através do algoritmo Costa-Netto, é ilustrado através da coloração do Mapa neural (Figura 5.11). Cada cor representa um agrupamento. Ao clicar num neurônio (círculo) uma nova tela aparece, contendo informações sobre quais dados de entrada estão relacionados com este neurônio (*Label*), qual sua posição (*Neuron number*) e qual o seu agrupamento (*Cluster ID*). O sistema ainda gera os Planos de Componentes (Figura 5.12) e a U-matriz (Figura 5.13).

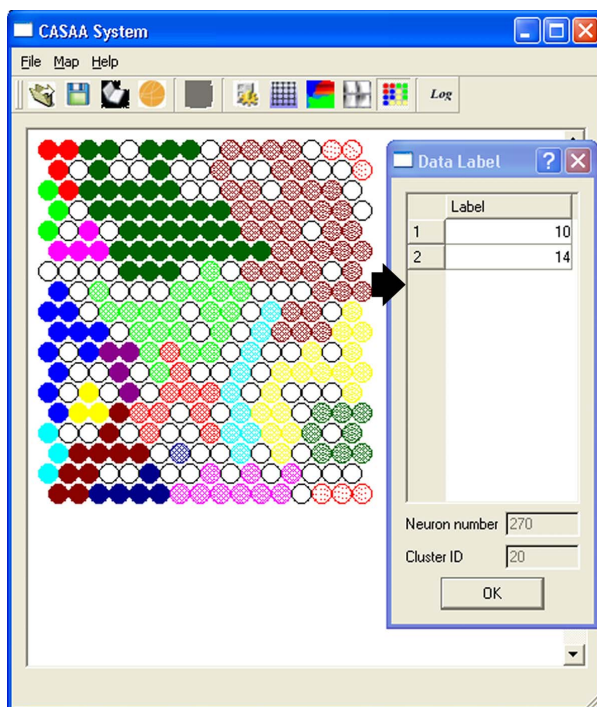


FIGURA 5.11 – Resultado do processo de segmentação do Mapa neural através do algoritmo Costa-Netto. O formulário *Data Label* informa, para cada neurônio, quais padrões de entrada estão relacionados com o mesmo, sua posição (*Neuron number*) e a qual agrupamento pertence (*cluster ID*).

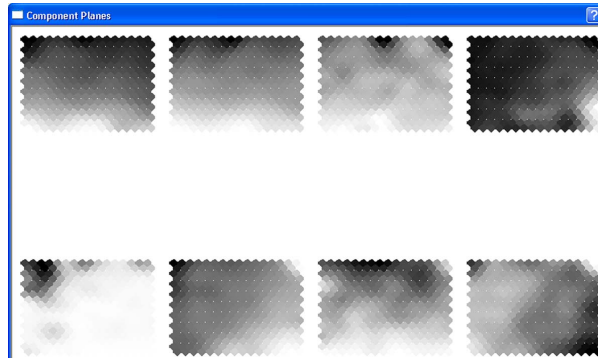


FIGURA 5.12 – Planos de Componentes gerados pelo sistema.

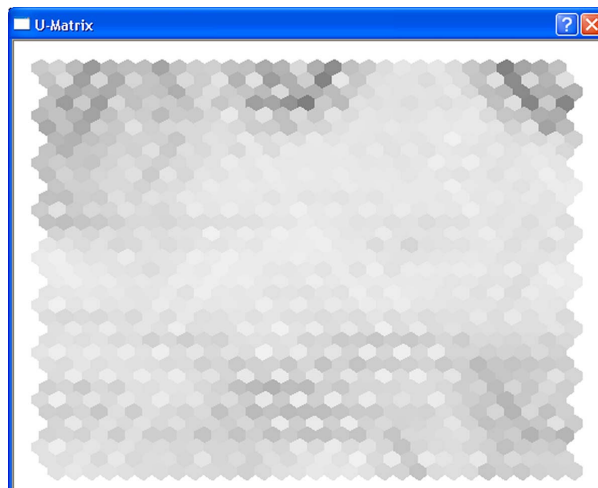


FIGURA 5.13 – U-matriz pelo sistema.

5.7 Sumário

Em função da necessidade de se integrar os algoritmos do SOM com a biblioteca Terralib foi necessário o projeto e programação do Mapa Auto-Organizável. Pacotes disponíveis e de código aberto como o SOM PAK e o SOM ToolBox atendem as necessidades de adaptações no SOM mas apresentam dificuldades de integração com a biblioteca TerraLib e de escalabilidade.

O projeto SOMLib baseou-se no paradigma Orientado a Objetos e em técnicas de programação como padrões de projeto, STL e programação genérica. O objetivo desse projeto foi construir uma biblioteca com alto nível de escalabilidade, facilidade de manutenção e de fácil integração com a TerraLib.

A partir das bibliotecas QT e SOMLib foi desenvolvido o sistema *CASA* - (*Connectionist Approach for Spatial Analysis of Areal Data*), ambiente gráfico que facilita o processo de configuração e uso dos algoritmos do SOM. Este sistema foi usado para a tarefa de treinamento da rede, análise de agrupamentos e comunicação com o banco de dados geográficos. Para visualização da U-matriz e dos Planos de Componentes foi usado o pacote SOM ToolBox.

CAPÍTULO 6

ESTUDO DE CASO: ANÁLISE ESPACIAL INTRA-URBANA EM SÃO JOSÉ DOS CAMPOS-SP

6.1 Estudo de Caso

As técnicas de análise exploratória de dados usando Mapas Auto-Organizáveis, apresentadas nos capítulos anteriores, foram aplicadas num problema de Análise Espacial de Dados em Área na escala Intra-Urbana. O problema consiste na análise exploratória de dados socioeconômicos multivariados, relativos ao estudo da exclusão/inclusão social intra-urbana, no município de São José dos Campos-SP.

A análise conduzida por Genovez (2002) tomou como base a metodologia de análise de exclusão/inclusão social do município de São Paulo-SP (Sposati, 2001). A metodologia consiste da coleta de dados socioeconômicos brutos de diversas fontes, definição de indicadores de exclusão/inclusão social, determinação de índices de avaliação de exclusão/inclusão a partir dos indicadores, determinação dos índices de Utopia, até que se chegue num índice composto de exclusão/inclusão social urbana (Iex) final para cada setor censitário. As Utopias, definidas em Sposati (1996), são convertidas em índices, que agregam informações relativas a determinadas variáveis do censo demográfico necessárias a metodologia de criação de medidas de exclusão/inclusão social. Foram usados quatro índices relativos às Utopias, são eles: Autonomia de Renda dos Chefes de Família, Desenvolvimento Humano, Qualidade de Vida e Equidade.

Para a aplicação da metodologia de Sposati (1996) nos setores censitários de São José dos Campos foram necessárias algumas alterações no método. Estas revisões tiveram como metas adequar o método às restrições quanto ao número de indicadores, uma vez que estavam disponíveis somente informações de censo do IBGE, e ajustar o método quantitativo usado para escalonamento dos valores brutos para a escala de inclusão/exclusão social urbana $[-1, 1]$, onde o valor -1 significa alto nível de exclusão social e 1 alto nível de inclusão social. Este método, desenvolvido por Genovez (2002), foi chamado de método revisto. O método revisto consiste de três fases. A primeira é responsável pela análise quantitativa dos dados brutos e composição dos índices. Na segunda fase aplicam-se métodos de análise estatística mono e multivariada sobre os índices calculados na fase anterior, para a geração de mapas síntese¹. Na terceira fase aplica-se iterativamente análise espacial de áreas para estudo da correlação espacial entre as áreas e os índices.

¹Mapa síntese é a espacialização dos índices síntese, computados a partir das variáveis mais significativas para o modelo de regressão usado.

Neste trabalho, aplicou-se a rede neural SOM na segunda e terceira fases do processo do método revisito, usando como fonte de informação os índices criados na fase 1 deste método. O objetivo foi o de verificar se a rede neural SOM chegaria a resultados e levaria a conclusões semelhantes às encontradas por Genovez (2002).

6.2 Seleção dos Dados e Pré-processamento

A seleção dos dados baseou-se no método revisito quantitativo (Genovez, 2002). A partir dos indicadores, valores percentuais, a autora definiu um método para transformação e composição de índices que possuem valores no intervalo $[-1, +1]$. Os índices indicam maior (+1) ou menor (-1) inclusão social no espaço urbano de São José dos Campos.

Dado um PRI (Parâmetro de Referência de Inclusão) para um determinado índice, mede-se a incidência dos percentuais acima e abaixo deste PRI. Para os índices compostos o procedimento é o mesmo, somando os percentuais acima e abaixo dos PRIs dos índices componentes. Quanto maior for a soma dos percentuais acima do PRI maior será o nível de inclusão. Estes cálculos já foram feitos em Genovez (2002).

Para este estudo selecionou-se os índices de Distribuição de Renda dos Chefes de Família (ARENDR), Desenvolvimento Educacional (DESEDUCR), Estímulo Educacional (ESTDUCR), Longevidade (LONGR), Qualidade Ambiental (QAMBR), Conforto Domiciliar (QDOMR), Mulheres não Alfabetizadas (MANAFR) e Concentração de Mulheres Chefes de Família (CMCHFR). Também usou-se as coordenadas planas (x, y) para avaliar o impacto da posição espacial do setor na análise exploratória dos dados, segundo a proposta da Seção 4.3.

Todo o conjunto de dados compreende um total de $n = 342$ padrões de dimensionalidade igual a $d = 8$, sem as coordenadas planas, e $d = 10$ com as coordenadas.

6.3 Configuração da Rede SOM

Ao longo deste trabalho demonstrou-se que o Mapa bidimensional, hexagonal, com função de vizinhança gaussiana, iniciação linear e aprendizagem em lote atendem aos requisitos necessários para tratar convenientemente o estudo de caso em questão. Restam, portanto, poucos parâmetros livres para definição por parte do usuário do algoritmo, são eles: as dimensões da rede, número total de épocas e raio inicial da função de vizinhança.

Definiu-se um conjunto de configurações de rede que serão avaliadas ao longo do processo de análise exploratória dos dados (Tabela 6.1). Para este mesmo conjunto de configurações de rede foram aplicados quatro mecanismos de aprendizagem, com número de épocas

TABELA 6.1 – Configurações de rede avaliadas.

| Id | M | N | m (número de neurônios) | raio inicial |
|----|----|----|-------------------------|--------------|
| 1 | 3 | 3 | 9 | 2 |
| 2 | 4 | 4 | 16 | 3 |
| 3 | 5 | 5 | 25 | 4 |
| 4 | 6 | 5 | 30 | 5 |
| 5 | 6 | 7 | 42 | 5 |
| 6 | 7 | 5 | 35 | 4 |
| 7 | 7 | 6 | 42 | 5 |
| 8 | 8 | 8 | 64 | 6 |
| 9 | 9 | 8 | 72 | 7 |
| 10 | 9 | 9 | 81 | 8 |
| 11 | 10 | 8 | 80 | 8 |
| 12 | 10 | 9 | 90 | 8 |
| 13 | 10 | 10 | 100 | 8 |
| 14 | 12 | 10 | 120 | 9 |
| 15 | 12 | 12 | 144 | 9 |
| 16 | 14 | 10 | 140 | 10 |
| 17 | 14 | 12 | 168 | 12 |
| 18 | 15 | 10 | 150 | 10 |
| 19 | 15 | 12 | 180 | 10 |
| 20 | 15 | 15 | 225 | 13 |
| 21 | 16 | 15 | 240 | 13 |
| 22 | 16 | 16 | 256 | 14 |
| 23 | 17 | 15 | 255 | 13 |
| 24 | 17 | 16 | 272 | 14 |
| 25 | 18 | 15 | 270 | 14 |
| 26 | 18 | 16 | 288 | 16 |
| 27 | 18 | 17 | 306 | 15 |
| 28 | 18 | 18 | 324 | 16 |
| 29 | 20 | 15 | 300 | 15 |
| 30 | 20 | 16 | 320 | 16 |
| 31 | 20 | 17 | 340 | 16 |
| 32 | 20 | 18 | 360 | 16 |
| 33 | 20 | 20 | 400 | 17 |
| 34 | 25 | 20 | 500 | 20 |
| 35 | 25 | 25 | 625 | 20 |
| 36 | 30 | 15 | 450 | 20 |
| 37 | 30 | 20 | 600 | 20 |
| 38 | 30 | 25 | 750 | 22 |
| 39 | 30 | 30 | 900 | 25 |

fixo em 1000, número de épocas igual a $3000 * m/n$ e treinamento em duas fases com número de épocas fixo para cada fase, 1000 para as duas fases do primeiro experimento em duas fases e $3000 * m/n$ ou 1000 para o segundo, a depender das dimensões da rede (Tabela 6.2). Os valores 1000 e 3000 são valores empíricos definidos segundo observações experimentais.

TABELA 6.2 – Experimentos conduzidos para uma rede neural SOM bidimensional, hexagonal, com função de vizinhança gaussiana e aprendizagem em lote.

| Experimento | n° de fases | n° de épocas (fase 1) | n° de épocas (fase 2) |
|-------------|-------------|-----------------------|---|
| 001 | 1 | 1000 (fixo) | - |
| 002 | 2 | 1000 (fixo) | 1000 (fixo) |
| 003 | 1 | $3000m/n$ | - |
| 004 | 2 | $3000m/n$ | se $3000m/n < 1000$ $3000m/n$ senão 1000 |

Os resultados comparativos entre os quatro experimentos estão ilustrados nos gráficos nas Figuras 6.1 a 6.6. Foram gerados os gráficos do erro de quantização, topológico, n° de agrupamentos encontrados pelo algoritmo Costa-Netto e dos índices de validação da partição de dados CDbw e Davies-Bouldin.

Todos os experimentos apresentaram curvas próximas para o erro de quantização (Figura 6.1), mostrando que o comportamento desta configuração neural não varia significativamente para alterações no número de épocas e fases de aprendizagem. Do gráfico relativo ao erro topológico (Figura 6.2) observou-se que a irregularidade de seu comportamento é mantida nos quatro experimentos, todavia, mantendo sempre as mesmas tendências. Devido à irregularidade do erro topológico é recomendável bastante critério na consideração desta métrica na avaliação final do desempenho das redes neurais avaliadas.

A aplicação do algoritmo Costa-Netto sobre os dados gerou, para cada experimento, quase o mesmo número de agrupamentos (Figura 6.3). Aqui também constata-se que há pouca interferência do número de épocas de treinamento neste processo. É evidente que isto ocorre a partir de um valor mínimo para o número de épocas, todavia este valor não foi pesquisado. O principal objetivo aqui foi, apenas, comparar quatro formas de aplicação do algoritmo de aprendizagem.

Embora o comportamento das curvas dos quatro experimentos para os índices CDbw e Davies-Bouldin (Gráficos 6.4 a 6.6) não sejam tão uniformes quantos os gráficos anteriores, constata-se que as variações encontram-se dentro de um determinado limite. Esta

interpretação visual permite afirmar que estes índices são mais sensíveis a variações no número de épocas e fases de aprendizagem do que os erros de quantização e topológico, mas os quatro experimentos seguem uma mesma tendência.

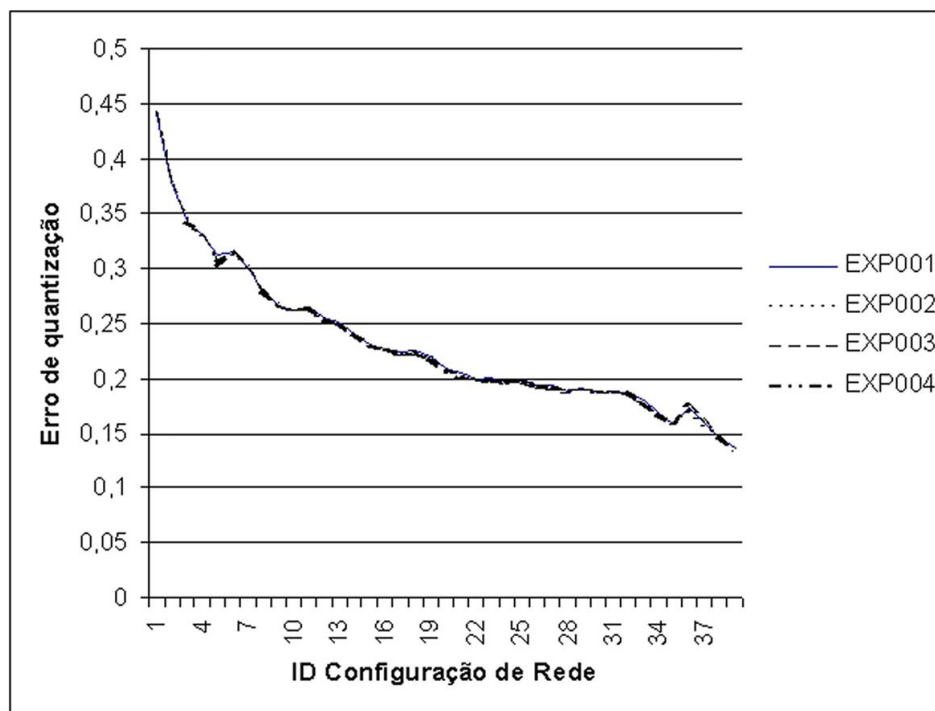


FIGURA 6.1 – Gráfico do erro de quantização.

Como o objetivo da análise exploratória é identificar tendências no conjunto de dados amostral, conclui-se que, uma vez fixado o algoritmo de aprendizagem em lote, a função de vizinhança gaussiana, a grade bidimensional hexagonal, os resultados das diversas configurações de dimensões e o raio inicial de aprendizagem são variam pouco em relação ao número de épocas e fases de aprendizagem, isto para o conjunto de dados amostral usado. Desta forma, a configuração mais simples das 4 pode ser usada como parâmetro de trabalho, o Experimento 001. Também observou-se, visualmente, que a U-matriz e os Planos de Componentes gerados pelos quatro experimentos são equivalentes, não sendo significativas as mudanças de um para outro experimento.

6.4 Identificando Dados Atípicos e Organização Geral da Estrutura dos Dados

A U-matriz, como visto na Seção 2.3.4, permite que a estrutura geral do conjunto de dados amostrais seja avaliada de maneira visual, inclusive permitindo que conjuntos de

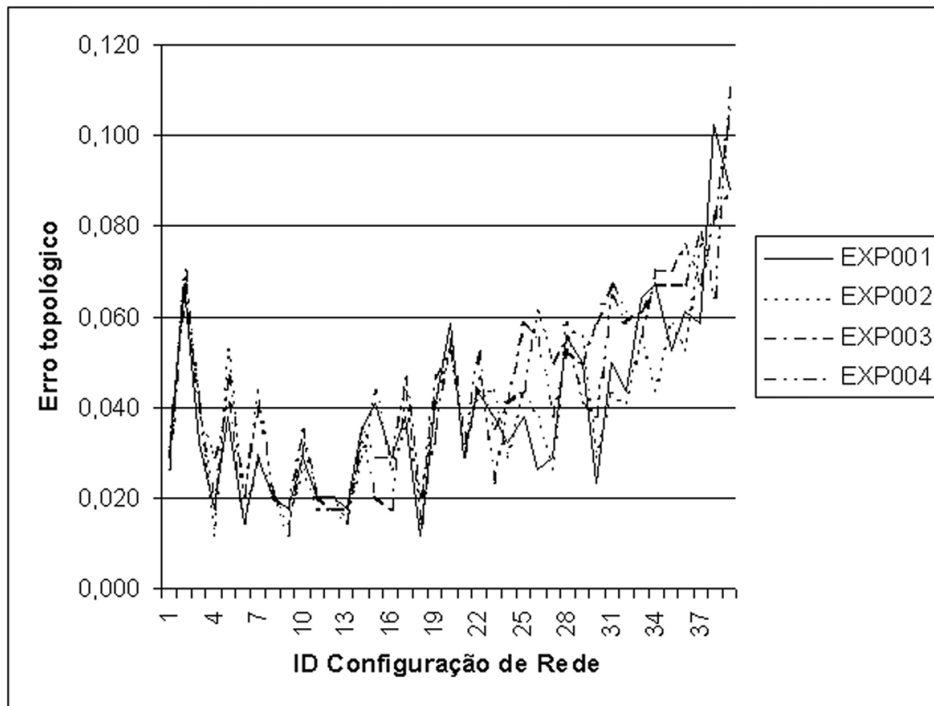


FIGURA 6.2 – Gráfico do erro topológico.

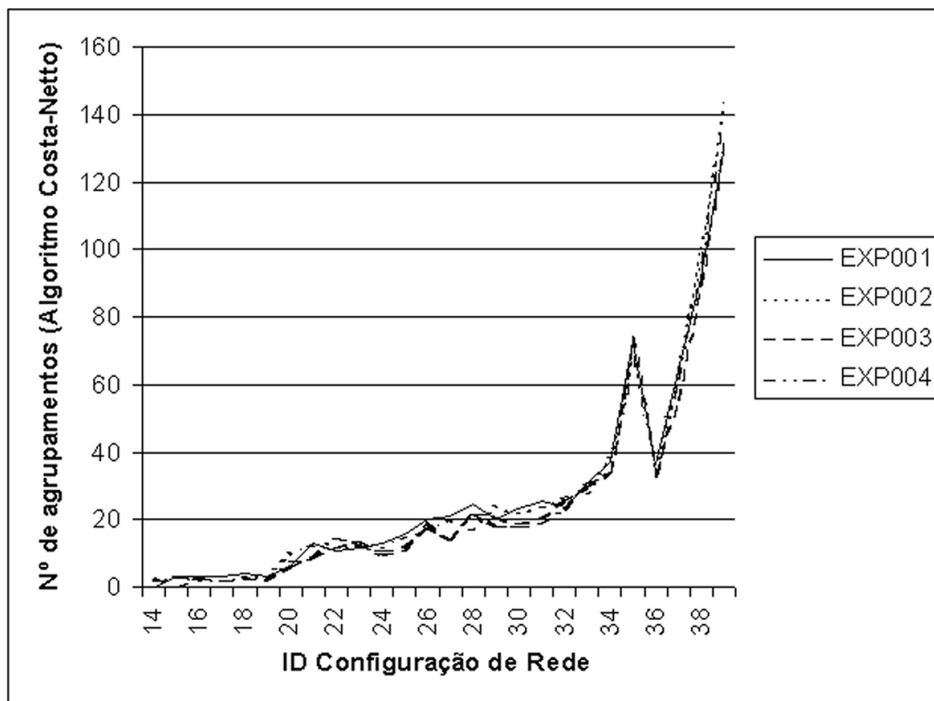


FIGURA 6.3 – Número de agrupamentos encontrados pelo algoritmo de segmentação Costa-Netto.

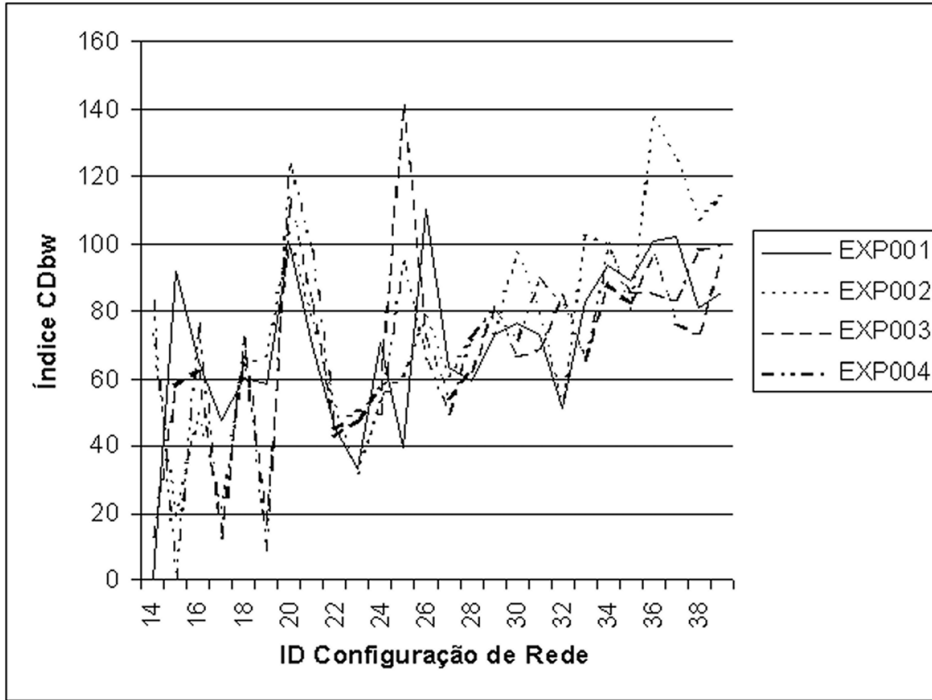


FIGURA 6.4 – Índice de validação CDbw.

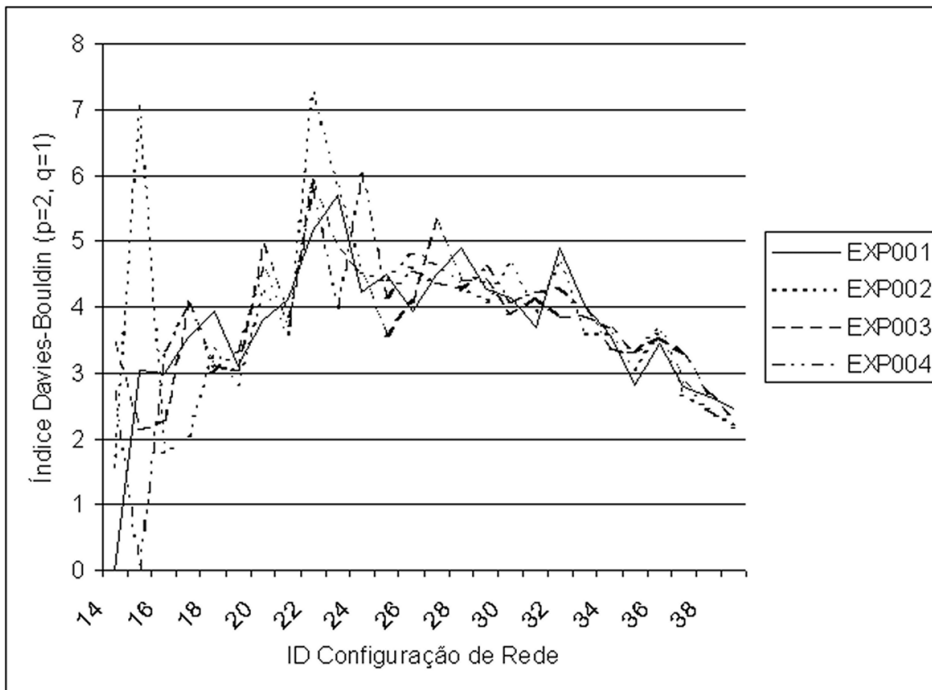


FIGURA 6.5 – Índice de validação Davies-Bouldin (p=2, q=1).

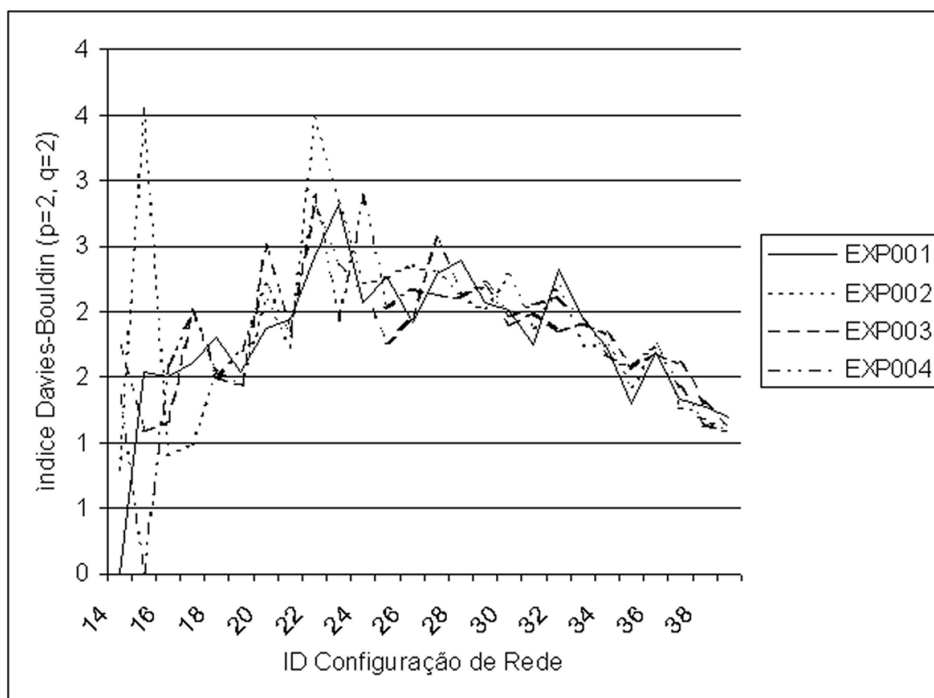


FIGURA 6.6 – Índice de validação Davies-Bouldin (p=2, q=2).

dados atípicos sejam facilmente identificados.

Avaliar cada configuração de rede da Tabela 6.1 não apresenta sentido prático, uma vez que a estrutura da U-matriz para os vários Mapas são semelhantes. A Figura 6.7 mostra que, para redes pequenas (5x5), a estrutura da U-matriz apresenta-se complexa e não fornece subsídios para a análise dos dados; já para redes muito grandes (50x30) percebe-se uma superespecialização do Mapa, representada pelos vários agrupamentos de dados observados. Esta superespecialização foi ilustrada através da plotagem do histograma do nível de atividade dos neurônios (em branco).

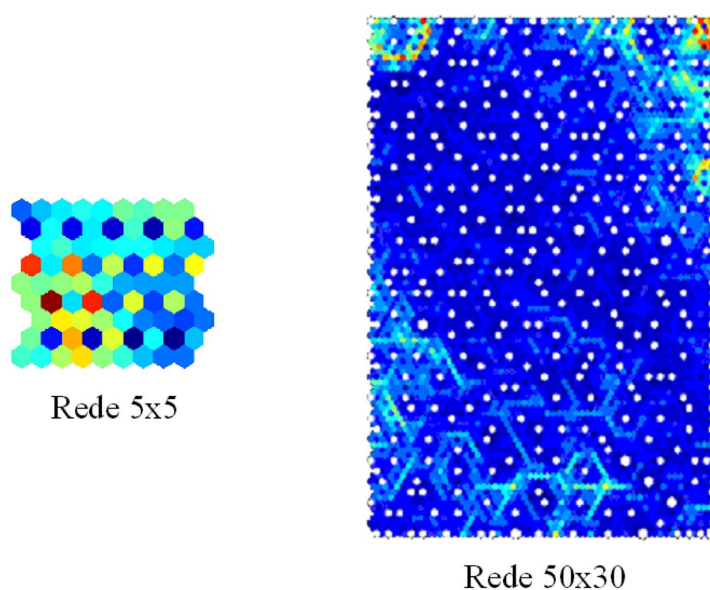


FIGURA 6.7 – U-matrizes geradas para as redes 5x5 e 50x30.

Analisando-se a curva dos erros de quantização e topológico (Figura 6.8) observa-se que a curva do erro topológico é irregular, porém levemente crescente para redes com $m/n > 1$; a curva do erro de quantização decai suavemente até, aproximadamente, $m/n = 1$. Logo, da análise visual da formação da U-matriz e dos gráficos do erro de quantização e topológico, optou-se pela configuração de rede com dimensão 20x15. Ou seja, uma rede com valor baixo para o erro de quantização, mas não com grandes dimensões, evitando a superespecialização do Mapa.

A U-matriz gerada pela rede 20x15 está ilustrada na Figura 6.9. Através desta U-matriz podem ser observados dois agrupamentos de dados bem definidos nos cantos superiores da imagem. Na parte inferior central da imagem há uma região candidata a agrupa-

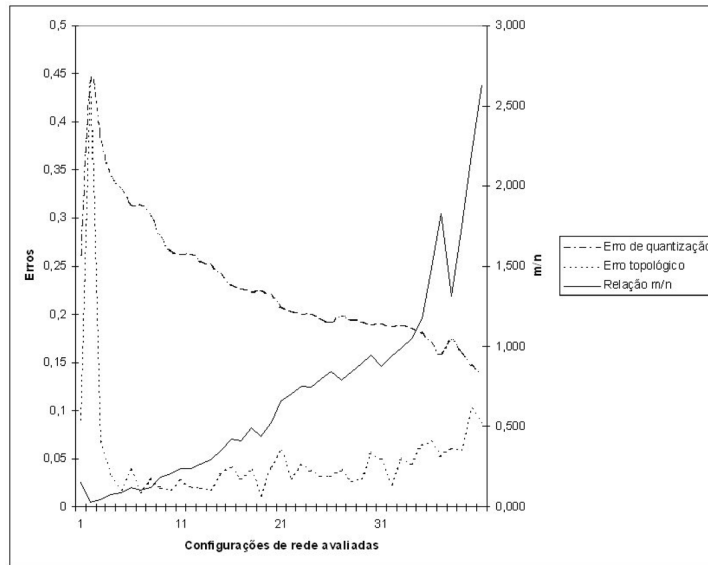


FIGURA 6.8 – Gráfico dos erros de quantização e topológico.

mento, 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 Grupo 1, e Grupo 2 para os setores relacionados com os neurônios do agrupamento do canto superior direito.

Usando o mapa dos setores censitários de São José dos Campos para mostrar quais são os setores dos Grupos 1 e 2, identifica-se os que correspondem a áreas sabidamente de exclusão social. Estas mesmas áreas foram encontradas por Genovez (2002), 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. Os mapas com os Grupos 1 e 2 estão ilustrados na Figura 6.10. Os setores dos Grupos 1 e 2 apresentam comportamento distinto dos demais setores de exclusão social. Alguns setores do Grupo 1, apesar de estarem na zona de exclusão social, possuem alto nível de propriedade. Os setores do Grupo 2 possuem alto nível de estímulo educacional (Genovez, 2002). Uma observação detalhada dos Planos de Componentes pode oferecer mais informações sobre como cada componente contribuiu para a diferenciação destes setores atípicos.

6.5 Análise 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 seguem o mesmo padrão que aqueles gerados pela rede 20x15 (Figura 6.11), com uma certa perda de resolução para a

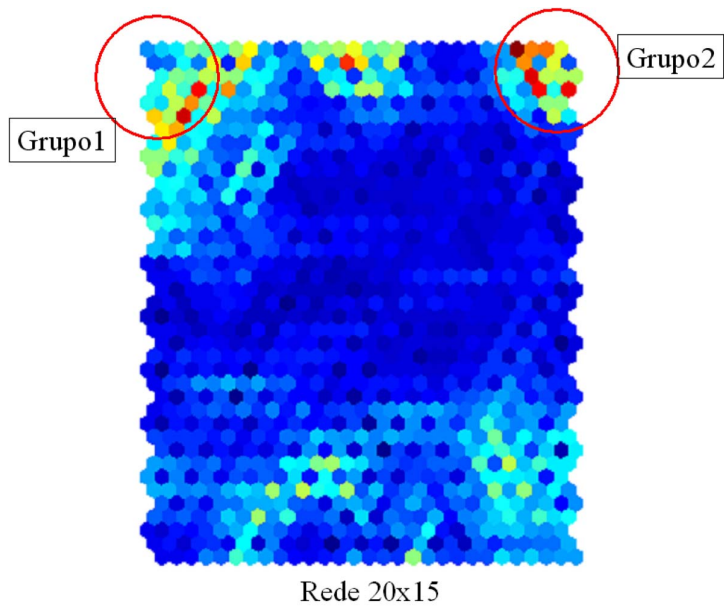


FIGURA 6.9 – U-matriz gerada para a rede 20x15.

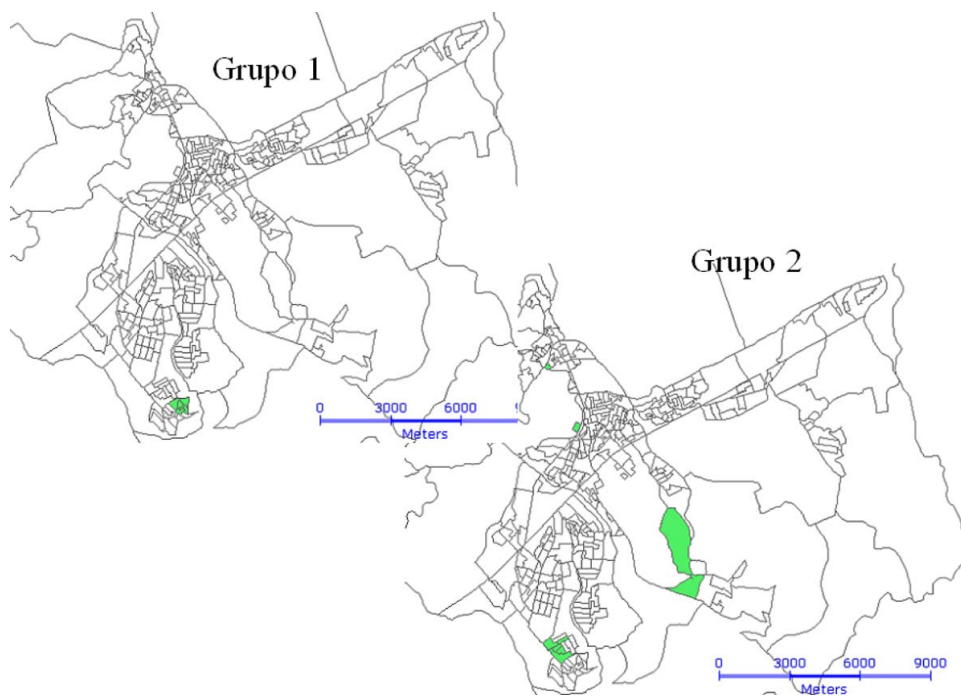


FIGURA 6.10 – Mapas dos setores censitários identificados como setores atípicos.

rede menor. Através da observação visual dos SOMs avaliados, constatou-se que o tamanho do Mapa não influencia significativamente na formação dos Planos de Componentes, embora Mapas muito pequenos acabem escondendo determinados comportamentos dos componentes. Assim, manteve-se a rede 20x15 para a análise dos Planos de Componentes.

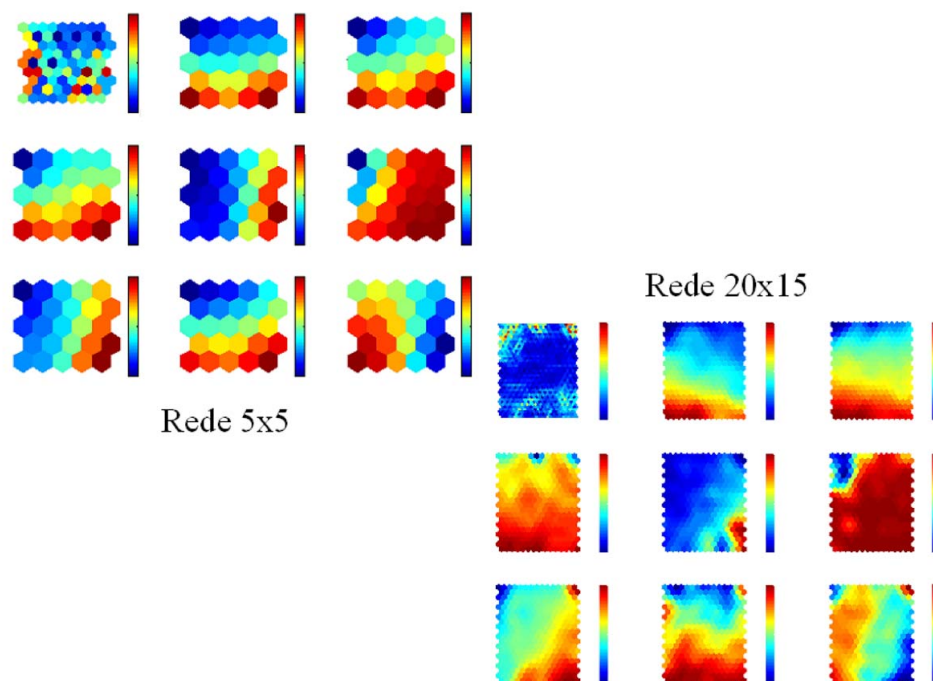


FIGURA 6.11 – Planos de Componentes. Tanto para redes pequenas (5x5), quanto para redes maiores (20x15), os planos de componentes são semelhantes.

A Figura 6.12 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 dos componentes, 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 a áreas do Mapa especializadas em setores censitários com alta inclusão social, inversamente, as regiões em azul estão especializadas em setores com alta exclusão social.

Da observação dos Planos de Componentes observa-se que:

- a) Quando as variáveis ARENDR e DESEDUCR são comparadas percebe-se que ambas possuem um elevado grau de similaridade visual, um forte indício que ambas podem estar estatisticamente correlacionadas. Também observa-se que existem mais setores com maior nível de inclusão na variável DESEDUCR do que na variável ARENDR. Todavia, existem mais setores no setor de exclusão

na variável ARENDR do que na variável DESEDUCR;

- b) Para a variável ESTEDUCR tem-se que há poucos setores identificados como de exclusão social. Estes setores estão posicionados na parte superior do plano de componente correspondente;
- c) As variáveis LONGR e QAMBR contribuem muito pouco para a diferenciação entre os setores censitários, uma vez que possuem grandes áreas homogêneas no Mapa, com destaque para a variável LONGR;
- d) As variáveis QDOMR e MCHFR possuem distribuição espacial nos Planos de Componentes horizontal, ou seja, distinta das demais variáveis, e não correspondentes entre si, o que sugere uma investigação mais detalhada para verificar por que zonas de alto nível de inclusão da variável QDOMR (canto inferior direito) correspondem a zonas de exclusão na variável MCHFR;
- e) Não existe correlação visual entre as variáveis MANALFR e MCHFR;
- f) No canto superior direito dos Planos de Componentes das variáveis ESTEDUCR, QDOMR, MANALFR e MCHFR há uma diferenciação tal que poderia explicar a formação do Grupo 2 na U-matriz da Figura 6.9.

A correlação estatística calculada por Genovez (2002), para as variáveis ARENDR e DESEDUCR foi $r = 0.946$. A baixa correlação das variáveis LONGR e QAMBR com as demais e o seu pouco impacto no processo de análise também foi observado pela autora.

6.6 Análise da Distribuição Espacial do Fenômeno

Da análise dos Planos de Componentes, rede 20x15, chega-se à conclusão de que existe um sentido exclusão-inclusão na distribuição do Mapa e que este é vertical. Usando a técnica de rotulação dos neurônios da Seção 4.4 tem-se o mapa da Figura 6.13. 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 também foi uma das conclusões do trabalho conduzido por Genovez (2002) 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. Observa-se, a partir da Figura 6.13, que o mapa gerado pelo SOM (imagem “a”) e o gerado pelo método revisto possuem distribuição semelhante, embora estejam com padrões de cores diferentes.

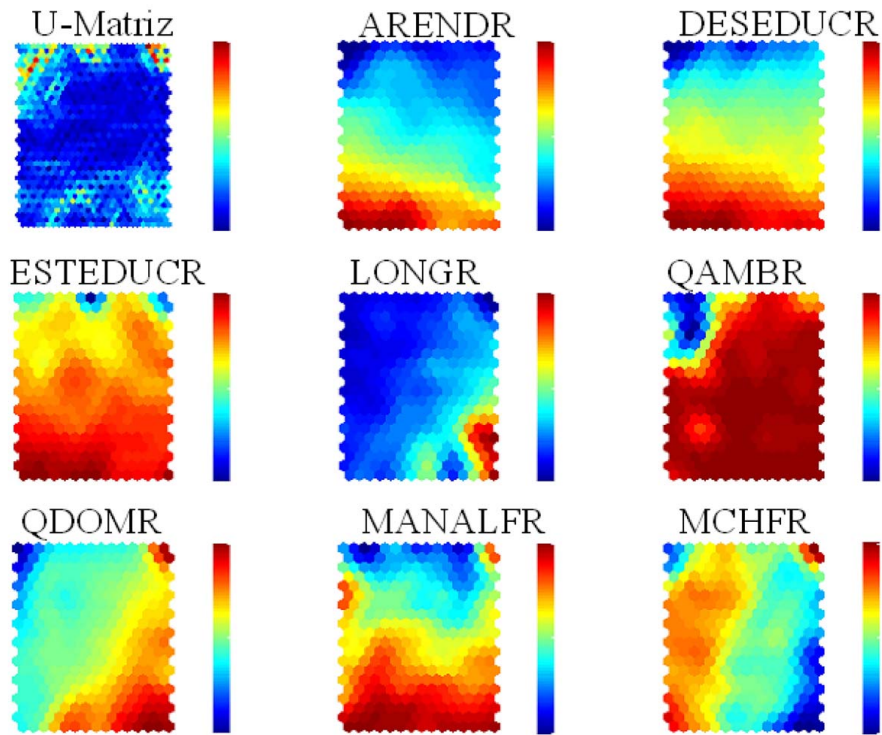


FIGURA 6.12 – Planos de Componentes para a rede 20x15.

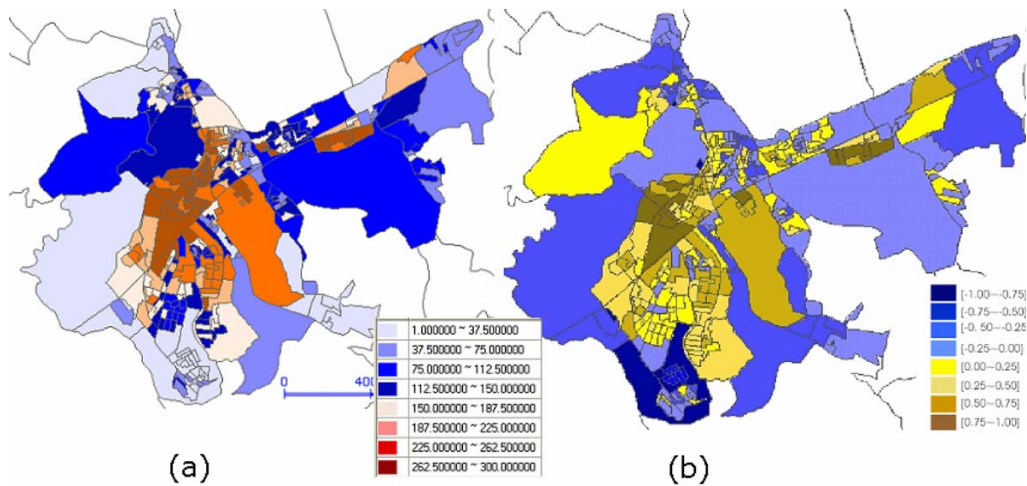


FIGURA 6.13 – Mapa gerado a partir da rotulação, no sentido vertical, da grade de neurônios, baseada na distribuição dos Planos de Componentes “a”. Mapa baseado no Iex revisto “b”. FONTE: (Genovez, 2002).

6.7 Avaliando a Inclusão das Coordenadas Planas (x, y) em x_k

Aqui avaliou-se o efeito da inclusão das coordenadas planas (x, y) , relativas ao centróide de cada setor censitário, na geração da U-matriz e dos Planos de Componentes. Observou-se que a inclusão simples destas coordenadas no vetor de característica x_k não contribuiu para a melhoria da definição da U-matriz (Figura 6.14), assim como para a análise dos Planos de Componentes (Figura 6.15).

Na U-matriz a inclusão das coordenadas planas teve efeito negativo, apagando as formações ou indicações de agrupamentos. A Figura 6.14 mostra que os agrupamentos nos cantos superiores e região central inferior da U-matriz foram apagados.

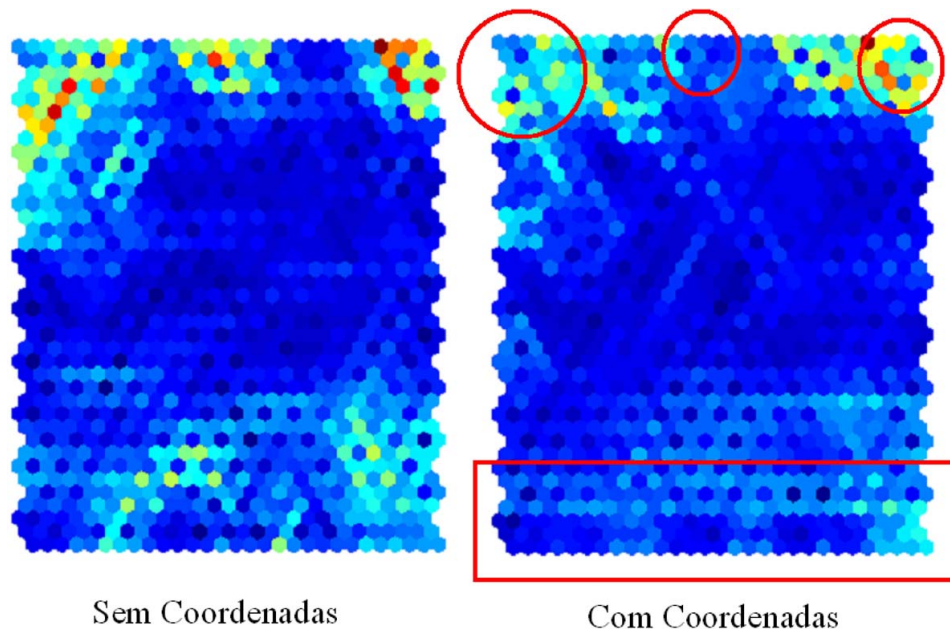


FIGURA 6.14 – Efeito, na U-matriz, da inclusão das coordenadas planas.

Nos Planos de Componentes a inclusão das coordenadas planas não alterou a formação dos demais componentes e não trouxe nenhum elemento novo para a análise de correlação e significância das variáveis. Pode-se atribuir isto ao fato de que as coordenadas planas fazem parte do fenômeno estudado, mas não participam diretamente da avaliação de exclusão/inclusão social.

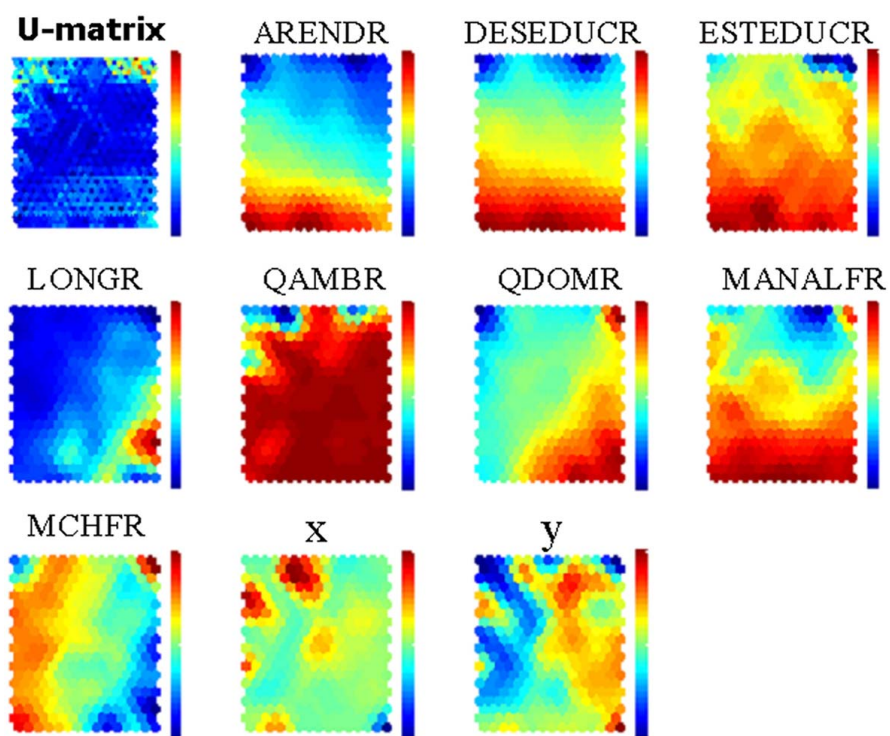


FIGURA 6.15 – Efeito, nos Planos de Componentes, da inclusão das coordenadas planas.

6.8 Descoberta de Agrupamentos e Análise da Dependência Espacial

O particionamento do conjunto de dados num número c de agrupamentos foi realizado através do algoritmo Costa-Netto, em duas fases (Figura 6.16). Primeiramente os dados são apresentados ao SOM, este é treinado e, então, seus vetores de código são particionados. Como cada padrão está associado a um vetor de código, seu BMU, pode-se particionar os dados a partir dos vetores de código particionados.

Para validação dos agrupamentos gerados usou-se os índices Davies-Bouldin, ($p = 2, q = 1$); ($p = 2, q = 2$), e o Cdbw. Para cada configuração de rede (Tabela 6.1), calculou-se os valores para o índice Davies-Bouldin e para o CDbw, todavia, aqui não estão incluídas algumas redes pequenas ($c = 1$).

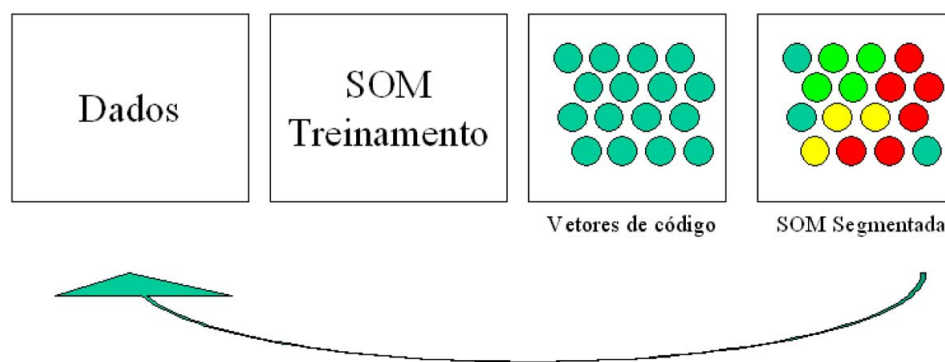


FIGURA 6.16 – Fases do processo de particionamento dos dados em c agrupamentos.

Do gráfico correspondente ao índice Davies-Bouldin (Figura 6.17), tem-se que a melhor partição é a da rede 14x10, com índices Davies-Bouldin 3,0 e 1,5, respectivamente, e $c = 3$. Porém, ao analisar o Mapa colorido (Figura 6.18), percebe-se que o particionamento não corresponde à realidade, uma vez que coloca, no mesmo grupo, neurônios especializados em setores de exclusão e inclusão.

Do gráfico correspondente ao índice CDbw (Figura 6.19), tem-se que a melhor partição é a da rede 18x16, com índice CDbw igual a 110,14 e $c = 20$. Da análise do Mapa particionado colorido (Figura 6.20), conclui-se que a partição obedece ao sentido da distribuição vertical do Mapa e que identifica claramente as zonas de dados atípicos. O mapa dos setores censitários da cidade de São José dos Campos foi colorido segundo esta partição do SOM (Figura 6.21), e demonstra coerência, tanto com os resultados obtidos anteriormente neste trabalho, quanto com os resultados obtidos por Genovez (2002), no sentido de identificação de áreas de inclusão e exclusão social urbana. Outra

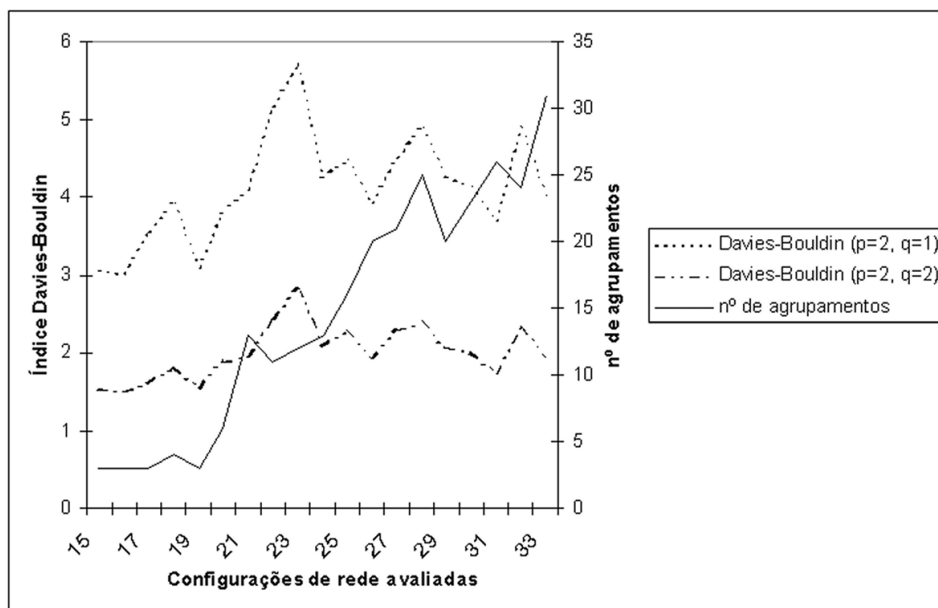


FIGURA 6.17 – Gráficos para o índice Davies-Bouldin.

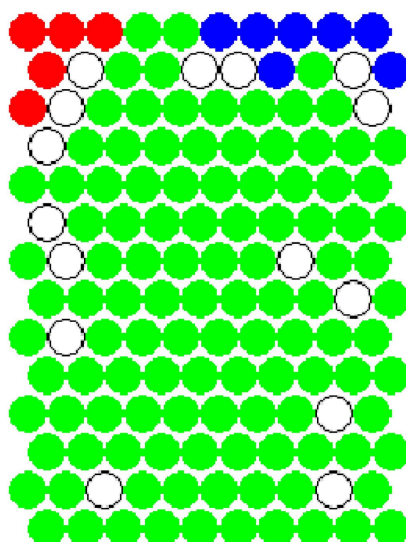


FIGURA 6.18 – Mapa neural particionado segundo o índice Davies-Bouldin.

observação é que o SOM evidencia forte presença de regimes espaciais significativos, bem definidos, para a área urbana como um todo, que podem ser caracterizados como fragmentos urbanos onde há uma forte relação entre os atributos e a posição espacial do setor. Para a avaliação quantitativa desta dependência espacial calculou-se o IRVE. Este índice, calculado para o mapa da Figura 6.13, foi de 0,66, ou seja, um alto grau de agregação dos agrupamentos. Isto significa que regiões próximas possuem indicadores de exclusão/inclusão social próximos. A Tabela 6.3 mostra os valores de p_i , q_i e $IRVE_i$ para cada agrupamento i .

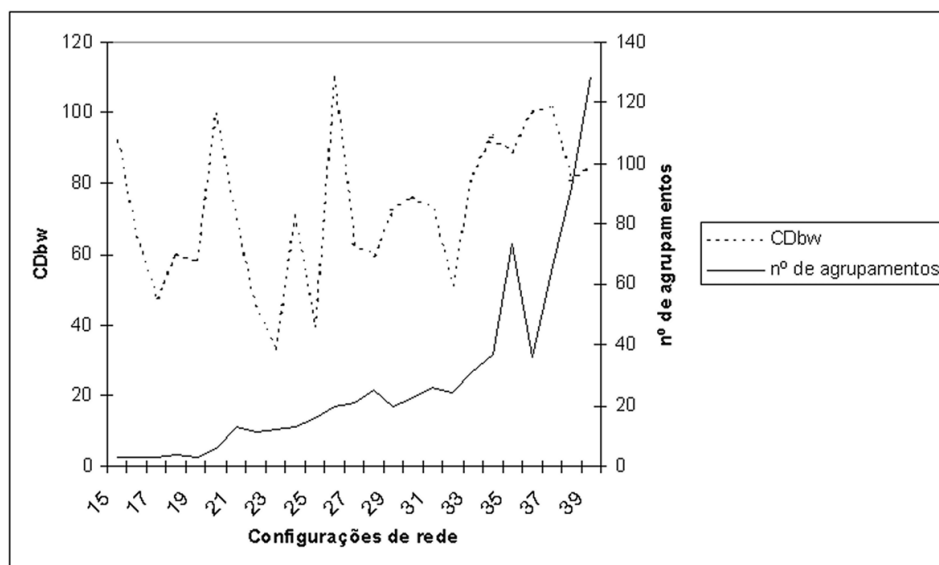


FIGURA 6.19 – Gráficos para o índice CDbw.

O índice IRVE não serve como mecanismo de subsídio à escolha da melhor configuração de rede porque não avalia a qualidade da partição dos dados. A aplicação do IRVE aqui restringiu-se à avaliação da dependência espacial para as redes com os melhores índices de validação do particionamento, em particular o índice CDbw. Observou-se que, para a maioria deles, o valor do índice está acima de 0,50 (Gráfico 6.22). Estes valores indicam um certo grau de dependência espacial a ser verificado pelo Índice Global de Moran (IGM). Segundo Genovez (2002) o IGM para o mesmo estudo de caso é igual a 0,7216, o que confirma o alto grau de dependência espacial.

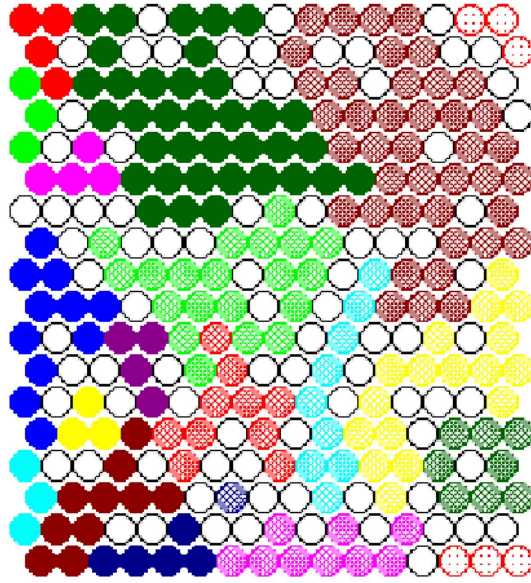


FIGURA 6.20 – Mapa particionado segundo o índice Cdbw.

TABELA 6.3 – Resultados para o índice IRVE do experimento 001, configuração de rede 26.

| Grupo | q_i | p_i | $IRVE_i$ |
|-------|-------|-------|----------|
| 1 | 3 | 11 | 0,82 |
| 2 | 3 | 4 | 0,50 |
| 3 | 10 | 19 | 0,53 |
| 4 | 4 | 7 | 0,57 |
| 5 | 5 | 5 | 0,00 |
| 6 | 6 | 6 | 0,00 |
| 7 | 4 | 18 | 0,83 |
| 8 | 18 | 68 | 0,75 |
| 9 | 6 | 8 | 0,38 |
| 10 | 5 | 8 | 0,50 |
| 11 | 6 | 15 | 0,67 |
| 12 | 14 | 34 | 0,62 |
| 13 | 10 | 12 | 0,25 |
| 14 | 6 | 18 | 0,72 |
| 15 | 11 | 22 | 0,55 |
| 16 | 13 | 61 | 0,80 |
| 17 | 2 | 10 | 0,90 |
| 18 | 3 | 3 | 0,00 |
| 19 | 5 | 8 | 0,50 |
| 20 | 1 | 5 | 1,00 |

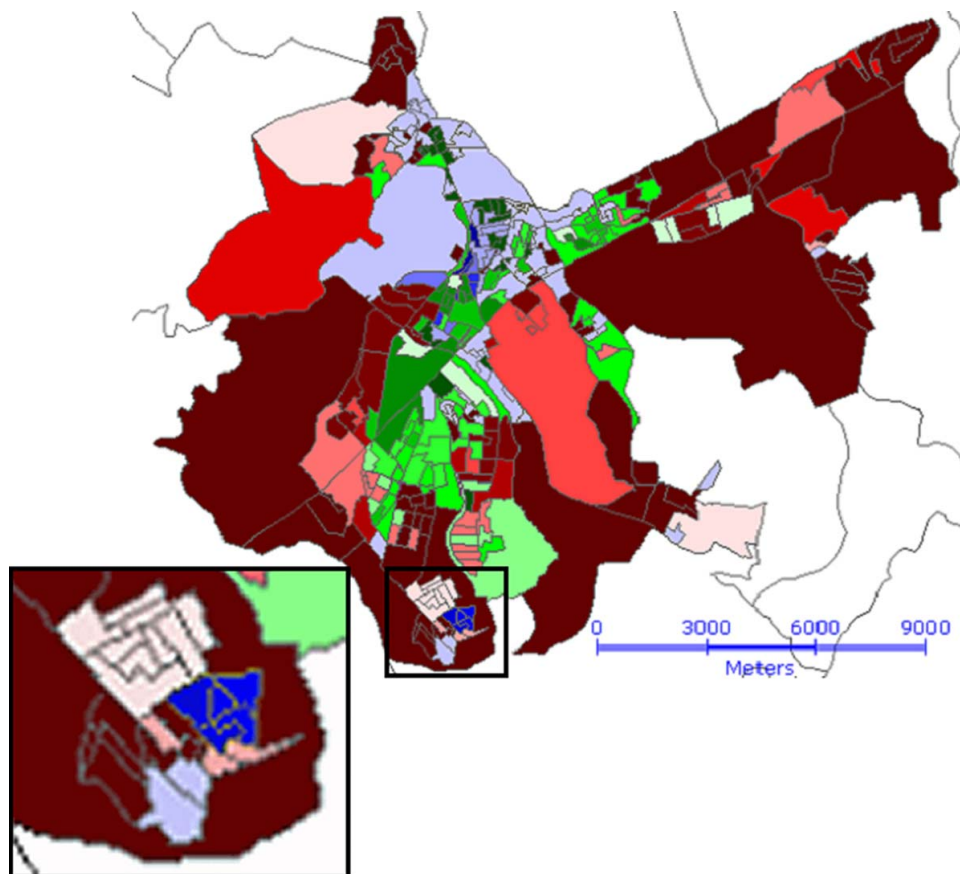


FIGURA 6.21 – Mapa dos setores censitários gerados a partir do SOM particionado segundo o algoritmo Costa-Netto e validação do índice CDbw. Em destaque o setor sul da área urbana onde pode-se verificar que o algoritmo identificou dentro de uma área de exclusão sub-agrupamentos que podem ser caracterizados como fragmentos urbanos.

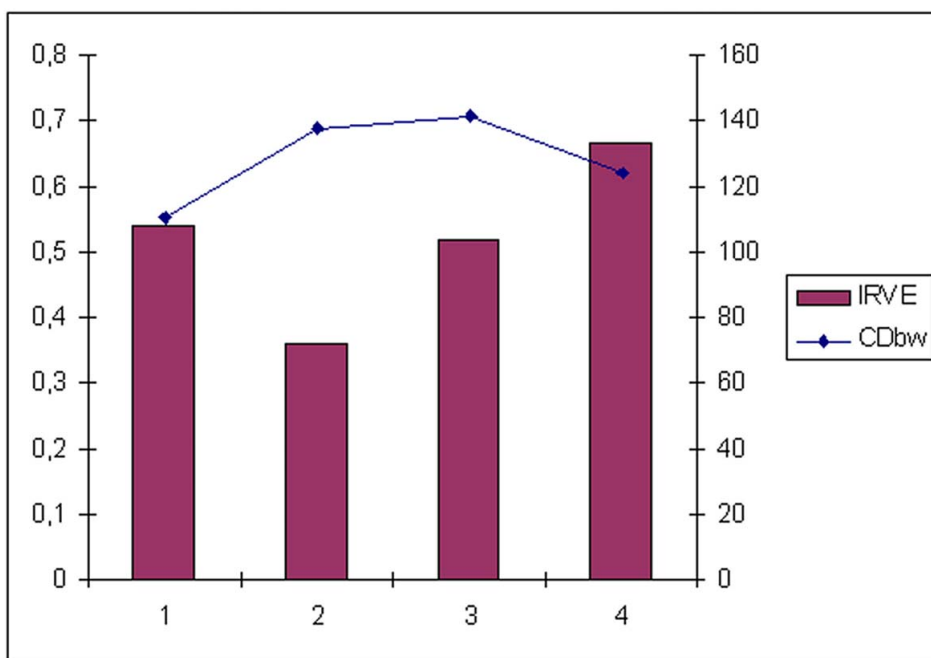


FIGURA 6.22 – Relação entre os índices IRVE e CDbw.

6.9 Sumário

Neste capítulo foram usadas as técnicas, os sistemas e os métodos apresentados nos capítulos anteriores no problema de mapeamento da exclusão/inclusão social urbana em São José dos Campos. O estudo baseou-se nos dados gerados por Genovez (2002). Foram analisados 342 setores censitários, cada setor associado a um conjunto de 8 variáveis: distribuição de renda dos chefes de família, desenvolvimento educacional, estímulo educacional, longevidade, qualidade ambiental, mulheres não-alfabetizadas e concentração de mulheres não-alfabetizadas chefes de família.

Foram analisadas 39 configurações de rede em 4 tipos de processos de aprendizagem. Des-tes experimentos observou-se que, para o conjunto de dados estudado, as redes pequenas (poucos neurônios) não conseguem extrair informações dos dados e as redes grandes (muitos neurônios) superespecializam-se. Observa-se, também, que o gráfico do erro de quantização apresenta decaimento contínuo a medida que o tamanho da rede aumenta, enquanto que o gráfico do erro topológico apresenta comportamento irregular mas com tendência a crescer a medida que o número de neurônios aumenta. Estes comportamentos sugerem que configurações intermediárias tendem a ser melhores opções.

A metodologia consiste no uso da U-matriz para visualização da estrutura geral dos dados e possível detecção de dados atípicos, no uso dos Planos de Componentes para análise de componentes e da distribuição espacial geral do fenômeno, uso do algoritmo de detecção automática de agrupamentos para análise de agregados espaciais e medida do grau de dependência espacial. Os resultados alcançados mostraram-se satisfatórios, além de serem compatíveis com os resultados obtidos por Genovez (2002).

CAPÍTULO 7

CONCLUSÕES

7.1 Considerações Finais

A estruturação de uma metodologia, congregando diversas abordagens de análise espacial com SOM mostrou-se eficaz para o caso estudado. Foram reunidas as técnicas de detecção visual de agrupamentos por meio da U-matriz, a análise de correlação e significância de componentes com os Planos de Componentes e a segmentação automática dos dados através do SOM. Foram adaptadas as técnicas relativas aos Planos de Componentes para a análise da distribuição espacial do fenômeno e a técnica de segmentação automática para a análise da dependência espacial através da visualização dos dados e do índice IRVE. Os experimentos mostraram que os resultados obtidos pelo SOM são particularmente sensíveis a variações nas dimensões da grade $M \times N$. O erro topológico contribui pouco para a avaliação da qualidade da rede neural, tendo comportamento bastante irregular em todos os casos estudados. O erro de quantização apresentou comportamento uniforme para os quatro experimentos e mostrou-se ser uma métrica confiável para a avaliação da qualidade do Mapa. Todavia, a escolha do Mapa ideal dependerá das várias observações combinadas, desde a U-matriz, passando pelos erros de quantização e topológico, Planos de Componentes, segmentação automática e cálculo do IRVE.

A aplicação da U-matriz e dos Planos de Componentes para, respectivamente, avaliar a presença de dados atípicos e analisar os componentes, mostrou-se eficaz uma vez que foram observados padrões distintos em relação a análise de agrupamentos e em relação a correlação e significância de variáveis.

No processo de segmentação automática, através do algoritmo Costa-Netto e validação pelo índice CDbw, o uso dos vetores de código como vetores de referência no cálculo do CDbw mostrou-se aplicável e com resultados coerentes. Embora nenhum comparativo com outras técnicas de cálculo dos vetores de referência tenha sido feita, a boa partição dos dados demonstrou a aplicabilidade do método. A partição dos dados para o melhor índice CDbw mostrou-se coerente com os resultados anteriores, separando áreas de exclusão e inclusão social, mostrando o sentido centro-periferia da distribuição espacial do problema e identificando regimes espaciais locais distintos. O índice de validação CDbw mostrou-se mais adequado para o estudo de caso avaliado quando comparado ao índice Davies-Bouldin.

O desenvolvimento da biblioteca SOMLib permitiu que o algoritmo SOM fosse integrado

à biblioteca **TerraLib** e que as simulações pudessem ser feitas diretamente sobre a base de dados geográficos. Os requisitos de manutenibilidade e portabilidade somente serão testados, efetivamente, a partir de novas versões da biblioteca SOMLib, quando as interfaces estiverem estabilizadas. Todavia, o desenho e a estrutura garantem independência entre as classes base, de aprendizagem e de topologia. Também foi garantida a separação total entre dados e algoritmos, permitindo que novos padrões de organização dos dados de entrada possam ser tratados pelos algoritmos da SOMLib. Um produto direto desta biblioteca é o sistema *CASA*, ferramenta visual para análise espacial de área com SOM.

Para a tarefa de análise exploratória espacial avaliou-se três técnicas: a inclusão das coordenadas planas do centróide de cada região de análise no vetor de características x_k ; a análise da distribuição espacial do fenômeno a partir do estudo dos Planos de Componentes; e a análise da presença de regimes espaciais através da detecção automática de agrupamentos e cálculo do IRVE.

O método de inclusão das coordenadas planas no vetor de características não contribuiu para a tarefa de análise exploratória dos dados. Como (x, y) não se constituem partes diretamente ligadas ao problema de exclusão/inclusão social, a sua adição não ajuda a separar os padrões. Logo, que outros métodos de inclusão das coordenadas devem ser considerados.

A análise da distribuição espacial do fenômeno a partir dos Planos de Componentes mostrou-se bastante útil e de fácil aplicação. Em função da pouca variabilidade da configuração dos Planos de Componentes aos parâmetros iniciais da rede pode-se obter bons resultados com pouco esforço de parametrização. Tanto o SOM, quanto as análises estatísticas, mostraram que existe um sentido para a distribuição espacial do problema e que este é centro-periferia.

A análise da dependência espacial através da segmentação automática do SOM mostrou que existe uma relação entre a variação nos atributos e a localização espacial dos setores censitários. Esta relação pode ser analisada visualmente através do mapa coroplético, particionado segundo a segmentação ou através do índice IRVE. Embora este índice não seja uma métrica de dependência espacial equivalente ao Índice Global de Moran (IGM), pode-se afirmar que o valor do IRVE para o experimento 001, $IRVE = 0,66$, confirma o alto grau de dependência espacial quando comparado ao IGM calculado por Genovez (2002), $IGM = 0,7216$, considerando o mesmo estudo de caso.

Conclui-se que os resultados obtidos pelo SOM foram bastante próximos dos resultados obtidos por Genovez (2002), considerando somente as questões relativas à análise

exploratória dos dados: detecção de dados atípicos, distribuição espacial do fenômeno, análise de correlação e significância de variáveis, análise de agrupamentos e dependência espacial.

7.2 Trabalhos Futuros

Quanto à metodologia de análise exploratória de dados geoespaciais com SOM, pode-se progredir a partir da adição de técnicas auxiliares para analisar os Planos de Componentes, verificar os resultados para o caso de uso de variantes do SOM com melhor formação do Mapa topológico e procurar um meio de inclusão das coordenadas planas no modelo neural. Faz-se necessário, também, a análise dos agrupamentos através de diferentes pontos de vista partindo-se de diferentes configurações do Mapa Auto-Organizável, incluindo-se neste contexto as redes com aprendizagem sequencial e com topologias dinâmicas, e a partir de diferentes algoritmos de detecção automática de agrupamentos através do SOM.

Quanto à biblioteca SOMLib, pode-se garantir que a estabilidade da mesma somente virá ao longo do tempo e a partir de seu uso por diferentes grupos de trabalho. Um avanço bastante importante seria a adição de técnicas de processamento de alto desempenho, para permitir que a biblioteca trabalhe com grandes massas de dados, como no caso das imagens de sensores remotos. O sistema *CASA* evoluirá de acordo com a evolução da biblioteca SOMLib, porém a conexão com maior nível de acoplamento entre o sistema *CASA* e TerraView pode contribuir para a adição de técnicas mais interativas de análise dos dados e geração de mapas coropléticos como, por exemplo, a conexão entre a navegação dinâmica pela U-matriz ou Planos de Componentes e coloração automática do mapa dos setores censitários.

Esperamos estar ampliando ainda mais o conjunto de possibilidades de trabalhar dados geográficos de natureza sócioeconômica de maneira territorializada, disponibilizando assim mais um instrumento de auxílio a recolocação do território na análise de políticas públicas para as cidades. Como disse Koga (2003, p. 266) “Entre o ‘fio da navalha’ da exclusão/inclusão social coloco em debate o papel do território enquanto um possível ‘fio da meada’ que possa dar início a uma nova trama de tecer as políticas públicas brasileiras em direção à justiça social”.

REFERÊNCIAS BIBLIOGRÁFICAS

- Babu, G. P. Self-organizing neural networks for spatial data. **Pattern Recognition Letters**, v. 18, n. 2, p. 133–142, February 1997. 59
- Bailey, T. C.; Gatrell, A. C. **Interactive spatial data analysis**. London: Longman Scientific and Technical, 1995. 413 p. 57, 58
- Bishop, C. **Neural networks for pattern recognition**. Oxford: Oxford University Press, 1995. 504 p. 32
- Blake, C.; Merz, C. **UCI repository of machine learning databases**. University of California, 1998. Disponível em:
<<http://www.ics.uci.edu/~mllearn/MLRepository.html>>. Acesso em: Abril 2004. 72
- Bollivier, M.; Dubois, G.; Maignan, M.; Kanevsky, M. Multilayer perceptron with local constraint as an emerging method in spatial data analysis. **Nuclear Instruments & Methods in Physics Research Section A**, v. 309, n. 1-2, p. 226–229, November 1997. 26
- Cai, Y. Artificial neural-network method for soil-erosion forecasting. **Bodenkultur**, v. 46, n. 1, p. 19–24, 1995. 26
- 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, December 2001. 27, 58, 59
- Câmara, G.; Monteiro, A. Geocomputation techniques for spatial analysis: is it the case for health data sets? **Caderno de Saúde Pública**, v. 17, n. 5, p. 1059–1081, September/October 2001. 25
- Câmara, G.; Neves, M.; Monteiro, A.; Souza, R.; Paiva, J. A.; Vinhas, L. SPRING and TerraLib: integrating spatial analysis and GIS. In: Specialist meeting on spatial data analysis software tools, 2002, Santa Barbara. **Proceedings...** Santa Barbara, CA: Center for spatially integrated social science, 2002. p. 65–78. 65
- Câmara, G.; Vinhas, L.; Souza, R.; Paiva, J.; Monteiro, A.; Carvalho, M.; Raoult, B. Design patterns in GIS development: the Terralib experience. In: Workshop Brasileiro de Geoinformática, 3., 2001, Rio de Janeiro. **Anais...** São José dos Campos: INPE, 2001. p. 89–98. 27, 70

Coplien, J. **Multi-paradigm design for C++**. Reading: Addison Wesley, 1998. 132 p. 65

Costa, J. A. F. **Classificação automática e análise de dados por redes neurais auto-organizáveis**. 1999. 345 p. Tese (Doutorado em Engenharia Elétrica) - Universidade Estadual de Campinas, Campinas. 1999. 15, 36, 37, 41, 47, 48, 50, 75

Costa, J. A. F.; Andrade Netto, M. L. Clustering of complex shaped data sets via Kohonen maps and mathematical morphology. In: Data Mining and Knowledge Discovery, 2001, Bellingham. **Proceedings...** Bellingham: SPIE, 2001. p. 16–27. 41, 47

———. Segmentação do SOM baseada em particionamento de grafos. In: Congresso Brasileiro de Redes Neurais, 6., 2003, São Paulo. **Anais ...** São Paulo: SBRN, 2003. 43, 54

Couclelis, H. Geocomputation in context. In: Longley, P.; Brooks, S.; McDonnell, R.; Macmillan, B. ed. **Geocomputation: a primer**. Chichester: John Wiley and Sons, 1998. Cap. 3, p. 145–168. 25

Davies, D. L.; Bouldin, D. W. A cluster separation measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 1, n. 2, p. 224–227, April 1979. 47, 48, 50

Demartines, P.; Blayo, F. Kohonen Self-Organizing Maps: is the normalization necessary? **Complex Systems**, v. 6, n. 2, p. 105–123, April 1992. 36

Erwin, E.; Obermayer, K. K.; Schuler, K. Self-Organizing Maps: stationary states, metastability and convergence rate. **Biological Cybernetics**, v. 67, n. 1, p. 35–45, July 1992. 34

Fausett, L. **Fundamentals neural networks: architectures, algorithms, and applications**. Englewood, NJ: Prentice Hall, 1994. 462 p. 32

Fischer, M.; Getis, A. **Recent developments in spatial analysis**. Heidelberg: Springer, 1996. 433 p. 25

Flexer, A. On the use of Self-Organizing Maps for clustering and visualization. **Intelligent Data Analysis**, v. 5, n. 5, p. 373–384, October 2001. 37

Foody, G. Applications of the Self-Organising Feature Map neural network in community data analysis. **Ecological Modelling**, v. 120, n. 2-3, p. 97–107, August 1999. 59

- Franzini, L.; Bolchi, P.; Diappi, L. Self Organizing Maps: a clustering neural method for urban analysis. In: *Recontres de Théo Quant*, 5., 2001, Besançon. **Proceedings ...** Besançon: Univ-FCOMTE, 2001. p. 1–15. 59, 60
- Gahegan, M. What is geocomputation? **Transactions in GIS**, v. 3, n. 3, p. 203–206, June 1999. 25
- 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, n. 4, p. 267–292, July 2002. 26, 59, 60
- Gamma, E.; Helm, R.; Johnson, R.; Vlissides, J. **Design patterns : elements of reusable object-oriented software**. Reading, MA: Addison Wesley, 1995. 395 p. 65, 66, 67
- Genovez, P. C. **Território e desigualdades**: análise espacial intra-urbana no estudo da dinâmica de exclusão/inclusão social no espaço urbano em São José dos Campos-SP. 2002. 325 p. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2002. 17, 27, 82, 90, 94
- Haese, K. Self-Organizing Feature Maps with self-adjusting learning parameters. **IEEE Transactions on Neural Networks**, v. 9, n. 6, p. 1270–1278, November 1998. 36
- Haese, K.; Goodhill, G. J. Auto-SOM: recursive parameter estimation for guidance of Self-Organizing Feature Maps. **Neural Computation**, v. 13, n. 3, p. 595–619, March 2001. 36
- Halkidi, M.; Vazirgiannis, M. Clustering validity assessment using multi representatives. In: *Hellenic Conference on Artificial Intelligence*, 2., 2002, Thessaloniki. **Proceedings ...** Thessaloniki: SETN, 2002. 47, 50, 52, 53
- Haykin, S. **Redes neurais**: princípios e práticas. Porto Alegre: Bookman, 2001. 900 p. 31, 32, 38
- Hewitson, B.; Crane, R. **Neural nets**: applications in geography. Dordrecht: Kluwer, 1994. 194 p. 25, 26
- Ji, C. Y. Land-use classification of remotely sensed data using self-organizing feature maps neural networks. **Photogrammetric Engineering & Remote Sensing**, v. 66, n. 12, p. 1451–1460, December 2000. 26, 58

- Kaski, S.; Kangas, J.; Kohonen, T. Bibliography of Self-Organizing Map (SOM) papers: 1981–1997. **Neural Computing Surveys**, v. 1, p. 102–350, 1998. 27
- Kaski, S.; Kohonen, T. Exploratory data analysis by the Self-Organizing Map: structures of welfare and poverty in the world. In: *Neural Networks in the Capital Markets*, 3., 1996, London. **Proceedings...** London: World Scientific, 1996. p. 498–507. 41, 44, 59, 60
- Kaski, S.; Venna, J.; Kohonen, T. Coloring that reveals high-dimensional structures. In: *Neural Information Processing*, 6., 1999, Perth. **Proceedings ...** Piscataway, NJ: IEEE Service Center, 1999. v. 2, p. 729–734. 39, 41
- Koga, D. **Medidas de cidades: entre territórios de vida e territórios vividos**. São Paulo: Cortez, 2003. 299 p. 25
- Kohonen, T. **Self-Organizing Maps**. Berlin: Springer, 2001. 501 p. 26, 27, 32, 33, 37, 38, 39, 41, 44, 65
- Kohonen, T.; Hynninen, J.; Kangas, J.; Laaksonen, J. **SOM_PAK: the self-organizing map program package**. Helsinki, April 1995. Disponível em: http://www.cis.hut.fi/research/som_lvq_pak.shtml. Acesso em: Dezembro 2002. 65
- Kropp, J. A neural network approach to the analysis of city systems. **Applied Geography**, v. 18, n. 1, p. 83–96, January 1998. 59
- Lawrence, S.; Tsoi, A. C.; Giles, C. L. Correctness, efficiency, extendability and maintainability in neural network simulation. In: *International Conference on Neural Networks*, 1996, Piscataway, NJ. **Proceedings ...** Piscataway, NJ: IEEE Press, 1996. p. 474–479. 66
- Lee, S.; Cho, S.; Wong, P. Rainfall prediction using artificial neural networks. **Journal of Geographic Information and Decision Analysis**, v. 2, n. 1-2, p. 233–242, 1998. 26
- Lo, Z.; Fujita, M.; Bavarian, B. Analysis of neighborhood interaction in Kohonen neural networks. In: *International Parallel Processing Symposium*, 6., 1991, Anaheim, CA, USA. **Proceedings ...** Anaheim, CA, USA: IEEE, 1991. p. 247–249. 34
- Lo, Z.; Yu, Y.; Bavarian, B. Analysis of the convergence properties of topology preserving neural networks. **IEEE Transactions on Neural Networks**, v. 4, n. 2, p. 207–220, March 1993. 34

- Longley, P. A.; Brooks, S. M.; McDonnell, R.; B., M. **Geocomputation: a primer**. Chichester: John Wiley, 1998. 290 p. 26
- Medeiros, J. **Banco de dados geográficos e redes neurais artificiais: tecnologias de apoio à gestão do território**. São Paulo, 1999. 255 p. Tese (Doutorado em Geografia) - Universidade de São Paulo, São Paulo. 1999. 26
- Morlini, I. Multivariate outliers detection with Kohonen networks: an useful tool for routine exploration of large data sets. In: *New Techniques and Technologies in Statistics*, 1998, Sorrento, Italy. **Proceedings...** Sorrento, Italy: NTTTS, 1998. p. 345–350. 43, 44
- Muñoz, A.; Muruzábal, J. Self-Organizing Maps for outlier detection. **Neurocomputing**, v. 18, n. 1-3, p. 33–60, January 1998. 43
- Musser, D. R.; Saini, A. **STL tutorial and reference guide**. Boston, MA: Addison-Wesley, 1996. 560 p. 65
- Nobrega, R. **Análise espacial “knowledge-driven” e “data-driven”**: o uso das lógicas booleana, fuzzy e redes neurais para geração de mapas de favorabilidade mineral na região Centro-Oeste da Bahia. Campinas, SP, 2001. 153 p. Dissertação (Mestrado em Geologia) - Universidade Estadual de Campinas, Campinas. 2001. 26
- Oja, M.; Kaski, S.; Kohonen, T. Bibliography of Self-Organizing Map (SOM) Papers: 1998-2001 Addendum. **Neural Computing Surveys**, v. 3, p. 1–156, 2003. 27
- Openshaw, S.; Abrahart, S. **Geocomputation**. London: Taylor & Francis, 2000. 436 p. 25
- Openshaw, S.; Blake, M.; Wymer, C. **Using neurocomputing methods to classify britain’s residential areas**. Leeds, 1994. (Working paper 94/17). 59
- Openshaw, S.; Openshaw, C. **Artificial intelligence in geography**. Chichester: John Wiley and Sons, 1997. 348 p. 25, 26
- Openshaw, S.; Turton, I. A parallel Kohonen algorithm for the classification of large spatial datasets. **Computers & Geosciences**, v. 22, n. 9, p. 1019–1026, November 1996. 26, 27, 36, 58, 59
- Park, Y.-S.; Cereghino, R.; Compin, A.; Lek, S. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. **Ecological Modelling**, v. 160, n. 3, p. 265–280, February 2003. 37, 50

- Purvis, M.; Zhou, Q.; Cranefield, S.; Ward, R.; Raykov, R.; Jessberger, D. Spatial information modelling and analysis in a distributed environment. **Ecological Modelling & Software**, v. 16, n. 5, p. 439–445, July 2001. 26
- Rosa, D. d. I.; Mayol, F.; Moreno, J. A.; Bonsón, T.; Lozano, S. An expert system/neural network model (ImpelERO) for evaluating agricultural soil erosion in Andalusia region, southern Spain. **Agriculture, Ecosystems and Environment**, v. 13, n. 3, p. 211–226, May 1999. 26
- Rumelhart, D.; Hinton, G.; Williams, R. Learning internal representations by error propagation. In: Rumelhart, D.; McClelland, J. ed. **Parallel distributed processing: explorations in the microstructure of cognition**. Cambridge, MA: MIT PRESS, 1986. v. 1: Foundations. 32
- Silva, N.; Rosa, A. Estimative of SOM learning parameters using genetic algorithms. In: World Multi-Conference on Systemics, Cybernetics and Informatics, 6., 2002, Orlando. **Proceedings ...** Orlando: SCI/ISAS, 2002. p. 14–19. 36
- Sposati, A. O. **Cidade em pedaços**. São Paulo: Brasiliense, 2001. 173 p. 81
- Stroustrup, B. **A Linguagem de programação C++**. Porto Alegre: Bookman, 2000. 823 p. 65
- Takatsuka, M. An application of the self-organizing map and interactive 3-D visualization to geospatial data. In: International Conference on GeoComputation, 6., 2001, Brisbane. **Proceedings ...** Brisbane, Australia: University of Queensland, 2001. 59, 60
- Tso, B.; Mather, P. M. **Classification methods for remotely sensed data**. London: Taylor & Francis, 2001. 272 p. 26
- Ultsch, A. Knowledge extraction from self-organizing neural networks. In: Opitz, O. ed. **Information and Classification**. Berlin: Springer, 1993. p. 301–306. 38, 40, 41, 44
- . Data mining and knowledge discovery with Emergent Self-Organizing Feature Maps for multivariate time series. In: Oja, E.; Kaski, S. ed. **Kohonen Maps**. Amsterdam: Elsevier, 1999. Cap. 3, p. 33–46. 43
- Vesanto, J. **Data mining techniques based on the Self-Organizing Map**. 1997. 245 p. Dissertação (Mestrado em Computação) - Helsinki University of Technology, Helsinki. 1997. 41

———. SOM based data visualization methods. **Intelligent Data Analysis**, v. 3, n. 2, p. 111–126, August 1999. 41

Vesanto, J.; Ahola, J. Hunting for correlations in data using the Self-Organizing Map. In: International ICSC Congress on Computational Intelligence Methods and Applications (CIMA '99), 1999, Rochester, NY. **Proceedings ...** Rochester, NY: ICSC Academic Press, 1999. p. 279–285. 35, 36, 41

Vesanto, J.; Alhoniemi, E. Clustering of the Self-Organizing Map. **IEEE Transactions on Neural Networks**, v. 11, n. 3, p. 586–600, May 2000. 50

Vesanto, J.; Himberg, J.; Alhoniemi, E.; Parhankangas, J. Self-Organizing Map in matlab: the SOM toolbox. In: Matlab DSP Conference, 1999, Espoo, Finland. **Proceedings ...** Espoo, Finland: Comsol Oy, 1999. p. 35–40. 65

Villmann, T.; Merényi, E.; Hammer, B. Neural maps in remote sensing image analysis. **Neural Networks**, v. 16, n. 3-4, p. 389–403, April-May 2003. 26, 58

Winter, K.; Hewitson, B. Self organizing maps - applications to census data. In: Hewitson, B.; Crane, R. ed. **Neural nets: applications in geography**. Kluwer, 1994. Cap. 4, p. 45–57. 27, 59, 60

Wu, S.; Chow, T. W. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. **Pattern Recognition**, v. 37, n. 2, p. 175–188, February 2004. 50, 52

Zell, A.; Mache, N.; Huebner, R.; Schmalzl, M.; Sommer, T.; Korb, T. **SNNS: Stuttgart Neural Network Simulator**. Stuttgart, 1992. 65

APÊNDICE A

INICIAÇÃO LINEAR DOS VETORES DE CÓDIGO DO SOM

Segundo Kohonen (2001) a iniciação linear dos vetores de código é mais recomendada que a iniciação randômica, devido ao fato de que a iniciação linear dispensa a fase de treinamento para ajuste inicial dos vetores de código.

Seja Ξ o conjunto dos vetores de entrada $x_k, k = 1, \dots, n$.

Calcula-se a matriz de correlação de Ξ, A' . Seja y os auto-vetores de A' e e os seus auto-valores, tem-se

$$A'y = ey \quad (\text{A.1})$$

Pegar os auto-vetores ortogonais a y_i correspondentes aos dois maiores auto-valores $e_i, i = 1, \dots, d'$, sendo d' a dimensão do Mapa Auto-Organizável.

A partir destes auto-vetores gera-se um sub-espço linear com centróide igual à média \bar{x} do conjunto Ξ . Portanto, sendo $w_{ij}(0)$ o vetor de código inicial do neurônio localizado nas posições i, j , para Mapas com $d' = 2$ tem-se

$$w_{ij}(0) = s \left[\left(i - \frac{N}{2} \right) y_1 + \left(j - \frac{M}{2} \right) y_2 \right] \quad (\text{A.2})$$

onde N, M são as dimensões da rede bidimensional e s uma constante selecionada de forma adequada.