MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
CURSO DE MESTRADO EM SISTEMAS E COMPUTAÇÃO

JOÃO CARLOS SANTIAGO FILHO

ESTUDO E APLICAÇÃO DE MAPAS AUTO-ORGANIZÁVEIS NA IDENTIFICAÇÃO E DELIMITAÇÃO DE REGIÕES DE INFLUÊNCIA DAS CIDADES BRASILEIRAS

Rio de Janeiro
2010

INSTITUTO MILITAR DE ENGENHARIA

JOÃO CARLOS SANTIAGO FILHO

ESTUDO E APLICAÇÃO DE MAPAS AUTO-ORGANIZÁVEIS NA IDENTIFICAÇÃO E DELIMITAÇÃO DE REGIÕES DE INFLUÊNCIA DAS CIDADES BRASILEIRAS

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Mestre em Sistemas e Computação.

Orientador: Prof. Ricardo Choren Noya D.Sc.
Co-orientador: Prof. Marcello Goulart Teixeira D.Sc.

Rio de Janeiro
2010

INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80-Praia Vermelha
Rio de Janeiro-RJ CEP 22290-270

004.62 Santiago, J. C.
S235

Estudo e Aplicação de Mapas Auto-Organizáveis na Identificação e Delimitação de Regiões de Influência das Cidades Brasileiras/ João Carlos Santiago Filho.
– Rio de Janeiro: Instituto Militar de Engenharia, 2010.
107 p.: il., tab.

Dissertação (mestrado) – Instituto Militar de Engenharia – Rio de Janeiro, 2010.

1. Redes. 2. Mapas, estudo e aplicação. I. Ttulo. II. Instituto Militar de Engenharia.

CDD 004.62

INSTITUTO MILITAR DE ENGENHARIA

JOÃO CARLOS SANTIAGO FILHO

ESTUDO E APLICAÇÃO DE MAPAS AUTO-ORGANIZÁVEIS NA IDENTIFICAÇÃO E DELIMITAÇÃO DE REGIÕES DE INFLUÊNCIA DAS CIDADES BRASILEIRAS

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Mestre em Sistemas e Computação.

Orientador: Prof. Ricardo Choren Noya D.Sc.

Co-orientador: Prof. Marcello Goulart Teixeira D.Sc.

Aprovada em 1 de setembro de 2010 pela seguinte Banca Examinadora:

Prof. Ricardo Choren Noya D.Sc. do IME - Presidente

Prof. Marcello Goulart Teixeira D.Sc. da UFRJ

Prof. Prof. Paulo Fernando Ferreira Rosa Ph.D, do IME

Profa. Marley Maria Bernardes Rebuzzi Vellasco - DSc. da PUC-RJ

Rio de Janeiro
2010

À Carla, minha parceira querida, sem a qual eu aqui, não teria sequer sonhado em chegar.

AGRADECIMENTOS

Aos meus pais, pela vida e pelo esforço em favor de minha educação, particularmente a minha mãe pela certeza de que existe um Deus.

Aos meus filhos, Ilana, Mikael, Leonardo e Marcelo, grandes motivos para eu ter desejado chegar até aqui.

Aos meus orientadores, Prof. Ricardo Choren Noya e Prof. Marcello Goulart Teixeira, pela confiança, amizade, apoio e orientações decisivas para que esse trabalho migrasse solidamente do plano das ideias para o mundo concreto.

Ao Profs. Antônio Rosa e Luiz Roberto Martins Bastos da Universidade Estácio de Sá, pelo incentivo e chancela à minha candidatura à uma vaga no curso de mestrado do IME.

Ao amigo Maurício Costa Reis pelo incentivo e por me fazer acreditar na possibilidade do mestrado no IME.

Aos amigos Anna Carolina Riqueza e Gustavo Ciarelli pelas palavras de apoio e constante preocupação com os destinos desse trabalho.

Aos amigos Azenclever Sancler, Eduardo Tadeu e Nilthon Candeia para os quais distância é uma palavra que não existe.

Ao Prof. Paulo Rosa do IME-RJ pela orientações fortes quanto ao único caminho possível, o estudo.

A Profa. Marley Vellasco da PUC-Rio pelo exemplo de dedicação ao ato de ensinar.

Aos meus coordenadores na Agência Nacional de Cinema - ANCINE, com os quais pude contar com apoio e compreensão ao longo do curso de mestrado.

Por fim, a todos os professores e funcionários do Departamento de Engenharia de Sistemas (SE/8) do Instituto Militar de Engenharia.

João Carlos Santiago Filho

'One person with passion is better than forty people merely interested.'

E. M. Forster

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SÍMBOLOS

ABREVIATURAS

ANAC	-	*Agência Nacional de Aviação Civil*
ANTT	-	*Agência Nacional de Transportes Terrestres*
BMU	-	*Best Match Unit*
CRISP-DM	-	*Crisp Data Mining*
DB-index	-	*Davies-Bouldin Index*
D-Matrix	-	*Matriz de Distâncias*
EQ	-	*Erro de Quantização*
ESOM	-	*Emergent Self Organizing Map*
ET	-	*Erro Topográfico*
IBGE	-	*Instituto Brasileiro de Geografia e Estatística*
IR	-	*Intensidade de Relacionamento*
KMCA	-	*Kohonen Multiple Correspondence Analysis*
LMS	-	*Least Mean Square*
MLPs	-	*Multilayer Perceptrons*
MPOG	-	*Ministério do Planejamento, Orçamento e Gestão*
PCA	-	*Principal Component Analysis*
REGIC	-	*Região de Influência das Cidades*
RNAs	-	*Redes Neurais Artificiais*
SOM	-	*Self Organizing Map*
SSE	-	*Sum of Square Error*
U-Matrix	-	*Unified Distance Matrix*

RESUMO

O estudo de redes complexas, como as urbanas e sociais, oferece subsídios para a compreensão da organização espacial da sociedade e ações de planejamento estratégico de empresas e de governos.

Nesta dissertação propõe-se uma metodologia que, diferentemente de abordagens como a análise de centralidade em redes segundo aspectos meramente estruturais, ou da extração de grupos em grafos em função da similaridade entre os vértices, concilia e estende o conceito de associação dominante entre cidades com as reconhecidas propriedades de projeção e quantização vetorial dos mapas auto-organizáveis ou SOM.

O modelo resultante é composto por uma topologia hierárquica e estruturalmente simplificada, mas capaz de simultaneamente representar adequadamente o mecanismo de associação observado em redes urbanas, oferecer suporte à determinação de centros locais e respectivas regiões de influência e complementarmente capturar padrões de grupos de cidades similares.

Apresenta-se uma opção de representação visual dos resultados obtidos e é investigado o potencial de captura de grupos e de localidades centrais, exclusivamente por meio da análise das rotas que ligam os centros urbanos que compõem a rede.

A metodologia aqui proposta foi aplicada ao conjunto das 5274 cidades brasileiras descritas por seus atributos sócio-econômicos e respectivos relacionamentos representados por dados oficiais do governo brasileiro sobre o transito inter-estadual de passageiros terrestres e aéreos. O resultados obtidos pela presente proposta quando comparados com àqueles resultantes por técnicas tradicionais demonstram um grande potencial desta abordagem, particularmente a análise de padrões de grupos determinados por conjuntos de conexões dominantes entre as cidades. O método, apesar de não testado explicitamente em outras redes que não as urbanas, possui propriedades que o torna plenamente adaptável para a análise de outros modelos de redes hierárquicas como as redes sociais.

ABSTRACT

The study of complex networks, such as urban and social, offers subsidies for understanding the spatial organization of society and,for defining strategic planning activities of companies and governments.

In this dissertation we propose a methodology that, unlike approaches such as network's centrality analysis, or the extraction of groups on graphs according to the similarity between the vertices, reconciles and extends the concept of dominant association among cities and the well known properties of projection and vector quantization of self-organizing maps, or SOM.

The resulting model consists of a structurally simplified hierarchical network, but able to simultaneously represent the mechanism of association observed in urban networks, support the establishment of local centers and their regions of influence and capture complementary patterns of clusters of similar cities.

It also presented a choice of visual representation of the results ans the results of a brief investigation on potential for capturing groups and central locations exclusively through the analysis of routes linking the urban centers that comprise the network.

The methodology proposed was applied to all the 5274 Brazilian cities described by their attributes and their socio-economic relationships represented by official data from the Brazilian government on the interstate transit of land and air passengers. The results obtained by this proposal when compared with those derived by traditional techniques show great potential of this approach, particularly the analysis of patterns of certain groups by dominant sets of connections between cities. The method, although not explicitly tested in other networks than urban ones, has properties that make it fully adaptable to the analysis of other models of hierarchical networks as the social ones.

1 INTRODUÇÃO

Redes complexas constituem um formalismo eficaz para muitos sistemas do mundo real como as redes sociais e urbanas (PIZZUTI, 2009). Seu estudo oferece não só subsídios para a compreensão das diferentes formas de organização espacial da sociedade, como também suporte para ações de planejamento estratégico de empresas e de governos (IBGE, 2008).

O modelo corrente de redes complexas aplicadas a redes urbanas e sociais é baseado na teoria dos grafos, onde os nós representam as entidades participantes e os vértices as relações existentes. Os primeiros trabalhos sobre essas aplicações remontam a década de 50 (BAVELAS, 1948), tendo como principais focos de investigação a detecção de comunidades e sua estrutura, onde o conceito de centralidade, sua mensuração e o alcance de sua influência são dependentes de aspectos posicionais em relação à topologia dessas redes (MUSIAL, KAZIENKO e BRÓDKA, 2009). Nesses contextos, a topologia da rede é pré-existente e analisa-se a performance e papel relativo dos nós participantes ao longo dessa estrutura.

As redes urbanas, formadas pelo conjunto de localidades geográficas, seus habitantes e interações delas decorrentes, acrescentam ao conceito de centralidade estrutural a noção de influência decorrente da ação de grandes metropóles como irradiadoras de fluxos de bens e de serviços sobre centros menores e dependentes destes grandes centros. Aqui, o número de habitantes, as dimensões do seu entorno e a distância geográfica são os fatores determinantes no desenvolvimento de uma hierarquia dos espaços urbanos (ULLMAN, 1941). Grandes metrópoles são nós altamente especializados e diversificados em torno dois quais centros menores se agrupam em função de aspectos econômicos, geográficos e políticos.

Outra abordagem possível para análise de redes urbanas e sociais deriva dos registros de fluxos de bens, serviços e informações entre os participantes dessas estruturas. A rede é n-dimensional, possui tantas expressões de conectividade quantas as dimensões sobre as quais é analisada. A rede, nessa visão, possui uma topologia virtual derivada dos fluxos e independente de ligações físicas (RABINO e OCCELLI, 2007). Os fluxos pré-existem e são estes que determinam as topologias que podem ou não ter associação com as ligações físicas, como as formadas pelo fluxos de pessoas entre duas cidades por meio das rotas

que as ligam.

A captura contínua de informação nos mais variados espectros da experiência humana, associada à explosão dos canais de interatividade social, acrescenta não só um número cada vez maior de dimensões sobre as quais uma rede pode se desenvolver, como coloca à nossa disposição um volume crescente de informações, aproximando o problema de estudos de centralidade da esfera de interesse e alcance de soluções da área de Mineração de Dados (NISBET, IV e MINER, 2009). A Mineração de Dados engloba em sua área de atuação um conjunto de práticas científicas e heurísticas que procura extrair de bases de dados, padrões desconhecidos a priori, e produzir conhecimento útil e relevante. Aprendizagem de Máquina é a ciência que oferece o suporte teórico e instrumental para diversas estratégias de análise adotadas pela Mineração de Dados. Na Mineração de Dados, técnicas tradicionais de modelagem analítica e análise discriminante são complementadas com metodologias nas quais os padrões descobertos automaticamente podem ser identificados, validados e adotados em modelos preditivos (WITTEN e FRANK, 2005).

Entre as diversas metodologias para captura automática de padrões, as redes neurais artificiais (RNAs) se destacam por sua reconhecida capacidade de aprendizado, armazenamento e representação do conhecimento adquirido, mesmo na presença de distribuições complexas e não-lineares (HAYKIN, 2001). Em processos de aprendizado não-supervisionado empregando RNAs, os mapas auto-organizáveis, redes neurais que aprendem por processos competitivos, oferecem a solução para inúmeros casos de aplicações bem sucedidas, entre elas, o reconhecimento de padrões e a análise de grupos (ULTSCH e MÖRCHEN, 2005).

Essa dissertação se motiva pela constatação de que é possível extrair topologias multidimensionais a partir de informações de fluxos de informações entre nós da rede e, simultaneamente, tratar a noção de similaridade na formação de agrupamentos dentro da rede extraída. O objetivo principal é oferecer subsídios para a compreensão dos mecanismos subjacentes e a construção de um modelo preditivo.

1.1 DESCRIÇÃO DO PROBLEMA

A determinação de regiões de influência é de importância crítica para o entendimento e a mensuração do alcance de fenômenos que afetem estruturas compostas de elementos que se interrelacionam. Em problemas dessa natureza, é buscada a compreensão dos mecanismos que regulam causas e efeitos, se possível a sua modelagem e reprodutibilidade.

No caso de redes urbanas, a observação do mundo real e diversos modelos teóricos (ULLMAN, 1941) preconizam que essas estruturas são intrinsecamente geradoras de agrupamentos que giram em torno de localidades centrais, cada uma por sua vez capaz de se tornar participante de grupos hierarquicamente superiores. Nessas redes, a densidade populacional tem papel decisivo e é fato gerador de fluxos de bens, serviços e informações através de relações complexas afetadas por fatores diversos.

As principais abordagens existentes para a descrição do papel desempenhado por localidades centrais em redes complexas é baseada em suas características estruturais e posicionais e buscam mensurar a influência de um nó sobre outros em função do papel deste como parte de um caminho ao longo da rede (FREEMAN, BORGATTU e WHITE, 1991). Estratégias de análise topológica têm sido aplicadas em problemas de redes sociais e urbanas. Nesse contexto, não são levadas em consideração as características intrínsecas de cada nó, e quando feitas, essas descrições usualmente se realizam por meio de relações lineares. Aqui, grupos são formados por nós próximos enquanto menos distantes, não por suas similaridades ou dissimilaridades.

Outras modelagens, como as dos fluxos prioritários, ultrapassam as limitações de considerações espaciais e procuram explicações para os mecanismos geradores de estruturas ao longo de redes urbanas, por meio da análise dos fluxos direcionados entre pares de nós da rede, impondo uma definição estrutural em função destes fluxos (RABINO e OCCELLI, 2007). Nestes casos, novamente, as características e similaridades entre os nós da rede são desconsideradas, oferecendo assim uma compreensão apenas parcial dos fenômenos que governam os comportamentos dessas estruturas.

Ambas abordagens oferecem pouco ou nenhum suporte para a descrição de comportamentos não-lineares e não têm por objetivo a definição de modelos preditivos e a captura de grupos por similaridade.

Dentro do escopo da revisão bibliográfica empreendida, identifica-se a ausência de metodologias que busquem, de forma combinada, capturar topologias de rede derivadas de fluxos multidimensionais levando em consideração a presença de grupos nessas estruturas. Esses grupos devem ser formados não por proximidade estrutural, mas sim similaridade característica.

1.2 OBJETIVOS DA DISSERTAÇÃO

O que se propõe no presente trabalho é a construção de um modelo simplificado e ágil, capaz de oferecer suporte à determinação de centros locais e respectivas regiões de influência vistos tanto como regiões nodais determinadas pelo método dos fluxos dominantes quanto grupos de nós que compartilham similaridades em redes urbanas e sociais.

O objetivo final é a construção de um modelo preditivo que permita estabelecer sobre um conjunto de atributos estruturais, característicos e relacionais, a topologia de uma rede hierárquica urbana, definir a extensão dos relacionamentos entre nós e identificar padrões de grupos que compartilhem similaridades.

Visando a reprodutibilidade, acompanhamento temporal e redução da subjetividade da presente proposta, foram definidas as seguintes estratégias:

- empregar exclusivamente dados disponíveis publicamente e passíveis de atualização periódica,
- utilizar preferencialmente dados de origem primária, ou seja, aqueles livres de pós-processamento e inferências por especialistas de domínio.

1.3 CONTRIBUIÇÕES

As principais contribuições desse trabalho são:

- Estabelecimento de parâmetros e de um processo que oriente a implementação de uma ferramenta para a extração de redes hierárquicas, orientada por grupos, bem como a geração de mapas de regiões de influência, levando-se em consideração os fluxos de informações entre nós da rede.
- Verificação da aplicabilidade de mapas auto-organizáveis em problemas cuja natureza relacional e distribuição estatística se assemelhem à existente no universo da rede urbana brasileira.
- Mensuração, no escopo do problema aqui posto, do potencial de indicadores selecionados para a identificação de partições *naturais* obtidas pelo modelo.
- Verificação do alcance representativo de mapas georreferenciados como suporte visual de resultados do modelo.

- Subsídios para trabalhos futuros baseados na análise de similaridade de rotas empregando mapas auto-organizáveis.

1.4 ORGANIZAÇÃO

Essa dissertação foi organizada segundo a estrutura dos estudos, experimentações e análises que nortearem a busca pela consecução dos objetivos traçados e se inicia no **Capítulo 2 - Fundamentação Teórica**, onde são apresentados de forma concisa e objetiva os fundamentos sob os quais repousa o presente trabalho, particularmente aspectos de centralidade estrutural e relacional, análise de grupos e mapas auto-organizáveis.

O **Capítulo 3 - Trabalhos Relacionados** percorre alguns dos trabalhos existentes cujos objetivos e estratégias guardam estreita relação com os métodos e soluções adotadas no neste trabalho.

O **Capítulo 4 - Um Modelo para Representação de Regiões de Influência** apresenta a metodologia proposta nessa dissertação para a reconstrução de redes hierárquicas, o emprego do SOM na determinação de agrupamentos nessas mesmas redes e o uso de indicadores para validação do resultados resultantes do emprego do modelo.

O **Capítulo 5 - Resultados** desenvolve um estudo de caso real empregando a metodologia proposta no capítulo anterior, empregando dados oriundos do estudo do IBGE, *Regiões de Influência das Cidades*, fluxos terrestres provenientes da ANTT e dados sobre trânsito aéreo de carga e passageiros compilados pela ANAC. Diversos experimentos são relatados e os respectivos resultados apresentados e discutidos.

A dissertação finaliza com o **Capítulo 6 - Considerações Finais** onde é feita um revisão dos resultados pretendidos e alcançados, e comenta-se sobre possíveis direcionamentos para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 CENTRALIDADE E REGIÕES DE INFLUÊNCIA

As ciências sociais contribuíram de forma definitiva para a consolidação do conceito de centralidade e sua importância para as relações humanas a partir de BAVELAS (1948), que ao estudar padrões de comunicação entre superiores e subordinados em ambientes corporativos, desenvolveu um modelo de representação topológica para as estruturas observadas, observando que a posição relativa de indivíduos centrais varia em função dos arranjos estruturais.

2.1.1 CENTRALIDADE ESTRUTURAL

Os principais conceitos, em torno dos quais propostas de mensuração da centralidade foram inicialmente construídas, se baseiam nos aspectos estruturais de redes sociais à luz da Teoria dos Grafos, a saber (FREEMAN, 1979):

- ser central enquanto próximo de outros, ou centralidade por proximidade (*Closeness*) e
- ser central enquanto entre outros, ou centralidade por intermediação (*Betweenness*).

Trabalhos recentes incorporaram outras visões às anteriores (PORTA, CRUCITTI e LATORA, 2006):

- ser central enquanto diretamente conectado a outros, ou centralidade por eficiência (*Eficiency Centrality*),
- ser central enquanto mais próximo de uma reta é o caminho que o liga a outros (*Straightness Centrality*) e
- ser central enquanto crítico para todos outros, ou centralidade da informação (*Information Centrality*).

Entre as medidas de centralidade revistas por Freeman está a introduzida por NIEMINEN (1973), onde a centralidade é simplesmente o grau de um nó p_k, ou número de

arestas que se conectam a ele,

$$C_D\left(p_k\right) = \sum_{i=1}^{n} a\left(p_i, p_k\right) \tag{2.1}$$

onde $a\left(p_i, p_k\right)$ é igual 1 se há uma conexão entre os pontos p_i e p_k e 0 caso contrário.

Normalizando C_D em relação ao maior grau possível, $n-1$, tem-se:

$$C'_D\left(p_k\right) = \frac{C_D}{n-1} \tag{2.2}$$

Proximidade ou *Closeness* leva em conta a soma das distâncias, medida pelo número de arestas ao longo de todos os caminhos que ligam dois nós, podendo ser definida na forma:

$$C_c\left(i\right) = \left[\sum_{j}^{n} d\left(i,j\right)\right]^{-1} \tag{2.3}$$

onde $d\left(i,j\right)$ é o número de arestas no caminho entre i e j.

Betweeness é uma expressão para a capacidade de intermediação de um nó, levando em conta a quantidade de vezes em que este se coloca no caminho entre dois outros nós. Expressa o potencial de influência ao se fazer parte do processo de comunicação entre nós da rede. Na figura 2.1, três caminhos ligam p_i e p_j, sendo que dois passam por p_k.

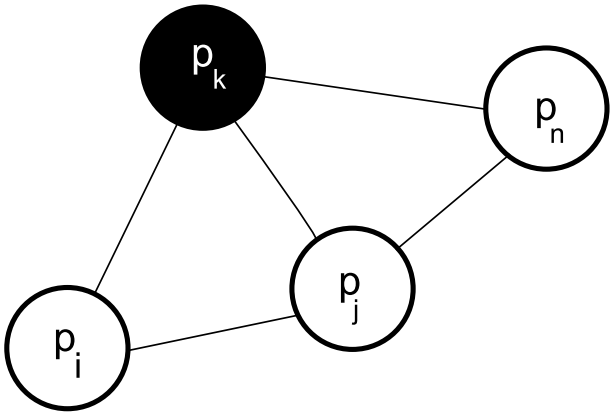


FIG. 2.1: Um grafo com quatro vértices e cinco arestas.

Sejam dois pontos quaisquer p_i e p_j ligados entre si por meio de g_{ij} caminhos, e que p_k seja um ponto situado em $g_{ij}\left(p_k\right)$ caminhos. A centralidade de p_k, em termos da capacidade de intermediação de um nó ou *Betweeness* em relação aos nós i e j, pode ser definida por

$$b_{ij}(p_k) = \frac{g_{ij}(p_k)}{g_{ij}} \tag{2.4}$$

A centralidade global de p_k em relação a todos os pares possíveis de pontos não-ordenados tais $i \neq j \neq k$ para um grafo com n pontos é:

$$C_B\left(p_k\right) = \sum_{i<j}^{n} \sum_{j}^{n} b_{ij}(p_k) \tag{2.5}$$

Freeman também demonstrou que o maior valor possível para $C_B\left(p_k\right)$ em um grafo qualquer de n pontos é obtido por um ponto central em configuração de estrela com o valor de $\frac{n^2-3n+2}{2}$. Logo a centralidade relativa de qualquer ponto pode ser definida pela razão entre a centralidade do ponto e a máxima possível como em:

$$C'_B\left(p_k\right) = \frac{2C_B\left(p_k\right)}{n^2 - 3n + 2} \tag{2.6}$$

A centralidade por eficiência é construída a partir do conceito de que eficiência na comunicação entre dois vértices i e j é inversamente proporcional ao menor caminho d_{ij} sendo comparada com a distância física $\widetilde{d_{ij}}$ entre os nós (LATOR e MARCHIORI, 2001). A rede, tratada como um grafo não-orientado e com pesos é então descrita por 2 matrizes, uma de adjacências e outra de distâncias euclidianas, onde centralidade por eficiência de um nó i qualquer é definida por:

$$C_i^E = \left(\sum_{j \in n; j \neq i} \frac{1}{d_{ij}} \right) / \left(\sum_{j \in n; j \neq i} \frac{1}{\widetilde{d_{ij}}} \right) \tag{2.7}$$

Centralidade por colinearidade (VRAGOVÌC, LOUIS, DÃAZ-GUILERA, CHICO e MARIA, 2005) é uma expressão do desvio do caminho entre dois nós em relação à reta que os une:

$$C_i^L = \left(\sum_{j \in n; j \neq i} \frac{\widetilde{d_{ij}}}{d_{ij}} \right) / \left(n - 1\right) \tag{2.8}$$

Já no campo da geografia econômica, a Teoria das Localidades Centrais de Walter Christaller ocupa lugar de destaque como referência primária a oferecer uma explicação para o desenvolvimento dos grandes centros urbanos e da região sob sua influência (CORRÊA (2010); EATON e LIPSEY (1982); BERRY e GARRISON (1958)). Nela, o número de habitantes e a distância geográfica ao longo de uma geometria específica são

os fatores determinantes no desenvolvimento de uma hierarquia dos espaços urbanos no entorno dos grandes centros, sendo proposta uma classificação para as localidades em função do seu tamanho, especialização e capacidade produtiva (ULLMAN, 1941).

Tendo por base essa mesma teoria e conceitos oriundos da Teoria dos Grafos, pode-se considerar a articulação entre os locais centrais como um tipo especial de rede, a rede urbana. A rede urbana seria uma rede hierárquica, tendo as cidades como vértices e as arestas ou conexões os diversos fluxos entre esses nós. Nessas redes, nós diferenciados ocupam lugar de destaque como destinos para nós inferiores e os fluxos mais intensos ocorrem entre nós de hierarquia mais alta (CORRÊA, 2010).

2.1.2 CENTRALIDADE RELACIONAL

Os conceitos de centralidade oriundos de estudos de sociologia estrutural encontraram a sua contraparte em aplicações de redes urbanas em propostas diversas, entre as quais aquelas que oferecem suporte à análise de estruturas espaciais baseadas em medidas de fluxos dos entes envolvidos sob suporte de estruturas de grafos. Nesses modelos, a referência espacial, parte integrante da descrição estrutural do sistema, é relevada em favor da análise topológica resultante dos fluxos de pessoas, informações, bens e serviços já não mais restritos exclusivamente às rotas físicas tradicionais. O método dos Fluxos Dominantes (NYSTUEN e DACEY, 1961) oferece um suporte para extração automática de estruturas hierárquicas de redes urbanas a partir de algumas hipóteses sobre a subordinação de centros urbanos em função dos fluxos de comunicação, bens e serviços entre metrópoles e cidades menores. O método se sustenta sobre hipóteses de geografia econômica (BERRY e GARRISON, 1958) derivadas dos conceitos propostos por Christaller sobre fluxos de bens e serviços no entorno dos grandes centros e a distribuição de novas localidades nessa região, ou *hinterland*. A partir da relação de subordinação capturada pela comparação da intensidade dos fluxos registrados em uma matrix de adjacências, o método dos fluxos dominantes determina a região sob influência do centro de maior importância, então representado em um nível hierárquico superior.

O algoritmo para determinação de localidades centrais pode ser descrito nos seguintes passos:

1 para cada coluna da matriz de adjacências, selecione a célula onde ocorre o fluxo mais intenso;

2 determine, segundo algum critério externo (ex. população) ou interno (soma total de fluxos), qual cidade é subordinada ou se são independentes. Caso o fluxo mais intenso ocorra entre uma cidade menor e outra maior, a primeira é identificada como subordinada à segunda, caso contrário é classificada como independente.

O algoritmo para a determinação da conexão dominante pode ser descrito nos seguintes passos:

Para cada coluna da matriz de adjacências,

a) Selecione a linha onde ocorre o maior valor de fluxo e identifique o nó associado a coluna correspondente como a conexão principal do nó corrente;

b) Compare os fluxos totais do nó corrente e de sua conexão principal;

c) Se o fluxo total do nó corrente é menor do que o fluxo total de sua conexão principal, marque-o como subordinado à conexão principal, caso contrário classifique-o como independente.

No exemplo demonstrado na tabela 2.1 os fluxos principais, neste caso bidirecionais, das cidades C1, C3, C4 e C5 ocorrem com C2 e o fluxo total de C2 é superior ao fluxo total de cada uma dessas cidades, que passam a ser consideradas subordinadas à esta última.

Um asterisco indica para cada coluna, a conexão dominante, caso essa exista. Na figura 2.2, é apresentado o grafo resultante da aplicação do método.

TAB. 2.1: Matriz de adjacências para um fluxo de passageiros de ônibus entre cidades

		Cidades destino					
		C1	C2	C3	C4	C5	C6
Cidades origem	C1	0	11250	2250	3000	4200	300
	C2	*1035	0	6750	*7500	1700	800
	C3	750	7650	0	1800	*****6000	0
	C4	285	10050	2100	0	4500	1050
	C5	105	6000	*7200	3900	0	*****2350
	C6	150	900	150	150	1500	0
Fluxo total		15150	35850	18450	16350	24900	4500

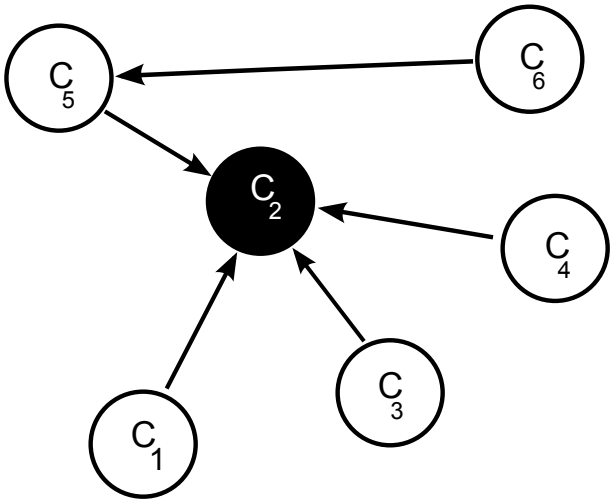


FIG. 2.2: Hierarquia de nós segundo os fluxos dominantes entre um conjunto de cidades.

As seguintes propriedades podem ser elencadas a partir da estrutura decorrente:

a) um nó é dito independente se o seu maior fluxo é dirigido a um nó "menor".

b) a relação de hierarquia é transitiva. Se o nó a é subordinado a b e este se subordina a c, implica em que a é subordinado a c.

c) um nó não pode ser subordinado a qualquer dos seus subordinados. A estrutura é acíclica.

Observa-se também que o método particiona o conjunto de nós da rede em grupos de estruturas nodais, cada qual contendo uma única localidade central (figura 2.3).

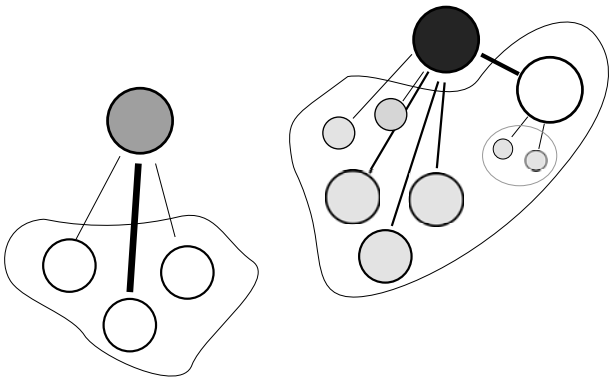


FIG. 2.3: Agrupamentos de localidades centrais

Uma extensão natural para o método do fluxos dominantes é a consideração de combinações lineares, ou não, de conjuntos de fluxos n-dimensionais que ocorram ao longo da

rede, apreciados segundo ponderações de pesos ou preferências. Na sua expressão linear, essa intensidade de relacionamento multi-modal pode descrita na forma:

$$IR_{ij} = \sum_{k=1}^{m} fluxo_{k_{ij}} * w_{k_{ij}} \tag{2.9}$$

onde IR é a *Intensidade de Relacionamento*(IR) para o par de nós ij, resultante do somatório de m fluxos, cada qual ponderado por um peso w_{ij}.

Uma questão que pode ser levantada nesse momento é que nós de uma rede urbana ou social, até agora representados por uma estrutura pontual, são na realidade objetos ou entidades n-dimensionais dotados de características que os descrevem em um determinado contexto de análise de domínio, e que a sua própria centralidade é descrita, ou melhor, é função dessa descrição.

No caso de redes urbanas fica evidente a relação entre os fluxos observados e o tamanho das populações envolvidas. Em redes sociais, essa obviedade pode advir de um fator como a escolaridade ou renda, mas outras relações subjacentes ocorrem simultaneamente nos mais variados planos de análise, tão mais complexas de detectar quanto maior for o número de atributos envolvidos. Sob nosso ponto de vista, a busca de agrupamentos formados pelas estruturas nodais extraídas pelo método dos fluxos dominantes é uma das estratégias capazes de revelar esses padrões (figura 2.3).

2.2 ANÁLISE DE AGRUPAMENTOS

O processo de classificação é uma atividade fundamental de prática científica na medida em que os sistemas classificatórios formam a estrutura de conceitos sobre a qual se apoiam as teorias dentro de uma ciência. A Análise de Agrupamentos engloba um conjunto de práticas envolvendo medidas de similaridade na obtenção de grupos que podem ser usados a fim de produzir uma classificação, geração e testes de hipóteses (ALDENDERFER e BLASHFIELD, 1984).

Diferentemente de áreas como reconhecimento de padrões, análises estatística e discriminante, as quais procuram identificar regras de classificação a partir de um conjunto previamente classificado, a análise de grupos objetiva a definição de um conjunto de classes ou grupos a partir dos dados. Esses grupos presumidamente refletem algum mecanismo emergente no domínio do qual as instâncias de dados são extraídas (WITTEN e FRANK, 2005).

Ao final de um processo de agrupamento (*clustering*) têm-se um conjunto de classes e uma lista de instâncias, além da sua relação de pertinência com essas classes, relação essa que pode ser rígida, sobreposta ou difusa. Na classificação rígida cada instância pode pertencer exclusivamente a uma única partição. Na sobreposta é permitido que um objeto pertença a mais de uma classe simultâneamente, enquanto na difusa cada objeto possui um grau de pertinência variável em relação à cada classe.

A análise de grupos tem sido objeto de estudo e interesse das mais variadas áreas de conhecimento, cada qual contribuindo com as respectivas visões conceituais e metodológicas, englobando, em seu conjunto, soluções formais e heurísticas, análise estatística e técnicas computacionais. A análise de grupos forma um universo vasto e complexo, sujeita ela mesma a um processo de classificação taxonômica, não existindo até hoje uma solução unificada que possa ser aplicada indistintamente a todos os problemas ou conjunto de dados.

A classificação proposta por JAIN, MURTY e FLYNN (1999) divide inicialmente os algoritmos usuais em hierárquicos e partitivos, os quais se subdividem depois em função de técnicas específicas.

Algoritmos hierárquicos tratam inicialmente instâncias individuais como grupos, a partir dos quais outros grupos em nível hierárquico superior são formados através da aglomeração das instâncias similares, e cujo resultado final é uma árvore. A definição do número de grupos é função da altura do corte nessa árvore, que resulta em tantas sub-árvores quantos grupos sejam identificados. No exemplo da figura 2.4 é possível a definição de partições contemplando de um a seis grupos dependendo da linha de corte adotada. No primeiro nível de corte, cada instância é um grupo, cortando-se a árvore no nível imediatamente acima, a mesma distribuição é então representada por 3 grupos, depois 2 e por fim como um único grupo.

Algoritmos partitivos partem do conjunto de instâncias à procura de partições ótimas segundo critérios de erro, distâncias inter e intra-grupos, densitométricos, entre outros. Entretanto, essa taxonomia não é capaz de representar, em sua plenitude, as várias dimensões sobre as quais é possível caracterizar as técnicas de *clustering*.

O processo envolvendo a análise de agrupamentos pode ser dividido em quatro etapas principais:

- seleção ou extração de atributos;

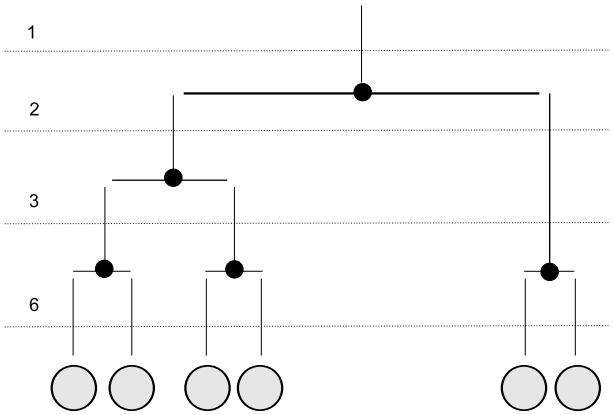


FIG. 2.4: Abordagem hierárquica de agrupamento.

- projeto ou seleção de algoritmo;
- obtenção dos grupos; e
- validação dos resultados.

A seleção de atributos, um campo de estudos em si, procura identificar entre todos os atributos registrados aqueles que melhor podem representar as propriedades estatísticas observadas sem uma perda significativa de informação, procurando reduzir a dimensionalidade e a complexidade do conjunto de padrões observados. Técnicas de seleção de atributos são capazes de oferecer subsídios na construção, classificação, validação e escolha de variáveis, assim como uma definição otimizada da função objetivo (GUYON e ELISSEEFF, 2003).

A extração de atributos visa a redução dimensional do problema por meio de transformações do espaço de entrada, de tal forma que um número reduzido de atributos possa explicar ou representar a maior parte da variância observada. A Análise dos Componentes Principais (*Principal Component Analysis - PCA*) (HOTELLING, 1933), aplicável em mapeamentos lineares, e a *Curvilinear Component Analysis - CCA* (DEMARTINES e HERAULT, 1997), baseada em mapas auto-organizáveis e projeções não-lineares são alguns dos exemplos frequentemente citados na literatura.

O projeto ou a seleção do algoritmo mais apropriado está intimamente associado ao espaço de entrada, o tamanho das amostras, dimensionalidade, características de atributos e a definição do conceito de proximidade ou similaridade que melhor seja capaz de evidenciar a emergência de padrões na forma de grupos.

A análise de grupos ao mesmo tempo que procura por estruturas e padrões de grupos em conjuntos de dados, é ela mesma impositora de padrões, pois é capaz de produzir partições mesmo quando estas originalmente não existam. A determinação do número de partições em um conjunto de dados é um problema ainda em aberto e objeto de pesquisa (FRALEY e RAFTERY, 1998). Para uma plena compreensão e validação dos resultados obtidos o concurso de critérios quantitativos e de especialistas de domínio se faz necessário.

2.2.1 K-MEANS

O *k-means* é apresentado nessa seção devido à sua simplicidade, eficiência computacional e principalmente por ser citado frequentemente como uma redução do processo de quantização vetorial durante o processo de aprendizado de mapas auto-organizáveis sem ajustes de vizinhança. O *k-means* também é empregado como método complementar na extração de grupos da grade de neurônios de um mapa auto-organizável.

O *k-means* proposto formalmente por MACQUEEN (1967) é um algoritmo de *clustering* divisivo para espaços métricos com raízes em trabalhos sobre quantização de mínimos quadrados aplicada em circuitos elétricos (LLOYD, 1982). Este método particiona um população de N amostras com d dimensões em K conjuntos, tendo cada conjunto um centro ótimo em função do somatório do erro mínimo quadrático, tomando-se por base a posição do centro espacial em relação a todos pontos que compõem cada partição. O algoritmo tem complexidade computacional da ordem de $\mathcal{O}(NKd)$, e, desde que o número de amostras seja bem maior do que o número de grupos e de dimensões, é capaz de tratar conjuntos de dados de grandes dimensões (XU e WUNSCH, 2005).

O algoritmo pode ser descrito segundo as seguintes etapas:

1 inicie com k grupos com centros com coordenadas escolhidas aleatoriamente.

2 aloque sequencialmente cada ponto da amostra ao grupo com centro mais próximo, recalculando a cada ciclo as coordenadas do centro como o ponto médio dos pontos adicionados.

3 em caso de convergência, parar, caso contrário retornar ao passo 2.

Critérios de convergência usualmente empregados são: não realocação de pontos após dois ciclos consecutivos ou um nível mínimo de redução do erro quadrático (JAIN, MURTY e FLYNN, 1999).

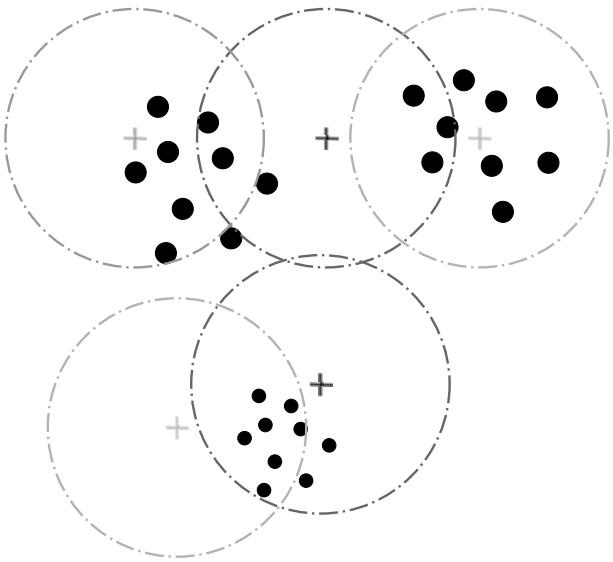


FIG. 2.5: k-means e o impacto da posição inicial dos centros na formação dos grupos.

O k-means tende a gerar grupos esféricos e tem como principal desvantagem a sua sensibilidade à alocação inicial dos centros de cada grupo, não havendo uma garantia de convergência para uma partição *ótima*, muito menos a *natural*. No exemplo apresentado na (figura 2.5) observa-se claramente que dado um critério de definição de pertinência a um grupo por proximidade, que grupos distintos são formados em função da posicão inicial do centro adotado. Diversas adaptações e extensões determinísticas (SU e DY, 2004) e estocásticas foram implementadas procurando aliviar essa limitação, e uma solução usual é a repetição do procedimento por um número fixo de vezes, selecionando-se a partição que produz o menor somatório do erro quadrático ou *Sum of Square Error* (SSE).

2.2.2 VALIDAÇÃO DE GRUPOS

2.2.2.1 *DAVIES-BOULDIN INDEX*

DAVIES e BOULDIN (1979) introduziram uma medida de separação entre grupos aplicada a espaços de entrada euclidianos onde seja possível a definição de uma função real de distância e, dados dois vetores $\mathbf{X_i}$ e $\mathbf{X_j}$, as seguintes propriedades sejam satisfeitas:

1 $d(\mathbf{X_i}, \mathbf{X_j}) \geq 0 \quad \forall\, \mathbf{X_i}, \mathbf{X_j} \in E_p$

2 $d(\mathbf{X_i}, \mathbf{X_j}) = 0 \quad$ se e somente se $\mathbf{X_i} = \mathbf{X_j}$

3 $d(\mathbf{X_i}, \mathbf{X_j}) = d(\mathbf{X_j}, \mathbf{X_i}) \quad \forall\, \mathbf{X_i}, \mathbf{X_j} \in E_p$

4 $d(\mathbf{X_i}, \mathbf{X_j}) \leq d(\mathbf{X_i}, \mathbf{X_k}) + d(\mathbf{X_k}, \mathbf{X_j}) \quad \forall\, \mathbf{X_i}, \mathbf{X_j}, \mathbf{X_k} \in E_p$

Medidas de distância que satisfaçam essas propriedades também são conhecidas como métricas.

O *Davies-Bouldin Index (DB-index)* é calculado levando-se em conta a dispersão interna ou intra-cluster e a distância inter-grupos. A dispersão interna é tomada como a média das distâncias de todos os vetores em relação ao centróide do grupo, e a separação inter-grupos é calculada como a distância entre os respectivos centros de grupo.

Dados dois grupos i e j quaisquer de uma partição resultante de um processo de agrupamento, define-se a distância inter-grupos por:

$$M_{ij} = \left\{ \sum_{k=1}^{N} |a_{ki} - a_{kj}|^p \right\}^{1/p} \tag{2.10}$$

onde a_{ki} é o k-ésimo componente do vetor característico a_i, centróide do grupo i. M_{ij} é também conhecida como métrica de Minkowski e que para $p = 2$ se reduz a distância euclidiana.

A dispersão de um grupo é definida pela expressão:

$$S_i = \left\{ \frac{1}{T_i} \sum_{j=1}^{T_i} |\mathbf{X_j} - \mathbf{A_i}|^q \right\}^{1/q} \tag{2.11}$$

sendo $\mathbf{A_i}$ o centróide do grupo i e T_i o número de elementos desse mesmo grupo.

R_i é o valor máximo de R_{ij} para $i \neq j$ e definido por:

$$R_i = max \left(R_{ij} = \frac{S_i + S_j}{M_{i,j}} \right) \tag{2.12}$$

O *DB-index* é, por fim, calculado como a média de similaridade de todos os grupos, comparando-se cada grupo com aquele que lhe é mais similar na forma:

$$DB = \frac{1}{N} \sum_{i=1}^{N} R_i \tag{2.13}$$

A melhor partição seria aquela que minimiza DB, já que quanto menor a razão mais compactos e distintamente separados estão os grupos.

Davies e Bouldin em experimentos com simulações de distribuições aleatórias de vetores bi-dimensionais, ou seja, sem a definição de grupos, encontraram valores de $DB \geq 0.6$. A princípio, valores inferiores a esse patamar indicariam a existência de grupos.

2.2.2.2 *SILHOUETTES*

Silhouttes é ao mesmo tempo um indicador e um suporte gráfico para a representação da coesão de grupos e partições em espaços euclidianos. Partindo de um definição sobre a média das proximidades, o método leva à construção de uma representação visual da qualidade de cada grupo que compõe uma partição e a partição como um todo.

Assuma uma partição P de n grupos onde $a(i)$ representa a dissimilaridade média da instância i para todas outras instâncias pertencentes ao grupo A.

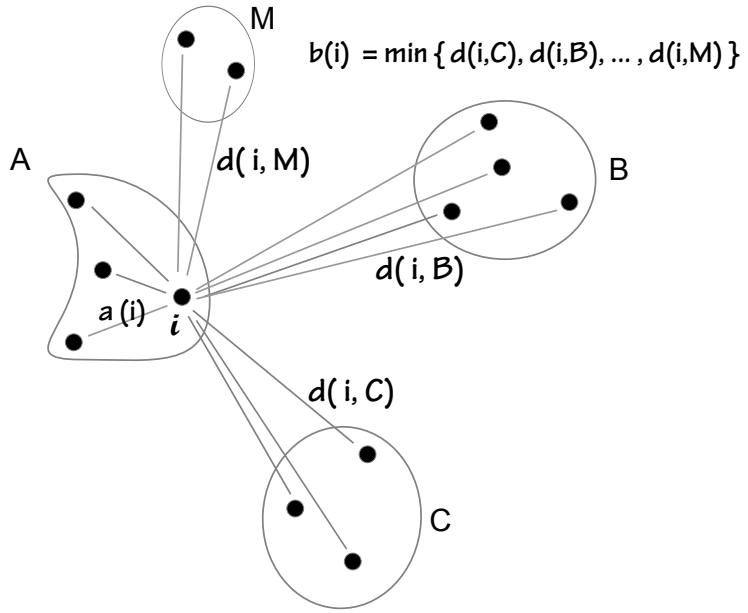


FIG. 2.6: Elementos envolvidos no cálculo de s(i), onde i pertence ao grupo A (adaptado de ROUSSEEUW (1987)).

Para qualquer outro grupo C tal que $C \neq A$, defina-se $d(i, C)$ como a média das dissimilaridades da instância i em relação a todos os componentes do grupo C (figura 2.6).

Por fim seja $b(i) = min\, d(i, C)$.

O *Silhoutte* de i é expresso por

$$s(i) = \frac{b(i) - a(i)}{max\,\{a(i), b(i)\}} \tag{2.14}$$

onde, no caso de grupos com apenas um elemento, $a(i)$ é definido como zero e $s(i) \in [-1\ 1]$.

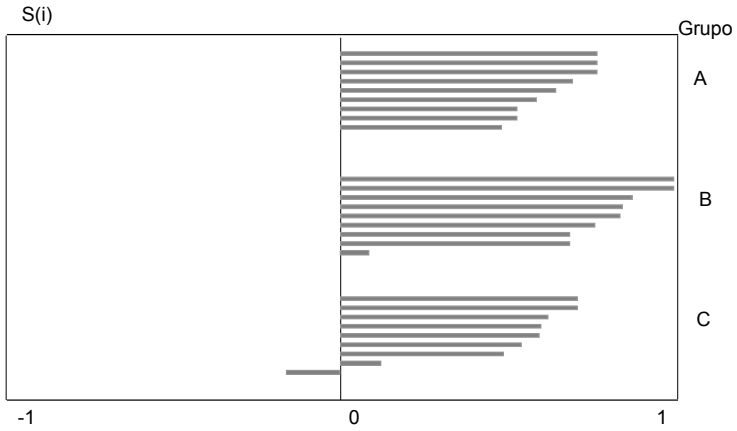


FIG. 2.7: *Silhouetes* de uma partição com 3 grupos.

A representação visual das silhuetas dos grupos de uma partição é obtida ao se desenhar ordenadamente, do maior para o menor, os valores de $s(i)$ para cada elemento de cada grupo (figura 2.7). Agrupamentos coesos tendem a gerar geometrias trapezoidais e quasi-retangulares no lado positivo do gráfico, com $s(i)$ se aproximando de 1.

O *Silhouettes* é particularmente sensível àquelas partições constituídas de grupos bem separados e esféricos como os potencialmente resultantes do k-means.

2.3 REDES NEURAIS

As Redes Neurais Artificiais (RNAs) têm uma recebido uma atenção crescente ao longo das últimas décadas por parte de pesquisadores face à sua capacidade de modelar problemas de alta complexidade, superando, em muitos casos, os resultados obtidos por metódos analíticos tradicionais.

RNAs podem ser definidas como "modelos computacionais" que possuem a propriedade de se adaptar ou aprender, generalizar, agrupar ou organizar dados, e cuja operação é baseada em processamento em paralelo. As RNAS são estruturas adaptativas que aprendem por experiência e são capazes de resumir ou capturar a essência dos sinais que lhes são inputados, fazer associações e generalizações ante estímulos oriundos do mesmo espaço sobre o qual foram treinadas.

Uma RNA é composta pelo arranjo de unidades elementares de processamento, os neurônios artificiais, conectados conforme uma arquitetura ou projeto, operando em par-

alelo e capazes de se adaptar em função de sinais recebidos ao longo de um processo de aprendizagem que pode, em última análise, armazenar e representar conhecimento.

Entre as características essenciais de uma RNA destacam-se:

- uma arquitetura de neurônios conectados ou não entre si;
- neurônios empregam uma função de ativação;
- há uma regra de propagação dos sinais entre os neurônios; e
- a informação é armazenada ao longo de processo de correção dos pesos sinápticos via uma regra ou algortimo de aprendizagem (KRÖSE e SMAGT, 1996)

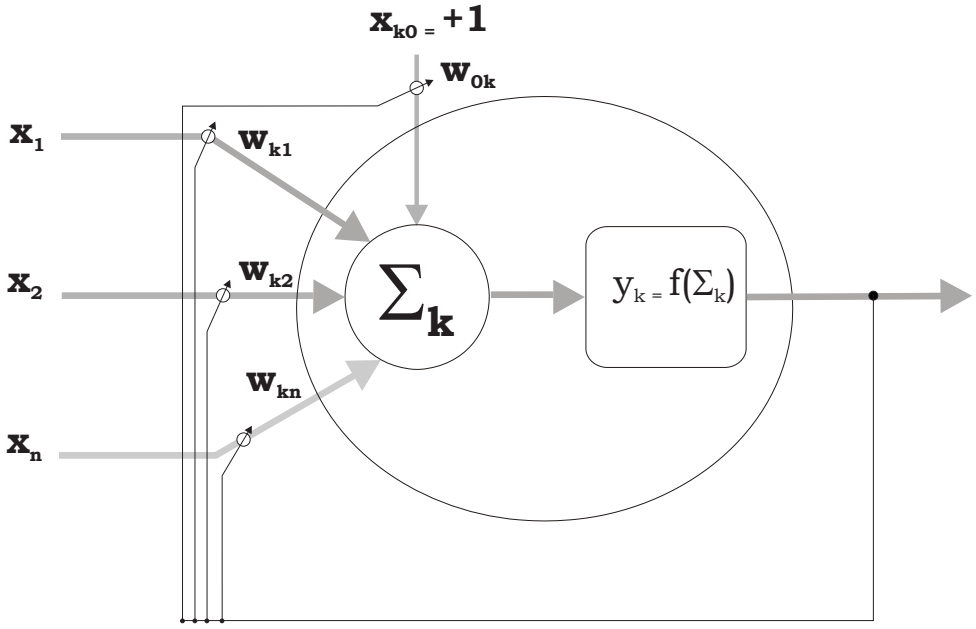


FIG. 2.8: Modelo de neurônio artificial baseado no Adaline (adaptado de WIDROW e LEHR (1990)).

A unidade elementar de uma RNA é um modelo simplificado do que é assumido ser a representação de um neurônio biológico, o neurônio artificial. O Perceptron de Rosenblatt, em 1958 e o Adaline de Widrow, em 1960, constituem os exemplos clássicos dessas estruturas e, salvo pequenas diferenças na forma como ajustam o sistema de pesos, ambos implementam os mesmos componentes e funcionalidades (figura 2.8): um conjunto de entradas, $\mathbf{x_1}, .., \mathbf{x_n}$, onde para cada uma existe um peso associado $\mathbf{w_{(k,1)}}, ..., \mathbf{w_{(k,n)}}$, um somador e uma função de ativação (VEELENTURF, 1995).

A função de ativação pode ser linear, ou não linear, e determina as condições de ativação ou disparo do neurônio (figura 2.9) enquanto um algoritmo ou regra de aprendizado é o responsável pela alteração dos pesos sinápticos ao longo de um processo interativo que procura de forma adaptativa convergir para um estado capaz de representar e reter informação sobre a distribuição do espaço de entrada apresentada ao neurônio.

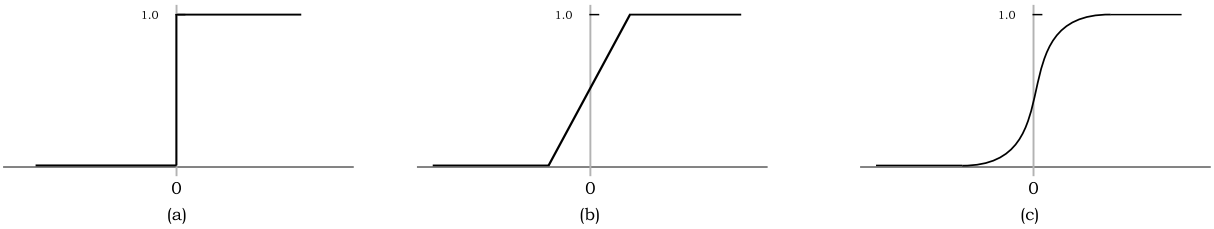


FIG. 2.9: Funções de ativação: (a) limiar, (b) linear por partes (b) e (c) logística.

Um neurônio artificial é um combinador linear adaptativo capaz de operar como um classificador linear. Após a fase de *treinamento*, os pesos sinápticos convergem para um estado no qual o neurônio é capaz de separar novos estímulos oriundos da mesma distribuição com probabilidade função do número de padrões de entrada N_p e do número de pesos N_w. WIDROW e LEHR (1990) reve a expressão para essa probabilidade demonstradas por COVER (1964) e BROWN (1964) sob a forma:

$$P_{sep} = \begin{cases} 2^{-(N_p-1)} \sum_{i=0}^{N_w-1} \binom{Np-1}{i}, & \text{se} \quad N_p > N_w \\ 1, & \text{caso} \quad N_p \leq N_w \end{cases} \tag{2.15}$$

Dois aspectos estão intimamente relacionados à capacidade das RNAs de resolver problemas bem mais complexos do que os oferecidos na classificação linear, um deles é a topologia ou arquitetura da rede e a correspondente regra de propagação e processo de aprendizado.

2.3.1 ARQUITETURAS

Quanto à sua arquitetura, as RNAs usualmente podem ser projetadas como redes redes recorrentes, alimentadas adiante em camada única ou alimentadas adiante em múltiplas camadas (figura 2.10). As *Multilayer Perceptrons* (MLPs) são o caso mais conhecido de topologia de múltiplas camadas, composto por neurônios, contempla uma camada de neurônios de entrada, uma de saída e uma ou mais camadas adicionais, denominadas

camadas escondidas, não existindo conexões entre neurônios da mesma camada. Nas MLPs, o fluxo de dados percorre a rede numa única direção, indo da camada de entrada para a de saída sem retroalimentação ou *feedback*. Diferentemente das redes de uma única camada, as MLPs podem ser treinadas para representar dados linearmente não-separáveis (HAYKIN, 2001).

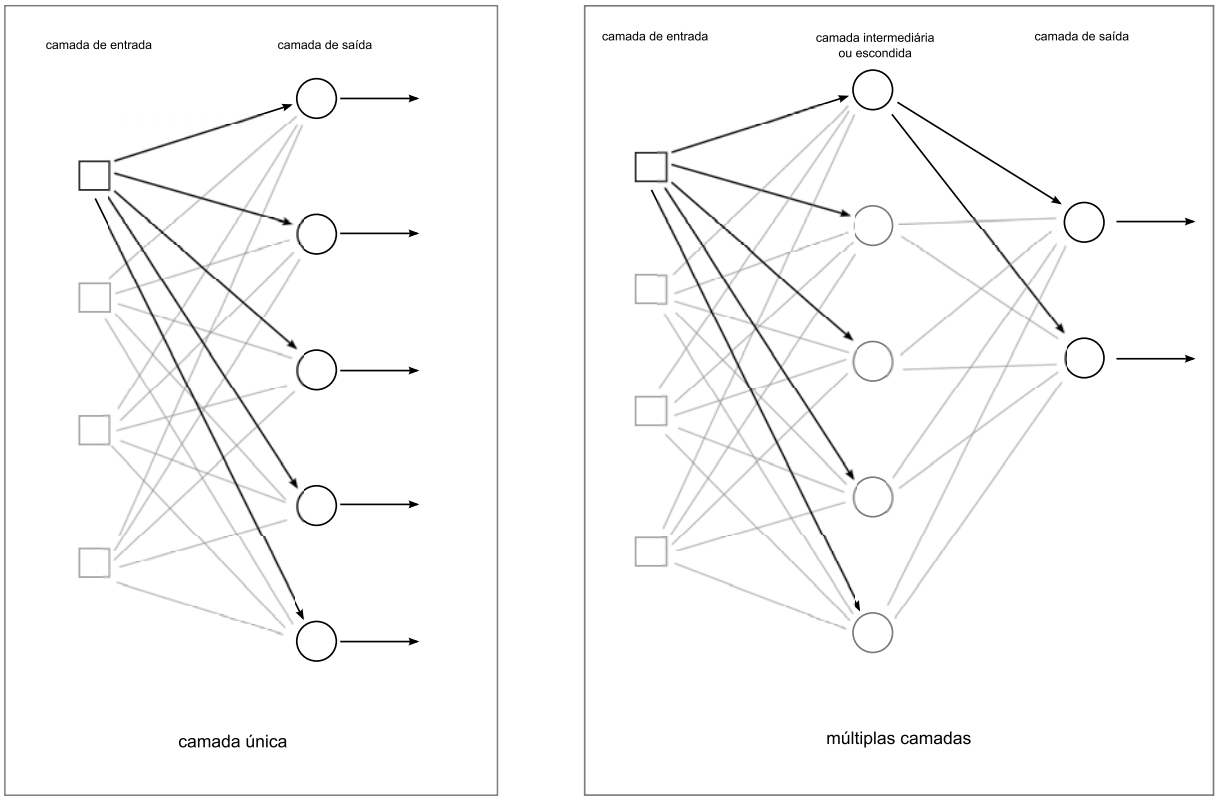


FIG. 2.10: Arquitetura de camada única e múltiplas camadas.

2.3.2 PROCESSO DE APRENDIZADO

HAYKIN (2001) cita cinco regras básicas de aprendizagem de uma RNA:

- por correção de erro;
- por memória;
- hebbiana;
- competitiva; e
- de Bolztamann.

.

No aprendizado supervisionado, método empregado nas MLPs, as saídas resultantes da rede são comparadas com valores de referência ou valores desejados, direcionando o processo de ajuste dos pesos sinápticos. Na aprendizagem não-supervisionada são os fatores inerentes à distribuição estatística do espaço de entrada que direcionam o ajuste da rede, a rede *aprende* exclusivamente a partir dos dados de entrada.

Segundo a regra Delta, os pesos sinápticos são atualizados iterativa e proporcionalmente ao negativo da derivada parcial do erro entre a saída do neurônio e o valor desejado, método conhecido como gradiente decrescente, e representado por:

$$\Delta_{ik} = \eta \, x_{ik} \, w_{ki} \, (y_{ref} - y_k) \tag{2.16}$$

onde $0 < \eta < 1$ é uma constante denominada taxa de aprendizagem, x_{ik} o valor de entrada i para o neurônio k, w_{ik} o peso, y_{ref} o valor desejado ou de referência e y_k a saída do neurônio k (figura 2.11).

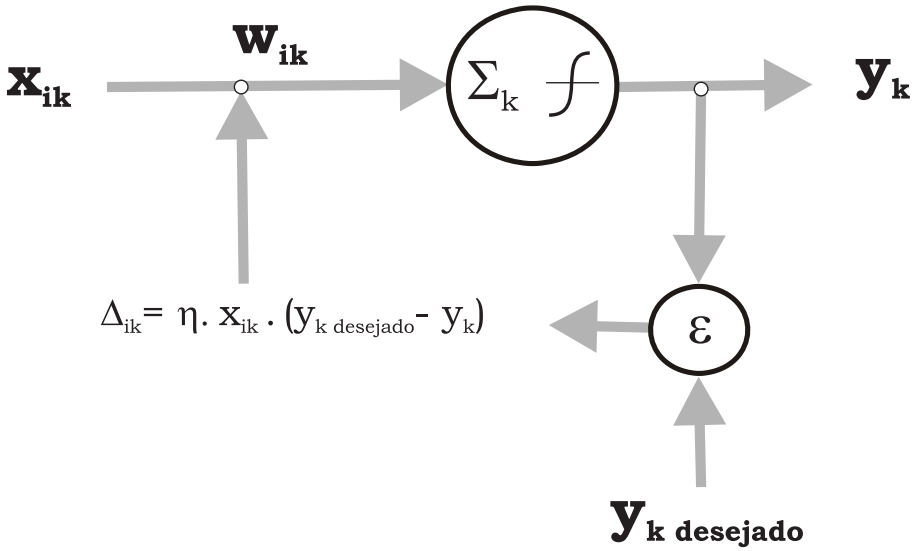


FIG. 2.11: Ajuste de pesos segundo a regra Delta.

O algoritmo para um ciclo de aprendizado pode ser descrito segundo os seguintes passos:

1 iniciar os pesos sinápticos com valores aleatórios reduzidos ou zero;

2 apresentar o padrão de entrada e respectivo valor desejado e calcular o valor de saída;

3 calcular o erro;

4 se o erro é zero, voltar ao passo 2, caso contrário atualizar pesos pela taxa definida por 2.16;

5 voltar ao passo 2 até que todos os padrões tenham sido apresentados a rede (VELLASCO, 2009).

Após a sua proposta inicial na década de 60, as RNs tiveram um longo período de ocaso, até que, em 1982, Parker redescobriu o algoritmo de treinamento por retropropagação ou *backpropagation*, proposto inicialmente por Werbos em sua tese de doutorado em 1974 (WIDROW e LEHR, 1990).

O algoritmo de retropropagação aplicado em MLPs resolve de modo elegante a correção dos pesos das camadas escondidas. Enquanto os dados percorrem a rede no sentido da camada de entrada para a camada de saída, os pesos são ajustados em função do erro no sentido oposto. Demonstrou-se que as MLPs com uma camada escondida são capazes de operar como aproximadores universais de funções contínuas (CYBENKO, 1989).

A capacidade de aprendizado do algoritmo de retropropagação é relacionada à quantidade de neurônios nas camadas escondidas e ao número de interações durante a fase de treinamento. O erro médio decresce à medida que o treinamento prossegue ao longo do tempo, ou ciclos de treinamento. Entretanto, não é desejável nem que haja um treinamento excessivo ou um número elevado de camadas intermediárias, a fim de se evitar um ajuste extremamente rígido aos dados do espaço de entrada, reduzindo, ou até mesmo eliminando a capacidade da rede de produzir generalizações para novos dados.

Uma estratégia complementar é expor a RNA a sub-conjuntos dos dados de entrada extraídos de modo a representar a distribuição das densidades existentes no universo disponível. Um conjunto destina-se à fase de treinamento, outro para validar os resultados da fase treinamento e, por fim, um conjunto de teste é usado para comprovar o resultado da rede com melhor desempenho. Validação cruzada, ou *cross-validation*, é outra estratégia de amostragem para o treinamento de RNAs, onde o conjunto de dados é dividido em n grupos, geralmente dez; n-1 subgrupos são usados na fase de treinamento e o enésimo restante para testes de validação, havendo um circuito de teste percorrendo todas n partições.

De relevância maior para o presente trabalho são os processos de aprendizagem competitiva, pois constituem o cerne do funcionamento dos Mapas Auto-Organizáveis. Como

o próprio nome descreve, é através de uma competição entre os neurônios que apenas um é selecionado para ter os respectivos pesos ajustados. Ao contrário das MLPs onde todos os neurônios são afetados, na aprendizagem competitiva somente o vencedor tem esse privilégio, numa estratégia conhecida como "o vencedor leva tudo" (*the winner takes it all*). Ao longo do ciclo de treinamento, o vetor de referência associado a cada neurônio, descrito pelo conjunto de pesos de suas conexões sinápticas, move-se na direção do centro de subconjuntos do espaço amostral, de modo semelhante ao que ocorre no k-means.

Um mapa auto-organizável adota o aprendizado competitivo com uma adaptação crítica se comparada a proposta original, nele não só os pesos do neurônio vencedor são ajustados, mas um conjunto de neurônios vizinhos como apresentado na próxima seção.

2.4 MAPAS AUTO-ORGANIZÁVEIS

Um Mapa Auto-Organizável (*Self-Organizing Maps*- SOM) é uma RNA de camada única com neurônios dispostos em uma grade capaz de se ordenar topológicamente capturando a distribuição dos dados de entrada de espaços n-dimensionais. Esse ordenamento é realizado por meio de aprendizagem competitiva. Contudo, diferentemente da estratégia tradicional, tanto o neurônio vencedor como os neurônios que compõem a sua vizinhança são afetados pelo processo de ajuste dos pesos sinápticos.

O SOM ou mapa de Kohonen, nome dado em referência ao seu criador, pode ser descrito como uma mapeamento suave, ordenado e não-linear de coletores de dados de alta dimensionalidade sobre os elementos de uma grade regular de baixa dimensionalidade (KOHONEN, 2000a).

Na grade do SOM cada neurônio tem associado a si um vetor real paramétrico $m_i = [\mu_{i1}, \mu_{i2}, ..., \mu_{in}]_T \in \Re^n$, chamado de modelo ou protótipo, com a mesma dimensão dos vetores do espaço de entrada $x = [\xi_1, \xi_2, ..., \xi_n]_T \in \Re^n$ (figura 2.12).

A competição para escolha do neurônio que mais se aproxima de um vetor de entrada $\mathbf{x}$ se dá pela comparação através de alguma métrica de distância com todos os vetores modelo $\mathbf{m_i}$, cada um correspondendo a um neurônio da grade. O neurônio vencedor de índice c, ou *Best Match Unit* (BMU), é aquele cujo modelo é mais próximo do vetor de entrada e satisfaz :

$$c = arg\ min_{i=1}^{k} \{d(\mathbf{x}, \mathbf{m_i})\} \tag{2.17}$$

Uma vez definido um BMU, o próximo passo do algoritmo é a fase de adaptação, ou

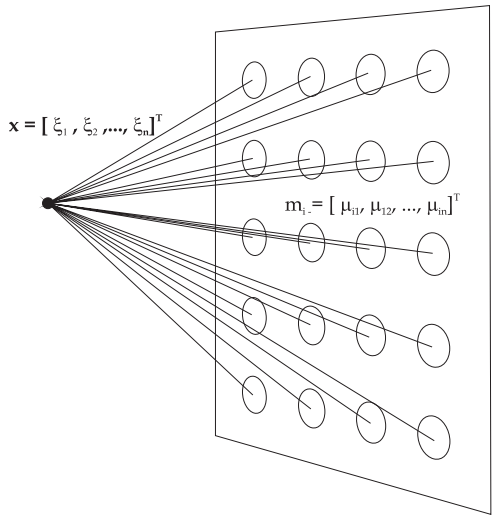


FIG. 2.12: Grade do SOM.

o ajuste da vizinhança do neurônio vencedor. O ajuste dos pesos é sujeito a uma taxa de aprendizado decrescente ao longo do tempo e ao longo da vizinhança do BMU segundo a expressão:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)], \tag{2.18}$$

onde t, um inteiro, é uma coordenada de tempo discreta associada aos passos do treinamento. $h_{ci}(t)$ é a função de vizinhança que deve satisfazer $h_{ci}(t) \to 0$ quando $t \to \infty$ para que haja convergência do algoritmo e pode ser descrita por:

$$h_{ci}(t) = \alpha(t)\, e^{\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right)}, \tag{2.19}$$

onde $\mathbf{r_c}$ e $\mathbf{r_i}$, vetores bi-dimensionais com as coordenadas dos neurônios na grade e $\alpha(t)$ é a taxa de aprendizado e $\sigma(t)$ define a largura do núcleo em torno do BMU, ambos decrescentes com o tempo (figura 2.13).

Essa fase de ajuste de vizinhaça é a responsável pela capacidade de projeção e mapeamento do SOM, sem a qual os seus resultados se reduzem aos obtidos pelo k-means.

O algoritmo pode ser resumido nos seguintes passos:

1 Inicialize os vetores de pesos com valores diferentes entre si. Essa inicialização pode ser aleatória originária dos próprios dados de entrada, aleatória com valores quaisquer, ou seguir algum conhecimento a priori sobre a distribuição dos dados.

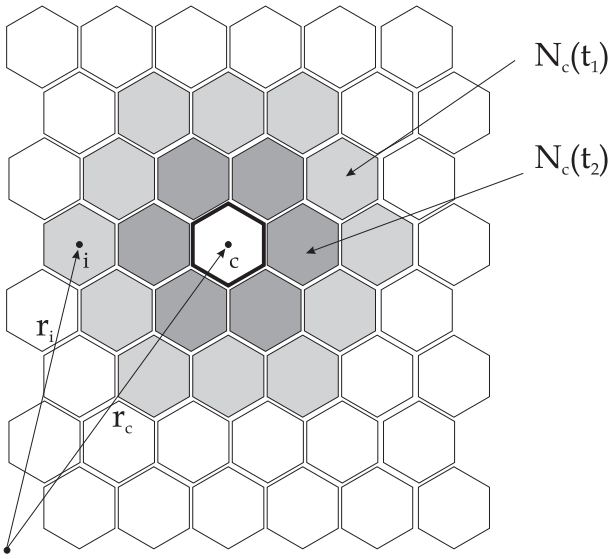


FIG. 2.13: BMU e regiào de vizinhança afetada pelo ciclo de treinamento do SOM.

2 Retire uma amostra aleatória $\mathbf{x}$ do espaço de entrada.

3 Selecione o BMU $\mathbf{m_i}$ correspondente segundo a expressão 2.17. Aloque $\mathbf{x}$ a $\mathbf{m_i}$.

4 Por um tempo t, ajuste os pesos do vetor modelo $\mathbf{m_i}$ e de todos em sua vizinhança pela expressão 2.18. Usualmente essa fase ocorre em duas etapas, uma para o ajuste grosseiro com taxa de aprendizado $\alpha(t)$ maior (preferencialmente $t_1 > 1000$), e outra de ajuste fino, ou fase de convergência, onde $t_2 > t1$, um vizinhança menor e com taxas de aprendizado bem menores e tendendo a zero.

5 Retornar ao passo 2 até que um critério de ordenação ou de qualidade tenha sido atingido.

Essa estratégia é conhecida como *treinamento sequencial* que pode ser substituída sem perda significativa de qualidade pelo *treinamento em lote*, convergindo mais rapidamente que o treinamento sequencial.

No *treinamento em lote*, o ajuste dos pesos sinápticos é feito ao final da varredura de todas as amostras do espaço e os vetores protótipos atualizados segundo a expressão:

$$m_i(t+1) = \frac{\sum_{j=1}^{n} h_{ci(j)}(t)\mathbf{x_j}}{\sum_{j=1}^{n} h_{ci(j)}(t)}, \tag{2.20}$$

onde $c(j)$ é o BMU de $\mathbf{x_j}$, $h_{ci(j)}$ a função de vizinhança e n o número de instâncias.

O BMU é atualizado como uma média ponderada sobre o espaço de entrada, tendo como pesos os valores correspondentes da função de vizinhança.

Na fase de ordenação, envolvendo pelo menos 1000 iterações, o tamanho do raio inicial N_c pode ser tão grande quanto a metade da largura da grade. A taxa de aprendizado $\alpha(t)$ em função do tempo não é um fator extremamente crítico, podendo ser iniciada com valores altos e decrescer a uma taxa inversamente proporcional ao tempo. Na fase de ajuste fino do treinamento, o número total de iterações não deve ser menor do 500 vezes o número de neurônios na grade.

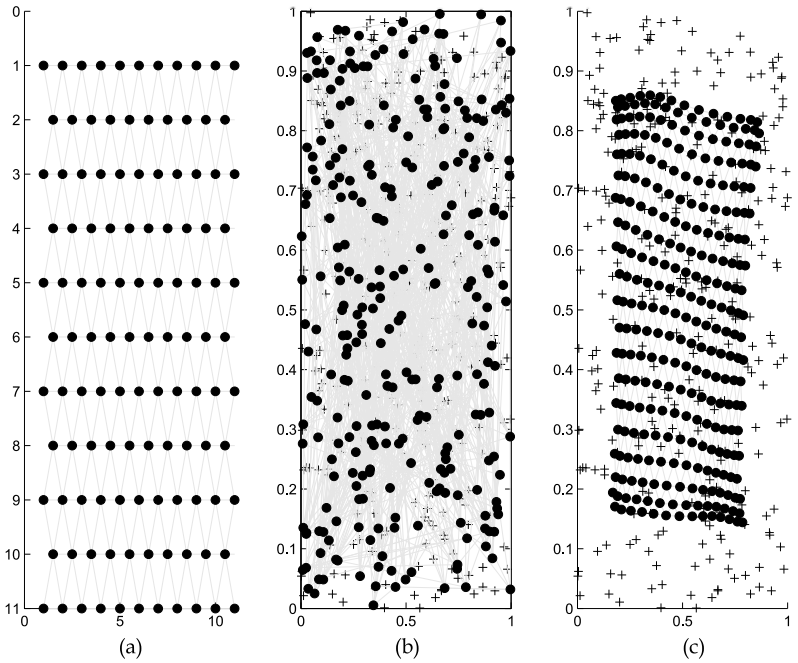


FIG. 2.14: Ordenação dos vetores modelo: (a) grade de neurônios, (b) vetores modelo antes e (c) depois do SOM (Gerado pelo SOM Toolbox, CIS (2010b), para o Matlab 6.5).

Após o treinamento, o conjunto de neurônios que compõem a grade exibe uma configuração espacial que reflete o espaço de entrada e, a cada BMU, é atribuída uma lista associada de instâncias que lhe são mais próximas. Uma vez a rede treinada, é simples encontrar qual o BMU que se encontra mais próximo no espaço de entrada de uma nova instância oriunda da mesma distribuição, bastando para isso empregar a mesma métrica utilizada na fase de treinamento e comparar com todos os BMUs. Assumindo-se que cada modelo tenha sido objeto de processo classificatório, a rede resultante do SOM terá *aprendido* a classificar padrões semelhantes.

Na figura 2.14(a) é possível ver a grade de neurônios e sua disposição no espaço

de saída. Em 2.14(b) é apresentado um conjunto exemplo de dados de entrada bi-dimensionais de distribuição aleatória indicados pelo sinal + e os vetores modelo em seu estado aleatório inicial. Ao final do processo de aprendizagem, os vetores modelos ocupam no espaço de entrada posições relacionadas às distribuições de densidades dos dados de treinamento (figura 2.14(c)).

Ao comparar o SOM com a análise de componentes principais (*Principal Componente Analysis* - PCA), Kohonen destaca que, ao contrário do PCA, que descreve as propriedades estatísticas globais da distribuição dos dados, o SOM poderia ser caracterizado como uma rede ou superfície bi-dimensional elástica de elementos finitos que se ajusta à distribuição das densidades do espaço de entrada.

Importante ressaltar que os neurônios do SOM, apesar de dispostos em grades regulares, frequentemente de 1 ou 2 dimensões, podem ser conectados aos neurônios adjacentes em estruturas variadas formando malhas. As forma da grade se relaciona com a geometria de suas células que podem ser ou hexagonais ou retangulares. Já a topologia da grade se refere ao modo como são tratadas as bordas da grade. Malhas em forma de folha possuem quatro bordas, as cilíndricas duas e as toroidais não possuem bordas (figura 2.15). A existência ou não de bordas impacta o cálculo e ajuste das vizinhanças de BMUs nessas regiões, devido ao fato de serem circundadas por um número menor e não-regular de células adjacentes.

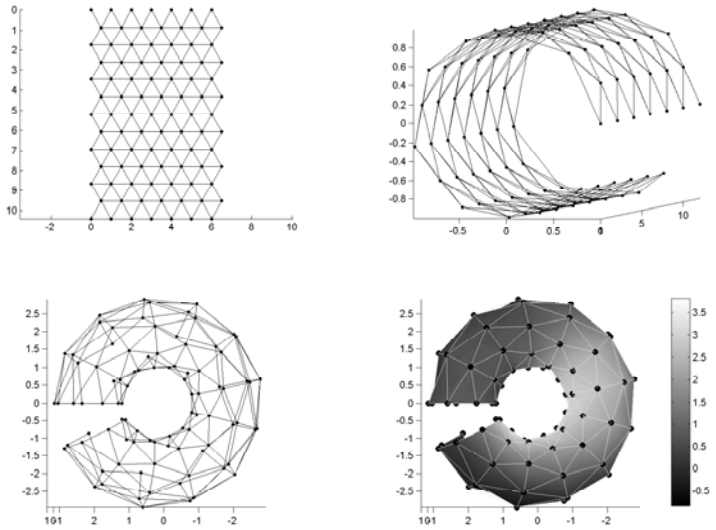


FIG. 2.15: Topologia de malhas para grade de neurônios do SOM.

O SOM tem sido aplicado extensa e intensivamente nas mais variadas áreas e em aplicações diversas, inclusive incorporado em circuitos eletrônicos. É extensa a lista de variantes e extensões e, até o ano de 2000, o autor contabilizava mais de 4000 artigos referentes ao tema. Em 2010, esse número já ultrapassa a casa dos 7000 (CIS, 2010a).

Na literatura há referências que procuram comparar os resultados do SOM com os obtidos com o k-means, contudo diversos trabalhos exibem resultados indicando a superioridade do primeiro em cenários de distribuições complexas. Um destes exemplos é o caso de dois grupos formados por anéis tri-dimensionais entrelaçados como na figura 2.16 (ULTSCH e VETTER, 1995).

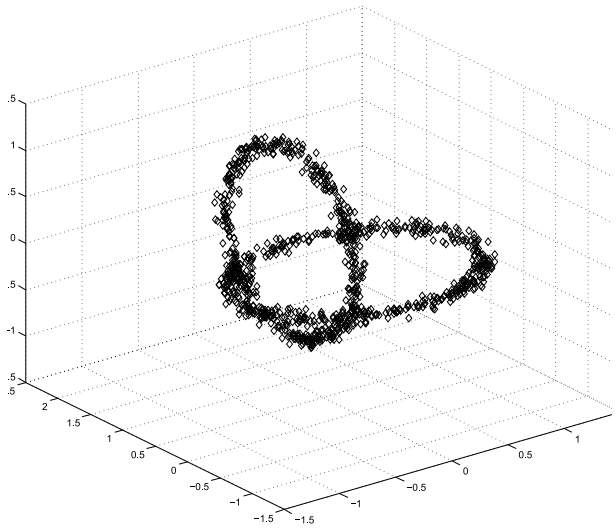


FIG. 2.16: Conjunto de dados Chainlik: 2 grupos no formato de anéis tri-dimensionais entrelaçados (ULTSCH e VETTER, 1995).

2.4.1 QUALIDADE DO SOM

Diversas medidas têm sido propostas para avaliar a qualidade dos resultados obtidos com o SOM. Tradicionalmente duas se destacam pela frequência com que são citadas e implementadas nas ferramentas abertas disponíveis: o erro de quantização e o erro topográfico.

O Erro de Quantização (EQ) é expresso pela distância média entre as instâncias ou vetores de entrada e os respectivos centróides ou protótipos.

$$\overline{EQ_j} = \frac{\sum \|\mathbf{x_i} - \mathbf{m_i}\|}{|C_{m_i}|}, \tag{2.21}$$

onde $|C_{m_i}|$ é o número de amostras alocadas ao BMU m_i.

O Erro de Quantização (EQ) tem a propriedade interessante de ser relacionado a outras medidas de quantização empregadas para avaliar diversos algoritmos de *clustering*, mas o seu valor descresce à medida que o número de unidades do mapa aumenta, tornando-o ineficaz para comparação entre mapas de tamanhos diferentes. O Erro de Quantização também não leva em conta nenhum aspecto referente à alocação topográfica das instâncias e respectivos modelos.

O Erro Topográfico (ET) é uma medida simples da alocação de instâncias aos respectivos protótipos. Aqui, contam-se violações de topografia, ou o número de vezes que o segundo *Best Match Unit* não se encontra em um neurônio adjacente ao primeiro BMU. Essa contagem é normalizada para o intervalo entre 0 e 1, onde 0 indica a melhor preservação de topologia. O Erro Topográfico não leva em consideração erros de quantização e é sensível ao tamanho do mapa (PÖLZLBAUER, 2008).

Multi-SOM é uma proposta recente que procura endereçar as dificuldades da avaliação da qualidade de diferentes SOM aplicados a um mesmo conjunto de dados. O método analisa os deslocamento topológico de pares de dados e pares de grupos, expresso entre um SOM de referência e outros SOMs (MAYER, NEUMAYER e RAUBER, 2009).

2.4.2 VISUALIZAÇÃO DO SOM

Uma ampla variedade de representações visuais de resultados do SOM baseia-se na sobreposição de símbolos ou colorização das células da grade de neurônios ou de *metáforas* mais sofisticadas, como as realizadas por meio de mapas cartográficos.

Na sua forma mais geral, a topologia da malha de neurônios não é limitada a um número específico de dimensões, mas ultrapassar o limite de 3 dimensões representa retornar ao dilema da incapacidade humana de visualizar espaços de dimensões superiores. Desta forma mapas bidimensionais são a opção mais frequentemente empregada.

Desenvolvido no Departamento de Ciência da Computação da Universidade de Helsinki, o pacote SOM Toolbox (CIS, 2010b) é recorrentemente citado na literatura como ferramenta utilizada em estudos de aplicações com o SOM e permite uma ampla gama de representações visuais como o planos dos componentes e a matrix de distâncias, além de outros métodos de projeção como Componentes Principais e o Mapeamento de Sammon.

O plano dos componentes exibe a distribuição das contribuições de cada atributo na coordenadas do espaço de entrada de cada neurônio. Um escala de referência nas dimensões

originais de cada plano é exibida como suporte para a interpretação dos resultados (figura 2.17).

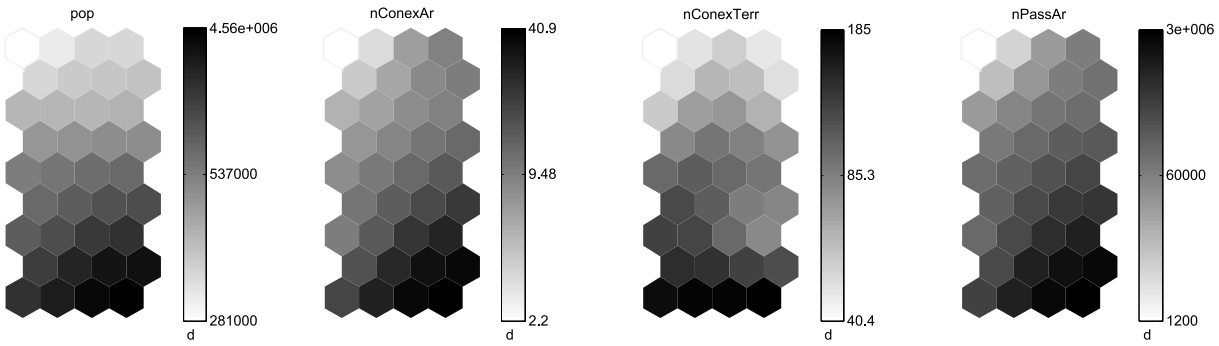


FIG. 2.17: Plano de componentes de uma malha SOM topologia de folha com células hexagonais.

A *Unified Distance Matrix*, ou U-matrix (ULTSCH e SIEMON, 1990) é um modelo de representação visual baseada na distância em coordenadas do espaço dos atributos de protótipos adjacentes no espaço de saída do SOM.

A distância u entre dois modelos mi e m_j é

$$u(mi, m_j) = \|\mathbf{mi} - \mathbf{mj}\| \tag{2.22}$$

O valor médio para cada modelo mi é a média das distâncias entre $\mathbf{m_i}$ e o conjunto P_i de protótipos adjacentes que compõem a sua vizinhança na grade.

$$\overline{u(mi)} = \frac{1}{|P|} \sum_{k \in P_i} u(mi, m_k) \tag{2.23}$$

Na construção da U-matrix, a representação simultânea da distância média em cada célula e das distâncias com as células adjacentes exige a inserção de celúlas extras com valores calculados por 2.22. Por fim, é aplicado um mapeamento de cores em relação às distâncias, de modo que tons mais escuros representem as maiores distâncias entre um neurônio e seus vizinhos.

Alternativamente, podemos contar com a matriz de distâncias, ou D-Matrix, que é realizada através do mapeamento da distância média em cores aplicadas às células do mapa em um mapeamento um a um, com os neurônios da grade (figura 2.18).

O resultado principal da observação da U-matrix é a identificação de potenciais agrupamentos, uma vez que surgem regiões de tons mais claros formadas por protótipos com

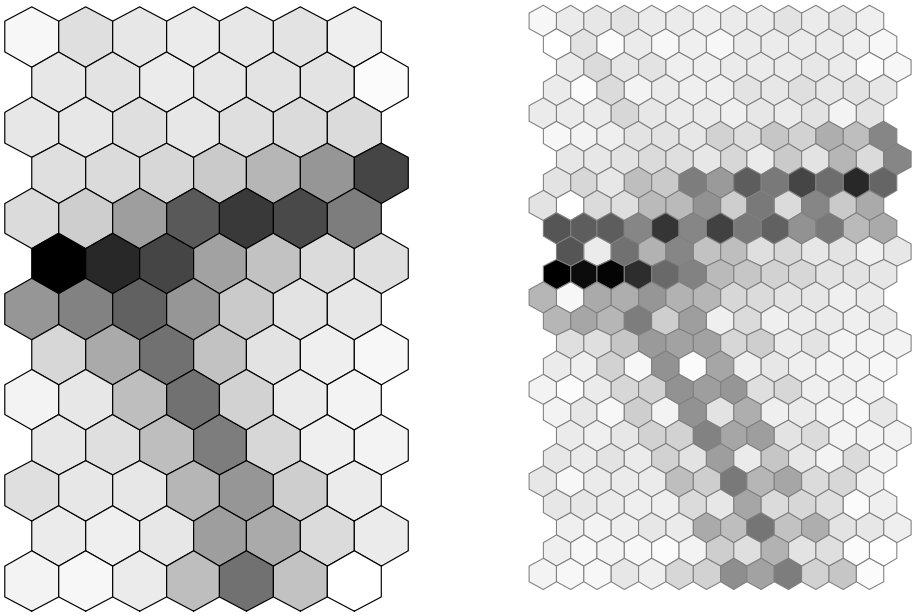


FIG. 2.18: (a) D-matrix, (b) U-matrix.

coordenadas próximas entre si circundadas por regiões escuras formando barreiras ou vales. A emergência dessas estruturas na U-matrix e a qualidade da definição das mesmas estão intimamente associadas ao tamanho do mapa, mas refletem a distribução das densidades do espaço de entrada captadas pela SOM.

2.5 SOM *CLUSTERING*

O SOM não é, em sua essência, um método de análise de agrupamentos. KOHONEN (2000b), seu criador, o categoriza como um tipo especial de adaptação, um fenômeno relacionado com regressão não-paramétrica, mas foi com o advento de representações mais elaboradas como a U-matrix que sua aplicabilidade em inúmeros problemas dessa natureza tornou-se evidente.

Como visto na seção anterior, os BMUs atuam como coletores de instâncias do espaço de entrada, agrupando aquelas que lhe são mais próximas. Nesse contexto, os BMUs podem ser vistos como *proto-clusters* e a observação da distribuição desses coletores na U-matrix enseja a percepção que fazem parte de estruturas maiores.

A estratégia primária de recuperação dessas estruturas que emergem na U-matrix é a manual. Nela um operador identifica os limites de grupos e recupera as instâncias alocadas aos BMUs que compõem as regiões formadas por protótipos com valores similares.

Obviamente, consiste num processo sujeito a interpretações diversas e inadequado para a automação de processos de identificação de *clusters*.

Outra abordagem consiste em aplicar algoritmos de *clustering* nos protótipos. Uma das referências mais citadas é o trabalho de VESANTO e ALHONIEMI (2000) que conduz uma série de experimentos procurando verificar a eficácia e o desempenho no uso combinado do SOM com algoritmos aglomerativos e divisivos, entre eles o k-means, com resultados bastante satisfatórios.

Emergent SOM (ESOM) é uma variante do SOM que preconiza que estruturas realmente diferenciadas só podem ser realmente captadas na U-matrix em mapas de grandes dimensões, acima de 4000 unidades.

Na figura 2.19, um sequência de mapas gerados com 80, 320, 720 e 1200 unidades de neurônios, para um mesmo conjunto de dados sintético composto por 210 instâncias, representando seis grupos esféricos e bem separados. Observa-se que um número reduzido de unidades no mapa dificulta a percepção de grupos. À medida que a quantidade de neurônios é aumentada, as linhas divisórias se tornam mais nítidas, mas, após um certo limiar, o excessivo número de células em comparação ao número de instâncias provoca o surgimento de nervuras secundárias que podem pertubar o mapa induzindo o surgimento de *falsas* sub-regiões.

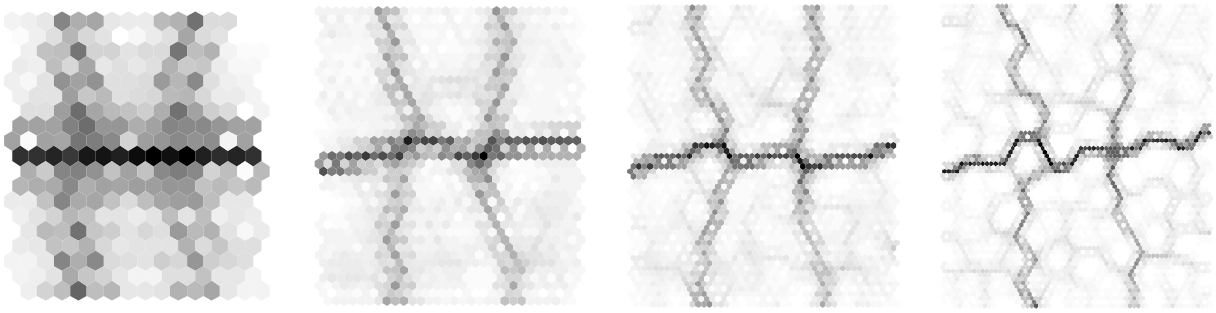


FIG. 2.19: U-matrix para conjunto de dados com seis grupos: (a) 80, (b) 320, (c) 720 e (d) 1200 unidades ou células.

O ESOM também se apóia no emprego de mapas toroidais, onde a borda superior se liga a inferior, assim como laterais entre si. Esse arranjo visa eliminar o efeito borda de mapas em formato de folha ou lâmina, mas cria uma dificuldade extra na visualização no plano do mapa resultante.

U*-matrix é uma variante da U-matrix concebida na esteira de estudos relacionados com o ESOM que adiciona considerações de densidade das distribuições nas bordas

de regiões formadoras de *clusters*. O conceito central é que não só as distâncias entre neurônios indicam os limites de grupos, mas também a sua densidade. Os núcleos dos grupos seriam regiões relativamente densas e homogêneas e as bordas regiões de baixa densidade e com valores bem distintos nos núcleos. O método emprega estimativas de densidades calculadas por meio de cálculo de probabilidades ao longo de hiper-esferas em torno dos protótipos (ULTSCH e MÖRCHEN, 2005).

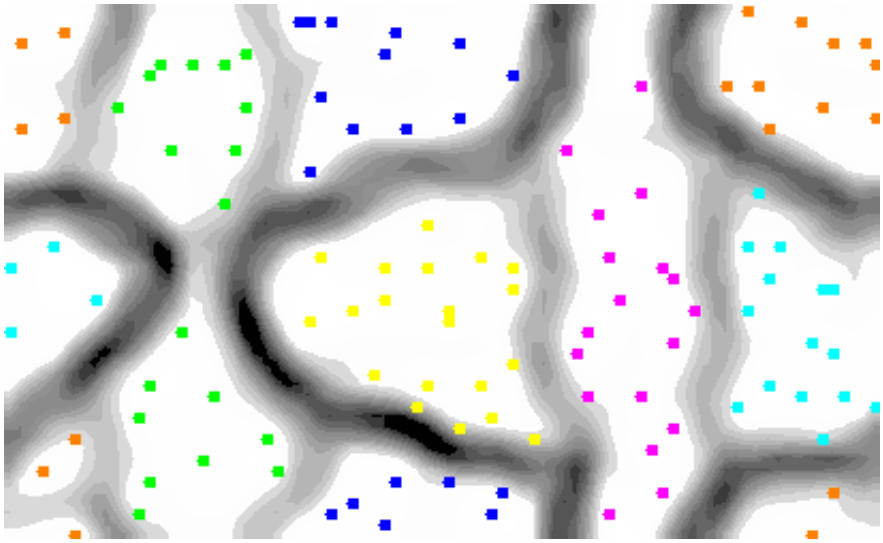


FIG. 2.20: U*-matrix para conjunto de dados com seis grupos. Nela se vê os BMUs como pontos separados por bordas determinadas pela análise distribuição das densidades.

Para a construção da U*-matrix é necessário o aporte de outra estrutura, o mapa de densidades ou P-matrix, cujas contribuições combinadas permitem a produção de mapas topográficos e inclusão de curvas de nível. Na figura 2.20 é apresentada a U*-matrix para o mesmo conjunto de dados da figura 2.19.

SHARMA e OMLIN (2006) abordam a questão das extração de grupos inseridos na estrutura da U-matrix, empregando algoritmos de otimização de enxame de partículas, e comparam esse método com a extração manual e uso do k-means, com bons resultados, mas ressaltam que a aplicabilidade é limitada a problemas de pequeno porte, face a degradação de performance à medida que as dimensões da grade aumentam.

Baseado no exposto, observa-se que é contínua a produção científica envolvendo a captura de agrupamentos a partir das estruturas adaptativas geradas pelo SOM, tendo sido demonstrada, em diversos estudos, a sua superioridade sobre outras estratégias em situações diversas. Esse contínuo interesse indica que o método mantém o seu apelo, alcance e aplicabilidade em processos de análise de grupos não-supervisionados.

3 TRABALHOS RELACIONADOS

Diversas abordagens procuram identificar a existência de grupos em redes descritas por meio de grafos, sendo que em sua grande maioria as características estruturais, como o grau dos vértices, a posição relativa dos nós e a distribuição das arestas, fornecem os subsídios para a determinação de partições ou da probabilidade destas ocorrerem.

Ainda no campo da Teoria dos Grafos é possível destacar estudos que procuram analisar a similaridade estrutural relativa de redes como um todo buscando detectar padrões estruturais globais na formação grupos de grafos e não de vértices (SCHAEFFER, 2007). Em direção diversa, mais ainda empregando grafos, WATTS e STROGATZ (1998), em um trabalho seminal, apontaram que diversos tipos de redes sociais, tecnológicas e biológicas compartilham características estruturais. O estudo de redes complexas, incluindo as formadas por *links* em páginas da internete, redes sociais, como as formadas por atores de cinema, e muitas redes biológicas, mostrou que, na maioria desses casos, o grau de distribuição segue, uma escala de lei de potência $P(k) = \frac{N_k}{N} \approx k_{-\lambda}$, onde k é o número de ligações de um nó, N o número total de nós e λ uma constante entre 2 e 3 (ALBERT e BARABÁSI, 2002).

Em sua quase totalidade, a linha de estudos derivada da Teoria dos Grafos , trata a similaridade entre vértices pelo aspecto grau do nó e visualiza grupos empregando critérios de distância relativa entre um nó de referência e nós vizinhos. Dentro do escopo da revisão bibliográfica empreendida nessa dissertação, não foram encontradas referências que oferecessem soluções para similaridades não-estruturais, empregando somente características das entidades representadas pelos nós ou vértices das redes.

No caso particular da análise de redes urbanas quanto a centralidade e similaridade, podemos destacar as seguintes grandes linhas de pesquisa:

- Análise de grupos formados por entidades urbanas onde as similaridades são definidas apenas por seus atributos característicos, entre estes, atributos exclusivos de sistemas georreferenciados como as coordenadas geográficas, que podem ser objeto de uma avaliação diferenciada se comparada com outros atributos.

- Análise de grupos formados por entidades urbanas com tratamento direto da con-

tribuição das coordenadas espaciais, tratando inicialmente como próximos entre si, aqueles que estão próximos espacialmente.

- Determinação da centralidade relativa de estruturas urbanas empregando conceitos derivados de redes sociais e teoria dos grafos.

Entre os trabalhos pesquisados que guardam estreita relação com a pesquisa desenvolvida nessa dissertação é possível destacar os seguintes eixos principais:

- identificação de centralidade estrutural aplicada a localidades geográficas por uso exclusivo de atributos estruturais;
- reconstrução estrutural de hierarquias urbanas via informações de fluxos; e
- captura de agrupamentos de similaridades empregando mapas auto-organizáveis.

PORTA, CRUCITTI e LATORA (2006), em *The network analysis of urban streets: a primal approach*, incorporam conceitos de sociologia estrutural na definição de uma métrica denominada *Multiple Centrality Assessment* baseada na incorporação de quatro métricas de centralidade: *Closeness*, *Betweeneess*, *Straightness* e *Infomation Centrality*. A aplicação dessa metodologia, que emprega grafos não-direcionados, permite reduzir a complexidade de mapas de ruas em centros urbanos, substituindo-os por uma projeção simplificada que facilita a identificação de padrões semelhantes em geometrias urbanas aparentemente distintas.

Ao mudar o foco do nó para as rotas, RABINO e OCCELLI (2007) estendem o conceito dos fluxos dominantes e passam a tratar o que denominam de *dyad*. Uma *dyad* é uma rota descrita tanto pelos fluxos que entram e saem nos dois extremos da conexão e que definem relações significativas diferenciadas para cada extremo. Esses fluxos significativos são depois descritos e quantificados em função da ocorrência entre nós e a sua respectiva hierarquia na rede.

(BEHNISCH e ULTSCH, 2008) aplicam o Emergent SOM, o ESOM, em problemas de descobertas de grupos de cidades alemãs que compartilham similaridades de crescimento ou redução populacional. ESOM é uma variante do SOM que emprega análises densitométricas para uma melhor detecção de grupos na U-matrix.

Referências diversas empregam o SOM em problemas de análise de grupos sobre dados ligados à geografia urbana. Esses trabalhos, em geral, têm enfoque apenas nas similaridades de atributos de entidades urbanas sem qualquer relação com coordenadas espaciais

ou com a ação recíproca desses agentes (TUIA, KAISER, DA CUNHA e KANEVSKI (2009)).

CHENG, ZHANG, HU e LI (2007), pesquisando a formações de relações espaço-temporais em fluxos de tráfegos de veículos ao longo de vias urbanas, emprega o SOM para definir e agrupar diferentes padrões de trânsito. O método é aplicado como um sistema preditivo para as condições de tráfego de redes urbanas em cidades.

Geo-SOM, uma variante do SOM voltada especificamente para problemas georreferenciados, altera o algoritmo orginal de modo que o aprendizado e a ordenação da grade na vizinhança do neurônio vencedor do processo competitivo ocorra em duas etapas, sendo que na primeira, vizinhos são aqueles neurônios cujas coordenadas geográficas sejam próximas. Na busca pelo BMU apenas os neurônios geograficamente mais próximos o padrão de dados devem ser pesquisados. A proximidade dos candidatos à BMU é definida por uma variável k, chamada de tolerância geográfica. Fazendo $k = 0$, apenas os BMUs com coordenadas adjacentes à instância de entrada são considerados. Aumentando-se a tolerância geográfica é possível tomar em consideração, unidades que estão dentro de um raio k (medido no espaço de saída). Caso k seja da ordem do tamanho do mapa, o algorimo se reduz ao SOM padrão. Na etapa seguinte, o processo de ordenação continua em função de outros descritores das instâncias de dados (BAÇÃO, 2005). Entretanto, também aqui, o objeto final é o agrupamento de instâncias geográficas similares sem outras considerações sobre as relações mútuas a que estão sujeitas.

4 UM MODELO PARA REPRESENTAÇÃO DE REGIÕES DE INFLUÊNCIA

O modelo proposto nessa dissertação para a representação de localidades centrais e respectivas regiões de influência é inovador na medida que incorpora e estende uma estratégia simples e direta de representação de estruturas hierárquicas. Diferentemente de outras estratégias, traz para o tratamento de redes descritas por grafos a possibilidade da busca de padrões de similaridade em função de características dos descritores de entidades representadas pelos nós de rede. O método permite ainda a adoção de uma ampla gama de funções de preferência para a determinação das conexões dominantes, que podem ser calculadas de forma simples, mas com um custo computacional proporcional a $n/2$, onde n é o número de nós da rede.

Ao empregar mapas auto-organizáveis na detecção de grupos, a metodologia proposta agrega a capacidade potencial de capturar de forma não-supervisionada, estruturas de entes similares que correspondam efetivamente às distribuições estocásticas existentes no problema descrito pelos dados de entrada, possuindo como características principais:

a) A extração de uma arquitetura de rede subsidiada por dados que expressam fluxos entre os seus nós.

b) A rede extraída é hierárquica, e a hierarquia é definida pela determinação da associação ou conexão dominante para cada nó da rede.

c) A função de dominância quer seja linear, não-linear ou difusa, pode ser livremente determinada por analistas de domínio que podem assim expressar preferências e percepções diferenciadas sobre os fatores determinantes para a formação de regiões de influência.

d) Regiões de influência de cada nó da rede podem ser determinadas diretamente pela análise da sub-árvore correspondente, cujo nó de interesse é o nó raiz.

e) Fluxos de informações inter e intra-níveis, assim como inter e intra-grupos, podem ser destacados e oferecer relevantes subsídios para a captura de padrões ainda desconhecidos.

f) Os nós são classificados e formam grupos por suas características comuns, entre elas indicadores de centralidade derivados de observações das respectivas associações dominantes.

g) Mapas auto-organizáveis são empregados na detecção de partições *naturais* formadas pelos nós da rede, caso essas existam.

A fim de exemplificar e verificar a aplicabilidade da metodologia proposta, foi desenvolvida uma série de experimentos empregando dados disponíveis em repositórios públicos de dados. Atenção especial é dada àqueles dados que, pela estreita relação com o presente trabalho, formam a base de dados do estudo desenvolvido pelo Instituto Brasileiro de Geografia e Estatística - IBGE, *Regiões de Influência das Cidades*, também conhecido como REGIC IBGE (2008).

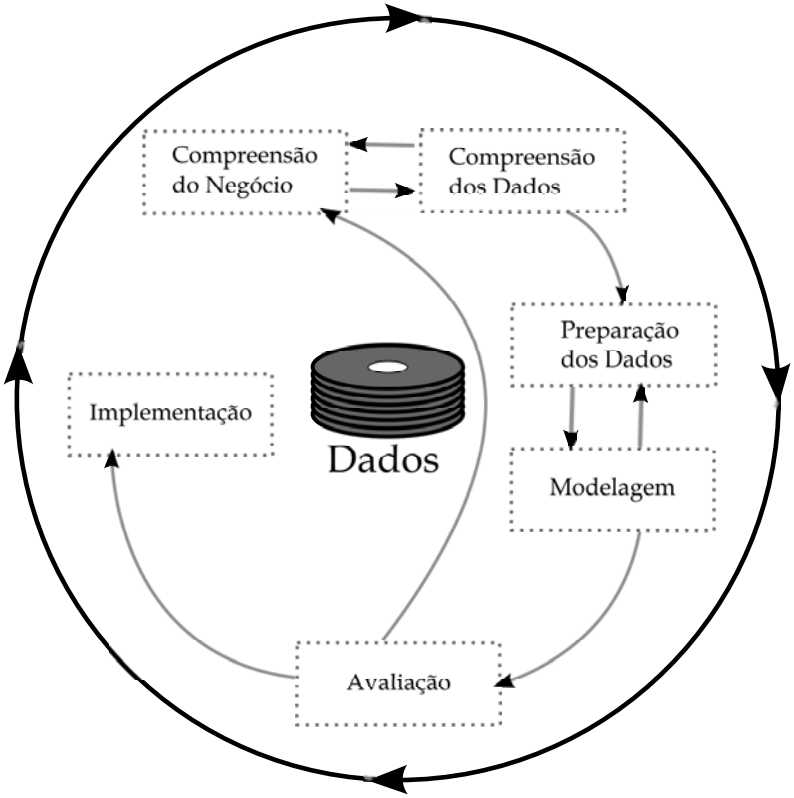


FIG. 4.1: Ciclo de vida do modelo Crisp-DM, adaptado de CHAPMAN (2000)

Onde aplicável e possível foi implementado o modelo conhecido como *Crisp Data Mining CRISP-DM* (CHAPMAN, 2000), que adota um ciclo de vida para o processo de mineração de dados composto de 6 fases: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implementação, além de pressupor

um processo iterativo e de refinamentos sucessivos até o alcance das metas estabelecidas pelo projeto (figura 4.1).

NISBET, IV e MINER (2009) estendem a fase de compreensão de dados do CRISP-DM desmembrando-a em três fases: aquisição, descrição e avaliação da qualidade dos dados.

4.1 AQUISIÇÃO E DESCRIÇÃO DE DADOS

As fontes de dados empregadas no presente trabalho foram: base de dados do estudo *Regiões de Influência das Cidades* ou IBGE, *Anuário Estatístico 2007 de transporte coletivo interestadual* compilado pela Agência Nacional de Transportes Terrestres - ANTT e o *Anuário do Transporte Aéreo - Vol. I - Dados Estatísticos 2007* produzido pela Agência Nacional de Aviação Civil a ANAC, descritas a seguir.

4.1.1 REGIÕES DE INFLUÊNCIA DAS CIDADES

Em 2007, a coordenação de geografia da diretoria de geociências do IBGE, órgão vinculado ao Ministério do Planejamento, Orçamento e Gestão - MPOG, publicou um trabalho com o título: *Regiões de Influência das Cidades*, que logo em sua introdução estabelece a pretensão de "subsidiar o planejamento estatal e as decisões quanto à localização das atividades econômicas de produção, consumo privado e coletivo, bem como prover ferramentas para o conhecimento das relações sociais vigentes e dos padrões espaciais que delas emergem". Os objetivos do trabalho desenvolvido pelo IBGE podem ser imediatamente incorporados como macro-objetivos do presente trabalho, na medida em que o mesmo pretende propor um modelo que possa ser uma ferramenta complementar em estudos de mesma natureza.

Nesse estudo, o IBGE estabelece uma hierarquia de centros urbanos e os classifica, em função de sua importância relativa, em metrópole, capital regional, centro sub-regional, centro de zona e centro local, e logo após identifica as respectivas zonas de influência de cada centro urbano.

Exemplos de resultados obtidos pelo IBGE na forma de redes urbanas e mapas de influência, podem ser vistos na figura 4.2.

Para a identificação do papel centralizador de cada centro, o IBGE investigou e mensurou diversos fatores, a exemplo da presença de gestão federal, gestão empresarial e

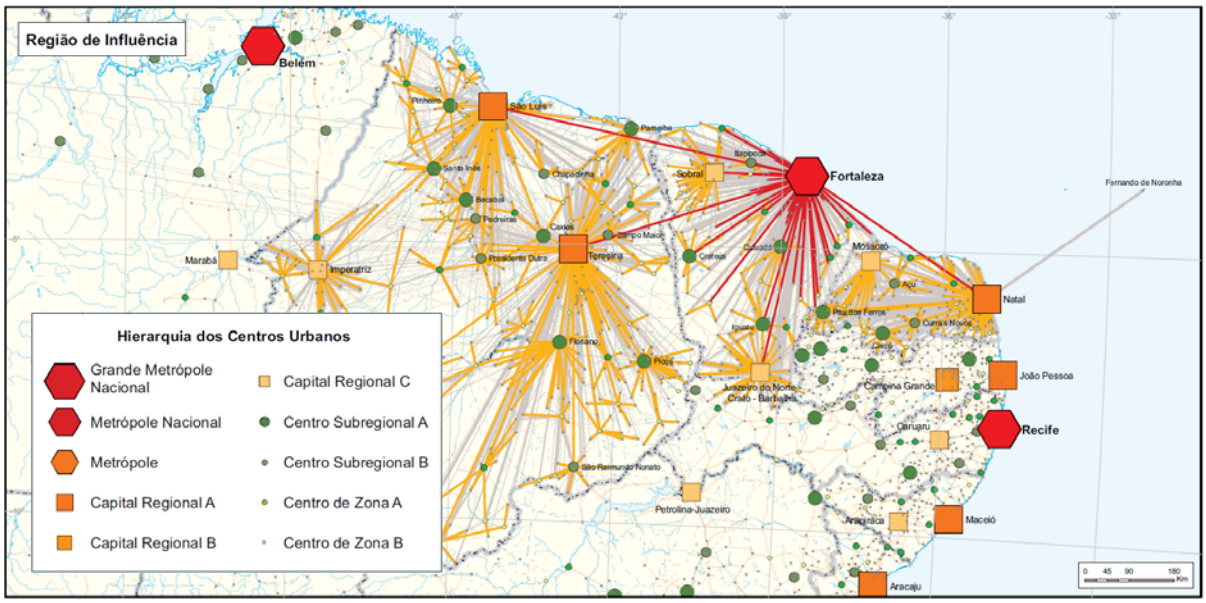


FIG. 4.2: Rede Urbana Brasileira, 2007. (IBGE, 2008)

equipamentos e serviços. As regiões de influência foram determinadas por dados de trocas comerciais e pesquisas de campo que identificaram destinos rodoviários, do lazer ao tratamento de saúde, entre outros, e os dados obtidos compuseram o cenário para o trabalho dos especialistas.

A representatividade e a extensão do estudo do IBGE ofereceram o substrato ideal aos experimentos realizados no presente trabalho, ainda que face aos objetivos secundários propostos, tenham também oferecido algumas dificuldades, a exemplo do grande número de atributos formados a partir de inferências de especialistas.

4.2 MODELAGEM

O modelo proposto procura estabelecer uma estratégia concisa e de relativamente baixa complexidade que possa orientar futuras implementações de produtos de software voltados a solução de problemas da mesma natureza a partir das seguintes premissas:

- Não são levadas em consideração as topologias das rotas físicas que ligam as cidades, mas sim aquelas decorrentes da existência de informações de fluxo de bens e passageiros. Uma cidade só está efetivamente conectada à outra se há informação significativa na dimensão estudada. Ao mesmo tempo, evita-se qualquer tipo de inferência ou adoção de modelo que trate da transferência de fluxos ao longo de

rotas e, consequentemente, são desconsideradas métricas de intermediação, caminhos mínimos, eficiência e outras análises típicas de soluções através da Teoria dos Grafos.

- Extensão da matriz de adjacências de Nystuen e Dacey, de modo que intensidade do relacionamento IR entre pares de nós seja calculada por uma combinação linear de fluxos apreciados por pesos. O objetivo é permitir que o analista de domínio tenha flexibilidade para determinar a contribuição relativa dos atributos selecionados.

- Não são consideradas as direções dos fluxos, mas somente a sua intensidade total.

- Não são feitas inferências quanto à transferência de fluxos de nós inferiores para nós superiores, o que implicaria no emprego de ponderação das contribuições de cada conexão que acomete cada nó.

- O IR representa vários papéis, o de suportar a representação estrutural da rede analisada, tornar-se uma dimensão adicional na caracterização de nós da rede, além de ser eventualmente componente em indicadores derivados de centralidade, por exemplo, o número de vezes que uma cidade é a conexão prioritária de outras.

- A rede passa a ser composta de nós enriquecidos de atributos que expressam suas características de estudo (população, PIB, formação, etc.), e aquelas relativas à centralidade.

Todo o processo de obtenção de uma rede urbana hierárquica e com nós classificados em grupos similares pode ser descrita segundo as seguintes etapas:

1 definição dos fluxos a serem considerados e valor da apreciação relativa a ser dado para cada um, por meio da determinação de pesos;

2 cálculo do índice de relacionamento para todos os pares de nós para os quais exista algum fluxo;

3 construção da matriz de adjacências e determinação das conexões prioritárias;

4 construção do vetor de entrada enriquecido com *features* de centralidade provenientes da matriz de adjacências expandida;

5 treinamento do SOM e delimitação de prováveis agrupamentos pela observação da U-matrix ou pela adoção de indicadores de qualidade;

6 alternativamente, treinamento do ESOM e delimitação de prováveis agrupamentos pela observação da P-matrix e U*-matrix;

7 clusterização dos BMUs em intervalo próximo ao número de grupos detectados na fase anterior;

8 seleção do melhor resultado e atribuição do nó ao grupo correspondente; e

9 representação descritiva e visual da rede.

Considerou-se essa metodologia capaz de oferecer subsídios para estudos de redes urbanas e sociais por agregar atributos estruturais e relacionais e simultaneamente possibilitar a visualização de processos de classificação não-supervisionada ao se procurar captar estruturas emergentes da distribuição subjacente.

4.3 PLANO DE EXPERIMENTOS

Os experimentos foram estruturados em quatro grandes grupos:

1 Experimentos introdutórios de classificação e de similaridade de resultados para dados sintéticos criados para esse estudo (3Clusters e 6Clusters) e de referência (IRIS, WINE e WDBC). Aqui foram testadas as estratégias de obtenção direta de grupos via k-means e do *clustering* dos BMUs resultantes do SOM. Procurou-se estabelecer a capacidade dos indicadores empregados, DB-index e Silhouette, de expressarem a classificação *original*.

2 Experimentos com os atributos característicos dos municípios e os indicadores de centralidade do próprio IBGE. Aqui a questão levantada é se as atribuições de centralidade e a classificação produzida por especialistas de domínio é passível de ser reproduzida pelo modelo proposto e em que extensão. As mesmas métricas do primeiro grupo de experimentos foram aplicadas. Além da amostra global com todos os municípios, amostras secundárias com cortes horizontais e verticais na massa de dados foram objeto de experimentação e registro dos resultados.

3 O terceiro grupo de experimentos é semelhante ao grupo 2 com os atributos característicos dos municípios e os indicadores de centralidade oriundos dos fluxos aéreos e terrestres adotados nesse trabalho.

4 Alternativamente e de forma exploratória para futuros trabalhos, procurou-se verificar a capacidade do SOM de extrair grupos de cidades em torno dos suas conexões dominantes diretamente, adotando-se como entrada o par origem-destino e seus atributos de fluxos.

Para os experimentos do grupo 2 foram selecionadas amostras em função de características conhecidas a priori a saber:

- A, o grupo de 802 cidades que segundo o REGIC exibiam as características mais evidentes de serem candidatas a localidades centrais;
- B, um subgrupo de A contendo 433 cidades que possuíam valores para todos os atributos, desprezando assim instâncias com valores desconhecidos ou *missing cases*;
- C, o grupo de 144 cidades que são destinos aéreos.

4.4 VISUALIZAÇÕES DO MODELO

Ao final do processo de modelagem é obtida uma relação de nós e uma relação de rotas dominantes que permite a representação georreferenciada sobre mapas.

Posteriormente, os nós são classificados em grupos segundo o critério adotado para definição do melhor grupo, seja pelo resultado das métricas ou pela observação das regiões e bordas de grupos da U-matrix. A estatística descritiva dos grupos é então empregada para a identificação das características comuns que auxiliem o entendimento e o estabelecimento de uma classificação.

As rotas dos fluxos dominantes, que oferecem suporte à representação da topologia da rede resultante, têm associadas a si um IR. Para facilitar a representação visual, destacando as diferenças, um indicador derivado foi construído a partir da discretização do IR em cinco faixas (1 a 5), denominado IREscala, com pontos de corte proporcionais à distribuição dos resultados, inversamente proporcional a intensidade, sendo 1 o relacionamento mais intenso e 5 o mais fraco.

A partir das coordenadas geográficas dos marcos geodésicos e do sistema de geocodificação do *Google Maps API* (GOOGLE INC., 2010), um aplicativo PHP (PHPGROUP, 2010) exibe sobre o mapa do Brasil os centros locais com cores proporcionais à classificação adotada e as rotas traçadas, via mapeamento de cores e espessuras proporcionais ao IREscala.

5 RESULTADOS

5.1 DESCRIÇÃO DOS DADOS

Os dados referentes ao estudo do IBGE, em formato de planilhas de Excel, estão disponibilizados no sítio do órgão, no endereço ftp://geoftp.ibge.gov.br/regic/banco_de_dados.

A base de dados do IBGE é composta de uma tabela resumo principal com 5274 registros e setenta e cinco atributos, além de tabelas secundárias relativas ao fluxo de bens e serviços obtidos por meio de questionários.

Uma análise preliminar procurou identificar as informações disponíveis e diferenciar:

- atributos definidos ou exclusivos apenas no contexto do REGIC, quer seja pelo processo de aquisição de dados, ou pela indicação de inferência de especialistas sem a devida descrição de metodologia que garanta sua reprodução;
- atributos relativos a dados primários, incluindo aqueles publicados periodicamente por algum orgão público fora do contexto do REGIC; e
- atributos contendo classificações diversas das instâncias.

Foram então selecionados os seguintes atributos para comporem uma tabela com 5274 registros (tabela 5.1).

Importantes para os experimentos de classificação e similaridade foram os atributos de centralidade produzidos pelos especialistas do domínio. Esses indicadores foram desenvolvidos com metodologia própria para cada dimensão estudada, parcialmente descrita no REGIC e incluída aqui apenas para verificação se o conjunto de atributos disponíveis seria capaz de produzir agrupamentos semelhantes empregando a estratégia de agrupamento via o k-means e SOM e clusterização dos BMUs com o k-means.

A classificação do REGIC é composta por 5 classes principais, sub-divididas em 11 sub-classes 1A, 1B, 1C, 2A, 2B, 2C, 3A, 3B, 4A, 4B e 5 (tabela 5.2).

Quanto aos fluxos, o REGIC apresenta não só informações sobre transportes aéreos e terrestres, como também o resultado de análises de questionários referentes a informações sobre os principais destinos de lazer, compras, saúde, entre outros, preenchidos pelos técnicos de agências regionais do IBGE.

TAB. 5.1: Características dos atributos selecionados do REGIC.

Código	Atributo	Unidade	Tipo	Brancos
POP	População Residente	Habitantes	Inteiro	não
AREA	Área do município	Km^2	Real	não
PIB	Produto Interno Bruto Municipal	R$	Real	não
VAGRO	Valor adicionado do setor agropecuária	R$	Real	não
VAIND	Valor adicionado do setor industrial	R$	Real	não
VASER	Valor adicionado do setor serviços	R$	Real	não
VAAP	Valor adicionado da administração pública	R$	Real	não
IMPIB	Imposto no PIB	R$	Real	não
NBANCOS	Número de bancos	Unidade	Inteiro	sim
VATIVOS	Volume de ativos	R$	Real	não
NMATRICULAS	Total de matrículas na graduação	Unidade	Inteiro	sim
DOMINIOS	Número de domínios de Internet	Unidade	Inteiro	sim
INTERNET	Presença de internet banda larga	-	sim/não	sim
CLASSE	Classificação	-	categórico	não

Em nosso estudo, apenas as informações de transporte coletivo interestadual e fluxos aéreos foram considerados face a existência dos anuários produzidos pela ANAC e ANTT, os quais permitem a reprodutibilidade e o acompanhamento evolutivo de resultados do modelo proposto. Os dados dos questionários foram descartados, por consistirem em um produto exclusivo do REGIC.

TAB. 5.2: Classificação e distribuição de objetos por classe do REGIC

classeNum	Classe	Classe Nominal	Qtd	Percentual
1	1A	Grande Metrópole Nacional	1	0.019
2	1B	Metrópole Nacional	2	0.038
3	1C	Metrópole	9	0.171
4	2A	Capital Regional A	11	0.209
5	2B	Capital Regional B	20	0.379
6	2C	Capital Regional C	39	0.739
7	3A	Centro Subregional A	85	1.612
8	3B	Centro Subregional B	79	1.498
9	4A	Centro de Zona A	192	3.641
10	4B	Centro de Zona B	364	6.902
11	5	Centro Local	4472	84.793
		Total	5274	

5.1.1 FLUXOS TERRESTRES

O fluxo correspondente ao transporte terrestre interestadual de passageiros teve como origem os dados disponíveis no sítio da ANTT em seu *Anuário Estatístico 2008 (Ano Base 2007)* (ANTT, 2008). As informações sobre o fluxo de passageiros terrestres disponibilizadas pela ANTT são baseadas em informações primárias fornecidas pelas permissionárias, formando uma tabela com 25 atributos e 65832 registros, entre os quais foram selecionados apenas aqueles descreviam o fluxo de passageiros e exibidos na tabela 5.3, aos quais foram acrescidos:

1 cod7 da cidade de origem, código de 7 dígitos que identifica os municípios brasileiros,

2 cod7 da cidade de destino.

O sistema de codificação dos municípios adotado pela ANTT, não compatível com aquele empregado pelo IBGE, exigiu um trabalho intenso de pré-processamento para limpeza, uniformização, recuperação e associação com o sistema de códigos de municípios adotado pelo instituto (tabela 5.4).

TAB. 5.3: Características de atributos selecionados do Anuário Estatístico 2008 - ANTT.

Ordem	Atributo	Unidade	Tipo	Brancos
1	Cidade Origem	-	Cadeia de Caracteres	não
2	Cidade Destino	-	Cadeia de Caracteres	não
3	Passageiros ida	unidade	Inteiro	não
4	Passageiros volta	unidade	Inteiro	não

TAB. 5.4: Extrato de tabela de rotas terrestres, adaptado de ANTT (2008).

no_razao_social	co_linha	no_linha	no_cidade_origem	no_cidade_destino	nr_passageiro_ida	nr_passageiro_volta
A. CANDIDO & CIA LTDA (EXPRESSO NACIONAL DE LUXO)	13042	CAMPINA GRANDE (PB) - SURUBIM (PE)	GRAVATA DO IBIAPINA (PE)	SURUBIM (PE)	554	511
A. CANDIDO & CIA LTDA (EXPRESSO NACIONAL DE LUXO)	13042	CAMPINA GRANDE (PB) - SURUBIM (PE)	GRAVATA DO IBIAPINA (PE)	VERTENTES (PE)	526	475
A. CANDIDO & CIA LTDA (EXPRESSO NACIONAL DE LUXO)	13042	CAMPINA GRANDE (PB) - SURUBIM (PE)	GRAVATA DO IBIAPINA (PE)	TAQUARITINGA DO NORTE (PE)	526	446
A. CANDIDO & CIA LTDA (EXPRESSO NACIONAL DE LUXO)	13042	CAMPINA GRANDE (PB) - SURUBIM (PE)	VERTENTES (PE)	SURUBIM (PE)	695	449
A. CANDIDO & CIA LTDA (EXPRESSO NACIONAL DE LUXO)	13042	CAMPINA GRANDE (PB) - SURUBIM (PE)	TAQUARITINGA DO NORTE (PE)	SURUBIM (PE)	486	402
A. CANDIDO & CIA LTDA (EXPRESSO NACIONAL DE LUXO)	13042	CAMPINA GRANDE (PB) - SURUBIM (PE)	TAQUARITINGA DO NORTE (PE)	VERTENTES (PE)	586	477
A. CANDIDO & CIA LTDA (EXPRESSO NACIONAL DE LUXO)	13042	CAMPINA GRANDE (PB) - SURUBIM (PE)	CAMPINA GRANDE (PB)	GRAVATA DO IBIAPINA (PE)	580	612
A. CANDIDO & CIA LTDA (EXPRESSO NACIONAL DE LUXO)	13042	CAMPINA GRANDE (PB) - SURUBIM (PE)	CAMPINA GRANDE (PB)	SURUBIM (PE)	1438	1288
A. CANDIDO & CIA LTDA (EXPRESSO NACIONAL DE LUXO)	13042	CAMPINA GRANDE (PB) - SURUBIM (PE)	CAMPINA GRANDE (PB)	VERTENTES (PE)	852	919
A. CANDIDO & CIA LTDA (EXPRESSO NACIONAL DE LUXO)	13042	CAMPINA GRANDE (PB) - SURUBIM (PE)	CAMPINA GRANDE (PB)	TAQUARITINGA DO NORTE (PE)	675	719
A. CANDIDO & CIA LTDA (EXPRESSO NACIONAL DE LUXO)	13042	CAMPINA GRANDE (PB) - SURUBIM (PE)	CAMPINA GRANDE (PB)	QUEIMADAS (PB)	889	864
A. CANDIDO & CIA LTDA (EXPRESSO NACIONAL DE LUXO)	13042	CAMPINA GRANDE (PB) - SURUBIM (PE)	CAMPINA GRANDE (PB)	RIACHO GRANDE (PB)	804	1757
A. CANDIDO & CIA LTDA (EXPRESSO NACIONAL DE LUXO)	13042	CAMPINA GRANDE (PB) - SURUBIM (PE)	QUEIMADAS (PB)	GRAVATA DO IBIAPINA (PE)	632	563

As rotas terrestres associadas às companhias com sedes em cidades distintas produziram casos onde o mesmo par origem-destino era eventualmente citado duas vezes, pois o mesmo caminho era percorrido por empresas diferentes em sentidos diferentes (ex. Rio-São Paulo com fluxos de ida e volta pela empresa A e São Paulo-Rio com fluxos nos dois sentidos pela empresa B) (figura 5.1). Esses casos foram detectados e incorporados aos fluxos não-direcionados representativos das cidades correspondentes.

Após o processo de criação de vínculos entre as cidades do estudo do IBGE, todas as rotas de todas as empresas que convergiam para o mesmo par origem-destino foram consolidadas em um fluxo único não-direcionado.

Nem todas as cidades constantes da base do REGIC tiveram uma contraparte de informação sobre fluxos terrestres devido ao fato das informações oriundas da ANTT se referirem ao transporte inter-estadual e a informação sobre transporte intermunicipal ser fragmentada, quando existente, e sob o controle de órgãos estaduais e municipais.

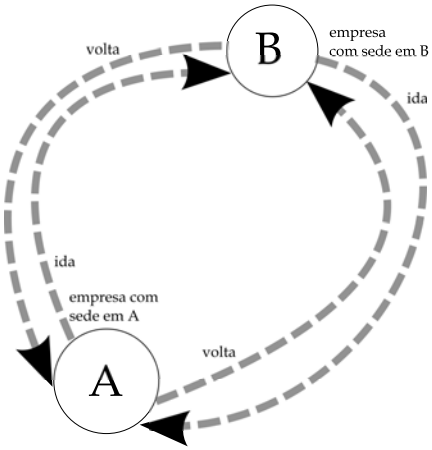


FIG. 5.1: Rotas terrestres entre duas cidades servidas por empresas distintas

5.1.2 FLUXOS AÉREOS

As informações sobre o tráfego aéreo incluem o número de passageiros e o total de carga transportada entre os principais aeroportos brasileiros. Essas informações tiveram como origem o *Anuário do Transporte Aéreo - Vol. I - Dados Estatísticos 2007* (ANAC, 2007), disponível no sítio da ANAC, autarquia federal vinculada ao Ministério da Defesa e responsável pela regulação do mercado de aviação civil no Brasil.

TAB. 5.5: Extrato de tabela de rotas aéreas, adaptado de ANAC (2007).

3.2 - TRÁFEGO DOMÉSTICO							
a – TRÁFEGO POR ORIGEM E DESTINO - ANO 2008							
ORIGEM		DESTINO		LIGACOES	PAX	CARGA (KG)	CORREIO
PA	**AGROPALMA**	PA	INT. VAL DE CAES	2	15	0	0
PA	**ALMEIRIM**	PA	INT. VAL DE CAES	2	17	0	0
MT	**ALTA FLORESTA**	AL	CAMPO DOS PALMARES	24	5	56	0
		SP	CONGONHAS	215	1,178	18,420	0
		DF	INT. DE BRASILIA	212	1,559	23,114	0
		SP	INT. DE SAO PAULO	9	256	4,336	0
		BA	INT. DOIS DE JULHO	32	47	866	0
		RS	INT. SALGADO FILHO	29	15	253	0
		MT	MARECHAL RONDON	366	21,184	240,163	0
		MT	SINOP	70	3,118	34,724	0

Como pode ser visto na tabela 5.5, a base de dados sobre o fluxo de cargas e passageiros aéreos tem a sua codificação de rotas centrada em nomes de domínio público para os principais aeroportos brasileiros, não existindo referências a um sistema de códigos para os destinos capaz de permitir a vinculação com a codificação de sete dígitos adotada pelo

IBGE. Procurou-se então extrair a relação de municípios a partir da associação com a *Lista de Aeródromos Públicos*, disponível no site da Agência. Entretanto, curiosamente, a descrição dos aeródromos não seguia a mesma descrição dos destinos empregada no *Anuário do Transporte Aéreo* desenvolvido pelo mesmo órgão. A fase de pré-processamento envolveu então o levantamento e a uniformização de todos os municípios associados aos aeródromos, emparelhamento por nome de municípios e incorporação da codificação IBGE correspondente.

Após o processo de vinculação de cidades origem e destino aéreos com as cidades do estudo do IBGE, todas as rotas de todas as empresas que convergiam para o mesmo par origem-destino foram consolidadas em um fluxo único não-direcionado e adotadas as colunas com dados anualizados para o ano de 2007, conforme descrito na tabela 5.6:

TAB. 5.6: Características de atributos selecionados do Anuário do Transporte Aéreo - 2007 - ANAC.

Ordem	Atributo	Unidade	Tipo	Brancos
1	Cidade Origem	-	Cadeia de Caracteres	não
2	Cidade Destino	-	Cadeia de Caracteres	não
3	Passageiros	unidade	Inteiro	não
4	Carga	Kg	Inteiro	não

Aos quais foram acrescidos:

1 cod7 da cidade de origem, código de 7 dígitos que identifica os muncípios brasileiros,

2 cod7 da cidade de destino.

Das 5274 cidades constantes da base do REGIC, apenas 144 possuem aérodromos e consequentemente informação sobre fluxos áereos.

5.1.3 LOCALIZAÇÃO GEOGRÁFICA

Para o suporte à representação visual dos resultados sobre uma estrutura georreferenciada realística, adotou-se como solução simplificadora as coordenadas geográficas dos marcos geodésicos associados a cada município brasileiro. Esses marcos não necessariamente se inserem na geografia urbana daquilo que hoje se conhece como *a cidade*, mas são

suficientemente precisos para o escopo do presente trabalho. Os dados foram gentilmente cedidos pelo IBGE após consulta ao serviço de atendimento do órgão e com os seus atributos descritos na tabela 5.7.

TAB. 5.7: Coordenadas de marcos geodésicos - IBGE.

Ordem	Atributo	Unidade	Tipo	Brancos
1	cod7	-	código de 7 dígitos	não
2	Município	-	Cadeia de Caracteres	não
3	Latitude do marco geodésico	graus	Real	não
4	Longitude do marco geodésico	graus	Real	não
5	Área ocupada pelo município	Km^2	Real	não

5.2 INTEGRAÇÃO E CARACTERIZAÇÃO DOS DADOS

Os dados oriundos das 4 tabelas devidamente normalizadas pela chave primária contendo o código da cidades foram integrados em uma base *MySql* para produção de consultas e extração de amostras.

O conjunto de dados foi caracterizado estatisticamente, primeiro em função da classificação proposta pelo IBGE, e subsequente mente em relação às amostras extraídas para os experimentos.

Na fase de exploração dos dados, dezenas de testes foram conduzidos para determinar o impacto extremos, normalizações e de diferentes amostragens. As normalizações experimentadas foram a por intervalo, logarítmica e a gaussiana também conhecida como padronização.

Os dados uma vez integrados, deram origem à amostras que buscavam agregar instâncias segundo alguns grupos inicialmente destacados no estudo do IBGE, entre eles aquele formado por um contingente de 802 municípios que exibiam de antemão potencial para se destacarem como localidades centrais.

Todo o universo de municípios e os vários subconjuntos foram estudados segundo a Análise de Componentes Principais, à procura de evidência de grupos a luz dos seus dois maiores componentes.

5.3 CARACTERIZAÇÃO DOS DADOS

Todos os atributos do REGIC foram analisados independentemente e registradas as suas estatísticas descritivas. Algumas são mostradas na tabela 5.8 a título de exemplo.

TAB. 5.8: Descritores estatísticos de alguns atributos do REGIC

	POP	AREA	PIB	VAGRO	VAIND	NBANCOS
Média	34886.18	1611.92	407136.8	19939.8938	102259.4	2.910607
Erro padrão	4852.0587	79.58504	92680.82	472.183812	21450.48	0.156449
Mediana	10281	436.83	47475.63	10591.6335	4551.299	2
Modo	4014	146.124	12087.97	N/A	N/A	1
σ	352367.71	5779.649	6730695	34291.0795	1557783	8.818214
σ^2	1.242E+11	33404345	4.53E+13	1175878132	2.43E+12	77.7609
Curtose	2056.8668	238.0649	3052.159	94.5842605	3237.331	1440.061
Assimetria	41.322679	13.08294	51.34242	7.61159945	52.20401	34.01462
Intervalo	19591467	159693.1	4.23E+08	735117.622	1E+08	404
Mínimo	804	2.859	3499.177	0	231.253	1
Máximo	19592271	159695.9	4.23E+08	735117.622	1E+08	405

Após a sua aquisição e preparação, os dados provenientes do REGIC, ANAC e ANTT foram normalizados por intervalo a fim de permitir a comparação adimensional de suas distribuições e descritores estatísticos. Para métricas euclidianas de distância, a normalização por intervalo ajusta o intervalo de valores entre zero e 1, mantendo as distâncias relativas entre as instâncias do conjunto de dados.

A normalização por intervalo para um grupo de valores x_1 a x_n é descrita por:

$$x^{'} = \frac{x_i - min(x_1..x_n)}{max(x_1..x_n) - min(x_1..x_n)} \tag{5.1}$$

Gráficos do tipo box plot (Figuras 5.2, 5.3 e 5.4) indicam uma concentração massiva de observações em torno de valores provenientes dos pequenos municípios brasileiros. Sob a ótica dessa descrição e, segundo estratégias padrão de tratamento de ruídos de dados

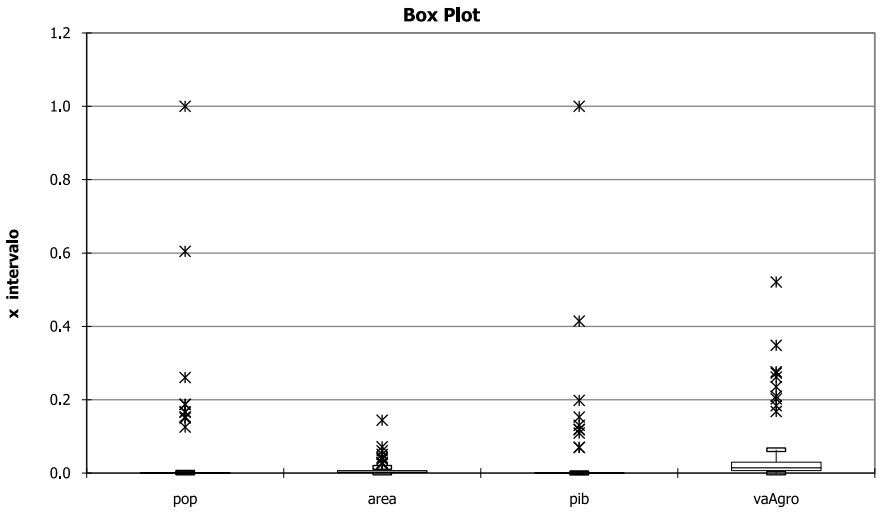


FIG. 5.2: Box Plot para os atributos população (pop), área (AREA), PIB (PIB) e valor agregado da agricultura (VAGRO).

extraídos de amostragens aleatórias, os trinta maiores municípios seriam extremos e potencialmente candidatos a serem descartados. Aqui, extremos representam a informação de alto valor agregado, pois carregam o conhecimento sobre a distribuição das grandes metrópoles e de seus relacionamentos ao longo da rede urbana. É assumido implicitamente que durante o processo de aquisição e tratamento de dados em sua fonte original, no caso IBGE, ANAC e ANTT, houve o tratamento adequado de casos extremos cuja causa esteja atrelada ao processo de amostragem ou de entrada de dados.

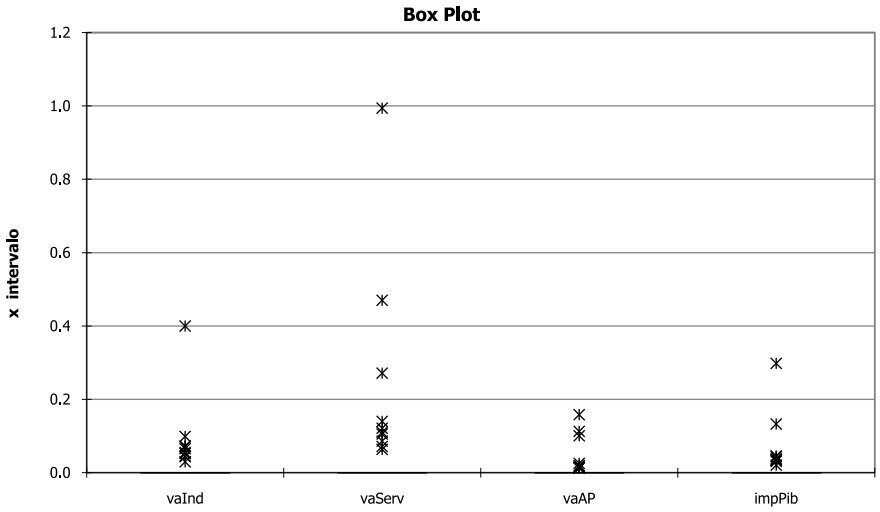


FIG. 5.3: Box Plot para valor agregado industrial (VAIND), de serviços (VASERV), da administração pública (vaAP) e impostos no PIB (impPib)

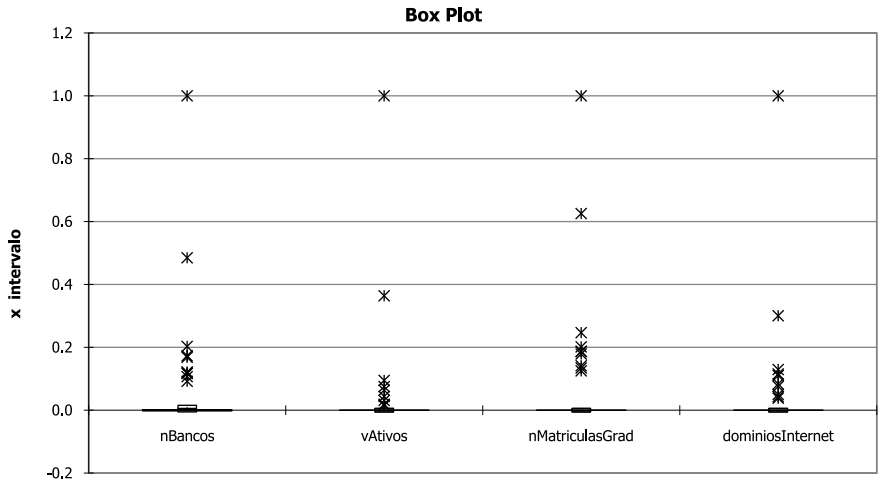


FIG. 5.4: Box Plot para número de bancos (NBANCOS), volume de ativos (VATIVOS), matrículas de graduação (NMATRICULAS) e domínios na internet (DOMINIOS)

5.3.1 DISTRIBUIÇÃO E GRUPOS NO REGIC

O Brasil, a exemplo de diversos países do mundo, possui um padrão de distribuição de densidade populacional fortemente concentrado em suas grande metrópoles. Em 2000, segundos dados do IBGE, 50 cidades representando $0,95\%$ do universo total de municípios concentravam em si, $49,28\%$ de toda população brasileira. O gráfico da figura 5.5 demonstra que a curva de distribuição da população brasileira segue um padrão de decaimento exponencial quando comparados número de cidades e número de habitantes.

A forte concentração populacional em relativamente poucas cidades brasileiras produz uma distribuição com queda exponencial dos valores referentes as cidades pequenas que representam mais de 95% do universo. Além disso, uma enorme variância é encontrada em todos os espectros analisados. Para processos de aprendizado baseados em redes neurais, esse tipo de distribuição ocasiona uma dificuldade complementar já que pouquíssimas instâncias contêm informação altamente relevante em relação ao espaço amostral. Heurísticas de compensação dessa deficiência crítica para a captura de conhecimento proveniente de padrões esparsos incluem metodologias de normalização do espaço amostral, duplicação de instâncias e a partição do conjunto de dados em dois sub-conjuntos, um de extremos e outro para valores mais próximos das tendências estatísticas centrais, cada um sujeito a análises independentes e posterior unificação de resultados.

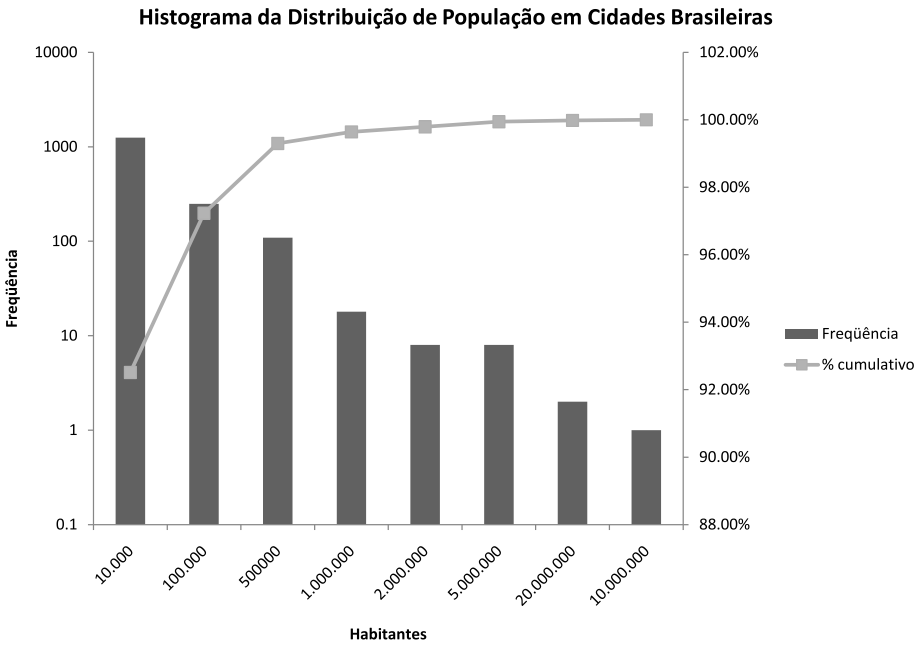


FIG. 5.5: Distribuição populacional das cidades brasileiras.

Em nossos testes comparamos a normalização por intervalo e a normalização logarítmica e o seu impacto na percepção de grupos na U-matrix resultante do SOM. A normalização logarítmica provoca uma re-distribuição da concentração dos dados originais e uma boa aproximação da curva normal para a maioria dos atributos (figura 5.6).

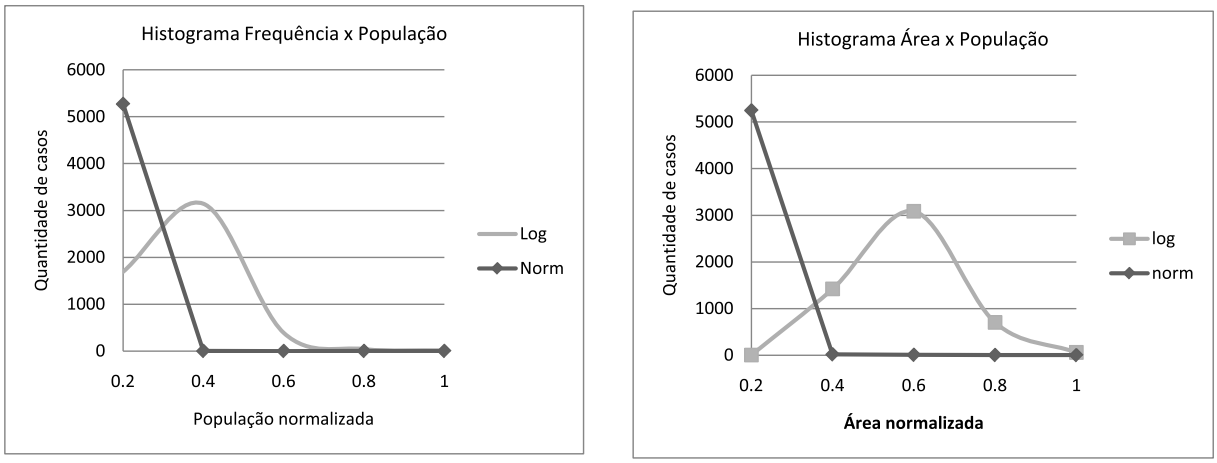


FIG. 5.6: Distribuição original normalizada por intervalo e logarítmica.

Outra descrição cabível do espaço amostral é o Coeficiente de Correlação entre o conjunto de multi-atributos representativos das instâncias associadas a cada município. O Coeficiente de Correlação expressa o grau de relacionamento linear entre 2 variáveis, assumindo valores entre -1 e 1, e é comumente calculado em diversos pacotes comerciais

segundo a expressão:

$$Correl(X,Y) = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2(y-\overline{y})^2}} \tag{5.2}$$

Valores positivos próximos a 1 indicam uma forte relação linear positiva, enquanto que uma relação linearmente negativa é indicada por um coeficiente próximo ou igual a -1. Entretanto, o fator de correlação deve ser usado com cautela em distribuições complexas e fortemente não lineares. A existência de linearidade entre dois fatores que contribuem para uma distribuição amostral não implica numa relação casual entre esses atributos, mas diversos autores advertem sobre o impacto aditivo em sistemas preditivos da contribuição de duas ou mais variáveis altamente relacionadas (PYLE, 1999). A tabela 5.9 exibe os fatores de correlação para os principais atributos descritores dos municípios brasileiros.

TAB. 5.9: Fatores de Correlação para os dez primeiros atributos do REGIC

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
[1] POP	**1.00**										
[2] AREA	0.05	**1.00**									
[3] PIB	**0.97**	0.04	**1.00**								
[4] VAAGRO	0.22	0.20	0.20	**1.00**							
[5] VAIND	**0.93**	0.03	**0.97**	0.18	**1.00**						
[6] VASERV	**0.97**	0.04	1.00	0.20	**0.95**	**1.00**					
[7] VAAP	0.76	0.06	0.76	0.23	0.63	0.80	**1.00**				
[8] IMPIB	**0.97**	0.04	**1.00**	0.18	**0.97**	**0.99**	0.72	**1.00**			
[9] NBANCOS	**0.96**	0.05	**0.96**	0.28	**0.93**	**0.95**	0.71	**0.95**	**1.00**		
[10] NBNAC	0.25	0.07	0.19	0.42	0.21	0.19	0.19	0.18	0.43	1.00	
[11] VATIVOS	0.83	0.04	0.92	0.16	0.89	**0.93**	0.72	**0.90**	0.84	0.10	1.00

5.3.2 ANÁLISE DE COMPONENTES PRINCIPAIS

Como técnica complementar e preparatória visando a identificação da existência de grupos, podemos dispor da Análise de Componentes Principais, ou *Principal Component Analysis (PCA)*. O PCA é uma técnica de redução dimensional baseada na projeção dos dados de seu espaço dimensional original sobre um novo espaço orientado na direção das maiores variâncias. O método gera um novo conjunto de vetores ortogonais entre si, os componentes principais, cada um resultante de uma combinação linear dos atributos originais. Uma vez definidos, os componentes principais são então ordenados de forma

decrescente em função da respectiva variância e, caso a soma total dos primeiros dois ou três componentes ultrapasse um determinado patamar, por exemplo 80%, são considerados uma boa representação do espaço de entrada para o uso em traçados de gráficos de duas e três dimensões. Um dos usos para o método é permitir um primeiro *insight* sobre a distribuição e a presença de grupos.

Como exemplo, submetemos o conjunto sintético 3Clusters ao PCA e registramos os resultados na tabela 5.10. Como descrito anteriormente, o 3Clusters é uma amostra gerada sinteticamente composta por 3 grupos tridimensionais bem separados e aproximadamente esféricos, mas com densidades variadas. O PCA resulta em 2 componentes capazes de explicar 96.4% da variância total e cujo gráfico bi-dimensional demonstra o potencial da projeção em representar o espaço de entrada de dimensão superior (figura 5.7). O PCA é particularmente indicado para problemas explicados pela combinação linear dos seus descritores ou atributos.

TAB. 5.10: Componentes principais e variância para amostra sintética com 3 grupos 3D - 3Clusters.

Componente	% de σ^2
1	84.066
2	12.332
3	3.602

Quando aplicado a um universo representativo dos 28 atributos dos 5274 municípios, os componentes principais resultantes do PCA (Tabela 5.11), sugerem serem capazes de produzir uma excelente representação acima de 4 ou 5 dimensões, o que inviabiliza a sua aplicabilidade para a representação visual em 2 dimensões. Entretanto, essa redução é usualmente aplicada e altamente recomendável como estratégia de simplificação de problemas complexos de análise de dados.

No gráfico resultante da projeção das instâncias sobre as duas dimensões formadas pelos componente principais 1 e 2, os extremos como Rio, São Paulo e Brasília e outras grandes metrópoles aparecem bem destacadas. Essas metrópoles são tratadas como grupos na classificação original, entretanto a separação representada pelos dois maiores componentes não é capaz de se aproximar sequer razoavelmente da previsão de 11 grupos (figura 5.8).

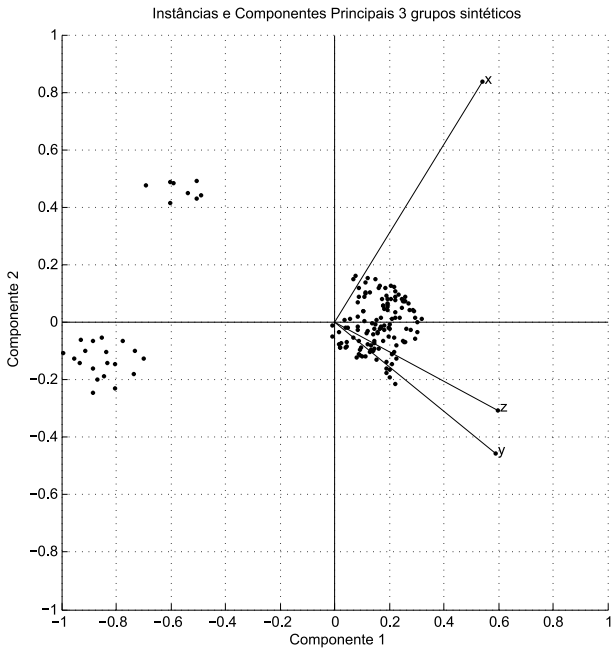


FIG. 5.7: Projeção de amostra tridimensional sobre campo bidimensonal dos componentes principais 1 e 2.

A normalização logarítmica aplicada antes do PCA desembaraça parcialmente os grupos formados pelas classes mais numerosas e de valores menos expressivos, entretanto torna a separação geral entre classes mais difusa (figura 5.8).

5.3.3 SILHOUETTES

Como citado na seção 2.2.2.2, o *Silhouettes* é um indicador e um suporte visual para mensuração da pertinência de instâncias às classes que lhes foram associadas. Adotamos aqui o termo *Silhouettes* para referenciar a representação visual resultante da aplicação do método e *Silhouette* ou *Si* na citação sobre o indicador numérico associado a um grupo ou partição.

Tanto o *Silhouettes* como *Si* foram determinados para os seguintes grupos:

- Dados sintéticos, os dois com boa separação inter-grupo e alta coesão intra-grupo. O primeiro, denominado 3Clusters, composto de 3 grupos, 88 instâncias, 2 dimensões e densidade variável (figura 5.10 a). O segundo denominado 6Clusters, composto por 6 grupos, 120 instâncias, 3 dimensões e igual distribuição de instâncias pelas classes (figura 5.10 b).

- Dados de referência disponíveis no UC Irvine Machine Learning Repository (FRANK,

TAB. 5.11: Dez primeiros componentes principais, % σ^2 para dados REGIC

Componente	% σ^2	% acumulado de σ^2
1	61.663	61.663
2	8.662	70.325
3	5.642	75.967
4	5.189	81.156
5	3.207	84.363
6	2.16	86.523
7	2.038	88.561
8	1.724	90.285
9	1.409	91.694
10	1.143	92.837
11	1.119	93.956
12	1.04	94.996
13	0.844	95.84
14	0.721	96.561

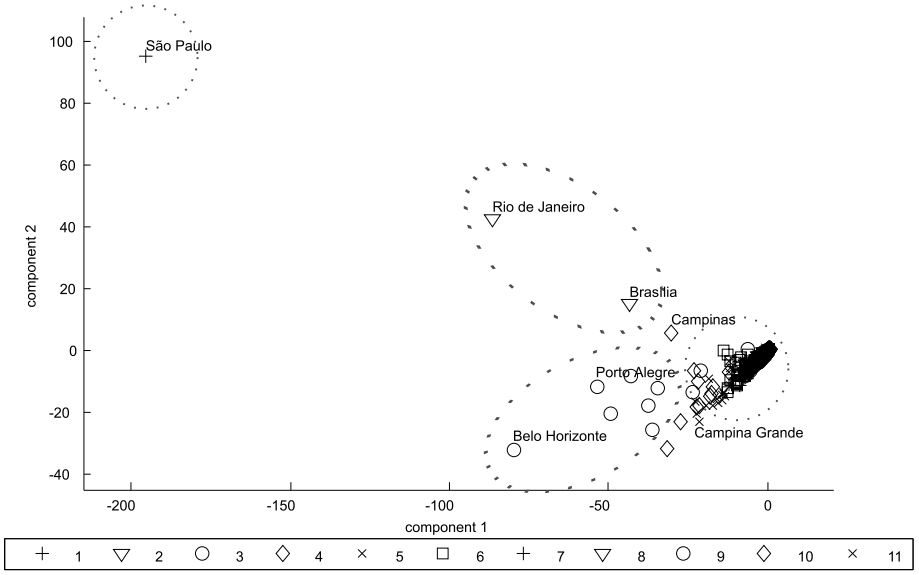
2010). Iris com 3 grupos, 4 atributos e 150 instâncias (figura 5.10 c). Wine com 3 classes, 13 atributos e 178 instâncias (figura 5.10 d). WDBC com 2 grupos, 569 instâncias e 32 atributos (figura 5.10 e).

- Dados do REGIC, com 5274 instâncias, 11 classes e 28 atributos, distribuição exponencial pelas classes (figura 5.10 f).

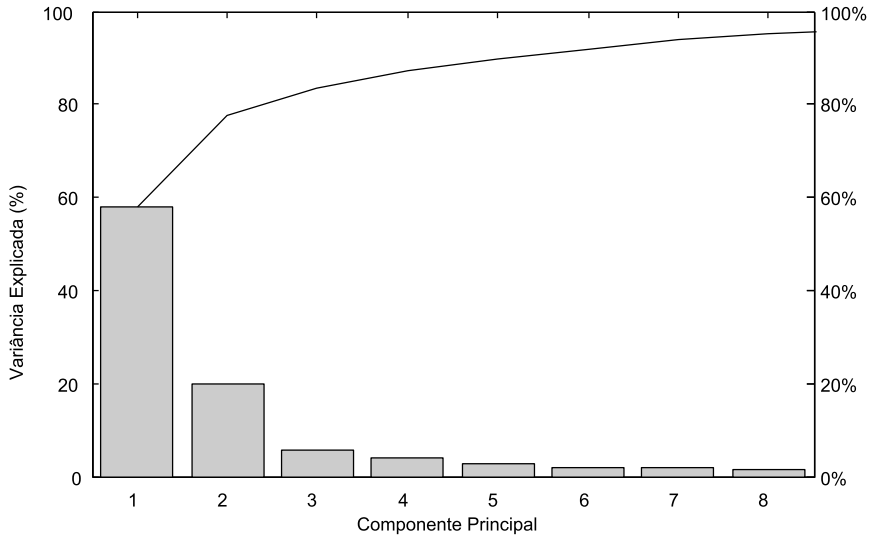
Uma atenção especial deve ser dada ao valor de Si quando expresso pela média global de todas instâncias e o representativo da média dos grupos. No caso particular dos dados do REGIC, a alta densidade de pontos do grupo mais homogêneo desloca a nota global para essa distribuição (tabela 5.12).

5.4 SOM, K-MEANS E SIMILARIDADE

Nesta secão são apresentados os resultados relativos a dados sintéticos, de referência e REGIC que foram submetidos ao SOM e posteriormente ao k-means. Os BMUs resultantes do SOM foram agrupados ao longo de um intervalo de número de grupos que era função da quantidade conhecida pela classificação original de cada amostra. A cada novo grupamento gerado foram registrados o *Davies-Bouldin index* e o *Silhouette Index*. Ao final foram traçados gráficos com os indicadores e comparou-se o Silhouette original com o Silhoutte determinado pelos melhores índices.



(a)



(b)

FIG. 5.8: (a) Grupos originais e sua distribuição segundo os principais componentes; (b) Percentual de variância explicada por componente.

O SOM foi treinado em sua configuração padrão adotada pelo pacote SOM Toolbox para uma varredura generalizada com os seguintes parâmetros:

- o tamanho do mapa é função do número de instâncias, enquanto que a razão entre largura-altura obedece a razão entre os dois maiores componentes resultantes do

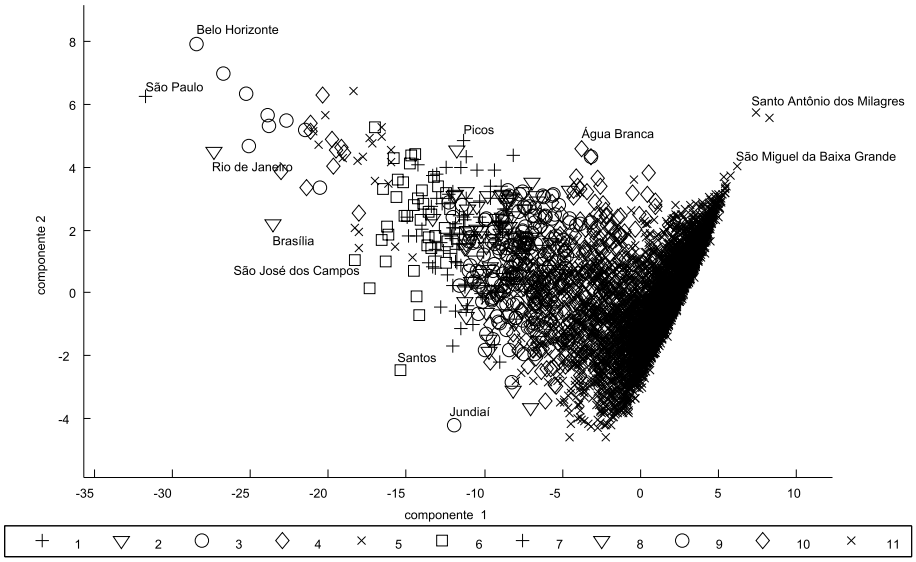


FIG. 5.9: Grupos e componentes principais para dados REGIC como normalização logarítmica.

PCA sobre o dados de entrada;

- células hexagonais;
- malha folha retangular;
- inicialização linear (lininit), baseada nos componentes principais dos dados de treinamento;
- função gaussiana de vizinhança com raio inicial definido por (a) $raio\_ini = max(1, maior\_lado\_mapa/8)$ e raio final por

TAB. 5.12: Si médio global e médio de grupos.

Conjunto de dados	Si dados	Si grupos
3Clusters	0.95	0.94
6Clusters	0.89	0.89
Iris	0.65	0.65
Wine	0.24	0.27
WDBC	0.60	0.49
REGIC	0.45	-0.22

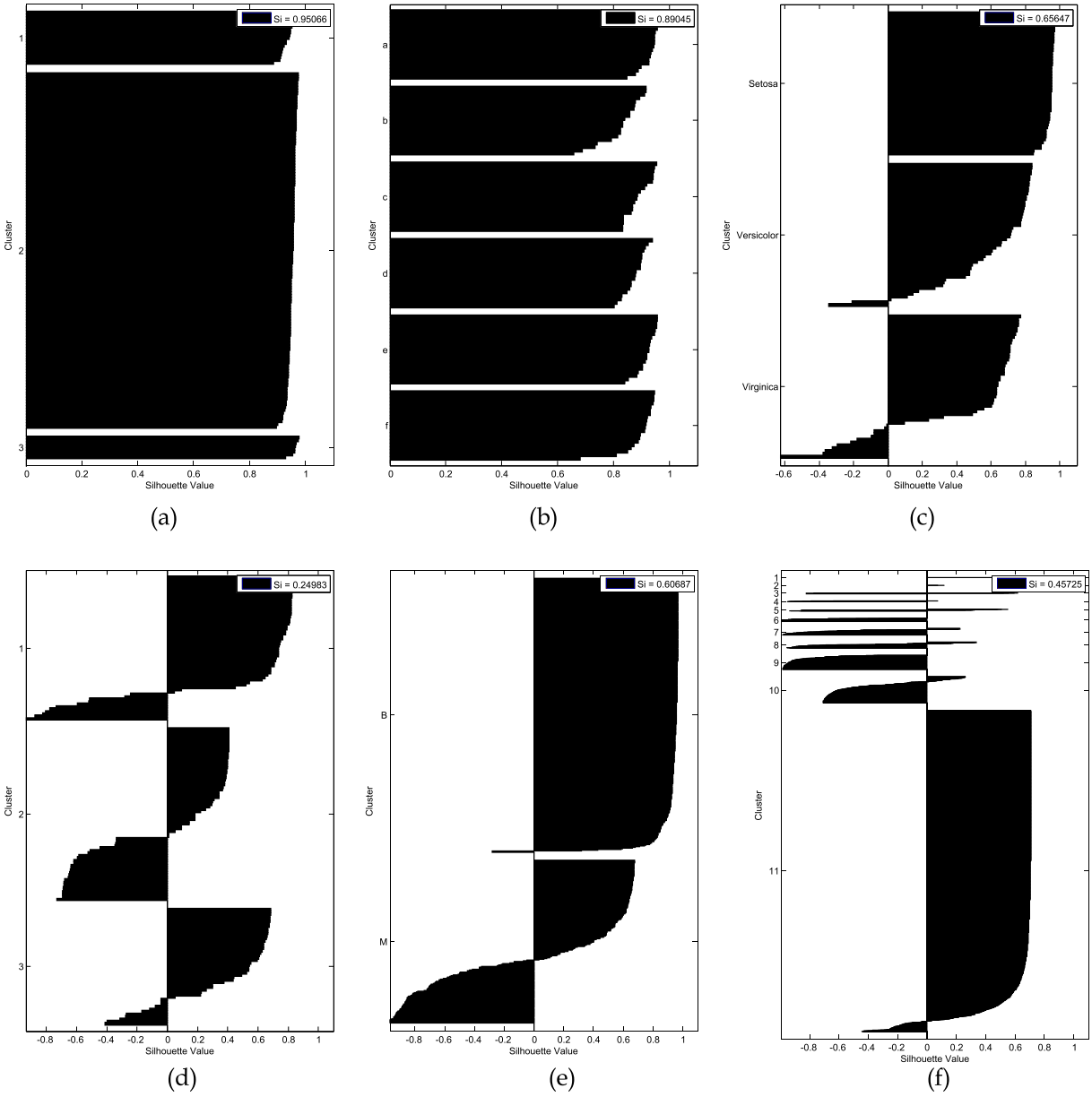


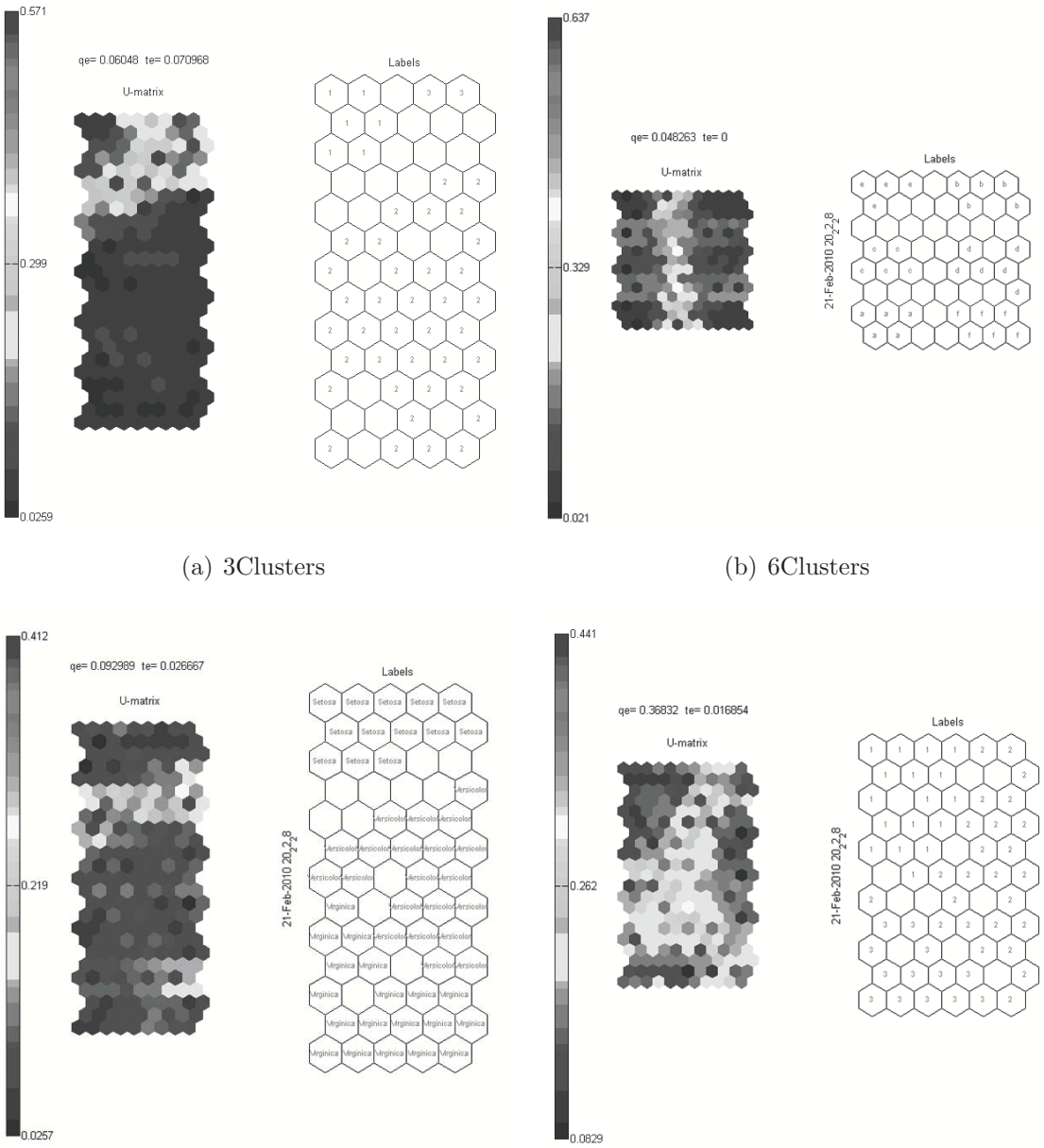
FIG. 5.10: Si referentes a conjunto de dados examinados: 3Clusters (a); 6Clusters (b); Iris (c); Wine (d); WDBC (e); e REGIC (f).

(b) $raio\_fim = max(1, maior\_lado\_mapa/4)$;

- treinamento em lote;
- taxa de aprendizado de 0.5 para ajuste inicial e 0.05 para ajuste fino;
- tempo de treinamento total, ou número de épocas de treinamento, é de 20 vezes a razão entre o número de unidades de neurônios na malha e o tamanho da amostra sendo treinada. 20% desse tempo aplicado na fase de ajuste bruto e 80% para o

ajuste fino.

Alguns gráficos de U-matrix resultantes do SOM para conjuntos sintéticos e de referência são apresentados na figura 5.11.



(a) 3Clusters (b) 6Clusters

(c) Iris (d) Wine

FIG. 5.11: U-matrix resultante de SOM.

Para cada conjunto de dados, registrou-se a sequência de varredura ao longo do intervalo de número de grupos ou *clusters* pesquisados. Em seguida, foi representado graficamente o *Silhouettes* da melhor formação segundo os indicadores adotados (figura 5.12).

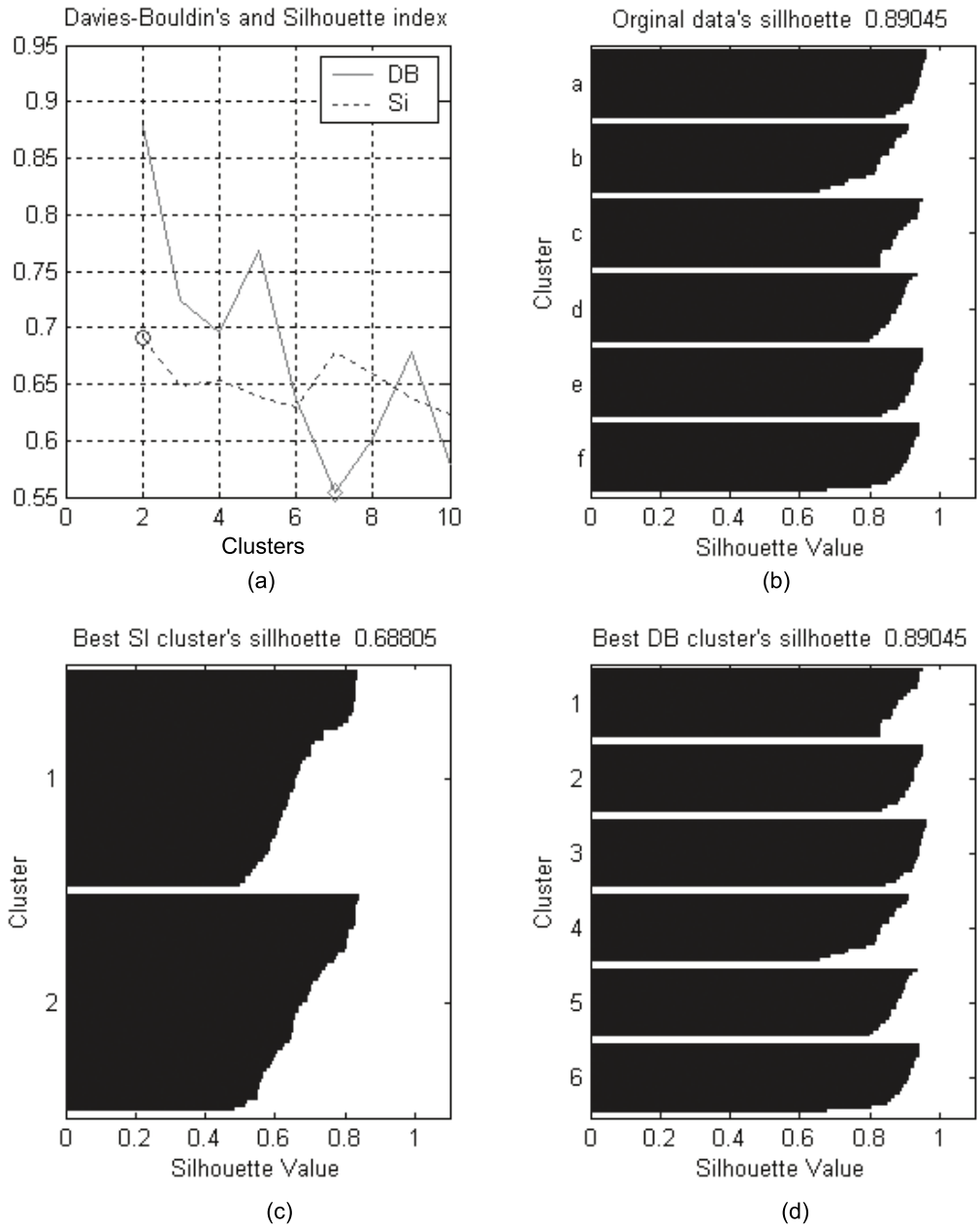


FIG. 5.12: Ciclo de busca de agrupamento ótimo para dados 6Clusters, série clusters DB e SI index (a), Silhoettes original (b), Silhouetes de melhor partição segundo Si index (c), Silhouetes de melhor partição segundo DB index (d).

O conjunto do REGIC foi apresentado ao SOM dentro das mesmas condições dos grupos de testes, variando-se apenas o tamanho da malha à procura de evidências de emergência de estruturas eventualmente obscurecidas em malhas menores (tabela 5.13).

TAB. 5.13: Resultados quantitativos de agrupamentos SOM e k-means

malha	EQ	ET	Clusters DB	Si	Clusters Si	Si
33x11	0.0361	0.0316	3	0.771	2	0.879
50x18	0.0280	0.0330	2	0.895	2	0.895
75x27	0.0228	0.0229	3	0.799	2	0.891
80x50	0.0180	0.0220	2	0.786	3	0.895

5.5 APLICANDO O MODELO

5.5.1 CONSTRUINDO A REDE

Para aplicar o modelo foram selecionadas as 50 maiores cidades brasileiras segundo o critério população e suas respectivas conexões.

Na parte de terrestre, de um conjunto de 31594 rotas terrestres relativas ao transporte de passageiro interestadual, 5540 envolvem as 50 maiores cidades. Segue-se aqui o padrão do REGIC que engloba como uma unidade o grupo de municípios limítrofes às grandes metrópoles que formam regiões conhecidas, a exemplo de *Grande São Paulo*, ou *Grande Belo Horizonte*. Quanto ao tráfego aéreo, 1327 rotas aéreas conectam 144 destinos.

As tabelas 5.14, 5.15 e 5.16 exemplificam o processo de extração de dados que deram origem aos indicadores de intensidade de relacionamento, necessários para a determinação das conexões dominantes.

TAB. 5.14: Extrato de lista das principais rotas aéreas envolvendo as 50 maiores cidades.

UF Origem	Origem	UF Destino	Destino	Passageiros	Carga (Kg)
SP	São Paulo	AM	Manaus	224824	35699858
AM	Manaus	SP	São Paulo	224824	35699858
SP	São Paulo	RJ	Rio de Janeiro	2310018	25765888
RJ	Rio de Janeiro	SP	São Paulo	2310018	25765888
SP	São Paulo	DF	Brasília	1097910	24769030
DF	Brasília	SP	São Paulo	1097910	24769030
SP	São Paulo	RS	Porto Alegre	1030538	21520816
RS	Porto Alegre	SP	São Paulo	1030538	21520816
SP	São Paulo	BA	Salvador	875627	20211236
BA	Salvador	SP	São Paulo	875627	20211236

TAB. 5.15: Extrato de lista dos principais destinos aéreos envolvendo as 50 maiores cidades.

Município	Qtd. Conexões	Passageiros	Carga
São Paulo	66	12114663	253292258
Brasília	50	5317024	116712519
Rio de Janeiro	56	6421894	98549392
Manaus	42	940838	65238523
Salvador	36	2685726	54603681
Recife	27	2051355	44586347
Fortaleza	27	1554093	40900152
Porto Alegre	41	2203756	40026203
Belo Horizonte	45	2555978	36743417
Curitiba	34	2026703	30360944
Belém	40	978248	25582855

TAB. 5.16: Extrato de lista dos principais destinos terrestres envolvendo as 50 maiores cidades

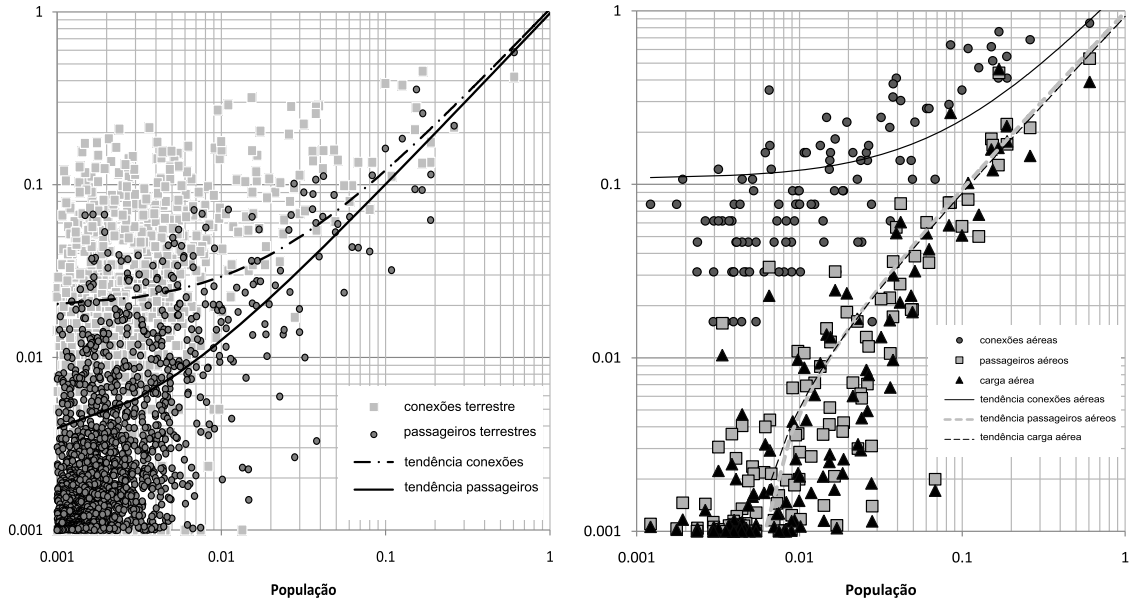
município	conexões	Passageiros
São Paulo	750	4664898
Brasília	341	1201796
Rio de Janeiro	315	2719165
Goiânia	289	751006
Campinas	283	856823
Cascavel	240	164368
Presidente Prudente	218	330086
Cuiabá	215	174264
Maringá	213	306151
Campo Grande	212	275686
Curitiba	210	1655905
Rondonópolis	203	83165

A rede desenvolvida a partir das conexões dominantes apresenta um alto grau de correlação entre o tamanho da população e o número de conexões terrestres, o fluxo de passageiros terrestres, o número de conexões aéreas e o fluxo de carga e passageiros aéreos (figura 5.13).

O gráfico exibido na figura 5.14 apresenta a distribuição dos graus de conexão ao longo da rede resultante, e sugere a existência de mecanismos de formação de *clusters* densos de cidades, *clusters* esses que se assemelham aos estudados em redes de relacionamento social e denominados *small words* (ALBERT e BARABÁSI, 2002).

TAB. 5.17: Extrato de lista de IR e IREscala

origem	destino	Pass. Aéreos	Carga Aérea	Pass. Terrestres	IR	iRelacEscala
Rio de Janeiro	São Paulo	1.000	0.722	1.000	2.722	1.000
São Paulo	Brasília	0.475	0.694	0.076	1.246	2.000
Brasília	São Paulo	0.475	0.694	0.076	1.246	2.000
Belo Horizonte	São Paulo	0.382	0.360	0.437	1.179	2.000
São Paulo	Curitiba	0.349	0.320	0.503	1.172	2.000
Manaus	São Paulo	0.097	1.000	0.000	1.097	2.000
São Paulo	Manaus	0.097	1.000	0.000	1.097	2.000
Porto Alegre	São Paulo	0.446	0.603	0.035	1.084	2.000
São Paulo	Salvador	0.379	0.566	0.029	0.974	2.000
Recife	São Paulo	0.264	0.472	0.023	0.759	3.000
Rio de Janeiro	Brasília	0.290	0.358	0.050	0.698	3.000
Brasília	Rio de Janeiro	0.290	0.358	0.050	0.698	3.000
Rio de Janeiro	Belo Horizonte	0.178	0.137	0.289	0.604	3.000
Fortaleza	São Paulo	0.172	0.378	0.011	0.561	3.000
São Paulo	Florianópolis	0.224	0.261	0.052	0.536	3.000



(a) Correlação população versus número de passageiros e conexões terrestres.

(b) Correlação população versus número de passageiros, conexões e carga aérea.

FIG. 5.13: Gráficos de correlação população e fluxos aéreos e terrestres.

Para determinação do IR foi adotada uma combinação simples não ponderada dos fluxos, havendo diversas opções a serviço do analista de domínio para apreciação das contribuições relativas dos fluxos, entre elas fluxo/habitante ou a incorporação dos fluxos indiretos de conexões subordinadas.

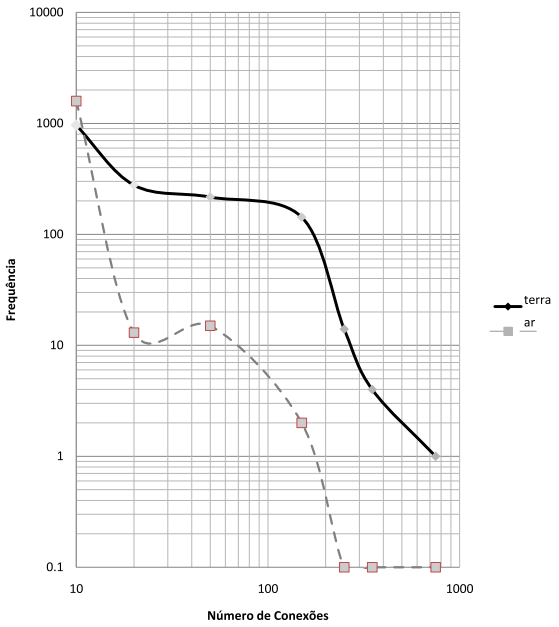


FIG. 5.14: Distribuição de graus de conexão ao longo da rede urbana.

A tabela 5.17 contém os dados de fluxos normalizados por intervalo, o IR resultante e o IR discretizado ($IREscala$) necessário para o traçado das rotas. Para cada destino são comparados todos pares de conexões e respectivos IRs, atribuindo-se como conexão dominante aquela com o maior valor.

Não é imposta a limitação de conexão de cidade maior-menor como na proposta original, o que tende a produzir estruturas cíclicas, mas ao mesmo tempo amplia a capacidade de introspecção na topologia da rede gerada. O resultado pode ser facilmente ampliado para representar para cada destino, por meio de uma sequência ordenada de conexões dominantes, sendo possível se referir então à primeira, segunda ou terceira conexão dominante, caso essas existam. A tabela 5.18 demonstra parcialmente o resultado obtido.

5.5.2 SIMILARIDADES ENTRE CENTROS

O passo seguinte no processo de construção deste modelo consiste na incorporação dos aspectos estruturais (número de conexões, número de vezes em que aquela conexão é dominante) e relacionais (fluxos) ao modelo de representação característico de cada instância representativa do nó da rede, no caso, as cidades. Essa nova coleção de nós expandidos deve então ser apresentada ao SOM para treinamento e identificação de grupos.

Em nosso estudo de caso, além dos atributos descritos na tabela 5.1, foram incluídos os dados sobre os fluxos aéreos e terrestres e número de conexões áereas e terrestres de

TAB. 5.18: Extrato de lista das conexões dominantes envolvendo as 50 maiores cidades.

Conexão Dominante	Rota	IR
São Paulo	Rio de Janeiro >>> São Paulo	2.72174
Rio de Janeiro	São Paulo >>> Rio de Janeiro	2.72174
São Paulo	Brasília >>> São Paulo	1.24554
São Paulo	Belo Horizonte >>> São Paulo	1.17877
São Paulo	Curitiba >>> São Paulo	1.17229
São Paulo	Manaus >>> São Paulo	1.09733
São Paulo	Porto Alegre >>> São Paulo	1.0837
São Paulo	Salvador >>> São Paulo	0.97378
São Paulo	Recife >>> São Paulo	0.75896
São Paulo	Fortaleza >>> São Paulo	0.56124
São Paulo	Florianópolis >>> São Paulo	0.53588
Brasília	Goiânia >>> Brasília	0.49626
Rio de Janeiro	Vitória >>> Rio de Janeiro	0.48228

cada cidade. Esse novo conjunto de dados foi denominado *REGIC Expandido*.

Inicialmente, adotando-se a normalização por intervalo, o treinamento do SOM foi realizado com malhas com quantidade crescente de neurônios à procura de padrões de agrupamentos na U-matrix. Os BMUs resultantes foram agrupados por k-means no intervalo de 2 a 14 grupos, registrando-se os respectivos primeiro e segundo melhores resultados para o DB-index e o Si-index (tabela 5.19). A cada partição no intervalo adotado, o k-means foi executado 50 vezes e selecionada a partição que produziu o menor erro quadrático.

Com exceção de uma pequena região com alta concentração de BMUs contendo instâncias representando extremos, a observação da U-matrix não permitiu a percepção da formação de grupos independentes do tamanho da malha adotada (figura 5.15).

Entretanto, a análise dos grupos obtidos revela que a preservação das distâncias, garantida pelo uso da normalização por intervalo, preserva as instâncias dos grupos das metrópoles próximas entre si, particularmente Rio de Janeiro, São Paulo e Brasília cujas instâncias são alocadas a BMUs próximos no limite inferior direito do mapa.

TAB. 5.19: Resultados de agrupamento de BMUs resultantes de SOM para REGIC Expandido com indicadores de centralidade envolvendo as 50 maiores cidades.

Malha	Visual	DB1	grupos DB1	DB2	grupos DB2	Si 1	grupos Si1	Si2	grupos Si2
20x10	não	0.974	2	0.594	11	0.974	2	0.871	4
50x18	não	0.982	2	0.91	5	0.982	2	0.91	5
75x27	não	0.992	2	0.938	3	0.992	2	0.938	3
100x65	não	0.994	2	0.943	3	0.994	2	0.943	3

Esse mesmo grupo de metrópoles passa a ser agrupado com cidades menores quando o conjunto de dados é tratado por meio de normalização logarítmica, efeito esse já demonstrado na Análise de Componentes Principais do conjunto original de dados.

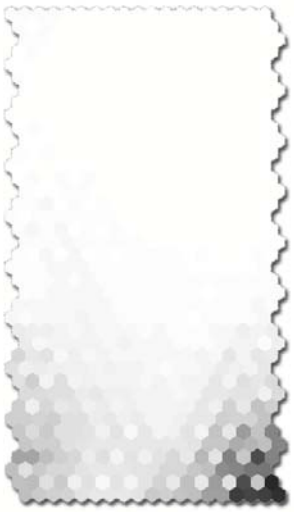


FIG. 5.15: U-matrix resultante de treinamento SOM com normalização por intervalo.

O treinamento foi repetido aplicando-se então a normalização logarítmica sobre o mesmo conjunto de dados de treinamento. Foi possível então observar-se a atenuação do impacto dos valores extremos e o surgimento na U-matrix de regiões e bordas em torno de grupos de BMUs com coordenadas similares, sugerindo a existência de grupos.

Na figura 5.16 são apresentadas quatro U-matrix resultantes de treinamentos realizados com número variado de neurônios na malha do SOM. É claramente perceptível a melhora na visualização de grupos à medida que a malha de neurônios é expandida de 250 (A)

para 1000(B) unidades. Acréscimos posteriores para 2000 (C) e 4000 (D) células não produziram alterações significativas nos padrões observados.

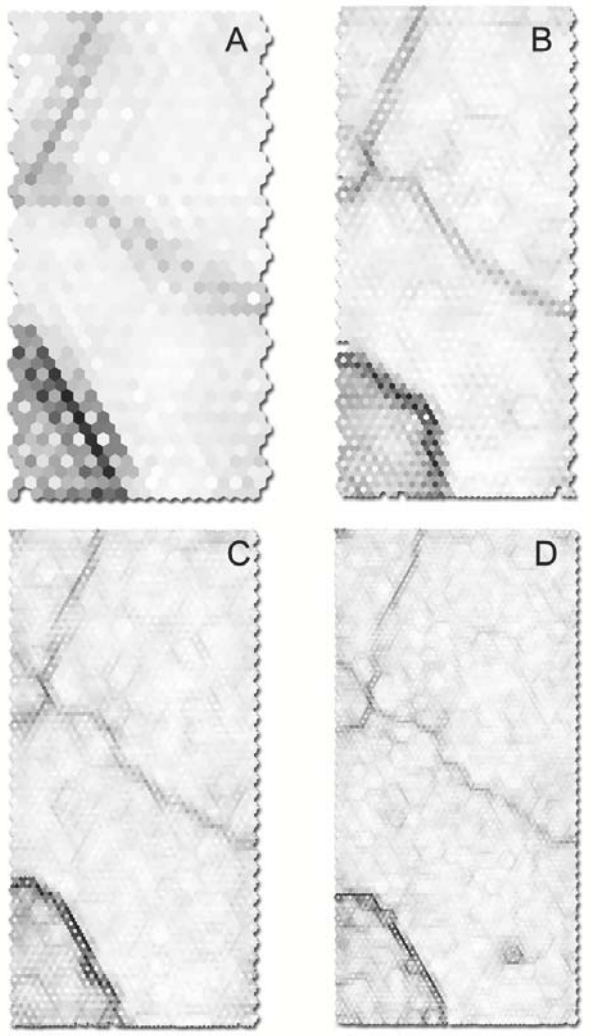


FIG. 5.16: U-matrix resultantes de treinamento SOM com 250 (A), 1000 (B), 2000 (C) e 4000 (D) neurônios.

Pela observação da U-matrix, avaliação do *Silhoutes* (figura 5.17 a) e do DB-index (figura 5.17 b), no caso de normalização logarítmica, a partição de 4 grupos é selecionada como aquela capaz de melhor representar a distribuição das densidades do espaço de entrada. Nesse caso, apesar de não haver uma clara distinção de grupos, eventualmente poderíamos indicar 3 grandes áreas distintas. Já com a normalização por intervalo, o melhor resultado quantitativo indicava 10 grupos, enquanto visualmente havia sugestão de apenas 5 classes. Importante salientar que os dois melhores resultados dos indicadores apontaram a partição de 2 grupos como a ideal. Nesta partição binária, um pequeno

grupo é composto invariavelmente por 5 a 15 instâncias, todas cidades do grupo original de grandes metrópoles, e o segundo grupo pelo restante da amostra.

Observa-se que nas duas melhores partições, tanto para normalização por intervalo quanto logarítmica, os melhores resultados qualitativos do primeiro e quantitativos do segundo método estão próximos ou correspondem ao número exato de partições iniciais apresentado pelo REGIC original, que é igual a cinco.

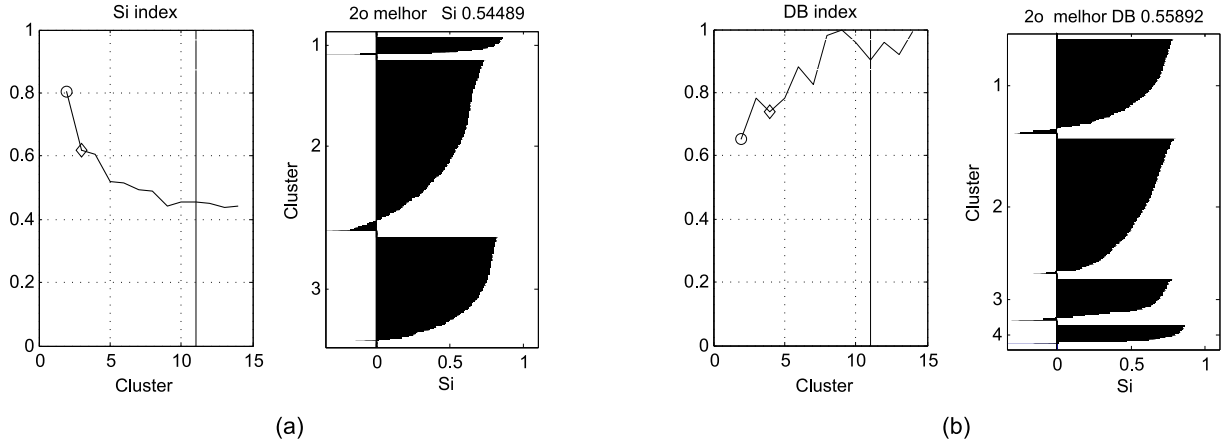


FIG. 5.17: Série de agrupamentos segundo o Si index e o segundo melhor Silhouettes (a), Série de agrupamentos segundo o DB index e o segundo melhor Silhouettes (b)

A tabela 5.20 é conhecida como Matriz de Confusão ou *Contigency Matrix* e tem como função representar a comparação entre duas partições distintas realizadas sobre uma mesma amostra. A matriz é composta de m linhas, uma para cada classe resultante da partição 1, e n colunas para as classes da partição 2. No caso de partições idênticas, a matrix de confusão é uma matriz $m\ x\ m$ contendo ocorrências em apenas uma única célula $a(i, j)$, que é o único elemento não-zero da linha i e coluna j em que este ocorre. A *Contigency Matrix* é aquela necessária para várias métricas de qualidade relativa entre classifição distintas de um mesmo conjunto de dados.

Na tabela 5.21 é apresentado um extrato da tabela de classificação resultante exibindo as 4 primeiras classes da classificação original e classificação obtida pela modelagem de fluxos dominantes, SOM e k-means.

A partir dos resultados desse processo classificatório, da distribuição geográfica dos municípios e das respectivas intensidades de relacionamentos nas rotas mapeadas é possível então desenvolver modelos de representação visual. Entre eles o traçado sobreo *framework* do *Google Maps API*.

TAB. 5.20: Matriz de Confusão, 11 classes originais e 10 classes do resultado do SOM.

	Classe Original											
Classe SOM	1	2	3	4	5	6	7	8	9	10	11	n. instâncias
1	0	0	0	7	7	7	3	0	1	0	0	25
2	0	1	1	0	0	0	0	0	0	0	0	2
3	0	0	8	4	0	0	0	0	0	0	0	12
4	0	0	0	0	0	5	48	45	118	184	913	1313
5	0	0	0	0	0	3	1	0	4	2	4	14
6	0	0	0	0	0	0	0	3	3	6	21	33
7	1	1	0	0	0	0	0	0	0	0	0	2
8	0	0	0	0	1	4	9	8	18	13	36	89
9	0	0	0	0	11	16	22	13	20	17	21	120
10	0	0	0	0	1	1	1	1	1	0	5	10
n. instâncias	1	2	9	11	20	36	84	70	165	222	1000	1620

5.6 REPRESENTAÇÃO VISUAL DOS RESULTADOS

Para o suporte à interpretação dos resultados, uma pequena aplicação *PHP* empregando bibliotecas do *Google Maps API* é empregada para os traçados das rotas e centros com ícones diferenciados em função dos resultados do SOM. Por meio de consultas a uma base *MySql* é possível limitar-se a janela de observação aos níveis, cidades e regiões desejadas.

Na figura 5.18 é exibida uma série de mapas correspondendo ao resultado do processo de modelagem e reconstrução da rede particionada por rotas nas quais participem as cidades do Rio de Janeiro e São Paulo. A aplicação permite limitar-se progressivamente quais centros e rotas são exibidos em função da intensidade de relacionamento coberta pela análise, permitindo assim ao analista de domínio evidenciar as relações existentes na rede urbana em níveis hierárquicos distintos.

5.7 SOM DE ROTAS

A fim de direcionar futuros trabalhos procurou-se verificar a capacidade de ordenação do SOM em relação às rotas entre as cidades e os respectivos indicadores de fluxos.

Do conjunto total de mais de 30 mil rotas cadastradas, extraiu-se uma amostra envolvendo rotas nas quais participam as cinco maiores cidades brasileiras segundo o número de habitantes, contendo os seguintes atributos:

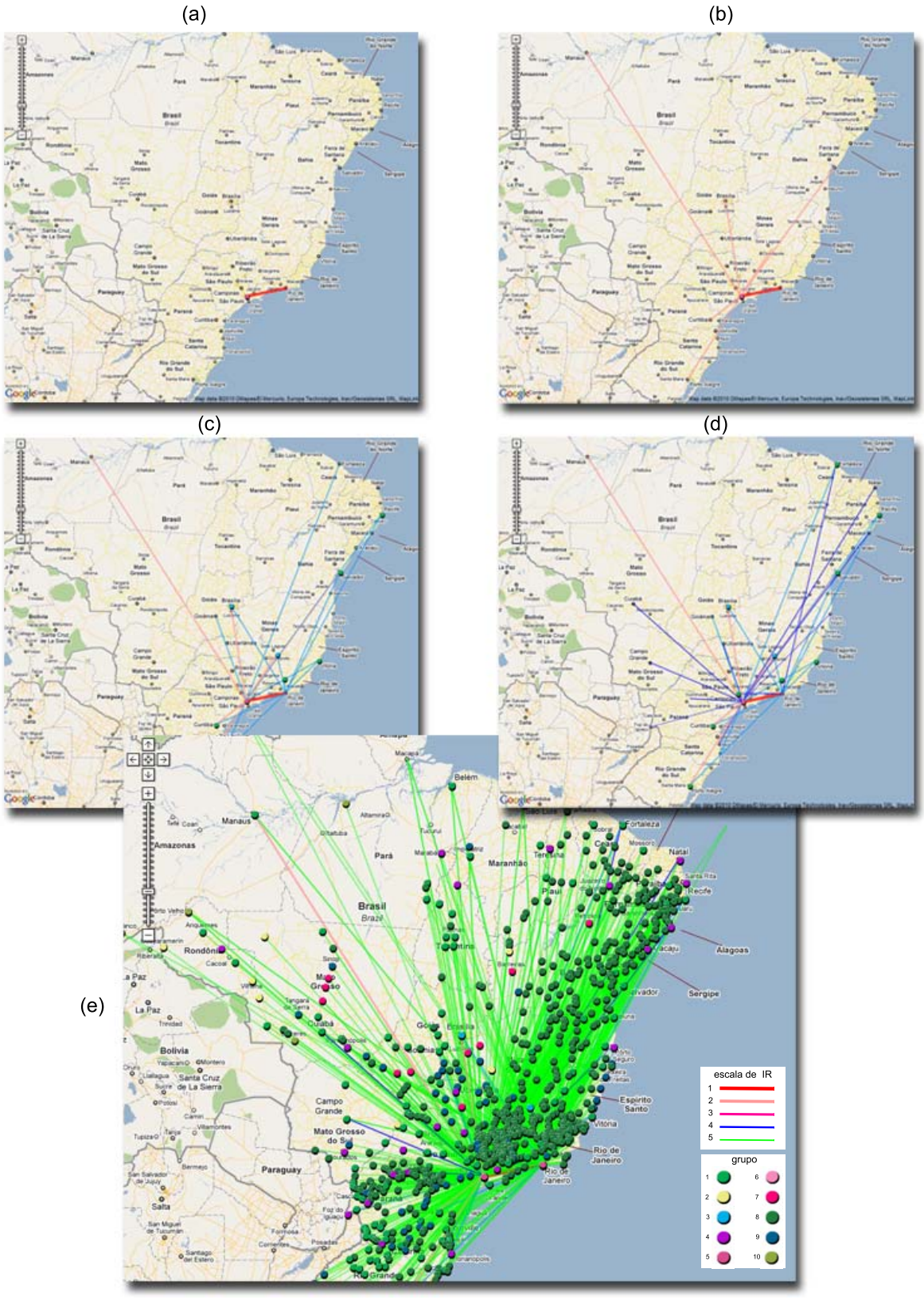


FIG. 5.18: Rotas e centros conectando as cidades do Rio de Janeiro e São Paulo, $IR = 1$ (a), $IR <= 2$ (b), $IR <= 3$ (c), $IR <= 4$ (d), $IR <= 5$ (e).

TAB. 5.21: Resultado comparativo do processo de classificação resultante do modelo com a classificação original.

Município	Classe original	Classe SOM
São Paulo	1	5
Rio de Janeiro	2	5
Brasília	2	3
Curitiba	3	1
Belo Horizonte	3	3
Goiânia	3	1
Recife	3	1
Porto Alegre	3	1
Fortaleza	3	1
Salvador	3	1
Belém	3	1
Manaus	3	1
Campinas	4	1
Vitória	4	1
Florianópolis	4	4
Teresina	4	4
Aracaju	4	4
Natal	4	4
Campo Grande	4	1
Maceió	4	4
João Pessoa	4	4
SãoLuís	4	4
Cuiabá	4	1

- código cidade origem
- código cidade destino
- fluxo anual de passageiros terrestres
- fluxo anual de passageiros aéreos
- fluxo anual de carga aérea,

e atributos de identificação. Cada instância de rota foi marcada com uma etiqueta contendo menemonicos para o nome da cidade acrescido do IRescala da conexão (tabela 5.22).

Adotando-se as etiquetas de rotas como classificadores foram identificados 18 grupos nessa amostra, os quais se distribuem conforme a figura 5.19.

TAB. 5.22: Extrato de tabela de rotas e fluxos ligando as cinco maiores cidades brasileiras.

Cod Orig	Cod Dest	NPassAr	NPassTerr	TCargaAr	IREscala	Tag	Origem	Destino
1302603	3550308	224824	0	35699858	2	SP2	Manaus	SãoPaulo
3304557	3550308	2310018	663128	25765888	1	SP1	RiodeJaneiro	SãoPaulo
5300108	3550308	1097910	50692	24769030	2	SP2	Brasília	SãoPaulo
4314902	3550308	1030538	23051	21520816	2	SP2	PortoAlegre	SãoPaulo
2927408	3550308	875627	18953	20211236	2	SP2	Salvador	SãoPaulo
3550308	2611606	609486	15418	16845334	3	REC3	SãoPaulo	Recife
2304400	3550308	398338	7313	13486522	3	SP3	Fortaleza	SãoPaulo
3106200	3550308	881616	289855	12852438	2	SP2	BeloHorizonte	SãoPaulo
5300108	3304557	670494	32892	12778590	3	RIO3	Brasília	RiodeJaneiro
4106902	3550308	805257	333686	11441632	2	SP2	Curitiba	SãoPaulo
4205407	3550308	516902	34184	9302029	3	SP3	Florianópolis	SãoPaulo

Os códigos das cidades, atributos categóricos, trazem um complicador extra para a determinação de distâncias entre as instâncias, uma vez que métricas euclidianas não podem ser aplicadas na ausência de uma escala de valores ordenados.

Uma das abordagens para o tratamento de atributos categóricos é a conversão da categoria original em uma representação binária com tantos bits quantos valores existentes na escala de atributos original. A título de exemplo, uma coleção de valores tais como $azul, amarelo, preto$ poderia ser representada pela sequência $001, 010, 100$, exigindo 2 novas colunas na descrição de cada instância. À medida que cresce a lista de descritores categóricos, esse método gera obstáculos extras para a manutenção, representação e interpretação dos resultados. COTTRELL, IBBOU e LETRÉMY (2004) oferecem soluções alternativas à representação binária, entre elas a denominada *Kohonen Multiple Correspondence Analysis, KMCA*.

Em nossos experimentos com rotas ligando as 50 maiores cidades brasileiras, foram obtidos resultados semelhantes entre a abordagem da conversão para a representação binária e o uso do código de 7 dígitos como um número real.

Pela observação da U-matrix na figura 5.20 e comparação com as atribuições feitas das rotas aos BMUs, percebe-se nitidamente a ordenação do mapa, separando categorias distintas, assim como diferenciando as intensidades dos fluxos. Nesse exemplo, a quantidade de neurônios ou de unidades na malha desempenhou papel relevante na percepção de padrões, seguindo a recomendação da literatura do ESOM, de que a malha deva conter mais unidades que o número de instâncias da amostra de treinamento (ULTSCH e

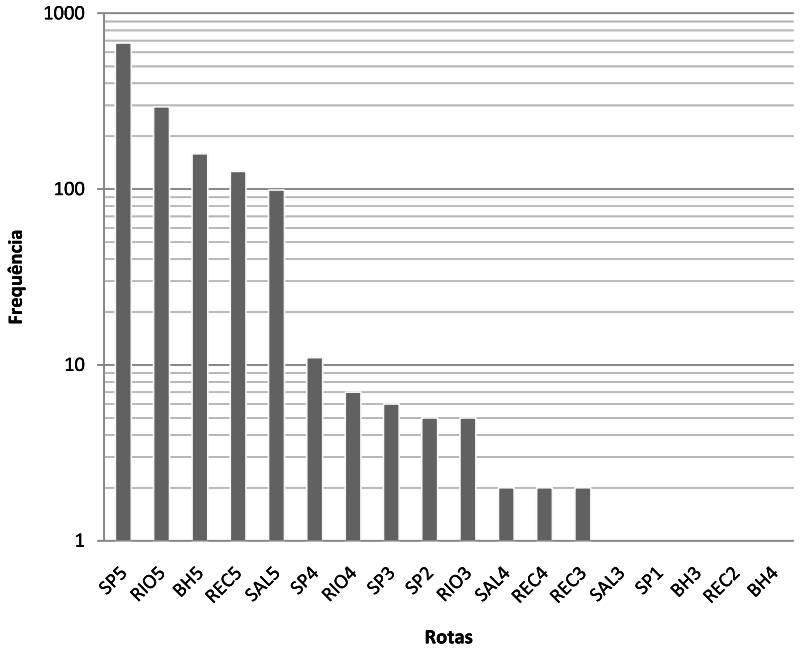


FIG. 5.19: Distribuição de rotas relativas às cinco maiores cidades.

MÖRCHEN, 2005).

Na figura 5.21 é apresentado o gráfico referente ao ciclo de varredura em busca do agrupamento com melhores índices (a), o *Silhouettes* do agrupamento original (b), e o *Silhouettes* do grupo com melhor *Davies-Bouldin index* (c).

5.8 DISCUSSÃO DOS RESULTADOS

A análise dos resultados obtidos deve ser necessariamente conduzida sobre diferentes eixos de observação levando-se em conta os principais objetivos pretendidos.

Centralmente, o modelo se propõe a oferecer uma abordagem de reconstrução de topologia de redes descritas por seus descritores de fluxos, indicadores de centralidade e similaridades entre os seus nós. Esse objetivo foi alcançado e demonstrada a possibilidade de extração de topologias que transcendem o padrão acíclico e hierárquico do método dos fluxos dominanes. A rede descoberta pelo modelo, a critério do analista de domínio, é capaz de contemplar ligações entre nós de altura distinta e a hierarquia é capturada a partir de nós enriquecidos por atributos que não somente aqueles derivadas dos fluxos e conexões dominantes.

Não menos importante foi a decisão de incorporar ao modelo o emprego de mapas auto-organizáveis, principalmente por sua reconhecida e aqui demonstrada capacidade de

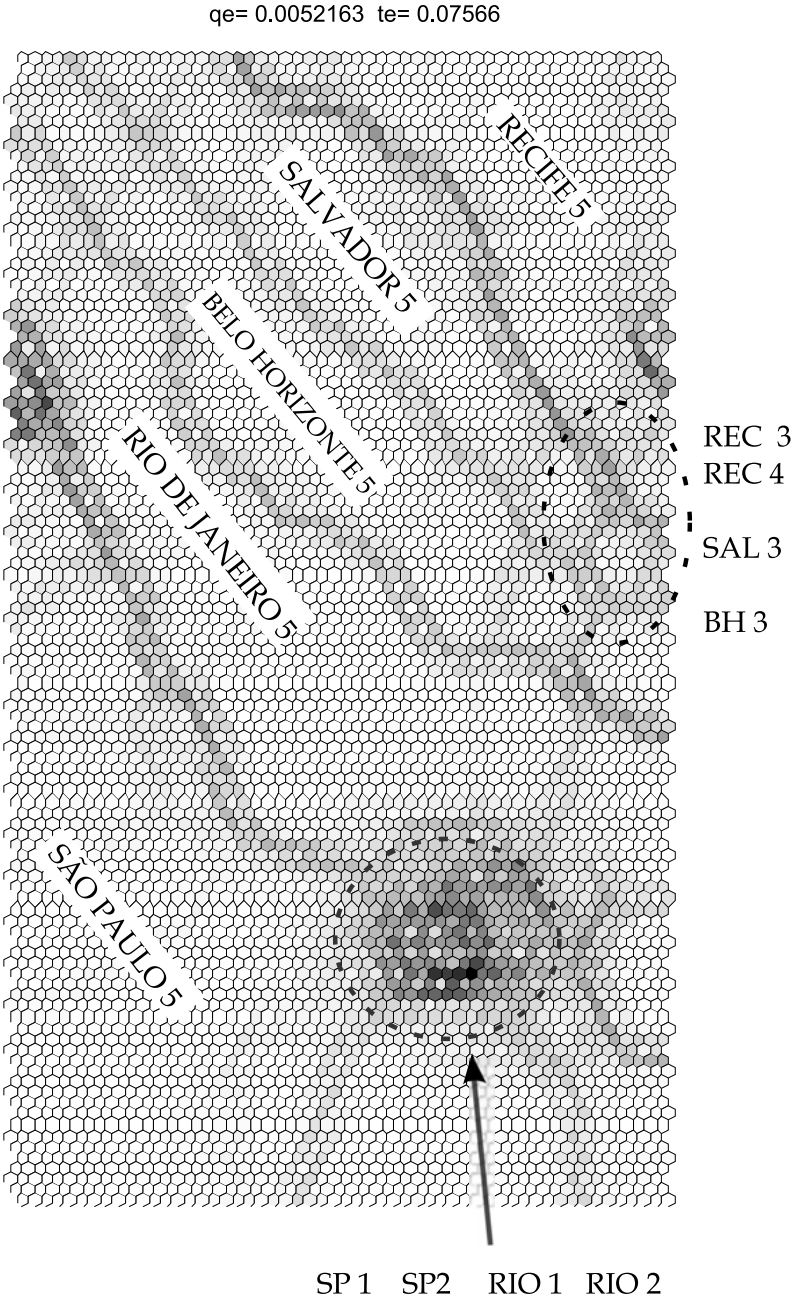


FIG. 5.20: U-matrix de rotas para um mapa com 700 unidades de neurônios.

aprender sobre distribuições de dados e, subsequentemente, operar como uma máquina de inferências de padrões para classificações de objetos desconhecidos, mas oriundos do mesmo espaço amostral. Uma vantagem secundária, mas não desprezível, do emprego do SOM é que o processo de aprendizagem e a captura de conhecimento é feita numa relação custo-benefício para o usuário final extremamente atraente se comparado à análises discriminantes tradicionais.

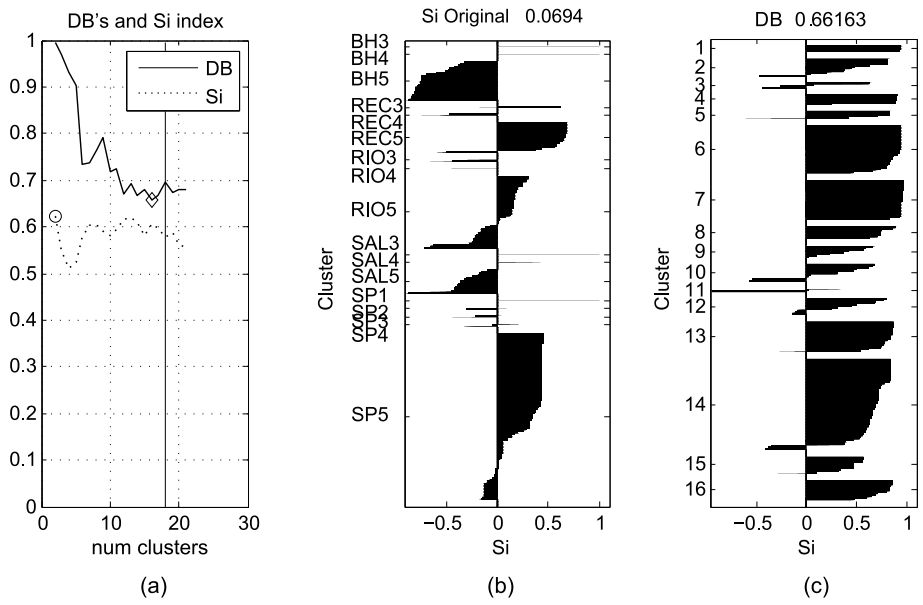


FIG. 5.21: (a) Ciclo de varredura de agrupamento do SOM, (b) Silhouette do agrupamento original, e (c) do melhor segundo DB-index

Quanto ao emprego do SOM na detecção de estruturas de agrupamentos em distribuições fortemente concentradas e com decaimento exponencial, o método sofre de limitações comuns a outras técnicas e exigem tratamentos e heurísticas diferenciadas caso a caso. Sob nosso ponto de vista, no caso particular do REGIC, uma estratégia cabível seria a condução da análise de agrupamentos em dois passos. Numa primeira etapa ocorre a partição em dois grandes grupos, um composto de extremos com alto valor agregado e outro o restante das instâncias do problema. No passo seguinte, cada sub-conjunto é submetido a análises em separado e os resultados combinados em uma classificação final.

Quando ao tratamento dos dados, normalizações devem ser consideradas com a devida cautela e sob a supervisão de outros indicadores complementares para avaliação dos resultados obtidos. O *Silhouttes* demonstrou ser um suporte visual extremamente útil para compreensão de classificações prévias e posteriores comparações com partições derivadas do mesmo conjunto de dados.

No que tange a percepção sobre a formação de grupos na U-matrix, os resultados indicaram que o tamanho da malha é um fator decisivo na emergência de padrões e que diferentes topologias devem ser exploradas, vide figura 5.16. Não foi possível no escopo desse trabalho explorar as potencialidades de análises densitométricas na detecção automática de bordas de clusters na U-matrix.

Os indicadores *DB-index* e *Silhouettes* devem ser adotados com parcimônia pelos ar-

gumentos levantados nas seções 2.2.2.1 e 2.2.2.2 e usados como um suporte complementar à tomada de decisão sobre a presença ou não de grupos e qual seria a partição *natural*. A figura 5.22 demonstra um possível painel virtual de uma ferramenta de suporte à análise em problemas dessa natureza, onde esse conjunto de informações é oferecido ao analista interessado para a tomada de decisão da partição ideal.

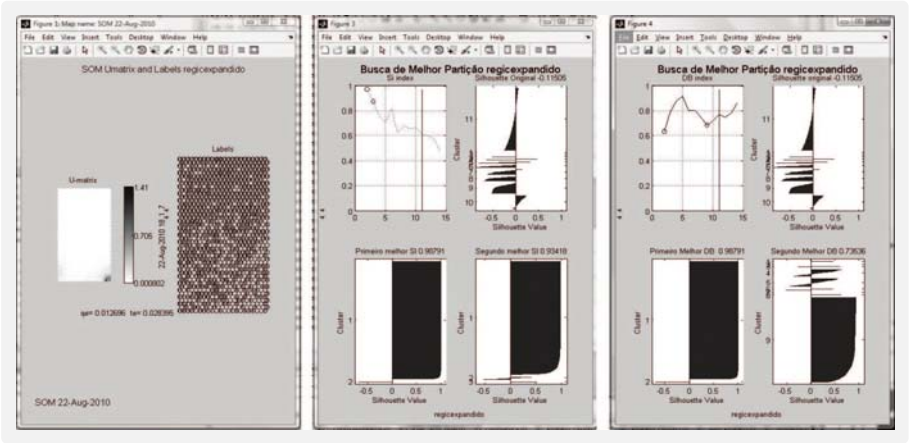


FIG. 5.22: Painel de análise de padrões de grupos

O estudo das distribuições de rotas pelo SOM exibe um potencial atraente pois é capaz de evidenciar a formação de cliques, ou grupos de objetos em torno de um destino comum, com aplicações interessantes em estudos de redes sociais. Um problema aqui é a relativa incapacidade da U-matrix oferecer suporte à representação satisfatória para um número elevado de grupos. Um estratégia possível seria o emprego de algoritmos de SOM hierarquizados que particionem o universo de rotas analisadas por grupos de destinos.

Sendo o SOM e a U-matrix exemplos claros do papel que o suporte visual possui para a interpretação de problemas de classificação, deve-se observar que a representação visual georreferenciada adotada para exibição parcial dos resultados (vide figura 5.18) apenas toca a superfície de um universo de possibilidades a serem exploradas.

A comparação dos resultados obtidos com a classificação original (vide (tabela 5.20) se justifica apenas parcialmente. O modelo aqui proposto prioriza a meta de empregar dados públicos e regularmente atualizáveis, o que por si só altera dramaticamente a capilaridade dos informações disponíveis. Enquanto no REGIC o IBGE dispunha de dados sobre trocas a nível intermunicipal e num espectro amplo de dimensões de observação (saúde, lazer, etc.), o presente trabalho contou apenas com os registros de rotas interestaduais, no caso terrestre, e apenas poucos casos de ligações intermunicipais, nas rotas aéreas. Ao mesmo

tempo, e mesmo apesar de outras diferenças metodológicas, as semelhanças obtidas para as grandes metrópoles (vide tabela 5.21) parecem demonstrar ser a densidade populacional a grande força geradora dos fluxos observados e a eficácia do modelo aqui apresentado.

6 CONSIDERAÇÕES FINAIS

Consideramos que os objetivos inicialmente estabelecidos para essa dissertação foram satisfatóriamente atingidos baseado na extensão dos experimentos realizados e resultados obtidos. O modelo se mostrou capaz de oferecer suporte à determinação de centros locais e respectivas regiões de influência, particularmente no que tange a extração de grupos de nós que compartilham similaridades.

Quanto às contribuições esperadas podemos destacar que:

a) A metodologia descrita, os relatos dos experimentos e as referências comentadas oferecem subsídios consistentes para a reprodução sistemática da abordagem descrita e guia para futuras implementações.

b) O SOM se mostrou plenamente aplicável ao tipo de problema proposto, salvo as devidas ressalvas quanto ao tratamento de extremos ou a decomposição em duas amostras.

c) Os indicadores adotados sugeriram que, assim como qualquer outra estratégia de mensuração de problemas dessa natureza, não devem ser adotados cegamente, observando-se sempre o padrão de distribuições subjacente.

d) Os mapas georreferenciados se mostraram um instrumento valioso no compartilhamento e representação dos resultados, entretanto suportes visuais complementares se fazem necessários, particularmente aqueles capaz de revelar estruturas hierárquicas e fluxos inter e intra-níveis.

6.1 TRABALHOS FUTUROS

A despeito da capacidade de aprendizagem a partir de exemplos, o SOM aplicado a problemas de *clustering* ainda carece de uma estratégia definitiva e totalmente automática para a extração de grupos formados pelos BMUs. Sendo essa, ao nosso ver, uma das direções a seguir em trabalhos futuros, particularmente as relacionadas com a análise geométrica e densitométrica da U-matrix (ULTSCH e HERRMANN, 2006) e métodos baseados em otimização por enxame de partículas (SHARMA e OMLIN, 2006).

À luz dos resultados obtidos, as oportunidades de captura de estruturas de centros locais e centros subordinados, empregando exclusivamente a descrição de rotas e os fluxos entre os nós participantes, mostraram-se bastante promissoras e merecem futuras investigações.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ALBERT, R. e BARABÁSI, A.-L. **Statistical mechanics of complex networks**. *Rev. Mod. Phys.*, 74(1):47–97, Jan 2002.

ALDENDERFER, M. S. e BLASHFIELD, R. K. K. ***Cluster Analysis***. Sage Publications, Inc, illustrated edition edition, novembro 1984. ISBN 0803923767.

ANAC. **Anuários Estatísticos do Transporte Aéreo**. Technical report, Agência Nacional de Aviação Civil, 2007. URL `http://www.anac.gov.br/estatistica/estatisticas1.asp`.

ANTT. **Anuário Estatístico Rodoviário - 2008 (DADOS PRELIMINARES)**. Technical report, Agência Nacional de Transportes Terrestres, 2008. URL `http://www.antt.gov.br/passageiro/anuarios/anuario2008/default.asp`.

BAÇÃO, F., LOBO, V. e PAINHO, M. **The self-organizing map, the Geo-SOM, and relevant variants for geosciences**. *Comput. Geosci.*, 31(2):155–163, 2005. ISSN 0098-3004.

BAVELAS, A. ***Some mathematical properties of psychological space***. Tese de Doutorado, Massachusetts Institute of Technology, 1948. URL `http://hdl.handle.net/1721.1/34690`.

BEHNISCH, M. e ULTSCH, A. **Urban Data Mining Using Emergent SOM**. Em PREISACH, C., BURKHARDT, H., SCHMIDT-THIEME, L. E DECKER, R., editores, *Data Analysis, Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, págs. 311–318. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-78246-9. URL `http://dx.doi.org/10.1007`. 10.1007.

BERRY, B. J. L. e GARRISON, W. L. **A Note on Central Place Theory and the Range of a Good**. *Economic Geography*, 34(4):304–311, outubro 1958. ISSN 00130095. URL `http://www.jstor.org/stable/142348`. ArticleType: primary_article / Full publication date: Oct., 1958 / Copyright © 1958 Clark University.

BROWN, R. J. ***Adaptatitve Multiple-Output Threshold Systems and Their Storage Capacities***. Tese de Doutorado, Stanford Electronics Laboratories, Stanford, CA, Junho 1964.

CHAPMAN, P., CLINTON, J., KERBER, R., KHABAZA, T., REINARTZ, T., SHEARER, C. e WIRTH, R. **CRISP-DM 1.0 Step-by-step data mining guide**. Technical report, The CRISP-DM consortium, August 2000. URL `http://www.crisp-dm.org/CRISPWP-0800.pdf`.

CHENG, Y., ZHANG, Y., HU, J. e LI, L. **Mining for Similarities in Urban Traffic Flow Using Wavelets**. Em *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE*, págs. 119 –124, sep. 2007.

CIS. **Bibliography of SOM papers**. http://www.cis.hut.fi/research/som-bibl/, 2010a. URL `http://www.cis.hut.fi/research/som-bibl/`.

CIS. **SOM Toolbox**, howpublished = http://www.cis.hut.fi/somtoolbox/, 2010b. URL `http://www.cis.hut.fi/somtoolbox/`.

CORRÊA, R. L. ***Trajetórias Geográficas***. Editora Bertrand Brasil LTDA., Rio de Janeiro, 4 edition, 2010.

COTTRELL, M., IBBOU, S. e LETRÉMY, P. **SOM-based algorithms for qualitative variables**. *Neural Netw.*, 17(8-9):1149–1167, 2004. ISSN 0893-6080.

COVER, T. M. ***Geometrical and Statistical Properties of Linear Threshold Devices***. Tese de Doutorado, Stanford Electronics Laboratories, Stanford, CA, Maio 1964.

CYBENKO, G. **Approximation by superpositions of a sigmoidal function**. *Mathematics of Control, Signals, and Systems (MCSS)*, 2:303–314, 1989. ISSN 0932-4194. URL `http://dx.doi.org/10.1007/BF02551274`. 10.1007/BF02551274.

DAVIES, D. L. e BOULDIN, D. W. **A Cluster Separation Measure**. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-1(2):224 –227, april 1979. ISSN 0162-8828.

DEMARTINES, P. e HERAULT, J. **Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets**. *Neural Networks, IEEE Transactions on*, 8(1):148 –154, jan 1997. ISSN 1045-9227.

EATON, B. C. e LIPSEY, R. G. **An Economic Theory of Central Places**. *The Economic Journal*, 92(365):56–72, março 1982. ISSN 00130133. URL `http://www.jstor.org/stable/2232256`. ArticleType: primary_article / Full publication date: Mar., 1982 / Copyright ÂⒸ 1982 Royal Economic Society.

FRALEY, C. e RAFTERY, A. E. **How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis**. *The Computer Journal*, 41:578–588, 1998.

FRANK, A. e ASUNCION, A. **UCI Machine Learning Repository**, 2010. URL `http://archive.ics.uci.edu/ml`.

FREEMAN, L. C., BORGATTU, S. P. e WHITE, D. R. **Centrality in valued graphs: A measure of betweenness based on network flow**. *Social Networks*, 13(2):154, 141, junho 1991. ISSN 03788733. URL `http://dx.doi.org/10.1016/0378-8733(91)90017-N`.

FREEMAN, L. **Centrality in social networks: Conceptual clarification**. *Social Networks*, 1:239, 215, 1979.

GOOGLE INC. **API do Google Maps - Google Code**. http://code.google.com/intl/pt-BR/apis/maps/index.html, 2010. URL `http://code.google.com/intl/pt-BR/apis/maps/index.html`.

GUYON, I. e ELISSEEFF, A. **An introduction to variable and feature selection**. *J. Mach. Learn. Res.*, 3:1157–1182, 2003. ISSN 1532-4435.

HAYKIN, S. ***Redes Neurais. Princípios e prática***. Bookman, 2 edition, julho 2001. ISBN 9788573077176.

HOTELLING, H. **Analysis of a complex of statistical variables into principal components**. *Journal of Educational Psychology*, 24(7):498–520, outubro 1933. ISSN 0022-0663.

IBGE. **Regiões de Influência das Cidades Brasileiras, 2007**. Technical report, Coordenação de Geografia, Diretoria de Geociências - Instituto Brasileiro de Geografia e Estatística - IBGE, 2008.

JAIN, A. K., MURTY, M. N. e FLYNN, P. J. **Data clustering: a review**. *ACM Comput. Surv.*, 31(3):264–323, 1999. ISSN 0360-0300.

KOHONEN, T. ***Self-Organizing Maps***, pág. 97. Springer, 3rd edition, dezembro 2000a. ISBN 3540679219.

KOHONEN, T. ***Self-Organizing Maps***, pág. 105. Springer, 3rd edition, dezembro 2000b. ISBN 3540679219.

KRÖSE, B. e SMAGT, P. ***An introduction to Neural Networks***. The University of Amsterdam, 8ed edition, 1996.

LATOR, V. e MARCHIORI, M. **Efficient Behavior of Small-World Networks**. *Physical Review Letters*, 87(19):198701, outubro 2001. URL `http://link.aps.org/doi/10.1103/PhysRevLett.87.198701`.

LLOYD, S. **Least squares quantization in PCM**. *Information Theory, IEEE Transactions on*, 28(2):129 – 137, mar 1982. ISSN 0018-9448.

MACQUEEN, J. B. **Some Methods for Classification and Analysis of MultiVariate Observations**. Em CAM, L. M. L. E NEYMAN, J., editores, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, págs. 281–297. University of California Press, 1967.

MAYER, R., NEUMAYER, R. e RAUBER, A. **Analytic comparison of self-organising maps**. Em *in Proceedings of the 7th Workshop on SelfOrganizing Maps (WSOM 09)*, págs. 8–10, 2009.

MUSIAL, K., KAZIENKO, P. e BRÓDKA, P. **User position measures in social networks**. Em *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, págs. 1–9, Paris, France, 2009. ACM. ISBN 978-1-60558-676-2. URL http://portal.acm.org/citation.cfm?id=1731011.1731017.

NIEMINEN, U. J. **On the centrality in a directed graph**. *Social Science Research*, 2 (4):371–378, dezembro 1973. ISSN 0049-089X.

NISBET, R., IV, J. E. e MINER, G. ***Handbook of Statistical Analysis and Data Mining Applications***. Academic Press, Har/DVD edition, junho 2009. ISBN 0123747651.

NYSTUEN, J. D. e DACEY, M. F. **A graph theory interpretation of nodal regions**. *Papers of the Regional Science Association*, 7(1):29–42, 1961. ISSN 1056-8190. URL http://deepblue.lib.umich.edu/handle/2027.42/45977.

PHPGROUP. **PHP: Hypertext Preprocessor**. http://www.php.net/, 2010. URL http://www.php.net/.

PIZZUTI, C. **Overlapped community detection in complex networks**. Em *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, págs. 859–866, Montreal, Québec, Canada, 2009. ACM. ISBN 978-1-60558-325-9. URL http://portal.acm.org/citation.cfm?id=1569901.1570019.

PÖLZLBAUER, G. ***Advanced data exploration methods based on Self-Organizing Maps***. Tese de Doutorado, Vienna University of Technology, 2008.

PORTA, S., CRUCITTI, P. e LATORA, V. **The network analysis of urban streets: a primal approach**. *Environment and Planning B: Planning and Design*, 33(5):705–725, 2006. URL http://www.envplan.com/abstract.cgi?id=b32045.

PYLE, D. ***Data Preparation for Data Mining***. Morgan Kaufmann, book & CD-ROM 1st edition, abril 1999. ISBN 1558605290.

RABINO, G. A. e OCCELLI, S. **Understanding spatial structure from network data : theoretical considerations and applications**, janeiro 2007. URL http://www.cybergeo.eu/index2199.html.

ROUSSEEUW, P. **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis**. *J. Comput. Appl. Math.*, 20(1):53–65, 1987. URL http://portal.acm.org/citation.cfm?id=38772.

SCHAEFFER, S. E. **Graph clustering**. *Computer Science Review*, 1(1):27 – 64, 2007. ISSN 1574-0137.

SHARMA, A. e OMLIN, C. W. **Determining cluster boundary using Particle Swarm Optimization**. Em *Proceedings of world academy of science engineering and technology*, volume 15, págs. 250–254, 2006.

SU, T. e DY, J. **A Deterministic Method for Initializing K-Means Clustering**. *Tools with Artificial Intelligence, IEEE International Conference on*, 0:784–786, 2004. ISSN 1082-3409.

TUIA, D., KAISER, C., DA CUNHA, A. e KANEVSKI, M. **Detection of Urban Socio-economic Patterns Using Clustering Techniques**. Em MURGANTE, B., BORRUSO, G. e LAPUCCI, A., editores, *Geocomputation and Urban Planning*, volume 176 of *Studies in Computational Intelligence*, págs. 19–36. Springer Berlin / Heidelberg, 2009. URL `http://dx.doi.org/10.10072`. 10.10072.

ULLMAN, E. **A Theory of Location for Cities**. *The American Journal of Sociology*, 46(6):853–864, maio 1941. ISSN 00029602. URL `http://www.jstor.org/stable/2769394`. ArticleType: primary_article / Full publication date: May, 1941 / Copyright © 1941 The University of Chicago Press.

ULTSCH, A. e HERRMANN, L. **Automatic Clustering with U*C**. Technical report, Dept. of Mathematics and Computer Science, Philipps-University of Marburg, 2006.

ULTSCH, A. e MÖRCHEN, F. **ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM**. Technical Report 46, Dept. CS,University of Marburg, Germany, 2005.

ULTSCH, A. e SIEMON, H. **Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis**. Em *Proc. INNC'90, Int. Neural Network Conf.*, págs. 308, 305. Kluwer, 1990.

ULTSCH, A. e VETTER, C. **Self-Organizing-Feature-Maps versus Statistical Clustering Methods: A Benchmark**. Technical Report 0994, FG Neuroinformatik & Künstliche Intelligenz, University of Marburg, Denmark, 1995.

VEELENTURF, L. P. J. ***Analysis and Applications of Artificial Neural Networks***. Prentice Hall, 1st edition, junho 1995. ISBN 013489832X.

VELLASCO, M. M. B. R. **Curso Redes Neurais I - cadeira do curso de pós-graduação nível mestrado em Engenharia Elétrica**. Notas de Aula, março-julho 2009.

VESANTO, J. e ALHONIEMI, E. **Clustering of the self-organizing map**. *IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council*, 11(3):586–600, 2000. ISSN 1045-9227. URL `http://www.ncbi.nlm.nih.gov/pubmed/18249787`. PMID: 18249787.

VRAGOVÌC, I., LOUIS, E., DÃAZ-GUILERA, A., CHICO e MARIA. **Efficiency of informational transfer in regular and complex networks**. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 71(3 Pt 2A):036122, março 2005. ISSN 1539-3755. URL `http://www.ncbi.nlm.nih.gov/pubmed/15903508`. PMID: 15903508.

WATTS, D. J. e STROGATZ, S. H. **Collective dynamics of 'small-world' networks.** *Nature*, 393(6684):440–442, June 1998. ISSN 0028-0836. URL http://dx.doi.org/10.1038/30918.

WIDROW, B. e LEHR, M. **30 years of adaptive neural networks: perceptron, Madaline, and backpropagation**. *Proceedings of the IEEE*, 78(9):1415 –1442, sep 1990. ISSN 0018-9219.

WITTEN, I. H. e FRANK, E. ***Data Mining: Practical Machine Learning Tools and Techniques, Second Edition***. Morgan Kaufmann, 2 edition, junho 2005. ISBN 0120884070.

XU, R. e WUNSCH, D. **Survey of clustering algorithms**. *Neural Networks, IEEE Transactions on*, 16(3):678, 645, 2005. URL http://dx.doi.org/10.1109/TNN.2005.845141.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo