**UNIVERSIDADE FEDERAL DE VIÇOSA**

Eduardo Martins Viana

# AGRUPAMENTO DE CONJUNTOS CONSUMIDORES DE ENERGIA ELÉTRICA UTILIZANDO MAPAS AUTO-ORGANIZÁVEIS

Viçosa – MG
2006

**UNIVERSIDADE FEDERAL DE VIÇOSA**

Eduardo Martins Viana

# AGRUPAMENTO DE CONJUNTOS CONSUMIDORES DE ENERGIA ELÉTRICA UTILIZANDO MAPAS AUTO-ORGANIZÁVEIS

Monografia apresentada ao Curso de Graduação em Engenharia Elétrica pela Universidade Federal de Viçosa, como requisito parcial à obtenção do título de Engenheiro Eletricista.

Orientador: Prof. Dr. David Calhau Jorge

Viçosa – MG
2006

Aos meus pais...

# AGRADECIMENTOS

A Deus, pela vida.

Aos meus pais Rubens e Suzana, que sempre me apoiaram em tudo, e que possibilitaram minha graduação. E aos demais familiares que sempre torceram e incentivaram.

Aos professores da Engenharia Elétrica da UFV, pelos conhecimentos transmitidos durante o curso de graduação. Em especial ao professor David, pela orientação, paciência e pelas idéias fornecidas para este trabalho.

Ao Maurício Sperandio da UFSC, que forneceu o banco de dados de empresas de energia elétrica e colaborou com idéias para realização deste trabalho.

Aos meus amigos de Ipatinga: Magnum, Maicon, Marcos, Valdeir, Warlei. Aos amigos que fiz durante o curso de Engenharia Elétrica, especialmente aos colegas da turma de 2001, pela parceria durante a graduação principalmente nos *Alambique's Parties e Alambicaretas*.

# RESUMO

## AGRUPAMENTO DE CONJUNTOS CONSUMIDORES DE ENERGIA ELÉTRICA UTILIZANDO MAPAS AUTO-ORGANIZÁVEIS

***Resumo:*** *O presente trabalho apresenta uma abordagem para o problema de agrupamento de conjuntos consumidores de energia elétrica. O processo de agrupamento apresentado é feito utilizando redes auto-organizáveis de Kohonen mais conhecidas como mapas auto-organizáveis (do inglês Self-Organizing Maps). Antes da utilização da rede de Kohonen, é feita a seleção de variáveis através de análise estatística, pela técnica da Análise Fatorial. A seleção de variáveis tem a finalidade de selecionar as que são mais representativas na base de dados, obtendo uma base de dados mínima para que o estudo possa ser feito. A obtenção de agrupamentos de consumidores de energia elétrica é útil para as concessionárias de energia elétrica para a definição das metas de qualidade de energia elétrica junto ao órgão regulador do setor energético, no caso brasileiro a ANEEL.*

***Palavras-chave:*** *energia elétrica, agrupamento de consumidores, Redes Neurais Artificiais, Mapas Auto-Organizáveis.*

## ABSTRACT

### ELECTRICAL ENERGY CONSUMERS GROUPING USING SELF-ORGANZING MAPS

***Abstract:*** *This work presents a new approach to the electrical consumer grouping problem. The grouping process proposed uses Kohonen's Self-Organizng Maps, one of the Artificial Neural Networks. Previous to utilize SOM networks the variable selection is processed using statistical analysis with Factor Analysis. The variable selection has the aim to select the most representatives variables in database, obtaining a minimum database helps to find the electrical energy consumer groups. The attainment of electrical energy consumers groups is useful for the electricity companies that have to define electricity quality index for electricity supply, with the regulator agency of electrical sector, in Brazil, ANEEL.*

***Keywords:*** *electrical energy, grouping consumers, Artificial Neural Networks, Self-Organizing Maps.*

**SUMÁRIO**

**LISTA DE FIGURAS**

**LISTA DE TABELAS**

# 1 INTRODUÇÃO

## 1.1 Motivação

O setor elétrico brasileiro vem passando, há algum tempo, por grandes transformações. Houve um processo de mudanças na sua estrutura, com a desverticalização das empresas de energia elétrica, separando as áreas de geração, transmissão e de distribuição. E posteriormente a essa ação de reestruturação, ocorreu o processo de privatização, devido à ausência de recursos do governo para investir na expansão do sistema. Essas mudanças no setor energético visam o aumento da eficiência e da qualidade da energia elétrica fornecida aos consumidores finais. A reestruturação ocorreu com os objetivos de assegurar os investimentos necessários para a expansão da oferta de energia e também para assegurar que o setor energético se tornasse mais eficiente. Para atingir estes objetivos, foram tomadas as seguintes medidas com relação ao setor elétrico (ZANINI, 2004):

- Desverticalização das empresas;
- Privatização;
- Aumento da eficiência na geração e distribuição;
- Despacho descentralizado;
- Livre acesso às redes de transmissão e distribuição.

Os grandes consumidores de energia passaram a ter livre acesso às redes de transmissão e distribuição. Assim, estes consumidores podem decidir de quem irão comprar a energia elétrica que consomem. O livre acesso às redes de transmissão e distribuição estimula a competitividade entre as empresas do setor. Ao contrário do consumidor de grande porte, o

consumidor de pequeno porte ou cativo está sujeito ao monopólio de uma empresa de distribuição e, portanto necessita de regulação. A regulação desse mercado é feita pela ANEEL (Agência Nacional de Energia Elétrica), criada justamente para esse fim.

A ANEEL determinou que as concessionárias deveriam atingir metas criadas por ela própria, a fim de melhorar a qualidade do fornecimento de energia elétrica. A qualidade do fornecimento é medida pelos indicadores DEC (Duração Equivalente de Interrupção por Unidade Consumidora), e FEC (Freqüência Equivalente de Interrupção por Unidade Consumidora), (DNAEE, 1978). As metas criadas eram determinadas para grupo de conjuntos consumidores, sendo cada grupo definido por cinco características de seu sistema elétrico e indicadores DEC e FEC passados. O cumprimento das metas determinadas pela ANEEL não foi possível em varias concessionárias, o que levou a ANEEL a emitir a resolução 024 de 27 de Janeiro de 2000, permitindo que as concessionárias proponha novos critérios para agrupamentos de conjuntos consumidores. A resolução 075 de 13 de Fevereiro de 2003 trata das metas de qualidade de fornecimento de energia.

## 1.2 Objetivos

Este trabalho tem o objetivo de obter agrupamentos de consumidores de energia elétrica utilizando redes SOM, posteriormente a uma análise estatística dos dados disponíveis. Esse estudo é feito a partir de dados disponibilizados pelas concessionárias de energia elétrica. Tendo em vista que nem todas as concessionárias têm muitos dados disponíveis, uma análise estatística dos dados será feita a fim de que seja determinada uma base de dados mínima capaz de representar o sistema. A análise estatística é feita através da Análise Fatorial.

Observa-se que as concessionárias de energia elétrica podem avaliar seus conjuntos consumidores obtidos através deste estudo e propor novas metas de qualidade de energia junto ao órgão regulador.

## 1.3 Organização do trabalho

No capitulo dois é feita uma introdução teórica sobre as ferramentas utilizadas no desenvolvimento da pesquisa. É feita uma descrição dos indicadores de qualidade da ANEEL e dos critérios adotados pela ANEEL para o agrupamento de conjuntos consumidores. Ainda no segundo capítulo uma breve explicação sobre Redes Neurais Artificiais e mapas auto-organizáveis. Os mapas auto-organizáveis serão utilizados como ferramenta de obtenção dos agrupamentos propostos neste trabalho.

O capítulo três trata dos métodos utilizados na pesquisa. Nesse item são mostradas as ferramentas estatísticas utilizadas para seleção de variáveis da base de dados. É mostrado também como foram obtidos os agrupamentos de consumidores de energia elétrica utilizando mapas auto-organizáveis.

O quarto capítulo faz a apresentação e discussão dos resultados. Por fim, no capítulo 5 encontram-se as conclusões retiradas da pesquisa e a sugestão para trabalhos futuros.

# 2 INTRODUÇÃO TEÓRICA

## 2.1 Distribuição de energia elétrica

Com a reestruturação do setor elétrico, foi necessária a criação da ANEEL, que tem dentre suas atribuições a competência de regular os serviços de energia elétrica, expedindo atos necessários ao cumprimento das normas estabelecidas pela legislação em vigor, e estimular a melhoria do serviço prestado. Em sua resolução 024/2000, a ANEEL define os termos: concessionária, unidade consumidora, e conjunto de unidades consumidoras, termos constantes nesse trabalho e que serão dados aqui apenas por conveniência.

Concessionária ou Permissionária é o agente titular de concessão ou permissão federal para explorar a prestação de serviços públicos de energia elétrica (ANEEL, 2000).

Unidade consumidora é um conjunto de instalações e equipamentos elétricos caracterizado pelo recebimento de energia elétrica em um só ponto de entrega, com medição individualizada e correspondente a um único consumidor (ANEEL, 2000).

Conjunto de unidades consumidoras é qualquer agrupamento de unidades consumidoras, global ou parcial, de uma mesma área de concessão de distribuição, definido pela concessionária ou permissionária e aprovado pela ANEEL (ANEEL, 2000).

### 2.1.1 Indicadores de Qualidade

A privatização do setor de distribuição de energia elétrica levou à necessidade de um maior controle da qualidade da energia elétrica fornecida aos consumidores finais. Para isso, torna-se importante o estabelecimento de índices de desempenho do fornecimento de modo

que seja possível o controle da qualidade de energia elétrica de forma objetiva (KAGAN, N., *et al*, 2005). São muitos os indicadores ligados à continuidade do fornecimento. Serão definidos aqui, cinco deles. Considerando as seguintes variáveis:

$C_a(i) \rightarrow$ Número de consumidores atingidos na interrupção $i$;

$C_s \rightarrow$ Número de consumidores existentes na área em estudo;

$t(i) \rightarrow$ Duração da interrupção de suprimento de energia elétrica $i$;

$T \rightarrow$ Período de estudo;

$N \rightarrow$ Número de ocorrências no período em estudo.

Define-se então para um determinado período, por exemplo, o mês, os índices operativos a seguir:

- Duração Equivalente por consumidor, (DEC): Exprime o tempo que, em média, cada consumidor da área considerada ficou privado da energia elétrica no período considerado. O DEC tem dimensão de tempo, geralmente hora ou minuto. Sendo o período de análise o mês, e a duração das contingências em minutos, o índice DEC representa quantos minutos o consumidor ficou sem energia elétrica no mês. Formalmente define-se a equação (2.1):

$$DEC = \frac{\sum_{i=1}^{N} C_a(i) \cdot t(i)}{C_S} \qquad (2.1)$$

- Freqüência equivalente de interrupção por consumidor, (FEC): Exprime o numero de interrupções que, em média, cada consumidor considerado sofreu. Este parâmetro é adimensional, representando o numero de interrupções sofridas pelo consumidor da área em estudo no período considerado. Formalmente define-se a equação (2.2):

$$FEC = \frac{\sum_{i=1}^{N} C_a(i)}{C_S} \qquad (2.2)$$

Com a necessidade de se ter um controle maior sobre cada unidade consumidora, o que é difícil realizar através de índices coletivos, DEC e FEC, são definidos outros três indicadores relacionados à duração e freqüência de interrupções de um dado consumidor. São eles:

- Duração de interrupção individual por unidade consumidora, DIC: Exprime o intervalo de tempo que no período de observação, em cada unidade consumidora ocorreu descontinuidade no fornecimento. Formalmente define-se a equação (2.3):

$$DIC = \sum_{i=1}^{N} t(i) \qquad (2.3)$$

- Freqüência de interrupção individual por unidade consumidora, FIC: Exprime o número de interrupções ocorridas no período considerado, em cada unidade consumidora. O FIC é simplesmente igual ao valor de N definido anteriormente. Assim o FIC é dado pela equação (2.4):

$$FIC = N \qquad (2.4)$$

- Duração máxima de interrupção contínua por unidade consumidora, DMIC: É o tempo máximo de interrupção contínua que uma unidade consumidora sofreu. Sendo definido pela equação (2.5):

$$DMIC = \max_{i=1,\ldots,N} [t(i)] \qquad (2.5)$$

É mensal o período de apuração do intervalo de tempo entre o início e o fim da contabilização das interrupções ocorridas no conjunto de unidades consumidoras considerado. A concessionária deve enviar a ANEEL os indicadores DEC e FEC de todos os seus conjuntos, até o último dia útil do mês subseqüente ao período de apuração (ANEEL, 2000). Os consumidores também recebem esses índices na sua fatura mensal de energia elétrica.

Com o objetivo de melhorar o fornecimento de energia elétrica aos consumidores, a ANEEL estabelece metas para os índices coletivos e individuais, DEC e FEC e DIC e FIC, respectivamente. Em caso de violação das metas estabelecidas no período de apuração as concessionárias deverão sofrer penalidades de acordo com o fato gerador. Particularmente para as metas individuais, a partir de 2005 os consumidores que tiveram suas metas transgredidas, têm direito a ressarcimento direto em sua conta de energia (KAGAN, N., *et al*, 2005).

## 2.2 Outras abordagens para o problema

As metas atualmente aplicadas pela ANEEL estão descritas em (TANURE, 2001). Foram considerados cinco atributos básicos, respeitando-se as atuais limitações das empresas para informar os dados de seus sistemas de forma confiável. Tais variáveis são:

1. Área de cada conjunto em $km^2$;
2. Extensão da rede primária em *km*;
3. Potência instalada em *kVA*;
4. Número de consumidores;
5. Consumo médio mensal de cada conjunto em *MWh*.

Em (TANURE, 2001) foram reunidos dados de 4135 conjuntos de 56 concessionárias de todo o Brasil. Para realizar o agrupamento dos conjuntos consumidores, foi utilizado o método estatístico *k-means* no qual é necessário definir *a priori* a quantidade de grupos a serem formados no final do processo. Os dados passaram por uma transformação logarítmica para reduzir a dispersão dos mesmos, e normalização. Com os agrupamentos formados, as metas da ANEEL relativas aos indicadores DEC e FEC foram determinadas.

Outra abordagem é estudada em (SPERANDIO, 2003). Nesse trabalho foram utilizados 16 atributos de cada conjunto consumidor de energia elétrica (entre eles os indicadores DEC e FEC) e foi utilizado o algoritmo dos mapas auto-organizáveis para agrupamentos dos dados. Em (SPERANDIO, 2004a), é realizada uma abordagem mais sistemática do problema. Nesse trabalho, foram utilizados 18 atributos divididos em variáveis de mercado e variáveis de sistema de cada conjunto consumidor de energia elétrica, não sendo incluído nesses atributos, os indicadores DEC e FEC. Foi utilizado o mapa auto-organizável em conjunto com o método de agrupamento *k-means*, onde o *k-means* foi utilizado como ferramenta de validação dos agrupamentos obtidos pelo algoritmo SOM. Ainda foi feito um cruzamento entre os dois tipos de variáveis, para se verificar se as metas da ANEEL para determinado agrupamento estavam de acordo com a capacidade do sistema de distribuição.

Em (SPERANDIO, 2004b) são utilizadas também 18 atributos, divididos em variáveis de mercado e variáveis de sistema, porém, neste trabalho, não foi utilizada a análise estatística.

## 2.3 Redes Neurais

As origens das redes neurais artificiais remontam no desejo de construir artefatos capazes de exibir comportamento inteligente, ou seja, na inteligência artificial. A inteligência artificial é definida como um campo da ciência que visa reproduzir por meios computacionais as características normalmente atribuídas à inteligência humana, tais como aprendizagem, reconhecimento de padrões e solução de problemas. Inspiradas nas redes neurais biológicas, as Redes Neurais Artificiais tem sido aplicadas com sucesso na solução de problemas em vários e distintos domínios, tais como: processamento de sinais, controle, reconhecimento de padrões, medicina, reconhecimento e produção de voz, e negócios. Uma de suas propriedades, de importância fundamental, é a capacidade de aprender a partir de seu ambiente e de aperfeiçoar sua performance através do aprendizado (HAYKIN, 2001). Convencionou-se chamar redes neurais artificiais a toda topologia de processamento de sinais constituída de vários elementos processadores simples altamente interconectados.

Uma rede neural artificial, ou simplesmente rede neural, é um sistema de processamento massivamente paralelo, composto por unidades simples com capacidade natural de armazenar conhecimento e disponibilizá-lo para uso futuro (HAYKIN, 2001). A figura 2.1 apresenta um modelo de neurônio biológico com os nomes de cada parte do neurônio. O fluxo da informação (corrente elétrica) é dos dendritos para o axônio. Ocorre um processo de soma dos estímulos de entrada, produzindo um impulso elétrico que se propaga através do axônio, caso a soma das entradas seja maior que um certo limiar.

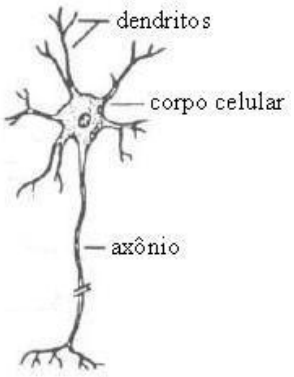



Figura 2.1 - Modelo do neurônio biológico.

A figura 2.2 mostra o modelo matemático de um neurônio, que é a unidade fundamental de processamento para a operação da rede neural. O modelo apresentado foi baseado no modelo biológico e proposto pelo psiquiatra e neuroanatomista McCulloch e pelo matemático Pitts em 1943.

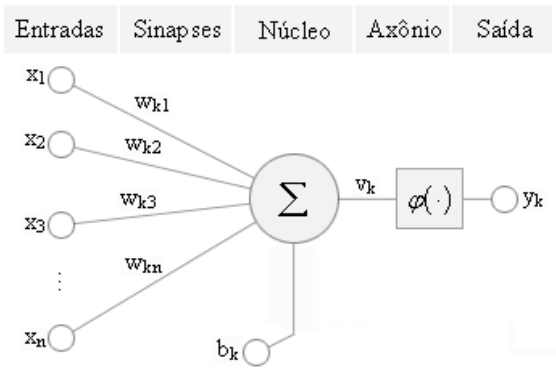



Figura 2.2 - Modelo matemático de um neurônio.

No modelo mostrado na figura 2.2, podem-se identificar três os elementos básicos no neurônio artificial:

- Sinapses: as sinapses são elos de conexão entre as entradas e o bloco somador do neurônio. Cada sinapse é caracterizada por um vetor peso próprio. Um sinal de entrada $x_j$ na entrada da sinapse $j$ conectada ao neurônio $k$ é multiplicada pelo peso sináptico $w_{kj}$.

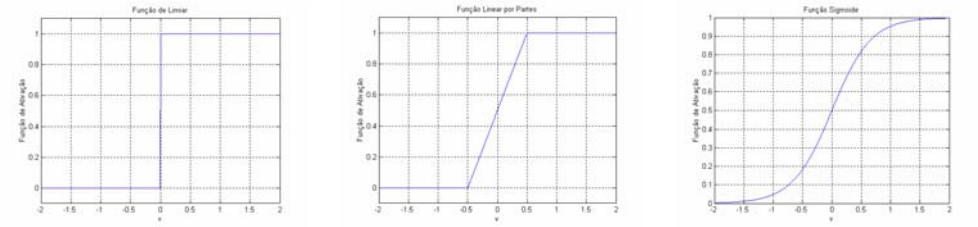
- Somador: este bloco é responsável por somar sinais de entrada, ponderados pelas respectivas sinapses. Na figura 2.2 ele está representado pelo bloco denominado *Núcleo*.

- Função de ativação: também denominada de função restritiva, tem como finalidade limitar a amplitude do sinal de saída de um neurônio. Na figura 2.2, é representado pelo bloco denominado *Axônio*. Geralmente o valor da amplitude de saída é normalizado estando no intervalo [0,1] ou ainda [-1,1]. As funções de ativação normalmente utilizadas são de três tipos:

1. Função de Limiar ou função de Heaviside: esse tipo de função de ativação é dada pelas seguintes condições $\varphi(v) = \begin{cases} 1, \text{se } v \geq 0 \\ 0, \text{se } v < 0 \end{cases}$ onde $v$ é o potencial de ativação. A função de limiar pode ser vista na figura 2.3(a).

2. Função Linear por Partes: a função linear por partes é descrita pelas seguintes condições $\varphi(v) = \begin{cases} 1, \text{se } v \geq k \\ v, \text{se } -k < v < k \\ 0, \text{se } v \leq -k \end{cases}$, onde $k$ é um número inteiro. Esta forma de função de ativação é uma aproximação de um amplificador não linear. A função linear por partes é mostrada na figura 2.3(b).

3. Função Sigmóide: essa é a forma mais comum de função de ativação utilizada na construção de redes neurais. O gráfico dessa função tem forma de S e pode ser visto na figura 2.3(c). A função sigmóide é definida por $\varphi(v) = \frac{1}{1+e^{-av}}$, onde $a$ é o parâmetro de inclinação da sigmóide. Note que a função sigmóide aproxima-se da função de limiar quando $\lim_{a \to \infty} \varphi(v)$.




(a) (b) (c)

Figura 2.3 - Funções de ativação: (a) Função de limiar; (b) Função linear por partes; (c) Sigmóide.

O modelo neuronal mostrado na figura 2.2 tem ainda uma outra entrada denominada ajuste, ou *bias*, representado por $b_k$. A entrada *bias* é aplicada externamente, e tem a finalidade de aumentar ou diminuir a entrada líquida da função de ativação. O *bias* pode ser positivo ou negativo.

Um neurônio *k* fica descrito então pelas equações (2.6), (2.7) e (2.8):

$$u_k = \sum_{j=1}^{n} w_{kj} x_j \qquad (2.6)$$

$$v_k = u_k + b_k \qquad (2.7)$$

$$y_k = \varphi(u_k + b_k) \qquad (2.8)$$

onde $x_1$, $x_2$,..., $x_n$ são os sinais de entrada, $w_{k1}$, $w_{k2}$,..., $w_{kn}$ são os pesos sinápticos do neurônio k, $u_k$ é a saída do combinador linear devido aos sinais de entrada, $b_k$ é o *bias* e $\varphi(.)$ é a função de ativação e $y_k$ é a saída do neurônio k. O uso do *bias* tem o efeito de modificar o potencial de ativação $v_k$ do neurônio k.

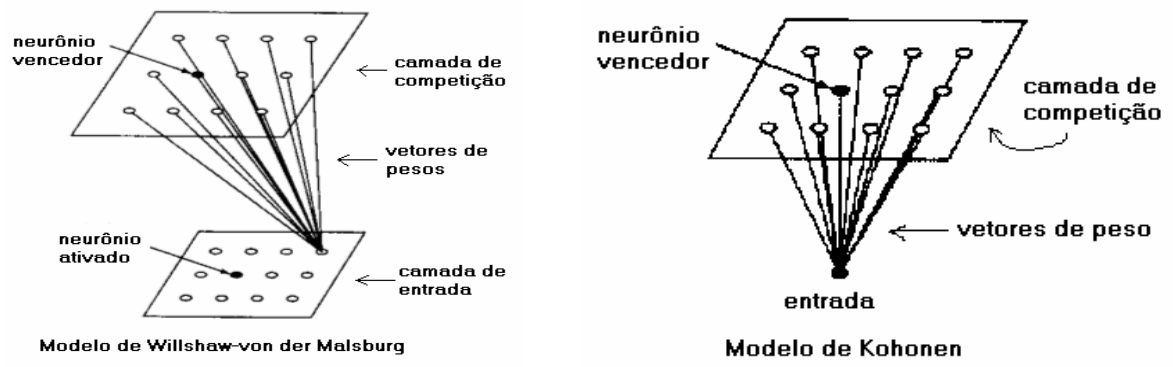
### 2.3.1 Mapas auto-organizáveis

O Mapa Auto-Organizável, ou *self-organizing map* (SOM), é um tipo de rede neural artificial baseada em aprendizado competitivo e não supervisionado, sendo capaz de mapear um conjunto de dados, de um espaço de entrada contínuo contido em $\Re^M$, em um conjunto finito de neurônios organizados em um arranjo normalmente bidimensional.

O SOM realiza uma transformação não-linear dos dados de entrada, em $\Re^M$, para o espaço de dados do arranjo, em $\Re^N$, executando uma redução dimensional (FLEXER, A., 2001). Como o espaço $\Re^N$ é normalmente bidimensional, então N = 2, como no caso da figura 2.4. Mapas de dimensionalidade mais alta também são possíveis, embora não sejam comuns. Ao realizar a transformação não-linear, $\Re^M \longrightarrow \Re^N$, (TÖRMÄ, 1994), o algoritmo tenta preservar ao máximo a topologia do espaço original, ou seja, procura fazer com que neurônios vizinhos no arranjo apresentem vetores de pesos que retratem as relações de vizinhança entre os dados. Para tanto, os neurônios competem para representar cada dado, e o neurônio vencedor tem seu vetor de pesos ajustados de modo a ficar mais próximo do dado de entrada. Esta redução de dimensionalidade com preservação topológica permite ampliar a capacidade de análise de agrupamentos dos dados pertencentes a espaços de elevada dimensão, o que é difícil de ser feito via métodos estatísticos convencionais. O SOM é utilizado numa diversidade de aplicações de agrupamento de dados e extração de conhecimento de uma base de dados (KOHONEN, 2000).

### 2.3.1.1 Modelo formal da Rede de Kohonen

O córtex cerebral talvez seja a estrutura mais complexa conhecida no universo. É impressionante a extensão que o córtex cerebral ocupa no cérebro. O que também é impressionante é o modo como diferentes entradas sensoriais, são mapeadas para áreas correspondentes do córtex de uma forma ordenada (HAYKIN, 2001). Isso significa que os neurônios tornam-se sensíveis a determinados estímulos em particular e a outros não, especializando-se no processamento de um determinado sinal, o que pode ser explicado pela separação dos canais nervosos que ligam os órgãos sensoriais ao cérebro. Este princípio forneceu a motivação neurobiológica para dois modelos de mapeamento de características, mostrados na figura 2.4.




(a) (b)

Figura 2.4 - Modelos de mapeamento de características.

O primeiro modelo, proposto por Willshaw e Von der Malsburg, é composto por duas grades bidimensionais separadas por neurônios conectadas entre si, uma delas se projetando sobre a outra. O segundo modelo, proposto por Kohonen, não pretende explicar os detalhes neurobiológicos, mas contém as características essenciais do mapeamento de caracteres. O modelo de Kohonen é mais geral que o modelo de Willshaw-Von der Malsburg, uma vez que ele é capaz de reduzir a dimensão dos dados de entrada.

Analisando-se mais especificamente estas regiões especializadas, há evidências de uma organização um pouco mais abstrata e complexa: suas células organizam-se e tornam-se sensíveis aos estímulos de acordo com uma ordem topológica que especifica uma relação de similaridade entre os sinais de entrada. Assim, os neurônios exibem uma ordenação física tal que estímulos semelhantes no espaço de dados são processados por neurônios fisicamente próximos entre si no córtex cerebral (ULTSCH, 1994). A formação de mapeamentos

topologicamente corretos é atribuída a uma diversidade de mecanismos, dos quais um em particular, a auto-organização, ou processo pelo qual estruturas com ordem global são obtidas através de interações locais entre os elementos (KOHONEN, 2000). Isto levou à proposição de vários modelos de mapas topográficos entre eles o modelo baseado em aprendizado competitivo. Neste modelo é que se baseia o SOM. O aprendizado competitivo pode ser sucintamente descrito desta forma:

- Uma dada amostra de dados de $\Re^M$ é apresentado a uma rede composta por neurônios organizados de forma específica, cada um com seu vetor de pesos;

- Para cada vetor apresentado à rede haverá um neurônio que se assemelha ao vetor de entrada. Este neurônio é conhecido como BMU, ou *Best Matching Unit*. Este neurônio é determinado por um processo competitivo.

- O BMU tem seu vetor de pesos ajustados, de modo a se aproximar ainda mais do dado de entrada da rede. Os vizinhos próximos a este também tem seus pesos ajustados. O ajuste nos pesos é feito para que se aumente a probabilidade do neurônio vencer a competição numa subseqüente apresentação do mesmo dado.

A idéia do mapeamento de características é a que neurônios próximos representem dados próximos, no conjunto de dados do espaço $\Re^M$. Para isso é utilizado um mecanismo excitatório de curto alcance e um mecanismo inibitório de longo alcance. Estes são os dois mecanismos de natureza local e são cruciais para a auto-organização (HAYKIN, 2001).

O principal objetivo do mapa auto-organizável é transformar um padrão de sinal incidente de dimensão arbitrária em um mapa discreto unidimensional ou bidimensional e realizar esta transformação adaptativamente de uma maneira topologicamente ordenada.

Considerando que M represente a dimensão do espaço de entrada, $\Re^M$, um padrão selecionado aleatoriamente nesse espaço é dado por (2.9).

$$x = [x_1, x_2, ..., x_m]^T \qquad (2.9)$$

O vetor peso sináptico de cada neurônio da grade tem a mesma dimensão do espaço de entrada $\Re^M$. O vetor peso sináptico do neurônio $j$ é representado por (2.10).

$$w_j = \left[w_{j1}, w_{j2}, ...w_{jm},\right]^T \qquad j = 1,2,...,l \qquad (2.10)$$

onde $l$ é o numero total de neurônios na grade. Para se determinar o neurônio que é mais parecido com a entrada $x$, basta comparar os produtos internos $w_j^T x$ para $j = 1,2,...,l$ e selecionar o maior deles. O critério de casamento do vetor $x$ com o neurônio $j$ baseado na maximização do produto interno é matematicamente equivalente a minimizar a distância euclidiana entre os vetores $x$ e $w$. A determinação do neurônio com maior produto interno determina a localização onde a vizinhança topológica dos neurônios excitados deve ser centrada. O índice $i(x)$ será usado para identificar o neurônio que melhor casa com o vetor de entrada $x$. Esse índice é dado por (2.11).

$$i(x) = \arg\min_j \left\| x - w_j \right\| \qquad j = 1,2,...,l \qquad (2.11)$$

O neurônio que satisfaz essa condição é chamado de neurônio vencedor para o vetor de entrada $x$. O neurônio vencedor localiza o centro de uma vizinhança topológica de neurônios cooperativos. É necessário então, definir uma vizinhança topológica que seja capaz de ajustar com mais ênfase os neurônios mais próximos ao centro da vizinhança e com menos ênfase os neurônios mais distantes. Para isso, a vizinhança topológica em torno do neurônio vencedor, deve ser simétrica em relação ao centro e decair suavemente com a distancia lateral. Considere então que $h_{j,i}$ represente a função vizinhança topológica centrada no neurônio vencedor $i$ e que contenha um conjunto de neurônios excitados, sendo um neurônio representado por $j$. A função vizinhança é tipicamente uma *gaussiana* (figura 2.5) dada pela equação (2.12).

$$h_{j,i(x)} = \exp\left( -\frac{d_{j,i}^2}{2\sigma^2} \right) \qquad (2.12)$$

onde:

$h_{j,i(x)} \rightarrow$ Função de vizinhança para o vizinho $j$ em relação ao vencedor $i$ para o

dado $x$, que determina o nível de cooperação;

$d_{j,i} \rightarrow$ Distância lateral entre o vizinho $j$ e o neurônio vencedor $i$;

$\sigma \rightarrow$ Largura efetiva da vizinhança na interação.

O parâmetro $d_{j,i}$, distância lateral entre o vizinho $j$ e o neurônio vencedor $i$, é calculado pela equação (2.13).

$$d_{j,i} = \left\| r_j - r_i \right\| = \sqrt{\sum_{j=1}^{M} \left\| r_j - r_i \right\|^2} \qquad (2.13)$$

onde o vetor discreto $r_j$ define a posição do neurônio excitado $j$ e $r_i$ define a posição discreta do neurônio vencedor $i$. A figura 2.5 mostra uma típica curva gaussiana, onde a amplitude determina o nível de cooperação entre o neurônio vencedor e seus vizinhos.

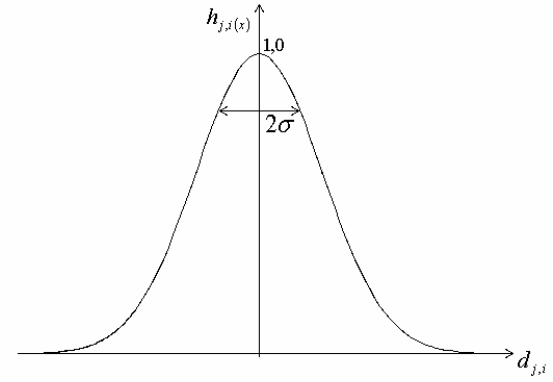



Figura 2.5 - Função de vizinhança gaussiana.

Outra característica do algoritmo SOM, é que o tamanho da vizinha topológica diminui com o tempo. Isso é feito diminuindo-se a largura efetiva da vizinhança, $\sigma$, com o tempo. Uma escolha para essa variação é o decaimento exponencial de $\sigma$ com o tempo discreto $n$, dada por (2.14).

$$\sigma(n) = \sigma_0 \exp\left( -\frac{n}{\tau_1} \right) \qquad (2.14)$$

onde $\sigma_0$ é o valor de $\sigma$ na inicialização do algoritmo SOM e $\tau_1$ é uma constante de tempo. A variação de $\sigma$ com o tempo discreto $n$, torna a função de vizinhança também variável com o tempo, como mostra a equação (2.15).

$$h_{j,i(x)}(n) = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2(n)}\right) \qquad (2.15)$$

Assim quando o número de iterações $n$ aumenta, a largura $\sigma(n)$ decresce a uma taxa exponencial e a função de vizinhança de maneira correspondente.

Para que o mapa seja auto-organizável, é necessário que o vetor de peso sináptico $w_j$ do neurônio $j$ da grade se modifique em relação ao vetor de entrada $x$. Para o isso o vetor peso sináptico $w_j(n)$ do neurônio $j$ no tempo discreto $n$ é ajustado pela equação (2.16).

$$w_j(n+1) = w_j(n) + \eta(n)h_{j,i(x)}(n)(x - w_j(n)) \qquad (2.16)$$

onde: $w_j(n)$ é o vetor de pesos sinápticos no tempo discreto $n$;

$w_j(n+1)$ é o vetor de pesos atualizado, no tempo $n+1$;

$\eta(n)$ é o parâmetro taxa de aprendizagem dado por (2.17);

$h_{j,i(x)}(n)$ é a função de vizinhança definida em (2.15);

$x$ é o vetor de entrada.

A equação tem o efeito de mover o vetor peso sináptico $w_i$ do neurônio vencedor $i$ em direção ao vetor $x$. A equação é aplicada a todos os neurônios que se encontram dentro da vizinhança topológica. O parâmetro taxa de aprendizagem $\eta(n)$ é variável com o tempo e ele deve diminuir gradualmente com o tempo. Para isso escolhe-se a taxa de aprendizagem como sendo exponencial (2.17).

$$\eta(n) = \eta_0 \exp\left(-\frac{n}{\tau_2}\right) \qquad n = 0,1,2,... \qquad (2.17)$$

onde $\tau_2$ é uma constante de tempo.

Os parâmetros apresentados nas equações podem receber uma infinidade de valores para formarem o processo de auto-organização, porém para representação ordenada dos padrões de entrada, esses parâmetros devem ser selecionados adequadamente. O processo de adaptação sináptica, mostrado pela equação (2.16), é composto de duas fases: *fase de ordenação* e *fase de convergência*. Durante a fase de organização, que exige aproximadamente 1000 iterações, a função de vizinhança deve incluir quase todos os neurônios da grade e a taxa de aprendizagem deve ser alta, para permitir uma aproximação mais rápida de neurônios semelhantes e já define a característica topológica do mapa. A fase de convergência é uma espécie de sintonia fina do mapa, de modo a produzir uma quantização estatística precisa do espaço de entrada. Essa fase dura no mínimo 500 vezes o número de neurônios da rede. Alguns parâmetros para treinamento da rede (HAYKIN, 2001) são dados a seguir.

- Fase de ordenação:

$\eta_0 = 0{,}1$;

$\tau_2 = 1000$;

$\sigma_0 =$ raio da grade;

$\tau_1 = \dfrac{1000}{\log(\sigma_0)}$.

- Fase de convergência:

$\eta(n) \rightarrow$ valor pequeno da ordem de 0,01;

$\sigma_0 \rightarrow$ valor pequeno comparado ao tamanho da grade;

$\tau_2 = 1000$;

$\tau_1 = \dfrac{1000}{\log(\sigma_0)}$.

#### 2.3.1.2 Visualização da rede SOM

A interpretação do resultado final do algoritmo de auto-organização é muito importante para a correta obtenção dos resultados, e para extrair o máximo de informações que esse método proporciona (SPERANDIO, 2004a).

Após o processo de auto-organização do mapa, um elemento da base de dados é representado por apenas um neurônio da grade. Porém, um neurônio poderá apresentar vários elementos. E há ainda neurônios que não estão associados a nenhum dado de entrada. Esses neurônios não conseguem vencer o processo competitivo descrito anteriormente e ficam adaptados ao espaço vazio que separa os grupos obtidos.

Quando se utiliza o mapa auto-organizável, duas ferramentas básicas são obtidas, os mapas componentes e a matriz de distâncias (*U-mat*). Os mapas componentes, mostram a influência de cada variável na organização final da rede. A propriedade de abstração do SOM está juntamente nos mapas componentes, segundo os quais, é possível determinar como as variáveis influenciaram na obtenção do resultado final. A matriz de distâncias contém a distância euclidiana entre ou neurônios vizinhos, possibilitando assim a definição dos agrupamentos de neurônios. Quando essas distâncias são colocadas em escala de cores, as maiores distâncias entre neurônios, tem cores mais fortes, representando assim a fronteira entre dois grupos.

Para esclarecer o conceito de mapas componentes e a matriz de distâncias, será dado um exemplo a seguir. A base de dados utilizada nesse exemplo, contém 50 amostras de três tipos de flores contendo quatro medidas diferentes, um total de 150 amostras. As medidas das amostras são largura e altura das sépalas e pétalas.

Primeiramente, o mapa é gerado, podendo ser aleatoriamente ou linearmente. Para um mapa 7 x 7 gerado aleatoriamente, a configuração inicial é mostrada na figura 2.6.

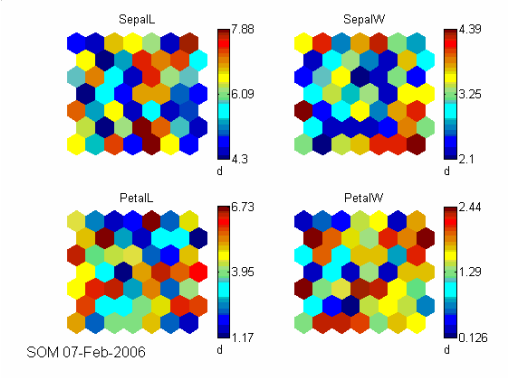



Figura 2.6 - Mapas componentes com a distribuição inicial dos dados.

Com o mapa gerado, procede-se o treinamento da rede com os dados, e após o processo de organização, o mapa ficou como mostrado na figura 2.7.

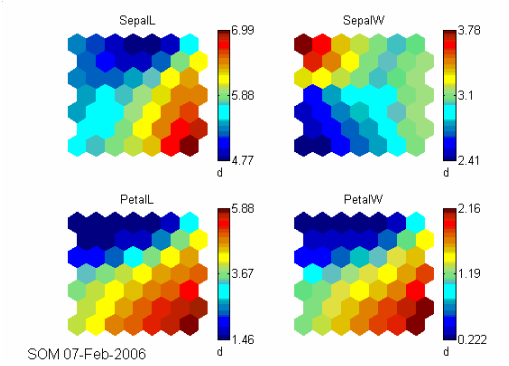



Figura 2.7 - Mapa treinado.

Na figura 2.7, têm-se os mapas componentes já treinados. Observa-se que a distribuição dos dados no mapa treinado tomou uma forma organizada, sendo que valores mais altos nas variáveis são representados pela cor vermelha, os valores médios pela cor amarela e verde, e os valores baixos pela cor azul. A posição de cada neurônio é a mesma em todos os mapas. Seus modelos é que são adaptados de maneira a formar um estado organizado da rede.

Para verificar se dados alocados em neurônios adjacentes estão próximos, é necessário observar a matriz de distâncias. A *U-mat*, quando apresentada na forma de um mapa, tem mais hexágonos que os mapas componentes, tendo a dimensão de $2 \cdot N_l - 1$ por $2 \cdot N_c - 1$, onde $N_l$ e $N_c$ são respectivamente número de linhas e de colunas de neurônios no mapa. O maior número de hexágonos na *U-mat* é devido ao fato de que a distância entre neurônios adjacentes é mostrada, junto com os neurônios do próprio mapa. Valores altos na

matriz de distâncias, representam grandes distâncias entre neurônios vizinhos, e isso indica a fronteira de um agrupamento de dados. Os agrupamentos são áreas uniformes com pequenas distâncias entre neurônios (VESANTO, 1999). Para o exemplo citado anteriormente, a representação da *U-mat* em forma de mapa está mostrada na figura 2.8.

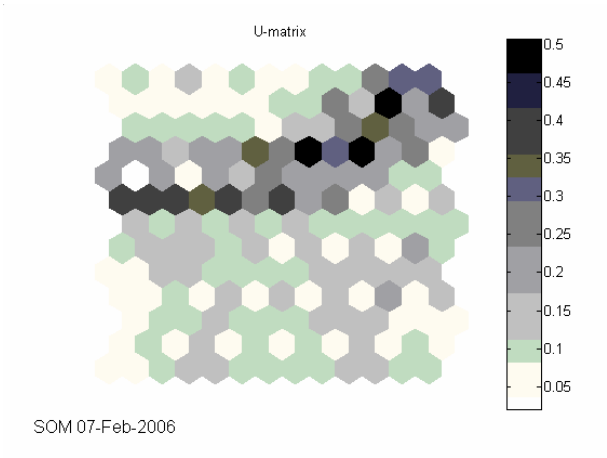



Figura 2.8 - *U-mat*.

Na figura 2.8, nota-se dois grupos principais, divididos pela faixa escura que desce da parte superior direita até a quase a região central esquerda. Isso indica a presença de pelo menos dois agrupamentos. Porém, observando-se a parte inferior do mapa, é possível ver uma parte branca na região esquerda e outra parte branca na parte direita. Isso indica a presença de outros dois grupos, totalizando três grupos.

Como citado anteriormente, os dados utilizados nesse exemplo consistem de medidas feitas em três tipos de flores de uma mesma espécie, assim os grupos já eram conhecidos. Na figura 2.9 é mostrada a distribuição já conhecida dos grupos sobre a *U-mat*.

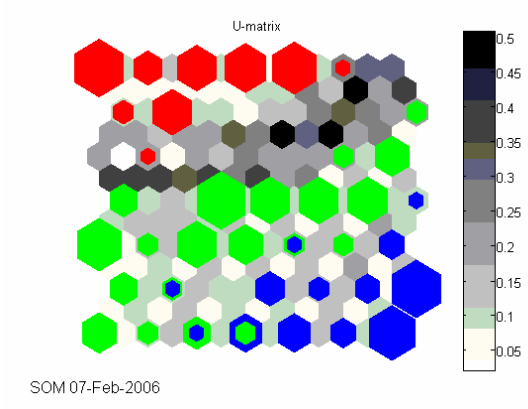



Figura 2.9 - Distribuição dos grupos já conhecidos sobre a *U-mat*.

Os hexágonos maiores que se sobrepõe ao mapa representam os três grupos que já eram conhecidos *a priori*, onde cada cor representa um grupo. Assim fica claro a classificação correta de uma base de dados com três grupos propositalmente já conhecidos, validando assim o uso do algoritmo SOM para agrupamento de dados.

Outro tipo visualização de agrupamentos, que será utilizado neste trabalho, foi proposto em (PAMPALK *et al*., 2002). A idéia principal nesse método é que agrupamentos de dados são áreas no espaço de dados com alta densidade probabilística. O método utiliza histogramas de freqüência suavizados (SDH) para estimar a densidade de probabilidade e determinar os centros dos agrupamentos de neurônios. Abaixo a representação de um histograma de freqüência suavizada.

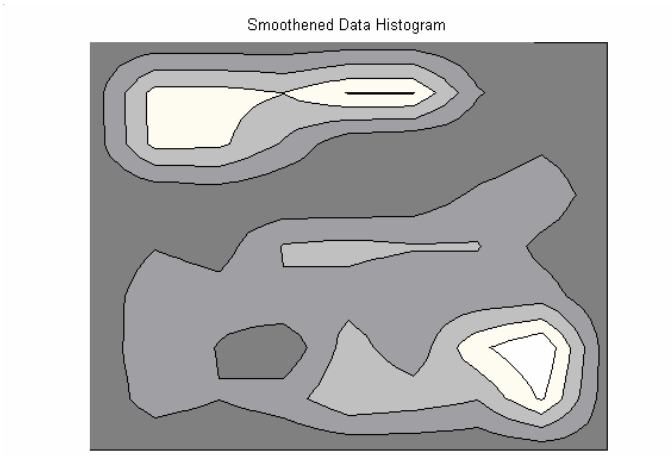



Figura 2.10 - Histograma de freqüência suavizada.

A figura 2.10 mostra curvas de nível que indicam as distribuições dos dados nos neurônios da rede. As cores mais claras mostram a localização de neurônios que representam muitos dados. Note que as cores brancas na figura 2.10 correspondem corretamente com os maiores hexágonos destacados na figura 2.9. As áreas mais escuras representam neurônios com poucos dados agrupados ou mesmo sem nenhum dado agrupado. Tal como na *U-mat*, é possível definir dois grandes grupos, um superior e outro inferior, e ainda dentro do grupo inferior é possível definir dois outros grupos, sendo ao todo três grupos.

# 3 METODOLOGIA

A metodologia proposta consiste em tomar uma base de dados com parâmetros diferentes dos utilizados pela ANEEL, fazer o tratamento estatístico dos dados e aplicá-los à rede SOM. Para isso utilizou-se uma base de dados composta por 16 atributos para cada conjunto consumidor considerado. Esses atributos foram separados em dois grupos de variáveis denominadas variáveis de mercado consumidor e variáveis técnicas. São elas:

As variáveis de Mercado Consumidor:

- Consumo Residencial anual em *MWh*;
- Consumo Industrial anual em *MWh*;
- Consumo Rural anual em *MWh*;
- Consumo Comercial anual em *MWh*;
- Consumo Público anual em *MWh*;
- Número de Consumidores Residenciais;
- Número de Consumidores Industriais;
- Número de Consumidores Rurais;
- Número de Consumidores Comerciais;
- Número de Consumidores Públicos.

As variáveis Técnicas:

- Número de Subestações;

- Número de Alimentadores;
- Capacidade de Manobra entre Alimentadores (*Sim / Não*);
- Distância da Sede do Município a SE em *km* ;
- Índice DEC;
- Índice FEC.

Apesar da base de dados não ser muito extensa, apenas 16 variáveis, ela pode conter dados redundantes, por isso é feita a análise estatística dos mesmos a fim de se encontrar dados que possam ser redundantes e conseqüentemente retirados da base de dados. Dados redundantes, não trazem benefícios aos resultados finais e ainda exigem maior esforço computacional na simulação. Assim, a avaliação das variáveis disponíveis, tem o objetivo de reduzir ainda mais a base de dados, obtendo um conjunto mínimo de variáveis representativas. Uma avaliação inicial das variáveis pode ser feita utilizando-se a matriz de correlações entre as variáveis. Observando o coeficiente de correlação entre as variáveis é possível perceber quais variáveis são influenciadas por outras, e o nível dessa influência. Valores positivos de correlação entre duas variáveis significam que um valor alto na primeira variável está associado a um valor alto na segunda variável. Valores negativos de correlação entre duas variáveis significam que um valor alto na primeira variável está associado a um valor baixo na segunda variável. Entretanto, somente esta análise não é suficiente.

Uma melhor maneira de se avaliar as variáveis é a técnica da Análise Fatorial, segundo a qual é possível descrever um grupo de *p* variáveis $X_1, X_2, ..., X_p$ em termos de um número reduzido de variáveis (MANLY, 1994). A Análise Fatorial é feita através da Análise de Componentes Principais que transforma as variáveis originais em variáveis novas, não-correlacionadas. Essa técnica reduz o número de variáveis, retendo o máximo de informação e ainda tornando as variáveis restantes mais significativas. É importante frisar que se as variáveis originais já forem não-correlacionadas, a extração de componentes principais não faz absolutamente nada. Com as informações obtidas da análise fatorial, podem-se escolher quais variáveis da base de dados serão utilizadas. As análises são feitas para os dois grupos de variáveis separadamente, variáveis de mercado consumidor, variáveis técnicas. Porém antes da análise das variáveis através de análise fatorial é feita uma análise descritiva dos dados, onde se busca compreender a distribuição dos mesmos.

Na análise descritiva dos dados são calculados os seguintes parâmetros: amplitude, média, desvio padrão, mediana e quartis. A média aritmética serve para determinar a

tendência da maioria das medidas de uma variável. Entretanto se no conjunto de dados existir um valor errôneo ou um dado discrepante, a média será fortemente influenciada e não mais representará a tendência dos valores. Para isso são tomadas medidas de tendências centrais como a mediana, que divide o grupo de valores ao meio. Isso significa que os valores são colocados em ordem crescente e o valor que representa a metade do grupo é tomado como mediana. Os quartis também são uma boa medida para se avaliar os dados já que estes dividem o conjunto de dados em quatro grupos, sendo que o *quartil 25%* contém os 25% de valores menores, o *quartil 50%* (mesmo valor da mediana) divide o conjunto ao meio, e por fim o *quartil 75%* contém 75% dos dados de menor valor.

## 3.1 Seleção de variáveis

### 3.1.1 Variáveis de Mercado Consumidor

Como citado anteriormente as variáveis estão separadas em dois grupos, variáveis de mercado consumidor e variáveis técnicas. Para o mercado consumidor, são mostradas as tabelas 01 e 02 com a análise descritiva dos dados.

Tabela 01 - Análise descritiva das variáveis de mercado consumidor (Consumo, em *MWh*/ano).

| Cons. *MWh* | Média | Desvio Padrão | Quartil 25% | Quartil 50% | Quartil 75% | Amplitude |
|---|---|---|---|---|---|---|
| Residencial | 13,0718 | 42,2402 | 0,7093 | 1,9468 | 6,6326 | 384,9615 |
| Industrial | 23,9329 | 102,0646 | 0,3073 | 1,8626 | 13,3153 | 1239,9912 |
| Comercial | 7,6502 | 27,6749 | 0,2932 | 0,9086 | 2,7046 | 268,3425 |
| Rural | 2,7218 | 3,5293 | 1,0103 | 1,9244 | 3,1971 | 27,1153 |
| Público | 3,8066 | 12,7848 | 0,3659 | 0,9082 | 2,0476 | 138,8078 |

Tabela 02 - Análise descritiva das variáveis de mercado consumidor (Número de consumidores).

| N. Cons. | Média | Desvio Padrão | Quartil 25% | Quartil 50% | Quartil 75% | Amplitude |
|---|---|---|---|---|---|---|
| Residencial | 6059,0000 | 16930,9574 | 385 | 1152 | 4098 | 147484 |
| Industrial | 233,6118 | 557,4602 | 19 | 48 | 171 | 5253 |
| Comercial | 611,9059 | 1778,6662 | 70 | 150 | 351 | 17363 |
| Rural | 772,5000 | 676,8187 | 353 | 614 | 1029 | 4853 |
| Público | 69,3353 | 103,7791 | 28 | 43 | 64 | 1049 |

A tabela com a análise descritiva dos dados foi separada em duas novas tabelas, uma contendo apenas dados de consumo das diversas classes em *MWh* / ano e a outra contendo dados referentes ao número de consumidores.

Na tabela 01, nota-se que as classes residencial, industrial, comercial e público têm a média acima da mediana, ou *quartil 50%*. Na verdade elas têm a média até maior que o *quartil 75%*. Isso indica a presença de municípios com grande consumo de energia, para essas classes, o que pode ser confirmado pela amplitude que é alta. A mediana para as diversas classes mostra-se mais homogênea que a média, significando uma distribuição de consumo parecida entre metade dos municípios estudados. Para as variáveis número total de consumidores de cada classe nota-se mais uma vez a grande diferença entre a média e o *quartil 50%*, exceto para a classe rural e pública.

Feita a análise descritiva, passa-se agora para a análise fatorial das variáveis, com o intuito de procurar variáveis que sejam redundantes e que possam ser retiradas da base de dados. Por exemplo, espera-se que variáveis como o consumo de cada classe seja proporcional ao número de consumidores da respectiva classe. Entretanto, veremos que essa relação não é tão direta como se imagina e que cada variável não está relacionada somente a ela, mas sim a uma série de outras variáveis. A verificação da dependência entre as variáveis é feita a partir da matriz de correlação, mostrada a seguir na forma de uma tabela.

Tabela 03 - Tabela de correlação das variáveis de mercado consumidor.

| | C. Res. | C. Ind | C. Com. | C. Rur. | C. Pub. | N. Res. | N. Ind | N. Com. | N. Rur. | N. Pub. |
|---|---|---|---|---|---|---|---|---|---|---|
| C. Res. | 1,0000 | - | - | - | - | - | - | - | - | - |
| C. Ind. | 0,6409 | 1,0000 | - | - | - | - | - | - | - | - |
| C. Com. | **0,9782** | 0,5487 | 1,0000 | - | - | - | - | - | - | - |
| C. Rur. | 0,0191 | 0,1492 | 0,0274 | 1,0000 | - | - | - | - | - | - |
| C. Pub. | **0,9653** | 0,5019 | **0,9659** | 0,0416 | 1,0000 | - | - | - | - | - |
| N. Res. | **0,9941** | 0,6376 | **0,9656** | 0,0223 | **0,9543** | 1,0000 | - | - | - | - |
| N. Ind | **0,8909** | **0,8222** | **0,8359** | 0,0386 | 0,7796 | **0,8990** | 1,0000 | - | - | - |
| N. Com. | **0,9857** | 0,5464 | **0,9814** | 0,0295 | **0,9718** | **0,9822** | **0,8397** | 1,0000 | - | - |
| N. Rur | 0,0073 | 0,1801 | -0,0021 | 0,5770 | 0,0391 | 0,0130 | 0,0425 | 0,0192 | 1,0000 | - |
| N. Pub | **0,9379** | 0,5099 | **0,9356** | 0,1069 | **0,9650** | **0,9434** | 0,7915 | **0,9564** | 0,1609 | 1,0000 |

A matriz de correlações é simétrica, então a parte superior foi omitida sem perda da generalidade. As variáveis que tem correlação significativa são as que têm correlação acima de 0,8 e estão destacadas na tabela. Como citado anteriormente, a relação entre as variáveis não é direta e pode ser vista na matriz de correlações. É interessante notar como o consumo residencial e o consumo comercial estão ligados a muitas outras variáveis, enquanto a classe rural não está correlacionada com nenhuma variável. Note-se que o consumo da classe pública está mais relacionado ao número de consumidores comerciais (0,9718) do que o próprio número de consumidores públicos (0,9650). O consumo da classe rural não está relacionado nem com o número de consumidores rurais, fato diferente do que se esperava. Isso ocorreu possivelmente porque os municípios que têm maiores consumos rurais não têm os maiores números de consumidores rurais.

Somente com a análise das correlações entre as variáveis não é possível determinar quais podem ser retiradas da base de dados. É feita então a Análise Fatorial através da extração de componentes principais para avaliação de quais dados mais influenciam a variância da amostra. A extração de componentes principais é feita a partir da matriz de correlação e consiste na determinação dos autovalores da matriz e na verificação de quanto cada um deles contribui para a amostra. A extração de componentes principais da matriz de correlação das variáveis de mercado produziu os resultados mostrados na tabela 04.

Tabela 04 - Análise Fatorial das variáveis de mercado consumidor.

| Fator | Autovalor | % do total | % Acumulada |
|---|---|---|---|
| 1 | 7,0111 | 70,1111 | 70,1111 |
| 2 | 1,6178 | 16,1785 | 86,2896 |
| 3 | 0,7525 | 7,5246 | 93,8142 |
| 4 | 0,4285 | 4,2849 | 98,0991 |
| 5 | 0,0852 | 0,8518 | 98,9510 |
| 6 | 0,0516 | 0,5162 | 99,4671 |
| 7 | 0,0251 | 0,2508 | 99,7179 |
| 8 | 0,0187 | 0,1874 | 99,9053 |
| 9 | 0,0076 | 0,0765 | 99,9818 |
| 10 | 0,0018 | 0,0182 | 100,0000 |

A tabela 04 contém dez fatores, que influenciam na base de dados, sendo que o primeiro deles representa 70,11% de variância dos dados. O número de fatores é igual ao número de variáveis. Entretanto, percebe-se que com apenas quatro fatores (98,0991%) é possível representar a base de dados sem perda significativa na informação.

Resta agora, determinar quais variáveis influenciam em cada um dos dez fatores acima referidos. Isso é feito através das cargas fatoriais, com os resultados mostrados na tabela 05.

Tabela 05 - Cargas fatoriais para variáveis de mercado consumidor.

| | Fator 01 | Fator 02 | Fator 03 | Fator 04 | Fator 05 | Fator 06 | Fator 07 | Fator 08 | Fator 09 | Fator 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| C. Res. | **0,9358** | -0,0237 | -0,3405 | -0,0045 | -0,0370 | 0,0623 | -0,0178 | -0,0125 | 0,0140 | 0,0433 |
| C. Ind. | 0,3519 | 0,0954 | **-0,9275** | 0,0711 | 0,0407 | -0,0019 | 0,0017 | 0,0016 | -0,0005 | -0,0002 |
| C. Com. | **0,9579** | -0,0310 | -0,2315 | 0,0121 | -0,0293 | 0,0610 | 0,1514 | -0,0121 | -0,0056 | -0,0004 |
| C. Rur. | 0,0139 | 0,2990 | -0,0518 | **0,9527** | 0,0004 | -0,0015 | 0,0001 | 0,0003 | -0,0001 | -0,0001 |
| C. Pub. | **0,9759** | 0,0153 | -0,1656 | 0,0156 | 0,0533 | -0,0060 | -0,0242 | 0,1263 | -0,0158 | -0,0049 |
| N. Res. | **0,9308** | -0,0183 | -0,3399 | -0,0028 | -0,0743 | 0,0211 | -0,0583 | -0,0480 | 0,0780 | 0,0014 |
| N. Ind | 0,7099 | -0,0069 | -0,6316 | -0,0019 | -0,3115 | -0,0039 | 0,0036 | -0,0075 | 0,0020 | 0,0004 |
| N. Com. | **0,9686** | -0,0066 | -0,2235 | 0,0068 | -0,0343 | 0,0388 | -0,0197 | -0,0768 | -0,0452 | -0,0289 |
| N. Rur | 0,0082 | **0,9502** | -0,0700 | 0,3034 | 0,0012 | -0,0044 | -0,0001 | 0,0007 | -0,0002 | -0,0001 |
| N. Pub | **0,9555** | 0,1326 | -0,1704 | 0,0471 | -0,0193 | -0,1924 | -0,0288 | 0,0053 | -0,0013 | -0,0004 |

Na tabela 05, nota-se que as variáveis consumo residencial, consumo comercial, consumo público, número de consumidores residenciais, número de consumidores comerciais e número de consumidores públicos influenciam diretamente o Fator 01. Esse fator, como mostrado na tabela 04, é responsável por 70,11% da variância dos dados. As variáveis que influenciam esse fator são variáveis tipicamente urbanas, capazes de definir o perfil de consumo urbano do município. Ainda nessa tabela vemos que somente quatro fatores são influenciados significativamente pelas variáveis disponíveis. A variável número de consumidores industriais, não influencia significativamente nenhum fator. A variável número de consumidores rurais influencia somente o Fator 02, indicando que tal fator deve representar municípios com características rurais.

Agora finalmente é possível escolher quais variáveis formarão a base de dados para a obtenção dos agrupamentos. Uma opção interessante é somar as variáveis de consumo: consumo residencial, comercial e público, criando uma nova variável consumo urbano, e somar as variáveis número de consumidores: o número de consumidores residenciais, comerciais e públicos obtendo uma nova variável número de consumidores urbanos. Essa escolha foi devida ao fato de que as novas variáveis são capazes de representar bem a característica urbana de determinado município. Para uma concessionária de energia é mais fácil fornecer os dados urbanos como um todo do que separá-los em categorias. Assim a base de dados a ser utilizada como variáveis de mercado consumidor será composta por cinco variáveis: consumo urbano, número de consumidores urbanos, consumo rural, número de consumidores rurais e consumo industrial. Selecionadas as novas variáveis, a Análise Fatorial é feita novamente para verificar a distribuição das variáveis na nova base. A tabela 06 mostra os resultados obtidos.

Tabela 06 - Analise Fatorial da nova base de mercado consumidor.

| Fator | Autovalor | % do total | % Acumulada |
|---|---|---|---|
| 1 | 2,5210 | 50,4205 | 50,4205 |
| 2 | 1,3919 | 27,8380 | 78,2584 |
| 3 | 0,6237 | 12,4735 | 90,7319 |
| 4 | 0,4530 | 9,0603 | 99,7922 |
| 5 | 0,0104 | 0,2078 | 100,0000 |

Ainda é possível retirar-se uma variável da base de dados e conseguir 99,79% da representatividade da amostra. Para escolher qual variável será retirada da base de dados, analisam-se as cargas fatoriais, mostradas na tabela 07.

Tabela 07 - Cargas fatoriais para nova base de mercado consumidor.

| Variável | Fator 01 | Fator 02 | Fator 03 | Fator 04 | Fator 05 |
|---|---|---|---|---|---|
| Cons. Urbano | **0,9701** | 0,0062 | -0,0246 | -0,2308 | 0,0700 |
| Cons. Ind. | 0,3960 | -0,0991 | -0,0600 | **-0,9109** | -0,0003 |
| Cons. Rural | 0,0301 | -0,1987 | **-0,9783** | -0,0496 | -0,0000 |
| N. Cons. Urb. | **0,9582** | 0,0046 | -0,0252 | -0,2751 | -0,0742 |
| N. Rural | -0,0100 | **-0,9764** | -0,2001 | -0,0808 | -0,0000 |

Nota-se pela tabela 07, que o Fator 01 é bastante dependente das variáveis consumo urbano e número de consumidores urbanos e representa as características urbanas do município. Uma dessas duas variáveis pode sair da base de dados. Poderia ser retirado o numero de consumidores urbanos, já que este influencia o Fator 01 em menor grau que o

consumo urbano. Entretanto, optou-se por manter o número de consumidores urbanos na base de dados para mostrar que a distribuição do consumo urbano e número de consumidores urbanos no resultado final é muito parecida.

### 3.1.2 Variáveis Técnicas

Na análise das variáveis do sistema elétrico optou-se por incluir indicadores DEC e FEC e não somente as características físicas do sistema. Essa escolha foi feita por não ser conhecido o fato de que esses índices influenciam ou não a determinação de agrupamentos consumidores. Os indicadores DEC e FEC escolhidos são relativos ao ano e estão dados em horas e número de interrupções, respectivamente. A análise que será feita aqui é a mesma feita anteriormente. Espera-se que os índices DEC e FEC possam trazer informações úteis para a obtenção dos agrupamentos de conjuntos consumidores. A análise descritiva é mostrada na tabela 08.

Tabela 08 - Análise descritiva das variáveis técnicas.

| Variável | Média | Desvio Padrão | Quartil 25% | Quartil 50% | Quartil 75% | Amplitude |
|---|---|---|---|---|---|---|
| D. Mun. – SE. | 18,1086 | 15,7326 | 4,2200 | 16.6500 | 26,1800 | 75,0000 |
| N. de SE's | 0,5765 | 0,9279 | 0 | 0 | 1,0000 | 6,0000 |
| N. Alim. | 3,3294 | 5,0169 | 1,000 | 2,0000 | 4,0000 | 41,0000 |
| Cap. Man | - | - | - | - | - | 1,0000 |
| DEC | 39,5529 | 16,3334 | 28,0000 | 39,0000 | 46,0000 | 96,0000 |
| FEC | 26,3176 | 9,2124 | 18,0000 | 26,0000 | 32,0000 | 61,0000 |

Na tabela 08 são dados, distância da sede do município a SE, número de SE's, número de alimentadores, capacidade de manobra e indicadores DEC e FEC. A capacidade de manobra é uma variável binária representando *sim* ou *não*, por isso ela não tem média, desvio padrão e *quartis*. A análise descritiva permite inferir que metade dos municípios estudados não tem SE, isso pode ser visto pelo *quartil 50%* ou mediana. No que se refere ao número de alimentadores, nota-se que a amplitude é grande, indicando a talvez a presença de municípios de grande porte, número bastante elevado se comparado à mediana. Também é alto o número de subestações se comparado à mediana. Possivelmente o alto valor do número de subestações esteja relacionado com o grande número de alimentadores. Os índices DEC e FEC parecem ser bem distribuídos, já que tem valores de média e mediana bastante próximos. Nota-se a presença de índices DEC e FEC de elevada amplitude, o que significa que na base

de dados utilizada há possivelmente a presença de pequenos municípios com sistemas de distribuição não muito eficientes.

Semelhante à análise das variáveis de mercado consumidor, é feita uma avaliação da correlação entre as variáveis técnicas e a tabela de correlações é mostrada.

Tabela 09 - Correlação das variáveis técnicas.

| | Dist. Mun.-SE | N. de SE's | N. de Alim. | Manobra (S/N) | DEC | FEC |
|---|---|---|---|---|---|---|
| D. Mun. - SE | 1,0000 | - | - | - | - | - |
| N. de SE's | -0,3551 | 1,0000 | - | - | - | - |
| N. de Alim. | -0,3358 | 0,6886 | 1,0000 | - | - | - |
| Cap. Man. | -0,1463 | 0,2137 | 0,2715 | 1,0000 | - | - |
| DEC | 0,5783 | -0,2429 | -0,3756 | -0,2284 | 1,0000 | - |
| FEC | 0,5055 | -0,2894 | -0,3766 | -0,2736 | **0,8160** | 1,0000 |

Na tabela 09, nota-se que há apenas duas variáveis fortemente correlacionadas, DEC e FEC, valor destacado na tabela. Há ainda valores negativos, que como já foi citado, significam a tendência de que um valor alto em uma variável produza um valor baixo em outra variável. Veja que, assim, os índices DEC e FEC diminuem se o número de subestações, o número de alimentadores for aumentado.

Com as informações obtidas da tabela de correlação, é de se esperar que não seja possível retirar nenhuma variável da base de dados. Isso é verificado realizando-se a Análise Fatorial, como feito para as variáveis de mercado consumidor. O resultado é apresentado na tabela 10.

Tabela 10 - Análise Fatorial das variáveis técnicas.

| Fator | Autovalor | % do total | % Acumulada |
|---|---|---|---|
| 1 | 2,9706 | 49,5096 | 49,5096 |
| 2 | 1,1518 | 19,1965 | 68,7061 |
| 3 | 0,8800 | 14,6662 | 83,3722 |
| 4 | 0,5314 | 8,8561 | 92,2283 |
| 5 | 0,2972 | 4,9539 | 97,1822 |
| 6 | 0,1691 | 2,8178 | 100,0000 |

Pela tabela de Análise Fatorial das variáveis técnicas, é possível ver que 97,18% da distribuição dos dados é conseguida com cinco fatores. Apesar de uma redução do número de variáveis de sistema de seis para cinco, ser uma redução muito pequena, esta redução é efetuada.

Tabela 11 - Cargas fatoriais para variáveis técnicas

| | Fator 01 | Fator 02 | Fator 03 | Fator 04 | Fator 05 | Fator 06 |
|---|---|---|---|---|---|---|
| Dist. Mun.-SE | 0,3027 | -0,1521 | -0,0465 | **0,9318** | 0,1123 | -0,0461 |
| N. de SE's | -0,1030 | **0,9213** | 0,0900 | -0,1523 | -0,3304 | 0,0104 |
| N. de Alim. | -0,1896 | 0,3740 | 0,1267 | -0,1225 | **-0,8900** | 0,0332 |
| Manobra (S/N) | -0,1238 | 0,0794 | **0,9831** | -0,0421 | -0,0991 | 0,0145 |
| DEC | **0,8202** | -0,0532 | -0,0898 | 0,3002 | 0,1542 | -0,4501 |
| FEC | **0,9462** | -0,1170 | -0,1299 | 0,2001 | 0,1313 | 0,1295 |

A análise das cargas fatoriais mostra que realmente é possível reduzir o número de variáveis, de seis para cinco. O Fator 01 é essencialmente um fator que traz informações sobre a qualidade do sistema de distribuição de energia, já que está fortemente relacionado aos indicadores DEC e FEC, e representa 49,50% da variância dos dados. Aqui não pode ser feito como nas variáveis de mercado consumidor, e somar os índices DEC e FEC, pois estes representam unidades diferentes. Então deve ser escolhido um deles para representar as características de qualidade de energia.

O indicador FEC foi o escolhido para representar a qualidade de energia do município, pois o Fator 01 é mais dependente deste (94,62%) do que do indicador DEC (82,02%). Então a nova base de dados das variáveis técnicas é formada pelas seguintes variáveis: distância do município à subestação, número de subestações, número de alimentadores, capacidade de manobra e o indicador FEC.

## 3.2 Aplicação do Algoritmo SOM para obtenção dos agrupamentos

Após a escolha das variáveis da base de dados, a mesma é aplicada à rede SOM para obtenção dos agrupamentos. A nova base de dados é uma matriz de 170 x 10, em que são analisados 170 municípios com as 10 variáveis que descrevem o sistema de distribuição e as características de consumo de energia de cada um. No banco de dados utilizado, os municípios estão numerados de 1 a 170 e têm cada atributo rotulado com os devidos nomes: Consumo Urbano, Consumo Industrial, Consumo Rural, Número de Consumidores Urbanos, Número de Consumidores Rurais, Distância Município - SE, Número de SE's, Número de Alimentadores, Capacidade de Manobra e o indicador FEC.

Nas simulações feitas neste trabalho foi utilizado o SOM Toolbox para MATLAB©. Esse Toolbox foi implementado pelo Centro de Pesquisas em Redes Neurais da Universidade de Helsinki, e disponibilizado como software livre (VESANTO *et al.*, 1999). O SOM Toolbox é uma biblioteca de códigos que exploram a flexibilidade da série de funções prontas do MATLAB©, especialmente as gráficas, para treinamento e visualização e extração de conhecimento dos mapas (SPERANDIO, 2004a).

Como já explicado, o algoritmo SOM é baseado na distância euclidiana entre os parâmetros de entrada. Sendo assim, a diferença de amplitude entre variáveis pode influenciar a organização do mapa, já que valores com grande amplitude podem dominar o processo de auto-organização do mesmo. No caso do banco de dados utilizado nesse trabalho, a diferença entre a amplitudes das variáveis é muito grande. Por exemplo, note-se a diferença entre a amplitudes da variável consumo urbano, da ordem de centenas, e a amplitude da variável número de subestações, cujo valor máximo é de seis unidades. Para resolver esse problema, a normalização dos dados foi feita, antes dos mesmos serem aplicados à rede SOM.

Os algoritmos utilizados nas simulações encontram-se nos apêndices B e C. No apêndice B é mostrado o algoritmo elaborado para a análise estatística dos dados, onde são feitos as análises descritivas, o cálculo das correlações e a Análise Fatorial. O apêndice C mostra o algoritmo utilizado para a obtenção dos mapas auto-organizáveis. Nele é feita a leitura dos dados que estão em formato ASCII, normalização, treinamento e visualização da rede. Os resultados obtidos são mostrados a seguir.

# 4 RESULTADOS

Como a intenção é de realizar agrupamentos, a opção para o tamanho inicial do mapa deve ser de aproximadamente um neurônio para cada quatro dados de entrada, o que tende a formar grupos concisos. A decisão sobre o tamanho do mapa depende de certa forma da análise descritiva dos dados. Se os dados tendem a se concentrar em um valor central, um mapa de menores dimensões será capaz de representar a totalidade dos dados. O tamanho inicial para a rede foi de 6 x 6, o que formou grupos bastante concisos com até 12 municípios representado por um só neurônio da rede. Foi feito então alteração no tamanho dos mapas e até se encontrar uma distribuição menos concisa no mapa. A tabela 12 mostra os resultados obtidos para alguns tamanhos de mapas.

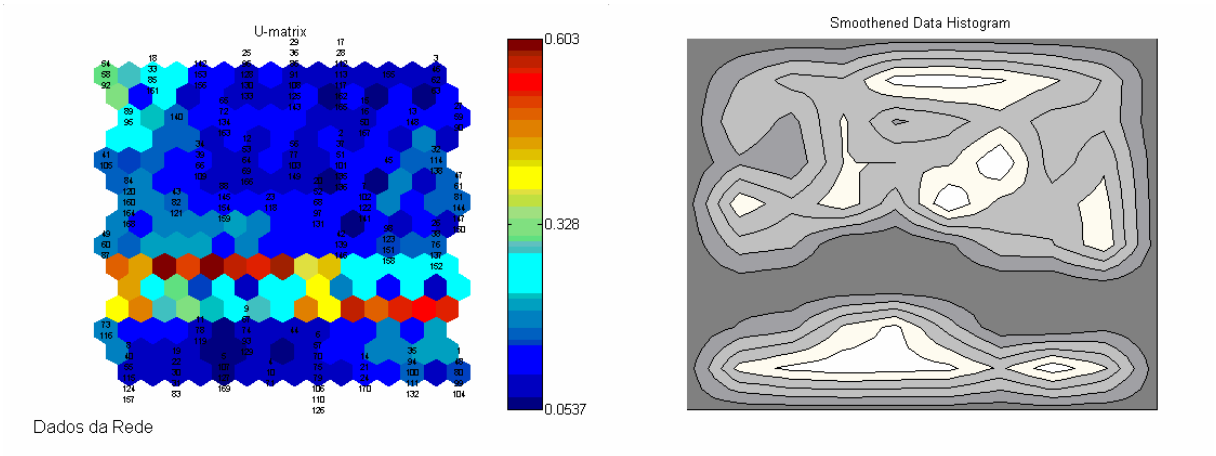
Tabela 12 - Erros de dispersão e topográficos de acordo com o tamanho da rede.

| Tamanho do Mapa | Erro de Dispersão | Erro Topográfico |
|---|---|---|
| 6 x 6 | 0,2206 | 0,0412 |
| 6 x 7 | 0,2081 | 0,0235 |
| 7 x 7 | 0,1944 | 0,0118 |
| 7 x 8 | 0,1897 | 0,0118 |
| 8 x 8 | 0,1880 | 0,0353 |
| 8 x 9 | 0,1733 | 0,0176 |
| 9 x 9 | 0,1682 | 0,0235 |
| 10 x 10 | 0,1572 | 0,0353 |
| 11 x 11 | 0,1481 | 0,0176 |

Para os diversos tamanhos simulados, foram calculados os erros de dispersão ou quantização e o erro topográfico. O primeiro valor é uma medida da resolução do mapa e calcula a distância média entre cada vetor de dados e o neurônio BMU. O segundo mede a

preservação da topologia dos dados e representa a proporção de todos os vetores de dados em que o primeiro e o segundo neurônio BMU não foram neurônios adjacentes.

O mapa selecionado foi o de tamanho 8 x 8 pois mapas maiores que esse começaram a apresentar muitos elementos dispersos, apenas um município por neurônio e longe de outros grupos. A figura 4.1 mostra a matriz de distâncias em forma de *U-mat* e em forma de histograma suavizado de freqüência.

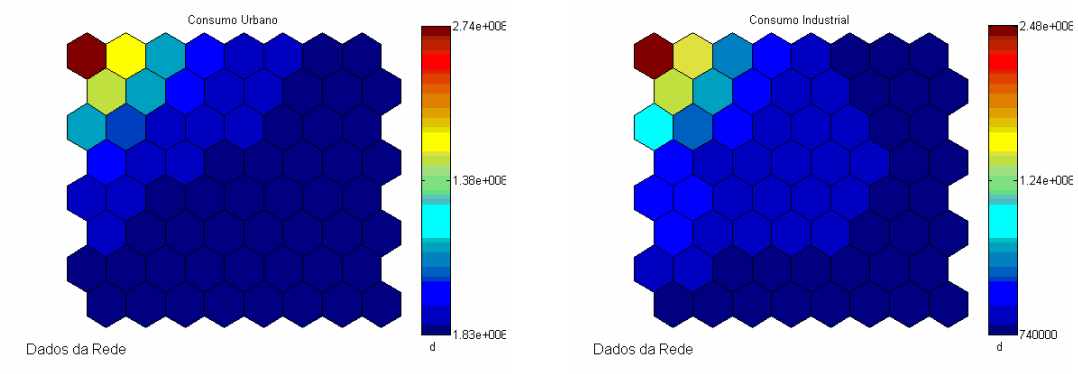



(a) *U-mat* (b) Histograma suavizados de freqüência

Figura 4.1 - Matriz de distâncias.

Pela figura 4.1(b) nota-se a presença de dois grandes grupos, um grupo maior na parte superior e um grupo menor na parte inferior. Na figura 4.1(a) também é possível perceber essa separação dos grupos, região mais escura do mapa. O histograma suavizado de freqüência indica a existência de cinco grupos menores dentro do grupo superior e dois grupos menores dentro do grupo inferior.

A separação desses grupos é explicada pela variação de cada variável conforme pode ser visto pelos mapas componentes, figura 4.2.



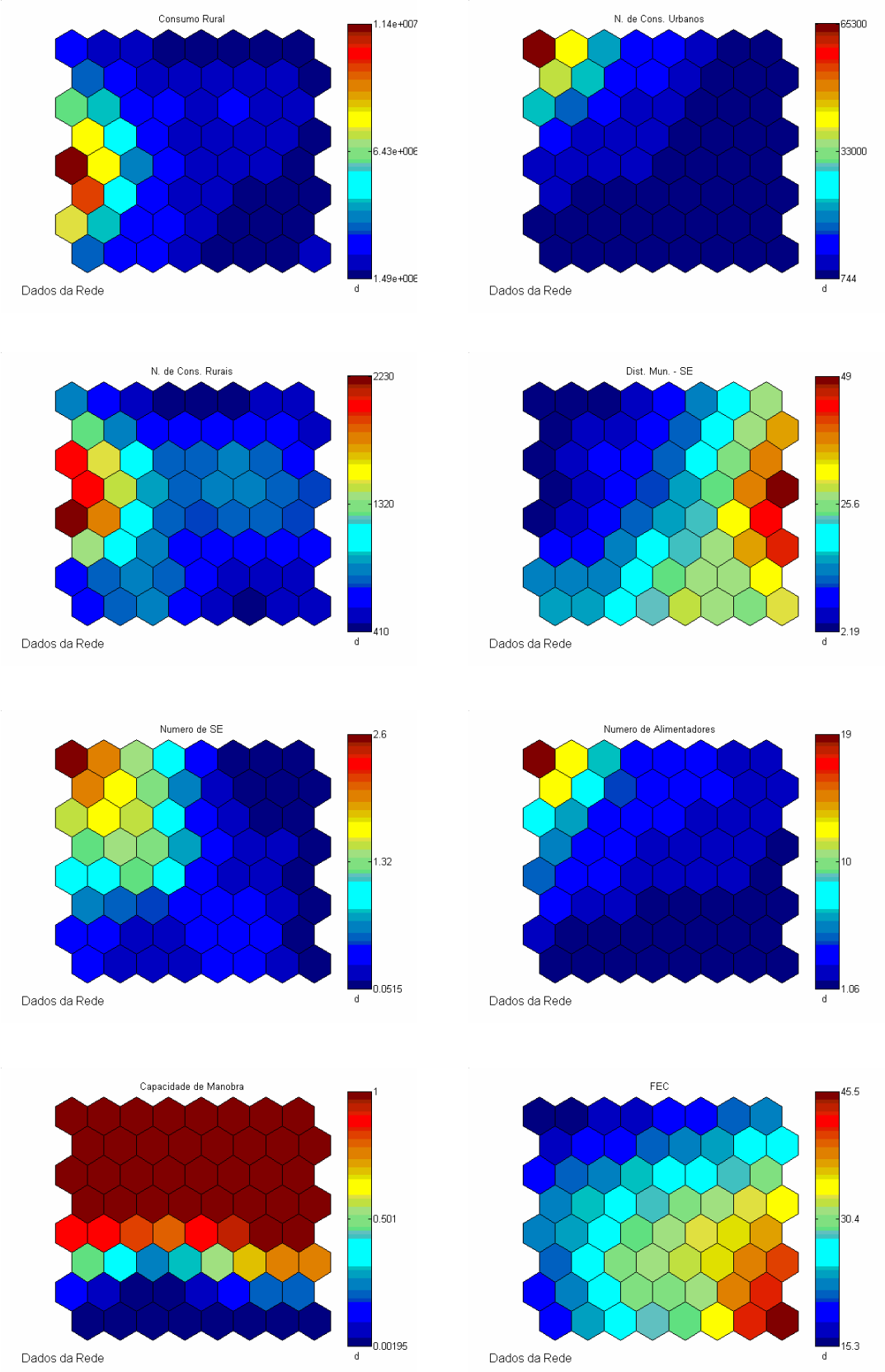



Figura 4.2 - Mapas componentes para as diversas variáveis.

Os mapas componentes mostrados na figura 4.2 permitem observar o resultado final obtido na classificação dos conjuntos consumidores de energia elétrica. Note como o consumo urbano é muito parecido com o consumo industrial. O maior consumo urbano e industrial é pelos municípios rotulados de 54, 58 e 92, que são representados pelo hexágono vinho no mapa. Estes mesmos municípios apresentam um baixo consumo rural. O número de consumidores urbanos, também é muito parecido com o consumo urbano, o que já era de se esperar e já havia sido verificado através do cálculo das correlações no item 3.1.1. Como exposto anteriormente, o número de consumidores urbanos poderia ser retirado da base de dados.

Uma observação interessante deve ser feita com base nos mapas componentes distancia município SE, número de SE's, número de alimentadores, capacidade de manobra e FEC. Os municípios com poucas SE's, que tem as SE's distantes das sedes do município e que não tem capacidade de manobra entre alimentadores, apresentam elevado índice FEC. Note-se o fato de como o índice FEC e distância entre município e as SE's estão fortemente relacionados. Observando-se as figuras 4.1, percebe-se a formação de dois grupos bem distintos. Comparando-se essas figuras com o mapa componente da capacidade de manobra, note-se que neste também há uma separação idêntica em dois grupos, indicando que a variável capacidade de manobra foi a responsável por esta separação na figura 4.1. Ainda na figura 4.1(a), é possível ver no canto superior esquerdo a separação de outro grupo, que é um grupo de características essencialmente urbanas, determinadas pelo consumo urbano e consumo industrial.

Como citado anteriormente, a figura 4.1(b) indica onde há os possíveis centros de agrupamentos. De acordo com os histogramas suavizados de freqüência, existem sete desses centros. A figura 4.3 mostra a *U-mat* com os neurônios que representam pelo menos um município, destacados em azul. Os neurônios são destacados em tamanho proporcional ao número de municípios que representam.

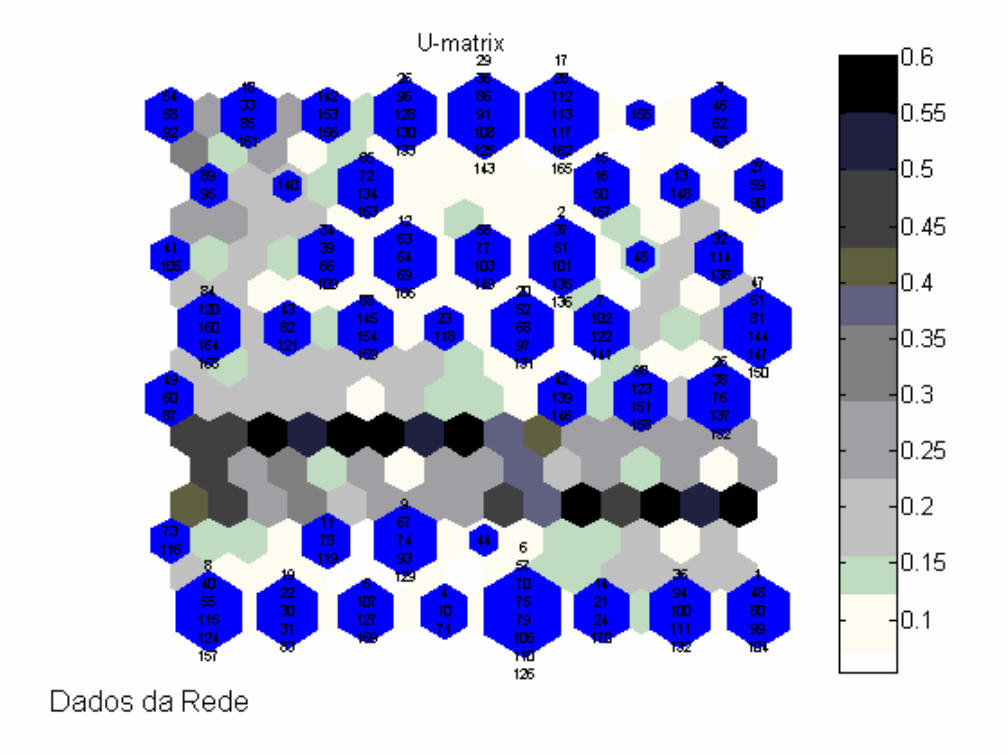



Figura 4.3 - *U-mat* com neurônios destacados.

Na figura 4.3 nota-se claramente onde estão os centros dos agrupamentos. Os hexágonos maiores, coincidem com as regiões mais claras na figura 4.1(b). Na parte superior central da figura 4.3 tem três hexágonos maiores, correspondendo à região branca na parte superior central da figura 4.1(b).

# 5 CONCLUSÕES

O algoritmo SOM é um poderoso método de agrupamento, classificação e extração de conhecimentos de uma base de dados. Ele fornece uma representação estrutural dos dados de entrada através de vetores de peso dos neurônios. Especulações sobre os mapas auto-organizáveis serem um método heurístico são infundadas, pois como se viu no exemplo dado no item 2.3.1.2 ele realmente agrupa e separa os dados.

A visualização dos agrupamentos obtidos via *U-mat* é difícil, sendo possível somente quando as distâncias entre neurônios vizinhos forem bem significativas ou quando o número de neurônios da rede for alto, produzindo assim, muitos elementos isolados. A visualização em histogramas suavizados de freqüência, mostrou-se uma ferramenta capaz de ajudar a determinar onde estão localizados os centros dos agrupamentos. Assim o SOM, por si só, pode não ser tão eficiente na visualização dos agrupamentos, e outros métodos de agrupamento podem ser utilizados em conjunto com o mapa auto-organizável. Portanto os grupos obtidos pelo SOM, podem ser validados ou rejeitados.

A separação entre variáveis de mercado e variáveis técnicas facilita a compreensão do problema. A inclusão dos indicadores DEC e FEC nas variáveis de sistema não trouxeram informação relevante na obtenção dos agrupamentos, visto que a qualidade do sistema de fornecimento é também de certo modo medida pelo número de SE's, e pela capacidade de manobra.

A seleção de varáveis a serem utilizadas também foi importante de modo a reduzir o trabalho computacional nas simulações e também para utilização de um banco de dados menor, visto que nem todas as empresas concessionárias de energia têm esses dados à disposição. Do ponto de vista das concessionárias, é melhor utilizar menos variáveis.

Com a metodologia proposta, destaca-se que as concessionárias de energia podem avaliar os agrupamentos obtidos através do mapa auto-organizável e propor esta metodologia junto ao órgão regulador, no caso do Brasil, a ANEEL. Vale lembrar que após janeiro de 2003 a ANEEL permitiu às concessionárias de energia, propor novos critérios para o agrupamento de consumidores. Fica claro que esses novos métodos devem possuir vantagens técnicas, econômicas e sociais em relação ao critério de agrupamento vigente do órgão regulador. Ao aplicar a metodologia proposta nesse trabalho, as concessionárias de energia podem agrupar conjuntos de consumidores que tenham características do sistema elétrico mais parecidas. Assim, as concessionárias podem rever as metas de qualidade de seus agrupamentos junto ao órgão regulador, requerendo metas menos rigorosas para conjuntos de consumidores com sistemas de distribuição de energia pouco eficientes, os quais eram anteriormente alocados em grupos que tinham metas muito rigorosas.

Como sugestão para trabalhos futuros, estão as escolhas de outras variáveis como parte do banco de dados, nesse caso podem ser variáveis da própria rede elétrica ou variáveis externas, tais como as variáveis climáticas que talvez tenham influência sobre a classificação dos agrupamentos. Uma opção também interessante é de utilização dos mapas auto-organizáveis para previsão de carga, ou ainda a expansão regional de consumo de energia elétrica, onde seria possível determinar grupos com forte tendência de crescimento. Outra opção com base neste trabalho é fazer o mesmo estudo, utilizando outros métodos de agrupamento tal como um método estatístico.

## REFERÊNCIAS BIBLIOGRÁFICAS


ANEEL, 2000, Resolução Normativa Nº. 024.

ANEEL, 2003, Resolução Normativa Nº. 075.

DNAEE, 1978, Portaria Nº. 046.

FLEXER, A., 2001: On the Use of Self-Organizing Maps for Clustering and Visualization, *Intelligent Data Analysis* (5), pp: 373–384.

HAYKIN, S., 2001, Redes Neurais : Princípios e Práticas, Bookman, 2ª ed.

KAGAN, N., de OLIVEIRA, C.C.B., ROBBA, E. J., 2005: Introdução aos sistemas de distribuição de energia elétrica, Edgard Blücher, 1ª ed.

KOHONEN, T., 2000, Self-Organizing Maps, Springer-Verlag, 3ª ed.

MANLY, B. F. J., 1994: Multivariate Statistical Methods – A primer, Chapman e Hall 2ª ed.

MOORE, A. W., 2001: K-means and Hierarchical Clustering, *School of Computer Science Carnegie Mellon University.*

PAMPALK, E., RAUBER, A., MERKL, D., 2002: Using Smoothed Data Histograms for Cluster Visualization in Self-Organizing Maps, *Proc. of the International Conference on Artificial Neural Networks*.

SPERANDIO, M., COELHO, J., QUEIROZ, H.L., *et al*., 2004: Avaliação de Conjuntos Críticos face à Nova Regulamentação da Qualidade de Fornecimento. In: Seminário Nacional de Distribuição de Energia Elétrica (SENDI), 2004, Brasília. Anais do XVI SENDI.

SPERANDIO, M., COELHO, J., QUEIROZ, H.L., 2003: Identificação de Agrupamentos de Consumidores de Energia Elétrica através de Mapas Auto-Organizáveis, Anais do V Seminário Brasileiro sobre Qualidade da Energia Elétrica (SBQEE), vol.2, pp: 439-443, Aracaju, Brasil.

SPERANDIO, M., COELHO, J., QUEIROZ, H.L., *et al*., 2004: Revisão dos Critérios para Agrupamentos de Conjuntos Consumidores de Energia Elétrica, IX Simpósio de Especialistas em Planejamento da Operação e Expansão Elétrica (SEPOPE), SP-34, Rio de Janeiro, Brasil.

SOM TOOLBOX, 2005: http://www.cis.hut.fi/projects/somtoolbox/

TÖRMÄ, M., 1994: Kohonen Self-Organizing Feature Map and its Use in Clustering, *Proc. of the International Society for Optical Engineering (SPIE)*, vol.2357, pp: 830 – 835, Ukraine.

ULTSCH, A., VETTER, C., 1994: “Self-Organizing Feature Maps versus Statistical Clustering Methods: A Benchmark.” Research Report No. 9, Dep. of Mathematics, University of Marburg.

VESANTO, J., 1999: SOM-Based Data Visualization Methods. *In Intelligent Data Analysis*, Vol.3, n.2, Elsevier Science, pp. 111-126.

VESANTO, J., HIMBERG, J., ALHONIEMI, E., PARHANKANGAS, J., 1999: “Self-Organizing Map in Matlab: the SOM Toolbox”, In Proc. of the Matlab DSP Conference, Espoo, Finland, pp: 35-40.

ZANINI, ALEXANDRE, 2004: Regulação econômica no setor brasileiro: uma metodologia para definição de fronteiras de eficiência e cálculo do fator X para empresas distribuidoras de energia elétrica, Rio de Janeiro: PUC, Departamento de Engenharia Elétrica.

## APÊNDICE A – O Método *K-means*

O *k-means* é um método de agrupamento utilizado para classificação não supervisionada. A idéia do método de agrupamento k-means é dividir um conjunto de N dados em K grupos tal que a distância métrica entre os centróides dos grupos seja minimizada. O algoritmo para o agrupamento de N dados em K grupos consiste em minimizar a soma quadrática dada pela equação A.1:

$$J = \sum_{j=1}^{k} \sum_{n \in S_j} \left| x_n - \mu_j \right|^2 \qquad \text{A.1}$$

onde $x_n$ é um vetor de dados representando o n-ésimo vetor e $\mu_j$ é o centróide do agrupamento $j$. O algoritmo *k-means* consiste nos seguintes passos:

1. Determine as posições iniciais dos K centróides no espaço representado pelos objetos a serem agrupados;
2. Aloque cada elemento ao grupo do centróide mais próximo;
3. Recalcule a posição dos K centróides com base nos elementos alocados;
4. Repita o segundo e o terceiro passo até que os centróides não se modifiquem mais ou até atingir o número máximo de iterações. Isso produz a separação dos objetos em grupos.

O algoritmo *k-means* apresenta como vantagens:

- O *k-means* é mais rápido que métodos de agrupamentos hierárquicos para uma base de dados com muitas variáveis ;
- O *k-means* produz grupos mais concisos do que os métodos hierárquicos.

No entanto as desvantagens são que:

- O número de agrupamentos a serem formados deve ser conhecido *a priori* o que pode levar a uma formação não ótima dos agrupamentos;
- A distribuição inicial dos centróides influencia na distribuição final dos agrupamentos.

## APÊNDICE B – Algoritmo para análise estatística

```
%Algoritmo para a analise estatistica.
clear                                                  %Limpando as variaveis.
clc                                                          %Limpando a tela.
format long                                        %Definindo o formato numerico.

tam = 16;                                           %Numero de colunas dos dados.

%Leitura dos dados do arquivo 'dados_rede.txt'.
dados = som_read_data('dados_rede.txt',tam);

d = dados.data(1:170,:);              %Tomando 170 cidades da base de dados.

dmer = d(:,1:10);   %Colunas de 1 a 10 representam as variaveis de mercado.
dsis = d(:,11:16);   %Colunas de 11 a 16 representam as variaveis tecnicas.

%Somando as variaveis consumo para obter nova variavel.
consu = d(:,1) + d(:,3) + d(:,5);
%Somando as variaveis numero de consumidores para obter a nova variavel.
nconsu = d(:,6) + d(:,8) + d(:,10);
%Nova base de dados das variaveis de mercado.
nd = [consu d(:,2) d(:,4) nconsu d(:,9)];

%ENTRE COM A MATRIZ PARA A QUAL DE DESEJA CALCULAR OS PARAMETROS.
sd = dsis;

%Inicio do calculo da analise descritiva
for j = 1:size(sd,2)
    ades(1,j) = mean(sd(:,j));                %Media.
    ades(2,j) = std(sd(:,j));                 %Desvio Padrao.
    ades(3,j) = prctile(sd(:,j),25);          %Quartil 25%.
    ades(4,j) = prctile(sd(:,j),50);          %Quartil 50% ou mediana.
    ades(5,j) = prctile(sd(:,j),75);          %Quartil 75%.
    ades(6,j) = max(sd(:,j));                 %Amplitude.
end
ades1 = ades';                    %Transpondo a matriz com os resultados.
%Fim do calculo da analise descritiva

%Inicio do calculo da analise fatorial
cor = corrcoef(sd);               %Calculando a correlacao.
[ave,av] = eig(cor);              %Calculando os autovalores e autovetores.

%Ordenando o vetor de autovalores
for i = 1:size(av,1)
    av1(i,1) = av(size(av,1)+1-i,size(av,1)+1-i);
end
ave1 = ave';                      %Transpondo o matriz de autovalores.

%Ordenando o autovetor
for i = 1:size(ave1,1)
    ave2(i,:) = ave1(size(ave1,1)+1-i,:);
end

%Calculando a tabela de variancia
for i = 1:size(cor,1)
    c(i,1) = cor(i,i);
end
traco = sum(c);
```

```
for i = 1:size(sd,2)
    table(i,1) = i;
    table(i,2) = av1(i,1);
    table(i,3) = av1(i,1)/traco*100;
end
table(1,4) = table(1,3);
for j = 1:(size(sd,2)-1)
    table(j+1,4) = table(j,4) + table(j+1,3);
end

%Calculando as cargas fatoriais
coef = [av1 ave2];
for i = 1:size(coef,1)
    for j = 1:(size(coef,2)-1)
        cf(i,j) = sqrt(coef(i,1))*coef(i,j + 1);
    end
end
cf1 = cf';                          %Transpondo a matriz de cargas fatoriais.
[MR,MRO] = varimaxTP(cf1);          %Varimax rotation.
%Fim do calculo da Analise Fatorial

ades1                    %Imprimindo na tela a tabela de analise descritiva.
cor                             %Imprimindo na tela a matriz de correlacao.
MRO                       %Imprimindo na tela a matriz de cargas fatoriais.
table                     %Imprimindo na tela a tabela de analise fatorial.
%Fim do algoritmo.
```

## APÊNDICE C – Algoritmo para simulação da rede SOM

```
%Algoritmo utilizado nas simulacoes do trabalho
%
%Nesse algoritmo foi utilizado o SOMTOOLBOX para MATLAB:
%
% SOM Toolbox
%   Version 2.0beta, May 30 2002
%   Copyright 1997-2000 by
%   Esa Alhoniemi, Johan Himberg, Juha Parhankangas and Juha Vesanto
%   http://www.cis.hut.fi/projects/somtoolbox/
%
%Utiliza tambem o SDH TOOLBOX, disponivel em:
%http://www.ofai.at/~elias.pampalk/sdh/index.html

%Inicio do algoritmo.
clear                                                  %Limpando as variaveis.
clc                                                          %Limpando a tela.
close all                                           %Fechando as figuras abertas.
format long                                        %Definindo o formato numerico.

tam = 16;                                           %Numero de colunas dos dados.

%Leitura dos dados do arquivo 'dados_rede.txt'.
dados = som_read_data('dados_rede.txt',tam);
%Tomando 170 cidades do banco de dados.
d = dados.data(1:170,:);
%Somando as variaveis consumo para obter nova variavel.
consu = d(:,1) + d(:,3) + d(:,5);
%Somando as variaveis numero de consumidores para obter a nova variavel.
nconsu = d(:,6) + d(:,8) + d(:,10);
%Nova base de dados utilizada.
nd = [consu d(:,2) d(:,4) nconsu d(:,9) d(:,11) d(:,12) d(:,13) d(:,14)
      d(:,16)];

%nome das componentes.
cnames = {'Consumo Urbano','Consumo Industrial','Consumo Rural',
          'N. de Cons. Urbanos','N. de Cons. Rurais','Dist. Mun. - SE',
          'Numero de SE','Numero de Alimentadores',
          'Capacidade de  Manobra','FEC'};

for i = 1:length(nd)
    clabels{i,1} = num2str(i);          %Criando os rotulos dos municipios.
end

%Criando a estrutura dos dados.
sDs = som_data_struct(nd,'comp_names',cnames,'labels',clabels);
%Normalizando os dados.
sDn = som_normalize(sDs,'range');

nl = 8;                                          %Numero de linhas de neuronios.
nc = 8;                                         %Numero de colunas de neuronios.

%Criando o mapa.
sMap = som_make(sDn,'lininit','batch','msize',[nl nc],'name',
                'Dados da Rede','training','long','comp_names',cnames);

%Rotulando o mapa.
sMap = som_autolabel(sMap,sDn);
```

```
for i = 1:size(nd,2)
    %Abrindo nova figura.
    figure(i)
    %Mostrando o mapa.
    som_show(sMap,'norm','d','comp',[i:i],'edge','on');
    %Alterando a cor do mapa.
    colormap(1-gray);
    %Alterando a cor da legenda.
    som_recolorbar('all','auto','denormalized');
end

figure(size(nd,2) + 1)                              %Abrindo uma nova figura.
som_show(sMap,'umat','all');                               %Plotando a U-mat.
colormap(1-gray);                                   %Alterando a cor do mapa.
som_recolorbar;                                  %Alterando a cor da legenda.
%Rotulando o mapa.
som_show_add('label',sMap,'Textsize',6,'TextColor','k');

figure(size(nd,2) + 2)                              %Abrindo uma nova figura.
som_show(sMap,'umat','all');                               %Plotando a U-mat.
colormap(1-gray);                                   %Alterando a cor do mapa.
som_recolorbar;                                  %Alterando a cor da legenda.
h = som_hits(sMap,sDn,'crisp')                  %Calculando os dados no mapa.
%Mostrando os dados sobre o mapa.
som_show_add('hit',h,'MarkerColor',[0 0 1]);
%Rotulando o mapa.
som_show_add('label',sMap,'Textsize',6,'TextColor','k');

%Calculando os histogramas suavizados.
S = sdh_calculate(sDn,sMap,'spread',3);
%Visualizando em historagramas.
sdh_visualize(S,'contour','levels',5);
%Colocando titulo na figura.
title('Smoothened Data Histogram')
%Fim do algoritmo.
```