

## **TOPOLOGIA DO PERFIL DE PAGAMENTO DOS CONSUMIDORES DE ALTA TENSÃO DE UMA DISTRIBUIDORA DE ENERGIA ELÉTRICA**

**Norma Alice da Silva Carvalho**

Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio  
Rua Marquês de São Vicente, 225, Gávea, Rio de Janeiro, CEP: 22451-041  
ncarvalho@aluno.puc-rio.br; nasc86@gmail.com

**Reinaldo Castro Souza**

Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio  
Rua Marquês de São Vicente, 225, Gávea, Rio de Janeiro, CEP: 22451-041  
reinaldo@ele.puc-rio.br

**Eugenio Kahn Epprecht**

Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio  
Rua Marquês de São Vicente, 225, Gávea, Rio de Janeiro, CEP: 22451-041  
eke@puc-rio.br

### **RESUMO**

As empresas distribuidoras de energia elétrica como agente concessionária ou permissionária de energia sofrem com perdas técnicas e perdas não-técnicas que impactam na remuneração do setor elétrico. Em decorrência disso, a gestão de perdas –em especial, a gestão da inadimplência relacionada a administração da capacidade de caixa das distribuidoras– necessita ser tratada em pesquisas desse setor. O presente estudo propõe uma metodologia topológica do perfil de pagamento dos consumidores de uma distribuidora considerando conhecimento extraído de base de dados e visando formular estratégias para combater a inadimplência. A metodologia proposta consiste em agrupar consumidores por meio de um modelo neural não-supervisionado e, então, identificar as características do grupo através da análise exploratória dos dados. A revisão da literatura e os resultados obtidos revelaram que esse estudo apresenta contribuições para o meio científico e empresarial.

**PALAVRAS CHAVE. Inadimplência, Redes Neurais Artificiais, Análise Exploratória dos Dados.**

**Área principal (PO na área de energia, PO em serviços, Estatística)**

### **ABSTRACT**

The electricity distribution utilities suffer from technical and non-technical losses that impact in the electricity sector remuneration. Therefore, researches about loss management –in particular, the default management related to administration of the cash capability of the distributions – are necessary in this sector. This study proposes a profile's topological methodology of payment of the consumers of an electricity distribution utility considering knowledge extracted from database with the aim of formulating strategies to combat the default. The methodology proposed consists in applying unsupervised neural model for grouping consumers and then identify of the characteristics of the group through exploratory data analysis. The literature review and results obtained show that this study bring contributions to the scientific and business community.

**KEYWORDS. Default, Artificial Neural Networks, Exploratory Data Analysis.**

**Main area (OR in energy, OR in services, Statistics)**

## 1. Introdução

A maioria das organizações tem coletado e armazenado grandes volumes de dados obtidos de suas operações diárias, entretanto, poucas utilizam o conhecimento escondido nesses dados para compreender fenômenos que envolvem sua própria atividade (Côrtes et al., 2002; Ngai et al., 2009).

Em sintonia com essa premissa, estudos recentes realizados no âmbito das distribuidoras de energia elétrica mostraram que o conhecimento armazenado em base de dados as auxiliaram no gerenciamento de seus processos –em particular, na gestão de perdas. (Calili, 2005; Carvalho, 2011; Carvalho et al, 2013; Medeiros, 2013; Ortega, 2008; Souza et al., 2009; Souza et al., 2013; Penin, 2008).

As distribuidoras de energia elétrica são agentes titulares de concessão ou permissão de energia que integram uma das atividades de valor<sup>1</sup> do Setor Elétrico (Carvalho, 2011) e seguem as diretrizes instituídas pelo agente regulador do Setor (Agência Nacional de Energia Elétrica – ANEEL) no exercício de sua função – a conexão e atendimento do consumidor no sistema de abastecimento de energia elétrica.

O equilíbrio econômico-financeiro das distribuidoras é impactado pela perda total de energia na rede de distribuição. Entende-se por perda total de energia, a “diferença entre a energia adquirida pelas distribuidoras e a efetivamente fornecida aos consumidores” (Instituto Acende Brasil, 2007, p. 2) que ocorre devido a componentes (Araújo, 2007; Medeiros, 2013): (i) técnicos–montante de energia elétrica dissipada por ações internas dos sistemas elétricos (cabos, condutores, transformadores, medidores e equipamentos) inerentes ao processo de transporte e distribuição de energia e (ii) não técnicos– montante de energia dissipada devido a inadimplência de consumidores e ao uso irregular da energia elétrica causado, principalmente, pela ação de terceiros (furto de energia e fraude no consumo de eletricidade) ou por equipamento defeituosos (erros de leitura, falha no faturamento).

Fonseca e Reis (2012, p. 93) mencionam que muitas distribuidoras sofrem “com níveis de perdas acima dos limites permitidos e passíveis de serem reconhecidos e admitidos na composição das tarifas”. Conforme Medeiros (2013), no Brasil, a perda total de energia representa cerca de 17% da energia consumida. Ainda conforme esse autor, o prejuízo com perdas não-técnicas é aproximadamente de R\$ 7 bilhões ao ano –isso equivale a 13% da energia consumida. A perda não-técnica ocorre devido a problemas na gestão comercial das distribuidoras de energia. Conforme Instituto Acende Brasil (2007), a perda não-técnica devido à inadimplência das unidades consumidoras chega a aproximadamente R\$ 1 bilhão ao ano. Em decorrência disso, Fonseca e Reis (2012) e Araújo (2007) destacam a necessidade de novos estudos que abordem ações, iniciativas e novas tecnologias para combatê-la.

Este artigo se insere nesse contexto e tem como objetivo usar conhecimento escondido na base de dados de uma distribuidora de energia elétrica para propor uma metodologia topológica do perfil de pagamento dos consumidores de uma distribuidora fazendo uso de método de agrupamento e da análise exploratória de dados visando formular estratégias para combater a inadimplência. O presente trabalho se estrutura em cinco seções: introdução, revisão da literatura, metodologia, resultados e conclusão.

## 2. Revisão da literatura

A gestão da inadimplência é vital para a capacidade de caixa das distribuidoras (Fonseca e Reis, 2012). A inadimplência caracterizada pelo “montante da receita faturada e não recebida pela distribuidora” (Instituto Acende Brasil, 2007, p. 3) é gerada por consumidores de diferentes setores. Medeiros (2013) afirma que o nível de inadimplência varia entre as distribuidoras de energia elétrica, isso pois, considera-se que a inadimplência está associada a fatores setoriais, sociais, econômicos, educacionais e culturais da área de concessão inerente a cada distribuidora. Entretanto, Fonseca e Reis (2012, p. 100) mencionam que “há uma necessidade de estudar mais as

---

<sup>1</sup> As atividades de valor são “atividades física e tecnologicamente distintas, através das quais uma empresa cria um produto valioso para os seus compradores” (Porter, 1998, p. 34).

causas da inadimplência, considerando os diferentes perfis dos devedores da área de concessão”. Tais perfis podem ser obtidos através de uma análise exploratória dos dados realizada após segmentação<sup>2</sup> do universo de consumidores por agrupamento.

## 2.1 Segmentação por agrupamento

A segmentação por agrupamento consiste na realização de uma tarefa de aprendizagem não supervisionada, na qual os elementos de dado universo são agrupados conforme características de similaridade, e envolve quatro etapas (Jain et al., 1999): **1ª etapa-** seleção de dados; **2ª etapa-** definição de padrão de proximidade entre os dados; **3ª etapa-** escolha do método, técnica e algoritmo de agrupamento e **4ª etapa-** interpretação e validação dos resultados.

A segmentação de clientes pode ser realizada por meio de técnicas estatísticas multivariadas (técnicas tradicionais) ou por técnicas *soft computing*. Atualmente, há uma tendência no uso das técnicas *soft computing* em diferentes problemas envolvendo a área de *marketing* e de negócios (Hiziroglu, 2013).

Hiziroglu (2013) realizou uma revisão da literatura sobre a aplicação de técnicas de *soft computing* em segmentação de clientes. Esse estudo bibliométrico foi realizado em base de dados e conduzido considerando somente artigos publicados em revistas científicas cujo conteúdo era predominantemente quantitativo com abordagem de técnicas de *soft computing*, especificamente, em segmentação. Nesses termos, foram obtidos 42 estudos com datas de publicação entre 1986-2012, sendo a maioria dos estudos publicados na revista *Expert Systems with Applications* e poucos, em revistas da área de *marketing*. Por essa pesquisa, observou-se que: (i) há pouca aplicação dessas técnicas em problemas de segmentação relacionados às áreas de negócios e gestão; (ii) a computação neural é a técnica mais utilizada nos estudos; (iii) cerca de 80% dos estudos usam as técnicas *soft computing* para realizar a tarefa de agrupamento. Por esse estudo bibliométrico, concluiu-se que, as técnicas *soft computing* têm crescido como área potencial em pesquisas de segmentação do mercado consumidor.

## 2.2 Modelo Neural não-supervisionado

As redes neurais artificiais são sistemas paralelos distribuídos compostos por unidades de processamento simples (nodos, neurônios) que são interligadas por um grande número de conexões— conhecidas como pesos sinápticos, os quais armazenam o conhecimento representado no modelo e o passa para a rede através de um processo de aprendizagem<sup>3</sup> (Carvalho et al., 1998; Haykin, 2001). Conforme Haykin (2001), a escolha do paradigma de aprendizado depende do tipo de tarefa que a rede neural deve executar. No caso da tarefa de agrupamento, foco deste estudo, aplica-se a aprendizagem não-supervisionada.

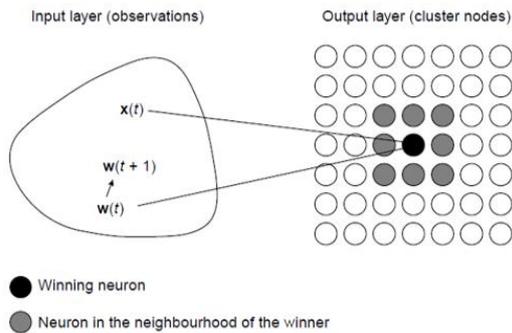
As Redes Neurais *Self-Organizing Maps* (Redes SOM), introduzidas na literatura por Teuvo Kohonen na década de 80, seguem o paradigma de aprendizado não-supervisionado e têm sido amplamente aplicadas em tarefas de agrupamento em diversas áreas, tais como: indústria, finanças, ciências naturais e linguística (Kohonen, 2013; Yang et al., 2012; Vesanto, 1999). A arquitetura da Rede SOM, ilustrada pela Figura 1, é formada por uma camada de entrada com  $x$  observações  $m$ -dimensionais e uma camada de saída representada por uma grade que consiste em  $k$  neurônios associados a  $w$  pesos  $m$ -dimensional (Everitt et al., 2011).

Nessa estrutura, cada vetor de entrada deve selecionar a unidade que melhor o representa, sendo tal unidade (neurônio vencedor) e suas unidades vizinhas (neurônio vizinho) modificadas a cada instante de tempo através de um processo iterativo (treinamento) até que uma melhor adequação dos dados de entrada seja alcançada (Kohonen, 2013).

---

<sup>2</sup> Entende-se, por segmentação, a partição de um grupo heterogêneo em vários grupos homogêneos menores.

<sup>3</sup> Trata-se de um processo iterativo de competição e cooperação entre neurônios e de adaptação dos seus pesos sinápticos através de um algoritmo de aprendizado que treina a rede.



**Figura 1- Rede SOM**

Fonte: Everitt (2011)

O treinamento pode ser realizado de duas maneiras distintas, a saber (Faria et al., 2010; Vesanto et al., 2000):

**Em batelada**, a atualização dos pesos sinápticos é realizada somente após a apresentação de todos os elementos do conjunto de dados utilizados

**Sequencial (incremental)**, a atualização dos pesos sinápticos é realizada toda vez que um exemplo é apresentado a rede.

Os passos básicos para aplicação do algoritmo, interpretação e validação dos resultados obtidos por esse método são descritos a seguir.

**Inicialização dos parâmetros**– A inicialização dos pesos do neurônio pode ser realizada por uma sequência de vetores bidimensionais gerados pelos dois maiores componentes principais de  $x$ , trata-se da inicialização linear (Kohonen, 2013). A taxa inicial de aprendizagem ( $\eta_0$ ) e o raio inicial entre neurônios na vizinhança topológica ( $\sigma_0$ ) também devem ser inicializados. Conforme Haykin (2001),  $\eta_0 \cong 0,1$  e  $\sigma_0 =$  valor igual ao "raio da grade".

**Processo de Competição**– O “espaço contínuo de vetores de entrada [são mapeados] para um espaço discreto de saída de neurônios por um processo de competição entre os neurônios da grade” (Haykin, 2001, p. 488). A ativação de um neurônio é estabelecida por uma relação de proximidade entre seu peso e o vetor de entrada. Na Rede SOM, a função de ativação é baseada na medida de distância Euclidiana, veja Equação 1 (Carvalho et al., 1998; Kohonen, 1990; Kohonen, 2013).

$$i(x) = \arg \min_j \|x - w_j\|, \quad j = 1, 2, \dots, l \quad (1)$$

Na eq. (1), tem-se:

- $i(x)$ : neurônio que melhor casa com o vetor de entrada  $x$ – isto é, neurônio vencedor;
- $x$ : vetor de entrada selecionado aleatoriamente do espaço de entrada com  $m$ -dimensão – representado, matematicamente, por  $x = [x_1, x_2, \dots, x_m]^T$ ;
- $w_j$ : vetor peso sináptico do neurônio  $j$  na grade com mesma dimensão do espaço de entrada– representado, matematicamente, por  $w_j = [w_{j1}, w_{j2}, \dots, w_{jm}]^T$ ;
- $l$ : número total de neurônios na grade.

**Processo de Cooperação**– O “neurônio vencedor determina a localização espacial de uma vizinhança topológica de neurônios excitados [que com ele interagem]” (Haykin, 2001, p. 487). Conforme esse autor, a vizinhança topológica ( $h_{ij}$ , onde:  $i$  representa neurônio vencedor e  $j$ , neurônio excitado) é uma função unimodal da distância lateral entre neurônio vencedor e neurônio excitado ( $d_{ij}$ ), sendo representada, tipicamente, pela função gaussiana, veja Equação 2.

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

Na eq. (2), tem-se:

$\sigma(t)$ : representa a “largura efetiva” da vizinhança topológica no instante  $t$ . Esse parâmetro deve diminuir com o tempo a fim de satisfazer uma das condições da função de vizinhança topológica;

$d_{j,i}^2$ : distância entre neurônio excitado  $j$  e neurônio vencedor  $i$  ao quadrado;

$t$ : instante de tempo.

Haykin (2001) menciona que a “largura efetiva” da vizinhança topológica pode ser descrita por uma função de decaimento exponencial como descrito pela Equação 3.

$$\sigma(t) = \sigma_0 \exp\left(-\frac{t}{\tau_1}\right), \quad t = 0,1,2,\dots \quad (3)$$

Na eq. (3),  $\sigma_0$  é o valor de  $\sigma$  na inicialização do algoritmo SOM e  $\tau_1$  uma constante de tempo.

Ainda em relação ao processo de cooperação, é válido ressaltar que, em uma Rede SOM, a vizinhança pode assumir vários formatos diferentes, tais como hexagonal e quadrado (Vesanto et al., 1999; Vesanto et al., 2000; Faria et al., 2010). Conforme Carvalho et al. (1998, p. 114), “a definição do formato mais adequado depende do problema atacado e da distribuição de seus dados. Geralmente, o formato da vizinhança é definido por tentativa e erro”.

**Processo de Adaptação**— Os “neurônios excitados [aumentam] seus valores individuais [da função de ativação] em relação ao padrão de entrada através de ajustes adequados aplicados a seus pesos” (Haykin, 2001, p. 487-488). O ajuste dos pesos é definido pela Equação 4 e consiste na atualização dos pesos de “todos os neurônios da grade que se encontram dentro da vizinhança topológica do neurônio vencedor  $i$ ” (Haykin, 2001, p. 492).

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \eta(t)h_{j,i(x)}(t)[\mathbf{x}(t) - \mathbf{w}_i(t)] \quad (4)$$

Na eq. (4),  $\eta(t)$  corresponde a taxa de aprendizagem do algoritmo.

A taxa de aprendizagem do algoritmo deve começar em um valor inicial  $\eta_0$  e, então, decrescer gradualmente com o aumento do tempo. Em geral, opta-se pela função de decaimento exponencial dada pela Equação 5.

$$\eta(t) = \eta_0 \exp\left(-\frac{t}{\tau_2}\right), \quad t = 0,1,2,\dots \quad (5)$$

Na eq. (5),  $\tau_2$  refere-se a uma constante de tempo do algoritmo de aprendizagem do SOM. Conforme Haykin (2001, p. 492), a eq. (4) “tem o efeito de mover o vetor peso sináptico  $\mathbf{w}_i$  do neurônio vencedor  $i$  em direção ao vetor de entrada  $\mathbf{x}$ ”.

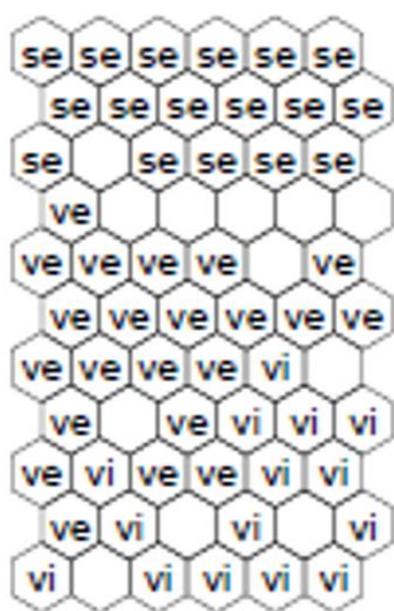
O processo adaptativo é realizado em duas etapas, a saber (Kohonen, 1997; Haykin, 2001; Carvalho et al., 1998): (i) **Fase de auto-organização ou de ordenação**, consiste na ordenação topológica dos vetores de peso. Essa etapa pode exigir 1000, ou mais, iterações do algoritmo SOM visando agrupar os padrões de entrada nos neurônios do mapa topológico; (ii) **Fase de convergência**: consiste no aprimoramento do agrupamento realizado. Haykin (2001) menciona que, como regra geral, o número de iterações nessa fase deve ser no mínimo 500 vezes o número de neurônios na rede. A taxa de aprendizagem nessa fase de ter um valor pequeno igual ou acima de 0,01. O raio da vizinhança do neurônio vencedor deve se reduzir a 1 ou zero neurônios vizinhos.

**Interpretação gráfica**—a visualização da Rede SOM pode ser realizada através de diferentes tipos de métodos, os quais podem ser classificados em três categorias (Vesanto, 1999):

**1ª categoria**— métodos que fornecem uma ideia global e possível da estrutura de grupo. Nesse caso, tem-se: a projeção realizada pelo próprio SOM (Figura 2, fornece a quantidade e a identificação do vetor de entrada por neurônio) e o hits (Figura 3e, quantifica o número de vetor de entrada por unidade do mapa através da proporção de preenchimento desta unidade.).

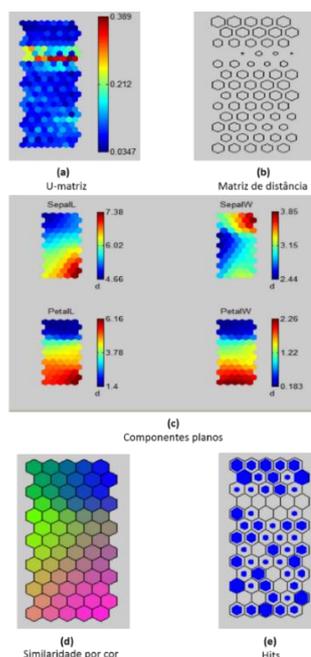
**2ª categoria**— métodos de análise das características dos grupos. Nesse caso, tem-se: a matriz de distância (Figura 3b, trata-se de um método de visualização das distâncias médias entre unidades vizinhas do espaço de saída. Nesse método, o tamanho de cada unidade de mapeamento é proporcional à distância média entre suas vizinhas) e os componentes planos (Figura 3c, refere-se a uma maneira de verificar a existência de correlação entre os  $m$ -dimensionais componentes do vetor de entrada  $x$ . A correlação entre pares de componentes são revelados como padrões semelhantes em posições idênticas dos planos de componentes).

**3ª categoria**— métodos de detecção de grupos. Nesse caso, tem-se: a U-matriz (Figura 3a, trata-se de uma matriz de distâncias dos pesos entre neurônios vizinhos do espaço de saída. Esse método permite a detecção visual das relações topológicas do mapa revelando possíveis grupos devido a colorações distintas determinadas pelo valor da distância entre os neurônios vizinhos do mapa) e a similaridade por cor (Figura 3d, consiste em atribuir cor as unidades do mapa a fim de verificar semelhanças entre elas. Esse método, baseia-se na distância entre vetor de entrada e cada vetor de protótipo. Assim, unidades com tons semelhantes são próximas no espaço de entrada).



**Figura 2 - Representação de uma Rede SOM**

Fonte: Vesanto et al. (2000)



**Figura 3- Métodos de visualização do SOM**

Fonte: Elaboração própria

**Validação**— A literatura aborda duas métricas para avaliação dos resultados do algoritmo SOM, a saber (Faria et al., 2010; Vesanto et al., 2000; Sassi, 2006):

**Erro de quantização** consiste em verificar o quão bem ajustado estão os vetores de entrada na rede. O erro de quantização é estimado pelas médias das distâncias entre cada

vetor de dados  $\mathbf{x}$  e o correspondente vetor de pesos do neurônio vencedor  $\mathbf{w}_i$ . Quanto menor o erro de quantização, melhor o ajuste.

**Erro topográfico** consiste em verificar a capacidade do mapa em representar a topologia dos dados de entrada. O erro topográfico é calculado considerando para cada vetor de entrada  $\mathbf{x}$  se seus primeiro e segundo neurônios vencedores são sempre vizinhos ou não, de modo que para cada vetor de entrada  $\mathbf{x}$  seja atribuído valor 1 se não forem adjacentes e 0, caso contrário. Tal métrica é descrita em Kiviluoto (1996, apud Vesanto et al., 2000).

A partir dos resultados obtidos pelas métricas verifica-se a ocorrência de sobreajuste e subajuste das Redes SOM. Conforme Sassi (2006), em geral: (i) EQ e ET diminuem com o aumento do número de neurônios na rede; (ii) o sobreajuste ocorre quando ambos valores de EQ e ET são muito baixos; (iii) o subajuste, por sua vez, ocorre quando valor de EQ é muito alto e de ET é muito baixo, o que pode ser justificado pela presença de poucos neurônios para representar um grande número de dados.

### 2.3 Análise exploratória de dados

Entende-se por análise exploratória de dados, o estudo descritivo de um conjunto de dados por meio de medidas numéricas ou gráficas cuja função é sintetizar informações contidas nos dados (Behrens, 1997; Seltman, 2013). Tal estudo pode seguir abordagem univariada ou multivariada, dependendo da quantidade de variáveis em questão. Em geral, utilizam-se as seguintes medidas para descrever um conjunto de dados:

**Medidas de tendência central**—refere-se a medidas que produzem um valor médio representativo de um conjunto de dados (Spiegel e Stephens, 2008), também denominadas por medidas de localização. A média aritmética, mediana e moda são as medidas mais comumente usadas para esse fim.

**Medidas de dispersão**— refere-se a medidas que representam o grau de variabilidade entre dados (Spiegel e Stephens, 2008). Nesse caso, as mais utilizadas são: amplitude, desvio padrão e variância.

**Histograma de frequência**— trata-se de um diagrama de barras no qual cada barra representa a frequência ou proporção de casos para um intervalo de valores (Seltman, 2013). A escolha dos intervalos é definida pelo pesquisador arbitrariamente de modo que as classes tenham a mesma amplitude e a frequência de observação dos dados sejam superiores a cinco.

## 3. Metodologia

O método proposto para a obtenção da topologia do perfil de pagamento das unidades consumidoras de energia elétrica consiste em utilizar Rede Neurais *Self-Organizing Maps* (SOM) —método de agrupamento adotado— sobre a matriz com os históricos de atrasos das unidades consumidoras obtidas através da base de dados da empresa em estudo e, então, identificar e analisar o perfil do grupo e das unidades consumidoras que o compõe através da análise exploratória de dados realizada sobre a saída do método de agrupamento adotado.

A base de dados que será utilizada nesse estudo foi extraída de uma empresa de distribuição de energia elétrica brasileira por Carvalho (2011), o qual a denominou *Energy Service*. As variáveis disponibilizadas pela empresa foram: nome do cliente, atividade econômica, município, demanda lida/contratada/ultrapassada, consumo, valor da fatura de energia, data de vencimento e de arrecadação do valor faturado, dados mensais de faturamento (jan/2005 a abr/2011) e dados mensais de arrecadação (jan/2009 a abr/2011). Embora a inadimplência esteja associada a  $n$  fatores (conforme descrito na seção anterior), o agrupamento das unidades consumidoras será realizado considerando somente os atrasos de pagamento do período jan/2009

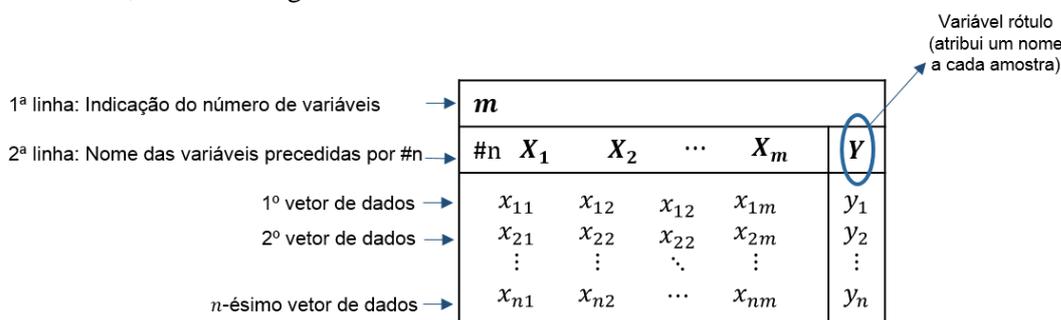
a dez/2010. O *software* MatLab, especificamente o Toolbox SOM<sup>4</sup>, foi utilizado para realizar a tarefa de agrupamento dos dados selecionados.

A aplicação do método se limitará a categoria de consumidores cativos<sup>5</sup>, com mínimo de 35 meses de histórico de dados, atendidos na alta tensão da empresa distribuidora de energia selecionada. Nessas condições, o quadro de consumidores da empresa se resume em 4.265 consumidores. Busca-se dessa maneira manter o universo utilizado na pesquisa realizada por Carvalho (2011).

Desta forma, conforme método proposto e taxionomia de Vergara (2013), quanto aos fins essa pesquisa é classificada como: descritiva (pois, expõe características de determinada população ou fenômeno), metodológica (pois, utiliza instrumentos de captação ou manipulação da realidade) e aplicada (pois, apresenta uma finalidade prática motivada pela necessidade de resolver problemas concretos). E, quanto aos meios de investigação, classifica-se em: bibliográfica (pois, faz uso de material publicado em livros, redes eletrônicas, jornais e revistas científicas) e documental (utiliza registros provenientes da base de dados de determinada empresa distribuidora de energia elétrica).

#### 4. Resultados

Os históricos de atrasos de pagamento das unidades consumidoras foram dispostos em uma matriz, conforme Figura 4.



1ª linha: Indicação do número de variáveis	$m$					
2ª linha: Nome das variáveis precedidas por #n	#n	$X_1$	$X_2$	...	$X_m$	$Y$
1º vetor de dados		$x_{11}$	$x_{12}$	$x_{12}$	$x_{1m}$	$y_1$
2º vetor de dados		$x_{21}$	$x_{22}$	$x_{22}$	$x_{2m}$	$y_2$
		$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n-ésimo vetor de dados		$x_{n1}$	$x_{n2}$	...	$x_{nm}$	$y_n$

**Figura 4- Matriz histórico de atrasos de pagamento**

Fonte: Elaboração própria

Os dados dessa matriz de atrasos foram normalizados a fim de permitir que os valores dos atributos tenham igual influência na obtenção do resultado. A normalização empregada foi a *range*<sup>6</sup>. O tamanho do mapa (número de neurônios da grade) foi determinado através de um processo de “tentativa e erro” na busca por uma boa configuração dos perfis das unidades consumidoras. A inicialização dos pesos neurais seguiu a abordagem linear e o treinamento ocorreu de modo sequencial. O número de épocas de treinamento foi definido considerando o tamanho do mapa adotado e conforme especificações descritas na seção 2.2 (1000 épocas de treinamento na fase ordenação e 3000 épocas de treinamento na fase convergência). Além disso, adotou-se: topologia de vizinhança hexagonal,  $\eta_0 = 0,5$  e  $\sigma_0 = 1$ .

O processo iterativo de aprendizagem da Rede SOM ocorreu antes do número de ciclos determinado. A fase de ordenação e convergência foram concluídas após cerca de, respectivamente, 500 e 1400 ciclos.

A Rede SOM gerada conforme configurações supracitadas é ilustrada pela Figura 5. Nessa figura, consta o nome de um dos clientes que integram cada neurônio. É importante ressaltar

<sup>4</sup> Trata-se de uma biblioteca de rotinas para implementação das Redes SOM no MatLab. Foi desenvolvido pelo Laboratório de Computação e Ciências da Informação da Universidade de Tecnologia de Helsinki – Finlândia.

<sup>5</sup> Conforme Carvalho (2011), clientes cativos são aqueles que somente podem comprar energia elétrica da empresa distribuidora que detém concessão do serviço para localidade em que ele esteja instalado.

<sup>6</sup> Transformação linear simples definida pela razão  $\frac{x - \min_A}{\max_A - \min_A}$ , onde:  $x$  representa o valor de determinado atributo A que desejamos transformar,  $\min$  e  $\max$  são, respectivamente, o valor mínimo e máximo desse atributo.

que, o rótulo atribuído a cada neurônio não expressa significados maiores é, apenas, o nome atribuído ao cliente da primeira amostra de dados, contida na base apresentada a rede, identificada para aquele neurônio.

Pela Figura 6, visualiza-se o quão denso de clientes é um neurônio. Isso permite que o usuário reflita se a quantidade de neurônio especificado para a geração do SOM é adequada ao problema que deseja modelar.

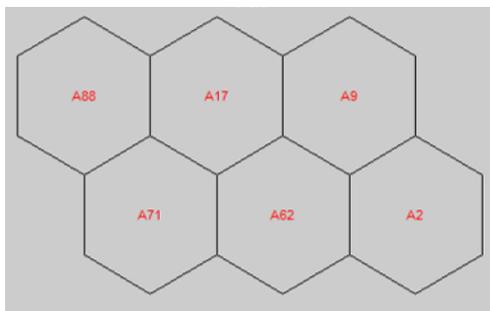


Figura 5 - Rede SOM

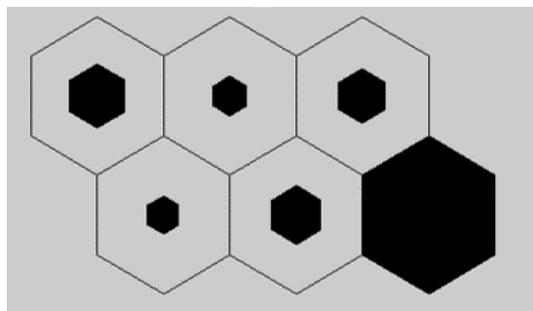


Figura 6- Número de clientes por neurônio: hits

Os dados numéricos relativos a Figura 6 são apresentados no Quadro 1 e serão discutidos posteriormente. A princípio, o que se pode dizer, considerando o estudo anterior realizado por Carvalho (2011), é que há grande chance do modelo neural estar representando a realidade, visto que no referido estudo uma amostra dos dados utilizados na pesquisa atual (os 250 maiores clientes da empresa) foi analisada e, por ela, verificou-se que grande parte dos clientes (cerca de 170) eram adimplentes. Assim, estima-se que isso se replique aqui (neurônio 6).

Quadro 1- Neurônio x quantidade de clientes

Neurônio	Quantidade de clientes
1	476
2	163
3	182
4	391
5	349
6	2704

A Figura 7 ilustra a distância entre os vetores de peso de toda a vizinhança do neurônio e o seu próprio vetor de peso. Note que, para cada tonalidade de cor uma escala de valor é definida de modo que grupos sejam delimitados por valores maiores da escala. Nessa Figura, pode-se a priori detectar a presença de cinco grupos.

A Figura 8 apresenta a distância média entre as unidades e seus vizinhos do espaço de saída sendo proporcional ao tamanho do neurônio.

Pela Figura 8, verifica-se que dos grupos definidos a priori na Figura 7, o grupo 1 talvez seja aquele em que os clientes apresentam em média comportamento mais semelhante.

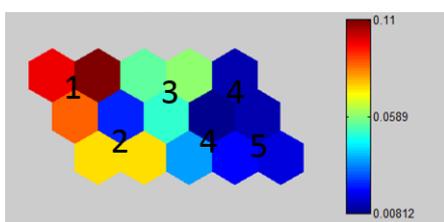


Figura 7- U-matriz do SOM gerado

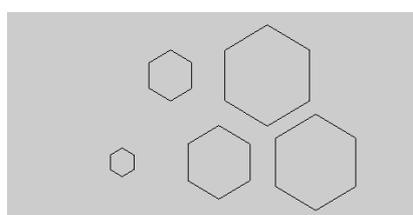


Figura 8- Matriz distância do SOM gerado

Embora, sabe-se que as variáveis em uso (histórico de atrasos) são auto correlacionadas, os componentes planos, usados para verificar correlação entre os pares de componentes, foram visualizados (veja Figura 9). Pela Figura 9, verifica-se que o comportamento de atrasos tende a se manter semelhante no decorrer do histórico dos clientes, o que era de se esperar.

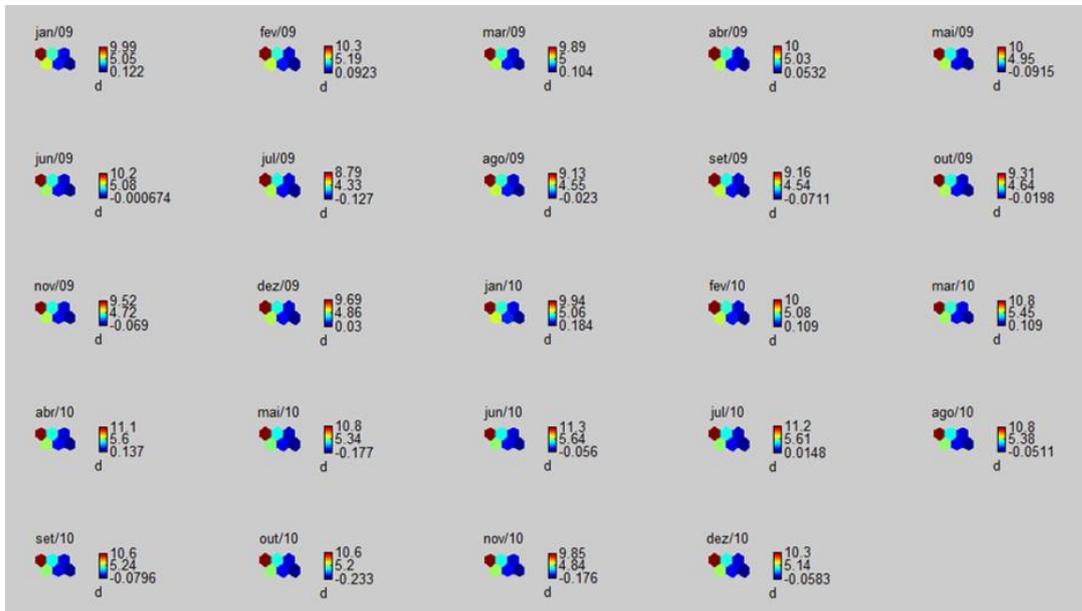


Figura 9- Componentes planos do SOM gerado

A Figura 10 nos informa que unidades com tons semelhantes são próximas no espaço de entrada (nesse estudo, isso significa que, apresentam comportamento de atrasos semelhantes). Através dela, os grupos identificados a priori ao analisar Figura 7 são confirmados.

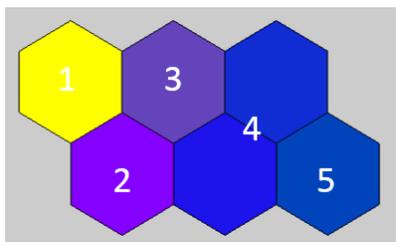


Figura 10- Similaridade por cor do SOM gerado

Nesse estudo, após interpretação gráfica do SOM gerado, uma análise exploratória dos grupos formados foi realizada para fins de identificação dos perfis de pagamento das unidades consumidoras. A Tabela 1 resume as análises realizadas.

Tabela 1- Características dos grupos formados

Grupos	Quantidade de clientes	Estatística descritiva dos atrasos por grupo			Moda da Frequência do comportamento de pagamentos		
		Média	Mediana	Desvio padrão	Antecipa	Na data	Atrasa
Neurônio 1	476	16,61	15,00	20,93	0	0	24
Neurônio 2	163	6,80	4,00	11,83	0	6	15
Neurônio 3	182	3,72	0,00	13,33	0	15	11
Neurônio 4 e 5	740	1,18	0,00	5,54	0	19	5
Neurônio 6	2704	-0,40	0,00	2,76	0	24	0

A Rede Neural SOM foi avaliada pelas métricas: erro de quantização e topológico. Os resultados das métricas foram, respectivamente, 0,060 e 0. Levando em consideração, o tamanho do conjunto de dados, as Figuras 6 e 9, o Quadro 1 e a Tabela 1, conclui-se que o ajuste do modelo aos dados foi satisfatório.

## 5. Conclusão

A busca por conhecimento em base de dados (em inglês, *Knowledge Discovery in Database*– KDD) tem sido cada vez mais solicitada dentro do ambiente organizacional devido necessidade das empresas em compreender os processos que a compõem.

A gestão da inadimplência nas distribuidoras de energia elétrica é uma atividade vital para todo Setor Elétrico pois, é a tarifa de fornecimento de energia recolhida mensalmente por elas, a fonte que o remunera. A implementação de metodologias para o combate a inadimplência nessas empresas tem sido, essencialmente, focadas para recuperação do faturamento perdido. Nessa área, poucos estudos têm explorado o desenvolvimento de uma metodologia preventiva.

Este estudo contribuiu com uma metodologia preventiva para gestão de inadimplência de empresas distribuidoras de energia. As tecnologias levantadas para o desenvolvimento metodológico dessa pesquisa tem mostrado ser campo potencial para muitos estudos aplicado à área gerencial. Os resultados obtidos com a implementação da metodologia proposta mostraram-se potenciais em explicar o comportamento de pagamento das unidades consumidoras. Pretende-se em um estudo futuro, usar os grupos gerados na formulação de regras para criar uma metodologia classificatória da capacidade de pagamento.

## Referências

- Araújo, A. C. M. de**, *Perdas e inadimplência na atividade de distribuição de energia elétrica no Brasil*, Tese de Doutorado, COPPE/UFRJ, Rio de Janeiro, 2007.
- Behrens, J. T.** (1997), Principles and procedures of exploratory data analysis, *Psychological Methods*, 2, n. 2, p. 131-160.
- Carvalho, A. C. P. L. F., Braga, A. P., Ludermir, T. B.**, *Fundamentos de redes neurais artificiais*, DCC/IM, COPPE/Sistemas, NCE/UFRJ, Rio de Janeiro, 1998.
- Carvalho, N. A. S., Souza, R. C., Frota, M. N.** (2013), Statistical analysis of the collectability index: a case study of an electricity distribution utility, *Anais do International Symposium on Operation Research and Applications – ISORAP2013*, p. 565-570.
- Carvalho, N. A. S.**, *Aplicação de modelos estatísticos para previsão e monitoramento da cobrabilidade de uma empresa distribuidora de energia elétrica no Brasil*, Dissertação de Mestrado, PUC, Rio de Janeiro, 2011.
- Calili, R. F.**, *Desenvolvimento de sistema para detecção de perdas comerciais em redes de distribuição de energia elétrica*, Dissertação de Mestrado, PUC, Rio de Janeiro, 2005.
- Côrtes, S. C., Porcaro, R. M., Lifschitz**, *Mineração de dados: funcionalidades, técnicas e abordagens*, PUC, Rio de Janeiro, 2002. Disponível em: [www.dbd.puc-rio.br/depto\\_informatica/02\\_10\\_cortes.pdf](http://www.dbd.puc-rio.br/depto_informatica/02_10_cortes.pdf). Acesso em: abr.2014.
- Everitt, B., Landau, S., Leese, M., Stahl, D.**, *Cluster analysis*, 5. ed., Wiley, Wiley Series in Probability and Statistics, [S.l.], 2011.
- Faria, E., Albuquerque, M. P., Alfonso, J. L. G., Albuquerque, M. P., Cavalcante, J. T. P.**, *Introdução ao toolbox de redes neurais de Kohonen*, Centro Brasileiro de Pesquisas Físicas, Rio de Janeiro, 2010.
- Fonseca, J. N. e Reis, L. B.**, *Empresas de distribuição de energia elétrica no Brasil: temas relevantes para gestão*, Synergia, Rio de Janeiro, 2012.
- Haykin, S.**, *Redes neurais*, tradução de Paulo Martins Engel, 2. Ed., Bookman, Porto Alegre, 2001.
- Hiziroglu, A.** (2013), Soft computing applications in customer segmentation: state-of-art review and critique, *Expert Systems with Applications*, 40, n. 16, p. 6491-6507.
- Instituto Acende Brasil**, *Cadernos de Política Tarifária: Análise do processo de revisão tarifária e da regulação por incentivos*, nº 5, dezembro 2007. Disponível em: [www.acendebrasil.com.br/site/secoes/Estudos.asp](http://www.acendebrasil.com.br/site/secoes/Estudos.asp). Acesso em: abr.2014.

- Jain, A. K., Murty, M. N., Flynn, P. J.** (1999), Data clustering: a review. *ACM Computing Surveys*, 31, n. 3, p. 264-323.
- Kiviluoto, K.** (1996), Topology preservation in self-organizing maps. *Proceedings of International Conference on Neural Networks*, p. 294-299.
- Kohonen, T.** (1990), The self-organizing map. *Proceedings of the IEEE*, 78, n. 9, p. 1464-1480.
- Kohonen, T.**, Self organizing maps, 2. ed., Springer, Springer series in Information Sciences, v. 30, Heidelberg, 1997.
- Kohonen, T.** (2013), Essentials of the self-organizing map. *Neural Networks*, 37, p. 52-65.
- Medeiros, A. L.**, *Alocação de equipes de campo para avaliação de perdas não-técnicas de energia elétrica: desenvolvimento de um Sistema de apoio à decisão*, Tese de Doutorado, UFLA, Lavras, 2013.
- Ngai, E. W. T., Xiu, L., Chau, D. C. K.** (2009), Application of data mining techniques in customer relationship management: a literature review and classification, *Expert Systems with applications*, 36, 2592-2602.
- Ortega, G. V. C.**, *Redes neurais na identificação de perdas comerciais do setor elétrico*, Dissertação de Mestrado, PUC, Rio de Janeiro, 2008.
- Penin, C. A. S.**, *Combate, prevenção e otimização das perdas comerciais de energia elétrica*, Tese de Doutorado, USP, São Paulo, 2008.
- Porter, M. E.**, *Vantagem competitiva: criando e sustentando um desempenho superior*, 11. ed., Campos, Rio de Janeiro, 1998.
- Sassi, R. J.**, *Uma arquitetura híbrida para descoberta de conhecimento em base de dados: teoria dos rough sets e redes neurais artificiais mapas auto-organizáveis*, Tese de Doutorado, Escola Politécnica da Universidade de São Paulo, São Paulo, 2006.
- Seltman, H. J.**, *Experimental design and analysis*, 2013. Disponível em: <http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf>. Acesso em: abr.2014.
- Souza, R. C., Musafir, J. R., Pessanha, J. F. M., Oliveira, F. L. C., Dantas, B. F., Domingues, R.** (2009), Índice de capacidade de pagamento dos consumidores residenciais BT: um método baseado em lógica fuzzy, *Anais do 42º Simpósio Brasileiro de Pesquisa Operacional*.
- Souza, R. C., Pessanha, J. F. M., Oliveira, F. L. C.** (2013), A residential consumer payment capability index based on fuzzy logic inference, *Journal of Intelligent & Fuzzy Systems*, 25, n. 3, 649-657.
- Spiegel, M. R., Stephens, L. J.**, *Schaum's outline of theory and problems of statistics*, 4. ed., McGraw-Hill, Schaum's outlines series, [S. 1.], 2008.
- Vergara, S. C.**, *Projetos e relatórios de pesquisa em administração*, 14. ed., Atlas, São Paulo, 2013.
- Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J.**, *SOM toolbox for Matlab 5*, Laboratory of Computer and Information Science – Helsinki University of Technology, Helsinki, 2000.
- Vesanto, J.** (1999), SOM-based data visualization methods, *Intelligent Data Analysis*, p. 111-126.
- Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J.** (1999), Self-organizing map in Matlab: the SOM toolbox, *Proceedings of the Matlab DSP Conference*, p. 35-40.
- Yang, M.-S., Hung, W.-L., Chen, D.-H.** (2012), Self-organizing map for symbolic data, *Fuzzy set and systems*, n. 203, p. 49-73.