# 4 DESCRIÇÃO DO MODELO DE DETECÇÃO DE PERDAS

## 4.1 INTRODUÇÃO

São apresentados nesta seção os dados utilizados e selecionados, tanto no âmbito do cadastro, disponibilizado pela concessionária de energia ELEKTRO, quanto no âmbito da pesquisa nesta empresa, os quais foram obtidos a partir de uma Pesquisa de Posses e Hábitos - PPH [29 - 33]. Destes dados disponibilizados apresentam-se também as variáveis que foram utilizadas, bem como o tratamento dado às mesmas.

Em seguida é apresentada a metodologia utilizada, a qual fez-se previamente um "tratamento" aos dados de entrada (dados de cadastro) e aos dados da PPH. É feita inicialmente uma "clusterização" utilizando rede de Kohonen para o banco de dados de cadastro disponibilizado pela distribuidora de energia. Tomando os grupos desta classificação prévia feita pela rede identificam-se quais e quantos destes tiveram PPH's realizadas. Para se ter a classificação de um grupo quanto a incidência de consumidores normais, inadimplentes e fraudulentos, o que é objetivo principal desta dissertação de mestrado, utiliza-se um processo de análise fuzzy, o qual identifica o percentual de consumidores em cada segmento.

## 4.2 DADOS SELECIONADOS

Os dados utilizados para construção do software foram divididos, como já dito, em duas categorias: os dados de entrada que são compostos pelos três arquivos disponibilizados pela concessionária de energia ELEKTRO (cadastro, medidor e consumo) e os arquivos referentes a Pesquisa de Posses e Hábitos dos consumidores (PPH). Neste item é explicado também como foram feitas as ligações dos dados ditos de entrada e dos dados da pesquisa.



### 4.2.1 Dados de Entrada

Os dados de entrada que foram enviados pela concessionária eram compostos, como dito anteriormente, por três arquivos, a saber: cadastro, medidor e consumo. Estes arquivos continham inúmeras variáveis das quais as mais relevantes foram selecionadas para serem utilizadas na construção de *clusters* ou famílias de clientes de mesmo perfil ou características descritivas. Após a seleção das variáveis, esses dados foram “linkados”, já que pertenciam a bancos de dados diferentes para obtenção de apenas um banco.

Escolhidas as variáveis a serem utilizadas na “clusterização” em cada um dos bancos de dados, o passo seguinte foi seccionar as variáveis por faixas, como descrito a seguir.

- **Banco de dados de Cadastro**

1. Classe Principal: 7 faixas, variando de 1 a 7.

2. Sub-grupo de Tensão: 4 faixas, variando de 1 a 4.

3. Tipo de Construção: 10 faixas, sendo: CD, DM, ED, ES, HP, HT, IN, OU, PR e RS.

4. Tipo de Fornecimento: 4 faixas, sendo: BI, MO, MR e TR.

- **Banco de dados de Medidor**

5. Status do Medidor: 3 faixas, sendo: A, C LTDA e P.

6. Marca: 11 faixas, sendo: GE, NANSEN, SCHLUMBERGER, ELO, ESB, AEG, SIEMENS, TOSHIBA, GEC, ALSTOM e OUTRAS.

- **Banco de dados de Consumo**

7. Média de Consumo: 4 faixas, sendo: $0 > x \geq 100$, $100 > x \geq 300$, $300 > x \geq 700$ e $x > 700$.



No arquivo cadastro, escolheu-se as variáveis que melhor identificariam os clientes em uma busca de perda. A **classe principal** foi segmentada nas sete faixas representando: residencial, comercial, rural, industrial, poder público, serviço público e iluminação pública. Já a variável **subgrupo de fornecimento** foi divida em quatro faixas onde cada uma delas, numeradas de 1 a 4 significando respectivamente residencial, rural, comercial e iluminação pública. O **tipo de construção** tem a seguinte segmentação: casa (CD), demolido (DM), escritório (ED), escola (ES),

hospital (HP), hotel (HT), prédio (PR), residência (RS), indefinido (IN) e outros (OU). A variável **tipo de fornecimento** refere-se ao tipo de ligação a que o ramal está ligado, sendo monofásico (MO), bifásico (BI), trifásico (TR) e MR.

Já para o arquivo medidor, duas variáveis foram escolhidas, pois desejava-se identificar se tinha alguma marca de medidor na qual fraude era mais incidente. As variáveis relacionadas ao medidor foram **status do medidor** e **marca do medidor**, que foram anteriormente descritas.

No banco consumo a variável utilizada foi a **média de consumo**, que era a média de consumo dos últimos 12 meses de dados disponibilizados pela empresa, que correspondiam ao consumo de fevereiro de 2003 a janeiro de 2004. A média foi dividida em 4 faixas como se mostrou anteriormente.

Como se pôde ver foram selecionadas 7 variáveis, as quais foram seccionadas em um total de 43 faixas. A faixa correspondente a um dado de um cliente pertencente a uma determinada faixa de uma variável recebe o valor “1” e para as demais faixas recebem o valor “-1”. Assim cria-se uma seqüência de “1” e “-1” que caracteriza esse cliente. Para exemplificar, selecionou-se a transformação do tipo de dado numérico da variável média de consumo. Essa é uma variável que foi seccionada em 4 faixas. Se o cliente pertence a faixa 2 (consumo entre 100 e 300 kWh), por exemplo, a seqüência numérica de “1” e “-1” para essa variável seria a seguinte:

| -1 | 1 | -1 | -1 |
|---|---|---|---|

**Figura 7: Caracterização de um cliente na variável média de consumo.**

Com isso, tem-se uma seqüência de “1” e “-1” que foi denominada de DNA do cliente, onde cada DNA possui 43 campos. A da rede de Kohonen teve por objetivo agrupar os clientes de mesma seqüência de “1” e “-1” no mesmo *cluster* ou *clusters* vizinhos. A figura a seguir mostra a configuração do chamado DNA.

| Classe Principal | | | | | | | Subgrupo de Tensão | | | | Tipo de Construção | | | | | | | | | | Tipo de Fornecimento | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 4 | CD | DM | ED | ES | HP | HT | IN | OU | PR | RS | BI | MO | MR | TR |
| **1** | **-1** | **-1** | **-1** | **-1** | **-1** | **-1** | **1** | **-1** | **-1** | **-1** | **-1** | **-1** | **-1** | **-1** | **-1** | **-1** | **-1** | **-1** | **1** | **-1** | **1** | **-1** | **-1** | **-1** |

| Status do Medidor | | | Marca do Medidor | | | | | | | | | | | Consumo | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | C | P | GE | NAN | SCH | ELO | ESB | AEG | SIE | TOS | GEC | ALS | OUT | Fx 1 | Fx 2 | Fx 3 | Fx 4 |
| **-1** | **-1** | **1** | **-1** | **-1** | **-1** | **-1** | **-1** | **-1** | **-1** | **1** | **-1** | **-1** | **-1** | **-1** | **1** | **-1** | **-1** |

**Figura 8: Configuração de um DNA.**

No DNA deste cliente consta que ele pertence à classe principal 1 (residencial) e ao subgrupo de fornecimento 1 (residencial BT), que seu tipo de construção é PR (prédio), que o tipo de fornecimento é BI (bifásico), que o status de seu medidor é P (principal), que a marca deste é TOS (Toshiba) e, finalmente, que faixa de consumo a que pertence é a 2 (248 kWh).

### 4.2.2 Dados de pesquisa

Os dados de pesquisa foram obtidos através da Pesquisa de Posses e Hábitos (PPH) feita na área de concessão da ELEKTRO nos meses de abril, maio e junho de 2004. Essa pesquisa foi feita no segmento residencial para clientes adimplentes (ou normais), inadimplentes e fraudulentos. Os questionários da PPH se encontram em anexo neste trabalho. Nesta pesquisa foram utilizados dois tipos de questionário, um que era destinado aos clientes adimplentes ou normais (anexo 1) e outro destinado aos clientes inadimplentes e fraudadores (anexo 2). Como dito anteriormente as variáveis de pesquisa foram utilizadas para o processo de classificação, por análise fuzzy, de um cliente como adimplente, inadimplente ou fraudulento.

As variáveis escolhidas compunham três bancos de dados, sendo esses para os consumidores residenciais normal, fraudulento e inadimplente. Como nesses bancos existiam questões (variáveis) não coincidentes, foram escolhidas apenas aquelas comuns aos três bancos citados anteriormente. Assim, foi possível fazer uma comparação dos perfis destas variáveis escolhidas nestes três segmentos. O critério de escolha dessas variáveis foi pelo diagnóstico daquelas que teriam melhor separação na “clusterização” dos três segmentos: normal, fraudulento e inadimplente. A seguir são mostrados os histogramas referentes à freqüência relativa de cada uma das questões (ou variáveis) escolhidas.

A primeira questão selecionada foi a que perguntava no domicílio entrevistado o número de pessoas que viviam nele, ela foi denominada por variável **pessoas**. As faixas escolhidas para se montar o histograma e que também foram utilizadas para se fazer a análise fuzzy foram: até 2 pessoas, de 3 a 5 pessoas e acima de 5 pessoas. Este histograma é mostrado na figura a seguir.

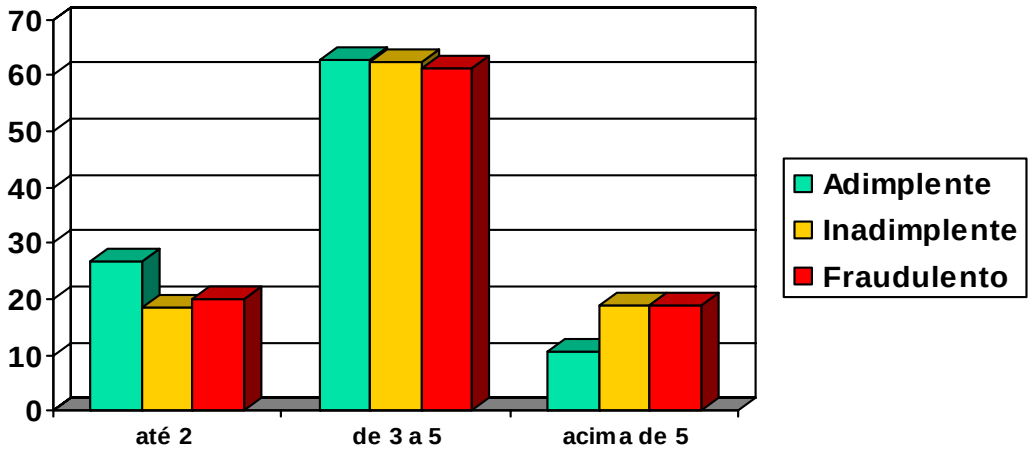


**Figura 9: Número de moradores em cada domicílio entrevistado**

Outra questão selecionada foi o tempo que o entrevistado morava no domicílio, que foi chamada de variável **tempo**. Esta variável foi seccionada nas seguintes faixas: menos de 2 anos, entre 2 e 7 anos e acima de 7 anos. Esta questão pode ser visualizada no histograma que se segue.

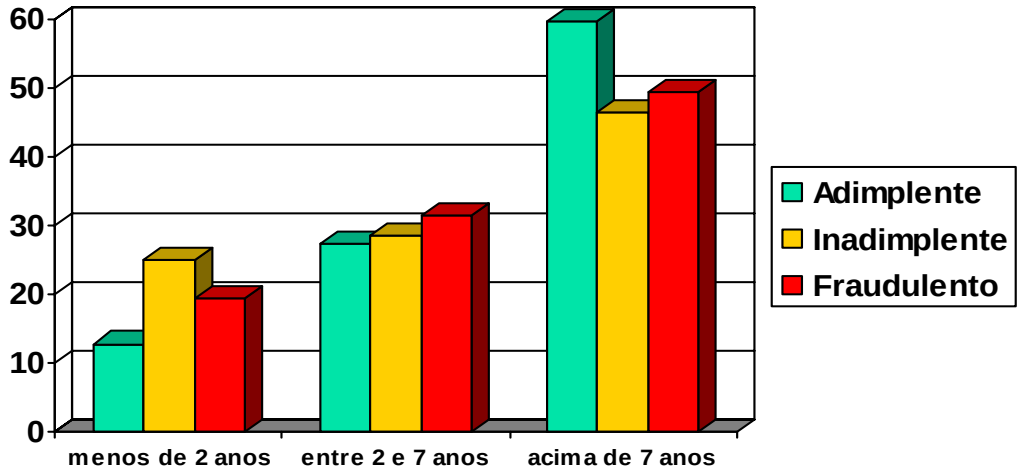


**Figura 10: Tempo que o entrevistado mora no domicílio**



Escolheu-se também a **área construída**, em $m^2$, como uma das variáveis para se fazer a análise fuzzy. Esta é uma variável de caracterização do domicílio pesquisado. Fez-se a seguinte segmentação desta variável: menos de 50 $m^2$, de 51 a 75 $m^2$, de 76 a 100 $m^2$, de 101 a 150 $m^2$, de 151 a 200 $m^2$ e acima de 200 $m^2$. O histograma referente a esta variável é mostrado na seqüência.

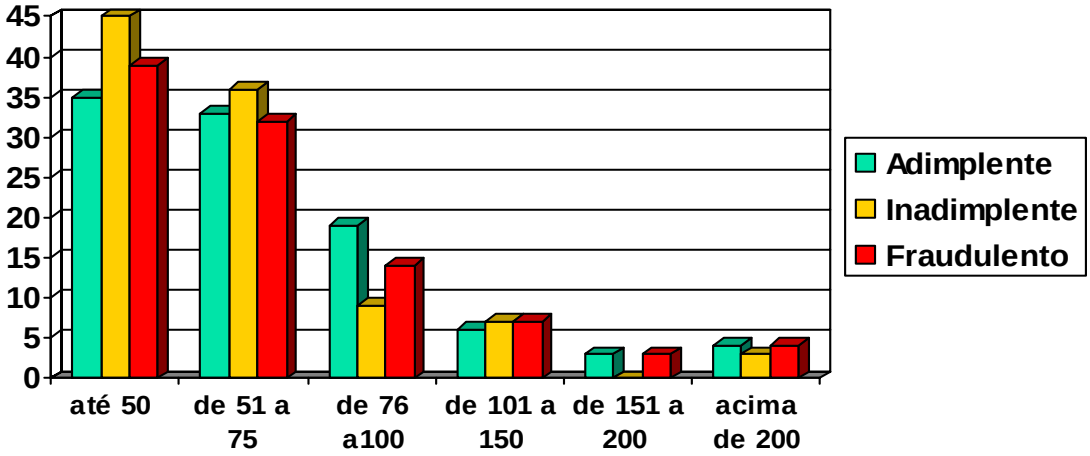


**Figura 11: Área construída do domicílio em $m^2$**

Escolheram-se também como variáveis para a análise fuzzy algumas questões relacionadas com a posse média de alguns aparelhos eletrodomésticos. Como a quantidade destes aparelhos era muito grande, fizeram-se alguns testes estatísticos para se escolher aqueles que deveriam ser escolhidos.

A seguir são mostradas as posses médias de alguns destes aparelhos nos três segmentos utilizados para análise. O primeiro aparelho, cujo histograma de posse média é mostrado, é a lâmpada. Foram mostradas as posses apenas para as lâmpadas fluorescentes e incandescentes. No entanto, vale frisar que as lâmpadas fluorescentes compactas foram inseridas ao montante de lâmpadas fluorescentes. A seguir tem-se o histograma deste equipamento.



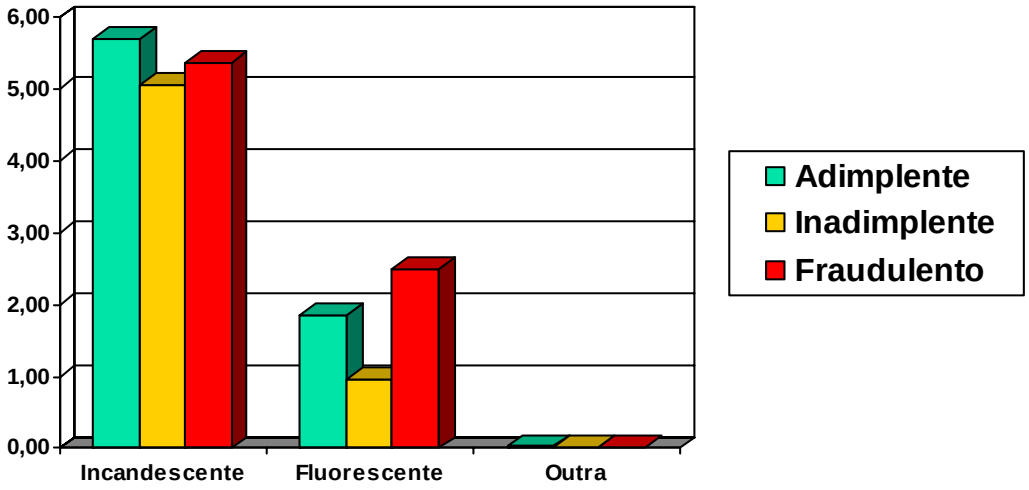


**Figura 12: Posse média de lâmpadas**

Na figura que se segue é mostrada a posse média de alguns dos aparelhos eletrodomésticos pesquisados. São eles: TV, refrigerador, freezer, ar condicionado e chuveiro elétrico.

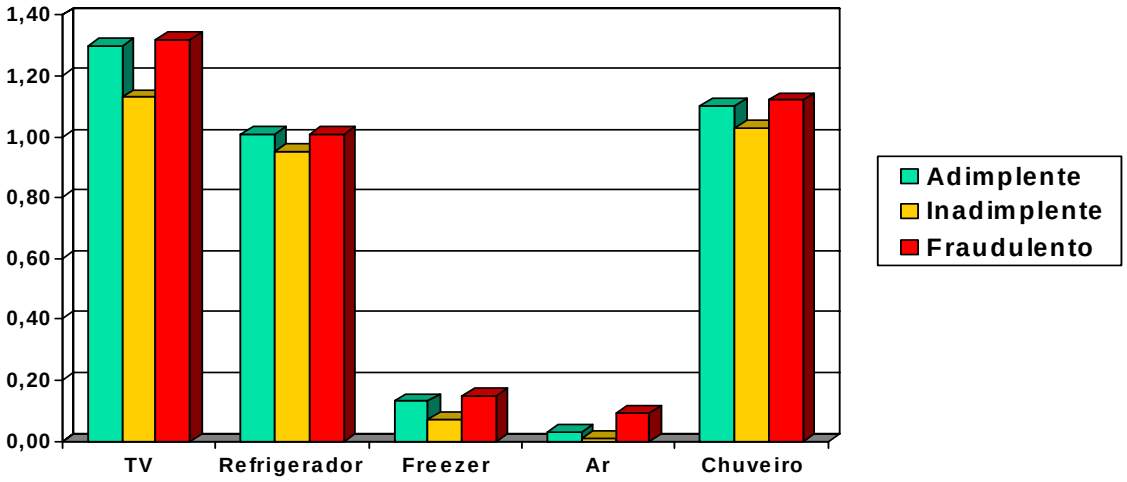


**Figura 13: Posse média de eletrodomésticos (1)**

Na seqüência, é apresentado o histograma com a posse média de outros aparelhos eletrodomésticos pesquisados. Estes aparelhos são: microondas, ventilador, liquidificador, vídeo e DVD. Percebe-se na figura que estes dois últimos foram somados e se fez apenas uma estatística. É importante dizer, que foram pesquisadas as posses de outros eletrodomésticos, mas devido a pouca incidência destes, eles não foram considerados para se fazer a análise fuzzy.

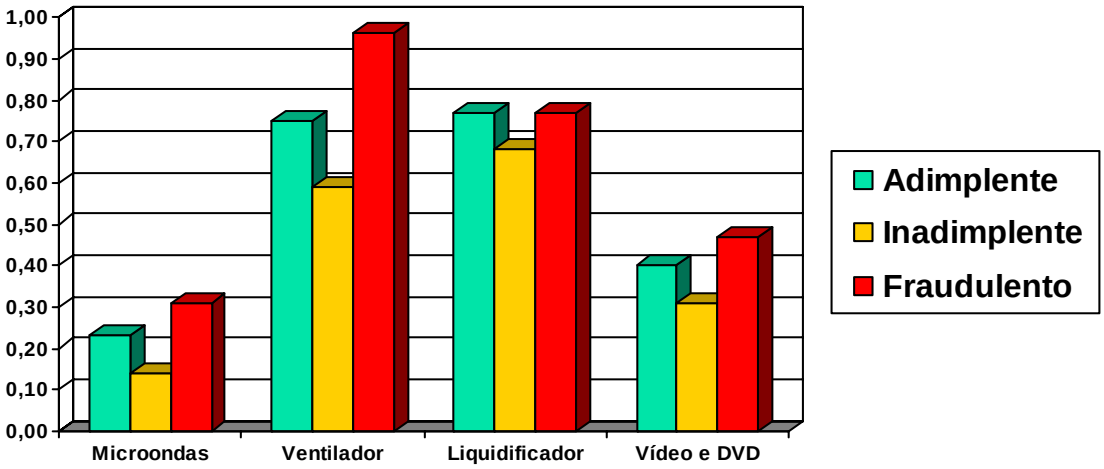


**Figura 14: Posse média de eletrodomésticos (2)**



Vê-se, pelos histogramas anteriores, que os clientes residenciais fraudulentos possuem, em média, mais aparelhos eletrodomésticos que os adimplentes e inadimplentes. E que estes últimos possuem menor quantidades de aparelhos. Visto que a quantidade de eletrodomésticos é uma variável que caracteriza bem estes três segmentos de clientes, procurou-se então utilizar algumas destas para a classificação. Nem todas foram selecionadas, pois isto em muito aumentaria número de variáveis e assim o tempo de processamento destas. Tomou-se então, variáveis que teriam mais significância quanto à separação por quantidades. Para se ter uma escolha mais acurada utilizou-se dois testes: o de Levene, que testa a igualdade de variâncias, e o teste-t, que faz uma comparação entre igualdades de médias de duas amostras [5].

As variáveis utilizadas nos testes foram as descritas a seguir: quantidades de lâmpadas (somou-se todos os tipos de lâmpadas encontrados no questionário da PPH, a saber: lâmpadas incandescentes,

lâmpadas fluorescentes tubulares e compactas e, finalmente, as lâmpadas dicróicas), quantidade de refrigeradores, de freezers, de condicionadores de ar, de televisores e quantidade de chuveiros. Os demais eletrodomésticos não foram utilizados para o teste, pois eles não eram de muita relevância no consumo de energia elétrica.

Para o teste de igualdade de médias, teste-t, obteve-se o seguinte resultado: para a amostra de clientes fraudulentos e inadimplentes a hipótese nula de que as médias das duas amostras selecionadas seriam iguais foi sempre rejeitada para todas as seis variáveis descritas anteriormente; já quando a comparação foi feita entre as médias das amostras de consumidores residenciais normais e inadimplentes, viu-se que houve uma aceitação da hipótese nula apenas para a variável quantidade de chuveiros, mas o valor encontrado para esta estatística (0,056) quase se aproximou do valor de significância do teste que é de (0,05) o que podia ser um indício de que estas médias podiam ser consideradas diferentes; por último quando se analisou as amostras de fraudulentos e normais, viu-se que as médias eram iguais, ou seja, aceitava-se a hipótese nula, para quase todas as variáveis, salvo a quantidade de condicionadores de ar na qual esta hipótese foi rejeitada.

Levando em consideração o teste de Levene, o qual compara se duas amostras têm igualdade de variâncias, obteve-se o seguinte resultado: quando se compararam as amostras de clientes residenciais fraudulentos e inadimplentes houve aceitação da hipótese nula, igualdade de variância, apenas para a variável quantidade de refrigeradores; o mesmo aconteceu

para a comparação entre as amostras de consumidores normais e inadimplentes, onde hipótese nula só foi aceita para a quantidade de refrigeradores; ao se efetuar este teste para a amostra de clientes fraudulentos e normais, a aceitação da hipótese nula ocorreu para três variáveis (quantidade de refrigeradores, de freezers e de televisores).

Os quadros abaixo ilustram melhor a aceitação ou a rejeição da hipótese nula para os testes acima citados, teste-t e teste de Levene.

**Tabela 2: Teste-t para a posse média de eletrodomésticos**

| **Teste-t (Ho: Igualdade de Médias)** | | | |
|---|---|---|---|
| | Fraude e Inadimplência | Normal e Inadimplência | Normal e Fraude |
| Quant. de lâmpadas | Rejeita Ho | Rejeita Ho | **Aceita Ho** |
| Quant. de refrigeradores | Rejeita Ho | Rejeita Ho | **Aceita Ho** |
| Quant. de freezers | Rejeita Ho | Rejeita Ho | **Aceita Ho** |
| Quant. de condicionadores de ar | Rejeita Ho | Rejeita Ho | Rejeita Ho |
| Quant. de televisores | Rejeita Ho | Rejeita Ho | **Aceita Ho** |
| Quant. de chuveiros | Rejeita Ho | **Aceita Ho** | **Aceita Ho** |

**Tabela 3: Teste de Levene para a posse média de eletrodomésticos**

| **Teste de Levene (Ho: Igualdade de Variâncias)** | | | |
|---|---|---|---|
| | Fraude e Inadimplência | Normal e Inadimplência | Normal e Fraude |
| Quant. de lâmpadas | Rejeita Ho | Rejeita Ho | Rejeita Ho |
| Quant. de refrigeradores | **Aceita Ho** | **Aceita Ho** | **Aceita Ho** |
| Quant. de freezers | Rejeita Ho | Rejeita Ho | **Aceita Ho** |
| Quant. de condicionadores de ar | Rejeita Ho | Rejeita Ho | Rejeita Ho |
| Quant. de televisores | Rejeita Ho | Rejeita Ho | **Aceita Ho** |
| Quant. de chuveiros | Rejeita Ho | Rejeita Ho | Rejeita Ho |

A escolha das variáveis de quantidades de aparelhos a serem utilizadas no modelo de decisão fuzzy partiu dos quadros acima. Foram escolhidas três variáveis a serem utilizadas, pois não se podia escolher

muitas variáveis deste tipo, já que se tinha uma quantidade bastante considerável destas. Os critérios de seleção foram: primeiramente a rejeição da hipótese nula no teste de Levene e depois a rejeição da hipótese nula no teste-t. Queria-se utilizar varáveis que pudessem ser discerníveis mais pelo seu grau de variabilidade dos componentes das amostras do que pelo seu grau de diferença entre médias.

A primeira variável a ser escolhida foi a **quantidade de condicionadores de ar**, pois esta rejeitava a hipótese nula para a todas as amostras tanto no teste de Levene quanto no teste-t. Depois escolheu-se a **quantidade de chuveiros**, pois esta variável tinha a hipótese nula rejeitada para todas as amostras no teste de Levene, que era o critério fundamental. Além disso, a variável rejeitava a hipótese nula para a comparação entre as amostras de clientes fraudulentos e inadimplentes e, como dito acima, aceitava a hipótese nula para a comparação entre as amostras de normais e inadimplentes como um valor muito próximo ao valor de significância do teste. Por esse mesmo critério dever-se-ia tomar a variável quantidade de lâmpadas, mas esta não foi utilizada devido ao grande número de tipos, o que aumentaria em muito o número de variáveis. A última variável a ser selecionada foi então a **quantidade de freezers**, pois esta tinha a hipótese nula rejeitada para duas das comparações de variâncias e médias realizadas. Da mesma forma que a quantidade de freezers a quantidade de televisores também possuía a hipótese nula rejeitada para duas das comparações de variâncias e médias realizadas. No entanto, esta variável

não foi a escolhida, pois ela ocorria em todas as classes de consumidores ao contrário do freezer que seleciona mais uma determinada classe.

Utilizou-se apenas o **grau de utilização de freezers**, apesar de ter se tentado tomar os graus de utilização das outras 2 variáveis também (refrigeradores e condicionadores de ar). No entanto, as tabelas de graus de utilização para ar condicionador e refrigerador nos questionários da PPH eram de uma complexidade tal que aumentariam em muito o número de variáveis.

A seguir são apresentadas as variáveis que foram escolhidas para a utilização no modelo de análise fuzzy e que estão ligadas a condições sócio-econômicas dos clientes da ELEKTRO pesquisados.

A primeira das variáveis, que será mostrada no histograma a seguir, é a **renda familiar** declarada pelo entrevistado, em salários mínimos (s.m.). Esta variável foi dividida nas seguintes faixas: de 1 a 2 , de 3 a 7 , de 8 a 10 e mais de 10 salários  mínimos.

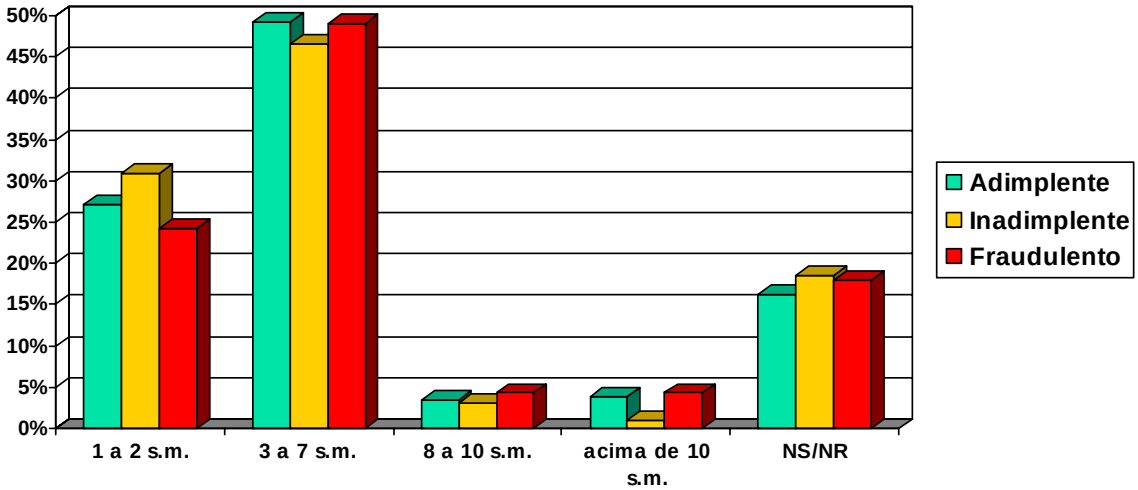


**Figura 15: Renda familiar declarada pelo entrevistado (em s.m.)**

Uma outra questão da PPH e que também foi escolhida como variável para se fazer a classificação de consumidores adimplentes, inadimplentes e fraudulentos foi a classificação da região do domicílio pelo pesquisador como de luxo ou de classe média alta ou de classe média baixa ou pobre, esta variável foi denominada de **região**. A seguir é mostrado o histograma referente a esta questão.

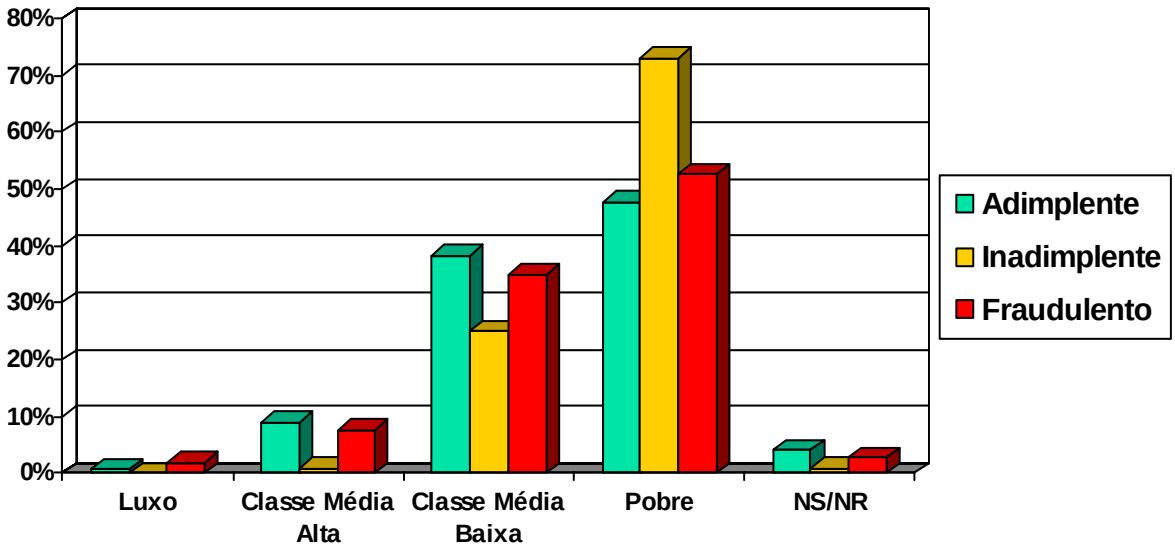


**Figura 16: Classificação do domicílio pelo pesquisador**

A questão ou variável seguinte, e que também foi selecionada, foi a que perguntava se o domicílio era próximo ou não da favela ou se encontrava na favela (variável **prox_fav**). As respostas da PPH podem ser visualizadas no histograma que se segue.

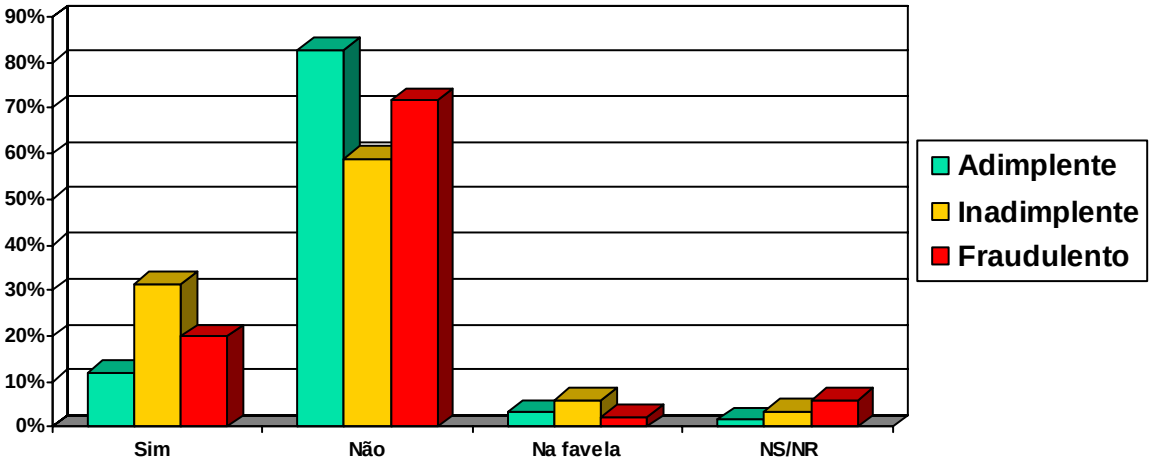


**Figura 17: Proximidade do domicílio em relação a favela**



A última questão dos questionários, que também foi escolhida, perguntava qual era o peso da conta de luz no orçamento dos entrevistados. Esta questão é mostrada na figura abaixo e a variável foi denominada por **orçamento**.

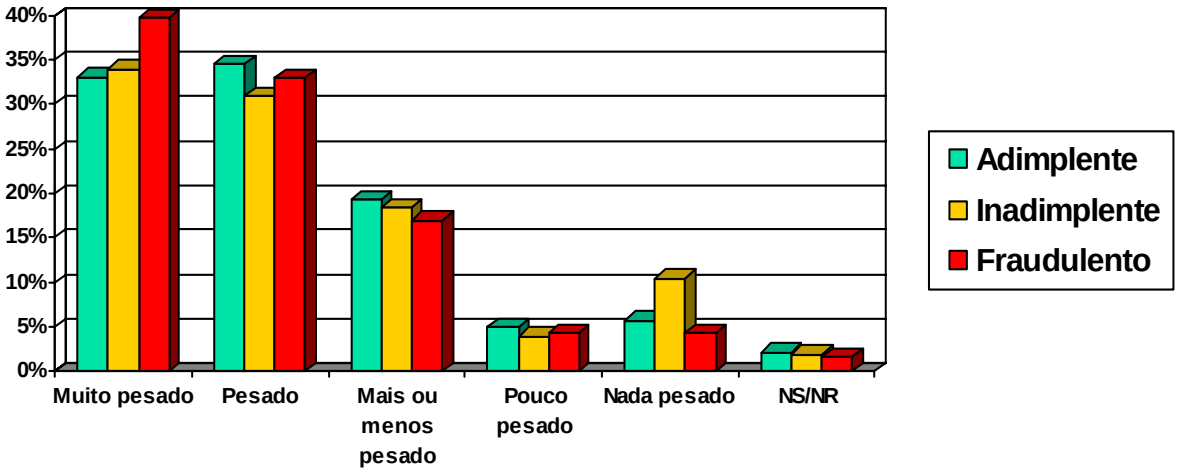


**Figura 18: Peso da conta de luz no orçamento do entrevistado**

Na tabela a seguir tem-se a relação destas 11 variáveis escolhidas, suas respectivas descrições e a questão que estas estão relacionadas aos questionários da pesquisa que se encontram nos anexos 1 e 2 deste trabalho.

**Tabela 4: Relação das variáveis selecionadas no questionário da PPH**

| Variável | Descrição | Questão |
|---|---|---|
| Pessoas | Número de pessoas que moram no domicílio | 1.10 |
| Tempo | Quanto tempo a família mora no domicílio | 1.12 |
| Área | Área construída do domicílio | 2.6 |
| Freezer_qu | Quantidade de freezers | 6.1 |
| Freezer_ut | Grau de utilização dos freezers | 6.1 |
| Ar_qu | Quantidade de condicionadores de ar | 7.1 |
| Chuv_qu | Quantidade de chuveiros elétricos | 10.2 |
| Renda | Renda familiar declarada | 11.2 |
| Região | Região do domicílio | 11.5 |
| Prox_fav | Proximidade à favela | 11.6 |
| Orçamento | Peso da conta de luz no orçamento | 11.7 |



Escolhidas as variáveis, o próximo passo foi o de seccionar estas variáveis por faixas, seguindo as separações mostradas nos histogramas apresentados anteriormente. No quadro a seguir tem-se: a variável em cada uma das faixas, a descrição de cada faixa e a qual questão que estas variáveis estão ligadas no questionário da PPH. Foram selecionadas, portanto, 11 variáveis e um total de 40 faixas onde o número de faixas para cada variável não obedeceu a um valor único. Por exemplo, a variável,

pessoas possuía apenas três faixas enquanto que a variável área possuía 6 faixas. Portanto, essas quantidades de faixas variavam de questão para questão.

**Tabela 5: Relação das variáveis selecionadas e suas respectivas faixas.**

| Variável e Faixa | Descrição da Faixa | Questão |
|---|---|---|
| Pessoas Fx1 | até 2 pessoas | 1.10 |
| Pessoas Fx2 | entre 3 e 5 pessoas | 1.10 |
| Pessoas Fx3 | mais que 5 pessoas | 1.10 |
| Tempo Fx1 | menos que 2 anos | 1.12 |
| Tempo Fx2 | entre 2 e 7 anos | 1.12 |
| Tempo Fx3 | mais que 7 anos | 1.12 |
| Area Fx1 | até 50 m2 | 2.6 |
| Area Fx2 | de 51 a 75 m2 | 2.6 |
| Area Fx3 | de 76 a 100 m2 | 2.6 |
| Area Fx4 | de 101 a 150 m2 | 2.6 |
| Area Fx5 | de 151 a 200 m2 | 2.6 |
| Area Fx6 | acima de 200 m2 | 2.6 |
| Freezer_qu Fx1 | Nenhum | 6.1 |
| Freezer_qu Fx2 | 1 | 6.1 |
| Freezer_qu Fx3 | mais que 1 | 6.1 |
| Freezer_ut Fx1 | pequena | 6.1 |
| Freezer_ut Fx2 | média | 6.1 |
| Freezer_ut Fx3 | Grande | 6.1 |
| Ar_qu Fx1 | Nenhum | 7.1 |
| Ar_qu Fx2 | 1 | 7.1 |
| Ar_qu Fx3 | mais que 1 | 7.1 |
| Chuv_qu Fx1 | Nenhum | 10.2 |
| Chuv_qu Fx2 | 1 | 10.2 |
| Chuv_qu Fx3 | mais que 1 | 10.2 |

**Tabela 5: Relação das variáveis selecionadas e suas respectivas faixas (continuação)**

| Variável e Faixa | Descrição da Faixa | Questão |
|---|---|---|
| Renda Fx1 | até 2 salários | 11.2 |
| Renda Fx2 | de 2 a 7 salários | 11.2 |
| Renda Fx3 | de 7 a 10 salários | 11.2 |
| Renda Fx4 | maior que 10 salários | 11.2 |
| Região Fx1 | Luxo | 11.5 |
| Região Fx2 | classe média alta | 11.5 |
| Região Fx3 | classe média baixa | 11.5 |
| Região Fx4 | Pobre | 11.5 |
| Prox_fav Fx1 | próximo à favela | 11.6 |
| Prox_fav Fx2 | longe da favela | 11.6 |
| Prox_fav Fx3 | na favela | 11.6 |
| Orçamento Fx1 | muito pesado | 11.7 |
| Orçamento Fx2 | Pesado | 11.7 |
| Orçamento Fx3 | mais ou menos pesado | 11.7 |
| Orçamento Fx4 | pouco pesado | 11.7 |
| Orçamento Fx5 | nada pesado | 11.7 |



### 4.2.3 Processo de Ligação dos Dados de Entrada com os Dados de Pesquisa

A variável Uc (Unidade consumidora) foi utilizada para que a ligação entre os dados de entrada e de pesquisa pudesse ser realizada. Esta variável, Uc, é uma das mais importantes, pois cada cliente da ELEKTRO possui este número. Além disto, este número não permite duplicidade, por ser único. Isto é importante, pois quando se liga pela Uc vários bancos de dados, o risco de serem ligados clientes com dados trocados é nulo.

A seguir tem-se um esquema que mostra a ligação dos arquivos de entrada (cadastro geral) e pesquisa PPH pela Uc.

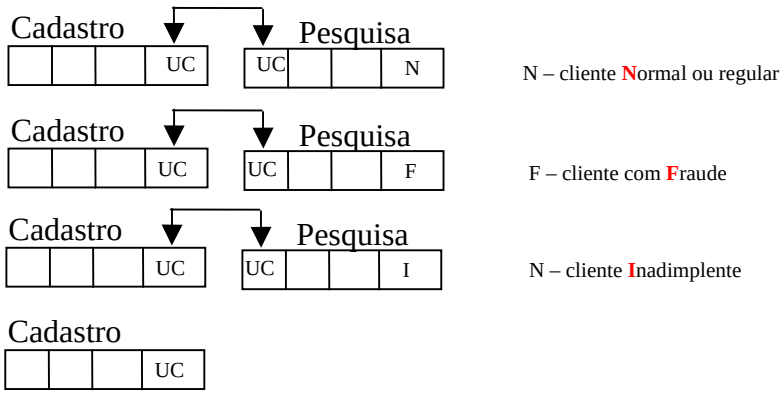


**Figura 19: Esquema de ligação entre os dados de entrada e dados de pesquisa pela Uc**

## 4.3 METODOLOGIA UTILIZADA

### 4.3.1 Introdução

Toda metodologia utilizada nesta dissertação de mestrado foi aplicada na construção de um *software* amigável que foi implantado na empresa de distribuição de energia elétrica ELEKTRO. Todos os resultados deste programa serão mostrados no tópico 5 que será apresentado mais tarde.

A metodologia foi dividida em duas partes. Na primeira etapa, foram utilizados os dados de cadastros fornecidos pela concessionária e que foram

transformados em DNA's, como mencionado no tópico 4.2.1, e classificados por uma rede de Kohonen. Depois de "clusterizados", os dados foram classificados por um sistema fuzzy que identificava a incidência de adimplência, inadimplência e fraude dentro de cada um dos *clusters* selecionados. Esta identificação foi feita a partir dos 1238 dados de pesquisa, cujas variáveis foram apresentadas na seção 4.2.

### 4.3.2 Processo de Formação dos *Clusters*



Depois de todo o tratamento realizado com os dados de entrada de cada cliente, a próxima etapa foi a classificação destes pela rede de Kohonen ou mapa auto-organizável. Em um mapa deste tipo, como visto anteriormente, os neurônios são colocados em nós de uma grade que é normalmente uni ou bidimensional. No caso deste trabalho esta rede foi de tamanho 4X4 totalizando 16 neurônios. Os neurônios se tornam seletivamente sintonizados a vários padrões de entrada, estímulos, no decorrer de um processo de aprendizagem. Esse processo é baseado na aprendizagem competitiva, ou seja, os neurônios de saída da grade competem entre si para serem ativados. O neurônio cujo vetor de pesos gera a menor distância Euclidiana com o vetor de entrada é o vencedor, ou seja, o cliente é classificado nesse neurônio. Todos os clientes classificados no mesmo neurônio pertencem ao mesmo *cluster*.

Utilizou-se um algoritmo não supervisionado na classificação dos dados de cadastro, pois não se sabia de antemão quais eram os grupos

naturais a serem formados e se queria visualizar a quantidade e a relação existente entre estes grupos naturais, muitas vezes denominados de *cluster*. Um outro fator que avaliou a utilização da rede de Kohonen foi que esta atua na projeção de um espaço multidimensional (as 40 variáveis selecionadas) a outro de dimensões visualizáveis (os 16 *clusters* gerados).

Além dos dados de entrada que se queria classificar, foram incluídos também os dados de entradas de clientes pesquisados na PPH nos três segmentos, fraudulentos, adimplentes e inadimplentes. Estes dados foram encontrados pela variável Uc (Unidade consumidora) de cada um destes clientes.

A seguir tem-se um esquema da formação dos *clusters* ou famílias ao se utilizar este tipo de mapa auto-organizável.

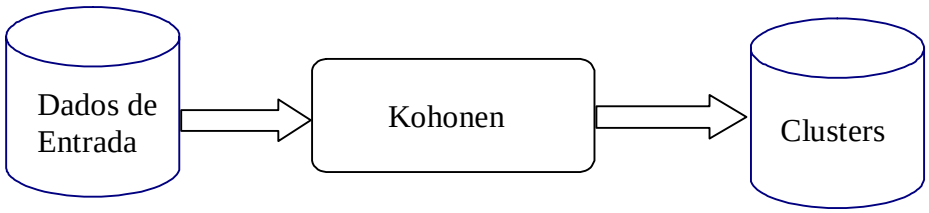


**Figura 20: Esquema de formação de *clusters*.**

A métrica utilizada neste método para se medir distância entre os vetores de entrada "**x**" e os de pesos "**w**" foi a Euclidiana. Vale observar que outras métricas poderiam ser utilizadas, no entanto esta é a mais usual e a que foi escolhida.

O que se esperava com este método é que em um *cluster* genérico todas as variáveis de entrada tivessem seqüências de "1" ou "-1" próximas

ou iguais, ou seja, um *cluster* ou família deveria ter seqüência parecida de DNA.

O software criado permitia que fossem utilizadas redes neurais bidimensionais quadradas. Por exemplo, uma rede bidimensional 4x4, totalizava 16 neurônios ou *clusters*. É importante frisar que, quanto maiores eram os números de neurônios na rede maior era também o tempo de processamento de classificação de dados, já que distância euclidiana deve ser feita a todos os neurônios.

Nos *clusters* que foram formados existiam alguns elementos que faziam parte da pesquisa de campo (PPH), pois como estes dados foram incluídos antes da "clusterização" ser realizada. A identificação destes era feita por suas Uc's. Com isso, era possível identificar quantos clientes pesquisados cada *cluster* possuía e, assim, passava-se para a fase seguinte que utilizava uma classificação por métodos de lógica fuzzy.

É importante dizer que, os *clusters* que não tivessem dados de pesquisa ou que os possuíssem em quantidades insuficientes, no caso menos de 20 pesquisas, não podiam participar do processo seguinte de classificação de clientes como normais, inadimplentes ou fraudulentos por técnicas Fuzzy.

### 4.3.3 Processo de Análise Fuzzy

As variáveis utilizadas nesta fase foram as de pesquisa de posses e hábitos de consumo (PPH) e o processo para escolhê-las foi o descrito na seção 4.2.2. Estas variáveis eram incluídas num banco de dados que possuía 1238 pesquisas de posses e hábitos de consumo, as quais 846 eram de clientes normais, 290 de fraudulentos e 102 de consumidores inadimplentes.

Depois de escolhidos os *clusters* que poderiam participar da classificação quanto ao número de clientes fraudulentos, adimplentes e inadimplentes, lembrando que estes tinham de ter pelo menos 20 clientes. O próximo passo foi fazer, para cada um dos três tipos de clientes citados anteriormente, as curvas de freqüência normalizada para as questões de pesquisa.

O procedimento para se obter estas curvas é descrito a seguir. A partir dos arquivos das PPH´s de clientes normais, fraudulentos e inadimplentes, já com apenas as 11 questões da pesquisa de posses e hábitos escolhidas e assim com as 40 variáveis (entende-se por variáveis as 11 questões em suas respectivas faixas), fez-se uma distribuição de freqüência para cada uma destas variáveis para cada um dos segmentos em análise. Resumindo, três distribuições de freqüência para as respostas dos clientes foram plotadas, uma para os clientes adimplentes, uma para os inadimplentes e outra para os fraudulentos.

Em seguida, para cada um dos três segmentos, fez-se o percentual da freqüência de cada variável em relação ao total. Ou seja, uma curva de freqüência relativa para cada um dos três “tipos” de clientes foi gerada.

Finalmente, fez-se uma normalização destas curvas pelo número máximo de respostas em cada variável, já que, como dito anteriormente, o número de consumidores pesquisados adimplentes, fraudulentos e inadimplentes pesquisados não era o mesmo. Para se ter uma idéia, o número de consumidores normais pesquisados era de 846, de fraudadores era de 290 e o de inadimplentes de apenas 102 clientes. Com isso, obteve-se o gráfico que é mostrado na figura 21.



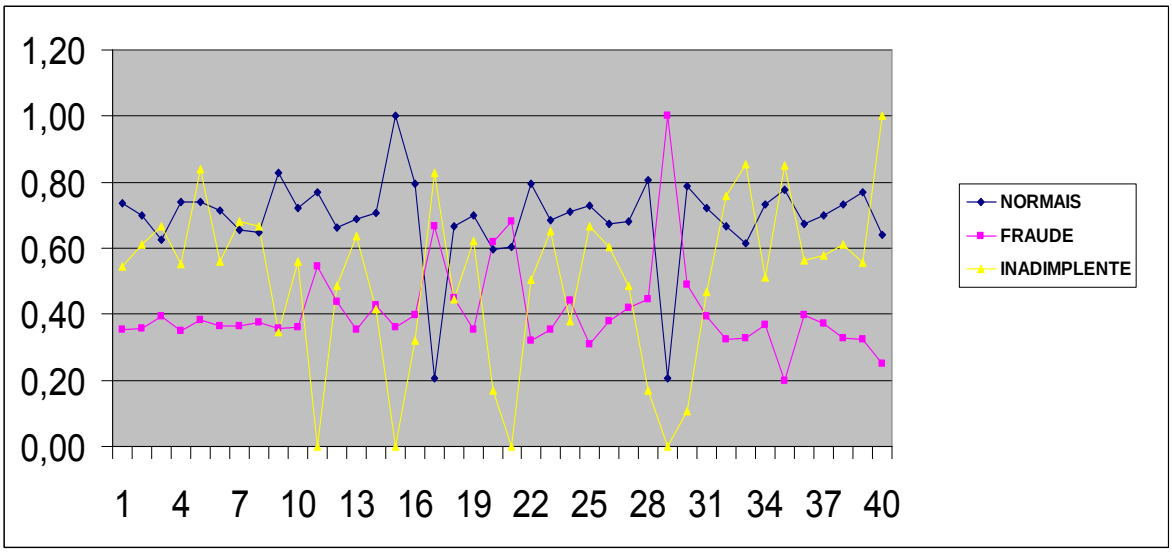


**Figura 21: Curva de freqüência normalizada das respostas para cada uma das 40 variáveis**

Na figura 21, vê-se que existem perguntas que têm o maior número de respostas que outras. O que se queria com esse método era tomar

aquelas variáveis que obtiveram um maior número de respostas para um determinado tipo de cliente, seja ele fraudulento ou inadimplente ou adimplente. Portanto, utilizou-se de diferentes níveis de cortes para que se pudessem analisar apenas as questões mais representativas para cada um dos três "tipos" de clientes. Ver-se-á que para cada um dos segmentos utilizaram-se níveis diferentes de cortes, objetivando escolher as perguntas (ou variáveis) que tinham maior freqüência de respostas para cada uma das três curvas. Pode-se notar que as configurações das três curvas são bem diferentes, isto faz com que se considerem níveis distintos para cada um dos cortes a serem considerados nestas curvas. Além disso, o corte também visa diminuir o número de curvas de pertinência na análise. As figuras na seqüência mostram com diferentes níveis de corte, realçados por cor vermelha, as variáveis mais representativas para cada um dos três segmentos de clientes. Pode-se notar que este foi um critério de corte, no entanto, o corte depende da sensibilidade do analista, cabendo a este escolhê-lo.

A primeira curva diz respeito às respostas dos clientes normais ou adimplentes, nesta foi estabelecido um nível de corte próximo de 0.8, onde foi selecionado um total de oito variáveis as quais são listadas a seguir: 9, 15, 16, 22, 28, 30, 35 e 39. Esta curva com seu respectivo corte pode ser vista na figura 22.

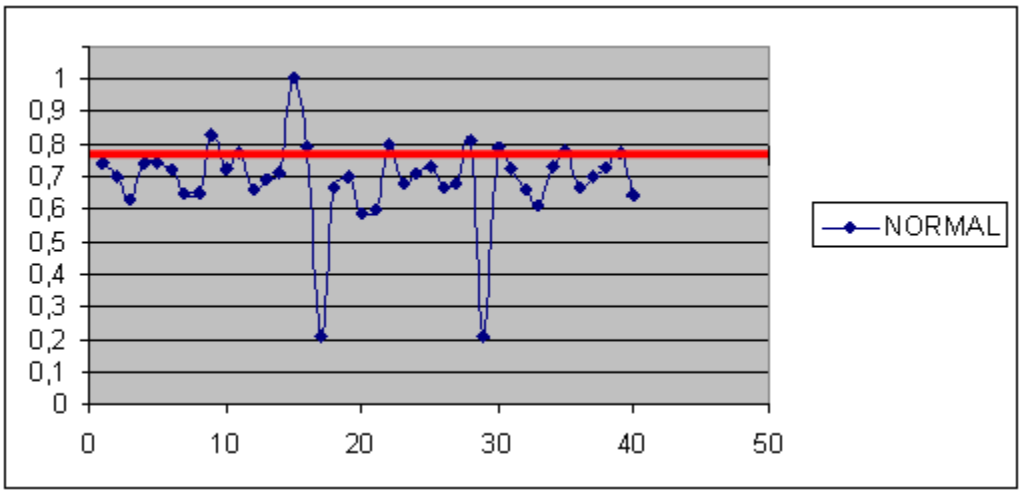


**Figura 22: Curva de freqüência respostas para os clientes normais com o corte**



Na seqüência, na figura 23 abaixo, é apresentada a curva de freqüência das respostas para os clientes fraudulentos. Nesta, o nível de corte escolhido foi de 0.4 e as variáveis selecionadas com esse corte foram: 11, 12, 14, 17, 18, 20, 21, 24, 27, 28, 29, 30 e 31. Somando assim, treze variáveis representativas.

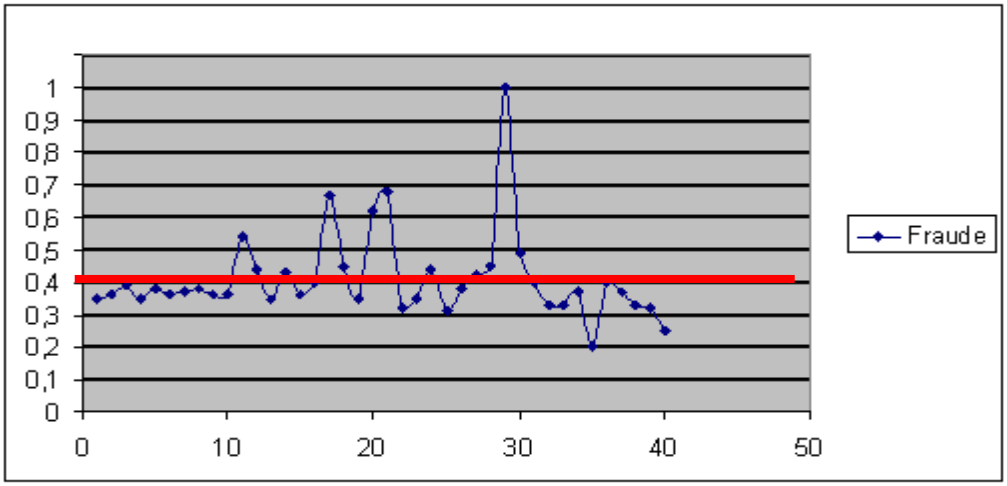


**Figura 23: Curva de freqüência respostas para os clientes fraudulentos com o corte.**

Já para a curva dos clientes inadimplentes as variáveis acima ou na linha de corte nível 0.6 foram selecionadas, eram 14 variáveis. Essas variáveis foram as de número 3, 5, 7, 8, 13, 17, 19, 23, 25, 32, 33, 35, 38 e 40. A figura 24 mostra a curva de respostas dos inadimplentes com o nível de corte sugerido.

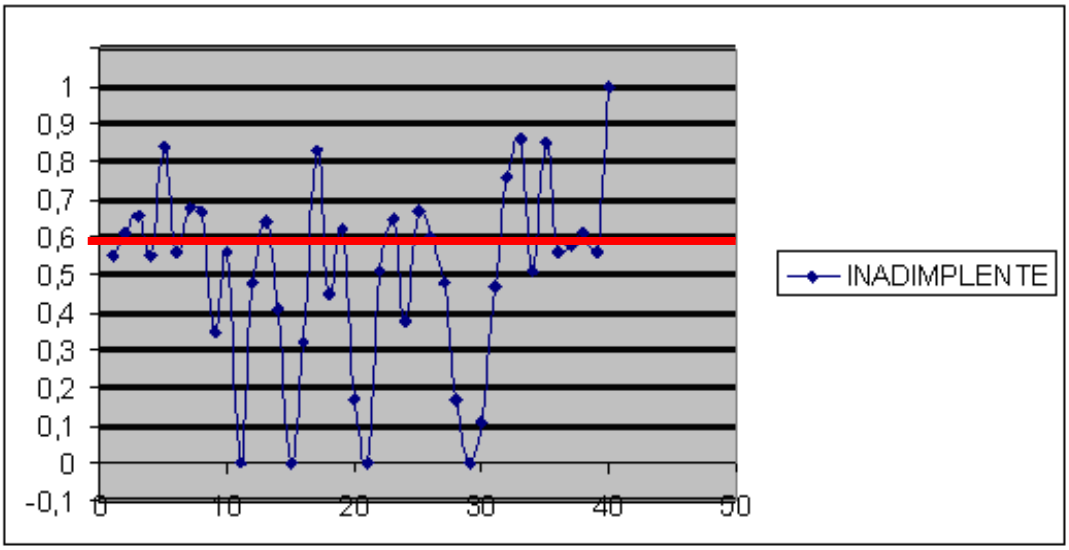


**Figura 24: Curva de freqüência respostas para os clientes inadimplentes com o corte.**

Pode-se notar que com os cortes escolhidos acima apenas as variáveis 28 e 30 foram selecionadas para os clientes adimplentes e fraudulentos, somente a variável 35 foi escolhida para os clientes inadimplentes e normais e, finalmente, apenas a variável 17 foi selecionada para os clientes fraudulentos e inadimplentes. Ou seja, a grande maioria das variáveis servia para caracterizar apenas um tipo de cliente, o que demonstra que as escolhas dos cortes foram bem dimensionadas. É percebível ainda que, existem variáveis que não foram selecionadas por

nenhum dos cortes adotados para os três tipos de clientes, são elas: 1, 2, 4, 6, 10, 26, 31, 34, 36 e 37. Pode-se dizer que estas variáveis não caracterizam bem os clientes da ELEKTRO quanto a adimplência, a fraude ou a inadimplência.

Selecionadas as variáveis que devem ser usadas para cada um dos "tipos" de clientes, o próximo passo foi tomar em cada um dos *clusters* escolhidos ou representativos (aqueles que têm no mínimo 20 pesquisas somadas) as variáveis escolhidas e montar a curva de pertinência para cada uma destas variáveis. As curvas de pertinência foram montadas com os graus de pertinência variando de 0 a 1, de modo que a análise fuzzy pudesse ser aplicada.

Para exemplificar, fez-se a "clusterização" via Kohonen dos dados de pesquisa, onde uma rede de 16 neurônios (4X4) foi utilizada para essa classificação. Com o mapa de Kohonen "rodado", apenas nove *clusters* relevantes foram obtidos, aqueles que possuíam no mínimo 20 pesquisas. Estes *clusters* eram: 1, 3, 4, 5, 8, 11, 13, 15 e 16.

Com intuito de que a montagem das curvas de pertinência fosse bem compreendida, mostra-se a seguir uma tabela com a primeira variável selecionada para os clientes normais (variável 9) e a última para os consumidores fraudulentos (variável 31). Deve-se lembrar que, estas variáveis são aquelas selecionadas pelos cortes de cada uma das curvas de freqüência relativas, os quais foram descritos previamente. Nesta tabela, se encontra o número de respostas positivas a uma determinada variável no cluster, o total de pesquisas no *cluster* e a pertinência da variável (ou

pergunta do questionário), que foi determinada pela razão entre as respostas positivas e o total de pesquisas.

**Tabela 6: Pertinência para as variáveis 9 dos clientes normais e 31 dos clientes fraudulentos**

| Normal - Variável 9 | | | |
|---|---|---|---|
| **Cluster** | **Pesquisas no cluster** | **Total de Pesquisas** | **Pertinência da variável** |
| 1 | 12 | 62 | **0,19** |
| 3 | 17 | 146 | **0,12** |
| 4 | 1 | 40 | **0,03** |
| 5 | 3 | 53 | **0,06** |
| 8 | 10 | 56 | **0,18** |
| 11 | 13 | 82 | **0,16** |
| 13 | 17 | 103 | **0,17** |
| 15 | 32 | 161 | **0,20** |
| 16 | 19 | 127 | **0,15** |

| Fraude - Variável 31 | | | |
|---|---|---|---|
| **Cluster** | **Pesquisas no cluster** | **Total de Pesquisas** | **Pertinência da variável** |
| 1 | 3 | 20 | **0,15** |
| 3 | 7 | 36 | **0,19** |
| 4 | 3 | 18 | **0,17** |
| 5 | 16 | 55 | **0,29** |
| 8 | 4 | 27 | **0,15** |
| 11 | 18 | 40 | **0,45** |
| 13 | 30 | 52 | **0,58** |
| 15 | 3 | 26 | **0,12** |
| 16 | 4 | 12 | **0,33** |

Vê-se na tabela 6, mostrada anteriormente, que existem valores totais de pesquisas para essas variáveis selecionadas que são menores que 20. É importante frisar que este valor é para apenas um segmento, e não a soma dos “três tipos” de clientes como o modelo propõe. Pode-se notar, por exemplo, que no *cluster* 16 há um total 12 entrevistas para clientes fraudulentos, mas existem 127 PPH’s para os clientes normais. Portanto, somando-se este valores o número encontrado é maior que 20, isso sem incluir os clientes inadimplentes.

Após o cálculo da pertinência da perguntas em cada *cluster*, partiu-se para o passo seguinte que foi o de plotar cada uma destas curvas. As figuras 25 e 26 a seguir mostram, respectivamente, as curvas das perguntas 9, para os clientes adimplentes, e 31, para os consumidores fraudadores.

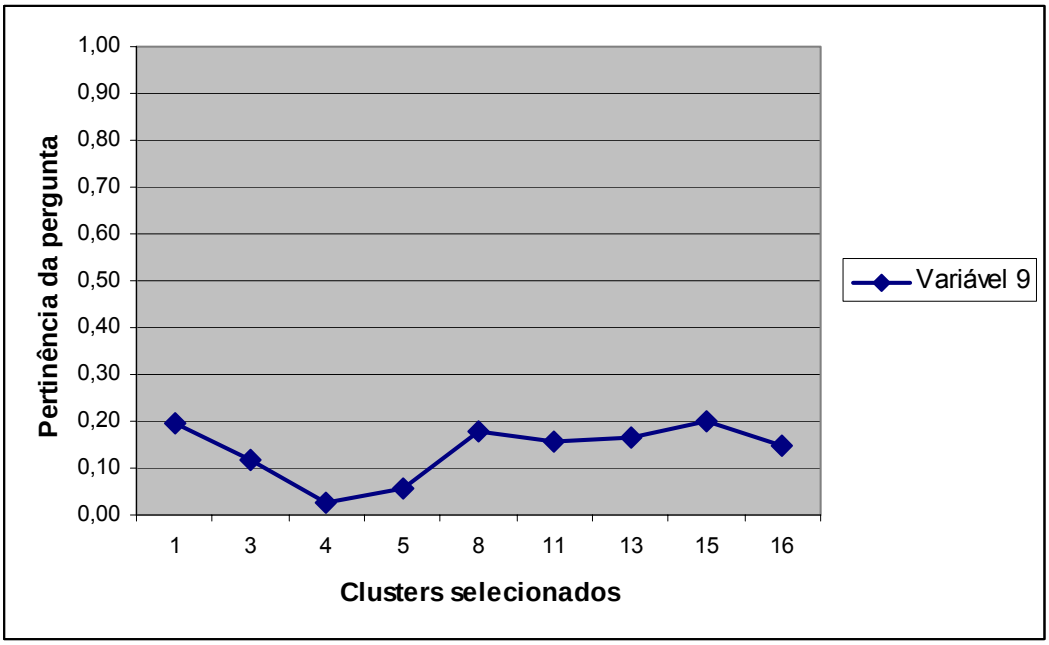


**Figura 25: Curva de pertinência para variável 9 dos clientes normais.**

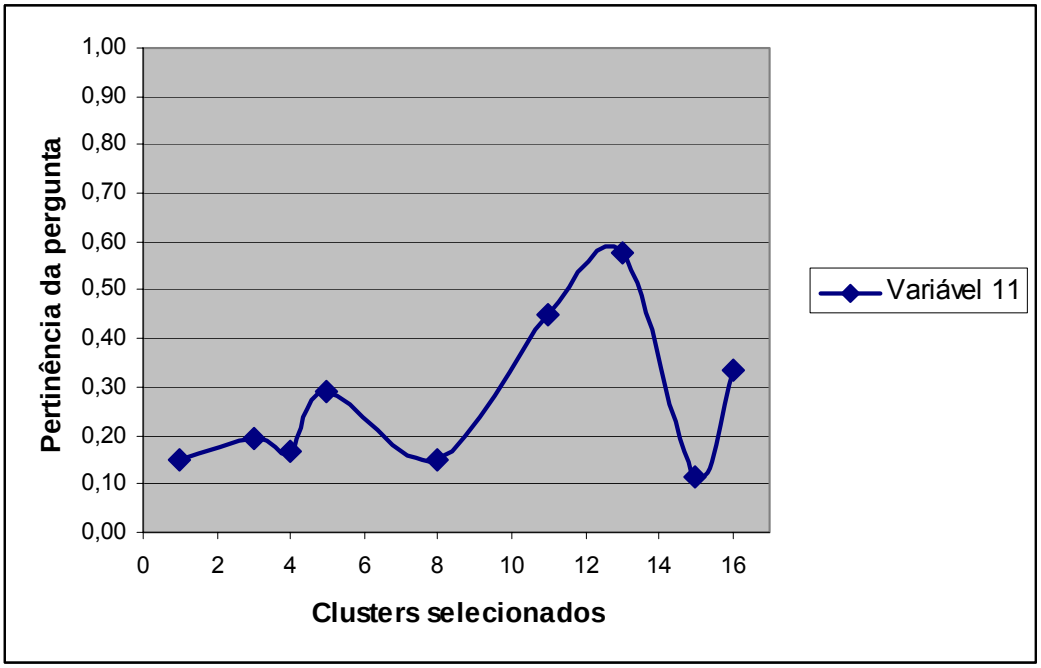


**Figura 26: Curva de pertinência para variável 31 dos clientes fraudulentos**

As curvas de pertinência para as variáveis escolhidas foram feitas para cada um dos “três tipos” de clientes (normais, inadimplentes e fraudulentos), para que se pudesse mais tarde ter uma comparação entre estas.

Definidas as variáveis pelo corte, selecionados os *clusters* a serem classificados e montadas as curvas de pertinências das variáveis selecionadas; o próximo passo foi escolher um operador fuzzy para se fazer a análise final das curvas pertinências selecionadas para cada um dos “três tipos” de clientes. O operador escolhido para definir o perfil dos clientes quanto a adimplência, fraude e inadimplência foi o **MAX** (máximo). Ou seja, o valor máximo entre todas as curvas de pertinência em um determinado *cluster* foi selecionado, gerando assim uma nova curva, denominada de **Max**.

A seguir são mostradas as curvas para cada uma das perguntas (variáveis) selecionadas para os clientes adimplentes, fraudulentos e inadimplentes, respectivamente. Além disso, é também mostrada a curva do operador máximo para cada um dos três “tipos” de consumidores de energia e que foi utilizada na solução deste problema. As curvas de máximo para cada um dos clientes se encontram nos mesmos gráficos que mostram as curvas de pertinência de cada um destes três segmentos de clientes com suas perguntas selecionadas pelo corte, o qual foi explicado anteriormente.

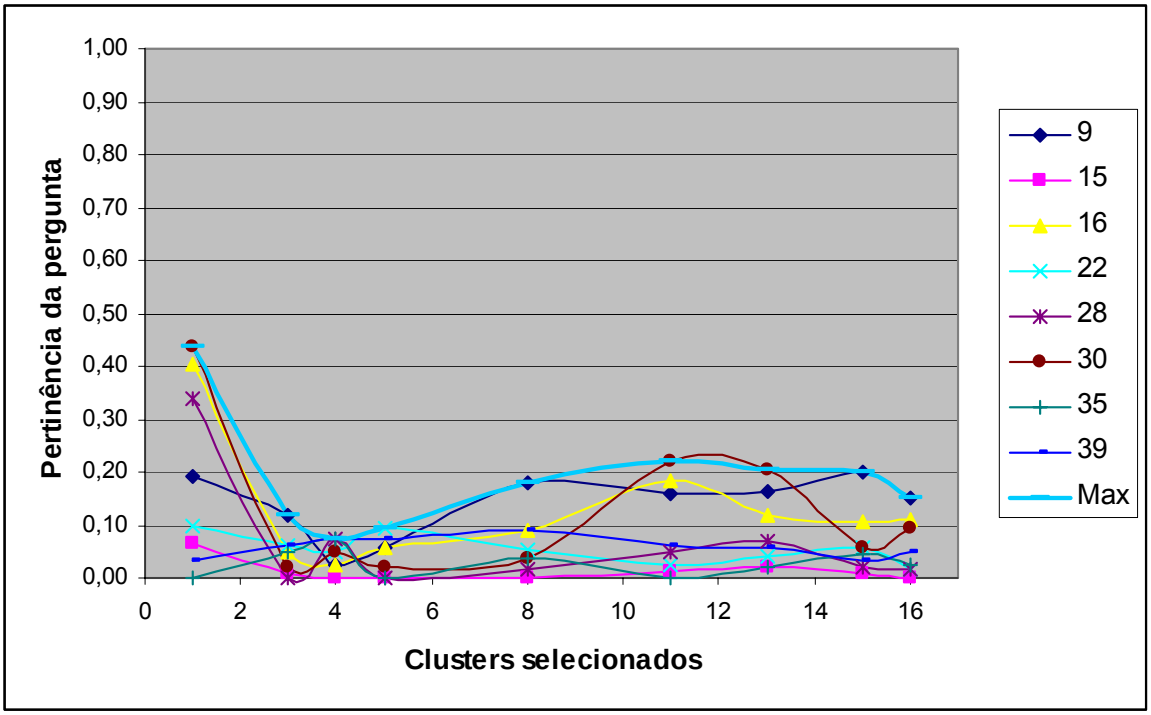


**Figura 27: Curvas de pertinência das respostas de clientes adimplentes às 8 perguntas selecionadas**

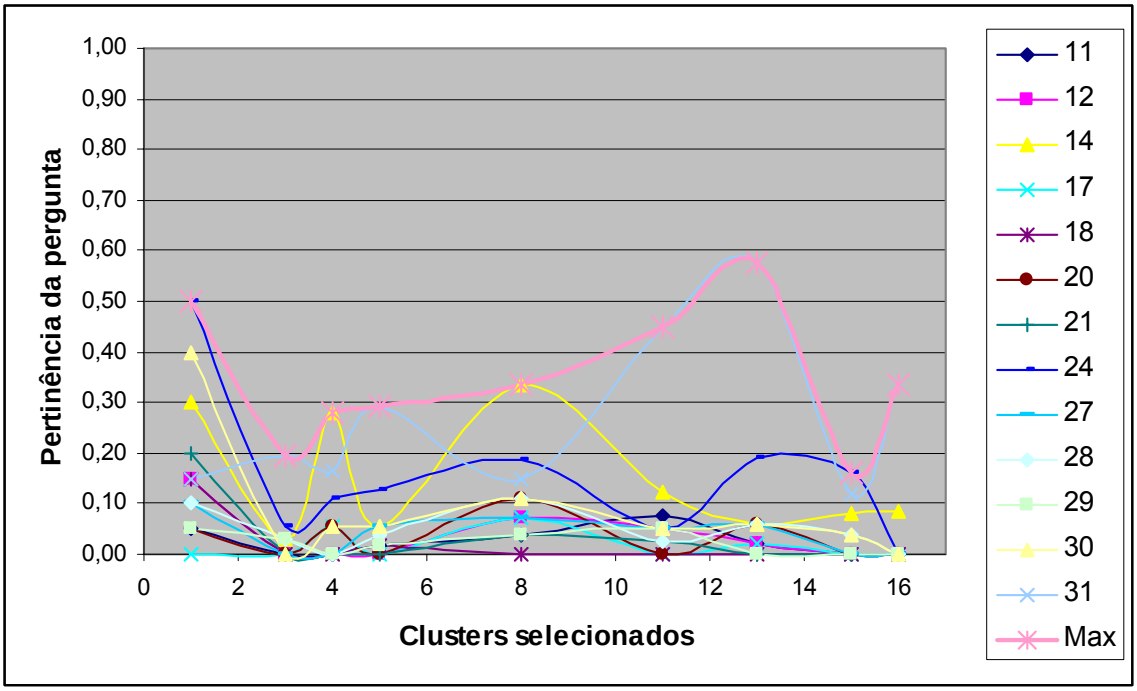


**Figura 28: Curvas de pertinência das respostas de clientes fraudulentos às 13 perguntas selecionadas**

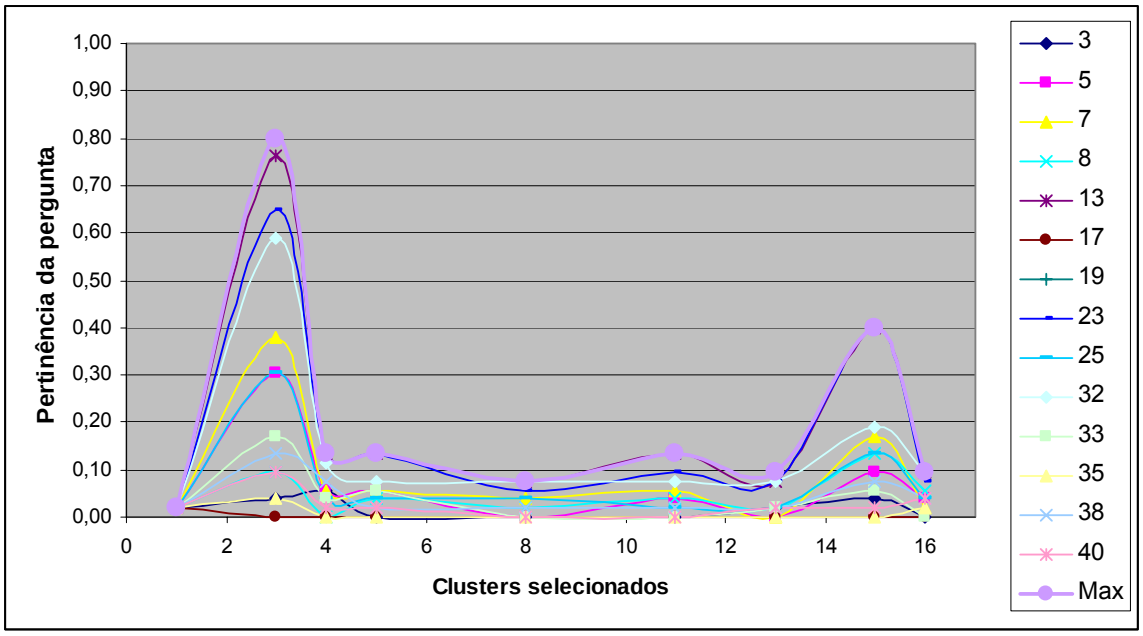


**Figura 29: Curvas de pertinência das respostas de clientes inadimplentes às 14 perguntas selecionadas**

A análise final é feita sobrepondo as curvas de máximo dos clientes normais, fraudulentos e inadimplentes que ressalta o diagnóstico de cada *status* dos clientes nos *clusters*, possibilitando realizar uma estratégia de se combater o problema. Poder-se-á ver nesta sobreposição de curvas que haverá *clusters* que a fraude e a inadimplência "saltarão aos olhos". Portanto, nestes *clusters* dever-se-á fazer o diagnóstico de quem são os clientes com rótulo de normal, pois estes têm todas as características de serem fraudulentos ou inadimplentes se seus *clusters* indicam um grande número de fraudes e inadimplência, respectivamente.

A figura 30 que se segue mostra a análise final para o exemplo proposto da "clusterização" dos clientes pesquisados.

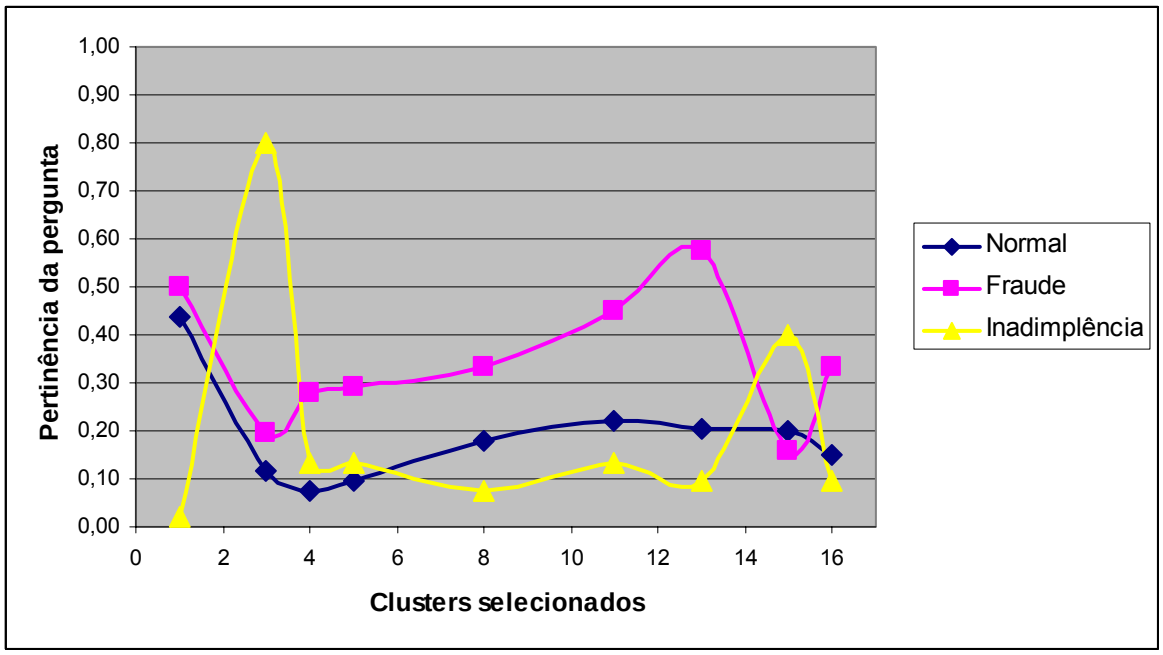


**Figura 30: Resultado final da classificação fuzzy dos *clusters*.**

Na figura 30 pode-se fazer as seguintes análises: os *clusters* 4, 5, 8, 11, 13 e 16 têm pouco problema de inadimplência; já grupo 3 apresenta um problema crônico de inadimplência, cujos clientes já são ou serão brevemente inadimplentes; o *cluster* 15 apresenta uma tendência menor do que o *cluster* 3 quanto à inadimplência; o grupo 1 apresenta sério problema de fraude que pode estar camuflado com um bom número de clientes adimplentes e para esse grupo não deve haver inadimplência a se considerar; assim como o *cluster* 1, os grupos 8, 11, 13 e 16 possuem um número de consumidores fraudulentos e um baixo índice de inadimplentes; os *clusters* 4, 5 e 8 apresentam basicamente a mesma proporção entre clientes normais, fraudulentos e inadimplentes e para esse grupo o importante primeiro é combater fraude.

Deve-se levar em consideração na hora da análise dos dados um fato que ficará mais a vista com a análise dos resultados. Ver-se-á no capítulo seguinte que os resultados para os clientes inadimplentes não são tão bons como para os consumidores fraudulentos. Neste trabalho isto se deveu, por ser o número de amostras da pesquisa de posses e hábitos de consumo para os clientes inadimplentes de apenas 102 clientes. Portanto, deve-se ter em mente que as curvas de respostas para os clientes inadimplentes indicam que em um determinado *cluster* há um número relevante de inadimplentes, no entanto, a chance de se conseguir detectá-lo é bem menor que para os clientes fraudulentos.