Aplicação de Técnicas de Mineração de Dados para detecção de Fraudes de Energia

J. L. Todesco[1] – tite@egc.ufsc.br
A. B. T. Morales[1] – aran@egc.ufsc.br
S. Rautenberg[1] - srautenberg@egc.ufsc.br
L. A. Garbelotto[2] – lagarbelotto@celesc.com.br
E. D. Athayde[2] – eduardodias@celesc.com.br

1. Laboratório de Engenharia do Conhecimento – Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento - Universidade Federal de Santa Catarina – (LEC/EGC/UFSC)
2. Centrais Elétricas de Santa Catarina S.A. – CELESC

***Resumo*- Fraude é uma das principais causas de perda de receita em muitos ramos de negócios. Empresas de distribuição de energia também sofrem com fraudes de seus consumidores. O objetivo deste trabalho é o de construir um sistema para identificação de possíveis fraudadores de energia elétrica, empregando o processo KDD. Para a etapa de mineração dos dados, foram utilizadas somente informações sobre o consumo e definiu-se uma medida chamada *score* acumulado. Trata-se de uma medida que calcula a diferença entre o consumo do mês atual e o consumo do mesmo mês no ano anterior para 12 meses, sendo acumulada no último mês. Consumidores com *score* acumulado acima de determinado valor limiar são candidatos à inspeção. Após ajuste do valor limiar, a taxa de acerto para o grupo de consumidores residenciais foi de 64% e a taxa média de acerto para o grupo de consumidores comerciais (padarias, lanchonetes e postos de gasolina) foi de 80%.**

***Palavras-chave*: Detecção de fraudes, Energia Elétrica e Knowledge Discovery in Database.**

I. INTRODUÇÃO

A energia elétrica é um bem de consumo cujo custo é calculado no montante consumido, ou seja, quanto mais se consome, mais se gasta. Quando a energia representa um alto custo no negócio ou na economia do lar, alguns consumidores partem para caminhos alternativos no desvio de energia, fazendo ponte no medidor ou invertendo a ligação na entrada da unidade de medição, por exemplo. Os caminhos da planejada "esperteza" são vários, todavia, qualquer um deles tem um nome: furto.

Segundo dados das Centrais Elétricas de Santa Catarina S.A. (CELESC), as perdas comerciais em 2003, 2004 e 2005 ficaram em 6,9%, 7,8% e 8%, respectivamente, [3]. Para se ter uma idéia do rombo mensal e da importância em maximizar a arrecadação, considerando-se uma perda comercial (desvios ou "gatos", fraudes e irregularidades na medição) na ordem de 3,2% e tendo-se um faturamento bruto mensal (dez/2005) de R$ 346.164.378,00, pode-se concluir que a perda de receita foi na ordem de R$ 11.077.260,10.

Ressalta-se ainda que desvios de energia a partir de postes, defeitos nas unidades de medição e falta de qualidade do cadastro de consumidores são exemplos de variáveis que corroboram perdas na arrecadação de uma empresa de distribuição de energia elétrica. Todavia, essas variáveis estão totalmente sob controle da área gerencial.

Mas, quando o desvio de energia ocorre dentro de uma unidade residencial ou comercial, este somente é detectado *in loco* por uma equipe de fiscalização. Nos anos de 2004 e 2005, para um contingente de 687.361 unidades de medição do grupo "B" (alimentados em baixa tensão) fiscalizadas, detectaram-se 6.393 fraudes. Houve uma receita adicional com as descobertas dessas fraudes no valor de R$ 24.142.413,94 (R$ 11.696.611,44 receita recuperada e R$ 12.445.802,50 receita agregada). Acredita-se que a quantidade de fraudadores pode ser ainda maior.

Contudo, para colaborar com a metodologia e os resultados, indicativos de desvio podem ser aferidos por um processo de investigação sazonal em bases de dados de consumidores e de faturas. Tal investigação direcionaria os esforços de uma equipe de fiscalização, a qual constataria *in loco* a existência ou não de fraude. Em virtude disso, a aplicação de tecnologia apropriada é encorajada. Nesse contexto, a Descoberta de Conhecimento em Bases de Dados, ou *Knowledge Discovery in Database* (KDD), é um campo multidisciplinar que compreende todo um processo não trivial de descoberta de conhecimento útil a partir de dados.

Atualmente, são encontrados na literatura científica trabalhos que relatam a aplicação do KDD na detecção de fraudes no consumo de energia elétrica. Dentre os trabalhos recentes, enumera-se:

- *Referência [6]*: empregaram *Rough Sets* para detecção de fraude de energia elétrica no Estado do Mato Grosso do Sul. Esse tipo de abordagem reduz a dimensionalidade da informação, permitindo

construir regras de predição de consumidores fraudadores. O objetivo da aplicação dessa abordagem é identificar, entre os consumidores, quais devem ser inspecionados. Para tal, foi utilizado um conjunto de atributos dos consumidores de maneira que o algoritmo pudesse assim classificá-los. O experimento empregou uma base inicial de 100 mil consumidores que, após a análise e a retirada de registros com ruídos, totalizou 40.600 consumidores, tendo 3.900 fraudadores e 36.700 não fraudadores. Após a aplicação do método, foram geradas 425 regras que aplicadas sobre um conjunto de testes identificaram um subconjunto de consumidores candidatos à inspeção. Desse subconjunto apenas 20% foram identificados como fraudadores.

- *Referência [8]*: empregaram o algoritmo C4.5 na geração de uma árvore de decisão para identificar consumidores candidatos à inspeção. Os autores do trabalho selecionaram 5 atributos dos consumidores para utilizar no algoritmo: classe do consumidor, atividade do consumidor, *electrical connection*, consumo mensal (12 meses) e consumo médio local. Os testes mostraram que, do conjunto identificado para inspeção, 40% eram de fato fraudadores. Contudo, muitos fraudadores não foram selecionados nos testes para inspeção (somente 25%). Isso demonstra que a identificação de possíveis fraudadores depende fortemente da qualidade e da quantidade de informações disponíveis dos consumidores.
- *Referência [9]*: aplicaram redes neurais artificiais e estatísticas para a detecção de perdas não técnicas (fraude) de energia elétrica na cidade de Sevilha, Espanha. A abordagem foi aplicada sobre um conjunto de consumidores residenciais e comerciais do ramo da hotelaria. Segundo os autores os resultados obtidos foram promissores, mas, após a aplicação das técnicas, foram identificados casos de fraudes em campo.

Neste artigo, expõem-se os esforços na construção de um modelo de detecção de fraudadores em consumidores de energia elétrica. Para tanto, foi desenvolvido um protótipo computacional baseado em KDD. O protótipo, aplicando métodos estatísticos e tratando a sazonalidade, é capaz de selecionar um subconjunto de consumidores candidatos à inspeção (consumidores com perfil potencial de fraudador) em dois segmentos: consumidores comerciais e consumidores residenciais.

Assim, como nos trabalhos correlatos citados, cabe ressaltar que o objetivo deste trabalho recai na perspectiva de planejar estratégias de auditoria para posterior inspeção *in loco*. Pontualmente, o trabalho é falho na garantia de que todos os fraudadores sejam inspecionados. Contudo, metodologicamente, o trabalho traz contribuições na otimização do tempo e dos recursos disponíveis à equipe de fiscalização.

Resultados preliminares apontaram que o processo estabelecido para a identificação de consumidores para inspeção mostrou-se bastante satisfatório. O protótipo desenvolvido atendeu às expectativas dos técnicos e está de acordo com as diretrizes da divisão e da empresa.

Para tanto, o artigo compreende, além da presente seção, seções reservadas aos assuntos pertinentes: fraude de energia elétrica, *Knowledge Discovery in Database*, processo de KDD empregado para consumidores comerciais e residenciais, utilização do processo de KDD, protótipo desenvolvido e resultados preliminares. Por fim, são apresentadas as conclusões e trabalhos futuros.

II. FRAUDES EM ENERGIA ELÉTRICA

O documento técnico do Comitê de Distribuição, [1], define as perdas técnicas como sendo "*a energia perdida no transporte, na transformação e nos equipamentos de medição de energia elétrica quando do fornecimento da mesma. As perdas comerciais são aquelas decorrentes da energia efetivamente entregue aos consumidores finais ou a outras concessionárias, mas não computadas na venda*".

Das perdas comerciais, como objeto relevante de estudo, têm-se as perdas decorrentes de fraude. Conceitua-se fraude no uso de energia elétrica o ato de má-fé praticado contra a empresa fornecedora de energia elétrica, o qual impede a correta medição e/ou faturamento, [5].

Segundo estimativas da CELESC, a empresa perde cerca de quatro milhões de reais por mês com ligações clandestinas, desvios e fraudes de energia elétrica. Ocorre uma perda em torno de 3% do consumo, ficando 0,5% nas favelas e 2,5% nos outros setores da sociedade, [10].

As irregularidades mais comuns ocorrem na adulteração de medidores e na instalação de circuito paralelo durante a construção do imóvel. Somente no ano de 2001, a CELESC precisou deslocar equipes em todas as agências regionais para recuperar o equivalente a R$ 1.967.716,90 em energia roubada. Foram 11.579.565 kwh de energia fraudada, [2].

O gerenciamento das perdas comerciais na CELESC é realizado através do acompanhamento mensal do total das perdas globais e da atuação sistemática no combate de fraudes e irregularidades na medição. O combate às perdas comerciais dá-se através de inspeções periódicas, informações fornecidas por leituristas, informações fornecidas por terceiros e relatórios de ocorrências do Sistema de Consumidores.

Assim como a fiscalização do grupo B, a fiscalização do grupo A é realizada pelas agências regionais. Reside no grupo B o foco de estudo do trabalho, no qual se detecta um subconjunto de candidatos com perfil potencialmente fraudador a serem inspecionados. Para tanto, o trabalho se respalda no processo de KDD, descrito brevemente a seguir.

III.KNOWLEDGE DISCOVERY IN DATABASE

Knowledge Discovery in Database (KDD) é um campo criado a partir do desenvolvimento de métodos e técnicas para dar "sentido" aos dados. O intuito geral do KDD é, a partir de um volumoso conjunto de dados de baixo nível, mapear/encontrar outras formas de representação acerca dos dados (informações) que sejam mais abstratas, compactas e úteis, [11]. Para tanto, KDD envolve uma intersecção de campos de pesquisa como Aprendizagem de Máquina, Reconhecimento de Padrões, Banco de Dados, Estatística, Inteligência Artificial, Aquisição de Conhecimento de Sistemas Especialistas, Visualização de Dados e Computação de Alto Desempenho.

Os objetivos do KDD são definidos na intenção de uso do sistema que implementa o processo de KDD. Genericamente, os objetivos podem ser distinguidos em *verificação* e *descoberta*, [11]. Na referência [4] detalha-se que a *verificação de hipóteses* é uma aproximação *top-down* na qual se verificam assertivas predefinidas nos dados. Ou seja, o especialista de domínio gera hipóteses que são aferidas perante os dados disponíveis. Já para a *descoberta de conhecimento*, os autores detalham-na como uma aproximação *bottom-up* que, a partir do conjunto de dados, tenta induzir algumas características relevantes. Nesse tipo de análise, os dados sugerem conjecturas sobre a sua semântica.

Independente da familiarização do domínio e do objetivo de aplicação, são passos do KDD (Figura 1), [11]: i) seleção de dados; ii) pré-processamento; ii) transformação de dados; iv) mineração de dados; e e) avaliação/interpretação de resultados.

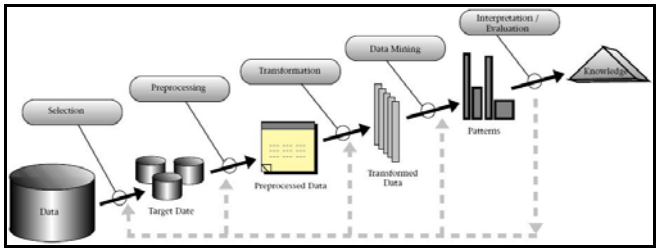

Figura 1: Passos do processo de KDD, [11]

A *seleção de dados* consiste em:

- definir as necessidades de dados para testar as hipóteses;
- localizar os dados, definir o modelo computacional para realizar as consultas às bases de dados (SQL's);
- definir quais são as informações relevantes, e selecionar os atributos relevantes para atingir os objetivos; e
- separar dados para treinamento e para testes das hipóteses levantadas.

No *pré-processamento,* são tomados alguns cuidados para garantir a qualidade da seleção de dados. Nessa fase, são despendidos esforços no sentido de:

- eliminar registros incompletos ou inconsistentes, bem como completar registros incompletos quando possível;
- remover colunas que não são pertinentes, que são redundantes, ou que contêm campos óbvios e desinteressantes;
- observar descrições de campo obscuras ou confusas. As descrições podem significar coisas diferentes que dependem da fonte. Por exemplo, a data de ordem pode significar a data em que a ordem foi enviada, carimbada, recebida ou teclada; e
- atentar para a atualidade de dados, que podem estar obsoletos. Os clientes podem, por exemplo, ter mudado de endereço.

A *transformação de dados* consiste em um conjunto de atividades que têm como objetivo gerar uma representação conveniente dos dados para os algoritmos de mineração. A transformação pode melhorar bastante o desempenho do modelo implementado durante o processo de KDD. Nesse sentido, os dados podem ser:

- agrupados em faixas;
- normalizados;
- criados, aplicando-se fórmulas matemáticas para tal; e
- compatibilizados com o formato requisitado pelos algoritmos de mineração de dados utilizados.

Durante a fase de *mineração de dados*, é construído um modelo que descreve os dados analisados. Isso é feito automaticamente através das técnicas e dos algoritmos escolhidos. Portanto, essa fase consiste em selecionar uma técnica, um algoritmo e uma ferramenta apropriada com base nas características dos dados selecionados e no objetivo do processo de KDD. Isto resulta na construção de um modelo a ser avaliado.

A fase de *avaliação* consiste em:

- selecionar e ordenar as descobertas interessantes;
- descartar as descobertas não interessantes;
- apresentar e visualizar os resultados;
- avaliar a precisão de um modelo, refinando sua compreensão e sua utilidade.

Por fim, na fase de *interpretação*, os resultados avaliados devem tomar consistência para serem utilizados nos processos de tomada de decisão de uma organização.

Na referência [4], comenta-se que o processo de KDD na tarefa de detecção de fraudes resume-se em construir modelos ou perfis do comportamento fraudulento, o qual pode ser útil em sistemas de suporte à decisão. Nesse sentido, a próxima seção descreve os procedimentos adotados no processo de KDD empregado na detecção de fraudes em consumo de energia elétrica, que vai ao encontro do planejamento de estratégias de auditoria (detecção de fraudes *a posteriori*).

IV. PROCESSO DE KDD EMPREGADO PARA CONSUMIDORES COMERCIAIS

O processo de KDD apresentado a seguir refere-se aos passos executados na construção do modelo de detecção de fraudes de energia elétrica de consumidores

comerciais e é subdividido segundo os passos do processo de KDD descritos na seção anterior.

Cabe ressaltar que o processo de KDD implementado tem o objetivo de *verificação de hipóteses*, por meio das quais, através da utilização de métodos estatísticos e de consultas apropriadas sobre uma base de dados, o conjunto de candidatos (consumidores comerciais suspeitos de fraude) é selecionado para inspeção.

A. Seleção de dados

O primeiro passo na construção do modelo de detecção foi a *seleção de dados*. Para tanto, foi disponibilizado um conjunto de 1.574.711 tuplas que contemplavam os dados descritos na TABELA I referentes a consumidores comerciais. Nos dados selecionados, encontravam-se consumidores fraudadores e não fraudadores.

TABELA I

DADOS DISPONIBILIZADOS DOS CONSUMIDORES (FRAUDADORES E NÃO FRAUDADORES)

Nome do campo	Descrição
nr_conta_cns	Número da conta do consumidor
dt_ref_fatura	Data de referência da fatura
qt_kwh_med	Kwh consumido no mês
nr_regional	Número da regional
Id_classe_cns	Identificador da classe do consumidor
id_ramo_atvdd_cns	Identificador do ramo de atividade do consumidor
id_sub_ramo_atvdd	Identificador do sub-ramo de atividade
po_tot_instalada	Potência total instalada
qt_dias_efet_fornt	Quantidade de dias de efetivo fornecimento no mês
id_tipo_cns	Identificador do tipo de conta
Dt_ligacao_cns	Data da ligação do consumidor
Nm_consumidor	Nome do consumidor
dt_ult_inspecao	Data da última inspeção
id_sub_classe_cns	Identificador da subclasse do consumidor
id_tipo_tributacao	Identificador do tipo de tributação
id_gpo_tensao_cns	Identificador do grupo de tensão
id_opcao_fatur	Identificador da opção de faturamento
cd_municipio	Código do município
nr_faturamento	Número do faturamento

B. Pré-processamento de dados

Como a seleção de dados foi disponibilizada em arquivos no formato CSV, na fase de *pré-processamento* organizou-se o conjunto de dados pela aplicação de estratégias de normalização pregadas na teoria de Banco de Dados. O modelo de dados resultante deste passo é ilustrado na Figura 2.

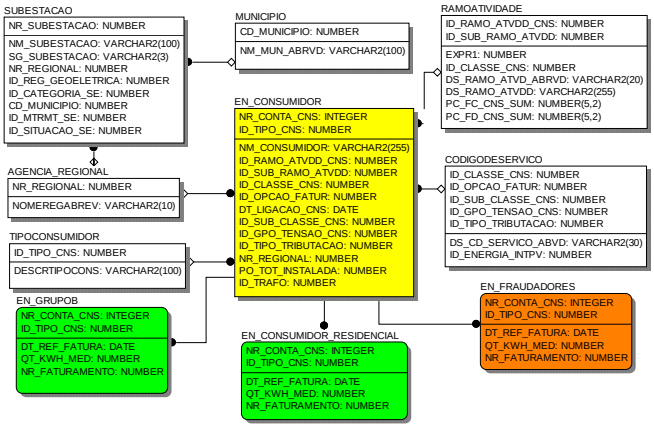

Figura 2: Modelo de dados de consumidores para o processo de KDD

Com o modelo de dados e consultas SQL apropriadas, foi possível abstrair alguns números de interesse para a continuação da construção do modelo: 91,31% dos consumidores eram caracterizados como não fraudadores e 8,69% como fraudadores, TABELA II. Como o número de fraudadores é muito inferior ao número de não fraudadores, a construção do modelo com o objetivo de verificar hipóteses e não descobrir conhecimento foi a opção mais viável.

TABELA II

DISTRIBUIÇÃO DOS DADOS DA AMOSTRA RECEBIDA

Tipo de consumidor	Quantidade
não fraudador	19.539 (91,31%)
Fraudador	1.860 (8,69%)
Total	21.399 (100%)

Analisando os dados organizados por ramo de atividade e contemplando os maiores subconjuntos com dados referentes aos dois tipos de consumidores, selecionaram-se todos os dados referentes aos seguintes ramos de atividade: Padaria, Postos de Combustíveis e Lanchonetes. Os subconjuntos selecionados são enumerados na TABELA III.

TABELA III

CONSUMIDORES COMERCIAIS SELECIONADOS PARA ANÁLISE

Ramo de atividade	Fraudador	Não fraudador
Lanchonetes	197	15.959
Padarias	51	2.098
Postos de Gasolina	38	1.482
Total	286	19.539

Após limpeza dos dados com ruídos, das 1.574.711 tuplas disponibilizadas inicialmente, somente 980.327 tuplas foram utilizadas para a construção do modelo.

A eliminação de colunas da seleção de dados também fez parte do passo de pré-processamento. Nesse sentido, são preservadas para a construção do modelo as colunas enumeradas na TABELA IV.

TABELA IV

DADOS SELECIONADOS PARA ANÁLISE

Nome do campo	Descrição
nr_conta_cns	Número da conta do consumidor
dt_ref_fatura	Data de referência da fatura
qt_kwh_med	Kwh consumido no mês
nr_regional	Número da regional
id_classe_cns	Identif. da classe do consumidor
id_ramo_atvdd_cns	Identif. do ramo de atividade do consumidor
id_sub_ramo_atvdd	Identif. do sub-ramo de atividade
id_tipo_cns	Identificador do tipo de conta
nm_consumidor	Nome do consumidor
id_sub_classe_cns	Identificador da subclasse do consumidor
cd_municipio	Código do município

C. Transformação de dados

Neste passo, traçou-se como objetivo a preservação da sazonalidade que ocorre em cada ramo e sub-ramo de atividade. Ou seja, para determinados sub-ramos de atividade, o consumo de energia elétrica no verão tipicamente é diferente do inverno. Estendendo-se essa observação, pode-se considerar também o outono e a primavera. Para garantir essa característica, foi criada uma medida chamada *score*. Trata-se de uma medida que calcula a diferença entre uma leitura (consumo) do mês atual com a leitura (consumo) do mesmo mês no ano anterior. A equação (1) mostra o cálculo da medida.

$$\text{score}_{\text{mensal}} = \text{inteiro}\left(\left(\frac{\text{leitura}_{\text{mês\_no\_ano\_anterior}} - \text{leitura}_{\text{mês\_atual}}}{\text{leitura}_{\text{mês\_no\_ano\_anterior}} + 1}\right) \times 10\right) \quad (1)$$

Na equação, o denominador tem a soma de 1 para evitar divisões por zero, que ocorre quando não houve consumo no mês do ano anterior. Para valores negativos, ou seja, o consumo do mês atual ser maior do que no ano anterior, define-se o *score* para o valor zero. Além disso, se o consumo no mês atual for menor do que 10% de aumento, o *score* é igual a zero. A TABELA V apresenta as faixas de valores do percentual de decréscimo para o estabelecimento do valor do *score*.

TABELA V

FAIXAS DE VALORES DOS SCORES

Score	Percentagem de decréscimo
0	[-∞, 10 %]
1	(10%, 20%]
2	(20%, 30%]
3	(30%, 40%]
4	(40%, 50%]
5	(50%, 60%]
6	(60%, 70%]
7	(70%, 80%]
8	(80%, 90%]
9	(90%, 100%]

A partir da definição do *score*, foi criado o *score* acumulado, que tem como objetivo acumular o valor do score dos 12 meses para o mês atual. A equação (2) apresenta o cálculo do score acumulado.

$$\text{score}_{\text{acumulado}} = \sum_{i=0}^{11} \text{score}_i, \quad \text{onde} \quad i = \text{variação retroativa do mês na janela anual} \quad (2)$$

Como pôde ser percebido, quanto maior o *score* acumulado durante o período de 12 meses, maior é o indicativo de fraude. Deve-se notar que esse *score*, por ser cumulativo, potencializa o indicativo de fraude mês a mês. Caso o mês em que o consumidor tenha uma tendência a ser fraudador seja uma anomalia (problema no medidor, por exemplo), no mês seguinte o *score* acumulado se estabiliza e tende a decrescer após 11 meses.

Assim, o modelo final de dados, após a transformação de dados, é incrementado em duas variáveis, ou seja, as colunas score_mensal e score_acumulado.

D. Mineração de dados

Para realizar a mineração de dados, no processo de KDD foram implementadas duas variáveis de auxílio na análise de consumidores suspeitos de fraude, sendo elas:

- *Corte*: na mineração de dados, os consumidores são tipificados como não suspeitos, indefinidos e suspeitos de fraude. Para cada tipo, um corte (faixa de valores) no domínio do *score* acumulado foi definido, e o corte $\in$ [0, 108]. Cabe ressaltar que quanto mais "relaxado" for o valor do corte, menos consumidores tornam-se suspeitos na verificação da hipótese.
- *Percentis*: como existe uma diversidade acentuada no consumo de energia elétrica em virtude do potencial econômico dos consumidores dentro de um sub-ramo de atividade (ex.: uma padaria na zona central × uma padaria num bairro pouco populoso), a inclusão de faixas de consumo através de percentis de consumo predefinidos se torna uma informação mais expressiva em detrimento ao uso de médias, desvio padrão e moda, por exemplo. Cabe ressaltar que o uso de percentis é feito por sub-ramo de

atividade, e foi uma evolução do uso inicial dos 1º e 3º quartis e da mediana. Empiricamente, foram preestabelecidos os percentis de consumo {5, 15, 25, 35, 45, 50, 55, 65, 75, 85, 95}. A necessidade de definição dos percentis dá-se pela necessidade de uma melhor análise de um consumidor em particular, ou seja, se o seu consumo é instável (graficamente, perpassa por diversos percentis em um curto espaço de tempo), isto é um indicativo acentuado de fraude (vide a Figura 9).

Cabe ainda ressaltar que o processo de mineração é promovido aplicando-se cortes e percentis num subconjunto de consumidores, tomando-se em consideração o mês de análise, a agência regional e o sub-ramo de atividade.

E. Avaliação

Uma vez selecionado o subconjunto da base de dados, os consumidores são ordenados decrescentemente pelo seu *score* acumulado, sendo evidenciados segundo os cortes previamente configurados. Ações de descarte de descobertas (consumidores que não devem ser inspecionados) e de visualizações pontuais são disponibilizadas na interface do protótipo implementado. Tais interfaces são apresentadas na seção VI. SIMULAÇÃO DO PROCESSO DE KDD.

F. Interpretação de resultados

O produto final da interpretação de resultados no processo de KDD é a composição de um relatório de inspeção. Um cabeçalho resumido do relatório também é evidenciado na seção VI. SIMULAÇÃO DO PROCESSO DE KDD.

V. PROCESSO DE KDD EMPREGADO PARA CONSUMIDORES RESIDENCIAIS

O processo de KDD empregado para consumidores residenciais seguiu a mesma metodologia empregada para consumidores comerciais, com diferenças apenas no tratamento de dados no passo de seleção e no pré-processamento. As diferenças são descritas a seguir.

A. Seleção de dados

Por ser considerável o número de consumidores residenciais da CELESC, optou-se pela seleção de consumidores residenciais da cidade de Florianópolis da agência regional Florianópolis. Para tanto, foi disponibilizado um conjunto de 4.263.538 tuplas.

B. Pré-processamento de dados

Como a seleção de dados foi disponibilizada em arquivos no formato CSV, na fase de *pré-processamento* organizou-se o conjunto de dados pela aplicação das mesmas estratégias de normalização citadas para o tratamento de consumidores comerciais.

Como resultados desse procedimento, são observados que os consumidores são divididos em 99,65% não fraudadores e apenas 0,35% fraudadores (TABELA VI). Como o número de fraudadores, tal qual evidenciado no conjunto de dados disponibilizados para análise dos consumidores comerciais, é muito inferior ao número de não fraudadores, a construção do modelo também ocorreu com o objetivo de verificar hipóteses, e não descobrir conhecimento.

TABELA VI

DISTRIBUIÇÃO DOS CONSUMIDORES RESIDENCIAIS

Tipo de consumidor	Quantidade
Não fraudador	359.938 (99,65%)
Fraudador	1.273 (0,35%)
Total	361.211 (100%)

Como os consumidores residenciais não têm sub-ramo de atividade como dimensão para caracterizá-los quanto ao consumo, buscou-se outra dimensão de interesse que conseguisse caracterizar os consumidores quanto à setorização de consumo. Nesse ponto, utilizou-se um trabalho de P&D que agrupava os transformadores (trafo) em dez famílias conforme o percentual de consumo por tipo de consumidor (residencial, industrial, comercial, rural, público e outros), [7]. Cabe ressaltar que a incorporação da dimensão "família" repercutiu na redução no volume de dados, como pode ser percebido na TABELA VII.

TABELA VII

TIPO DE TRAFOS E CONSUMIDORES LIGADOS

Tipo de trafo	Fraudador	Não fraudador
Residencial 1	43	10.153
Residencial 2	310	42.110
Residencial 3	277	46.282
Residencial 4	129	25.090
Residencial 5	47	16.741
Não Residencial 1	0	161
Não Residencial 2	62	7.394
Não Residencial 3	2	2.085
Não Residencial 4	0	20
Não Residencial 5	1	376
Total	871	150.412

Agregando a dimensão trafo ao modelo de dados e eliminando as dimensões referentes ao ramo de atividade, o modelo apropriado para análise de consumidores residenciais tem sua estrutura de dados final conforme evidenciado na TABELA VIII.

TABELA VIII

ESTRUTURA DE DADOS FINAL

Nome do campo	Descrição
nr_conta_cns	Número da conta do consumidor
dt_ref_fatura	Data de referência da fatura

qt_kwh_med	Kwh consumido no mês
nr_regional	Número da regional
id_classe_cns	Identificador da classe do consumidor
id_trafo	Identificador do trafo em que o consumidor está ligado
Id_tipo_cns	Identificador do tipo de conta
nm_consumidor	Nome do consumidor
id_sub_classe_cns	Identificador da subclasse do consumidor
cd_municipio	Código do município
score_mensal	Medida de desvio de consumo de um determinado mês em relação ao mês no ano anterior
Score_acumulado	Medida acumulativa anual do score mensal

Após limpeza dos dados com ruídos, das 4.263.538 tuplas disponibilizadas inicialmente, somente 4.002.454 tuplas foram utilizadas para a construção do modelo de consumidores residenciais.

VI. SIMULAÇÃO DO PROCESSO DE KDD

O protótipo implementado opera sobre duas visões distintas de consumidores: comerciais e residenciais. A Figura 3 mostra a interface de seleção das visões. Já a Figura 4 demonstra o início do processo iterativo de KDD. Ressalta-se que as interfaces demonstradas são passos no processo de KDD para consumidores comerciais. Como o processo de KDD para consumidores residenciais é análogo, este não será demonstrado.

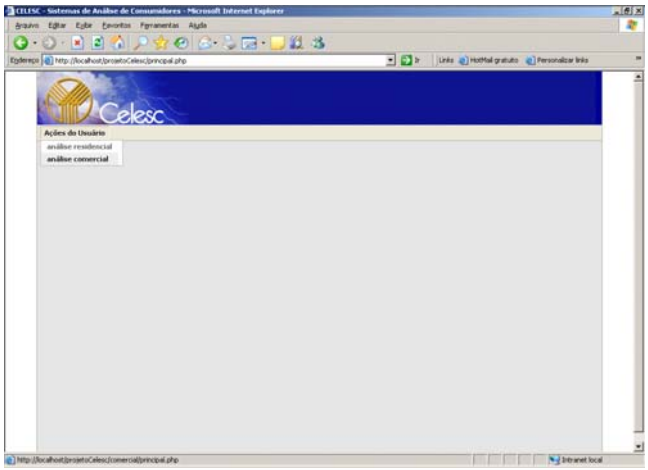

Figura 3: tela inicial

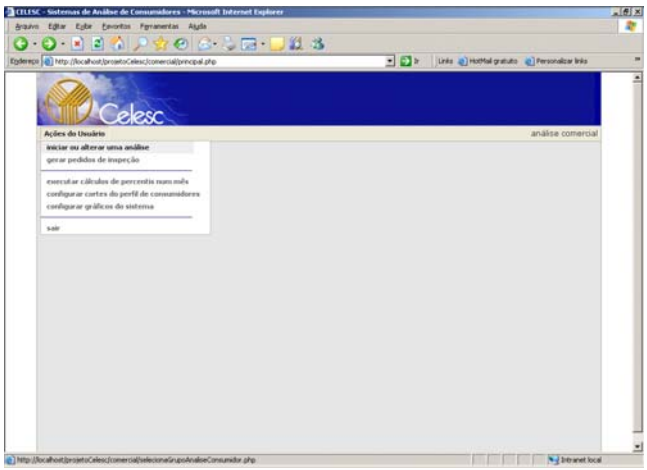

Figura 4: menu principal

Contudo, antes de executar qualquer processo de KDD o protótipo necessita estar previamente configurado, conforme a subseção a seguir.

A. Configurando o protótipo

Existem dois procedimentos de configuração para o protótipo. O primeiro (Figura 5) refere-se à configuração dos cortes do *score* acumulado, delimitando os consumidores não suspeitos, indefinidos e suspeitos de fraude no consumo de energia elétrica.

O segundo procedimento refere-se à forma de visualização dos gráficos de consumo de energia elétrica e do acompanhamento do comportamento do *score* acumulado de um consumidor em particular. Para tanto, conforme a Figura 6 é necessário informar quais percentis devem compor o gráfico de consumo e quantos meses de consumo são considerados.

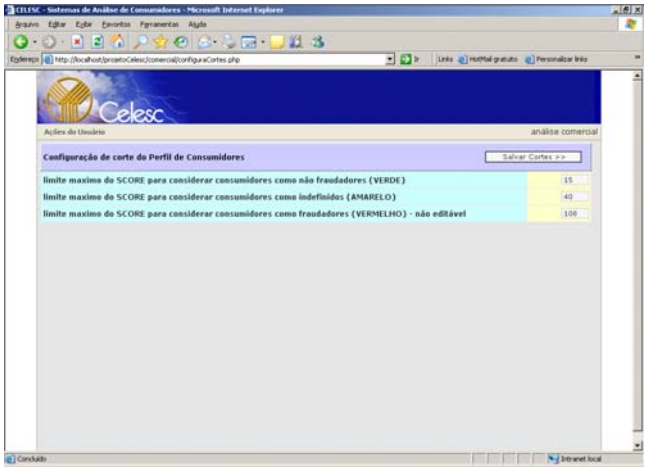

Figura 5: configuração dos cortes para o *score* acumulado x tipo consumidor

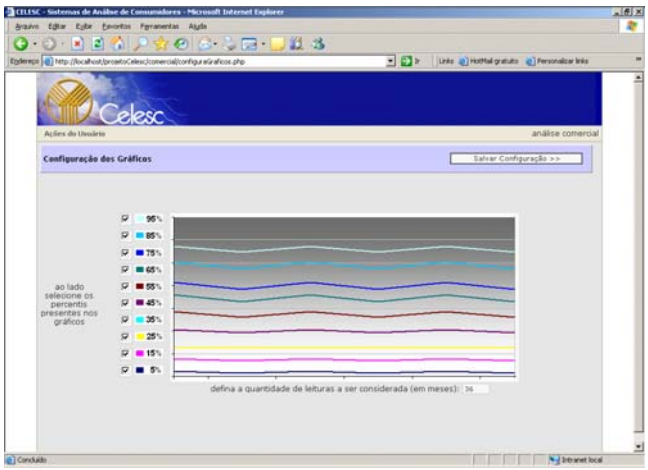

Figura 6: configuração dos gráficos

Sendo o protótipo configurado, pode-se executar o processo de KDD.

B. Executando um teste de hipótese

O primeiro passo no processo de KDD propriamente dito é selecionar um subconjunto de consumidores ao informar o mês de análise, a agência regional e o ramo de atividade para os consumidores comercias, (Figura 7).

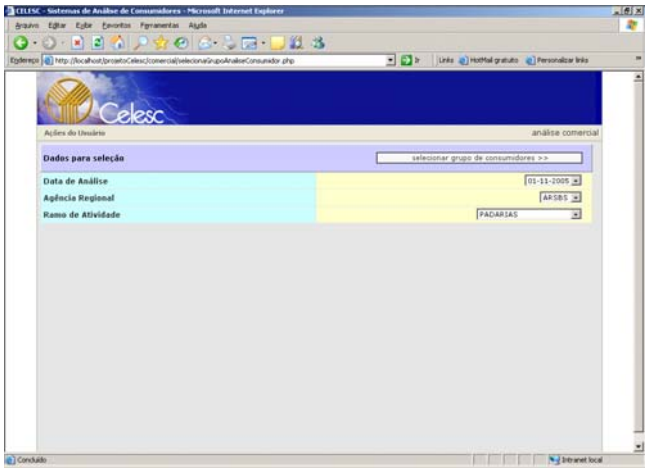

Figura 7: tela de seleção dos consumidores a serem analisados

Neste passo reside a diferença principal para a efetivação da análise dos consumidores residenciais. Nesse sentido, são informados o mês de análise, a agência regional, a família do trafo e o trafo de ligação.

Diante da seleção, o protótipo faz um *ranking* através do *score* acumulado previamente configurado dos consumidores potencialmente suspeitos de fraude. Nesse sentido, os consumidores suspeitos são destacados na zona vermelha da tela, os consumidores indefinidos, na zona amarela, e os consumidores não suspeitos, na zona verde, (Figura 8). A Figura 9 mostra um consumidor particular, apresenta o consumo de forma histórica e o *score* acumulado no período de análise. Essa visualização colabora no processo decisório de escolha do consumidor a ser inspecionado.

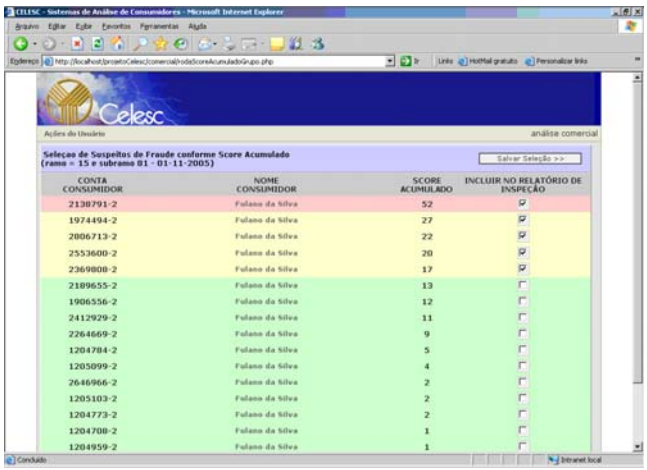

Figura 8: tela de seleção de consumidores para inspeção

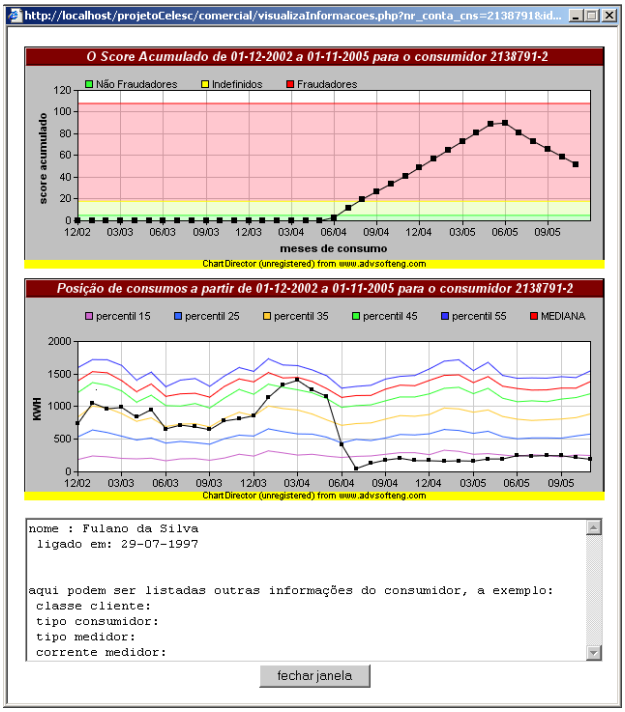

Figura 9: visualização do consumo de determinado consumidor

A partir da Figura 8, os consumidores com *score* acumulado acima de determinado limiar aparecem na faixa vermelha e podem ser marcados para inspeção. A escolha dos limiares (faixa verde até 5; faixa amarela entre 5 e 18; e faixa vermelha acima de 18) foi baseada em estudos com a massa de consumidores fraudadores, utilizando estudos anteriores da empresa. A Figura 10 mostra um conjunto de consumidores escolhidos para inspeção, chamado de relatório de inspeção. Em caso de uma dúvida pontual, o especialista pode visualizar o consumo histórico e o comportamento do *score* acumulado de um consumidor em particular (Figura 9), colaborando no processo decisório.

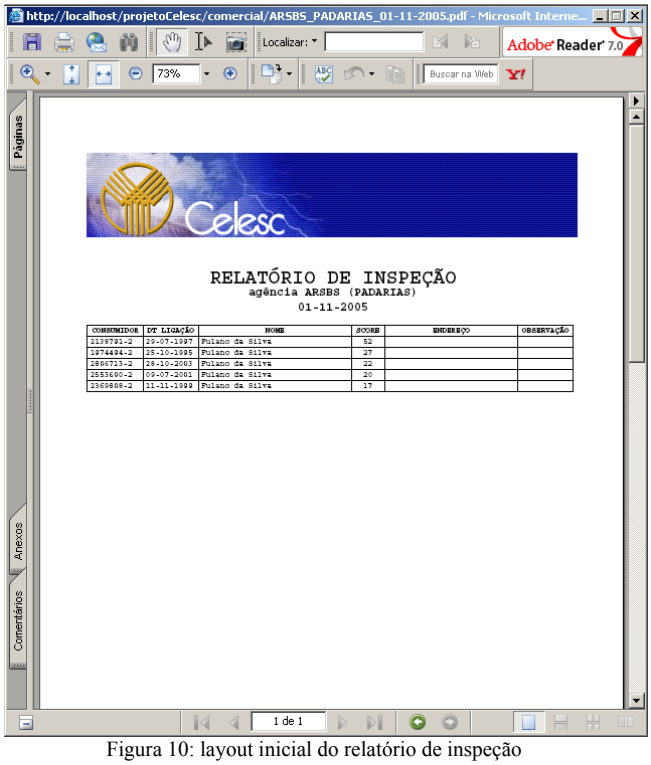

Figura 10: layout inicial do relatório de inspeção

VII. RESULTADOS PRELIMINARES

Para a realização dos testes, foi identificado primeiramente um conjunto de consumidores residenciais fraudadores que pudessem ser agrupados em uma das dez famílias de trafos. Posteriormente, foi calculado o *score* acumulado desses consumidores dentro do período posterior a 24 meses, uma vez que para o cálculo dessa medida são necessários no mínimo essa quantia de meses para apurar a sazonalidade de consumo para um ano corrente. Assim, para o conjunto de 30 meses disponíveis para os testes, foi calculado o *score* acumulado para os últimos seis meses. Num teste realizado, como regra considerou-se que consumidores com *score* acumulado superior ao limiar 18 eram classificados como fraudador. Aplicando o teste numa amostra de dados que contava com 11 consumidores fraudadores, 7 destes foram identificados, ou seja, uma taxa de acerto de 63,64%. Ressalta-se que uma nova amostra está sendo preparada para a realização de um teste em campo.

Para os consumidores comerciais, foi considerada uma amostra de dados que continha 286 consumidores fraudadores (51 padarias, 197 lanchonetes e 38 postos de gasolina). Dessa amostra, 230 consumidores tiveram *score* acumulado acima de 18, indicando serem candidatos à inspeção. Essa relação indica uma taxa de acerto de 80,42%, sendo 80,39% para as padarias, 83,76 para as lanchonetes e 63,16 para os postos de gasolina.

Percebe-se que há os casos de falsos fraudadores, ou seja, possíveis problemas em equipamentos, mudanças de ramo de atividade, estabelecimentos fechados por determinados períodos que possam suscitar indicações errôneas. Mas, é importante destacar que o protótipo apresenta a relação de consumidores a serem inspecionados, e não a indicação de fraude. A fraude só poderá ser comprovada *in loco*. Além disso, o protótipo disponibiliza uma janela (veja Figura 9), com informações correntes de um consumidor em particular, o que permite que sejam identificadas visitas por problemas no equipamento, mudança de ramo ou outras observações registradas.

VIII. CONCLUSÃO

As técnicas de mineração de dados vêm sendo empregadas com sucesso no domínio da extração e descoberta em diversos setores produtivos e para variados fins.

Neste artigo, foram apresentados os aspectos relacionados ao processo de KDD para a identificação de possíveis fraudadores de energia elétrica. Foram resgatados os elementos teóricos das etapas empregadas, e mostrada a metodologia empregada no processo, a medida usada para caracterização dos consumidores e o protótipo sugerido.

Observou-se que o cadastro de consumidores residenciais é bastante pobre, ou seja, informações importantes como a área total construída, o número de moradores e padrão do imóvel não estão disponíveis. Essas informações poderiam ser incorporadas ao cadastro de maneira incremental, sendo uma ação obrigatória para as novas ligações e para os consumidores antigos, devendo-se estabelecer um prazo para tal. Certas informações poderiam ser capturadas no ato da medição pelo próprio leiturista.

No cadastro de consumidores comerciais, também se percebe que informações sobre o ramo de atividade e a classe de consumidores não são precisas, ou melhor, não são atualizadas. Isso ocorre quando em um determinado estabelecimento há uma sala comercial e é feita uma mudança no ramo de atividade. Essa mudança por vezes não é registrada, o que faz com que o cadastro não seja confiável.

O uso dos dados de consumo mensal disponíveis, tanto comercial quanto residencial, já foi possível obter resultados satisfatórios para o estudo, embora inspeções em campo ainda devam ser feitas. O protótipo apresentado poderá ainda ser incrementado com outras informações do consumidor, mas o mais importante seria incrementar os dados cadastrais dos consumidores. Com certeza, com informações cadastrais mais ricas seria possível fazer mais cruzamentos, estabelecer padrões de consumos e obter mais precisão na identificação de consumidores para a inspeção.

IX. AGRADECIMENTOS

Às Centrais Elétricas de Santa Catarina - CELESC, por apoiarem esta pesquisa a partir do Programa P&D e por possibilitarem a sua abertura para a condução da pesquisa, e também à Universidade Federal de Santa Catarina - UFSC, através do departamento de Informática.

X. REFERÊNCIAS BIBLIOGRÁFICAS

[1] ABRADEE, "Perdas Comerciais". Associação Brasileira de Distribuidores de Energia Elétrica. Brasília, BR, Relatório Técnico. ABRADEE-08.05. 1997.

[2] BEC, "Boletim estatístico comercial de dezembro de 2001", CELESC, Florianópolis, SC, Boletim. 2001.

[3] CELESC/DPSC/DVMD, "Relatório Técnico da Divisão de Medição", Divisão de Medição – CELESC, Florianópolis, SC, Relatório Técnico. 2005.

[4] F. Bonchi, F. Giannotti, G. Mainetto, D. Pedreschi, "A classification-based methodology for planning audit strategies in fraud detection" in Proc. 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, United State, 1999.

[5] J. A. de BETTIO, "Constatação de procedimentos irregulares e deficiência no medidor ou demais equipamentos de medição". CELESC, Florianópolis, SC, Apostila de Curso. 2001.

[6] J. E. Cabral, J. O. P. Pinto, E. M. Gontijo, J. Reis Filho, "Fraud Detection in Electrical Energy Consumers Using Rough Sets" in Proc. 2004 IEEE International Conference on Systems, Man and Cybernetics, Hague, Netherlands, 2004.

[7] J. L. Todesco, A. Morales, R. C. Pacheco, F. J. S. Pimentel, "Previsão de Demanda de Energia usando Famílias de Circuitos e Rede Neural Artificial" in Proc. 8th Brazilian Symposium on Neural Networks, São Luiz, Brazil, 2004.

[8] J. Reis Filho, E. M. Gontijo, E. Mazina, J. E. Cabral, J. O. P. Pinto, A. C. Delaíba, "Fraud Identification In Electricity Company Costumers Using Decision Tree" in Proc. 2004 IEEE International Conference on Systems, Man and Cybernetics, Hague, Netherlands, 2004.

[9] I. Monedero, F. Biscarri, C. León, J. Biscarri, R. Millán, "MIDAS: Detection of Non-technical Losses in Electrical Consumption Using Neural Networks and Statistical Techniques", Lecture Notes in Computer Science, vol. 3984, pp. 725-734, May. 2006.

[10] N. Pavei, "Gatos fraudam SC em R$ 4 milhões", Jornal Diário Catarinense - Caderno de Economia, pp. 18, Florianópolis, Brazil, 17 jun. 2001.

[11] U. Fayyad, G. Piatetsky-Shapiro, P. Smith, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, vol. 17(3): Fall, pp. 37-54, 1996.