MINERAÇÃO DE DADOS SOBRE PESQUISA DE HÁBITOS DE CONSUMO DE ENERGIA NO SUL DO BRASIL

COSTA, Francisco Daniel de Oliveira
GREEF, Ana Carolina
TSUNODA, Denise Fukumi

RESUMO
Informações e decisões baseiam-se em dados sobre contextos sociais e organizacionais, armazenados em bases, desde que organizados e alinhados às necessidades informacionais nos mesmos contextos. O *data mining* contribui para a obtenção dos subsídios ao processo decisório, ao passo que permite a aplicação de algoritmos de clusterização, associação e classificação a bases de dados – conforme a necessidade previamente estabelecida, originando regras inerentes às situações a que eles se referem. Descreve-se então a aplicação do Processo de Descoberta de Conhecimento em Base de Dados (KDD), do qual a mineração de dados faz parte, a uma base de dados relativa à posse de eletrodomésticos e hábitos de consumo por cidadãos brasileiros da Regão Sul do país, oriundos do Programa Nacional de Conservação de Energia Elétrica (Procel). O estudo foi proposto, em caráter exploratório, por pesquisador em eficiência energética, com o intuito de verificar quais resultados poderiam ser extraídos da base por meio do *data mining*, sem necessariamente basear-se em atributos meta. Após o levantamento bibliográfico pertinente a regras de associação, ao sistema Suporte-Confiança e ao algoritmo *Apriori*, apresenta-se a metodologia conduzida para obtenção dos resultados do estudo: a operacionalização das etapas de seleção, limpeza, integração e transformação da base de dados. Em seguida, resume-se a etapa de mineração de dados em si, realizada com auxilio do software WEKA, com base no algoritmo *Apriori* – associativo. A intepretação das regras geradas aponta relações limitadas entre as variáveis (dados) que compõem a base, como a caracterização das residências e a ausência de problemas de consumo de energia nestas, além de ressaltar deficiências nos processos de coleta e tabulação dos dados, inclusive do instrumento utilizado para a pesquisa. Enfim, os dados originais mostram-se excessivamente dispersos para a obtenção de regras completas sobre o contexto estudado, resultando em contribuições mínimas para processos decisórios a respeito da distribuição, da valoração e do consumo de energia no Sul do Brasil, no âmbito do Procel, por parte de governos, concessionárias e até mesmo consumidores.

PALAVRAS-CHAVE
Procel. Regras de associação. *Apriori*.

1. INTRODUÇÃO

A transformação de dados em informações úteis para resolução de problemas e decisões, em qualquer contexto, é facilitada à medida que o conhecimento dos indivíduos no mesmo ambiente, e as Tecnologias de Informação e Comunicação, voltam-se aos mesmos objetivos. Às pessoas cabe o planejamento, a estruturação e o controle de coleta, tratamento, armazenamento, recuperação, uso e descarte ou realimentação de dados, bem como a interpretação de resultados por eles compostos, ao passo que à tecnologia são atribuídos os papéis de auxiliar em cada uma dessas tarefas, e de integradora do processo por elas composto.

A Descoberta de Conhecimento em Bases de Dados (KDD), contribui para que esses insumos tornem-se informações e interpretados por seus usuários, gerando novos conhecimentos acerca do contexto em questão. Permeado pela ação desse público em cada etapa, o KDD depende da mesma para gerar resultados relevantes, seja de classificação, de clusterização ou de associação entre dados.
Apresenta-se as características dessa última possibilidade, componentes de regras de associação entre dados, o esquema suporte-confiança e o algoritmo *Apriori*, que se utiliza de parâmetros de suporte e de confiança para extrair informações de bases de dados. Seus resultados, regras no formato *"Se-então"*, em várias combinações dos dados em questão, como antecedentes e consequentes.
A pesquisa não conta com cliente propriamente dito, sendo realizada a título de experiência, como sugestão de pesquisador em eficiência energética, e fornecedor da base de dados utilizada: registros de moradores de todo o Brasil, relativos a um questionário sobre posse de eletrodomésticos e hábitos de consumo de energia, aplicado pelo Programa Nacional de Conservação de Energia Elétrica (Procel).

2. MINERAÇÃO DE DADOS E REGRAS DE ASSOCIAÇÃO

A análise e o processamento de dados visando obter informações e conhecimentos úteis, de forma automatizada, realizados no âmbito da mineração de dados (*data mining*), são tidos como tarefa essencial para os processos decisórios em quaisquer ambientes. Essa mineração consiste na extração consciente de informações implícitas, previamente desconhecidas e de uso potencial em organizações, a partir de bases de dados. (LEE; SIAU, 2001, tradução nossa).
Contemplado pelo Processo de Descoberta de Conhecimento em Bases de Dados (KDD), o *data mining* representa a etapa de aplicação de algoritmos de associação, clusterização ou classificação aos dados dos quais se pretende extrair as referidas informações, precedida por: seleção dos dados a serem minerados; limpeza dos mesmos para eliminação de ruídos, dados irrelevantes e duplicidades; integração com outras bases de dados, com objetivo de agregar abrangência e confiabilidade ao resultado esperado; e transformação de valores, por exemplo, em categorias. Para compreensão dos resultados da mineração de dados, e uso desses em processos decisórios, a etapa de interpretação encerra o KDD. (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).
As decisões advindas desse processo são viabilizadas mediante o entendimento do propósito estabelecido para o *data mining*, e suportam a melhoria contínua em atividades organizacionais, no âmbito privado, público e terceiro setor. (ZHANG; ZHOU, 2004, tradução nossa).

2.1 Regras de associação

Regras de associação, conforme Zhang e Wu (2011, tradução nossa), têm papel fundamental na descoberta de conhecimento, devido à sua representação simples e facilidade de compreensão. Utilizadas comumente na identificação de comportamentos de usuários / clientes, sua mineração visa descobrir correlações de interesse entre subconjuntos de atributos de bases de dados, cujos valores têm condições que ocorrem de maneira conjunta com certa frequência.
Cada regra de associação obedece ao formato $X \rightarrow Y$ (sendo Y pelo menos uma variável), representando, por exemplo, que se um cliente adquire o produto ou tem a característica X, comporta-se de maneira Y. Essa estrutura assemelha-se a regras lógicas no estilo *"Se-então"*, diferenciando-se por sua característica probabilística por natureza. A validação de regras de

associação obtidas por meio de *data mining* se dá mediante a correspondência dessas a um suporte ($X \ U \ Y$) e a uma confiança ($X \rightarrow Y$) necessariamente iguais ou superiores aos estabelecidos pelo cliente ou pesquisador do processo. (ZHANG; WU, 2011, tradução nossa).
Essa prática visa a avaliação e compreensão prévia de características de novos registros incorporados à base de dados minerada. Apesar da simplicidade da extração de regras por meio de métodos associativos entre atributos, essa tarefa perpassa desafios inerentes ao processo de descoberta de conhecimento, entre eles:

a) a mineração em bases de dados extensas, com variáveis múltiplas e heterogêneas, cujos dados podem mostrar-se incompletos e não padronizados;
b) a identificação de metas e teorias para o processo de mineração dos referidos dados, que deve adequar-se às necessidades do cliente em questão ou do pesquisador, e basear-se em critérios previamente estabelecidos de seleção de regras;
c) o tratamento de restrições, como a precisão à qual as regras devem obedecer e limites temporais do KDD.

Os tópicos ressaltam a importância das fases iniciais desse último processo, ou seja: a seleção de dados relevantes para a mineração e, entre eles, de atributo(s) meta; limpeza e transformação do conteúdo da base de dados em questão, padronizando-os e obtendo formatos cuja interpretação por algoritmo é facilitada.

2.2 Suporte e confiança

As regras de natureza associativa são extraídas de bases de dados por meio de algoritmos que compreendem a testagem das ocorrências conjuntas de atributos (colunas) em cada registro (linhas) do composto de dados em questão:
Cada um dos registros componentes da base de dados (normalmente relativos a transações) representa um subconjunto da primeira, cujo tamanho ou comprimento varia conforme o número de atributos existentes na base, e aquele associado entre si no registro em questão. As ocorrências individuais e associadas entre atributos, por sua vez, são representadas por uma medida estatística, o suporte, que refere-se à proporção de registros da base que contêm o referido subconjunto – aos quais a regra se aplica. Por exemplo, o suporte de $X \rightarrow Y$ configura a frequência de $XY \ / \ (X \cup Y)$, em que X e Y representam subconjuntos que ocorrem simultaneamente em registros de dados. (WITTEN; FRANK, 2005; ZHANG; WU, 2011, tradução nossa).
Conforme Zhang e Wu (2011), as regras validadas são descritas no formato *"Se-então"* ($X \rightarrow Y$), sendo testadas todas as possibilidades de distribuição de seus componentes como antecedente e consequente. A confiança da regra representa a razão do número de registros que contemplam todos os seus componentes, sobre aqueles que contemplam o antecedente $(p(XY)/p(X))$, ou seja, a "força" da regra em si.
Os valores mínimos de suporte e de confiança para validação das regras identificadas pelo algoritmo em questão, são definidos pelo usuário / cliente ou pesquisador que realiza o processo de mineração.

2.3 Algoritmo *Apriori*

Entre os algoritmos utilizados para extração de regras de associação, de bases de dados, o *Apriori* é frequentemente utilizado, fato que, segundo Zhang e Wu (2011, tradução nossa), se deve aos seguintes parâmetros por ele obedecidos:

a) subconjuntos de dados da base em questão somente são considerados frequentes caso seus componentes também ocorram com frequência, individualmente;
b) os primeiros subconjuntos *S* frequentes identificados são utilizados para identificação de candidatos a *S*, também frequentes, buscados na base de dados.

Esses, por sua vez, são testados quanto a suporte e confiança, e validados ou não mediante correspondência aos respectivos valores;

c) esse processo de identificação é iterativo, até a inviabilidade de extrair *S* da base de dados em sua totalidade*;*
d) todos os atributos da base são contemplados na identificação de regras, ou seja, inexiste um atributo meta, tido como "objetivo" ou consequente obrigatório de cada regra.

Desse modo, as operações realizadas pelo *Apriori* para extração de regras resumem-se a: cálculo do suporte dos valores compreendidos na base de dados, individualmente e constituindo subconjuntos *S* frequentes. Existindo um conjunto *C* de subconjuntos *S*, frequentes, *Vn* representam valores de *Sn*, inicialmente verificados como também frequentes. Para geração de novos *Sn*, são verificados valores não vazios nele compreendidos, candidatos a frequentes, gerando todos os subconjuntos frequentes possíveis na base em questão. Nesse processo, cada *S* deve atender ao suporte e à confiança mínimos, estabelecidos pelo usuário / cliente ou pesquisador do processo.

Subconjuntos de maior pertinência em relação a ambos os parâmetros são, enfim, selecionados como fundamento para extração de regras válidas do conjunto total de dados. O algoritmo, enfim, constitui e retorna as regras validadas a partir dos subconjuntos *S*. (AGRAWAL; SRIKANT, 1994, tradução nossa).

3. EFICIÊNCIA ENERGÉTICA NO BRASIL: O PROJETO PROCEL

O Programa Nacional de Conservação de Energia Elétrica (Procel), criado em dezembro de 1985, está subordinado à Eletrobrás, empresa estatal de capital aberto, controlada pelo Ministério de Minas e Energia, do Brasil. Seu objetivo é promover a racionalização da produção e do consumo de energia elétrica, minimizando desperdícios e reduzindo custos nesse Setor, com aporte a metas consideradas essenciais: desenvolvimento tecnológico, eficiência econômica, redução nas perdas técnicas das concessionárias, mudança no comportamento dos cidadãos e redução de impactos ambientais. (PROGRAMA NACIONAL DE CONSERVAÇÃO DE ENERGIA ELÉTRICA, 2011).

São públicos interessados em ações no âmbito do Programa, portanto: usuários de energia elétrica, concessionárias das redes de energia em todo o país, e governos reguladores das mesmas concessões.

Na atual gestão, uma das principais metas do Procel tem sido a redução de 10% em perdas técnicas na transmissão e distribuição de energia entre concessionárias. Segundo a Conferência Das Nações Unidas Sobre Mudança Do Clima (2011), uma economia de aproximadamente 2.158 GWh de energia é atribuída à mesma meta, até o ano de 2009. Isso por meio de atividades como a adoção do Selo Procel de eficiência energética em eletrodomésticos, gerando conscientização da população sobre o consumo de cada equipamento, e estimulando a indústria a melhorar o desempenho energético de seus produtos. (PROGRAMA NACIONAL DE CONSERVAÇÃO DE ENERGIA ELÉTRICA, 2011).

4. PROCEDIMENTOS METODOLÓGICOS: PROCESSO DE DESCOBERTA DE CONHECIMENTO (KDD) SOBRE BASE DE DADOS DO PROCEL – REGIÃO SUL

Este estudo, em caráter exploratório, foi realizado com uma base de dados sobre posse de eletrodomésticos e hábitos de consumo por cidadãos brasileiros, realizada em intervalos de cinco anos, no âmbito do Procel. Uma vez que os resultados mais recentes da pesquisa (ano de 2010) não foram divulgados até então, o estudo baseou-se nos dados do ano de 2005.

Aplicou-se o algoritmo *Apriori* à base de dados, considerando a ausência de questão de pesquisa para este estudo e a proposta de descoberta de regras embutidas nos dados, cuja noção prévia seria inviável.
Selecionou-se, da base original, os dados relativos à Região Sul do Brasil, operacionalizando em seguida as etapas do KDD: Limpeza, em que parte das colunas originais da base foi excluída devido a seu formato, ausência de registros ou irrelevância; Integração, ignorada pelo fato de a base em questão referir-se ao ano de 2005; Transformação, na qual colunas com valores numéricos dispersos foram convertidos em faixas, cabeçalhos foram adaptados e dados foram concatenados em colunas únicas; Mineração em si, aplicando o *Apriori* por meio do software WEKA, para geração de regras; e Interpretação dessas últimas.
Além da base de dados propriamente dita, foi disponibilizado para o estudo o questionário utilizado para sua obtenção, contendo 102 questões relativas a: perfil socioeconômico; posse de eletrodomésticos; valores e períodos de consumo; atitudes para economia de energia e relacionamento do cliente com a respectiva concessionária. O mesmo foi aplicado pelo Procel a moradores de cidades selecionadas, em todo o Brasil.
O instrumento apresenta questões discursivas, objetivas, de atribuição de nota, de ordenação de prioridades e tabelas, constituindo uma ausência de padronização que torna complexas a coleta e a tabulação dos dados, que, além disso, são realizadas manualmente. A coleta, por sua vez, se utiliza do questionário impresso, sendo os dados posteriormente transcritos para planilhas, que não necessariamente seguem o padrão do questionário (questões em formato de tabela, por exemplo, são ignoradas devido à complexidade de transcrição, e questões com mais de uma alternativa possível são tabuladas em mais de uma coluna). O número de colunas das planilhas tabuladas, assim, excede aquele de questões do instrumento de coleta.
A tabulação manual incorre, ainda, na possibilidade de manipulação dos dados no momento da tabulação, da realização dessa sem um compromisso com a completude dos registros, e de componentes tabulados não verificados no questionário original. Ou seja, a certificação de veracidade dos dados e sua utilização em processos de mineração, por exemplo, são dificultados mediante a possibilidade de existência de dados errôneos e/ou faltantes nos resultados da pesquisa.
Apesar dessas questões, partiu-se do pressuposto da veracidade e coerência dos dados tabulados na base do Procel. O processo de descoberta de conhecimento aplicado para extração de regras de associação da mesma, portanto, foi conduzido conforme o proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996): etapas de Seleção, Limpeza, Pré-processamento / Integração, Transformação, Mineração dos dados em si e Interpretação / Avaliação, descritas a seguir.

4.1 Seleção

Visto que o resultado de pesquisa referente a todo o Brasil contempla características distintas de tabulação para cada Região, optou-se por conduzir a mineração de dados de registros relativos somente ao Sul do Brasil (estados do Paraná, Santa Catarina e Rio Grande do Sul). Constituiu-se assim um recorte da base original, com 1000 registros, e todas as 190 colunas da primeira.
As etapas demonstradas a seguir são passíveis de aplicação aos dados relativos às demais regiões do país, com a ressalva de que, para cada Região, a etapa de limpeza deve adaptar-se aos respectivos dados.

4.2 Limpeza

A limpeza dos dados iniciou com a verificação da legibilidade dos cabeçalhos já das colunas da base, por exemplo *"mud_cons"*. Esse formato, nem sempre inteligível, foi alterado comparando-se o cabeçalho e valores de cada coluna às questões do questionário, com

objetivo de compreender a relação entre ambos. A coluna *"mud_cons"*, por exemplo, foi renomeada para *"mudaria hábitos de consumo"*, e assim sucessivamente, até a renomeação de todos os atributos.
Identificada a relação das 190 colunas originais da base com o questionário a ela relacionado, definiu-se os critérios para limpeza de dados, eliminando atributos e respectivos valores, conforme o Quadro 1:

CRITÉRIO	**COLUNAS ELIMINADAS**
Colunas com registros únicos e/ou irrelevantes	Número do questionário; Número do cliente; Nome do entrevistador; Nome do morador; Endereço do morador; Telefone do cliente; Hora de início da resposta; Hora de término da resposta; Dia de realização da pesquisa; Nome do digitador; Ano do automóvel mantido na residência;
Coluna com mais de 90% de valores iguais	Região; Inexistência de forma de aquecimento solar ou outra;
Colunas vazias	Outro tipo de cobertura; Outro tipo de esquadria de janelas; Outra condição de ocupação do domicílio; Trabalho domiciliar para ser comercializado – questões3 a 5 e 8;Outra razão para continuação do uso de lâmpadas de determinado consumo; Descrição de ação para redução de consumo realizada em relação aos eletrodomésticos – questões: geladeira, ar condicionado, freezer, chuveiro, lava roupas, *stand by*, microondas, lâmpadas, outro, qual outro eletrodoméstico; Outra variação de qualidade de vida causada pelo racionamento; Quais problemas de energia ocorrem; Outro tipo de domicílio;
Coluna cuja variação é tal que se aproxima de registros únicos	Bairro;
Colunas com registros textuais não padronizados	Outro tipo de piso; Responsabilidade de iluminação por outro órgão;
Várias colunas relativas a uma mesma questão (formato alternativas ou tabela), de difícil interpretação	Fonte de acesso à internet própria, comunitária, no trabalho, amigos, outro, qual, não sabe; Trabalho doméstico para ser comercializado – questões 1, 2, 6, 7, 9; Equipamentos utilizados no trabalho – questões 1, 2, 3, 4; Fontes de informação sobre eficiência energética – questões televisão, revistas, jornais, internet, contatos, amigos, lojas, outros, quais outros, não sabe ou não respondeu;
Colunas com menos de 80% de dados	Outro tipo de forro; Outro tipo de parede externa; Número de referência de refrigerador; Consumo de refrigerador novo; Opção por refrigerador com menor consumo; Gostaria de manter mais um refrigerador na residência; Interesse em freezer com baixo consumo de energia; Gostaria de manter mais um freezer na residência; Não sabe/não respondeu a respeito de formas de aquecimento de água; Mudaria sistema de aquecimento para utilização de gás; Mudaria sistema de aquecimentopara energia solar; Utilização de GNV por automóvel(is) na residência; Quantos automóveis utilizam GNV na residência; Motivo de uso do GNV 1, 2, 3, 4, 5, 6, 7, outro; Medida de racionamento – questões 1, 2, 3; Substituiria lâmpadas; Descrição

	de lâmpadas – questões 1, 2, 3, 4, 5, 6, 7; Quantidade de lâmpadas; Continua utilizando lâmpadas de determinado consumo; Razão pela qual continua utilizando lâmpadas de determinado consumo;Identifica selo do Procelem eletrodomésticos;
Colunas com valores não relacionados no questionário	*"Cla_con"; "And_loc";* Abastecimento de água - questão 2; Água para banho – questões 1, 2, 3; Como avalia o racionamento de energia; Número de ação para redução de consumo realizada em relação aos eletrodomésticos – questões: geladeira, ar condicionado, freezer, chuveiro, lava roupas, *stand by*, microondas, lâmpadas, outro;

QUADRO 1 – CRITÉRIOS E COLUNAS ELIMINADAS
FONTE: Os Autores (2011).

Manteve-se a coluna "Salário de empregada doméstica", embora apresentasse 70% de registros vazios, por compor o perfil socioeconômico das residências registradas na base, atribuindo-se os vazios ao fato de a residência em questão não manter empregada doméstica.
Da primeira triagem, portanto, restaram 74 das 190 colunas originais da base.

4.3 Integração

Uma vez que a pesquisa para obtenção dos dados foi realizada em 2005, e se obteve acesso à base integral da pesquisa, não houve proposta de integração da mesma com outras bases.

4.4 Transformação

As colunas "Tempo de moradia em anos" e "Tempo de moradia em meses", relativas ao período de residência no local em anos e meses, foram unificadas, resultando no período em meses de moradia na residência em questão. Ou seja, transformou-se os valores da primeira (anos) em meses, somando-os aos valores da segunda.
Em seguida, identificou-se as colunas com valores textuais (por exemplo "Cidade"), formatando-os sem espaçamento entre palavras (de "Rio Negrinho" para "RioNegrinho", por exemplo), acentuação e em caixa baixa, e inserindo o valor "VAZIO" em registros em branco.
Os valores das colunas numéricas foram comparados às respectivas perguntas no questionário, conferindo-se a correspondência entre dados registrados e alternativas. Constatada a existência de valores "99" e "888" nas referidas colunas, entendeu-se que o primeiro representou a opção "Não sabe ou Não respondeu", existente nas questões cujas colunas apresentavam o valor. O segundo, por sua vez, referiu-se a "Vazio", ou seja, resposta não aplicável ao registro, não tabulada ou inexistente. Manteve-se essa valoração e, a células ainda vazias nas colunas numéricas, atribuiu-se o valor "888".
A coluna "NR_quest", relativa à alternativa "Não sabe ou não respondeu", da questão "Nota para serviços prestados por empresas de água, telefonia e energia", apresentava valor somente quando mesma alternativa era assinalada pelo respondente. Consequentemente, as demais colunas relativas à mesma questão apresentavam valor vazio, nesse caso. Portanto, os registros da primeira coluna foram convertidos em "99" e incorporados às demais relacionadas à mesma questão, constituindo o padrão numérico citado acima.
As colunas "Tempo de moradia no local em meses", e "Tempo de construção da residência", tratando-se de períodos, foram categorizadas em faixas de valores. Para tanto, identificou-se os valores máximo e mínimo de cada coluna, sendo o segundo subtraído do primeiro, e o resultado, dividido por 5 (visando constituir cinco categorias para cada coluna – nomeadas pelas letras de A até E). Ao resultado obtido, somou-se o valor mínimo por quatro vezes consecutivas, constituindo os patamares máximos compreendidos nas quatro primeiras faixas, e sendo a quinta, composta por valores acima da anterior. À segunda coluna supracitada

acrescentou-se uma sexta faixa, contemplando somente valores relativos a "Não sabe ou não respondeu", nomeados pela letra F. Comparou-se enfim os conteúdos originais das colunas aos compreendidos pelas respectivas faixas, substituindo-os.
De forma semelhante, os valores dos atributos "Salário de empregada doméstica" e "Quantidade de lâmpadas mantidas na residência", foram categorizados, reduzindo sua variabilidade e, portanto, elevando a probabilidade de identificação de regras neles baseadas. Nestes casos, as categorias foram defnidas arbitrariamente, ou seja, sem a utilização de função para distribuição dos valores. Aqueles do primeiro atributo foram substituídos pelas letras A, B, C, D, E e F, respectivamente contemplando salários menores ou iguais a: R$ 100, R$ 200, R$ 300, R$ 400, R$ 500, e valor "888", ou seja, vazio. Já os valores do segundo atributo (originalmente entre 0 e 70, não necessariamente contemplando todos os números entre os extremos) foram listados e categorizados como A, B, C, D, E, F e G, respectivamente contemplando valores de 0 a 4, 5 a 8, 9 a 12, 13 a 16, 17 a 20, 21 a 21, e iguais ou maiores de 25.
Os cabeçalhos das colunas, até então apresentados em forma de texto, foram substituídos pelo número de sua referida questão, por extenso (por exemplo, a coluna relativa à pergunta "1.2" foi nomeada como "umDois", e assim sucessivamente). Desse modo, eventuais problemas de leitura de variáveis, pelo software de mineração, foram eliminados.
Enfim, após esta etapa de transformação, a base a ser submetida à mineração constituiu-se de 72 atributos.

4.5 Mineração de Dados

Para esta tarefa utilizou-se o software WEKA (acrônimo de *Waikato Environment for Knowledge Analysis*), na versão 3.6.4 – mais recente, para Windows x86[1]. O WEKA contempla diversos algoritmos de mineração (associação, clusterização e classificação) em um pacote de software *stand-alone*, baseado em JAVA. (MIKUT; REISCHL, 2011, tradução nossa).
Algoritmos nele embutidos são passíveis de aplicação a bases de dados, desde que apresentadas em formato *Attribute-Relation File Format* (ARFF): lista dos atributos contidos na base a ser minerada e seus respectivos valores, separados por vírgulas e já tratados pelas tarefas anteriores do KDD. (WITTEN; FRANK, 2005, tradução nossa). Para tanto, a base do Procel, até o momento em formato tabular, foi convertida para o formato *Comma Separated Values* (CSV) e, em editor de texto, salva como *.arff*, contemplando: nome da base (procel), precedido do termo *@relation*, lista de cada atributo nela contido e respectivos valores possíveis, entre colchetes, precedido do termo *@attribute*, e os valores em si, separados por vírgula, sendo seu conjunto precedido do termo *@data*.
O desconhecimento por parte dos pesquisadores quanto a necessidades dos públicos interessados na pesquisa do Procel, levou à indefinição de atributo meta para a mineração da base, e consequente seleção da estratégia associativa. Entre os algoritmos de associação disponíveis no WEKA, portanto, selecionou-se o *Apriori*, viabilizando a verificação de associações genéricas entre os atributos da base.
Logo, a mineração em si foi realizada em processador *Intel Core* i3, de 2,53 GHz e memória RAM de 3 GB, com sistema operacional *Windows 7* (versão *Home Basic*), de 64 bits.
Para a seleção das regras a serem apresentadas pelo WEKA, utilizou-se como métrica o suporte e a confiança, conforme os seguintes parâmetros:

[1] http://prdownloads.sourceforge.net/weka/weka-3-6-4jre.exe

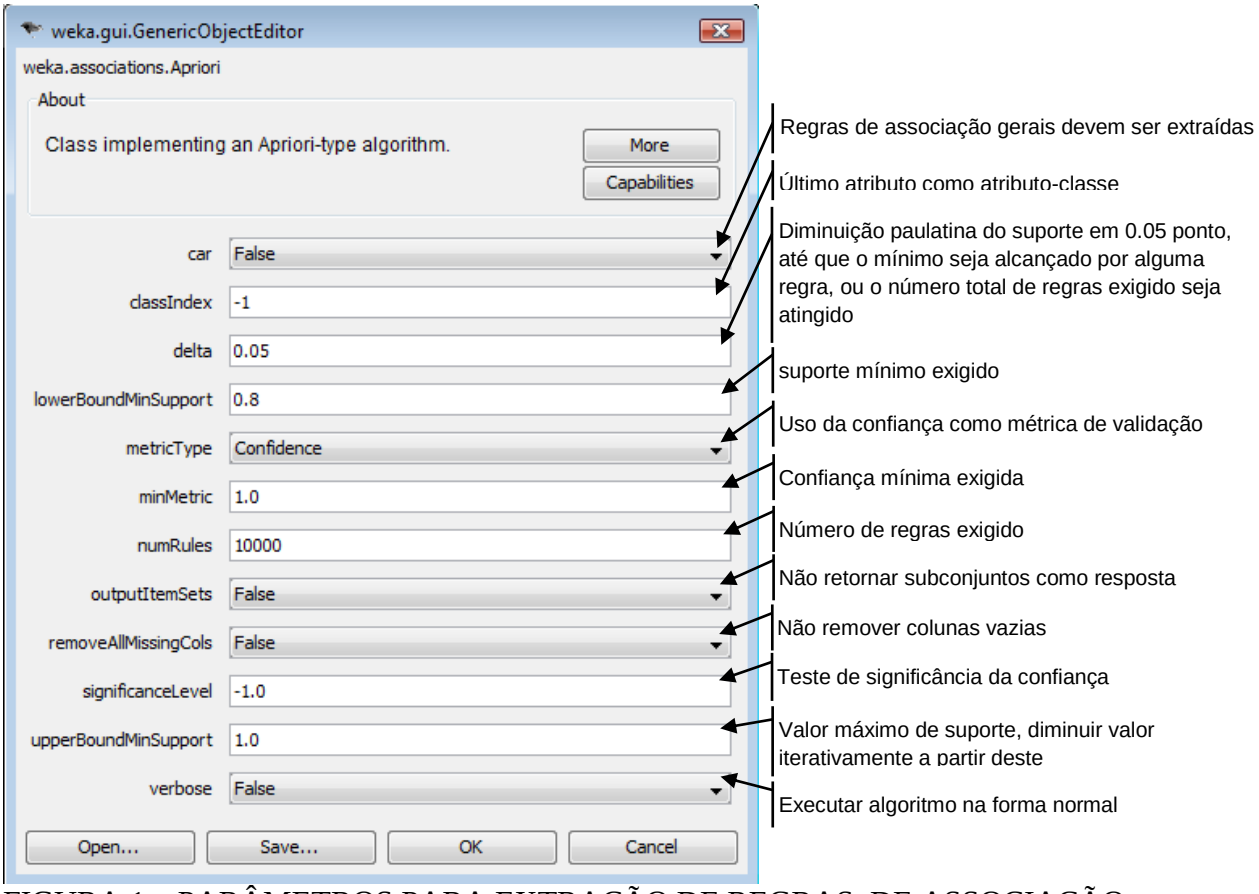

FIGURA 1 – PARÂMETROS PARA EXTRAÇÃO DE REGRAS DE ASSOCIAÇÃO
FONTE: Os Autores (2011).

O software apresentou as regras extraídas em menos de um minuto e, das 10000 exigidas (visando verificar o máximo de regras passível de recuperação com suporte de 0,8 e confiança de 100%), o WEKA retornou 1754 regras. O suporte máximo obtido foi de 0,952, e o mínimo, de 0,8, sendo a confiança igual a 100% em todas as proposições.
Ressalta-se que o citado valor mínimo de suporte se deve ao fato de a base do Procel conter dados demográficos que deveriam levar a decisões concretas quanto a distribuição, consumo e valoração de energia envolvendo altos montantes financeiros. Além disso, a variação de formato dos dados, seu processo de coleta e o e baixo número de registros, contribuíram para a definição, como ideal, do suporte mínimo exigido.
Para obtenção dos resultados, o WEKA rodou 4 ciclos de geração de conjuntos frequentes e, apesar do total de 1754 regras, as mesmas foram compostas apenas por 19 dos 72 atributos da base, cuja ocorrência é demonstrada no Quadro 2, em ordem de questão:

NOME DA COLUNA UTILIZADO NO WEKA	**QUESTÃO**	**OCORRÊNCIA EM REGRAS**
umOnze	Domicílios atendidos pelo medidor de consumo de energia	685
doisTres	Tipo de parede externa da residência	524
doisQuatroI	Vidros coloridos ou com películas	296
doisOito	O domicílio possui sistema de abastecimento de água	711

NOME DA COLUNA UTILIZADO NO WEKA	QUESTÃO	OCORRÊNCIA EM REGRAS
tresUm	Houve falta de energia na residência nos últimos 15 dias	196
tresDoisA	Duas ou mais queimas de lâmpadas nos últimos 3 meses	307
tresDoisB	Desligamento / Queima de disjuntor nos últimos 3 meses	826
tresDoisC	Queda de tensão de iluminação nos últimos 3 meses	368
tresDoisD	Choque elétrico em eletrodomésticos nos últimos 3 meses	1688
tresDoisE	Aquecimento de parede nos últimos 3 meses	1697
tresDoisF	Ocorrência de outro problema nos últimos 3 meses	61
quatroDois	Respondente conhece lâmpadas fluorescentes	915
cinco	Quantidade de refrigeradores na residência	716
dezSete	Respondente conhece aquecedores a gás para banho	617
DezOito	Respondente conhece aquecedores solares para banho	35
dezNove	Respondente considera sistema de aquecimento utilizado eficiente	184
onzeUmC	Quantidade de empregadas domésticas na residência	147
onzeSeis	Residência é próxima à favela	372
dozeUm	São adotadas medidas de economia de energia na residência	330

QUADRO 2 – ATRIBUTOS DA BASE CONSTANTES NAS REGRAS
FONTE: Os Autores (2011).

São exemplos de regras extraídas pelo WEKA, da base Procel:
Em primeiro, com 952 ocorrências: "Se não houve Desligamento / Queima de disjuntor, tampouco Choque elétrico em eletrodomésticos nos últimos 3 meses, na residência, então não houve Aquecimento de parede no mesmo período" (Resultado do software: *1. tresDoisB=2 tresDoisD=2 952 ==> tresDoisE=2 952 conf:(1)*).
Com 926 ocorrências: "Se a parede externa da residência tem revestimento externo de alvenaria, e não houve Choque elétrico em eletrodomésticos nos últimos 3 meses, então não houve Aquecimento de parede no mesmo período" (Resultado do software: *9. doisTres=1 tresDoisD=2 926 ==> tresDoisE=2 926 conf:(1)*).
Com 917 ocorrências: "Se não houve Choque elétrico em eletrodomésticos nos últimos 3 meses e o Respondente conhece aquecedores a gás para banho, então não houve Aquecimento de parede nos últimos 3 meses" (Resultado do software: *11. tresDoisD=2 dezSete=1 917 ==> tresDoisE=2 917 conf:(1)*).
Notou-se que, independentemente de outros atributos no antecedente, os fatores "Parede externa da residência em alvenaria", "Não ocorrência de Desligamento / Queima de disjuntor nos últimos 3 meses", e "Conhecimento por parte do respondente, de aquecedores a gás para banho", incorreram sempre em consequente "Não ocorrência de Aquecimento de parede nos últimos 3 meses". Das 1754 regras, 1626 tiveram esse último como consequente, e em 13

dessas, teve como complemento o "Conhecimento, por parte do respondente, de lâmpadas fluorescentes" (atributo quatroDois).
Do total de regras, 110 apresentaram o "Conhecimento, por parte do respondente, de lâmpadas fluorescentes" como único consequente, cuja primeira apresentou 844 ocorrências na base: "Se não houve Duas ou mais queimas de lâmpadas nos últimos 3 meses na residência e o Respondente conhece aquecedores a gás para banho, então o Respondente conhece lâmpadas fluorescentes" (Resultado do software: *336. tresDoisA=2 dezSete=1 844 ==> quatroDois=1 844 conf:(1)*).
Das 110 regras supracitadas, somente duas não apresentaram a não ocorrência de "Duas ou mais queimas de lâmpadas nos últimos 3 meses na residência" (atributo tresDoisA) como antecedente, apresentando, ambas, os valores "Não houve falta de energia na residência nos últimos 15 dias", "Apenas 1 refrigerador na residência" e "Respondente considera sistema de aquecimento utilizado eficiente" como antecedentes. A primeira delas, com 813 ocorrências na base, e a segunda, dela derivada, com 804 ocorrências.
O WEKA retornou, ainda, 18 regras com consequente "Não houve Choque elétrico em eletrodomésticos nos últimos 3 meses, na residência", a primeira com 837 ocorrências na base: "Se os vidros da residência não são coloridos ou têm película, Não houve desligamento / Queima de disjuntor, queda de tensão, tampouco aquecimento da parede nos últimos 3 meses, então Não houve Choque elétrico em eletrodomésticos nos últimos 3 meses, na residência" (Resultado do software: *1282. doisQuatroI=2 doisOito=1 tresDoisB=2 tresDoisC=2 tresDoisE=2 cinco=1 809 ==> tresDoisD=2 809 conf:(1)*). As demais 17 regras derivam desta primeira.
À exceção dos consequentes das regras acima, os atributos do Quadro 2 não citados, foram apresentados como antecedentes das primeiras.
A titulo de teste de retorno de regras, repetiu-se o processo com suporte mínimo de 0,7, exigindo, desta vez, 20000 regras, sendo todas obtidas. Contudo os consequentes por elas apresentados mantiveram-se os mesmos da primeira rodagem do algoritmo e, em uma seleção dos 20000 resultados, verificou-se que os mesmos derivaram dos anteriores, exemplificados acima.

4.6 Interpretação

As regras obtidas por meio da aplicação do *Apriori* são relativas a perfis de consumidores de energia, deixando de contemplar os perfis de consumo de energia, o uso de eletrodomésticos e o relacionamento dos consumidores com as respectivas concessionárias. Desse modo, sua utilidade limita-se a caracterizar residências da Região Sul do Brasil e reforçar a ausência de problemas de consumo de energia nas mesmas.
Acredita-se que esses resultados se devem ao baixo número de registros da base (apenas 1000) relativos a toda a Região Sul do Brasil, que comprometem a obtenção de regras generalistas e mais abrangentes, para tomada de decisão por parte dos interessados no processo. A redefinição da amostragem utilizada para coleta dos dados auxiliaria na resolução dessa questão.
O tratamento da base integral, relativa a todo o Brasil, alternativa para a extração de regras mais consistentes, entretanto, excederia o limite de tempo para realização desta pesquisa. Ainda que fosse executado, entretanto, seria comprometido pelas peculiaridades regionais dos dados, retornando regras que dificilmente poderiam ser generalizadas.
O valor mínimo de suporte exigido para validação das regras, mesmo quando reduzido na testagem final, originou resultados pouco diferenciados dos inicialmente obtidos, reforçando a interpretação de que os dados do Procel são demasiado dispersos para a extração de regras completas sobre o perfil dos respondentes.

Para cumprimento dos objetivos do próprio Programa, ou seja, racionalização da produção e do consumo de energia elétrica, minimizando desperdícios e custos, bem como para pesquisas em eficiência energética, a base de dados utilizada apresenta contribuições mínimas, levando à conclusão de uma necessária revisão de ações para coleta, transcrição e uso dos dados em si.

5. CONSIDERAÇÕES FINAIS

O processo decisório estruturado depende essencialmente da organização de dados que o fundamentam, bem como de planejamento e controle das atividades por meio das quais são coletados. Da mesma forma, a mineração de dados retorna resultados pertinentes quando baseada em dados estruturados e cuja coleta é voltada a um propósito.

O formato de condução da pesquisa à qual os dados aqui apresentados se referem demonstra a ausência desse propósito, dificultando a melhoria contínua, citada por Zhang e Zhou (2004), do conhecimento por parte dos públicos interessados a respeito do consumo de energia no Sul do Brasil.

As falhas de condução da pesquisa verificadas incorrem em tratamento e interpretação, custosos, dos dados e resultados advindos de sua mineração. No caso da base Procel, observa-se que planejamento, pré-testagem e validação do questionário; treinamento de aplicadores e inserção de tecnologias no processo de coleta e tabulação dos registros, necessitam de aprimoramento.

Especialmente o não uso de tecnologias para constituição da base de dados em si, compromete sua credibilidade e, portanto, dificulta a adequação de suas regras a ações de distribuição, valoração e consumo de energia, respectivamente por governos, as concessionárias de energia e os próprios cidadãos.

Os desafios mencionados por Zhang e Wu (2011) quanto à extração de regras de associação, concretizaram-se durante esta pesquisa: a extensão e variação de dados de colunas da base, dificultou a consecução do KDD, desde a interpretação prévia dos mesmos em relação ao questionário utilizado para sua obtenção, devido à desestruturação do mesmo. Desse modo, a identificação de metas para a tarefa de mineração também foi comprometida, uma vez que não se pode identificar necessidades pré-estabelecidas em relação aos dados.

Sugere-se a aplicação das etapas acima descritas aos dados das demais regiões do Brasil, na mesma base, e a atualização das regras com dados de 2010 – pesquisa mais recente realizada pelo Procel, quando disponíveis.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R.. Fast algorithms for mining association rules in large databases. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 20, 1994, Santiago. **Proceedings…** p. 478-499, Santiago: 1994. Disponível em: <http://rakesh.agrawal-family.com/papers/vldb94apriori.pdf>. Acesso em: 01 jun. 2011.

CONFERÊNCIA DAS NAÇÕES UNIDAS SOBRE MUDANÇA DO CLIMA. Participação do Brasil na COP15. Disponível em: <http://www.cop15brasil.gov.br/pt-BR/>. Acesso em: 01 jun. 2011.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery in databases**. American Association for Artificial Intelligence, 1996. Disponível

em: <http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf>. Acesso em: 29 mai. 2011.

LEE, S. J.; SIAU, K..A review of data mining techniques, **Industrial management and data systems**, v. 101, n. 1, p. 41-46, 2001. Disponível em:<http://www.emeraldinsight.com/journals.htm?issn=0263-5577&volume=101&issue=1&articleid=850015&PHPSESSID=8alcdeo1t9ql84bm0taqesa6s7>. Acesso em: 21 mai. 2011.

MIKUT, R.; REISCHL, M..Data mining tools, **Advanced review**, v. 00, jan./feb.2011. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/widm.24/pdf>. Acesso em: 22 mai. 2011.

PROGRAMA NACIONAL DE CONSERVAÇÃO DE ENERGIA ELÉTRICA – Procel. Disponível em: <http://www.eletrobras.com/elb/procel/>. Acesso em: 29 mai. 2011.

WITTEN, I. H.; FRANK, E.. **Data mining**: practical machine learning tools and techniques. 2. ed. San Francisco: Elsevier, 2005.

ZHANG, S.; WU, X.. Fundamentals of association rules in data mining and knowledge discovery, **Overview**, v. 1, n. 2, p. 97-116, mar./abr. 2011.Disponível em:<http://onlinelibrary.wiley.com/doi/10.1002/widm.10/pdf>. Acesso em: 22 mai. 2011.

ZHANG, D.; ZHOU, L.. Discovering golden nuggets: data mining in financial application, **IEEE transactions on systems, man, and cybernetics—part c: applications and reviews**, v. 34, n. 4, nov. 2004. Disponível em: <http://suraj.lums.edu.pk/~cs631s05/Papers/financial.pdf>. Acesso em: 31 mai. 2011.