Controladoria-Geral da União (CGU)

ESPECIALIZAÇÃO EM AUDITORIA E CONTROLE GOVERNAMENTAL

SELEÇÃO DE OBRAS PARA FISCALIZAÇÃO DO PROGRAMA LUZ PARA TODOS UTILIZANDO FERRAMENTAS COMPUTACIONAIS.

JOSÉ FERNANDO DE FARIA LUCENA DANTAS

ADRIANO AUGUSTO DE SOUZA (MESTRE)

Brasília – DF
2011

SELEÇÃO DE OBRAS PARA FISCALIZAÇÃO DO PROGRAMA LUZ PARA TODOS UTILIZANDO FERRAMENTAS COMPUTACIONAIS.

.

JOSÉ FERNANDO DE FARIA LUCENA DANTAS

Orientador: ADRIANO AUGUSTO DE SOUZA (MESTRE)

Artigo apresentado ao Instituto Serzedello Corrêa – ISC/TCU, como requisito parcial à obtenção do grau de Especialista em Auditoria e Controle Governamental.

BRASILIA – DF
2011

RESUMO

Desde 2004, o governo federal, já desembolsou mais de 11 bilhões de reais para execução do Programa Luz para Todos – PLPT. A Controladoria-Geral da União – CGU, em sede de trabalhos de auditoria realizadas em obras deste programa, evidenciou impropriedades na sua execução, ensejando a necessidade de realização de ações de controle para a sua avaliação. O presente trabalho trata do desenvolvimento de uma estratégia para a seleção de obras a serem auditadas, através da utilização de ferramentas computacionais, utilizando a base de dados do PLPT obtida nas Centrais Elétricas Brasileiras - ELETROBRAS que é composta de mais de 345 mil obras. Como resultado final deste trabalho é apresentado uma proposta de metodologia que busca propiciar a obtenção de indícios de problemas que permitam a diminuição do esforço empreendido, servindo como opção metodológica, na análise de dados, processos ou obras que estejam fora do padrão, principalmente quando da realização de ações de controle especiais. A principal motivação deste trabalho é buscar otimizar o esforço dos analistas e técnicos da CGU na realização dos seus trabalhos de campo, oferecendo uma alternativa as técnicas tradicionais de acompanhamento da execução de programas de governo.

Palavras-Chave: Programa Luz para Todos, Metodologia para auditoria, Modelagem matemática de obras, Mineração de dados.

ABSTRACT

Since 2004, the federal government has disbursed over 11 billion dollars for implementing the "Programa Luz para Todos – PLPT". The "Controladoria Geral da União – CGU" at audits performed in this program revealed inadequacies in its implementation, entailing the need to carry out control measures for their evaluation. The present work deals with the development of an estrategy for selecting projects to be audited through the use of computational tools using the database obtained in "Centrais Elétricas Brasileiras – ELETROBRAS" which is composed of more than 345 000 works. As a final result of this work is presented a proposal for a methodology that seeks to facilitate the obtaining of evidence of problems that allow the reduction of the effort made by serving as a methodological option, the analysis of data, processes, or works that are outside the norm, especially when the performing special control actions. The main motivation of this paper is to seek to optimize the efforts of analysts and technicians in the CGU conduct their research field, providing an alternative to traditional techniques for monitoring the implementation of government programs.

Keywords: Programa Luz para Todos, Methodology Audit, Mathematical Model Methodology, Data Mining.

Índice de Figuras

Índice de Tabelas

SUMÁRIO

1) INTRODUÇÃO:

A Controladoria-Geral da União – CGU possui, conforme estabelecido no inciso II do art. 74 da Constituição Federal do Brasil de 1988 – CF88, a atribuição de avaliar a utilização dos recursos públicos federais, para isso, utiliza-se de diferentes tipos de ações de controle: o acompanhamento da execução de programas de governo e a realização de auditorias especiais.

No primeiro tipo, frequentemente é necessário que os servidores da CGU sejam instados a manipular grandes volumes de dados, que podem estar organizados em planilhas derivadas de ordens bancárias para pagamentos, planilhas orçamentárias, medições de obras,

planilhas de aplicação de insumos e caderneta de campo de topografia. Estes conjuntos de dados são compostos, muitas vezes, de um grande número de planilhas com 40.000 a 100.000 linhas, muitas vezes, com 10 a 50 colunas de informações entre datas, valores, CNPJ, CPF, município, UF, quantidades, preços, dentre outros.

Este artigo foca o segundo tipo de ação de controle, pois procura desenvolver uma estratégia que permita mapear a identificação de riscos (indícios) que facilite a geração de ordens de serviço através das auditorias especiais.

Neste trabalho é proposto um estudo de caso do Programa Luz para Todos - PLPT, tendo por objetivo o desenvolvimento de uma estratégia para seleção de obras com indícios de problemas deste programa com o uso de técnicas computacionais, em especial a mineração de dados, buscando otimizar a seleção de obras com indícios de desconformidade em relação a um padrão. O resultado obtido com este trabalho poderá ser validado quando da realização de ordens de serviço pela CGU.

Para atingir o objetivo do desenvolvimento da estratégia de seleção de obras (ou ainda mapeamento e identificação de riscos) é necessário estabelecer os seguintes objetivos: Importação e ajuste da base de dados; aplicação de ferramenta de clusterização para organização dos dados e aplicação da ferramenta regressão linear para seleção de obras do LpT.

A partir de ação de controle desenvolvido pela Coordenação-Geral de Auditoria da área de Minas e Energia – DIENE, foram obtidos diversas planilhas em Excel que foram importadas para uma base de dados em ACCESS, Após essa importação da base de dados é realizado uma etapa de limpeza, daí são realizados vários passos para o desenvolvimento da estratégia, através do uso de . técnicas de Mineração de Dados, em especial o agrupamento e a regressão linear.

A partir dos resultados obtidos é consolidada uma estratégia que pode ser empregada para seleção de obras com indícios de problemas para serem fiscalizadas em ordens de serviço a serem executadas pelos analistas e técnicos da CGU.

2) REFERENCIAL TEÓRICO:

Buscando oferecer um suporte teórico ao desenvolvimento da estratégia de seleção de obras para fiscalizaçãoé apresentado no tópico 2.1 o Programa Luz para Todos – LpT.

No tópico 2.2 são apresentados os processos de avaliação da execução de programas de governo e de auditorias especiais adotados pela CGU e no tópico 2.3 a técnica de

mineração de dados (Data Mining).

2.1) O Programa Luz para Todos - LpT:

O programa Luz para Todos – LpT, do governo federal, é regulamentado pelo Decreto nº 4.873, de 11 de novembro de 2003, e alterações e tem por objetivo, até o ano de 2014, atender com energia elétrica à parcela da população do meio rural brasileiro que ainda não tem acesso a esse serviço público, consistindo de execução de obras de instalação de redes rurais para o atendimento de unidades consumidoras isoladas ou em distritos que não possuem energia elétrica e que não possuem capacidade econômica para o pagamento pelo investimento na infraestrutura, objetivando, conforme o manual do programa "garantir o acesso ao serviço público de energia elétrica à parcela da população do meio rural brasileiro que ainda não possui acesso a esse serviço público" (MANUAL DE OPERACIONALIZAÇÃO, 2009, p. 6).

De acordo com art. 3º do Decreto nº 4.873/03, o Programa é coordenado pelo Ministério de Minas e Energia - MME e operacionalizado com a participação das Centrais Elétricas Brasileiras S.A. - ELETROBRÁS e das empresas que compõem o Sistema ELETROBRÁS.

A operacionalização do programa se dá pela concessão de subsídio e empréstimo aos Agentes do Programa LpT, que são as Concessionárias ou Permissionárias de Distribuição de Energia Elétrica (que são denominadas de “Distribuidoras ao longo do texto) . Os recursos para concessão dos subsídios são obtidos no fundo setorial Conta de Desenvolvimento Energético - CDE e os recursos para concessão de empréstimos são obtidos no fundo setorial da Reserva Global de Reversão – RGR geridos pela Centrais Elétricas Brasileiras S.A – ELETROBRAS. A utilização de recursos destes fundos setoriais tem como objetivo a mitigação do impacto tarifário que os investimentos representam (MANUAL DE OPERACIONALIZAÇÃO, 2009, p. 6).

Para acessar os recursos as Distribuidoras submetem a ELETROBRÁS um “programa de obras”, onde são informadas as composições de custo unitárias dos diversos itens de custo que são utilizados para a realização das obras, além de “metas de atendimento”. Após análise pelos técnicos da ELETROBRÁS deste “plano de obras“, a partir da ponderação de custos dos itens, é obtido o valor a ser repassado às Distribuidoras, sendo celebrado o

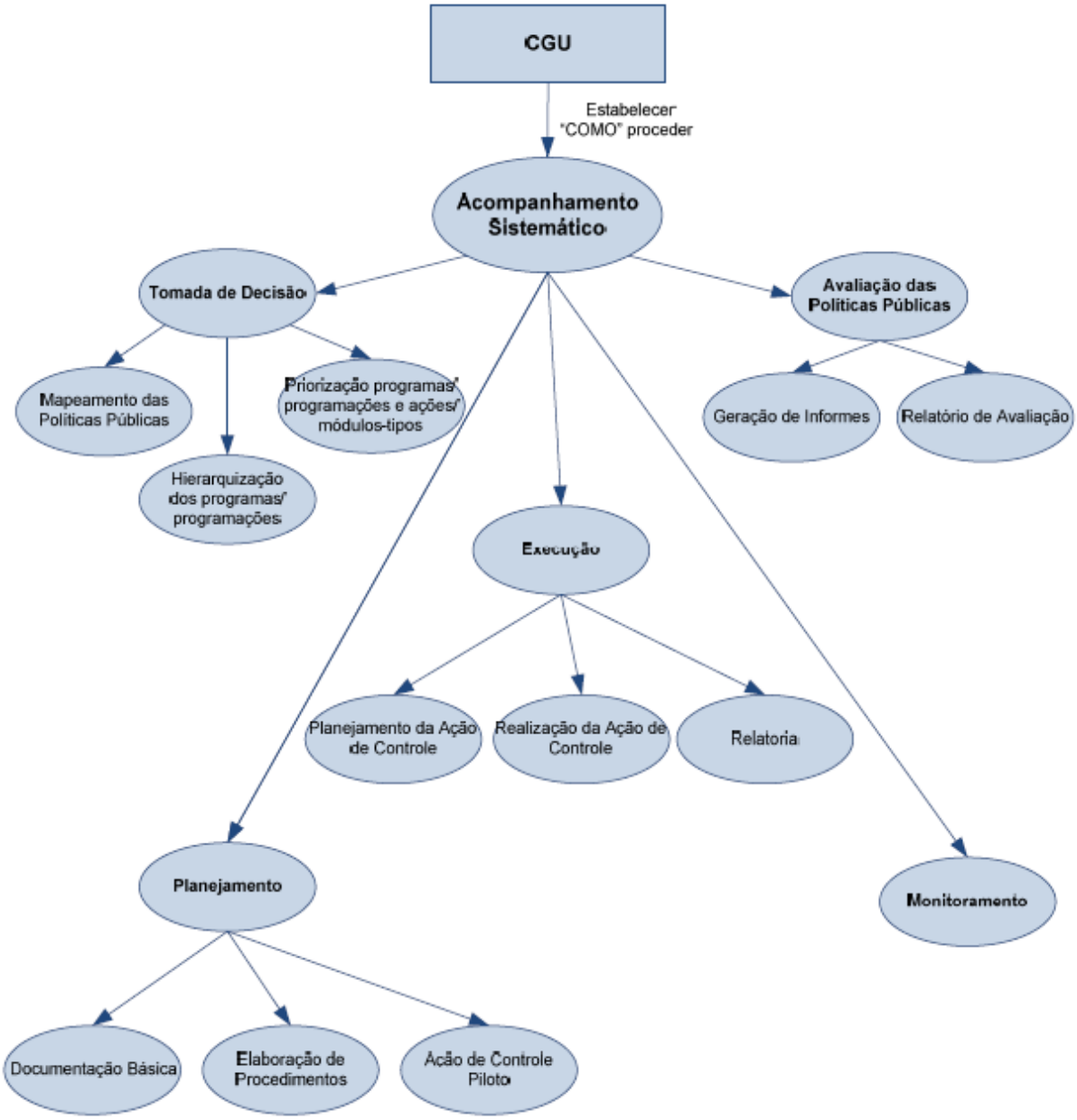
contrato em que são estabelecidos alguns parâmetros tais como o custo por km de rede, o número de consumidores por km de rede e o custo por consumidor, e metas de atendimento, geralmente em número de consumidores a serem atendidos (MANUAL DE OPERACIONALIZAÇÃO, 2009, p. 9).

Cada Distribuidora passa então a realizar as obras, devendo registrar no sistema gerenciado pela ELETROBRAS, para cada obra, os dados que permitam a esta entidade averiguar o avanço físico da execução do contrato e com isso planejar as fiscalizações a serem realizadas, com o intuito de avaliar a qualidade dos trabalhos realizados pelas Distribuidoras.

2.2) O processo de acompanhamento da execução de programas de governo e as auditorias especiais:

O processo de acompanhamento da execução de programas de governo é realizado pela CGU, de acordo com o manual "Metodologia para o Acompanhamento Sistemático da Execução de Programas de Governo", (MANUAL AEPG, 2010, p.12), sendo composto de cinco macro-processos: Tomada de Decisão; Planejamento; Execução das Ações de Controle; Monitoramento e Avaliação, apresentadas no diagrama a seguir:

Figura 1 – Diagrama de macroprocessos do acompanhamento da execução de programas de governo AEPG.



Fonte: MANUAL AEPG 2010, página 12.

O fluxo apresentado imprime as diversas coordenações da CGU alguns desafios, como por exemplo, a obtenção de informações precisas sobre os programas a serem avaliados, em termos de abrangência e profundidade para o melhor planejamento da ação de controle, o desenvolvimento de procedimentos adequados.

O macroprocesso "Planejamento" é composto de vários processos, na elaboração de documentação básica são gerados os seguintes documentos:

- "Relatório de Situação - RS", onde se busca o levantamento amplo de dados,

buscando o conhecimento detalhado da ação governamental em relação a sua estrutura de funcionamento, suas normas e mecanismos de planejamento, gerenciamento, execução e controle, assim como seu histórico recente de desempenho, restrições e avaliações (MANUAL AEPG, 2010, p.19);

- "Plano Estratégico - PE": documento que apresenta a atuação da CGU para emissão de opinião sobre o programa de governo, definindo: os pontos críticos a serem avaliados; o universo de dados a serem trabalhados; a necessidade de utilização de ferramentas estatísticas para seleção de amostras e a segmentação da ação de controle visando à viabilização da estrutura operacional da CGU para a execução das Ordens de serviço (MANUAL AEPG, 2010, p.24);
- "Plano Operacional - PO": documento gerado para cada das "divisões" do PE que tem por função a estruturação das ações de controle com a definição dos agentes responsáveis por etapas da execução da ação de governo, visando o estabelecimento de ações de controle que gerem resultados objetivos (MANUAL AEPG, 2010, p.26).

Embora se tenha todo um esforço para a busca da avaliação dos programas de governo de forma planejada, é necessária a realização de ações de controle com o objetivo de verificar patologias ou ainda padrões que fogem a normalidade da execução do programa de governo, ensejando a realização das denominadas "auditorias especiais", que, conforme definição da IN SFC 01/2001 (BRASIL, 2001, p. 32), consiste:

> *"V. Auditoria Especial: objetiva o exame de fatos ou situações consideradas relevantes, de natureza incomum ou extraordinária, sendo realizadas para atender determinação expressa de autoridade competente. Classificam-se nesse tipo os demais trabalhos auditoriais não inseridos em outras classes de atividades.*"

Com foco nas auditorias especiais, a seguir é apresentada a técnica de mineração de dados, que aplicada em uma base de dados permite a seleção de registros que possam conter informações que destoam de um padrão estabelecido.

Destaca-se que a cada dia mais a mineração de dados vem sendo utilizada por pesquisadores em diversas áreas, tais como para a seleção de atributos determinantes no preço da habitação (BATISTA, 2010), ou ainda para avaliação do Programa Nacional de Informática na Educação (BRANDÃO et al, 2003), indicando ser esta uma ferramenta apropriada ao estudo neste artigo.

2.3) Mineração de Dados (Data Mining):

Neste tópico são apresentadas as definições de mineração de dados (tópico 2.3.1),

sendo no tópico 2.3.2 abordadas as etapas da mineração de dados.

No tópico 2.3.3 é buscado conceituar funcionalidade e técnicas em mineração de dados.

Buscando estudar as técnicas a serem utilizadas neste trabalho são apresentados nos tópicos 2.3.4 a técnica de análise de agrupamento e no 2.3.5 a técnica de regressão linear.

2.3.1) Definições de Mineração de dados:

Na literatura tem-se várias definições para Mineração de dados, ou "Data Mining", dos quais pode-se citar (CÔRTES, PORCARO, LIFSCHITZ, 2002, p. 1):

- "*Mineração de dados é a busca de informações valiosas em grandes bancos de dados. É um esforço de cooperação entre homens e computadores. Os homens projetam bancos de dados, descrevem problemas e definem seus objetivos. Os computadores verificam dados e procuram padrões que casem com as metas estabelecidas pelos homens (WEIS & INDURKHYA,1999).*
- *Mineração de dados é a exploração e análise de dados, por meios automáticos ou semi-automáticos, em grandes quantidades de dados, com o objetivo de descobrir regras ou padrões interessantes (BERRY & LINOFF,1977).*
- *Mineração de dados, em poucas palavras, é a análise de dados indutiva (MENA, 1999).*
- *Mineração de dados é o processo de proposição de várias consultas e extração de informações úteis, padrões e tendências, freqüentemente desconhecidos, a partir de grande quantidade de dados armazenada em bancos de dados (THURAISINGHAM ,1999).*
- *Mineração de dados, de forma simples, é o processo de extração ou mineração de conhecimento em grandes quantidades de dados (HAN & KAMBER, 2001).*
- *mineração de dados é um processo altamente cooperativo entre homens e máquinas, que visa a exploração de grandes bancos de dados, com o objetivo de extrair conhecimentos através do reconhecimento de padrões e relacionamento entre variáveis, conhecimentos esses que possam ser obtidos por técnicas comprovadamente confiáveis e validados pela sua expressividade estatística."*

Alem destas definições, tem-se descrita na literatura (RAMEZ & NAVATHE, 2010, P. 625) que a mineração de dados é parte de um processo de descoberta de conhecimento, ou ainda de busca de Conhecimento em Banco de Dados (Knowledge Discovery in Database - KDD).

2.3.2) Etapas do processo de Mineração de dados

O processo de mineração de dados é composto das seguintes fases: Seleção de dados, Limpeza, Enriquecimento, Transformação ou codificação, data mining (mineração de dados) e apresentação das informações obtidas.

- Seleção de dados: é a fase de obtenção das informações e sua colocação (importação) em um banco de banco de dados;
- Limpeza: consiste na aplicação de regras de verificação de consistência nos dados obtidos, verificando, por exemplo, se os códigos de CEP estão compatíveis com os nomes de cidades informados, ou ainda se os dígitos verificadores de cpf e cnpj informados estão corretos;
- Enriquecimento: são incorporadas ao banco de dados outras informações que possam ser correlacionadas com os dados presentes ao Banco de dados, como exemplo, a população dos municípios, o número de domicílio, a área do município e assim por diante;
- Transformação ou codificação: busca reduzir o número de dados, através da manipulação dos campos do banco de dados, agrupando ou desagrupando informações, estruturando campos, criando índices e etc.
- Mineração de dados: é a etapa que se busca extrair padrões e regras de relacionamento entre os dados. O produto da mineração de dados é então formatado para apresentação ao usuário final.

Os resultados obtidos a partir da mineração de dados podem ser utilizados para diversas finalidades, dentre as quais: predição de valores futuros de atributos de dados; identificação de padrões que podem indicar anomalias; classificação de dados, permitindo a sua partição em diferentes classes ou categorias; ou ainda na otimização de dados, buscando utilizar de forma ótima determinado recurso disponível, tal como memória ou tempo de CPU.

2.3.3) Funcionalidades e técnicas em Mineração de dados:

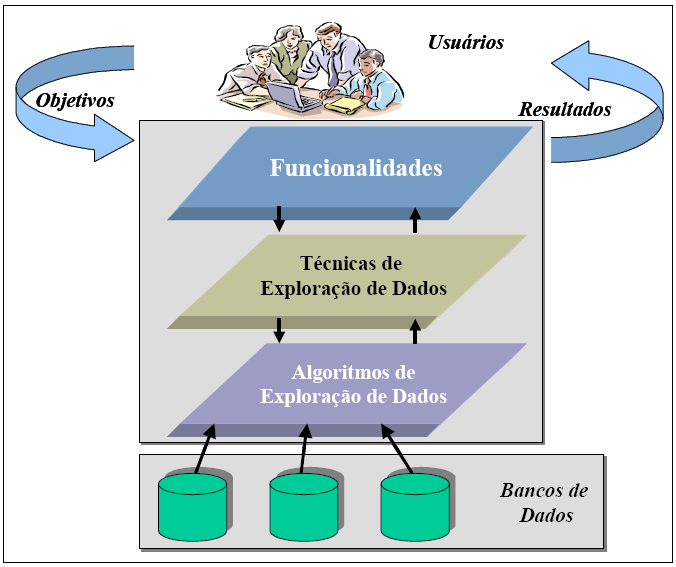
Não é consenso a definição entre técnicas e funcionalidades em mineração de dados, na literatura são encontradas algumas definições para as funcionalidades de mineração de dados, que pode-se destacar (CÔRTES, PORCARO, LIFSCHITZ, 2002, p. 3):

- *"Descoberta de conhecimento e Predição (ADRIAANS & ZANTINGE,1996)*
- *Classificação, Estimação, Predição, Afinidade em grupos, Agrupamentos (clustering) e Descrição (BERRY & LINOFF,1977).*
- *Classificação, Detecção de seqüência, Análise de dependência de dados e Análise de desvio (THURAISINGHAM ,1999).*
- *Previsão, Identificação, Classificação e Otimização (RAMEZ & NAVATHE, 2010, P. 625)*
- *Descrição e Predição(HAN & KAMBER, 2001)*

- *Predição, Classificação, Agrupamento (clustering), Segmentação, Associação, Visualização e Otimização (MENA, 1999)*
- *Classificação, Estimação, Segmentação e Descrição (WESTPHAL & BLAXTON, 1998)*
- *Predição, Detecção de desvio, Segmentação, Agrupamento (clustering), Análise de ligações e Regras de associação, Sumarização e Visualização e Garimpagem em textos (WEIS & INDURKHYA,1999)"*

Funcionalidade é a camada mais próxima do usuário, onde se busca selecionar os resultados a serem alcançados, conforme figura a seguir:

Figura 2 - relação entre funcionalidades, técnicas e algoritmos de mineração de dados.

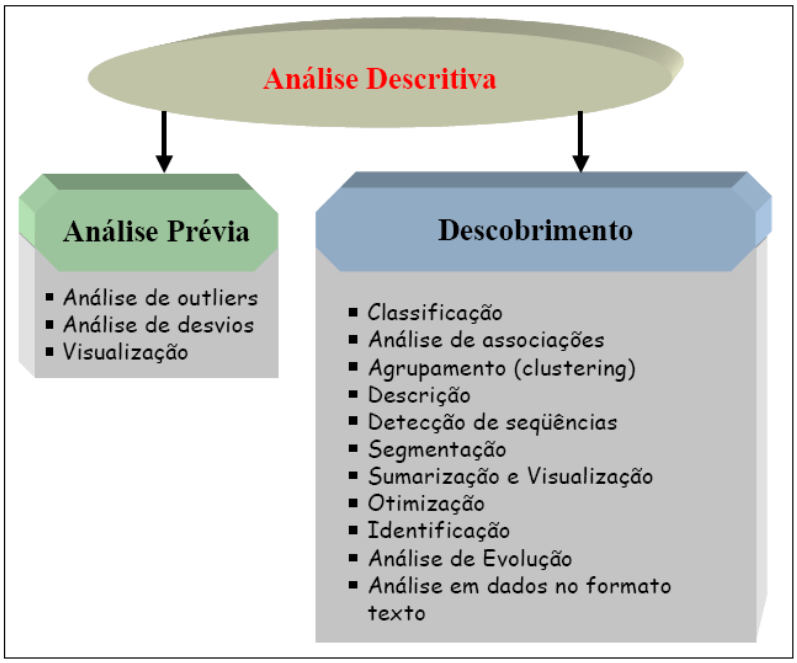


Fonte: (CÔRTES, PORCARO, LIFSCHITZ, 2002, p. 3)

As funcionalidades, segundo Cortes et AL, (CÔRTES, PORCARO, LIFSCHITZ, 2002, p. 4), pode ser agrupadas em dois conjuntos (sub-funcionalidades): Análise Descritiva (que busca investigar os dados para identificar anomalias e descobrir novas relações entre os dados) e Análise de Prognóstico (onde se busca inferir resultados a partir dos padrões disponíveis, ou ainda prever o comportamento para um novo conjunto de dados).

Para a análise descritiva pode-se agrupar as técnicas conforme a sua utilização em análise prévia ou para descobrimento conforme figura a seguir:

Figura 3 - Agrupamentos de técnicas de Análise Descritiva.



Fonte: (CÔRTES, PORCARO, LIFSCHITZ, 2002, p. 5)

A análise de Prognósticos, por sua vez, pode ser subdividida em três processos (CÔRTES, PORCARO, LIFSCHITZ, 2002, p. 11):

- *"Estimação – é o processo de predizer algum valor, baseado num padrão já conhecido. Por exemplo, conhecendo-se o padrão de despesas e a idade de uma pessoa, estimar seu salário e seu número de filhos.*
- *Predição – é o processo de predizer um comportamento futuro, baseado em vários valores. Por exemplo, baseado na formação escolar, no trabalho atual e no ramo de atividade profissional de uma pessoa, predizer que seu salário será de um certo montante até um determinado ano.*
- *Classificação – é o processo para predizer algum valor para uma variável categórica. Por exemplo, podemos num banco financeiro, determinar que conjunto de clientes oferecem risco ou não para contrair um empréstimo pessoal."*

Alternativamente pode-se ainda usar como definição de técnicas de mineração de dados,conforme Jaison Carlos Scheel (SCHEEL, 2008) "a fase de mineração possui várias técnicas de caráter genérico, onde podemos citar: redes neurais, indução de regras, árvores de decisão, análises de séries temporais, visualização, classificação estimativa, previsão, análise de afinidade e análise de agrupamento (clusterring).

- Redes Neurais – Essa técnica é tratada pelos processos de data mining como uma caixa preta, pois essa técnica constrói representações internas de modelos achados nos dados, porém essas representações não são mostradas aos usuários, por isso que os analistas de negócios não usam essa técnica, pois não podem explicar os resultados finais.

- Indução de Regras – Nesta técnica vários algoritmos e índices são colocados para executar esse processo, onde a maioria deste processo é feito pela máquina e só uma parte insignificante é feita pelo usuário. Na indução de regras, essas regras são apresentadas aos usuários como uma lista chamada de "não encomendada".
- Análises de Séries Temporais – É a fundação de todas as outras tecnologias, essa técnica é envolvida fortemente com o usuário e exige engenheiros experientes para construir modelos que detalhem o comportamento do dado através de métodos matemáticos.
- Classificação – É uma das tarefas de conhecimento humano mais usada no auxílio à compreensão do ambiente onde vivemos e é uma das técnicas mais utilizadas. No data mining essa técnica é utilizada na classificação de clientes, como por exemplo, classificá-los quanto ao seu nível social.
- Estimativa – Esta técnica é utilizada para determinar um valor aproximado de um índice através de dados que foram passados ou de dados adquiridos de outros índices semelhantes, sobre os quais se tem conhecimento.
- Previsão – Esta técnica tem por objetivo a avaliação de um valor de um índice ainda não identificado, baseando-se em dados adquiridos através do comportamento deste índice.
- Análise de Afinidade – Como o nome já diz, essa técnica determina que fatos ocorram simultaneamente com probabilidade razoável, ou então que itens de dados estão presentes juntos com certa chance. Um exemplo disso, como o autor cita é a de um carrinho de supermercado, através dele podem-se extrair informações para que a disposição dos produtos do supermercado agrade aos consumidores, colocando produtos próximos uns aos outros, produtos comprados em conjunto com outro.
- Análise de Agrupamentos – Nessa técnica os dados são agrupados conforme sua classificação e grupo em que pertencem e esses grupos são construídos com base na semelhança que há entre os elementos desse grupo." Pode-se, também utilizar a seguinte definição de análise de agrupamento (BRANDÃO et al, 2003, p. 369): *"A análise de agrupamento é uma técnica multivariada cujo objetivo primário é agrupar objetos ou entidades com base em características comuns, utilizando um determinado critério de seleção. Os*

agrupamentos resultantes (clusters) apresentam alta homogeneidade interna e alta heterogeneidade externa [Hair & Black, 2000]".

2.3.4) Análise de agrupamentos:

A análise de agrupamento pode ser classificada na funcionalidade "Descobrimento" como uma sub-funcionalidade (CÔRTES, PORCARO, LIFSCHITZ, 2002, p. 12), para o qual tem-se disponíveis diversas técnicas: Métodos de particionamento; Métodos hierárquicos; Métodos baseados em densidade; Métodos baseados em grid; Métodos de clustering baseados em modelos – abordagem estatística e redes neurais e Análise de outliers.

A análise de agrupamentos, ou clustering, (HAN & KAMBER, 2011, cap. 8 slide 3) pode ser utilizada como etapa de pré-processamento de um conjunto de dados, visando a aplicação de outra técnica para extração das informações ou pode ser a técnica utilizada diretamente para obtenção das informações pretendidas.

Pode-se, matematicamente, definir um problema de clusterização como (OCHI et AL, 2004, p. 2):

"Dado um conjunto com n elementos X = {X1, X2, ..., Xn}, o problema de clusterização consiste na obtenção de um conjunto de k clusters, C = {C1, C2, ..., Ck}, tal que os elementos contidos em um cluster Ci possuam uma maior similaridade entre si do que com os elementos de qualquer um dos demais clusters do conjunto C. O conjunto C é considerado uma clusterização com k clusters caso as seguintes condições sejam satisfeitas:

$$\bigcup_{i=1}^{k} C_i = X \quad (1)$$

$$C_i \neq \emptyset, para\ 1 \leq i \leq k \quad (2)$$

$$C_i \cap C_j = \emptyset, para\ 1 \leq i, j \leq k\ e\ i \neq j \quad (3)$$

O valor de k pode ser conhecido ou não. Caso o valor de k seja fornecido como parâmetro para a solução, o problema é referenciado na literatura como "problema de kclusterização" (FASULO, 1999, p. 1) . Caso contrário, isto é, caso o k seja desconhecido, o problema é referenciado como "problema de clusterização automática" e a obtenção do valor de k faz parte do processo de solução do problema, como em (DOVAL et al, 1999, p. 1)."

Existem diversos algoritmos específicos para solucionar os problemas de kclusterização, em especial o k-means (ou k-médias), que pode ser definido como (PIMENTEL et AL, 2003, p. 3): "*O objetivo deste algoritmo é encontrar a melhor divisão de P dados em K grupos Ci, i = 1, ... K, de maneira que a distância total entre os dados de um grupo e o seu respectivo centro, somada por todos os grupos, seja minimizada."*

Para os problemas de clusterização automática, onde o valor de k é desconhecido a priori, têm-se também diversos algoritmos em especial o Expectation Maximization – EM,

desenvolvido por Dempster, Laird and Rubin, (DEMPSTER, LAIRD and RUBIN, 1977) (BORMAN, 2011) (DELLAERT, 2002). Este algoritmo é um procedimento iterativo eficiente para calcular a máxima Estimativa de Verossimilhança (Likelihood), com a utilização de dois passos a cada iteração:

- No passo E, os dados que faltam são estimados, considerando os dados observados e a estimativa atual dos parâmetros do modelo, isto é conseguido usando a expectativa condicional;
- No passo M, a função de verossimilhança é maximizada sob a suposição de que os dados que faltam são conhecidos, sendo utilizados os dados obtidos no passo E ao invés dos dados reais em falta.

Convergência deste algoritmo é garantida pelo aumento da estimativa de verrosimilhança a cada iteração. O resultado da execução deste algoritmo é uma distribuição dos dados de entrada em K clusters, resultando no valor de estimação de verossimilhança.

Para garantir valores elevados da estimativa de verossimilhança pode-se aplicar uma transformação nos dados de entrada para garantir que possuam média zero e desvio padrão unitário, esta transformação é denominada de "Estandardização" (MATOS, 2011, p. 8) que é realizada através de aplicação do seguinte procedimento a um conjunto de dados X = {X1, X2, ..., Xn} :

- Cálculo do valor médio: $\bar{X} = \sum_{i=1}^{n} X_i$ (4)
- Cálculo do desvio padrão: $S = \sqrt[2]{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ (5)
- Cálculo dos novos vetor de valores X' = {X'1, X'2, ..., X'n}, onde os seus elementos são dados por $X'_{\iota} = \frac{X_{\iota} - \bar{X}}{S}$ (6)

2.3.5) Regressão Linear:

Para a funcionalidade Estimação ou Predição tem-se disponíveis diversas técnicas (CÔRTES, PORCARO, LIFSCHITZ, 2002, p. 13): Regressão Linear; Regressão Múltipla; Regressão não linear; Regressão Logística e Regressão de Poisson.

A seguir será focada a Regressão Linear, por ser a técnica a ser utilizada neste trabalho para desenvolvimento da estratégia de seleção de obras do programa LpT .

Pode-se destacar como definição de Regressão Linear (MATOS, 1995, p.3-4):

"A regressão nasce da tentativa de relacionar um conjunto de observações de certas variáveis, designadas genericamente por Xk (k=1..p), com as leituras de uma certa grandeza Y. No caso da regressão linear, está subjacente uma relação do tipo:

Y= a + b1 X1 + b2 X2 + ... + bp Xp

onde a, b1, b2, ... bp seriam os parâmetros da relação linear procurada. O objetivo pode ser explicativo (demonstrar uma relação matemática que pode indicar, mas não prova, uma relação de causa-efeito) ou preditivo (obter uma relação que nos permita, perante futuras observações das variáveis Xk, prever o correspondente valor de Y, sem necessidade de realizar a medição).

...

As variáveis Xk são muitas vezes designadas por variáveis explicativas, uma vez que tentam explicar as razões da variação de Y.

Supondo que se dispõe de n conjuntos de medidas com as correspondentes observações, a utilização do modelo incluirá sempre uma parcela de erro. Utilizando o índice i (i=1..n) para indicar cada conjunto, ter-se-á então:

yi = a + b1 xi1 + b2 xi2 + ... + bp xip+ ei i=1..n

A regressão linear que aqui se apresenta consiste em estimar os valores dos parâmetros a, b1, b2, ... bp, através da minimização da soma dos quadrados dos desvios. Daí o nome de método dos mínimos quadrados que às vezes se utiliza nomeadamente para a regressão simples (p=1). O termo multi-regressão é usado para explicitar o caso p>1."

Para a obtenção de equações de regressão linear seria desejável que as variáveis possuam escalas e dispersões não muito diferentes, para isso pode-se utilizar as seguintes transformações para gerar novas variáveis (MATOS, 1995, p.8-9):

- Centralização ("centered"): a nova variável passa a ter média 0;
- Estandardizadas ("standardized"): a nova variável passa a ter média 0 e desvio padrão unitário (conforme apresentadao no tópico anterior); e
- Norma unitária ("unit lenght"): onde a nova variável W=[w1 w2 .. wp], é tal que a matriz W'.W apresenta diagonal unitária e os demais elementos Wij correspondem à correlação entre Xi e Xj.

Para medir a qualidade da fórmula de regressão linear obtido pode-se utilizar de alguns critérios (MATOS, 1995, p.10):

- Erro quadrático: soma quadrada do erro de estimação, ou seja: $\sum_i r_i^2 = \sum_i \left(Y_i - \hat{Y}_i\right)^2$, onde Yi é o valor da variável independente e Ŷi o seu valor estimado;
- Variância do erro: é dado por $\hat{\sigma}^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-p-1}$, onde n-p-1 é o grau de liberdade da regresão
    - Coeficiente de Determinação: $R^2 = \frac{SSR}{(SSR+SSE)}$, onde TSS = SSM + SSR + SSE, sendo:
    - $TSS = \sum_i y_i^2$, a Soma quadratíca total (Total Sum of Squares);
    - $SSM = n.\overline{Y}$, a soma quadrática devida à média (Sum of Squares

due to the Mean);

- SSR $= \sum_i (\hat{y}_i - \hat{Y}_i)^2$ a soma quadrática devida à regressão (Sum of Squares due to the Regression); e
- SSE $= \sum_i r_i^2$ a soma quadrática devida ao erro (Sum of Squares due to the Error).

3) ESTUDO DE CASO:

Nesta seção é apresentada a estratégia proposta, sendo para o seu desenvolvimento adotado como estudo de caso o programa Luz para Todos – LpT.

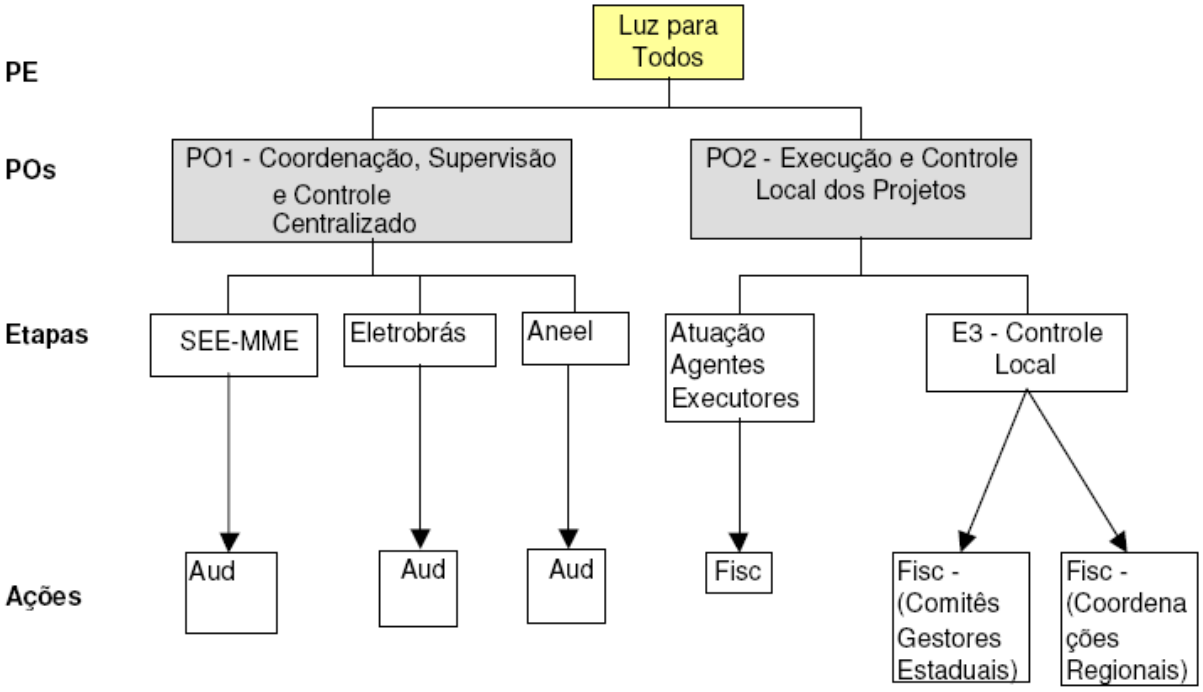
No tópico 3.1, é apresentada a base de dados do programa LpT, utilizada para avaliação de diferentes estratégias para seleção das obras a serem relacionadas na elaboração de ordens de serviço de fiscalizações de obras do programa.

No tópico 3.2 é proposta a estratégia para seleção das amostras, pela utilização de técnicas de Data Mining (mineração de dados), sendo utilizadas as técnicas de agrupamento e de regressão linear para obtenção de um rol de obras que o custo estaria divergente das demais obras no agrupamento.

3.1) A base de dados do programa Luz para Todos – LpT:

No Programa LpT, a Coordenação Geral de Auditoria das Áreas de Minas e Energia – DIENE, seguindo a metodologia de acompanhamento da execução de programas de governo – AEPG,elaborou um "Plano Estratégico" que contempla dois "Planos Operacionais", conforme diagrama a seguir:

Figura 4 - Planejamento para acompanhamento do Programa LpT.



Fonte: Fonte: Manual AEPG 2010, página 28.

Para a seleção das obras do programa LpT a serem fiscalizadas pode-se adotar a amostragem aleatória simples, neste caso a fórmula para determinação do número de amostras a serem fiscalizadas (n) para a amostragem aleatória Simples é :

$$n = \frac{Z_c \times p \times q \times N}{e^2 \times (N-1) + Z_c^2 \times p \times q} \qquad (7)$$

Onde:

N = tamanho da População;

e = Erro amostral em percentual;

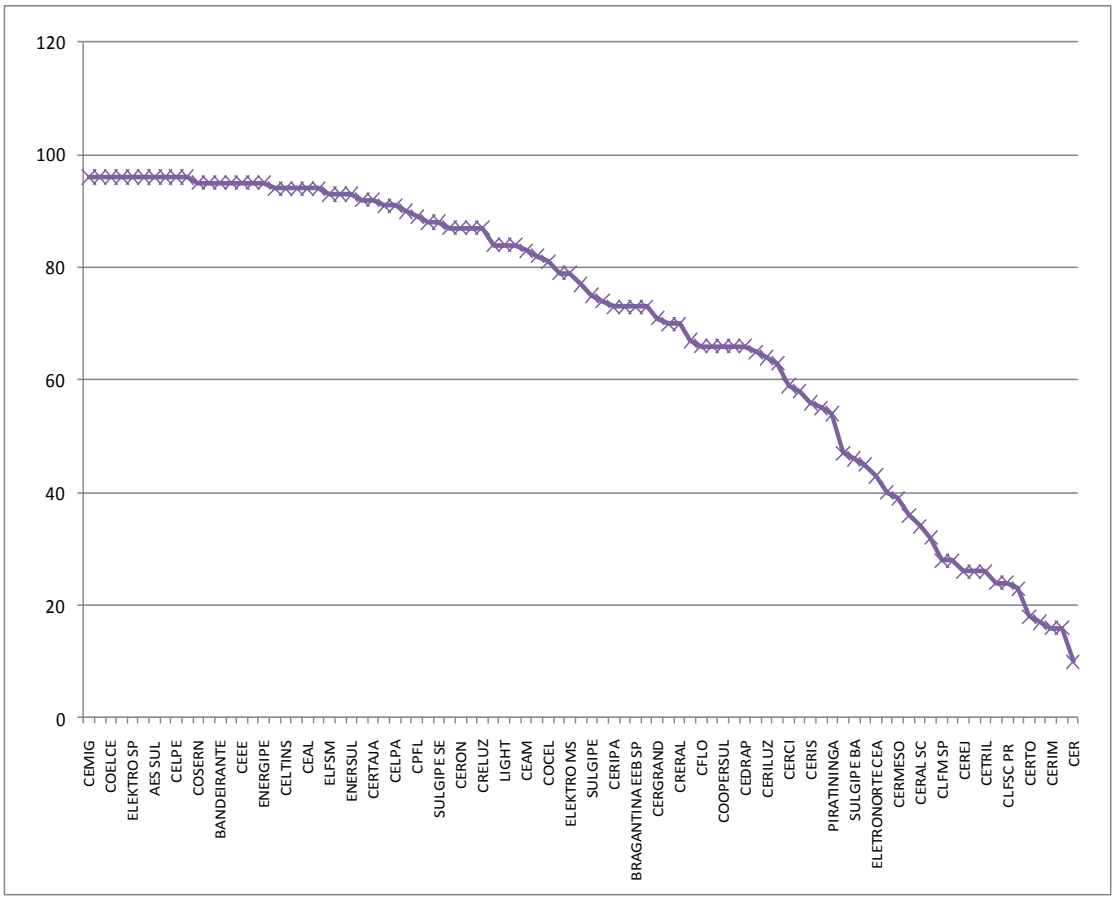
Z_c = abscissa da normal padrão (obtido a partir da definição do Intervalo de confiança);

p = Probabilidade de ocorrência positiva

q = Probabilidade de ocorrência de falha (q = 1 –p)

No Programa LpT tem-se no total 91 Distribuidoras que possuem contratos com a ELETROBRAS para utilização de recursos do programa Luz para Todos, utilizando a amostragem aleatória simples, com intervalo de confiança de 95% e erro amostral de 10% têm-se a distribuição do número de amostras por Distribuidora conforme gráfico a seguir:

Figura 5 - Número de obras a serem fiscalizadas por distribuidora.

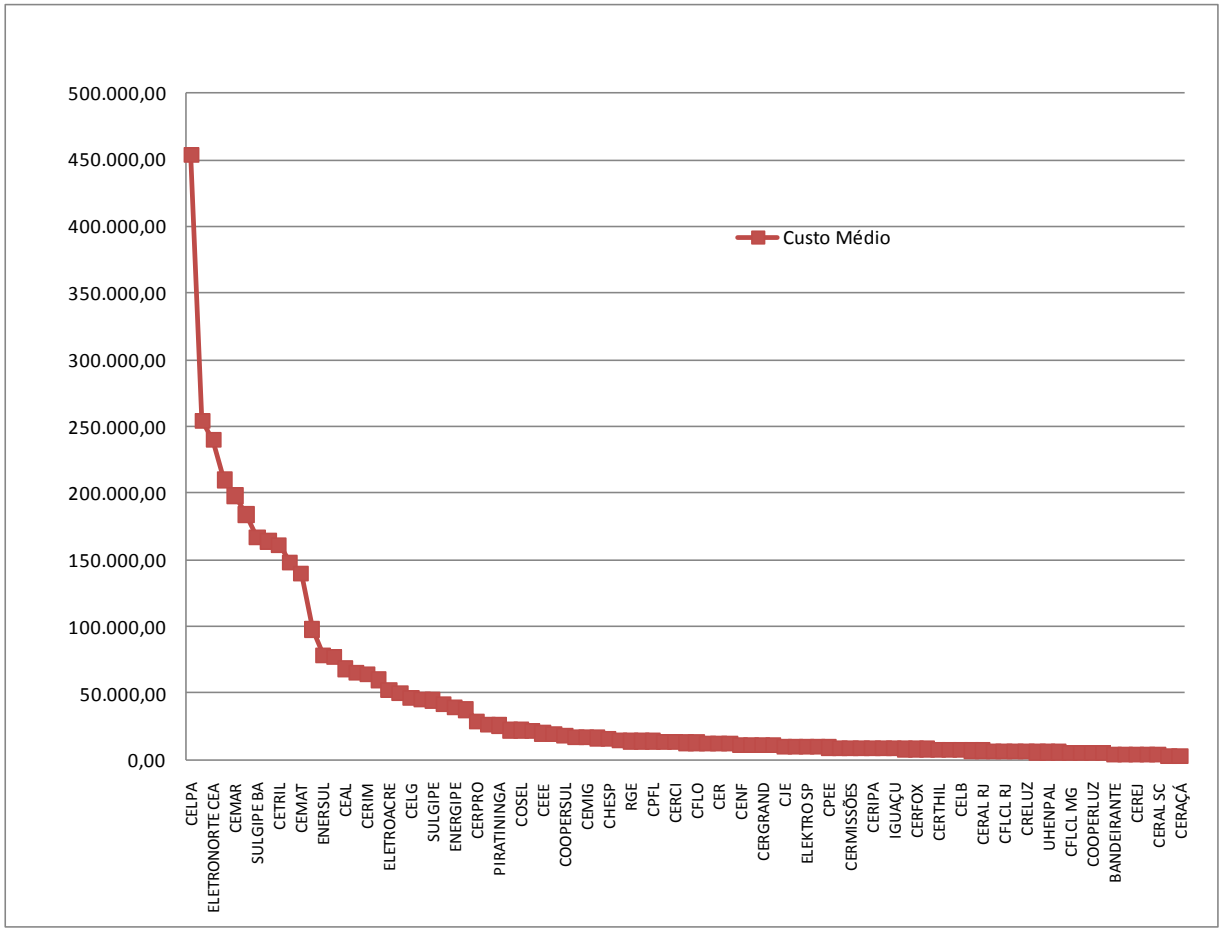


Fonte: Processamento dos dados.

Portanto, ao realizar a fiscalização de aproximadamente 10 a 96 obras por Distribuidora pode-se avaliar a utilização dos recursos públicos alocados a cada uma dessas distribuidoras, resultando na elaboração de 6.343 Ordens de Serviço para fiscalização de obras.

Pode-se gerar um gráfico com o valor do custo médio das obras por distribuidora, a partir dos custos informados pelas distribuidoras ao sistema ELETROBRÁS, resultando na figura a seguir:

Figura 6 - Custo médio de obras por concessionária.



Fonte: Processamento dos dados.

Neste gráfico é evidenciado que uma grande gama de valores médios de obras estão no intervalo de R$2.344,73 a R$453.490,30, portanto, diferentes entre si em torno de 193 vezes, isso indica que cada distribuidora adota seus próprios critérios para "dimensionamento" do tamanho das suas obras.

Portanto, o universo considerado (todas as obras do Programa LpT) não é homogêneo, pois cada obra possui "tamanho" diferente das demais e estão dispersas em áreas rurais situadas, muitas vezes, distantes da capital do estado ou mesmo de grandes centros urbanos, possuindo custos díspares.

Considerando que as obras estão dispersas na área de concessão das distribuidoras, o que implica em custos de deslocamento da equipe de fiscalização da CGU, e que o tempo necessário para fiscalizar uma obra é altamente correlacionado com o seu "tamanho".

A utilização da amostragem aleatória simples para avaliação do programa Luz para Todos - LpT pode ter implicações de custos para a CGU que devem ser muito bem avaliadas,

sendo necessário buscar alternativas que minimizem o custo de deslocamentos até as obras e também que sejam selecionadas obras que sejam críticas para a sua fiscalização.

Neste trabalho é proposto como alternativa a realização de fiscalizações para avaliação da execução de programas de governo com a utilização de amostragem aleatória simples, a adoção de auditorias especiais buscando-se avaliar o Programa LpT através um "conjunto de obras selecionadas" através de da utilização da mineração de dados, buscando reduzir as fiscalizações apenas as obras que contenham indícios de problemas

3.1.1) Importação de dados:

A CGU recebe periodicamente da ELETROBRAS os dados informados pelas Distribuidoras sobre as obras conforme tabela, a seguir:

Tabela 1 - Metadados encaminhado pela ELETROBRÁS.

Campo	Formato	Descrição
TIPO	Texto 50 caracteres	Indica se a ODI é uma obra de "SUBESTAÇÃO" ou de rede "RURAL"
CONCESSIONARIA	Texto 50 caracteres	Indica o nome abreviado da concessionária/permissionária
TRANCHE	Número Inteiro longo	Indica a qual Tranche (contrato) a ODI pertence
ECFS	Texto 50 caracteres	Indica o número do contrato celebrado entre a ELETROBRAS e a concessionária/permissionária
ANOECFS	Texto 50 caracteres	Indica o ano do contrato celebrado entre a ELETROBRAS e a concessionária/permissionária
ODI	Texto 50 caracteres	Indica o número da Ordem de Imobilização (ODI), associada à obra, é um número que identifica a obra na distribuidora.
PROJETO	Texto 255 caracteres	Indica a localização da obra no município
MUNICIPIO	Texto 255 caracteres	Indica o nome do município que foi realizada a obra
UF	Texto 10 caracteres	Indica a Unidade da Federação que foi realizada a obra
INIOBRA	Data abreviada (dd/mm/aaaa)	Indica a Data de Início da Obra
FIMOBRA	Data abreviada (dd/mm/aaaa)	Indica a Data de Final da Obra
DATAENVIO	Data abreviada (dd/mm/aaaa)	Indica a Data de digitação dos dados da obra no sistema LpT

Campo	Formato	Descrição
CONSUMIDORES	Número Inteiro longo	Indica o Número de consumidores atendidos pela ODI
KIT	Número Inteiro longo	Indica o Número de Kits de instalação interna instalados na ODI
CUSTO	Número Real	Indica o custo total da ODI
REDE AT	Número Real	Indica o comprimento total da Rede de Alta Tensão[1] instalada na ODI, em kilometros
REDE BT	Número Real	Indica o comprimento total da Rede de Baixa tensão e a rede conjugada Alta/Baixa Tensão instalados na ODI, em kilometros
POSTE	Número Inteiro longo	Indica o número de postes instalados na ODI
MEDIDOR	Número Inteiro longo	Indica o número de medidores instalados na ODI
TRAFO	Número Inteiro longo	Indica o número de transformadores instalados na ODI
EQUIP	Número Inteiro longo	Indica o número de equipamentos de rede instalados na ODI
LOTE	Número Inteiro longo	Indica o número do lote que a ODI faz parte para fins de fiscalização pela ELETROBRAS
AMOSTRA	Boleano	Indica se a ODI foi fiscalizada pela ELETROBRAS
VF	Numérico	Indica o valor da execução física da obra, que é um índice que, em termos gerais, indica a participação da obra na meta física de execução do contrato

Fonte: Análise dos dados recebidos da ELETROBRAS.

Os dados fornecidos periodicamente pela ELETROBRÁS são importados para uma base de dados de dados ACCESS, sendo colocados em uma tabela resultando em torno de 345.000 registros.

3.1.2) Limpeza de dados:

As variáveis comprimento da Rede de Alta Tensão (REDE AT), comprimento da rede de Baixa Tensão (REDE BT), número de transformadores (TRAFO), número de postes (POSTE), número de Kits de instalação interna (KIT), número de medidores (MEDIDOR) e número de equipamentos de rede (EQUIP) caracterizam as obras do programa LpT.

[1] A rigor as redes utilizadas no programa Luz para Todos são de 13,8 a 34 KVolts, portanto, a nomenclatura correta seria rede de "Média Tensão", porém, neste trabalho é utilizado o termo "Alta Tensão" para essas redes, devido à ampla utilização do termo pelos auditores da CGU, que, via de regra, não são engenheiros eletricistas.

Em normas técnicas de empresas de distribuição (COPEL, 2002) tem-se a seguinte classificação de obras de eletrificação rural: “Projeto de Rede Nova”; “Projeto de Melhoria de Rede”; “Projeto de Reforço de Rede” e “Projeto de Ampliação de Rede”.

A partir das definições acima se pode classificar as obras do Programa LpT. Esta classificação pode ser aprofundada de acordo com o componente da rede (Rede de Alta Tensão ou de Baixa Tensão) e o número de consumidores atendidos pela obra (atendimento individual ou coletivo). Na tabela a seguir é totalizado o total de obras (ou ainda de Ordem de Investimento - ODI) na base de dados utilizando estas classificações:

Tabela 2 - Classificação do “tipo de obras”.

Tipo de obra		Rede AT	Rede BT	Trafo	Poste	Kit	Medidor	Eqp	Total de ODI (0bras)
1	Reforço de Rede - Recondutoramento/adição de fases AT	>0	0	0		0	0		1.646
2	Reforço de Rede - Recondutoramento/adição de fases BT	0	>0	0		0	0		514
3	Rede Nova - Atendimento a consumidor individual		0	1			1	0	114.412
4	Rede Nova - Ligação de Novos consumidores BT	0	0				>0		59.927
5	Ampliação de Rede - Expansão de Rede AT	>0		>0	>0		>0		90.455
6	Ampliação de Rede - Expansão de Rede BT	0	>0				>0		72.920
7	Melhoria de Rede - Reforma de rede AT ou BT						>0		3.504
8	Melhoria de Rede - Adição de Equipamentos de rede	0	0	0		0	0	>0	1.529

Fonte: Análise dos dados recebidos.

Os “tipos de obras” 1, 2 e 8, por envolverem operações mais complexas, tais como

recondutoramento, adição de fases e adição de equipamentos, exigem para fiscalização que a equipe seja composta de engenheiros eletricistas, equipados inclusive com EPI, no caso de necessidade de inspeção do diâmetro dos cabos nos postes. Estes tipos de obras não serão utilizados nesse estudo de caso, pois não são factíveis de serem fiscalizadas por equipes da CGU com características de não especialistas.

O tipo de obra 7 possui valores de custos muito variáveis e sua execução é mais difícil de ser fiscalizada por auditores que não são engenheiros eletricistas, pois são obras em que são realizados acréscimos e ajustes a redes já existentes, sendo feito todo tipo de modificação, desde acréscimo de transformadores, postes, Redes AT ou Redes BT, até a simples movimentação de transformador, motivo pelo qual também serão excluídas desse estudo de caso.

Portanto, vai-se restringir, neste estudo de caso, as situações 3 a 6, que são as obras mais adequadas a serem fiscalizadas pela CGU, totalizando 337.714 obras, que corresponde a 98% do total de obras classificadas, ou seja, 344.907.

Será ainda adotado como critério de seleção, nas obras do tipo 4 a 6, que não sejam instalados equipamentos de rede, pois o custo dos equipamentos de rede afetam muito o custo global da obra, podendo distorcer os resultados obtidos neste estudo, com isso, o total de obras disponíveis para realizar o estudo é de 337.636 obras, ou seja, 97,8% do total de obras classificadas.

Ao realizar a classificação dos tipos de obras foram verificadas algumas situações que podem ser decorrentes do registro inadequado das informações no sistema da ELETROBRAS, relacionados na tabela a seguir:

Tabela 3 - Registros com inconsistências nos dados.

Critério	Total de registros
Data de Fim da obra igual à Data de Início da Obra	24.072
Numero de Kits de Instalação Interna maior do que número de Medidores	138
Número de Medidores diferente do número de Consumidores	293
Inconsistências entre os valores das variáveis de descrição da obra (mais transformadores do que postes ou mais transformadores que medidores ou kits)	210
Total de registros com indícios de problemas	**24.713**

Fonte: Processamento dos dados

A existência destas inconsistências na base de dados importada enseja a necessidade da realização de ação de controle específica para avaliar os mecanismos de controle adotados

no sistema da ELETROBRÁS para garantir a consistência dos dados registrados na sua base de dados. Registra-se que esta ação de controle não será tema deste artigo.

Como resultado desta etapa tem-se que dos 345.117 registros importados inicialmente 24.713 registros (7,2%) contem inconsistências nos seus dados e 7.193 (2,1%) registros são de obras que não são fiscalizáveis, resta, portanto, após realizar esta etapa de limpeza da base de dados 313.133 (90,7%) registros de obras para serem utilizadas no estudo de caso.

3.1.3) Enriquecimento dos dados:

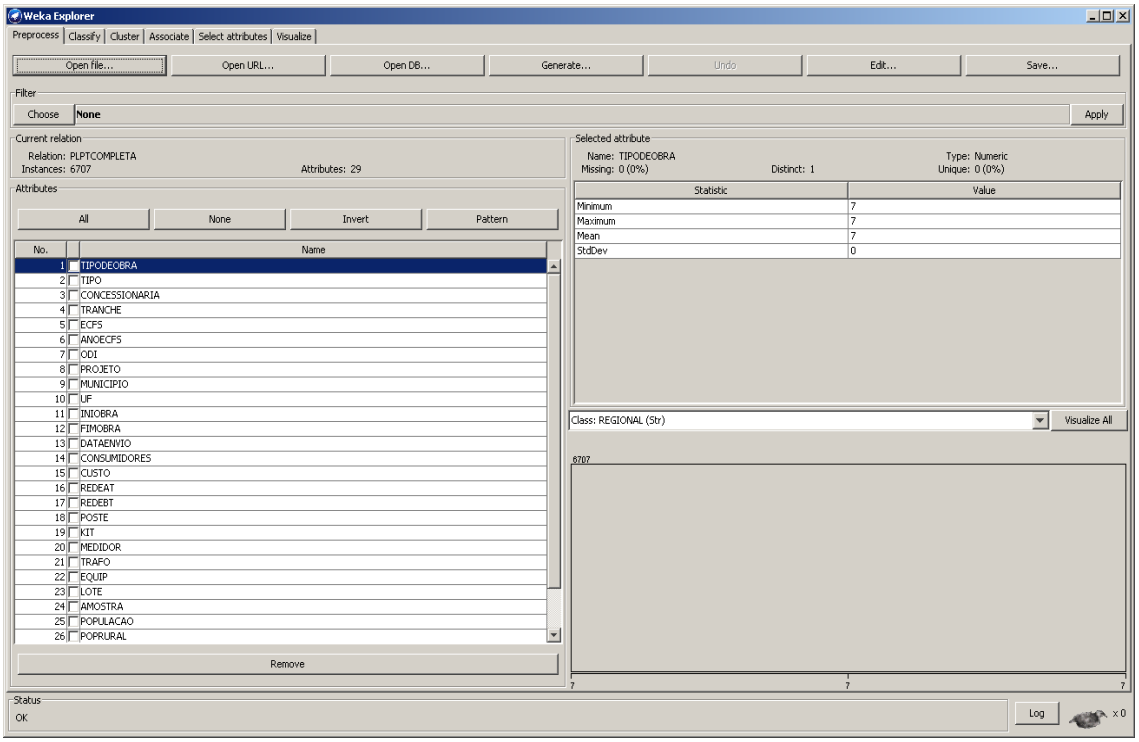
Foi buscado no site do IBGE outras informações relacionadas aos municípios que possam agregar informações a base de obras do PLPT, a informação mais apropriada foi obtida na "Tabela 2432 - Domicílios particulares permanentes por número de famílias, situação do domicílio e adequação da moradia", no site: http://www.sidra.ibge.gov.br/bda/tabela/listabl.asp?z=cd&o=13&i=P&c=2432, em relação ao censo demográfico do ano de 2000, que foi importado para a base em ACCESS. Foram importados a variável "total de domicílios permanentes na zona rural de municípios".

3.2) utilização da Mineração de Dados para seleção de obras:

A utilização de ferramentas de mineração de dados para

Para o desenvolvimento da estratégia de seleção de obras com o uso de técnicas de mineração de dados será utilizado um software livre, o "Waikato Environment for Knowledge Analysis – WEKA" (HALL et al, 2009). Este software contém diversas ferramentas para mineração de dados, permitindo a sua aplicação a um arquivo de dados que fora lido, o seu ambiente é bastante simples conforme tela a seguir, onde na parte superior se tem abas que permitem a aplicação das ferramentas, agrupadas em pre-processamento, classificação, agrupamento, seleção de atributos e visualização.

Figura 7 - Tela do software WEKA.



Fonte: Software WEKA.

Nos tópicos a seguir são desenvolvidos a aplicação das técnicas de análise de agrupamento e regressão linear aos dados importados do programa LpT, sendo apresentados os resultados obtidos em cada etapa do desenvolvimento

3.2.1) Análise de agrupamento ou Cluster (clusterização):

Para realizar o agrupamento de dados no WEKA se emprega a análise de cluster ou clusterização, que, conforme Jain and Dubes (1988) é um método que utiliza o aprendizado não supervisionado ou auto-organizável, ou seja, não há um "professor" ou "crítico" que lhe indique o que cada padrão representa.

Um algoritmo de clusterização bastante empregado é o Expectation-Maximization – EM, desenvolvido por Dempster, Laird and Rubin (Dempster et al, 1977), que consiste na busca do agrupamento dos dados de forma iterativa, buscando a máxima verossimilhança.

3.2.2) Agrupamento natural das obras por tipo, tranche e distribuidora:

Conforme o manual de operacionalização do Programa LpT (MANUAL DE OPERACIONALIZAÇÃO, 2009) cada distribuidora encaminha para análise pela ELETROBRÁS um "plano de obras" para cada "tranche" (contrato de subvenção e financiamento) a ser celebrado, isso implica que o custo de uma mesma obra é diferente para cada distribuidora, e que para uma mesma distribuidora os custos de uma mesma obra são diferentes entre os contratos celebrados com a ELETROBRAS, Na tabela a seguir, são

relacionados os valores médio, mínimo e máximo de custo de uma obra com 20 m de Rede de AT, 1 Transformador, 1 poste e 1 medidor (tipo de obra 3 – Atendimento a consumidor único) para as diferentes distribuidoras, sendo relacionado apenas as que para todas as tranches foram realizadas pelo menos 3 obras com aquelas características.

Tabela 4 - Custos de uma mesma obra para diferentes distribuidoras e tranches.

DISTRIBUIDORA	NÚMERO DA TRANCHE	CUSTO MÉDIO	CUSTO MÍNIMO	CUSTO MÁXIMO	NÚMERO DE OBRAS
CELESC	1	3.234,43	2.747,25	3.863,83	3
	2	4.019,02	2.839,56	5.421,84	20
	3	4.611,56	4.579,20	4.670,59	3
CELG	1	5.918,94	3.084,75	12.878,97	12
	2	6.082,02	3.177,74	10.795,22	35
CEMIG	1	3.496,64	2.578,48	4.397,81	104
	2	3.588,32	2.696,26	5.118,73	123
	3	4.270,90	3.569,49	5.240,94	12
ELEKTRO SP	1	3.827,54	2.712,75	4.399,16	10
	2	4.740,57	3.582,19	7.203,20	21
	3	5.350,36	4.142,52	7.860,80	17
	4	5.717,17	4.089,86	8.149,89	10
	5	5.976,64	4.906,77	7.903,55	6
ESCELSA	1	2.431,76	1.976,09	3.385,74	97
	2	3.210,58	2.285,07	10.018,73	141
	3	3.474,07	2.556,70	8.809,81	207
	4	3.602,66	2.937,90	4.337,75	13

Fonte: Processamento dos dados

Na tabela acima fica evidenciado que uma mesma obra possui custos que variam entre distribuidoras e mesmo entre tranches numa mesma distribuidora, portanto, pode-se eleger dois níveis de agrupamentos: Concessionária e tranche.

Na etapa de limpeza da base de dados foi realizada uma classificação das obras a partir da configuração dos valores das suas 6 variáveis de descrição, podem ser calculados os custos médios de cada um destes tipos de obras, resultando na tabela a seguir:

Tabela 5 - Custo médio para diferentes tipos de obras.

Tipo de obra (a)		Custo total (b)	Desvio padrão (c)	Variação % (d=c/f)	OBRAS (e)	Custo médio (f)
3	Atendimento a consumidor individual	778.065.265,51	7.669,71	104,14%	105.642	7.365,11
4	Ligação de Novos consumidores BT	126.281.036,18	22.544,33	910,98%	51.028	2.474,74
5	Expansão de Rede AT	7.532.593.244,46	194.651,26	228,54%	88.439	85.172,75

6	Expansão de rede BT	406.191.157,41	23.816,33	398,85%	68.024	5.971,29
	TOTAIS	**8.843.130.703,56**	**62.170,41**	**410,62%**	**313.133**	**25.245,97**

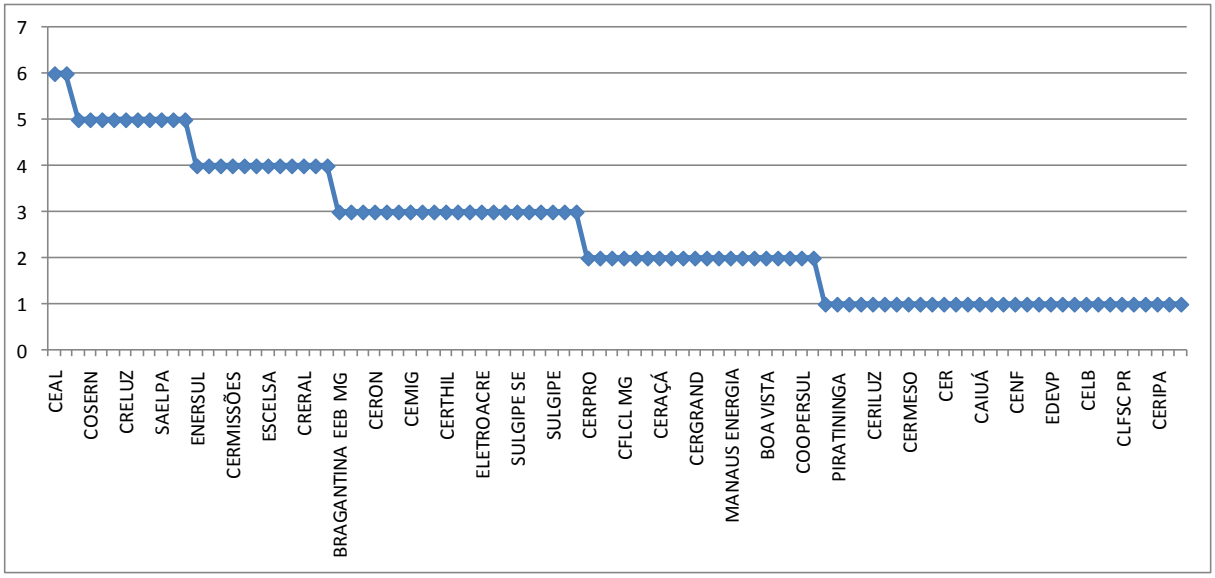
Fonte: Processamento dos dados

A partir deste quadro pode-se verificar que o tipo de obra 3 representa o maior número de obras realizadas e que os dados se encontram menos dispersos, enquanto o tipo de obra 5 apresenta os maiores custos médios e totais de execução e que os tipos de obras 4 e 6 são aqueles em que os custos totais são pouco significativos (6%), possuem muita variação nos custos e muitas obras, portanto, a partir desta tabela pode-se sugerir que sejam elaboradas estratégias distintas de análise, apropriadas a cada um dos tipos de obra.

Na base de dados tem-se 91 distribuidoras, para cada uma foram celebrados de 1 a 6 contratos de subvenção e financiamento (tranche) com suas próprias composições de custo unitário, para cada uma das tranches existem obras de todos os tipos, portanto, ao adotar 3 níveis de agrupamentos se teriam 4.368 agrupamentos (91 x 6 x 8). Considerando que só interessa no presente estudo os agrupamentos que possuam obras fiscalizáveis (tipos 3 a 6), então o universo já seria reduzido a 2.184 agrupamentos (91 x 6 x 4).

De acordo com o manual de operacionalização do Programa LpT (MANUAL DE OPERACIONALIZAÇÃO, 2009) para que uma distribuidora possa receber recursos de um novo contrato de subvenção e financiamento (tranche) é necessário que o "avanço físico" da tranche anterior seja de pelo menos 60%, isso implica que na base de dados se tenha poucas distribuidoras com muitas tranches contratadas, o que pode ser visto no gráfico a seguir:

Tabela 6 - Número de tranches por distribuidora.



Fonte: Base de dados importada.

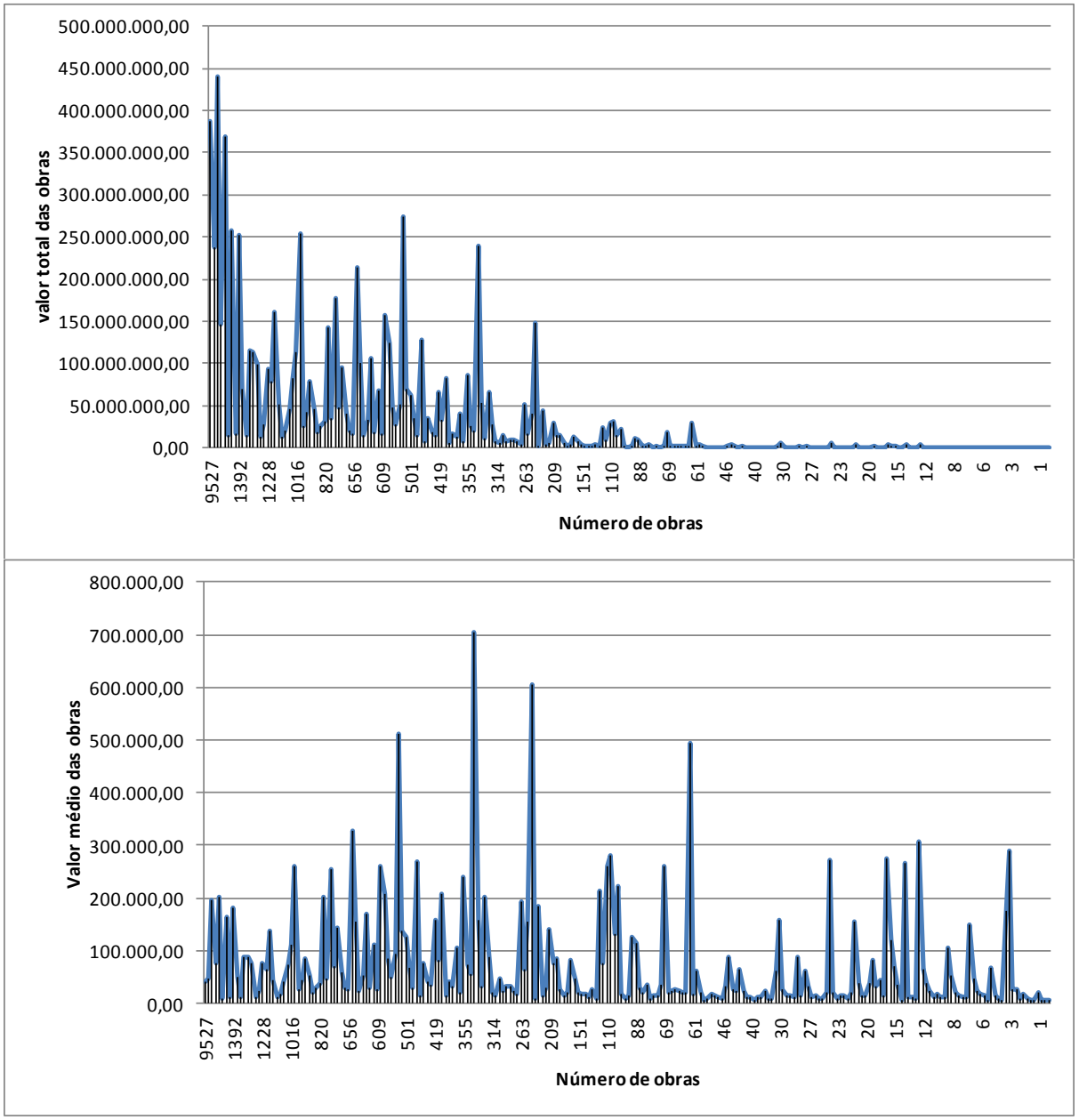
Adotando, portanto, 3 níveis de agrupamento (concessionária, tranche e tipo de obra) têm-se 852 agrupamentos disponíveis para desenvolver a estratégia de uso da ferramenta agrupamento do software WEKA para seleção de obras cujo custo esteja inconsistente em relação às demais obras no cluster.

3.2.3) A seleção de obras de expansão de rede AT:

Conforme apontado no tópico anterior, as obras de expansão de rede AT (obras do tipo 5) possuem os maiores custos médios ensejando a necessidade de desenvolvimento de uma estratégia para seleção de obras a serem fiscalizadas específica para este tipo de obra, já que devido a esse custo elevado é, por hipótese, mais elevado para a CGU a execução de fiscalização deste tipo de obra, envolvendo uma maior despesa para deslocamentos e alocação de equipes.

Pode-se selecionar nos 852 agrupamentos os que possuem somente o tipo de obra 5 (expansão de rede AT), totalizando 236 agrupamentos, a seguir tem-se um gráfico com os valores totais das obras em cada agrupamento e no outro gráfico os valores médios das obras, sempre ordenados pelo número de obras de forma decrescente.

Figura 8 - Gráficos de Valor total e valor médio de obras para os agrupamentos.



Pode-se calcular a média, o desvio padrão para os dados apresentados nestes gráficos, resultando na tabela a seguir:

Tabela 7 - Medidas de valores centrais e dispersão.

	CUSTO TOTAL	CUSTO MÉDIO
Valor Médio – VM	31.917.767,99	73.122,80
Desvio Padrão – DP	66.702.519,58	99.790,52
Coeficiente de variação (DP/VM) em percentual	208,98%	136,47%

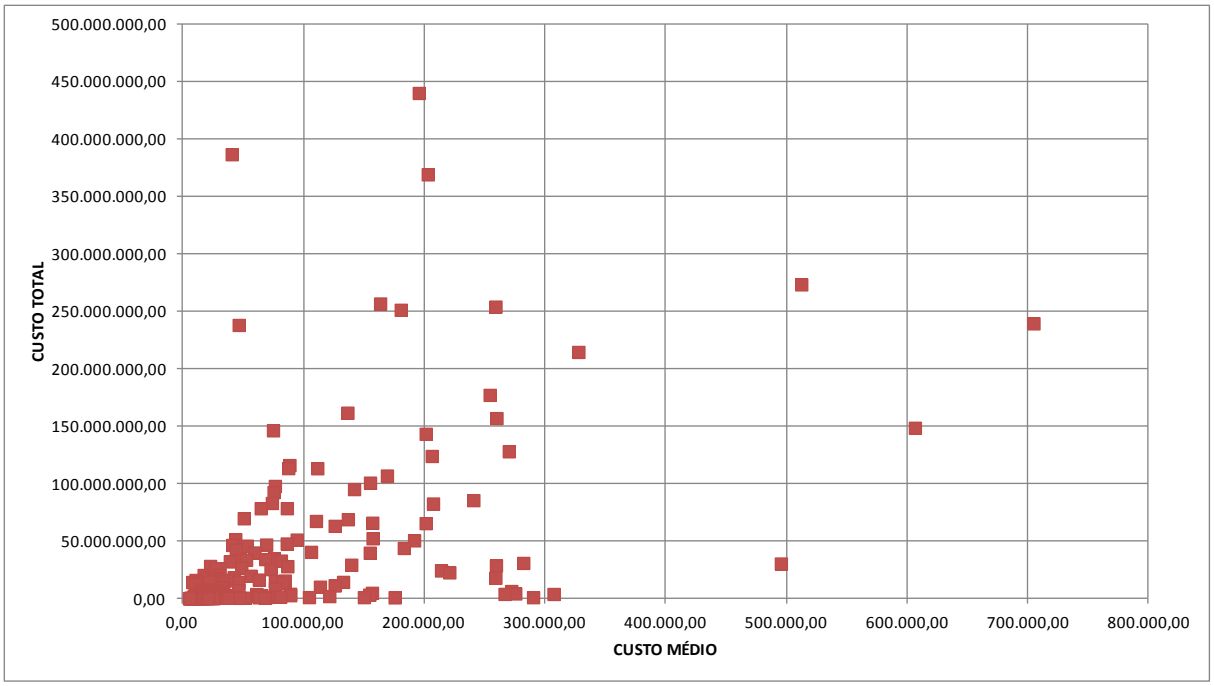
Fonte: Processamento dos dados

Nesta tabela evidencia-se que tanto o custo total como o custo médio dos

agrupamentos se encontram bastante dispersos, dificultando o uso de ferramentas estatísticas para seleção dos agrupamentos para o desenvolvimento da estratégia.

Buscando estabelecer uma estratégia para avaliar os 236 agrupamentos de obras de expansão de rede AT, pode-se gerar um gráfico de dispersão, comparando os custos totais de cada agrupamento com os seus custos médios, conforme gráfico a seguir:

Figura 9 - Gráfico de dispersão com custo total x custo médio.



Fonte: processamento dos dados.

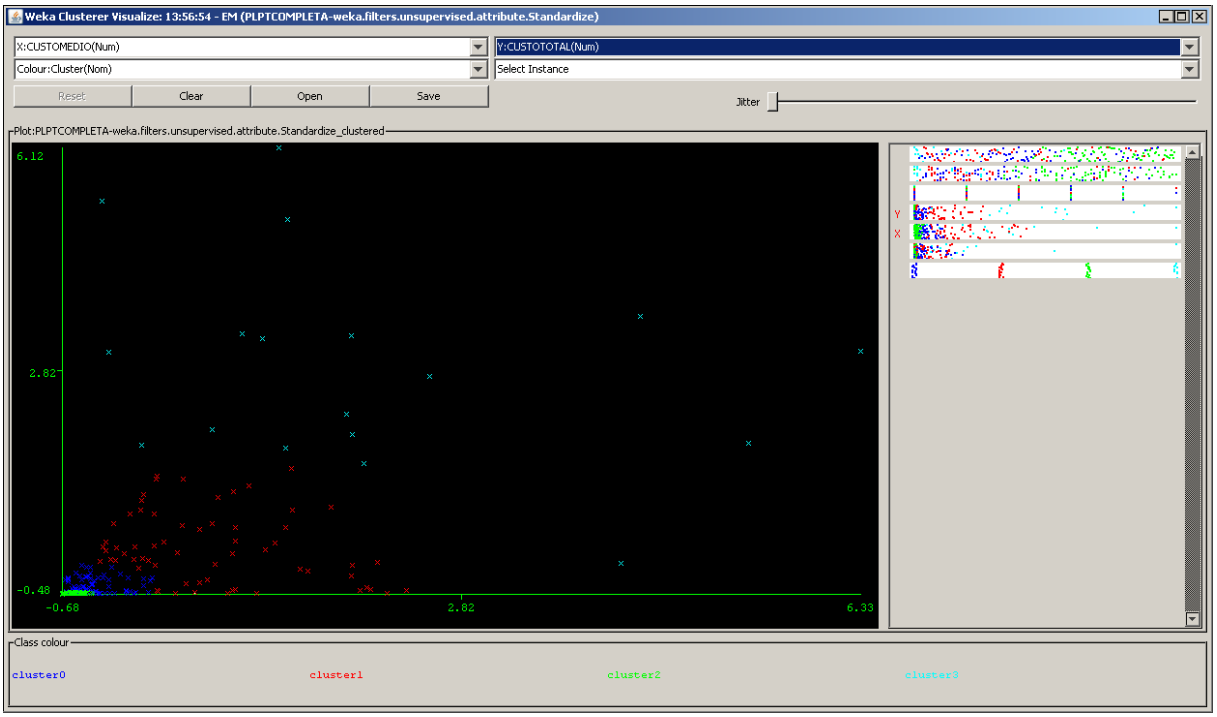
A partir deste gráfico pode-se verificar que grande parte dos agrupamentos possui os custos médios abaixo de R$300.000,00 e os custos totais abaixo de R$200.000.000,00, apontando a possibilidade de dividir os agrupamentos em pelo menos 2 conjuntos, sendo um conjunto formado pelos agrupamentos que estão mais dispersos e o outro conjunto formado com os agrupamentos mais homogêneos.

Pode-se utilizar a ferramenta de cluster do WEKA para estabelecer os conjuntos, para isso são realizados os seguintes passos:

- Aplicação de uma ferramenta de "estandartização" aos dados, visando diminuir a diferença entre as escalas dos dados de custos totais e custos médios para favorecer a aplicação do software EM;
- Aplicação da ferramenta de clusterização EM do WEKA adotando como "semente" o valor 0.0 e permitindo a escolha do número de clusters automático, resultando no índice logaritmo de verossimilhança (log likelihood) de 0,04095;

- Plotagem dos resultados obtidos, em um gráfico de dispersão com os valores do custo total no eixo Y e os valores dos custos médios no eixo X, tendo o cluster a que pertence o agrupamento identificado com as cores azul (cluster 0), vermelho (cluster 1), verde (cluster 2) e ciano (cluster 3).

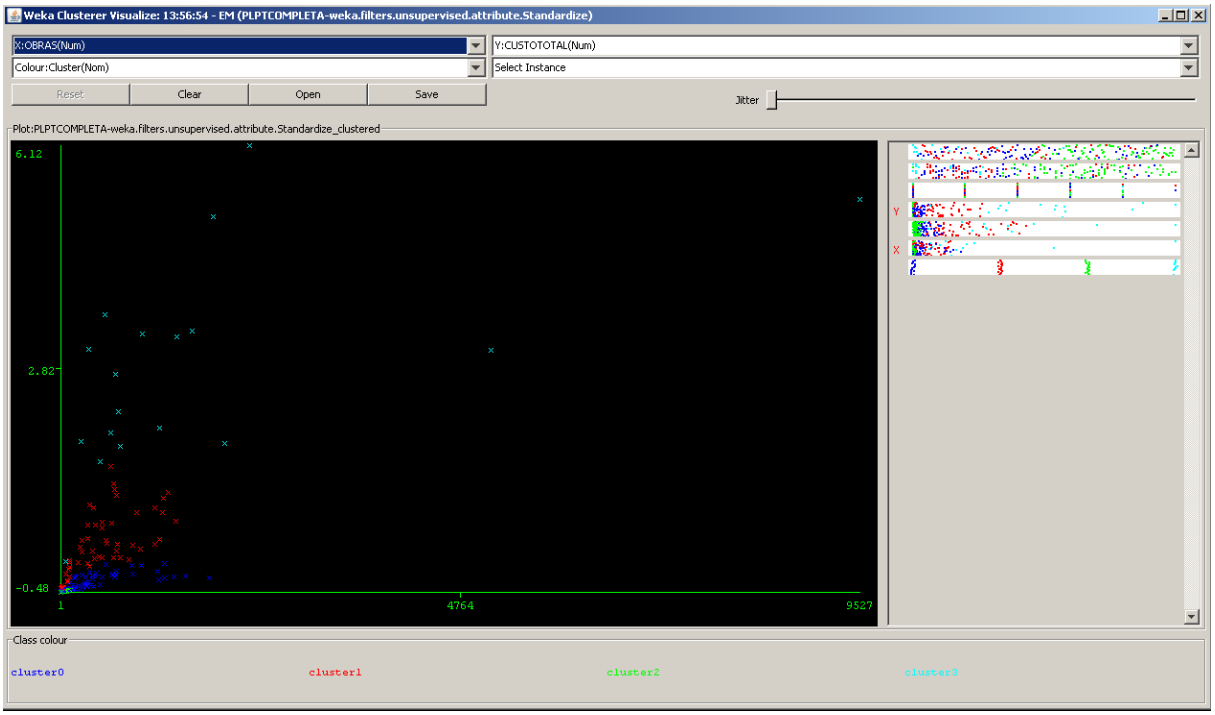
Figura 10 - Gráfico de dispersão com indicação dos clusters associados a cada agrupamento.



Fonte: Tela capturada do software WEKA.

Conforme gráfico acima, os clusters 1 e 3 formam um conjunto em que os valores dos custos totais e médios estão mais heterogêneos do que os agrupamentos alocados nos clusters 0 e 2. Pode-se ainda comparar o número de obras com o custo total das obras, resultando no gráfico a seguir:

Figura 11 - Gráfico de dispersão do Número de obras pelo cuso total com indicação do cluster associado.



Fonte: Tela capturada do software WEKA.A partir deste gráfico pode-se observar que no cluster 3 se tem os agrupamentos com maior número de obras, Também pode ser verificado que os agrupamentos no cluster 0 e 1 possuem número de obras bastante similares, sendo a grande diferença entre eles os custos totais. Para melhor caracterizar os clusters a seguir são apresentadas as medidas dos valores centrais e da dispersão dos seus dados, além do número de obras em cada cluster:

Tabela 8 - Medidas de valores centrais e dispersão dos clusters.

CLUSTER	AGRUP (%)	NÚMERO DE OBRAS (MÍNIMO E MÁXIMO)	CUSTO TOTAL			CUSTO MÉDIO		
			Média	Desvio Padrão	Coef variação	Média	Desvio Padrão	Coef variação
0	71 (30%)	5/1773	-0,3480	0,1160	-33,3%	.-0,3818	0,2018	-52,9%
1	63 (27%)	3/1376	0,1997	0,5448	272,8%	0,5807	0,7356	126,7%
2	84 (36%)	1/123	-0,4714	0,0068	-1,4%	-0,5845	0,0645	-11,0%
3	18 (08%)	61/9527	2,6345	1,6538	62,8%	1,9703	1,8040	91,6%
TOTAL	**236(100%)**							

FONTE: Processamento dos 236 clusters no software WEKA.

Da tabela e dos gráficos acima pode-se verificar que os agrupamentos no cluster 2 possuem os menores custos totais e médios, além de poucas obras em cada agrupamento, o

que implica em ser um grupo mais homogêneo do que os demais, embora possua um número maior de agrupamentos (84),

Também pode ser observado que a principal diferença entre os clusters 0 e 1 são os custos totais e médios das obras, pois estes dois clusters possuem agrupamentos com número de obras bastante similares.

O cluster 3 possui agrupamentos com custos totais, custos médios e número de obras bem mais elevados, se comparados aos demais, o que pode ser um indício de que essas obras possuem características que impliquem em investimentos maiores, ou ainda que exista sobrepreço, sendo portando o cluster que possui maior possibilidade de existência de problemas, portanto, dado a avaliação das características dos clusters obtidos após a aplicação da ferramenta de clusterização EM pode-se selecionar para estudo os 18 agrupamentos alocados nesse cluster 3, conforme tabela a seguir:

Tabela 9 - Agrupamentos selecionados para estudo.

INSTÂNCIA	DISTRIBUIDORA	TRANCHE	CUSTO TOTAL	CUSTO MÉDIO	OBRAS
0	CEMIG	2	386.994.442,78	40.620,81	9.527
1	CEMIG	1	238.186.602,67	46.412,04	5.132
2	COELBA	4	440.394.066,79	195.556,87	2.252
3	CELG	2	146.338.046,32	74.814,95	1.956
4	COELBA	5	369.450.626,59	202.994,85	1.820
6	COELBA	2	256.875.196,63	163.510,63	1.571
8	COELBA	3	251.411.246,94	180.611,53	1.392
18	COELBA	1	161.708.957,27	136.463,26	1.185
25	CEMAR	3	254.154.576,23	259.077,04	981
33	CEPISA	2	143.313.274,43	201.282,69	712
35	CEMAR	2	177.220.019,52	254.261,15	697
41	CEMAR	4	214.662.144,81	327.728,47	655
49	CEMAT	2	156.981.069,70	259.902,43	604
54	CELPA	1	273.564.596,84	512.293,25	534
59	CEMAT	4	128.029.117,01	270.103,62	474
75	CELPA	3	239.570.687,97	704.619,67	340
91	CELPA	2	148.570.183,70	606.408,91	245
135	CEMAR	5	30.220.221,67	495.413,47	61
VALOR TOTAL			**4.017.645.077,87**		**30.138**
VALOR MÉDIO				**274.004,20**	**1.674**

Fonte: Processamento dos dados

Resta ainda estabelecer uma estratégia para selecionar, dentre as obras em cada um dos agrupamentos acima que devem ser fiscalizadas por possuírem características que divergem de um padrão.

Para selecionar as obras em cada agrupamento do quadro acima é adotada como estratégia a utilização da ferramenta regressão linear do WEKA, sendo realizados os seguintes passos:

- Seleção para cada um dos 18 arquivos de agrupamentos listados na tabela 8 as

6 variáveis de descrição da obra e a variável do custo, portanto removendo das demais variáveis do arquivo;

- Executar a ferramenta Regressão Linear do WEKA (linha de comando: weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8), utilizando a variável "custo das obras" como valor a ser estimado;
- Copiar a equação da regressão linear do custo, para o quadro a seguir:

Tabela 10 - Equações de Regressão Linear para os agrupamentos no cluster 3.

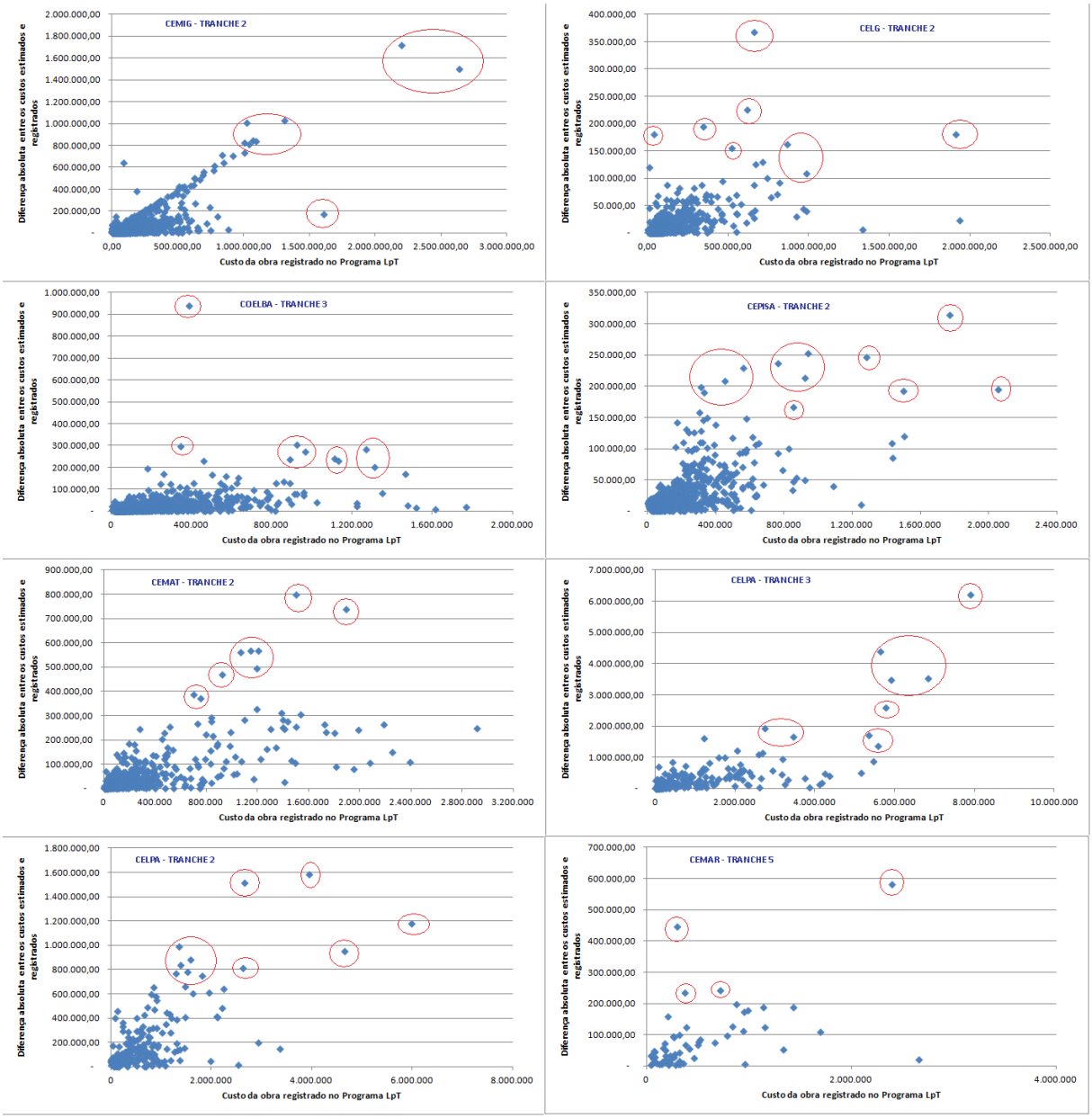
DISTRIBUIDORA	TRANCHE	EQUAÇÃO DE REGRESSÃO LINEAR
CEMIG	2	7332,8484 * REDEAT + 86179,7777 * REDEBT + 242,1752 * POSTE + 574,1964 * KIT - 1972,8897 * MEDIDOR + 4738,3073 * TRAFO - 17,4591
CEMIG	1	5346,9728 * REDEAT + 3290,7219 * REDEBT + 502,1661 * POSTE + 534,2085 * KIT + 538,4167 * MEDIDOR + 2312,0967 * TRAFO + 764,524
COELBA	4	3581,8579 * REDEAT - 9287,9123 * REDEBT + 1787,0964 * POSTE + 235,3865 * KIT + 723,023 * MEDIDOR + 5427,5035 * TRAFO + 2922,6888
CELG	2	488,4759 * REDEAT + 2796,9585 * REDEBT + 1169,865 * POSTE - 155.757 * KIT + 2056,406 * MEDIDOR + 1726,4461 * TRAFO - 14.3906
COELBA	5	10440,7665 * REDEAT + 6831,3414 * REDEBT + 1006,8041 * POSTE – 2278,5519 * KIT + 4664,0000 * MEDIDOR + 485,3566
COELBA	2	-5446,9508 * REDEAT - 27984.5113 * REDEBT + 2675.7382 * POSTE + 1387.4898 * MEDIDOR + 2247.9898 * TRAFO - 7039.6215
COELBA	3	7751,4504 * REDEAT + 2990,156 * REDEBT + 1405,9709 * POSTE + 738,0569 * MEDIDOR – 744,0706 * TRAFO + 10410,1631
COELBA	1	10665,5374 * REDEAT + 4842,3760 * REDEBT + 635,7502 * POSTE + 452,9057 * KIT + 496,2734 * MEDIDOR + 5475,3485 * TRAFO - 2715.2122
CEMAR	3	4199,4771 * REDEAT + 7057,9978 * REDEBT + 876,2972 * POSTE + 380,0539 * KIT + 314,2782 * MEDIDOR + 4456,7878 * TRAFO - 132.7688
CEPISA	2	5006,5866 * REDEAT – 6338,9981 * REDEBT + 855,2558 * POSTE + 419,2071 * MEDIDOR + 4962,4828 * TRAFO – 15586,977
CEMAR	2	7044,9280 * REDEAT – 18900,7340 * REDEBT + 2138,7200 * POSTE + 501,6517 * MEDIDOR - 3843.6887 * TRAFO - 9387.1387
CEMAR	4	6916,9738 * REDEAT + 12088,7382 * REDEBT + 1138,3115 * POSTE – 876,4409 * KIT + 1376,8018 * MEDIDOR + 2555,7238 * TRAFO - 15338.4539
CEMAT	2	20686,6283 * REDEBT + 1308,2043 * POSTE + 1909,2648 * TRAFO + 1231,3904
CELPA	1	-6184,3754 * REDEAT -21126,2944 * REDEBT +2633,1073 * POSTE +

		869,2421 * MEDIDOR – 1065,1969 * TRAFO – 13636,8381
CEMAT	4	1073,6649 * REDEAT + 33853,8951 * REDEBT + 867,6289 * POSTE - 1318.0693 * KIT + 1851,0700 * MEDIDOR + 3474,0858 * TRAFO - 6345.2292
CELPA	3	5303,6289 * REDEAT + 1245,6674 * POSTE + 5043,4198 * KIT - 4098.2354 * MEDIDOR + 3247,2391 * TRAFO - 11231.8798
CELPA	2	-7299,0505 * REDEAT – 35974,8672 * REDEBT + 2640,6317 * POSTE – 7191,6685 * KIT + 7626,7369 * MEDIDOR + 44446,6436
CEMAR	5	-6079,9548 * REDEAT + 2751,972 * POSTE + 1820,8939 * TRAFO – 28982,5557

Fonte: Processamento dos dados no WEKA com regressão Linear

A partir da tabela acima foram sorteados, de forma aleatória, 8 agrupamentos para elaboração de um gráfico de dispersão, onde se coloca no eixo Y a diferença absoluta entre o valor estimado e o custo da obra e no eixo X o custo da obra, obtendo os gráficos a seguir, onde são destacadas com círculos vermelhos as obras a serem selecionadas para fiscalização:

Figura 12 - seleção de obras para fiscalização.



Fonte: Processamento de dados.

Portanto, a partir dos gráficos acima se pode evidenciar que é possível à utilização de técnicas de mineração de dados para seleção de obras para fiscalização no programa Luz para Todos - LpT.

A partir dos gráficos acima se pode inferir que um critério para escolha das obras a serem fiscalizadas seja a razão entre os valores do eixo Y (valor absoluto da diferença entre o custo estimado e o custo registrado) e os valores do eixo X (valor do custo registrado) maior do que um determinado valor de referencia.

Para a busca do valor de referencia para escolha das obras seria necessário obter os gráficos para os demais agrupamentos, buscando estabelecer um valor que seja uma relação

de compromisso entre o número de obras selecionadas de cada agrupamento, o que pode ser objeto de estudo em trabalho futuro.

3.3) A estratégia proposta:

A partir dos resultados obtidos nos tópicos anteriores podem-se listar os seguintes passos adotados para a seleção de obras a serem fiscalizadas:

1) Importar os dados a partir de planilhas ou Banco de Dados;
2) A partir do banco de dados importado, realizar a limpeza, que consiste nas seguintes etapas:
    a. Classificação das obras, para seleção da que possam ser "fiscalizáveis";
    b. Exclusão das obras que possuam inconsistências nos dados;
3) Fazer um enriquecimento do banco de dados com dados dos municípios;
4) A partir do banco de dados "prontos", fazer agrupamentos tomando como base nos campos "concessionária", "tranche" e "tipo de obra";
5) Selecionar um tipo de obra para reduzir o total de agrupamentos a serem trabalhados;
6) Rodar a ferramenta cluster EM no software WEKA para selecionar o cluster com os maiores valores de custo total e custo médio para estudo;
7) Rodar a ferramenta de equações de regressão linear do software WEKA para obter o valor estimado para variável "custo", com base nas variáveis: "Rede AT", "Rede BT", "Kit"; "Medidor"; "Poste" e "Trafo" para cada um dos agrupamentos do cluster selecionado;
8) Gerar um gráfico de dispersão com os valores absolutos da diferença entre o custo estimado e o custo registrado no eixo Y e no eixo X com o custo da obra.
9) Selecionar para fiscalização as obras em que a razão entre a diferença absoluta entre o custo estimado e o custo registrado e o custo registrado for significante;

4) RESULTADOS OBTIDOS:

A base de dados do programa LpT importada para o ACCESS é composta de 345.117 obras. Como resultado da etapa de limpeza da base de dados tem-se que 24.713 registros (7,2%) contem inconsistências nos seus dados.

Considerando apenas as obras que são "fiscalizáveis" (tipos 4 a 7), se tem disponíveis para o estudo o total de 313.133 obras, que corresponde a 90,7% do total de obras importadas (7.193 registros (2,1%) não são fiscalizáveis, por se tratar de obras que necessitam de

equipamentos e mão de obra especializada para a sua fiscalização).

Ao programar a estratégia proposta na base de dados do programa LpT foram obtidos 852 agrupamentos, que ordenados por valores tem-se no máximo R$440.394.066,79 e no mínimo R$510,72, ou ainda, ordenando por ODI, o máximo de 17.447 ODIs e o mínimo de 1 ODIs em cada agrupamento.

Buscando desenvolver a estratégia foi eleito o tipo de obras nº 05 (expansão de rede AT) para o estudo de caso, sendo reduzido o elenco de agrupamentos a serem trabalhados para 236 agrupamentos.

Para a avaliação da estratégia aqui sugerida é necessário que sejam eleitas uma concessionária, uma tranche e um tipo de obra para que sejam selecionadas as obras com indícios de problemas, em seguida, é necessário que sejam elaboradas Ordens de Serviço para a fiscalização das obras e a partir do resultado da fiscalização comparar o índice de acertos da metodologia proposta.

5) CONSIDERAÇÕES FINAIS:

Inicialmente é necessário a fazer um mapeamento pela CGU dos riscos dos programas do governo federal para a escolha do processo de acompanhamento de programas de governo mais apropriado, seja através dos trabalhos de avaliação de programas de governo ou de processos de auditorias especiais.

Para o processo de auditorias especiais a principal contribuição deste trabalho é a proposição de uma estratégia para seleção de um conjunto de obras para instrumentalizar ordens de serviço para fiscalização do Programa Luz para Todos – LpT, com o foco em reduzir drasticamente o esforço para obtenção de evidências em ações de controle.

Como trabalho futuro prevê-se que seja selecionada uma distribuidora para que seja executada a estratégia aqui proposta e, após a realização das Ordens de Serviço, seja possível a sua validação.

Um aperfeiçoamento que deve ser considerado é a utilização de um conjunto de dados de “treino” para obtenção das fórmulas de regressão, para daí a sua aplicação a massa de dados. O desafio é a necessidade de validação de um conjunto de dados para que seja extraída uma fórmula de regressão que permita a comparação de custos de forma adequada.

A estratégia aqui proposta pode ser expandida para o desenvolvimento de uma metodologia que possa ser utilizada para avaliação do ambiente competitivo de processos licitatórios, através do mapeamento de descontos obtidos nos diversos processos licitatórios,

considerados os portes das obras e o número de licitantes presentes, bem como o ganho de escala nos processos licitatórios em que houve competição.

A partir do desenvolvimento de uma metodologia pode-se projetar como aplicações a identificação de indícios de inconsistências entre itens correlacionados, tais como, cotação de preços unitários muito menores ou muito maiores em determinados itens, mantendo o preço global competitivo, indicando que determinado licitante poderia deter o conhecimento de quais itens seriam mais utilizados, quando da execução do contrato, o que constitui uma burla a isonomia entre os licitantes.

Uma metodologia baseada na mineração de dados pode ser utilizada para identificar ordens bancárias ou diárias/passagens emitidas que apontam, por exemplo, viagens com periodicidade definida (em finais de semana ou feriados) ou ainda coincidência com outros eventos tais como visitas de autoridades, terceiros a administração pública (empreiteiros, lobistas) ou mesmo apresentação/votação de projetos de leis no congresso ou em assembleias estaduais, constituindo indícios que podem ser utilizados para abertura de ordens de serviço.

Pode-se ainda visualizar como aplicação futura da metodologia ou mesmo da estratégia proposta, no âmbito das coordenações da CGU, o seu uso para comparação de custos unitários em planilhas com grande número de itens e diversos concorrentes em licitações, buscando evidenciar jogo de planilhas; correlação entre deslocamentos de servidores e liberação de pagamento de empenhos a empresas contratadas ou ainda a liberação de parcelas de convênios, buscando evidenciar práticas de tráfico de influência e extração de informações sobre prazos para realização das etapas de julgamento de licitações, correlação entre preços entre processos licitatórios e análise de vínculos entre empresas licitantes buscando evidenciar contratações inadequadas.

6) REFERENCIAS BIBLIOGRÁFICAS:

1. MANUAL DE OPERACIONALIZAÇÃO**, Programa Nacional de Universalização do Acesso e Uso da Energia Elétrica,** ELETROBRAS/MME**,** Brasília-DF**,** Revisão nº 6, 2009.

2. MANUAL AEPG, **Metodologia para o Acompanhamento Sistemático da Execução de Programas de Governo**, CGU/PR, Brasília-DF, 2010.

3. BRASIL, **Instrução Normativa SFC n.º 01**, CGU/PR, Brasília-DF, 2001.

4. RAMEZ, Elmasri & NAVATHE, Shamkant B., **Sistemas de Banco de Dados**, Editora Pearson, São Paulo, 4ª edição, 2010

5. ADRIAANS, P. & ZANTINGE, D. **Data Mining**, Editora Addison-Wesley, 1996

6. SCHEEL, Jaison Carlos. **Técnicas do Data Mining**, 2008, Disponível em http://pt.shvoong.com/books/1778553-t%C3%A9cnicas-data-mining-dataminig-minera%C3%A7%C3%A3o/\". Acesso em 03/07/2011

7. BRANDÃO, M.F.R., TRÓCOLI, B.T., RAMOS, C.R.S., **Análise de agrupamento de escolas e Núcleos de Tecnologia Educacional: mineração na base de dados de avaliação do Programa Nacional de Informática na Educação**, in: XIV Simpósio Brasileiro de Informática na Educação - NCE - IM/UFRJ, 2003, p. 366-374.

8. HAN, Jiawei & KAMBER, Micheline, **Data Mining – Concepts and Techniques,** Morgan Kaufmann Publishers, Inc, 2001, Disponível em <http://www.cs.sfu.ca>. Acesso em 01/10/2011

9. OCHI, L. S., DIAS, C. R., SOARES, S. F., **Clusterização em Mineração de Dados**. Programa de Pós-Graduação em Computação - UFF, RJ, 2004. Obtido em <http://www.ic.uff.br/~satoru/conteudo/artigos/ERI-Minicurso-SATORU.pdf> Acesso em 18/10/2011.

10. FASULO, D. **An Analysis of Recent Work on Clustering Algorithms,** Technical Report, Dept. of Computer Science and Engineering, Univ. of Washington, 1999

11. DOVAL, D., MANCORIDIS, S. & MITCHELL, B. S. **Automatic Clustering of Software Systems using a Genetic Algorithm**. In 1999 International Conference on Software Tools and Engineering Practice (STEP '99), 1999.

12. PIMENTEL, E. P., FRANÇA, V. F., & OMAR, N. , **A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização**. In Anais do Simpósio Brasileiro de Informática na Educação, Rio de Janeiro, RJ. SBC, 2003

13. DEMPSTER, A. P.; LAIRD, N. M. and RUBIN D. B., **Maximum likelihood from in-complete data via the EM algorithm**. Journal of the Royal Statistical Society: Series B, 39(1):1–38, November 1977.

14. BORMAN, Sean, **The Expectation Maximization Algorithm - A short tutorial**, Disponivel em <http://www.seanborman.com/publications/>, Acesso em 10/11/2011.

15. DELLAERT, Frank, **The Expectation Maximization Algorithm - Technical Report number GIT-GVU-02-20**, emitido em fevereiro de 2002. Disponível em <http://www.cc.gatech.edu/~dellaert/em-paper.pdf >. Acesso em 16/11/2011,

16. MATOS, Manuel António, **Manual Operacional para a Regressão Linear**, FEUP, 1995. Disponível em <http://paginas.fe.up.pt/~mam/regressao.pdf>. Acesso em 16/11/2011.

17. CÔRTES, Sérgio da Costa, PORCARO, Rosa Maria and LIFSCHITZ, Sérgio, Mineração de Dados – Funcionalidades, Técnicas e Abordagens, PUC-RioInf.MCC10/02 Maio, 2002, Disponível em ftp://139.82.16.194/pub/docs/techreports/02_10_cortes.pdf. Acesso em 12/11/2011.

18. WEIS, Sholom M. & INDURKHYA, Nitim, **Predict Data Mining,** Morgan Kaufmann Publishers, Inc, 1999.

19. BERRY, Michael J. A. & LINOFF, Gordon, **Data Mining Techiques for Marketing, Sales, and Customer Support**, John Wiley & Sons, Inc., 1997.

20. MENA, Jesus, **Data Mining Your Website**, Digital Press, 1999

21. THURAISINGHAM, Bhavani, **Data Mining**, CRC Press, 1999

22. WESTPHAL, Christopher & BLAXTON, Teresa, **Data Mining Solutions – Methodos and Tools for Solving real-Word Problems**, John Wiley & Sons, Inc., 1998.

23. COPEL, **NTC 831001- Projeto de Redes de Distribuição Rural**, COPEL DISTRIBUIÇÃO, 4ª edição, 2002.

24. BATISTA, P. R. L., **Data Mining na Identificação de Atributos Valorativos da Habitação**, Secção Autónoma de Ciências Sociais Jurídicas e Políticas, Universidade de Aveiro, 2010.

25. HALL, Mark, EIBE, Frank, GEOFFREY, Holmes, BERNHARD, Pfahringer, PETER, Reutemann & IAN H. Witten; **The WEKA Data Mining Software: An Update**, SIGKDD Explorations, Volume 11, Issue 1, 2009.