UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE
ÁREA: CIÊNCIA DA COMPUTAÇÃO

DELANO BRANDES MARQUES

SISTEMA INTEGRADO DE MONITORAMENTO E CONTROLE DA QUALIDADE DE COMBUSTÍVEL

São Luís
2004

DELANO BRANDES MARQUES

SISTEMA INTEGRADO DE MONITORAMENTO E CONTROLE DA QUALIDADE DE COMBUSTÍVEL

Dissertação apresentada ao curso de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Sofiane Labidi

São Luís
2004

Marques, Delano Brandes

Sistema Integrado de Monitoramento e Controle da Qualidade de Combustível/ Delano Brandes Marques. – São Luis, 2004.

142 f.: il.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Maranhão, 2004.

1. Combustível – Análise. 2. Inteligência Artificial. I. Título

CDU 662.66:004.8

DELANO BRANDES MARQUES

SISTEMA INTEGRADO DE MONITORAMENTO E CONTROLE DA QUALIDADE DE COMBUSTÍVEL

Dissertação apresentada ao curso de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Dissertação aprovada em / /

BANCA EXAMINADORA

_____

Prof. Dr. Sofiane Labidi (Orientador)
Universidade Federal do Maranhão

_____

Prof. Dr. Edson Nascimento
Universidade Federal do Maranhão

_____

Prof. Dr. Rubens Nascimento Melo
Pontifícia Universidade Católica do Rio de Janeiro

Só sabemos com exatidão quando sabemos pouco;
à medida em que vamos adquirindo
conhecimentos instala-se a dúvida!

Autor: Goethe, Johann

Aos meus pais.

AGRADECIMENTOS

A DEUS pela vida e por me proporcionar momentos tão importantes como este;

Aos meus pais, Edmar e Aldaléa pelo amor incondicional e pelo exemplo de vida que são;

Às minhas queridas irmãs, Milena e Suelen, por sempre acreditarem em mim;

A minha namorada, Marília, pelo carinho, amor e compreensão por minhas ausências neste período;

Ao Prof. Dr. Sofiane Labidi pelos ensinamentos, incentivos e orientação transmitidos como orientador e amigo durante esses anos;

Aos meus companheiros de curso, que partilharam das dificuldades e alegrias desta caminhada;

Aos colegas de laboratório, Antônio, Bruno, Bysmarck. Daniel, Davi, Eduardo e Valdeci pelo companherismo e amizade;

Aos professores do curso de mestrado, Sofiane Labidi, Edson Nascimento e Rosário Girardi, com os quais tive excelente oportunidade de aprendizado e experiência;

A Coordenadora do Programa de Pós-Graduação em Engenharia de Eletricidade, Profa. Dra. Maria da Guia, pelo apoio, incentivo e suporte geral oferecido.

Aos funcionários da Coordenadoria de Pós-Graduação e em especial ao Alcides, pelos bons serviços oferecidos e que foram fundamentais para a realização deste trabalho.

SUMÁRIO

IV APLICAÇÃO DO PROCESSO KDD EM BANCOS DE DADOS DE ANÁLISE DE COMBUSTÍVEIS

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AEAC – Álcool Etílico Anidro Carburante.

AEH – Álcool Etílico Hidratado Comum.

AG – Algoritmos Genéticos.

ANP – Agencia Nacional de Petróleo.

ASTM – American Society for Testing and Materials.

CAT-RN-LEC – Capacitação e Assistência Técnica a Laboratórios da Rede Nacional de Ensaios para Monitoramento da Qualidade de Combustíveis.

CTPETRO – Programa de Apoio à Ciência e Tecnologia voltado para o setor de Petróleo e Gás Natural.

DEE – Departamento de Engenharia de Eletricidade.

DEQUI – Departamento de Química.

DETQI – Departamento de Tecnologia Química.

FNDCT – Fundo Nacional de Desenvolvimento Científico e Tecnológico.

GCA – Gasolina C Aditivada.

GCC – Gasolina C Comum.

GPS – Global Positioning System.

IA – Inteligência Artificial.

INMETRO – Instituto Nacional de Metrologia, Normalização e Qualidade Industrial.

KDD – Knowledge Discovery in Databases.

LAPQAP – Laboratório de Análises e Pesquisas em Química Analítica de Petróleo.

LSI – Laboratório de Sistemas Inteligentes.

MB-ABNT – Métodos Brasileiros da Associação Brasileira de Normas Técnicas.

MCT – Ministério da Ciência e Tecnologia.

MOLAP – Multidimentional On Line Analytical Processing.

NBR – Normas Brasileiras.

OBC – Óleo Diesel B Comum.

ODS – Operational Data Store.

OLAP – Online Analytical Processing.

PCA – Principal Component Analysis.

PGH – Projeto Genoma Humano.

PMQC – Programa de Monitoramento da Qualidade de Combustíveis.

RNA – Rede Neural Artificial.

ROLAP – Relational On Line Analytical Processing.

SAD – Sistemas de Apoio a Decisão.

SIG – Sistema de Informação Geográfica.

SIMCO – Sistema Integrado para Monitoramento da Qualidade de Combustível.

SQL – Structured Query Language.

TI – Tecnologia da Informação.

UFMA – Universidade Federal do Maranhão.

UNIFACS – Universidade Salvador.

RESUMO

O presente trabalho apresenta estudos que visam a implantação de um Sistema Integrado que, além de permitir um melhor monitoramento, praticidade e eficiência, possibilite o controle e otimização de problemas relacionados à indústria de petróleo. Para garantir qualidade e normalização do combustível, é indispensável o desenvolvimento de ferramentas eficientes que permitam o seu monitoramento de qualquer ponto e para qualquer tipo de combustível. Considerando a variedade dos critérios, uma tomada de decisão deve ser baseada na avaliação dos mais variados tipos de dados espaciais e não espaciais. Para isto, é utilizado o Processo de Descoberta de Conhecimento, onde são enfatizadas as etapas de *Data Warehouse* e *Data Mining* aliadas ao conceito de um Sistema de Informação Geográfica. O sistema tem por objetivo abranger várias regiões de monitoramento de combustíveis. A partir do levantamento e análise das diferentes informações usadas nos bancos de dados da ANP foi proposto um modelo de *data warehouse*. Na seqüência foram aplicadas técnicas de mineração de dados (Análise de Componentes Principais, Análise de Agrupamento e Regressão) visando à obtenção de conhecimento (padrões).

Palavras-Chave: Análise de Combustíveis, Processo KDD, Mineração de Dados, Data Warehouse, Sistema de Informação Geográfica.

ABSTRACT

This work aims the implantation of an Integrated System that, besides allowing a better, more efficient and more practical monitoring, makes possible the control and optimization of problems related to the oil industry. In order to guarantee fuel's quality and normalization, the development of efficient tools that allow it's monitoring of any point (anywhere) and for any type of fuel is indispensable. Considering the variety of criteria, a decision making should be based on the evaluation of the most varied types of space data and not space data. In this sense, Knowledge Discovery in Databases process is used, where the Data Warehouse and Data Mining steps allied to a Geographic Information System are emphasized. This system presents as objective including several fuel monitoring regions. From different information obtained in the ANP databases, an analysis was carried out and a Data Warehouse model proposed. In the sequel, Data Mining techniques (Principal Component Analysis, Clustering Analysis and Multiple Regression) were applied to the results in order to obtain knowledge (patterns).

Keywords: Fuel Analysis, KDD process, Data Warehouse, Data Mining, Geographic Information Systems.

I. INTRODUÇÃO

1 INTRODUÇÃO

Neste capítulo serão apresentados, o contexto da dissertação, objetivos, justificativa e relevância, assim como se encontra organizada.

1.1 Contexto da Dissertação

Ultimamente, vem se observando um crescente aumento na quantidade de dados armazenados em meios magnéticos. A automação das atividades de negócio produz uma grande quantidade de dados, uma vez que, até simples transações como ligações telefônicas, uso de cartões de crédito, exames médicos, entre outras, são normalmente armazenadas em computadores (Piatetsky et al., 1991).

Outro exemplo do crescente aumento de dados são os bancos de dados governamentais e científicos, podendo-se citar a aeronáutica norte americana que já possui mais dados do que pode analisar; os satélites de observação da terra, que tiram em média uma foto a cada segundo, que estima-se que armazenem cerca de 1 terabyte de dados por dia, o que levaria para uma pessoa vários anos, trabalhando dia e noite, incluindo finais de semana, somente para olhar as fotos geradas em um dia. O Projeto Genoma Humano (PGH), no qual participa a área de Biologia da UFMA, e que é caracterizado pelo sequenciamento das 3 bilhões de letras (bases nitrogenadas Timina, Adenina, Guanina, Citosina) do código genético humano, distribuídas em 23 pares de cromossomos, armazena milhares de bytes para cada uma das bilhares de bases genéticas.

Essa enorme quantidade de dados nos leva a uma pergunta inevitável: O que fazer com esses dados? Os bancos de dados constituem a memória dos sistemas de Informação atuais. Mas, a memória tem pouco uso sem a inteligência. A inteligência humana permite que o homem analise sua memória e estabeleça modelos, relações e formulações de novas idéias para fazer previsões sobre o futuro (Vasconcelos, 2002). A leitura e análise destes dados, produzidos e armazenados em larga escala, claramente é inviável a especialistas através de métodos manuais tradicionais (Piatetsky,1991), tais como, planilhas de cálculos e relatórios informativos operacionais. Por outro lado, sabe-se que grandes quantidades de dados equivalem a um maior potencial de informação. Entretanto, as informações contidas nos dados não estão caracterizadas explicitamente, uma vez que, sendo dados operacionais, não interessam quando estudados individualmente.

A capacitação de bancos de dados com mecanismos de inteligência é, portanto, fator fundamental para podermos inferir algum conhecimento dos dados por eles armazenados.

O processo capaz de descobrir este conhecimento em banco de dados chama-se KDD (*Knowledge Discovery in Databases*). O processo KDD foi proposto em 1989 para referir-se às etapas que produzem conhecimentos a partir dos dados e, principalmente, à etapa de mineração dos dados, que é a fase que transforma dados em informações (Fayyad et al., 1996a). É neste contexto que se baseia a pesquisa proposta neste trabalho.

A UFMA foi contemplada com recursos do CTPETRO (Programa de Apoio à Ciência e Tecnologia voltado para o setor de Petróleo e Gás Natural), tendo como um dos principais benefícios a implantação do Laboratório de Análises e Pesquisas em Química Analítica de Petróleo – LAPQAP/DETQI-DEQUI/UFMA. Este laboratório

tem como principal atividade o monitoramento de combustíveis do Estado do Maranhão, objeto de contrato entre a Agência Nacional do Petróleo (ANP) e a UFMA.

A exemplo de outras universidades, os laboratórios LAPQAP e LSI (Laboratório de Sistemas Inteligentes – DEE/UFMA) estão iniciando uma parceria. A parceria proposta a partir deste projeto é do maior interesse tanto do LAPQAP como da ANP no que diz respeito às aspirações do laboratório LAPQAP, quanto ao processo de credenciamento junto ao INMETRO (Instituto Nacional de Metrologia, Normalização e Qualidade Industrial).

O LAPQAP dispõe de uma base de dados que compreende resultados de um ano de monitoramento de combustíveis (análises físico químicas), regulamentado pela ANP. Como se trata de um programa pela primeira vez implementado no país, é do maior interesse o conhecimento do perfil destes dados. Neste sentido, o presente trabalho propõe o início de estudos voltados para o desenvolvimento de um sistema de apoio à tomada de decisão, vinculado ao Sistema Integrado para Monitoramento da Qualidade de Combustíveis – SIMCO (Labidi, 2003).

1.2 Objetivos da Dissertação

Este trabalho tem como objetivo principal aplicar o processo de Descoberta de Conhecimento em Bancos de Dados (KDD), visando o apoio à tomada de decisão, na base de dados do LAPQAP. Pretende-se com isto colaborar com o programa de monitoramento de combustíveis da UFMA quanto à tomada de decisão em relação ao processo de análise de combustíveis.

Tendo por base que a presente pesquisa utiliza a tecnologia de IA (Inteligência Artificial), KDD, os seguintes objetivos específicos foram propostos:

- Análise diagnóstica dos dados armazenados no banco de dados do LAPQAP desde o início do programa de monitoramento de combustíveis, objeto do contrato ANP-UFMA;
- Pesquisa, análise e escolha de técnicas atuais de IA aplicáveis à extração de conhecimento em bancos de dados, sendo aplicadas ou não em pesquisas semelhantes;
- Aplicação das técnicas selecionadas ao banco de dados do LAPQAP;
- Através do uso de um aparelho GPS, foram obtidas as coordenadas geográficas de alguns postos do estado para armazenamento dos mesmos na base de dados em questão;
- Além do desenvolvimento de um *data warehouse* e aplicação de técnicas de *data mining*, este trabalho propôs o inicio da implantação de um SIG – Sistema de Informação Geográfica.

1.3 Justificativa e Relevância

O Programa de Monitoramento da Qualidade de Combustíveis (PMQC) nasceu com a própria ANP em 1998 e, ambos, são decorrentes da quebra do monopólio do setor petróleo e gás no país.

Um dos primeiros objetivos da ANP é atender a uma das principais demandas do governo federal quanto ao estabelecimento a médio prazo de uma política nacional de combustíveis que abranja duas vertentes: o da fiscalização (que está relacionado com adulteração de combustíveis) e o da qualidade do combustível

distribuído e comercializado, nacionalmente (que está ligado à satisfação do consumidor).

O PMQC é, portanto, um estudo novo, e devido ao tamanho do país, muitos laboratórios foram instalados em todos os estados, gerando uma enorme quantidade de dados. Estes laboratórios estão, em sua maioria, vinculados às universidades.

A parceria entre as áreas de química ou engenharia química e a de informática, nos âmbitos acadêmico-científico e tecnológico tem sido um importante mecanismo para o melhor aproveitamento destes dados, bem como o conhecimento do perfil dos combustíveis no Brasil. Estes objetivos foram alcançados a partir de ferramentas que utilizam como metodologia, o processo KDD, que é o tema principal do presente trabalho.

A presente proposta de pesquisa propõe, como ponto de partida, a possível aplicação de técnicas de Descoberta de Conhecimento em Bancos de Dados, juntamente com outras tecnologias de IA, aos dados do LAPQAP (Marques et al., 2003b). Estes dados estão divididos por 4 regiões geográficas no Estado do Maranhão, e compreendem cerca de 400 postos de combustíveis. Por outro lado, diversas análises físico químicas são realizadas em três tipos diferentes de combustíveis (gasolina, óleo diesel e álcool). Este será, portanto, o universo a ser estudado no presente trabalho.

1.4 Organização da Dissertação

Esta dissertação está organizada em cinco partes, divididas em seis capítulos. A Parte I é constituída por esta introdução, Capítulo 1, onde este trabalho, seus objetivos, justificativa e relevância são contextualizados e explicitados.

Na Parte II, espera-se alcançar o primeiro objetivo deste trabalho: fazer a revisão bibliográfica e apresentar o referencial teórico necessário para o entendimento da solução proposta. Esta parte é dividida em três capítulos: O Capítulo 2 apresenta uma visão do Processo KDD como um todo – definição, etapas, aplicações e técnicas; o Capítulo 3 apresenta uma visão detalhada da tecnologia de *Data Warehouse*, etapa essencial para o sucesso do processo que lida com enormes quantidades de dados; O Referencial Teórico é finalizado com o Capítulo 4, que detalha a etapa mais importante do processo KDD, a etapa de *Data Mining*.

Na Parte III, é apresentado o projeto do qual o presente trabalho faz parte, o projeto SIMCO. No Capítulo 5 são mostrados seus objetivos e metas, e algumas das tecnologias utilizadas.

Na Parte IV, são apresentados os principais objetivos: a aplicação do processo KDD no banco de dados do LAPQAP. Nesta parte são mostrados o processo de análise de combustíveis no Capítulo 6, o *data warehouse* desenvolvido neste trabalho e os resultados da aplicação de técnicas de *data mining* realizados neste trabalho no Capítulo 7.

Por fim, Na parte V, são apresentadas as conclusões sobre este trabalho de pesquisa e as sugestões para trabalhos futuros no Capítulo 8.

II. REFERENCIAL TEÓRICO

2 O PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS

Para inicio do trabalho, é necessário à apresentação dos principais conceitos referentes ao Processo de Descoberta de Conhecimento (KDD), uma vez que este processo é o foco principal deste trabalho.

2.1 Introdução

Ultimamente, em diversas áreas, dados estão sendo coletados e acumulados em enorme quantidade e velocidade. Devido a este fato, existe uma crescente necessidade de uma nova geração de teorias e ferramentas computacionais que possam auxiliar o homem na extração de informação útil (conhecimento) deste crescente volume de dados digitais. De acordo com (Fayyad et al., 1996a), estas ferramentas e teorias são os temas deste campo emergente que é a Descoberta de Conhecimento em Banco de Dados (Knowledge Discovery in Databases - KDD). O grande aumento no volume de dados, decorrente dos avanços tecnológicos, fez com que fosse gerado um enorme volume de dados, superior a capacidade humana de processá-los, o que leva o homem a recorrer aos computadores (de princípio causadores desta sobrecarga) por uma solução.

Dentre as várias definições do processo KDD, uma das mais esclarecedoras, é a de Fayyad que diz o seguinte: “É o processo não trivial de identificação de padrões válidos, desconhecidos, potencialmente úteis e compreensíveis, de dados” (Fayyad et al., 1996a).

John Naisbett descreve bem o momento que passamos através da frase:

"We are drownning in information but starving for knowledge".

Ou seja, possuímos grandes quantidades de dados, de informações, mas estamos lutando para que destes dados possamos extrair algo que faça sentido e que possa nos passar conhecimento. Como um dos motivos para esse grande aumento no volume de dados, pode-se citar os avanços nas tecnologias de armazenamento de dados, tal como a velocidade de acesso aos dados, grande capacidade de armazenamento e barateamento de dispositivos de armazenamento. Estes fatos, juntamente com os atuais sistemas gerenciadores de bancos de dados, têm permitido aumentar ainda mais as grandes quantidades de dados existentes, transformando-as em montanhas de dados armazenados.

Considerando este enorme aumento na quantidade de dados, a incapacidade destes dados serem analisados por seres humanos, devido a sua quantidade e também complexidade dos mesmos, e à necessidade de ferramentas que automatizem ou facilitem o processo de análise destes dados. O Processo KDD é uma área que está emergindo rapidamente tanto em pesquisas como em aplicações. A Primeira Conferência Internacional de Descoberta de Conhecimento em Bancos de Dados e *Data Mining* foi presidida em 1995 em Montreal, Canadá (Hunt, 2000). Desde então, o número de tecnologias de descoberta de conhecimento em uso vem crescendo cada vez mais. Algumas destas técnicas são genéricas, enquanto outras são específicas a um domínio (Wright, 1998).

KDD é o processo de extração de conhecimento a partir de massas de dados, constituído de várias etapas, com o objetivo de se obter sentido e conseqüente entendimento dos dados assim como adquirir novos conhecimentos, sendo a Mineração de Dados (*Data Mining*) referente a um passo particular deste processo (aplicação de técnicas/algoritmos específicos para extração de padrões dos dados).

Os passos adicionais do processo KDD mostrados na figura 01, tais como Pré-Processamento dos dados, *Data Mining* e Pós-Processamento é a garantia essencial de que o conhecimento é derivado dos dados.

O sucesso do processo pode ser alcançado através da identificação de estruturas, características, tendências, anomalias e relacionamentos encontrados entre os dados pelo algoritmo aplicado na fase de *Data Mining*. Este é um processo bem complexo, pois consiste de uma tecnologia composta de um grupo de modelos técnicos e matemáticos de software que são utilizados para encontrar padrões e regularidades nos dados (Decker et al., 1995).

A exemplo de outros processos, o processo KDD é formado pela interseção de diferentes áreas. As áreas mais relacionadas à descoberta de conhecimento são: *Machine Learning* (Langley, 1996), (Shavlik, 1990), Inteligência Computacional, Estatística (Elder, 1996) e visualização dos dados (Lee, 1995). Na área de Inteligência Computacional, em particular, as técnicas mais utilizadas são: Redes Neurais Artificiais -RNA (Haykin, 1994), (Rumelhart et al., 1986), Indução de regras (Nilsson, 1980) e Algoritmos Genéticos – AG (Goldberg, 1989). As técnicas utilizadas nesta dissertação são relacionadas à área de Estatística. Estas técnicas são detalhadas no Capítulo 3 que se refere à *Data Mining*.

2.2 Etapas

O processo KDD consiste de uma série de passos, etapas (Figura 01) que, como já dito, objetivam obter conhecimento dos dados. Essas etapas incluem a limpeza dos dados, seleção de amostras, transformação dos dados, dedução de conhecimento das amostras através de algoritmos de *data mining* e a estimativa da

qualidade do conhecimento extraído. A complexidade do processo depende destas variáveis referidas.

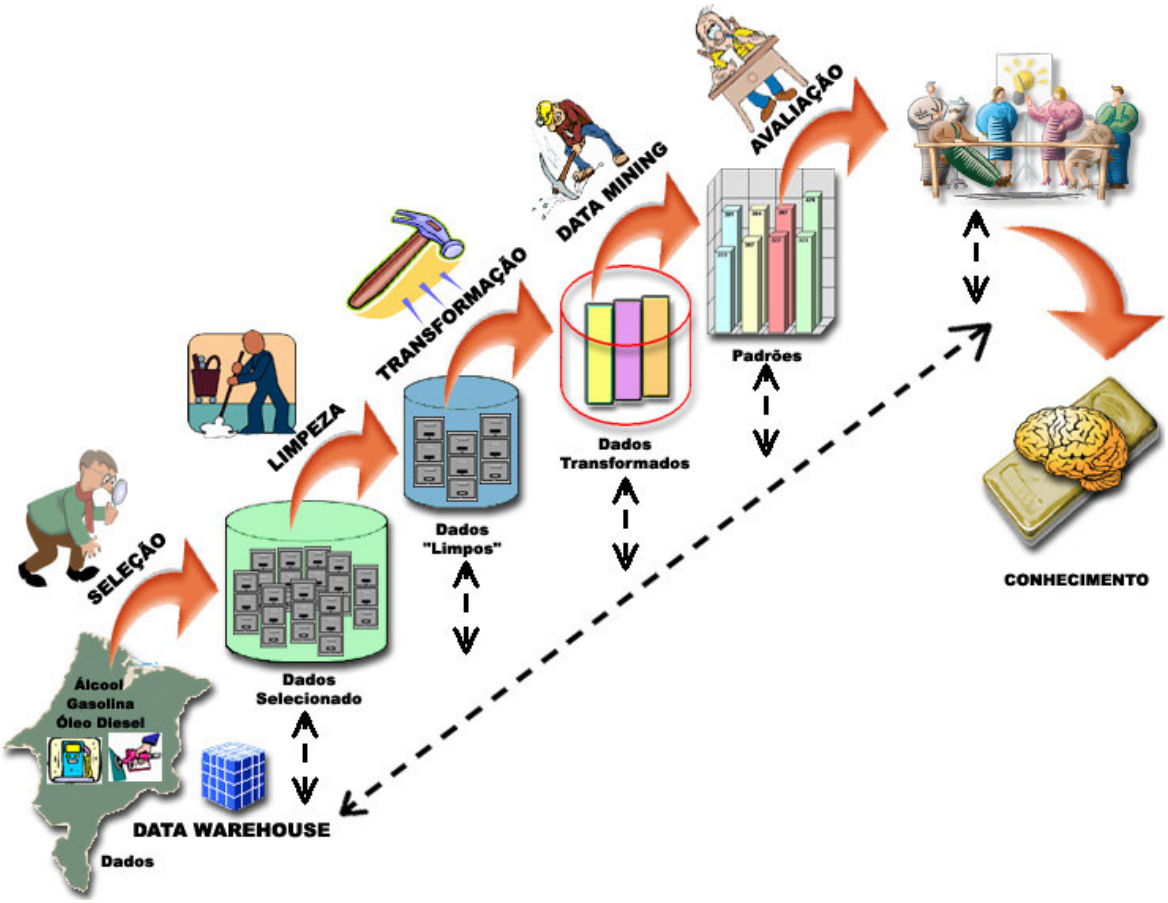

Figura 01: Etapas do Processo KDD.

No processo KDD cada fase pode possuir uma interseção com as demais. Desse modo, os resultados produzidos em uma fase podem ser utilizados para melhorar os resultados das fases seguintes. Esse cenário indica que o processo KDD é interativo e iterativo (Berry, 2000), ou seja, depende da constante interferência do usuário, técnico especialista, e se repetem de acordo com a necessidade.

Vale ressaltar que a participação do usuário é essencial para o sucesso do processo, uma vez que a definição do problema, fundamental para o processo, requer que a pessoa que solicita a tarefa de KDD entenda perfeitamente o problema existente

e tenha um objetivo bem especificado, ou seja, aquilo que se deseja conhecer ou extrair. Para isso, é necessário uma interação com o solicitador da tarefa de modo que seja exposto tudo o que se relaciona com o problema. Tendo sido definido o problema, podem-se fixar metas para os objetivos da tarefa de KDD.

2.2.1 Data Warehouse

A aplicação do processo KDD deve iniciar a construção de um *Data Warehouse*. Este é um meio efetivo de organizar grandes volumes de dados para sistemas de suporte à decisão e aplicações KDD. Pode-se definir um *data warehouse* como um repositório integrado, orientado para análise, histórico, com dados apenas para leitura, designado para ser utilizado como base para suporte à decisão e sistemas KDD (Inmon, 1993), (V.Poe, 1996). Um *data warehouse* funciona como uma base de dados para dar suporte à decisão, mantido separadamente das bases de dados operacionais da organização. Geralmente integra dados de diversas origens heterogêneas e por isso necessita de uma estrutura flexível que suporte consultas e geração de relatórios analíticos.

Em um processo KDD, a fase de *data warehouse* não é absolutamente necessária, podendo, ser executada pelo usuário do sistema conforme a necessidade de dados para a técnica de *data mining* a ser utilizada. Entretanto o uso do *data warehouse* é importante para agilizar, organizar e aumentar qualitativamente o processo KDD.

É importante notar que não existe um sistema que implemente todo este processo. Existem sistemas intermediários, controlados por um usuário, cada um

desses sistemas é bem definido e com o seu objetivo delineado conforme a tarefa solicitada.

2.2.2 Pré-Processamento dos Dados

Nesta etapa ocorrem vários passos para a construção de uma base de dados consistente. Ela é responsável por consolidar as informações relevantes para o algoritmo minerador buscando reduzir a complexidade do problema, e inclui 3 sub-fases: seleção dos dados, limpeza dos dados e transformação ou codificação dos dados. A execução destas fases não necessita de uma ordem, podendo ser executadas durante a construção de um *Data Warehouse*.

- **Seleção dos Dados**: Nesta fase é feita a seleção dos dados apropriados para a análise. Seu principal objetivo é a escolha dos atributos relevantes do conjunto de atributos do banco de dados. Aqui são identificados quais dados deverão ser extraídos para a mineração. Em suma, essa seleção consiste da escolha de um subconjunto de atributos disponíveis para o processo KDD que sejam relevantes para o objetivo da tarefa. O subconjunto selecionado é então fornecido para a técnica de *data mining*. A finalidade dessa seleção é otimizar o tempo de processamento da etapa de mineração, visto que ela apenas trabalhará com um subconjunto de atributos, desse modo diminuindo o seu espaço de busca.
- **Limpeza dos Dados**: A limpeza dos dados envolve uma verificação da consistência das informações, a correção de possíveis erros e o preenchimento ou a eliminação de valores nulos e redundantes. A execução dessa fase corrige a base de dados eliminando consultas desnecessárias que seriam executadas pela etapa de *data mining* e que afetariam o seu processamento. Os métodos de limpeza dos dados são herdados e dependentes do domínio da aplicação e, desse modo, a participação do analista de dados torna-se essencial.

- **Transformação dos Dados**: Existem vários tipos de técnicas de *data mining*, sendo que cada uma necessita de uma entrada específica. A finalidade desta fase é transformar os dados, de modo que eles se tornem compatíveis com as entradas dessas diversas técnicas. Em resumo essa fase converte os dados na forma mais adequada para a construção e interpretação do modelo. Esta fase é potencialmente a tarefa onde há a necessidade de grande habilidade no processo KDD. Pois geralmente exige a experiência do analista de dados e do seu conhecimento nos dados em questão. Embora o processo KDD possa ser executado sem essa fase, nota-se que quando efetivada os resultados obtidos são mais intuitivos e valiosos, além de que, na maioria das vezes, facilita a construção do modelo (Mannino et al., 1988), (Dougherty et al., 1995).

2.2.3 Data Mining

O *Data Mining* é uma das fases mais importantes do processo KDD (Fayyad et al., 1996a), onde é feita a escolha e a aplicação da técnica e algoritmo (ou algoritmos) de mineração de dados. Pode-se dizer que esta é a fase que transforma dados em informações. Caracteriza-se pela existência do algoritmo de mineração que diante da tarefa especificada será capaz de extrair eficientemente conhecimento (padrões/modelos) implícito e útil dos dados.

A tecnologia de mineração de dados possui a vantagem de extrair informação que não seria possível de ser obtida através de consultas tradicionais em bancos de dados. As ferramentas tradicionais são limitadas a simples questões feitas pelos usuários, assim, a mineração de dados, extraindo preciosas informações das bases de dados, pode auxiliar o usuário na tomada de decisões. Apesar disto, esta tecnologia possui alguns obstáculos a superar, como a necessidade de grandes

volumes de dados, complexidade das ferramentas, desafios na preparação dos dados para a mineração, dificuldade de realizar uma análise custo/benefício do projeto de descoberta e disponibilidade das ferramentas (OWG, 2000).

No contexto de mineração de dados, conhecer as diferenças entre dado, informação e conhecimento é muito importante. Sendo assim é de grande importância a definição destes termos. Dado é um conjunto de símbolos que tomado isoladamente não contém nenhum significado claro. Informação é todo dado trabalhado por pessoas ou por recursos computacionais, com valor significativo agregado a ele e com sentido lógico para quem usa a informação. Conhecimento é um conjunto de informações que permite articular os conceitos, os juízos e o raciocínio, usualmente disponíveis em um domínio particular de atuação (Junior et al., 2000).

Nessa fase, necessita-se definir a técnica e o algoritmo a ser utilizado em função da tarefa proposta. Uma vez escolhido o algoritmo a ser utilizado, deve-se implementá-lo e adaptá-lo ao problema proposto.

Para finalizar essa fase, deve-se executar o algoritmo a fim de obter resultados que serão analisados na fase seguinte, pós-processamento.

2.2.4 Pós-Processamento

Nesta etapa as seguintes questões devem ser respondidas: o conhecimento gerado é relevante e acionável? Se a resposta não for satisfatória, então será necessário repetir todo ou parte do processo KDD.

Quando um padrão é identificado na etapa de *data mining*, ele deve ser examinado para se determinar se ele é novo, relevante e “correto” por uma medida padrão. A etapa de pós-processamento pode envolver mais interação com o usuário

ou algum agente do usuário que possa fazer determinações relevantes. Quando o padrão é julgado relevante e útil, ele pode ser considerado conhecimento. O conhecimento então deve ser armazenado na base de conhecimentos para subseqüentes interações.

Essa fase envolve a interpretação do conhecimento descoberto, ou algum processamento desse conhecimento. Esse pós-processamento deve ser incluído no algoritmo de *data mining*, porém algumas vezes é vantajoso implementá-lo separadamente. Em geral, a principal meta dessa fase é melhorar a compreensão do conhecimento descoberto pelo algoritmo minerador, validando-o através de medidas da qualidade da solução e da percepção de um analista de dados. Esses conhecimentos serão consolidados em forma de relatórios demonstrativos com a documentação e explicação das informações relevantes ocorridas em cada etapa do processo KDD.

Uma maneira genérica de obter a compreensão e interpretação dos resultados é utilizar técnicas de visualização (Decker et al., 1995). Existem também outros tipos de técnicas de pós-processamento criados especialmente para um dado tipo de algoritmo de *data mining*, ou para uma dada tarefa de KDD.

2.3 Aplicações do Processo KDD

Muitas empresas que possuem visão do futuro, estão adotando a utilização do Processo KDD na análise de seus bancos de dados em busca de padrões úteis e interessantes. Dentre os benefícios adquiridos por essas empresas podemos citar a máxima utilização dos dados corporativos, descoberta de novos conhecimentos,

geração de modelos preditivos e exploratórios e identificação de dados essenciais e irrelevantes. A seguir exemplos de empresas que utilizaram o processo:

- *American Airlines* (EUA): pesquisa seu banco de dados em busca de seus passageiros mais freqüentes para encontrar os melhores clientes, fazendo a eles promoções especiais e personalizadas;
- *Farm Journal* (EUA): analisa seu banco de dados de assinantes e utiliza técnicas avançadas de impressão para construir centenas de edições particularizadas a determinados grupos de assinantes;
- *General Motors* (EUA): está utilizando um banco de dados de relatórios de problemas automotivos para derivar sistemas especialistas de diagnóstico para vários modelos de automóveis;
- Bancos: utilizam padrões descobertos em seus bancos de dados, na aprovação de empréstimos, identificação de transações fraudulentas e históricos de crédito. Com isso melhorou os métodos de aprovação de empréstimos e predições de corrupção;
- Supermercados: estão utilizando sistemas para varrer seus dados e medirem os efeitos de suas promoções e procurar por padrões de compras.

2.4 Técnicas

Embora haja muitas abordagens para o processo KDD, seis elementos comuns e essenciais às qualificam como uma técnica de descoberta de conhecimento. A seguir estão os aspectos básicos que todas as técnicas de KDD compartilham (Wright, 1998):

- Todas lidam com grandes quantidades de dados, o que é necessário para se prover informação suficiente e então produzir conhecimento adicional;

- Eficiência, que é requerida devido a grande quantidade de dados;
- Precisão é um elemento essencial para garantir que o processo de descoberta de conhecimento seja válido
- Todas requerem o uso de uma linguagem de alto nível, pois os resultados devem ser apresentados de uma maneira que seja entendida por humanos;
- Todas utilizam alguma forma de aprendizagem automatizada, uma das principais premissas do processo KDD é que o conhecimento seja descoberto utilizando-se técnicas inteligentes de aprendizado que filtram os dados em um processo automatizado;
- Todas produzirem resultados interessantes, para que uma técnica seja considerada útil em termos de descoberta de conhecimento, o conhecimento descoberto deve ser interessante, deve ter um valor potencial ao usuário;

Existem várias e diferentes abordagens que são classificadas como técnicas KDD. Existem abordagens quantitativas, tais como a probabilística e estatística, as que utilizam técnicas de visualização, as abordagens de classificação tais como a classificação bayesiana, lógica indutiva, limpeza de dados/descoberta de padrões, e análise de árvores de decisão. Outras abordagens incluem análises de desvio de tendências, algoritmos genéticos, redes neurais e abordagens híbridas que combinam duas ou mais técnicas. Por causa do modo como estas técnicas podem ser utilizadas e combinadas, existe uma desavença em relação à categorização destas técnicas. Por exemplo, a abordagem Bayesiana pode ser logicamente classificada como uma abordagem probabilística, de classificação ou de visualização. Devido a este fato, as abordagens descritas a seguir são incluídas nos grupos em que melhor parecerem se encaixar.

- **Probabilística** - Esta família de técnicas KDD utiliza modelos de representação gráfica para comparar diferentes representações de conhecimento. Estes

modelos são baseados em probabilidades e independência de dados. São úteis em aplicações que envolvem incerteza e em aplicações estruturadas onde a probabilidade pode ser atribuída a cada “saída” ou pedaço de conhecimento descoberto. Técnicas de probabilidade podem ser utilizadas em sistemas de diagnóstico e em sistemas de planejamento e controle.

- **Estatística** - A abordagem estatística utiliza a descoberta de regras e é baseada no relacionamento de dados. Um algoritmo de aprendizagem indutivo pode automaticamente selecionar *join paths* e atributos para construir regras de um banco de dados com vários relacionamentos. Este tipo de indução é utilizado para generalizar padrões nos dados e construir regras dos padrões encontrados. O *Online Analytical Processing* (OLAP) é um exemplo de uma abordagem orientada estatisticamente. As ferramentas OLAP são aplicações que os usuários finais têm acesso para extraírem os dados de suas bases com os quais geram relatórios capazes de responder as suas questões gerenciais. Elas surgiram juntamente com os sistemas de apoio à decisão – SAD, para fazerem a extração e análise dos dados contidos nos *Data Warehouses* e *Data Marts*.
- **Classificação** – A classificação é provavelmente a mais antiga e mais utilizada de todas as abordagens KDD. Esta abordagem agrupa os dados de acordo com similaridades ou classes. Existem vários tipos de técnicas de classificação e inúmeras ferramentas disponíveis. A abordagem Bayesiana ao KDD, é um modelo gráfico que utiliza arcos diretos exclusivamente para formar um grafo acíclico direto. Apesar desta abordagem utilizar probabilidade e meios gráficos de representação, ela também é um tipo de classificação. As redes bayesianas são geralmente utilizadas quando a incerteza associada a uma das saídas pode ser expressa em termos de probabilidade. Esta abordagem se baseia em conhecimento de domínio codificado (*encoded domain knowledge*) e tem sido utilizada em sistemas de diagnostico. Outras aplicações de reconhecimento de

padrões, incluindo *Hidden Markov Model* podem ser modeladas utilizando-se a abordagem bayesiana. As técnicas de descoberta de padrões e limpeza de dados também são tipos de classificação que sistematicamente reduzem um grande banco de dados em poucos e informativos registros. Se dados redundantes e sem interesse são eliminados, a tarefa de descoberta de padrões nesses dados é simplificada. Esta abordagem trabalha em cima de um provérbio que diz "menos é mais" (*less is more*). Outra técnica que se enquadra na abordagem em questão é a de Árvores de Decisão, que utiliza regras de produção, constrói um grado acíclico baseado em premissas de dados, e os classifica de acordo com seus atributos. Este método requer que as classes de dados sejam discretas e pré-definidas. A utilização primária desta abordagem é para modelos preditivos que são apropriados para ambas às técnicas de classificação ou regressão.

- **Análise e Desvio de Tendências** - A detecção de padrões através da filtragem de importantes tendências é a base desta abordagem. Técnicas de análise de desvio e tendências são normalmente aplicadas a bancos de dados temporais. Uma boa aplicação para este tipo de KDD é a análise de tráfego em grandes redes de telecomunicações. A empresa de telecomunicações americana AT&T utiliza um sistema para localizar e identificar circuitos que apresentam desvios (comportamento faltoso). O volume de dados com comportamento fora do padrão que requerem análise faz com que o uso de uma técnica automatizada seja imperativo. Análises do tipo das tendências podem ser de grande utilidade para dados de astronomia e oceanografia, já que são dados volumosos e baseados em tempo.
- **Outras Abordagens** - Existem técnicas que, por poderem ser atribuídas a mais de um tipo de abordagem, não chegam a ser classificadas por muitos autores. Como exemplo, podem-se citar as redes neurais e algoritmos genéticos. As redes neurais podem ser utilizadas como método de descoberta de

conhecimento. As RNA são particularmente úteis no reconhecimento de padrões, e muitas vezes são agrupadas com abordagens de classificação. São técnicas computacionais que propõem um modelo matemático baseado na estrutura neural de organismos inteligentes, mais especificamente o cérebro humano, e tem como principais características: capacidade de aprender, generalização, associação e abstração. Uma RNA procura por relacionamentos, constrói modelos automaticamente e os corrige de modo a diminuir seu próprio erro (Tafner et al., 1996). Os Algoritmos Genéticos, também utilizados na abordagem de classificação, são modelos estocásticos e probabilísticos de busca e otimização, inspirados na evolução natural e na genética, aplicados a problemas complexos de otimização (Mitchell, 1998). Na mineração de dados, os AG também têm sido empregados em tarefas de descrição de registros de uma base de dados, além da classificação e da seleção de atributos de bases de dados que melhor caracterizem o objetivo da tarefa de KDD proposta. Na classificação de registros, os modelos de algoritmo genético geram regras que exprimem uma realidade do domínio de aplicação. Essas regras são de fácil interpretação, o que incentiva o uso dessa técnica (Aurélio et al., 1999). Os AG podem ser vistos como uma técnica sub-simbólica, pois utilizam uma abordagem baseada na evolução humana, mas o conhecimento gerado é representado em forma de regras.

- **Híbrida -** Uma abordagem híbrida do processo KDD combina mais de uma abordagem e é também chamada de abordagem multi-paradigmática. Embora a implementação seja mais difícil, ferramentas híbridas podem combinar os pontos fortes de várias abordagens. Alguns dos métodos comumente utilizados combinam técnicas de visualização, indução, redes neurais e sistemas baseados em regras para alcançar o descobrimento do conhecimento desejado. Como exemplo podemos citar os Bancos de dados dedutivos e algoritmos genéticos que têm sido utilizados em abordagens híbridas.

2.5 Conclusão

O processo KDD é um campo que se expande rapidamente com promessas de grande aplicabilidade. A descoberta de conhecimento tem em vista ser a nova tecnologia de banco de dados nos anos por vir.

Também se pode concluir que este é um processo de certa complexidade uma vez que ele é constituído de várias etapas, e se inicia com o acesso de dados, continua com o pré-processamento (seleção, limpeza e transformação), *data mining* e pós-processamento. Durante o processo, várias técnicas trabalham de forma cooperativa, sendo estas técnicas distinguidas pela representação de conhecimento aplicada em cada uma delas e pelo tipo de raciocínio adotado em cada uma. A identificação da técnica mais apropriada a cada operação é dependente dos dados e características do domínio.

Outro fato importante é o de o processo KDD ser altamente interativo, uma vez que o usuário, especialista, está presente no processo não só na avaliação final, mas no decorrer de todo ele.

A completa utilização de dados armazenados depende do uso de técnicas de descoberta de conhecimento, uma vez que o processo KDD prove a capacidade de descobrir novas e significantes informações através de dados existentes e rapidamente supera a capacidade humana de analisar grandes conjuntos de dados. Pois a quantidade de dados que requer processamento e análise em um grande banco de dados excede as habilidades humanas e a dificuldade da exatidão na transformação de dados em conhecimento ultrapassa os limites dos bancos de dados tradicionais.

Apesar do seu rápido crescimento, o Processo KDD ainda é uma área de conhecimento emergente. Não existe ferramenta que atendam todos os requisitos do processo. Alguns aspectos onde se pode evoluir:

- Integração de diferentes técnicas;
- Maior integração com o banco de dados;
- Suporte tanto para especialistas em análise quanto para usuários inexperientes;
- Gerenciamento de dados em constante mudança;
- Maior integração entre as ferramentas existentes.

3 DATA WAREHOUSE

Esta seção apresenta e descreve diversos conceitos relativos à tecnologia de Data Warehouse, fase indispensável para a obtenção do sucesso no processo KDD aplicado a grandes bancos de dados.

3.1 Introdução

Além da sobrecarga de dados em que vivemos, e que tende a crescer devido aos avanços tecnológicos, outro fator que dificulta o acesso e análise desses dados é o fato de serem oriundos de diversas fontes heterogêneas. A tecnologia de *Data Warehouse* vem como uma solução para esses, dentre outros problemas, uma vez que duas de suas principais características são a organização de grandes volumes de dados e o suporte a diferentes fontes de dados.

3.2 Conceitos Básicos

Abaixo serão mencionados alguns conceitos básicos sobre a tecnologia em questão afim de um melhor entendimento do tema.

3.2.1 Data Warehouse

- Um *data warehouse* pode ser definido como um conjunto de dados baseado em assuntos, integrado, não volátil, e variável em relação ao tempo, para dar

suporte ao processo gerencial de tomada de decisão (Inmon, 1997), (Inmon et al., 1997b).

- Griffiths (1995), defini o *data warehouse* como um banco de dados projetado para apoiar o processo decisório. Onde os gerentes das organizações podem encontrar as informações para dirigir e decidir sobre o rumo dos negócios com maior probabilidade de êxito.
- Barquini (1996), defini o *data warehouse* como uma coleção de técnicas e tecnologias que juntas disponibilizam um enfoque pragmático e sistemático para tratar com o problema do usuário final de acessar informações que estão distribuídas em vários sistemas da organização.
- Harjinder (1996), defini o *data warehouse* como um processo em andamento que aglutina dados de fontes heterogêneas, incluindo dados históricos e dados externos para atender as necessidades de consultas estruturadas e ad-hoc, relatórios analíticos e de suporte à decisão.
- Henrique (1998), defini o *data warehouse* como um conjunto de arquiteturas e/ou sistemas de informação que viabilizam processos de tomada de decisões em diversos níveis organizacionais. Tais processos, que ocorrem em plataformas, segregados do ambiente transacional, são baseados em grandes volumes de dados, principalmente históricos, que manipulam dados no nível analítico e/ou sintético, relacionais ou multidimensionais, entrelaçados ou não, através de consultas invariavelmente não-previsíveis.
- Kimball (1996), defini o *data warehouse* como uma fonte de dados consultáveis da organização, formada pela união de todos os *data marts* correspondentes.
- De acordo com (Taurion, 1997), a grande vantagem de um *data warehouse* é permitir a tomada de decisões baseadas em fatos que acontecem dentro da área de negócios da empresa. Esta tecnologia começa a se delinear como uma ferramenta de gestão obrigatória para as empresas sobreviverem nos próximos anos. Pois o ambiente de negócios é crescentemente dinâmico, e à medida que

as regras de negócio são incorporadas às aplicações, exige-se uma rapidez cada vez maior nas respostas.

3.2.2 Dados Operacionais Vs. Dados Multidimensionais

Visando o esclarecimento do entendimento sobre *data warehouse* faz-se necessário um estudo comparativo entre o conceito tradicional de banco de dados e *data warehouse*. Um banco de dados é uma coleção de dados operacionais armazenados e utilizados pelo sistema de aplicações de uma empresa específica (Batini et al., 1986). Os dados no banco de dados são referidos como "dados operacionais", distinguindo-os de dados de entrada, dados de saída e outros tipos de dados. Baseado na definição anterior sobre dados operacionais pode-se dizer que um *data warehouse* é, na verdade, uma coleção de dados derivados dos dados operacionais para sistemas de suporte à decisão. Estes dados derivados são, geralmente, referidos como dados "gerenciais", "informacionais" ou "analíticos" (Inmon et al., 1997b).

Os bancos de dados operacionais mantêm armazenadas as informações necessárias para viabilizar operações cotidianas, como cadastros, alterações e exclusões de registros, sofrendo mudanças constantemente. Por não ocorrer redundância nos dados, e as informações históricas não ficarem armazenadas por muito tempo, este tipo de banco de dados não exige grande capacidade de armazenamento.

Por outro lado, um *data warehouse* armazena dados analíticos, destinados às necessidades da gerência no processo de tomada de decisões. O que pode envolver consultas complexas que necessitam acessar um grande número de

registros. Por isso, é importante a existência de muitos índices criados para acessar às informações da maneira mais rápida possível. Portanto um *data warehouse* armazena informações históricas de muitos anos, o que implica em uma grande capacidade de processamento e armazenamento dos dados que se encontram de duas maneiras, detalhados e resumidos.

Um *data warehouse* é formado por um resumo de dados extraídos de uma ou mais fontes de dados, normalmente utilizadas há vários anos e que continuam em operação. São construídos para que os dados possam ser armazenados e acessados de forma que não sejam limitados por tabelas e linhas estritamente relacionais.

3.2.3 Elementos Básicos de um Data Warehouse

Domenico (2001) comenta em seu trabalho que desde o início dos anos 90, surgimento da área de *data warehouse*, vem ocorrendo uma perda da precisão na definição de seus termos. A Tabela 1 apresenta os conceitos dos principais elementos componentes de um *data warehouse*, segundo a visão integrada de (Kimball et al., 1998).

Tabela 1 – Conceitos dos principais elementos componentes de um *data warehouse.*

ELEMENTO BÁSICO	**DEFINIÇÃO**
Sistemas de Origem	Sistema operacional de registros cuja função é capturar as transações do negócio.
Staging Area	Área de armazenamento e conjunto de processos que limpam, transformam, combinam, retiram duplicações, retêm, arquivam e preparam os dados fonte para uso no *data warehouse*.
Servidor de Apresentações	Máquina física de destino onde estão armazenados e organizados os dados do *data warehouse* para consultas diretas dos usuários finais, dos geradores de relatórios e

	de outras aplicações.
Modelo Dimensional	Disciplina específica para modelagem de dados que é uma alternativa ao modelo de entidades-relacionamento (modelo E/R).
Processos do Negócio	Conjunto coerente das atividades do negócio da organização, que fazem sentido aos usuários de negócio do *data warehouse*.
Data Mart	Um subconjunto lógico do *data warehouse* completo.
Armazenamento de Dados Operacionais (ODS)	Ponto de integração com os sistemas operacionais da organização. Criados para integrar em nível operacional os diferentes sistemas da organização, sem, contudo, incluir consultas gerenciais, que ficam ao nível do *data warehouse*.
OLAP	Atividade genérica de consultar e apresentar dados textuais ou numéricos de um *data warehouse*, bem como uma forma dimensional especifica de consultar e apresentar que é exemplificado por um número de 'vendedores OLAP'. Trata-se de uma tecnologia não-relacional e geralmente baseada em cubos multidimensionais de dados.
ROLAP (OLAP Relacional)	Conjunto de interfaces ao usuário e de aplicações que dão características multidimensional a bancos de dados relacionais.
MOLAP (OLAP Multidimensional)	Conjunto de interfaces ao usuário, aplicações com base de dados proprietária que são fortemente multidimensionais.
Aplicação para Usuário Final	Coleção de ferramentas que consultam, analisam e apresentam informações desejadas com vistas às necessidades do negócio da organização.
Ferramenta de Controle de Acesso aos Dados para Usuário Final	Uma ferramenta de controle de acesso aos dados para o usuário final. Pode ser simples como sistemas de consultas ad-hoc ou complexas e sofisticadas com mineração de dados ou aplicações de modelagem.
Ferramentas de Consultas Ad-hoc	Tipo específico de ferramenta dos dados que induz o usuário final a formar suas próprias consultas, manipulando diretamente tabelas relacionais e suas funções.

Aplicações de Modelagem	Tipo sofisticado de ferramenta cliente do *data warehouse* com capacidades analíticas de transformar ou compreender as saídas do *data warehouse* (e.g. *data mining*, modelos de previsão, modelos de comportamento, etc).
Metadados	Toda informação no ambiente do *data warehouse* que não é dado real em si mesmo.

3.3 Características do Data Warehouse

De acordo com (Inmon, 1997), o conjunto de dados constituintes do *data warehouse* possui algumas peculiaridades, este conjunto deve ser classificado por assunto, possuir uma integração de suas representações, representar um bom período de tempo e não serem modificados. A seguir, uma breve abordagem de cada uma destas características.

3.3.1 Orientação por Assunto

Os sistemas tradicionais são projetados em torno das principais atividades envolvendo determinado contexto, são orientados a processos desenvolvidos para manter as transações realizadas diariamente. Contrastando com um *data warehouse*, que armazena dados importantes de acordo com o interesse das pessoas que irão fazer uso desta informação. Um exemplo citado por Inmon que caracteriza este agrupamento de dados, informações por assuntos, é o de em uma empresa de seguros, onde suas atividades lidam com automóveis, saúde, vidas, e acidentes. Os principais assuntos que envolvem as atividades desta empresa, podem ser a apólice de seguro de seus clientes e as ocorrências de acidentes com seus clientes. Outro exemplo é o de um fabricante, os principais assuntos que envolvem suas atividades

podem ser produto, pedido, vendedor, preço dos materiais e bens. Cada contexto possui seu próprio conjunto de assuntos que o melhor representam.

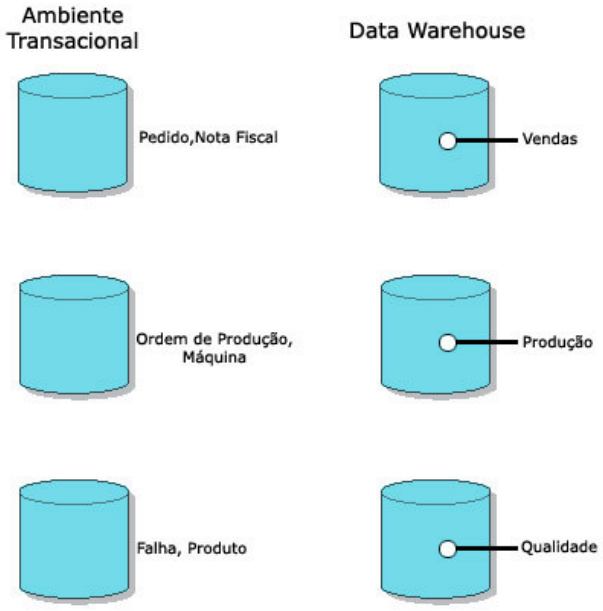

Figura 02. Orientação por assunto. Dados baseados em assuntos de negócio.

É de grande importância que ao inicio de um projeto de *data warehouse* seus usuários finais deixem claro quais os seus objetivos, quais são as informações importantes para o processo de análise, para o sucesso do projeto.

3.3.2 Integração

De acordo com Inmon, de todas as características do *data warehouse* a integração é a principal, segundo a qual se define a representação única para os dados procedentes de diversas e diferentes fontes que irão compor o *Data Warehouse*. O resultado desta integração, é que uma vez alocados no *Data Warehouse* esses dados passam a possuir uma única e incorporada imagem.

A Figura 03, retirada de (Inmon, 2002), demonstra a integração que ocorre quando os dados são passados de aplicações em ambientes operacionais para um *data warehouse*. A integração envolve padronizar atributos, formatos e convenções de nomes, além de remoção de inconsistências.

Outro ponto destacado por Inmon é o fato de os projetos feitos no passado, e até hoje, serem realizados sem nenhuma preocupação com a possibilidade de existir a necessidade de suas aplicações serem integradas com outros dados. Por esta razão, o que encontramos são inúmeras aplicações sem nenhum padrão de codificação, nomenclatura, atributos físicos, medidas de atributos, etc. Cada analista de sistemas define a mesma estrutura de dados de diversas maneiras, ocasionando que os dados que representam a mesma informação possuam representações diferentes dentro dos sistemas utilizados ao longo do tempo.

A carga dos dados no *data warehouse* é feita de tal maneira que as inconsistências encontradas no nível operacional são desfeitas. Na Figura 03, levando em consideração a codificação de sexo, pouco importa se os dados no *data warehouse* estão codificados como M/F ou 1/0. O que deve ser levado em consideração, não importando o método ou aplicação de origem, é que a codificação ao *data warehouse* seja feita de forma consistente.

Conforme os dados são levados para o *data warehouse*, eles são convertidos em um estado uniforme, ou seja, sexo é codificado apenas de uma forma. Da mesma maneira, se um elemento de dado é medido em centímetros em uma aplicação, em polegadas em outra, ele será convertido para uma representação única ao ser colocado no *data warehouse*.

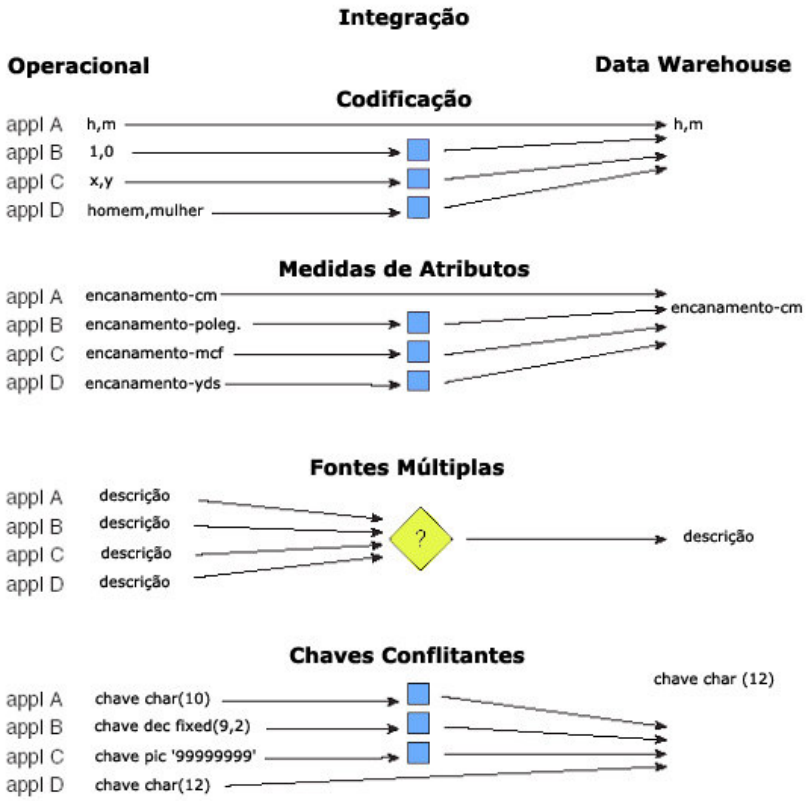

Figura 03. Integração de dados heterogêneos.

3.3.3 Variação no Tempo

Um *data warehouse* contém dados históricos, variantes no tempo (geralmente por um período de muitos anos). A variação no tempo de cada unidade de dado representa um resultado operacional em algum momento no tempo, o momento em que foram armazenados. De acordo com (Machado, 2000), os dados em um *data warehouse* representam um conjunto estático de registros de uma ou mais tabelas, capturadas em um momento de tempo predeterminado, implicando na não possibilidade de atualização desses dados.

Inmon destaca a diferença de horizonte de tempo de diferentes sistemas, caracterizando um horizonte de tempo como os parâmetros de tempo representados

em um ambiente, sendo que o de um *data warehouse* é significantemente maior que o de um sistema operacional, onde 60-90 dias é um horizonte de tempo normal contrastando com os 5-10 anos de uma *data warehouse*.

A estrutura chave de dados operacionais, pode ou não, conter algum elemento que represente tempo, como ano, mês, dia, e assim por diante. Já a estrutura chave de um *data warehouse* sempre possuirá algum elemento que represente tempo.

3.3.4 Não Volatilidade

Outra característica de um *data warehouse* é a não volatilidade (Figura 04), isto significa que quando os dados são carregados, não sofrem alterações. O que ocorre são a carga inicial e as consultas posteriores. As alterações que ocorrerem nas bases de dados que servirem de fontes para o *data warehouse*, não acarretam em alterações no mesmo, e sim em novas cargas de dados. As atualizações no *data warehouse* geralmente consistem na inserção de novos dados num período pré-determinado de tempo, por exemplo, uma vez por semana.

Um *data warehouse* é uma base apenas para leitura, onde o usuário obtém a informação desejada executando consultas pré-definidas que fazem junções entre as tabelas de fatos e dimensões (ver item 3.8).

No ambiente operacional, ao contrário, os dados são, em geral, atualizados registro a registro, em múltiplas transações, para que estejam disponíveis aos usuários para acesso. Esta volatilidade requer um trabalho considerável para assegurar integridade e consistência através de atividades de *rollback*, recuperação de falhas,

commits e bloqueios. Um *data warehouse* não requer este grau de controle típico dos sistemas orientados a transações (Campos, 2003).

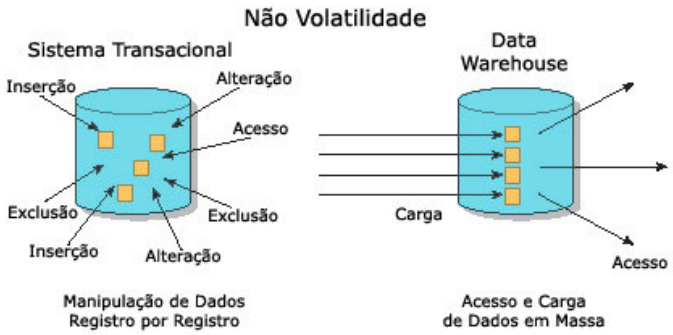

Figura 04. Não Volatilidade dos dados.

3.4 Granularidade

A granularidade se refere ao nível de detalhe ou de resumo em que as unidades de dados se encontram no *data warehouse*. Quanto maior o nível de detalhes, menor o nível de granularidade. Esta é uma questão fundamental no projeto de um *data warehouse* pois afeta diretamente seu volume de dados armazenados, e ao mesmo tempo, o tipo de consulta que pode ser respondida, havendo portanto uma troca a ser considerada entre estes dois aspectos.

Quando se tem um nível de granularidade muito alto, o espaço em disco e o número de índices necessários se tornam bem menores, entretanto há uma correspondente diminuição da possibilidade de utilização dos dados para atender a consultas detalhadas dos usuários. Por outro lado, com um nível baixo de granularidade, pode-se responder a qualquer consulta. Entretanto em um ambiente de *data warehouse* dificilmente um evento isolado é examinado, é mais comum ocorrer sempre à utilização de visões de conjunto dos dados (Inmon, 1997).

3.5 Metadados

Um importante aspecto do ambiente de *data warehouse* diz respeito aos metadados. Singh (1997) os descreve como sendo o principal componente do *data warehouse*. Na literatura, a descrição mais comum que se encontra sobre o assunto é que eles representam "dados sobre dados" (Inmon, 1997). Complementado esta descrição, pode-se dizer que o metadado é a "descrição do dado, do ambiente onde ele reside, como ele é manipulado e para onde é distribuído". Outra forma, mais sucinta e direta, é definir metadado como "documentação" (Tronchin, 1998).

Em um *data warehouse* os metadados assumem um papel de grande importância. Duas comunidades diferentes são atendidas por dados operacionais e dados do *data warehouse*. Os dados operacionais são utilizados por profissionais de TI (Tecnologia da Informação) e usuários especializados, capazes de se localizar nos sistemas em função de seu treinamento e experiência. Entretanto, o *data warehouse* atende à comunidade de usuários de negócio, com funções táticas e gerenciais (Analistas de SAD – Sistemas de Apoio a Decisão). O analista de SAD é, geralmente, acima de tudo, um profissional especializado em uma determinada área de negócio. Na comunidade de analistas de SAD, não há, normalmente, um alto grau de especialização em computadores. Os analistas de SAD precisam de tanta ajuda quanto possível para usar eficientemente o ambiente de *data warehouse*, e os metadados se prestam muito bem para este fim (Domenico, 2001).

A importância dos metadados é destacada por (Hurwitz, 1996) que diz que "quando você está iniciando um projeto de *data warehouse*, deve começar com os metadados". Os metadados têm o papel de ocultar do usuário a complexidade de

acessar informações distribuídas, enquanto facilita a atualização e sincronização de vários bancos de dados.

Todas as fases de um projeto de *data warehouse* geram metadados. Tipicamente, os aspectos sobre os quais os metadados mantêm informações são: sobre a estrutura dos dados, segundo a visão do programador e segundo a visão do analista de SAD. Mantêm ainda informações sobre o modelo de dados, especificação dos arquivos (chaves e atributos), histórico de extrações, controle de acesso, etc.

Sem metadados, os dados não têm significado. São exemplos de metadados as descrições de registros em um programa de aplicação ou o esquema de um banco de dados descrito em seu catálogo ou ainda as informações contidas em um dicionário de dados.

3.6 Data Mart

No processo de desenvolvimento de um *data warehouse*, existem diversos fatores que afetam sua complexidade como, por exemplo, a construção do projeto que é lenta e cara. Objetivando o equilíbrio dos gastos e a obtenção de resultados em prazos mais curtos, é possível construir *data marts*, uma alternativa mais modesta aos *data warehouses*. Menores e mais baratos, os *data marts* podem ser vistos como pequenos *data warehouses* que se limitam a um determinado contexto, *data marts* “departamentais” (Inmon et al., 1997a).

Oneil (1997) diz que existem diferentes alternativas de implementação de um *data mart.* Sendo que a proposta original é a aquela onde os *data marts* são desenvolvidos a partir de um *data warehouse* central. Neste caso, os usuários acessam diretamente os *data marts* de seus respectivos “departamentos”. No *data*

warehouse são realizadas somente as análises que necessitam de uma visão global do contexto em que se encontra o *data warehouse*. Os *data marts* se diferenciam do *data warehouse* pelos seguintes fatores (Inmon et al., 1997a):

- São personalizados: Atendem às necessidades de um grupo de usuários específico;
- Menor volume de dados: Por atenderem a um único "departamento", armazenam um menor volume de dados;
- Histórico limitado: Os *data marts* raramente mantém o mesmo período histórico que um *data warehouse*;
- Dados sumarizados: Os *data marts* geralmente não mantém os dados no mesmo nível de granularidade do *data warehouse*, ou seja, os dados são, quase sempre, sumarizados quando passam do *data warehouse* para os *data marts*.

Um dos problemas encontrados nos *data marts* é o grande risco de desvio do modelo original, podendo acarretar em um crescimento desestruturado. Outro problema que pode ocorrer é a replicação das mesmas informações em vários locais por ser muito utilizado e estar em constante aperfeiçoamento, o que dificulta uma futura integração de todos os *data marts* em um único *data warehouse*.

3.7 Arquitetura do Data Warehouse

Orr (1996) introduz uma arquitetura para o *data warehouse* com 8 camadas, como um modo de representar a estrutura global dos dados, comunicação, processamento e apresentação que existem para computação do usuário final.

A arquitetura de um *data warehouse* é constituída de várias camadas interconectadas como visto na figura 05. Abaixo são listadas as principais características de cada uma delas.

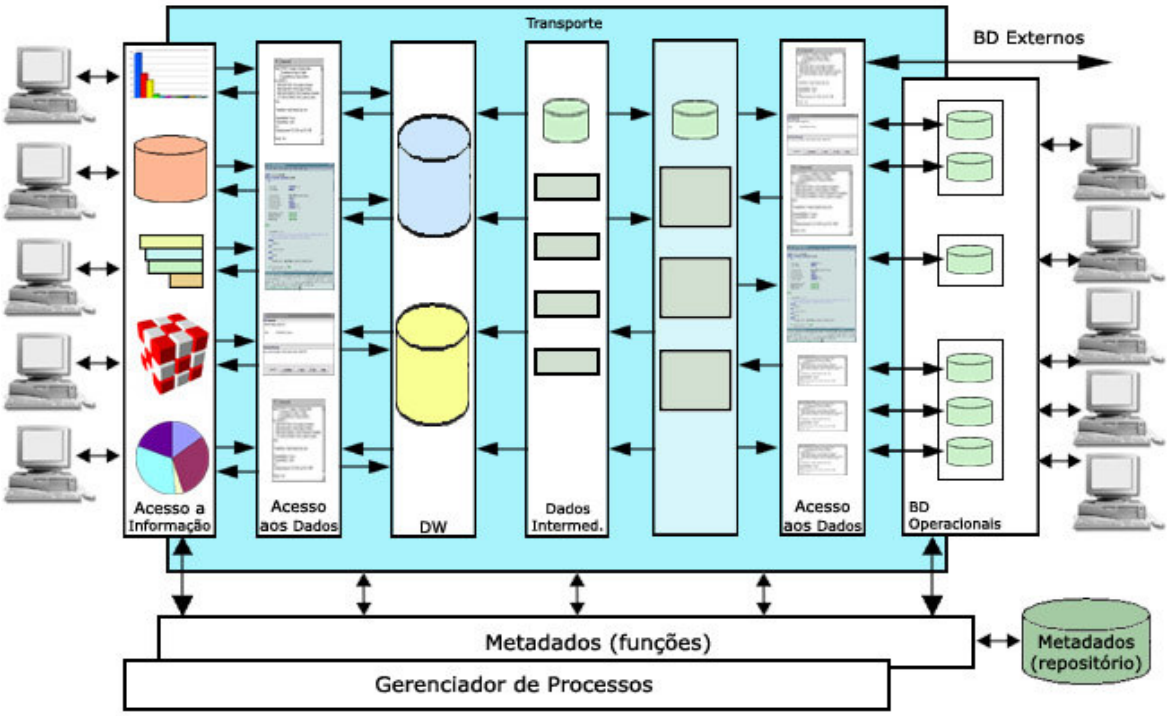

Figura 05. Arquitetura de um *data warehouse*.

3.7.1 Camada de Acesso à informação

A camada de acesso à informação da arquitetura de um *data warehouse* é a camada com a qual o usuário final tem contato direto. Ela representa as ferramentas que usuário final normalmente utiliza no dia-a-dia, como Excel, Lotus 1-2-3, Access, etc. Esta ferramenta também inclui hardwares e softwares envolvidos na visualização e impressão de relatórios, planilhas e gráficos para análise e apresentação.

Nas duas últimas décadas, a camada de acesso à informação sofreu uma enorme expansão, uma vez que hoje em dia, os usuários finais passaram a utilizar cada vez mais computadores e redes de computadores. Outro ponto que favorece este acontecimento, é sem dúvidas, o fato de que as ferramentas para manipular, analisar e

apresentar os dados estão cada vez mais sofisticadas e não impõem mais restrições no momento de implementar o *data warehouse*.

3.7.2 Camada de Acesso a Dados

A camada de acesso a dados do *data warehouse* permite que a camada de acesso à informação interaja com a camada operacional. Atualmente, a linguagem comumente utilizada para este acesso é a linguagem SQL (*Structured Query Language*), desenvolvida pela IBM como uma linguagem de consulta, mas que nos últimos 20 anos vem se tornando a linguagem padrão na troca de dados.

Esta camada é responsável pela interface entre ferramentas de acesso a informações e os bancos de dados operacionais, em alguns casos isto é tudo o que os usuários finais necessitam. Ela se comunica com diversos sistemas de bancos de dados, sistemas de arquivos e fontes sob diferentes protocolos de comunicação, o que é conhecido como acesso universal de dados.

3.7.3 Camada de Metadados

Para se obter acesso universal aos dados, é absolutamente necessário, manter alguma forma de diretório de dados ou repositório de metadados de informação. Este é o papel da camada de metadados. Metadados são dados sobre dados. Em um *data warehouse* é necessário que exista uma grande variedade de metadados, pois os metadados fornecem subsídios para que os usuários não se preocupem com a localização dos dados ou com a forma na qual estes estão armazenados.

3.7.4 Camada de Gerenciamento de Processos

A camada de gerenciamento de processos está envolvida com o controle dos vários processos que devem ocorrer para construção e manutenção do *data warehouse*.

Esta camada é responsável por priorizar e seqüenciar as cargas de dados que mantêm as informações do *data warehouse* atualizadas e consistentes. O administrador do *data warehouse* geralmente possui ferramentas para interagir com esta camada.

3.7.5 Camada de Transporte

A camada de transporte, também conhecida como *application messaging* ou *middleware*, diz respeito ao transporte das informações através da rede. Esta camada é responsável por gerenciar o transporte de informações como a coleta de mensagens e os dados gerados pelas transações.

A camada de transporte também é usada para isolar as aplicações operacionais das aplicações de apoio à decisão. É de fundamental importância para o *data warehouse* que esta camada seja atendida por uma ferramenta adequada. Deficiências nesta camada podem impedir que informações importantes sejam disponibilizadas no *data warehouse*.

3.7.6 Camada de Dados Operacionais e Dados Externos

Os sistemas aplicativos processam dados com a finalidade de dar suporte às necessidades operacionais críticas. Para que isto ocorra, os bancos de dados

operacionais proporcionam uma estrutura eficiente de processamento para armazenamento de um número relativamente pequeno de transações bem definidas. De qualquer modo, devido ao propósito limitado dos sistemas aplicativos, os bancos de dados projetados para dar suporte a esses sistemas, encontram uma certa dificuldade para acessar os dados com outras finalidades, como de gerenciamento e informacional. Esta dificuldade é ampliada pelo fato de vários sistemas aplicativos terem idades, muitas vezes, de 10 a 15 anos, implicando que a tecnologia de acesso a dados utilizada nestes sistemas se encontra ultrapassada.

A camada de dados operacionais e externos é composta pelos dados dos sistemas aplicativos de determinado contexto, por exemplo, uma empresa, e por informações procedentes de fontes externas que são integradas para compor o *data warehouse*. Um exemplo de fonte de dados externo é, as informações bancárias enviadas pelo banco de forma eletrônica para uma empresa (Gonçalves, 2003).

3.7.7 Camada do Data Warehouse

É nesta camada que se encontram armazenados os dados utilizados para se obter informações, é o *data warehouse* propriamente dito. Em alguns casos pode-se imaginar o *data warehouse* como sendo uma simples visão lógica ou virtual dos dados, podendo não envolver necessariamente o armazenamento dos mesmos. Um *data warehouse* físico pode armazenar muitas cópias de dados operacionais e externos em um formato de fácil acesso e altamente flexível.

3.7.8 Camada de Dados Intermediários (camada de gerenciamento de replicação)

A camada de dados intermediários, conhecida também como *staging area* ou camada de gerenciamento de cópias, inclui todos os processos necessários para

selecionar, editar, sumarizar, combinar e carregar dados de bancos de dados operacionais e/ou externos no *data warehouse*.

3.8 Modelo de dados

Os dados que se encontram armazenados em um *data warehouse* são geralmente organizados de modo a facilitar sua análise por um usuário especializado. Uma organização geralmente utilizada, é a de armazenar informações quantitativas (por exemplo, vendas de produtos) em grandes tabelas, chamadas tabelas de fatos, e dados qualitativos, informação descritiva (por exemplo, atributos do produto) armazenados em pequenas tabelas, chamadas tabelas de dimensão. Este modelo é chamado estrela, pois um simples objeto (tabela de fatos) está no centro do modelo conectado a um número de objetos (tabela de dimensão) radialmente. Além desse modelo, também se utiliza o modelo *Snowflake* (ou Floco de neve), que é na verdade um refinamento do modelo estrela onde a hierarquia dimensional é representada explicitamente pela normalização das tabelas de dimensão. Abaixo, mais detalhes sobre os dois modelos.

3.8.1 Modelo estrela

O modelo estrela se caracteriza por possuir uma tabela centralizada, chamada de tabela de fatos, que é cercada por tabelas normalizadas chamadas de tabelas de dimensões, ou apenas dimensões. Seu nome reflete o fato de se parecer com uma estrela. As dimensões assemelham-se a pontos que cercam a tabela fato, como ilustrado na Figura 06.

Os atributos de um modelo estrela geralmente incluem (URL 1):

- Uma tabela de fatos relativamente grande, contendo milhões ou bilhões de linhas contendo os fatos mensuráveis ou aditivos, como transações relativas a vendas ou eventos;
- Tabelas de dimensões relativamente pequenas e altamente normalizadas contendo dados descritivos sobre os fatos da tabela central de fatos, tais como informações referentes a clientes ou de localização;
- Uma tabela de fatos central que é dependente das tabelas de dimensões que a cercam, utilizando uma relação pai/filho, onde a tabela de fatos é a tabela filho e as dimensões a tabela pai.

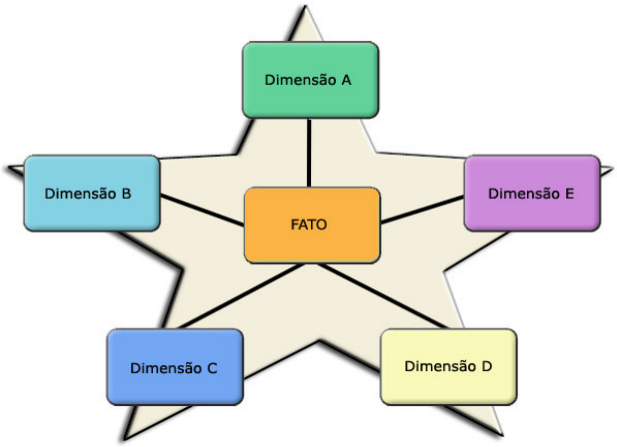

Figura 06. Modelo Estrela.

3.8.2 Variação do Modelo Estrela (Snowflake)

Outro tipo de estrutura bastante comum, conforme abordado em (Campos et al., 1997), é o modelo de dados floco de neve ou "*Snowflake*" (Floco de Neve), que consiste em uma extensão do modelo estrela, onde cada uma das "pontas da estrela" passa a ser o centro de outras estrelas.

O modelo *Snowflake* (figura 07) é o resultado da decomposição de uma ou mais dimensões que possuem hierarquias entre seus membros. Podemos definir

relacionamentos muitos para um entre os membros de uma dimensão, formando por meio desses relacionamentos entre entidades dimensão, uma hierarquia.

O modelo surge da des-normalização e redução de cardinalidade do modelo estrela "quebrando-se" a tabela original ao longo de hierarquias existentes em seus atributos. Este modelo pode ser ilustrado imaginando-se a classificação de um automóvel, onde a dimensão do produto possui uma hierarquia definida: categoria se divide em marca, e marca se divide em produtos. Kimball (1996) aconselha os projetistas a resistirem à tentação de transformar modelo Estrela em modelos Floco de Neve, devido ao impacto da complexidade deste tipo de estrutura sobre o usuário final, enquanto que o ganho em termos de espaço de armazenamento seria pouco relevante.

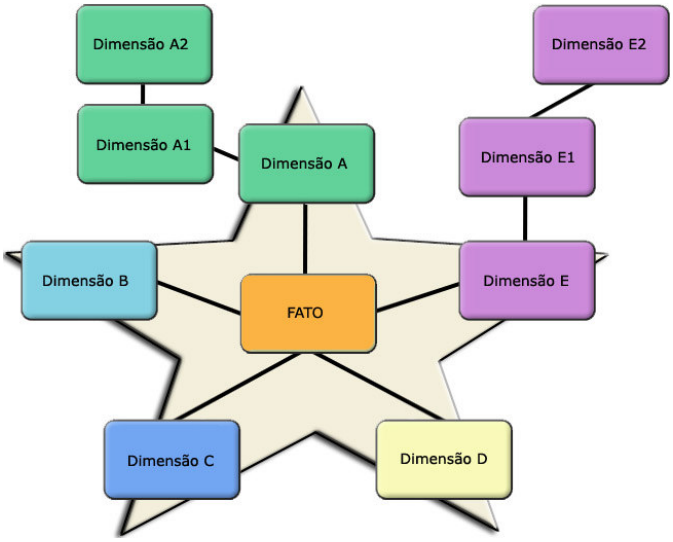

Figura 07. Modelo *Snowflake*.

3.9 Tipos de Implementação

A tecnologia de *data warehouse* se encontra em um estado de evolução, que pode ser considerado como uma resposta à crescente complexidade deste

ambiente e às dificuldades de integração entre todos os componentes. Uma das principais preocupações que os desenvolvedores deste ambiente tem, é em como integrar o *data warehouse* às diversas fontes heterogêneas e externas, aos *data marts*, aplicações servidoras, "WEB" e "*data mining*", entre outros tipos de ferramentas disponíveis (Firestone, 1998).

Basicamente a constituição de um *data warehouse* pode ser feita com três tipos de implementação: *top-down*, *bottom-up* e distribuída (ou combinada). A escolha do tipo de implementação é fator importante na seleção da tecnologia apropriada para o desenvolvimento e a implantação deste ambiente. Atualmente, considera-se que os problemas em um *data warehouse* estão mais relacionados com implementação e arquitetura do que com a tecnologia disponível (Melo, 1997). A seguir, comenta-se cada uma destas abordagens.

3.9.1 Implementação Top-Down

Este tipo de implementação é conhecido como padrão inicial do conceito de *data warehouse* (Inmom et al., 1997a). A Figura 08 exemplifica esta implementação que se inicia coma a extração, transformação, migração e carregamento de dados oriundos dos sistemas legados, transacionais e/ou de fontes externas. No processo de extração, transformação e migração, os dados são retirados de suas origens e armazenados na *Staging Area*. Em seguida os dados e os metadados são carregados para o *data warehouse*. A partir do *data warehouse* os dados e metadados são extraídos para os *data marts*. Nos *data marts*, as informações se encontram em um maior nível de sumarização e, normalmente, não apresentam o nível histórico encontrado no *data warehouse* (Inmom, 1998).

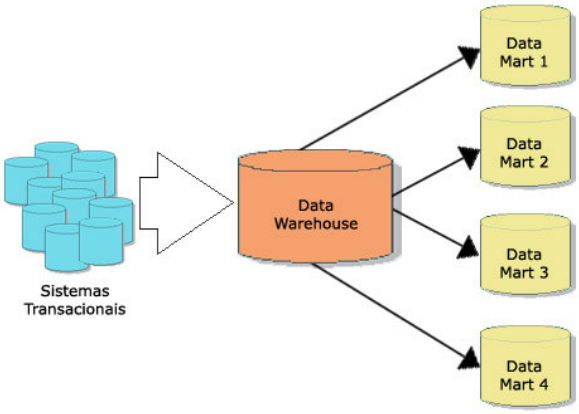

Figura 08. Implementação *Top-Dow*

Na arquitetura *top-down*, a integração entre o *data warehouse* e os *data marts* é automática, desde que se mantenha uma disciplina na construção, partindo da premissa de que os *data marts* são subconjuntos do *data warehouse*. Existem algumas ferramentas para gerar *data marts* a partir do *data warehouse*, um exemplo é a *Microstrategy's DSS Server*.

A seguir são listadas as principais vantagens e desvantagens da implementação *top-down* de acordo com (Hackney, 1998):

As principais vantagens:

- Herança de arquitetura - Todos os *data marts* originados a partir de *um data warehouse*, utilizam a arquitetura e os dados deste *data warehouse*, permitindo uma fácil manutenção;
- Visão de empreendimento - O *data warehouse* concentra um visão geral do contexto, sendo possível a partir dele extrair níveis menores de informações;
- Repositório de metadados centralizado e simples - O *data warehouse* provê um repositório de metadados central para o sistema. Esta centralização permite manutenções mais simples do que aquelas realizadas em múltiplos repositórios;
- Controle e centralização de regras - A implementação *top-down* garante a existência de um único conjunto de aplicações para extração, limpeza e

integração dos dados, além de processos centralizados de manutenção e monitoração.

As principais desvantagens:

- Implementação muito longa – um *data warehouse* geralmente, é desenvolvido de modo iterativo, por áreas de assuntos, como por exemplo, vendas, finanças e recursos humanos. Mesmo assim, são necessários, em média, 15 ou mais meses para que a primeira área de assunto entre em produção, dificultando a garantia de apoio político e orçamentário;
- Alta taxa de risco - Não existem garantias para o investimento neste tipo de ambiente;
- Heranças de cruzamentos funcionais - É necessário uma equipe de desenvolvedores e usuários finais, altamente capacitados para avaliar as informações e consultas que garantam à empresa habilidade para sobreviver e prosperar na arena de mudanças de competições políticas, geográficas e organizacionais;
- Expectativas Relacionadas ao Ambiente - A demora do projeto e a falta de retorno pode induzir expectativas nos usuários.

3.9.2 Implementação Bottom-Up

Devido a algumas desvantagens da implementação *top-down* como, ser politicamente difícil de ser definida e muito cara, requerer um tempo grande para implementação, investimento e sem apresentar retorno rápido, a implementação *bottom-up* vem se popularizando. A Figura 09 apresenta a referida arquitetura (Firestone, 1998).

A idéia central desta implementação é a construção do *data warehouse* de forma incremental, a partir do desenvolvimento de *data marts* independentes. Na literatura, essa implementação foi introduzida por Ralph kimball. Esse processo se inicia com a extração, transformação e a integração dos dados para um ou mais *data marts*.

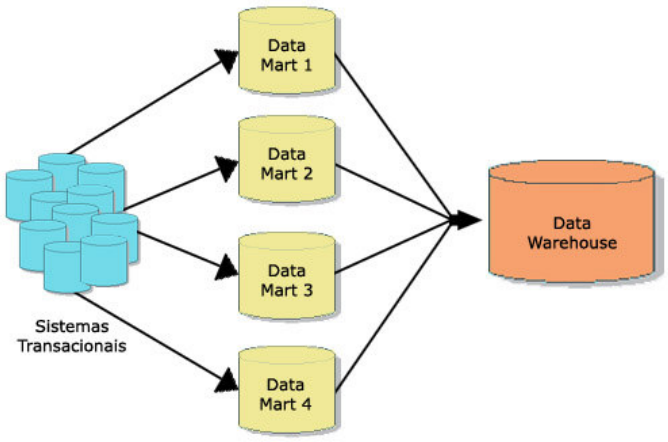

Figura 09. Implementação *Bottom-up*

A seguir são listadas as principais vantagens e desvantagens da implementação *top-down* de acordo com (Hackney, 1998).

As principais vantagens:

- Implementação rápida - A construção dos *data marts* é altamente direcionada, permitindo um rápido desenvolvimento. Normalmente, um *data mart* pode ser colocado em produção em um período de seis a nove meses;
- Retorno Rápido - A implementação baseada em *data mart* com incremento demonstra rapidamente seu valor, permitindo uma base para investimentos adicionais, com um nível mais elevado de confiança;
- Manutenção do Enfoque da Equipe - Um dos maiores desafios do desenvolvimento de um *data warehouse* é a manutenção do mesmo enfoque por toda a equipe. A elaboração de *data marts* incrementais, permite que os

principais negócios sejam enfocados inicialmente, sem que haja gastos no desenvolvimento de áreas que não são essenciais ao problema;

- Herança Incremental - A estratégia de *data marts* incrementais obriga a entrega de recursos de informação, passo a passo. Isto permite a equipe crescer e aprender, reduzindo os riscos. As avaliações de ferramentas, tecnologias, consultores e vendedores só devem ser realizadas uma vez, a não ser que existam restrições que impeçam o reaproveitamento.

As principais desvantagens:

- Perigo de *LegaMarts* - Um dos maiores perigos no *data warehouse* é a criação de *data marts* independentes. O advento de ferramentas de "*drag-and-drop*" facilitou o desenvolvimento de soluções individuais, de acordo com necessidades especificas. Estas soluções podem não considerar a arquitetura de forma global. Desta forma, os *data marts* independentes transformam-se em *data marts* legados, ou *LegaMarts*. Os *LegaMarts* dificultam, quando não inviabilizam futuras integrações. Eles são parte do problema e não da solução;
- Desafio de Possuir a Visão de Empreendimento - Durante a construção dos *data marts* incrementais é necessário que se mantenha um rígido controle do negócio como um todo. Este controle requer um maior trabalho ao extrair e combinar as fontes individuais do que utilizar um *data warehouse*;
- Administrar e Coordenar Múltiplas Equipes e Iniciativas - Normalmente, esse tipo de implementação emprega o desenvolvimento de *data marts* em paralelo. Isto pode conduzir a uma rígida administração tentando coordenar os esforços e recursos das múltiplas equipes, especialmente nas áreas de regras e semântica empresariais;
- A Maldição de Sucesso - A arquitetura com *data marts* incrementais carrega a "maldição de sucesso". Nestes casos, os usuários finais do *data mart* encontram-se felizes querendo mais informação para seus *data marts*. Ao

mesmo tempo, outros usuários de outros *data marts* aguardam o incremento de seus *data marts*. Isto conduz a equipe de *data marts* a vencer desafios políticos, de recurso e de administração.

3.9.3 Implementação Combinada

Além das duas formas de implementação já vistas, *top-down* e *bottom-up*, existe ainda a implementação combinada que tem o propósito de integrar estes dois tipos de implementação. Nesta implementação, primeiramente é efetuada a modelagem dos dados do *data warehouse*, e em seguida a implementação de partes desse modelo, as quais são escolhidas por área de interesse e constituem os *data marts*. Cada *data mart* gerado a partir do macromodelo de dados do *data warehouse* é integrado ao modelo físico do *data warehouse* (Figura 10).

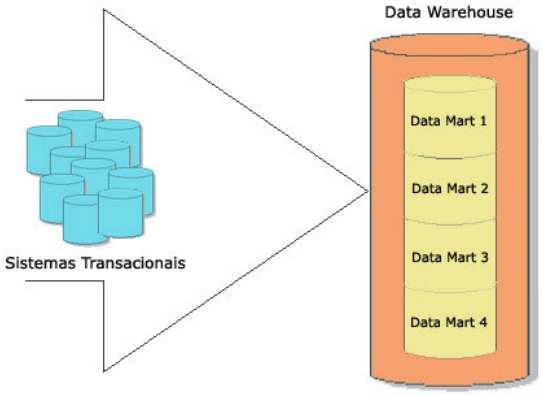

Figura 10. Implementação Combinada.

A principal vantagem deste tipo de implementação é a garantia da consistência dos dados. Esta garantia é obtida em virtude de o modelo de dados para os *data marts* ser único, possibilitando realizar o mapeamento e o controle dos dados.

3.10 Passos para Implantação de um Data Warehouse

A tecnologia de *data warehouse* não é um produto, mas sim um projeto que envolve uma série de etapas de análise e implementação, com a participação de diversas tecnologias. Isto que faz com que possua um certo grau de complexidade, por isso a observação de alguns passos, que facilitem a obtenção de sucesso em um projeto de *data warehouse* é de suma importância.

Os passos a seguir foram retirados de (META Group,1996), onde o autor sugere sete etapas para a criação de um *data warehouse*, que pode ser inicialmente um *data mart*.

- Passo 1 – Os primeiros resultados devem estar disponíveis a curto prazo. É importante traduzir rapidamente as necessidades encontradas em uma especificação que possa ser construída em etapas. Essa abordagem minimiza riscos e o tempo de apresentação dos resultados iniciais.
- Passo 2 – Mesmo para especialistas, construir um *data warehouse* pode ser um desafio de integração de sistemas. Os dados de produção e de fontes externas precisam ser mapeados para o modelo de dados do *data warehouse*. Precisa haver um sincronismo entre os dados operacionais e os dados de tomada de decisão. Para análises multidimensionais, os dados precisam ser transferidos e sincronizados em um banco de dados multidimensional.
- Passo 3 – A escolha do banco de dados de suporte ao *data warehouse* precisa ser criteriosa. Alguns critérios que devem ser analisados são: desempenho na carga e indexação dos dados, tempo de resposta, capacidade de armazenamento, paralelismo e escalabilidade.
- Passo 4 – deve-se considerar as ferramentas disponíveis no mercado que normalmente ajudam a compor um ambiente *data warehouse* com a finalidade

de prover: interfaces amigáveis, geração de relatórios, análises multidimensionais, acesso via *Web* e *data mining*.

- Passo 5 – Deve-se construir um *data warehouse* que possa ser expandido, mantendo níveis aceitáveis de desempenho até centenas de gigabytes.
- Passo 6 – O ambiente de *data warehouse* deve ser aberto para permitir que os componentes ou ferramentas identificadas no passo 4 possam ser substituídas por outras mais atuais e eficientes.
- Passo 7 – Deve-se considerar o sistema de armazenamento que fisicamente gerencia o tráfego, alocação, backup e restauração dos dados.

3.11 Conclusão

A tecnologia de *data warehouse* vem se tornado, cada vez mais, uma grande área de estudo e aplicação, tanto no universo cientifico e acadêmico, como em empresas. Apesar de se tratar de uma tecnologia nova e se encontrar em estágio precoce de desenvolvimento, o *data warehouse* está cativando cada vez mais interessados na área. Talvez isto ocorra pela fascinação ocasionada pela possibilidade de se acessar informações confiáveis com boa velocidade e a garantia de qualidade de dados oriundos de grandes volumes de dados gerados e acumulados durante um grande período, entre outras características da tecnologia.

Por outro lado, este desenvolvimento precoce faz com que existam relatos tanto de sucessos de desenvolvimento desta tecnologia, assim como de fracassos.

No próximo capítulo veremos a tecnologia de mineração de dados,(ou *data mining*) responsável por vasculhar os dados organizados pelo *data warehouse*, a procura de conhecimento (padrões).

4 DATA MINING

Aqui serão apresentados conceitos e características desta tecnologia que representa a fase mais importante do processo KDD, a mineração de dados ou data mining. Nosso principal objetivo neste capítulo é caracterizar as técnicas utilizadas nesse trabalho.

4.1 Introdução

Historicamente diversos nomes foram atribuídos à idéia de se encontrar padrões úteis em dados, entre eles o de *data mining* (Fayyad et al., 1996). A frase *Knowledge Discovery in Databases* foi cunhada pela primeira vez no primeiro *workshop* de KDD em 1989 (Piatetsky et al., 1991) para enfatizar que o conhecimento é o produto final de uma descoberta direcionada aos dados. Fayyad se refere a KDD como sendo o processo de descoberta de conhecimento útil a partir de dados como um todo e *data mining* como uma etapa particular deste processo. *Data Mining* é a aplicação de algoritmos específicos para a extração de padrões de dados.

Quando se fala em *Data Mining* não está se considerando apenas consultas complexas e elaboradas que visam ratificar uma hipótese gerada por um usuário em função dos relacionamentos existentes entre os dados, e sim da descoberta de novos fatos, regularidades, restrições, padrões e relacionamentos.

A errônea aplicação de métodos de *data mining* pode acarretar à descoberta de padrões inválidos e sem sentido.

4.2 Técnicas de Data Mining

A mineração de dados possui uma grande quantidade de técnicas, algoritmos e procedimentos. Nas seções a seguir são mostrados, resumidamente, os fundamentos das técnicas utilizadas neste trabalho.

4.2.1 Análise de Componentes Principais

A análise de componentes principais ou (em inglês) *Principal Component Analysis* – PCA (Box et al., 1978), (Manly, 1986), (Aitchison, 1986), (Gower, 1966), (Joreskog, 1976), (Reyment et al., 1996), (Zhou, 1989) é uma técnica utilizada para redução do número de variáveis em uma base multidimensional de dados. Uma vez reduzida a dimensionalidade deste conjunto de variáveis fica mais fácil a interpretação de interdependências existentes entre elas. Os componentes principais são combinações lineares das variáveis originais cuja consideração de seus valores, permite um melhor entendimento do conjunto de dados. Ou seja, a análise de componentes principais é um mecanismo de simplificação de dados através da redução do número de variáveis.

Considerando uma base de dados com .p. variáveis (X_1, X_2, ..., X_p), o objetivo da análise de componentes principais é produzir .p. componentes principais, (Z_1, Z_2,..., Z_p) que não sejam correlacionadas. O fato de não serem correlacionadas indica que as componentes principais estão medindo dimensões diferentes nos dados.

As componentes principais devem ser ordenadas de forma que a primeira tenha a maior variação, a segunda tenha a segunda maior variação, e assim por diante. Quando se realiza este tipo de análise espera-se que a variância da maioria

das componentes principais seja tão pequena que possa ser desprezada. Nestes casos, a variação da base de dados pode ser descrita por um pequeno número de componentes principais que possuam variâncias não desprezíveis. Assim, a variação das variáveis originais pode ser calculada através de um número menor de outras variáveis, que são as componentes principais.

Para que uma análise de componentes principais resulte na redução de um grande número de variáveis originais, é necessário que as variáveis originais sejam altamente correlacionadas, positivamente ou negativamente. Sendo assim, os componentes principais importantes serão de grande interesse como medidas das dimensões fundamentais dos dados.

Uma análise de componentes principais começa com dados de p variáveis para n indivíduos ou objetos (Tabela 2).

Tabela 2 Exemplo de Base de Dados.

Objetos	**X_1**	**X_2**	**...**	**X_p**
1	6.00	13.0	140	1.00
2				
...				
N				

O primeiro componente principal é a combinação linear das variáveis X_1, X_2, ..., X_p,

$$Z_1 = a_{11} X_1 + a_{12} X_2 + ... + a_{1p} X_p$$

que varia tanto quanto possível para os objetos, sujeito à seguinte condição:

$$a_{11}^2 + a_{12}^2 + ... + a_{1p}^2 = 1$$

Esta restrição é introduzida porque senão a variância de Z_1 poderia crescer ao se aumentar qualquer valor de a_{1j} .

$$Z_2 = a_{21} X_1 + a_{22} X_2 + ... + a_{2p} X_p$$

O segundo componente principal, apresenta a variância tão grande quanto possível sujeito à restrição de que e também à condição de que Z_1 e Z_2 não são correlacionados.

$$a_{21}^2 + a_{22}^2 + ... + a_{2p}^2 = 1$$

Os componentes principais subseqüentes são obtidos da mesma forma. Se existirem p variáveis então existirão até p componentes principais, embora nem todos representem as informações contidas nos dados.

Uma análise do componente principal implica basicamente na determinação dos autovalores da matriz de covariância. Esta matriz é simétrica e tem a seguinte forma:

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ c_{p1} & c_{p2} & \cdots & c_{pp} \end{bmatrix}$$

Onde o elemento da diagonal C_{11} é a variância de X_i e C_{ij} é a covariância de X_i e X_j.

As variâncias dos componentes principais são os autovalores da matriz C. Existem p autovalores, e alguns deles podem ser iguais a zero; porém autovalores negativos não existem para a matriz C. Assumindo-se que os autovalores são ordenados como $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, então $\lambda_i \geq 0$ correspondente ao i-ésimo componente principal.

$$Z_i = a_{i1} X_1 + a_{i2} X_2 + ... + a_{ip} X_p$$

Conseqüentemente, a variância de Z_i é igual a λ_i e as constantes a_{i1}, a_{i2}, ... , a_{ip} são os elementos do autovetor correspondente.

Uma propriedade importante dos autovalores de uma matriz é que sua soma é igual ao traço da matriz, que é a soma dos elementos de sua diagonal. Por conseguinte:

$$\lambda_1 + \lambda_2 + ... + \lambda_p = c_{11} + c_{22} + \ldots + c_{pp}$$

Como c_{ii} é a variância de X_i e λ_i é a variância de Z_i, a soma das variâncias dos componentes principais é igual à soma das variâncias das variáveis originais. Pode-se então dizer que os componentes principais computam toda a variação dos dados originais.

De modo a se evitar que uma variável tenha uma influência excessiva nos componentes principais, é comum, ao início da análise, normalizar as variáveis originais para que apresentem média igual a zero e variância igual a um. A matriz C toma então a seguinte forma:

$$C = \begin{bmatrix} 1 & c_{12} & \cdots & c_{1p} \\ c_{21} & 1 & \cdots & c_{2p} \\ . & . & \cdots & . \\ c_{p1} & c_{p2} & \cdots & 1 \end{bmatrix}$$

Onde $c_{ij} = c_{ji}$ é a correlação entre X_i e X_j. Pode-se considerar que a análise de componentes principais é executada na matriz de correlação. Neste caso, a soma dos termos da diagonal, e portanto a soma dos autovalores, é igual ao número de variáveis.

O algoritmo da análise de componentes principais está ilustrado na Figura 11.

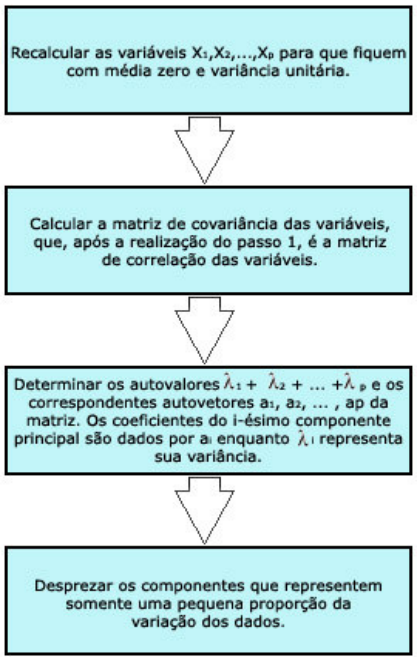

Figura 11. Algoritmo da analise de componentes principais.

4.2.2 Análise de Agrupamento

A técnica de análise de agrupamento ou (em inglês) *cluster analysis* (Anderberg, 1973), (Dilly, 1995), (Groth,1998), (Everitt, 1980), (Davis,1986), (Monteiro et al., 2000) é um termo usado para descrever diversas técnicas numéricas cujo propósito fundamental é classificar os valores de uma matriz de dados sob estudo em grupos discretos. Ela consiste em agrupar tipos similares de dados ou identificar exceções de acordo com valores em comum. Quando esta técnica é aplicada, ainda não é conhecida nenhuma classe, sendo que o objetivo da função de análise de agrupamento é produzir uma segmentação do conjunto de registros de entrada de acordo com algum critério. Este critério é estabelecido por uma ferramenta de análise de agrupamento e as funções podem produzir descrições implícitas e explicitas.

A técnica classificatória multivariada da análise de agrupamentos pode ser utilizada quando se deseja explorar as similaridades entre indivíduos (modo Q) ou entre variáveis (modo R) definindo-os em grupos, considerando simultaneamente, no primeiro caso, todas as variáveis medidas em cada indivíduo e, no segundo, todos os indivíduos nos quais foram feitas as mesmas mensurações. Segundo esse método, desenvolvido, inicialmente em Zoologia por taxonomistas numéricos, procura-se por agrupamentos homogêneos de itens representados por pontos num espaço n-dimensional em um número conveniente de grupos relacionando-os através de coeficientes de similaridades ou de correspondências.

Métodos de Classificação

Segundo (Davis,1986), os diversos métodos para a análise de agrupamentos podem ser enquadrados em quatro tipos gerais:

a. Métodos de partição: procuram classificar regiões no espaço, definido em função de variáveis, que sejam densamente ocupados em termos de observações daqueles com ocupação mais esparsa;
b. Métodos com origem arbitrária: procuram classificar as observações segundo “k” conjuntos previamente definidos; neste caso “k” pontos arbitrários servirão como centróides iniciais e as observações irão se agrupando, por similaridade, em torno desses centróides para formar agrupamentos;
c. Métodos por similaridade mútua: procuram agrupar observações que tenham uma similaridade comum com outras observações; inicialmente uma matriz n x n de similaridades entre todos os pares da observação é calculada; em seguida, as similaridades entre colunas são repetidamente recalculadas; colunas representando membros de um único agrupamento tenderão apresentar intercorrelações próximas a 1 e valores menores com não membros;

d. Métodos por agrupamentos hierárquicos: são as técnicas mais comumente usadas em Geologia; a partir da matriz inicial de dados obtém-se uma matriz simétrica de similaridades e inicia-se a detecção de pares de casos com a mais alta similaridade, ou a mais baixa distância; para essa combinação, segundo níveis hierárquicos de similaridade, escolhe-se entre os diversos procedimentos aglomerativos de tal modo que cada ciclo de agrupamento obedeça a uma ordem sucessiva no sentido do decréscimo de similaridade.

Metodologia para agrupamentos hierárquicos

Partindo de uma matriz inicial de dados [n*p], onde "n" linhas representam casos, espécimes ou amostras, no sentido geológico, e as "p" colunas as variáveis, feitas às comparações, usando um coeficiente de similaridade qualquer entre linhas, obtém-se uma matriz inicial de coeficiente de similaridade de tamanho $[\ ]\ n\ n\ *$, que será utilizada no modo Q (similaridade entre indivíduos). Se a comparação for entre colunas, obter-se-á uma matriz inicial de coeficientes de similaridade inicial $[\ ]\ p\ p\ *$, que será utilizada no modo R (similaridade entre variáveis). Embora diversas medidas de similaridade tenham sido propostas, somente duas são geralmente usadas: o coeficiente de correlação de Pearson e a medida de distância euclideana. Se as variáveis forem padronizadas a partir da matriz inicial de dados, dando o mesmo peso a cada uma delas, qualquer um desses coeficientes poderá ser diretamente transformado no outro.

Na matriz inicial de coeficientes de similaridade estes representam o grau de semelhança entre pares de objetos e os mesmos deverão ser arranjados de acordo com os respectivos graus de similaridade de modo a ficarem agrupados segundo uma disposição hierárquica. Os resultados quando organizados em gráfico, do tipo dendrograma mostrarão as relações das amostras agrupadas.

Várias técnicas de agrupamentos têm sido propostas, e os métodos mais comumente usados são: “ligação simples” (*single linkage method* ou *nearest neighbor*); “ligação completa” (*complete linkage method* ou *farthest neighbor*); ”agrupamento pareado proporcionalmente ponderado" (*weighted pair-group method*, WPGM); “agrupamento pareado igualmente ponderado” (*unweighted pair-group method*, UPGM); “variância mínima” (*minimum variance clustering* ou *Ward’s method of sum-of-squares method*).

No método de ligação simples os grupos iniciais são determinados pelos mais altos coeficientes de associação mútua. Para admissão de novos membros aos grupos é suficiente encontrar quais os que representam os maiores coeficientes de associação com um dos elementos de determinado grupo. A ligação será estabelecida a esse nível de associação com todo o grupo.

No método de ligação completa os grupos são determinados pelos mais baixos coeficientes de associação mútua. Ambos são os métodos mais simples, mas também os que apresentam os resultados mais distorcidos. Com o uso dos métodos de ligações completas espera-se obter resultados mais rigorosos.

No método de agrupamento pareado procura-se também inicialmente pelos mais altos coeficientes de associação mútua. Em seguida esses pares de casos fornecerão valores médios originando um novo elemento singular. No "método de agrupamento pareado igualmente ponderado" para o cálculo dos valores médios atribui-se sempre o mesmo peso aos dois elementos que estão sendo integrados. No método de agrupamento pareado proporcionalmente ponderado para cada agrupamento é dado um peso proporcional ao número de objetos que o constitui, de tal modo que a incorporação e um novo elemento a um grupo baseiam-se no nível médio de similaridade desse elemento com todos os que fazem parte do grupo. Tanto

num caso como no outro, alternativamente, em vez de obter valores médios entre os casos podem ser utilizados centróides e verificados as distâncias entre os mesmos.

Dendrograma

A forma gráfica mais usada para representar o resultado final dos diversos agrupamentos é o dendrograma (Figura 13). Nele estão dispostas linhas ligadas segundo os níveis de similaridade que agruparam pares de espécimes ou de variáveis. Como este gráfico é uma simplificação em duas dimensões de uma relação n-dimensional é inevitável que algumas distorções quanto à similaridade apareçam. A medida de tal distorção pode ser obtida por um coeficiente de correlação, entre os valores da matriz inicial de similaridade e aqueles derivados do dendrograma.

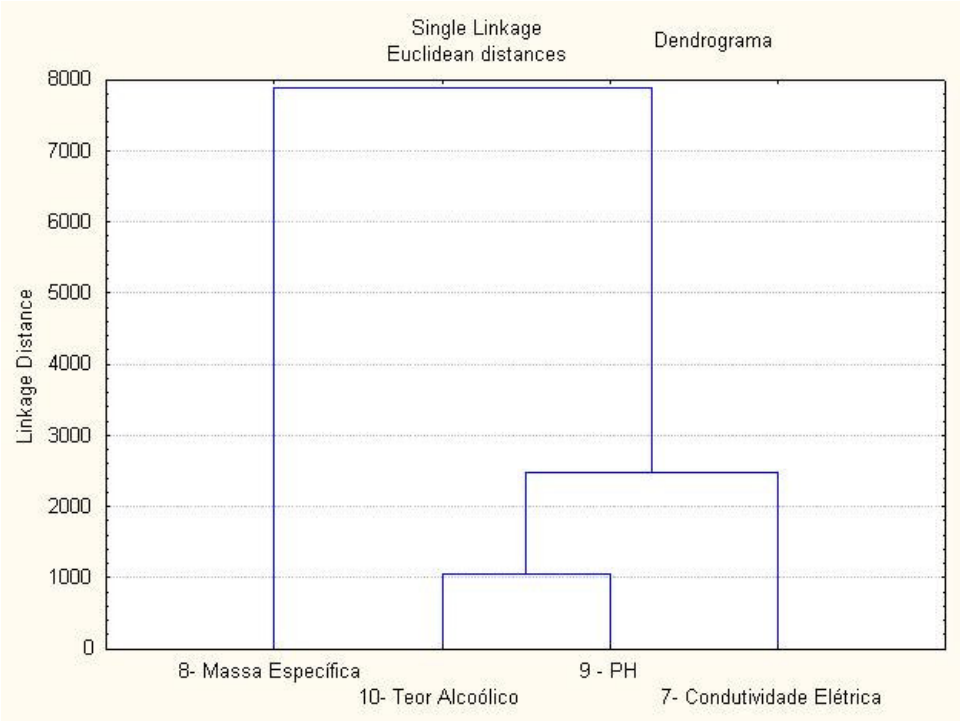

Figura 12. Exemplo de dendrograma.

Visualmente isso pode ser também verificado por meio da construção de um sistema de eixos ortogonais. Nele os valores dos coeficientes de similaridade originais estarão na abscissa e os coeficientes de similaridade a partir do dendrograma em

ordenada. Se ambas as matrizes forem idênticas os pontos cairão sobre uma linha reta que passa pela origem do sistema. Desvios dos pontos em relação a essa reta indicarão as distorções. Se situadas acima da reta indicarão coeficientes de similaridade apontados pelo dendrograma mais altos que os originais e vice-versa.

Tabela 3 – Matriz de correlação do exemplo acima.

	7	8	9	10
7 – Condutividade	0	7886	3162	2490
8 – Massa Específica	7886	0	9975	8932
9 – PH	3162	9975	0	1044
10 – Teor Alcoólico	2490	8932	1044	0

O dendrograma pode ser construído a partir da matriz euclidiana de correlação. O exemplo acima foi calculado a partir da matriz mostrada na tabela 3.

No dendrograma mostrado acima podemos classificar esses dados em grupos, representando uma reta que corte o dendrograma em quantas partes quiser.

4.2.3 Regressão Múltipla (Multiple Regression)

O principal objetivo da análise de regressão é o de previsão. Nosso propósito na análise de regressão é o desenvolvimento de um modelo estatístico que possa ser utilizado para prever os valores de uma variável dependente ou variável de resposta, com base nos valores de pelo menos uma variável independente ou explicativa. No caso da regressão múltipla utilizam-se diversas variáveis explicativas (X_1, X_2, ..., X_3), para prever uma variável de resposta Y_1 (Bacci, 2000), (Li, 1964), (Levine et al., 2000).

A análise de regressão é freqüentemente usada para:

- Determinar como a variável de resposta muda com as mudanças das variáveis independentes (explicativas);

- Predizer o valor da variável de resposta para todo o valor da variável independente, ou a combinação dos valores das variáveis independentes.

Alguns cuidados que se deve tomar quando da utilização da análise de regressão:

a. as relações entre as variáveis devem ser lineares;
b. evitar um número inferior de casos em relação ao número de variáveis consideradas, sendo recomendado que tal relação seja da ordem de 10 a 20 vezes superior;
c. evitar variáveis independentes redundantes, isto é, que tenham um alto coeficiente de correlação entre si;
d. verificar, utilizando resíduos, a presença de valores anômalos.

A Equação da Regressão

A equação da regressão é uma representação algébrica da regressão linear e é usada para descrever o relacionamento entre a variável de resposta Y_1 e as variáveis explicativas (X_1, X_2, ..., X_3). Seu modelo geral é representado por

$$\text{Resposta } (Y_1) = \text{constante} + \text{coeficiente}(X_1) + \ldots + \text{coeficiente}(X_k) + \xi_i$$

A condição inicial, como na regressão linear simples, é descrita por

$$Y = b_0 + b_1(X_1) + e_1$$

Onde:

- A resposta (Y) - é o valor da resposta.
- Constante (b_0) - é o valor da variável da resposta quando as variáveis (X_k) são zero. A constante é chamada também de intercepção porque determina onde a linha de regressão intercepta o eixo y.
- (X_1 ... X_k) - é o valor das variáveis explicativas.

- Os coeficientes (b_1, b_2, ..., b_k) representam a mudança estimada na resposta média para cada mudança de unidade no valor de X. Ou seja é a mudança em Y que ocorre quando X aumenta por uma unidade.
- e_1 é o erro, a variabilidade de Y não explicada pela relação linear.

A variável que, em seguida, mais reduz a variabilidade do erro é em seqüência adicionada de tal modo que: $Y = b_0 + b_1(X_1) + b_2(X_2) + e_1$, sendo b_0, b_1 e b_2 calculados e $e_2 < e_1$.

O processo segue por etapas até que o comportamento de todas as variáveis independentes em relação à dependente seja verificado. Os coeficientes "b_i" são conhecidos como parciais de regressão porque cada um deles fornece a taxa de mudança na variável dependente correspondente à respectiva variável independente, mantendo constantes as demais variáveis independentes.

Uma das mais importantes aplicações da análise de regressão múltipla é a escolha, entre diversas variáveis explicativas, daquelas mais úteis na previsão de Y e, para tanto, o método "passo a passo" (*stepwise multiple regression*) é o mais recomendado.

A variância total de Y é em parte "explicada" pelas diversas variáveis X's e o restante pela variabilidade devido ao erro (e1). É claro que o termo "explicada" tem apenas um significado numérico não implicando necessariamente em um conhecimento causa-efeito sobre o porquê da relação existente.

Os tamanhos relativos dessas duas componentes de variância são obviamente de grande interesse quando da aplicação da análise de regressão múltipla. A proporção da variância dos Y´s observados "explicadas" por uma equação de regressão ajustada é representada pelo coeficiente de determinação R^2.

$$R^2 = \frac{(\text{variância de Y explicada pela análise de regressão})}{(\text{variância total})}$$

Os Valores de R^2 irão dispor-se no intervalo [0-1], fornecendo uma medida dimensional de quantidade do ajuste do modelo de regressão múltipla aos dados. Se o valor de R^2 for próximo de 1 isso significa que as diversas variáveis X's medidas são responsáveis quase que totalmente pela variabilidade de Y. Caso contrário, R^2 apresentará um valor próximo a zero. Como os coeficientes de regressão são parciais devem ser obtidas as porcentagens explicadas da soma de quadrados de Y segundo 2^k - 1 combinações, onde k é o número de variáveis independentes. Finalmente verifica-se a contribuição pura de cada variável independente por comparações sucessivas entre os diversos resultados.

Embora a regressão múltipla seja multivariada no sentido de que mais de uma variável é medida simultaneamente em cada observação, trata-se na realidade de uma técnica univariada, pois o estudo é apenas em relação à variação da variável dependente Y, sem que o comportamento das variáveis independentes, X's, seja objeto de análise.

4.3 Conclusão

O uso do *data mining* como fase do processo KDD, constitui uma poderosa ferramenta, uma vez que, automatiza, em parte, a captura e análise de dados, simplificando o processo de exploração de grandes massas de dados.

Existem diversas técnicas de *data mining* que podem ser utilizadas e que devem ser escolhidas de acordo com os objetivos do problema a ser resolvido. Apesar dessa gama de técnicas existentes, vale ressaltar que embora as técnicas atuais sejam capazes de processar grandes quantidades de dados e encontrar padrões válidos e novos, ainda não se tem uma solução eficaz em relação aos padrões valiosos.

Devido a isso, a etapa de *data mining* ainda é muito dependente da experiência dos analistas humanos realizando a mineração, os principais responsáveis pela determinação do valor dos padrões encontrados. Do mesmo modo, a exploração dos dados também é dirigida por analistas humanos, fator que não pode ser desprezado em nenhuma aplicação do processo KDD onde se deseje a obtenção de sucesso. No capítulo seguinte, será apresentado o projeto SIMCO, onde se encontram os principais objetivos desta dissertação, a aplicação do processo KDD no banco de dados do LAPQAP.

III. SISTEMA INTEGRADO PARA MONITORAMENTO DA QUALIDADE DE COMBUSTÍVEL - SIMCO

5 SIMCO

Neste capítulo, é descrito o ambiente SIMCO de monitoramento da qualidade de combustível (Labidi, 2003). O objetivo é fornecer uma visão clara do projeto, de modo que se possa delimitar e justificar a contribuição deste trabalho dentro do SIMCO.

5.1 Introdução

A abertura do mercado de exploração das reservas nacionais de óleo e gás foi decorrente da crescente pressão sobre a competitividade internacional das empresas produtoras, o que resultou na quebra do monopólio do setor petróleo. No Brasil, o marco referencial desta mudança foi a criação da ANP (URL 2) em 1998. A ANP foi criada, graças à Lei nº 9.478/97 que estabeleceu o CTPETRO (URL 3). A sustentação financeira do CTPETRO dá-se por meio dos recursos oriundos dos *royalties* do petróleo repassados pela ANP ao Ministério da Ciência e Tecnologia - MCT, conforme previsto na Lei no 9.478/97. Estes são transferidos ao Fundo Nacional de Desenvolvimento Científico e Tecnológico – FNDCT.

Antes, na década de 70, após as duas crises mundiais de petróleo, o Brasil decidiu investir em grandes programas alternativos de energia e combustíveis, a exemplo do programa pró-alcool em 1977 do programa Biodiesel.

Apesar dos avanços e da importância de outras fontes alternativas de energia, o petróleo continuará sendo ainda por cerca de 100 anos uma das maiores fontes de energia consumida no mundo (Schuchardt et al., 2001). No Brasil, a biomassa é reconhecida por muitos como o principal substituto do petróleo (Schuchardt et al., 2001).

O país tem reconhecido, através de vários acontecimentos mundiais recentes, a importância de se investir em tais tecnologias, tanto do ponto de vista de soberania como por ser considerada uma estratégia de desenvolvimento econômico, especialmente no futuro. Entretanto, a manutenção da qualidade de programas como o CTPETRO e outros investimentos no setor, é da máxima importância para a atual geração.

A ANP tem desempenhado um importante papel social nos últimos 5 anos, no que diz respeito à qualidade de combustíveis e combate à fraude e sonegação de impostos. Qualquer que seja o tipo de combustível: diesel, gasolina, álcool, etc. a sua qualidade e conformidade com as normas técnicas é um fator fundamental para seu uso com segurança. Uma das recomendações mais importantes, formulada pela ANP, tem sido o desenvolvimento de laboratórios de combustível que são estruturas que permitem a realização de ensaios de rotina referentes à qualidade dos combustíveis, regulamentados por normas vigentes (ASTM e NBR), estabelecidas em Portaria pela ANP cujo objetivo é a coleta e análise de amostras de combustíveis (Contrato UFMA-ANP. Contrato nº 4.067/2001 de 27/08/2001). Além do monitoramento de combustíveis, a ANP tem investido no apoio a estes laboratórios, a exemplo do Programa CAT-RN-LEC (Capacitação e Assistência Técnica a laboratórios da rede nacional de ensaios para o monitoramento da qualidade de combustíveis) (URL 4).

Atualmente, a nível nacional há dezoito (18) laboratórios de combustível, sendo um destes no Maranhão. Todos em fase de preparação para credenciamento. Este número é considerado irrisório pela própria ANP (URL 4). Como exemplo, o Programa de Monitoramento da Qualidade dos Combustíveis Automotivos comercializados no Estado do Maranhão, estabelecido através do Contrato

ANP/UFMA (Contrato nº 4.067/2001 de 27/08/2001), iniciou suas atividades em dezembro de 2001. Mensalmente são coletadas e analisadas amostras de 400 postos em todo o Estado. Os dados são armazenados e enviados a ANP através do programa MQC (Software de Monitoramento da Qualidade de Combustível). Porém, o MQC não permite nenhum tipo de análise estatística, apoio à tomada de decisão, ajuda na busca de informações, etc.

Para garantir qualidade e normalização, é indispensável o desenvolvimento de ferramentas eficientes que permitam o monitoramento do combustível de qualquer ponto do estado e para qualquer tipo de combustível. Considerando a variedade dos critérios, uma tomada de decisão para uma otimização e controle da produção, *scheduling* e roteamento de veículos, etc. devem ser baseadas na avaliação dos mais variados tipos de dados espaciais (ex.: características da rede de distribuição, etc.) e não espaciais (ex.: qualidade do combustível, tipo do combustível, etc.).

Neste contexto, o presente projeto propõe o início de estudos para criação de um Sistema Integrado (SIMCO) que alem de permitir um melhor monitoramento, praticidade e eficiência, possibilite o controle e otimização de problemas relacionados à industria do petróleo no Estado do Maranhão. Para isto serão utilizadas técnicas de Geoprocessamento e de IA baseadas em agentes inteligentes e uma abordagem de *Data Warehouse* e *Data Mining* para apoiar à tomada de decisão. O sistema poderá posteriormente ser estendido para abranger outras regiões do país.

Assim, convém afirmar que o ambiente SIMCO proposto pode ser localizado nas categorias dos sistemas computacionais de Controle, Normalização e Regulamentação Técnica e de Apoio à Tomada de Decisão em sua vertente que

utiliza métodos e técnicas provenientes da Inteligência Artificial Distribuída através de uma Abordagem de Sistemas Multiagentes, como uma alternativa para se modelar sistemas de informação geográfica aplicada ao petróleo.

5.2 Objetivos e Metas

O projeto tem como objetivo principal a realização e implantação de um Sistema de Informação Geográfica baseado em recursos multimídia e técnicas de Inteligência Artificial com base em múltiplos agentes dispostos na arquitetura aberta da *Web*.

Pretende-se com isto, integrar pesquisadores, agentes da ANP e o Sistema Computacional em um espaço que sirva para promover o monitoramento dos combustíveis e o apoio à otimização e tomada de decisão em problemas de petróleo no estado do Maranhão.

Portanto pretende-se, essencialmente, desenvolver-se ao longo de quatro grandes metas:

I. Aquisição, organização e publicação de dados espaciais e não espaciais relevantes para monitoramento da qualidade de combustível no Maranhão;

II. Análises estatísticas para teste de qualidade e consistência dos dados com aplicação de técnicas de *Data Warehouse* e *Data Mining*;

III. Desenvolvimento do SIG adaptado para apoiar a otimização e tomada de decisões nos problemas de petróleo (distribuição, trajetória e roteamento de veículos,otimização e controle da produção de poços, etc.);

IV. A nível avançado do Sistema Computacional, estaríamos preocupados com aspectos inerentes à concepção e desenvolvimento da sociedade de agentes, levando em conta tanto questões internas (modelo de agentes, formas de organização dos agentes, linguagens e protocolos de interação e comunicação entre agentes), quanto externas, no caso das interações com seu ambiente.

5.3 Metodologia

O desenvolvimento do projeto contará como ponto de partida, com resultados e experiências, relacionados ao tema em pauta, de trabalhos anteriores na área, oriundos de investimentos por parte integrantes do grupo de pesquisa. Várias atividades serão executadas no desenvolvimento deste projeto. Estas atividades incluem: (a) uma revisão bibliográfica conclusiva e um estudo nos temas de paradigma SIG, *Data Warehouse* e *Data Mining*; (b) proposta de modelagem e implantação do ambiente; (c) avaliação da implementação inicial e posterior manutenção.

Estudaremos formas mais eficientes de modelar processos de otimização e tomada de decisão cooperativa onde aplicaremos resultados das ciências cognitivas para aprimorar a concepção do sistema a partir de uma abordagem multiagentes. Nessa perspectiva, alguns desafios para desenvolvimento de nossa pesquisa são: Detecção e organização do *data warehouse*; definição das tecnologias de coleta e tratamento das informações espaciais e não espaciais (inicialmente iremos usar os dados do software MQC); e a definição das técnicas de mineração de dados. A vantagem destas tecnologias, é de prover múltiplas visões

da informação e um acesso à fonte de dados não previamente relacionadas (fontes independentes). O uso da tecnologia de agentes de software nos permite uma melhor ajuda e controle no processo de tomada de decisão.

Aplicação – Um dos beneficiados com o projeto inicialmente será o Maranhão que, através do único laboratório de combustível do estado, terá ferramentas avançadas de alto desempenho para melhor monitorar o combustível no estado. Teremos também outras aplicações interessantes na otimização e tomada de decisões.

Avaliação – As experimentações e avaliações iniciais dos resultados obtidos acontecerão, no âmbito da UFMA, com o apoio do laboratório de combustível desta instituição. Será, portanto observado o desempenho das nossas ferramentas e analisaremos o impacto de nossa abordagem sobre a pratica de monitoramento de combustível bem como na melhoria nos problemas de otimização como na distribuição, controle de produção, planejamento de rota, etc.

A equipe é interdisciplinar e contará com a colaboração dos pesquisadores da área de petróleo na UFMA, principalmente a Coordenadora Geral Profa. Dra. Aldaléa Lopes Brandes Marques e o Coordenador Técnico do Laboratório de Combustível (LAPQAP) o Prof. Dr. Henrique Tadeu Castro Cárdias.

5.4 Resultados e Impactos Esperados

Pretende-se, com a consolidação deste projeto, dispor de um ambiente sofisticado, implementando um sistema integrado de gestão de informação geográfica na área de monitoramento de combustível e apoio à tomada de decisão.

O SIMCO permite uma consulta aos dados do cadastro de resultados de análises de combustível representados na forma de mapas digitais disponibilizados

na *WEB* para quem tiver interesse e direito de atuar na área. Poderão ser executadas consultas que estão organizadas na forma de camadas (*layers*). Cada uma destas camadas possui uma representação gráfica dos dados.

De um ponto de vista acadêmico esperamos conseguir publicar artigos em revista e em congressos importantes no tema em questão. Ainda nessa linha, contribuiremos na formação de recursos humanos através, por exemplo, do desenvolvimento dos participantes em programas de pesquisa, tanto em nível de graduação, quanto de pós-graduação. Do ponto de vista da aplicação prática, teremos um modelo computacional pronto (testado em situações reais) permitindo também extensões integrando novas funcionalidades em maior abrangência de uso.

5.5 Colaboração

A colaboração do presente trabalho neste projeto pode ser enquadrada nas metas II e III acima citadas, referentes à aplicação das técnicas de data warehouse e data mining, juntamente com o desenvolvimento do sistema de informação geográfica.

Portanto, esta colaboração pode ser resumida em 3 itens:

- Aplicação do processo KDD nos dados do LAPQAP. Esta tarefa por sua vez pode ser dividida em dois principais itens:
    - A construção de um data warehouse, alimentado com o banco de dados da ANP, LAPQAP;
    - Aplicação da mineração de dados nos resultados obtidos do data warehouse, afim de descobrir novos padrões.

- Disponibilizar toda informação obtida através do Processo de Descoberta de Conhecimento em um Sistema de Informação Geográfica, exibido em uma página *web*.

5.6 Conclusão

SIMCO, é um projeto inovador, uma vez que esta é uma área relativamente nova e trabalhos semelhantes a este, processo KDD aplicado a bancos de dados de análise de combustível, estão em fase embrionária. Espera-se que este trabalho possa dar uma boa contribuição e um bom início ao o projeto, já que este é seu trabalho inicial.

A colaboração deste trabalho para o projeto SIMCO consiste no desenvolvimento de um *data warehouse*, na aplicação de técnicas de *data mining* nos resultados obtidos e o início de desenvolvimento de um SIG, onde serão disponibilizados os resultados obtidos na aplicação do processo KDD.

No capitulo seguinte, é visto o processo de análise de combustível. Contexto em que se encontra o desenvolvimento deste trabalho.

IV. APLICAÇÃO DO PROCESSO KDD EM BANCOS DE DADOS DE ANÁLISE DE COMBUSTÍVEIS

6 O PROCESSO DE ANÁLISE DE COMBUSTÍVEIS - LAPQAP

Neste capítulo são apresentados o processo de análise de combustível, o software responsável pela alimentação do banco de dados, que é o objeto principal do nosso trabalho, e o próprio banco de dados da ANP.

6.1 Introdução

Aqui são apresentados os procedimentos de coleta e análises adotados pelo Programa de Monitoramento de combustíveis do LAPQAP da UFMA, de acordo com o relatório entregue a ANP referente às atividades de monitoramento de combustíveis do mês de junho de 2003 (Marques et al., 2003a) utilizado como material de estudo nesta pesquisa.

Para fins de execução das atividades do Programa de Monitoramento de Combustíveis no Maranhão foi definido o universo de amostragem, composto pelos postos revendedores do estado cadastrados na ANP. A metodologia de coleta dos combustíveis nos postos revendedores, constitui-se basicamente de três etapas:

1. Seleção aleatória dos postos revendedores a serem visitados e das amostras de combustíveis a serem coletadas;
2. Coleta das amostras de combustíveis propriamente dita;
3. Colocação das etiquetas e armazenamento criterioso das amostras.

A seleção aleatória dos postos revendedores e amostras de combustíveis é feita uma vez por mês pelo Coordenador ou Supervisor do Laboratório de Combustível, a partir da relação de postos revendedores fornecida pela Superintendência de Qualidade de Produtos da ANP e atualizado mensalmente pelo próprio laboratório. O roteiro a ser seguido é discutido e otimizado com o

amostrador. Os procedimentos adotados para a seleção dos postos e de coleta de amostras é informado mensalmente à ANP (Marques et al., 2003a).

A recepção dos amostradores nos postos revendedores de combustíveis é boa, o que vem prejudicando algumas visitas é o fato de alguns postos revendedores não apresentarem a nota fiscal de aquisição dos combustíveis impedindo a obtenção de dados como: preços, data de compra e número da nota fiscal. As principais alegações para estes casos são:

- O gerente/proprietário do posto não se encontra e somente ele tem acesso à nota fiscal;
- Não se encontram com a nota fiscal no posto, pois a mesma está com o contador ou mesmo na matriz da empresa;
- As notas fiscais não se encontram no posto, simplesmente.

Os postos do Estado do Maranhão não comercializam gasolina *premium*, além de ser raro encontrar gasolina aditivada, principalmente no interior do Estado. Há então necessidade de substituir essas amostras por gasolina comum ou óleo diesel, conforme procedimento de coleta indicado pela ANP. Portanto, são coletadas amostras de óleo diesel para substituir tanto a gasolina *premium* quanto à gasolina aditivada, uma vez que a maioria dos postos possui mais tanques de óleo diesel do que de gasolina comum. Exemplificando esta substituição, podemos citar os dados apresentados no relatório do mês de junho entregue a ANP (Marques et al., 2003a). Das 28 amostras de óleo diesel coletadas a mais, 20 amostras substituem as 20 amostras de gasolina aditivada coletadas a menos; as 8 amostras de gasolina *premium* não encontradas foram substituídas pelas 8 amostras de óleo diesel coletadas a mais. Foram coletadas 4 amostras de gasolina C comum a mais, referente a postos que apresentam mais de dois tanques de gasolina C Comum; conforme pode ser verificado na Tabela 4.

Tabela 4. Distribuição de amostras coletadas no mês de junho de 2003 e diferença entre amostras coletadas e previsão de coleta.

Produto	Previsão de Coleta (A)	Amostras coletadas (B)	Diferença (B – A)
Gasolina Comum	80	84	4
Gasolina Aditivada	32	12	-20
Gasolina *Premium*	8	0	-8
Óleo Diesel B Comum	40	67	27
Óleo Diesel B Aditivado		1	1
Álcool Hidratado Combustível	20	20	0
Total	**180**	**184**	**4**

Nas Tabelas 4 e 5 é apresentado o resumo quantitativo das regiões e dos municípios do Maranhão monitorados no mês de junho de 2003.

Tabela 5. Relação das Regiões do Estado, monitoradas no mês de junho de 2003.

Região	Número de Postos	Postos Fechados	Postos em Reforma	Postos em Atividade	Postos Monitorado	% de postos*
1	98	8	0	90	20	22,2
2	124	7	3	114	20	17,5
3	119	8	0	111	20	18,0
4	119	3	0	116	20	17,2
Total	460	26	3	431	80	18,6

(*) Percentagem de postos monitorados em relação ao número de postos em atividade

O número de postos de cada região é atualizado a cada mês. O amostrador é orientado a coletar amostras de postos que não constem na relação de coleta, permitindo assim ser verificado, junto ao banco de dados da ANP/MQC (software de coleta de dados do PMQC, detalhado no decorrer deste trabalho), a situação dos referidos postos e, desta forma, feitas as atualizações.

A defasagem entre os dados de cadastramento dos postos revendedores fornecidos pela ANP e os dados reais tem causado problema na amostragem, pois alguns postos constantes na relação fornecida pela ANP como em funcionamento, estão fechados ou não existem, outros tiveram os dados alterados (como razão

social, CNPJ e endereço). Para evitar que o amostrador retorne sem coletar o número de amostras previsto, o mesmo leva consigo, a relação de todos os postos da região (Tabela 6), além dos previamente selecionados a mais que constam na planilha para substituir os selecionados para coleta. O amostrador ainda está orientado a coletar amostras em todos os postos que não conste em nenhuma das relações; o que permitirá a obtenção, num futuro próximo, de dados atualizados sobre todos os postos do estado.

Tabela 6. Relação dos municípios monitorados no Maranhão no mês de junho de 2003.

N° da Região	Município (nome)	N° de Postos	N° de Postos Monitorados	% Postos monitorados
1	São Luís	90	16	17,78
1	Raposa	02	01	50,00
1	São José de Ribamar	10	03	30,00
2	Cantanhede	02	01	50,00
2	Coroatá	07	03	42,86
2	Peritoro	03	03	100,00
2	Esperantinópolis	01	01	100,00
2	Igarapé grande	01	01	100,00
2	Bacabal	09	06	66,67
2	São Luís Gonzaga do Maranhão	02	02	100,00
2	Lago dos Rodrigues	01	01	100,00
2	Porção de Pedras	01	01	100,00
2	Timbiras	01	01	100,00
3	Balsas	07	06	85,71
3	Carolina	04	03	75,00
3	Estreito	03	03	100,00
3	Riachão	03	02	66,67
3	São Raimundo das Mangabeiras	03	02	66,67
3	Fortaleza dos Nogueiras	03	02	66,67
3	São João do Paraíso	02	02	100,00
4	Arari	02	01	50,00
4	Matinha	02	01	50,00
4	Miranda do Norte	04	04	100,00
4	Santa Rita	02	02	100,00
4	São Bento	03	01	33,33
4	Viana	03	04	133,33
4	Vitória do Mearim	03	02	66,67
4	São Vicente Ferrer	02	02	100,00
4	Penalva	01	01	100,00
4	Cajapió	01	01	100,00
4	Olinda Nova do Maranhão	01	01	100,00

6.2 Análises Físico-Químicas dos Combustíveis

As amostras de combustíveis coletadas estão sendo analisadas através dos ensaios previstos no Convênio ANP/UFMA. Todas as amostras de combustíveis coletadas são analisadas através dos ensaios regulares listados na Tabela 7.

Tabela 7. Ensaios regulares realizados e métodos utilizados.

Ensaio	Método	Gasolina	Ó. Diesel	AEHC
Aparência – Aspecto e/ou Cor (Método Visual)	Visual	X	X	X
Cor ASTM	ASTM D1500		X	
Composição quanto ao tipo de hidrocarbonetos	APG*	X		
Teor de AEAC	NBR 13992	X		
Teor de AEAC	APG*	X		
Massa específica 20°C	ASTM D4052	X	X	
Massa Específica e Teor Alcoólico	NBR 5992			X
Destilação	ASTM D86	X	X	
MON*	APG*	X		
RON*	APG*	X		
Condutividade Elétrica	NBR 10547			X
Teor de Gasolina	NBR 13993			X
PH	NBR 10891			X
Enxofre	ASTM D4294		X	
Benzeno	APG*	X		
Índice de Cetano	ASTM D4737		X	
Ponto de Fulgor	ASTM D 56		X	

* APG - Pelo Analisador Portátil de Gasolina (Grabner – Modelo IROX), segundo metodologia sugerida pelo fabricante.

Os ensaios apresentados na Tabela 7 são realizados de acordo com as Normas Brasileiras (NBR), Métodos Brasileiros da Associação Brasileira de Normas Técnicas (MB-ABNT), Normas da *American Society for Testing and Materials* (ASTM).

Para os ensaios realizados com o analisador portátil de gasolina, dada a ausência de metodologia específica, seguem-se as instruções fornecidas pelo fabricante do equipamento.

Em cada relatório mensal apresentado pelo LAPQAP a ANP são apresentados e discutidos os resultados dos ensaios obtidos em amostras coletadas nos postos revendedores do Estado do Maranhão no mês em questão. Para as amostras coletadas no mês faz-se uma avaliação estatística, onde é calculada a média, o desvio padrão e feito o histograma para cada ensaio. Além dos dados sobre as amostras, é efetuado um acompanhamento estatístico do índice de não conformidade obtido durante os últimos meses de monitoramento.

6.3 Resultados de Ensaios Regulares

Após o encerramento das análises de cada região, os resultados dos ensaios são conferidos, digitados e enviados para a ANP através do programa MQC. Este programa gera as planilhas dos resultados dos ensaios em planilhas do excel, de todos os postos com registro regular ou provisório junto a ANP. Na Tabela 8 é apresentado um resumo das não conformidades por região.

Tabela 8. Resumo das amostras analisadas em junho de 2003, por combustível e por região.

Região	Gasolina C Comum				Gasolina C Aditivada				AEHC				Óleo Diesel				TOTAL			
	Nº de amostras				Nº de amostras				Nº de amostras				Nº de amostras				Nº de amostras			
	T	NC	I	100-I	T	NC	I	100-I	T	NC	I	100-I	T	NC	I	100-I	T	NC	I	100-I
C	84	66	78,6	21,4	12	6	50,0	50,0	20	10	50,0	50,0	68	2	2,9	97,1	184	84	45,7	54,3
1	24	15	62,5	37,5	7	1	14,3	85,7	5	1	20,0	80	13	0	0,0	100,0	49	17	34,7	65,3
2	20	20	100,0	0,0	1	1	100,0	0,0	5	3	60,0	40	19	0	0,0	100,0	45	24	53,3	46,7
3	20	17	85,0	15,0	3	3	100,0	100,0	5	2	40,0	60	17	0	0,0	100,0	45	22	48,9	51,1
4	20	14	70,0	30,0	1	1	100,0	0,0	5	4	80,0	20	19	2	10,5	89,5	45	21	46,7	53,3

T= Total, NC= Não conforme, I=Índice de NÃO CONFORMIDADE, C= completa , ou seja, o universo do contrato 100-I= índice de conformidade

6.4 Tipos de Não Conformidades

Em todo relatório mensal são apresentados os tipos de não conformidades encontradas nas amostras de combustível monitorados durante o mês. Na Tabela 9 é apresentado um resumo das não conformidades por produto.

Tabela 9. Número de amostras por tipo de não-conformidade (junho 2003).

Gasolina Comum	***Número de não conformidades***	**% NC**
Teor de AEAC	30	35,3
Destilação (PFE)	2	2,4
Resíduo	1	1,2
MON	50	58,8
CNPJ inválido	2	2,4
Total	85	100,0
Gasolina Aditivada	***Número de não conformidades***	
MON	4	40,0
Teor de AEAC	6	60,0
Total	10	100,0
Álcool	***Número de não conformidades***	
Massa Específica	7	38,9
Teor Alcoólico	7	38,9
PH	4	22,2
Total	18	100,0
Diesel Tipo B	***Número de não conformidades***	
CNPJ inválido	2	100,0
Total	2	100,0

Outro item apresentado no relatório é o histórico de não conformidades, onde são apresentados e avaliados os dados das amostras coletadas nos últimos 7 meses: total por região e por produto.

6.5 Histograma dos ensaios regulares dos combustíveis

Um histograma de cada ensaio realizado no mês é gerado, ou seja, o gráfico que relaciona valores de resultados com seu número de ocorrências. Os

dados dos ensaios são tratados pelo programa ORIGIN 5.0. O pacote produz um histograma dos dados de cada ensaio com uma curva normal para cada conjunto de dados (ensaio). O seguinte procedimento foi realizado:

- ✓ Fixou-se os limites mínimo e máximo dos valores. Para os dados do LAPQAP foi fixado o menor e o maior valor obtido.
- ✓ Fixou-se o valor do incremento dos valores (Bin size) para agrupamento dos valores (Tabela 10). Ex.: fixando-se os limites entre 50 e 90 com um incremento de 5, significa que o gráfico gerado (histograma) mostrará o número de ocorrências dos resultados obtidos, agrupados em intervalos de 5 unidades.
- ✓ Obteve-se a gaussiana (*gaussian fitting*) cujo ajuste matemático fornece o valor médio para o conjunto de valores e desvio padrão (α) correspondente.

Estatisticamente, valores contidos em um intervalo de $\pm 3\alpha$ (chamado intervalo de confiança), têm 99,7% de probabilidade de pertencerem ao conjunto de valores que refletem o comportamento de um determinado parâmetro.

Para um determinado parâmetro (densidade, MON, IAD, etc.) os valores contidos no intervalo de confiança irão refletir o perfil das amostras analisadas. Amostras com valores fora deste intervalo de confiança terão um perfil diferenciado do restante. Apesar de estarem conformes, ou seja, dentro dos limites estabelecidos nas portarias da ANP, elas têm um perfil diferenciado, sendo consideradas atípicas.

Quando este comportamento diferenciado se dá em vários parâmetros, isto pode ser um reflexo de diferentes razões como diferentes origens das amostras, adulterações, dentre outras.

O estudo mais minucioso das amostras típicas (com valores dentro do intervalo de confiança) pode trazer informações importantes para a determinação dos perfis de diferentes regiões e até mesmo de sua origem, entre outras.

O estudo das amostras atípicas pode auxiliar na identificação de amostras que tiveram uma adulteração mais elaborada, fazendo com que sua composição diferenciada resulte em uma amostra dentro dos limites das especificações.

Tabela 10. Valores dos Incrementos para a obtenção dos Histogramas.

ENSAIOS PARA GASOLINAS	*Bin Size*
Densidade	0,0005
Temperaturas Equivalentes a 10, 50 e 90%	0,5
Ponto final de Evaporação	0,5
Resíduo	0,15
Teor Alcoólico	1,0
MON, RON e IAD	0,5
% Oxigenados, Aromáticos, Olefinas e Saturados	0,5
% Benzeno	0,05
ENSAIOS PARA ÓLEO DIESEL	
Massa Específica	0,5
Temperaturas Equivalentes a 50 e 85 %	0,5
Ponto Final de Evaporação	0,5
Cor	0,5
Enxofre	0,02
Índice de Cetano	1,0
Ponto de Fulgor	2,0
ENSAIOS PARA AEHC	
Massa Específica	0,5
PH	0,1
Condutividade	5
Teor Alcoólico	0,1

A seguir, um exemplo de histograma dos ensaios regulares de combustível realizados encontrado em Marques et al. (2003a) (Figura 13):

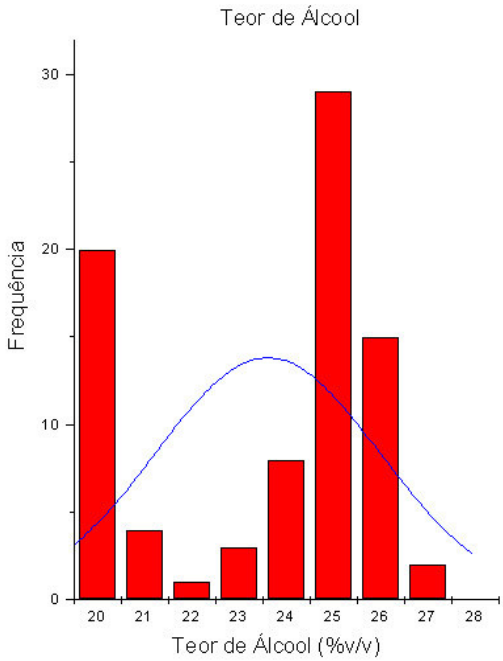

Figura 13. Histograma do teor de AEAC na GCC, coletada em postos revendedores do MA em junho de 2003. Gaussiana (gaussian fitting) com média = 23,610 e desvio padrão = 2,377.

6.6 O Software MQC

Aqui é apresentado o Software para Coleta de Dados do Programa de Monitoramento da Qualidade de Combustíveis da Agência Nacional do Petróleo, denominado de Software MQC. Este Software foi desenvolvido pela Universidade de Salvador para a Superintendência de Qualidade de Produtos da Agência Nacional do Petróleo.

O software visa suportar a entrada dos dados coletados pelos conveniados do PMQC da ANP. Denominado de Software MQC, ele foi desenvolvido dentro do convênio celebrado entre UNIFACS (Universidade Salvador) e a ANP para o desenvolvimento de novas tecnologias para exploração inteligente de dados no PMQC. O software MQC é de uso exclusivo da ANP e seus conveniados no PMQC.

O objetivo principal do MQC é permitir a ANP processar de forma eficiente os dados coletados pelos conveniados do PQMC. O software visa:

- Padronizar os dados enviados pelos conveniados do PMQC a ANP;
- Consolidar todos os dados coletados em uma única base de dados na ANP;
- Prover através deste banco de dados uma plataforma única para extração e análise de dados do PMQC;
- Minimizar problemas de digitação dos dados nos conveniados;
- Manter bases de dados locais padronizadas em todos os conveniados.
- Além dos itens listados acima, o projeto do software é norteado pelos seguintes condicionantes:
    - Facilitar ao máximo o processo de entrada de dados nos conveniados;
    - Rodar em uma plataforma presente em todos os conveniados do programa;
    - Ser de fácil instalação e uso nos conveniados.

Para cumprir parte dos condicionantes acima o MQC roda em plataformas MS Windows e usa arquivos comprimidos para troca de dados entre ANP e conveniados.

Os arquivos de instalação do software MQC podem ser obtidos ou em um CD-ROM fornecido pela UNIFACS ou através do Portal de Suporte ao Software de Coleta de Dados. O software roda em todas as versões do sistema operacional Windows.

Todos os postos monitorados devem estar devidamente cadastrados no sistema. Cada um deles é identificado unicamente por um número de registro que é fornecido pela ANP. Caso um posto ainda não esteja registrado na ANP, este posto é cadastrado no sistema com um número de registro temporário, até que o registro seja oficializado na ANP e ela informe seu número de registro definitivo. Para evitar

duplicidades e facilitar o reconhecimento destes postos, seus números de registro temporários serão sempre negativos. A seguir, as Figuras de 14 a 16 mostram algumas telas do software:

Figura 14. Tela de Entrada do Sistema.

Figura 15. Tela de Cadastro dos Dados Físicos dos Postos.

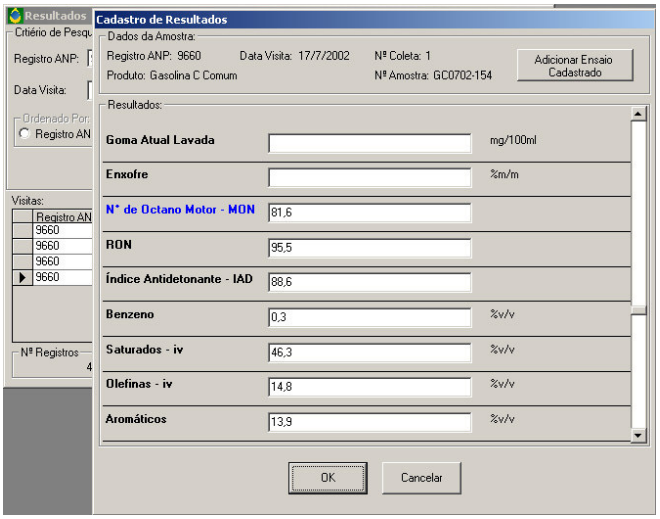

Figura 16. Tela de Cadastro de Resultados.

6.7 O Banco de Dados da ANP

O principal objeto de estudo desta pesquisa, o banco de dados da ANP, é alimentado pelo Software de Coleta de Dados – MQC, visto anteriormente no Item 6.6. É um banco de dados relacional que se encontra no Microsoft Access e é constituído de 25 tabelas (Relação abaixo e Figura 17):

1. TANP_MQC_UNIVERSIDADE
2. TANP_MQC_VISITA
3. TANP_MQC_DISTR
4. TANP_MQC_POSTO
5. TANP_MQC_SUBAREA
6. TANP_MQC_AREA
7. TANP_MQC_ESTADO
8. TANP_MQC_ALTERACAO_POSTO
9. TANP_LOCALIDADE
10. TANP_LOGRA
11. TANP_NOME_LOGRA
12. TANP_TITULO
13. TANP_BAIRRO
14. TANP_MQC_PRODUTO_VISITA
15. TANP_MQC_PROD_POSTO_VISITA
16. TANP_SIT_OPER_INST
17. TANP_PRODUTO
18. TANP_MQC_PROP_PROD
19. TANP_MQC_AMOSTRA
20. TANP_MQC_OPCAO
21. TANP_MQC_PROPRIEDADE
22. TANP_MQC_VISITA_PROP_POSTO
23. TANP_MQC_PROPRIEDADE_POSTO
24. TANP_MQC_OPCAO_PROP_POSTO
25. TANP_MQC_PROP_POSTO_GRUPO

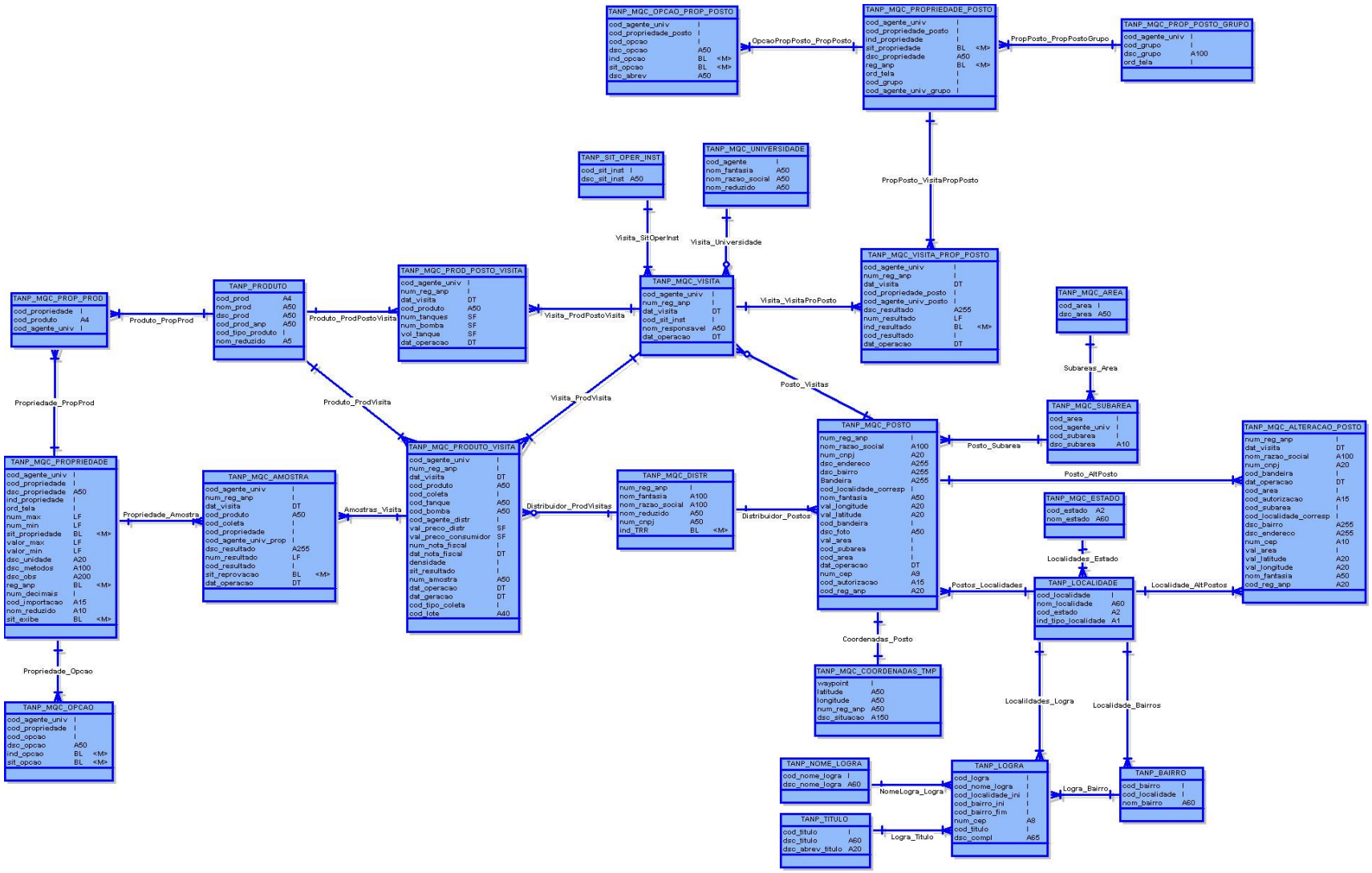

Figura 17. Modelo relacional do banco de dados da ANP.

7 APLICAÇÃO

Neste capítulo é vista a colaboração deste trabalho no projeto SIMCO. Também é apresentado o data warehouse criado para o banco de dados da ANP, juntamente com a aplicação das técnicas de mineração de dados aplicadas e o SIG.

7.1 *Data Warehouse*

Como já apresentado nos objetivos, este trabalho visa o desenvolvimento de um *data warehouse*. No *data warehouse* em questão, foi utilizado o modelo estrela (Figura 18), por sua simplicidade e por possuir um melhor acesso às informações.

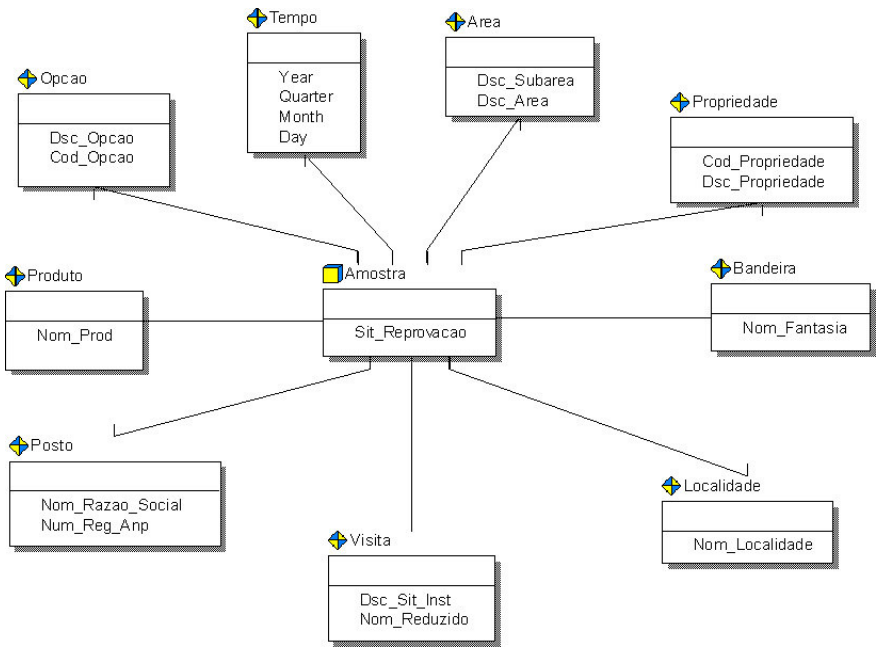

Figura 18. Modelo estrela do *data warehouse* proposto.

O modelo estrela é composto de:

- Tabela de fatos Amostra: onde contam os registros das amostras coletadas no estado do maranhão e como medida, a situação de cada amostra, se ela foi aprovada ou não;
- Dimensão Área: dados referentes à localização dos postos no que diz respeito as áreas e subáreas;
- Dimensão Propriedade: dados das propriedades analisadas para cada tipo de combustível;
- Dimensão Bandeira: dados dos distribuidores de combustíveis;
- Dimensão Localidade: nomes dos municípios onde se encontram os postos;
- Dimensão Visita: nomes dos responsáveis pelas visitas e a situação em que se encontravam os postos quando ela foi realizada;
- Dimensão Posto: razão social do posto e seu registro junto a ANP;
- Dimensão Produto: tipo de combustível sendo analisado;
- Dimensão Opção: tipos de resultados que as propriedades dos combustíveis podem ter;
- Dimensão Tempo: dados referentes a quando ocorreram as visitas.

A ferramenta usada no desenvolvimento do *Data Warehouse* foi o *Analysis Manager* que possui o *Cube Editor* para edição do *Data Warehous*e. As Figuras 19 e 20 apresentam as telas correspondentes à ferramenta usada.

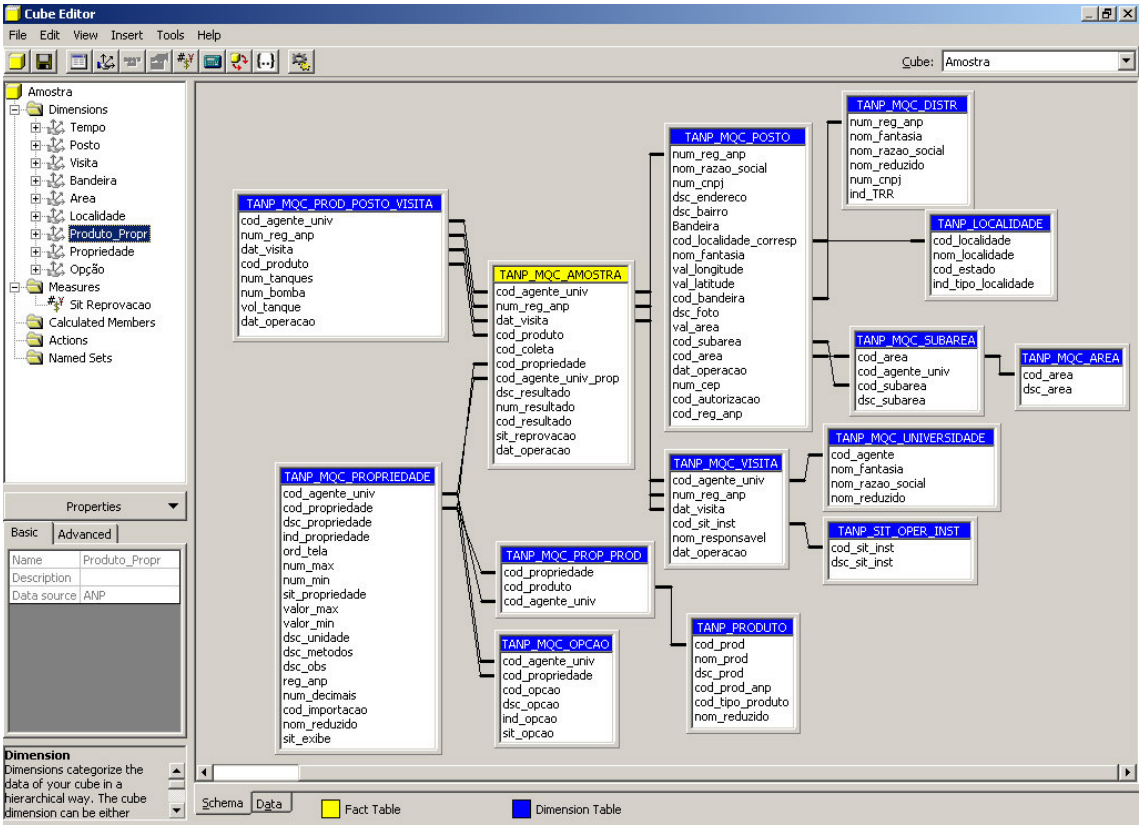

Figura 19.Cube editor – Tela 1.Tabelas que contem os campos selecionados para fazerem parte das dimensões do *data warehouse*.

Foram feitas seleções e limpezas no Banco de Dados da ANP e utilizadas somente variáveis quantitativas. Também foram descartadas as variáveis que não possuíam obrigatoriedade de acordo com as normas da ANP, ou pela falta de dados de alguns postos. Assim, essas variáveis ficaram de fora da matriz de dados utilizada para análise através das técnicas de mineração de dados selecionadas.

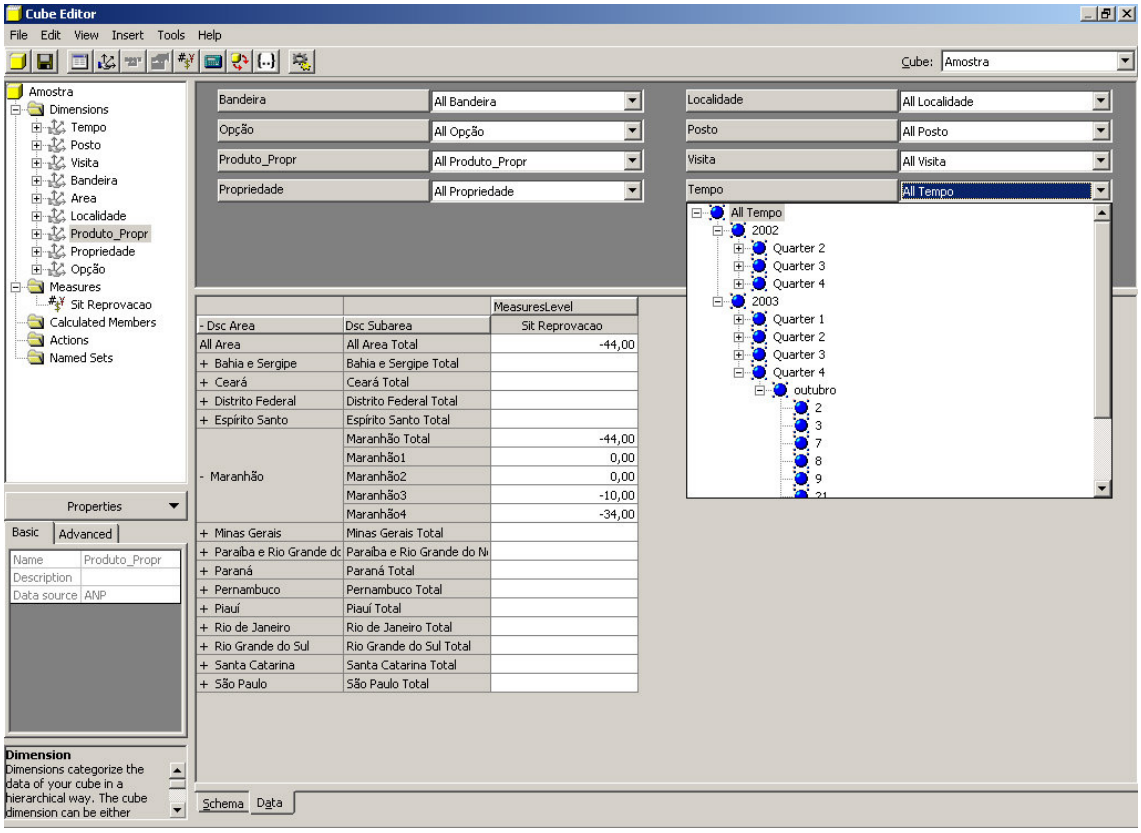

Figura 20. Cube Editor – tela 2. Navegação pelos dados do *data warehouse*.

7.2 *Data Mining*

Neste item são apresentados os resultados obtidos através da aplicação das técnicas de *data mining* selecionadas no *data warehouse* acima mencionado. Cada subitem é composto do tipo de combustível analisado e os resultados das técnicas aplicadas.

7.2.1 AEH - Álcool Etílico Hidratado Comum

O grupo AEH, contém 8 variáveis conforme mostra a Tabela 11, sendo utilizadas apenas 4 variáveis, pelos critérios de limpeza e seleção citados na página 119.

Tabela 11.Variáveis do AEH.

Código	Propriedade	Unidade	Observação
* 6	Aparência	Sem unidade	Não Quantitativa
7	Condutividade Elétrica	$\mu S / m$	Utilizada
8	Massa Específica	Kg/m^3	Utilizada
9	Potencial Hidrogênio (pH)	pH	Utilizada
10	Teor Alcoólico	°INPM	Utilizada
* 11	Teor de Gasolina	%v/v	Não Requerida pela ANP
* 175	Cor	Sem unidade	Não Quantitativa
* 197	Teor de Hidrocarboneto	%v/v	Falta de Dados

* Propriedades não utilizadas nas analises de dados

A Tabela 12 apresenta os dados estatísticos do grupo AEH a partir de uma matriz de dados gerada pela análise de 154 amostras e 4 variáveis.

Tabela 12.Dados estatísticos do AEH.

Variável	N	Média	Valor Médio	Desv. Padrão	Variação	Min.	Max.	Min.(ANP)	Max.(ANP)
7	154	197,65	152	170,00	28899,9	39	1215	0	500
8	154	810,99	810,85	10,82	117,26	691,4	847	805	811
9	154	7,31	7,4	0,54	0,29	5,4	9,4	6	8
10	154	91,43	92,25	2,47	6,11	74,4	96	94,7	92,6

Amostras – Número de amostras analisadas.

Média – Média padrão das propriedades

Valor Médio – Valor mais encontrado nas amostras de cada propriedade

Desvio Padrão – Distancia comum entre as amostras

Variação – Variação das propriedades

Mínimo- Valor Mínimo encontrado

Máximo- Valor Máximo encontrado.

Na Figura 21 mostra-se a variabilidade dos resultados em função das variáveis analisadas do AEH. Nota-se uma grande variação na propriedade 7 sendo muito difícil uma estimativa.

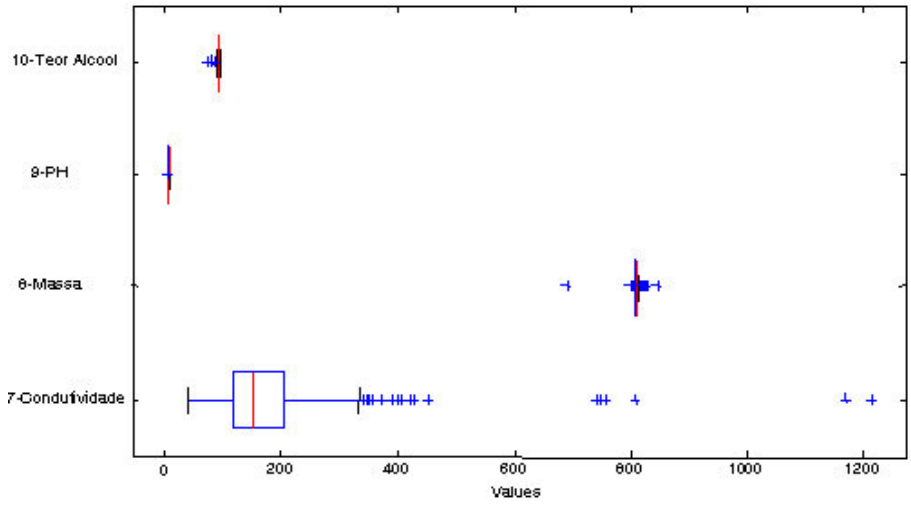

Figura 21. Variabilidade dos resultados em função das variáveis analisadas do AEH.

A Tabela 13 apresenta a correlação entre as variáveis do AEH.

Tabela 13. Correlação entre as variáveis do AEH.

	7	8	9
8	0,287		
9	0,105	0,138	
10	-0,318	-0,017	-0,029

7.2.1.1 Análise de Componentes Principais

Para a Análise de Componentes Principais do AEH utilizou-se o software MINITAB, em um conjunto de 154 amostras e 4 variáveis. Primeiro, foram determinados os autovalores (*Eigenvalues*) que são utilizados para determinar o número de componentes principais, cujo modo de determinação está baseado no

tamanho dos autovalores. Este número pode ser determinado se os autovalores forem maior que 1. No caso do AEH os autovalores são apresentados na Tabela 14.

Tabela 14. Autovalores (AEH).

	PC1	PC2	PC3	PC4
Eigenvalue	**1,5804**	**1,0280**	**0,8153**	**0,5763**
Proporção	0,395	0,257	0,204	0,144
Cumulativo	0,395	0,652	0,856	1,000

Pode-se observar que das 4 Componentes Principais possíveis apenas 2 Componentes têm autovalores maior que 1. Então, o conjunto de variáveis do AEH pode ser representado com apenas 2 componentes principais. Este comportamento é apresentado na Figura 22.

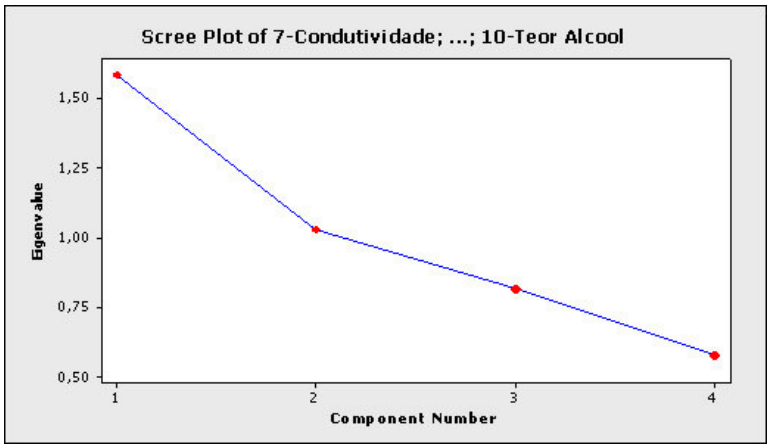

Figura 22. Número de componentes principais do AEH.

Coeficientes

Os Coeficientes indicam o peso da relação entre os componentes principais e as variáveis do AEH. Abaixo se encontra a tabela 15 gerada pelo MINITAB (Tabela 15).

Tabela 15. Coeficientes (AEH).

Variáveis	**PC1**	**PC2**	**PC3**	**PC4**
7-Condutividade	-0,629	0,200	0,188	-0,727
8-MassaEsp	-0,487	-0,482	0,582	0,438
9-PH	-0,425	-0,470	-0,773	0,039
10-Teor Álcool	0,433	-0,712	0,169	-0,527

Analisando as duas componentes principais, vemos que na 1ª temos valores negativos altos, sendo que essa componente melhor representa as variáveis 7 – Condutividade, 8 – Massa e 9 – pH, enquanto que na 2ª estes valores correspondem as variáveis 8 – Massa, 9 – pH e 10 – Teor Alcoólico. A figura 23 mostra este comportamento.

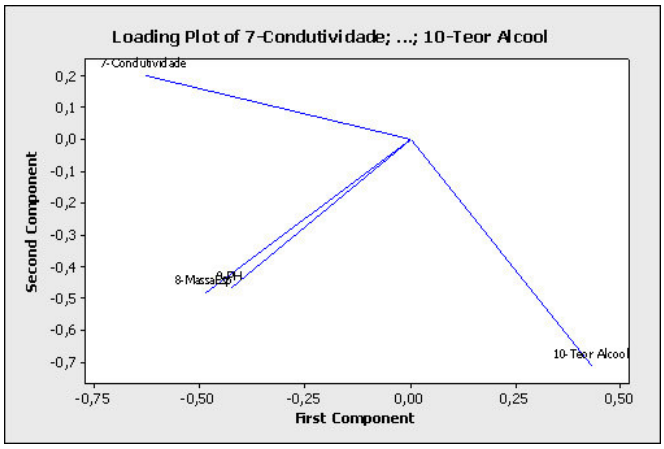

Figura 23. Análise das duas componentes principais (AEH).

A figura 24 mostra o gráfico das 2 componentes principais em relação as amostras do AEH, já que essas componentes representam a maior variância dos dados. Com esse gráfico é possível detectar outliers, clusters, e tendência das amostras; no caso das amostras do AEH, é possível verificar outliers (pontos afastados) e uma tendência uniforme de dados próximos do zero.

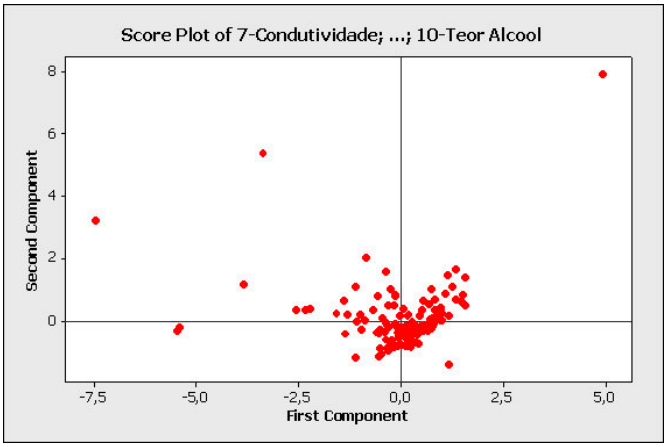

Figura 24. Gráfico das 2 Componentes X Amostras (AEH).

7.2.1.2 Análise de Agrupamento

A Análise de Agrupamentos é utilizada com o objetivo de classificar observações similares em grupos, quando, inicialmente, não há grupos conhecidos. Para este tipo de análise, utilizou-se o software MINITAB e o Método de Agrupamento Hierárquico.

Análise das Variáveis

Foram analisadas apenas 4 variáveis para o AEH. Devido ao pequeno número de variáveis foram divididos em apenas 2 grupos distintos. Na Figura 25 temos o dendrograma onde se podem visualizar os grupos formados.

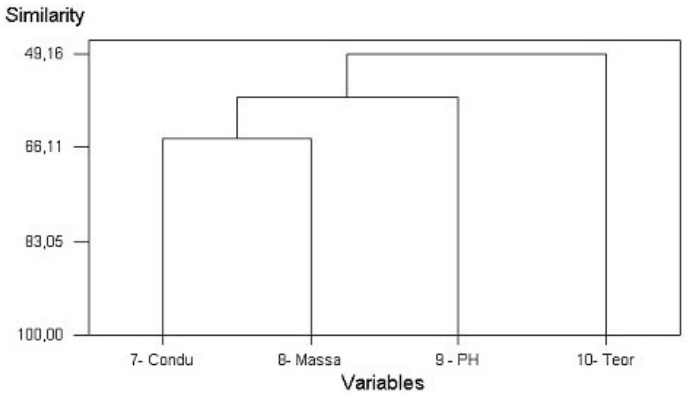

Figura 25. Dendrograma das variáveis do AEH.

O grupo 1 compreende as variáveis 7 – Condutividade, 8 – Massa e 9 – pH, enquanto o grupo 2 compreende apenas a variável 10 – Teor Alcoólico. Com isso, conclui-se que essa variável (10) tem pouca (ou nenhuma) similaridade com as demais propriedades do AEH.

Análise das Amostras

Foram analisadas 154 amostras do AEH, devido a grande quantidade de amostras, ficou um pouco difícil agrupá-las, no plano das Componentes Principais

pode-se verificar algumas variáveis um pouco distantes umas das outras, entretanto procurou-se formar 6 grupos de similaridade, mas como foram consideradas as amostras (outliers) fora do padrão pode-se aproveitar apenas 3 grupos. Isto pode ser observado na Figura 26.

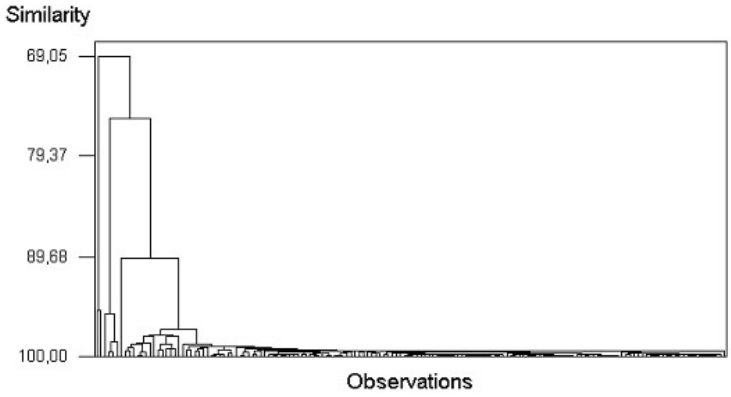

Figura 26. Dendrograma das amostras de AEH.

7.2.1.3 Regressão Múltipla

Para análise de Regressão do AEH (Tabela 16), definimos a variável 7 – Condutividade Elétrica como a variável de Resposta e as demais como variáveis explicativas.

A equação de regressão obtida da matriz de dados do AEH é a seguinte:

Condutividade = -1443 + 4,28 (Massa) + 15,0 (PH) - 21,2 (Teor Alcoólico)

Na Tabela 16 observa-se que o valor de P para as variáveis: 8 – Massa e 10 – Teor Álcool são iguais a 0,000, portanto menor que α-level = 0.05. Isso indica que essas variáveis têm significativo valor para a Resposta 7 – Condutividade. Já a variável 9 – pH tem um valor P muito alto, sendo assim, ela não é significativa no modelo de regressão e, portanto poderia ser descartada da análise. As variáveis

explicativas em questão explicam de 18,4% a 16,7% a variável resposta (7 – Condutividade).

Tabela 16. Regressão das variáveis (AEH).

	Coef	SE Coef	T	P
Constante	-1443	1050	-1,38	0,171
8 – Massa Específica	4,279	1,165	3,67	0,000
9 – PH	15,03	19,16	0,78	0,434
10 – Teor Alcoólico	-21,226	5,029	-4,22	0,000
S = 155,016 R-Sq = 18,4% R-Sq(adj) = 16,7%				

A Figura 27 mostra os gráficos da regressão do AEH.

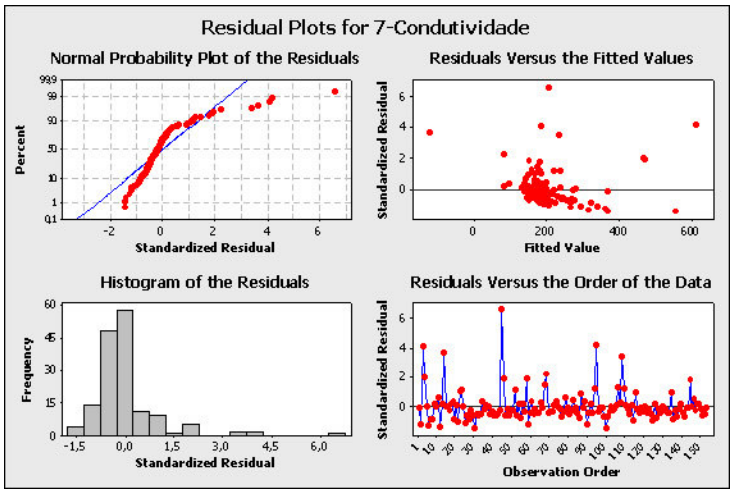

Figura 27.Gráficos da Regressão (AEH).

Os gráficos da regressão do AEH (Figura 27) são discutidos a seguir:

- **Histograma dos resíduos**: O Histograma do AEH indica alguns valores com simetria, e também algumas barras afastadas indicando a presença de Outilers (Amostras fora do padrão).
- **Gráfico da Probabilidade Normal dos Resíduos:** Este gráfico traça os resíduos contra seus valores previstos quando a distribuição é normal. Os resíduos da análise devem normalmente ser distribuídos. Na prática, para dados com um grande número de observações, os desvios moderados da

normalidade não afetam seriamente os resultados. No Gráfico do AEH observam-se algumas amostras outilers e alguns desvios.

- **Resíduos Vs. Ajustes:** Este gráfico analisa a distribuição dos resíduos. Baseado neste gráfico, os resíduos não parecem ser distantes da linha de ajuste, do zero. Há evidência de alguns outliers.
- **Resíduos Vs. Dados:** Este gráfico traça o conjunto de dados em função dos resíduos. No caso das amostras do álcool etílico, foi possível a identificação de alguns outliers.

7.2.2 GCC – Gasolina Comum

O grupo GCC contém 28 variáveis (Tabela 17), sendo que apenas 13 foram utilizadas para as análises.

Tabela 17. Variáveis da GCC.

Código	Propriedade	Unidade	Observação
*79	Cor	Sem unidade	Não Quantitativa
*80	Aspecto	Sem unidade	Não Quantitativa
81	AEAC	%v/v	Utilizada
*82	Densidade Relativa a 20ºC/4ºC	Sem unidade	Dados Irrelevantes
*83	Destilação – PI	°C	Dados Irrelevantes
*84	Destilação – 5%	°C	Falta de Dados
85	Destilação – 10%	°C	Utilizada
*86	Destilação – 20%	°C	Dados Irrelevantes
*87	Destilação – 30%	°C	Dados Irrelevantes
*88	Destilação – 40%	°C	Dados Irrelevantes
89	Destilação – 50%	°C	Utilizada
*90	Destilação – 60%	°C	Falta de Dados
*91	Destilação – 70%	°C	Falta de Dados
*92	Destilação – 80%	°C	Falta de Dados
93	Destilação – 90%	°C	Utilizada
*94	Destilação – 95%	°C	Falta de Dados
95	Destilação – PFE	°C	Utilizada
96	Resíduo	%v/v	Utilizada
97	Nº de Octano Motor – MON	Sem unidade	Utilizada
98	Índice Antidetonante – IAD	Sem unidade	Utilizada
*99	Pressão de Vapor a 37,8 ºC	kPa	Dados Irrelevantes
*100	Goma Atual Lavada	mg/100ml	Falta de Dados
*101	Enxofre	%m/m	Falta de Dados
102	Benzeno	%v/v	Utilizada
*103	AEAC – iv	%v/v	Falta de Dados
104	Saturados – iv	%v/v	Utilizada
105	Olefinas – iv	%v/v	Utilizada

106	Aromáticos	%v/v	Utilizada
107	RON	Sem unidade	Utilizada

* Propriedades não utilizadas nas analises de dados

A Tabela 18 apresenta os dados estatísticos do grupo GCC a partir de uma matriz de dados gerada pela análise de 413 amostras e 13 variáveis.

Tabela 18. Dados Estatísticos GCC.

Propriedades	**81**	**85**	**89**	**93**	**95**	**96**	**97**	**98**	**102**	**104**	**105**	**106**	**107**
Amostras	413	413	413	413	413	413	413	413	413	413	413	413	413
Média	24,38	54,18	72,45	168,92	209,53	1,32	82,31	88,97	0,03	47,44	13,98	13,90	95,63
Valor Médio	25	54	73	169	208	1	82	89	0	47	14	14	96
Desvio Padrão	2,01	1,06	0,84	6,01	10,20	0,49	1,04	1,18	0,18	4,16	3,54	1,37	1,66
Variação	4,06	1,13	0,71	36,19	104,14	0,24	1,10	1,39	0,03	17,36	12,56	1,89	2,76
Mínimo	17	50	69	152	188	1	81	87	0	19	0	9	75
Máximo	35	57	74	192	271	3	86	92	1	58	27	19	99

Analisando o gráfico da Figura 28, juntamente com a Tabela 19, nota-se a variabilidade das propriedades da Gasolina Comum e percebe-se um percentual maior na propriedade 95 – Destilação–PFE.

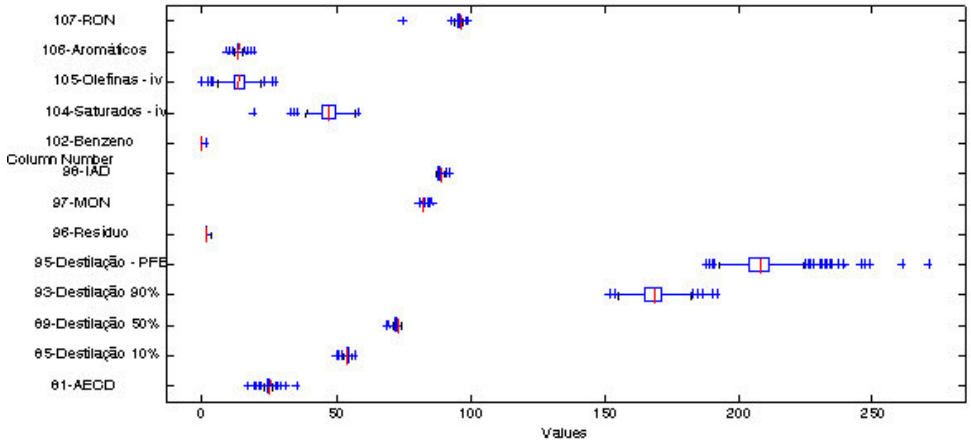

Figura 28. Variabilidade dos resultados em função das variáveis analisadas (GCC).

Tabela 19. Correlação das variáveis do GCC.

	81-AECD	85-Desti	89-Desti	93-Desti	95-Desti	96-Resíd	97-MON	98-IAD
85-Desti	0,490							
89-Desti	0,496	0,760						
93-Desti	-0,141	0,371	0,494					
95-Desti	-0,132	0,230	0,296	0,723				
96-Resíd	-0,038	0,070	0,067	0,293	0,502			
97-MON	0,105	0,056	0,058	0,041	-0,075	-0,054		
98-IAD	0,109	0,085	0,091	0,104	-0,090	-0,057	0,902	

102-Benz	-0,268	-0,021	-0,058	-0,062	-0,032	-0,025	0,027	0,048
104-Satu	-0,267	-0,300	-0,308	-0,222	0,033	0,132	-0,134	-0,295
105-Olef	-0,099	0,055	0,051	0,330	0,045	-0,098	0,077	0,237
106-Arom	-0,312	0,085	0,090	0,102	0,027	-0,007	-0,022	0,028
107-RON	0,047	0,082	0,088	0,128	-0,076	-0,030	0,646	0,751

	102-Benz	104-Satu	105-Olef	106-Arom
104-Satu	0,098			
105-Olef	-0,098	-0,792		
106-Arom	0,343	-0,157	-0,029	
107-RON	0,027	-0,343	0,326	0,089

Apesar de se observar um certo grau de variabilidade para algumas variáveis (Ex.: 75 e 76), observa-se também que a linearidade apresentada em todos os parâmetros pode indicar alguma tendência ou comportamento possivelmente útil.

7.2.2.1 Análise de Componentes Principais

Para Análise de Componentes Principais da GCC utilizou-se o software MINITAB, em um conjunto de 413 amostras e 13 variáveis. Inicialmente foram determinados os autovalores (eigenvalues). Na Tabela 20 se encontram autovalores da GCC.

Tabela 20. Autovalores (GCC).

Eigenvalue	**3,2137**	**2,5004**	**1,8700**	**1,5347**	**1,3786**	**0,6977**	**0,6447**	**0,3470**
Variação	0,247	0,192	0,144	0,118	0,106	0,054	0,050	0,027

Eigenvalue	**0,2922**	**0,2401**	**0,1376**	**0,0726**	**0,0706**
Variação	0,022	0,018	0,011	0,006	0,005

Podemos observar que das 13 componentes principais, 5 componentes têm autovalor maior que 1 (Figura 29). Expressando que essas componentes são significativas na representação do conjunto de variáveis da GCC. Sendo que as duas primeiras componentes representam muito bem o modelo.

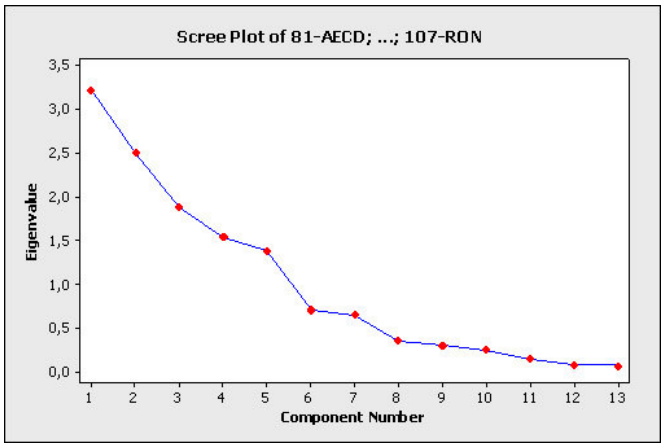

Figura 29. Número de componentes principais do GCC.

Os Coeficientes que indicam o peso da relação entre os componentes principais e as variáveis da GCC, encontram-se na Tabela 21.

Tabela 21. Coeficientes (GCC).

Variáveis	PC1	PC2	PC3	PC4
81-AECD	-0,196	0,079	0,595	0,158
85-Destilação - 10%	-0,324	0,308	0,249	0,090
89-Destilação - 50%	-0,343	0,340	0,220	0,094
93-Destilação - 90%	-0,298	0,367	-0,313	-0,029
95-Destilação - PFE	-0,134	0,437	-0,337	0,147
96-Resíduo	-0,029	0,288	-0,259	0,295
97-MON	-0,324	-0,358	-0,084	0,378
98-IAD	-0,385	-0,364	-0,107	0,261
102-Benzeno	0,053	-0,075	-0,300	0,055
104-Saturados - iv	0,382	0,043	-0,078	0,511
105-Olefinas - iv	-0,299	-0,071	-0,121	-0,586
106-Aromáticos	-0,046	0,012	-0,328	-0,129
107-RON	-0,372	-0,319	-0,142	0,119

Analisando as quatro primeiras componentes principais, vê-se que na 1ª componente predominam os valores negativos. Essas Componentes, melhor representam as variáveis em azul (Tabela 21), é claro que essa análise é um pouco subjetiva, mas importante para analisar parâmetros, principalmente químicos.

No gráfico das componentes principais (Figura 30) vê-se melhor essa representação das variáveis. Podemos ver que algumas variáveis têm tendência negativa e outras positivas para as duas Componentes Principais.

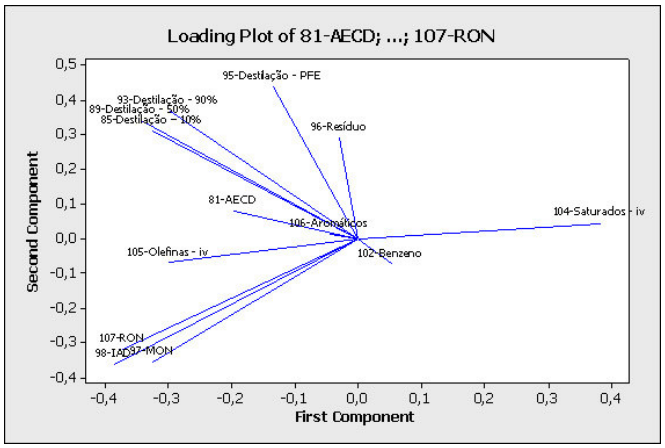

Figura 30. Análise das componentes principais (GCC).

A Figura 31 mostra o gráfico das 2 componentes principais em relação às amostras do GCC. No caso das amostras da GCC é possível verificar outliers, pontos afastados, e uma tendência de dados próximo do zero.

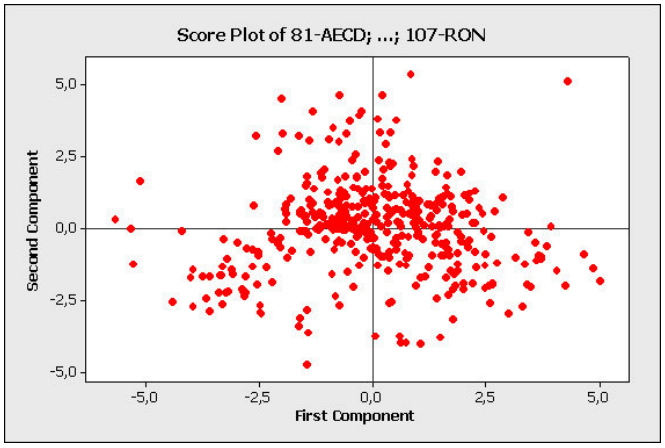

Figura 31. Gráfico das 2 Componentes X Amostras (GCC)

7.2.2.2 Análise de Agrupamento

Utilizou–se o software MINITAB com 13 variáveis de GCC para geração do dendrograma (Figura 32). Foi verificada a possibilidade da formação de até 5

grupos com boa similaridade cada um. No caso das 413 amostras utilizadas para geração do dendrograma (Figura 33), foram formados 13 grupos.

Análise de Agrupamento das Variáveis

A Figura 32 representa o dendrograma geral das variáveis do GCC. Nota-se a possibilidade de dividi-lo em até 2 grupos, entretanto a similaridade entre as variáveis ficaria muito baixa. A divisão em 5 grupos torna-se ideal, pois possui um grau de similaridade razoável e grupos bem definidos.

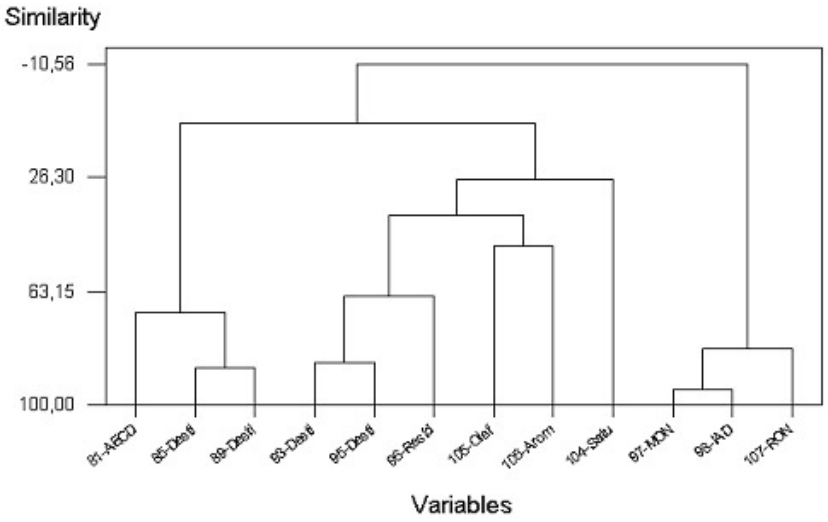

Figura 32. Análise de agrupamento das variáveis.(GCC)

Análise de Agrupamento das Amostras do GCC

Procurou-se o máximo de coerência para o agrupamento das amostras devido ao seu grande número (413) e inconsistência de algumas delas. Apesar dessas dificuldades bons resultados foram encontrados.

No dendrograma geral das amostras (Figura 33) pode se observar que, apesar da dificuldade de visualização devido a grande quantidade de amostras, pelo método de ligação simples, ficou coerente a divisão das amostras de GCC em 13 grupos.

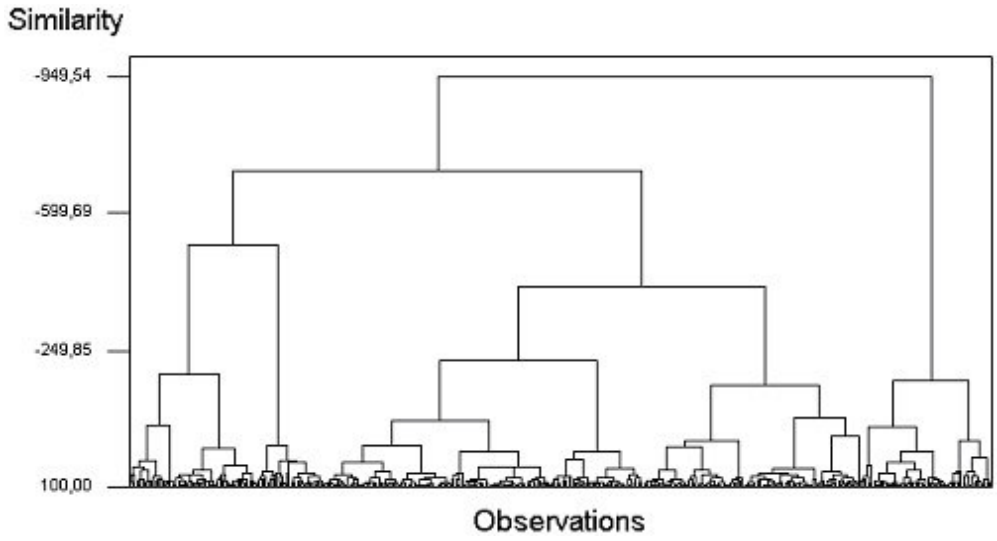

Figura 33. Análise de agrupamento das amostras (GCC).

7.2.2.3 Regressão Múltipla

Para análise de regressão do GCC (Tabela 22), foi definida a variável 95 – Destilação–PFE como a variável de resposta e as demais como variáveis explicativas.

A equação de regressão obtida da matriz de dados do GCC é a seguinte:

(Destilação–PFE) = 163 - 0,103 (AECD) + 0,145 (Destilação–10%) - 1,07 (Destilação-50%) + 1,27 (Destilação-90%) + 5,68 (Resíduo) + 1,38 (MON) - 1,85 (IAD) + 1,69 (Benzeno) - 0,204 (Saturados–iv) - 0,526 (Olefinas–iv) - 0,460 (Aromáticos) - 0,318 (RON)

Na Tabela 23 observa-se que o valor de P para as variáveis: 93, 96, 97, 98, 105 são menores que α-level =0.05. Isso indica que essas variáveis tem significativo valor estatístico para a resposta (95 – Destilação–PFE). Já as demais variáveis um valor P muito alto em relação ao α-level, sendo assim, elas não são significativas no modelo de regressão, portanto poderiam ser descartadas da

análise. As variáveis explicativas em questão explicam de 66,0% a 64,9% a resposta (95-Destilação-PFE).

Tabela 22.Regressão das variáveis (GCC).

	Coef	SE Coef	T	P
Constante	163,14	45,65	3,57	0,000
81. AECD	-0,1034	0,2808	-0,37	0,713
85. Destilação-10%	0,1451	0,4501	0,32	0,747
89. Destilação-50%	-1,0662	0,6688	-1,59	0,112
93. Destilação-90%	1,27473	0,07820	16,30	0,000
96. Resíduo	5,6850	0,6610	8,60	0,000
97. MON	1,3847	0,7052	1,96	0,050
98. IAD	-1,8524	0,7168	-2,58	0,010
102. Benzeno	1,687	1,777	0,95	0,343
104. Saturados-iv	-0,2039	0,1822	-1,12	0,264
105. Olefinas-iv	-0,5258	0,2078	-2,53	0,012
106. Aromáticos	-0,4603	0,3068	-1,50	0,134
107. RON	-0,3179	0,2811	-1,13	0,259
S = 6,04256 R-Sq = 66,0% R-Sq(adj) = 64,9%				

A Figura 34 mostra os gráficos da regressão a GCC.

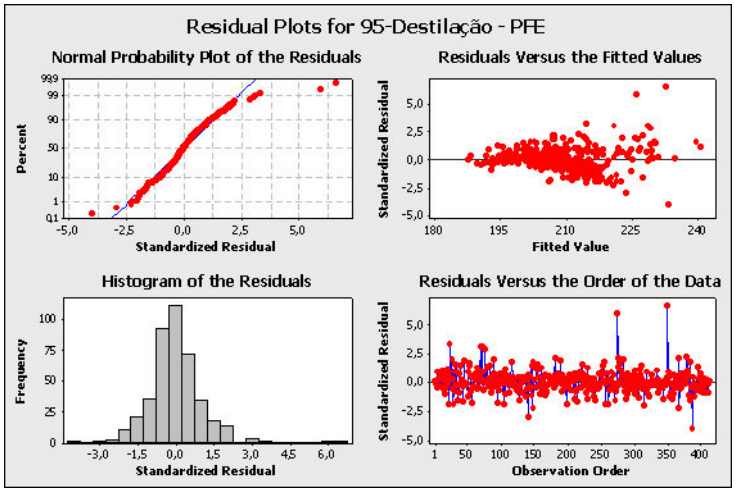

Figura 34. Gráficos da regressão (GCC).

Os gráficos da regressão da GCC (Figura 34) são discutidos a seguir:

- **Histograma dos resíduos**: O histograma de resíduos da GCC indica alguns valores com simetria, e também algumas barras separadas indicando a presença de Outilers.

- **Gráfico da Probabilidade Normal dos Resíduos**: Neste gráfico observa-se algumas amostras fora do padrão e alguns desvios.
- **Gráfico do Resíduo Vs. Ajustes**: Baseado neste gráfico, observa-se alguns pontos distantes de zero podendo ser considerados outliers.
- **Resíduos Vs. Dados**: Neste gráfico observam-se grandes oscilações em relação aos resíduos representando outliers.

7.2.3 GCA – Gasolina Aditivada

O grupo GCA contém 27 variáveis conforme mostra a Tabela 23, sendo que utilizadas apenas 12 variáveis.

Tabela 23. Variáveis da GCA.

Código	Propriedade	Unidade	Observação
52	Cor	Sem unidade	Não Quantitativa
53	Aspecto	Sem unidade	Não Quantitativa
54	AEAC	%v/v	Utilizada
55	Densidade Relativa a 20ºC/4ºC	Sem unidade	Dados Irrelevantes
56	Destilação – PI	°C	Dados Irrelevantes
57	Destilação – 5%	°C	Falta de Dados
58	Destilação – 10%	°C	Utilizada
59	Destilação – 20%	°C	Dados Irrelevantes
60	Destilação – 30%	°C	Dados Irrelevantes
61	Destilação – 40%	°C	Dados Irrelevantes
62	Destilação – 50%	°C	Utilizada
63	Destilação – 60%	°C	Falta de Dados
64	Destilação – 70%	°C	Falta de Dados
65	Destilação – 80%	°C	Falta de Dados
66	Destilação – 90%	°C	Utilizada
67	Destilação – 95%	°C	Falta de Dados
68	Destilação – PFE	°C	Utilizada
69	Resíduo	%v/v	Utilizada
70	Nº de Octano Motor – MON	Sem unidade	Utilizada
71	Índice Antidetonante – IAD	Sem unidade	Utilizada
72	Enxofre	kPa	Dados Irrelevantes
73	Benzeno	mg/100ml	Falta de Dados
74	Álcool Etílico Combustível – iv	%m/m	Falta de Dados
75	Saturados – iv ou ir	%v/v	Utilizada
76	Olefinas – iv	%v/v	Utilizada
77	Aromáticos – iv	%v/v	Utilizada
78	RON – iv	%v/v	Utilizada

A Tabela 24 apresenta os dados estatísticos do grupo GCA a partir de uma matriz de dados gerada pela análise de 102 amostras e 12 variáveis.

Tabela 24. Matriz de dados (GCA).

Propriedades	**54**	**58**	**62**	**66**	**68**	**69**	**70**	**71**	**75**	**76**	**77**	**78**
Amostras	102	102	102	102	102	102	102	102	102	102	102	102
Média	24,17	54,20	72,47	168,53	207,58	1,30	82,54	89,24	47,38	13,90	14,07	95,92
Valor Médio	25	54	73	169	207	1	82	89	48	14	14	96
Desv. Padrão	2,01	1,06	0,84	6,01	10,20	0,49	1,04	1,18	0,18	4,16	3,54	1,37
Variação	3,69	1,35	1,00	28,74	48,38	0,23	1,41	1,75	22,04	16,56	2,40	1,93
Mínimo	19	51	68	153	188	1	81	87	32	1	9	93
Máximo	27	58	74	184	238	3	86	93	61	29	19	99

A Figura 35 mostra a variabilidade das propriedades da GCA, pode-se observar a maior variabilidade na propriedade 69 – Destilação–PFE e também na 66 – Destilação 90%.

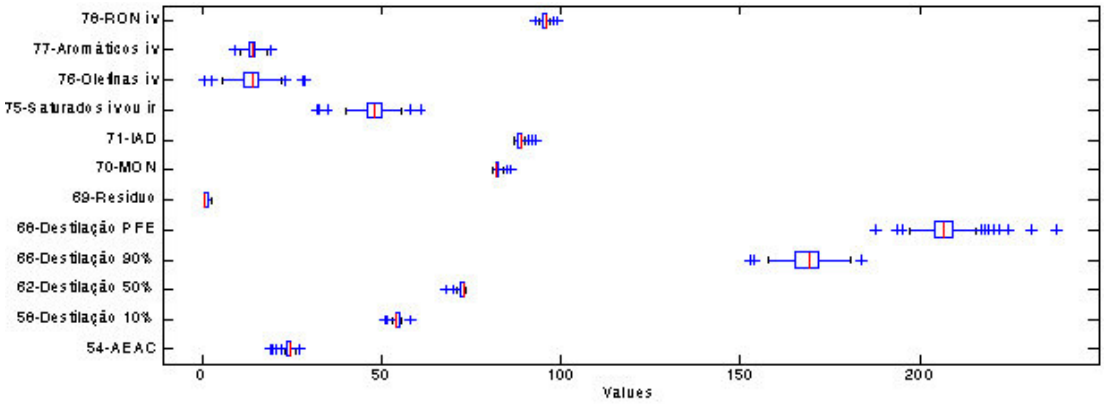

Figura 35. Variabilidade das propriedades do GCA.

Observando a matriz de correlação (Tabela 25) Observa-se por exemplo que a propriedade 75 tem uma força 0,892 com a propriedade 76 e tem direção negativa, isto é, quando uma delas tende a diminuir a outra também diminui.

Tabela 25. Matriz de correlação.

	54	58	62	66	68	69	70	71
58	0,431							
62	0,507	0,774						
66	0,026	0,561	0,538					
68	-0,103	0,309	0,376	0,693				
69	-0,080	-0,289	-0,114	0,055	0,312			
70	0,035	-0,082	0,014	0,012	-0,004	0,034		
71	0,127	0,038	0,144	0,186	0,057	0,068	0,837	
75	-0,218	-0,330	-0,302	-0,355	-0,042	0,118	0,074	-0,267
76	-0,061	0,130	0,055	0,368	0,058	0,030	-0,085	0,199
77	-0,260	0,166	0,161	0,028	0,083	-0,283	-0,125	-0,029
78	0,109	0,077	0,119	0,184	0,025	0,036	0,720	0,914

	75	76	77
76	-0,892		
77	-0,253	0,062	
78	-0,394	0,336	0,081

7.2.3.1 Análise de Componentes Principais

A Análise de componentes principais da GCA compreendeu um conjunto de 102 amostras, observa-se ai um o menor conjunto de amostras analisadas dos 4 combustíveis analisados, e 12 variáveis. Os autovalores referentes a este combustível encontram-se na Tabela 26.

Tabela 26. Autovalores (GCA).

Eigenvalue	**3,3836**	**2,1656**	**1,7502**	**1,5958**	**1,0533**	**0,7761**	**0,4357**	**0,3155**
Variação	0,282	0,180	0,146	0,133	0,088	0,065	0,036	0,026

Eigenvalue	**0,2026**	**0,1609**	**0,1371**	**0,0236**
Variação	0,017	0,013	0,011	0,002

Podemos observar que das componentes principais, 5 tem autovalor maior que 1. Isto indica que essas componentes são significativas na representação do conjunto de variáveis da GCA. Sendo que as duas primeiras possuem um maior grau de representação do modelo. Este comportamento pode ser observado na Figura 36.

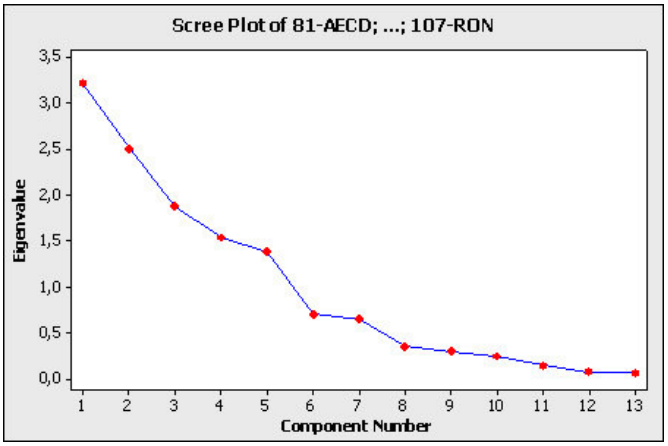

Figura 36. Componentes principais da GCA.

Os Coeficientes que indicam o peso da relação entre os componentes principais e as variáveis da GCA, encontram-se na Tabela 27.

Tabela 27. Coeficientes (GCA).

Variáveis	PC1	PC2	PC3	PC4
54-AEAC	-0,163	-0,229	-0,119	-0,575
58-Destilação 10%	-0,372	-0,370	-0,007	-0,163
62-Destilação - 50%	-0,382	-0,338	0,077	-0,231
66-Destilação - 90%	-0,421	-0,141	0,202	0,181
68-Destilação - PFE	-0,283	-0,131	0,473	0,284
69-Residuo	-0,083	0,122	0,496	0,177
70-MON	-0,100	0,445	0,242	-0,409
71- IAD	-0,269	0,425	0,132	0,025
75-Saturados - iv ou ir	0,391	-0,109	0,456	-0,136
76-Olefinas - iv	-0,317	0,201	-0,415	0,286
77-Aromáticos - iv	-0,108	-0,024	-0,117	0,326
78-RON	-0,279	0,461	-0,017	-0,265

Analisando as 4 primeiras Componentes principais, vemos que na 1ª predominam os valores negativos. Essas componentes melhor representam as variáveis em azul.

A Figura 37 mostra este comportamento em relação as duas primeiras componentes principais. Neste gráfico vê-se a representação das variáveis. Pode-se notar que algumas têm maior tendência negativa e outras positivas para as duas

componentes principais. Observar-se também a formação de grupos, sendo que a propriedade 75 está distante das demais.

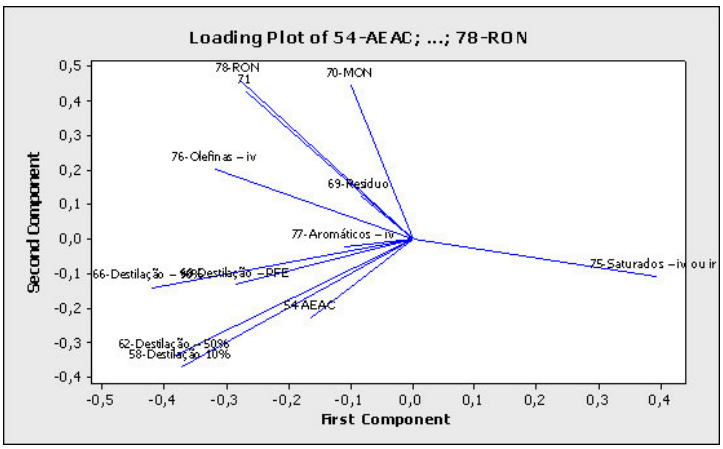

Figura 37. Análise das componentes principais (GCA).

Na figura 38, pode-se ver o gráfico das 2 componentes principais em relação as amostras do GCA, já que essas componentes representam a maior variância dos dados. No caso das amostras do GCA, é possível a verificação de outliers, e uma tendência de dados próximos do zero, com algumas dispersões.

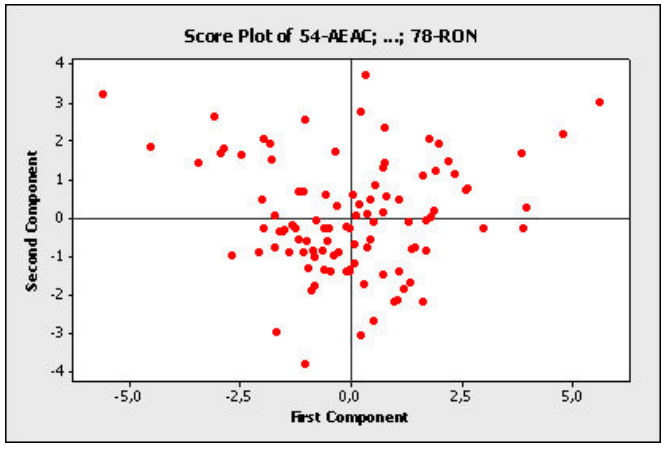

Figura 38. Gráfico das 2 Componentes X Amostras

7.2.3.2 Análise de Agrupamento

Utilizou–se o software MINITAB com 13 variáveis de GCA para geração do dendrograma (Figura 39). Foi verificada a formação de até 5 grupos, cada um com boa similaridade. No caso das 102 amostras utilizadas para geração do dendrograma (Figura 40) foram formados 13 grupos

Agrupamento das Variáveis do GCA

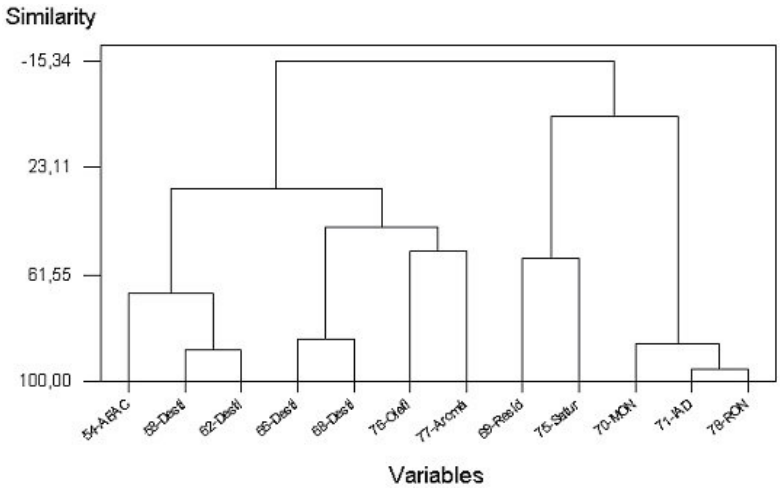

Figura 39. Análise de agrupamento das variáveis do GCA.

Devido a grande diferença entre essas propriedades tornou-se difícil um bom agrupamento, mas pelo método de ligação simples, os 5 grupos formados mostraram um bom grau de similaridade.

Análise de Agrupamento das Amostras do GCA

Dentro de um grupo de 102 amostras de GCA, procurou-se o máximo de coerência para o agrupamento dessas amostras. A Figura 40 mostra o dendrograma geral das amostras da GCA. Nele é possível observar uma grande diversidade nas amostras, sendo difícil seu agrupamento. Pelo método de ligação simples é possível classificar essas amostras de GCA em 7 Grupos.

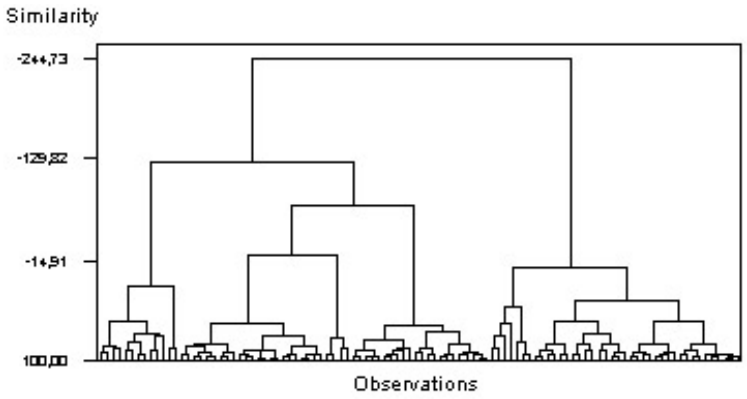

Figura 40. Análise de agrupamento das amostras do GCA.

7.2.3.3 Regressão Múltipla

Para análise de regressão do GCA, a variável 95 – Destilação–PFE foi definida como a variável de resposta e as demais como variáveis explicativas.

A equação de regressão obtida da matriz de dados da GCA é a seguinte:

(Destilação–PFE) = 67,8 - 0,551 (AEAC) - 0,144 (Destilação 10%) + 0,486 (Destilação 50%) + 0,971 (Destilação 90%) + 4,73 (Resíduo) + 0,605 (MON) - 0,70 (IAD) - 0,317 (Saturados– iv ou ir) - 0,665 (Olefinas–iv) + 0,401 (Aromáticos– iv) - 0,138 (RON)

Na tabela 28 observa-se que o valor de P para as variáveis: 66, 69 são menores que α-level = 0.05. Isso indica que essas variáveis tem significativo valor estatístico para a resposta (68-Destilação-PFE). Já as demais variáveis possuem um valor P muito alto em relação ao α-level, sendo assim elas não tem significativo valor para este modelo de regressão, portanto poderiam ser descartadas da análise. As variáveis explicativas em questão explicam de 64,1% a 59,7% a resposta (68-Destilação-PFE).

Tabela 28. Regressão das variáveis (GCA).

	Coef	SE Coef	T	P
Constante	67,82	71,53	0,95	0,346
54. AEAC	-0,5509	0,4659	-1,18	0,240
58. Destilação-10%	-0,1435	0,7060	-0,20	0,839
62. Destilação-50%	0,4863	0,8463	0,57	0,567
66. Destilação-90%	0,9708	0,1287	7,54	0,000
69. Resíduo	4,731	1,065	4,44	0,000
70. MON	0,6054	0,8442	0,72	0,475
71. IAD	-0,698	1,198	-0,58	0,562
75. Saturados-iv ou ir	-0,3165	0,4543	-0,70	0,488
76. Olefinas-iv	-0,6649	0,4822	-1,38	0,171
77. Aromáticos-iv	0,4010	0,5010	0,80	0,426
78. RON	-0,1384	0,9166	-0,15	0,880
S = 4,41654 R-Sq = 64,1% R-Sq(adj) = 59,7%				

A Figura 41 mostra os gráficos da regressão da GCA.

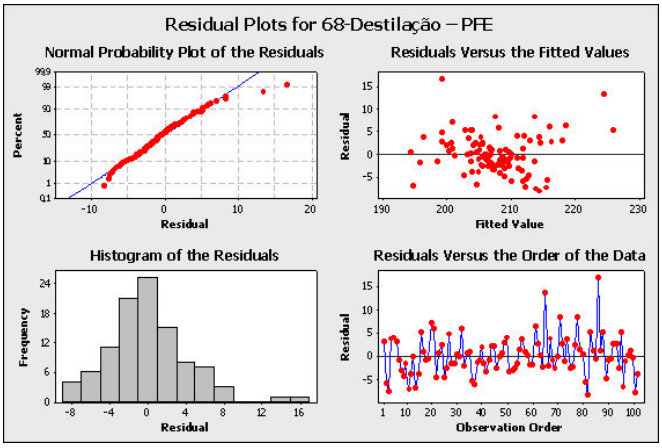

Figura 41. Gráficos da regressão (GCA).

Os gráficos da regressão da GCA (Figura 41) são discutidos a seguir:

Histograma dos resíduos: O histograma de resíduos da GCA indica boa simetria, não existindo barras separadas.

Gráfico da Probabilidade Normal dos Resíduos: Neste gráfico observam-se algumas amostras fora do padrão e alguns desvios.

Gráfico do Resíduo Vs. Ajustes: Baseado neste gráfico, observa-se alguns pontos distantes de zero que podem ser considerados Outliers.

Resíduos Vs. Dados: Observam-se alguns pontos distantes de zero que podem ser considerados outliers.

7.2.4 OBC – Óleo Diesel Comum

O Grupo OBC contém 23 variáveis (Tabela 29), sendo que apenas 3 variáveis foram utilizadas.

Tabela 29. Propriedades (OBC).

Código	Propriedade	Unidade	Observação
*12	Aspecto	Sem unidade	Não Quantitativa
*13	Cor	Sem unidade	Não Quantitativa
*14	Composição – Enxofre (S)	%m/m	Falta de Dados
*15	Densidade a 20ºC/4ºC	Sem unidade	Não Obrigatório
*16	Ponto de Entupimento	Sem unidade	Falta de Dados
17	Índice de Cetano	°C	Utilizada
*18	Destilação – PI	°C	Falta de Dados
*19	Destilação – 5%	°C	Falta de Dados
*20	Destilação – 10%	°C	Falta de Dados
*21	Destilação – 20%	°C	Falta de Dados
*22	Destilação – 30%	°C	Falta de Dados
*23	Destilação – 40%	°C	Falta de Dados
24	Destilação – 50%	°C	Utilizada
*25	Destilação – 60%	°C	Falta de Dados
*26	Destilação – 70%	°C	Falta de Dados
*27	Destilação – 80%	°C	Falta de Dados
28	Destilação – 85%	°C	Utilizada
*29	Destilação – 90%	°C	Utilizada
*30	Destilação – 95%	°C	Utilizada
*31	Destilação – PFE	Sem unidade	Utilizada
*182	Ponto de Fulgor	°C	Falta de Dados
*186	Cor Visual	Sem unidade	Não Quantitativa
*190	Corante	Sem unidade	Não Quantitativa
*194	Massa específica a 20ºC	%v/v	Falta de Dados

* Propriedades não utilizadas nas analises de dados

A Tabela 30 apresenta os dados estatísticos do grupo OBC a partir de uma matriz de dados gerada pela análise de 387 amostras e 3 variáveis.

Tabela 30. Matriz de dados do OBC.

Propriedades	**17-Índice de Cetano**	**24-Destilação – 50%**	**28-Destilação – 85%.**
Amostras	387	387	387
Média	51,1085	283,132	336,313

Valor Médio	50	283	336
Desvio Padrão	3,72133	6,07899	6,82254
Variação	13,8483	36,9541	46,5471
Mínimo	44	266	305
Máximo	65	303	359

O gráfico da Figura 42 mostra a variabilidade das propriedades ao longo das amostras de OBC. Percebe-se uma maior variação na propriedade 28 – Destilação–85%.

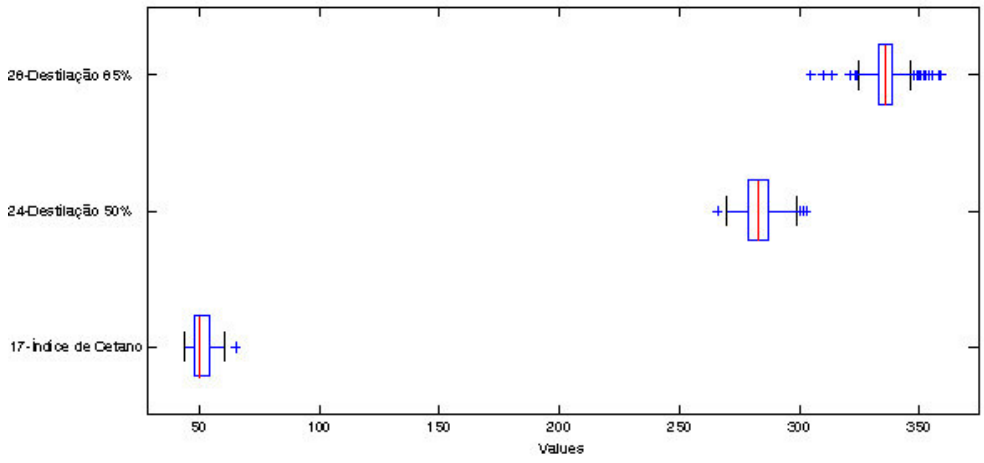

Figura 42. Variabilidade dos resultados em função das variáveis analisadas da OBC.

Dentre as propriedades do OBC, pela matriz de correlação (Tabela 31) observa-se que a propriedade 24 – Destilação–50% tem maior relação com a propriedade 28 – Destilação–85%. Esta relação é medida pela força 0,636 que tem direção positiva, ou seja, quando a propriedade 24 tende a aumentar a 28 também aumenta.

Tabela 31. Matriz de correlação (OBC)

	17-Índic	24-Desti
24-Desti	0,500	
28-Desti	0,065	0,636

7.2.4.1 Análise de Componentes Principais

Para Análise de Componentes Principais utilizou-se o software MINITAB, em um conjunto de 387 amostras, observa-se, que dentre os grupos de combustíveis analisados, este possui o menor conjunto de variáveis, apenas 3. Isso se deu pela falta de dados das demais variáveis, pois sem as 387 amostras de cada propriedade não foi possível colocar essas variáveis na análise. Apesar disso, foi possível a representação do conjunto de dados com 3 variáveis na análise de componentes principais. Os autovalores do OBC são apresentados na Tabela 32.

Tabela 32. Autovalores (OBC).

	PC1	PC2	PC3
Eigenvalue	1,8415	0,9366	0,2219
Variação	0,614	0,312	0,074

Pode-se observar que das 3 possíveis componentes principais, apenas uma possui autovalor maior que 1. Isto significa que esta componente pode representar o conjunto de variáveis do OBC. Entretanto, a 2ª componente também pode representar o modelo, mas com menos expressão. Este comportamento é visto na Figura 43.

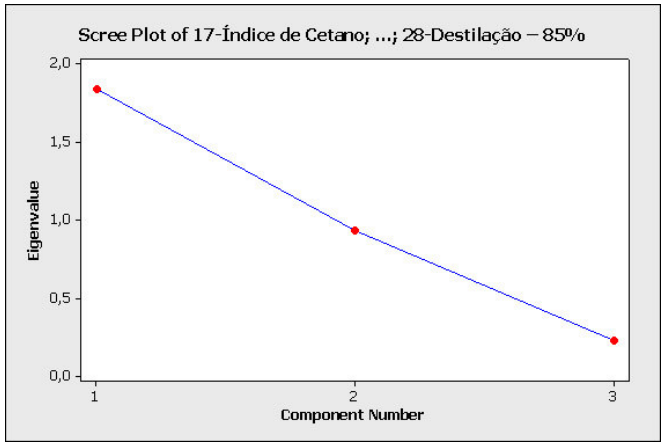

Figura 43. Componentes Principais (OBC).

Os Coeficientes que indicam o peso da relação entre os componentes principais e as variáveis da OBC, encontram-se na Tabela 33.

Tabela 33. Coeficientes (OBC).

Variáveis	**PC1**	**PC2**	**PC3**
17-Índice de Cetano	-0,455	-0,787	-0,416
24-Destilação – 50%	-0,693	0,019	0,721
28-Destilação – 85%	-0,559	0,616	-0,554

Analisando as 3 primeiras componentes principais (Tabela 33), vemos que na 1 Componente predominam os valores negativos. Sendo que essa componente melhor representa as variáveis em azul. Esse comportamento em relação às duas primeiras componentes principais é visto na Figura 44. Neste gráfico vê-se melhor essa representação das variáveis. Nota-se que algumas variáveis têm maior tendência negativa para a primeira componente e outras positivas para a segunda.

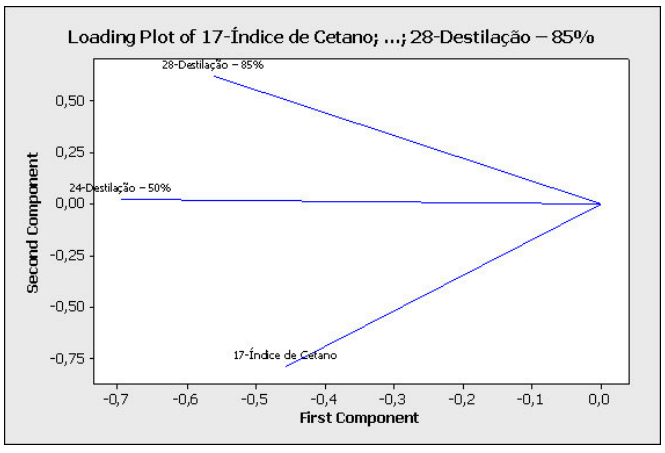

Figura 44. Análise das componentes principais (OBC).

A Figura 45 mostra o gráfico das 2 componentes principais em relação às amostras do OBC, já que essas componentes representam a maior variância dos

dados. No caso das amostras do OBC é possível verificar outliers e uma tendência de dados próximo do zero.

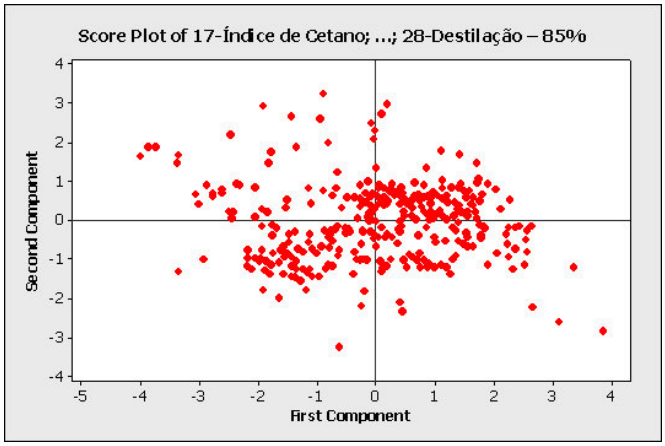

Figura 45. Gráfico das componentes X amostras (OBC)

7.2.4.2 Análise De Agrupamento

Utilizou – se o software MINITAB com 3 variáveis de OBC para geração do dendrograma (Figura 46). Foi verificada a possibilidade de formação de até 2 grupos com boa similaridade cada um. No caso das 387 amostras de OBC utilizadas para geração do dendrograma (Figura 47), foram formados 8 grupos.

Agrupamento das variáveis do OBC

A Figura 46 representa o dendrograma geral das variáveis do OBC. Nota-se a possibilidade de dividi-las em até 2 grupos, sendo que um grupo ficaria com apenas uma variável, o 17 – Índice de Cetano, e o outro com duas variáveis, a 24 – Destilação–50% e 28 – Destilação–85%.

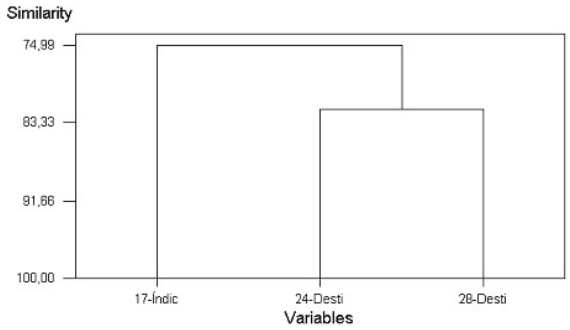

Figura 46. Análise de agrupamento das variáveis do OBC.

Agrupamento das Amostras do OBC

A Figura 47 mostra o dendrograma geral das amostras do OBC. Tendo em vista a boa distribuição dos dados, o formação de 8 grupos é suficiente.

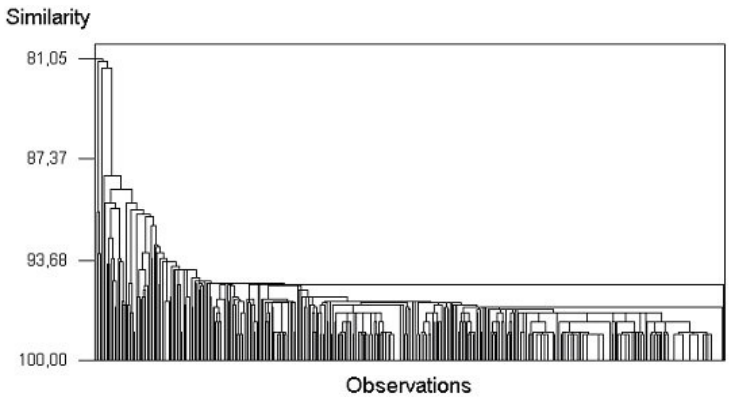

Figura 47. Análise de agrupamento das amostras do OBC.

7.2.4.3 Regressão Múltipla

Para análise de regressão do OBC, a variável 28 – Destilação–85% foi definida como a variável de resposta e as demais como variáveis explicativas.

A equação de regressão obtida da matriz de dados do OBC é a seguinte:

(Destilação–85%) = 112 - 0,618 (Índice de Cetano) + 0,903 (Destilação–50%)

Na Tabela 34 observa-se que o valor de P para as variáveis 17 e 24 são menores que α-level = 0.05. Isso indica que essas variáveis têm significativo valor estatístico para a resposta (28 – Destilação–50%). As variáveis explicativas em questão explicam de 49,0% a 48,7% a Resposta (28 – Destilação–50%).

Tabela 34. Regressão das variáveis (OBC).

	Coef	SE Coef	T	P
Constante	112,21	11,90	9,43	0,000
17. índice de Cetano	-0,61757	0,07715	-8,00	0,000
24. Destilação-50%	-0,90298	0,04723	-19,12	0,000
S = 4,88551 R-Sq = 49,0% R-Sq(adj) = 48,7%				

A Figura 48 mostra os gráficos da regressão do OBC.

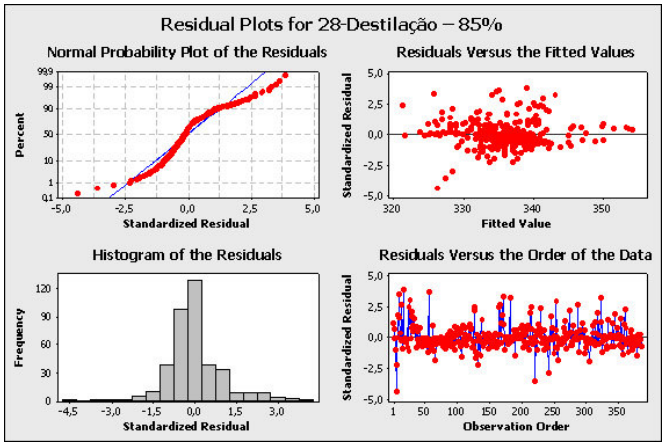

Figura 48. Gráficos da Regressão (OBC).

Os gráficos da regressão do OBC (Figura 48) são discutidos a seguir:

Histograma dos resíduos: No Histograma de Resíduos da OBC indica boa Simetria, mas existem algumas barras separadas, o que indica a presença de Outilers.

Gráfico da Probabilidade Normal dos Resíduos: Neste gráfico observam-se algumas amostras fora do padrão e alguns desvios.

Gráfico do Resíduo Vs. Ajustes: Observam-se alguns pontos distantes de zero que podem ser considerados outliers.

Resíduos Vs. Dados: Nas observações em relação aos resíduos notam-se algumas oscilações que representam outliers.

7.3 Sistema de Informação Geográfica – SIG

Além do desenvolvimento de um *data warehouse* e aplicação de técnicas de mineração de dados, este trabalho propõe a abordagem de uma das tecnologias de informação que cada vez mais está acessível às empresas brasileiras. Os Sistemas de Informação Geográfica (SIG) têm apresentado um crescimento grande nos EUA, em torno de 20% ao ano. No Brasil, apesar de não existirem estatísticas específicas, alguns especialistas estimam que existe um crescimento na ordem de 30% ao ano. Este crescimento está fortemente ligado ao fato de que entre 70% e 80% de todos os dados estarem geograficamente referenciados.

Além disso, outra razão pela qual a integração desse sistema é de grande utilidade, é que no roteamento de veículos, este tipo de sistema é fundamental, pois permite ao usuário visualizar as rotas que foram geradas a partir de um algoritmo. Isto é bastante útil na formação de rotas para coleta de amostras de combustíveis no Estado do Maranhão.

Na Figura 49 é visualizado o mapa digital da cidade de São Luis, destacando-se dois postos de combustíveis. As coordenadas de cada posto utilizadas para a implantação do SIG foram obtidas no banco de dados da ANP.

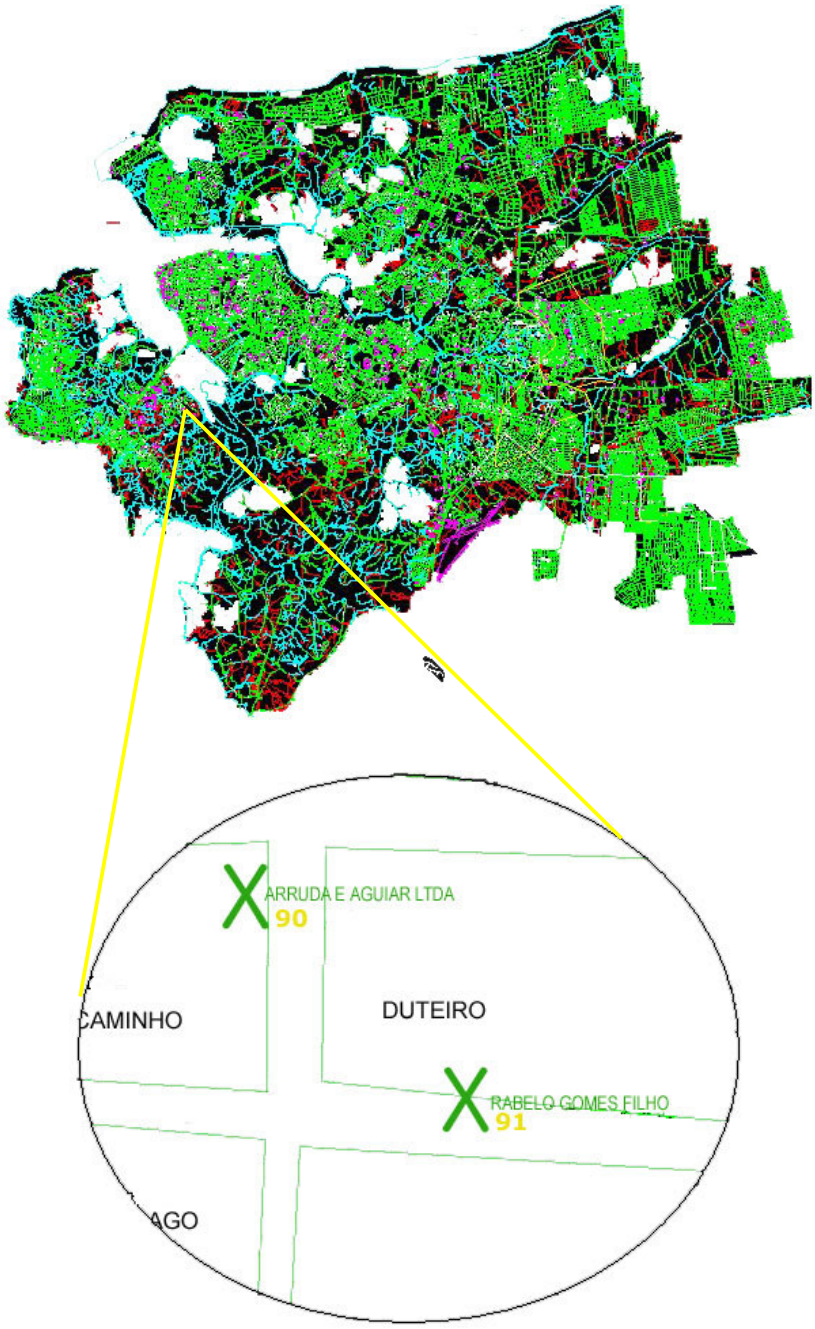

Figura 49. Sistema de Informação Geográfica.

O objetivo deste trabalho é o de propor e apenas dar início ao desenvolvimento de um SIG. Esta proposta visa uma futura integração entre o data warehouse com as informações sobre os postos de combustíveis, além dos

resultados obtidos durante todo o processo KDD, e as informações espaciais dos postos do estado.

7.4 Ferramentas

Existe uma série de ferramentas disponibilizadas para obtenção do sucesso na construção de *data warehouses* e aplicações de técnicas de mineração de dados. Na Tabela 35 são listadas as que foram utilizadas neste trabalho.

Tabela 35. Ferramentas.

Ferramenta	Empresa	Site	Etapa
Analysis Manager	Microsoft	http://www.microsoft.com	*Data Warehouse*
Minitab v.14	Minitab	http://www.minitab.com	*Data Mining*
Statistica 6.0	Statsoft	http://www.statsoft.com	*Data Mining*
Matlab 6.5	Mathworks	http://www.mathworks.com	*Data Mining*
The Unscrambler v. 7.01	camo	http://www.camo.no	*Data Mining*
AutoCAD Map 2000	Autodesk	http://www.autodesk.com	*SIG*

7.5 Conclusão

Em todas as análises feitas procurou-se fazer uma padronização dos dados, durante a etapa de *data warehouse*, já que estas estão com unidades e valores diferentes. Em seguida foram aplicadas as técnicas de mineração de dados, visando compreender ou reduzir a dimensão dos dados analisando a estrutura de covariância dos dados. As técnicas aplicadas foram:

- Análise de Componentes Principais - usados para reduzir os dados em um menor número de componentes.
- Análise de Agrupamento (Amostras) – usado para classificar observações similares em grupos quando os grupos forem inicialmente desconhecidos.

- Análise de Agrupamento (Variáveis) – usado para classificar variáveis similares em grupos similares quando os grupos forem inicialmente desconhecidos.
- Regressão Múltipla - Tem como objetivo a previsão de uma variável de resposta em relação a um conjunto de variáveis.

V. CONCLUSÃO E TRABALHOS FUTUROS

8 CONSIDERAÇÕES FINAIS

Esta dissertação foi feita com a finalidade de apresentar o projeto SIMCO. Nela foi mostrada a aplicação do processo KDD em um banco de dados de análise de combustíveis. Foram enfatizadas, neste trabalho, as etapas de *data warehouse* e mineração de dados, além do inicio da implantação de um SIG.

A seguir apresenta-se de forma resumida os resultados da aplicação das técnicas de mineração de dados usadas, os quais expressam a conclusão do presente trabalho.

Análise de Componentes Principais

A análise de dados utilizando a técnica de Análise de Componentes Principais foi bastante proveitosa em relação à redução do número de variáveis. Dentre os quatro tipos de combustíveis analisados, apenas em um (OBC) foram encontradas dificuldades. A dificuldade encontrada reside no fato de haverem poucas variáveis disponíveis, apenas três. Neste caso, apesar de uma componente principal representar bem as demais variáveis, optou-se por duas componentes principais, uma vez que não é possível o gráfico das componentes somente com uma variável. De qualquer forma, a segunda componente tem uma representatividade razoável em relação às amostras.

Em geral, foi possível a redução de variáveis em componentes principais que representam bem o modelo. Analisando as componentes principais formadas, e as variáveis por elas representadas, tem-se uma visão mais clara das relações existentes entre as variáveis. No Contexto das amostras a existência de alguns outilers pôde ajudar na Análise de Agrupamentos, feita posteriormente, como também mostrou uma tendência centralizada das amostras.

Na tabela 36, se pode ver um resumo da aplicação desta técnica aos dados referentes às amostras coletadas e aos tipos de combustíveis analisados.

Tabela 36. Resumo da aplicação da Análise de Componentes Principais.

Combustível	Nº de Variáveis	Nº de Componentes Principais
AEH	4	2
GCC	13	5
GCA	12	4
OBC	3	2

Análise de Agrupamento

Foram realizadas análises de agrupamento tanto no universo das amostras como no das variáveis de cada tipo de combustível. Em ambas foram gerados dendrogramas. Onde foi possível a identificação de grupos, formados devido a similaridades encontradas, tanto nas amostras como nas variáveis. Este tipo de análise é muito subjetivo, e no caso deste trabalho, poderiam ser formados mais grupos, entretanto menos significativos. Por esta razão, foi-se bem rígido quanto ao grau de similaridade, entre as amostras e variáveis, visando a formação de agrupamento melhores ajustados.

De modo geral a Análise de Agrupamento foi bastante eficiente, possibilitando futuros estudos das amostras e variáveis estudadas, como também facilitando na análise de Hipóteses e Tomada de Decisão.

A tabela 37 mostra um resumo da aplicação desta técnica aos dados referentes às amostras coletadas e aos tipos de combustíveis analisados.

Tabela 37. Resumo da aplicação da Análise de Agrupamento.

Combustível	Nº de Amostras	Nº de Grupos Formados	Nº de Variáveis	Nº de Grupos Formados
AEH	154	6	4	2
GCC	413	13	13	5
GCA	102	13	12	5
OBC	387	8	3	2

Regressão Múltipla

A análise de regressão tem como principal objetivo prever uma variável de resposta em relação às demais variáveis de um determinado conjunto de dados através de um modelo matemático. O critério utilizado neste trabalho para escolha da variável resposta foi o nível de variação das variáveis ao longo das amostras. Foi escolhida a variável com maior variação por representar melhor o conjunto. Vale lembrar que qualquer outra variável poderia ser escolhida.

A tabela 38 mostra um resumo da aplicação desta técnica aos dados. A variável resposta de cada tipo de combustível, juntamente com suas respectivas variáveis independentes.

Tabela 38. Resumo da aplicação da Regressão.

Combustível	Variável Resposta	Variáveis Independentes
AEH	7 (Condutividade Elétrica)	8 (Massa Especifica); 9 (pH) ; 10 (Teor Alcoólico).
GCC	95 (Destilação-PFE)	13
GCA	68 (Destilação-PFE)	54 (AEAC); 58 (Destilação-10%); 62 (Destilação-50%); 66 (Destilação-90%); 69 (Resíduo); 70 (MON); 71 (IAD); 75 (Saturados-iv ou ir); 76 (Olefinas-iv); 77 (Aromáticos-iv); 78 (RON-iv).
OBC	28 (Destilação-85%)	17 (Índice de Cetano); 24 (Destilação–50%).

Com base no exposto, espera-se que os resultados obtidos neste trabalho possam contribuir com o processo de análise de combustíveis, tendo em vista a grande quantidade de dados acumulados. Os resultados poderão auxiliar na interpretação de resultados, bem como na predição de comportamentos funcionais

dos diversos parâmetros físico-químicos que compõem o programa de monitoramento da ANP.

Propostas de Trabalhos Futuros:

- Propor novas modelagens para o *data warehouse;*
- Melhoramento do SIG;
- Integração do *data warehouse* com o SIG, formando um *data warehouse* espacial;
- Aplicação de outras técnicas de *data mining*;
- Utilização da tecnologia de agentes inteligentes, possibilitando maior automação no processo;

REFERÊNCIAS

(Aitchison, 1986) Aitchison, J. **The statistical analysis of compositional data**. Chapman & Hall. 1986.

(Anderberg, 1973) Anderberg, D. **Cluster Analysis for Applications**. 1 ed. New York: Academic Press, 1973.

(Aurélio et al., 1999) Aurélio, M., Vellasco, M., and Lopes, C. H. (1999). **Descoberta de conhecimento e mineração de dados**. Apostila.ICA - Laboratório de Inteligência Computacional Aplicada, DEE, PUC-Rio.

(Bacci, 2000) BACCI, D. L. C. **Vibrações geradas pelo uso de explosivos no desmonte de rochas: avaliação dos parâmetros físicos do terreno e dos efeitos ambientais.** Tese de Doutorado, Programa em Geociências e Meio Ambiente, I.G.C.E., UNESP, Rio Claro. 2000.

(Barquini, 1996) Barquini, Ramon. **Planning and designing the Warehouse**. New Jersey, Prentice-Hall, 1996. 311p.

(Batini et al., 1986) Batini, C., Lenzerini, M. **Comparative Analysis Of Methodologies For Database Schema Integration**, ACM Computing Surveys. New York, v.18, nº 4, pág.323 - 364, Dez/1986.

(Berry, 2000) Berry, M. and Linoff, G., **Mastering Data Mining á Art and Science of Customer Relationship Management**, Ed. Wiley, 2000.

(Box et al., 1978) Box, G. E. P., Hunter, W. G. & Hunter, J. S., **Statistic for Experimenters**. 1 ed. New York, John Wiley & Sons, 1978.

(Campos et al., 1997) Campos, Maria Luiza e Rocha, Arnaldo V. **Data Warehouse**. XVII Congresso da Sociedade Brasileira de Computação, XVI Jornada de Atualização em Informática, Rio de Janeiro, 1997.

(Campos et al., 2003) Campos, Maria Luiza e Rocha, Arnaldo V. **Características do Data Warehouse**. Disponível em: http://genesis.nce.ufrj.br/dataware/tutorial/cbasicos.html#tg_I5.Acessado em 09 de Setembro de 2003.

(Davis,1986) Davis, J.C. **Statistics and Data Analysis in Geology**. 2nd. ed., John Wiley and Sons, Inc. 1986.

(Decker et al., 1995) Decker, K.M. and Focardi, S. (1995). **Technology Overview: A Report on Data Mining**. Technical Report CSCS-ETH TR-95-02, Swiss Scientific Computing Center.

(Domenico, 2001) Domenico, Jorge Antonio Di. **Definição de um Ambiente de Data Warehouse em uma Instituição de Ensino Superior**. Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina como requisito parcial para obtenção do título de Mestre em Engenharia de Produção. Aprovada em 28 de Fevereiro de 2001.

(Dougherty et al., 1995) J. Dougherty, R. Kohavi and M. Sahami. **Supervised and unsupervised discretization of continuos features**. Proc. 12th Int. Conf. Machine Learning, 194-202. 1995.

(Dilly, 1995) Dilly, Ruth. **Data Mining – An Introduction**. Parallel Computer Centre. Queen's University of Belfast. 1995.

(Elder, 1996) J. F. Elder IV and D. Pregibon. **A statistical perspective on knowledge discovery in data bases**. In: U. M. Fayyad et al. (Ed.) Advances in Knowledge Discovery and Data Mining, 83-113. AAAI/MIT Press, 1996.

(Everitt, 1980) Everitt, B. **Cluster Analysis**. 2nd ed., Gower Publishing Co. 1980.

(Fayyad et al., 1996a) Fayyad, U. (Ed.); Piatetski-Shapiro (Ed.). **Advances in Knowledge Discovery in Databases and Data Mining**, Massachusetts: AAAI Press, The MIT Press, 1996a. 611 p.

(Fayyad et al., 1996b) Fayyad, U.M.; Piatetsky-Shapiro,G and Smyth, P. **From Data Mining to Knowledge Discovery in Databases** in AI Magazine, 17:3, pp. 37-54, 1996.

(Firestone, 1998) Firestone, Joseph M. (1998) **Architectural Evolution in Data Warehousing and Distribuited Knowledge Management Architecture**. White Paper No. Eleven, Executive Information Systems, Inc.

(Goldberg, 1989) D. E. Goldberg. **Genetic Algorithms in Search, Optimization, and Machine Learning**. Reading, MA: Addison-Wesley, 1989.

(Gonçalves, 2003) Gonçalves, Marcio. **Extração de Dados Para Data Warehouse**. Axcel, 2003. 160 p.

(Gower, 1966) Gower, J. C. **Some distance properties of latent root and vector methods used in multivariate methods**. Biometrika, 55: 325-338. 1966.

(Griffiths, 1995) Griffiths, S. **Data Warehousing – What, Where, Why and How**. Johannesburg, África do Sul, The Data Warehousing Conference, 1995.

(Groth,1998) Groth, Robert. **Data mining: a hands-on approach for business professionals**. Prentice Hall, New Jersey, 1998.

(Hackney, 1998) Hackney, Douglas. **Data Warehouse Delivery: Who Are You?** – Part I. DM Review Magazine, Vol. 8, No 2, Fevereiro, 1998.

(Haykin, 1994) Haykin, S., **Neural Networks: A Comprehensive Foundation**, Macmillan College Publishing Company, New York, NY, 1994.

(Harjinder et al., 1996) Harjinder, G. e Rao, P. C. **The Official Guide to Data Warehousing**. Que Corporation, 1996.

(Henrique, 1998) Henrique, M. **Data Warehouse : da realidade ao estado da arte**, 1998. Disponível em: http://www.wmc.com.br/revista/dataw.htm

(Hunt, 2000) Hunt,John. **Knowledge Discovery in Databases**. 12 de Novembro de 2000. Disponível em: http://www.jaydeetechnology.co.uk/expertsystems/KDD7.pdf

(Hurwitz, 1996) Hurwitz, Judit. **Preparing for the Warehouse** – DBMS Magazine, April 1996 – disponível em http://www.dbmsmag.com/9604d04.html

(Inmon, 1997) Inmon, W. H. **Como Construir o Data Warehouse**, Campus, Rio de Janeiro. 1997. 387 p.

(Inmom, 1998) Inmom, W.H. **Data Mart Does Not Equal Data Warehouse**. Disponível na INTERNET via http://www.dmreview.com/issues/1998/may/articles/may98_38.htm. Artigo nconsultado em 1998.

(Inmom et al., 1997a) Inmom, W.H & Hackathorn, Richard D. **Como Usar o Data Warehouse**. Rio de Janeiro: Editora Infobook, 1997. 277p.

(Inmon et al., 1997b) Inmon, W. H., J.D. Welch, and K.L. Glassey. **Managing the Data Warehouse**. New York, NY: John Wiley &; Sons, 1997. 400 p.

(Junior et al., 2000) Junior, C. L. N. and Yoneyama, T. (2000). **Inteligência Artificial em Controle e Automação**. Edgard Blücher Ltda.1ª edition.

(Joreskog, 1976) Joreskog, K.G., Klovan, J.E. & Reyment, R.A. **Geological factor analysis**. Elsevier. 1976.

(Kimball, 1996) Kimball, Ralph. **Data Warehouse Toolkit**. John Wiley & Sons Inc., New York, 1996. 388 p.

(Kimball et al.,1998) Kimball, Ralph; Reeves Laura; Ross Margy; Thornthwaite Warren. **The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses**. John Wiley & Sons Inc., New York, 1998. 771 p.

(Labidi, 2003) Labidi, Sofiane. **SIMCO: Realização e Implantação de um Sistema Integrado para Monitoramento da Qualidade de Combustível**. Projeto submetido ao CNPq dentro do edital Procet. 2003.

(Langley, 1996) P. Langley. **Elements of Machine Learning**. Morgan Kaufmann, 1996.

(Lee, 1995) H-Y. Lee, H-L. Ong and L-H. Quek. **Exploiting visualization in knowledge discovery**. Proc. 1st Int. Conf. Knowledge Discovery and Data Mining (KDD-95), 198-203. AAAI, 1995.

(Levine et al., 2000) Levine, David M.; Berenson, Mark L.; Stephan, David. **Estatística: Teoria e Aplicações Usando Microsoft Excel** (Tradução). Livros Técnicos e Científicos Editora – 2000. Pg 580-620.

(Li, 1964) LI, C.C. **Introduction to Experimental Statistics**. McGraw Hill, Inc. 1964.

(Machado, 2000) Machado, Felipe N. R. **Projeto de Data Warehouse – Uma Visão Multidimensional**. Editora Érica. 2000. 248 p.

(Manly, 1986) Manly, B. F. J., **Multivariate Statistical Methods**. A Primer. 1 ed. London, Chapman & Hall, 1986.

(Mannino et al., 1988) M.V. Mannino, P. Chu and T. Sager, **Statistical Profile Estimation in Database Systems**. Computing Surveys 20, 3 (Sep. 1988), 191-221.

(Marques et al., 2003a) Marques, Aldaléa L. Brandes; Cárdias, Henrique T. Castro; dos Santos, Antonio Araújo. **Relatório Mensal de Coleta e Análise de Combustíveis Automotivos no Estado do Maranhão - RELATÓRIO TÉCNICO Nº 06/03 Junho de 2003.** Convênio ANP/UFMA Contrato Nº 4.067/01-ANP-009.215 de 27 de Agosto DE 2001.

(Marques et al., 2003b) Marques, Delano Brandes; Labidi, Sofiane. **Sistema De Apoio a Tomada de Decisão Em Monitoramento e Controle da Qualidade de Combustível**. XII ENQA – Encontro Nacional de Química Analítica. Workshop Petróleo. São Luis (UFMA), 14 à 17 de outubro de 2003.

(Melo, 1997) Melo, Rubens Nascimento. **Data Warehousing (Tutorial)**. In: XIII SBBD - Simpósio MEBrasileiro de Banco de Dados, Salvador, 1997.

(META Group,1996) META Group. **How to Build An Effective Data Warehouse**. Industry Overview: New Insights in Data Warehousing Solutions. Information Week, outubro de 1996, 1-27HP.

(Mitchell, 1998) Mitchell, M. (1998). **An Introduction to Genetic Algorithms**. MIT Press.

(Monteiro et al., 2000) Monteiro, R. C.; Bernardes, E.V.; Masson, M.R. & Landim, P.M.B. **Análise estatística multivariada para materiais cerâmicos**. VIII Simp.Quant.Geociências, Bol.Res. Expandidos, 163-166. 2000

(Nilsson, 1980) N. J. Nilsson. **Principles of Artificial Intelligence**. Palo Alto, CA: Tioga, 1980.

(Oneil, 1997) Oneil, B. **Oracle Data Warehousing**. Indianapolis, Sams Publishing, 1997.

(Orr, 1996) Orr, Ken. **Data Warehousing Technology**. The Ken Orr Institute, A white paper, 1996. Disponível em http://www.kenorrinst.com/dwpaper.html.

(OWG, 2000) OWG (2000). **Data mining. OWG - Smart Business. Smart Solutions**. Disponível on-line em: http://www.dwbrasil.com.br/html/dmining.html.

(Piatetsky,1991) G. Piatetsky-Shapiro. **Knowledge discovery in real databases: A report on the IJCAI-89 Workshop**. AI Magazine, Vol. 11, No. 5, Jan. 1991, Special issue, 68-70.

(Piatetsky et al., 1991) W. Frawley, G. Piatetsky-Shapiro, and C. Matheus. **Knowledge discovery in databases: an overview**. In G. Piatetsky-Shapiro and W. Frawley, editors, Knowledge Discovery in Databases, pages 1--27. Cambridge, MA: MIT Press, 1991.

(Poe, 1996) Poe,V. **Building a Data Warehousing for Decision Suport**. Prentice-Hall, 1996.

(Reyment et al., 1996) Reyment, R. A. & Jöreskog, K. G. **Applied Factor Analysis in the Natural Sciences**. Cambridge University Press, second printing. 1996.

(Rumelhart et al., 1986) D. Rumelhart and McClelland. (Eds.) **Parallel Distributed Processing: Explorations in the Microstructure of Cognition**. Cambridge, MA: MIT Press, 1986.

(Shavlik , 1990) J. W. Shavlik and T. G. Diettrich. (Eds.) **Readings in Machine** Learning. San Mateo, CA: Morgan Kaufmann, 1990.

(Schuchardt et al., 2001) Schuchardt, Ulf, Ribeiro, Marcelo L. & Gonçalves, Adilson R. **A indústria petroquímica no próximo século: como substituir o petróleo como matéria-prima?**. Quím. Nova, Abr 2001, vol.24, no.2, p.247-251. ISSN 0100-4042.

(Singh, 1997) Singh, Harry. **Data Warehousing: Concepts, Technologies, Implementations, and Management**. Upper Saddle River, NJ: Prentice Hall, 1997.

(Tafner at al., 1996) Tafner, M. A., de Xerez, M., and Filho, I. W. R. **Redes Neurais Artificiais: Introdução e Princípios de Neurocomputação**. Blumenau: EKO: Editora da FURB, 1996. 199p.

(Taurion, 1997) Taurion, C. **Data Warehouse: Estado de Arte e Estado de Prática**. Developers' Magazine,1997. ano 1, n. 6, p. 10-11, fev.

(Tronchin, 1998) Tronchin, Valsoir. **Análise, Modelagem e Implementação de Data Warehouses**. São Paulo: Fenasoft/98 em 20/07/98.

(Vasconcelos, 2002) Vasconcelos, Benitz de Souza. **Mineração de Regras de Classificação com Sistemas de Bancos de Dados Objeto-Relacional. Estudo de Caso: Regras de Classificação de Litofácies de Poços de Petróleo**. Dissertação de Mestrado em Ciência da Computação – UFPB. 2002.

(Wright, 1998) Wright, Peggy. **Knowledge Discovery In Databases: Tools and Techniques**. Work is funded by U.S. Army Corps Engineers Waterways Experiment Station,Vicksburg,EUA. 1998. Disponível em: http://www.acm.org/crossroads/doc/indices/features.html#Artificial%20Intelligence

(Zhou, 1989) Zhou, D. **ROPCA: A FORTRAN Program for Robust Principal Components Analysis**. Computers & Geosciences, 15:59-78. 1989

URLs

(URL 1) Página de artigos da IBM. 2003. URL: http://www-919.ibm.com/developer/db2/documents/star/star2.html

(URL 2) Página da Agência Nacional do Petróleo. 2003. URL: http://www.anp.gov.br

(URL 3) Plano Nacional de Ciência e Tecnologia de Petróelo e Gás Natural – CTPETRO. 2004. URL: http://www.anp.gov.br/NDT/index.htm

(URL 4) Página sobre o Programa de Capacitação e Assistência Técnica a laboratórios da rede nacional de laboratórios de ensaios para o monitoramento da qualidade de combustíveis. 2004. URL: http://www.seminariogestao.ufsc.br/arquivos/anais/Arruda/Biblion-DesenvolvimentoImplementa%E7%E3oDeUmaFerramenta.pdf

(URL 5) Página da Microsoft. 2004. URL: http://www.microsoft.com

(URL 6) Página da Minitab. 2004. URL: http://www.minitab.com

(URL 7) Página da Statsoft. 2004. URL: http://www.statsoft.com

(URL 8) Página da Mathworks. 2004. URL: http://www.mathworks.com

(URL 9) Página da Camo. 2004. URL: http://www.camo.no

(URL 10) Página da Autodesk. 2004. URL: http://www.autodesk.com