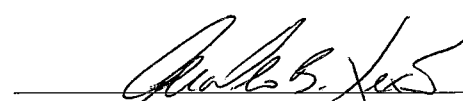
CONTROLE DE QUALIDADE EM BANCO DE DADOS

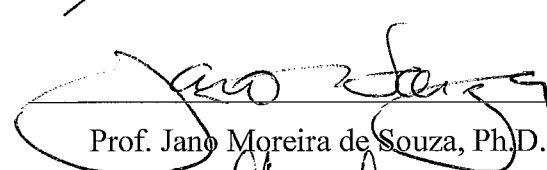
Flávio de Brito Pinheiro

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO

Aprovada por:



Prof. Geraldo Bonorino Xexéo, Ph.D.



Prof. Jano Moreira de Souza, Ph.D.

Prof. José Antônio Moreira Xexéo, Ph.D.

Profª. Vera Maria Benjamin Werneck, Ph.D.

RIO DE JANEIRO, RJ – BRASIL

MARÇO DE 2002

PINHEIRO, FLÁVIO DE BRITO

Controle de Qualidade em Banco de Dados [Rio de Janeiro] 2002

VI, 97 – p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia de Sistemas e Computação, 2001)

Tese – Universidade Federal do Rio de Janeiro, COPPE

1. Qualidade de Dados
2. Banco de Dados

I. COPPE/UFRJ II.Título (série)

Dedicatória

A Deus, minha esposa Andréa e meu filho Guilherme que sempre me apoiaram nos momentos bons e naqueles que me faltava esperança.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

CONTROLE DE QUALIDADE EM BANCOS DE DADOS

Flávio de Brito Pinheiro

Março / 2002

Orientadores: Geraldo Bonorino Xexéo
Jano Moreira de Souza
Programa: Engenharia de Sistemas e Computação

Este trabalho implementa um sistema que utiliza os conceitos de Qualidade de Dados associados ao Controle Estatístico de Processos e a Banco de Dados, para manipulação dos dados em SGBDRS.

Ele permite que a qualidade dos dados seja analisada por pessoas sem qualificação técnica específica, utilizando tutores especializados, que apóiam a navegação do usuário pelo sistema, de forma amigável, robusta e de fácil compreensão, fornecendo orientações técnicas a respeito do nível de qualidade encontrado nos Bancos de Dados.

Propomos também, um modelo capaz de manipular, de forma eficiente e robusta, as incertezas ligadas aos dados em diversos tipos de SGBDRS.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

QUALITY CONTROL IN DATABASE SYSTEMS

Flávio de Brito Pinheiro

March / 2002

Advisors: Geraldo Bonorino Xexéo
Jano Moreira de Souza
Department: Computing and Systems Engineering

This work implements a system that uses the concepts of Data Quality associate with Statistical Process Control and Database Systems to manipulate information in RDBMS.

It permits that data be manipulated by persons without specific technical qualification, using specialized Wizards, that support user navigation through the system in a strong, easy and friendly way, supplying technical orientation about the level of quality found in Databases.

We also propose a capable model to manipulate , in a strong and efficient way, the uncertainty on data in many RDBMS types.

Índice

Capítulo 1 - Introdução

O homem que parava para admirar a natureza e dela extrair regras a respeito do tempo, hoje procura através de grandes bancos de dados aprender mais e mais, com fenômenos cada vez mais complexos.

Apesar dos benefícios da disseminação crescente e sem volta da informática é lugar comum vermos exemplos de excesso de dados e informação armazenados em mídia digital que poderiam ser utilizados para auxiliar, no processo de planejamento, escolas, hospitais e empresas e não são utilizados. Mesmo assim, outro ponto importante é a confiabilidade dos dados armazenados. Será que os dados que estão armazenados têm confiabilidade para serem extrapolados da amostra para a "população"? Acreditamos que não.

A exploração estatística dos dados fornece um substrato capaz de apontar possíveis erros e, mais, mostra-nos a possibilidade da investigação e aprendizado contínuo do fenômeno estudado. Assim sendo, dominando as diversas técnicas de análise de dados, podemos começar a tecer maiores comentários a respeito deste ou de outro fenômeno.

Mas como podemos explorar estes dados? Até onde podemos confiar neles? Só especialistas "falam" e "entendem" os dados?

1.1. Motivação

Em um país como o Brasil, carente de informações sociais, fica cada vez mais difícil a justiça social quando não temos elementos para que organismos de defesa social possam acompanhar as ações governamentais em prol da população. Enquanto os países que mais carecem de informação a renegam como um meio de democratizar a informação e a melhoria da formação social, os países ditos desenvolvidos fazem dela bom uso para melhorar a vida de seus cidadãos.

Esta tese pretende contribuir , fornecendo uma ferramenta, mesmo que embrionária, que possa facilitar o acesso à informação, orientar a análise da qualidade dos dados e suas

críticas de consistência, possibilitando ao não-especialista ensaiar algumas análises que até então, mesmo que básicas, fogem a compreensão dos que realizam estes estudos.

O objetivo deste nosso trabalho é consorciar técnicas de CEQ (*Controle Estatístico de Qualidade*) e CEP (*Controle Estatístico de Processos*) à área de Banco de Dados. Estas técnicas difundidas por Deming (DEMING, 1986) e Juran (JURAN, 1991) são de largo uso e aplicação. Contudo, como vimos durante o levantamento bibliográfico, poucos textos e poucas empresas abordam este assunto. Nosso objetivo é contribuir, adaptando de forma clara estas técnicas, associadas a ferramentas capazes de prover o usuário do dado ou seu administrador de um controle efetivo do que está sendo produzido.

Neste capítulo iremos abordar alguns estudos de caso onde coletamos informações a respeito das dificuldades apresentadas pelos usuários e administradores de dados, apresentando ao final dos estudos, um resumo das principais necessidades levantadas. Já no capítulo 2 apresentaremos as tendências e trabalhos elaborados pelos cientistas do setor, a fim de consolidarmos nossa opinião quanto às técnicas que serão apresentadas, apresentadas no Capítulo 3. No Capítulo 4, a arquitetura proposta para suportar a aplicação descrita no Capítulo 5 será apresentada e cada um de seus módulos terão uma análise detalhada sob os aspectos da funcionalidade, integração e expansão.

1.2. Estudo de Casos

Durante a análise dos problemas de qualidade de dados, identificamos alguns casos que exemplificam a necessidade da construção e adaptação de técnicas de QD (Qualidade de Dados), possibilitando ao usuário uma compreensão facilitada dos seus dados. Vários sistemas carecem de controles especializados, por parte de peritos, pois necessitam que seus operadores possuam formação acadêmica específica a fim de operá-los. Isto se deve a profunda complexidade dos métodos utilizados por estes.

Realizamos alguns estudos de casos que validaram a necessidade da democratização do acesso ao dado e a falta de especialistas, tornando quase que obrigatória de busca por soluções mais adequadas para estes problemas. Selecionamos três casos para pesquisa: um instituto de pesquisa com grande volume de dados, neste caso o IBGE, um órgão público de saúde – SUS e por fim um pesquisador. Através desta estrutura poderemos perceber como a qualidade do dado influencia, de forma equivalente, segmentos diferentes da sociedade. O primeiro estudo de casos que iremos abordar é do Instituto Brasileiro de Geografia e Estatística (IBGE), analisando seus aspectos. Em seguida apresentaremos o caso SUS – Sistema Único de Saúde e por fim, apresentaremos um caso na área científica.

1.2.1. IBGE

Descrição Geral

Nosso primeiro estudo foi realizado junto ao IBGE (Instituto Brasileiro de Geografia e Estatística), órgão este regulamentador e principal instrumento de análise e produção de estatísticas brasileiras. O caso IBGE começa com o número cada vez mais reduzido de especialistas estatísticos no seu quadro, devido a falta de renovação e aos constantes pedidos de aposentadoria. Desta forma, a cada novo projeto desenvolvido pela entidade, os especialistas alocados devem ser redistribuídos entre o projeto novo e os já existentes. Assim sendo, foram avaliados três tipos de pesquisas - as de curto prazo, as de médio e as de longo prazo. As pesquisas classificadas como curto prazo mobilizam

bases alimentadas com dados oriundos de coletas feitas pelas unidades do IBGE espalhadas pelo país, que posteriormente são consolidadas em bases maiores para futuras análises.

Durante nosso levantamento, observamos que o preenchimento das fichas de cadastro das pesquisas, seja manual (papel) ou automatizado (disquetes), não é feito por pessoal treinado pelo Órgão e , no caso da indústria, por aqueles diretamente envolvidos com a informação da produção industrial, dificultando assim a precisão das respostas obtidas e muitas vezes obrigando a ida de um entrevistador, na tentativa de correção dos dados preenchidos. Outro fato que nos preocupou neste estudo foi a baixa escolaridade daqueles que realizam os trabalhos nos setores de produção das pesquisas. Geralmente, os grupos de pesquisas são divididos em direção, técnicos e produção[1]. Assim, temos uma pirâmide com vértice superior muito afinado, que corresponde à hierarquia do poder e conseqüentemente à dificuldade na democratização das ações, ou seja, a tomada de decisão é completamente verticalizada, obrigando a produção, em muitos casos, a esperar os estudos da equipe técnica para a tomada de decisão a respeito dos casos ditos problemáticos. Este modelo de trabalho seria oneroso em uma empresa privada. Contudo, a falta de renovação nos quadros do IBGE verticaliza cada vez mais a tomada de decisões, pois existem, a cada dia, menos especialistas e mais pesquisas que demandam ações descentralizadas. Como se trata de uma empresa estatal, o seu modelo acaba estável – pois não há competidores na geração destas informações.

Para dinamizar suas ações e tentar descentralizá-las ao máximo, sob pena de não poder realizar contratações de pessoal efetivo, o IBGE hoje com a falta de material humano, necessita mudar de modelo produtivo, automatizando e informatizando ao máximo suas frentes de trabalho com o intuito de garantir a qualidade dos dados e minimizar possíveis erros no processo produtivo.

Dados

Em um processo de tabulação de pesquisas, se todos os questionários estivessem completos e consistentes seria desnecessário o processamento de crítica destes dados.

[1] esta divisão é bem simplificada, para não nos determos em especificidades de cada uma das pesquisas

Pela experiência, isto não ocorre. Por melhor que seja planejada, uma pesquisa sempre irá demandar um processo de crítica de dados devido a questionários incompletos, errados ou inconsistentes.

Percebemos que estes fatos vêm sendo estudados ao longo dos anos a fim de minimizar os impactos dos erros nas estimativas calculadas.

Usualmente ao falarmos em erro nas pesquisas de campo, alguns podem pensar em falha ou equívoco, mas em Estatística o sentido de erro é mais objetivo. Isto é, erro é a diferença entre a estimativa e o verdadeiro valor populacional que se deseja estimar.

O erro amostral permite construir um intervalo (chamado de intervalo de confiança), que estabelece limites em torno da estimativa obtida, e afirmar que, com determinado nível de confiança, o valor real da população está nele contido.

Pode-se estabelecer níveis de confiança maiores ou menores, dependendo do grau de certeza que se deseja ter sobre os resultados. Quando se diz que o nível de confiança da pesquisa é de 95%, significa que se 100 amostras forem tiradas da população em estudo, em 95 delas os resultados ficariam dentro do intervalo calculado com os dados obtidos na única pesquisa realizada. Em suma, há uma probabilidade conhecida de que o valor verdadeiro da população esteja contido nesse intervalo. O nível de confiança mais usual é o de 95%, mas em alguns casos, é necessário que se trabalhe até com limite de 99%.

Nota-se que nas pesquisas podemos considerar dois aspectos referentes aos erros não-amostrais (erros que não podem ser calculados, mas podem ser controlados e minimizados):

i. Erros de cobertura - cuja ocorrência se dá quando há falha de cobertura nas unidades de pesquisa.

ii. Erros de conteúdo - cuja ocorrência se dá na inconsistência e/ou erros de preenchimento das informações.

iii. Erros de não-resposta - relacionado à recusa no fornecimento ou na impossibilidade de se obter a informação, mesmo que parcial ou totalmente.

Estes erros provocam grandes impactos na pesquisa realizada. Objetivando minimizar estes problemas, torna-se necessário à criação e aplicação de métodos para se assegurar o menor impacto possível das informações "ruidosas" sobre as estimativas calculadas.

Pela nossa observação, durante as pesquisas a maior preocupação é verificar a falta de aderência das informações às estimativas calculadas, preservando a qualidade da informação, fazendo com que os dados sejam mais próximos do valor verdadeiro à ser medido.

Pesquisa

Avaliamos a PIA - Produto (Pesquisa Industrial Anual de Produtos) que forma, em conjunto com a PIA - Empresa, o núcleo central das estatísticas industriais. Além de proporcionarem uma gama ampla de informações importantes em si mesmas, estas pesquisas são a base de referência para o desenho de pesquisas conjunturais (como a PIM-PF e a PIM-DG) e de pesquisas satélites.

O objetivo da PIA - Produto é gerar informações, de valor e quantidade, dos produtos e serviços industriais produzidos e/ou vendidos em determinado ano, tendo 1998 como ponto inicial da série.

Por se tratarem de informações de produção: quantidade produzida, quantidade vendida, valor de venda e outras, um filtro crítico no sistema local (disquete) não conseguiria parâmetros suficientes para compararmos os dados digitados com os valores apresentados por outras empresas do mesmo segmento. Além do mais, cada empresa está inserida na sua realidade demográfica, dificultando sobremaneira uma comparação simplesmente linear. Embora estes fatos fossem de conhecimento dos especialistas, algumas situações dificultavam intensamente o desenvolvimento de curto prazo da solução.

Sendo uma pesquisa nova e de curto prazo, não havia outra referência para servir de modelo temporal comparativo às demais, isto é, somente depois da primeira, novas técnicas de análise temporal e sazonal poderiam ser empregadas. Além disso, como o número de especialistas é pequeno para a demanda crescente, caberia à equipe de produção o controle da filtragem e análise das situações de erro, embora a maioria das

pessoas que desenvolveria este trabalho não possuía qualificativo técnico para se tornarem independentes da figura do especialista.

Visto desta forma, alguns poderiam pensar que o movimento de terceirização poderia em parte resolver o problema, mas normalmente a contratação de pessoal para os projetos de curto prazo não possui este enfoque. A solução, a princípio, poderia estar em uma ferramenta de análise da qualidade dos dados a fim de permitir que não-especialistas pudessem usa-la de forma mais intuitiva, visto que o IBGE já possuía ferramentas altamente especializadas que deixaram de ser usadas devido à falta de material humano capacitado para sua operação.

1.2.2. SUS – Sistema Único de Saúde

Descrição Geral

"18:22 30/10/2001
Agência JB

MPF investiga fraude milionária no SUS no Amapá

AMAPÁ - MPF investiga uma fraude milionária no Sistema Único de Saúde (SUS) do Amapá; já foram detectados desvios de R$ 1,5 milhão, e pelo menos um funcionário da Secretaria de Estado da Saúde, que tem salário de apenas R$ 500, já foi identificado por fraudar o sistema informatizado de pagamento do governo federal para o sistema de saúde"

"Jornal do Commercio
Recife - 12.07.2000
Quarta-feira

Saúde e PF investigam fraudes no SUS

BRASÍLIA – O Ministério da Saúde investiga 4.530 novos casos de fraude no Sistema Único de Saúde (SUS) e pelo menos dois deles foram encaminhados ontem à Polícia Federal pelo ministro José Serra. Entre os casos já descobertos está o da Casa de Saúde Miguel Couto, em Nova Iguaçu, na Baixada

Fluminense. Pessoas que foram ao hospital procurar emprego tiveram os dados de seus currículos usados para forjar internações."

Na área de saúde temos o exemplo do SUS – que depende fortemente de parâmetros coletados nos diversos postos de saúde e hospitais por todo o Brasil, a fim de serem tomadas medidas objetivas na condução do processo de saúde do povo brasileiro. Neste caso, a sede do SUS em Brasília – DF, solicita aos estados mensalmente os dados relativos a cada centro de saúde e no caso dos municípios, os PAMs e hospitais ligados à rede SUS. Estes devem informar não só os gastos contábeis, como também o número de internações, patologias e o perfil demográfico dos pacientes, assim como o nível de absenteísmo apresentado. Uma outra preocupação dos centros médicos é a ausência de continuidade dos tratamentos, neste caso, provocando gastos desnecessários, resultando em um atendimento de baixa qualidade do ponto de vista social.

Um dos maiores problemas apresentados é o fato da não haver comunicabilidade intra-centros e entre centros, isto é, não há estrutura de redes de computadores ampla e definida para a automação de tarefas capazes de estabelecer um controle sobre os gastos e ações a serem tomadas de acordo com as mudanças no perfil dos usuários e centros médicos, acarretando re-trabalhos e tomadas de ações com base em dados que nem sempre correspondem à verdade, visto que por vezes, sofrem interpretações erradas por falta de estrutura , informação e normatização do processo de coleta e definição de necessidades.

Os poucos centros que possuem alguma infra-estrutura de informática muitas vezes necessitam de ferramental para a análise da qualidade dos dados, pois por esta não ser sua atividade fim, não possuem especialistas para o controle de qualidade de dados.

Dados

Os dados oriundos dos processos clínicos são coletados por diversos profissionais o que, pôde-se perceber, multiplica as chances de erros em todo o processo. Além disso, a tarefa de coleta de dados nem sempre passa pela mão do profissional médico, isto é, existe uma triagem que define os primeiros dados a serem coletados, realizados por pessoal sem formação em saúde.

Durante este processo, pôde-se perceber que os casos do tipo, erro de idade, peso, altura, patologia clínica entre outros, não podem ser revisados completamente devido a falta de material humano para esta atividade.

Durante nosso levantamento foi apontada a necessidade de um sistema de controle de qualidade de dados coletados e armazenados para que não só o SUS, através do DATASUS (órgão de informática do SUS), tivesse uma real imagem da estrutura de atendimento hospitalar, bem como o próprio PAM, para controlar sua demanda de serviços.

O caso estudado foi no departamento de reabilitação motora que atende cerca de 200 pacientes semanalmente e sofre com o absenteísmo dos pacientes e a descontinuidade do processo terapêutico, resultando em anamneses nem sempre precisas ou conclusivas. Mensalmente cabe a este setor informar o número de tratamentos realizados, tempo de duração médio e quantos estão ausentes ou abandonaram definitivamente o processo terapêutico. O maior problema relatado é a falta de recursos de sistemas capazes de possibilitar a mineração dos dados e o seu controle junto à fonte primária destes. Uma ferramenta genérica, independente do contexto abordado, poderia ser de grande ajuda no processo de controle da qualidade da informação divulgada por estes órgãos.

1.2.3. Ensaios Acadêmicos

Descrição Geral

O estudo do tempo de reação realizado pelo professor Nei Calvano Gonçalves, do Curso de Psicologia da UFRJ(comunicação pessoal), analisa a percepção, discriminação e escolha de estímulos variados que podem ser obtidos quantitativamente através do tempo de reação. Podemos definir tempo de reação como o intervalo de tempo que transcorre entre a apresentação de um estímulo e o início de uma resposta dada por um organismo, ao qual se apresenta tal estímulo.

A utilização de estímulos visuais e auditivos, em pesquisas, sobre tempo de reação é freqüente porque permite um melhor controle experimental. Outros estímulos sensoriais podem ser usados desde que se encontrem soluções técnicas para assegurar que o organismo está respondendo ao estímulo em questão. Os estímulos olfativos e gustativos apresentam maior dificuldade para controle experimental.

O tempo de reação pode variar em função das características do estímulo (quantidade, intensidade, tipo, ritmo de apresentação, etc.) e das características do organismo (aprendizagem, atenção, idade, cansaço, etc.).

Dados

O trabalho da equipe do professor Nei, na realidade, é desenvolver um software que possa substituir os equipamentos existentes, que são os cronoscópios e/ou cronômetros de alta precisão, no sentido de, através do uso de computador, este possa efetuar as mesmas tarefas da medida de tempo de reação e simultaneamente, tratar os dados estatisticamente produzidos pelos sujeitos experimentais.

i. Este "software" controlará os estímulos dos eventos e as medidas de Tempo de Reação a estímulos luminosos e sonoros para serem utilizados em laboratórios de Psicologia Experimental pelos alunos de graduação em aulas práticas específicas de Percepção e Sensação, permitindo também a utilização deste como complemento às variáveis dinâmicas do comportamento humano.

1.2.4. Conclusão dos Estudos

Durante o nosso estudo no IBGE, constatamos que os problemas de pessoal e de qualificação da mão-de-obra deverão se agravar, visto que a demanda por novos dados de qualidade e mais informações deverão conduzir o IBGE a adotar uma nova política e o uso de sistemas especialistas, onde o usuário poderá interferir no processo sem a necessidade de uma alta qualificação em métodos estatísticos.

Com relação ao caso SUS o Governo poderia economizar recursos e diminuir o afastamento de pacientes e fraudes caso houvesse mecanismos de fiscalização exercidos

pela sociedade, garantindo que os dados possuíssem a qualificação desejada, e ainda poder-se-ia adotar formas de controle de produtividade que realmente viessem refletir a realidade da saúde brasileira.

Estes estudos apontam para uma tendência que se verifica a cada dia: a necessidade de ferramentas automatizadas inteligentes para o controle de qualidade da produção de dados, como no caso dos ensaios acadêmicos, visto que o foco é a pesquisa por parte do pesquisador e não o desenvolvimento de uma ferramenta de análise de dados. Estas ferramentas serão de grande valia ao processo produtivo nos diversos segmentos de mercado e produção do conhecimento, visto que o uso em larga escala de SGBDRs e o desenvolvimento de sistemas sem o controle da qualidade dos dados é crescente, dificultando sobremaneira os trabalhos de exploração de dados realizados tanto por analistas quanto por investigadores que na esperança de obter vários cenários com a mesma base de dados por muitas vezes não detêm conhecimento ou tempo suficientes para a limpeza e adequação das bases de dados disponíveis. Estes fatos se transformam de preocupação em necessidade de se possuir sistemas inteligentes que possam apontar eventuais problemas e até propor novas formas de análise a fim de minimizar os problemas de qualidade de dados. Identificamos após estes estudos de caso, que nossa ferramenta deverá contemplar as seguintes características desejadas pelos usuários:

i. Facilidade de aprendizado/operação
ii. Clareza na interface homem/máquina
iii. Análise de Metadados
iv. Tutores amigáveis para a exploração de dados
v. Técnicas compreensíveis de análise exploratória
vi. Análise fortemente baseada em gráficos
vii. Análise de regras de negócios
viii. Versatilidade no acesso as bases de dados
ix. Tratamento de erros
x. Visualização gráfica dos dados de forma rápida

Capítulo 2 - Qualidade de Dados

O Que é Qualidade?

De acordo com o dicionário Aurélio (FERREIRA, 1985) – Qualidade é...

i. "Propriedade, atributo ou condição das coisas ou das pessoas capaz de distingui-las das outras e de lhes determinar a natureza",
ii. "Numa escala de valores, qualidade permite avaliar e, conseqüentemente, aprovar, aceitar ou recusar, qualquer coisa:",

Já a ISO 9000:2000 (ABNT, 2000) a define como

iii. "... o grau no qual um conjunto de características inerentes satisfaz a requisitos"

De acordo com o terceiro conceito, poderíamos apontar a Qualidade dos Dados como sendo uma medida de concordância entre as visões do dado, apresentada pelos sistemas de informação e o mesmo dado no mundo real.

A este respeito, Laudon (LAUDON, 1986), descreveu a dependência da sociedade de Sistemas de Informação e examina a qualidade dos dados sobre a ótica de um Sistema Inter-organizacional.

O problema de manter registros de alta qualidade em um sistema de informações é magnificado ao tratar-se de um Sistema Inter-organizacional. Alguns Sistemas Inter-organizacionais permitem somente o compartilhamento de informações disponibilizadas voluntariamente, como em um boletim eletrônico (BBS). No exemplo estudado por Laudon, não existia autoridade centralizadora ou gerencial, poucos padrões para o conteúdo e formato da informação, regras para incentivar reclamações a respeito das regras do sistema e pouca precisão.

Outros Sistemas Inter-organizacionais são baseados no modelo de uma biblioteca central, onde existe um forte controle sobre a entrada de dados e gerenciamento ativo dos arquivos, fortes incentivos aos participantes para seguirem as regras do sistema e sanções para o não cumprimento destas.

Os sistemas também diferem em termos de sua relação com o público. Alguns influenciam criticamente na vida humana, outros envolvem conveniências individuais.

Pode-se notar que os sistemas podem variar em termos da sua "visibilidade" e abertura ao público. Um sistema de reserva aérea, por exemplo, opera de forma que uma informação errônea é avistada facilmente, mas sistemas de registros criminais, por exemplo, não são igualmente tão "visíveis" para os indivíduos envolvidos – Quais são as informações registradas nos arquivos e quem as usa?

Laudon,(LAUDON, 1986) comenta ainda, que existem três métodos geralmente utilizados para examinar a qualidade dos dados em grandes bancos de dados:

i. Entrevista com usuários finais do sistema ou clientes,
ii. Amostragem de registros e
iii. Amostras de casos ativos ou correntes.

As pesquisas com usuários finais e clientes, tipicamente, medem a percepção sobre a qualidade dos dados e são carregadas de problemas de "lembrança"(LAUDON,1986), enquanto que as pesquisas em arquivos são realizadas por grandes empresas, contendo também erros inerentes, pois muitas pessoas cadastradas nos arquivos não estão mais "ativas", outras faleceram. Assim sendo, o método irá falhar na contagem de pessoas para terem um determinado benefício, por exemplo. A solução encontrada pelo autor foi para cada sistema examinado, tomar uma amostra de casos correntes e os registros comparados com a cópia impressa.

No tocante às dimensões da qualidade dos dados, Laudon identificou as seguintes dimensões da qualidade dos registros em sistemas de registros criminais:

i. **Registros Incompletos** – alguns registros não continham todas as informações e decisões apresentadas.
ii. **Registros Incorretos** – Os registros são incorretos quando as informações apresentadas não conferem com as informações contidas na base de dados
iii. **Registros Ambíguos** – Muitas alterações descaracterizando o perfil do dado – ex.: datas que não correspondem

Desta forma, foi sugerida uma fórmula objetivando tornar possível a realização de inferências estatísticas, com base nas amostras coletadas, para a população. O intervalo de confiança (LAUDON, 1986), é dado por:

$$ic = p \pm 1{,}96 \times \sqrt{((p \times q) / n)}$$

Onde

p = a proporção observada na amostra

q =1 –p

1,96 = valor de z (medida de erro padrão) apropriado ao intervalo de confiança de 95%

n= Número de respostas para uma dada medida

Com base nos dados coletados, Laudon (LAUDON, 1986) recomendou a criação de sistemas de controle de dados incompletos e que auditorias periódicas fossem feitas nas bases de dados (BD).

Wang (WANG, 2000) envereda por outro ponto de vista, afirmando que com relação aos problemas de Qualidade de Dados (QD), existe uma preocupação com grandes Sistemas de Informação (WANG, 2000). Assim, ele propõe uma conceitualização a respeito do tema qualidade de dados.

As pesquisas sobre Banco de Dados (BD) têm o propósito de assegurar a qualidade do dado dentro do Banco de Dados. Na área de QD existem pesquisas investigando as definições de QD (WANG e STRONG, 1996), modelagem (BALLOU *et al.*, 1996) e controle (LIEPINS *et al.*, 1990). Com poucas exceções, a QD é tratada como um conceito intrínseco, independente do contexto em que o dado é usado ou produzido. Este foco sobre os problemas de QD na armazenagem do dado falha em problemas, quando se trata de organizações mais complexas. Wang (WANG, 2000) atribui esta falha, em parte, à falta de uma ampla conceitualização a respeito da QD.

Contrastando com este ponto de vista, a qualidade não pode ser conseguida independentemente dos consumidores que escolhem ou usam os produtos. De maneira

semelhante, a qualidade do dado não pode ser conseguida independentemente das pessoas que irão utilizar o dado – consumidores de dados.

Usando análises qualitativas, Wang (WANG e KON 1993) examinou projetos de QD de três grandes empresas e identificou problemas comuns com relação ao tema. Esses padrões apareceram devido a utilização de uma abordagem conceitual sobre QD. Desta forma foi possível fazer recomendações aos profissionais de SI a fim de melhorar a qualidade do dado pela perspectiva dos seus consumidores.

Poderíamos olhar o conceito de transformação/produção de dados em informação como um sistema de manufatura onde a matéria prima seria do dado e no qual identificaríamos três perfis: gerador (pessoas, grupos ou outras fontes que geram dados); guardião (pessoas que fornecem e gerenciam computacionalmente recursos para armazenagem e processamento dos mesmos) e consumidores de dados (pessoas ou grupos que utilizam os dados) (BALOU, *et al.*, 1996) (WANG e KON, 1993). Cada perfil é associado a um processo ou uma tarefa: produtores de dados são associados com processos de produção de dados; guardião dos dados com armazenagem de dados, manutenção e segurança e consumidores de dados com os processos de utilização dos dados que podem envolver ainda outros processos, de agregação e de integração.

Desta forma, definiu-se dado de alta qualidade como aquele que atende ao uso dos consumidores de dados. Isto significa que a capacidade de uso do dado e a capacidade de seu uso integral são aspectos importantes na qualidade. Usando estas definições, Wang (WAND e WANG, 1996) propôs uma tabela contendo as características e as quatro categorias na definição de dados de alta qualidade: intrínseca, de acessibilidade, de contexto e aspectos representacionais.

QD –Categoria	QD – Dimensões
QD Intrínseca	Exatidão (Precisão), Objetividade, Confiança e Reputação
QD Acessibilidade	Acessibilidade, Acesso Seguro
QD Contextual	Relevância, Valor Agregado, Temporalidade, Totalidade e Quantidade de Dados
QD Respresentacional	Interpretabilidade, Facilidade de Entendimento, Representação Concisa e Consistente

Wang (WAND e WANG, 1996) define o problema de QD como sendo qualquer dificuldade encontrada ao longo de uma ou mais dimensões da qualidade e que tornam o dado completamente, ou em grande parte, não adequado para o uso. Além disso, quando se trata de um projeto de QD, tem-se também em mente ações organizacionais a fim de resolver o problema de QD visto o reconhecimento da baixa qualidade dos dados produzidos pela organização.

Após um levantamento feito em 42 organizações, Wang (WAND e WANG, 1996) observou os seguintes padrões:

QD Intrínseco

Desigualdades em torno de dados da mesma fonte são as causas mais comuns a respeito de QD Intrínseco. A princípio, os consumidores de dados não conhecem a origem dos dados a fim de lhe atribuir problemas de qualidade. Eles sabem somente que o dado é conflitante. Assim, esta preocupação inicialmente parece estar relacionada com problemas de confiabilidade. Com o passar do tempo, a informação a respeito das causas das desigualdades irá acumular, a partir de uma avaliação da exatidão das diferentes origens, conduzindo a uma reputação ruim as origens desses dados. Assim, quando é senso comum que a reputação da origem dos dados é ruim, estas origens passam a não possuir mais valor agregado para a organização, trazendo como resultado o seu pouco uso ou o abandono.

QD Acessibilidade

Problemas de acessibilidade em QD são caracterizados pela: falta de preocupação com acessibilidade técnica (segurança no acesso ou acesso ruim – poucos recursos ou dados), representação dos dados interpretados pelos seus consumidores como sendo problemas de acessibilidade (necessidade da interpretação de um especialista e dificuldade no resumo das informações) e grande volume de dados (grande quantidade de dados para serem processados, muitos dados acumulados no tempo).

QD Contextual

Foram observadas três causas para as queixas dos usuários com respeito aos dados que não estão disponíveis às suas aplicações: ausentes ou incompletos, dados inadequadamente definidos ou medidos e dados que não podem ser agregados adequadamente.

A fim de resolver o problema contextual de QD, projetos específicos foram iniciados com o objetivo de fornecer dados relevantes que agreguem valor às tarefas dos seus usuários.

Concluiu-se que as pesquisas existentes mantêm seu foco nos aspectos intrínsecos de QD e elas falham em relação aos consumidores dos dados. Além da importância com relação aos aspectos intrínsecos, as organizações também começam a iniciar projetos endereçados à acessibilidade e contexto em QD. Acessibilidade em QD inclui a preocupação com a facilidade de acesso aos dados, sua compreensão e ainda, a visão contextual do assunto. O foco da preocupação está em quão fiéis são os dados em relação ao contexto.

Em recente discussão (WANG, 2000), além dos atributos de qualidade propostos (precisão, objetividade, confiança, acessibilidade, relevância, valor agregado, objetividade e entre outros), foi fortemente enfatizada a necessidade do gerenciamento de dados como um produto e estipulou-se que o processo de geração de dados para consumo pode ser gerenciado da mesma forma que a produção de qualquer outro produto. Isto vem confirmar as observações e tendências apresentadas neste trabalho desde o seu o primeiro estudo. Além disso, os conceitos de Gerenciamento Total da Qualidade (*TQM - Total Quality Management*) introduzidos por Deming (DEMING,

1986) e os princípios de Controle Estatístico de Processos podem ser aplicados ao processo de produção de dados como uma nova alternativa na busca do conhecimento da qualidade dos dados armazenados pelas empresas.

A estrutura de TQM proposta por ele foi baseada em "Definir, Medir, Analisar e Melhorar", uma nova maneira de ver o PDCA da Qualidade.

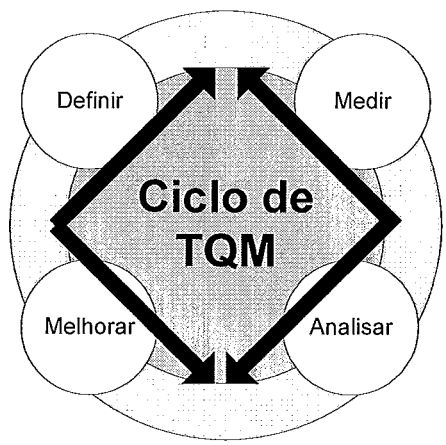


Fig. 1 - DMMA – Novo PDCA da Qualidade

Além disso, Wilks (WILKS, 2000) levanta a necessidade de sistemas de controle de qualidade de dados para grandes bases, a fim de minimizar as lacunas identificadas pelos usuários com relação aos dados ofertados. Ademais, ele coloca a necessidade de: 1) trazer o conhecimento do usuário para o problema; 2) A importância de uma resposta regular e sistemático dos usuários e 3) Manter todos os dados brutos e logs alinhados para que se possa traçar a origem das lacunas e explicar as anormalidades. Colaborando com a visão de minimizar os desvios de qualidade dos dados, Thornton (THORNTON e ANN, 2000) define a importância da qualidade da informação através do desenvolvimento de metadados de alta qualidade e dicionários de dados (*metadata quality*) como ponto central da discussão sobre o conhecimento da qualidade dos dados.

Larry (ENGLISH, 2000b) apresenta em seu estudo de caso que já estaríamos vivendo na era Industrial amadurecida, quando os processos de qualidade tais como Melhoria Contínua de Processos (MCP), TQM e Kaizen transformaram os processos de manufaturas, eliminando custos e re-trabalho. Agora, diz English (ENGLISH, 2000a), estamos vendo a maturidade da Era da Informação como o resultado da aplicabilidade

dos mesmos processos de qualidade aos produtos de informação – a nova moeda da economia moderna.

O que percebemos, através dos seus pensamentos, é que a fronteira econômica se dará pela velocidade de processamento e de entendimento na Era da Informação, simbolizando a necessidade da confiança nos dados e fontes.

Para Orr (ORR, 1998) o principal papel de um sistema de informação nesta Era é apresentar visões do mundo real para que as pessoas, na organização, possam criar produtos ou tomar decisões. Se essas visões não concordam substancialmente com o mundo real por um período de tempo qualquer, então este sistema é ruim e em último caso, como uma ilusão psicótica, a organização irá atuar irracionalmente.

Pela representação do modelo de Controle de Respostas (*Feedback-Control System – FCS*), torna-se mais fácil a definição de Qualidade de Dados. QD é uma medida de concordância entre as visões do dado, apresentada pelos sistemas de informação e o mesmo dado no mundo real. Sistemas com 100% de QD indicarão, por exemplo, que nossas visões a respeito dos dados estão em perfeita concordância com o mundo real; por outro lado, sistemas com 0% de QD mostram a total discordância.

Nenhum sistema de informação, realmente operacional, possui dados com 100% de qualidade. A preocupação com a qualidade do dado não é garantir que ela seja perfeita, mas sim, que a qualidade do dado no sistema de informação seja: exata o bastante, temporal o bastante e consistente o bastante para que a organização possa sobreviver e tomar decisões razoáveis.

Ultimamente, a maior dificuldade em QD é a mudança. O dado em nossos bancos de dados é estático, mas no mundo real ele se mantém sob constante mudança. Mesmo que o nosso sistema possuísse um BD com 100% de concordância com o mundo real no momento t_0, no momento t_1 ele poderia estar ligeiramente desatualizado e no momento t_2 ele poderia estar mais desatualizado. A teoria do FCS declara que, se um sistema pretende acompanhar o mundo real, deverá existir um mecanismo que sincronize o dado do sistema com as mudanças do mundo real – Uma resposta se faz é necessária.

Mas de onde vem esta resposta? A resposta clássica dos desenvolvedores de sistemas de informação é que o retorno sobre a qualidade do dados é de responsabilidade de seus usuários. Na prática, os desenvolvedores de IS constroem os sistemas que vêm somente ao encontro às exigências dos usuários. Eles dizem que é trabalho do usuário garantir que o dado no BD seja mantido de forma exata e temporal.

Os usuários, por outro lado, historicamente têm o sentimento que eles carregam a responsabilidade pela qualidade dos dados em um sistema de informação que eles não entendem. Sistemas que são difíceis de se fazer correções e sistemas em que resultam certos tipos alterações não aplicáveis.

Para Orr (ORR, 1998), duas coisas devem ocorrer em todo BD para estarem sincronizados com o mundo real:

i. Alguém ou algo (pessoas ou sensores automáticos) deve comparar as visões dos dados com o mundo real e
ii. Qualquer desvio destes deve ser corrigido e novamente digitado.

Ainda sim, freqüentemente, os desenvolvedores de sistemas possuem visões muito simplistas a respeito de como os sistemas são organizados. Para eles, os sistemas não passam de um processo de Entrada - Processamento e Saída (EPS) (*Input-Process-Output (IPO)*), vide figura 2. Porém, este tipo de abordagem falha quando se trata de contextos mais amplos, vide figura 3.

FIGURA 2
MODELO DE ENTRADA - PROCESSAMENTO - SAÍDA

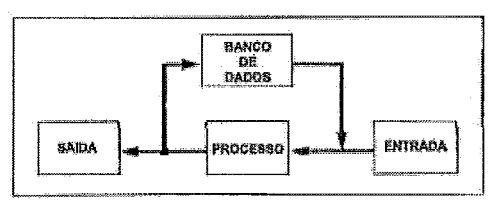


FIGURA 3
MODELO DE ENTRADA - PROCESSAMENTO - BANCO DE DADOS - SAÍDA

Fig. 2 e 3 – Modelo de Entrada – Processamento e Saída e Modelo de Entrada – Processamento – Banco de Dados e Saída

Em sistemas de informações reais, o BD atua como mediador entre o processo de Entrada e Saída, onde a entrada e saída: (1) ocorrem em momentos diferentes (tempo) e /ou (2) representam diferentes visões a respeito do mundo real. Esta ampla visão de um sistema torna possível o entendimento completo do modelo FCS, onde os sistemas de informação se ajustam de acordo com as ações tomadas no mundo real.

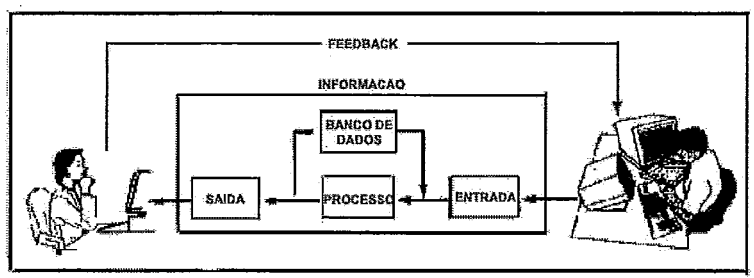


FIGURA 4

Fig. 4 – Modelo FCS

No modelo FCS proposto, por Orr (ORR, 1998), o dado é introduzido no sistema por vias externas. Posteriormente, passa por um processo e é armazenado no BD, que então é processado para produzir as saídas que serão usadas no (comparadas com o) mundo real. Por fim, novas entradas são produzidas (e re - alimentadas), de forma que o BD permaneça exato/preciso. Sem este loop final, o sistema irá falhar na manutenção de seu BD e saídas corretas. Desta forma o modelo final de FCS nos permite entender mais amplamente o verdadeiro problema de qualidade de dados – QD – quanto melhor nosso sistema de informação se ajusta ao mundo real, melhor será a qualidade do nosso dado; quanto pior ele se ajustar, pior será a qualidade de nosso dado. Orr (ORR, 1998) afirma que com o modelo FCS, torna-se mais fácil identificar se a organização usa ou não o dado, e exemplifica, comparando com um órgão que se não for usado, se atrofiará, dizendo que se ninguém usar um dado, o sistema ficará insensível àquele dado.

Em uma entrevista feita por Orr (ORR, 1998) , um gerente de dados de uma grande companhia relatou que 60% dos dados transferidos para o sistema de *Data Warehouse* falharam na passagem pelos testes de regras de negócios que os operadores de sistemas realizaram. Isto mostra quão pouco usados eles eram. A necessidade de dados de

qualidade. dia-a-dia, está ganhando maior atenção dos gerentes, tendo em vista o pobre estado de qualidade da maioria dos dados organizacionais.

Orr (ORR, 1998) recomenda que o foco primário de projetos envolvendo QD esteja no aumento de processos de controle internos envolvendo edição e digitação de dados e que o único caminho de se melhorar verdadeiramente a qualidade dos dados é aumentar o uso destes.

Outro autor que se preocupa com a criação de uma conscientização a respeito dos impactos da falta QD e a resolução deste problema é Redman (REDMAN, 1998). Ele acredita que os mais importantes problemas, encontrados pelos executivos hoje em dia, têm em sua origem a falta de qualidade dos dados.

Desta forma, ele categorizou os problemas de QD em;

i. Questões associadas a "Visões" (modelos do mundo real capturados pelos dados), bem como relevância, granularidade e nível de detalhes;
ii. Questões associadas aos conteúdos -valores- dos dados, tais como exatidão/precisão, consistência, moeda e totalidade;
iii. Questões associadas com a apresentação do dado, tais como formato apropriado, facilidade de interpretação e assim por diante;
iv. Outras questões tais como privacidade, segurança e propriedade.

A ciência de QD ainda não avançou para o ponto onde existam métodos de medidas padrões para todas estas questões. Além do mais, são poucas as empresas que rotineiramente medem a qualidade de seus dados. O que se conhece, empiricamente, é que no nível de atributos, a amplitude dos erros varia entre 5%-30% em um BD. Naturalmente, existem dificuldades de se comparar essas taxas de erros. Para este artigo, Redman (REDMAN, 1998) faz as seguintes declarações:

- A menos que a empresa tenha realizado grandes esforços na área de QD, é de se esperar que as taxas de erro dos dados (atributos) sejam de aproximadamente de 1-5% (onde Taxa de Erro = nº de atributos com erro/número total de atributos.)

- O que não tem medida, então não tem controle, pois as empresas esperam que existam outros problemas de QD mais sérios também. As empresas estão baseadas em BDs redundantes, inconsistentes que não possuem os dados que elas realmente precisam.

Levando isso em consideração, dados de baixa qualidade na verdade aumentam o custo operacional, visto que o tempo e outros recursos são despendidos na detecção e correção de erros. Ao nível operacional, dados de baixa qualidade baixam o moral dos empregados dificultando sobremaneira seus trabalhos.

Outrossim, dados com baixa qualidade são causadores de grandes problemas nas tomadas de decisão, obedecendo à máxima de que *as decisões não são melhores que os dados em que estão baseadas*. Ao nível tático, os dados são extremamente difíceis de sofrerem reengenharia, visto que um dos maiores objetivos da reengenharia é colocar o dado correto no lugar correto no tempo correto, para melhor servir o usuário. Mas como foi dito anteriormente, o dado não pode servir ao consumidor quando não está correto.

Do ponto de vista estratégico, o impacto pode comprometer grandemente as estratégias organizacionais, já que, normalmente, são feitas a longo prazo e necessitam de dados externos de qualidade muitas vezes incerta e com freqüência defasados. Contudo, como o planejamento estratégico está ligado aos resultados obtidos, este é freqüentemente modificado de acordo com os resultados.

Deste modo, este artigo (REDMAN, 1998) sinalizou que a questão de conscientização é o primeiro obstáculo para se implementar QD dentro de uma organização e que, posteriormente, as técnicas devem ser aperfeiçoadas tornando-se disponíveis para aplicação.

Uma nova fronteira de estudos está se abrindo, comenta Hipp (HIPP *et al.*, 2001a). De acordo com as técnicas de TQM, nós definimos a qualidade como sendo "busca consistente das expectativas dos desejos dos consumidores" (ENGLISH, 1999). Como já foi mencionado, o dado de baixa qualidade é sempre um problema em aplicações práticas de KDD. Isto pode parecer uma surpresa para alguns, mas a explicação é simples. O dado normalmente não é originário de aplicações onde o principal objetivo é

a mineração. Isto é, nós lidamos com sistemas que produzem dados tal qual produtos, sem a preocupação com o design e implementações, voltados para a qualidade do dado a longo prazo. Com a introdução de técnicas de Mineração da Qualidade dos Dados espera-se estimular a pesquisa sobre a importância e o potencial deste novo campo de aplicação.

Hipp (HIPP *et al.*, 2001b). sugere estes quatro importantes aspectos:

i. Emprego de métodos de mineração de dados para medir e explicar as deficiências dos dados
ii. Empregar os métodos de mineração de dados para corrigir as deficiências dos dados
iii. Expandir os modelos dos processos de KDD para refletir o potencial da DQM
iv. Desenvolvimento de modelos de processos especializados para "DQM pura"

Como mencionamos a DQM não está necessariamente baseada nos processo de KDD. Isto é. É fundamental o melhoramento da qualidade dos dados ser visto como um objetivo em si, fora do contexto da análise de dados. Isto implica na necessidade do desenvolvimento de processos de DQM especializados, que reflitam a mudança de escopo da análise pura de dados para o melhoramento e mensuração da qualidade dos dados.

Conclusão

Podemos perceber aos aspectos contextuais de QD, dos quais Wang (WANG, 2000) é o maior defensor, abrindo uma frente de análise mais voltada para a qualificação dos dados. Este aspecto pode acabar engessando as métricas, visto que se tentarmos metrificar os dados com base nos aspectos contextuais, seremos induzidos a utilizar pesos ou outros fatores de ponderação. Isto acabaria mudando o foco de nossa tese, visto que um dos objetivos a ser alcançado é o uso de técnicas que facilitem a medição da qualidade dos dados e a não necessidade de especialistas o tempo todo presentes. Contudo, Wang (WANG, 2000) defende o uso de TQM para os casos de quantificação de dados com problemas. Wilks (WILKS, 2000), Orr (ORR, 1998) e Redman (REDMAN, 1998) enfatizam a necessidade da investigação de técnicas capazes de

controlar as lacunas entre os dados corretos e os dados com desvios e técnicas de feedback entre os usuários e os sistemas, a fim de garantir um maior controle das operações com dados.

Os textos que surgiram em 2001, vide Hipp (HIPP *et al.*, 2001b), apresentam uma tendência interessante no uso de KDD com metadados e técnicas voltadas à quantificação dos problemas com os dados. Hoje, a grande preocupação é a velocidade altíssima que as massas de dados estão sendo geradas, pois muitas empresas já estão perdendo a capacidade de análise destes dados. A tendência deverá se basear em técnicas de exploração de dados que misturem KDD, metadados, inteligência artificial e técnicas de TQM. Desta forma, nossa aplicação deverá contemplar as seguintes características baseadas nas indicações técnicas destes autores:

i. Controle de metadados
ii. Tutores para facilitar a interface com o usuário não plenamente qualificado
iii. Métodos para medir e explicar a deficiência dos dados
iv. Métodos para definir, medir, analisar e melhorar a qualidade dos dados

Capítulo 3 - Técnicas, Métodos e Filosofia de Qualidade de Dados Aplicados a Banco de Dados

Melhorar a qualidade dos dados de uma organização é uma tarefa bastante complexa. As organizações possuem enormes quantidades de dados espalhados pelos seus departamentos e divisões, associados a diferentes tecnologias. Um programa de Qualidade de Dados é essencial para a melhoria da qualidade dos dados usados pela organização. Redman (REDMAN, 1996) afirma que um bom plano de QD deve possuir os seguintes pontos:

i. Deve possuir regras e objetivos claros
ii. Associar responsabilidade aos dados e garantir que os responsáveis devam ter as ferramentas adequadas para isso
iii. Ter um plano operacional de melhoria que especifique que métodos devam ser aplicados em que dados
iv. Estabelecer um programa de administração

A estrutura de TQM (*Total Quality Management*) para melhoria da qualidade dos dados foi proposta por Dvir e Evans (DVIR e EVANS, 1996), em vista da necessidade de traduzir as precisões dos consumidores de dados em métricas. Eles orientam o uso das técnicas de Controle Estatístico de Processos, tais como Diagrama de Pareto e Cartas de Controle para serem ser aplicadas à medição, acompanhamento (auditoria) e melhoria da qualidade dos dados.

Muitos profissionais de IS encaram o dado como mero objeto de Entrada/Processamento/Saída. Esta visão simplificada do problema não colabora no resgate da qualidade da informação. Isto porque, quanto mais se sabe a respeito de um processo, mais se pode controlá-lo. Desta forma, propomos olhar os dados produzidos pela organização como um produto que deva necessariamente ter um consumidor. O dado que for gerado sem objetivo de consumo, não deverá ser de preocupação e monitoração constante. Os profissionais de SI não podem transferir o problema da garantia de qualidade dos dados ao consumidor, visto que este apenas deseja que o seu produto atenda às especificações e que o satisfaça.

Visto assim, compararemos o dado produzido a um produto e sua produção a uma linha de montagem, onde um produto (dado) só tem razão de existir a fim de satisfazer a demanda de seus consumidores. O objetivo imediato de uma linha de produção é garantir que a manufatura do produto tenha a menor variabilidade possível, isto é, em qualquer processo de produção de dados, apesar de bem desenhados ou cuidadosamente mantidos, certa quantidade inerente ou natural de variabilidade sempre irá existir. Essa variabilidade natural ou "ruído", quando é relativamente pequena, usualmente é considerada aceitável dentro do nível de performance. Exemplificando, caso uma secretária grafe três ou quatro palavras erradas em um texto de 60 laudas, podemos considerar uma variabilidade mínima. Contudo se em uma pesquisa de opinião com 400 questionários com 20 itens, alguns digitadores interpretam errado um parâmetro, pode contribuir para um diagnóstico errado do problema pesquisado. O que podemos perceber nestes dois exemplos é que dado o contexto é fundamental podermos adaptar a ferramenta de controle ao problema apresentado.

Nem sempre quem produz o dado consegue avaliar o seu erro, assim sendo, se o consumidor antes de divulgar sua análise a respeito da informação obtida tivesse um ferramental que pudesse executar uma análise na qualidade do dado e ponderar a fonte de informação, o índice de "defeito" seria menor.

O principal intuito do controle Estatístico de Processos associado à BD é detectar rapidamente a ocorrência de causas associadas ou deslocamento de processos, a fim de que a investigação e as ações corretivas possam ser tomadas antes de vários dados não conformes serem produzidos. O desenvolvimento matemático das fórmulas e gráficos poderão ser encontrados no livro de Montgomery (MONTGOMERY, 1991).O gráfico de Cartas de Controle (*Control Chart*) (MONTGOMERY, 1991) é uma técnica de controle de processos on-line amplamente utilizada na indústria com este propósito. As Cartas de Controle podem ser utilizadas na estimativa de parâmetros numéricos do processo, relacionadas à produção do dado e através desta informação determinar a capacidade do processo. Além disso, a Carta de Controle também pode prover informações para a melhoria do processo, visto que se torna claro quem produziu aquela informação e quando. Mas lembrando, o principal objetivo do Controle Estatístico de Processo é a eliminação da variabilidade dentro do processo – ou seja garantir a qualidade da informação coletada. Pode não ser possível a total eliminação da

variabilidade, mas as cartas de controle são uma ferramenta valiosa na redução da variabilidade.

3.1. Bases Estatísticas da Carta de Controle

Uma típica carta de controle é um controle gráfico da qualidade das características mensuradas ou calculadas a partir de uma amostra, versus o número amostral ou o tempo. A linha central do gráfico representa o valor médio da qualidade característica correspondendo ao estado de controle. As outras duas linhas horizontais, chamadas Upper Control Limit (*Limite de Controle Superior – UCL*) e Lower Control Limit (*Limite de Controle Inferior - LCL*) representam os limites de controle. Desta forma, o processo é dito sob controle quando todos os pontos da amostra estão entre eles. Enquanto os pontos permanecerem dentro dos limites de controle, nenhuma ação é necessária.

Contudo, se o ponto estiver fora destes limites, ele é interpretado como evidência do descontrole do processo. Assim sendo, ações investigativas e corretivas serão necessárias, para identificar e eliminar a causa associada ou causas responsáveis por este comportamento.

Mesmo que todos os pontos estejam dentro dos limites de controle, se eles descreverem comportamentos não aleatórios, existe uma indicação de que o processo está fora de controle.

Existe uma relação próxima entre as Cartas de Controle e o Teste de Hipóteses. Essencialmente, a Carta de Controle é o teste de hipótese cujo processo está estatisticamente sob controle. O ponto plotado dentro dos limites de controle é equivalente a aceitar a hipótese de controle estatístico, enquanto que os pontos fora dos limites de controle equivaleriam a rejeitar a hipótese de controle estatístico. Tal como no Teste de hipóteses, podemos pensar que a probabilidade do erro do Tipo I da Carta de Controle (concluindo que o processo está fora de controle quando ele não está) e a probabilidade do *erro Tipo II* da Carta de Controle (concluindo que o processo está sobre controle quando ele não está). Ocasionalmente é aconselhável utilizar a curva característica de operação de uma Carta de Controle, a fim de mostrar a probabilidade

de erro do tipo II. Isso indicará ao usuário a habilidade da Carta de Controle em identificar mudanças no processo de diferentes magnitudes.

Definimos um modelo geral de Carta de Controle, onde w representa a amostra estatística que mede a qualidade de alguma característica de nosso interesse e supomos que a média de w é representada por μ_w e o desvio padrão de w é σ_w. Desta forma, a linha central , a linha superior e inferior serão definidas como:

$$UCL = \mu_w + k\,\sigma_w$$

$$\text{Linha Central} = \mu_w$$

$$LCL = \mu_w - k\,\sigma_w$$

Onde o parâmetro *k* corresponde à "distância" dos limites de controle a partir da linha central, expresso em unidades de desvio padrão . Essa teoria geral de Carta de Controle foi proposta por Dr. Walter Shewart e as Cartas de Controle de acordo com estes princípios são também chamadas de *Shewart Control Charts*.

A Carta de Controle, com certeza, é um dispositivo capaz de descrever de forma precisa o que se entende por controle estatístico, isto é, ela pode ser usada de várias formas, inclusive em processos on-line. Por exemplo, amostras de dados são coletadas e usadas a fim de construir uma Carta de Controle e se os valores da amostra coletada de $\bar{x}$ caírem dentro dos limites de controle e não exibirem qualquer padrão sistemático, dizemos que o processo está sob controle ao nível indicado pela Carta.

O uso mais importante da Carta de Controle é na melhoria do processo, onde encontramos geralmente:

i. Muitos processos não operando em um modo de controle estatístico.
ii. Conseqüentemente a rotina e o uso atento de Carta de Controle identificarão as causas associadas. Se essas causas puderem ser eliminadas do processo, a variabilidade será reduzida e o processo melhorado. (essa atividade de melhoria do processo usando a Carta de Controle é ilustrada nas figuras 5 e 6)
iii. A Carta de Controle somente detecta as causas associadas.

Na identificação e eliminação das causas possíveis é importante encontrar a fonte primária do problema e atacá-la. Uma solução cosmética não resultará em nenhum melhoramento a longo termo do processo. Desenvolver um sistema efetivo para a ação corretiva é um componente essencial para implantação de SPC.

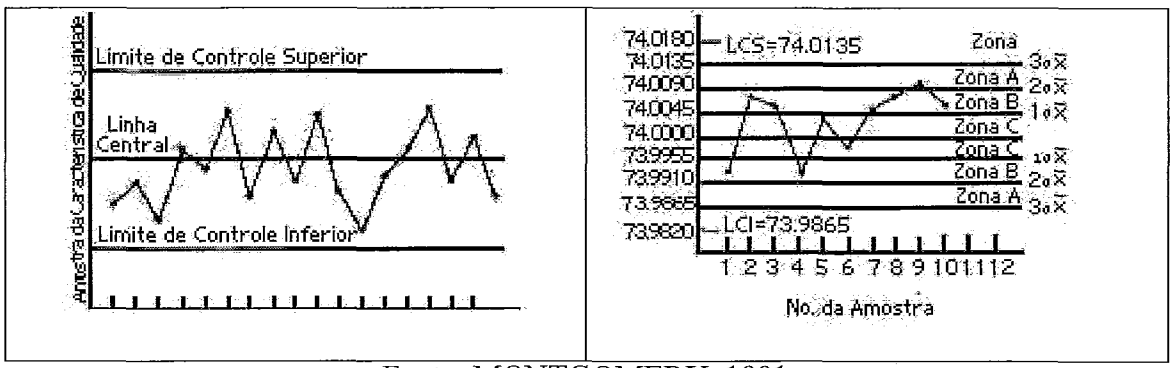
Podemos também usar a Carta de Controle como um dispositivo de estimação, isto é, a partir da Carta de Controle podemos estimar certos parâmetros de processos, tais como: a média, o desvio padrão, a fração de não conformidade ou falha. Essas estimativas podem ser usadas na determinação da capacidade do processo de produzir dados aceitáveis. Esses estudos de capacidade de processo têm consideráveis impactos nos problemas de gerenciamento de decisão que ocorrem no ciclo de produção, incluindo decisões de compra e venda e acordos contratuais com consumidores ou vendedores.

Assim sendo, as cartas de controle podem ser classificadas em dois tipos gerais. Se a qualidade característica pode ser mensurada e expressa como um número em alguma escala contínua de medida, então usualmente é chamada variável. Nesse caso, é conveniente descrever a característica de qualidade como uma medida de tendência central e uma medida de variabilidade. As cartas de controle de tendência central e variabilidade são chamadas de Cartas de Controles de Variáveis. O gráfico $\bar{x}$ é o mais utilizado no controle de tendência central – gráficos baseados em faixas amostrais ou desvio padrão amostral, discutiremos mais à frente esses tipos de gráficos.

Muitas características de qualidade não são mensuradas em uma escala contínua ou quantitativa, nesses casos nós podemos julgar cada unidade do produto como estando em conformidade ou não conformidade, isto é, nós podemos contabilizar o número de não conformidades (*defeitos*) que aparecem em uma unidade de produto, neste caso os gráficos para controle dessas características de qualidade são chamados Cartas de Controles de atributos, definidos mais à frente.

Não podemos esquecer que, além das cartas de controle, existem outras ferramentas que nos ajudam a resolver problemas relacionados ao SPC. Uma das mais conhecidas é o diagrama de Pareto. Este diagrama relaciona de forma simples a distribuição de freqüência (ou histograma) de um atributo de acordo com categorias. Contudo, devemos

notar que o diagrama de Pareto (figura 7) não identifica automaticamente os defeitos mais importantes, mas sim os que ocorrem com maior freqüência.

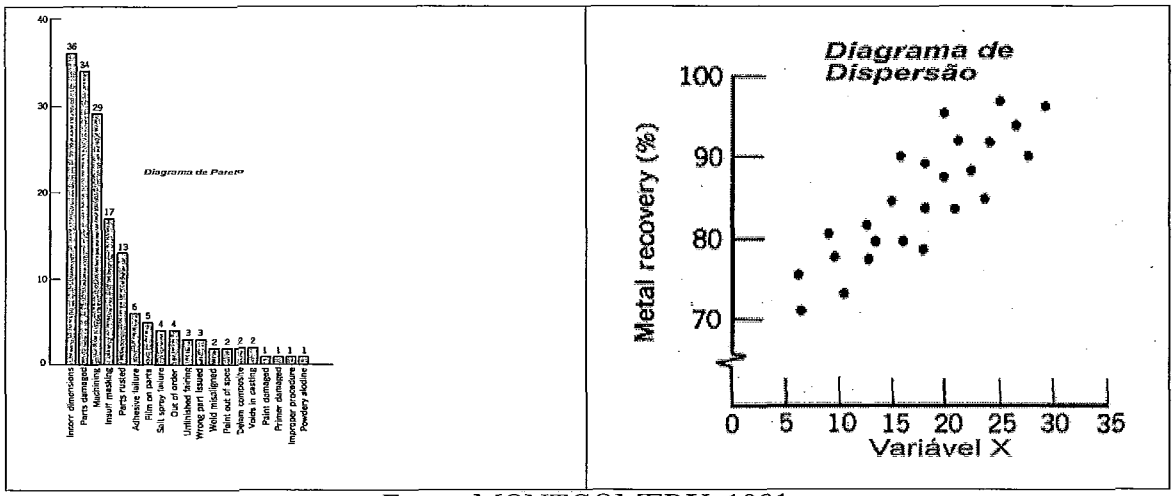


Fonte: MONTGOMERY, 1991

Fig. 5 e 6 – Modelo de Carta de Controle e Exemplo de Carta de Controle

Outra importante ferramenta utilizada na manutenção de controle de qualidade é o diagrama de dispersão. Quando identificamos problemas de relacionamento entre duas variáveis, podemos usar o diagrama de dispersão com base nos dados coletados em pares, usando as duas variáveis; desta forma plotamos a variável x contra a variável *y*. A forma do diagrama de dispersão nos indica que tipo de relação pode ocorrer entre as duas variáveis. Isto é, se o diagrama indicar uma correlação positiva teremos a informação de que se houver um incremento da variável *x,* haverá também um incremento da variável *y*.

Contudo, este tipo de pensamento é perigoso porque correlação não significa necessariamente casualidade. Este relacionamento aparente pode ser causado por alguma coisa diferente. Desta forma, recomendamos o uso criterioso deste diagrama, indicando-o para identificação de potenciais relacionamentos.



Fonte: MONTGOMERY, 1991

Fig. 7 e 8 – Diagrama de Pareto e Diagrama de Dispersão (Scatter Plot)

3.1.1.Tipos de Carta de Controle

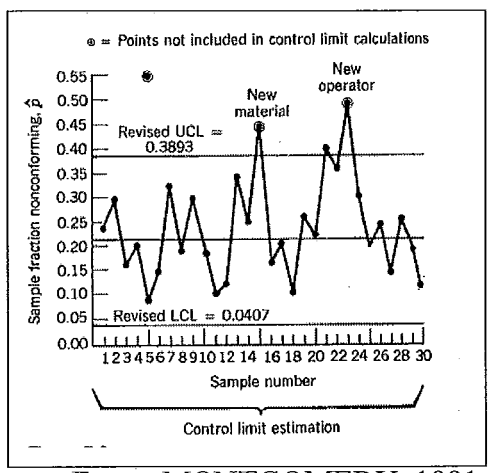
3.1.1.2. Carta de Controle de Atributos

Muitas características de qualidade não podem ser representadas numericamente. Nestes casos, nós usualmente classificamos cada item inspecionado como dentro das especificações ou não conforme (*defeituoso*) de acordo com as especificações relacionadas àquela característica.

O termo "defeituoso" e "não defeituoso" é mais utilizado na identificação destas duas classificações de produtos. Este tipo de característica de qualidade é chamado de *atributo*. Alguns exemplos das características de qualidade que são atributos incluem a ocorrência de engrenagens empenadas de carros conectadas à barra de direção em um dia de produção ou a ocorrência de dados clínicos preenchidos com erro e assim por diante.

Apresentaremos as três Cartas de Controle mais usadas. A primeira tem relação com a fração de não conformidade ou defeito em produtos criados pelos processos de manufatura – esta Carta de Controle é chamada de Carta de Controle de Fração de Não-Conformidade ou *P* Chart. Em algumas situações é mais conveniente lidar com o número de defeitos ou não conformidades, observados em lugar da fração de não conformidade. O segundo tipo de Carta de Controle que apresentaremos chama-se Carta

de Controle de Não Conformidades ou *C* Chart - desenhada para lidar com este caso. E finalmente, apresentaremos a Carta de Controle para Não Conformidades por Unidade ou *U* Chart, que é útil em situações onde o número médio de não conformidades por unidade é a base mais conveniente para o Controle do Processo.



Fonte: MONTGOMERY, 1991

Fig. 9 – Exemplo de Carta de Controle de Processos

3.1.1.2.1. Carta de Controle para Frações de Não-Conformidades

A fração de não conformidade é definida com a relação entre o número de itens fora de especificação em uma população e o número total de itens desta população. Os itens podem ter várias características de qualidade que são examinadas simultaneamente pelo inspetor. Se o item não estiver conforme o padrão estabelecido em uma ou mais destas características, será classificado como não conforme ou fora das especificações. Nós usualmente expressamos a fração de não conformidade na forma decimal, embora ocasionalmente, o percentual de não conformidade seja utilizado. Ao apresentarmos o gráfico de controle para o público, o percentual de não conformidade é mais usual, visto que apresenta uma forma mais intuitiva. Conquanto seja habitual trabalhar com a fração de não-conformidade, podemos também analisar a fração como um rendimento do processo.

Os princípios estatísticos associados ao Gráfico de Controle de Frações de não Conformidade são baseados na distribuição binomial. Suponhamos que um processo

seja operado de maneira estável, tal que a probabilidade de qualquer unidade que não atenda às conformidades seja p e que sucessivas unidades produzidas sejam independentes, então, cada unidade produzida é a efetivação de variável aleatória de Bernoulli de parâmetro p. Se uma amostra aleatória de n unidades de um produto é selecionada e D é o número de unidades de um produto que não estão conforme as especificações, então, D tem distribuição binomial com parâmetros n e p, isto é,

$$P\{D = x\} = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{x=0,1,....,n}$$

A fração amostral de não conformidade é definida como a razão entre o número de unidades fora da especificação amostra D e o tamanho da amostra n, isto é,

$$\hat{p} = \frac{D}{n}$$

A distribuição da variável aleatória $\hat{p}$ pode ser obtida a partir da binomial. Além disso, a média e a variância de $\hat{p}$ são

$$\mu = p$$

e

$$\sigma^2_{\hat{p}} = \frac{p(1-p)}{n}$$

respectivamente. Veremos como essa teoria pode ser aplicada no desenvolvimento de Cartas de Controle de Frações de Não Conformidade. Devido ao gráfico de controle, o processo de frações de não conformidade p é chamado de P Chart.

Desenvolvimento e Operação do Gráfico de Controle

Discutiremos os princípios estatísticos gerais em que este gráfico de controle encontra-se baseado. Se *w* é uma estatística que mede a característica de qualidade e se a média de w é μ_w e a variância de *w* é σ_w^2, então o modelo geral de Carta de Controle de Shewhart é o seguinte:

$$UCL = \mu_w + k\,\sigma_w$$
$$\text{Linha Central} = \mu_w \quad (1)$$
$$LCL = \mu_w - k\,\sigma_w$$

Onde *k* é a distância entre os limites de controle e a linha central, em múltiplos de desvio padrão de *w*. É habitual a escolha de *k*=3.

Suponhamos que a verdadeira fração de não conformidade p no processo produtivo é conhecida ou tem valor padronizado especificado pelo administrador. Como vimos em (1), a linha central e os limites de controle da fração de não conformidade da Carta de Controle serão:

$$UCL = p + 3\sqrt{\frac{p(1-p)}{n}}$$
$$\textit{Linha}\ \text{Central} = p$$
$$LCL = p - 3\sqrt{\frac{p(1-p)}{n}}$$

A operacionalização real deste gráfico consiste na coleta de amostras subseqüentes de *n* unidades , calculando a fração amostral de não conformidade $\hat{p}$ e plotando a estatística $\hat{p}$ no gráfico. Desde que $\hat{p}$ permaneça dentro dos limites de controle e a seqüência de pontos plotados não exiba padrão não aleatório sistemático, podemos concluir que o processo está sob controle ao nível de *p*. Se a pontos forem plotados fora dos limites de controle ou se padrões não aleatórios no gráfico forem observados, pode-se concluir que a fração de não conformidade do processo trocou para um novo nível e que o processo está fora de controle.

Quando a fração de não conformidade p do processo não for conhecida, ela deve ser estimada a partir dos dados observados. O procedimento usual é selecionar preliminarmente m amostras, com tamanho n cada. Como regra geral, *m* pode ser 20 ou 25. Então, se existirem D_i unidades fora de especificação na amostra *i* , calcularemos a fração de não conformidade da i-ésima amostra como

$$\hat{p}_i = \frac{D_i}{n} \qquad i = 1,2,....,m$$

e a média desta fração de não conformidade individual como

$$\bar{p} = \frac{\sum_{i=1}^{m} D_i}{mn} = \frac{\sum_{i=1}^{m} \hat{p}_i}{m}$$

A estatística $\bar{p}$ estima a fração de não conformidade p desconhecida. A linha central e os limites de controle central deste gráfico de fração de não conformidade são calculados como

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$Linha\ \text{Central} = \bar{p}$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Nós consideramos os limites de controle obtidos como "trial control limits" – limites de teste. Eles permitem determinar onde o processo estava sob controle quando as m amostras iniciais foram selecionadas. Para testar a hipótese de controle pretérito, plote a fração de não conformidade de cada amostra e analise os resultados apresentados. Se todos os pontos estiverem dentro dos limites de controle e não for encontrado nenhum padrão sistemático, dizemos que o processo estava sob controle no passado. Assim, podemos utilizar estes parâmetros para o controle da produção corrente ou futura.

Suponhamos que exista mais de uma estatística $\hat{p}_i$ plotada, fora de controle quando comparada com os limites de controle iniciais. Se os limites de controle para a produção atual ou futura forem significativos, devem ser baseados nos dados oriundos do processo que está sob controle. Contudo, quando a hipótese do controle pretérito é rejeitada, é necessário fazer a revisão dos limites de controle iniciais.

Se a Carta de Controle for baseada em parâmetros - p conhecidos ou padronizados, não há necessidade de se calcular os limites de controle iniciais.

Montgomery (MONTGOMERY,1991) sugeriu que o tamanho da amostra deveria ser grande o suficiente para que tivéssemos 50% de chance de detectarmos uma mudança de uma determinada ordem. Assim, se δ é a magnitude da mudança do processo, então, n deve satisfazer.

$$\delta = k\sqrt{\frac{p(1-p)}{n}}$$

$$\text{logo, } n = \left(\frac{k}{\delta}\right)^2 p(1-p)$$

$$LCL = p - k\sqrt{\frac{p(1-p)}{n}} > 0 \text{ , o que implica em } n > \frac{(1-p)}{p}k^2$$

3.1.1.2.2. Carta de Controle np

Também é possível basear a Carta de Controle no número de não conformidades ao invés da fração de não conformidade, chamada Gráfico de Controle *np* .Os parâmetros deste gráfico são

$$UCL = np + 3\sqrt{np(1-p)}$$
$$\text{Linha Central} = np$$
$$LCL = np - 3\sqrt{np(1-p)}$$

Se o valor padrão de p não for disponível, então $\bar{p}$ pode ser utilizado para estimar p. Muitas pessoas que não possuem treinamento estatístico acham o Gráfico *np* mais fácil de interpretar que a Carta de Controle de fração de não conformidade.

Amostra de Tamanhos Variáveis

Algumas aplicações de Carta de Controle para fração de não conformidade se baseiam em amostras de diferentes tamanhos, visto que o número de unidades produzidas pode ser diferente a cada período. Existem diferentes abordagens para se construir e operacionalizar uma Carta de Controle com amostra de tamanho variável.

A primeira e talvez a mais simples abordagem é determinar os limites de controle para cada amostra. Isto é, se a i-ésima amostra tem tamanho n_i e os limites de controle são

$$p \pm 3\sqrt{\frac{p(1-p)}{n_i}}$$

Note que a largura dos limites de controle é inversamente proporcional à raiz quadrada do tamanho da amostra.

Devemos ter cuidado ao analisar os testes ou padrões aparentemente fora da normalidade ao utilizar amostras de tamanhos variáveis. O problema é que uma fração de não conformidade da amostra $\hat{p}$ deve ter sua interpretação com base no tamanho da amostra. Ex.: suponha que p=0.20 e que duas frações de não conformidade sucessivas sejam $\hat{p}_i = 0.28$ e $\hat{p}_{i+1} = 0.24$. À primeira vista, existe uma indicação de que a segunda possui menos qualidade que a primeira, visto que $\hat{p}_i > \hat{p}_{i+1}$. Contudo, suponhamos que os tamanhos das amostras são n_i=50 e n_{i+1}=250. Em unidades com desvios padronizados, o primeiro ponto corresponde a 1.89 unidades acima da média, enquanto o segundo ponto →2.11 acima. Isto é, o segundo ponto representa um grande desvio do padrão de p=0.20 tanto quanto o primeiro, embora o segundo ponto seja o menor dos dois.

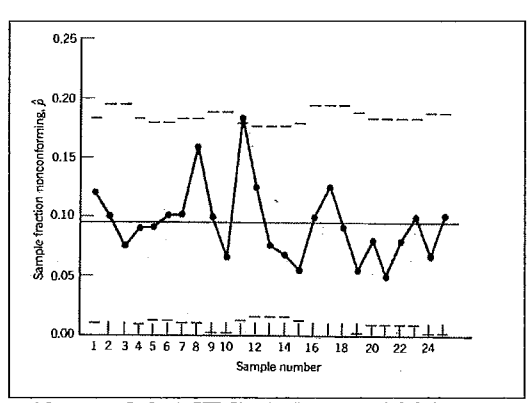
Uma solução para este problema é utilizar a Carta de Controle "padronizada" , onde os gráficos sejam plotados em unidades de desvios padronizados. Tal controle tem Linha Central zero e limites de controle superior e inferior de +3 e –3 respectivamente. A variável plotada no gráfico é

$$Z_i = \frac{\hat{p}_i}{\sqrt{\frac{p(1-p)}{n_i}}}$$

onde p (ou $\bar{p}$ caso nenhum padrão seja fornecido) é a fração de não conformidade do processo em estado de controle. Testes para reconhecimento de padrões podem ser utilizados com segurança aplicando-se este gráfico, visto que todas unidades estão na mesma medida.

O gráfico de controle padronizado não é mais difícil de se construir ou manter que outros procedimentos anteriormente descritos. Conceitualmente, todavia, seu entendimento e interpretação do processo podem trazer dificuldade para o pessoal de operação.

Todavia, se existirem grandes variações do tamanho amostral, os métodos para os testes e o reconhecimento de padrões poderão ser aplicados com segurança somente utilizado Cartas de Controle Padronizadas. Neste caso, recomenda-se manter a Carta de Controle com limites de controle individuais para cada amostra, no caso do pessoal de operação, e simultaneamente manter uma Carta de Controle Padronizada para os engenheiros de qualidade.



Fonte: MONTGOMERY, 1991

Fig. 10 –Carta de Controle para Amostras de Controle Variáveis

Para Aplicações Não Fabris

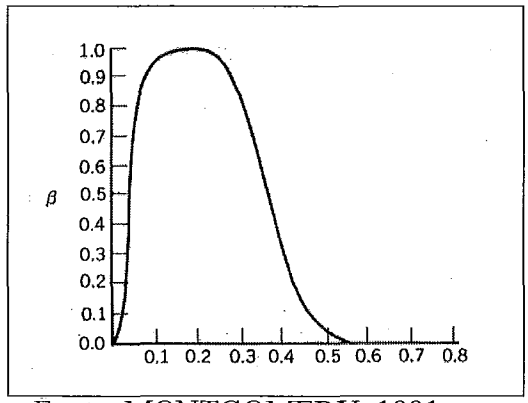
O Gráfico de Controle ou Carta de Controle para fração de não conformidade é amplamente utilizado em aplicações não fabris de controle estatístico de processos. Nos ambientes não fabris, muitas características de qualidade podem ser observadas sob as bases da conformidade e não conformidade. Alguns exemplos incluiriam o número de cheques de pagamentos de empregados que possuem erro de preenchimento ou são distribuídos posteriormente à data de pagamento, o número de cheques não pagos dentro do ciclo padrão de pagamento e o número de entregas realizadas atrasadas por um fornecedor.

Muitas aplicações não manufaturadas de Carta de Controle de fração de não conformidade envolverão tamanhos de amostras variáveis. Por exemplo, o número de cheques requisitados durante um ciclo de pagamentos é mais provável que não seja constante e desde que as informações a respeito do período do processo de todas as requisições dos cheques estejam disponíveis, pode-se calcular $\hat{p}$ como sendo uma razão de todos os cheques atrasados sobre o total de cheques processados durante o período.

Curva de Função Característica de Operação e Tamanho Médio das Provas

A Curva de Função Característica de Operação (ou OC) da Carta de Controle da fração de não conformidade é um gráfico contendo a probabilidade de aceitação da hipótese estatística incorreta (i.e., tipo II ou erro β) contra o processo de fração de não conformidade . A curva OC fornece uma medida de sensibilidade da Carta de Controle , isto é, sua habilidade está na detecção de mudanças na fração de não conformidade do processo a partir do valor nominal de $\overline{p}$ até algum valor de p. A probabilidade do erro tipo II para a Carta de Controle da fração de não conformidade pode ser calculada como

$$\beta = P\{\hat{p} < UCL \mid p\} - P\{\hat{p} \leq LCL \mid p\} = P\{D < nUCL \mid p\} - P\{D \leq nLCL \mid p\}$$



Fonte: MONTGOMERY, 1991

Fig. 11 - Curva de Função Característica de Operação

Desde que D seja uma variável aleatória binomial com parâmetros n e p, o erro β definido acima pode ser obtido a partir da função de distribuição binomial acumulada. Podemos também, calcular o tamanho médio das provas (*Average Run Lengths – ARL*) para a fração de não conformidade da Carta de Controle.

$$ARL = \frac{1}{P(\text{pontos amostrais fora de controle})}$$

Deste modo, se o processo estiver sob controle, o valor de ARL será:

$$ARL = \frac{1}{\alpha} \text{ e se estiver fora de controle, } ARL = \frac{1}{1-\beta}$$

Estas probabilidades (α,β) podem ser calculadas diretamente a partir da função de distribuição binomial ou lidas a partir da curva OC .
Então, se o processo estiver realmente sob controle, nós experimentaremos um "alarme falso" – um sinal de fora de controle sobre todas as $1/\alpha$ amostras.

O aumento no tamanho da amostra resultará num valor de baixo β e um pequeno ARL fora de controle. Outra abordagem seria reduzir o intervalo entre as amostras. Isto é, se coletarmos amostras a cada hora, levaremos cerca de 7 horas, em média, para

detectarmos uma mudança, e se coletarmos a cada meia hora, levaremos cerca de 3,5 horas, em média, para detectarmos a mudança.

3.1.1.2.3. Carta de Controle para Não Conformidades (Defeitos)

Um item fora das especificações é uma unidade de produção que não satisfaz uma ou mais especificações para aquele produto. Este ponto, em específico,, em que uma especificação não é satisfeita, resulta em um defeito ou não conformidade. Conseqüentemente, um item não conforme possuirá pelo menos uma não conformidade.

Todavia, dependendo de natureza e severidade, é bastante provável que uma unidade possua várias não conformidades e então seja classificada como defeituosa. Como exemplo, suponhamos que manufaturamos computadores pessoais. Cada unidade pode ter mais de uma pequena falha e desde que estas falhas não afetem seriamente o funcionamento, a unidade será considerada conforme. No entanto, se existirem tantas falhas que afetem o funcionamento da unidade, o computador será considerado defeituoso. Existem muitas situações na prática, nas quais é preferível trabalhar com o número de efeitos ou não conformidades, em vez de fração de não conformidade.

É possível desenvolver Cartas de Controle tanto para o número de defeitos quanto para a média de defeitos por unidade. Estes gráficos, freqüentemente, assumem que a ocorrência de defeitos em uma amostra de tamanho constante, melhor modelada pela distribuição de Poisson.

Procedimento para amostras de tamanho constante

Considere a ocorrência de defeitos em uma unidade inspecionada de um produto. Na maior parte dos casos, a unidade inspecionada é um único produto, apesar de necessariamente não sê-lo sempre. A unidade inspecionada é simplesmente uma entidade para qual é conveniente manter registros. Podem ser grupos de 5 unidades ou 10 unidades e assim por diante. Suponhamos que estes defeitos ocorram na unidade inspecionada de acordo com o a distribuição de Poisson, isto é

$$p(x) = \frac{e^{-c}c^{x}}{x!} \qquad x = 0,1,2,...$$

onde x representa o número de defeitos e c>0 é o parâmetro da distribuição Poisson. A distribuição Poisson possui média e variância c. Então, a Carta de Controle para não conformidades com limites 3-sigmas[2] será dada por

$$UCL = c + 3\sqrt{c}$$
$$\text{Linha Central} = c$$
$$LCL = c - 3\sqrt{c}$$

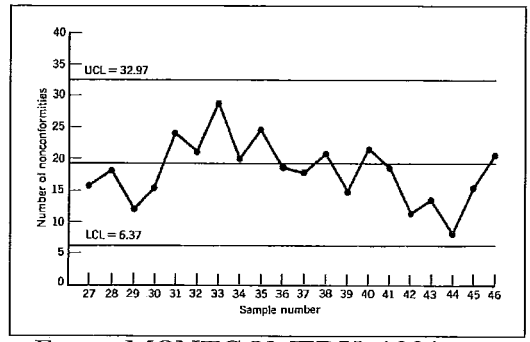
assumindo que o valor padrão para c seja conhecido. Estes cálculos podem render valores negativos para LCL, fixe LCL=0.

Se nenhum padrão for dado, então c deve ser estimado como o número médio de defeitos observados em uma amostra preliminar de unidades inspecionadas → $\bar{c}$. Neste caso, a Carta de Controle terá os seguintes parâmetros

$$UCL = \bar{c} + 3\sqrt{\bar{c}}$$
$$\text{Linha Central} = \bar{c}$$
$$LCL = \bar{c} - 3\sqrt{\bar{c}}$$

Quando nenhum padrão for dado, os limites de controle deverão ser considerados como limites de controle iniciais e as amostras preliminares examinadas como falta de controle. A Carta de Controle para Não Conformidades também é chamada de c Chart.

[2] O risco a para 3-sigmas não é igualmente alocado acima de UCL e abaixo de LCL, por causa da assimetria apresentada pela distribuição Poisson. Alguns autores recomendam o uso dos limites de probabilidade para este gráfico, particularmente quando o valor de c for pequeno.



Fonte: MONTGOMERY, 1991

Fig. 12 – Carta de Controle de Não Conformidades – c Chart

Análises Adicionais de Não Conformidades

Os dados de defeitos ou não conformidades são mais informativos que a fração de não conformidade, isto porque existem vários tipos diferentes de não conformidades. Analisando as não conformidades por tipo, podemos ganhar com freqüência uma considerável percepção da causa. Logo, poderíamos usar o Diagrama de Pareto a fim de sabermos qual o defeito com maior freqüência e ao isola-lo ou elimina-lo, aumentaríamos sobremaneira a capacidade do processo. Note que as não conformidades seguem a distribuição de Pareto, isto é, a maioria dos defeitos é atribuída a poucos tipos causas.

Outra técnica útil para uma análise adicional é o diagrama de causa e efeito. Contudo, este foge do escopo de nosso trabalho.

3.1.1.2.3. Sistemas de Deméritos

Com produtos complexos tais como automóveis, computadores entre outros, freqüentemente vários tipos de não conformidades ou defeitos podem ocorrer. Nem todos os tipos de defeitos são igualmente importantes. Uma unidade de um produto contendo um sério defeito pode ser classificada como não conforme, enquanto uma unidade contendo vários defeitos pequenos pode não ser necessariamente classificada como defeituosa. Para estas situações, precisamos de um método de classificação de não

conformidades ou defeitos de acordo com a severidade, para ponderar os vários tipos de defeitos de forma razoável.

Um esquema de demérito possível seria:

i. **Defeitos Classe A – Muito Sério.** A unidade é completamente imprópria para o serviço ou falhará em serviço de tal forma que não será fácil corrigi-la em uso;

ii. **Defeitos Classe B – Sérios.** A unidade possivelmente sofrerá uma falha do tipo Classe A ou certamente causará algum problema operacional sério menor;

iii. **Defeitos Classe C – Moderadamente Sério.** A unidade possivelmente falhará em serviço ou causará problema de falha, menor que uma falha de operação ou prejudicará a qualidade do trabalho;

iv. **Defeito Classe D – Menor.** A unidade não falhará em serviço, mas possuirá pequenos defeitos.

Façamos c_A, c_B, c_C e c_D representarem o número de defeitos de Classe A, B,C e De, respectivamente, em uma unidade inspecionada. Assumiremos que cada classe de defeitos é independente e a ocorrência de defeitos em cada classe é melhor modelada pela distribuição de Poisson. Assim definimos o número de deméritos na unidade inspecionada como

$$D=100\,c_A+50\,c_B+100\,c_C+c_D$$

Os pesos da Classe A-100, Classe B-50, Classe C-10 e Classe D-1 são amplamente utilizados na prática. Contudo, qualquer conjunto razoável de pesos pode ser definido apropriadamente para um problema específico.

Suponhamos que a amostra de n unidades inspecionadas é usada. O número de deméritos por unidade é

$$u=\frac{D}{n}$$

onde D é o número total de deméritos em todas as n unidades inspecionadas. Já que u é uma combinação linear de variáveis aleatórias de Poisson, a estatística u pode ser plotada em uma Carta de Controle com os seguintes parâmetros:

$$UCL = \bar{\bar{u}} + 3\hat{\sigma}_u$$

Linha Central $= \bar{\bar{u}}$ onde $\bar{\bar{u}} = 100\bar{u}_A + 50\bar{u}_B + 10\bar{u}_C + \bar{u}_D$ e

$$LCL = \bar{\bar{u}} - 3\hat{\sigma}_u$$

$$\hat{\sigma}_u = \sqrt{\left[\frac{(100)^2 u_A + (50)^2 u_B + (10)^2 u_C + u_D}{n}\right]},$$

$\bar{u}_A, \bar{u}_B, \bar{u}_C$ e $\bar{u}_D$ representam o número médio de defeitos Classe A,B,C e D por unidade. Outras variações desta idéia são possíveis. Como por exemplo, classificar as não conformidades como aparentes ou não aparentes.

Curva de Função Característica de Operação

A curva de função característica (OC) para Gráficos u e c pode ser obtida a partir da distribuição de Poisson. Para o *c* Chart, a curva OC plota a probabilidade do erro tipo II β contra a valor médio verdadeiro de defeitos *c*. A expressão para o valor de β é

$$\beta = P\{x < UCL \mid c\} - P\{x \leq LCL \mid c\}$$

onde x é uma v.a.. (variável aleatória) Poisson de parâmetro c.

Para o *u* Chart, podemos generalizar a curva OC a partir de

$$\begin{aligned}\beta &= P\{x < UCL \mid c\} - P\{x \leq LCL \mid c\} \\ &= P\{c < nUCL \mid u\} - P\{c \leq nLCL \mid u\} \\ &= P\{nLCL < c \leq nUCL \mid u\} \\ &= \sum_{c=<nLCL>}^{[nUCL]} \frac{e^{-nu}(nu)^c}{c!}\end{aligned}$$

onde $<nLCL>$ denota o menor valor inteiro maior ou igual a *nLCL* e $[nUCL]$ corresponde a maior inteiro menor ou igual a *nUCL*. Os liites de controle seguem o fato

de que o número de não conformidades observadas na amostra de n unidades deva ser inteiro.

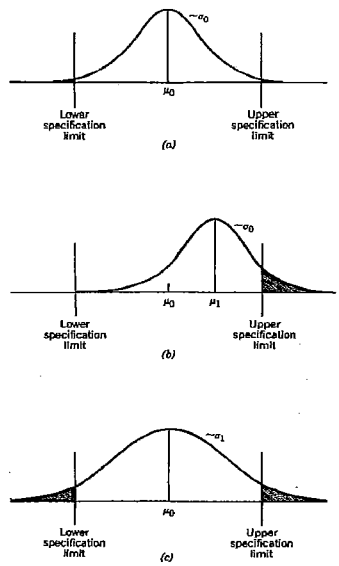
Aplicações Não Fabris

As Cartas de Controle – C e U são amplamente usadas em aplicações não manufaturadas de controle estatístico de processos. Na realidade, nós podemos tratar erros no ambiente de não manufaturas da mesma forma que tratamos os defeitos ou não conformidades no mundo manufaturado. Por exemplo, podemos plotar os erros em plantas de engenharia e documentos e erros associados a programas de computador em Gráficos - U e C.

3.1.1.3.Cartas de Controle de Variáveis

Muitas características qualitativas podem ser expressas em termos de uma medida numérica. Por exemplo, o diâmetro de uma engrenagem pode ser medido com um micrômetro e expresso em milímetros. Uma característica mensurável única, como, dimensão, peso, ou volume é chamada: variável. Carta de Controle de Variáveis são usadas intensivamente. Elas freqüentemente resultam em processos de controle mais eficientes e fornecem mais informações a respeito da performance do processo que as Cartas de Controle de Atributos.

Quando lidamos com uma característica de qualidade que é uma variável, uma prática comum é controlar tanto a média da característica de qualidade quanto sua variabilidade. Para controlar a média do processo ou o nível médio de qualidade é usualmente utilizada a Carta de Controle para Médias, ou $\overline{X}$ Chart. A variabilidade dos processos ou dispersões pode ser controlada tanto pela Carta de Controle para Desvios Padrão, chamado S Chart quanto pela Carta de Controle de Faixas ou Amplitudes, chamada R Chart. O R Chart é amplamente utilizado. Habitualmente, as Cartas de Controle $\overline{X}$ e R são mantidas para cada característica de qualidade que nos interesse. Os gráficos $\overline{X}$ e R (ou S) estão entre os mais importantes e úteis nas técnicas de Controle Estatístico de Processos.



Fonte: MONTGOMERY, 1991

Fig. 13 – Curvas de Distribuição

Na figura (a) acima tanto a média μ e o desvio padrão σ estão sob controle em seus valores nominais (digo μ_0 e σ_0). Contudo, na figura (b) acima o valor da média foi empurrado para um valor de $\mu_0 > \mu_1$, resultando em um produto com alta fração de não conformidade. Na figura c, acima, o desvio do processo foi empurrado para o valor de $\sigma_1 > \sigma_0$. Isto resulta em uma alta discrepância, embora a média do processo permaneça em seu valor nominal.

3.1.2.3.1. Cartas de Controle para $\overline{X}$ e R

Bases Estatísticas para os Gráficos

Suponha que a característica de qualidade é normalmente distribuída com média μ e desvio padrão σ, onde μ e σ sejam conhecidos. Se $x_1, x_2, ..., x_n$ é a amostra de tamanho n, a média desta amostra será

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

e sabemos que $\bar{x}$ é normalmente distribuído com média μ e desvio padrão $\sigma_{\bar{x}} = \sigma / \sqrt{n}$. Além disso, a probabilidade será 1-α que qualquer amostra caia entre

$$\mu + Z_{\alpha/2}\sigma_{\bar{x}} = \mu + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{e} \quad \mu - Z_{\alpha/2}\sigma_{\bar{x}} = \mu - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Então, se μ e σ são conhecidos, as equações acima podem ser usadas como limite de controle superior e limite de controle inferior de uma Carta de Controle de Médias. É costume substituir o valor de $Z_{\alpha/2}$ por 3, de tal forma que os limites 3-sigma sejam empregados. Se a média amostral cair fora dos limites, isso é a indicação que a média do processo não mais eqüivale a μ.

Nós assumimos que a distribuição da característica de qualidade é normal. Contudo, os resultados acima continuam aproximadamente corretos mesmo que a distribuição subjacente não seja normal, por causa do teorema do limite central.

Na prática, comumente, não sabemos o valor de μ e σ. Assim, eles devem ser estimados a partir de amostras preliminares coletadas quando pensamos que o processo está sob controle. Estas estimativas, devem ser baseadas em amostras de 20 a 25 unidades pelo menos. Suponhamos que a média das amostras esteja disponível, cada uma contendo n observações da característica de qualidade; tipicamente, o tamanho de n será pequeno, freqüentemente 4, 5 ou 6. Estes pequenos grupos amostrais resultam na construção de subgrupos, isto pelo fato de do custo de amostragem e inspeção associados as variáveis mensuradas ser alto. Façamos $\bar{x}_1, \bar{x}_2, ..., \bar{x}_m$ serem as médias de cada uma das amostras. Então o melhor estimador de μ , a média do processo, é a média principal, definida como

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + ... + \bar{X}_m}{m}$$

Deste modo, $\overline{\overline{x}}$ seria usado como a linha central do Gráfico $\overline{X}$.

Para construir os limites de controle, precisaremos de um estimador para o desvio padrão σ. Podemos calcular σ tanto pelos desvios padrão quanto pelas faixas (ou amplitude) das média amostras. Por agora, nos concentraremos no método das faixas. Se $x_1, x_2,....,x_n$ é uma amostra de tamanho n, a amplitude amostral será a diferença entre o maior e o menor valor das observações, isto é,

$R = x_{max} - x_{min}$

Existe uma relação conhecida entre a amplitude de uma amostra e a distribuição normal e o desvio padrão desta distribuição. A variável aleatória W=R/ σ, chamada *amplitude relativa* . Os parâmetros da distribuição de W são uma função do tamanho de amostra n. A média de w é d2. Consequentemente, um estimador para σ é

$$\hat{\sigma} = \frac{R}{d_2}$$

Façamos $R_1, R_2,...,R_3$ serem as amplitudes de m amostras. A amplitude média é

$$R = \frac{R_1 + R_2 + ... + R_m}{m}$$

Então, o estimador de s será calculado como:

$$\hat{\sigma} = \frac{\overline{R}}{d_2}$$

Se o tamanho da amostra for relativamente pequeno, renderá um bom estimador da variância como o estimador quadrático usual (a variância amostral S^2)

Para valores de n moderados, digo n≥10, a amplitude perde eficiência rapidamente, visto que ignora todas as informações da amostra entre $x_{max} - x_{min}$. Contudo, para

pequenos tamanhos de amostras freqüentemente empregados em Carta de Controle de Variáveis (n=4,5, ou6), é completamente satisfatória.

Se usarmos $\overline{\overline{x}}$ como um estimador de μ e $\overline{R}/d_2$ como um estimador de σ, então os parâmetros do Gráfico $\overline{X}$ são

$$UCL = \overline{\overline{x}} + \frac{3}{d_2\sqrt{n}}\overline{R}$$
$$\text{Linha Central} = \overline{\overline{x}}$$
$$LCL = \overline{\overline{x}} - \frac{3}{d_2\sqrt{n}}\overline{R}$$

Notemos que a quantidade

$$A_2 = \frac{3}{d_2\sqrt{n}}$$

é uma constante que depende somente do tamanho da amostra, logo é possível re-escrever

$$UCL = \overline{\overline{x}} + A_2\overline{R}$$
$$\text{Linha Central} = \overline{\overline{x}}$$
$$LCL = \overline{\overline{x}} - A_2\overline{R}$$

Vimos que a amplitude amostra é relacionada com o desvio padrão do processo. Então, a variabilidade processual pode ser controlada plotando os valores de R a partir das amostras sucessivas na Carta de Controle . Este gráfico de controle é chamado R Chart.

Os parâmetros de R podem ser facilmente determinados. A linha central será $\overline{\overline{R}}$

3.2.Conclusão

As técnicas (MONTGOMERY, 1991) apresentadas neste capítulo, representam os requisitos necessários para que possamos articular, de forma eficiente e objetiva, o

emprego de técnicas estatísticas no uso de SGBDRs a fim de garantir uma melhor análise de dados quanto a sua qualidade. Definimos as técnicas de CEP (Controle Estatístico de Processos), através das Cartas de Controle – X-Bar, R Chart, P Chart, np Chart, C Chart e U Chart para o controle quantitativo e qualitativo das não-conformidades apresentadas pelos dados.

Todos os métodos aqui apresentados deverão fazer parte da nossa aplicação, bem como o uso de metadados no controle de atributos definidos nos projetos das bases de dados. Além disso, nossa aplicação deverá contar com o uso de árvores de decisão, em conjunto com banco de dados, como solução ao Tutor que auxiliará os usuários na escolha da melhor técnica para o seu problema.

Capítulo 4 - Arquitetura de Qualidade de Dados para Controle de Qualidade em Bancos de Dados

Neste capítulo iremos apresentar a uma proposta de arquitetura para o controle de qualidade em banco de dados. Nossa proposta contempla a utilização de tutores inteligentes que orientem os usuários na escolha das técnicas apresentadas anteriormente na busca de soluções.

Esta arquitetura compreende a utilização de árvores de decisão a fim de proporcionar uma estrutura de escolha, baseada em perguntas e respostas, encaminhando o usuário dito "leigo" para a melhor técnica de análise da qualidade dos dados associados aos SGBDRs, além de medidas de controle estatístico na busca da garantia de qualidade.

A proposta compreende, ainda, a definição de uma camada para implementação de novas técnicas que poderão ser implementadas em futuros trabalhos, garantindo assim uma estrutura de independência e flexibilidade à aplicação. Assim sendo, estamos prevendo o uso de uma interface gráfica amigável e independente para as interações usuário-máquina e a realização de consultas e extrações de dados.

4.1. Concepção

A concepção da arquitetura proposta nesta teste está dividida em macro funções, conforme o diagrama abaixo. Uma das funções executa a gerência dos recursos do banco de dados, formada pela camada de controle do banco de dados; uma outra é responsável pelo assistente orientador de métodos e outra pelo assistente estatístico.

Pelo ambiente proposto o usuário poderá utilizar a ferramenta para consultar estruturalmente o banco de dados, bem como explorar seu conteúdo, além dos descritores de informações a respeito de cada uma das bases armazenadas, organizadas pelo sistema de metadados.

Pode-se utilizar filtros para a seleção de informações, de acordo com as dimensões desejadas - espaciais e temporais, assim como a ferramenta através de um módulo de análise de base de dados, onde teremos 5 funcionalidades determinantes:

i. Manipulação de associatividades entre a camada do banco de dados e o banco em sistema de informação – Camada de Acesso
ii. Exploração visual da estrutura da base de dados
iii. Exploração visual da estrutura do conteúdo de dados
iv. Exploração de metadados
v. Tutores

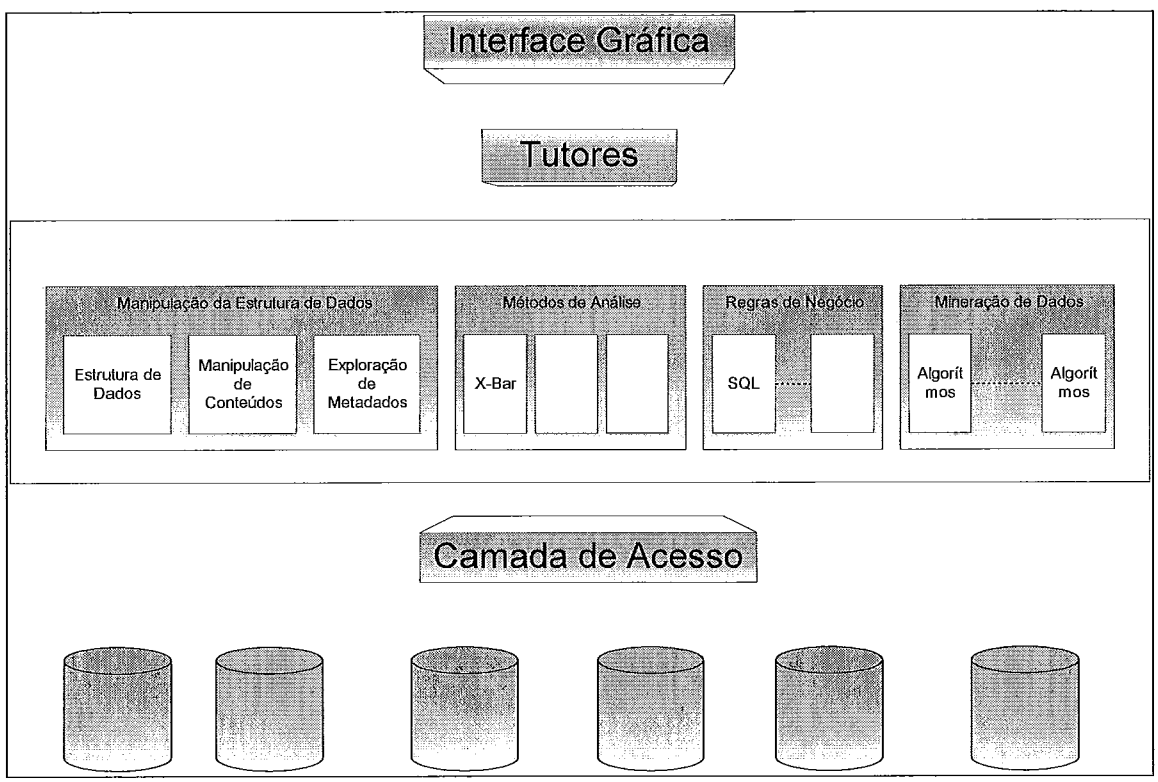


Fig. 14 - Arquitetura de Qualidade de Dados

4.1.1. Camada de Acesso

Durante as pesquisas a respeito de como tornar mais fácil o controle da camada entre a aplicação e o SGBDR, percebemos que alguns autores fazem referência ao uso de bancos de dados específicos, isto é, ou só utilizam um tipo de SGBDR, ou desenvolvem

filtros de importação de dados. Todavia, esta prática limita o escopo de nossa aplicação, pois se desejamos controlar a qualidade dos dados advindos de quaisquer áreas, não podemos limitar o acesso a um determinado SGBDR nem tão pouco duplicar os dados através de filtros. Além disso, para o escopo desta tese não justificaria escrevermos uma camada que fizesse o tratamento de dados entre a aplicação e o SGBDR, nem do ponto de vista local nem da forma cliente/servidor, já que nosso objetivo é o controle de qualidade dos dados para usuários leigos.

Visto desta forma, apresentamos uma camada de uso consagrado que nos possibilita um conjunto rico e robusto de funcionalidades na assistência do desenvolvimento de aplicações cliente/servidor de Banco de Dados. Sua arquitetura de conexão de banco de dados nos permite incluir vários serviços compartilhados utilizados pelas conexões de bancos de dados e outras funções. Adotando esta arquitetura, poderemos trabalhar com conexões Microsoft ODBC quando necessário, além do acesso a servidores de bancos de dados como Informix, DB2, Oracle e Sybase. Juntamente com suas conexões de banco de dados e uma biblioteca de funções consistente, esta Camada fornece aos desenvolvedores de aplicações para Microsoft Windows 9x e NT um acesso direto, rápido, limpo, separado e compartilhado de múltiplas fontes de dados.

4.1.2. Exploração Visual da Estrutura da Base de Dados

Esta funcionalidade torna-se de fundamental importância em nossa arquitetura, visto que o usuário final de uma aplicação nem sempre consegue visualizar os dados por ele inseridos em um sistema, de tal forma que a garantia posterior deste dado fica comprometida se observarmos que a maioria das aplicações desenvolvidas não possui módulos ou operações de controle de qualidade de dados.

Pensando nisto, propomos esta funcionalidade que capacitará a investigação exploratória através da escolha de suas bases de dados associadas a pseudônimos, a fim de facilitarmos a gerência por parte dos administradores de bancos de dados e demais analistas. Neste módulo, o usuário poderá verificar a definição dos atributos e seus domínios, podendo dimensionar algumas ocorrências de erros a partir de problemas de definição dos atributos e de como estes se relacionam no contexto da aplicação.

Acompanhando esta funcionalidade, forneceremos ao usuário conhecedor de SQL uma opção investigatória dos dados sem contudo possibilitarmos a manipulação destes. Além disso, em associatividade com a camada de conexão com bancos de dados, forneceremos um maior controle sob os aspectos da estrutura de controle e acesso aos dados e a performance dos SGBDRs.

4.1.3. Exploração Visual do Conteúdo Base de Dados

Esta funcionalidade se faz presente, pois nem sempre se torna possível o acesso através das aplicações primárias, o conhecimento do dado diretamente de sua fonte. Muitas vezes, ao tentarmos utilizar aplicações exploratórias, esbarramos com problemas de duplicação da base e de utilização de filtros de importação associados a aplicações de terceiros. Assim sendo, recomendamos a exploração visual dos dados de duas formas.

A primeira, consiste em apresentações gráficas explorando e investigando os dados e suas associatividades com outras variáveis. Nesta funcionalidade, desenvolveremos o uso de ferramentas gráficas a fim de introduzirmos o usuário leigo no uso das demais ferramentas de análise mais específicas. O usuário deverá poder selecionar, a partir do banco de dados, quais atributos ele irá explorar visualmente através de gráficos.

Outra forma de exploração do conteúdo da fonte de dados é a análise das tabelas do banco de dados de forma tabular, apresentando algumas informações de conteúdo relevante no auxílio e verificação do registro destes dados, feitos por outros sistemas. Neste caso, a arquitetura do sistema deverá permitir o crescimento de novas funcionalidades com o intuito de oferecer novos mecanismos de investigação, proporcionando uma análise rápida e concisa de todas as informações registradas pelas aplicações de terceiro junto aos bancos de dados.

4.1.4. Exploração de Metadados

Ao falarmos em qualidade de dados, estamos preocupados em explorar toda e qualquer informação a respeito dos dados em questão. Para isto, é necessário que tenhamos como armazenar estas informações.

Desenvolvemos o conceito de integração dos metadados à nossa arquitetura com o objetivo de não só explorar os dados univariadamente, mas colocá-los em um contexto espaço-temporal. A aplicação de controle de qualidade de dados deverá conter um controle de metadados, que poderá ser de uso coletivo ou específico do administrador do sistema, que poderá atualizar a base de dados para a realização de consultas, podendo gravar no final uma nova base de dados, a base do usuário, composta com objetos definidos pelos atributos selecionados por ele.

A base de consultas do usuário é utilizada na armazenagem dos critérios para a seleção de informações. Desta forma, o registro destas informações permitirá sua re-utilização em diversas consultas. A definição destas listas deverá ser de forma intuitiva, explorando a capacidade do sistema de instruir o usuário em suas operações.
As consultas deverão ser elaboradas em seus módulos de origem e depois documentadas pelo usuário na metabase.

A metabase do sistema deverá conter informações a respeito dos dados, incluindo as descrições dos nomes das relações, domínios dos atributos(tipos de dados), nomes dos atributos, chaves - primária, secundária e estrangeira e outros tipos de restrições, além de descrições a respeito dos filtros e visões.

Nossa arquitetura pode comportar outras metabases:

i. Informações a respeito das chaves
RELATION_KEYS→ [REL_NAME, KEY_NUMBER, MEMBER_ATTR]

ii. Informações sobre os índices
RELATION_INDEXES→ [REL_NAME, INDEX_NAME,
MEMBER_ATTR, INDEX_TYPE, ATTR_NO, ASC_DESC]

iii. Informações a respeito dos filtros e visões
VIEW_QUERIES→ [VIEW_NAME, QUERY]
VIEW_ATTRIBUTES→ [VIEW_NAME, ATTR_NAME, ATTR_NUM],
onde armazena os nomes dos atributos da visão, onde ATTR_NUM >0 especificando a correspondência de cada atributo da visão aos atributos do resultado da query.

Exemplo da metabase principal:

REL_NAME	ALIAS	DATA_CR	USUARIO	DESCRIÇÃO	OBS
EMPREGADO	ORACLE2	12/02/2001	DBA	Tabela de empregados da empresa EXXON Ltda.	Atualização mensal
PROJETO	INFORMIX3	20/01/2000	FLAVIO	Pesquisa sobre índices ao consumidor IBGE	Atualização quinzenal

4.1.5. Tutores

O mecanismo de tutores, proposto em nossa arquitetura, foi motivado pelo fato da não especialização técnica do usuário quando se trata da investigação e adequação da qualidade dos dados. Percebemos que em muitos casos, o usuário final, possui apenas o conhecimento de uma aplicação, isto é, para ele, toda técnica está armazenada em uma caixa preta, onde há uma entrada de dados e uma saída. Nossa primeira proposta ao desenvolver uma aplicação piloto, foi anexar um guia como um manual, a fim de capacitar o usuário. Contudo, percebemos que muitos conceitos deveriam ser passados, visto que, para cada abordagem, poder-se-ia utilizar uma técnica.

Descartamos o modelo de arquitetura que utilizasse um manual de operações, em prol de uma arquitetura mais dinâmica e robusta através da qual pudéssemos explorar as funcionalidades do conhecimento humano, conduzindo o usuário para a melhor técnica.

Após abraçarmos a idéia do uso de técnicas de aprendizado de máquina, cabe-nos escolher qual abordagem técnica se encaixaria melhor no nosso problema.

Problema: usuário leigo com pouca experiência de informática, sem os conceitos das técnicas de investigação da Qualidade dos Dados.

Visando tal meta, dividimos nosso problema em dois: um dos Tutores deverá identificar qual a melhor técnica para cada problema levantado junto ao usuário. Serão feitas perguntas que encaminharão o usuário para um conjunto de técnicas ou uma técnica específica, ou ainda, a solicitar o apoio de uma nova ferramenta. Assim, estaremos

garantindo que na falta do especialista, os passos deste usuário no sistema serão acompanhados pela figura do especialista virtual, a partir de agora denominado - Tutor de Métodos.

O Tutor de Métodos possuirá uma arquitetura expansível, isto é, novos conhecimentos poderão ser introduzidos por um especialista com o intuito de tornar mais precisas suas orientações.

Deveremos associar o Tutor de Métodos à tecnologia de banco de dados, porque, este método irá nos proporcionar maior rapidez, flexibilidade e adequação do Tutor de Métodos às futuras implementações que poderão ser feitas no sistema.

Definido o escopo do nosso problema, pesquisamos diversas técnicas e apontamos uma que atende perfeitamente às características de análise de qualidade de dados.

Identificamos que o nosso problema poderia ser melhor resolvido através de uma técnica de aprendizado de máquina, visto serem técnicas que possibilitam à máquina "melhorar" com a experiência adquirida. Muitas aplicações desenvolvidas nesta área comprovam, com sucesso, sua aplicabilidade em vários segmentos → mineração de dados , filtragem e investigação de informações associadas às preferências dos usuários, veículos autônomos etc.

Na arquitetura proposta cabe um sistema especialista. Contudo, o processo de Árvore de Decisão foi o método escolhido para nos auxiliar na implantação do Tutor de Métodos de nossa arquitetura, já que atende perfeitamente às características de análise de qualidade de dados. A implementação deste método poderá ser expandida para outras técnicas de IA de acordo com o grau de complexidade do projeto. Para o escopo desta tese, implementaremos o Tutor de Métodos baseado na Árvore de Decisão.

Árvores de Decisão

O método de árvores de decisão (MITCHEL, 1997) é um dos métodos mais extensamente utilizado para inferência indutiva. Este é um método de aproximação de funções de valores discretos em funções de aprendizado representadas por uma árvore de decisão.

Este método é amplamente utilizado em vários segmentos de pesquisa: área médica, industrial, P&D etc (MITCHEL, 1997) .

As árvores de decisão classificam as instâncias ordenando de cima para baixo na árvore, isto é, da raiz ao nó da folha, que fornece a classificação da instância. Cada nó na árvore fornece uma especificação de um teste de algum atributo da instância e sua descendência → nó correspondente a um possível valor deste atributo. A instância é classificada inicialmente a partir da raiz, testando o atributo especificado por este nó, descendo pelo ramo correspondente ao valor do atributo dado.

De forma geral, as árvores de decisão representam a disjunção das conjunções das restrições dos valores dos atributos das instâncias. Cada caminho da árvore, a partir da raiz, corresponde a conjunção dos atributos de teste e a árvore propriamente, à disjunção destas conjunções.

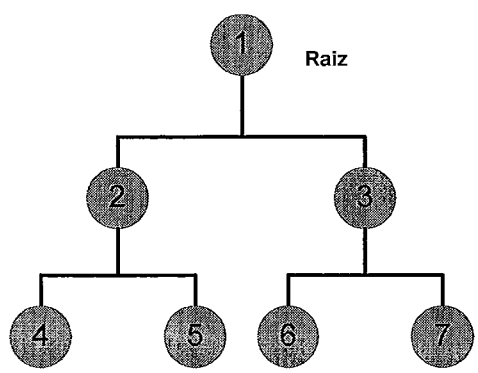


Fig. 15 - Árvore de Decisão

Problemas Apropriados para Árvore de Decisão

Existe uma variedade de métodos de aprendizagem de árvores de decisão (MITCHEL, 1997) , que se diferenciam uns dos outros, baseados nos seus requerimentos e diferentes capacidades.

- As instâncias são representadas pelos pares de atributos e descritas a partir de um conjunto fixo de atributos (ex.: Temperatura) e seus valores (ex.: Quente). A situação mais simples para o aprendizado da Árvore de Decisão é quando cada atributo possui poucos valores disjuntos possíveis (Ex: Quente, Aprazível e Frio)
- Função alvo possui valores de saída discretos. A Árvore de Decisão determina valores booleanos (ex.: sim ou não) para cada exemplo. Os métodos de Árvore de Decisão podem estender as funções de aprendizado para mais de dois valores de resposta possíveis.
- Descrições disjuntas podem ser necessárias. Como acima, Árvores de Decisão naturalmente representam expressões disjuntivas.
- Os dados de treinamento podem conter erros. Os métodos de aprendizado por Árvores de Decisão são robustos com relação a erros.

Os dados de treinamento podem conter missing values nos atributos. Os métodos de Árvores de Decisão podem ser utilizados até quando existem falta de valores em alguns exemplos de treinamento.

A cada nó da árvore de decisão será associado um código de pergunta.Este, por sua vez, será associado aos ramos filhos e assim por diante. Desta forma, ao se deparar com uma pergunta, o usuário poderá optar por n possíveis respostas. Não existe reposta certa ou errada, pois o que desejamos é a orientação do usuário para um conjunto de técnicas.

Durante a seleção de possíveis respostas, serão apresentados ao usuário, textos complementares explicativos a respeito da pergunta ou técnica escolhida.

A estrutura geral do Tutor de Métodos está associada à um banco de dados, onde contemplaremos as seguintes tabelas:

i. TAB_ARVORE
ii. TAB_QUESTÕES

Na tabela TAB_ARVORE associaremos os seguintes atributos:

i. NO - nó origem da pergunta

ii. RAMO - o ramo da árvore associado à pergunta correspondente

Exemplo

Nó	Ramo
1	2
1	3
2	4
2	5
3	6

No exemplo acima, podemos perceber que não existe nenhum número inferior ao do nó 1 em nossa estrutura de Árvore, simbolizando a sua raiz. A partir deste ponto, o usuário será argüido com mais duas questões - números 2 e 3. Caso opte pela de número 2, o usuário será levado para o ramo 2 da Árvore onde deverá selecionar uma das 2 possíveis respostas - números 4, 5 e assim sucessivamente. No exemplo abaixo, poderemos ver uma representação gráfica desta abordagem.

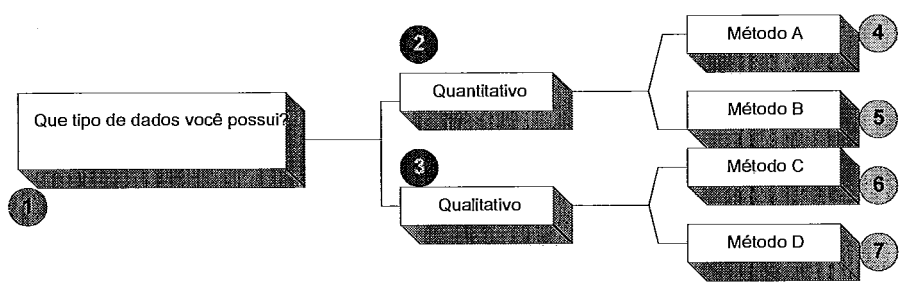


Fig. 16 - Tutor de Métodos

Além da tabela de nós da Árvore de decisão - TAB_ARVORE, implementamos a TAB_QUESTOES. Esta última, corresponde às questões apresentadas ao usuário, além das explicações e orientações quanto ao conteúdo de cada uma.

A apresentação da tabela, vide Anexo A, serve como guia e referência a respeito da implementação e objetivos destas técnicas. Deste jeito, conseguiremos expandir o

conteúdo do Tutor de Métodos de maneira rápida, ampliando assim o grau de fidelidade das respostas.

A funcionalidade do Tutor de Métodos não poderia ser utilizada mesma forma para o acompanhamento das análises feitas pelos usuários empregando as técnicas estatísticas da Qualidade dos Dados propostas nesta tese. A solução será a construção de um Tutor Estatístico, que terá a função técnica de analista das estatísticas geradas pelos métodos de Qualidade dos Dados. Nossa arquitetura permite que o Tutor Estatístico aconselhe o usuário leigo no que diz respeito aos valores obtidos por cada técnica e, ainda, na escolha de uma dada técnica, este indique sua aplicabilidade.

Com relação às técnicas de Qualidade de Dados aqui apresentadas, implementaremos filtros de condições junto às técnicas apresentadas, assim sendo, poderemos não só enriquecer a análise, bem como conduzir a análise dos dados usuário, evitando assim erros comuns, como o emprego de técnicas erradas ou, até mesmo, conclusões errôneas.

No caso dos métodos de Qualidade de Dados foi proposto o uso de controle de bandas, onde serão criados limites superiores e inferiores de controle, a fim de assegurarmos o controle dos processos que serão analisados. Sendo assim, deveremos ao utilizar esta técnica, orientar o usuário no momento em que um determinado ponto esteja fora das especificações técnicas, para que a partir deste momento ele possa identificar a raiz do problema, implementar uma ação corretiva externa, e por fim realizar novas verificações e acompanhamentos.

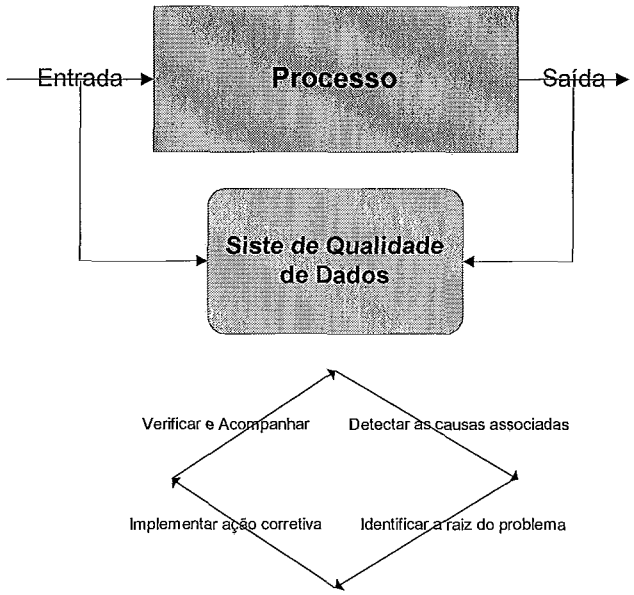
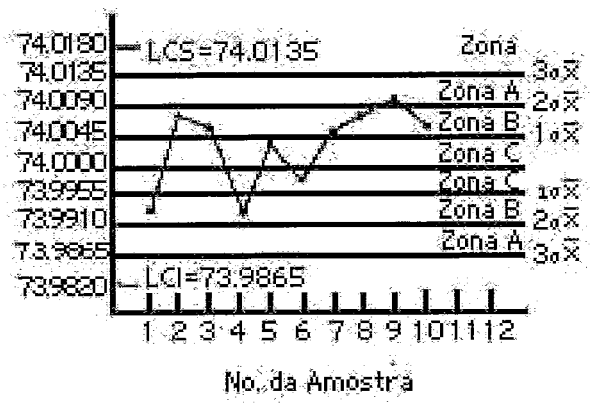


Fig. 17 - Melhoria de processo utilizando as técnicas de Qualidade de Dados

Alguns analistas sugerem a utilização de dois conjuntos de limites de controle, como mostraremos na figura abaixo. Os limites externos, ditos 3-sigma, são usualmente limites de ação, isto é, quando um ponto fica fora destes limites, uma busca da causa associada deve ser feita e uma ação corretiva deve ser tomada, se necessário. Os limites internos, usualmente 2-sigma, são chamados de limites de aviso. Caso um ou mais pontos caiam entre os limites de aviso e os limites de controle, ou muito próximos dos limites de aviso, podemos suspeitar que o processo não está adequadamente operacional.

Limites de aviso aumentam a sensibilidade dos gráficos de controle. Sua desvantagem é que eles não possuem uma interpretação precisa e podem confundir o usuário, mas esta não é uma objeção séria.

É comum nos EUA, independentemente da distribuição de qualidade característica, determinar os limites de controle como valores múltiplos do desvio padrão. O múltiplo usualmente escolhido é 3. Conseqüentemente, os limites 3-sigma são normalmente empregados nos gráficos de controle, independente do tipo do gráfico utilizado.

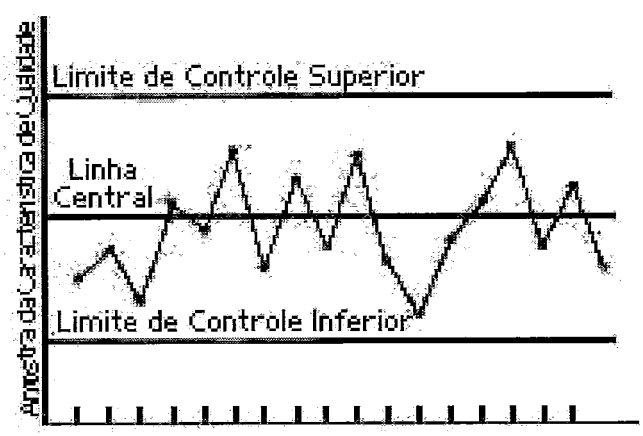


Fonte: MONTGOMERY, 1991

Fig. 18 - Zonas de Limite de Controle e Avisos - Padrão Americano e Inglês

O Reino Unido e partes da Europa utilizam limites de probabilidade, cujo limite padrão de probabilidade é 0,001.

Nós justificamos o uso dos limites de controle 3-sigma com base nos resultados obtidos na prática, visto que, a verdadeira distribuição da característica de qualidade não é bem conhecida a fim de calcularmos os limites de probabilidade exatos.



Fonte: MONTGOMERY, 1991

Fig. 19 - Limites de Controle Adotados 3-sigma

Critérios de Controle

Muitos critérios podem ser aplicados de forma simultânea em um gráfico de controle para determinar quando o processo está fora de controle. O critério básico implementado é quando um ou mais pontos estiverem fora dos limites de controle. Existem outros critérios suplementares que poderão aumentar a sensibilidade do controle estatístico dos dados, visto que poderia responder mais rapidamente à mudança às causas associadas. Algumas das regras de sensibilidade mais comuns são:

i. Um ou mais pontos fora do limite de controle
ii. Uma sucessão de pelo menos 8 pontos, onde o tipo de tendência pode ser positivo ou negativo, acima ou abaixo da linha central ou ainda, acima ou abaixo da mediana,
iii. Dois de três pontos consecutivos fora da linha de 2-sigmas, mas dentro dos limites de controle,
iv. Quatro de cinco pontos consecutivos além dos limites de 1-sigma,
v. Padrões incomuns ou não aleatórios dos dados,
vi. Um ou mais pontos próximos aos limites de aviso e de controle.

Estas regras são empíricas. Com o hábito da observação, podemos aplicar algumas destas situações a fim de aprimorarmos nossa análise. Contudo, como este conjunto de regras não tem comprovação científica, e conseqüentemente é de difícil análise para o leigo, implementamos o Tutor Estatístico com base nos Limites de Controle Superior (LCS) e inferior (LCI) para os controles de banda, indicativos da performance do processo analisado.

4.1.6. Regras de Negócios

A regra de negócio foi definida (DATE, 2000) como "uma frase que define ou restringe alguns aspectos do negócio. Ela pretende expressar a estrutura do negócio ou controlar ou influenciar o comportamento dos negócios".

Quando uma regra de negócio é elaborada por uma pessoa, é comum expressa-la ambiguamente, de forma não rigorosa. Nesta situação, cada regra poderia ser decomposta em outras sub-regras. Assim sendo, quando decompostas em sua forma mais elementar se tornariam atômicas (indivisíveis), podendo simbolizar pensamentos completos.

Em resumo, as regras de negócios devem ser:

i. Declarativas (i.e. não procedurais),
ii. Atômicas,
iii. Expressas em linguagem natural,
iv. Orientadas a negócio e não tecnologia

Date (DATE, 2000) propõe a utilização de SQL como padrão para a construção de regras de negócios, justificando sua facilidade de aprendizado e disseminação, entre outros fatores.

A construção de regras de negócios em nossa arquitetura está baseada em SQL a fim de facilitar o aprendizado e a velocidade da construção destas, visto que a estrutura da linguagem é bem próxima a linguagem natural. Um dos problemas enfrentados pelos usuários recai justamente na falta de ferramentas associadas aos seus dados capazes de filtrar e orientar suas análises de forma objetiva e sintática. A disseminação e recuperação destas regras, em SQL, torna-se mais fácil, pois sua estrutura é escrita e estática, possibilitando uma armazenagem externa associada a metadados, gerando a "metadata quality" (THORNTON, 2000).

A arquitetura proposta neste trabalho tem possui características robustas no que se refere à adaptabilidade de novos módulos ou funcionalidades. O módulo de Tutores (*Wizards*) tem como característica fundamental a facilidade de composição e expansão, garantindo-nos um rápido crescimento. A camada do assistente está conectada diretamente a outros dois módulos – Métodos de Análise e Regras de Negócio. O módulo de Métodos de Análise capacita nossa arquitetura com os métodos científicos de análise da qualidade dos dados, como por exemplo: As cartas de controle de atributos e variáveis. Já no módulo de Regras de Negócio, podemos controlar o uso e armazenagem das regras referente à cada banco, habilitando ou não seu acesso a usuário. Através da

armazenagem de estruturas SQL, possibilitamos o uso e re-aproveitamento de regras. Estes dois módulos, em conjunto com o módulo Assistente, servem como base a futuros módulos de Mineração de Dados associados. Utilizamos uma interface entre os SGBDRS e a aplicação, objetivando proporcionar ao usuário contatos com mais bancos, sem a necessidade de importação de bases. Todos os módulos apresentados se conectam ao Módulo de Interface Gráfica garantindo acesso facilitado as funcionalidades do sistema.

Capítulo 5 - Aplicação

Para podermos validar a implementação do estudo feito, criamos uma aplicação exemplo, onde podemos verificar e demonstrar todo seu funcionamento.

A aplicação é uma ferramenta de controle de qualidade de dados na qual foram criadas interfaces que facilitam o acesso do usuário à informação.

5.1. Características

A aplicação exemplo possui uma característica que a difere de outras aplicações de Qualidade de Dados, que se baseiam em questionários é que à primeira vista possuem um grau maior de complexidade para o seu entendimento. Esta aplicação tem como objetivo principal conquistar o usuário pela sua facilidade de uso, visto que em todos os estudos de casos relatados no decorrer do nosso trabalho, observamos que a falta de especialistas em análise de dados, com os conhecimentos necessários para validar estatisticamente o volume crescente dos dados que são diariamente armazenados, nos transformará em meros depositários de dados, mas não de informações.

5.1.1. Introdução a Aplicação – DBMINER

O software **DBMINER** foi desenvolvido objetivando a aplicação das técnicas defendidas nesta tese. Sua estrutura obedece aos parâmetros apresentados e especificados, cujo embasamento teórico está presente entre os primeiros capítulos desta tese.

A aplicação **DBMINER** inclui:

i. **DBMINER** – Estrutura
ii. **DBMINER** – Visualizador de dados
iii. **DBMINER** – Gerenciador de Regras de Negócios
iv. **DBMINER** – Metadados - Explorerador
v. **DBMINER** – Metadados Busca Banco
vi. **DBMINER** – Metadados Busca Estrutura
vii. **DBMINER** – Metadados Cadastra Banco

viii. **DBMINER** – Metadados Cadastra Estrutura

ix. **DBMINER** – Metadados Alias

x. **DBMINER** – Tutores de Métodos

xi. **DBMINER** – Análise Qualidade de Dados

xii. **DBMINER** – Análise Gráfica de Dados

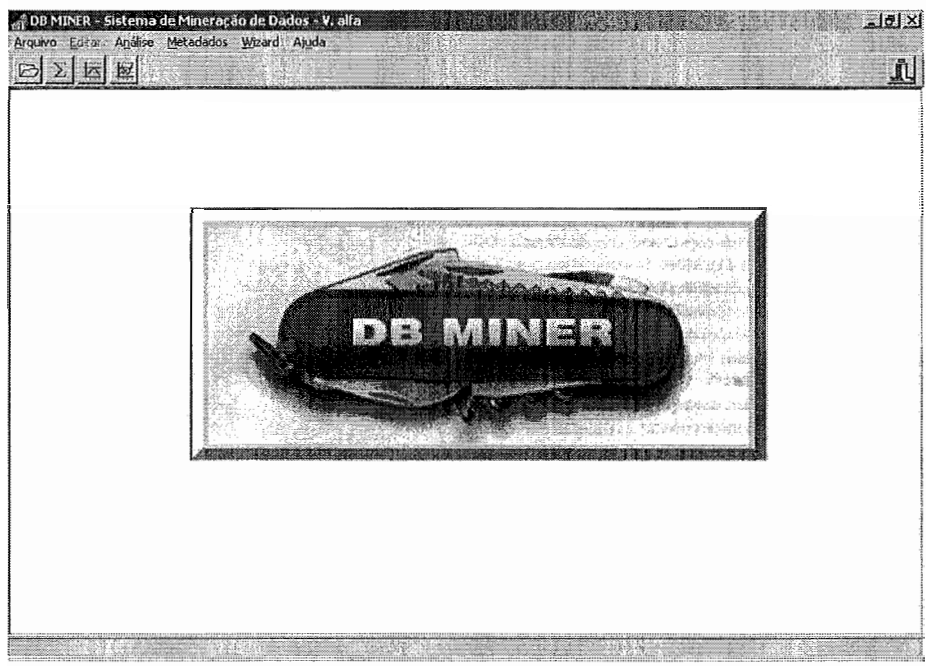


Fig. 20 - Tela de Abertura do Sistema

Nela o usuário poderá encontrar algumas facilidades, como a busca de informações a respeito das configurações dos bancos de dados, a estrutura de suas tabelas, bem como armazenar dados tanto a respeito dos atributos das tabelas e suas relações, como dos bancos.

Visualizador Eficiente de Estrutura de Tabelas

O **DBMINER** – Estrutura de Tabelas, fornece ao usuário a oportunidade de investigação rápida da estrutura das tabelas dos bancos de dados locais ou em rede, Através dele, torna-se possível descobrir o nome dos atributos, seu tamanho, tipo e a qual tabela estes estão vinculados.

Abaixo, encontramos a imagem do módulo, onde o usuário irá selecionar na caixa "Banco de Dados" o banco de dados que deseja identificar. Após esta seleção, aparecerão as tabelas vinculadas, bastando um clique para selecionar a tabela para visualização. No quadro, à direita, irão aparecer listados, obedecendo a ordem de criação, os atributos da tabela selecionada, como também , o seu tamanho.

Para acessar o **DBMINER** – Estrutura de Tabelas, basta clicar em Arquivos → Explorar Dados→Base de Dados.

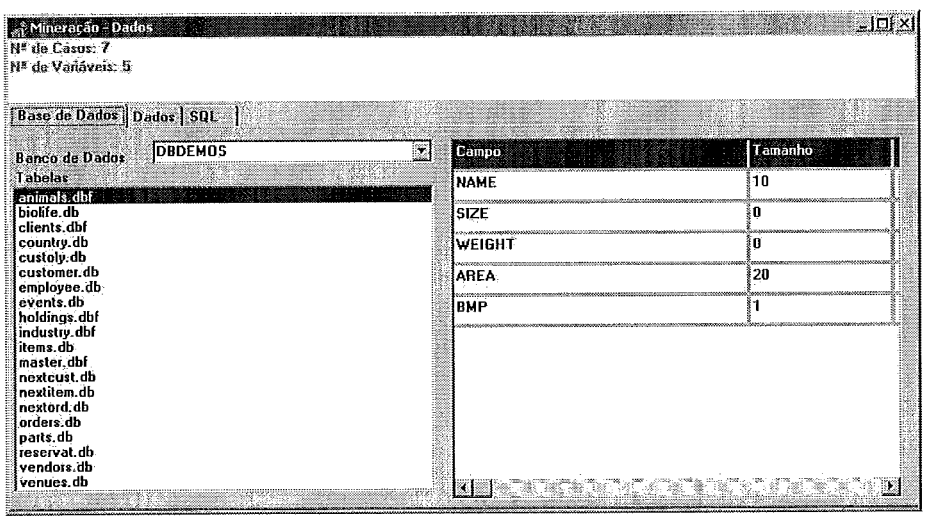
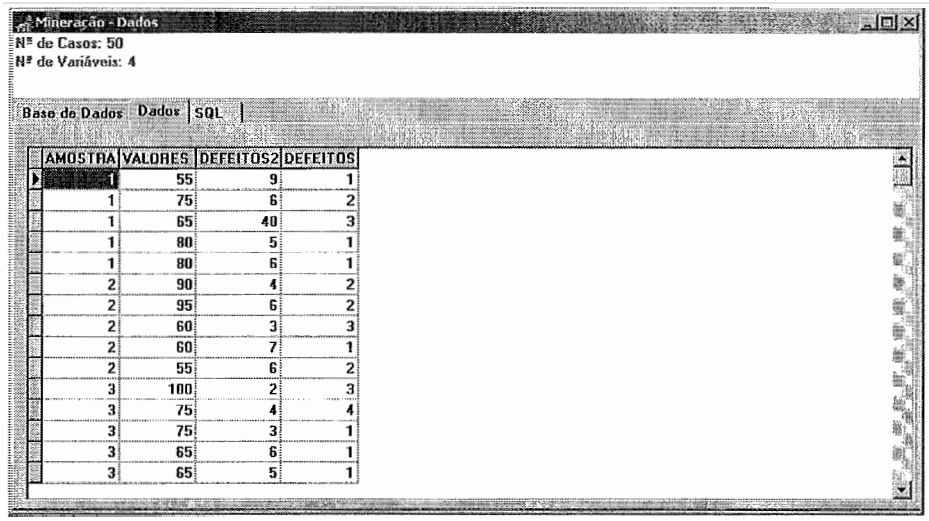


Fig. 21 - Estrutura das Tabelas

Visualizador Prático de Dados

Já o **DBMINER** – Visualizador de Dados permite ao usuário uma visão instantânea de tudo o que estiver acontecendo com os dados de cada uma das tabelas selecionadas. Através do menu Arquivos → Explorar Dados→ Dados, o usuário terá acesso às tabelas, somente com a opção de leitura. Isto se justifica por ser um sistema para pessoas ditas leigas. Desta forma, estas operações não trariam nenhum dano aos bancos e tabelas disponíveis. Veja figura abaixo.



AMOSTRA	VALORES	DEFEITOS2	DEFEITOS
1	55	9	1
1	75	6	2
1	65	40	3
1	80	5	1
1	80	6	1
2	90	4	2
2	95	6	2
2	60	3	3
2	60	7	1
2	55	6	2
3	100	2	3
3	75	4	4
3	75	3	1
3	65	6	1
3	65	5	1

Fig. 22 - Visualização dos Dados

Gerenciador de Regras de Negócios

Este é mais um facilitador o **DBMINER** – Regras de Negócios, uma ferramenta capaz de ajudar o usuário no uso da linguagem SQL para comunicação entre a aplicação e os bancos de dados. Através dele, o usuário poderá montar a sua regra de negócio, visualizando os atributos das tabelas selecionadas, facilitando enormemente suas operações. Para acessá-lo, basta clicar no menu Arquivos → Explorar Dados→ SQL. Aparecerá para o usuário uma lista de campos da tabela, em **Variáveis**, previamente selecionada através do módulo anterior – vide **DBMINER** – Estrutura de Tabelas, um Editor de Regras, e caso interesse ao usuário um ajuste fino em sua regra, poderá faze-lo livremente, além dos botões **Adiciona** e **Roda Query**. O primeiro se propõe a colar as informações dos atributos selecionados no Editor de Consulta, podendo o usuário selecionar a ordem de entrada de qualquer atributo. O botão Roda Query, executa a consulta montada pelo usuário presente no Editor de Consultas.

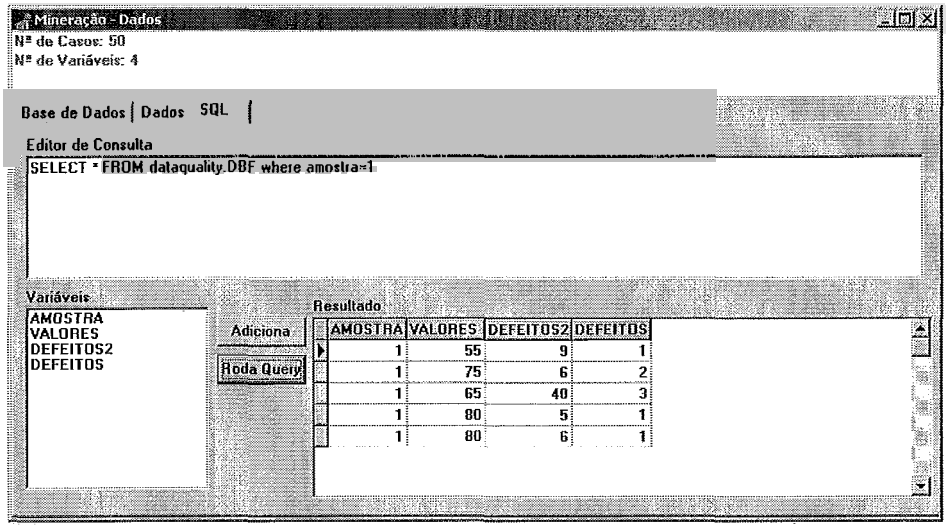


Fig. 23 - Assistente de Consultas - Busca de Dados

Metadados – Visão da Informação

O menu Metadados, apresenta a entrada para os demais módulos do DBMINER. Torna-se possível ao usuário acessar:

i. **DBMINER** – Metadados Explorer
ii. **DBMINER** – Metadados Busca Banco
iii. **DBMINER** – Metadados Busca Estrutura
iv. **DBMINER** – Metadados Cadastra Banco
v. **DBMINER** – Metadados Cadastra Estrutura

Fig. 24 - Metadados - Menu

Explorar as Facilidades

O **DBMINER** – Metadados Explorer permite ao usuário o total conhecimento da estrutura organizacional do seu banco, visualizando o nome de tabelas, atributos e índices que a compõem. Isto se verifica extremamente funcional, ao querermos conhecer como o banco e suas tabelas se relacionam. Formado por uma estrutura de árvore, o **DBMINER** – Metadados Explorer permite navegação hierarquizada, bastando apenas um clique no objeto de interesse.

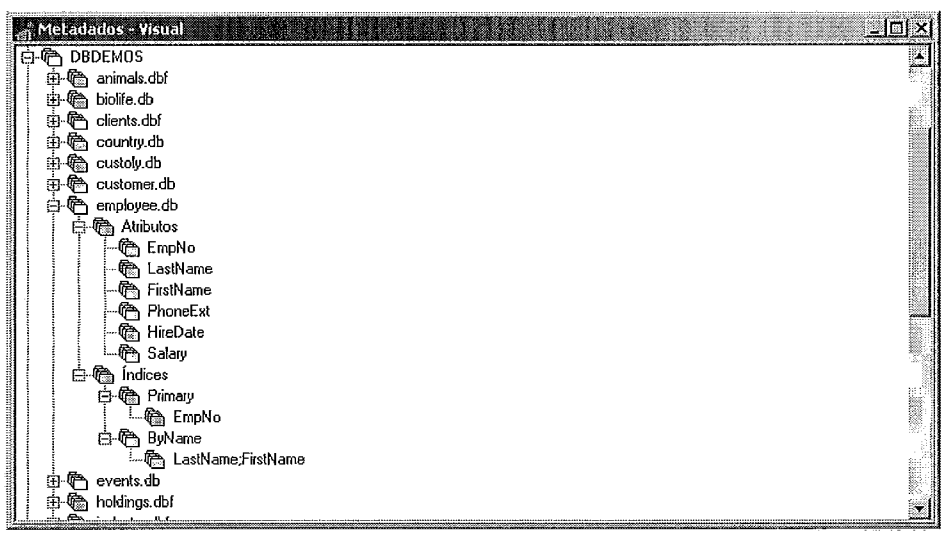


Fig. 25 - Metadados - Explorer

Busca por Dados

O módulo **DBMINER** – Metadados Busca Banco, permite-nos investigar o nome das relações com o banco, o alias definido pela camada do banco, o nome do SGBDR associado, a data de criação, o usuário responsável e a descrição. Todos conforme tabela abaixo:

Campos	Informações
REL_NAME	Nome das relações
ALIAS	Alias associado ao BD
BANCO	SGBDR

DATA_CR	Data de criação do banco
USUÁRIO	DBA
DESCRIÇÃO	Descrição da Relação
OBS	Observações gerais

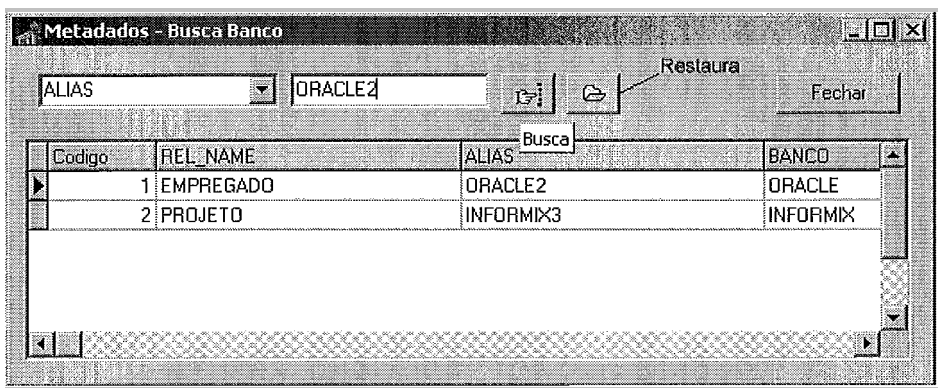


Fig. 26 - Metadados – Busca Dados do Banco

A primeira caixa de combinação oferece o tipo de informação que o usuário selecionar para investigar. Na caixa de texto ao lado, o usuário deverá digitar o dado que procura. Para executar a procura, bastar clicar sobre o ícone com a figura de uma mão para que se proceder a busca. Ao final da busca, caso seja de interesse do usuário realizar uma nova consulta, basta clicar no ícone **Restaura,** apontado na figura acima.

Metadados - Investigando a Estrutura

O módulo **DBMINER** – Metadados Busca Estrutura, fornece ao usuário uma forma rápida de investigação da estrutura associada a uma Relação, isto é, seu nome, seus atributos, tipo, tamanho, se o atributo pertence à chave primária ou a uma chave estrangeira e neste último caso a relação associada. Veja tabela abaixo:

Dados	**Informações**
ALIAS	Alias associado ao BD
REL_NAME	Nome da relação
ATTR_NAME	Nome do atributo

ATTR_TYPE	Tipo do atributo
ATTR_SIZE	Tamanho do atributo
MEMBER_OF_FK	Membro de chave estrangeira
MEMBER_OF_PK	Membro de chave primária
FK_RELATION	Relação associada a chave estrangeira

Abaixo, vemos a figura representativa deste módulo, onde as regras de busca e restauração de dados obedece as mesmas características do módulo anterior.

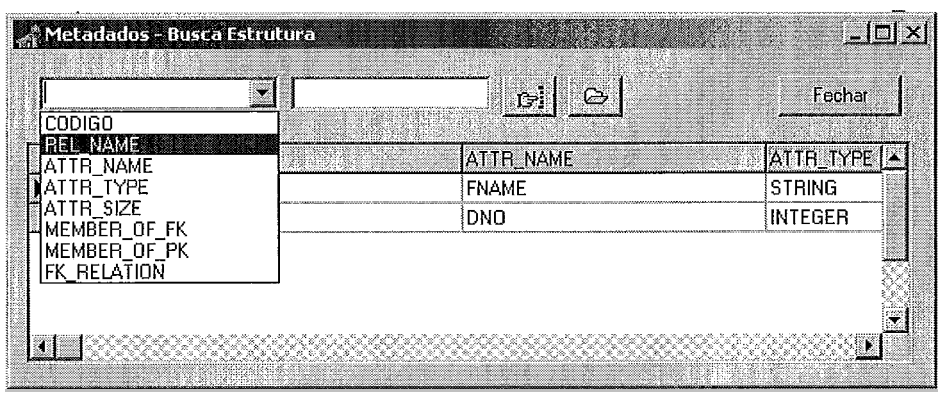


Fig. 27 - Metadados – Busca Estrutura

Metadados – Cadastramento de Bancos

Com o **DBMINER** – Metadados Cadastra Banco, torna-se possível cadastrar e alterar todos os dados cadastrados na base de metadados do DBMINER. Através dele podemos incluir dados a respeito do nome da relação, o alias associado ao banco, a sua data de criação, o seu usuário responsável – DBA, a descrição de suas características e demais observações a respeito. Vide exemplo abaixo. Nele criamos um registro contendo uma relação chamada **PROJETO**, associada a um alias chamado **Informix3**, com data de criação de **20/01/2000**, pelo usuário **Flávio**, tendo como descrição: **"PESQUISA SOBRE ÍNDICES AO CONSUMIDOR IBGE"**, cuja observação é **"ATUALIZAÇÃO QUINZENAL"**.

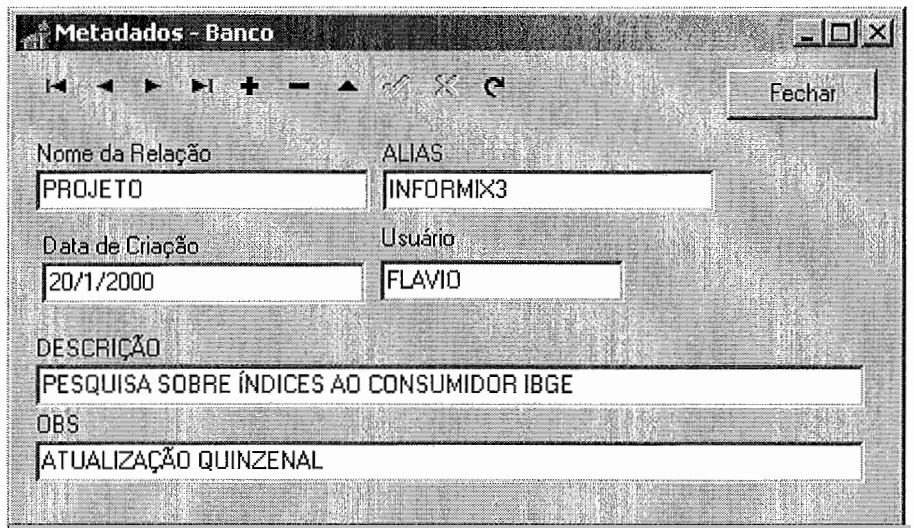


Fig. 28 - Metadados – Cadastramento de Banco

Metadados – Cadastramento de Estrutura

Através do módulo **DBMINER** – Metadados Cadastra Estrutura, o usuário poderá cadastrar dados com relação ao nome da relação, nome do atributo, tipo do atributo, tamanho do atributo, sua relação com o nome da sua relação com a chave estrangeira, além de sua participação na chave primária ou estrangeira. A navegação pelos registros foi simplificada através do uso de controles do tipo VCR, de conhecimento bastante comum. Para se efetivar uma alteração nos dados, basta que o usuário mova um dos controles ou use o último símbolo do navegador para que os dados se atualizem. Para apagar algum registro, basta clicar sobre o sinal de menos ("-") no navegador e o registro será excluído. O sinal de mais ("+"), adicionará um novo registro pronto para ser preenchido.

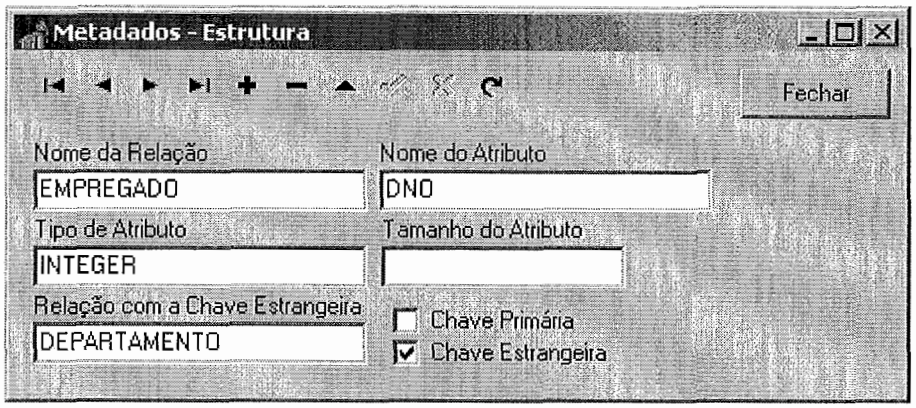


Fig. 29 - Metadados – Cadastra Estrutura

Parte Técnica do Banco

O usuário poderá verificar o dados técnicos de cada banco de dados cadastrado no sistema. Através do **DBMINER** – Metadados Alias, as informações do nome de servidor, o nome do usuário, o tipo de abertura do banco e demais características técnicas a respeito do banco estarão disponíveis aos usuários que delas precisarem. Na figura abaixo, à esquerda, temos os Alias associados aos bancos e à direita as informações a respeito de sua conexão, bem como suas características.

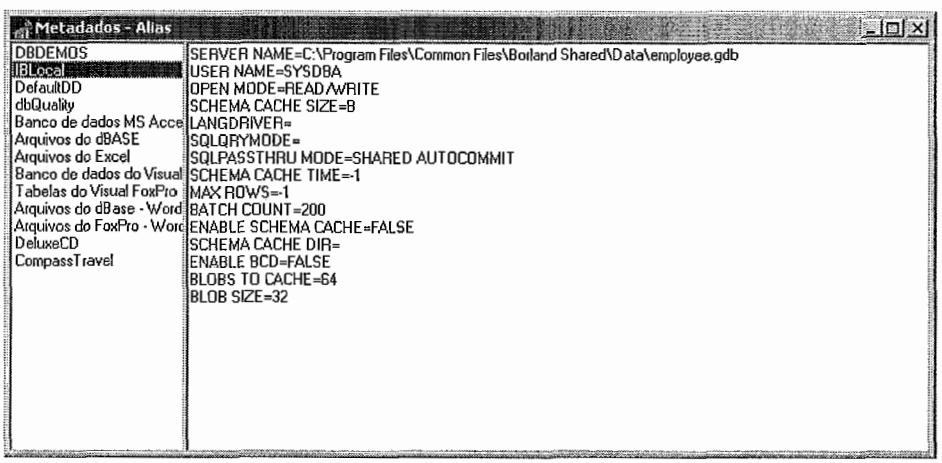


Fig. 30 - Metadados – Pseudônimos (*Alias*)

Tutores Técnicos

Abaixo traremos os módulos referentes aos Tutores que vêm auxiliar o usuário não especialista que deseja descobrir qual técnica utilizar para resolver o seu problema. Este

assistente permite ao administrador cadastrar perguntas e respostas em base de dados de forma ilimitada, oferecendo ao usuário um caminho de perguntas e respostas. O usuário ao iniciar o **DBMINER** – Tutor Métodos, no menu **Wizard** irá carregar a janela abaixo que irá lhe fazer questionamentos, a princípio genéricos e posteriormente mais específicos. Aparecendo uma pergunta para o usuário, um capo de explicações a respeito de seu objetivo será mostrado. O usuário só poderá selecionar com um clique do mouse uma caixa de verificação, como na figura abaixo.

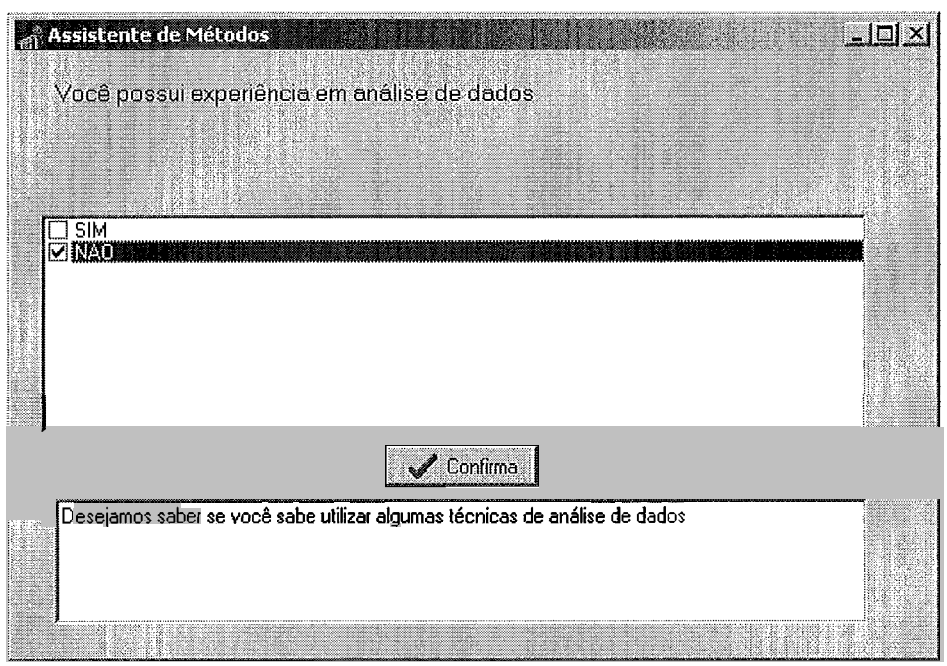


Fig. 31 - Tutor de Métodos

Ao selecionar continuar, outra pergunta irá aparecer até a solução técnica fornecida pelo especialista, conforme abaixo.

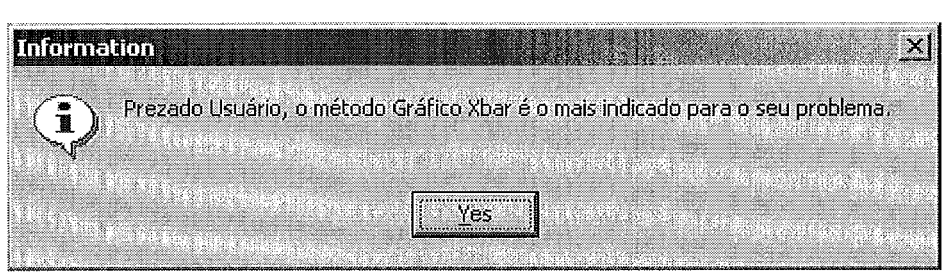


Fig. 32 – Tutor de Métodos – Resposta Técnica

Análise da Qualidade dos Dados

Através do menu apresentado na figura abaixo, obteremos o acesso às técnicas de Qualidade de Dados e Visualização Gráfica de Dados.

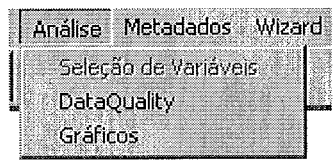


Fig. 32 - Menu - Análise

O Menu **Data Quality** apresentará as técnicas cujas referências técnicas se encontram no escopo desta tese, trazendo a implementação das 6 (seis) principais técnicas de controle estatístico e qualidade e processos associadas a bancos de dados.

Fig. 33 - Análise – Qualidade de Dados

Na figura acima o usuário poderá clicar em uma das técnicas e posteriormente selecionar – Abrir Dados, a fim de escolher a base de dados de sua pesquisa. Após esta seleção, a seguinte tela se apresentará:

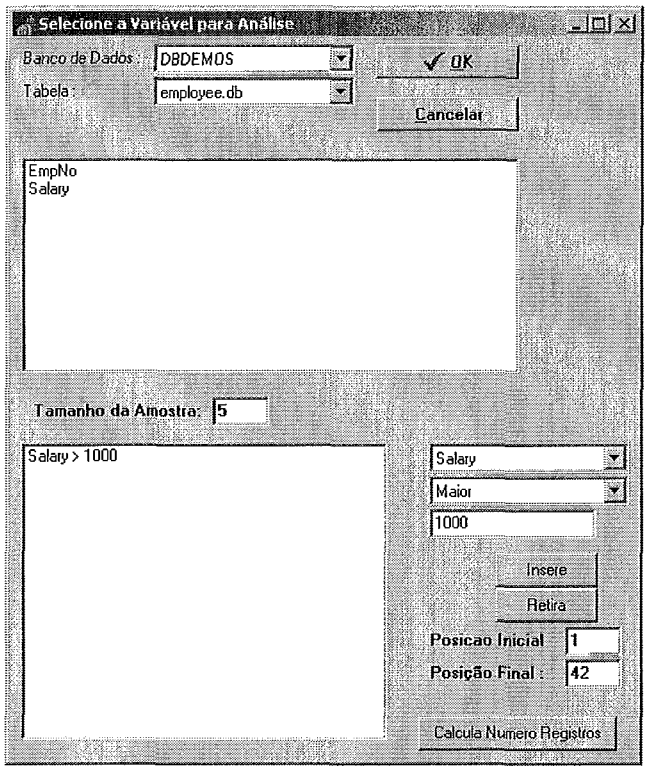


Fig. 34 - Qualidade de Dados – Seleção de Dados

Na caixa Banco de Dados, o usuário deverá escolher o Alias do banco de dados que se refere, na caixa Tabela, e selecionar a tabela que deseja investigar. Logo após a seleção das tabelas, o sistema irá informar somente os campos numéricos. Este é um controle interno do sistema, visto que nem sempre o usuário irá passar pelos módulos de exploração de dados. O tamanho da amostra deverá ser especificado, seu padrão é 2. O **DBMINER** – Análise – Qualidade de Dados permite ao usuário realizar filtros momentâneos em suas bases, a fim de dinamizar cortes e visões de dados. Para isto, basta que ele especifique a variável de corte, seu padrão: Igual, Maior, Menor e Diferente e o valor do ponto de corte. Além disso, o **DBMINER** – Análise – Qualidade de Dados (Análise - *Data Quality*) - permite ainda que o usuário filtre o número de registros em operação, facilitando um corte temporal nas investigações de dados. Para que a próxima fase da análise se proceda é necessário que o usuário clique no botão "Calcula Número de Registros" para que a aplicação capture todas as informações a respeito destes.

Métodos

X-Chart, R-Chart, P-Chart, NP-Chart, C-Chart e U-Chart

Escolheremos para demonstração o método X-Chart - o primeiro dos 6 métodos disponíveis de investigação estatística de dados. Neste manual não apresentaremos sua justificativa técnica, vide tese.

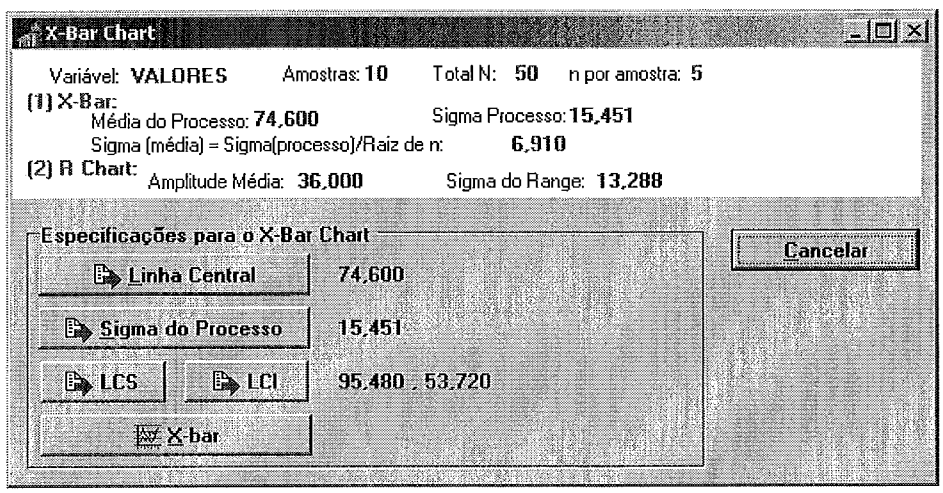


Fig. 35 - Estatísticas

As variáveis apresentadas aqui estão presentes nos demais gráficos de controle.

Obtemos através dele as seguintes informações:

i. Variável
ii. Amostras
iii. Total N
iv. N por Amostras
v. Média do Processo
vi. Sigma do Processo
vii. Sigma da média
viii. Amplitude (Range) Média
ix. Sigma da Amplitude (Range)

Tutor Estatístico

Ao clicar o botão X-bar no método anterior, o usuário irá invocar o Tutor Estatístico, um assistente que investigará os possíveis desvios dos dados apresentados. Este assistente irá pesquisar problemas de variações nas bases de dados e informará ao usuário, graficamente onde ocorreram os problemas como na forma e tabela, onde a primeira coluna irá apresentar o número da amostra pesquisada, a Média de cada uma das amostras e o LSC e LIC, bem como o seu parecer sobre a amostra de dados.

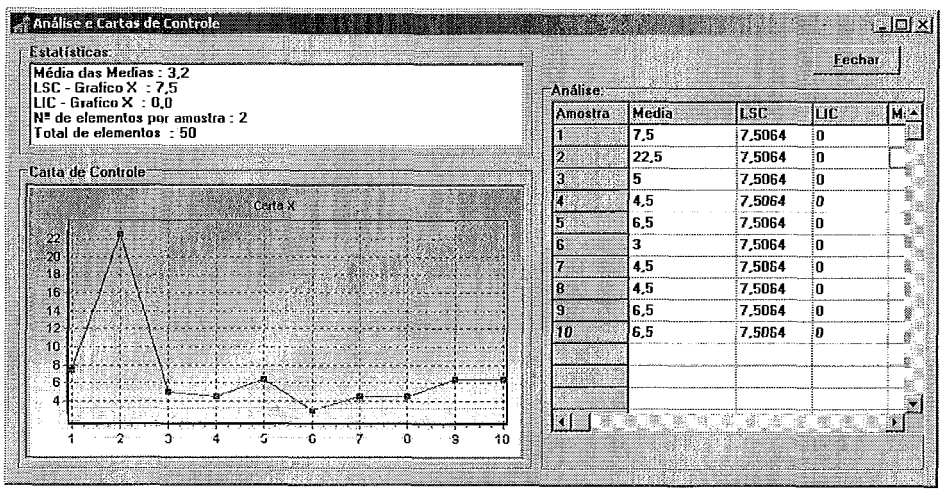


Fig. 36 – Assistente Estatístico

Caso a última coluna apareça com uma marcação em uma das amostras o usuário deverá proferir uma investigação minuciosa na amostra problemática. Na figura abaixo, mostraremos um exemplo desta situação.

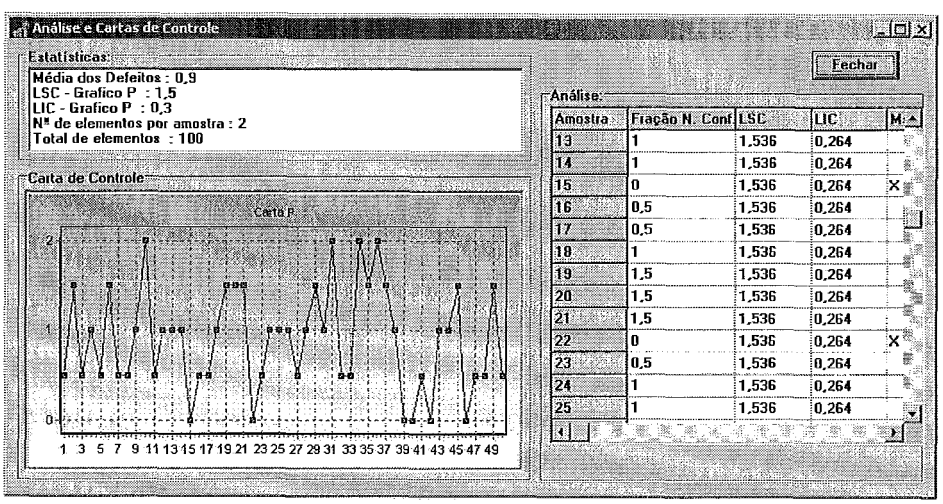


Fig. 37 – Avisos – Assistente Estatístico

Visualizador Gráfico de Dados

O **DBMINER** – Análise Gráfica de Dados é um módulo bastante funcional que permite ao usuário completo acesso visual aos dados. Através dele o usuário poderá plotar seus dados e correlacioná-los com outras variáveis, além da aplicação de filtros de corte, visualização instantânea e opções de impressão e titulação do gráfico.

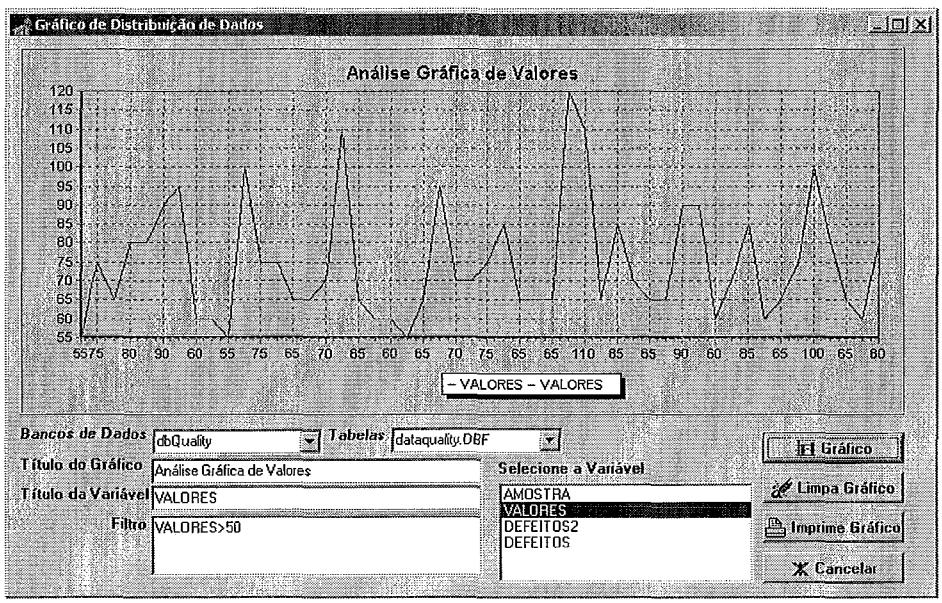


Fig. 38 – Análise Gráfica

No exemplo acima, selecionamos um alias da base de dados – dbQuality e a tabela dataquality.dbf. Fornecemos um título ao gráfico:" Análise Gráfica de Valores". O título da variável é automático, podendo ser atualizável; por fim, aplicamos um filtro para serem plotados somente os valores maiores que 50.

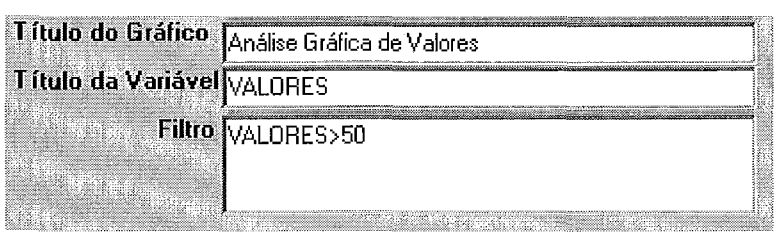


Fig. 39 – Filtros

Desta forma, o usuário pode contar com uma ferramenta bastante intuitiva e objetiva, de fácil manuseio, cuja manutenção é mínima. Acreditamos assim ter conseguido uma aplicação bem próxima do escopo desta tese.

Capítulo 6 - Conclusão

Esta tese pretende contribuir, fornecendo uma ferramenta, mesmo que embrionária, que possa facilitar o acesso à informação, orientar a análise da qualidade dos dados e suas críticas de consistência, possibilitando ao não-especialista ensaiar algumas análises de qualidade dos dados, que até então, mesmo sendo básicas, fugiam à compreensão dos que realizam estes estudos.

Através da exploração estatística dos dados, criamos um substrato capaz de apontar possíveis erros, e mais, mostrar-nos a possibilidade da investigação e aprendizado contínuo do fenômeno estudado. Demonstramos que a arquitetura proposta possibilita a junção de novos módulos operacionais, garantindo a expansão das técnicas de controle da qualidade dos dados, facilitando sua adequação a diferentes problemas. Portanto, dominando as diversas técnicas de análise de dados, podemos começar a tecer maiores comentários a respeito deste ou de outro fenômeno.

Nossa tese propõe uma metodologia capaz de ser implementada e expandida baseada no uso das técnicas de CEP (Controle Estatístico de Processos), visto tratarem de forma genérica e robusta os dados de quaisquer contextos.

Apesar de muitos trabalhos terem sido publicados recentemente, notamos a ausência de implementações no segmento das métricas da Qualidade de Dados, onde nossa arquitetura proposta se destaca, preenchendo esta carência. Alguns autores indicam o uso dos algoritmos de CEP, sem orientar a sua aplicação. Sentimos também a ausência de ferramentas capazes de traduzir de forma simples e direta o verdadeiro objetivo da Qualidade de Dados – verificar a qualidade de dados – para usuários não especialistas.

Assim, preenchemos uma lacuna importante oferecendo ao leitor:

- Uma visão de casos reais onde os conceitos de Qualidade de Dados são de extrema importância e que nos permitiram levantar requisitos para nossa arquitetura e implementação.
- Uma revisão do Controle Estatístico de Qualidade, permitindo uma maior compreensão das técnicas sugeridas na literatura

- Uma arquitetura com ferramentas de Controle da Qualidade de Dados, suportando diferentes técnicas de análise, sugeridas na literatura, como a exploração visual, a exploração estatística e o uso de regras de negócios.
- Uma implementação dessa arquitetura com foco nas técnicas de Controle Estatístico de Qualidade
- O uso de tutores que permitam ao usuário não especialista utilizar a ferramenta de forma adequada.

O sistema desenvolvido atendeu às expectativas propostas, apresentando como resultados: flexibilidade, clareza de interface, facilidade no aprendizado e grande capacidade de expansão técnica. Os exemplos apresentados mostram sua utilidade dentro das áreas de análise e exploração do conhecimento, demonstrando que o não-especialista também, pode analisar a qualidade do dado, se houver uma ferramenta que o oriente.

A contribuição dada pela arquitetura desenvolvida e apresentada nesta tese contempla as seguintes características anteriormente propostas:

i. Facilidade de aprendizado/operação. Essa característica é garantida pelo uso dos tutores que apresentam as técnicas e os passos à serem dados pelo usuário do sistema.

ii. Clareza na interface homem/máquina. Essa característica é garantida através da camada interface da arquitetura, que é modularizada e flexível, permitindo navegação fácil por parte do usuário, como demonstramos no protótipo construído.

iii. Análise de Metadados. Essa característica é garantida pelo processo de armazenamento e busca de informações a respeito dos bancos de dados, suas relações e regras de negócios armazenadas.

iv. Tutores amigáveis para a exploração de dados. Essa característica é garantida pelo uso de tutores especializados, capazes de orientar o usuário na busca da melhoria da qualidade dos seus dados.

v. Técnicas compreensíveis de análise exploratória de dados. Essa característica é garantida pelo uso das técnicas de CEP e análise exploratória de dados apresentadas no capítulo 3 deste trabalho.

vi. Análise fortemente baseada em gráficos. A arquitetura apresentada é fortemente baseada em ambiente gráfico, gerando assim, um facilitador para as operações dos usuários.

vii. Análise de regras de negócios. Essa característica é garantida através da armazenagem das regras de negócios em bases de dados próprias, podendo ser acessadas a qualquer momento pelos usuários do sistema como visto no capítulo 4.

viii. Versatilidade no acesso as bases de dados. Essa característica é garantida pelo uso de uma camada de acesso que controla os acessos aos SGBDRS de forma transparente ao usuário.

ix. Tratamento de erros. Essa característica é garantida pelo processo de tratamento de erros não só na camada de acesso aos SGBDRS como também nas demais camadas da Arquitetura.

x. Visualização gráfica dos dados de forma rápida. Esta característica é garantida através da inclusão de módulos de análise gráfica capazes de plotar dados univariados e multivariados.

Como tema para futuros trabalhos, sugerimos contribuições aos módulos de regras de negócio, análise exploratória e ferramentas de análise. Desta forma, poderíamos compor uma ferramenta mais abrangente, a fim de dar maior suporte o usuário leigo.

Além disso, seria muito interessante obter resultados de campo na análise de qualidade de dados de sistemas em produção.

Anexos

Anexo A

TAB_QUESTOES – Exemplo da estrutura das questões do Tutor de Métodos

RAMO	ORIGEM	DESCRICAO	ORIENTACOES
1	1	Você possui experiência em análise de dados	Desejamos saber se você sabe utilizar algumas técnicas de análise de dados
2	1	SIM	Neste caso, apresentaremos as técnicas enumeradas e você escolherá a que melhor resolver o seu tipo de problema
3	1	NÃO	Neste caso, iremos apresentar alguns conceitos elementares para que você possa potencializar sua análise.
4	3	Você sabe o que são variáveis?	Este conceito deve ser bem entendido para que você possa associar as técnicas ao tipo de variável apresentado
5	4	SIM	Próximo conceito
6	4	NÃO	Variáveis são coisas que medimos, controlamos ou manipulamos em uma pesquisa. Elas diferem em vários aspectos → o papel dela em nossa pesquisa e o tipo de medida
7	6	Você conhece as escalas de medidas de variáveis?	Desejamos saber se você sabe reconhecer a escala de dados
8	7	SIM	Próximo conceito
9	7	NÃO	Uma característica pode ser pesquisada e medida utilizando

			escalas nominais, ordinais, intervalares e numéricas. **Nominais** Variáveis nominais são aplicadas em classificações qualitativas. Isto é, elas podem ser medidas somente se os itens individuais pertencerem a alguma característica distintivamente diferente, mas não pode quantificar ou até ordenar estas categorias. Escalas nominais são algumas vezes chamadas de escalas categóricas ou dados categóricos. Ex.: Qual o sexo do funcionário? Masculino - 1 Feminino - 2 Descreva o tipo de câncer de pulmão Célula pequena 1 Célula grande 2 Célula em forma de aveia 3 Ordinais Se existir uma ordem inerente a entre as categorias, os dados parecem ter sido obtidos de uma escala ordinal. Qual o seu nível de escolaridade? 1 - Nenhum

			2 - 1° grau incompleto 3 - 1° grau completo 4 - 2° grau incompleto 5 - 2° grau completo 6 - superior incompleto 7 - superior incompleto 8 - pós-graduação **Intervalares** As variáveis intervalares nos permitem não só ordenarmos os itens medidos, mas também quantificar e comparar os tamanhos das diferenças entre eles. Por exemplo: Temperatura → medida em graus Celcius ou Fahrenheit , constitui uma escala de intervalos. Podemos dizer que a temperatura de 40° C é maior que a de 30° C. **Numéricas** As variáveis numéricas são similares às variáveis intervalares, só que possuem um zero absoluto identificável. Os dados numéricos podem ser contínuos → altura, peso, idade ou discretos → número de visitas ao hospital, número de falhas.
10	2	Que tipo de dados você irá	Dependendo do tipo de dado, a

		analisar?	família de técnicas muda.
11	10	Qualitativo	Algumas características não podem ser medidas da mesma forma que altura, peso, idade. Muitas características só podem ser categorizadas. Nestes casos, chamamos de dados qualitativos e utilizamos as técnicas de contagem para investigá-los. Ex: contar o número de pacientes admitidos no hospital durante um dia, com base em cada diagnóstico apresentado.
12	10	Quantitativo	Este tipo de variável é o mais comum. Podemos obter medidas das alturas dos adultos do sexo masculino, o peso de crianças do pré-escolar. Estes são exemplos de variáveis quantitativas.

Bibliografia

ABNT., 2000, *NBR ISO 9000 versão 2000* [online]. Disponível: http://www.abnt.org.br. Acesso em 12 de janeiro de 2002.

BALLOU, D.P, PAZER, H.L., 1985, *Modeling data and process quality in multi-input, multi-output information systems*. Manage Sci, 31,2 , pp. 150-162.

BALLOU, D.P., TAYI, K.G., 1989, *Methodology for Allocating Resources for Data Quality Enhancement* . Commun ACM 32, 3, pp. 320-329.

BALLOU, D.P., WANG, R.Y., PAZER, H., TAYI, K.G, 1996, *Modeling information manufacturing systems to determine information product quality*. Manage Sci

BAOHUA GU, BING LIU, FEIFANT HU, HUAN LIU., 2001, *Efficiently Determine the Starting Sample Size for Progressive Sampling*. Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD

COCHINWALA, M. 2000. *Data Quality and Reconciliation*. Affiliates Workshop on Data Quality: Challenges for Computer Science and Statistics. National Institute of Statistical Sciences

DATE, C.J., 2000, *What Not How – The Business Rules Approach to Application Development* – Addison-Wesley

DEMING, E. W., 1986, *Out of Crisis*, MIT Center for Advanced Engineering Study, Cambridge, Mass

DING, Q. PERIZO, W. DING, Q. , ROY, A., 2001, *On Mining Satellite and Other Remotely Sensed Images*. Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD

DOMINGOS, P., HULTEN, G. 2001, *Catching Up with the Data: Research Issues in Mining Data Streams*. Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD

DVIR, R., EVANS, S. 1996, A TQM Approach to the Improvement of Information Quality. MIT Center for Advanced Engineering Study, Cambridge, Mass

ENGLISH, L. P., 1999, *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits.*. John Wiley & sons, New York, EUA

ENGLISH, L. P., 2000a, *Information Quality Processes and Technologies: Information Quality in Practice.* Affiliates Workshop on Data Quality: Challenges for Computer Science and Statistics. National Institute of Statistical Sciences

ENGLISH, L. P., 2000b, *Record Linkage Methods.* Affiliates Workshop on Data Quality: Challenges for Computer Science and Statistics. National Institute of Statistical Sciences

ENGLISH, L. P., 2000c, *Seven Deadly Misconceptions about Information Quality.* Affiliates Workshop on Data Quality: Challenges for Computer Science and Statistics. National Institute of Statistical Sciences

FERREIRA, Aurélio B, de Hollanda., 1999, *Aurélio: Século XXI :* O *dicionário da língua portuguesa.* 3 ed. Rio de Janeiro: Nova Fronteira,.

HIPP, J. GÜNTZER, U., GRIMMER, U., 2001a, *Integrating Association Rule Mining Algorithms with Relational Database Systems.* ICEICS, Setúbal, Portugal

HIPP, J. GUNTZER, U., GRIMMER. , U. 2001b, *Data Quality Mining - Making a Virtue of Necessity.* Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD

JIAN PEI, ANTHONY K.H. TUNG, E JIAWEI HAN. *Fault-Tolerant Frequent Pattern Mining: Problems and Challenges.* Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD 2001

JUDSON, D. H., 2000, *The Statistical Administrative Records System: System Design, Successes and Challenges.* Affiliates Workshop on Data Quality: Challenges for Computer Science and Statistics. National Institute of Statistical Sciences

JURAN, J. M., 1991. *Juran, Planejando para a Qualidade.* 2 ed. São Paulo, MacGraw-Hill

LAUDON, K.C., 1986, *Data Quality and due processs in large interorganizacional record systems.* Commun, ACM 29, 1, pp. 4-11

LIEPINS, G. E., UPPULURI, V. R. R, EDS, 1990, *Data Quality Control: Theory and Pragmatics.* Nova York, N.Y., Marcel Dekker,

MITCHEL T. M., 1997, *Machine learning*, McGraw Hill

MONTGOMERY, D.C. 1991. *Introduction to Statistical Quality Control.* 2 ed. New York: John Wiley and Sons

ORR, KEN. , 1998, *Data Quality and Systems.* 1998, Vol.41, No. 2. ACM

PADHRAIC SMYTH., 2001, *Breaking Out of the Black-Box: Research Challenges in Data Mining.* Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD

REDMAN, T. C., 1998, *The Impact of Poor Data Quality on the Typical Enterprise.* Vol.41, No.2 ACM

SOFTWARE, S., 1999, *VISION:Solutions for Data Warehousing* DMReview

SOFTWARE, T. , 2001, *Microsoft Data Quality Case Study*, A division of Harte-Hanks, Inc. CONDEX, EUA

SOFTWARE, T., 1999, *Achieving Enterprise Data Quality* DMReview

STAIR R. M, 1998, *Princípios de Sistemas de Informação – uma abordagem gerencial*, São Paulo, LTC Editora

TECHNOLOGY, V., 1999, *Five Common Excuses For Not Re-engineering Legacy Data.* DMReview

TECHNOLOGY, V., 1999, *The Five Legacy Data Contaminants You Will Encounter in Your Warehouse Migration* DMReview

THORNTON, ANN, D., 2000, *Challenges in Improving Information Quality.* Affiliates Workshop on Data Quality: Challenges for Computer Science and Statistics. National Institute of Statistical Sciences

VASSILIOU, Y., 2000, *Developing Data Warehouses with Quality in Mind.* Affiliates Workshop on Data Quality: Challenges for Computer Science and Statistics. National Institute of Statistical Sciences

VERYKIOS, VASSILIS, 2000, *A Decision Model for Cost-Optimal Record Matching* Affiliates Workshop on Data Quality: Challenges for Computer Science and Statistics. National Institute of Statistical Sciences

WAND, Y.,WANG, R. Y., 1996, *Anchoring Data Quality Dimensions in Ontological Foundations.* Commun, ACM, Vol 39, No. 11, pp-86-95.

WANG R.Y., KON, H. B., 1993, *Towards Total Data Quality Management (TDQM). Information Technology in Action: Trends and Perspectives.* Prentice Hall, Englewood Cliffs, NJ.

WANG, R. Y., STOREY, V. C. , FIRTH. C. P., 1995, *A Framework for Analysis of Data Quality Research.* IEEE Trans. Know, Data Eng, pp. 623-640.

WANG, R. Y., STRONG, D. M., 1996, *Beyond accuracy; What Data Quality Means to Consumers.* J. Manage. Info Syst., pp. 5-34.

WANG, R., 2000, *Raising the Bar for Data Quality in the New Millenium.* Affiliates Workshop on Data Quality: Challenges for Computer Science and Statistics. National Institute of Statistical Sciences

WEBER, R. EDP, 1998, *Auditing: Conceptual Foundations and Practices*. G. B. Davis, McGraw-Hill, New York, N.Y

WENZ, D., 1999, *Implementing an AS/400 Data Warehouse* DMReview

WILKS, Allan R., 2000, Data Quality for Large Transaction Streams. Affiliates Workshop on Data Quality: Challenges for Computer Science and Statistics

WINKLER, E. WILLIAN, 2001, *Quality of Very Large Databases.. U.S. Census Bureau, Statistical Research Division. EUA. Website*

WINKLER, W. E., 2000, *Machine Learning, Information Retrieval and Record Linkage* Affiliates Workshop on Data Quality: Challenges for Computer Science and Statistics. National Institute of Statistical Sciences

WIZRULE:, 1999, *A New Approach to Data Cleansing* DMReview

XIONG WANG, 2001, *Mining Protein Surfaces.* Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD